Geometry & Topology

msp

GEOMETRY & TOPOLOGY

msp.org/gt

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

Duality between Lagrangian and Legendrian invariants

TOBIAS EKHOLM

YANKI LEKILI

Consider a pair (X, L) of a Weinstein manifold X with an exact Lagrangian submanifold L, with ideal contact boundary (Y, Λ), where Y is a contact manifold and $\Lambda \subset Y$ is a Legendrian submanifold. We introduce the Chekanov–Eliashberg DG–algebra, $\mathrm{CE}^*(\Lambda)$, with coefficients in chains of the based loop space of Λ, and study its relation to the Floer cohomology $\mathrm{CF}^*(L)$ of L. Using the augmentation induced by L, $\mathrm{CE}^*(\Lambda)$ can be expressed as the Adams cobar construction Ω applied to a Legendrian coalgebra, $\mathrm{LC}_*(\Lambda)$. We define a twisting cochain $\mathfrak{t} \colon \mathrm{LC}_*(\Lambda) \to \mathrm{B}(\mathrm{CF}^*(L))^{\#}$ via holomorphic curve counts, where B denotes the bar construction and # the graded linear dual. We show under simple-connectedness assumptions that the corresponding Koszul complex is acyclic, which then implies that $\mathrm{CE}^*(\Lambda)$ and $\mathrm{CF}^*(L)$ are Koszul dual. In particular, $\mathfrak{t}$ induces a quasi-isomorphism between $\mathrm{CE}^*(\Lambda)$ and $\Omega \mathrm{CF}_*(L)$, the cobar of the Floer homology of L.

This generalizes the classical Koszul duality result between $C^*(L)$ and $C_{-*}(\Omega L)$ for L a simply connected manifold, where ΩL is the based loop space of L, and provides the geometric ingredient explaining the computations given by Etgü and Lekili (2017) in the case when X is a plumbing of cotangent bundles of 2–spheres (where an additional weight grading ensured Koszulity of $\mathfrak{t}$).

We use the duality result to show that under certain connectivity and local-finiteness assumptions, $\mathrm{CE}^*(\Lambda)$ is quasi-isomorphic to $C_{-*}(\Omega L)$ for any Lagrangian filling L of Λ.

Our constructions have interpretations in terms of wrapped Floer cohomology after versions of Lagrangian handle attachments. In particular, we outline a proof that $\mathrm{CE}^*(\Lambda)$ is quasi-isomorphic to the wrapped Floer cohomology of a fiber disk C in the Weinstein domain obtained by attaching $T^*(\Lambda \times [0, \infty))$ to X along Λ (or, in the terminology of Sylvan (2019), the wrapped Floer cohomology of C in X with wrapping stopped by Λ). Along the way, we give a definition of wrapped Floer cohomology via holomorphic buildings that avoids the use of Hamiltonian perturbations, which might be of independent interest.

1 Introduction

In this introduction we first give an overview of our results. The overview starts with a review of well-known counterparts of our constructions in algebraic topology. We then introduce our Legendrian and Lagrangian invariants in Sections 1.1 and 1.2, respectively, and discuss the connection between them and applications thereof in Section 1.3. Among these the most central role is played by the *Chekanov–Eliashberg algebra with based loop space coefficients*, denoted as CE^*. As we show, any other invariant can be obtained from CE^* by algebraic manipulation. Finally, in Section 1.4 we give detailed calculations of the invariants introduced, in the simple yet illustrative example of the Legendrian Hopf link filled by two Lagrangian disks intersecting transversely in one point.

The starting point for our study is a construction in classical topology. Consider the following augmented DG–algebras over a field $\mathbb{K}$ associated to a based, connected, topological space (M, pt):

$$C^*(M) \to \mathbb{K}, \quad C_{-*}(\Omega M) \to \mathbb{K},$$

where $C^*(M)$ is the singular cochain complex equipped with the cup product and $C_{-*}(\Omega M)$ is the singular chain complex of the based (Moore) loop space of M equipped with the Pontryagin product. (We use cohomologically graded complexes throughout the paper so that all differentials increase the grading by 1.) In the case of singular cohomology, the inclusion $i: \mathrm{pt} \to M$ gives the augmentation $i^*: C^*(M) \to C^*(\mathrm{pt}) = \mathbb{K}$ and in the case of the based loop space, the augmentation is given by the trivial local system $\pi_1(M, \mathrm{pt}) \to \mathbb{K}$.

If M is of finite-type (for example, a finite CW–complex), then it is well known that one can recover the augmented DG–algebra $C^*(M)$ from the augmented DG–algebra $C_{-*}(\Omega M)$ by the Eilenberg–Moore equivalence

$$C^*(M) \simeq \mathrm{RHom}_{C_{-*}(\Omega M)}(\mathbb{K}, \mathbb{K}).$$

In the other direction, if M is *simply connected*, then the Adams construction gives a quasi-isomorphism

$$C_{-*}(\Omega M) \simeq \mathrm{RHom}_{C^*(M)}(\mathbb{K}, \mathbb{K}),$$

and in this case $C^*(M)$ and $C_{-*}(\Omega M)$ are said to be *Koszul dual* DG–algebras. Koszul duality is sometimes abbreviated and simply called *duality*. For more general M, using

the method of acyclic models, Brown [13] constructed a twisting cochain

$$\mathfrak{t}: C_{-*}(M) \to C_{-*}(\Omega M).$$

This is a degree 1 map that induces a DG–algebra map $\Omega C_{-*}(M) \to C_{-*}(\Omega M)$, where $\Omega C_{-*}(M)$ is the cobar construction applied to chains on M; see Section 2.2.1. By definition, $\mathfrak{t}$ is a quasi-isomorphism when duality holds, and this can be detected by an associated *Koszul complex*, which is acyclic if and only if duality holds. In the general case, $\Omega C_{-*}(M)$ is a certain completion of $C_{-*}(\Omega M)$ and consequently $C_{-*}(\Omega M)$ is a more refined invariant of M than $\Omega C_{-*}(M)$.

In this paper, we pursue this idea in the context of invariants associated to Lagrangian and Legendrian submanifolds. Here the role played by simple connectedness in the above discussion has two natural counterparts: one corresponds to a generalized notion of simple connectedness for intersecting Lagrangian submanifolds and the other is the usual notion of simple connectedness for Legendrian submanifolds.

We start with the geometric data of a Liouville domain X with convex boundary Y and an exact Lagrangian submanifold $L \subset X$ with Legendrian boundary $\Lambda \subset Y$. We assume that $c_1(X) = 0$, that the Maslov class of L vanishes (for grading purposes) and that L is relatively spin (to orient certain moduli spaces of holomorphic disks). Assume that L is subdivided into embedded components intersecting transversely $L = \bigcup_{v \in \Gamma} L_v$, and that Λ is subdivided into connected components $\Lambda = \bigsqcup_{v \in \Gamma} \Lambda_v$. To avoid notational complications, we take both parametrized by the same finite set Γ and assume that the boundary of L_v is Λ_v. We use a base field $\mathbb{K}$ and define the semisimple ring

$$\boldsymbol{k} = \bigoplus_{v \in \Gamma} \mathbb{K} e_v,$$

generated by mutually orthogonal idempotents e_v. Also, we fix a partition

$$\Gamma = \Gamma^+ \cup \Gamma^-$$

into two disjoint sets, and choose a basepoint $p_v \in \Lambda_v$ for each $v \in \Gamma^+$.

For simplicity, let us restrict, in this introduction, to the following situation:

- X is a subcritical Liouville domain.
- If $v \in \Gamma^-$, then the corresponding Legendrian Λ_v is an embedded *sphere*.

From a technical point of view, these restrictions are unnecessary. We make them in order to facilitate the explanation of our constructions from the perspective of Legendrian surgery. (Note that the topology of Λ_v is unrestricted when $v \in \Gamma^+$.)

We write X_Λ for the completion of the Liouville sector obtained from X by attaching critical Weinstein handles along Λ_v for each $v \in \Gamma^-$, and attaching cotangent cones $T^*(\Lambda_v \times [0, \infty))$ along Λ_v for each $v \in \Gamma^+$. If $\Gamma^+ = \varnothing$, X_Λ is an ordinary Liouville manifold. In this case Gromov compactness is ensured by convexity of the boundary. When $\Gamma^+ \neq \varnothing$, we also have part of the boundary that can be identified with a neighborhood of the zero section in the cotangent bundle $\bigcup_{v \in \Gamma^+} T^*(\Lambda_v \times [T, \infty))$, for some $T > 0$. This is a geometrically bounded manifold, hence Gromov compactness [38] still holds, and holomorphic curve theory is well behaved.

In X_Λ, for $v \in \Gamma^-$ there is a closed exact Lagrangian submanifold $S_v = L_v \cup D_v$, the union of the Lagrangian L_v in X and the *Lagrangian core disk* D_v of the Weinstein handle attached to Λ_v, and for $v \in \Gamma^+$ there is a noncompact Lagrangian obtained by attaching the cylindrical boundary $\Lambda_v \times [0, \infty)$ to L_v for $v \in \Gamma^+$, which we will still denote by L_v, by abuse of notation, even when we view them now in X_Λ. Dually, for each $v \in \Gamma^-$ we obtain (noncompact) exact Lagrangian disks C_v, the *Lagrangian cocore disks* of the Weinstein handles attached to Λ_v on X, and for each $v \in \Gamma^+$ we construct *dual* Lagrangians disks C_v intersecting L_v once and asymptotic to a Legendrian meridian of L_v — these can be constructed as the cotangent fiber at the point (p_v, t), $t > 0$, in $T^*(\Lambda_v \times [0, \infty)) \subset X_\Lambda$, where p_v is the basepoint on Λ_v.

The invariants we will construct are associated to the unions of Lagrangian submanifolds

$$L_\Lambda := \bigcup_{v \in \Gamma^+} L_v \ \cup \ \bigcup_{v \in \Gamma^-} S_v \quad \text{and} \quad C_\Lambda := \bigcup_{v \in \Gamma} C_v.$$

The Lagrangian L_Λ will be referred to as a *Lagrangian skeleton* of X_Λ; it is a union of Lagrangian submanifolds which intersect transversely. The dual Lagrangian C_Λ is the union of Lagrangian disks which can be locally identified with cotangent fibers to irreducible components of L_Λ.

We will study two algebraic invariants associated to $(X_\Lambda, L_\Lambda, C_\Lambda)$. The first one is the *Legendrian A_∞–algebra*, LA^*. It corresponds to the endomorphism algebra of L_Λ considered in the infinitesimal Fukaya category of X_Λ (Theorem 63). The second one is the *Chekanov–Eliashberg DG–algebra*, CE^*. It corresponds to the endomorphism algebra of C_Λ considered in the partially wrapped Fukaya category of X_Λ (Theorem 83). However, we will take the pre-surgery perspective as in Bourgeois, Ekholm and Eliashberg [11] and construct all these invariants by studying Legendrian invariants of $\Lambda \subset X$ rather than Floer cohomology in X_Λ. From this perspective, the

case $\Gamma^+ \neq \varnothing$ is a new construction, which generalizes the theory from [11] in a way analogous to how the partially wrapped Fukaya categories of Sylvan [62] generalize the wrapped Fukaya categories of Abouzaid and Seidel [3].

The invariants LA^* and CE^* come equipped with canonical augmentations to the semisimple ring $\boldsymbol{k}$, and it is easy to see by construction that LA^* is determined by CE^* via the equivalence

$$LA^* \simeq \mathrm{RHom}_{CE^*}(\boldsymbol{k}, \boldsymbol{k}).$$

The duality which would recover CE^* from LA^* holds in the "simply connected" case; see Section 2.3. In the topological case discussed above, this is analogous to the simple-connectedness assumption on M, which makes the augmented algebras $C^*(M)$ and $C_{-*}(\Omega M)$ Koszul dual. In fact, the topological case is a special case of our study for the Weinstein manifold T^*M, with the Lagrangian skeleton $L_\Lambda = M$ given by the zero section, and the dual Lagrangian C_Λ given by a cotangent fiber T_p^*M. This is because the wrapped Floer cohomology complex of a cotangent fiber is quasi-isomorphic to $C_{-*}(\Omega M)$ by Abouzaid [2] and the Floer cohomology complex of the zero section is quasi-isomorphic to $C^*(M)$ (Fukaya and Oh [35]) as augmented A_∞–algebras.

We next sketch the definition of our version of the Chekanov–Eliashberg DG–algebra without any assumption of simple connectedness; see Section 3 for details. This is the DG–algebra over $\boldsymbol{k}$ called CE^* above. Its underlying $\boldsymbol{k}$–bimodule is the unital $\boldsymbol{k}$–algebra generated by Reeb chords between components of Λ and chains in $C_{-*}(\Omega_{p_v}\Lambda_v)$ for $v \in \Gamma^+$. (This is the crucial distinction between Γ^+ and Γ^-.)

We use the cubical chain complex (cf Serre [60]) $C_{-*}(\Omega_{p_v}\Lambda_v)$ for $v \in \Gamma^+$ — see Section 3.1 for a discussion of other possible choices of chain models — to express CE^* as a free algebra over $\boldsymbol{k}$ generated by Reeb chords c and generators of $C_{-*}(\Omega_{p_v}\Lambda_v)$ for $v \in \Gamma^+$. The differential on CE^* is determined by its action on generators. On a generator of $C_{-*}(\Omega_{p_v}\Lambda_v)$ we simply apply the usual differential. On a generator c_0 which is a Reeb chord, the differential is determined by moduli spaces of holomorphic disks in the symplectization $\mathbb{R} \times Y$ which asymptotically converge to c_0 on the positive end and chords $c_1, \ldots, c_i$ at the negative end as follows. Consider the moduli space of J–holomorphic maps $u \colon D \to \mathbb{R} \times Y$, where D is a disk with $k+1$ boundary punctures $z_j \in \partial D = S^1$ that are mutually distinct with $(z_0, z_1, \ldots z_k)$ respecting the counterclockwise cyclic order of S^1, and u sends the boundary component (z_{j-1}, z_j) of $S^1 \setminus \{z_0, \ldots, z_k\}$ to $\mathbb{R} \times \Lambda$ and is asymptotic to c_j near the puncture at z_j for $j = 1, \ldots, k$ and to c_0 near the puncture at z_0 (as usual these disks may be anchored
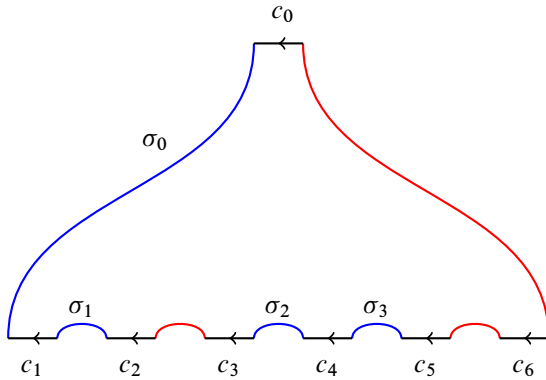
Figure 1: The differential in CE^*: the word $\sigma_0 c_1 \sigma_1 c_2 c_3 \sigma_2 c_4 \sigma_3 c_5 c_6$ appears in dc_0.

in X). The moduli space, which is naturally a stratified space with manifold strata that carries a fundamental chain, comes with evaluation maps to $\Omega_{p_v} \Lambda_v$ for $v \in \Gamma^+$. The image of the fundamental chain determines a word in our chain model of $C_{-*}(\Omega_{p_v} \Lambda_v)$. Reading these together with the Reeb chords in order gives the differential of c_0.

We remark that loop space coefficients have been used in the context of Lagrangian Floer cohomology before; see Barraud and Cornea [7] and Fukaya [34]. See also Abouzaid [2] and Cieliebak and Latschev [19] for uses of high-dimensional moduli spaces in Floer theory.

While CE^* with loop space coefficients is a powerful invariant, it is in general hard to compute as it involves high-dimensional moduli spaces of disks. As mentioned above, duality in the Legendrian Λ will also play a role. More precisely, we define another DG–algebra $CE^*_\parallel$ related to CE^* via a Morse-theoretic version of Adams cobar construction whose definition involves taking parallel copies of Λ but uses only 0–dimensional moduli spaces; see Section 3.4. In fact, we prove that the two DG–algebras are quasi-isomorphic when all Λ_v for $v \in \Gamma^+$ are simply connected.

Theorem 1 *There exists a DG–algebra map*

$$CE^* \to CE^*_\parallel,$$

which is a quasi-isomorphism when the Λ_v are simply connected for all $v \in \Gamma^+$.

Theorem 1 is restated and proved as Theorem 51 in Section 3.4.

1.1 Partially wrapped Fukaya categories by surgery

Let $\Lambda = \bigsqcup_{v \in \Gamma} \Lambda_v$ be a Legendrian submanifold and $\Gamma = \Gamma^+ \cup \Gamma^-$ be as above. Furthermore, we use the notation above for cocore disks and write $CE^* = CE^*(\Lambda)$. An important result that is implicit in [11, Remark 5.9] is the following:

Theorem 2 *Suppose* $\Gamma = \Gamma^-$. *Then there exists a surgery map defined via a holomorphic disk count that gives an* A_∞*–quasi-isomorphism between the wrapped Floer cochain complex* $CW^* := \bigoplus_{v,w \in \Gamma^-} CW^*(C_v, C_w)$ *and the Legendrian DG–algebra* CE^*.

We prove Theorem 2 in Section B.2 following [11], referring to Ekholm [25] for the necessary technical results omitted there. Section B.1 also contains a construction of wrapped Floer A_∞–algebras that uses only purely holomorphic disks (without Hamiltonian perturbation), and a proof that this agrees with the more standard version defined in Abouzaid and Seidel [3], which uses Hamiltonian perturbations.

One of the main guiding principles for the results in this paper is that Theorem 2 remains true when Γ^+ is nonempty, provided the Lagrangians C_v are considered as objects of the *partially wrapped* Fukaya category of X_Λ, where the noncapped Legendrians Λ_v for $v \in \Gamma^+$ serve as *stops*; cf Sylvan [62]. The full proof of this result when Γ^+ is nonempty can be reduced to the standard surgery result, Theorem 2, and will appear elsewhere. Here we give an outline of a somewhat different and more topological proof; see Section B.3. We will use the geometric intuition provided by this viewpoint, and our constructions of Legendrian invariants provide a rigorous "working definition" of CE^* even in the case that Γ^+ is nonempty, as well as a starting point for the study of "partially wrapped Fukaya categories" via Legendrian surgery (extending the scope of [11] considerably). For future reference, we state this result as a conjecture:

Conjecture 3 *There exists a surgery map defined via moduli spaces of holomorphic disks which gives an* A_∞*–quasi-isomorphism between the partially wrapped Floer cochain complex* $CW^* := \bigoplus_{v,w \in \Gamma} CW^*(C_v, C_w)$ *and the DG–algebra* CE^*.

While writing this paper, we learned that Sylvan [61] independently considered a similar conjecture in relation with his theory of partially wrapped Fukaya categories [62].

1.2 Augmentations and infinitesimal Fukaya categories

We keep the notation above and now consider an exact Lagrangian filling L in X of Λ. Such a filling gives an augmentation

$$\epsilon_L : \mathrm{CE}^* \to \mathbf{k}.$$

For chords on components Λ_v with $v \in \Gamma^-$, this is well known and given by a count of rigid disks with one positive puncture and boundary on L_v.

For components Λ_v with $v \in \Gamma^+$, we define an augmentation using the same disks. More formally, we define a chain map

$$\beta_L : \mathrm{CE}^* \to \bigoplus_{v \in \Gamma^+} C_{-*}(\Omega L_v),$$

which acts on chains in $\bigoplus_{v \in \Gamma^+} C_{-*}(\Omega \Lambda_v)$ by the inclusion and on Reeb chords c as the chain of loops carried by the moduli spaces of holomorphic disks with boundary on L_v (for each v) and a positive puncture at c. The augmentation ϵ_L is then this map followed by the augmentation on $\bigoplus_{v \in \Gamma^+} C_{-*}(\Omega L_v) \to \mathbf{k}$ that takes higher-dimensional chains to zero and takes any loop in L_v to e_v.

This allows us to write

$$\mathrm{CE}^* = \Omega\mathrm{LC}_*$$

for an A_∞–coalgebra $\mathrm{LC}_* = \mathrm{LC}_*(\Lambda)$ that we call the *Legendrian A_∞–coalgebra* (which depends on ϵ_L). Here Ω is the Adams cobar construction. Writing $\mathrm{LA}^* := (\mathrm{LC}_*)^\#$ for the A_∞–algebra which is the linear dual of LC_*, the following result recovers the Floer cochain complex of L in X_Λ:

Theorem 4 *There is an A_∞–quasi-isomorphism between* $\mathrm{CF}^* := \mathrm{CF}^*(L_\Lambda)$, *the Floer cochain complex in the infinitesimal Fukaya category of X_Λ, and the A_∞–algebra* LA^*.

By the general properties of bar–cobar constructions (see Section 2.2.1), the algebra $\mathrm{RHom}_{\Omega\mathrm{LC}_*}(\mathbf{k}, \mathbf{k})$ is quasi-isomorphic to the graded $\mathbf{k}$–dual of the bar construction on the algebra $\Omega\mathrm{LC}_*$, which can be computed as

$$(1) \qquad\qquad (B\Omega\mathrm{LC}_*)^\# \cong (\mathrm{LC}_*)^\# = \mathrm{LA}^*.$$

Remark 5 If Γ^+ is empty, the A_∞–algebra LA^* is obtained from the construction in Civan, Koprowski, Etnyre, Sabloff and Walker [20] and Bourgeois and Chantraine [10], known as the Aug_ category, by adding a copy of $\mathbf{k}$, making it unital.

If Γ^- is empty, the A_∞–algebra $LA^* \approx (BCE^*)^\#$ — see (1) — is the endomorphism algebra of Λ with the augmentation ϵ_L in the Aug_+ category of Ng, Rutherford, Shende, Sivek and Zaslow [55]. In the setting of microlocal sheaves, a related result was obtained by Nadler [53, Theorem 1.6].

1.3 Duality between partially wrapped and infinitesimal Fukaya categories

We study duality in the setting of the two categories described above: the partially wrapped Fukaya category and the infinitesimal Fukaya category of X_Λ (after surgery) or, equivalently, the augmented DG–algebra ΩLC_* and the augmented A_∞–algebra LA^* (before surgery).

As we have seen in Theorem 4, the augmented DG–algebra ΩLC_* determines the augmented (unital) A_∞–algebra CF^*. Now, a natural question is to what extent the quasi-isomorphism type of the A_∞–algebra CF^* determines the quasi-isomorphism type of the augmented Legendrian DG–algebra ΩLC_*.

We emphasize here the phrase "quasi-isomorphism type": even though it is possible to construct chain models of the A_∞–algebra LA^* (which is A_∞–quasi-isomorphic to CF^*) and the DG–algebra ΩLC_* by counting "the same" holomorphic disks interpreted in different ways, the two algebras are considered with respect to different equivalence relations, and the resulting equivalence classes can be very different. In particular, it is *not* generally true that $\mathfrak{f}\colon \mathscr{C} \to \mathscr{D}$ being a quasi-isomorphism of A_∞–coalgebras implies that $\Omega\mathfrak{f}\colon \Omega\mathscr{C} \to \Omega\mathscr{D}$ is a quasi-isomorphism.

We will study this question by (geometrically) constructing a *twisting cochain*

$$\mathfrak{t}\colon LC_* \to (BCF^*)^\#,$$

where B stands for the bar construction and $\#$ is the graded $\boldsymbol{k}$–dual. See Section 2.1.4. This twisting cochain induces a map of DG–algebras,

$$\Omega LC_* \to \mathrm{RHom}_{CF^*}(\boldsymbol{k}, \boldsymbol{k}),$$

which is a quasi-isomorphism if and only if $\mathfrak{t}$ is a *Koszul* twisting cochain. For example, we will prove the following result:

Theorem 6 *Suppose that LC_* is a locally finite, simply connected $\boldsymbol{k}$–bimodule. Then the natural map $\Omega LC_* \to \mathrm{RHom}_{CF^*}(\boldsymbol{k}, \boldsymbol{k})$ is a quasi-isomorphism.*

This is an instance of *Koszul duality* between the A_∞–algebras ΩLC_* and CF^*. It has many useful implications; for example, it implies an isomorphism between Hochschild cohomologies,

$$HH^*(\Omega LC_*, \Omega LC_*) \cong HH^*(CF^*, CF^*).$$

When $\Gamma^+ = \varnothing$, an isomorphism defined via a surgery map [11] was described between *symplectic cohomology*, $SH^* = SH^*(X_\Lambda)$, and the Hochschild cohomology $HH^*(\Omega LC_*, \Omega LC_*)$. Therefore, when duality holds (ie t induces an isomorphism), we obtain a more economical way of computing SH^*.

In the case of cotangent bundles T^*M of simply connected manifolds M, this recovers a classical result due to Jones [42], which gives

$$H_{n-*}(\mathcal{L}M) \cong HH^*(C_{-*}(\Omega M), C_{-*}(\Omega M)) \cong HH^*(CF^*(M), CF^*(M)),$$

where M is a simply connected manifold of dimension n and $\mathcal{L}M$ denotes the free loop space of M.

In Section 6, we give several concrete examples where the duality holds beyond the case of cotangent bundles. For example, the duality holds for plumbings of simply connected cotangent bundles according to an arbitrary plumbing tree; see Theorem 68.

In another direction, combining duality and Floer cohomology with local coefficients, we establish the following result for relatively spin exact Lagrangian fillings $L \subset X$ with vanishing Maslov class of a Legendrian submanifold $\Lambda \subset Y$.

Theorem 7 Let $\Gamma = \Gamma^-$ and assume that $SH^*(X) = 0$ and that Λ is simply connected. If $CE^*(\Lambda)$ is supported in degrees < 0, then L is simply connected. Moreover, if Λ is a sphere, then $CE^*(\Lambda)$ is isomorphic to $C_{-*}(\Omega \overline{L})$, where $\overline{L} = L \cup_\Lambda D$, for a disk D with boundary $\partial D = \Lambda$.

In general, duality between ΩLC_* and CF^* does not hold — as can be seen for example by looking at cotangent bundles of non-simply-connected manifolds, or letting Λ be the standard Legendrian trefoil knot in S^3 filled by a punctured torus. However, there are cases when duality holds even if LC_* is not simply connected, for instance because of the existence of an auxiliary weight grading (see Etgü and Lekili [32]), or, for an example in the 1–dimensional case, see Lekili and Polishchuk [48]. It is a very interesting open question to find a geometric characterization of when duality holds.

Remark 8 Constructions of Legendrian and Lagrangian holomorphic curve invariants require the use of perturbations to achieve transversely cut out moduli spaces. For our main invariant CE^*, all moduli spaces used can be shown to be transverse by classical techniques (see Theorem 74) except for the rigid holomorphic planes in X_Λ with a single positive end that are used to anchor the disks (in the terminology of [11]). These are also relevant for defining the wrapped Floer cochain complex CW^* without Hamiltonian perturbations and for constructing the surgery map. In all cases, there is a distinguished boundary puncture in the main disk that determines asymptotic markers on the split-off planes. Taking this marker into account removes symmetries of the planes, and a specific perturbation scheme for transversality of the resulting moduli spaces was constructed in [25]. We will use that perturbation scheme here; see Section A.2 for details.

1.4 An example: the Hopf link

In this section, we study the example of the Hopf link in order to illustrate our results in a simple and computable example. Some of the algebraic constructions used here are explained in detail only later; see Section 2.

Let $\Lambda \subset S^3$ be the standard Legendrian Hopf link. We work over $\boldsymbol{k} = \mathbb{K}e_1 \oplus \mathbb{K}e_2$ and with the Lagrangian filling L given by two disks in D^4 that intersect transversely in a single point. We choose the partition $\Lambda = \Lambda^+ \cup \Lambda^-$. This means that after attaching a Weinstein 2–handle to Λ^- and $T^*(S^1 \times [0, \infty))$ to Λ^+, we obtain the symplectic manifold X_Λ with Lagrangian skeleton

$$L_\Lambda = S^2 \cup T^*_{pt} S^2 \subset T^* S^2,$$

or, in the terminology of [62], X_Λ is $T^* S^2$ with wrapping stopped by a Legendrian fiber sphere. The DG–algebra $CE^* = CE^*(\Lambda)$ of Λ has coefficients in

$$C_{-*}(\Omega \Lambda^+) e_1 \oplus \mathbb{K}e_2 \cong \mathbb{K}[t, t^{-1}] e_1 \oplus \mathbb{K}e_2.$$

A free model for $\mathbb{K}[t, t^{-1}]$ is given by the tensor algebra $\mathbb{K}\langle s_1, t_1, k_1, l_1, u_1 \rangle$ with $|s_1| = |t_1| = 0$, $|k_1| = |l_1| = -1$, $|u_1| = -2$ and the differential

$$dk_1 = e_1 - s_1 t_1, \quad dl_1 = e_1 - t_1 s_1, \quad du_1 = k_1 s_1 - s_1 l_1.$$

The natural map $\mathbb{K}\langle s_1, t_1, k_1, l_1, u_1 \rangle \to \mathbb{K}[t, t^{-1}]$ sending $t_1 \to t$ and $s_1 \to t^{-1}$ is a quasi-isomorphism. The subscripts indicate that as $\boldsymbol{k}$–module generators, s_1, t_1, k_1, l_1 and u_1 are annihilated by the idempotent e_2.
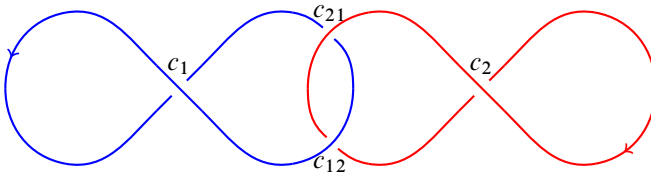
Figure 2: Hopf link when both Γ^+ and Γ^- are nonempty: the blue component lies in Γ^+ and the red in Γ^-.

Next, incorporating the Reeb chords, with notation as in Figure 2, we get the free algebra

$$\boldsymbol{k}\langle c_{12}, c_{21}, c_1, c_2, s_1, t_1, k_1, l_1, u_1 \rangle$$

with gradings

$$|u_1| = -2, \quad |c_1| = |c_2| = |k_1| = |l_1| = -1, \quad |c_{12}| = |c_{21}| = |s_1| = |t_1| = 0$$

and differential

$$dc_1 = e_1 + s_1 + c_{12}c_{21}, \quad dc_2 = c_{21}c_{12},$$
$$dk_1 = e_1 - s_1 t_1, \quad dl_1 = e_1 - t_1 s_1, \quad du_1 = k_1 s_1 - s_1 l_1.$$

The only augmentation to $\boldsymbol{k}$ is given by $\epsilon(s_1) = \epsilon(t_1) = -e_1$ and $\epsilon(c_1) = \epsilon(c_2) = \epsilon(c_{12}) = \epsilon(c_{21}) = \epsilon(k_1) = \epsilon(l_1) = \epsilon(u_1) = 0$. After change of variables, $s_1 \to s_1 - e_1$ and $t_1 \to t_1 - e_1$, we obtain the free algebra

$$\boldsymbol{k}\langle c_{12}, c_{21}, c_1, c_2, s_1, t_1, k_1, l_1, u_1 \rangle$$

with nonzero differential on generators

(2)
$$dc_1 = s_1 + c_{12}c_{21}, \quad dc_2 = c_{21}c_{12},$$
$$dk_1 = s_1 + t_1 - s_1 t_1, \quad dl_1 = s_1 + t_1 - t_1 s_1, \quad du_1 = l_1 - k_1 + k_1 s_1 - s_1 l_1.$$

On the other hand, we can compute the Floer cochains $\mathrm{CF}^* = \mathrm{CF}^*(L_\Lambda)$ of L_Λ as

$$\mathrm{CF}^* = \boldsymbol{k} \oplus \mathbb{K}a_{12} \oplus \mathbb{K}a_{21} \oplus \mathbb{K}a_2, \quad \text{where } |a_2| = 2, \ |a_{12}| = |a_{21}| = 1.$$

The cohomology level computation follows easily from the geometric picture and general properties of Floer cohomology: L_Λ is a union of a disk D^2 and a sphere S^2 that intersect transversely in one point, and we have

$$\mathrm{HF}^*(D^2, D^2) = \mathbb{K}e_1, \quad \mathrm{HF}^*(S^2, S^2) = \mathbb{K}e_2 \oplus \mathbb{K}a_2,$$
$$\mathrm{HF}^*(D^2, S^2) = \mathbb{K}a_{12}, \quad \mathrm{HF}^*(S^2, D^2) = \mathbb{K}a_{21}.$$

The only nontrivial product that does not involve idempotents is $\mathfrak{m}_2(a_{12}, a_{21}) = a_2$. For degree reasons, the only possible nontrivial higher products are

$$\mathfrak{m}_{2k}(a_{12}, a_{21}, \ldots, a_{12}, a_{21}) = ca_2 \quad \text{for some } k > 1 \text{ and } c \in \mathbb{K}.$$

It turns out that one can take $c = 0$ for all $k > 1$. Indeed, assuming that the $A_\infty-$ structure is strictly unital (which can be arranged up to quasi-isomorphism), consider the A_∞–relation that involves the term

$$\mathfrak{m}_2(\mathfrak{m}_{2k}(a_{12}, a_{21}, \ldots, a_{12}, a_{21}), e_2).$$

By induction on $k > 1$, this term has to vanish, implying $m_{2k}(a_{12}, a_{21}, \ldots, a_{12}, a_{21})$ has to vanish for all $k > 1$. Let us confirm this by using the quasi-isomorphism

$$\mathrm{CF}^* \cong \mathrm{RHom}_{\mathrm{CE}^*}(\boldsymbol{k}, \boldsymbol{k}).$$

We introduce the counital A_∞–coalgebra

$$\mathrm{LC}_* = \boldsymbol{k} \oplus \mathbb{K}c_{12} \oplus \mathbb{K}c_{21} \oplus \mathbb{K}c_1 \oplus \mathbb{K}c_2 \oplus \mathbb{K}s_1 \oplus \mathbb{K}t_1 \oplus \mathbb{K}k_1 \oplus \mathbb{K}l_1 \oplus \mathbb{K}u_1$$

with $|u_1| = -3$, $|c_1| = |c_2| = |k_1| = |l_1| = -2$ and $|c_{12}| = |c_{21}| = |s_1| = |t_1| = -1$, for which $\Delta_i = 0$ except for $i = 1$ or 2, where there are the nonzero terms

$$\Delta_1(c_1) = s_1, \quad \Delta_1(k_1) = s_1 + t_1, \quad \Delta_1(l_1) = s_1 + t_1, \quad \Delta_1(u_1) = l_1 - k_1.$$

Write $\Delta_2(x) = 1 \otimes_{\boldsymbol{k}} x + x \otimes_{\boldsymbol{k}} 1 + \overline{\Delta}_2(x)$. Then

$$\overline{\Delta}_2(c_1) = c_{12}c_{21}, \quad \overline{\Delta}_2(c_2) = c_{21}c_{12},$$

$$\overline{\Delta}_2(k_1) = -s_1t_1, \quad \overline{\Delta}_2(l_1) = -t_1s_1, \quad \overline{\Delta}_2(u_1) = k_1s_1 - s_1l_1,$$

where the A_∞ coalgebra operations on LC_* are defined so that $\Omega\mathrm{LC}_*$ is isomorphic to CE^*. Thus, $\mathrm{RHom}_{\mathrm{CE}^*}(\boldsymbol{k}, \boldsymbol{k})$ can be computed as the graded dual of LC_* which is the A_∞–algebra

$$\mathrm{LA}^* = \boldsymbol{k} \oplus \mathbb{K}c_{12}^\vee \oplus \mathbb{K}c_{21}^\vee \oplus \mathbb{K}c_1^\vee \oplus \mathbb{K}c_2^\vee \oplus \mathbb{K}s_1^\vee \oplus \mathbb{K}t_1^\vee \oplus \mathbb{K}k_1^\vee \oplus \mathbb{K}l_1^\vee \oplus \mathbb{K}u_1^\vee,$$

with gradings

$$|u_1^\vee| = 3, \quad |c_1^\vee| = |c_2^\vee| = |k_1^\vee| = |l_1^\vee| = 2, \quad |c_{12}^\vee| = |c_{21}^\vee| = |s_1^\vee| = |t_1^\vee| = 1,$$

where $c^\vee$ is the linear dual of the generator c of LC_*. The differential is

$$\mathfrak{m}_1(s_1^\vee) = c_1^\vee + k_1^\vee + l_1^\vee, \quad \mathfrak{m}_1(t_1^\vee) = k_1^\vee + l_1^\vee, \quad \mathfrak{m}_1(k_1^\vee) = -u_1, \quad \mathfrak{m}_1(l_1^\vee) = u_1,$$

and the products that do not involve idempotents are

$$\mathfrak{m}_2(c_{12}^\vee, c_{21}^\vee) = c_2^\vee, \quad \mathfrak{m}_2(c_{21}^\vee, c_{12}^\vee) = c_1^\vee, \quad \mathfrak{m}_2(t_1^\vee, s_1^\vee) = -k_1^\vee,$$
$$\mathfrak{m}_2(s_1^\vee, t_1^\vee) = -l_1^\vee, \quad \mathfrak{m}_2(k_1^\vee, s_1^\vee) = u_1^\vee, \quad \mathfrak{m}_2(s_1^\vee, l_1^\vee) = -u_1^\vee.$$

All the higher products vanish. We claim that this A_∞–algebra is quasi-isomorphic to the algebra

$$\boldsymbol{k} \oplus \mathbb{K}a_{12} \oplus \mathbb{K}a_{21} \oplus \mathbb{K}a_2, \quad \text{where } |a_2| = 2, \ |a_{12}| = |a_{21}| = 1,$$

with the only nontrivial product (not involving idempotents) given by

$$\mathfrak{m}_2(a_{12}, a_{21}) = a_2.$$

Indeed, it is easy to show that the map defined by

$$c_{12}^\vee \to a_{12}, \quad c_{21}^\vee \to a_{21}, \quad c_2^\vee \to a_2 \quad \text{and} \quad c_1^\vee, s_1^\vee, t_1^\vee, k_1^\vee, l_1^\vee, u_1^\vee \to 0$$

is a DG–algebra (hence also an A_∞–algebra) map, which induces an isomorphism at the level of cohomology.

Dually, we can construct a DG–algebra map

$$\mathrm{CE}^* \to \mathrm{RHom}_{\mathrm{CF}^*}(\boldsymbol{k}, \boldsymbol{k}).$$

The Floer cochain complex CF^* has a unique augmentation given by projection to $\boldsymbol{k}$, and we compute

$$\mathrm{RHom}_{\mathrm{CF}^*}(\boldsymbol{k}, \boldsymbol{k}) \cong \Omega\mathrm{CF}_*,$$

where CF_* is the coalgebra dual to CF^*. This is the free coalgebra

$$\boldsymbol{k}\langle a_{12}^\vee, a_{21}^\vee, a_2^\vee \rangle$$

with $|a_{12}^\vee| = |a_{21}^\vee| = 0$ and $|a_2^\vee| = -1$, and the only nontrivial differential not involving counits is

$$\Delta_2(a_2^\vee) = a_{21}^\vee a_{12}^\vee.$$

We have a twisting cochain

$$\mathrm{t} \colon \mathrm{LC}_* \to \Omega\mathrm{CF}_*$$

given by

$$\mathrm{t}(c_2) = a_2^\vee, \qquad\qquad \mathrm{t}(c_{12}) = a_{12}^\vee, \qquad \mathrm{t}(c_{21}) = a_{21}^\vee,$$
$$\mathrm{t}(c_1) = 0, \qquad\qquad \mathrm{t}(s_1) = -a_{12}^\vee a_{21}^\vee, \quad \mathrm{t}(t_1) = a_{12}^\vee a_{21}^\vee,$$
$$\mathrm{t}(l_1) = \mathrm{t}(k_1) = a_{12}^\vee a_2^\vee a_{21}^\vee, \qquad\qquad\qquad \mathrm{t}(u_1) = a_{12}^\vee a_2^\vee a_2^\vee a_{21}^\vee.$$

This means that $\mathfrak{t}$ satisfies the equations

$$dt(c_1) = t(s_1) + t(c_{12})t(c_{21}),$$
$$dt(c_2) = t(c_{21})t(c_{12}),$$
$$dt(k_1) = t(s_1) + t(t_1) - t(s_1)t(t_1),$$
$$dt(l_1) = t(s_1) + t(t_1) - t(t_1)t(s_1),$$
$$dt(u_1) = t(l_1) - t(k_1) + t(k_1)t(s_1) - t(s_1)t(l_1).$$

Hence, it induces a DG–algebra map

$$\Omega LC_* \to \Omega CF_*.$$

We have not checked whether this is a quasi-isomorphism, or equivalently whether $\mathfrak{t}$ is a *Koszul* twisting cochain. Note, however, that the DG–algebra map $\Omega CF_* \to \Omega LC_*$ defined by

$$a_2^\vee \to c_2, \quad a_{12}^\vee \to c_{12}, \quad a_{21}^\vee \to c_{21}$$

shows that $\mathfrak{t}$ is a retraction, and ΩCF_* is a retract of ΩLC_*.

2 Algebraic preliminaries

In this section, we review the homological algebra we use in our study of various invariants associated to Legendrian submanifolds and their Lagrangian fillings. Most of this material is well established; see [47] and also [45; 57; 56; 39; 49; 50]. Note though that our sign conventions follow [59]; see Remark 9.

2.1 A_∞–algebras and A_∞–coalgebras

In this section we will discuss the basic algebraic objects we use. These are modules over a ground ring k of the following form. Fix a coefficient field $\mathbb{K}$ (of arbitrary

characteristic) and let k be a semisimple ring of the form:

$$k = \bigoplus_{v \in \Gamma} \mathbb{K} e_v,$$

where $e_v^2 = e_v$ and $e_v e_w = 0$ for $v \neq w$, and where the index set Γ is finite.

We will use $\mathbb{Z}$–graded k–bimodules. If $M = \bigoplus_i M^i$ is such a module then we call M *connected* if $M^0 \cong k$ and either $M^i = 0$ for all $i > 0$, or $M^i = 0$ for all $i < 0$. We call M *simply connected* if, in addition, in the former case $M^{-1} = 0$, and in the latter $M^1 = 0$. Further, we say that M is *locally finite* if each M^i is finitely generated as a k–bimodule.

We have the usual shifting and tensor product operations on modules. If $M = \bigoplus_i M^i$ is a graded k–bimodule and s is an integer, then we let the corresponding shifted module $M[s] = \bigoplus_i M[s]^i$ be the module with graded components

$$M[s]^i = M^{i+s}.$$

If $N = \bigoplus_i N^i$ is another graded k–bimodule, then $M \otimes_k N = \bigoplus_k (M \otimes_k N)^k$ is naturally a graded k–bimodule with

$$(M \otimes_k N)^k = \bigoplus_{i+j=k} M^i \otimes_k N^j.$$

For iterated tensor products we write

$$M^{\otimes_k r} = \underbrace{M \otimes_k \cdots \otimes_k M}_{r}.$$

Our modules will often have further structure as $\mathbb{Z}$–graded A_∞–algebras and A_∞–coalgebras over k; see Sections 2.1.1 and 2.1.2. The modules are then in particular chain complexes with a differential, and we will use *cohomological* grading throughout; that is, the differential *increases* the grading by 1. For example, if L is a topological space then its cohomology complex $C^*(L)$ is supported in nonnegative grading, while the homology complex $C_{-*}(L)$ is supported in nonpositive degrees. To be consistent with this, we denote the grading as a subscript (resp. superscript) when the underlying chain complex has a coalgebra (resp. algebra) structure.

2.1.1 A_∞–**algebras** An A_∞–*algebra* over k is a $\mathbb{Z}$–graded k–module $\mathscr{A}$ with a collection of grading-preserving k–linear maps

$$\mathfrak{m}_i : \mathscr{A}^{\otimes_k i} \to \mathscr{A}[2-i]$$

for all integers $i \geq 1$ satisfying the A_∞–relations

$$(3) \quad \sum_{i,j} (-1)^{|a_1|+\cdots+|a_j|-j} \mathfrak{m}_{d-i+1}(a_d, \ldots, a_{j+i+1}, \mathfrak{m}_i(a_{j+i}, \ldots, a_{j+1}), a_j, \ldots, a_1) = 0$$

for all d.

Remark 9 We follow the sign conventions of [59]. Even though $\mathfrak{m}_i$ is written on the left of $(a_{j+i}, \ldots, a_{j+1})$, the sign convention is so that $\mathfrak{m}_i$ acts from the right. To be consistent, we will insist that all our operators act on the right independently of how they are written. This convention and the usual Koszul sign exchange rule applied with respect to the shifted grading $\mathscr{A}[1]$ determine the signs that appear in our formulas.

A *DG–algebra* over $\boldsymbol{k}$ is an A_∞–algebra $\mathscr{A}$ such that $\mathfrak{m}_i = 0$ for $i \geq 3$. In this case, we call the first two operations the *differential* and the *product*, respectively, and use the following adjustments to obtain an (ordinary) differential graded algebra:

$$(4) \qquad da = (-1)^{|a|} \mathfrak{m}_1(a) \quad \text{and} \quad a_2 a_1 = (-1)^{|a_1|} \mathfrak{m}_2(a_2, a_1).$$

In particular, the product is then associative and the graded Leibniz rule for d holds:

$$(5) \qquad d(a_2 a_1) = (da_2) a_1 + (-1)^{|a_2|} a_2 (da_1).$$

An A_∞–*map* $\mathfrak{e} : \mathscr{A} \to \mathscr{B}$ between A_∞–algebras $\mathscr{A}$ and $\mathscr{B}$ over $\boldsymbol{k}$, with operations $\mathfrak{m}_i$ and $\mathfrak{n}_i$ for $i \geq 1$, respectively, is a collection of $\boldsymbol{k}$–linear grading-preserving maps

$$\mathfrak{e}_i : \mathscr{A}^{\otimes_{\boldsymbol{k}} i} \to \mathscr{B}[1-i], \quad i \geq 1,$$

satisfying the relations

$$\sum_{i,j} (-1)^{|a_1|+\cdots+|a_j|-j} \mathfrak{e}_{d-i+1}(a_d, \ldots, a_{j+i+1}, \mathfrak{m}_i(a_{j+i}, \ldots, a_{j+1}), a_j, \ldots, a_1)$$

$$= \sum_{\substack{1 \leq j \leq d \\ 0 < i_1 < \cdots < i_j < d}} \mathfrak{n}_j(\mathfrak{e}_{d-i_j}(a_d, \ldots, a_{d-i_j}), \ldots, \mathfrak{e}_{i_2-i_1}(a_{i_2}, \ldots, a_{i_1+1}), \mathfrak{e}_{i_1}(a_{i_1}, \ldots, a_1)).$$

An A_∞–map $\mathfrak{e} : \mathscr{A} \to \mathscr{B}$ is called an A_∞–*quasi-isomorphism* if the map on cohomology $H^*(\mathscr{A}) \to H^*(\mathscr{B})$ induced by $\mathfrak{e}^1$ is an isomorphism.

We say that an A_∞–algebra $\mathscr{A}$ is *strictly unital* if there is an element $1_{\mathscr{A}} \in \mathscr{A}$ such that $\mathfrak{m}_1(1_{\mathscr{A}}) = 0$, $\mathfrak{m}_2(1_{\mathscr{A}}, a) = \mathfrak{m}_2(a, 1_{\mathscr{A}}) = a$ for any $a \in \mathscr{A}$, and $\mathfrak{m}_i$ for $i > 2$ annihilates any monomial containing $1_{\mathscr{A}}$ as a factor. Any A_∞–algebra $\mathscr{A}$ which has a cohomological unit, ie a cocycle representing the identity element in $H^*(\mathscr{A})$, is quasi-isomorphic to a strictly unital A_∞–algebra [56, Section 7.2].

An *augmentation* of a strictly unital A_∞–algebra is an A_∞–map $\epsilon \colon \mathscr{A} \to \mathbf{k}$, where $\mathbf{k}$ is considered as a strictly unital A_∞–algebra in degree 0 with trivial differential and higher A_∞–products, and is such that $\epsilon_1(1_\mathscr{A}) = 1_\mathbf{k}$ and ϵ_i for $i > 1$ annihilates any monomial containing $1_\mathscr{A}$. An augmentation is called *strict* if $\epsilon_i = 0$ for $i > 1$. The category of augmented, strictly unital A_∞–algebras is equivalent to the category of strictly augmented, strictly unital A_∞–algebras; see [56, Section 7.2].

2.1.2 A_∞–coalgebras   An A_∞–*coalgebra* $\mathscr{C}$ over $\mathbf{k}$ is a $\mathbb{Z}$–graded $\mathbf{k}$–module with a collection of $\mathbf{k}$–linear grading-preserving maps

$$\Delta_i \colon \mathscr{C} \to \mathscr{C}^{\otimes_\mathbf{k} i}[2-i]$$

for all integers $i \geq 1$, with the following properties. The maps satisfy the co-A_∞–relations

(6)
$$\sum_{i=1}^{d} \sum_{j=0}^{d-i} (\mathbf{1}^{\otimes_\mathbf{k}(d-i-j)} \otimes_\mathbf{k} \Delta_i \otimes_\mathbf{k} \mathbf{1}^{\otimes_\mathbf{k} j}) \Delta_{d-i+1} = 0,$$

where

$$\mathbf{1}^{\otimes_\mathbf{k}(d-i-j)} \otimes_\mathbf{k} \Delta_i \otimes_\mathbf{k} \mathbf{1}^{\otimes_\mathbf{k} j} (c_{d-i+1}, \dots, c_1)$$
$$= (-1)^{|c_1|+\cdots+|c_j|-j}(c_{d-i+1}, \dots, c_{j+2}) \otimes_\mathbf{k} \Delta_i(c_{j+1}) \otimes_\mathbf{k} (c_j, \dots, c_1)$$
$$\in \mathscr{C}^{\otimes_\mathbf{k}(d-i-j)} \otimes_\mathbf{k} \mathscr{C}^{\otimes_\mathbf{k} i} \otimes_\mathbf{k} \mathscr{C}^{\otimes_\mathbf{k} j}.$$

Furthermore, the degree 1 map

$$\mathscr{C}[-1] \to \prod_{i=1}^{\infty} \mathscr{C}[-1]^{\otimes_\mathbf{k} i},$$

with i^{th} component equal to Δ_i, factorizes through the natural inclusion

$$\bigoplus_{i=1}^{\infty} \mathscr{C}[-1]^{\otimes_\mathbf{k} i} \to \prod_{i=1}^{\infty} \mathscr{C}[-1]^{\otimes_\mathbf{k} i}$$

of the direct sum into the direct product.

A *DG–coalgebra* over $\mathbf{k}$ is an A_∞–coalgebra such that $\Delta_i = 0$ for $i \geq 3$. In this case, we call the first two operations the *differential* and the *coproduct*, respectively, and use the following adjustments to obtain an (ordinary) differential graded coalgebra:

(7)
$$\theta c = (-1)^{|c|}\Delta_1(c) \quad \text{and} \quad \Delta(c) = \sum (-1)^{|c_{(2)}|}c_{(1)} \otimes_\mathbf{k} c_{(2)},$$

where we write $\Delta_2(c) = \sum c_{(1)} \otimes_\mathbf{k} c_{(2)}$.

In particular, the coproduct is coassociative, ie $(\Delta \otimes_k \mathbf{1}) \circ \Delta = (\mathbf{1} \otimes_k \Delta) \circ \Delta$, and the graded co-Leibniz rule holds:

$$(8) \qquad \Delta\theta(c) = \sum (-1)^{|c_{(1)}|} c_{(1)} \otimes_k \theta(c_{(2)}) + \theta(c_{(1)}) \otimes_k c_{(2)}.$$

An A_∞–*comap* $\mathfrak{f}: \mathscr{C} \to \mathscr{D}$ between A_∞–coalgebras $\mathscr{C}$ and $\mathscr{D}$ over k, with operations Δ_i and Θ_i for $i \geq 1$, respectively, is a collection of k–linear grading-preserving maps

$$\mathfrak{f}_i: \mathscr{C} \to \mathscr{D}^{\otimes_k i}[1-i], \quad i \geq 1,$$

satisfying the relations

$$\sum_{i=1}^{d} \sum_{j=0}^{d-i} (\mathbf{1}^{\otimes_k (d-i-j)} \otimes_k \Theta_i \otimes_k \mathbf{1}^{\otimes_k j}) \mathfrak{f}_{d-i+1}$$

$$= \sum_{\substack{1 \leq j \leq d \\ 0 < i_1 < i_2 < \cdots < i_j < d}} (\mathfrak{f}_{d-i_j} \otimes_k \cdots \otimes_k \mathfrak{f}_{i_2-i_1} \otimes_k \mathfrak{f}_{i_1}) \Delta_j,$$

where

$$\mathbf{1}^{\otimes_k (d-i-j)} \otimes_k \Theta_i \otimes_k \mathbf{1}^{\otimes_k j} (d_{d-i+1}, \ldots, d_1)$$
$$= (-1)^{|d_1|+\cdots+|d_j|-j} (d_{d-i+1}, \ldots, d_{j+2}) \otimes_k \Theta_i(d_{j+1}) \otimes_k (d_j, \ldots, d_1)$$
$$\in \mathscr{D}^{\otimes_k (d-i-j)} \otimes_k \mathscr{D}^{\otimes_k i} \otimes_k \mathscr{D}^{\otimes_k j}.$$

Furthermore, the degree 0 map

$$\mathscr{C}[-1] \to \prod_{i=1}^{\infty} \mathscr{D}[-1]^{\otimes_k i},$$

with i^{th} component equal to $\mathfrak{f}_i$, factorizes through the natural inclusion

$$(9) \qquad \bigoplus_{i=1}^{\infty} \mathscr{D}[-1]^{\otimes_k i} \to \prod_{i=1}^{\infty} \mathscr{D}[-1]^{\otimes_k i}$$

of the direct sum into the direct product.

An A_∞–comap $\mathfrak{f}: \mathscr{C} \to \mathscr{D}$ is called an A_∞–*quasi-isomorphism* if the map on cohomology $H^*(\mathscr{C}) \to H^*(\mathscr{D})$ induced by $\mathfrak{f}^1$ is an isomorphism.

We say that an A_∞–coalgebra is *strictly counital* if there exists a k–linear map $\epsilon: \mathscr{C} \to k$ such that $(\epsilon \otimes \mathbf{1})\Delta_2 = (\mathbf{1} \otimes \epsilon)\Delta_2 = \mathbf{1}$ and $(\mathbf{1}^{\otimes_k (i-j)} \otimes_k \epsilon \otimes_k \mathbf{1}^{\otimes_k j-1})\Delta_i = 0$ for all $i \neq 2$ and j. Any A_∞–coalgebra $\mathscr{C}$ which has a cohomological counit, ie a cocycle representing the counit in $H^*(\mathscr{C})$, is quasi-isomorphic to a strictly counital A_∞–coalgebra; see [56, Section 7.5].

A *coaugmentation* of a strictly counital A_∞–coalgebra $\mathscr{C}$ is an A_∞–comap $\eta\colon \boldsymbol{k} \to \mathscr{C}$, where $\boldsymbol{k}$ is considered as a vector space in degree 0 with the trivial A_∞–coalgebra structure, and is such that $\epsilon\eta_1 = 1_{\boldsymbol{k}}$ and $(\boldsymbol{1}^{\otimes_k (i-j)} \otimes_{\boldsymbol{k}} \epsilon \otimes_{\boldsymbol{k}} \boldsymbol{1}^{\otimes_k j-1})\eta_i = 0$ for all $i > 1$ and j. The coaugmentation is called *strict* if $\eta_i = 0$ for $i \geq 2$.

A DG–coalgebra $\mathscr{C}$ is called *conilpotent* (also called *cocomplete*) if for any $c \in \mathscr{C}$, there exists an $n \geq 2$ such that c is in the kernel of the iterated comultiplication map defined recursively by $\Delta^{(2)} = \Delta$, and $\Delta^{(n)} = (\boldsymbol{1}^{\otimes_k (n-2)} \otimes_{\boldsymbol{k}} \Delta) \circ \Delta^{(n-1)}$ for $n > 2$. When considering coaugmented DG–coalgebras, conilpotency is enforced only on the coaugmentation ideal $\mathrm{coker}(\eta)$.

2.1.3 Graded dual

We next discuss the graded dual of a graded $\boldsymbol{k}$–module. Since we are working with bimodules over the ring $\boldsymbol{k}$, there are two $\boldsymbol{k}$–linear duals [8].

If $\mathscr{A}$ is a graded $\boldsymbol{k}$–bimodule, $\mathscr{A} = \bigoplus_i \mathscr{A}_i$, then the *graded duals* $\mathscr{A}^\# = \bigoplus_i (\mathscr{A}^\#)_i$ and $^\#\mathscr{A} = \bigoplus_i (^\#\mathscr{A})_i$ are defined as follows. The graded components $(\mathscr{A}^\#)_i$ of $\mathscr{A}^\#$ are left $\boldsymbol{k}$–module maps

$$\hom_{\boldsymbol{k}-}(\mathscr{A}_{-i}, \boldsymbol{k}),$$

and the $\boldsymbol{k}$–bimodule structure on $\mathscr{A}^\#$ is given as follows: if $e_v, e_w \in \boldsymbol{k}$, $a \in (\mathscr{A}^\#)_i$ and $c \in \mathscr{A}_{-i}$, then

$$(10) \qquad\qquad (e_v \cdot a \cdot e_w)(c) = a(ce_v)e_w.$$

The graded components $(^\#\mathscr{A})_i$ of $^\#\mathscr{A}$ in degree i are right $\boldsymbol{k}$–module maps, which we write as

$$\hom_{-\boldsymbol{k}}(\mathscr{A}_{-i}, \boldsymbol{k}),$$

and the $\boldsymbol{k}$–bimodule structure is given by: if $e_v, e_w \in \boldsymbol{k}$, $a \in (^\#\mathscr{A})_i$ and $c \in \mathscr{A}_{-i}$, then

$$(11) \qquad\qquad (e_v \cdot a \cdot e_w)(c) = e_v a(e_w c).$$

Both canonical maps $\mathscr{A} \to {}^\#(\mathscr{A}^\#)$ and $\mathscr{A} \to (^\#\mathscr{A})^\#$ are $\boldsymbol{k}$–bimodule maps, which are isomorphisms if $\mathscr{A}$ is locally finite.

If $V_1, V_2, \ldots, V_n$ are $\boldsymbol{k}$–bimodules, there is a natural map

$$V_n^\# \otimes_{\boldsymbol{k}} V_{n-1}^\# \otimes_{\boldsymbol{k}} \cdots \otimes_{\boldsymbol{k}} V_1^\# \to (V_1 \otimes_{\boldsymbol{k}} V_2 \otimes_{\boldsymbol{k}} \cdots \otimes_{\boldsymbol{k}} V_n)^\#$$

given by

$$(12) \quad (a_n \otimes a_{n-1} \otimes \cdots \otimes a_1)(c_1 \otimes c_2 \otimes \cdots \otimes c_n) := a_1(c_1 a_2(c_2 \cdots a_n(c_n) \cdots)).$$

Similarly, there is a natural map

$$^\#V_n \otimes_{\pmb{k}} {}^\#V_{n-1} \otimes_{\pmb{k}} \cdots \otimes_{\pmb{k}} {}^\#V_1 \to {}^\#(V_1 \otimes_{\pmb{k}} V_2 \otimes_{\pmb{k}} \cdots \otimes_{\pmb{k}} V_n)$$

given by

$$(13) \quad (a_n \otimes a_{n-1} \otimes \cdots \otimes a_1)(c_1 \otimes c_2 \otimes \cdots \otimes c_n) := a_n(\cdots a_2(a_1(c_1)c_2)\cdots c_n).$$

These give the graded duals $\mathscr{C}^\#$ and $^\#\mathscr{C}$ of a coaugmented A_∞–coalgebra $\mathscr{C}$ the structure of augmented A_∞–algebras, with structure maps defined by

$$(14) \qquad \mathfrak{m}_i(a_i, \ldots, a_1)(c) := (-1)^{|c|}(a_i \otimes \cdots \otimes a_1)\Delta_i(c).$$

Note that to get a nonzero product, we must have $|\mathfrak{m}_i(a_i, \ldots, a_1)| = |c|$, hence the sign $(-1)^{|c|}$ equals the sign $(-1)^{|a_1|+\cdots+|a_i|-i}$.

In general, there is no natural way of equipping the graded dual of an augmented A_∞–algebra with an A_∞–coalgebra structure. However, if the grading on $\mathscr{A}$ is locally finite (ie $\mathscr{A}_i$ are finitely generated as $\pmb{k}$–bimodules), it follows that

$$\mathscr{A}^\# \otimes_{\pmb{k}} \mathscr{A}^\# \otimes_{\pmb{k}} \cdots \otimes_{\pmb{k}} \mathscr{A}^\# \cong (\mathscr{A} \otimes_{\pmb{k}} \mathscr{A} \otimes_{\pmb{k}} \cdots \otimes_{\pmb{k}} \mathscr{A})^\#,$$

$$^\#\mathscr{A} \otimes_{\pmb{k}} {}^\#\mathscr{A} \otimes_{\pmb{k}} \cdots \otimes_{\pmb{k}} {}^\#\mathscr{A} \cong {}^\#(\mathscr{A} \otimes_{\pmb{k}} \mathscr{A} \otimes_{\pmb{k}} \cdots \otimes_{\pmb{k}} \mathscr{A}).$$

Using these isomorphisms, the graded duals $\mathscr{A}^\#$ and $^\#\mathscr{A}$ of an augmented A_∞–algebra $\mathscr{A}$ with locally finite grading can be naturally equipped with the structure of a coaugmented A_∞–coalgebra by using the formulas

$$\Delta_i(c)(a_i \otimes_{\pmb{k}} \cdots \otimes_{\pmb{k}} a_1) = (-1)^{|c|}c(\mathfrak{m}_i(a_i, \ldots, a_1)).$$

2.1.4 Twisting cochains Let $(\mathscr{C}, \Delta_\bullet)$ be an A_∞–coalgebra and let $(\mathscr{A}, \mathfrak{m}_1, \mathfrak{m}_2)$ be a DG–algebra. A *twisting cochain* is a $\pmb{k}$–linear map $\mathfrak{t} \colon \mathscr{C} \to \mathscr{A}$ of degree 1 that satisfies

$$(15) \qquad \mathfrak{m}_1 \circ \mathfrak{t} - \mathfrak{t} \circ \Delta_1 + \sum_{d \geq 2}(-1)^d \mathfrak{m}_2^{(d)} \circ \mathfrak{t}^{\otimes_{\pmb{k}} d} \circ \Delta_d = 0,$$

where $\mathfrak{m}_2^{(2)} := \mathfrak{m}_2$ and $\mathfrak{m}_2^{(d)} := \mathfrak{m}_2 \circ (\mathrm{Id}_{\mathscr{A}} \otimes_{\pmb{k}} \mathfrak{m}_2^{(d-1)})$. For $c \in \mathscr{C}$, note that $\Delta_i(c) \neq 0$ for only finitely many i, and hence the potentially infinite sum in (15) is actually finite when it acts on c.

If the coalgebra $\mathscr{C}$ is coaugmented by $\eta \colon \pmb{k} \to \mathscr{C}$ and the algebra $\mathscr{A}$ is augmented $\epsilon \colon \mathscr{A} \to \pmb{k}$, we require in addition that its twisting cochains $\mathfrak{t}$ are compatible in the sense that

$$(16) \qquad\qquad\qquad\qquad \mathfrak{t} \circ \eta = \epsilon \circ \mathfrak{t} = 0.$$

We denote the set of twisting cochains from $\mathscr{C}$ to $\mathscr{A}$ by $\mathrm{Tw}(\mathscr{C}, \mathscr{A})$.

Let $\mathfrak{t} \in \mathrm{Tw}(\mathscr{C}, \mathscr{A})$ be a twisting cochain. Consider the twisted tensor product $\mathscr{A} \otimes_{\boldsymbol{k}}^{\mathfrak{t}} \mathscr{C}$ as a chain complex with differential $d^{\mathfrak{t}} \colon \mathscr{A} \otimes_{\boldsymbol{k}}^{\mathfrak{t}} \mathscr{C} \to \mathscr{A} \otimes_{\boldsymbol{k}}^{\mathfrak{t}} \mathscr{C}$ defined by

$$(17) \quad d^{\mathfrak{t}} = \mathfrak{m}_1 \otimes_{\boldsymbol{k}} \mathrm{Id}_{\mathscr{C}} + \mathrm{Id}_{\mathscr{A}} \otimes_{\boldsymbol{k}} \Delta_1$$
$$+ \sum_{d \geq 2} (\mathfrak{m}_2^{(d)} \otimes \mathrm{Id}_{\mathscr{C}}) \circ (\mathrm{Id}_{\mathscr{A}} \otimes_{\boldsymbol{k}} \mathfrak{t}^{\otimes_{\boldsymbol{k}} d-1} \otimes_{\boldsymbol{k}} \mathrm{Id}_{\mathscr{C}}) \circ (\mathrm{Id}_{\mathscr{A}} \otimes_{\boldsymbol{k}} \Delta_d).$$

Here the differential squares to zero, $d^{\mathfrak{t}} \circ d^{\mathfrak{t}} = 0$, since $\mathfrak{t}$ satisfies (15). This complex is the *Koszul complex* associated with $\mathfrak{t}$. It is called acyclic if the projection to $\boldsymbol{k}$ is a quasi-isomorphism.

One also has an analogous complex of the form $\mathscr{C} \otimes_{\boldsymbol{k}}^{\mathfrak{t}} \mathscr{A}$.

The $\mathbb{K}$–vector space of $\boldsymbol{k}$–bimodule morphisms $\hom_{\boldsymbol{k}\text{-}\boldsymbol{k}}(\mathscr{C}, \mathscr{A})$ carries an A_∞–algebra structure with operations $\mathfrak{n}_d$ for $d \geq 1$, given by

$$\mathfrak{n}_1(t) = \mathfrak{m}_1 \circ t + (-1)^{|t|} t \circ \Delta_1,$$
$$\mathfrak{n}_d(t_d, t_{d-1}, \dots, t_1) = (-1)^{d(|t_d|+\cdots+|t_1|)} \mathfrak{m}_2^{(d)} \circ (t_d \otimes t_{d-1} \otimes \cdots \otimes t_1) \circ \Delta_d \quad \text{for } d \geq 2,$$

where the composition $(t_d \otimes t_{d-1} \otimes \cdots \otimes t_1) \circ \Delta_d$ is defined componentwise. Thus, if $\Delta_d(c) = c_d \otimes \cdots \otimes c_1$, then

$$(t_d \otimes_{\boldsymbol{k}} t_{d-1} \otimes_{\boldsymbol{k}} \cdots \otimes_{\boldsymbol{k}} t_1) \Delta_d(c) = (-1)^{\dagger} t_d(c_d) \otimes_{\boldsymbol{k}} t_{d-1}(c_{d-1}) \otimes_{\boldsymbol{k}} \cdots \otimes_{\boldsymbol{k}} t_1(c_1),$$

where $\dagger = \sum_{j=2}^{d} \sum_{i=1}^{j-1} |c_i| |t_j|$. In this setting, a twisting cochain $\mathfrak{t} \colon \mathscr{C} \to \mathscr{A}$ corresponds to a solution of the Maurer–Cartan equation

$$(18) \qquad\qquad \sum_{i \geq 1} \mathfrak{n}_i(\mathfrak{t}, \mathfrak{t}, \dots, \mathfrak{t}) = 0.$$

(As before, this sum is effectively finite since, for any $c \in \mathscr{C}$, $\Delta_i(c) \neq 0$ only for finitely many i.)

A twisting cochain $\mathfrak{t} \colon \mathscr{C} \to \mathscr{A}$ defines a twisted A_∞–structure on $\hom_{\boldsymbol{k}}(\mathscr{C}, \mathscr{A})$, with operations $\mathfrak{n}_d^{\mathfrak{t}}$ given by

$$\mathfrak{n}_d^{\mathfrak{t}}(t_d, t_{d-1}, \dots, t_1) = \sum_{l_i \geq 0} \mathfrak{n}_{d+l_0+l_1+\cdots+l_d} (\overbrace{\mathfrak{t}, \dots, \mathfrak{t}}^{l_d}, t_d, \overbrace{\mathfrak{t}, \dots, \mathfrak{t}}^{l_{d-1}}, t_{d-1}, \dots, t_1, \overbrace{\mathfrak{t}, \dots, \mathfrak{t}}^{l_0}).$$

We will denote this twisted A_∞–structure by $\hom_{\boldsymbol{k}}^{\mathfrak{t}}(\mathscr{C}, \mathscr{A})$.

There are direct analogues of the above construction if we instead consider a DG–coalgebra $(\mathscr{C}, \Delta_1, \Delta_2)$ and an A_∞–algebra $\mathscr{A}$ with operations $\mathfrak{m}_i$. The module

$\mathrm{hom}_{\boldsymbol{k}\text{-}\boldsymbol{k}}(\mathscr{C}, \mathscr{A})$ has the structure of an A_∞–algebra with operations $\mathfrak{n}_d$ given by

$$\mathfrak{n}_1(t) = \mathfrak{m}_1 \circ t + (-1)^{|t|} t \circ \Delta_1,$$

$$\mathfrak{n}_d(t_d, t_{d-1}, \ldots, t_1) = \mathfrak{m}_d \circ (t_d \otimes t_{d-1} \otimes \cdots \otimes t_1) \circ \Delta_2^{(d)} \quad \text{for } d \geq 2.$$

To make sense of the twisting cochain (18), one needs to make additional assumptions to ensure the convergence of the infinite sum. This holds, for example, if $\mathscr{C}$ is conilpotent.

We remark that if both $\mathscr{C}$ and $\mathscr{A}$ are A_∞–(co)algebras, then defining a twisting cochain is a more complicated matter; cf [57, Introduction]. We will not need this here.

2.2 Bar–cobar duality for A_∞–(co)algebras

In this section we first introduce the bar and cobar constructions and then discuss basic relations between them.

2.2.1 Bar and cobar constructions Let $(\mathscr{A}, \{\mathfrak{m}_j\}_{j\geq 1})$ be a strictly unital A_∞–algebra with a strict augmentation $\epsilon \colon \mathscr{A} \to \boldsymbol{k}$. Define the *augmentation ideal* $\overline{\mathscr{A}} = \ker(\epsilon)$. If we are given a nonunital A_∞–algebra $\overline{\mathscr{A}}$, we can turn it into a strictly unital A_∞–algebra $\mathscr{A} := \boldsymbol{k} \oplus \overline{\mathscr{A}}$ with an augmentation given by projection to $\boldsymbol{k}$.

We next recall the construction of the (reduced) *bar construction* $\mathrm{B}\mathscr{A}$. For any augmented A_∞–algebra $\mathscr{A}$, $\mathrm{B}\mathscr{A}$ is a coaugmented conilpotent DG–coalgebra. As a coaugmented coalgebra, $\mathrm{B}\mathscr{A}$ is defined as

$$\mathrm{B}\mathscr{A} = \boldsymbol{k} \oplus \overline{\mathscr{A}}[1] \oplus \overline{\mathscr{A}}[1]^{\otimes_{\boldsymbol{k}} 2} \oplus \cdots,$$

where [1] denotes the downwards shift by 1. We write a typical monomial using Eilenberg and Mac Lane's notation

$$[a_d | a_{d-1} | \cdots | a_1] = s a_d \otimes_{\boldsymbol{k}} s a_{d-1} \otimes_{\boldsymbol{k}} \cdots \otimes_{\boldsymbol{k}} s a_1,$$

where for $a \in \overline{\mathscr{A}}$, $sa \in \overline{\mathscr{A}}[1]$ denotes the corresponding element in $\overline{\mathscr{A}}[1]$ with degree shifted down by 1.

The differential $b \colon \mathrm{B}\mathscr{A} \to \mathrm{B}\mathscr{A}$ is defined to vanish on $\boldsymbol{k} \subset \mathscr{A}$, so $b_{|\boldsymbol{k}} = 0$, and defined on monomials by

$$b([a_d | a_{d-1} | \cdots | a_1])$$
$$= \sum_{i,j} (-1)^{|a_1| + \cdots + |a_j| - j} [a_d | \cdots | a_{j+i+1} | \mathfrak{m}_i(a_{j+i}, \ldots, a_{j+1}) | a_j | \cdots | a_1].$$

The coproduct $\Delta_2\colon \mathrm{B}\mathscr{A} \to \mathrm{B}\mathscr{A} \otimes_{\boldsymbol{k}} \mathrm{B}\mathscr{A}$ is defined by

$$\Delta_2([a_d|a_{d-1}|\cdots|a_1]) = \sum_{i=0}^{d} (-1)^{|a_i|+\cdots+|a_1|-i}[a_d|a_{d-1}|\cdots|a_{i+1}] \otimes_{\boldsymbol{k}} [a_i|a_{i-1}|\cdots|a_1].$$

The slightly unusual sign $(-1)^{|a_i|+\cdots+|a_1|-i}$ appears as a consequence of the following two facts:

(i) The equation $b^2 = 0$ is equivalent to the A_∞–relations (3) for $(\mathfrak{m}_i)_{i\geq 1}$.

(ii) The pair (b, Δ_2) satisfies the co-A_∞–relations (6).

Redefining (b, Δ_2) to (θ, Δ) using (7) removes the sign in Δ_2, and (θ, Δ) becomes a (usual) coassociative DG–coalgebra, where the co-Leibniz rule (8) holds. The coaugmentation $\eta\colon \boldsymbol{k} \to \mathrm{B}\mathscr{A}$ is defined by letting η_1 be the inclusion of $\boldsymbol{k}$ and $\eta_i = 0$ for $i > 0$.

There is an increasing, exhaustive and bounded below (hence, complete Hausdorff) filtration on the complex $\mathrm{B}\mathscr{A}$,

$$\boldsymbol{k} = \mathcal{F}^0\mathrm{B}\mathscr{A} \subset \mathcal{F}^1\mathrm{B}\mathscr{A} \subset \cdots \subset \mathrm{B}\mathscr{A}, \quad \text{where } \mathcal{F}^p\mathrm{B}\mathscr{A} := \boldsymbol{k} \oplus \bar{\mathscr{A}}[1] \oplus \cdots \oplus \bar{\mathscr{A}}[1]^{\otimes_{\boldsymbol{k}} p}.$$

This induces the *word-length spectral sequence* with

$$E_1^{p,q} = H^{p+q}(\mathcal{F}^p\mathrm{B}\mathscr{A}/\mathcal{F}^{p-1}\mathrm{B}\mathscr{A})$$

converging strongly to

$$E_\infty^{p,q} = \mathcal{F}^p H^{p+q}(\mathrm{B}\mathscr{A})/\mathcal{F}^{p-1}H^{p+q}(\mathrm{B}\mathscr{A})$$

by the classical convergence theorem [64, Theorem 5.5.1]. It can be proved using this spectral sequence that if an A_∞–map $\mathfrak{e}\colon \mathscr{A} \to \mathscr{B}$ is a quasi-isomorphism, then the naturally induced DG–coalgebra map $\mathrm{B}\mathfrak{e}\colon \mathrm{B}\mathscr{A} \to \mathrm{B}\mathscr{B}$ is a quasi-isomorphism; see [49, Proposition 2.2.3].

There is a universal twisting cochain $\mathsf{t}_{\mathscr{A}}\colon \mathrm{B}\mathscr{A} \to \mathscr{A}$ which is nonzero only on $\bar{\mathscr{A}}[1] \subset \mathrm{B}\mathscr{A}$ and is given by the inclusion map $\bar{\mathscr{A}}[1] \to \mathscr{A}$. The twisting cochain $\mathsf{t}_{\mathscr{A}}$ gives rise to a free $\mathscr{A}$–bimodule resolution of $\mathscr{A}$ obtained as a twisted tensor product

$$\mathscr{A} \otimes_{\boldsymbol{k}}^{\mathsf{t}_{\mathscr{A}}} \mathrm{B}\mathscr{A} \otimes_{\boldsymbol{k}}^{\mathsf{t}_{\mathscr{A}}} \mathscr{A},$$

with the differential d given by the formula

$$(19) \quad d = \mathfrak{m}_1 \otimes_{\boldsymbol{k}} \mathrm{Id}_{\mathrm{B}\mathscr{A}} \otimes_{\boldsymbol{k}} \mathrm{Id}_{\mathscr{A}} + \mathrm{Id}_{\mathscr{A}} \otimes_{\boldsymbol{k}} b \otimes_{\boldsymbol{k}} \mathrm{Id}_{\mathscr{A}} + \mathrm{Id}_{\mathscr{A}} \otimes_{\boldsymbol{k}} \mathrm{Id}_{\mathrm{B}\mathscr{A}} \otimes_{\boldsymbol{k}} \mathfrak{m}_1$$

$$+ \left(\sum_{d\geq 2}(\mathfrak{m}_d \otimes_{\boldsymbol{k}} \mathrm{Id}_{\mathrm{B}\mathscr{A}}) \circ (\mathrm{Id}_{\mathscr{A}} \otimes_{\boldsymbol{k}} \mathsf{t}^{\otimes_{\boldsymbol{k}} d-1} \otimes_{\boldsymbol{k}} \mathrm{Id}_{\mathrm{B}\mathscr{A}}) \circ (\mathrm{Id}_{\mathscr{A}} \otimes_{\boldsymbol{k}} \Delta_2^{(d)}) \right) \otimes_{\boldsymbol{k}} \mathrm{Id}_{\mathscr{A}}$$

$$+ \mathrm{Id}_{\mathscr{A}} \otimes_{\boldsymbol{k}} \left(\sum_{d\geq 2}(\mathrm{Id}_{\mathrm{B}\mathscr{A}} \otimes_{\boldsymbol{k}} \mathfrak{m}_d) \circ (\mathrm{Id}_{\mathrm{B}\mathscr{A}} \otimes_{\boldsymbol{k}} \mathsf{t}^{\otimes_{\boldsymbol{k}} d-1} \otimes_{\boldsymbol{k}} \mathrm{Id}_{\mathscr{A}}) \circ (\Delta_2^{(d)} \otimes_{\boldsymbol{k}} \mathrm{Id}_{\mathscr{A}}) \right).$$

This can be used to compute Hochschild homology and cohomology of $\mathscr{A}$ with coefficients in an $\mathscr{A}$–bimodule $\mathscr{M}$.

Consider instead a strictly counital A_∞–coalgebra $\mathscr{C}$ with operations Δ_i and with a strict coaugmentation $\eta\colon \boldsymbol{k} \to \mathscr{C}$. Let $\overline{\mathscr{C}} = \operatorname{coker}(\eta)$ be the coaugmentation ideal. We next recall the *cobar construction*, which associates a DG–algebra $\Omega\mathscr{C}$ to $\mathscr{C}$. As an augmented algebra, $\Omega\mathscr{C}$ is

$$(20) \qquad \Omega\mathscr{C} = \boldsymbol{k} \oplus \overline{\mathscr{C}}[-1] \oplus \overline{\mathscr{C}}[-1]^{\otimes_{\boldsymbol{k}} 2} \oplus \cdots .$$

As before, we write a typical monomial as

$$[c_d|c_{d-1}|\cdots|c_1] = s^{-1}c_d \otimes_{\boldsymbol{k}} s^{-1}c_{d-1} \otimes_{\boldsymbol{k}} \cdots \otimes_{\boldsymbol{k}} s^{-1}c_1,$$

where for $c \in \overline{\mathscr{C}}$, $s^{-1}c \in \overline{\mathscr{C}}[-1]$ denotes the corresponding element in $\overline{\mathscr{C}}[-1]$ with degree shifted up by 1. The differential $\mathfrak{m}_1$ on $\Omega\mathscr{C}$ vanishes on $\boldsymbol{k}$, so $\mathfrak{m}_1|_{\boldsymbol{k}} = 0$, and acts on monomials as

$$\mathfrak{m}_1([c_m|\cdots|c_1]) = \sum_{i,j}(-1)^{|c_1|+\cdots+|c_i|-i}[c_m|\cdots|c_{i+2}|\Delta_j(c_{i+1})|c_i|\cdots|c_1].$$

Here, by abuse of notation, we write Δ_j for the induced coproduct $\overline{\mathscr{C}}[-1] \to \overline{\mathscr{C}}[-1]^{\otimes j}$. The product $\mathfrak{m}_2\colon \Omega\mathscr{C} \otimes \Omega\mathscr{C} \to \Omega\mathscr{C}$ is given by

$$\mathfrak{m}_2([c_m|\cdots|c_{i+1}],[c_i|\cdots|c_1]) = (-1)^{|c_1|+\cdots|c_i|-i}[c_m|\cdots|c_{i+1}|c_i|\cdots|c_1].$$

The slightly unusual sign $(-1)^{|c_1|+\cdots+|c_i|-i}$ appears as a consequence of the following two facts:

  (i)   The equation $\mathfrak{m}_1^2 = 0$ is equivalent to co-A_∞–relations (6) for $(\Delta_j)_{j\geq 1}$.

 (ii)   The pair $(\mathfrak{m}_1, \mathfrak{m}_2)$ satisfies the A_∞–relations (3).

Redefining $(\mathfrak{m}_1, \mathfrak{m}_2)$ to $(d, \cdot)$ using (4) removes the sign in $\mathfrak{m}_2$, and $(d, \cdot)$ becomes a (usual) associative DG–algebra, where the Leibniz rule (5) holds. The augmentation $\epsilon\colon \Omega\mathscr{C} \to \boldsymbol{k}$ is given by letting ϵ_1 be the projection to $\boldsymbol{k}$ and $\epsilon_i = 0$ for $i > 0$.

There is a decreasing, exhaustive, bounded above filtration on the complex $\Omega\mathscr{C}$,

$$\Omega\mathscr{C} = \mathcal{F}^0\Omega\mathscr{C} \supset \mathcal{F}^1\Omega\mathscr{C} \supset \cdots ,$$

given by

$$\mathcal{F}^p\Omega\mathscr{C} := \overline{\mathscr{C}}[-1]^{\otimes_{\boldsymbol{k}} p} \oplus \overline{\mathscr{C}}[-1]^{\otimes_{\boldsymbol{k}} (p+1)} \oplus \cdots .$$

This gives the *word-length spectral sequence* with

$$E_1^{p,q} = H^{p+q}(\mathcal{F}^p\Omega\mathscr{C}/\mathcal{F}^{p+1}\Omega\mathscr{C}).$$

Unlike the case of the word-length filtration on the bar construction, for the cobar construction, in general, convergence may fail. Thus, we introduce completions. We define the *completed cobar construction* to be

$$\widehat{\Omega}\mathscr{C} = \varprojlim_{s}(\Omega\mathscr{C})/(\mathcal{F}^s\Omega\mathscr{C}).$$

The length filtration on $\Omega\mathscr{C}$ induces a filtration $\widehat{\mathcal{F}}$ on $\widehat{\Omega}\mathscr{C}$ defined by

$$\widehat{\mathcal{F}}^p\widehat{\Omega}\mathscr{C} = \varprojlim_{s}(\mathcal{F}^p\Omega\mathscr{C})/(\mathcal{F}^s\Omega\mathscr{C}),$$

which is decreasing, exhaustive, bounded above and complete Hausdorff. The spectral sequence associated to the filtration $\widehat{\mathcal{F}}$ on $\widehat{\Omega}\mathscr{C}$ is isomorphic to the length spectral sequence associated with the filtration $\widehat{\mathcal{F}}$ on $\Omega\mathscr{C}$ and converges conditionally to $H^*(\widehat{\Omega}\mathscr{C})$; see [9, Theorem 9.2]. It converges strongly to $H^*(\widehat{\Omega}\mathscr{C})$ if the spectral sequence is regular, ie only finitely many of the differentials $d_r^{p,q}$ are nonzero for each p and q; see [9, Theorem 7.1]. This holds, for example, if $\Omega\mathscr{C}$ is locally finite.

We say that $\Omega\mathscr{C}$ is complete if the natural map $\Omega\mathscr{C} \to \widehat{\Omega}\mathscr{C}$ is a quasi-isomorphism. For example, it is easy to see that this is the case if $\mathscr{C}$ is locally finite and simply connected.

If $\mathfrak{f}\colon \mathscr{C} \to \mathscr{D}$ is an A_∞–comap which is a quasi-isomorphism of A_∞–coalgebras, and if $\Omega\mathscr{C}$ and $\mathscr{D}$ are complete, then $\Omega\mathfrak{f}$ is a quasi-isomorphism. (This follows from [21, Theorem 7.4]; see also [64, Theorem 5.5.11].) The completeness assumptions are necessary and are related to the completeness of the word-length filtration. A counterexample when the completeness assumptions are dropped can be found in [49, Section 2.4.1].

There is a universal twisting cochain $\mathfrak{t}^\mathscr{C}\colon \mathscr{C} \to \Omega\mathscr{C}$ given by the composition of canonical projection $\mathscr{C} \to \overline{\mathscr{C}}[-1]$ and the canonical inclusion $\overline{\mathscr{C}}[-1] \to \Omega\mathscr{C}$.

2.2.2 Bar–cobar adjunction Suppose that $\mathscr{C}$ is a coaugmented A_∞–coalgebra and $\mathscr{A}$ is an augmented DG–algebra. Then we have a canonical bijection

(21) $$\hom_{\mathrm{DG}}(\Omega\mathscr{C}, \mathscr{A}) \to \mathrm{Tw}(\mathscr{C}, \mathscr{A})$$

given by $\phi \mapsto \phi \circ \mathfrak{t}^\mathscr{C}$. Similarly, if $\mathscr{C}$ is a coaugmented conilpotent DG–coalgebra and $\mathscr{A}$ is an augmented A_∞–algebra, then we have a canonical bijection

(22) $$\hom_{\mathrm{coDG}}(\mathscr{C}, \mathrm{B}\mathscr{A}) \to \mathrm{Tw}(\mathscr{C}, \mathscr{A}),$$

given by $\phi \mapsto \mathfrak{t}_\mathscr{A} \circ \phi$; see [57, lemme 3.17].

Therefore, when $\mathscr{C}$ is a coaugmented conilpotent DG–coalgebra, and $\mathscr{A}$ is an augmented DG–algebra, we have the bar–cobar adjunction

$$\hom_{\mathrm{DG}}(\Omega\mathscr{C}, \mathscr{A}) \cong \hom_{\mathrm{coDG}}(\mathscr{C}, \mathrm{B}\mathscr{A}).$$

Moreover, the natural DG–maps

(23) $$\Omega\mathrm{B}\mathscr{A} \to \mathscr{A} \quad \text{and} \quad \mathscr{C} \to \mathrm{B}\Omega\mathscr{C}$$

are quasi-isomorphisms for any DG–algebra $\mathscr{A}$ and conilpotent DG–coalgebra $\mathscr{C}$; see [56, Section 6.10]. It is also true that for any A_∞–algebra $\mathscr{A}$, the A_∞–algebra map

$$\mathscr{A} \to \Omega\mathrm{B}\mathscr{A}$$

given by the adjunction map $\mathrm{B}\mathscr{A} \to \mathrm{B}\Omega\mathrm{B}\mathscr{A}$ is an A_∞–quasi-isomorphism; see [47, lemme 2.3.4.3]. Note that any A_∞–quasi-isomorphism is invertible up to homotopy [59, Corollary 1.4].

Similarly, for any A_∞–coalgebra $\mathscr{C}$, the A_∞–comap

$$\mathrm{B}\Omega\mathscr{C} \to \mathscr{C}$$

given by the adjunction map $\Omega\mathrm{B}\Omega\mathscr{C} \to \Omega\mathscr{C}$ is an A_∞–quasi-isomorphism.

However, an A_∞–quasi-isomorphism for a general A_∞–coalgebra is not usually a convenient notion since, as we remarked above, a quasi-isomorphism of A_∞–coalgebras between $\mathscr{C}$ and $\mathbb{C}'$ does not necessarily induce a quasi-isomorphism of DG–algebras $\Omega\mathscr{C}$ and $\Omega\mathbb{C}'$.

For this reason, one considers the category of *conilpotent A_∞–coalgebras*. Let $\mathscr{C}$ be a coaugmented A_∞–coalgebra generated over $\boldsymbol{k}$ by variables $(c_i)_{i \in I}$, with I some countable index set, such that there exists a total ordering

$$c_{\sigma(1)} < c_{\sigma(2)} < \cdots,$$

where $\sigma: I \to I$ is a bijection. This produces an increasing filtration

$$\mathcal{F}^0 = \boldsymbol{k} \subset \mathcal{F}^1 \subset \cdots \subset \Omega\mathscr{C}$$

by setting $\mathcal{F}^p = \boldsymbol{k}\langle c_{\sigma(1)}, \ldots, c_{\sigma(p)}\rangle$. Suppose that the structure maps $(\Delta_i)_{i \geq 1}$ are compatible with this filtration, in the sense that $\Delta_i(c_{\sigma(p)}) \subset \mathcal{F}^{p-1}$ for all i and p. Then we call $\mathscr{C}$ a conilpotent A_∞–coalgebra. (More generally, homotopy retracts of such A_∞–coalgebras are called conilpotent [56, Sections 6.10 and 9]. This notion is called *finite type* in [46].). Given two such A_∞–coalgebras $\mathscr{C}$ and $\mathbb{C}'$, one considers

filtered A_∞–comaps between them. In the case of a conilpotent DG–coalgebra $\mathscr{C}$ there exists an increasing filtration on $\Omega\mathscr{C}$ given by the subalgebras $\mathrm{Ker}(\Delta^{(n)})$ that plays the same role; see [47, lemme 1.3.2.3].

We next state the following elementary lemma for later convenience.

Lemma 10 *Let $\mathscr{A}$ be an augmented A_∞–algebra such that the $\boldsymbol{k}$–bimodule structures on $\mathscr{A}$ and $\mathrm{B}\mathscr{A}$ are locally finite. Then there are quasi-isomorphisms of augmented DG–algebras*

$$\Omega(\mathscr{A}^\#) \to (\mathrm{B}\mathscr{A})^\# \quad and \quad \Omega(^\#\mathscr{A}) \to {}^\#(\mathrm{B}\mathscr{A}).$$

Note that the assumption is satisfied when $\mathscr{A}$ is locally finite and simply connected. We shall briefly consider the case when $\mathscr{A}$ is only assumed to be locally finite and connected, in which case we have:

Lemma 11 *Let $\mathscr{A} = \bigoplus_i \mathscr{A}^i$ be a connected, locally finite $\boldsymbol{k}$–bimodule equipped with an augmented A_∞–algebra structure. Then there are maps of DG–algebras*

$$\Omega(\mathscr{A}^\#) \to (\mathrm{B}\mathscr{A})^\# \quad and \quad \Omega(^\#\mathscr{A}) \to {}^\#(\mathrm{B}\mathscr{A})$$

which become quasi-isomorphisms, after completion,

$$\widehat{\Omega}(\mathscr{A}^\#) \to (\mathrm{B}\mathscr{A})^\# \quad and \quad \widehat{\Omega}(^\#\mathscr{A}) \to {}^\#(\mathrm{B}\mathscr{A}).$$

2.3 Koszul duality

Suppose $\mathscr{C}$ is a coaugmented conilpotent A_∞–coalgebra and $\mathscr{A}$ is an augmented DG–algebra. Via the bijection (21), any twisting cochain $\mathfrak{t} \in \mathrm{Tw}(\mathscr{C}, \mathscr{A})$ is of the form $\mathfrak{t} = \phi \circ \mathfrak{t}^{\mathscr{C}}$ for some unique $\phi \in \hom_{\mathrm{DG}}(\Omega\mathscr{C}, \mathscr{A})$. Similarly, if $\mathscr{C}$ is a coaugmented conilpotent DG–coalgebra and $\mathscr{A}$ is an augmented A_∞–algebra, any twisting cochain $\mathfrak{t} \in \mathrm{Tw}(\mathscr{C}, \mathscr{A})$ is of the form $\mathfrak{t} = \mathfrak{t}_{\mathscr{A}} \circ \phi$ for some $\phi \in \hom_{\mathrm{coDG}}(\mathscr{C}, \mathrm{B}\mathscr{A})$.

Definition 12 In either case above we call $\mathfrak{t}$ a *Koszul twisting cochain* if ϕ is a quasi-isomorphism, and we denote the set of Koszul twisting cochains by $\mathrm{Kos}(\mathscr{C}, \mathscr{A})$.

The terminology of Koszul twisting cochains is taken from [49]. They are also called *acyclic twisting cochains* in other sources [47; 56]. This terminology is due to the well-known fact that, under various local-finiteness assumptions, a twisting cochain $\mathfrak{t}$ is Koszul if and only if the Koszul complex (17) associated to $\mathfrak{t}$ is acyclic; see [56, Appendix A].

Informally, if $t \in \mathrm{Kos}(\mathscr{C}, \mathscr{A})$, then, depending on whether we write $t = \phi \circ t^{\mathscr{C}}$ or $t = t_{\mathscr{A}} \circ \phi$, either $\mathscr{A}$ can be used in place of $\Omega\mathscr{C}$, or $\mathscr{C}$ can be used in place of $B\mathscr{A}$ in various resolutions. This, in turn, may lead to smaller complexes to compute with. For example, one can compute Hochschild homology and cohomology of $\mathscr{A}$ and $\Omega\mathscr{C}$ using the $\mathscr{A}$–bimodule resolution of $\mathscr{A}$ given by the complex

$$\mathscr{A} \otimes_{\boldsymbol{k}}^{t} \mathscr{C} \otimes_{\boldsymbol{k}}^{t} \mathscr{A}$$

with the differential as in (19); see [39].

Suppose that $\mathscr{A}$ is an A_∞–algebra with an augmentation $\epsilon : \mathscr{A} \to \boldsymbol{k}$. The augmentation ϵ makes $\boldsymbol{k}$ into a left $\mathscr{A}$–module, or equivalently, a right $\mathscr{A}^{\mathrm{op}}$–module.

Definition 13 The *Koszul dual* of an augmented A_∞–algebra $\mathscr{A}$ is the DG–algebra of left $\mathscr{A}$–module maps from $\boldsymbol{k}$ to itself,

$$E(\mathscr{A}) := \mathrm{RHom}_{\mathscr{A}}(\boldsymbol{k}, \boldsymbol{k}).$$

Recall that for a unital A_∞–algebra $\mathscr{A}$ over a field $\mathbb{K}$ (or a semisimple ring such as $\boldsymbol{k}$), any A_∞–module is both h–projective and h–injective; that is, if M is an A_∞–module over $\mathscr{A}$ and N is an acyclic A_∞–module over $\mathscr{A}$, then the complexes $\mathrm{RHom}_{\mathscr{A}}(M, N)$ and $\mathrm{RHom}_{\mathscr{A}}(N, M)$ are acyclic [59, Lemma 1.16]. Hence, the DG–algebra $\mathrm{RHom}_{\mathscr{A}}(\boldsymbol{k}, \boldsymbol{k})$ can be computed as the A_∞–module homomorphisms from $\boldsymbol{k}$ to itself. (More generally, this holds if $\mathscr{A}$ is h–projective as a complex of $\boldsymbol{k}$–modules, which implies that $\boldsymbol{k}$ is h–projective as an A_∞–module over $\mathscr{A}$.) Therefore, we have the following:

Proposition 14 *If* $\mathscr{A}$ *(resp.* $\mathscr{A}^{\mathrm{op}}$*) is an augmented unital* A_∞*–algebra, then*

$$\mathrm{RHom}_{\mathscr{A}}(\boldsymbol{k}, \boldsymbol{k}) \cong (B\mathscr{A})^{\#} \quad (\textit{resp. } {}^{\#}(B\mathscr{A})).$$

Proof Recall that $\mathscr{A} \otimes_{\boldsymbol{k}} B\mathscr{A}$ is quasi-isomorphic to $\boldsymbol{k}$ as an $\mathscr{A}$–module. Hence, by the hom-tensor adjunction, we have $\mathrm{RHom}_{\mathscr{A}}(\mathscr{A} \otimes_{\boldsymbol{k}} B\mathscr{A}, \boldsymbol{k}) \cong \mathrm{RHom}_{\boldsymbol{k}}(B\mathscr{A}, \boldsymbol{k})$. Since $\mathscr{A}$ is h–projective as a complex of $\boldsymbol{k}$–modules, so is $B\mathscr{A}$; hence the latter is computed by $(B\mathscr{A})^{\#}$. $\qquad\square$

In this model of $E(\mathscr{A})$, the $\boldsymbol{k}$–bimodule structure on $E(\mathscr{A})$ can be seen as in (10), since $\boldsymbol{k}$ is viewed as a left $\boldsymbol{k}$–module induced from its structure as a left $\mathscr{A}$–module. If, instead, we have an augmentation of $\mathscr{A}^{\mathrm{op}}$, then we view $\boldsymbol{k}$ as a right $\mathscr{A}$–module, and the $\boldsymbol{k}$–bimodule structure on $\mathrm{RHom}_{\mathscr{A}}(\boldsymbol{k}, \boldsymbol{k})$ would be given by (11).

The cohomology of $E(\mathscr{A})$ is a graded algebra,

$$\mathrm{Ext}_{\mathscr{A}}(k, k) := H^*(\mathrm{RHom}_{\mathscr{A}}(k, k)) \cong H^*((\mathrm{B}\mathscr{A})^{\#}).$$

Dually, we also have the derived tensor product $k \widehat{\otimes}_{\mathscr{A}} k$, which can be computed by the complex $\mathrm{B}\mathscr{A}$. The cohomology is a graded coalgebra

$$\mathrm{Tor}_{\mathscr{A}}(k, k) := H^*(k \widehat{\otimes}_{\mathscr{A}} k) \cong H^*(\mathrm{B}\mathscr{A}).$$

In particular, if k is a field, we have that $\mathrm{Ext}_{\mathscr{A}}(k, k) \cong (\mathrm{Tor}_{\mathscr{A}}(k, k))^{\#}$ by the universal coefficient theorem.

Remark 15 If $\mathscr{A}$ is a *commutative* algebra (or more generally an E_2–algebra), then $\mathrm{Tor}_{\mathscr{A}}(k, k)$ also has a graded algebra structure, defined via

$$\mathrm{Tor}_{\mathscr{A}}(k, k) \otimes \mathrm{Tor}_{\mathscr{A}}(k, k) \to \mathrm{Tor}_{\mathscr{A} \otimes \mathscr{A}}(k \otimes k, k \otimes k) \to \mathrm{Tor}_{\mathscr{A}}(k, k),$$

induced by the algebra map $\mathscr{A} \otimes \mathscr{A} \to \mathscr{A}$ (which exists since $\mathscr{A}$ is commutative). This should not be confused with the natural coalgebra structure above.

Note that $\mathscr{A}$ itself can be viewed as a left $\mathscr{A}$–module and the map $\epsilon : \mathscr{A} \to k$ is a map of left $\mathscr{A}$–modules; hence, it induces a map of left $E(\mathscr{A})^{\mathrm{op}}$–modules

$$\tilde{\epsilon} : \mathrm{RHom}_{\mathscr{A}}(k, k)^{\mathrm{op}} \to \mathrm{RHom}_{\mathscr{A}}(\mathscr{A}, k),$$

which can in turn be viewed as an augmentation of $E(\mathscr{A})^{\mathrm{op}} = \mathrm{RHom}_{\mathscr{A}}(k, k)^{\mathrm{op}}$, since $\mathrm{RHom}_{\mathscr{A}}(\mathscr{A}, k)$ can again be identified with k as it is the Yoneda image of k as an $\mathscr{A}$–module. Hence, k can be viewed as a right $E(\mathscr{A})$–module.

Definition 16 The double dual of $\mathscr{A}$ is defined to be $E(E(\mathscr{A})) := \mathrm{RHom}_{E(\mathscr{A})}(k, k)$.

There is a natural map from $\mathscr{A}$ to its double dual,

$$\Phi : \mathscr{A} \to \mathrm{RHom}_{E(\mathscr{A})}(k, k),$$

defined via viewing the right $E(\mathscr{A})$–module k as $\mathrm{RHom}_{\mathscr{A}}(\mathscr{A}, k)$ and acting on the left by $\mathscr{A} \cong \mathrm{RHom}_{\mathscr{A}}(\mathscr{A}, \mathscr{A})$.

Definition 17 We say that $\mathscr{A}$ and $E(\mathscr{A})$ are Koszul dual if $\Phi : \mathscr{A} \to \mathrm{RHom}_{E(\mathscr{A})}(k, k)$ is a quasi-isomorphism.

One standard situation in which Koszul duality holds is the following:

Theorem 18 *Suppose $\mathscr{C} = \bigoplus_{i \leq 0} \mathscr{C}^i$ is a locally finite, simply connected $\boldsymbol{k}$–bimodule equipped with an A_∞–coalgebra structure and the coaugmentation $\boldsymbol{k} \cong \mathscr{C}^0 \to \mathscr{C}$. Let $\mathscr{A} = \Omega\mathscr{C}$, which is an augmented connected DG–algebra. Then $E(\mathscr{A}) \cong \mathscr{C}^{\#}$, and $\mathscr{A}$ and $\mathscr{C}^{\#}$ are Koszul dual. In other words, the natural morphism*

$$\Omega\mathscr{C} \to \mathrm{RHom}_{E(\mathscr{A})}(\boldsymbol{k}, \boldsymbol{k})$$

is a quasi-isomorphism.

Proof First, observe that indeed $E(\mathscr{A}) \cong (\mathrm{B}\mathscr{A})^{\#} \cong (\mathrm{B}\Omega\mathscr{C})^{\#} \cong \mathscr{C}^{\#}$ by (23) and because $\mathrm{Hom}_{\boldsymbol{k}-}(-, \boldsymbol{k})$ preserves quasi-isomorphisms. Next, we have that

$$\mathrm{RHom}_{E(\mathscr{A})}(\boldsymbol{k}, \boldsymbol{k}) \cong {}^{\#}(\mathrm{B}(\mathscr{C}^{\#})) \cong \Omega\mathscr{C},$$

where we applied Lemma 10 to $\mathscr{C}^{\#}$ and used the fact that ${}^{\#}(\mathscr{C}^{\#}) \cong \mathscr{C}$ since $\mathscr{C}$ is locally finite. $\qquad\square$

Rather than making the grading assumptions on $\mathscr{C}$ as in Theorem 18, which guarantee that $\mathrm{B}\mathscr{C}^{\#}$ is locally finite, one can directly assume that the grading on the cohomology $H^*(\Omega\mathscr{C})$ is locally finite. This assumption is harder to check in practice but Koszul duality still holds under this assumption, which one can prove by combining the above argument with the homological perturbation lemma; see for example [43, Theorem 2.8].

In the case that $\mathscr{C} = \bigoplus_{i \leq 0} \mathscr{C}^i$ is a locally finite, connected (but not simply connected) $\boldsymbol{k}$–bimodule, Lemma 10 no longer applies. We instead use Lemma 11 to deduce the following weaker duality result:

Proposition 19 *Let $\mathscr{C} = \bigoplus_{i \leq 0} \mathscr{C}^i$ be a connected, locally finite $\boldsymbol{k}$–bimodule, equipped with an A_∞–coalgebra structure and coaugmentation $\boldsymbol{k} \cong \mathscr{C}^0 \to \mathscr{C}$, and let $\mathscr{A} = \Omega\mathscr{C} = \boldsymbol{k} \oplus \bigoplus_{j \geq 1} (\overline{\mathscr{C}}[-1])^{\otimes_{\boldsymbol{k}} j}$, which is an augmented DG–algebra where augmentation is given by projection to $\boldsymbol{k}$. Then $E(\mathscr{A}) \cong \mathscr{C}^{\#}$ and there is a quasi-isomorphism*

$$\hat{\Omega}\mathscr{C} \to \mathrm{RHom}_{E(\mathscr{A})}(\boldsymbol{k}, \boldsymbol{k}).$$

Note that in Proposition 19, $\mathscr{A} = \Omega\mathscr{C}$ is not connected, and may admit other augmentations $\epsilon \colon \mathscr{A} \to \boldsymbol{k}$ than that induced by the cobar construction. Such augmentations will be considered below. For example, suppose that $\mathscr{C} \cong \boldsymbol{k} \oplus \overline{\mathscr{C}}$ is a coaugmented A_∞–coalgebra such that $\overline{\mathscr{C}} = \mathbb{K}\langle c \,|\, c \in \mathcal{R} \rangle$ is generated by elements c from an indexing set $\mathcal{R}$ and that $\epsilon \colon \Omega\mathscr{C} \to \boldsymbol{k}$ is an augmentation, which is induced by a map $\mathscr{C} \to \boldsymbol{k}$ since $\Omega\mathscr{C}$ is free. Now we can consider the coaugmented A_∞–coalgebra $\mathscr{C}^{\epsilon} = \boldsymbol{k} \oplus \overline{\mathscr{C}}^{\epsilon}$ such that

$$\overline{\mathscr{C}}^{\epsilon} = \mathbb{K}\langle c - \epsilon(c)1_{\boldsymbol{k}} \,|\, c \in \mathcal{R} \rangle.$$

Then $\Omega\mathscr{C}$ and $\Omega\mathscr{C}^{\epsilon}$ are quasi-isomorphic as nonaugmented DG–algebras, and the augmentation on $\Omega\mathscr{C}^{\epsilon}$ induced by the cobar construction coincides with the given augmentation ϵ on Ω.

Remark 20   When $\mathscr{C}$ is not simply connected, the proof of duality fails precisely because $B\mathscr{C}^{\#}$ is not locally finite. Nevertheless, the duality result can still be proved in certain cases where an extra *weight* grading (internal degree, or Adams degree) is available; see [50; 56, Appendix A.2; 39]. We will not study this situation systematically in this paper, but it is important as it extends the range of applicability of Koszul duality theory. In the setting of Chekanov–Eliashberg DG–algebras, such a situation was considered in [32].

3   Legendrian (co)algebra

In this section we introduce our Legendrian invariants. We start by discussing a model for loop space coefficients in Section 3.1. In Section 3.2 we define the Chekanov–Eliashberg algebra with loop space coefficients using moduli spaces of disks of all dimensions, and in Section 3.4 we give a more computable version, which uses only rigid disks and which carries the same information if the Legendrian submanifold is simply connected.

3.1   Coefficients

Before defining our Legendrian invariants, we describe chain models for their coefficients $C_{-*}(\Omega_{p_v}\Lambda_v)$ for $v \in \Gamma^+$. (Notation is as above, Λ_v is a + decorated connected component of the Legendrian Λ.) We work over a field $\mathbb{K}$.

Let $\Omega_{p_v}\Lambda_v$ denote the topological monoid of Moore loops based at p_v, where the monoid structure comes from concatenation of loops; see [5]. Write $C_{-*}(\Omega_{p_v}\Lambda_v)$ for the cubical chain complex (graded cohomologically). Since $\Omega_{p_v}\Lambda_v$ is a topological monoid, the complex $C_{-*}(\Omega_{p_v}\Lambda_v)$ becomes a DG–algebra using the natural product map $\times$ on cubical chains, where the DG–algebra product is given as

$$C_{-*}(\Omega_{p_v}\Lambda_v) \otimes C_{-*}(\Omega_{p_v}\Lambda_v) \xrightarrow{\times} C_{-*}(\Omega_{p_v}\Lambda_v \times \Omega_{p_v}\Lambda_v) \xrightarrow{\circ} C_{-*}(\Omega_{p_v}\Lambda_v).$$

We point out that the $\times$–map

$$\times \colon C_{-*}(\Omega_{p_v}\Lambda_v) \otimes C_{-*}(\Omega_{p_v}\Lambda_v) \to C_{-*}(\Omega_{p_v}\Lambda_v \times \Omega_{p_v}\Lambda_v),$$

when both sides are equipped with the Pontryagin product, is a DG–algebra map.

In what follows, we shall also make use of an inverse to the $\times$, known as the Serre diagonal [60], and the cubical analogue of the Alexander–Whitney map,

$$(24) \qquad \eta \colon C_{-*}(\Omega_{p_v}\Lambda_v \times \Omega_{p_w}\Lambda_w) \to C_{-*}(\Omega_{p_v}\Lambda_v) \otimes C_{-*}(\Omega_{p_w}\Lambda_w).$$

To define this map consider the n–cube I^n with coordinates $(x_1, \ldots, x_n)$. For an ordered j–element subset $J \subset \{1, 2, \ldots, n\}$, $J = (i_1, \ldots, i_j)$ with $i_1 < \cdots < i_j$, and for $\epsilon \in \{0, 1\}$, let $\iota_J^\epsilon \colon I^j \to I^n$ be the map given in coordinates $y = (y_1, \ldots, y_j)$ by

$$x_{i_r}(\iota_J(y)) = y_r \qquad \text{and} \qquad x_m(\iota_J(y)) = \epsilon \quad \text{if } m \notin J.$$

Consider a cubical chain $(\sigma, \tau) \colon I^n \to \Omega_{p_v}\Lambda_v \times \Omega_{p_w}\Lambda_w$. If J is an ordered subset of $\{1, \ldots, n\}$, let J' denote its complement ordered in the natural way. Define η by

$$\eta(\sigma, \tau) = \sum_J (-1)^{JJ'} (\sigma \circ \iota_J^0) \otimes (\tau \circ \iota_{J'}^1),$$

where the sum ranges over all ordered subsets J, and $(-1)^{JJ'}$ is the sign of the permutation JJ'. This is a strictly associative chain map inducing a quasi-isomorphism. Note also that there are obvious extensions of η to several products of loop spaces.

As the cubical chain complex $C_{-*}(\Omega_{p_v}\Lambda_v)$ is very large, it is not the most effective complex for computation. We next discuss smaller models. Starting with a 0–reduced simplicial set X with geometric realization $|X| = \Lambda_v$, an explicit economical model for $C_{-*}(\Omega_{p_v}\Lambda_v)$ is obtained by taking normalized chains on the Kan loop group GX; see [44]. We will not say much about this, but point out that GX is a free simplicial group, whose geometric realization $|GX|$ is homotopy equivalent to $\Omega|X|$; see [37, Corollary 5.11]. Hence, by the monoidal Dold–Kan correspondence [58], the normalized chains on GX give a (weakly) equivalent model of $C_{-*}(\Omega_{p_v}\Lambda_v)$. (Another similar construction is sketched in [46], and leads to a free model.)

Alternatively, one can work with CW–complexes. We start with the simply connected case: for a 1–reduced (unique 0–cell and no 1–cells) CW–structure on Λ_v, the Adams–Hilton construction [5] gives a free DG–algebra model for $C_{-*}(\Omega_{p_v}\Lambda_v)$ as follows. Denote the k–cells of Λ_v by e_k^i for $k \geq 2$ and $i = 1, \ldots, m_k$. The Adams–Hilton construction gives a CW–monoid with a single 0–cell, and generating cells $\bar{e}_k^i$ in dimension $k - 1$, which is quasi-isomorphic to $\Omega_{p_v}(\Lambda_v)$ as a monoid; see [14]. This gives a DG–algebra structure on the free algebra,

$$A(\Lambda_v) := \mathbb{K}\langle \bar{e}_2^1, \ldots, \bar{e}_2^{m_2}, \bar{e}_3^1, \ldots, \bar{e}_3^{m_3}, \ldots, \ldots, \rangle, \quad \text{with } |\bar{e}_k^i| = 1 - k,$$

and a DG–algebra map

$$A(\Lambda_v) \xrightarrow{\Psi} C_{-*}(\Omega_{p_v}\Lambda_v),$$

which is a quasi-isomorphism. The differential d on $A(\Lambda_v)$ is generally not explicit. It is defined recursively as follows. For every 2–cell e_2^i, we have $d(\bar{e}_2^i) = 0$. In general, assuming that d_{k-1} and Ψ_{k-1} have been defined on the k–skeleton $\Lambda_v^{(k)}$ of Λ_v, then for each $(k+1)$–cell e, with attaching map $f \colon S^k \to \Lambda_v^{(k)}$, define $d_k\bar{e} = c$ so that $(\Psi_{k-1})(c) = (\Omega f)_*(\xi)$, where ξ a generator of $H_{k-1}(\Omega S^k)$, and define $\Psi_k(\bar{e})$ to be the k–chain of loops in e (which then depends on earlier choices along the boundary of $\bar{e}$). We remark that $A(\Lambda_v)$ can be identified isomorphically with $\Omega C_*^{\mathrm{CW}}(\Lambda_v)$ for a suitable A_∞–coalgebra structure on the cellular chain complex $C_*^{\mathrm{CW}}(\Lambda)$.

This construction can be generalized to the non-simply-connected case as follows.[1] Begin with a 0–reduced CW–structure on Λ_v. Denote the k–cells by e_k^i for $i = 1, \ldots, m_k$. For each k–cell e_i^k with $k \geq 2$, we have a free variable in degree $1-k$, which we again denote by $\bar{e}_k^i$. For each 1–cell e_1^j with $j = 1, \ldots, m_1$, we have two variables t_j and t_j^{-1} in degree 0 such that $t_j t_j^{-1} = 1 = t_j^{-1}t_j$. Thus, the underlying algebra is the "almost free" algebra of the form

$$A(\Lambda_v) := \mathbb{K}\langle t_1^{\pm 1}, \ldots, t_{m_1}^{\pm 1}, \bar{e}_2^1, \ldots, \bar{e}_2^{m_2}, \bar{e}_3^1, \ldots, \bar{e}_3^{m_3}, \ldots \rangle.$$

This presentation is often more efficient than the presentation one gets from the Kan loop group construction using a simplicial set presentation of Λ_v. However, the differential in the Adams–Hilton model is not easy to describe explicitly. Note that we have

$$d(t_j) = d(t_j^{-1}) = 0$$

for degree reasons. For every 2–cell e_2^i, we have

$$d(\bar{e}_2^i) = 1 - c_i,$$

where $c_i \in \mathbb{K}\langle t_j^{\pm 1} \mid j = 1, \ldots, m_1 \rangle$ represents the class of the attaching map of e_2^i. The differential on higher-dimensional cells is generally harder to compute and is exactly as in the simply connected case discussed above.

Augmentations $\epsilon \colon A(\Lambda_v) \to \mathbb{K}$ correspond to solutions of the equations

$$\begin{cases} \epsilon(t_j)\epsilon(t_j^{-1}) = 1 & \text{for } j = 1, \ldots, m_1, \\ \epsilon(d\bar{e}_2^i) = 0 & \text{for } i = 1, \ldots, m_2. \end{cases}$$

---

[1] See [40; 41]: a generalization was given earlier in [33], however that paper contains an error.

Since $\mathbb{K}\langle t_j^{\pm 1}, j = 1, \ldots, m_1 \mid d\bar{e}_2^i, i = 1, \ldots, m_2 \rangle$ is a presentation of the fundamental group algebra $\mathbb{K}[\pi_1(\Lambda_v, p_v)]$, augmentations correspond exactly to local systems $\pi_1(\Lambda_v, p_v) \to \mathbb{K}$.

We will use the cubical chain complex $C_{-*}(\Omega_{p_v}\Lambda_v)$ to define Legendrian invariants below. Cubical chains work uniformly for all spaces Λ_v and are convenient for showing that the fundamental classes of moduli spaces of pseudoholomorphic disks $\mathcal{M}^{\mathrm{sy}}$, via evaluation maps, take values in the chain complex. The Legendrian invariants can also be studied using any of the smaller models discussed above. It is however important to note that in the non-simply-connected case, we only have either weak equivalence in the homotopy category of DG–algebras, or Morita equivalence [40; 41] of these models and the cubical chain complex $C_{-*}(\Omega_{p_v}\Lambda_v)$.

In the case that Λ_v is simply connected, we can use a DG–algebra map

$$\Phi \colon C_{-*}(\Omega_{p_v}\Lambda_v) \to A(\Lambda_v)$$

that goes in the opposite direction to the Adams–Hilton map to pass to a more economical quasi-isomorphic model. Such a homotopy equivalence Φ is constructed in two steps: first construct, as in [54] using Eilenberg–Moore methods, a DG–algebra quasi-isomorphism

(25) $$C_{-*}(\Omega_{p_v}\Lambda_v) \to \Omega C_*(\Lambda_v),$$

where in both instances C_* refers to the normalized singular chains. Second, using the standard A_∞–coalgebra quasi-isomorphism between the DG–coalgebra of singular chains $C_*(\Lambda_v)$ and the A_∞–coalgebra $C_*^{\mathrm{CW}}(\Lambda_v)$ of normalized cellular chains, one obtains a DG–algebra quasi-isomorphism

$$\Omega C_*(\Lambda_v) \to \Omega C_*^{\mathrm{CW}}(\Lambda_v) = A(\Lambda_v),$$

since we assumed that the complexes C_* and C_*^{CW} are simply connected. (In Section 3.5, we also give a more geometric construction of a DG–algebra quasi-isomorphism Φ corresponding to (25) landing in Morse chains, using Morse flow trees.)

Similarly, if Λ_v is homotopy equivalent to an Eilenberg–Mac Lane space $\mathrm{K}(\pi_1, 1)$, then the singular chains can be replaced with the group algebra $\mathbb{K}[\pi_1]$: there exists a quasi-isomorphism of DG–algebras

$$C_{-*}(\Omega_{p_v}\Lambda_v) \to \mathbb{K}[\pi_1]$$

given by sending a 0–chain to its homology class, and sending all higher-dimensional chains to 0. Note that this DG–algebra map exists for any space Λ_v, but is a quasi-isomorphism only in the case that Λ_v is homotopy equivalent to $K(\pi_1, 1)$.

It is often convenient to use a cofibrant (or free) replacement for $\mathbb{K}[\pi_1]$. For example, if $\Lambda_v = S^1$, then $\mathbb{K}[\pi_1] \cong \mathbb{K}[t, t^{-1}]$ and a cofibrant replacement is given by the free graded algebra

$$\mathbb{K}\langle s_1, t_1, k_1, l_1, u_1 \rangle, \quad \text{where } |s_1| = |t_1| = 0, \ |k_1| = |l_1| = -1, \ |u_1| = -2,$$

with the differential

$$dk_1 = 1 - s_1 t_1, \quad dl_1 = 1 - t_1 s_1, \quad du_1 = k_1 s_1 - s_1 l_1.$$

A DG–algebra defined over $\mathbb{K}[t, t^{-1}]$ can be pulled back to a weakly equivalent DG–algebra over this cofibrant replacement. (See [63] for background in model categories on DG–algebras that we are using in a very simple case here.)

3.2 Construction of Legendrian invariants

As above, let X be a Liouville domain with $c_1(X) = 0$ (for $\mathbb{Z}$–grading) and $\partial X = Y$ its contact boundary. Let $\Lambda = \bigsqcup_{v \in \pi_0(\Lambda)} \Lambda_v$ be a Legendrian submanifold in Y, where Λ_v is a connected component of Λ. Assume that Λ is relatively spin and that its Maslov class vanishes. Let each connected component Λ_v be decorated with a sign and write Λ^+ and Λ^- for the union of the components decorated accordingly. (Our different treatment of Λ^+ and Λ^- is natural from the point of view of handle attachments; recall from the introduction that when Λ^- is a union of spheres, we attach usual Lagrangian disk-handles to Λ^- and handles with cotangent ends to Λ^+.) When we have an exact Lagrangian filling L of Λ (relatively spin and with vanishing Maslov class), L can also be decomposed into embedded components $L = \bigcup_{v \in \Gamma} L_v$. These embedded components are not disjoint: they are allowed to intersect transversely at finitely many points. There is a bijection between Γ and the embedded components of L.

We require that if two components Λ_{w_1} and Λ_{w_2} are boundary components of the same embedded component L_v, then either both belong to Λ^- or both to Λ^+. Using this property, we get a decomposition $\Gamma = \Gamma^+ \sqcup \Gamma^-$, corresponding to the decomposition $\Lambda = \Lambda^+ \sqcup \Lambda^-$.

Let $\boldsymbol{k}$ be the semisimple ring generated by mutually orthogonal idempotents $\{e_v\}_{v \in \Gamma}$. If we are not given a filling of Λ, then the index set Γ is taken to be the connected

components, $\pi_0(\Lambda)$, instead. If we need to distinguish between the two choices, we will denote them as $\boldsymbol{k}_\Lambda$ and $\boldsymbol{k}_L$. Note that there is an injective ring map $\boldsymbol{k}_L \to \boldsymbol{k}_\Lambda$ which takes the idempotent e_v corresponding to an embedded component L_v to the sum $e_{w_1} + \cdots + e_{w_r}$ of idempotents of its boundary components Λ_{w_j}. In particular, this map turns any $\boldsymbol{k}_\Lambda$–bimodule into a $\boldsymbol{k}_L$–bimodule.

Let $\mathcal{R}$ denote the set of nonempty Reeb chords of Λ. This is a graded set: the grading of a chord $c \in \mathcal{R}$ is given by $|c| = -\mathrm{CZ}(c)$, where $\mathrm{CZ}(c)$ is the Conley–Zehnder grading; see Appendix A. (With this convention, the unique chord c of the standard Legendrian unknot in $\mathbb{R}^3$ has $|c| = -2$ and for the corresponding Legendrian unknot in $\mathbb{R}^{2n-1}$ with one Reeb chord c, we have $|c| = -n$. See also Remark 30.)

Note that the vector space generated by $\mathcal{R}$ is a $\boldsymbol{k}$–bimodule, where $e_v \mathcal{R} e_w$ corresponds to the set of Reeb chords from Λ_v to Λ_w. The underlying algebra of the standard Chekanov–Eliashberg DG–algebra is generated freely by $\mathcal{R}$ over $\boldsymbol{k}$. We need to modify this in the case that Λ^+ is nonempty to incorporate chains in the based loop space of Λ_v for $v \in \Gamma^+$. Let us first do this using cubical chains.

For each $v \in \Gamma^+$, consider the cubical chains $C_{-*}(\Omega_{p_v} \Lambda_v)$ as a $\boldsymbol{k}$–algebra by requiring that the left or right action of e_w is trivial except if $w = v$, when it acts as identity. Let CE^* be the algebra over $\boldsymbol{k}$ given by adjoining elements of $\mathcal{R}$ to the union of $C_{-*}(\Omega_{p_v} \Lambda_v)$ for $v \in \Gamma^+$. Thus an element of CE^* is a sum of alternating words in Reeb chords, $\sigma_1 c_1 \sigma_2 c_2 \cdots \sigma_m c_m \sigma_{m+1}$, where c_j are Reeb chords and σ_j chains of based loops in the component of the Legendrian where the adjacent Reeb chord lies.

Now the differential on CE^* is defined by extending the differential on the cubical complexes $C_{-*}(\Omega_{p_v} \Lambda_v)$ for $v \in \Gamma_+$. We describe the differential on a single Reeb chord and extend it by the graded Leibniz rule. The differential d on a Reeb chord decomposes to a sum

$$d = \sum_{i \geq 0} \Delta_i,$$

where for any Reeb chord c_0 only finitely many $\Delta_i(c_0)$ are nonzero. The operations $\Delta_i(c_0)$ are defined as follows.

Consider moduli spaces of holomorphic disks with positive puncture at c_0; for definitions and notation see Appendix A. More precisely, consider Reeb chords $c_i, \ldots, c_1$ such that $c_0 c_i \cdots c_1$ is a composable word and let $\boldsymbol{c} = c_0^+ c_i^- \cdots c_1^-$. Consider the space of disks D_{i+1} with one distinguished positive puncture and i negative punctures (across which the boundary numbering is constant, in the terminology of Appendix A).

Consider the moduli space $\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c})$. As we use a translation-invariant almost complex structure on the symplectization, $\mathbb{R}$ acts by this moduli space by translation. Write

$$(26) \qquad\qquad \mathcal{M}^{\mathrm{sy}\mathbb{R}}(\boldsymbol{c}) := \mathcal{M}^{\mathrm{sy}}(\boldsymbol{c})/\mathbb{R}$$

for the quotient. Theorems 74 and 75 imply that $\mathcal{M}^{\mathrm{sy}\mathbb{R}}(\boldsymbol{c})$ is a smooth orientable manifold, with a natural compactification as a stratified space that carries a fundamental chain. It follows, via the evaluation map at a point in the boundary arcs of D_{i+1}, that $\mathcal{M}^{\mathrm{sy}\mathbb{R}}(\boldsymbol{c})$ parametrizes a chain of paths in the $(i+1)$–fold product $\Lambda^{\times(i+1)}$.

We transform these chains of paths to chains of based loops as follows. On each component Λ_v pick reference arcs connecting all Reeb chord endpoints to the basepoint. Let $U_v \subset \Lambda_v$ be a disk which is a regular neighborhood of these arcs. For convenience we take the disk to be smooth. Then a collar neighborhood on its boundary gives a smooth map $\theta_v \colon (\Lambda_v, *_v) \to (\Lambda_v, *_v)$ such that $\theta_v(D_v) = *_v$ and $\theta_v|_{\Lambda_v \setminus D_v} \colon \Lambda_v \setminus D_v \to \Lambda_v \setminus \{*_v\}$ is a diffeomorphism. To get a chain of loops parametrized by $\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c})$ we compose its chains of paths with the maps θ_v. The resulting chain of paths then takes all Reeb chord endpoints in component Λ_v to the basepoint $*_v$. Thus by composition with θ_v, the moduli space parametrizes a chain of loops in $(\Omega_p \Lambda)^{\times(i+1)}$.

We treat two cases separately. First, if all boundary components of D_{i+1} map to components in Λ^-, then we let

$$(27) \qquad\qquad [\mathcal{M}^{\mathrm{sy}\mathbb{R}}(\boldsymbol{c})] = \begin{cases} nc_i \cdots c_1 & \text{if } \dim(\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c})) = 1, \\ 0 & \text{if } \dim(\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c})) \neq 1, \end{cases}$$

where n is the algebraic number of $\mathbb{R}$ components in the moduli space. Second, if some boundary component maps to a component in Λ^+, then we write $[\mathcal{M}^{\mathrm{sy}\mathbb{R}}(\boldsymbol{c})]$ for the chain of paths in $(\Omega_p \Lambda)^{\times(i+1)}$, where we separate the components in the product by the Reeb chords $\boldsymbol{c}' = c_i \cdots c_1$:

$$[\mathcal{M}^{\mathrm{sy}\mathbb{R}}(\boldsymbol{c})] = \sigma_{i+1} c_i \sigma_i \cdots \sigma_2 c_1 \sigma_1,$$

where σ_j are the components of the fundamental chain $\sigma \colon \mathcal{M}^{\mathrm{sy}\mathbb{R}}(\boldsymbol{c}) \to (\Omega\Lambda)^{\times(i+1)}$. Further, we write e_v for each boundary component that maps to a component in Λ^- in between the Reeb chords $c_i \cdots c_1$ as above.

A subtle point here is that the moduli space $\mathcal{M}^{\mathrm{sy}\mathbb{R}}(\boldsymbol{c})$ naturally gives rise to a chain σ in $C_{-*}(\Omega\Lambda^{\times(i+1)})$ rather than in $C_{-*}(\Omega\Lambda)^{\otimes(i+1)}$. Note that σ_i are simply components of σ, they are not considered as chains. To separate these out we apply the cubical Alexander–Whitney map

$$\eta \colon C_{-*}((\Omega\Lambda)^{\times(i+1)}) \to C_{-*}(\Omega\Lambda)^{\otimes(i+1)},$$

recalled in Section 3.1. With these conventions we then define for $i \geq 0$,

$$\Delta_i(c_0) := \sum_{c = c_0^+ c_i^- \cdots c_1^-} \eta[\mathcal{M}^{\mathrm{sy}}(c)],$$

where we separate the components of the tensor product by the Reeb chords in c', in analogy with the notation for the product chain and where η is the Serre diagonal from equation (24). The output of $\Delta_i(c_0)$ is thus a sum of alternating words of chains of loops in $C_{-*}(\Omega\Lambda)$ and Reeb chords, and Δ_i is an operation of degree $2 - i$ on $\mathrm{LC}_*(\Lambda)$. We point out that if there are Λ^+ components, then higher-dimensional moduli spaces contribute to the differential (unlike the case when $\Lambda = \Lambda^-$). Note also that it is possible to have holomorphic disks contributing to Δ_0, which means that the chord c_0 is the positive puncture of a disk without negative punctures.

Our next result shows that the operations Δ_i give a differential on CE^*. The proof uses boundaries of moduli spaces of holomorphic disks. By SFT compactness [12] and standard gluing results — see eg [31, Appendix A; 23, Appendix B] — the boundary of a moduli space $\mathcal{M}^{\mathrm{sy}}(c)$ consists of several level holomorphic buildings of curves with top level in $\mathcal{M}^{\mathrm{sy}}(c')$ and lower levels in $\mathcal{M}^{\mathrm{sy}}(c'')$, where the positive puncture of a curve in a lower level is attached at a negative puncture of a curve above it. In terms of $\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c)$, standard gluing results imply that in a neighborhood of several-level curves where positive and negative punctures are joined at d Reeb chords, the moduli space $\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c)$ is C^1–diffeomorphic to

$$(28) \qquad\qquad [0,1)^d \times \prod_{j=1}^{d} \mathcal{M}^{\mathrm{sy}\mathbb{R}}(c_j),$$

where the product runs over positive punctures in the holomorphic building which are not the positive puncture of the curve in $\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c)$.

We will use the compact notation $\star$ to denote all such broken configurations and write simply

$$\partial \mathcal{M}^{\mathrm{sy}\mathbb{R}}(c) = \mathcal{M}^{\mathrm{sy}\mathbb{R}}(c') \star \mathcal{M}^{\mathrm{sy}\mathbb{R}}(c'').$$

We next need to consider the fundamental chain of loops $[\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c') \star \mathcal{M}^{\mathrm{sy}\mathbb{R}}(c'')]$ carried by $\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c') \star \mathcal{M}^{\mathrm{sy}\mathbb{R}}(c'')$, or in other words the codimension 1 boundary of $[\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c)]$. If the dimension of $\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c)$ is d then its boundary gives $(d-1)$–dimensional chains of loops in Λ. Consider a several-level building with moduli space components $\mathcal{M}_j^{\mathrm{sy}}$ of dimension $d_j \geq 1$, $j = 1, \ldots, m$. Then, by SFT compactness, $d + 1 = \sum_{j=1}^{m} d_j$. A boundary component of a several-level disk that consists of boundary segments from k disks in $\mathcal{M}_{j_1}^{\mathrm{sy}\mathbb{R}}, \ldots, \mathcal{M}_{j_k}^{\mathrm{sy}\mathbb{R}}$ will then carry a chain of loops in Λ of dimension

$\sum_{r=1}^{k}(d_{j_r} - 1) \leq d - 1$, with equality only if the broken configuration consists of only two levels. It follows that only two level curves contribute to $[\mathcal{M}^{\text{syR}}(c') \star \mathcal{M}^{\text{syR}}(c'')]$. More precisely, the codimension 1 boundary of $\mathcal{M}^{\text{syR}}(c_0^+ c_m^- \cdots c_1^-)$ corresponding to curves joined at only one Reeb chord contributes with top-dimensional stratum of the boundary in the form of a product,

$$\mathcal{M}^{\text{syR}}(c_0^+ c_m^- \cdots c_j^- b^- c_{j-k}^- \cdots c_1^-) \times \mathcal{M}^{\text{syR}}(b^+ c_{j-1}^- \cdots c_{j-k+1}^-).$$

In particular, the chains of loops along the two-level boundary segments of the two-level curve are given by the Pontryagin product of the two adjacent chains of one level segments that form the two-level segment. In the two-level moduli space above, if $k > 1$ there are two two-level boundary segments: the segments between c_j^- and b^- in the upper-level curve joined to the segment between b^+ and c_{j-1}^-, and the segment between c_{j-k+1}^- and b^+ in the lower level is joined to the segment between b^- and c_{j-k}^- in the upper level. If the upper-level moduli space parametrizes the chains of loops in $C_{-*}(\Omega \Lambda^{\times(m-k)})$ with components given by

$$\sigma_m c_m \sigma_{m-1} \cdots c_j \beta'_{j-1} b \beta'_{j-k} c_{j-k} \cdots \sigma_1 c_1 \sigma_0,$$

and the lower-level the chain in $C_{-*}(\Omega \Lambda^{\times(k-1)})$ with components given by

$$\beta''_{j-1} c_{j-1} \sigma_{j-2} \cdots \sigma_{j-k+1} c_{j-k+1} \beta''_{j-k},$$

and if $*$ denotes the constant chain and $\cdot$ the Pontryagin product, then the chain in $C_{-*}(\Omega \Lambda^{\times(m+1)})$ that contributes to the boundary has components

$$(29) \quad (\sigma_m \cdot *) \, c_m \, (\sigma_{m-1} \cdot *) \cdots (\sigma_j \cdot *) \, c_j \, (\beta'_{j-1} \cdot \beta''_{j-1}) \, c_{j-1} \, (* \cdot \sigma_{j-2})$$
$$\cdots (* \cdot \sigma_{j-k+1}) \, c_{j-k+1} \, (\beta'_{j-k} \cdot \beta''_{j-k}) \, c_{j-k} \, (\sigma_{j-k-1} \cdot *) \cdots (\sigma_1 \cdot *) \, c_1 \, (\sigma_0 \cdot *).$$

In the case that $k = 1$, the lower-level curve lies in $\mathcal{M}^{\text{syR}}(b^+)$ and has no negative punctures. In this case the boundary contribution is

$$(30) \quad (\sigma_m \cdot *) \, c_m \cdots (\sigma_j \cdot *) \, c_j (\alpha'_{j-1} \cdot \beta'' \cdot \gamma'_{j-1}) \, c_{j-1} \, (\sigma_{j-2} \cdot *) \cdots (\sigma_1 \cdot *) \, c_1 \, (\sigma_0 \cdot *),$$

where $\alpha' \cdot \beta'' \cdot \gamma'$ denotes the chain of loops parametrized by

$$\mathcal{M}^{\text{syR}}(c_0^+ c_m \cdots b \cdots c_1) \times \mathcal{M}^{\text{syR}}(b^+),$$

which at $(s,t) \in \mathcal{M}^{\text{syR}}(c_0^+ c_m \cdots b \cdots c_1) \times \mathcal{M}^{\text{syR}}(b^+)$ is the loop $\alpha'(s) \cdot \beta''(t) \cdot \gamma'(s)$, where $\cdot$ denotes concatenation.

Proposition 21 *Let $d \colon \text{CE}^* \to \text{CE}^*$ be the map extended to CE^* by the graded Leibniz rule. Then d is a differential, $d^2 = 0$. We call CE^* with the differential d the* **Chekanov–Eliashberg DG–algebra**.

Remark 22 When $\Lambda = \Lambda^-$, CE* was called LCA^* in [32]; this is the cohomologically graded version of the usual Legendrian homology algebra LHA$_*$ in [11]. By definition, we have LHA$_* = $ CE^{-*}.

Proof When there are only components in Λ^- involved, the result follows from standard arguments involving the boundary of 1–dimensional moduli spaces; see eg [28; 23; 11]. Consider therefore the case when there are chains in the loop space involved.

Let $c = c_0^+ c_m^- \cdots c_1^-$. The d–dimensional moduli space $\mathcal{M}^{\mathrm{sy}}(c)$ contributes to dc_0. The codimension 1 strata of its boundary consists of broken curves with one level of dimension $d - k$ and one of dimension k for $0 < k < d$. We find, with ∂ denoting the natural tensor extension of the boundary operator in singular homology over boundary components involved in Λ^+, that

$$\partial[\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c)] = [\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c') \star \mathcal{M}^{\mathrm{sy}\mathbb{R}}(c'')],$$

where $\star$ is as explained above and

$$c' = c_0^+ c_m^- \cdots c_j^- b^- c_{j-k}^- \cdots c_1^- \quad \text{and} \quad c'' = b^+ c_{j-1}^- \cdots c_{j-k+1}^-.$$

We next apply the cubical Alexander–Whitney map η to this formula to deduce

$$\partial \circ \eta[\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c)] = \eta \circ \partial[\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c)] = \eta[\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c') \star \mathcal{M}^{\mathrm{sy}\mathbb{R}}(c'')]$$
$$= \eta[\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c')] \cdot \eta[\mathcal{M}^{\mathrm{sy}\mathbb{R}}(c'')],$$

where $\cdot$ is the Pontryagin product (see (29) and (30)) and we used that η is a chain map and is compatible with the product. The fact that η is a chain map is well known. We verify that it is compatible with the product below. It follows that the terms contributing to d^2 which arise from the differential acting on chains and acting on Reeb chords cancel.

It remains to check that η is compatible with the product. By the explicit product formulas for boundary contributions (29) and (30), we need to check that the compositions

$$C_{-*}(\Omega_{p_u}\Lambda_u \times \Omega_{p_v}\Lambda_v) \otimes C_{-*}(\Omega_{p_v}\Lambda_v \times \Omega_{p_w}\Lambda_w)$$
$$\xrightarrow{\times} C_{-*}(\Omega_{p_u}\Lambda_u \times \Omega_{p_v}\Lambda_v \times \Omega_{p_v}\Lambda_v \times \Omega_{p_w}\Lambda_w)$$
$$\xrightarrow{1 \times \cdot \times 1} C_{-*}(\Omega_{p_u}\Lambda_u \times \Omega_{p_v}\Lambda_v \times \Omega_{p_w}\Lambda_w)$$
$$\xrightarrow{\eta} C_{-*}(\Omega_{p_u}\Lambda_u) \otimes C_{-*}(\Omega_{p_v}\Lambda_v \times \Omega_{p_w}\Lambda_w)$$
$$\xrightarrow{1 \otimes \eta} C_{-*}(\Omega_{p_u}\Lambda_u) \otimes C_{-*}(\Omega_{p_v}\Lambda_v) \otimes C_{-*}(\Omega_{p_w}\Lambda_w)$$

and

$$C_{-*}(\Omega_{p_u}\Lambda_u \times \Omega_{p_v}\Lambda_v) \otimes C_{-*}(\Omega_{p_v}\Lambda_v \times \Omega_{p_w}\Lambda_w)$$
$$\xrightarrow{\eta \otimes \eta} C_{-*}(\Omega_{p_u}\Lambda_u) \otimes C_{-*}(\Omega_{p_v}\Lambda_v) \otimes C_{-*}(\Omega_{p_v}\Lambda_v) \otimes C_{-*}(\Omega_{p_w}\Lambda_w)$$
$$\xrightarrow{1 \otimes \times \otimes 1} C_{-*}(\Omega_{p_u}\Lambda_u) \otimes C_{-*}(\Omega_{p_v}\Lambda_v \times \Omega_{p_v}\Lambda_v) \otimes C_{-*}(\Omega_{p_w}\Lambda_w)$$
$$\xrightarrow{1 \otimes \cdot \otimes 1} C_{-*}(\Omega_{p_u}\Lambda_u) \otimes C_{-*}(\Omega_{p_v}\Lambda_v) \otimes C_{-*}(\Omega_{p_w}\Lambda_w)$$

agree. This is easily checked by evaluating them on a test chain (σ, τ). Note that this uses the fact that the cubical chain complex is a quotient; namely, degenerate cubical chains are divided out. □

Remark 23 As discussed in Appendix A, the moduli spaces in the definition of the differential on CE^* above are defined in terms of anchored moduli spaces $\mathcal{M}^{sy}$, ie moduli spaces of disks with additional interior punctures where holomorphic planes in the filling with asymptotic markers are attached. We point out that in order to calculate the differential one need only take into account rigid such holomorphic planes of dimension zero. For higher-dimensional moduli spaces of planes of dimension $d_0 > 0$, the dimension of the curves in the symplectization is $d + 1 - d_0$ and does not contribute to the d–dimensional chain $[\mathcal{M}^{sy}]$.

Remark 24 As mentioned in the introduction, we relate CE^* as defined above to a parallel copies version of the same algebra, which is defined solely in terms of rigid moduli spaces. In order to do so it is convenient to use a topologically simpler but algebraically more complicated model of CE^*, defined as follows. The generating set of our algebra is extended to chains in the product $(\Omega\Lambda)^{\times(i+1)}$, where we separate the coordinate functions by Reeb chords. We define the product of two such chains by taking the Pontryagin product of the chains at adjacent factors, giving an operation

$$C_{-*}((\Omega\Lambda)^{\times(i+1)}) \otimes C_{-*}((\Omega\Lambda)^{\times(j+1)}) \to C_{-*}((\Omega\Lambda)^{\times(i+j+1)}).$$

See (29) and (30) for explicit formulas. The differential on this version of CE^* is then defined by the singular differential on the chain, and as

$$\Delta_i(c_0) := \sum_{\mathbf{c} = c_0^+ c_i^- \cdots c_1^-} [\mathcal{M}^{sy}(\mathbf{c})]$$

on Reeb chord generators. (In other words, we define it as above but disregard the diagonal approximation.) It follows from the Künneth formula that the two versions of CE^* are quasi-isomorphic.

Although the definition of CE* given above works generally, from a computational perspective it is hard to get our hands on, as the cubical chain complexes $C_{-*}(\Omega_{p_v}\Lambda_v)$ have uncountably many elements.

Next, we provide a modification of the definition, which gives a quasi-isomorphic DG–algebra under the assumption that for each $v \in \Gamma^+$, there exists a DG–algebra quasi-isomorphism

$$\Phi \colon C_{-*}(\Omega_{p_v}\Lambda_v) \to \mathbb{K}\langle\mathcal{E}_v\rangle,$$

where $\mathbb{K}\langle\mathcal{E}_v\rangle$ is a DG–algebra structure on a free algebra generated by a graded finite set $\mathcal{E}_v$. For example, as discussed in Section 3.1, such a DG–algebra map exists when Λ_v is simply connected. (Or, if Λ_v is a $K(\pi_1, 1)$ space, one can first work with the group ring $\mathbb{K}[\pi_1]$ and base-change to a cofibrant replacement of it.)

We define a graded quiver $\mathcal{Q}_\Lambda$ with vertex set $\mathcal{Q}_0 = \Gamma$ and arrows in correspondence with

$$\mathcal{Q} := \mathcal{R} \cup \bigcup_{v \in \Gamma^+} \mathcal{E}_v.$$

More precisely, there are arrows from vertex v to w corresponding to the set of Reeb chords from Λ_v to Λ_w. In addition, for each $v \in \Gamma^+$, there are arrows from v to v corresponding to the elements in $\mathcal{E}_v$. Let $\mathrm{LC}_*(\Lambda)$ be the graded $\boldsymbol{k}$–bimodule generated by $\mathcal{Q}$. Thus, there is one generator for each arrow in $\mathcal{Q}$ and an idempotent e_v for each vertex $v \in \mathcal{Q}_0$. We write $\overline{\mathrm{LC}}_*(\Lambda)$ for the submodule without the idempotents.

Let $\mathrm{CE}^*(\Lambda)$ be $\boldsymbol{k}$–algebra given by the tensor algebra

$$\mathrm{CE}^*(\Lambda) = \boldsymbol{k} \oplus \bigoplus_{i=1}^\infty \overline{\mathrm{LC}}_*(\Lambda)[-1]^{\otimes_{\boldsymbol{k}} i}.$$

Recall that the path algebra of a quiver is defined as a vector space having all paths in the quiver as basis (including, for each vertex v an idempotent e_v), and multiplication given by concatenation of paths. Thus, the $\boldsymbol{k}$–bimodule $\mathrm{CE}^*(\Lambda)$ is the path algebra of the quiver $\mathcal{Q}_\Lambda$, where the grading of each arrow is shifted up by 1. Just like in the cobar construction, we write elements in $\mathrm{CE}^*(\Lambda)$ as

$$[x_m|\cdots|x_1] = s^{-1}x_m \otimes_{\boldsymbol{k}} \cdots \otimes_{\boldsymbol{k}} s^{-1}x_1 \in \mathrm{CE}^*(\Lambda),$$

where $x_j \in \overline{\mathrm{LC}}_*(\Lambda)$.

Next, we equip the $\boldsymbol{k}$–algebra $\mathrm{CE}^*(\Lambda)$ with a differential using the moduli spaces of holomorphic disks defined in Appendix A. This differential is induced by operations

$$\Delta_i \colon \overline{\mathrm{LC}}_*(\Lambda) \to \overline{\mathrm{LC}}_*(\Lambda)^{\otimes_{\boldsymbol{k}} i}, \quad i = 0, 1, \ldots,$$

where $\overline{\mathrm{LC}}_*(\Lambda)^{\otimes_k 0} = k$, which give $\mathrm{LC}_*(\Lambda)$ the structure of an A_∞–coalgebra if $\Delta_0 = 0$.

Consider a generator of $\overline{\mathrm{LC}}_*(\Lambda)$. If it is a generator $\sigma \in \mathcal{E}_v$ of the free model of $C_{-*}(\Omega_{p_v}\Lambda_v)$ for some component $\Lambda_v \subset \Lambda^+$, then we define

$$\tag{31} \Delta_i \sigma = d_i \sigma,$$

where d_i is the coproduct that corresponds to i^{th} homogeneous piece of the differential in the free model $\mathbb{K}\langle\mathcal{E}_v\rangle$ of $C_{-*}(\Omega_{p_v}\Lambda_v)$. If it is a Reeb chord c_0 then we define $\Delta_i(c_0)$ as before using moduli spaces $\mathcal{M}^{\mathrm{sy}}(c)$ but now take the image of all the singular chains in $C_{-*}(\Omega_{p_v}\Lambda_v)$ under the map $\Phi\colon C_{-*}(\Omega_{p_v}\Lambda_v) \to \mathbb{K}\langle\mathcal{E}_v\rangle$. Since the map Φ is a DG–algebra map, the proof that d is a differential on CE^* is the same. Furthermore, since Φ is a quasi-isomorphism, we get a quasi-isomorphic chain complex CE^* if we use $\mathbb{K}\langle\mathcal{E}_v\rangle$ coefficients instead of $C_{-*}(\Omega_{p_v}\Lambda_v)$.

From now on, unless otherwise specified, we will always assume that we work with a free (over k) model of CE^*.

If there exists an augmentation $\epsilon\colon \mathrm{CE}^*(\Lambda) \to k$, then there is a change of coordinates which turns $\mathrm{LC}_*(\Lambda)$ into a A_∞–coalgebra. More precisely, consider the restriction of ϵ, $\epsilon_1\colon \mathrm{LC}_*(\Lambda) \to k$, where we think of $\mathrm{LC}_*(\Lambda)$ as the degree 1 polynomials in $\mathrm{CE}^*(\Lambda)$. Define

$$\mathrm{LC}_*^\epsilon = k \oplus \ker(\epsilon_1).$$

Note that $\ker(\epsilon_1)$ is generated by idempotents e_v and by $c - \epsilon(c)$, where c ranges over the generators of $\overline{\mathrm{LC}}_*(\Lambda)$. Let

$$\phi_\epsilon\colon \bigoplus_{i\geq 0} \mathrm{LC}_*^{\otimes_k i} \to \bigoplus_{i\geq 0} \mathrm{LC}_*^{\otimes_k i}$$

be the k–algebra automorphism defined on generators as $\phi_\epsilon(c) = c + \epsilon(c)$. Define the operations $\Delta_i^\epsilon\colon \mathrm{LC}_*^\epsilon(\Lambda) \to \mathrm{LC}_*^\epsilon(\Lambda)^{\otimes_k i}$ by

$$\Delta_i^\epsilon = \phi_\epsilon \circ \Delta_i \circ \phi_\epsilon^{-1}.$$

Theorem 25 *The operations $(\Delta_i)_{i\geq 1}$ satisfy the A_∞–coalgebra relations, and with these operations, $\mathrm{LC}_*^\epsilon(\Lambda)$ is a coaugmented conilpotent coalgebra.*

Proof Let d denote the differential on $\mathrm{CE}^*(\Lambda)$ and let c be a generator of $\overline{\mathrm{LC}}_*(\Lambda)$. Since ϵ is a augmentation, $\epsilon(dc) = 0$ and it follows that $\Delta_0^\epsilon = 0$. The A_∞–coalgebra

relations then follow by combining the equation $d^2 = 0$ from Theorem 25 with the automorphism ϕ_ϵ.

The coaugmentation is simply the inclusion of k. The fact that LC^ϵ_* is conilpotent follows from Stokes' theorem: the sum of the actions of the Reeb chords at the negative end of a disk contributing to the differential is bounded above by the action of the Reeb chord at the positive end. This gives the desired finiteness. □

Remark 26 If the original operation Δ_0 on $\mathrm{LC}_*(\Lambda)$ equals 0, then the map ϵ which takes all generators of $\overline{\mathrm{LC}}_*(\Lambda)$ to 0 is an augmentation. In this case $\mathrm{LC}_*(\Lambda)^\epsilon = \mathrm{LC}_*(\Lambda)$ by construction.

Remark 27 If there is an augmentation $\epsilon: \mathrm{CE}^*(\Lambda) \to k$, then $\mathrm{CE}^*(\Lambda)$ can be expressed as the cobar construction of a coalgebra: by construction,

$$\mathrm{CE}^*(\Lambda) = \Omega(\mathrm{LC}^\epsilon_*(\Lambda)).$$

We next consider the k–linear dual $\mathrm{LA}^*_\epsilon(\Lambda) := (\mathrm{LC}^\epsilon_*(\Lambda))^\#$ of $\mathrm{LC}^\epsilon_*(\Lambda)$. It follows from Section 2.1.3 that this is an augmented A_∞–algebra. We call it the *Legendrian A_∞–algebra*.

Remark 28 In the case $\Lambda = \Lambda^-$, it can be shown that this A_∞–algebra can be obtained from the endomorphism algebra of the augmentation ϵ in the Aug_- category of [10] by adjoining a unit to it, but is, in general, different from the endomorphism algebra in the Aug_+ category of [55].

Definition 29 Given an augmentation $\epsilon: \mathrm{CE}^*(\Lambda) \to k$, we define the *completed Chekanov–Eliashberg DG–algebra* to be $\widehat{\mathrm{CE}}^*_\epsilon := \mathrm{B}(\mathrm{LA}^*_\epsilon)^\#$. The underlying k–algebra is the completed tensor algebra

$$\widehat{\mathrm{CE}}^*(\Lambda) = \varprojlim_i \mathrm{CE}^*(\Lambda)/I^i = k\langle\langle \overline{\mathrm{LC}}^\epsilon_*(\Lambda)[-1]\rangle\rangle,$$

where $\overline{\mathrm{LC}}^\epsilon_*(\Lambda)$ is the ideal determined by the natural augmentation.

Note that there is a natural chain map

$$\Phi: \mathrm{CE}^*(\Lambda) \to \widehat{\mathrm{CE}}^*_\epsilon(\Lambda).$$

Remark 30 To illustrate the various gradings, the unique Reeb chord for the standard Legendrian unknot Λ in $\mathbb{R}^{2n-1}$ has degree $-n$ in LC_*, n in LA^* and $-(n-1)$ in CE^* (while it is $n-1$ in LHA_*). Therefore, we have the graded isomorphisms $H_*(\mathrm{LC}) \cong H_{-*}(S^n)$, $H^*(\mathrm{LA}) \cong H^*(S^n)$ and $H^*(\mathrm{CE}) \cong H_{-*}(\Omega S^n)$.

3.3 Parallel copies

In this section we will describe the perturbation scheme we will use to define various versions of Lagrangian Floer cohomology. For exact Lagrangian submanifolds with Legendrian boundary we get an induced perturbation scheme for the Legendrian boundary that will allow us to define a simpler version of the Chekanov–Eliashberg algebra which is isomorphic to it when the Legendrian is simply connected.

Let X be a Weinstein manifold and let $L \subset X$ be an exact Lagrangian submanifold with Legendrian boundary Λ. We assume that Λ is embedded but allow L to be a several-component Lagrangian with components that intersect transversely. Assume that the components of L are decorated with signs and write L^+ and L^- for the union of the components decorated with $+$ and $-$, respectively. We will use specific families of Morse functions to shift Lagrangian and Legendrian submanifolds off of themselves in order to relate holomorphic curve theory to Morse theory, and to perform Floer cohomology calculations without Hamiltonian perturbations. Before we discuss the details of this we recall some general results for Morse flow trees.

3.3.1 General results for flow trees In this section we recall several basic results for Morse flow trees from [22]. Morse flow trees live in a neighborhood of a given Lagrangian or Legendrian and are thus defined in the corresponding cotangent bundle or the 1–jet space. In this paper, we will consider only the case of graphical Lagrangians and Legendrians, in the cotangent bundle and 1–jet space, respectively. That corresponds to a simple special case of the more general situation considered in [22], where the nearby Lagrangians and Legendrians are allowed to have singularities when projected to the zero section.

Let M be a smooth manifold with cylindrical ends of the form $\partial M \times [0, \infty)$. Consider the cotangent bundle T^*M and the 1–jet space $J^1 M$. We consider graphical Lagrangians and associated Legendrians $\Gamma_{dF} \subset T^*M$ and $\Gamma_{j^1 F} \subset J^1 M$. At the ends our functions will have the form $F = e^t f(q) + c$, where $t \in [0, \infty)$, $f : \partial M \to \mathbb{R}$ and c is a constant. Let $L_1, \ldots, L_m$ be a collection of graphical Lagrangians in T^*M and $\widetilde{L}_j$ be a Legendrian lift of L_j. As in [22, Section 2.2.2], $\widetilde{L}_j$ defines local gradients as well as cotangent and 1–jet lifts of paths in M. Furthermore, [22, Lemma 2.8] shows that there are maximal flow lines, and as in [22, Definition 2.9] we define their flow orientation. We define flow trees of $\widetilde{L} = \bigcup_j \widetilde{L}_j$ as in [22, Definition 2.10] and we will also use partial flow trees, which are flow trees with "free" 1–valent vertices, not necessarily at a critical point.

We next discuss transversality for flow trees following [22, Section 3]. There are two concepts of dimension of a flow tree Γ involved here: the formal dimension $\dim(\Gamma)$ (see [22, Definition 3.4]) and the geometric dimension $\operatorname{gdim}(\Gamma)$ (see [22, Definition 3.5]). In the graphical case considered here these can be described as follows. The formal dimension $\dim(\Gamma)$ is the dimension of the space of flow trees around a tree Γ without degeneracies (ie only trivalent internal vertices and nonzero length flow lines at positive punctures not at a minimum and negative punctures not at a maximum) assuming transverse intersections of flow manifolds at each vertex. The geometric dimension, on the other hand, is the dimension of a flow trees near Γ with fixed degeneracies (higher-valence vertices, etc). It is then clear that $\operatorname{gdim}(\Gamma) \le \dim(\Gamma)$.

We will use a transversality result that says that for generic geometric data, we have:

(FT) *Every flow tree Γ comes in a smooth family of dimension $\operatorname{gdim}(\Gamma)$. If Γ is degenerate then there is a natural Whitney stratification of the $\dim(\Gamma)$–dimensional space of flow trees around Γ with strata of dimension $\operatorname{gdim}(\Gamma)$.*

This result follows from [22, Proposition 3.14]. We next discuss the adaption (simplification, actually) in the current set-up of the results from [22, Section 3] that lead to [22, Proposition 3.14]. First, since all Lagrangians considered here are graphical, their front projections are smooth with empty singular locus of the front, and the preliminary transversality conditions of [22, Section 3.1.1] hold trivially. This absence of singularities also means that all the results [22, Lemmas 3.9–12] guaranteeing finitely many vertices for trees in the presence of front singularities hold automatically. Then [22, Proposition 3.14] follows readily and shows that for an open dense set of graphical Lagrangians or Legendrians, (FT) holds.

We say that a finite collection of functions $F_1, \ldots, F_k$ on M is flow-tree generic provided (FT) holds. It is a consequence of [22, Proposition 3.14] that any collection of functions can be made flow-tree generic by an arbitrarily small perturbation and furthermore that if $F_1, \ldots, F_{k-1}$ is already flow-tree generic then $F_1, \ldots, F_{k-1}, F_k$ can be made flow-tree generic by an arbitrarily small perturbation of F_k.

3.3.2 Systems of parallel copies In this section we describe how to choose systems of parallel copies for Lagrangians and Legendrians in such a way that higher product and coproduct operations on the Morse complexes can be directly defined (without mapping telescopes of continuation maps, typically used in Hamiltonian Floer cohomology).

Let L be a Lagrangian with Legendrian boundary Λ. Then a neighborhood of L in X looks like T^*L, and along the cylindrical end $[0, \infty) \times \Lambda$, the vector tangent to T^*L in

the direction of the dual of the $[0, \infty)$–direction corresponds to the Reeb direction. We consider a collection of parallel copies L_j for $j = 0, 1, 2, \ldots,$ with $L = L_0$. Here L_j is the graph in T^*L of the differentials dF_j of a Morse function $F_j \colon L \to \mathbb{R}$. The Morse functions F_j will have critical points in the compact part of L and in the cylindrical ends they will look like Morsifications of the Reeb push-off; see below for details.

We next discuss the main strategy without all technicalities: the first Morse function $H_1 = F_1$ gives the first parallel copy at small distance $\epsilon > 0$ from $L_0 = L$. We define L_1 as the graph of the differential of ϵF_1. We want all other copies to be good approximations of L_1 as seen from L_0, so that flow lines between L_j and L_0 and between L_1 and L_0 are sufficiently close that the corresponding spaces of flow lines can be canonically identified. Let $L_j^1 = L_1$ for $j > 1$.

We next construct $L_2 = L_2^2$ as the graph of the differential of a function $\epsilon^2 H_2$ over $L_2^1 = L_1$. For small $\epsilon > 0$, L_2 is then well approximated by L_1 as seen from L_0, and flow lines between L_0 and L_1 can be identified with flow lines between L_0 and L_2. We also want spaces of flow lines between L_1 and L_2 to be identified with flow lines between L_0 and L_1. This holds provided H_2 is a sufficiently good approximation of H_1. Thus we take

$$H_2 = F_1 + \epsilon H_2 = H_1 + \epsilon H_2,$$

where H_2 is sufficiently close to H_1 that the following further condition holds. The Lagrangians L_0, L_1 and L_2 together also define flow trees with three punctures. We take H_2 so that (FT) holds for L_0, L_1 and L_2. It follows from [22, Proposition 3.14] that this can be achieved by an arbitrarily small perturbation of H_2.

The construction now proceeds in the same manner. First, preliminarily, set $L_j^2 = L_2$, $j > 2$. Then let $L_3 = L_3^3$ be the graph of the function $\epsilon^3 H_3$ over $L_3^2 = L_2$. In order for L_3 to look like L_1 from the point of view of L_2 and like L_2 from the point of view of L_1, we take

$$H_3 = F_2 + \epsilon^2 H_3 = H_1 + \epsilon H_2 + \epsilon^2 H_3.$$

For $\epsilon > 0$ sufficiently small we may then identify flow lines and flow trees of any three of the functions, and after an arbitrarily small perturbation of H_3, condition (FT) holds for L_j for $j = 0, 1, 2, 3$. Continuing like this we get

$$(32) \qquad H_k = F_{k-1} + \epsilon^{k-1} H_k = \sum_{k=1}^{k} \epsilon^{k-1} H_k.$$

The corresponding collection $L_0, \ldots, L_k$ of parallel copies then has the following properties: flow trees with boundary on any increasing collection $L_{i_1}, \ldots, L_{i_k}$ are arbitrarily close to flow trees of $L_0, \ldots, L_{k-1}$, and condition (FT) holds for $L_0, \ldots, L_k$.

In order to get the system of parallel copies, we need first to set conventions for the description of the ends. Along the ends our Lagrangians L look like cylinders over Legendrians Λ. A small neighborhood of Λ in the contact boundary can be identified with a small neighborhood of the zero section in the 1–jet space of Λ. We think of this as the intersection of $(-\delta, \delta) \times T^*\Lambda$ with a small neighborhood of the zero section in the cotangent bundle factor, and the contact form is $ds - p \, dq$, where s is a coordinate on $(-\epsilon, \epsilon)$. Along this end the Lagrangian is $[0, \infty) \times \Lambda$ and the corresponding neighborhood is $(-\epsilon, \epsilon) \times [0, \infty) \times T^*\Lambda \subset T^*[0, \infty) \times \Lambda$. We observe then that the result of moving Λ ϵ units along the Reeb flow is the graph of the differential of the function $B(t, q) = \epsilon t$, $(t, q) \in [0, \infty) \times \Lambda$. We will Morsify this Bott situation by considering graphs of $F(t, q) = \epsilon(t + f(q))$, where $f(q)$ is a Morse function. Then the Reeb chords inside the neighborhood of L at infinity between the graph of F and $[0, \infty) \times \Lambda$ are in natural one-to-one correspondence with critical points of f. In this set-up, with infinite ends, there are also flow trees with positive punctures asymptotic to Reeb chords at infinity. In the compactification of the space of flow trees there are flow trees entirely in the $\mathbb{R}$–invariant end, $T^*\mathbb{R} \times \Lambda$. Along the end we have $dF = \epsilon(dt + (\partial f/\partial q) \, dq)$ and the results about Morse flow trees from Section 3.3.1 follow readily from the corresponding results for flow trees of f on Λ.

We now turn to a more detailed description of the construction of parallel copies such that flow trees of ordered subcollections of parallel copies can be identified as discussed above. Write $[0, \infty) \times Y$ and $[0, \infty) \times \Lambda$ for the ends of X and L, respectively. We use coordinates $(t, q) \in [0, \infty) \times \Lambda$. Consider a collection of pairs of Morse functions (F_j, f_j) such that $F_j : L \to \mathbb{R}$ and $f_j : \Lambda \to \mathbb{R}$, $j = 1, 2, \ldots$, are related at the ends by

$$F_j(t, q) = \epsilon(t + f_j(q)) + b \quad \text{for } 1 \gg \epsilon > 0 \text{ and } b > 0,$$

and F_j does not have any local maxima. We next discuss further restrictions related to critical points.

Let (F_1, f_1) be any pair of positive Morse functions as above. Let $z_1^1, \ldots, z_1^m$ be the critical points of F_1 and let $x_1^1, \ldots, x_1^l$ be the critical points of f_1. Fix disjoint coordinate balls $B_j^1 \subset L$ around z_j^1 and $D_j^1 \subset \Lambda$ of x_j^1 such that F_1 and f_1 are given

by quadratic polynomials in these coordinates. Fix small $\sigma > 0$ such that

$$|dF_1| > \sigma_1 = \sigma \quad \text{on } L - \bigcup_{j=1}^{m} B_1^j \quad \text{and} \quad |df_1| > \sigma_1 = \sigma \quad \text{on } \Lambda - \bigcup_{j=1}^{l} D_1^j.$$

Let (F_2, f_2) be another pair of positive Morse functions with m and l critical points $z_2^1, \ldots, z_2^m$ and $x_1^1, \ldots, x_2^l$, respectively, where

$$z_2^j \in B_1^j \quad \text{with } \mathrm{index}(z_2^j) = \mathrm{index}(z_1^j),$$
$$x_2^j \in D_1^j \quad \text{with } \mathrm{index}(x_2^j) = \mathrm{index}(x_1^j).$$

Let $\sigma_2 < \sigma\sigma_1$ and fix coordinate balls $B_2^j \subset B_1^j$ and $D_2^j \subset D_2^j$ such that

$$|dF_2| > \sigma_2 \quad \text{on } L - \bigcup_{j=1}^{m} B_2^j \quad \text{and} \quad |df_2| > \sigma_2 \quad \text{on } \Lambda - \bigcup_{j=1}^{l} D_2^j.$$

Finally, we make sure that $F_2 < \sigma F_1$ and $f_2 < \sigma f_1$, which we obtain by overall scaling. Note that we might have to shrink σ_2 after scaling.

We continue inductively and construct a family of pairs (F_k, f_k), $k = 1, 2, \ldots$ of positive Morse functions with the following properties. Each F_k has m critical points $z_k^1, \ldots, z_k^m$, each f_k has l critical points $x_k^1, \ldots, x_k^l$. There are $\sigma_k > 0$ and disjoint coordinate balls B_k^j around z_k^j and D_k^j around x_k^j such that

$$|dF_k| > \sigma_k \quad \text{on } L - \bigcup_{j=1}^{m} B_k^j \quad \text{and} \quad |df_k| > \sigma_k \quad \text{on } L - \bigcup_{j=1}^{l} D_k^j.$$

Furthermore, the following hold:

- $B_k^j \subset B_{k-1}^j$ and $D_k^j \subset D_{k-1}^j$.
- $\sigma_k \leq \sigma\sigma_{k-1}$.
- $F_k < \sigma F_{k-1}$ and $f_k \leq \sigma f_{k-1}$.

We next take into account the sign decoration. Assume that L^- is nonempty. In this case we first construct functions $\{(G_j, g_j)\}_{j=1}^{\infty}$ exactly as above on all components of L. The actual functions $\{(F_j, f_j)\}$ on (L, Λ) are then $(F_j, f_j) = (G_j, g_j)$ on L^+ and $(F_j, f_j) = (-G_j, -g_j)$ on L^-.

Consider now the Morse functions

$$\left(H_k = \sum_{j=1}^{k} F_j, \, h_k = \sum_{j=1}^{k} f_j \right).$$

(Compare (F_j, f_j) to the function $\epsilon^k F_k$ in the discussion preceding (32).) Define the system of parallel copies $L_0, \ldots, L_k, \ldots$ by letting L_k be the graph Γ_{dH_k} of the differential of H_k. Then we have the following:

Lemma 31 *For generic choice of functions (F_j, f_j) and all sufficiently small $\sigma > 0$ in the construction above, the resulting system of parallel copies $\{L_j\}_{j=0}^{\infty}$ has the following properties:*

- *Intersection points $L_{k_0} \cap L_{k_1}$ are transverse and are in natural one-to-one correspondence with intersection points of $L_0 \cap L_1$ (or, in terms of L only, critical points of F_1 and self-intersection points of $L = L_0$).*

- *On L_+, if $k_0 < k_1$, then Reeb chords from Λ_{k_0} to Λ_{k_1} are in natural one-to-one correspondence with Reeb chords from Λ_0 to Λ_1 (or, in terms of Λ only, critical points of f_1 and Reeb chords of $\Lambda = \Lambda_0$).*

- *For all ordered finite subcollections $L_{k_0}, L_{k_1}, \ldots, L_{k_m}$ with $k_0 < k_1 < \cdots < k_m$ of parallel copies, flow-tree transversality (FT) holds. Furthermore, for any two such ordered collections $L_{k_0}, L_{k_1}, \ldots L_{k_m}$ and $L_{j_0}, L_{j_1}, \ldots L_{j_m}$, the spaces of flow trees are canonically isomorphic.*

Proof Intersection points $L_{k_0} \cap L_{k_1}$ correspond to intersections between the components of L and critical points of $H_{k_1} - H_{k_0}$. Since

$$H_{k_1} - H_{k_0} = F_{k_0+1} + \mathcal{O}(\sigma F_{k_0+1}),$$

we find that, from the point of view of L_{k_0}, L_{k_1} can be viewed as a small perturbation of L_{k_0+1}. In particular, if $k_0 \neq k_1$, then $L_{k_0} \cap L_{k_1}$ is transverse and there is a unique intersection point near each intersection point in $L \cap L_1$ which corresponds to critical points of F_1 and self-intersections of L. The statement on Reeb chords follows similarly. The last statement follows from the special case of [22, Proposition 3.14] as described above. □

Remark 32 In Sections B.2 and B.3, we will also apply this construction to Lagrangian submanifolds $C \subset W$ where W is a Weinstein cobordism with both positive and negative ends. Our Lagrangian submanifolds C that will be equipped with systems of parallel copies will, however, have only positive ends in that case, and the above discussion applies without change. See Remark 65 for a version when both positive and negative ends are perturbed.

3.3.3 Holomorphic disks and flow trees

In this section we discuss results relating holomorphic disks and Morse flow trees that are used in computations with parallel copies.

Let $L \subset X$ be a Lagrangian with cylindrical end $\mathbb{R} \times \Lambda \subset \mathbb{R} \times Y$ and let $\bar{L}(\sigma) = \{L_j(\sigma)\}_{j=0}^{\infty}$ be a system of parallel copies for L constructed as in Section 3.3, where $\sigma > 0$ is the scaling parameter. (Roughly, $L_0 = L$, and for $k = 1, 2, 3, \ldots$ $L_k(\sigma)$ is at distance σ^k from $L_{k-1}(\sigma)$.)

We first consider the relation between local holomorphic disks and Morse flow trees. We have the following result for words $\boldsymbol{a}$ and $\boldsymbol{c}$ of Reeb chords and intersection points corresponding to critical points of the shifting functions (F_1, f_1); see Section 3.3.

Lemma 33 *If $\bar{L}$ is flow-tree generic and $\kappa = (\kappa_0, \kappa_1, \ldots, \kappa_m)$ is an increasing (or decreasing) boundary numbering, then for all $\sigma > 0$ sufficiently small there is a natural one-to-one correspondence between rigid holomorphic disks in $\mathcal{M}^{\mathrm{fi}}(\boldsymbol{a}, \kappa)$ and rigid flow trees of $L_{\kappa_0}, \ldots, L_{\kappa_m}$ with asymptotics according to $\boldsymbol{a}$: there is a neighborhood of the cotangent lift of each rigid tree that contains the boundary of a unique rigid holomorphic disk that is transversely cut out, and each rigid disk has boundary in the neighborhood of some rigid tree. Similarly, there are natural one-to-one correspondences between rigid disks in $\mathcal{M}^{\mathrm{co}}(\boldsymbol{c}; \kappa)$ and $\mathcal{M}^{\mathrm{sy}}_{\parallel}(\boldsymbol{c}, \kappa)$ and corresponding rigid flow trees determined by $L_{\kappa_0}, \ldots, L_{\kappa_m}$.*

Proof This is a consequence of the main results in [22], namely Theorems 1.2 and 1.3, which show, for compact L, as $\sigma \to 0$, that any sequence of rigid disks converges to a rigid flow tree ("compactness") and also that near any rigid flow tree in the limit there is a unique rigid holomorphic disk for all sufficiently small $\sigma > 0$ ("gluing"). It is essential for this one-to-one correspondence to hold that there be no multiply covered disks. In the present case the increasing (or decreasing) condition guarantees no disk is multiply covered. The modifications necessary for the case of cylindrical ends and corresponding Lagrangians of the form $\Lambda \times \mathbb{R} \subset T^* \Lambda \times \mathbb{R}$ are straightforward; see eg [29] for flow tree results in a related setting.   □

The second result concerns a mixed picture where the disks do not lie entirely in the cotangent bundle. In this case holomorphic disks on a system of parallel copies admit a description with holomorphic disks on the underlying Lagrangian with flow trees attached along their boundaries. Such configurations were considered and the main correspondence was worked out in the setting of knot contact homology in [26].

Consider a Lagrangian L and a system of parallel copies $\bar{L} = \{L_j\}_{j=0}^{\infty}$ as above, and let κ be an increasing (or decreasing) boundary decoration. A quantum flow tree of $\bar{L}$ is a finite collection of holomorphic disks $D_1, \ldots, D_m$ with boundaries on subdecorations $\kappa^1 \subset \kappa^m$ of κ with flow trees emanating from their boundaries with boundary subdecorations $\theta^1, \ldots, \theta^n$, such that inserting flow-tree domains in the disks gives a disk, and such that inserting the cotangent lifts of the flow tree at the insertion points we get a boundary condition respecting the decoration κ.

In order to establish the desired correspondence between rigid quantum flow trees and rigid holomorphic disks, we need additional transversality properties of the shifting Morse function that controls the interface of holomorphic disks and flow trees. The argument is the following. Start with a system of parallel copies that satisfies flow-tree transversality, and perturb the almost complex structure so that moduli spaces of holomorphic disks with decreasing (or increasing) boundary decoration (that cannot be multiply covered) are transversely cut out. Then perturb the shifting Morse functions slightly so that partial flow trees are transverse to the boundary evaluation maps of the transversely cut out holomorphic curves. We say that parallel copies and almost complex structures with this transversality property are quantum flow-tree transverse. Arguing as for flow-tree transversality it is straightforward to show that flow-tree transversality holds after arbitrarily small perturbation of the shifting Morse functions.

Remark 34 One feature of using parallel copies that all approximate a single push off is that, for a rigid configuration of quantum disks, disk components can be connected only by Morse flow lines (not trees). To see this, consider a rigid configuration with three disks connected by a tree with a trivalent vertex. As $\sigma \to 0$ the tree converges to a flow line and all three disks have to intersect it. This will generically not happen for a rigid configuration, ie such configurations can appear only when the formal dimension is at least 1.

Lemma 35 *If $\bar{L}$ is quantum flow-tree generic and $\kappa = (\kappa_0, \kappa_1, \ldots, \kappa_m)$ is an increasing (or decreasing) boundary numbering, then for all $\sigma > 0$ sufficiently small there is a natural one-to-one correspondence between rigid holomorphic disks in $\mathcal{M}^{\mathrm{fi}}(\boldsymbol{a}, \kappa)$ and rigid quantum flow trees of $L_{\kappa_0}, \ldots, L_{\kappa_m}$ with asymptotics according to $\boldsymbol{a}$: there is a neighborhood of the cotangent lift of each rigid quantum tree that contains the boundary of a unique rigid holomorphic disk that is transversely cut out, and each rigid disk has boundary in the neighborhood of some rigid quantum tree. Similarly, there are natural one-to-one correspondences between rigid disks in $\mathcal{M}^{\mathrm{co}}(\boldsymbol{c}; \kappa)$ and $\mathcal{M}^{\mathrm{sy}}_{\parallel}(\boldsymbol{c}, \kappa)$ and corresponding rigid quantum flow trees determined by $L_{\kappa_0}, \ldots, L_{\kappa_m}$.*

Proof   As for flow trees there are two main ingredients: "compactness", ie as $\sigma \to 0$, any sequence of rigid disks converges to a rigid quantum flow tree, and "gluing", ie near any rigid flow tree in the limit there is a unique rigid holomorphic disk for all sufficiently small $\sigma > 0$. The technically most difficult point is gluing. It is present already in the flow tree case in [22] discussed above. The full correspondence was worked out with all details in the case of knot conormals in [26, Sections 5.3 and 5.4]. The case considered here can be established in the same way, as follows.

For compactness, the first step is straightforward: if the positive puncture of a disk maps to a Morse Reeb chord of length $\mathcal{O}(\eta)$, then the whole disk lies in a small neighborhood of the Lagrangian, there is no holomorphic disk part, and the correspondence between disks and flow trees in [22] applies; compare [26, Lemma 5.7]. The second step is to show, via an action/area argument, that for any sequence of disks there are neighborhoods of the punctures mapping to short Reeb chords where the disks converge to flow trees. This argument follows the usual steps in flow-tree convergence once segments of action $\mathcal{O}(\eta)$ near such punctures have been found; see [26, Lemma 5.8 and Corollary 5.9]. The third step uses the fact that there is only one positive puncture to show that there can be only one big disk component in the limit; see [26, Lemma 5.11]. The final step is to show that the flow-tree limits that end at punctures meet the big disk in the limit. This follows from an action argument; see [26, Lemma 5.13]. This establishes flow-tree convergence. (Although the arguments in [26, Section 5.3] are written in the case where the Lagrangian is 2–dimensional, the arguments work unchanged in any dimension n.)

For gluing, the first step is to arrange, by standard transversality arguments for curves and perturbation of Morse functions, transversality of the Morse flow data and evaluation maps of holomorphic curves; see [26, Section 5.4.1]. Then there are finitely many rigid flow-tree configurations. The metric and Lagrangian is adapted near the flow tree parts and the points where they meet the boundary of the disk so that there are explicit holomorphic curves near large parts of the flow trees. Flow tree parts and big disk parts are joined over finite regions in the domain and we construct a weight function which is of size 1 in the finite joining regions and exponentially growing along the parts of the domain where we have solutions. The Floer gluing scheme is applied in this setting and surjectivity of gluing is established; see [26, Section 5.4.3]. Again, [26] works with a 2–dimensional Lagrangian, but most of the arguments above are dimension independent and local models generalize to general dimension in a straightforward way.   □

Remark 36 In the setting of Lemma 35, the action integral $\int d\alpha$ of a holomorphic disk in $\mathcal{M}^{\mathrm{sy}}_{\|}(c, \kappa)$ is positive since the almost complex structure J is compatible with the contact form. By Stokes' theorem the action integral equals the difference between the action of the Reeb chord at the positive puncture and the sum of the actions at the negative punctures. It follows in particular that if the positive puncture is a Morse chord, then all negative punctures are Morse chords as well and the moduli spaces are controlled already by Lemma 33.

3.4 Chekanov–Eliashberg algebra with parallel copies

We will relate the Legendrian invariants LA^* and LC_* to Floer (co)homology of exact Lagrangian submanifolds. When studying Lagrangian Floer (co)homology we employ the technique of parallel copies. In this setup no holomorphic disk under consideration is multiply covered, and transversality is achieved by perturbation near punctures as in [28, Lemma 4.5]; see Theorem 74 for an outline of the argument. It will be convenient to express the Legendrian invariants in the same language. As it turns out, in the case that the Legendrians are simply connected this technique leads to a simpler formulation of the theory which incorporates a model of chains on the based loop space automatically. Recall that Lagrangian fillings of Legendrian submanifolds induce augmentations, which after a change of variables lead to noncurved Legendrian A_∞–coalgebras. Geometrically, this means that one uses anchored holomorphic disks. We will assume that an augmentation of Λ has been fixed in this section and all disks considered will be anchored with respect to this augmentation. We now turn to the description of this theory in the Legendrian setting.

Let Λ be as above with decomposition $\Lambda = \Lambda^+ \sqcup \Lambda^-$. Fix a Morse function $f : \Lambda \to \mathbb{R}$ which is positive on Λ^+ and negative on Λ^-. Use it as described in Section 3.3 to construct a system $\overline{\Lambda} = \{\Lambda_j\}_{j=1}^{\infty}$ of parallel copies of $\Lambda = \Lambda_0$.

Let $\mathcal{Q}_0$, $\boldsymbol{k}$ and $\mathcal{R}$ be as above. Let $\mathcal{R}^+$ denote the Reeb chords connecting Λ_0 to Λ_1 that lie in a small neighborhood of Λ_0. By construction there is then a natural one-to-one correspondence between $\mathcal{R}^+$ and the set of critical points of f on Λ^+. (Since f shifts Λ^- in the negative Reeb direction there are no such short chords near Λ^-.) Write

$$\mathcal{R}_{\|} = \mathcal{R} \cup \mathcal{R}^+,$$

and think of chords in $\mathcal{R}^+$ as connecting a component Λ_v to itself. Again the set $\mathcal{R}_{\|}$ is a graded set: Reeb chords $c \in \mathcal{R}$ are graded as above, $|c| = -\mathrm{CZ}(c)$, and the grading

of a short chord $c \in \mathcal{R}^+$ equals the negative of the Morse index of the critical point of f corresponding to c.

We define a graded quiver $\mathcal{Q}_{\|,\Lambda}$ with vertex set $\mathcal{Q}_0 = \Gamma$ and arrows in correspondence with

$$\mathcal{Q}_{\|} := \mathcal{R}_{\|}.$$

More precisely, there are arrows from vertex v to w corresponding to the set of Reeb chords from Λ_v to Λ_w if $v \neq w$, and corresponding to short Reeb chords from $\Lambda_v = \Lambda_{v0}$ to Λ_{v1} if $v = w$.

Let $\mathrm{LC}_*^{\|}(\Lambda)$ be the graded k–bimodule generated by $\mathcal{Q}_{\|}$. We define an A_∞–coalgebra structure on $\mathrm{LC}_*^{\|}(\Lambda)$ given by operations Δ_i as follows. Given a chord c_0 (input) and chords $c_i, \ldots, c_1$ (outputs), we consider the disk D_{i+1} with distinguished puncture at c_0 and a *strictly decreasing* boundary decoration κ. Let $\boldsymbol{c} = c_0^+ c_i^- \cdots c_1^-$. Consider the moduli space $\mathcal{M}_{\|}^{\mathrm{sy}}(\boldsymbol{c}; \kappa)$. We write $|\mathcal{M}_{\|}^{\mathrm{sy}}(\boldsymbol{c}; \kappa)|$ for the algebraic number of $\mathbb{R}$ components in this moduli space provided $\dim(\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c}; \kappa)) = 1$, and $|\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c}; \kappa)| = 0$ otherwise. Define, for $i > 0$,

$$\Delta_i(c_0) := \sum_{\boldsymbol{c} = c_0^+ c_i^- \cdots c_1^-} |\mathcal{M}_{\|}^{\mathrm{sy}}(\boldsymbol{c}; \kappa)| \boldsymbol{c}',$$

where $\boldsymbol{c}' = c_i \cdots c_1$. This gives an operation of degree $2 - i$ on $\mathrm{LC}_*^{\|}(\Lambda)$. Note that $\Delta_0 = 0$ trivially, since the decoration κ is strictly increasing.

It is not a priori clear that the maps Δ_i are well-defined, as the sum may involve infinitely many terms. This can be avoided easily if Λ_v are simply connected for $v \in \Gamma^+$. Namely, the simple connectedness of Λ_v guarantees that the short chords cost index (recall the grading of a Morse chord of Morse index p is graded by $-p$, which means that its input in the dimension formula is $-p + 1$, and in the simply connected case $-1 \leq -p + 1 \leq -\dim(\Lambda) + 1$), whereas the long chords cost energy. In the non-simply-connected case, we will only consider a completed version of $\mathrm{LC}_*^{\|}$ where infinite expressions are allowed. We define the parallel copies algebra first in the simply connected case and turn to the non-simply-connected case in Section 3.4.2.

3.4.1 The parallel copies DG–algebra in the simply connected case
When the operations Δ_i defined above have suitable finiteness properties they define a coalgebra. More precisely, we have the following:

Lemma 37 *Suppose Λ is simply connected or, more generally, that Δ_i factorizes through the natural inclusion*

$$\bigoplus_{i=1}^{\infty} \mathrm{LC}_*^{\|}[-1]^{\otimes_k i} \to \prod_{i=1}^{\infty} \mathrm{LC}_*^{\|}[-1]^{\otimes_k}$$

for all $i \geq 1$. Then $\mathrm{LC}_*^{\|}$ equipped with the operations $(\Delta_i)_{i \geq 1}$ is an A_∞–coalgebra.

Proof Recall that the moduli spaces involved in the definition of the operations Δ_i are moduli spaces of anchored disks, where we use the augmentation induced by the filling L. Consider a 1–dimensional moduli space $\mathcal{M}^{\mathrm{sy}}(c; \kappa)$ and note that its boundary consists of two-level rigid disks by Theorems 75 and 74. There are two cases, either the levels are joined at a chord connecting the same copy of Λ to itself, or the chord connects distinct copies. Since we count anchored disks, the first type of breaking cancels algebraically; compare with [23, Lemma B.6; 29, Section 1.2; 24, Section 3.4; 6, Section 6.1]. (In the usual treatment of Chekanov–Eliashberg algebras this is the statement that augmentations are chain maps.) Theorem 76 then shows that there are canonical identifications between moduli spaces for different increasing numberings and the proof follows since the breakings of the second type are exactly what contribute to the coalgebra relations and are then in algebraic one-to-one correspondence with the endpoints of an oriented compact 1–manifold. $\square$

We will define the parallel copies version of the Chekanov–Eliashberg algebra as the reduced bar construction of the coalgebra above. To simplify matters we show that for a generic system of parallel copies, the coalgebra has a strict counit. (For more general systems of parallel copies one can instead use the definition in Section 3.4.2.) Consider a generator of $\mathrm{LC}_*^{\|}$ which is a small Reeb chord z that corresponds to a critical point of the Morse function f. The coalgebra operations $\Delta_i(z)$ then count holomorphic disks with positive puncture at z which, by an action argument, must lie in a small neighborhood of $\mathbb{R} \times \Lambda$ and rigid such holomorphic disks are in natural correspondence with rigid Morse flow trees; see Remark 36. In the simply connected case, ie for Morse functions without critical points of index 1 and $n - 1$, we have the following result:

Lemma 38 *Suppose Λ is simply connected. If the Morse functions for parallel copies described above are sufficiently close to the first function (ie if $\epsilon > 0$ in the construction of shifting copies is sufficiently small) then, if x_v is the minimum of the Morse function on the component Λ_v, the following holds: $\Delta_1(x_v) = 0$, $\Delta_2(x_v) = x_v \otimes x_v$*

and $\Delta_i(x_v) = 0$ for $i > 2$. *Furthermore, if c is any other generator corresponding to an arrow from v to w, then $\Delta_2(c) = c \otimes x_v + (-1)^{|c|} x_w \otimes c + D_2(c)$, where $D_2(c)$ does not contain any x_v factor, and $\Delta_i(c)$ does not contain any factor x_v for $i \neq 2$.*

Proof Consider a system $\overline{\Lambda}(\sigma) = \{\Lambda_j(\sigma)\}$ of parallel copies as in Section 3.3.2. Note that x_v is a Morse chord of action $\mathcal{O}(\sigma)$. Hence, disks with one positive puncture at x_v can have only other Morse chords as negative punctures. For sufficiently small $\sigma > 0$, Lemma 33 then shows we can compute $\Delta_i(x_v)$ by counting all flow trees with a positive puncture at x_v and Lemma 31 shows that the flow trees are independent of increasing boundary decoration. The equation $\Delta_1(x_v)$ follows since there is no (negative) gradient flow line emanating from a minimum. For the equation $\Delta(x_v) = x_v \otimes x_v$ we consider three copies L_0, L_1 and L_2 and observe that there is a unique flow tree with positive puncture at x_v and two negative punctures at x_v, this flow tree consists simply of two flow lines starting at the minimum chord connecting L_0 to L_2 and ending at the minimum chords connecting L_0 to L_1 and L_1 to L_2, respectively.

To see the equations $\Delta_i(x_v) = 0$ for $i > 2$, we start from a general limiting argument for flow trees of parallel copies. Consider a flow tree with positive puncture at a Morse chord a and negative punctures at Morse chords $b_1, \ldots, b_m$. As we take the limit $\sigma \to 0$, all shifting functions approach multiples of the same Morse function and the flow tree limits to a broken flow line starting at a connecting to b_{i_1}, then from b_{i_1} to b_{i_2}, continuing in this way until all negative punctures have been met.

Consider now a tree with positive puncture at x_v. In the limit this converges to a flow line emanating from x_v, which must then be constant. This shows that all negative punctures must be x_v as well. Since the dimension of a tree with positive puncture at x_v and i negative punctures at x_v is $i - 2$, it follows that $\Delta_i(x_v) = 0$ if $i > 2$.

We next consider the properties of $\Delta_i(c)$ for $c \neq x_v$. The flow trees contributing $c \otimes x_v + (-1)^{|c|} x_w \otimes c$ to $\Delta_2(c)$ are easily found. Considering three parallel copies L_0, L_1 and L_2, the flow trees consist of a single flow line from either one of the endpoints of the chord c to the minimum x_v or x_w of the corresponding component.

We next show that these are the only contributions with negative punctures at minima. We start in the case when c is a Morse chord. Consider a tree with positive puncture at some Morse chord c which does not limit to a single flow line to the minimum as $\sigma \to 0$, and assume that one of the negative punctures is x_v. Consider first the case when x_v is at the last puncture corresponding to the smallest function difference. Assume that the

negative punctures are x_v and a word of punctures b. Then $\|c\| - \|b\| - \|x_v\| - 1 = \|c\| - \|b\| = 0$, where $\|y\| = \text{index}(y) - 1$. Consider the limit when this function difference goes to zero. Then the flow tree goes to a flow tree with a flow line to x_v attached. The remaining flow tree has dimension $\|c\| - \|b\| - 1 = -1$ and hence does not exist by the flow-tree transversality condition (FT) in Section 3.3.1 for the subset of parallel copies obtained by forgetting the last copy.

Consider next the case that x_v is at some other function difference. Then we have negative punctures b before x_v and a after x_v and thus a are Morse chords of smaller function differences. Consider the limit when all these smaller function differences shrink. In the limit we find a flow tree with negative punctures at (b, x_v) with a partial flow tree with negative punctures at a attached. The evaluation dimension of the latter tree (ie the dimension of the partial tree with a free positive puncture) is

$$(n-1) - \|a\| < n - 1,$$

where we use the simple connectedness to get strict inequality. Applying the degeneration above to the remaining tree we get a tree with a flow line to x_v attached, and its evaluation dimension is

$$\|c\| - \|b\|.$$

Now, $\|c\| - \|b\| - \|a\| - 1 = -1$ so these two trees do not meet by condition (FT) in Section 3.3.1.

The remaining possibility is that the small tree with negative punctures a intersects the flow line towards x_v. However, such a tree can be viewed as the original partial tree merging with a flow line from the minimum and then continuing. For $\sigma > 0$, at the scale of the tree with punctures c and b (ie σ^k for some k) the flow line from the minimum and the flow at the positive puncture of the partial tree attached are very close to parallel (nonparallel only at order σ^{k+l} for $l > 0$), therefore the evaluation map at the positive puncture of the partial tree with negative punctures (x_v, a) is arbitrarily close to the evaluation map of the original partial tree with negative punctures a and taking the limit $\sigma \to 0$, the dimension count $\|c\| - \|b\| - \|a\| - 1 = -1$ above shows that these do not intersect if (FT) in Section 3.3.1 holds and $\sigma > 0$ is sufficiently small.

We finally consider the case when c is not a Morse chord, and a disk which in the limit $\sigma \to 0$ does not converge to a constant disk with a flow line attached and which has a puncture at x_v. Such a disk must have a nonconstant disk component in the limit and by Lemma 35 it converges to this disk with flow trees attached in the limit. Let b denote

the negative punctures of the disk in the limit and a all Morse chord negative punctures except x_v. If a flow line to x_v is directly attached to the disk then the dimension of the quantum flow tree obtained by removing this flow line is $\|c\| - \|b\| - \|a\| - 1 = -1$ and hence it does not exist by quantum flow-tree transversality; see Lemma 35. If this is not the case then x_v is one of the negative punctures in a flow tree attached to the disk. Now that partial flow tree with positive puncture constrained to the evaluation map of the disk must be rigid, and arguing exactly as for the trees above, we see that quantum flow-tree transversality shows that no such configuration exists for sufficiently small $\sigma > 0$. □

Remark 39 The simple connectedness is used in the above proof to ensure that cutting with a small tree really reduces dimension. Here, cutting means intersecting and starting a flow from the intersection locus. In the case that there are index 1 critical points one could have $|a| = 0$ in the above, and indeed there are trees with arbitrarily many punctures at index 1 critical points and then a puncture at x_v.

Lemma 38 shows that, in the simply connected case, there is a strict coaugmentation

$$(33) \qquad \eta\colon \boldsymbol{k} \to \boldsymbol{k}_- \oplus \mathrm{LC}_*^{\|}, \quad \text{with } \eta(e_v) = x_v,$$

where η is defined by

$$(34) \qquad \eta(e_v) = \begin{cases} x_v & \text{if } \Lambda_v \subset \Lambda^+, \\ e_v & \text{if } \Lambda_v \subset \Lambda^-. \end{cases}$$

Definition 40 If Λ is simply connected, the parallel copies Chekanov–Eliashberg DG–algebra is

$$\mathrm{CE}_{\|}^* = \Omega(\boldsymbol{k}_- \oplus \mathrm{LC}_*^{\|}).$$

3.4.2 The parallel copies DG–algebra in the non-simply-connected case

In the non-simply-connected case, the operations Δ_i defined counting holomorphic curves are not necessarily finite. To get a workable definition we will instead start from an algebra structure on the dual LA^* of LC_*. More precisely, we proceed as follows.

Let $\mathrm{LA}_{\|}^*(\Lambda)$ be the graded $\boldsymbol{k}$–bimodule generated by $\mathcal{Q}_{\|}$. We define an A_∞–algebra structure on $\mathrm{LA}_{\|}^*(\Lambda)$ given by operations Δ_i' as follows. Given chords $c_i, \ldots, c_1$ (inputs) and a chord c_0 (output), we consider the disk D_{i+1} with distinguished puncture at c_0 and a *strictly increasing* boundary decoration κ. As above, let $\boldsymbol{c} = c_0^+ c_i^- \cdots c_1^-$ and consider $\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c}; \kappa)$. Define, for $i > 0$,

$$\Delta_i'(\boldsymbol{c}') := \sum_{\boldsymbol{c} = c_0^+ c_i^- \cdots c_1^-} |\mathcal{M}_{\|}^{\mathrm{sy}}(\boldsymbol{c}; \kappa)| c_0,$$

where $\boldsymbol{c}' = c_i \cdots c_1$. This gives an operation of degree $2 - i$ on $\mathrm{LA}_\parallel^*(\Lambda)$. Note that $\Delta_0' = 0$ trivially, since the decoration κ is strictly increasing.

Lemma 41 $\mathrm{LA}_\parallel^*$ *equipped with the operations* $(\Delta_i)_{i \geq 1}'$ *is an* A_∞–*algebra.*

Proof This is identical to the proof of Lemma 37. □

In order to define the parallel copies DG–algebra, consider the minimum x of a Morse function on a component of Λ_+. Write $u_x \in \mathrm{LA}_\parallel^*$ for the corresponding generator. Add idempotents e_v to $\mathrm{LA}_\parallel^*$, one for each component $\Lambda_v \subset \Lambda^-$. We then get the algebra $\boldsymbol{k}_- \oplus \mathrm{LA}_\parallel^*$. Equip it with the trivial augmentation ϵ' which is the projection to $\boldsymbol{k}$. Define

$$\widetilde{\mathrm{CE}}_\parallel^* = (\mathrm{B}(\boldsymbol{k}_- \oplus \mathrm{LA}_\parallel^*))^\#,$$

and let $\mathscr{I}$ denote the subalgebra of $\widetilde{\mathrm{CE}}_\parallel^*$ defined as the space of functionals which vanish on monomials not containing u_x for some minimum chord $x \in \Lambda$.

Lemma 42 *The subalgebra* $\mathscr{I}$ *is closed under the differential.*

Proof To see that $\mathscr{I}$ is closed under the differential we check that if u_x is an output puncture of the differential then u_x is also an input puncture. Here the output puncture is the positive puncture of the holomorphic disk. The minimum corresponds to a Morse chord which is of smaller action than any Reeb chord of Λ. This means that all negative punctures in a disk with positive puncture at x must also be Morse chords and that the corresponding disks correspond to Morse flow trees. Since there is no negative gradient flow line that starts at a minimum, any gradient flow tree with positive puncture at x must also have a negative puncture at x. This shows that u_x is an output in a differential disk only when it is also an input. □

Remark 43 In the simply connected case, $\mathrm{LA}_\parallel^* = (\mathrm{LC}_*^\parallel)^\#$ and there is a natural restriction map $\rho \colon \widetilde{\mathrm{CE}}_\parallel^* \to \Omega\mathrm{LC}_*^\parallel$. Since u_x are strict idempotents by Lemma 38, this is a chain map. The kernel of ρ is $\mathscr{I}$ and, consequently, $\Omega\mathrm{LC}_*^\parallel = \widetilde{\mathrm{CE}}_\parallel^*/\mathscr{I}$.

Guided by Remark 43 we define the parallel copies DG–algebra as follows, in the non-simply-connected case.

Definition 44 If Λ is not simply connected, then we define, with notation as above, the completed DG–algebra

$$\widehat{\mathrm{CE}}_\parallel^*(\Lambda) = \widetilde{\mathrm{CE}}_\parallel^*/\mathscr{I}.$$

3.5 Isomorphism between Chekanov–Eliashberg algebras in the simply connected case

We next show that if Λ is simply connected then $\mathrm{CE}^*_\parallel(\Lambda)$ is in fact isomorphic to $\mathrm{CE}^*(\Lambda)$. To this end we first establish a Morse-theoretic version of the Adams result mentioned in the introduction, which here corresponds to the purely local situation of the zero section in a 1–jet space.

Let Q be a *simply connected* smooth manifold with a basepoint $q \in Q$. Fix a system of positive Morse functions $\bar{f} = \{f_j\}_{j=1}^\infty$ as in Section 3.3, and assume that the functions have only one minimum and no index 1 critical points. (This can always be arranged by handle cancellation if $\dim(Q) \geq 5$; see Remarks 50 and 52 for the lower-dimensional case.) We will first discuss a Morse flow tree model for chains on Q, which we denote by $\mathrm{CM}_{-*}(Q)$. Our treatment of Morse flow trees follows [22]; see Sections 3.3.1 and 3.3.2. We first recall the details of the flow tree definitions from [22, Section 2] in the special case needed here.

Consider a strip $\mathbb{R} \times [0, m]$ or half-strip $[T, \infty) \times [0, m]$ with coordinates $s + i\tau$ and with $m-1$ slits along $[a_j, \infty) \times j$ for $j = 1, \ldots, m-1$, and $T \leq a_j$ in the half-strip case. In the half-strip case the vertical segment $T \times [0, m]$ is a finite end that will be used as an input, and we do not consider it as a part of the boundary of the strip with slits. In the strip case, the input is at the puncture $-\infty \times [0, m]$, and in both cases we call punctures at $+\infty$ "output". Order the boundary components according to the positive boundary orientation of the disk with punctures starting from the input and decorate its boundary components by a strictly increasing sequence of positive integers $\kappa_1 < \kappa_2 < \cdots < \kappa_m$. Let $\kappa = \{\kappa_i\}_{i=1}^m$ denote this decoration. Cutting the strip by line segments $a_j \times [0, m]$ for $j = 1, \ldots, m-1$, subdivides it into *strip regions* of the form $[s_0, s_1] \times [\tau_0, \tau_1]$, where $s_0 \in \{-\infty, a_1, \ldots, a_m\}$, $s_1 \in \{a_1, \ldots, a_m, \infty\}$ and $\tau_0, \tau_1 \in \{0, 1, \ldots, m\}$, and with a numbering κ_j on each boundary component $[s_0, s_1] \times \{\tau_0\}$ and $[s_0, s_1] \times \{\tau_1\}$.

Definition 45 [22, Definition 2.10] A flow tree is a continuous map from a strip with slits into Q which in each strip region $[s_0, s_1] \times [\tau_0, \tau_1]$ depends only on the first coordinate $s \in [s_0, s_1]$, and there satisfies the gradient equation

$$\dot{x}(s) = -\nabla(f_{\kappa_i} - f_{\kappa_j})(x(s)),$$

where κ_i is the numbering of the upper horizontal boundary of the strip region and κ_j that of the lower.

A partial flow tree is defined analogously except that the domain is a half-strip with slits $[T, \infty) \times [0, m]$.

If y is a critical point of f_1 then we let $|y| = -\text{index}(y)$ denote the negative Morse index of y. If $\boldsymbol{y} = y_0 y_1 \cdots y_m$ is a word of critical points of f_1 then the space of flow trees $\mathcal{T}(\boldsymbol{y})$ with input puncture at y_0 and output punctures $\boldsymbol{y}' = y_1 \cdots y_m$, in the order induced by the boundary orientation, has dimension (formal dimension in the language of Section 3.3.1)

$$(35) \qquad \dim(\mathcal{T}(\boldsymbol{y})) = |\boldsymbol{y}'| - |y_0| + (m-2).$$

For a sufficiently small perturbing system of Morse functions $\bar{f}$, the space of flow trees is independent of the increasing boundary decoration κ; see Lemma 31.

Let $\text{CM}_{-*}(Q)$ denote the $\mathbb{K}$–module generated by critical points of f_1 and equip $\text{CM}_{-*}(Q)$ with the structure of a coalgebra with operations Δ_i given by

$$\Delta_i(y_0) = \sum_{|\boldsymbol{y}'| = |y_0| - (i-2)} |\mathcal{T}(\boldsymbol{y})| \boldsymbol{y}',$$

where the sum ranges over all $\boldsymbol{y}'$ of word length i. It is not hard to see that the boundary of a 1–dimensional space of flow trees consists of broken rigid flow trees from which it follows that the operations Δ_i satisfy the coalgebra relations; compare Lemma 37. Furthermore, by Lemma 38, the coalgebra has a natural counit, the critical point which is the minimum of f_1. We will call this critical point the *counit critical point*. We say that a flow tree with no puncture mapping to the counit critical point is *counit-free*.

The coalgebra $\text{CM}_{-*}(Q)$ agrees with the Floer coalgebra $\text{CF}_*(Q)$, where we view Q as the zero section in its own cotangent bundle T^*Q as follows. Let $\bar{Q}(\eta) = \{Q_j(\eta)\}_{j=0}^{\infty}$ be the system of parallel copies of the zero section $Q \subset T^*Q$ corresponding to the system of functions $\bar{f}$, where $\eta > 0$ gives the size of the perturbation, $|f_{k+1} - f_k|_{C^s} = \mathcal{O}(\eta^{k+1})$; see Section 3.3.2. Then, by Lemma 33, there is, for all sufficiently small shifts, a natural one-to-one correspondence between rigid holomorphic disks with boundary on $\bar{Q}$ and Morse flow trees in Q determined by $\bar{f}$. This gives a chain isomorphism $\text{CM}_{-*}(Q) \to \text{CF}_*(Q)$.

We now return to the Morse-theoretic approach to Adams' result. We will define a map

$$\phi : C_{-*}(\Omega Q) \to \Omega \text{CM}_{-*}(Q)$$

in terms of the operation of attaching counit-free partial flow trees to Moore loops $\sigma_v : [0, r_v] \to Q$, $r_v \geq 0$, based at $q \in Q$, parametrized by a simplex $v \in \Delta$. To define this
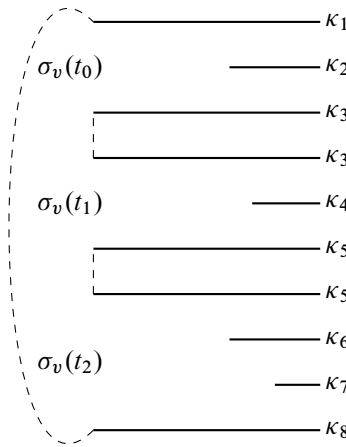
Figure 3: A configuration with three partial flow trees attached to σ_v at the points $\sigma_v(t_0)$, $\sigma_v(t_1)$ and $\sigma_v(t_2)$. The numbers on the right determine the gradient equation at that end. The dashed part represents the loop σ_v.

operation we will use the following notion: we say that a partial flow tree parametrized by a half-strip $\gamma \colon [T, \infty) \times [0, m] \to Q$ starts at a point $p \in Q$ if its input puncture maps to p, where $\gamma(T \times [0, m]) = p$.

Attaching partial flow trees to $\sigma_v \colon [0, r_v] \to Q$ then means fixing points $0 \le t_0 \le t_1 \cdots \le t_m \le r_v$ and partial flow trees Γ_j for $j = 1, \ldots, m$ that start at $\sigma_v(t_j)$. Our map ϕ takes values in $\Omega\mathrm{CM}_{-*}(Q)$ and will accordingly be defined by attaching flow trees which have no output at the counit.

Take the system of parallel copies $\bar{Q}(\eta)$ to be flow-tree generic (to satisfy (FT) of Section 3.3.1). Then the set of positive punctures of minimum free partial flow trees for a fixed increasing boundary numbering constitutes a codimension-two subset in Q and that the corresponding subset for any numbering lies in an $\mathcal{O}(\eta^2)$–neighborhood of it. We pick our Morse functions for the flow trees so that the basepoint q does not lie on any minimum free partial flow tree.

If $\sigma_v \colon [0, r_v] \to Q$ is a loop with flow trees attached at $0 \le t_0 \le t_1 \le \cdots \le t_m \le r_v$, we also introduce a numbering of the components of $[0, r_v] - \{t_1, \ldots, t_m\}$ induced by the flow trees attached as follows. The rightmost interval $(t_m, r_v]$ is numbered by κ_0. The right boundary segment of the strip with slits attached at $\sigma_v(t_m)$ is numbered by κ_0 as well, whereas the left boundary segment of its domain is numbered by κ_{k_m}. We number the boundary segment (t_{m-1}, t_m) and the right boundary segment of the domain of the

flow tree attached at $\sigma_v(t_{m-1})$ by κ_{k_m} as well. The left boundary segment of the flow tree attached at $\sigma_v(t_{m-1})$ is then numbered by $\kappa_{k_{m-1}}$, which determines the numbering of the segment (t_{m-2}, t_{m-1}) as well as the right boundary segment in the flow tree at $\sigma_v(t_{m-2})$, etc. We view the end result of this process as the domain for a loop with flow trees attached with numbering κ that decreases; see Figure 3.

Note next that if $\sigma \colon I^d \to \Omega Q$ is a d–dimensional cube in general position with respect to $\bar{Q}$ (ie transverse to the stratified space of the partial puncture of all minimum free partial flow trees), the set of σ_v with $v \in I^d$ for which a single partial flow tree can be attached is at most $(d-1)$–dimensional. Attaching more partial flow trees, the dimension decreases further, by at least one for each flow tree. We say that the loops in σ with flow trees attached which form a 0–dimensional family are the rigid loops with flow trees in σ. If σ is a cubical simplex in ΩQ and if y is a word of critical points, then we let $\mathcal{T}(\sigma; y)$ denote the space of loops with flow trees in σ, where the critical points at punctures of the flow trees read in order give the word y. The formal dimension of $\mathcal{T}(\sigma; y)$ is then

$$\dim(\mathcal{T}(\sigma; y)) = |y| + \dim(\sigma) + (\ell(y) - 1),$$

where ℓ is the word length, and for chains transverse to the system of parallel copies the formal dimension equals the actual dimension.

Note that if the set of loops in σ with flow trees is transversely cut out, then, by construction of the system of parallel copies, loops with flow trees corresponding to different decreasing numberings are canonically diffeomorphic; see the proof of Theorem 76. We define the map ϕ by counting rigid loops with flow trees in cubical simplices σ:

$$(36) \qquad \phi(\sigma) = \sum_{\dim(\mathcal{T}(\sigma;y))=0} |\mathcal{T}(\sigma; y)| y.$$

Remark 46 We sketch technical aspects of the definition of the map ϕ in (36). In order to get a suitable chain model for ΩQ on which the map ϕ is defined we equip Q with a Riemannian metric and consider piecewise smooth loops in Q. It is shown in [52, Section 17] that the inclusion of piecewise smooth loops into all loops is a homotopy equivalence and we will work with piecewise smooth loops. In [52, Section 16] it is shown that if E is the energy functional on the space of piecewise smooth loops, then the preimage $E^{-1}(b)$ for any noncritical value b is compact and is a deformation retract of the corresponding subset of the space of piecewise geodesic loops, which has a natural structure of a finite-dimensional manifold. Furthermore, as we increase

the energy level the spaces of piecewise geodesics are naturally included in piecewise geodesics with finer subdivision. In this way we get a directed system of inclusions

$$M_{E_0} \subset M_{E_1} \subset \cdots \subset M_{E_j} \subset \cdots$$

such that $E^{-1}(b) \cap M_{E_j} \subset E^{-1}(b) \cap M_{E_k}$ for $b < E_j < E_k$ is a deformation retract. To define the map ϕ we can, for example, use chains of simplices in a sufficiently fine triangulation of the finite-dimensional manifolds M_{E_j} that are suitably transverse to the system of parallel copies $\bar{Q}(\eta)$.

Since the shifting Morse functions do not have any index 1 critical points a partial flow tree has at most $\dim(M)$ punctures. Consider the natural evaluation map on partial flow trees that takes a flow tree to the location of its positive puncture discussed above. The image of this map for partial flow trees not involving the minimum is a stratified space of codimension two and by construction of parallel copies, the corresponding set for partial flow trees defined by distinct boundary numberings lie $\mathcal{O}(\eta^2)$–close to each other. The map ϕ above is defined for chains in M_{E_j} (chains of piecewise smooth loops) with evaluation maps that are transverse to this stratified subset. It is straightforward to see that the chains of simplices in M_{E_j} can be made transverse without destroying transversality for chains at earlier energy levels. This means that we can define the map on the direct limit of chains which is a chain model for the based loop space.

In order to connect this to the path loop fibration we consider a similar map

$$\hat{\phi} \colon C_{-*}(\mathrm{P}Q) \to \Omega\mathrm{CM}_{-*}(Q) \otimes^{\mathrm{t}} \mathrm{CM}_{-*}(Q),$$

where t denotes the canonical twisting cochain of the cobar construction and $\mathrm{P}Q$ the based path space. This map can be described geometrically as follows. The chain complex $C_{-*}(\mathrm{P}Q) \to \Omega\mathrm{CM}_{-*}(Q) \otimes^{\mathrm{t}} \mathrm{CM}_{-*}(Q)$ can be thought of as generated by words of critical points in which the last letter is distinguished and may be the minimum x; in other words, the words are either minimum free, or the last letter (and only the last) is the minimum. The differential counts rigid flow trees as usual and also here only the last letter may be the minimum. To define the map $\hat{\phi}$ we consider chains of paths. As above we attach counit-free partial flow trees to paths in such a chain at interior points and also attach a partial flow tree with last puncture distinguished at the endpoint of the path. Also here, only the distinguished (ie the last puncture in the tree at the endpoint of the path) may be the minimum. The map $\hat{\phi}$ then counts rigid paths with flow trees attached as described.

Lemma 47 *The maps ϕ and $\hat{\phi}$ are chain maps.*

Proof For the chain map property of $\hat{\phi}$, we consider 1–dimensional moduli spaces of chains of paths with flow trees attached as described above, including flow trees at the endpoint. This moduli spaces form oriented 1–manifolds with a natural compactification consisting of the following 0–dimensional configurations:

   (i)   spaces of paths with flow trees attached which is obtained by attaching trees to the boundary of the original chain of paths, and

  (ii)   spaces of paths with flow trees attached where one of the flow trees is broken.

The configuration (i) contributes to the composition $\hat{\phi} \circ d$ and the configuration (ii) to $d \circ \hat{\phi}$. Since the algebraic number of boundary points of a 1–dimensional oriented manifold equals zero we conclude the chain map equation, $d \circ \hat{\phi} = \hat{\phi} \circ d = 0$.

The chain-map property of ϕ is proved applying the same argument to 1–dimensional spaces of chains of loops with flow trees attached. $\qquad\square$

Remark 48 The codimension-one boundary of $\mathcal{T}(\sigma; y)$ corresponds either to the loop or path moving to the boundary of σ or to the breaking of a flow tree at a critical point. Instances when two trees are attached at the same point are naturally interior points of the moduli space where the disks with slits join to a new disk with a slit of width equal to the sum of the widths. See Figure 4.
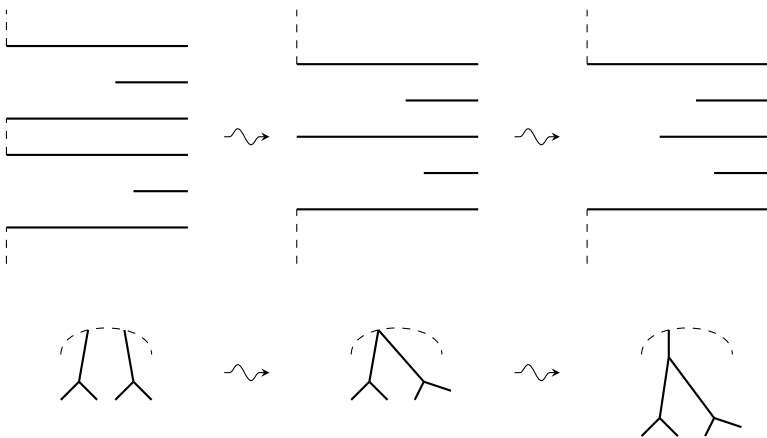

Figure 4: Flow trees attached at the same point are interior points: in the source, top, and in the target, bottom.

With this established we can now prove Adams' result:

Theorem 49 *The flow tree map $\phi: C_{-*}(\Omega Q) \to \Omega \mathrm{CM}_{-*}(Q)$ is a quasi-isomorphism.*

Proof The first observation is that the chain complex $\Omega \mathrm{CM}_{-*}(Q) \otimes^{\mathrm{t}} \mathrm{CM}_{-*}(Q)$ is acyclic. To see that, note that for each critical point y there are exactly two flow trees with positive puncture y and two negative punctures, one at the counit x and one at y; see Lemma 38. Exactly one of these — the tree in which the negative punctures are in the order y followed by x — contributes to the differential on $\Omega \mathrm{CM}_{-*}(Q) \otimes^{\mathrm{t}} \mathrm{CM}_{-*}(Q)$.

Add a constant to the Morse function f used to build the parallel copies so that the minimum x lies at level 0 and all other critical points at positive levels. We then filter by action of the parallel copies $\bar{f} = \{f_j\}$; more precisely, we associate to a word $y_1 \cdots y_m$ of critical points the action $\sum_j (f_j - f_{j-1})(y_j)$. Then, by definition of a flow tree for the flow-tree map, the differential does not increase action. (To see this, recall that the flow-tree map uses flow trees with only one positive puncture which decrease action since the value of a Morse function decreases along negative gradient flow; see [22, Lemma 2.3, equations (2-2)–(2-3)] for the calculation.) Since all flow trees except those involving the counit decrease action, we find that the differential on the associated graded complex acts only on the last (distinguished) letter in the word and it acts there as $y \mapsto yx$ if $y \neq x$ and $x \mapsto 0$. Since this is an isomorphism from words not ending with the counit x to those which end with x, the desired acyclicity follows. Clearly, $C_{-*}(PQ)$ is also acyclic.

Consider next the stratification of Q induced by the stable manifolds of the Morse function f and the corresponding filtration on $C_{-*}(PQ)$ induced by evaluation at the endpoint. The corresponding filtration on $\Omega \mathrm{CM}_{-*}(Q) \otimes^{\mathrm{t}} \mathrm{CM}_{-*}(Q)$ is filtration by degree of the distinguished (the last) critical point and the map $\hat{\phi}$ respects these filtrations.

The corresponding E_1–terms with induced maps are

$$C_{-*}(Q; H_{-*}(\Omega Q)) \to H_*(\Omega \mathrm{CM}_{-*}(Q)) \otimes \mathrm{CM}_{-*}(Q).$$

To see this, consider first the left-hand side. The associated graded complex can be represented by chains of paths in Q with endpoint in a Morse stratum of fixed dimension, divided by such chains of paths with endpoints in Morse strata of lower dimension, and the differential in the associated graded complex acts as the singular differential on such paths. The resulting homology is then the homology of the chain with endpoints in cells of fixed dimension modulo their boundary, which gives $C_{-*}(Q; H_*(\Omega Q))$.

Consider the right-hand side: here the associated graded complex can thought of as the direct sum of complexes with fixed distinguished last critical point. The differential attaches minimum-free flow trees at all nondistinguished critical points. The resulting homology is clearly $H_*(\Omega CM_{-*}(Q)) \otimes CM_{-*}(Q)$.

The E_2–terms are then

$$(37) \qquad H_{-*}(Q; H_{-*}(\Omega Q)) \to H_{-*}\big(Q; H_*(\Omega CM_{-*}(Q))\big).$$

To see this, on the left-hand side the part of the boundary operator on a chain of loops that remains after passing to the E_1–level corresponds to the endpoint going to the boundary of the Morse chain in Q. On the right-hand side, it remains to add flow trees at the last distinguished critical points. The homologies of these differentials are then clearly as stated. Equation (37) together with Zeeman's comparison theorem [51, Section 3.3] then establishes the result. □

Remark 50 If $\dim(Q) = 4$ then the Morse function may have critical points of index one. In this case we use stabilization as follows. Multiply Q by $\mathbb{R}^N$ for any $N \geq 2$, and consider the function $F(q, x) = f(q) + x^2$. Then F has the same critical points as f and $-\nabla F$ is inward pointing at infinity. In $Q \times \mathbb{R}^N$ there is room to cancel 1–handles and the above applies. In this case we define $CM_{-*}(Q)$ to be $CM_{-*}(Q \times \mathbb{R}^N)$, which is a 1–reduced version of the original complex. Noting that $C_{-*}(Q \times \mathbb{R}^N)$ and $C_{-*}(Q)$ are canonically isomorphic, the result follows also in this case.

We next show that $CE_\parallel^*(\Lambda)$ and $CE^*(\Lambda)$ are isomorphic if Λ is simply connected. This is a more or less a direct consequence of the description of rigid disks on a Legendrian with parallel copies in Lemma 35 and the isomorphism in Theorem 49. Since components in Λ_- are not affected by this choice of $CE_\parallel^*$ versus CE^*, we disregard them and assume that $\Lambda = \Lambda_+$ in what follows.

Recall the definition of $CE^*(\Lambda)$ given in Remark 24, which is generated by chains $C_{-*}((\Omega_p\Lambda)^{\times(i+1)})$ in the product of the based loop space of Λ with factors separated by Reeb chords. Here the differential of a chain is just the usual differential of the chain whereas the differential of a Reeb chord is a sum over all moduli spaces of disks with one positive puncture at the chord and any number of negative punctures. Such a moduli space carries a fundamental chain and the contribution to the differential is an alternating word of chains in the based loop space corresponding to the boundary arcs of the disk carried by the fundamental chain and Reeb chords at the negative puncture,

$$dc_0 = \sum_{c'} [\mathcal{M}^{\mathrm{sy}}(c)],$$

where $c = c_0 c'$ and c' is a word of Reeb chords. Here we use the diagonal in the product of loop spaces; see Remark 24.

We next consider a system of parallel copies $\overline{\Lambda}(\eta) = \{\Lambda_j(\eta)\}_{j=0}^{\infty}$ defined by a system of positive Morse functions (Section 3.3.2), where $\Lambda_0 = \Lambda$. Recall that the generators of the algebra $\mathrm{CE}_{\parallel}^*(\Lambda)$ are the Reeb chords connecting Λ_0 to Λ_1, and that these can be long, corresponding to Reeb chords of Λ, and short, corresponding to critical points of f_1 except for the minimum. The differential counts rigid disks with one positive puncture in $\mathcal{M}_{\parallel}^{\mathrm{sy}}(\boldsymbol{b}; \kappa)$ where κ is a decreasing boundary numbering, $\boldsymbol{b} = b_0 \boldsymbol{b}'$.

We next consider the map

$$\phi \colon \mathrm{CE}^*(\Lambda) \to \mathrm{CE}_{\parallel}^*(\Lambda)$$

which takes every Reeb chord to itself and takes a chain σ in the based loop space to $\phi(\sigma)$, where ϕ is as in (36) and where we identify the critical points of f_1 with the corresponding Reeb chords connecting Λ_0 to Λ_1.

Theorem 51 *The map ϕ is a DG–algebra map and if Λ^+ is simply connected then ϕ is a quasi-isomorphism.*

Proof The fact that ϕ is a chain map follows from Lemma 35. Filter the algebras by action of Reeb chords on the left-hand side and actions of long Reeb chords on the right-hand side. The map respects this filtration. The E_2–pages are obtained by acting by the differential on the chains on the based loop space only on the left-hand side and on Morse chords only on the right-hand side. The result is words of Reeb chords separated by homology classes in the based loop space and by homology classes in the (reduced) bar construction on the Morse coalgebra on the left- and right-hand sides, respectively. On these E_2–pages the map ϕ induces an isomorphism by Theorem 49. Since the sum of actions of the Reeb chords at the negative end is bounded by that at the positive end, the spectral sequences converge. The theorem follows. $\qquad\square$

Remark 52 The isomorphism in Theorem 51 is compatible with the stabilization of Remark 50. To see this we multiply the ambient contact manifold Y with contact form α by $T^*\mathbb{R}^N$ and consider $\Lambda \times \mathbb{R}^N \subset Y \times T^*\mathbb{R}^N$ with contact form $\theta = (\alpha - y\,dx)$. The Reeb chords of $\Lambda \times \mathbb{R}^N$ then come in $\mathbb{R}^N$–families, one for each Reeb chord of Λ. Consider the contact form $e^{x^2}\theta$ and note that with respect to this contact form the Reeb chords of $\Lambda \times \mathbb{R}^N$ are in natural one-to-one correspondence with those of Λ and there is a canonical isomorphism between $\mathrm{CE}^*(\Lambda)$ and $\mathrm{CE}^*(\Lambda \times \mathbb{R}^N)$. In fact the disks in the differential are canonically identified. It follows in particular that Theorem 51 holds also if $\dim(Q) \leq 4$.

4 Lagrangian (co)algebra

As before, X is a Liouville manifold with $c_1(X) = 0$ and L is an exact relatively spin Lagrangian in X with vanishing Maslov class and ideal boundary given by the Legendrian Λ.

We will associate several chain-level structures to L. To begin with, let us first assume that L is an embedded Lagrangian. Since L has boundary, in classical topology, one can consider either $C^*(L)$ or $C^*(L, \partial L)$. In our case, these two choices are reflected in the choices of $+$ or $-$ decorations on L, respectively. More generally, let L^v, $v \in \Gamma$, be the (irreducible) components of L. As with the Legendrian submanifolds in Section 3.2, we assume these components of L are decorated by signs and we write $L = L^+ \cup L^-$ for the corresponding decomposition. Let $F \colon L \to \mathbb{R}$ be a Morse function with prescribed behavior at infinity (depending on the $+$ or $-$ decoration) as explained in Section 3.3. We use this to construct a system of parallel copies $\bar{L} = \{L_j\}_{j=1}^\infty$, as in Section 3.3, shifted at infinity along the Reeb flow either in the positive or negative direction on L^+ and L^-, respectively.

Now, using the parallel copies, $\{L_j\}_{j=1}^\infty$, we define a graded quiver $\mathcal{Q}_L$ as follows. The parallel copies $\{L_j\}_{j=1}^\infty$ give rise to following sets, for fixed $i_1 < i_2$ positive integers, and $v, w \in \Gamma$ with $v \neq w$:

- Intersection points $L_{i_1}^v \cap L_{i_2}^v$ in bijection with the union of critical points of $F|_{L_v}$. These critical points may depend on the $+$ or $-$ decoration on L_v, one can for example turn a $-$ decorated component L^v into a $+$ decorated one, by introducing critical points corresponding to the topology of ∂L_v; see Figure 5.

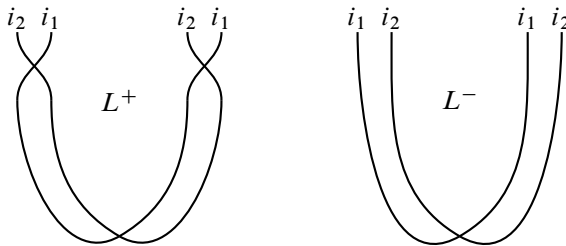- Intersection points $L_{i_1}^v \cap L_{i_2}^w$ in bijection with $L^v \cap L^w$.

Figure 5: Difference between $+$ and $-$ generators for $i_1 < i_2$. Both the left- and the right-hand sides depict shifts corresponding to Morse functions with a maximum. One of the intersection points in L_+ is the minimum and corresponds to the unit for the Floer cohomology product.
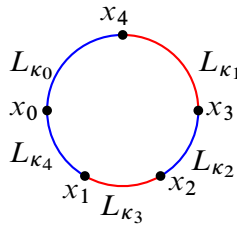
Figure 6: An example of a disk with labelings. The blue labeled Lagrangians are perturbed with $+$ perturbations and red labeled Lagrangians are perturbed with $-$ perturbation.

Furthermore, by the construction in Section 3.3, there are canonical bijections between the above sets associated to any pairs (i_1, i_2) and (i_1', i_2') with $i_1 < i_2$ and $i_1' < i_2'$. So, fix a pair (i_1, i_2) such that $i_1 < i_2$, and define a graded quiver $\mathcal{Q}_L$ with vertex set Γ and with an arrow connecting v to w (possibly equal to v) for each element of the above sets. Let $\mathcal{I}$ denote the set of arrows.

Alternatively, one can describe the generators as the set of intersection points in $L_0 \cap L_1$, between the original L and the first shifted copy.

Let $\mathrm{CF}^*(L)$ be the graded $\boldsymbol{k}$–bimodule generated by $\mathcal{I}$. Thus, there is one generator x_{vw} in degree $|x_{vw}|$ for each arrow in $\mathcal{Q}_L$ from v to w. We endow $\mathrm{CF}^*(L)$ with the structure of an A_∞–algebra. Let x_0 be an intersection point generator and let $\boldsymbol{x}' = x_i \cdots x_1$ be a word of intersection points. Consider the disk D_{i+1} with $i+1$ boundary punctures and with a decreasing numbering of its boundary arcs κ. Let $\boldsymbol{x} = x_0 x_i \cdots x_1$. Consider the moduli space $\mathcal{M}^{\mathrm{fi}}(\boldsymbol{x}; \kappa)$; see Appendix A for notation. Define the operations $\mathfrak{m}_i$ by

$$\mathfrak{m}_i(\boldsymbol{x}') = \sum_{\substack{\ell(\boldsymbol{x}')=i \\ |x_0|=|\boldsymbol{x}'|+(2-i)}} |\mathcal{M}^{\mathrm{fi}}(\boldsymbol{x}; \kappa)| x_0,$$

where $\ell(\boldsymbol{y})$ denotes the word length of $\boldsymbol{y}$ and $|\mathcal{M}^{\mathrm{fi}}(\boldsymbol{x}; \kappa)|$ denotes the algebraic number of points in the oriented 0–dimensional manifold.

Lemma 53 *The operations $\mathfrak{m}_i$ satisfy the A_∞–algebra relations and are independent of the decreasing boundary labeling κ.*

Proof This follows by the usual argument: after noting that the decreasing boundary numbering ensures that there is no boundary bubbling, one observes that the terms in the A_∞–algebra relations count the ends of a 1–dimensional oriented compact manifold by Theorems 74 and 75. The operations compose because of Theorem 76. □

We call $\mathrm{CF}^*(L)$ the Lagrangian Floer cohomology algebra of L. Let u_v denote the generator corresponding to the minimum on the component $L_v \subset L^+$. If L_v is simply connected, then, by Lemma 38, u_v is a strict idempotent. We write $\mathbf{k}_- \oplus \mathrm{CF}^*(L)$ for the augmented algebra where we adjoined an idempotent e_w for each component $L_w \subset L^-$. (On these components the shifting function is decreasing at infinity and has a maximum rather than a minimum in the compact part.) This is a connected algebra over $\mathbf{k}$.

Remark 54 The two different choices of perturbations at infinity corresponding to $+$ and $-$ are the two extremal constructions, where one pushes the copies either always in the positive direction or always in the negative direction. One can also choose perturbations at infinity to depend on the topology of the manifold at infinity; see, for example, Section 4 of [1]. All our constructions should extend meaningfully to this more general setting, but we have not pursued this direction in this paper.

We next consider various linear duals of $\mathrm{CF}^*(L)$ and associated algebraic objects. The simplest case occurs when $\mathrm{CF}^*(L)$ is simply connected. In this case the linear dual $\mathrm{CF}_*(L)$ is a coalgebra with operations Δ_i dual to $\mathfrak{m}_i$, and as before we can adjoin $\mathbf{k}_-$ so that $\mathbf{k}_- \oplus \mathrm{CF}_*(L)$ is coaugmented over $\mathbf{k}$. Then we define the *Adams–Floer DG–algebra*

$$\Omega(\mathbf{k}_- \oplus \mathrm{CF}_*(L))$$

by applying the cobar construction.

In the non-simply-connected case, we replace this object by the *completed Adams–Floer DG–algebra*

$$\left(\mathrm{B}(\mathbf{k}_- \oplus \mathrm{CF}^*(L))\right)^{\#}.$$

Example 55 Let L be the standard Lagrangian D^n filling of the standard Legendrian unknot $\Lambda \subset S^{2n-1}$. The Floer cohomology can be computed as

$$\mathrm{CF}^*(L) = \begin{cases} \mathbb{K}x & \text{with } |x| = 0 \text{ if } L \text{ is decorated with } +, \\ \mathbb{K}c & \text{with } |c| = n \text{ if } L \text{ is decorated with } -. \end{cases}$$

Alternatively, if we want compatibility with the inclusion $C^*(D^n, \partial D^n) \to C^*(D^n)$, it can be computed as

$$\mathrm{CF}^*(L) = \begin{cases} \mathbb{K}c \oplus \mathbb{K}y \oplus \mathbb{K}x & \text{with } |c| = n, |y| = n-1, |x| = 0 \text{ and } dy = c \\ & \qquad\qquad\qquad\qquad\qquad \text{if } L \text{ is decorated with } +, \\ \mathbb{K}c & \text{with } |c| = n \text{ if } L \text{ is decorated with } -. \end{cases}$$

In Section B.1, we introduce a model for wrapped Floer cohomology without a Hamiltonian term and prove it is quasi-isomorphic to the usual wrapped Floer cohomology.

We refer there for details and give only a short description here. The chain complex underlying $\mathrm{CW}^*(L)$ is the following. Let $L = L_0$ and shift L off itself to L_1 by a Morse function that is positive at infinity (as in the definition of parallel copies when L is decorated $+$). The generators of $\mathrm{CW}^*(L)$ are then Reeb chords connecting L_1 to L_0 and intersection points in $L_0 \cap L_1$.

There is an A_∞–functor, often called the *acceleration functor*,

$$\mathrm{CF}^*(L) \to \mathrm{CW}^*(L).$$

If L is decorated $+$, it can be shown that this functor is unital.

5 Maps relating Legendrian and Lagrangian (co)algebras

We continue with our usual set-up, where X is a Liouville manifold with $c_1(X) = 0$ and L is an exact Lagrangian in X with vanishing Maslov class and ideal boundary given by the Legendrian Λ. Let Γ be the set of embedded components of L subdivided into $\Gamma^+ \cup \Gamma^-$. Let Θ be the set of components of Λ with induced subdivision $\Theta^+ \cup \Theta^-$.

In this section we will define twisting cochains and associated DG–algebra maps relating the parallel copies version $\mathrm{CE}^*_\|(\Lambda)$ and the Floer cohomology $\mathrm{CF}^*(L)$. Since L is an exact filling, we have an augmentation $\epsilon_L \colon \mathrm{CE}^*_\|(\Lambda) \to \boldsymbol{k}$. As in Section 3.2 we use this augmentation throughout to change coordinates in such a way that $\Delta_0 = 0$.

As explained in Section 3.4, the definition of $\mathrm{CE}^*_\|$ differs depending on whether or not the components of Λ in Θ^+ are simply connected. We will start in the simply connected case and turn later to the non-simply-connected case, using the definitions in Section 3.4.2.

Assume thus that all components of Λ in Θ^+ are simply connected. As usual, let $\boldsymbol{k}_- \oplus \mathrm{LC}^\|_*(\Lambda)$ denote the coalgebra corresponding to $\mathrm{CE}^*_\|(\Lambda)$ augmented by the Lagrangian filling, with counits e_v adjoined to all components Λ_v in Θ^-. As Λ^+ has simply connected components, by Lemma 38, this is a counital coalgebra with counit

$$\sum_{v \in \Theta^+} x_v + \sum_{v \in \Theta^-} e_v.$$

Let $\eta \colon \boldsymbol{k} \to \boldsymbol{k}_- \oplus \mathrm{LC}^\|_*(\Lambda)$ denote the coaugmentation

$$(38) \qquad \eta(e_v) = \begin{cases} x_v & \text{if } v \in \Theta^+, \\ e_v & \text{if } v \in \Theta^-, \end{cases}$$

(see (33)), so that $\mathrm{CE}^*_\| = \Omega(\boldsymbol{k}_- \oplus \mathrm{LC}^\|_*)$. This means that x_v is traded for e_v for $v \in \Theta^+$.

Consider the Floer cohomology A_∞–algebra $\mathrm{CF}^*(L)$. If all components of L^+ are simply connected there exists a strict idempotent $u_v \in \mathrm{CF}^*(L)$ for each $v \in \Gamma^+$ corresponding to the minimum of the shifting Morse function, and we make $\mathrm{CF}^*(L)$ unital by adding an idempotent e_w for each $w \in \Gamma^-$. We write the strictly unital algebra as $\boldsymbol{k}_- \oplus \mathrm{CF}^*(L)$. Let $\epsilon : \boldsymbol{k}_- \oplus \mathrm{CF}^*(L) \to \boldsymbol{k}$ be the augmentation that maps u_v to e_v for $v \in \Gamma^+$ and e_w to e_w for $w \in \Gamma^-$. Consider the dual of the bar construction,

$$\tag{39} \mathscr{A} = \big(\mathrm{B}(\boldsymbol{k}_- \oplus \mathrm{CF}^*(L))\big)^{\#},$$

or in other words the completed Adams–Floer DG–algebra. In what follows we will represent $\mathscr{A}$ as a quotient in way that generalizes to the non-simply-connected case in analogy with the construction in Section 3.4.2. In the non-simply-connected case we introduce strict idempotents by hand as follows.

Consider adding extra idempotents e_v for $v \in \Gamma^+$ to $\boldsymbol{k}_- \oplus \mathrm{CF}^*(L)$. This gives $\boldsymbol{k} \oplus \mathrm{CF}^*(L)$ and we equip it with the trivial augmentation ϵ' which is the projection to $\boldsymbol{k}$. Let

$$\mathscr{A}' = \big(\mathrm{B}(\boldsymbol{k} \oplus \mathrm{CF}^*(L))\big)^{\#},$$

and let $\mathscr{I}$ denote the subalgebra of $\mathscr{A}'$ given by the space of functionals which vanish on monomials not containing u_v for some $v \in \Gamma^+$. Let $\rho : \mathscr{A}' \to \mathscr{A}$ denote the restriction map induced by the inclusion $\boldsymbol{k}_- \oplus \mathrm{CF}^*(L) \to \boldsymbol{k} \oplus \mathrm{CF}^*(L)$.

Lemma 56 *The subalgebra $\mathscr{I}$ is closed under the differential. In the simply connected case, the map ρ is a chain map with kernel $\mathscr{I}$ and consequently $\mathscr{A}$ is quasi-isomorphic to $\mathscr{A}'/\mathscr{I}$.*

Proof Similar to Lemma 42. To see that $\mathscr{I}$ is closed under the differential we check that if u_v is an output of the differential, then u_v is also an input. Here the output is the positive puncture of the holomorphic disk, or equivalently flow tree; see Lemma 33. Since there is no negative-gradient flow line that starts at a minimum, any gradient flow tree with positive puncture at a minimum must also have a negative puncture at that minimum. This shows that if u_v is an output then it is an input as well.

In the simply connected case, monomials not containing u_v come from $\mathrm{B}(\boldsymbol{k}_- \oplus \mathrm{CF}^*)$, therefore the kernel of ρ is contained in $\mathscr{I}$ and, conversely, any element in $\mathscr{I}$ restricts to zero on $\mathrm{B}(\boldsymbol{k}_- \oplus \mathrm{CF}^*)$. Thus, in this case $\mathscr{I}$ is the kernel of ρ. $\qquad\square$

In the general case we define $\mathscr{A} = \mathscr{A}'/\mathscr{I}$. Lemma 56 shows that in the simply connected case this definition agrees with the alternative definition of $\mathscr{A}$ given in (39).
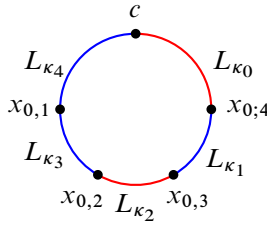
Figure 7: An example of a disk contributing to t'. The blue labeled Lagrangians are perturbed with $+$ perturbations and red labeled Lagrangians are perturbed with $-$ perturbation.

Remark 57 The somewhat artificial construction of $\mathscr{A}$ as $\mathscr{A} = \mathscr{A}'/\mathscr{I}$ is used to adapt the bar construction to a not necessarily strictly unital algebra.

We next define a map t' on generators of $\mathrm{CE}^*_\parallel(\Lambda)$ which then gives a map

$$\mathsf{t}' \colon \mathrm{LC}^\parallel_*(\Lambda) \to \mathscr{A}'$$

in the simply connected case, and in that case it will induce a twisting cochain

$$\mathsf{t} \colon \boldsymbol{k}_- \oplus \mathrm{LC}^\parallel_*(\Lambda) \to \mathscr{A}.$$

The map t' is defined by the following curve count for generators of $\mathrm{LC}^\parallel_*(\Lambda)$. Fix systems of parallel copies $\overline{L}$ of L. Recall that the components labeled with a $+$ sign are shifted by a positive Morse function and the components labeled with a $-$ sign are shifted by a negative Morse function.

Let c be a Reeb chord of $\overline{\Lambda}$ and let $\boldsymbol{x}_0 = x_{0;1} \cdots x_{0;j}$, $j > 0$, be a nonempty word of intersection points of $\overline{L}$. Let

$$\boldsymbol{c} = c\boldsymbol{x}_0$$

and define

(40) $$\mathsf{t}'(c) = \sum_{|\boldsymbol{x}_0| = |c| + (1-j)} |\mathcal{M}^{\mathrm{fi}}(\boldsymbol{c})|\boldsymbol{x}_0,$$

where we interpret $\boldsymbol{x}_0$ as an element in $\mathscr{A}'$.

Remark 58 In the non-simply-connected case we use the same formula to define $\mathsf{t}'(c)$ and note that the sum in the definition may be infinite.
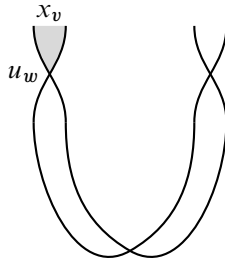
We have the following:

Figure 8: Minimum x_v is sent to the minimum u_w.

Proposition 59 *If $v \in \Theta^+$ is such that Λ_v is a boundary component of L_w for $w \in \Gamma^+$, then*

$$\mathfrak{t}'(x_v) = u'_w,$$

where u'_w is the dual of the minimum u_w. Furthermore, $\mathfrak{t}'$ satisfies the equation of a twisting cochain.

Proof For the first property we need to understand rigid holomorphic disks with positive puncture at the Reeb chord x_v. By small action such a holomorphic disk must lie in a neighborhood of L_w and is hence given by Morse flow trees. There is only one flow line emanating from the minimum in Λ_v and the flow line generically ends at the minimum of the shifting of L_w; see Figure 8. The first equation follows.

To see the twisting cochain equation, we need to check that

$$\mathfrak{m}_1 \circ \mathfrak{t}' - \mathfrak{t}' \circ \Delta_1 + \sum_{d \geq 2} (-1)^d \mathfrak{m}_2^{(d)} \circ \mathfrak{t}'^{\otimes_k d} \circ \Delta_d = 0,$$

where $\mathfrak{m}_2^{(2)} := \mathfrak{m}_2$ and $\mathfrak{m}_2^{(i)} := \mathfrak{m}_2 \circ (\mathrm{Id}_{\mathscr{A}} \otimes_k \mathfrak{m}_2^{(d-1)})$. To this end, we consider the boundary of the 1–dimensional moduli space $\mathcal{M}^{\mathrm{fl}}(c_0 x)$. By Theorem 75 this corresponds to two-level curves which by Theorems 74 and 76 form the boundary of an oriented compact 1–manifold. $\square$

Proposition 59 shows that $\mathfrak{t}'$ maps the submodule generated by x_v for $v \in \Theta^+$ into $\mathscr{I} \subset \mathscr{A}'$. Hence, by letting $\mathfrak{t}(e_v) = 0$, $\mathfrak{t}'$ induces a map

$$\mathfrak{t} \colon \boldsymbol{k}_- \oplus \mathrm{LC}^{\parallel}_*(\Lambda) \to \mathscr{A}'/\mathscr{I} = \mathscr{A}.$$

Note that, if $\eta \colon \boldsymbol{k} \to \boldsymbol{k}_- \oplus \mathrm{LC}^{\parallel}_*(\Lambda)$ is the coaugmentation in (38) and $\epsilon \colon \mathscr{A} \to \boldsymbol{k}$ is the trivial augmentation, then $\epsilon \circ \mathfrak{t} = \mathfrak{t} \circ \eta = 0$.

Corollary 60 *The map $\mathfrak{t}$ is a twisting cochain.*

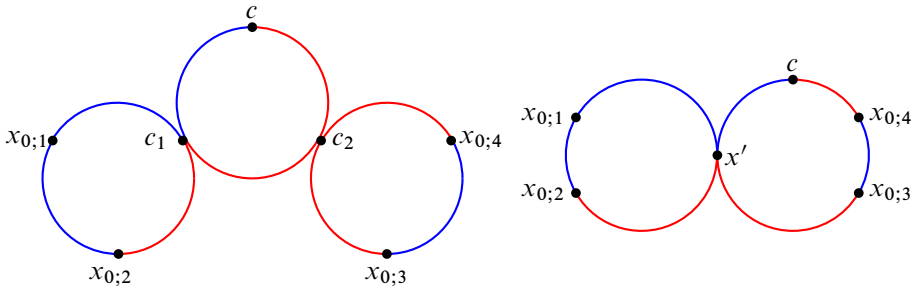**Proof** Since $\mathfrak{t}'$ satisfies the twisting cochain equation, so does $\mathfrak{t}$. $\square$

Figure 9: Left: a two-story disk contributing to the term $\mathfrak{m}_2 \circ \mathfrak{t}^{\otimes 2} \circ \Delta_2$ applied to c. There are similar disks in the compactification of the 1–dimensional moduli space, with 2 replaced by n. Right: a two-story disk which contributes to the term $\mathfrak{m}_1 \circ \mathfrak{t}$ applied to c.

This twisting cochain is always defined, and determines a map

$$(41) \qquad\qquad \mathrm{CE}^*_\parallel(\Lambda) \to \mathscr{A}.$$

We next consider the question whether $\mathfrak{t} \in \mathrm{Kos}(\mathrm{LC}^\parallel_*(\Lambda), \mathscr{A})$. The following theorem gives a sufficient condition for $\mathfrak{t}$ to be Koszul:

Theorem 61 *Suppose that* $\mathrm{LC}^\parallel_*(\Lambda)$ *is a locally finite, simply connected $\boldsymbol{k}$–bimodule. Suppose further that* $\mathrm{HW}^*(L) = 0$. *Then $\mathscr{A}$ is quasi-isomorphic to $\Omega(\boldsymbol{k}_- \oplus \mathrm{CF}_*(L))$ and* $\mathfrak{t}: \mathrm{LC}^\parallel_*(\Lambda) \to \mathscr{A}$ *is a Koszul twisted cochain. In other words, the induced DG–algebra map*

$$\mathrm{CE}^*_\parallel(\Lambda) \to \mathscr{A} \approx \Omega(\boldsymbol{k}_- \oplus \mathrm{CF}_*(L))$$

is a quasi-isomorphism.

Corollary 62 *In the situation of Theorem 61, suppose that L is connected and decorated by $-$ and that $\partial L = \Lambda$ is diffeomorphic to a sphere S^{n-1}. Writing $\overline{L} = L \cup_\partial D^n$, there exists a quasi-isomorphism of DG–algebras*

$$\mathrm{CE}^*(\Lambda) \to C_{-*}(\Omega \overline{L}),$$

where $\Omega \overline{L}$ is the based loop space of $\overline{L}$.

Proof We first note that there exists a quasi-isomorphism $\boldsymbol{k} \oplus \mathrm{CF}^*(L) \to C^*(\overline{L})$ since L is an exact Lagrangian; this is well known and can be deduced eg from Lemma 33. We next use Theorem 61 and Adams' cobar equivalence [4]

$$\Omega(C_*(\overline{L})) \simeq C_{-*}(\Omega \overline{L}),$$

which holds for the simply connected topological space $\overline{L}$. $\qquad\square$

Theorem 61 is obtained as a corollary of Theorem 18 and the following result. For the notion of wrapped Floer cohomology, see [3] and Section B.1.

Theorem 63 *Suppose that* $HW^*(L) = 0$. *Then there exists a quasi-isomorphism of augmented* A_∞–*algebras*

$$\mathfrak{e} \colon CF^*(L) \to LA_\parallel^*(\Lambda)$$

such that

$$\mathfrak{e}(u_v) = \sum_{\Lambda_w \subset \partial L_v} x_w \quad \text{for } v \in \Gamma^+.$$

Note that $HW^*(L) = 0$ if X is subcritical [3], or more generally a flexible Weinstein manifold. (A Weinstein manifold is flexible if the attaching spheres of all top handles are loose and hence have trivial Chekanov–Eliashberg algebras [18]; see Section B.2 for the vanishing of wrapped Floer cohomology.)

Note also that if there is no bijection between connected components of L and connected components of Λ, then we have to work over the semisimple ring of idempotents determined by the Lagrangian.

Proof We construct an A_∞–map $\mathfrak{e} = (\mathfrak{e}_i)_{i\geq 1}$, where the maps

$$\mathfrak{e}_i \colon (CF^*(L))^{\otimes_\kappa i} \to LA_\parallel^*(\Lambda)$$

are constructed by dualizing the components of the map $\mathfrak{t}'$. More explicitly, given c_0 and $\boldsymbol{x}_0 = x_{0;n}, \dots, x_{0;1}$, write $\boldsymbol{c} = c_0 \boldsymbol{x}_0$ as in (40), and define

$$\mathfrak{e}_i(\boldsymbol{x}_0) = \sum_{c_0 \in \mathcal{R}} |\mathcal{M}^{\mathrm{fi}}(\boldsymbol{c})| c_0.$$

The proof that $(\mathfrak{e}_i)_{i\geq i}$ is an A_∞–map follows as in the proof of Proposition 59.

Now, we need to check that $\mathfrak{e}_1$ is a quasi-isomorphism. We point out that $\mathfrak{e}_1$ concerns only strips and is defined using only two parallel copies. In the case that $L = L^-$, this is a consequence of the exact sequence for wrapped Floer cohomology induced by the subdivision of the complex into high- and low-energy generators,

$$0 \to CW_0^*(L) \to CW^*(L) \to CW_+^*(L) \to 0,$$

as follows. In terms of the version of wrapped Floer cohomology presented in Section B.1.1, the low-energy subcomplex $CW_0^*(L)$ is generated by Lagrangian intersection points in $L_0 \cap L_1$, where $L = L_0$ and L_1 is the first parallel copy of L, shifted by a

positive Morse function f that increases at the end; see Section 3.3. The differential on $\mathrm{CW}_0^*(L)$ counts holomorphic strips which are incoming along L_1 and outgoing along L_0 at the output puncture. Similarly, the high-energy quotient $\mathrm{CW}_+^*(L)$ is generated by Reeb chords connecting Λ_1 to Λ_0, and the differential counts holomorphic strips interpolating between such Reeb chords. Since $\mathrm{CW}^*(L)$ is acyclic it follows that the connecting homomorphism $\mathrm{HW}_+^*(L) \to \mathrm{HW}_0^{*+1}(L)$ is an isomorphism. In order to connect this to $\mathrm{CF}^*(L)$ and $\mathrm{LA}_\parallel^*(\Lambda)$, renumber the parallel copies so that L_1 now lies in the negative Reeb direction of L_0 at infinity and the shifting function f is replaced by $-f$. Then note that since L is labeled by $-$ it holds that:

- The linear dual of $\mathrm{CW}_+^{*-1}(L)$ is canonically identified with $\mathrm{LA}_\parallel^*(\Lambda)$ as a chain complex. Note that $\mathrm{CW}_+^{*-1}(L)$ also has an A_∞–coalgebra structure as defined in [30] and this should dualize to the A_∞–algebra structure on LA^*, but we do not need that here.

- The linear dual of $\mathrm{CW}_0^*(L)$ is canonically identified with $\mathrm{CF}^*(L)$ as a chain complex.

- The linear dual of the connecting homomorphism can be canonically identified with the map $\mathfrak{e}_1 : \boldsymbol{k}_- \oplus \mathrm{CF}^*(L) \to \boldsymbol{k}_- \oplus \mathrm{LA}_\parallel^*(\Lambda)$ on critical points which counts strips with an input puncture at $L_0 \cap L_1$ and output puncture at a Reeb chord and is the canonical map on $\boldsymbol{k}_-$.

Since the connecting homomorphism is an isomorphism so is its linear dual. (The argument here is originally due to Seidel; compare [29, Theorem 7.2].)

In the case that $L^+ \neq \varnothing$, the argument just given applies after a certain deformation, which we describe next. For components in L^+, we define $\mathrm{CF}^*(L)$ and $\mathrm{LA}_\parallel^*(\Lambda)$ via parallel copies shifted in the positive Reeb direction at infinity. To connect to the previous case we consider a Lagrangian cobordism of two cylinders: $\mathbb{R} \times \Lambda_0$, which is constant, and $\mathbb{R} \times \Lambda_1$, which is the trace of an isotopy pushing Λ_1 across Λ_0 in the negative Reeb direction. This can be arranged so that the intersection points of the two cylinders are in natural one-to-one correspondence between the short Reeb chords between Λ_0 and Λ_1. Furthermore, there is exactly one transverse holomorphic strip connecting each intersection point to the corresponding Reeb chord at the negative end of the cobordism. To see this, note that $\mathbb{R} \times \Lambda_1$ can be obtained from a graphical Lagrangian in a cotangent neighborhood of $\mathbb{R} \times \Lambda_0$ that intersects the zero section $\mathbb{R} \times \Lambda_0$ cleanly along $\Lambda_0 \times \{0\}$. More precisely, it is the graph of the pullback of a 1–form on the $\mathbb{R}$–factor with a single transverse zero at $0 \in \mathbb{R}$. The corresponding

Lagrangians are then a product of the zero section in $T^*\Lambda_0$ and a 1–dimensional Lagrangian in $T^*\mathbb{R}$, and of the zero section and a curve that intersects the zero section once transversely at 0. The transverse holomorphic strips after perturbation by a Morse function on Λ_0 are products of constants at the critical points of the Morse function and the obvious strips for the 1–dimensional Lagrangian. Transversality is a consequence of transversality for the components of the product. We call such curves *basic strips*.

Adding these cylinders to L^+ we get a 1–parameter family of pairs of Lagrangian submanifolds $\widehat{L}_0^\rho$ and $\widehat{L}_1^\rho$, where $\rho > 0$ is a gluing parameter that measures the length of the trivial cobordism between L^+ and the added cylinders. The wrapped Floer cohomology $\mathrm{CW}^*(\widehat{L}_0^\rho, \widehat{L}_1^\rho)$ between Lagrangians $\widehat{L}_0^\rho$ and $\widehat{L}_1^\rho$ vanishes: it is isomorphic to the wrapped Floer cohomology $\mathrm{CW}^*(L)$ by Hamiltonian deformation invariance. Write $\mathrm{CW}^*(\widehat{L}^\rho) := \mathrm{CW}^*(L_0^\rho, L_1^\rho)$. This complex is then acyclic and is generated by the set of long Reeb chords C^+ from L_0 to L_1, the set of intersection points I between the cylinders, and intersection points P in L. Let C^- denote the short Reeb chords connecting L_0 to L_1 and recall the natural one-to-one correspondence $C^- \approx I$ above. Let $\rho > 0$. We claim that the following sets are in natural one-to-one correspondence for all sufficiently large ρ:

(i) Rigid strips of $\widehat{L}^\rho$ with input puncture at $c \in C^+$ and output puncture at $p \in P$ and rigid strips of L with input puncture at $c \in C^+$ and output puncture at $p \in P$.

(ii) Rigid strips of $\widehat{L}^\rho$ with input puncture at $c \in C^+$ and output puncture at $q \in I$ and rigid-up-to-translation strips of $\mathbb{R} \times \Lambda$ with input puncture in $c \in C^+$ and output puncture at $q \in C^-$.

(iii) Rigid strips of $\widehat{L}^\rho$ with input puncture at $p \in P$ and output puncture at $q \in I$ and rigid strips of L with input puncture at $p \in P$ and positive puncture at $q \in C^-$.

To see this note first that the strips in (i) are unaffected by adding the almost trivial cobordism: the strips are transversely cut out and therefore solutions for small variations of the boundary data are canonically identified. Taking ρ sufficiently large the boundary data of the disks can be made arbitrarily close.

For (ii), in the limit $\rho \to \infty$ the disks limit to an anchored disk with a positive puncture and a puncture at the intersection point. Gluing to it the basic strip (see above for this notion) connecting the intersection point to the short Reeb chord gives a 1–dimensional moduli space. The other boundary component of this moduli space consists of a rigid strip in the cobordism and a disk or plane of dimension one at either symplectization end. We next argue that the other boundary component of the moduli space must be

a trivial strip followed by a strip connecting c to q at the negative symplectization end. To see this we note that the cobordism is obtained by a very small perturbation of the trivial cobordism $\mathbb{R} \times \Lambda$. It is well known that the only rigid strips of the trivial cobordism are the trivial strips; nontrivial curves have dimension ≥ 1. Therefore, a rigid strip limits either to a trivial strip or does not leave a small neighborhood of the trivial cobordism $\mathbb{R} \times \Lambda$. In this neighborhood holomorphic strips correspond to Morse flow lines of the shifting function, and thus the only rigid strips in the cobordism with negative ends at Reeb chords are either close to trivial strips or a basic strip. Our assertion follows. For (iii) we note that every rigid strip must break under stretching into two rigid strips. Since the only rigid strips in the upper part are the basic strips, the claim follows.

Observe that the strips in (i) and (iii) contribute to $\mathfrak{t}'$, the strips in (ii) to the differential on $\mathrm{LC}_*^{\|}(\Lambda)$. The vanishing of the wrapped Floer cohomology of $\widehat{L}^\rho$ then implies that $\mathfrak{e}_1$ is a quasi-isomorphism. The last statement follows from Proposition 59. □

Proof of Theorem 61 The A_∞–quasi-isomorphism given in Theorem 63 induces a quasi-isomorphism of DG–algebras

$$\Phi\colon \mathrm{B}(\boldsymbol{k}_- \oplus \mathrm{CF}^*(L)) \to \mathrm{B}(\mathrm{LA}_\|^*(\Lambda))$$

by an application of [21, Theorem 7.4] with respect to length filtrations on the bar construction.

By the local-finiteness and simple-connectedness assumptions, each of these bar constructions is locally finite. So we can apply the linear dual operation to get a quasi-isomorphism of DG–algebras

$$(42) \qquad\qquad \Phi^\#\colon \mathrm{B}(\mathrm{LA}_\|^*(\Lambda))^\# \to \mathrm{B}(\boldsymbol{k}_- \oplus \mathrm{CF}^*(L))^\#.$$

The result then follows as in Theorem 18, where local-finiteness of the grading enabled us to appeal to Lemma 10. Therefore, the quasi-isomorphism given in (42) induces the required quasi-isomorphism

$$\Omega(\mathrm{LC}_*^{\|}(\Lambda)) \to \Omega(\boldsymbol{k}_- \oplus \mathrm{CF}_*(L)). \qquad\qquad □$$

We next turn to the non-simply-connected case, where we use $\mathrm{CE}_\|^*(\Lambda)$ as defined in Section 3.4.2 directly without using the corresponding coalgebra. Note that the A_∞–algebras $\boldsymbol{k}_- \oplus \mathrm{CF}^*(L)$ and $\mathrm{LA}_\|^*(\Lambda)$ are finite-rank $\boldsymbol{k}$–bimodules (in particular, they are locally finite), thus we can consider their $\boldsymbol{k}$–duals, which are, by definition,

the A_∞–coalgebras $\boldsymbol{k}_- \oplus \mathrm{CF}_*(L)$ and $\mathrm{LC}_*^{\|}(\Lambda)$. However, unless we have the simple-connectedness assumption, the A_∞–quasi-isomorphism

$$\mathfrak{e} \colon \boldsymbol{k}_- \oplus \mathrm{CF}^*(L) \to \mathrm{LA}_{\|}^*(\Lambda)$$

does not necessarily dualize to an A_∞–comap

$$\mathfrak{f} \colon \mathrm{LC}_*^{\|}(\Lambda) \to \boldsymbol{k}_- \oplus \mathrm{CF}_*(L),$$

because A_∞–comaps are required to factorize through inclusion of the corresponding direct sum into the direct product as in (9). This is to ensure that a A_∞–coalgebra map $\mathfrak{f}$ induces a DG–algebra map $\Omega\mathfrak{f}$ on the cobar construction.

If we drop this condition, the A_∞–quasi isomorphism dualizes to DG–algebra map

$$\widetilde{\mathrm{CE}}_{\|}^*(\Lambda) \to \mathscr{A},$$

and this is just the map (41) induced by the twisting cochain $\mathfrak{t}$. Furthermore, by Proposition 59 this gives a DG–algebra map

$$\widehat{\Omega}(\mathfrak{f}) \colon \widehat{\mathrm{CE}}_{\|}^*(\Lambda) \to \widehat{\Omega}(\boldsymbol{k}_- \oplus \mathrm{CF}_*(L)).$$

Now, since $\mathfrak{f}$ is a quasi-isomorphism, by using the length filtration on $\widehat{\Omega}$, and appealing to [21, Theorem 7.4], we can conclude the following:

Theorem 64 *Suppose that* $\mathrm{HW}^*(L) = 0$. *Then there exists a quasi-isomorphism of DG–algebras*

$$\widehat{\mathrm{CE}}_{\|}^*(\Lambda) \to \widehat{\Omega}(\boldsymbol{k}_- \oplus \mathrm{CF}_*(L)).$$

Note that the completion $\widehat{\mathrm{CE}}_{\|}^*$ is in general a cruder invariant than both $\mathrm{CE}_{\|}^*(\Lambda)$ and $\mathrm{CE}^*(\Lambda)$. Nevertheless, we always have a map

$$\mathrm{CE}^*(\Lambda) \to \widehat{\mathrm{CE}}_{\|}^*(\Lambda).$$

Theorem 64 can be used to compute $\widehat{\mathrm{CE}}_{\|}^*$ in a variety of cases. For example, if L is a connected Lagrangian filling decorated by $-$ of a Legendrian Λ diffeomorphic to a sphere S^{n-1}, then writing $\bar{L} = L \cup_\partial D^n$, we have a quasi-isomorphism $\boldsymbol{k} \oplus \mathrm{CF}_*(L) \simeq C_*(\bar{L})$ since $\bar{L}$ is an exact Lagrangian. Hence, we have a map

$$\mathrm{CE}^*(\Lambda) \to \widehat{\Omega}(C_*(\bar{L})).$$

Here the right-hand side can often be computed; in particular, $H^0\big(\widehat{\Omega}(C^*(L))\big)$ is the group ring of the unipotent completion of the fundamental group $\pi_1(L)$; see [16].

In particular, any information on the completion map $\mathrm{CE}^*(\Lambda) \to \widehat{\mathrm{CE}}^*_\parallel(\Lambda)$ can help to obtain information about $\mathrm{CE}^*(\Lambda)$. We will see an application of this idea in the next section.

We end this section with a discussion of the twisting cochains constructed above from an after-surgery perspective. Assume that all components of Λ^- are spheres and recall the surgery isomorphism

$$\Theta \colon \mathrm{CW}^*(C) \to \mathrm{CE}^*(\Lambda)$$

of Conjecture 89. Let $\widetilde{\Theta} = \phi \circ \Theta$, where ϕ is the identity map on components in Λ_- and the map ϕ of Theorem 51. We next note that there is a natural A_∞–algebra map

$$\Psi \colon \mathrm{CW}^*(C) \to \mathrm{B}(\boldsymbol{k} \oplus \mathrm{CF}^*(L))^\# = \mathscr{A}' \to \mathscr{A} = \mathscr{A}'/\mathscr{I},$$

where we identify $\boldsymbol{k}_- \oplus \mathrm{CF}^*(L)$ with the Floer cohomology $\mathrm{CF}^*(L')$ of the manifold after surgery obtained by capping off all boundary spheres in Λ^- by disks. (Note that in the simply connected case, the shifting Morse function then extends with a unique minimum in each disk which gives an idempotent corresponding to e_v.)

The map Ψ is defined by a curve count. Fix systems of parallel copies $\overline{C}$ of C and $\overline{L}'$ of L'. Let $\boldsymbol{c}' = c_1 \cdots c_i$ be a word of Reeb chords of C and let $\boldsymbol{x}_0 = x_{0;1} \cdots x_{0;j}$ be a word of intersection points of L'. Let

$$\boldsymbol{c} = \boldsymbol{c}' z_v \boldsymbol{x}_0 z_w$$

and define

$$\Psi(\boldsymbol{c}') = \sum_{|\boldsymbol{x}_0| = |\boldsymbol{c}'| + (1-i)} |\mathcal{M}^{\overline{\mathrm{co}}}(\boldsymbol{c})| \boldsymbol{x}_0.$$

Remark 65 We require here that the parallel copies $\overline{L}'$ give a system of parallel copies of Λ near the surgery region in such a way that, for the components of Λ_v labeled by $-$ (resp. $+$), the induced parallel copies $\overline{\Lambda}_v$ lie in the negative (resp. positive) Reeb direction. Compare with Figure 5.

Theorem 66 *The map Ψ is a map of A_∞–algebras and $\Psi(z_v) = u'_v$, where z_v is the strict unit in $\mathrm{CW}^*(C_v)$ and u'_v is the dual of the unit in $\mathbb{K}e_v \oplus \mathrm{CF}^*(L_v)$ for each v.*

Proof This follows as usual by identifying terms contributing to the A_∞–relations with the oriented boundary of an oriented 1–manifold and Theorems 74–76.

To compute $\Psi(z_v)$ note that we can represent z_v as the minimum of the shifting Morse function of C and there is a unique flow line from this minimum to the intersection

point $C_v \cap L'_v$, and a unique flow line in L'_v from the intersection point to the minimum in L_v. The corresponding holomorphic disk starts at the intersection point between C_0 and C_1, has two corners at $C_0 \cap L'_0$ and $C_1 \cap L'_1$, and ends at the intersection point in $L'_0 \cap L'_1$ corresponding to the minimum of the shifting Morse function. $\qquad\square$

The pre-twisting cochain $\mathfrak{t}' \colon \mathrm{LC}_*^\parallel(\Lambda) \to \mathscr{A}$ can now be seen to arise via SFT stretching as follows. Consider the first component Ψ_1 of the A_∞–map above and a holomorphic disk contributing to it. Now stretch the lower end of the cobordism. Then by SFT compactness each curve contributing to Ψ_1 breaks up into a curve contributing to the map $\phi \circ \Theta$ followed by the twisting cochain at each negative puncture. The z_v are the only low-energy generators, so it follows that the map induces a map of the high-energy quotient into $\mathrm{coker}(\eta)$, and we can write the induced map Ψ_1^+ as $\Psi_1^+ = (\Omega\mathfrak{t}) \circ \phi \circ \Theta^+$.

6 Examples and applications

6.1 Concrete calculations

In this section we compute Legendrian and Lagrangian invariants in a number of concrete examples.

6.1.1 The unknot
Consider the Legendrian unknot $\Lambda \subset S^{2n-1}$ for $n > 1$ with its standard filling $L = D^n \subset D^{2n}$. Then Λ can be represented as a standard unknot in a small Darboux chart which has effectively only one Reeb chord with respect to the standard Reeb flow on S^{2n-1}; see [11, Section 7.1].

We work over a field $\mathbb{K}$. Consider first the case when L is decorated by $-$. Then

$$\mathrm{LC}_*(\Lambda) = \mathbb{K} \cdot 1 \oplus \mathbb{K} \cdot c \quad \text{with } |c| = -n,$$

with all coalgebra maps $(\Delta_i)_{i \geq 1}$ trivial, except Δ_2, for which we have

$$\Delta_2(1) = 1 \otimes 1 \quad \text{and} \quad \Delta_2(c) = 1 \otimes c + c \otimes 1$$

by the counitality. Using a Morse function on D^{2n} with a unique local maximum a and which decreases along the end corresponding to a shift in the negative Reeb direction, we have

$$\mathbb{K} \oplus \mathrm{CF}_*(L) = \mathbb{K} \cdot 1 \oplus \mathbb{K} \cdot a \quad \text{with } |a| = -n.$$

Let $\mathscr{A} = \Omega(\mathbb{K} \oplus \mathrm{CF}_*(L))$ be the Adams–Floer algebra, where the degree of a is now shifted up by 1. Then we have the twisting cochain

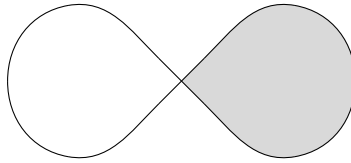$$\mathfrak{t} \colon \mathrm{LC}_*(\Lambda) \to \mathscr{A}, \quad \mathfrak{t}(c) = a.$$

Figure 10: Computation of t for the Legendrian unknot. The disk drawn lies in an $(n-1)$–dimensional family that sweeps the filling once.

Here the disk with input c and output a corresponds to the family of disks with positive puncture at c that sweeps L once.

The case of $n = 2$ is drawn in Figure 10.

The Koszul complex is generated over $\mathbb{K}$ by a^k, $a^k c$ with $k \geq 0$. We can compute the nontrivial part of the differential to be

$$d^{\mathrm{t}}(a^k c) = a^{k+1} \quad \text{for all } k \geq 0;$$

hence, t is acyclic. This is consistent with the classical Koszul duality between the algebras $C^*(S^n)$ and $C_{-*}(\Omega S^n)$ for $n > 1$.

Consider next the case when L is decorated by $+$. Since S^{n-1} is simply connected, $\mathrm{CE}^*(\Lambda) \approx \mathrm{CE}_\parallel^*(\Lambda)$ and we will use the parallel copies version in our calculation. Choose a Morse function on Λ with a single minimum and a single maximum. Denote the corresponding Reeb chords by x (the counit chord) and y. Then

$$\mathrm{LC}_*^\parallel(\Lambda) = \mathbb{K} \cdot x \oplus \mathbb{K} \cdot y \oplus \mathbb{K} \cdot c, \quad \text{with } |x| = 0, \ |y| = -(n-1), \ |c| = -n.$$

Here

$$\Delta_1(c) = y, \quad \Delta_2(x) = x \otimes x + +(-1)^{n-1} x \otimes y + y \otimes x + (-1)^n x \otimes c + c \otimes x,$$

and all other operations are trivial. It follows that

$$\mathrm{CE}_\parallel^*(\Lambda) = \Omega(\mathrm{LC}_*^\parallel(\Lambda)) \simeq \mathbb{K}.$$

This is in line with Conjecture 3, which says that $\mathrm{CE}^*(\Lambda) \simeq \mathrm{CE}_\parallel^*(\Lambda)$ is isomorphic to $\mathrm{CW}^*(C)$, where C is the cotangent fiber in the manifold obtained by attaching a cotangent end $T^*(S^{n-1} \times [0, \infty))$ to the ball along Λ. The manifold that results from this attachment is simply $T^*\mathbb{R}^n$, and the wrapped Floer cohomology of the cotangent fiber C has rank 1 and is generated by the minimum in the disk C, in accordance with the above calculations.
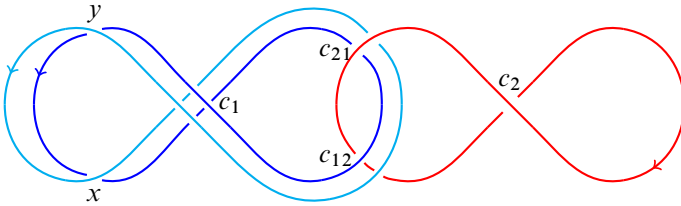
Figure 11: Hopf link with one (blue) marked $+$ and one (red) marked $-$ component.

Finally, the twisting cochain $\mathfrak{t}$ in the $+$ case is the canonical map $\mathbb{K} \to \mathbb{K}$, and again the Koszul complex is acyclic. As explained in Section 5 this map is induced by a map $\mathfrak{t}' \colon \mathrm{LC}^{\parallel}_*(\Lambda) \to (\mathrm{BCF}^*(L))^{\#}$. To define $\mathrm{CF}^*(L)$ we use a Morse function on L with a single local minimum and which is increasing along the end corresponding to a shift in the positive Reeb direction. Denote the generator of $\mathrm{CF}^*(L)$ corresponding to the minimum by u, $|u| = 0$. Then

$$\mathfrak{t}'(x) = u',$$

where u' is the dual of u and the holomorphic strip is the thin strip corresponding to a rigid Morse flow line from the minimum u to the minimum y in the boundary.

6.1.2 Geometric twisting cochain for the Hopf link In this section we carry out the geometric calculation of the twisting cochain for the Hopf link. As explained we cannot directly calculate the twisting cochain into the Legendrian coalgebra with coefficients in chains of the based loop space. We can however calculate the corresponding twisting cochain when we replace chains on the based loop space with the Morse theory of parallel copies for the components decorated by a positive sign. To carry out the calculation we pick a Morse function on the component Λ^+ with on minimum x and a maximum y. We place them on the circle and choose parallel copies as shown in Figure 11. The parallel copies algebra $\mathrm{CE}^*_{\parallel}(\Lambda)$ is

$$\boldsymbol{k} \langle x, y, c_1, c_{12}, c_{21}, c_2 \rangle,$$

where we use notation for Reeb chords and Floer cohomology generators as in Section 1.4. The differential is

$$dc_1 = xc_1 + c_1 x + y + c_{12}c_{21}, \qquad dx = xx, \qquad dy = xy - yx,$$
$$dc_{12} = -xc_{12}, \qquad\qquad\qquad dc_{21} = c_{21}x, \quad dc_2 = c_{21}c_{12}.$$

Passing to $\mathrm{CE}^*_{\parallel}(\Lambda)$ means dividing out by the cokernel of the coaugmentation that takes e_1 to x. This gives the algebra

$$\boldsymbol{k} \langle y, c_1, c_{12}, c_{21}, c_2 \rangle,$$

and the differential becomes

$$dc_1 = y + c_{12}c_{21}, \quad dc_2 = c_{21}c_{12}.$$

The twisting cochain $\mathfrak{t}$ is induced from a map $\mathfrak{t}' : \mathrm{LC}^{\|}_*(\Lambda) \to \big(\mathrm{B}(\boldsymbol{k}_- \oplus \mathrm{CF}^*(L))\big)^{\#}$ that counts holomorphic disks with one positive puncture and boundary on L, and with several punctures at Lagrangian intersection points in the compact part; see (40). In the current example it is straightforward to find these disks. Note first that, by general properties (see Proposition 59),

$$\mathfrak{t}'(x) = a_1^{\vee},$$

where a_1 is the idempotent corresponding to the minimum of the shifting Morse function on L_1. The holomorphic disk corresponds to a Morse flow line connecting x to u_1. We next consider $\mathfrak{t}'(c_1)$ and $\mathfrak{t}'(c_2)$. Consider first the moduli spaces $\mathcal{M}^{\mathrm{fi}}(c_j)$ of holomorphic disks with a positive puncture at c_j and boundary on L_j. As in the case of the unknot this moduli space sweeps L_j. On both L_1 and L_2 the shifting functions have one critical point, on L_1 it is a minimum and on L_2 a maximum. The maximum is constraining for the map into the linear dual of $\mathrm{CF}^*(L_j)$, whereas the minimum is not. We find that

$$\mathfrak{t}'(c_1) = 0 \quad \text{and} \quad \mathfrak{t}'(c_2) = a_2^{\vee}.$$

The spaces $\mathcal{M}^{\mathrm{fi}}(c_j)$ give further information of the twisting cochain as follows. Note that as the evaluation map hits the double point one can glue on a constant disk. These broken disks are also the boundary of the 1–dimensional moduli spaces $\mathcal{M}^{\mathrm{fi}}(c_1 a_{12} a_{21})$ and $\mathcal{M}^{\mathrm{fi}}(c_2 a_{21} a_{12})$. The other end of these moduli spaces correspond to broken disks with one level in the symplectization, a disk in $\mathcal{M}^{\mathrm{sy}}(c_1 c_{12} c_{21})$ in the former case and in $\mathcal{M}^{\mathrm{sy}}(c_2 c_{21} c_{12})$ in the latter, and two disks one in $\mathcal{M}^{\mathrm{fi}}(c_{12} a_{12})$ and one in $\mathcal{M}^{\mathrm{fi}}(c_{21} a_{21})$ attached at its negative end. The last disks contribute to $\mathfrak{t}'$ and we conclude that

$$\mathfrak{t}'(c_{12}) = a_{12}^{\vee} \quad \text{and} \quad \mathfrak{t}'(c_{21}) = a_{21}^{\vee}.$$

Finally, we compute $\mathfrak{t}'(y)$. Since y is a small chord corresponding to a critical point at infinity of the shifting Morse function the only contributions to $\mathfrak{t}'(y)$ come from small holomorphic disks that are controlled by the Morse theory. It is straightforward to check that the only rigid disk corresponds to a flow line in L_1 connecting y to the intersection point and that this flow line corresponds to a holomorphic triangle in $\mathcal{M}^{\mathrm{fi}}(y a_{12} a_{21})$. It follows that
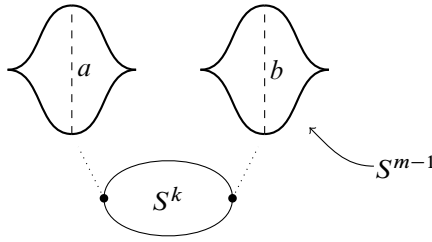
$$\mathfrak{t}'(y) = a_{21}^{\vee} a_{12}^{\vee}.$$

Figure 12: Front of products of spheres.

The actual twisting cochain takes the cokernel $\overline{LC}_*(\Lambda)$ of the coaugmentation into the cokernel of the counits. Concretely, this means disregarding x and $a_1^\vee$, and we get

$$\mathfrak{t}(c_1) = 0, \quad \mathfrak{t}(c_2) = a_2^\vee, \quad \mathfrak{t}(c_{12}) = a_{12}^\vee, \quad \mathfrak{t}(c_{21}) = a_{21}^\vee, \quad \mathfrak{t}(y) = a_{12}^\vee a_{21}^\vee.$$

Remark 67   The parallel copies algebra $CE_\parallel^*(\Lambda)$ is defined using a fixed augmentation (in the current example the zero augmentation) on the one copy version of $CE^*(\Lambda)$. Here this is reflected in the change of variables $t - e_1 = y$.

6.1.3 Products of spheres   We consider a Legendrian embedding $\Lambda \subset \mathbb{R}^{2(m+k)-1}$, where the ambient space is standard contact $(2(m-k)-1)$–space with coordinates $(x, y, z) \in \mathbb{R}^{m+k-1} \times \mathbb{R}^{m+k-1} \times \mathbb{R}$ and contact form $dz - y\,dx$. We will define it by describing its front in $\mathbb{R}^{m+k-1} \times \mathbb{R}$. To this end consider first the following construction of the front of the Legendrian unknot in $\mathbb{R}^n \times \mathbb{R}$. Take a disk D^n lying in $\mathbb{R}^n$. Think of it as having multiplicity two and lift one of the sheets up in the auxiliary $\mathbb{R}$–direction (with coordinate z) keeping it fixed along the boundary. In this way we construct the front of the standard unknot with Reeb chord at the maximum distance between the two sheets and a cusp edge along the boundary of D^n. Consider now instead the base $\mathbb{R}^{m+k-1}$ and the standard embedding of the k–dimensional sphere S^k into this space. A tubular neighborhood of this embedding has the form $S^k \times D^{m-1}$ with fibers D^{m-1}. Now take two copies of this tubular neighborhood and repeat the above construction for the D^{m-1} in each fiber. The result is the front of a Legendrian $S^k \times S^{m-1}$ with an S^k Bott family of Reeb chords corresponding to the maxima in fibers. Figure 12 shows this front after Morse perturbation. The resulting Legendrian Λ has only two Reeb chords. We denote them by a, with grading $|a| = -(m+k)$, and b, with grading $|b| = -m$. Note also that Λ has an exact Lagrangian filling $L \approx D^m \times S^k$.

Consider first the case when L is decorated by $-$. If d is the differential in $CE^*(\Lambda)$, we have

$$da = 0 \quad \text{and} \quad db = 0.$$

The Floer cohomology of L is defined by choosing a shifting function which is decreasing at infinity and we find that CF* has two generators M, with $|M| = m + k$, and S, with $|S| = m$. As in the case of the unknot $\mathcal{M}^{\mathrm{fi}}(a)$ sweeps L and $\mathcal{M}^{\mathrm{fi}}(b)$ sweeps $D^m \times \mathrm{pt}$. It follows that the twisting cochain satisfies

$$\mathfrak{t}(a) = M^\vee \quad \text{and} \quad \mathfrak{t}(b) = S^\vee,$$

and duality holds.

Consider second the case when L is decorated by $+$. In this case there are additional generators of $\mathrm{LC}^\|_*(\Lambda)$ corresponding to the Morse theory of Λ. We have in addition to a and b above also

$$x, \; s_1, \; s_2, \; y, \quad \text{with} \quad |x| = 0, \; |s_1| = -k, \; |s_2| = -(m-1), \; |y| = -(m+k-1).$$

It follows that $\mathrm{CE}^*_\|(\Lambda)$ is generated by s_1, s_2, y, a and b. Using the flow tree description of moduli spaces $\mathcal{M}^{\mathrm{sy}}$ one verifies that if d is the differential on $\mathrm{CE}^*_\|(\Lambda)$ then

$$da = y - ((-1)^{km} bs_1 + s_1 b), \quad dy = s_1 s_2 + (-1)^{k(m-1)} s_2 s_1,$$
$$db = s_2, \quad ds_1 = 0, \quad ds_2 = 0.$$

The Floer cohomology of L is defined by choosing a shifting function which is increasing at infinity and we find that CF* has two generators M, with $|M| = 0$, and S, with $|S| = k$, where M is the unit. It follows that $(\mathrm{BCF}^*(L))^\# \simeq \Omega\mathrm{CF}_*(L))$ is generated by $S^\vee$ with $|S^\vee| = -k$ and the twisting cochain is

$$\mathfrak{t}(a) = 0, \quad \mathfrak{t}(y) = 0, \quad \mathfrak{t}(b) = 0, \quad \mathfrak{t}(s_2) = 0, \quad \mathfrak{t}(s_1) = S^\vee,$$

and duality holds.

6.1.4 Plumbings of simply connected cotangent bundles

Let $\mathcal{T}$ be a tree with vertex set Γ. For each $v \in \Gamma$, let M_v be a compact simply connected manifold of dimension $n \geq 3$. We will see that duality holds between the wrapped and the compact Fukaya categories of the symplectic manifold $X_\mathcal{T}$ obtained by plumbing the collection of $T^* M_v$ according to the tree $\mathcal{T}$.

As usual, we take the pre-surgery perspective. Hence, consider a handle decomposition of each M_v with a unique top-dimensional n–handle. Removing this handle, we get manifolds L_v with spherical boundary Λ_v. Let $W_\mathcal{T}$ be the subcritical Weinstein manifold obtained by plumbing the cotangent bundles $T^* L_v$ according to the tree $\mathcal{T}$. We take the plumbing region to be away from the boundary of L_v. Write $\Lambda = \bigsqcup_v \Lambda_v$ for the Legendrian in the boundary of the subcritical Weinstein manifold $W_\mathcal{T}$ which is

filled by the Lagrangian $L = \bigcup_v L_v$. Equip the components of Λ with either $+$ or $-$ labeling.

Theorem 68 *If $n \geq 3$ and L_v is simply connected for each $v \in \Gamma$, then $\mathrm{CE}^*(\Lambda)$ and $\mathrm{CF}^*(L)$ are Koszul dual.*

Proof Consider first the case $n \geq 5$. Pick a handle decomposition of L_v without 1– and $(n-1)$–handles. The existence of such a handle decomposition is equivalent to simple connectedness in high dimensions by the work of Smale. Consider the corresponding Weinstein handle decomposition of T^*L_v. Attaching a k–handle alters the Legendrian boundary by surgery and adds Reeb chords in the cocore sphere of the handle, of grading $\leq -(n-k)$. It follows that all Reeb chords of Λ_v have grading ≤ -2. To construct $\Lambda \subset \partial W_{\mathcal{T}}$, we perform a version of boundary connected sum as follows. For each edge in $\mathcal{T}$ we pick a $2n$–ball B with two transversely intersecting Lagrangian disks $D \subset B$ that intersect the boundary sphere ∂B in a standard Legendrian Hopf link Δ. We then make the boundary connected sum adding (B, Δ) to join the T^*L_v according to the tree. This adds Reeb chords of index $\leq -(n-1)$ in the boundary connected sum handles and further Reeb chords corresponding to the Reeb chords of each Hopf link, which effectively has four Reeb chords: two Reeb chords connecting the unknot components to themselves of grading $-n$, and two mixed Reeb chords connecting distinct components. We can pick gradings so that the gradings of these two mixed chords are $-d$ and $-(n-d)$ for any d, where the first Reeb chord goes from the component closest to the a priori fixed root of the tree $\mathcal{T}$ to the one further from it and the other in the opposite direction. Taking d between 2 and $n-2$, we find that $\mathrm{LC}_*(\Lambda)$ is simply connected, as is Λ. The result then follows from Theorem 61.

For $n = 4$, we can stabilize by multiplying by $\mathbb{R}^N$ as described in Remarks 50 and 52 to get 1–reduced versions of the Legendrian and Lagrangian algebras, then use the above argument.

Finally, for $n = 3$, the assumptions of Theorem 61 do not hold, but we recall from Remark 20 that to apply the duality result from Theorem 61 all we really need is that $\mathrm{B}(\mathrm{LA}^*)$ is locally finite. This is easily seen to be the case in our case, since the plumbing is according to a tree (which by definition has no cycles) and for any word of Reeb chords that do not consist of connecting Reeb chords going away from the root of the tree, the grading has to increase with the size of the word. Alternatively, in this case we know by Perelman's theorem that we can take Λ as a link of standard

Legendrian spheres linked according to the tree as Hopf links, and $B(LA^*)$ can be seen to be locally finite directly. $\square$

Remark 69 The case $n = 2$ corresponds to plumbing of copies of T^*S^2. This case was studied in [32] and a version of the duality result still holds, at least when $\text{char}(\mathbb{K}) = 0$. However, the above argument fails in that case and a more complicated argument using an additional grading is used in [32]. Also a set of examples for $n = 1$ are studied in [48], where the plumbing tree is a star and the corresponding symplectic manifold is a punctured torus. The duality still holds in this case, even though this is a plumbing of T^*S^1 copies (not simply connected). The proof of duality given in [48] uses homological mirror symmetry.

6.1.5 The trefoil Consider the standard Legendrian trefoil $\Lambda \subset S^3$ described in Figure 13. Let us first consider the case when Λ is marked $-$. With respect to the standard choice of orientation datum, the Chekanov–Eliashberg DG–algebra CE^* is then given by the free algebra

$$\mathbb{K}\langle c_1, c_2, b_1, b_2, b_3 \rangle, \quad \text{with } |c_1| = |c_2| = -1 \text{ and } |b_1| = |b_2| = |b_3| = 0,$$

and the nontrivial part of the differential can be read from Figure 13 as

$$dc_1 = 1 + b_1 + b_3 + b_3 b_2 b_1, \quad dc_2 = -1 - b_1 - b_3 - b_1 b_2 b_3.$$

It is well known that Λ has an exact Lagrangian torus filling. (In fact, there are at least five of them; see [29].) Any of these can be obtained by doing surgery (pinch move) at b_1, b_2 and b_3 in some order. Corresponding augmentations $\epsilon_L : \text{CE}^* \to \mathbb{K}$ are given by their values $\epsilon(b_1), \epsilon(b_2), \epsilon(b_3) \in \mathbb{K}$ subject to the condition

$$1 + \epsilon(b_1) + \epsilon(b_3) + \epsilon(b_1)\epsilon(b_2)\epsilon(b_3) = 0.$$

Note that since CE^* is not simply connected, Theorem 61 does not apply here. In fact, duality does not hold in this example. However, Theorem 64 shows that there is a
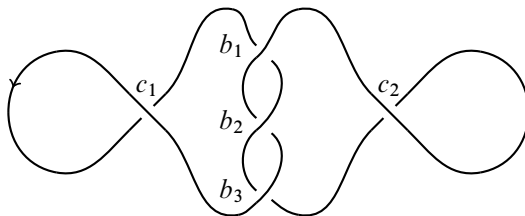

Figure 13: Trefoil.

quasi-isomorphism of completions

$$\widehat{\mathrm{CE}}^* \cong \hat{\Omega} C_*(T^2) \cong \mathbb{K}[\![u, v]\!],$$

where the latter is a power series ring in two commuting variables concentrated in degree 0.

Thus, the completion map composed with the twisting cochain gives an algebra map

$$H^0(\mathrm{CE}^*) = \mathbb{K}\langle b_1, b_2, b_3 \rangle / \langle 1 + b_1 + b_3 + b_3 b_2 b_1, 1 + b_1 + b_3 + b_1 b_2 b_3 \rangle \to \mathbb{K}[\![u, v]\!].$$

We claim that this map is injective. Indeed, observe that $H^0(\mathrm{CE}^*)$ is a commutative algebra,

$$d(c_1 + c_2 + b_3 b_2 c_1 + c_1 b_2 b_3) = b_3 b_2 - b_2 b_3,$$

and thus, b_2 commutes with b_3 in homology. Using this, one shows similarly that b_1 commutes with b_2 and b_3; see [15]. Hence, we have a completion map from a commutative ring to its completion,

$$H^0(\mathrm{CE}^*) = \mathbb{K}[b_1, b_2, b_3]/(1 + b_1 + b_3 + b_1 b_2 b_3) \to \mathbb{K}[\![u, v]\!].$$

It is a well-known theorem in commutative algebra, the Krull intersection theorem, that, for any commutative Noetherian ring which is an integral domain, the completion map is injective. Thus, even though duality fails in this example, partial information can still be obtained by considering completions.

We next describe the twisting cochain $t: \mathrm{LC}_*(\Lambda) \to \mathrm{BCF}^*(L)^\#$ for one of the Lagrangian fillings L of Λ. More precisely, we choose the filling which is obtained by pinching first at b_1, then at b_2, then at b_3, and finally filling the resulting unknots with Lagrangian disks; see [29, Section 8.1]. We think of the Lagrangian filling as two disks connected by three twisted bands corresponding to the three pinchings. We next consider moduli spaces of holomorphic disks with boundary in L. Here [29, Sections 4–5] gives a description in terms of Morse flow trees which gives the following:

- $\mathcal{M}^{\mathrm{fi}}(b_1)$ consists of one disk, δ_1^1.
- $\mathcal{M}^{\mathrm{fi}}(b_2)$ consists of two disks, δ_2^1 and δ_2^2.
- $\mathcal{M}^{\mathrm{fi}}(b_3)$ consists of two disks, δ_3^1 and δ_3^2.

The boundaries of these disks are as follows:

- $\partial \delta_1^1$ is a fiber in the first twisted band.

- $\partial \delta_2^1$ is a fiber in the second twisted band, and $\partial \delta_2^2$ runs across the first and the second twisted band.

- $\partial \delta_3^1$ is a fiber in the third twisted band, and $\partial \delta_3^2$ runs across the second and the third twisted bands.

We next describe the moduli spaces $\mathcal{M}^{\mathrm{fi}}(c_2)$ and $\mathcal{M}^{\mathrm{fi}}(c_1)$ in a completely analogous manner. The space has four components, all diffeomorphic to intervals, as follows:

- θ_1 with one boundary point the disk in $\mathcal{M}^{\mathrm{sy}}(c_2 b_1)$ with δ_1^1 attached, and the other the disk in $\mathcal{M}^{\mathrm{sy}}(c_2 b_1 b_2 b_3)$ with δ_1^1, δ_2^1 and δ_3^2 attached.

- θ_2 with one boundary point the disk in $\mathcal{M}^{\mathrm{sy}}(c_2 b_3)$ with δ_3^1 attached, and the other the disk in $\mathcal{M}^{\mathrm{sy}}(c_2 b_1 b_2 b_3)$ with δ_1^1, δ_2^2 and δ_3^1 attached.

- θ_3 with one boundary point the disk in $\mathcal{M}^{\mathrm{sy}}(c_2 b_3)$ with δ_3^2 attached, and the other the disk in $\mathcal{M}^{\mathrm{sy}}(c_2 b_1 b_2 b_3)$ with δ_1^1, δ_2^2 and δ_3^2 attached.

- θ_4 with one boundary point the disk in $\mathcal{M}^{\mathrm{sy}}(c_2)$ and the other the disk in $\mathcal{M}^{\mathrm{sy}}(c_2 b_1 b_2 b_3)$ with δ_1^1, δ_2^1 and δ_3^1 attached.

Here the disks in θ_4 sweep the right-hand disk of L, whereas the evaluation maps of θ_j for $1 \leq j \leq 3$ do not map surjectively onto either disk. The moduli space $\mathcal{M}^{\mathrm{fi}}(c_1)$ also has four components, only one of which sweeps the left-hand disk of L.

In order to compute the Floer cohomology $\mathrm{CF}^*(L)$ we pick a Morse function with one maximum M and two saddle points S_1 and S_2. Letting the maximum lie in the right-hand disk we find

$$\mathfrak{t}(c_1) = 0,$$

since the only way to rigidify a disk of dimension one is that its boundary passes the maximum M.

The twisting cochain can now in principle be computed from the moduli spaces $\mathcal{M}^{\mathrm{fi}}$ described above by attaching flow trees. To get an algebraically feasible twisting cochain we first arrange the perturbation scheme so that

$$\mathfrak{t}(b_1) = S_1^\vee + S_1^\vee S_2^\vee,$$

where the first term is δ_1^1 with a flow line and the second with a flow tree, and next so that

$$\mathfrak{t}(b_3) = -S_1^\vee,$$

the disk δ_2^2 with a flow line contributes, while other contributions cancel (the disk δ_2^2 with two flow lines and the same disk with a flow tree, the two distinct disks with

one flow line). Remaining parts of the twisting cochain are now determined from the coproduct in the Floer homology

$$dM^\vee = S_1^\vee S_2^\vee - S_2^\vee S_1^\vee,$$

the twisting cochain equation, and the sweeping property of θ_4, as

$$\mathfrak{t}(c_2) = M^\vee\left(\frac{1 + S_1^\vee}{1 + S_1^\vee + S_1^\vee S_2^\vee}\right) \quad \text{and} \quad \mathfrak{t}(b_2) = \frac{S_2^\vee}{(1 + S_1^\vee + S_1^\vee S_2^\vee)}.$$

It is possible to check that these power series agree with the geometric count. We leave out the details but describe the mechanism. In order to arrange that only one flow line can be attached to δ_1^1 we order the stable manifolds of the flow line of the parallel copies so that if a flow line (or flow tree) between copy j and $j - l$ is attached then following δ_1^1 we already passed all intersections with stable manifolds between copy $j - l$ and k for $k < j - l$. Now, if the disk intersects the collection of stable manifolds in the opposite direction this means that we can jump down in all ways, which then gives the desired power series.

We finish this section with a brief discussion of the case of L decorated by $+$ and parallel copies. As usual this introduces two extra generators in addition to the Reeb chords above in LC_*, namely x, with $|x| = 0$, and y, with $|y| = -1$, where x is the counit corresponding to the minimum of the shifting function and y is the maximum. In the reduced coalgebra (disregarding x) we get the new differential (using the augmentation ϵ_L above to change coordinates)

$$dc_1 = b_1 + b_3 - b_3 b_2 + b_3 b_2 b_1, \quad dc_2 = -y - b_1 - b_3 + b_2 b_3 - b_1 b_2 b_3.$$

In this case CF^* is defined instead by choosing a shifting function that increases at infinity, and CF^* is generated by the unit u, with $|u| = 0$, and s_1 and s_2, with $|s_j| = 1$. The new twisting cochain is

$$\mathfrak{t}(c_1) = \mathfrak{t}(c_2) = 0, \ \mathfrak{t}(y) = (s_1^\vee s_2^\vee - s_2^\vee s_1^\vee)\left(\frac{1 + s_1^\vee}{1 + s_1^\vee + s_1^\vee s_2^\vee}\right),$$

and $\mathfrak{t}(b_j)$ is exactly as above, after the substitutions $s_j \to S_j$ for $j = 1, 2$.

6.1.6 Mirror of 7_2

We next discuss an example where Koszulity fails and also the completion map fails to be injective, even though CE^* is supported in nonpositive degrees. This example was shown to us by Lenhard Ng.

Consider the Legendrian knot Λ drawn in Figure 14 decorated $-$. It is easy to see that two pinch moves indicated by dashed lines in Figure 14 give a Legendrian unknot,
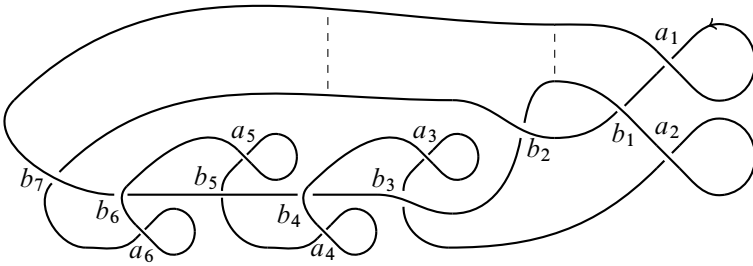
Figure 14: Mirror of 7_2.

hence Λ has a Lagrangian torus filling, call it L. Thus, we again have a completion map

$$H^0(\mathrm{CE}^*) \to \mathbb{K}[\![u, v]\!],$$

where $K[\![u, v]\!]$ is the commutative power-series algebra in two variables.

CE^* is given by the free algebra

$$\mathbb{K}\langle a_1, a_2, a_3, a_4, a_5, a_6, b_1, b_2, b_3, b_4, b_5, b_6, b_7 \rangle,$$

with $|a_i| = -1$ and $|b_i| = 0$.

The differential is given by

$$da_1 = -1 + (1 + b_1 b_2) b_7 + b_1 (1 + b_4 b_3)(1 + b_6 b_5), \quad da_2 = 1 - b_3(1 + b_2 b_1),$$

$$da_3 = 1 + b_3 b_4, \quad da_4 = 1 + b_5 b_4, \quad da_5 = 1 + b_5 b_6, \quad da_6 = 1 + b_7 b_6.$$

Taking the quotient of $H^0(\mathrm{CE}^*)$ by letting $b_4 = b_6$, $b_3 = b_5 = b_7$, $b_1 = 1$ and $b_2 = -1 - b_4$ gives

$$\langle b_3, b_4 \rangle / \langle 1 + b_3 b_4 \rangle,$$

which is a noncommutative algebra.

Thus, the completion map cannot be injective in this case. Otherwise, $H^0(\mathrm{CE}^*)$, and thus any quotient of it, would have been commutative.

6.2 Simply connected Legendrian submanifolds

Let $\Lambda \subset Y$ be a Legendrian $(n-1)$–submanifold with $\pi_1(\Lambda) = 1$ in the boundary Y of a Weinstein $2n$–manifold X that bounds an exact Lagrangian $L \subset X$. Assume that $c_1(X) = 0$ and that the Maslov class of L vanishes and that L is relatively spin. Decorate L by $-$. Our next result shows that if the symplectic homology of X vanishes and if all Reeb chords of Λ have negative grading as generators of $\mathrm{CE}^*(\Lambda)$, then

$CE^*(\Lambda)$ is determined by the topology of L, and conversely. If Λ is a sphere then we write $\bar{L} = L \cup_\partial D^n$ for the closed manifold obtained by adding a disk to L along Λ.

Theorem 70 *Suppose that $\Lambda = \Lambda^-$ is simply connected. Assume that $SH^*(X) = 0$ and that $CE^*(\Lambda)$ is supported in degrees ≤ -1. Then L is simply connected. Moreover, if Λ is a sphere, then $CE^*(\Lambda)$ is isomorphic to $C_{-*}(\Omega \bar{L})$.*

Proof Consider the wrapped Floer cohomology $HW^*_{\pi_1}(L)$ of L with coefficients in $\mathbb{Z}[\pi_1(L)]$. Using our model for wrapped Floer cohomology in Section B.1, a chain complex $CW^*_{\pi_1}(L)$ which calculates $HW^*_{\pi_1}(L)$ can be described as follows. Let $L = L_0$ and let L_1 be a parallel copy of L shifted in the negative Reeb direction at infinity. The complex $CW^*_{\pi_1}(L)$ is then generated over $\mathbb{Z}[\pi_1]$ by the intersection points in $L_0 \cap L_1$, which we call the Morse generators, and the Reeb chords starting on Λ_0 and ending on Λ_1. The differential of a generator counts the usual rigid holomorphic strips, keeping track of the homotopy class of the loop obtained from the boundary component of the disk in L_1 completed by the reference paths connecting Reeb chord endpoints and intersection points to the basepoint. We point out that since there are no Reeb chords of degree 0 the augmentation induced by L is trivial and the high-energy part of the differential on $CW^*_{\pi_1}$ counts honest holomorphic strips (without extra negative punctures at augmented Reeb chords). As for usual wrapped Floer homology, $HW^*_{\pi_1}(L)$ is naturally a module over symplectic cohomology $SH^*(X)$ and hence vanishes.

We next describe a geometric version of the complex $CW^*_{\pi_1}(L)$ that we call $CW^*_{\tilde{p}}$ and that also computes $HW^{\pi_1}(L)$. Let $\tilde{p}: \tilde{L} \to L$ denote the universal covering of L and let $\tilde{\Lambda} = \tilde{p}^{-1}(\Lambda)$. Pick a Morse function $f: \tilde{L} \to \mathbb{R}$ such that

- f has exactly one local maximum M,
- f has no index $n - 1$ critical points,
- f has no local minima,
- f is constant on $\tilde{\Lambda}$ where it attains its global minimum and if ν is the unit normal vector field along $\tilde{\Lambda}$ then $df(\nu) = -1$.

The generators of $CW^*_{\tilde{p}}$ are of two types:

(i) The preimages of endpoints of Reeb chords $L_0 \to L_1$ in $L \approx L_1$ under $\tilde{p}$ graded as the corresponding Reeb chord in $CE^*(\Lambda)$.

(ii) The critical points of the Morse function $f: \tilde{L} \to \mathbb{R}$ graded by the negative of the Morse index.

Let M_{-*} denote the Morse chain complex of f with cohomological grading, with generators as in (ii) and differential δ which counts negative gradient flow lines. Then M_{-*} is supported in degrees d with $-n \le d \le -1$ and $M_{-(n-1)} = 0$. Let C^* denote the complex generated by the generators of type (i) and equip it with the differential ∂ that counts lifts of the boundary of holomorphic strips in the symplectization interpolating between Reeb chords. (This corresponds naturally to the high-energy part of the differential on $\mathrm{CW}^*_{\pi_1}$.) By our assumption on Reeb chord grading, the grading of C^* is supported in degrees d where $d \le -1$.

We now define the complex $\mathrm{CW}^*_{\widetilde{p}} = C^* \oplus M_{-*}$, with differential

$$d = \begin{pmatrix} \partial & \phi \\ 0 & \delta \end{pmatrix},$$

where δ and ∂ are the differentials on M^* and C^*, and where ϕ counts rigid lifts of disks with flow lines of f attached. (This is the linear part of the map ϕ in (36).)

The homology of d is then isomorphic to $\mathrm{HW}^*_{\pi_1}(L)$. To see this note that we can describe $\mathrm{CW}^*_{\pi_1}(L)$ exactly as $\mathrm{CW}^*_{\widetilde{p}}(L)$ just replacing the Morse function f above with a Morse function $h \circ \widetilde{p}$, where $\widetilde{p}$ is a Morse function on L without minimum and with the required boundary behavior. Thus, passing from $\mathrm{CW}^*_{\pi_1}(L)$ to $\mathrm{CW}^*_{\widetilde{p}}(L)$ corresponds to deforming the Morse function on $\widetilde{L}$, and it is well known that this induces a homotopy of complexes. In particular, $\mathrm{CW}^*_{\widetilde{p}}(L)$ is acyclic.

We next want to show that $\pi_1(L) \approx 1$ or, equivalently, that the map $\widetilde{L} \to L$ has degree one. To show this we first observe that since there are no Reeb chords of grading 0 the augmentation of $\mathrm{CE}^*(\Lambda)$ is trivial and the differential on C^* counts honest holomorphic strips in the symplectization. This in turn means that the whole boundary of any holomorphic strip contributing to ∂ actually lies in $\Lambda \times \mathbb{R}$ and therefore cannot pick up any nontrivial $\mathbb{Z}[\pi_1]$–coefficient.

Consider the part of the chain complex $C^* \oplus M_{-*}$ given by

$$\cdots \to C^{-(n+1)} \to C^{-n} \oplus M_{-n} \to C^{-(n-1)} \to \cdots ,$$

where we use that $M_{-k} = 0$ for $k = n-1$ and $k > n$. It follows from the above discussion and the vanishing of the wrapped homology $\mathrm{HW}^*(L)$ with trivial coefficients that the cohomology $\mathrm{HW}^{-n}_{\pi_1}(L)$ in degree $-n$ has one generator for each nontrivial element in π_1. On the other hand, $\mathrm{HW}^{-n}_{\pi_1}(L) = 0$, and we conclude that $\pi_1 = 1$.

The statement about the isomorphism class of $\mathrm{CE}^*(\Lambda)$ follows from Corollary 62. □

Appendix A   Basic results for moduli spaces

Consider as above a Weinstein manifold with an exact Lagrangian submanifold (X, L), which outside a compact set agrees with the positive part of the symplectization of the contact manifold with Legendrian submanifold (Y, Λ). We assume that the Maslov class of L vanishes and that L is relatively spin. We will consider several versions of punctured holomorphic spheres and disks with boundary on L. The most basic disks we consider will lie either in X or in the symplectization $\mathbb{R} \times Y$. We call the former *filling curves* and the latter *symplectization curves*. We will also consider a more general cobordism setting where, like in the symplectization, disks may have both positive and negative punctures. Here we assume that (W, K) is a Weinstein cobordism with negative end $(-\infty, 0] \times (Y, \Lambda)$ and positive end $[0, \infty) \times (Z, \Gamma)$, where Z is a contact manifold and Γ is a Legendrian submanifold. We call disks in W *cobordism disks*.

When we want to consider the relation of our "pre-surgery" invariants to "post-surgery" invariants, we will consider the case when K decomposes as a Lagrangian $C \subset W$ with positive end Γ and empty negative end and a Lagrangian L with negative end Λ and empty positive end. We further assume that there is a natural one-to-one correspondence between the components of L^v of L and C^v of C for $v \in Q_0$, and that corresponding components L^v and C^v intersect transversely at one point z^v and that $L^v \cap C^w = \varnothing$ if $v \neq w$.

We will describe a geometric setup that covers the cases considered below. Let Y be the contact boundary of the Weinstein manifold X, where $c_1(X) = 0$. Let $\Lambda \subset Y$ be a Legendrian with connected components $\Lambda_1, \ldots, \Lambda_m$. Let D denote the unit disk in $\mathscr{C}$ and let $z_1, \ldots, z_r$ be boundary punctures and $\zeta_1, \ldots, \zeta_k$ interior punctures. Let each component of $\partial D \setminus \{z_1, \ldots, z_r\}$ be decorated with a component Λ_j. The boundary punctures come in two types, positive and negative; all interior punctures are negative. Following [23], we make the further requirement that the disk be *admissible*:

> Any arc in D that connects two boundary arcs in $\partial D \setminus \{z_1, \ldots, z_r\}$ subdivides the boundary punctures into two subsets. If both these subsets contain positive punctures, then the labels of the two boundary segments at the endpoints of the arc are different.

Remark 71   When we consider parallel copies of Lagrangians and Legendrians, boundary arcs labeled by different numbers in the numbering of parallel copies lie on copies shifted by different Morse functions, and correspond to distinctly labeled boundary conditions in the current discussion.
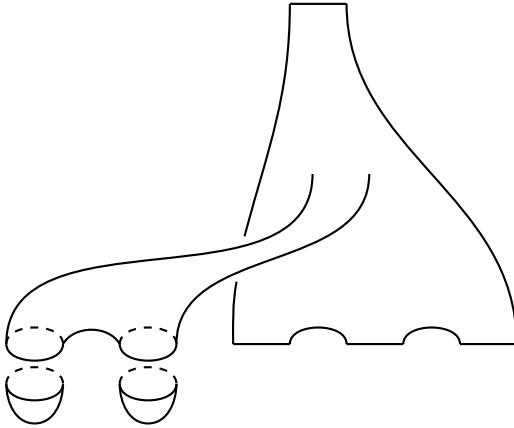
Figure 15: Anchored disks.

We also assume that the disk has one distinguished boundary puncture. Note that using a conformal model where the distinguished positive puncture lies at $1 \in \partial D$ and an interior puncture ζ_j lies at the origin, the positive real axis determines an asymptotic marker at ζ_j for each j. In the conformal model of the upper half-plane with the distinguished puncture at infinity, this marker at any interior puncture is that determined by the vertical axis.

A.1 Moduli spaces of spheres for anchoring, and compactifications of moduli spaces of disks

Following [11], all symplectization and cobordism disks we consider will be anchored. This means that the actual disks we consider have, aside from their boundary punctures, also additional interior punctures, where the maps are asymptotic to Reeb orbits at the negative end. An anchored disk is such a disk completed by holomorphic planes in X at all its negative interior punctures; see Figure 15.

When defining our version of the Chekanov–Eliashberg algebra CE^*, use moduli spaces of anchored disks to parametrize chains of boundary paths. When defining our versions of the Legendrian coalgebra LC_* and the wrapped Floer A_∞–algebra CW^*, we will need to consider disks in the symplectization with additional interior and boundary punctures, completed by rigid planes in X and disks in (X, L), respectively. We will also call such disks "anchored disks".

Although standard arguments using classical methods allow us to prove transversality for the disks with boundary punctures that we consider, anchored disks require transversality

and gluing also for holomorphic planes in (X, L), and that requires a more general perturbation scheme. Necessary perturbations for such curves were constructed in [25].

To state the relevant result let Y denote the contact boundary of X and consider a Reeb orbit γ in Y with a marker (ie a point $p \in \gamma$) on it. We write γ' for the orbit γ with a marker. Let $\mathcal{M}(\gamma')$ denote the moduli space of holomorphic planes in X with positive puncture with an asymptotic marker where the curve is asymptotic to γ, with the asymptotic maker mapping to the marker on γ. As in [25, Theorem 1.1], we define perturbation data λ so that $\mathcal{M}^\lambda(\gamma')$ is a transversely cut out space of solutions to a perturbed Cauchy–Riemann equation $\overline{\partial}_{J_\lambda} u = 0$, where J_λ is a domain-dependent almost complex structure that is allowed to depend also on the map u in the neighborhood of $\mathcal{M}(\gamma')$. The moduli space $\mathcal{M}^\lambda(\gamma')$ furthermore has a natural compactification $\overline{\mathcal{M}}(\gamma')$ as a manifold with boundary with corners, where boundary strata correspond to several level spheres. Here levels not in moduli spaces of the form $\mathcal{M}^\lambda(\beta')$ just discussed lie in a moduli space $\mathcal{M}^{\mathrm{sy},\lambda}(\beta', \eta')$, where β' is a Reeb orbit with marker and $\eta' = \eta'_1 \cdots \eta'_k$ is a word of Reeb orbits with markers. Elements in $\mathcal{M}^{\mathrm{sy},\lambda}(\beta', \eta')$ are maps $u: S \to \mathbb{R} \times Y$ of a punctured spheres S into the symplectization. There are fixed cylindrical ends $S^1 \times [0, \infty)$ near the punctures in S that are compatible with breaking in the sense of [30, Section 2.1]. The map u takes $(1, \infty)$ at a puncture to the marker of the corresponding Reeb orbit and u again solves a perturbed Cauchy–Riemann equation $\overline{\partial}_{J_\lambda} u = 0$, where J_λ is domain dependent and only depends on the angular coordinate along the ends near the punctures, has a positive puncture at β' and negative punctures at the orbits in η'.

We refer to [25, Section 2.4] for more details on $\overline{\mathcal{M}}(\gamma')$. Here we only point out that the asymptotic marker at the positive puncture of a curve in the compactification determines asymptotic markers at all negative punctures and that the level structure is compatible with this in the sense that the asymptotic marker at the positive puncture in a lower-level curve agrees with the asymptotic marker at the negative puncture where it is attached.

We next consider holomorphic disks in a cobordism (Z, K) with positive and negative ends $(\partial_\pm Z, \partial_\pm K)$. We include also the case when the cobordism (Z, K) is trivial, ie the symplectization $(\mathbb{R} \times Y, \mathbb{R} \times \Lambda)$, with $(\partial_\pm Z, \partial_\pm K) = (Y, \Lambda)$. Let c be a word of Reeb chords of $\partial_+ K$. Let $\gamma = \gamma_1 \cdots \gamma_k$ be a word of Reeb orbits in $\partial_- Z$. We define $\mathcal{M}^{\mathrm{neg}}(c, \gamma')$ to be the moduli space of punctured holomorphic disks in Z with boundary on K, with boundary punctures mapping to Reeb chords in the word c and one distinguished boundary puncture, and with additional negative interior punctures at

$\zeta_1, \ldots, \zeta_k$ mapping to Reeb orbits in the word $\boldsymbol{\gamma}$. Note that the distinguished boundary puncture determines an asymptotic marker at each interior puncture ζ_j that determines a marker on the corresponding γ_j. Let $\boldsymbol{\gamma}'$ denote the corresponding word or Reeb orbits with markers. Below we will show that such moduli spaces of disks with interior negative punctures with markers that are relevant to our study cannot contain multiple covers and are transversely cut out for a generic almost complex structure.

Recall the symplectic filling X of the negative end Y of the cobordism above. We will use punctured sphere curves in X in the compactification of the moduli spaces $\overline{\mathcal{M}}^\lambda(\gamma')$ to fill the interior punctures of the disks and treat them as disks with only boundary punctures. For this purpose, we define the moduli space of anchored disks $\mathcal{M}^{\mathrm{anc}}(\boldsymbol{c})$ as

$$\mathcal{M}^{\mathrm{anc}}(\boldsymbol{c}) = \bigcup_{\boldsymbol{\gamma}'} \left(\mathcal{M}^{\mathrm{neg}}(\boldsymbol{c}, \boldsymbol{\gamma}') \times \prod_{\gamma_j' \in \boldsymbol{\gamma}'} \mathcal{M}^\lambda(\gamma_j') \right),$$

where markers on Reeb orbits are induced from the distinguished boundary puncture. Here the topology on the moduli space of anchored curves is the product topology. This means in particular that the dimension of the boundary evaluation map equals the dimension $\dim(\mathcal{M}^{\mathrm{anc}}(\boldsymbol{c}))$ only on components $\mathcal{M}^{\mathrm{neg}}(\boldsymbol{c}, \boldsymbol{\gamma}') \times \prod_{\gamma_j' \in \boldsymbol{\gamma}'} \mathcal{M}^\lambda(\gamma_j')$ where $\dim(\mathcal{M}^\lambda(\gamma_j')) = 0$ for all j.

We consider next the case when the cobordism is trivial $(\mathbb{R} \times Y, \mathbb{R} \times \Lambda)$ and when all punctures mapping to chords in $\boldsymbol{c}$ are positive. The above construction then gives a stratification of the moduli space $\mathcal{M}(\boldsymbol{c})$ of holomorphic disks in (X, L) as follows. First, since all moduli spaces $\mathcal{M}^{\mathrm{neg}}(\boldsymbol{c}, \boldsymbol{\gamma}')$ are transversely cut out, the corresponding moduli spaces $\mathcal{M}^{\mathrm{neg},\lambda}(\boldsymbol{c}, \boldsymbol{\gamma}')$, where a small perturbation near the negative ends corresponding to the perturbation λ of holomorphic planes with asymptotic marker has been turned on, is canonically diffeomorphic to $\mathcal{M}^{\mathrm{neg}}(\boldsymbol{c}, \boldsymbol{\gamma}')$. Gluing the curves in $\mathcal{M}^{\mathrm{neg},\lambda}(\boldsymbol{c}, \boldsymbol{\gamma}')$ to the curves in $\mathcal{M}^\lambda(\gamma_j')$ and extending the perturbation, we get a compactification of the moduli space $\mathcal{M}(\boldsymbol{c})$, with boundary given by tree configurations of anchored disks. Near a broken configuration the moduli space is a manifold with boundary with corners, with corner structure induced by the gluing parameters; compare [31, Sections 6.4–6.6].

For example, if $\boldsymbol{c} = c$ is a single Reeb chord, then $\mathcal{M}(c)$ has a natural compactification with boundary of the form

$$\mathcal{M}^{\mathrm{anc}}(c, \boldsymbol{b}) \times \prod_{b_j \in \boldsymbol{b}} \mathcal{M}^{\mathrm{anc}}(b_j).$$

All moduli spaces of disks considered in this paper will be anchored, and we will drop the superscript "anc" from the notation.
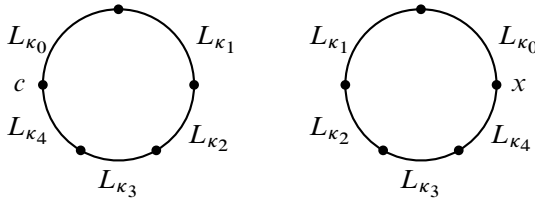
Figure 16: Strictly decreasing (left) and strictly increasing (right).

A.2 Moduli spaces of anchored disks

Consider a system of parallel copies $\overline{L} = \{L_j\}_{j=0}^{\infty}$, where $L_0 = L$ is an embedded Lagrangian, as in Section 3.3. This induces a system $\overline{\Lambda} = \{\Lambda_j\}_{j=0}^{\infty}$ of parallel copies of Λ in Y. We first discuss numberings that determine the boundary conditions of the holomorphic disks that we use in the various theories considered. Let D_m denote the unit disk in the complex plane with m boundary punctures $\zeta_1, \ldots, \zeta_m$. One of the boundary punctures is distinguished. We choose notation so that ζ_1 is distinguished. The m punctures subdivide the boundary of D_m into m boundary arcs. We will consider disks with numbered boundary arcs where the numbers correspond to the parallel copies L_j. We will consider two types of numberings, *increasing* and *decreasing*. Traversing the boundary of the disk across a boundary puncture in the positive direction, the numbering increases, remains constant, or decreases as we pass the puncture. We call punctures increasing, constant, and decreasing, accordingly. We call a disk increasing (resp. decreasing) if all its nondistinguished punctures are either increasing (resp. decreasing) or constant. Then its distinguished puncture is decreasing (resp. increasing) or constant.

When defining operations Δ_k for the Legendrian A_∞–coalgebra LC_*, we count anchored increasing disks in the symplectization asymptotic to Reeb chords at all punctures. When defining the operations for the Lagrangian A_∞–algebras CF^* and CW^*, we count decreasing disks in X. When defining the twisting cochain $\mathrm{t}\colon \mathrm{LC}_*^{\|} \to \Omega(\mathbf{k}_- \oplus \mathrm{CF}_*)$, we count increasing disks in X which are asymptotic to a Reeb chord at the distinguished puncture, and to Lagrangian intersection points at the other punctures.

We next consider asymptotic conditions at the boundary punctures. There are two basic forms of asymptotics: a puncture is either asymptotic to a Lagrangian intersection point or to a Reeb chord. The former case is the standard form of asymptotics in Lagrangian Floer theory, and the latter in Legendrian DG–algebras. More precisely, we choose an almost complex structure on X which along the cylindrical end $\mathbb{R} \times Y$ is invariant under

$\mathbb{R}$–translation, leaves the contact planes invariant and is compatible with the symplectic form induced on the contact planes by the contact form. Furthermore, it pairs the $\mathbb{R}$–direction with the Reeb direction. This means in particular that the Reeb chord strip, which is the product of a Reeb chord and $\mathbb{R}$, is holomorphic. "Reeb chord asymptotics" means we study holomorphic disks with boundary punctures that are asymptotic to these Reeb chord solutions, while "Lagrangian intersection point asymptotics" means we study holomorphic disks that are asymptotic to constant strips at intersection points.

First we consider disks in the filling (X, L). Consider a disk D_m as above with strictly increasing or decreasing numbering $\kappa = (\kappa_1, \ldots, \kappa_m)$ and let $\boldsymbol{a} = a_1 \cdots a_m$ be a word of Reeb chords and Lagrangian intersection points in $L_0 \cap L_1$. We let $\mathcal{M}^{\mathrm{fi}}(\boldsymbol{a}; \kappa)$ denote the moduli space of holomorphic disks $u \colon (D_m, \partial D_m) \to (X, \bar{L})$ such that

- u takes the boundary component labeled by κ_j to the Lagrangian L_{κ_j}, and

- u is asymptotic to the unique Reeb chord or Lagrangian intersection point of L_{κ_j} and $L_{\kappa_{j+1}}$ near a_j at ζ_j, where we let $\kappa_{m+1} = \kappa_1$.

We next consider disks in the symplectization. Consider the disk D_{m+k} with increasing or decreasing boundary numbering κ' and punctures $\zeta_1, \ldots, \zeta_{m+k}$. We next note that in the symplectization there are two possible Reeb chord asymptotics, positive or negative according to the sign of the t–coordinate near the puncture. Let $\boldsymbol{c}' = c_1^{\sigma_1} \cdots c_{m+k}^{\sigma_{m+k}}$ be a word of signed Reeb chords of $\Lambda_0 \cup \Lambda_1$, where $\sigma \in \{+, -\}$ is a sign. If $m > 1$ then we require that the Reeb chords c_r at all constant punctures connect Λ_0 to Λ_0 and that their signs are all negative, $\sigma_r = -1$. (These constant punctures will be capped by augmentation disks.) We let $\mathcal{M}_{\parallel}^{\mathrm{sy},\circ}(\boldsymbol{c}'; \kappa')$ denote the moduli space of anchored holomorphic disks $v \colon (D_m, \partial D_m) \to (\mathbb{R} \times Y, \mathbb{R} \times \Lambda)$ such that

- v takes the boundary components labeled by κ_j to the Lagrangian Λ_{κ_j}, and

- v is asymptotic at positive or negative infinity, according to the sign of σ_j, to the unique Reeb chord between Λ_{κ_j} and $\Lambda_{\kappa_{j+1}}$ near c_j at a puncture ζ_j.

If $\boldsymbol{c}$ is a word of strictly increasing (resp. decreasing) Reeb chords then we define the moduli space $\mathcal{M}_{\parallel}^{\mathrm{sy}}(\boldsymbol{c}; \kappa)$ by anchoring also at constant boundary punctures,

$$\mathcal{M}_{\parallel}^{\mathrm{sy}}(\boldsymbol{c}; \kappa) = \bigcup_{\boldsymbol{c} \subset \boldsymbol{c}'} \left(\mathcal{M}_{\parallel}^{\mathrm{sy},\circ}(\boldsymbol{c}'; \kappa') \times \prod_{c_r \in \boldsymbol{c}' \setminus \boldsymbol{c}} \mathcal{M}(c_r) \right),$$

where the union runs over all words $\boldsymbol{c}'$ extending $\boldsymbol{c}$ by constant punctures.
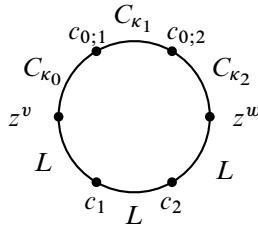
Figure 17: Disk contributing to $\mathcal{M}^{\mathrm{co}}(\boldsymbol{c};\kappa)$.

We will also consider a simpler version with only one copy of the Legendrian $\Lambda = \Lambda_0$ in the case that exactly one puncture in $\boldsymbol{c}$ is positive and all others are negative. In the language above, all punctures of such a disk are constant and we write $\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c})$ for the moduli space of such disks. The (constant) negative punctures of disks in this moduli space are typically not filled by augmentation disks.

Finally, we consider disks in the cobordism and in the filled cobordism. We start with the cobordism disks without filling. Recall that we assume a decomposition $K = C \cup L$, where C has no negative end and L has no positive end. Consider a system of parallel copies $\overline{C} = \{C_j\}_{j=0}^{\infty}$. Consider the disk D_{i+j+2} where we fix two punctures that subdivide the boundary of the disk into two arcs, upper and lower. Let κ be a decreasing boundary numbering of the boundary components in the upper arc and extend it to a constant numbering in the lower arc. Let $\boldsymbol{c}_0 = c_{0;1} \cdots c_{0;j}$ be a composable word of Reeb chords connecting Λ_v to Λ_w, and let $\boldsymbol{c}' = c_i \cdots c_1$ be a word of Reeb chords of Γ. Consider the word of Reeb chords and intersection points

$$\boldsymbol{c} = c_{0;1} \cdots c_{0;j} z^w c_i \cdots c_1 z^v,$$

and let

$$\mathcal{M}^{\mathrm{co}}(\boldsymbol{c};\kappa)$$

denote the moduli space of holomorphic disks $u\colon (D_{i+j+2}, \partial D_{i+j+2}) \to (W, \overline{C} \cup L)$ such that:

- u is asymptotic to the Reeb chord $c_{0;r}$ at its rth constant puncture, takes adjacent boundary arcs to L, and neighboring punctures to the unique intersection point near z^w in $L \cap C_{\kappa_i}^w$ and near z^v in $L \cap C_{\kappa_i}^v$, respectively.

- On remaining boundary arcs and punctures, the boundary maps as described by $\boldsymbol{c}$ and the numbering κ, exactly as above.

The disks in the filled cobordism are entirely analogous. Here we assume that L is a Lagrangian submanifold in $X = X_0 \cup W$. Consider again a system of parallel copies

$\bar{C} = \{C_j\}_{j=0}^{\infty}$ and also a system of parallel copies $\bar{L} = \{L_j\}_{j=0}^{\infty}$ of L. Consider the disk D_{i+j+2} where we fix two punctures that subdivide the boundary of the disk into two arcs, upper and lower. Let κ be a decreasing boundary numbering of the boundary components in the upper and lower arcs. Let $\boldsymbol{x}_0 = x_{0;1} \cdots x_{0;j}$ be a word of intersection points of L and let $\boldsymbol{c}' = c_i \cdots c_1$ be a word of Reeb chords of Γ. Consider the word of Reeb chords and intersection points

$$\boldsymbol{c} = x_{0;1} \cdots x_{0;j} z^w c_i \cdots c_1 z^v,$$

and let

$$\mathcal{M}^{\overline{co}}(\boldsymbol{c}; \kappa)$$

denote the moduli space of holomorphic disks $u \colon (D_{i+j+2}, \partial D_{i+j+2}) \to (X, \bar{C} \cup \bar{L})$ such that:

- u is asymptotic to the intersection point $x_{0;r}$ at its rth puncture in the lower arc, takes adjacent boundary arcs to L, and neighboring punctures to the unique intersection point near z^w in $L_{\kappa_j}^w \cap C_{\kappa_j+1}^w$ and the unique intersection point near z^v in $K_{\kappa_1} \cap C_{\kappa_i+j}^v$.
- On remaining boundary arcs and punctures, the boundary maps as described by $\boldsymbol{c}$ and the numbering κ, exactly as above.

The disks in the cobordism without filling will be used to map into the DG–algebra of the negative end, whereas the disks in the filled cobordism will be used to map into the Floer cohomology of a Lagrangian. This is why we use parallel copies in one case but not the other.

The formal dimension of the moduli spaces above is computed in terms of the negative of a Conley–Zehnder index CZ of the Reeb chords. Recall CZ(a) of a Reeb chord a of Λ as defined for example in [11, Section 2.1]: we pick paths connecting basepoints in the boundary of the various components of L, and paths connecting Reeb chord endpoints to the basepoints. We define CZ(a) to be the Maslov index of this path closed up by a positive rotation in the contact plane, and the grading $|a| = -\text{CZ}(a)$. For a Lagrangian intersection x between L^1 and L^2 we similarly pick paths connecting to the basepoints and use these to form a loop γ starting in L^2 and ending in L^1, and define CZ(x) to be the Maslov index of the loop of Lagrangian planes that results from closing up the path of Lagrangian planes along γ by a positive rotation, and $|x| = -\text{CZ}(x)$. The Conley–Zehnder index is independent of the basepoint paths since the Maslov class vanishes.

Remark 72 The above gradings are related to the grading $|\cdot|_{\mathrm{Leg}}$ in the Legendrian contact homology algebra [28; 11] as

$$|c|_{\mathrm{Leg}} = -|c| - 1.$$

Gradings generally depend on the choice of paths connecting endpoints to the basepoint: two such choices differ by a loop and the grading is shifted by the Maslov index of that loop. In particular, if the Maslov class vanishes the grading is well defined. Also, the paths connecting tangent planes at basepoints in different components are defined only up to choice. Changing the homotopy class shifts the Maslov potential between components and indices of mixed Reeb chords accordingly.

Lemma 73 *The formal dimension of the moduli space $\mathcal{M}^{\mathrm{fi}}(\boldsymbol{a};\kappa)$ is given by*

$$\dim \mathcal{M}^{\mathrm{fi}}(\boldsymbol{a};\kappa) = (n-3) - \sum_{j=1}^{m}(|a_j| - (n-2)).$$

The formal dimension of the moduli spaces $\mathcal{M}^{\mathrm{sy}}_{\|}(\boldsymbol{c};\kappa)$ is given by

$$\dim \mathcal{M}^{\mathrm{sy}}_{\|}(\boldsymbol{c};\kappa) = (n-3) + \sum_{\sigma_j=-1}(|c_j| + 1) - \sum_{\sigma_j=+1}(|c_j| - (n-2)).$$

The formal dimension of the moduli spaces $\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c})$ is given by

$$\dim \mathcal{M}^{\mathrm{sy}}(\boldsymbol{c}) = (n-3) + \sum_{\sigma_j=-1}(|c_j| + 1) - \sum_{\sigma_j=+1}(|c_j| - (n-2)).$$

The formal dimension of the moduli space $\mathcal{M}^{\mathrm{co}}(\boldsymbol{c};\kappa)$ is given by

$$\dim \mathcal{M}^{\mathrm{co}}(\boldsymbol{c};\kappa) = 1 - \sum_{r=1^i}(|c_r| - (n-2)) + \sum_{s=1}^{j}(|c_{0;s}| + 1).$$

The formal dimension of the moduli space $\mathcal{M}^{\overline{\mathrm{co}}}(\boldsymbol{c};\kappa)$ is given by

$$\dim \mathcal{M}^{\overline{\mathrm{co}}}(\boldsymbol{c};\kappa) = 1 - \sum_{r=1^i}(|c_r| - (n-2)) - \sum_{s=1}^{j}(|x_{0;s}| - (n-2)).$$

Proof See [17, Theorem A.1]. □

We next study topological properties of the moduli spaces just defined. It turns out to be comparatively simple because of two key features. First, since we require our disks to switch copies at punctures "in the same direction" they cannot be multiply covered,

and second, for the same reason there can be no boundary splitting. As in [23], the first property allows us to prove transversality by perturbing the almost complex structure, and the second shows that the moduli spaces admit compactifications consisting only of punctured curves joined at Reeb chords or Lagrangian intersection points. Precise formulations of these results are as follows.

Theorem 74 *For a generic almost complex structure J the moduli spaces $\mathcal{M}^{\mathrm{fi}}(\boldsymbol{a};\kappa)$, $\mathcal{M}^{\mathrm{sy}}_{\|}(\boldsymbol{c};\kappa)$, $\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c})$, $\mathcal{M}^{\mathrm{co}}(\boldsymbol{c};\kappa)$ and $\mathcal{M}^{\overline{\mathrm{co}}}(\boldsymbol{c};\kappa)$ are transversely cut out manifolds of respective dimensions $\dim\mathcal{M}^{\mathrm{fi}}(\boldsymbol{a};\kappa)$, $\dim\mathcal{M}^{\mathrm{sy}}_{\|}(\boldsymbol{c};\kappa)$, $\dim\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c})$, $\dim\mathcal{M}^{\mathrm{co}}(\boldsymbol{c};\kappa)$ and $\dim\mathcal{M}^{\overline{\mathrm{co}}}(\boldsymbol{c};\kappa)$.*

Proof A well-known argument gives transversality for disks that are somewhere injective on the boundary by perturbing the almost complex structure: any element in the cokernel of the linearized operator must be zero on the set of injectivity and then identically zero by unique continuation. Our disks are not necessarily somewhere injective but the disks cannot be multiple covers and there is a region with the property of the region of injectivity above. We briefly recall the argument for this from [28, Lemma 4.5]. Fix a puncture of u.

Consider first the more difficult case when this puncture maps to a Lagrangian intersection. Pick coordinates so that the intersection point lies at the origin in $\mathbb{C}^n$ and so that the two Lagrangians correspond to $\mathbb{R}^n$ and $i\mathbb{R}^n$. Let $(x_1 + iy_1, \ldots, x_n + iy_n)$ be standard coordinates on $\mathbb{C}^n$. Consider the complex hyperplanes $H_{\pm\epsilon} = \{x_1 + iy_1 = \pm\epsilon(1+i)\}$. Looking at the Fourier expansion of u near the puncture it is clear that for suitable coordinates (such that the leading Fourier coefficient of u lies in the direction of the first coordinate) the number of intersection points of the image of u and $H_{\pm\epsilon}$ near the puncture have different parities depending on the sign of ϵ. With more details: choose coordinates in $\mathbb{C}^n$ for which the complex structure J agrees with the standard almost complex structure at the origin and coordinates $[0,\infty) \times [0,1]$ around the puncture in the domain. Then the argument in the proof of [25, Lemma 2.1] shows that u is conjugate to a standard holomorphic map $\widetilde{u}$ with boundary condition $\mathbb{R}^n$ and $i\mathbb{R}^n$. That map has Fourier expansion

$$\widetilde{u}(s+it) = \sum_{k\geq 0} c_k e^{-(k\pi+\frac{1}{2}\pi)(s+it)}, \quad \text{with } c_k \in \mathbb{R}^n.$$

For the intersection with $H_{\pm\epsilon}$ pick coordinates so that the first Fourier coefficient c_k which is nonzero has the form $c_k = (a, 0, \ldots, 0)$, with $a \neq 0$. By analytic continuation,

other disks and half-disks mapping there either have images agreeing completely in the ball or there are injective points of the disk near the puncture. In the case where they agree completely, we note that the parity of the number of intersection points in each local sheet not containing the puncture with $H_{\pm\epsilon}$ is independent of the sign of $\pm\epsilon$. We then find that we can achieve transversality by perturbing the complex structure near $H_{\pm\epsilon}$: because the sheet with the puncture intersects only one of $H_{\pm\epsilon}$, if the contributions of the sheets mapping to $H_{-\epsilon}$ cancel then those mapping to $H_{+\epsilon}$ cannot cancel and vice versa, by unique continuation.

Once transversality is achieved the statement that solutions form manifolds follows from a well-known argument; see eg [28, Proposition 2.3]. □

Theorem 75 *The moduli space $\mathcal{M}^{\mathrm{fi}}(\boldsymbol{a}; \kappa)$ admits a compactification consisting of several-level disks joined at Reeb chords and intersection points, where some levels may lie in the symplectization.*

The moduli spaces $\mathcal{M}^{\mathrm{sy}}_{\|}(\boldsymbol{c}; \kappa)$ and $\mathcal{M}^{\mathrm{sy}}(\boldsymbol{c})$ admit compactifications consisting of several-level disks joined at Reeb chords.

The moduli space $\mathcal{M}^{\mathrm{co}}(\boldsymbol{a}; \kappa)$ admits a compactification consisting of several-level disks joined at Reeb chords. There is one level of disks in the cobordisms and the remaining levels are in the symplectization ends.

The moduli space $\mathcal{M}^{\overline{\mathrm{co}}}(\boldsymbol{a}; \kappa)$ admits a compactification consisting of several-level disks joined at Reeb chords and intersection points.

Proof The boundary conditions on our punctured disks have the following property: any arc in a disk with more than one positive puncture that subdivides the source into two components with a positive puncture in each must connect boundary components numbered with distinct numbers. This shows that there can be no boundary splitting. The theorem then follows from SFT compactness; see [22, Appendix B.1] for the curve with boundary version of [12]. □

We next discuss orientations of moduli spaces following [27]. We fix capping operators at all Reeb chords and Lagrangian intersection points so that the two capping operators there glue to a disk with the Fukaya orientation; see [36]. Recall that the relative spin structure on the Lagrangian submanifold induces an orientation on the determinant bundle over the space of disks with boundary condition in the Lagrangian; see [27] or

[28, Section 4.4]. As in [27] we see that these choices then induce a system of coherent orientations on the moduli spaces.

We will use one more property of the moduli spaces, which says that they are effectively independent of the increasing or decreasing boundary labeling κ.

Theorem 76 *Let κ and κ' be two increasing (decreasing) boundary numberings. Then there are canonical orientation-preserving diffeomorphisms*

$$\mathcal{M}^{\mathrm{fi}}(\boldsymbol{a};\kappa) \approx \mathcal{M}^{\mathrm{fi}}(\boldsymbol{a};\kappa'), \quad \mathcal{M}^{\mathrm{sy}}_{\parallel}(\boldsymbol{c};\kappa) \approx \mathcal{M}^{\mathrm{sy}}_{\parallel}(\boldsymbol{c};\kappa'), \quad \mathcal{M}^{\mathrm{co}}(\boldsymbol{c};\kappa) \approx \mathcal{M}^{\mathrm{co}}(\boldsymbol{c};\kappa').$$

Proof Let $\mathcal{M}(\kappa)$ denote either one of the above moduli spaces. This moduli space is the transverse zero set of a Fredholm section in a Banach bundle. Changing the numbering from κ to κ' corresponds to an arbitrarily small isotopy, which induces an arbitrarily small deformation of the section. The theorem follows. □

Appendix B   Wrapped Floer cohomology and Legendrian surgery

In this section we present the argument that establishes the isomorphism between $\mathrm{CE}^*(\Lambda)$ and $\mathrm{CW}^*(C)$, where C is the cocore disk of the surgery. Our proof is a generalization of the corresponding result under Lagrangian handle attachment explained in [11], and uses the technical results on relevant moduli spaces in [25].

We first define a version of wrapped Floer cohomology using only purely holomorphic disks and show that the resulting theory agrees with the usual version defined in terms of holomorphic disks with a Hamiltonian term. Second we discuss the surgery isomorphism in [11], and third we discuss how to generalize that argument to partially wrapped Floer cohomology calculations.

B.1   Wrapped Floer cohomology without Hamiltonian

Let X be a Weinstein manifold and L be an exact Lagrangian. Fix a system of shifting Morse functions that are positive at infinity and let $\overline{L} = \{L_j\}_{j=0}^{\infty}$ be the corresponding family of parallel Lagrangian submanifolds. Define $\mathrm{CW}^*(L)$ to be the chain complex generated by Reeb chords of L and intersection points $L_0 \cap L_1$. We define operations $\mathfrak{m}_i$ on $\mathrm{CW}^*(L)$ using what we call *partial holomorphic buildings*.

We start in the simplest case when the output of $\mathfrak{m}_i$ is an intersection point c_0. Consider i generators $c_i, \ldots, c_1$ and consider a disk D_{i+1} with a decreasing boundary numbering κ,

distinguished negative (output) puncture and remaining punctures positive (inputs). Let $c' = c_i \cdots c_1$ and $c = c_0 c_i \cdots c_1$. Define

$$\mathfrak{m}'_i(c') = \sum_{|c_0| = |c'| + (2-i)} |\mathcal{M}^{\mathrm{fi}}(c; \kappa)| c_0.$$

Here we use the temporary notation $\mathfrak{m}'_i$ to denote the summand of the full operation $\mathfrak{m}_i$ that takes values in intersection points. We next turn to the more complicated definition of the part $\mathfrak{m}''_i$ of the operation that takes values in Reeb chord generators, and to this end we introduce the notion of a partial holomorphic building.

The domain of a partial holomorphic building is a possibly broken disk D_{i+1} with decreasing boundary numbering κ. The partial holomorphic buildings we consider always have exactly one disk in the symplectization. We call it the primary disk of the building. We require that the distinguished puncture is increasing and is a negative puncture of this primary disk. If the distinguished puncture is the only negative puncture of the primary disk then the partial building consists only of its primary component. If on the other hand the primary disk has additional negative punctures then we require that at each additional negative puncture (which is decreasing or constant) there is a disk in the filling with decreasing boundary condition that is attached at its distinguished increasing or constant puncture to the additional negative puncture. We call these disks the secondary disks of the partial building. The resulting partial holomorphic building is then a disk with domain a broken D_{i+1}, with distinguished puncture a negative puncture at a Reeb chord and with remaining i punctures either Reeb chords or intersection points. See Figure 18.

Remark 77 At additional negative punctures there may be holomorphic disks with one positive puncture and boundary on L attached. These are the usual augmentation disks, or disks on L used as anchoring disks in the definition of $\mathcal{M}^{\mathrm{sy}}_{\parallel}(c, \kappa)$.

We write the punctures of the partial holomorphic disk building as $c = c_0 c_i \cdots c_1$, where c_0 is the distinguished puncture. Write

$$\mathcal{M}^{\mathrm{pb}}(c; \kappa)$$

for the moduli space of partial holomorphic disk buildings with boundary condition on $\overline{L}$ according to κ. Using this we define for generators $c' = c_i \cdots c_1$ the operation

$$\mathfrak{m}''_i(c') = \sum_{|c_0| = |c'| + (2-i)} |\mathcal{M}^{\mathrm{pb}}(c; \kappa)| c_0,$$
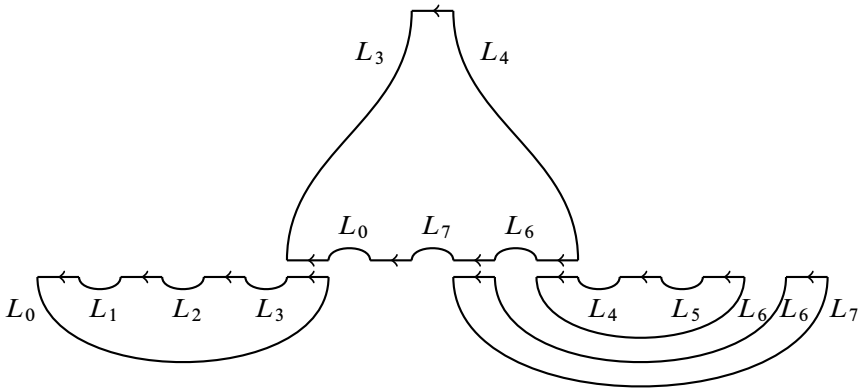
Figure 18: The domain of a partial building contributing to the operation $\mathfrak{m}_7$ of $\mathrm{CW}^*(L)$ with a possible decoration. The map sends the tensor product of chords $L_0 \leftarrow L_1, L_1 \leftarrow L_2, \ldots, L_6 \leftarrow L_7$ to a chord $L_0 \leftarrow L_7$.

where the sum ranges over Reeb chords c_0 with grading as indicated. Finally we define the total operation $\mathfrak{m}_i$ as the sum

$$\mathfrak{m}_i(c') = \mathfrak{m}_i'(c') + \mathfrak{m}_i''(c').$$

Lemma 78 *The A_∞–relations hold for the operations $\mathfrak{m}_i$.*

Proof First, Theorem 76 shows that the operations compose and that they are independent of the choice of decreasing boundary numbering. To see that the relations hold, we will as usual identify the terms contributing to them with the boundary of an oriented 1–dimensional compact manifold.

To this end we first consider 1–dimensional moduli spaces $\mathcal{M}'$ of the form $\mathcal{M}' = \mathcal{M}^{\mathrm{fi}}(c; \kappa)$, where the distinguished puncture c_0 is an intersection point. As usual, the boundary numbering precludes boundary bubbling and we find that the boundary consists of broken disks that either break at an intersection point, in which case the holomorphic parts both have dimension zero, or break into a partial holomorphic building with a rigid disk attached at its negative puncture, in which case the primary component of the partial building has dimension one. We find the boundary points of $\mathcal{M}'$ are in one-to-one correspondence with disks contributing to compositions of $\mathfrak{m}_i'$ and $\mathfrak{m}_j'$ (disks breaking at intersection points) and disks contributing to $\mathfrak{m}_i''$ and $\mathfrak{m}_j'$.

The remaining contributions to the A_∞–relations correspond to compositions of $\mathfrak{m}_i''$ and $\mathfrak{m}_j''$. We show that all contributions to this composition constitute the boundary of

an oriented 1–manifold. The contributions are of two forms: either the output puncture of the first operation (which lies in the primary disk of the corresponding partially broken configuration) is glued to an input puncture of the primary disk in the partially broken configuration of the second operation, or it is glued to an input puncture in a secondary disk.

The configurations of the former type correspond to a part of the boundary of the moduli space $\mathcal{M}_{\parallel}^{\mathrm{sy}}(\boldsymbol{b}; \kappa)$ of dimension two with distinguished negative puncture and decreasing boundary numbering (after we divide out the natural $\mathbb{R}$–action this is a 1–dimensional space) capped off by rigid disks in $\mathcal{M}^{\mathrm{co}}(\boldsymbol{a})$ at all nondistinguished negative punctures. This is the part of the boundary where the distinguished puncture belongs to the lower level disk.

The configurations of the second type correspond to the part of the boundary of the 1–dimensional moduli space $\mathcal{M}^{\mathrm{co}}(\boldsymbol{a}; \kappa)$, with a distinguished increasing positive puncture where a negative puncture in the primary disk of the second operation is attached (other negative punctures in the primary disk of the second operation are capped off as usual). The part of the boundary containing the distinguished positive puncture lies in the rigid disk in the cobordism.

Finally, the remaining part of the boundary in the first case is two-level buildings in $\mathcal{M}_{\parallel}^{\mathrm{sy}}(\boldsymbol{b}; \kappa)$, where the distinguished negative puncture belongs to the top-level curves. These are exactly the configurations that we get from the remaining parts of the boundary in the second case (ie configurations where the distinguished puncture belongs to the component in the symplectization) when we glue to it the primary disk of the second operation.

We conclude that also the composition of $\mathfrak{m}_i''$ and $\mathfrak{m}_j''$ cancels. The lemma follows. $\square$

B.1.1 Isomorphism with the Hamiltonian version

In this section we show that the above definition of wrapped Floer cohomology agrees with the standard theory. Similar results can be found in [30; 29]. Here we will give a sketch. We keep the geometric setting as above and write $\mathrm{CW}_{\mathrm{Ham}}^*(L)$ for the usual version of Hamiltonian wrapped Floer cohomology. We give a brief recollection of the definition.

We define the wrapped Floer cohomology complex $\mathrm{CW}_{\mathrm{Ham}}^*(L)$ of L as follows. Write $X = \overline{X} \cup [0, \infty) \times Y$, where $\overline{X}$ is a compact domain and $[0, \infty) \times Y$ the positive end of the Weinstein manifold X. Consider time-dependent Hamiltonians $H_a \colon X \times [0, 1] \to \mathbb{R}$

which are perturbations of functions that equal 0 on $\overline{X}$ and are linear of the form

$$(t, y) \mapsto ae^t + b \quad \text{on } [0, \infty) \times Y,$$

where a is not in the chord and orbit spectrum or the contact form on Λ. We choose these Hamiltonians in such a way that if $a_0 < a_1$ then $H_{a_0} < H_{a_1}$ on X. After a small perturbation, Hamiltonian time 1 chords and Hamiltonian time 1 orbits are nondegenerate.

Define the chain complex $\mathrm{CW}^*_{\mathrm{Ham}}(L, H_a)$ to be generated by Hamiltonian chords $\gamma: [0, 1] \to X$ of C of action

$$\mathfrak{a}(\gamma) = \int_0^1 \big(\lambda(\dot\gamma(t)) - H_a(\gamma(t))\big) \, dt < a.$$

The differential on $\mathrm{CW}^*(L, H_a)$ is defined by counting solutions of the perturbed Cauchy–Riemann equation over the strip with coordinates $s + it \in \mathbb{R} \times [0, 1]$:

$$(du + X_{H_a} \otimes dt)^{0,1} = 0.$$

Choosing an increasing interpolation between H_{a_0} and H_{a_1} we get continuation maps

$$\mathrm{CW}^*_{\mathrm{Ham}}(L, H_{a_0}) \mapsto \mathrm{CW}^*_{\mathrm{Ham}}(L, H_{a_1}),$$

and we define the wrapped Floer cohomology complex as the direct limit

$$\mathrm{CW}^*_{\mathrm{Ham}}(L) = \varinjlim_a \mathrm{CW}^*_{\mathrm{Ham}}(C, H_a).$$

The wrapped Floer cohomology $\mathrm{HW}^*_{\mathrm{Ham}}(C)$ is the homology of this complex. Writing $\mathrm{HW}^*_{\mathrm{Ham}}(L, H_a)$ for the homology of $\mathrm{CW}^*_{\mathrm{Ham}}(L, H_a)$ we then have

$$\mathrm{HW}^*_{\mathrm{Ham}}(L) = \varinjlim_a \mathrm{HW}^*_{\mathrm{Ham}}(L, H_a),$$

by exactness of direct limits.

A well-known argument shows that $\mathrm{CW}^*(L)$ with differential $\mathfrak{m}_1$ is quasi-isomorphic to the wrapped Floer cohomology by a geometrically defined chain map [29]. We recall the argument here.

The filling L of Λ gives an augmentation of $\mathrm{CE}^*(\Lambda)$ and we define $\mathrm{CW}^*(L)$ (as a chain complex disregarding higher product operations) without Hamiltonian as the "Morse extended linearized Chekanov–Eliashberg complex" with respect to this augmentation as generated by Reeb chords and the critical point of a Morse function on L with a unique minimum, and take the differential to count unperturbed augmented and anchored

holomorphic strips. We also introduce the subcomplexes $CW^*(L, a)$ generated by chords of action $< a$. Then, by definition,

$$CW^*(L) = \varinjlim_a CW^*(L, a).$$

The isomorphism $CW^*(L) \to CW^*_{Ham}(L)$ is now constructed by interpolating exactly as in the continuation maps above from the zero Hamiltonian (ordinary Cauchy–Riemann equation) to the Hamiltonians H_a above. Choosing the interpolations compatibly, we get the commutative diagram

$$
\begin{array}{ccccccc}
CW^*(L, a_0) & \longrightarrow & CW^*(L, a_1) & \longrightarrow \cdots \longrightarrow & CW^*(L, a_j) & \longrightarrow \cdots \\
\downarrow & & \downarrow & & \downarrow & \\
CW^*_{Ham}(L, H_{a_0}) & \longrightarrow & CW^*_{Ham}(L, H_{a_0}) & \longrightarrow \cdots \longrightarrow & CW^*_{Ham}(L, H_{a_j}) & \longrightarrow \cdots
\end{array}
$$

Here all vertical arrows are chain isomorphisms by the standard argument — see for instance [30, Section 6] — and taking limits we find a chain isomorphism

$$CW^*(L) \to CW^*_{Ham}(C).$$

We extend this chain map to an A_∞–map, then the standard spectral sequence argument establishes the desired A_∞–quasi-isomorphism.

We follow the approach in [30], where similar isomorphisms between contact and symplectic differential graded algebras were constructed. More precisely, we construct a splitting compatible nonnegative field of 1–forms with values in Hamiltonian vector fields, and further a 1–parameter family of such forms interpolating between the zero Hamiltonian at the positive end and the Hamiltonian used to define wrapped Floer cohomology at the negative end [30, Section 2]. We then define the corresponding moduli spaces over the deformation interval. Keeping the notation from [30] we write

$$\mathcal{F}_{\mathbb{R}}(\boldsymbol{a}, b).$$

In order for the asymptotics at infinity of these maps to make sense we need to include the parallel copies of the Lagrangians according to boundary numbering, and in particular also to incorporate this in the description of wrapped Floer cohomology. More precisely, as in the case above we will have moduli spaces of Floer holomorphic disks with boundary in distinct Lagrangians that are arbitrarily close. The analogue of Theorem 76 holds by the same argument and the corresponding moduli spaces are canonically isomorphic for sufficiently small perturbations. Using these observations

we then define $\Phi\colon \mathrm{CW}^*(L) \to \mathrm{CW}^*_{\mathrm{Ham}}(L)$ by

$$\Phi(\boldsymbol{a}) = \sum_{\dim \mathcal{F}_{\mathbb{R}}(\boldsymbol{a},b)=0} |\mathcal{F}_{\mathbb{R}}(\boldsymbol{a},b)|\, b.$$

Lemma 79 *The map Φ is an A_∞–homomorphism.*

Proof To see this we note again that the disks which contributes to the A_∞ relations correspond exactly to the ends of 1–dimensional moduli space. $\square$

Lemma 80 *The map Φ is a quasi-isomorphism.*

Proof The map respects the word-length filtration and is the standard isomorphism from the linearized Legendrian cohomology to the wrapped Floer cohomology, discussed above, on the E_2–page. $\square$

B.2 Wrapped Floer cohomology and Lagrangian handle attachment

In this subsection we prove the results in [11] giving a Legendrian surgery description of the wrapped Floer cohomology of a cocore disk in a Weinstein manifold obtained by Lagrangian handle attachment along a Legendrian sphere, referring to [25] for the results on holomorphic curves missing in [11]. To state this result we first introduce notation.

Suppose that X_0 is a Weinstein $2n$–manifold with ideal boundary the contact $(2n-1)$–manifold Y_0. Let $\Lambda = \Lambda_1 \cup \cdots \cup \Lambda_m$ be a Legendrian submanifold such that all of its components Λ_j are parametrized $(n-1)$–spheres. Let X be the Weinstein manifold that results from attaching Lagrangian handles H to Λ. Here $H = H_1 \cup \cdots \cup H_m$, where each component H_j is a disk subbundle of the cotangent bundle T^*D of the n–disk D, and where H_j is attached to Λ_j. Then X contains m cocore disks corresponding to the cotangent fibers at the center of the disk in each H_j. We let $C_j \subset X$ denote the cocore disk in H_j, let $\Gamma_j \subset Y$ denote its Legendrian boundary inside the contact boundary Y of X, and write $\Gamma = \Gamma_1 \cup \cdots \cup \Gamma_m$.

As a first step in the calculation of the wrapped Floer cohomology of C we describe the generators of the underlying chain complex. By definition — see Section B.1 — generators of $\mathrm{CW}^*(C)$ are of two kinds: Lagrangian intersection points and Reeb chords. Here the Lagrangian intersection points are easily understood: pick the shifting Morse function so that it has one minimum on each component of C and no other critical
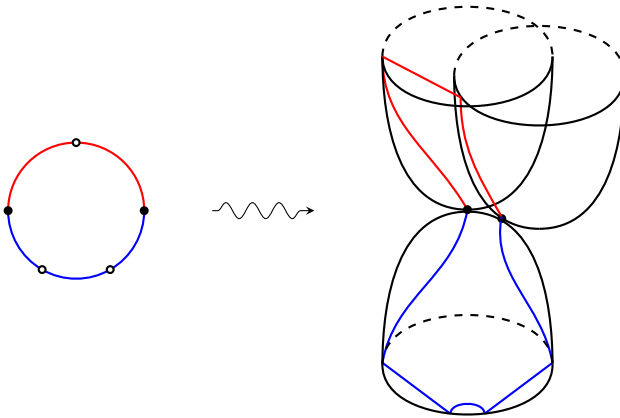
Figure 19: A picture illustrating a curve contributing to Φ^1 of the A_∞–functor Φ.

points, then there is exactly one intersection point for each component of C. We denote the intersection point of C_j by m_j and we denote the subcomplex generated by the m_j by $\mathrm{CW}^*_0(\Gamma)$. Remaining generators are Reeb chords of Γ we write $\mathrm{CW}^*_+(\Gamma)$ for the quotient complex $\mathrm{CW}^*(\Gamma)/\mathrm{CW}^*_0(\Gamma)$ and note that CW^*_+ is generated by Reeb chords.

Consider the link Λ and let all components be decorated by minus, $\Lambda^- = \Lambda$. Consider $\mathrm{CE}^*(\Lambda)$ as a chain complex, generated by composable words of Reeb chords with differential d and with product $\cdot$ given by concatenation if the words are composable and zero otherwise. Let $\epsilon > 0$ denote the size of the attaching region, ie the size of the tubular neighborhood of Λ where H is attached. We then have the following:

Lemma 81 *For any $A > 0$ there exists $\epsilon_0 > 0$ such that if $\epsilon < \epsilon_0$ then there is a natural one-to-one correspondence between the generators of $\mathrm{CW}^*_+(\Gamma)$ (Reeb chords of Γ) of action $< A$ and the generators of $\mathrm{CE}^*(\Lambda)$ (words of Reeb chords of Λ) of action $< A$.*

Proof This is [25, Theorem 1.2]. $\qquad\square$

We will next define the surgery map, which is an A_∞–morphism

$$\Phi \colon \mathrm{CW}^*(\Gamma) \to \mathrm{CE}^*(\Lambda)$$

that counts certain holomorphic disks. See Figure 19.

As in Appendix A, consider the disk D_{i+j+2} with two special punctures subdividing the boundary into an upper and a lower arc with i and j punctures, respectively, and with a boundary numbering in the upper arc. Let $\boldsymbol{c}_0 = c_{0;1} \cdots c_{0;j}$ be a composable

word of Reeb chords connecting Λ_v to Λ_w, and let $c_i \cdots c_1$ be a word of generators of $\mathrm{CW}^*(C)$. Consider the word of Reeb chords and intersection points

$$c = c_{0;1} \cdots c_{0;j} z^w c_i \cdots c_1 z^v.$$

Define $\Phi_i : \mathrm{CW}^*(C)^{\otimes i} \to \mathrm{CE}^*(\Lambda)$ by

$$\Phi_i(c') = \sum_{|c_0| = |c'| + i(n-2)} |\mathcal{M}^{\mathrm{co}}(c)| c_0.$$

Remark 82 If m^v is the minimum of the Morse function on C^v as above, then

$$\Phi_1(m^v) = e_v,$$

because of the unique holomorphic disk corresponding to the flow line from the minimum to the intersection point between $C^v \cap L$; for the parallel copies this gives a triangle with corners at $m^v = C_0^v \cap C_1^v$, at $C_0^v \cap L$ and at $C_1^v \cap L$, and since there are no negative punctures the output is e_v. Also, if a word c' of generators of $\mathrm{CW}^*(C)$ contains a generator m^v and has length $i > 1$, then

$$\Phi_i(c') = 0,$$

as this corresponds — see Lemma 35 — to a holomorphic disk with a flow line from the minimum attached, and such a configuration cannot be rigid unless the disk is constant.

Theorem 83 *The maps Φ_i give an A_∞–map $\mathrm{CW}^*(C) \to \mathrm{CE}^*(\Lambda)$, which is an A_∞–quasi-isomorphism.*

Proof In order to see the A_∞–relations we study the boundary of the moduli space $\mathcal{M}^{\mathrm{co}}(c)$. As usual the boundary numbering guarantees that there is no boundary splitting on C. The fact that there is no boundary splitting on L follows from Stokes' theorem: such a splitting would give a disk without positive puncture. The boundary of the moduli space thus consists of the following configurations:

   (i)   Two level curves with one level in the cobordism and one in either symplectization end.

  (ii)   Curves which split at the intersection point $C \cap L$.

Splitting (i) corresponds to the map followed by the operation d in $\mathrm{CE}^*(\Lambda)$ when the symplectization disk lies in the negative end, and to an operation in $\mathrm{CW}^*(C)$ followed by the map when the symplectization disk lies in the positive end. Splitting (ii) corresponds to the tensor product of the map followed by the product operation $\cdot$ in $\mathrm{CE}^*(\Lambda)$. The A_∞–relations follow.

To see that Φ is a quasi-isomorphism we argue as follows. We first fix an action cut-off $A > 0$ and show that Φ_1 induces an isomorphism on homology below action A by constructing algebraically one holomorphic disk interpolating between a Reeb chord of Γ and the corresponding word of Reeb chords of Λ. For a complete proof we refer to [25, Theorem 1.3]; the argument is roughly as follows. One starts from unique and uniformly transversely cut out such disks for single chord words obtained by a straightforward explicit geometric construction. Gluing such disks at their Lagrangian intersection punctures in $L \cap C$ and using small action to rule out all breakings except one, we find that there is algebraically one disk interpolating between a chord on Γ and the corresponding word of chords of Λ. Together with Remark 82, which shows $\Phi_1(m^v) = e^v$, the existence of such disks implies that the map Φ_1 has a triangular matrix with respect to the action filtration and hence is a chain isomorphism (compare [11, Section 6.2]): since Φ_1 is an isomorphism on the subquotients of the action filtration (and the isomorphism on generators is given by the bijection given in Lemma 81), Φ_1 is an isomorphism below action A for any A. The A_∞–isomorphism below action $A > 0$ then follows from the usual spectral sequence argument.

To see that we get an isomorphism on the full complex we show that the isomorphisms discussed are compatible with action limits. More precisely, in order to increase the action limit $A > 0$ for the one-to-one correspondence between Reeb chords of Γ of action $< A$ and words of Reeb chords of Λ of action $< A$, we must shrink the size $\delta > 0$ of the handle attached. Consider attaching a handle of size $\delta > 0$ to Λ and denote the resulting new Weinstein manifold by X_δ and the cocore disk by $C_\delta \subset X_\delta$.

If $\delta_0 > \delta_1$, then by the isomorphism in Lemma 80 and standard results for wrapped Floer cohomology — see eg [30, Section 5.5] — there is a cobordism map

$$\mathrm{CW}^*_+(C_{\delta_0}) \to \mathrm{CW}^*_+(C_{\delta_1}),$$

which is a quasi-isomorphism. Moreover, by the surgery description of chords for any $A > 0$ there exists $\delta_1 > 0$ such that the above map has ± 1 on the diagonal (with respect to the identification of generators in Lemma 81) for all chords of Γ and words of chords of Λ of action $< A$.

Consider the directed system

$$(43) \qquad \mathrm{CW}^*_+(C_{\delta_0}) \to \mathrm{CW}^*_+(C_{\delta_1}) \to \cdots \to \mathrm{CW}^*_+(C_{\delta_j}) \to \cdots,$$

where $\delta_j \to 0$, and let

$$\overline{\mathrm{CW}}^*_+(C) = \varinjlim_\delta \mathrm{CW}^*_+(C_\delta).$$

Then the homology $\overline{HW}{}^*_+(C)$ of $\overline{CW}{}^*_+(C)$ satisfies

$$\overline{HW}{}^*_+(C) = \varinjlim_\delta HW^*_+(C_\delta) = HW^*_+(C_{\delta_j}) \quad \text{for any fixed } j.$$

Here the last equality follows since all the arrows in the directed system of homology groups of (43) are isomorphisms.

Consider next the Chekanov–Eliashberg algebra $CE^*(\Lambda)$ of Λ. We define the action-truncated subcomplex $CE^*(\Lambda, a)$ generated by words of chords of total action $< a$. Viewing $CE^*(\Lambda)$ as a chain complex generated by words of chords, we then have

$$CE^*(\Lambda) = \varinjlim_a CE^*(\Lambda, a).$$

For each a_j the surgery map gives $\delta_j > 0$ such that the map

$$\overline{CW}{}^*_+(C_{\delta_j}) \to CE^*(\Lambda, a_j)$$

is a chain isomorphism with ± 1 on the diagonal below action a_j. By definition of surgery and cobordism maps, the diagram

(44)
$$
\begin{array}{ccccccc}
\overline{CW}{}^*_+(C_{\delta_0}) & \longrightarrow & \overline{CW}{}^*_+(C_{\delta_1}) & \longrightarrow \cdots \longrightarrow & \overline{CW}{}^*_+(C_{\delta_j}) & \longrightarrow & \cdots \\
\downarrow & & \downarrow & \vdots & \downarrow & & \\
CE^*(\Lambda, a_0) & \longrightarrow & CE^*(\Lambda, a_1) & \longrightarrow \cdots \longrightarrow & CE^*(\Lambda, a_j) & \longrightarrow & \cdots
\end{array}
$$

commutes, where $\delta_{j+1} < \delta_j$ and $a_j < a_{j+1}$. Taking limits of the sequences we get a chain map

(45)
$$\overline{CW}{}^*_+(C) \to CE^*(\Lambda).$$

Taking the limits of the sequence (44) on the homology level and using that all vertical arrows are homology isomorphisms then gives homology isomorphisms in the limit, and (45) is a quasi-isomorphism inducing an isomorphism

$$\overline{HW}{}^*(C) \approx HCE^*(\Lambda).$$

The above gives a homology isomorphism of chain complexes. To consider also products one uses the exact same argument. The product operation on $\overline{CW}{}^*(C)$ is induced from the action-truncated version

$$\overline{CW}{}^*(C, a_1) \otimes \cdots \otimes \overline{CW}{}^*(C, a_m) \to \overline{CW}{}^*(C, a_1 + \cdots + a_m),$$

and similarly, on CE^*,

$$CE^*(C, a_1) \otimes CE^*(C, a_2) \to CE^*(C, a_1 + a_2). \qquad \square$$

Remark 84 In the above proof we obtain the isomorphism by taking smaller and smaller handles. To see that such a procedure is necessary note that the correspondence between words of chords and chords is true only below an action limit determined by the size of the handle. For larger actions there are Reeb flows before the surgery that hit the neighborhood of the Legendrian without being close to a chord and which could give chords after the surgery.

Remark 85 There is also an "upside-down" perspective on the surgery just described. Namely, one can start from the contact manifold Y and produce the contact manifold Y_0 by doing so-called $+1$–surgery on Γ. In complete analogy with the above, one shows that Reeb chords on Λ are in natural one-to-one correspondence with words of Reeb chords on Γ, and one can construct an upside-down surgery map of A_∞–coalgebras,

$$\mathrm{BCW}^*(C) \to \mathrm{LC}_*(\Lambda).$$

A similar argument also shows that this map is a quasi-isomorphism. Alternatively, one can prove this from the original surgery map using only algebra as follows. First write $\mathrm{CE}^*(\Lambda) = \Omega\mathrm{LC}_*(\Lambda)$. Then

$$\mathrm{BCW}^*(C) \simeq \mathrm{B}\Omega\mathrm{LC}_*(\Lambda) \simeq \mathrm{LC}_*(\Lambda),$$

since $\mathrm{LC}_*(\Lambda)$ is conilpotent; see Section 2.2.2.

B.3 Legendrian surgery and stopped wrapping

In this section we outline a surgery approach to the computation of wrapped Floer cohomology in a Weinstein manifold X with wrapping stopped by a Legendrian Λ in its boundary. We will use the following model for the ambient manifold. Fix a tubular neighborhood of Λ in the contact boundary Y of X. Attach a disk-bundle neighborhood of the zero section in $T^*([0,\infty) \times \Lambda)$ along the boundary $T^*([0,\infty) \times \Lambda)|_{0\times\Lambda}$, just like in Lagrangian handle attachment. We use a Liouville vector field on this domain that agrees with the standard Liouville vector field pointing outwards along fibers in the cotangent bundle over $[T,\infty) \times \Lambda$ for some $T > 0$. Let the components of Λ be denoted by Λ_v for $v \in Q_0$. Fix a basepoint $p_v \in \Lambda_v$ for each v. Let $C^{v;\tau}$ denote the cotangent fiber $T^*_{(p_v,\tau)}([0,\infty) \times \Lambda)$. We compute the wrapped Floer cohomology of $C^\tau = \bigcup_{v\in Q_0} C^{v;\tau}$ for sufficiently large τ using a surgery approach. A straightforward monotonicity argument shows that the noncompactness of the cotangent bundle $T^*\Lambda \times [0,\infty)$ does not interfere with the compactness results for holomorphic curves used in the definition of wrapped Floer cohomology.

We first consider the surgery map into the Chekanov–Eliashberg algebra with loop space coefficients. Consider all components of Λ decorated by a positive sign $\Lambda = \Lambda^+$ and consider $\mathrm{CE}^*(\Lambda)$, which now involves, aside from Reeb chords, also chains $C_{-*}(\Omega\Lambda)$ on the based loop space. We define an A_∞–map

$$\Phi\colon \mathrm{CW}^*(C) \to \mathrm{CE}^*(\Lambda),$$

where the A_∞–structure on the right-hand side is the standard DG–algebra structure induced by concatenation and the Pontryagin product (as defined in this paper). As in Appendix A, consider a disk D_{i+j+2} with two dividing punctures that subdivides the boundary into two arcs, lower and upper. Let the upper arc contain i boundary punctures and be equipped with a decreasing boundary numbering κ, and the lower arc have j boundary punctures. Let $c' = c_i \cdots c_1$ be Reeb chords of C and let $c_0 = c_{0;1} \cdots c_{0;j}$ be Reeb chords of Λ. Let

$$c = c_{0;1} \cdots c_{0;j} z^v c_i \cdots c_1 \cdots z^w,$$

and consider $\mathcal{M}^{\mathrm{co}}(c;\kappa)$, again as in Figure 19.

Theorems 74 and 75 show that this moduli space carries a fundamental chain. We view this chain as parametrizing chains of paths in Λ connecting the Reeb chord endpoints in c_0. We write $[\mathcal{M}^{\mathrm{co}}(c)]$ for the alternating word of chains of loops and Reeb chords and view it as an element in $\mathrm{CE}^*(\Lambda)$. Define the maps

$$\Psi_i\colon \mathrm{CW}^*(C)^{\otimes_\kappa i} \to \mathrm{CE}^*(\Lambda) \quad \text{by } \Psi_i(c') = \sum_{c_0}[\mathcal{M}^{\mathrm{co}}(c)].$$

Theorem 86 *The map $\Psi\colon \mathrm{CW}^*(C) \to \mathrm{CE}^*(\Lambda)$ is an A_∞–map.*

Proof To see that the A_∞–relations hold we look at the boundary of the moduli space $\mathcal{M}^{\mathrm{co}}(c)$ of dimension d. The codimension-one boundary consists of three splitting types:

  (i)   A one-dimensional curve splits off in the positive symplectization end.
  (ii) A curve splits off at the negative end.
  (iii) Splitting at one of the intersection points z^v.

In order for splittings of the form (i) to contribute to the codimension-one boundary of the moduli space, the part of the holomorphic building in W consists of rigid disks with only positive punctures attached at one puncture to a negative puncture of the disk in the positive end and a disk of dimension $d - 1$ in $\mathcal{M}^{\mathrm{co}}(b)$ attached at the remaining negative puncture. (Splittings where the dimension of the components of the

holomorphic building are distributed differently have higher dimension along the disk C and correspond to "hidden faces" from the point of view of $C_*(\Omega\Lambda)$.) Assembling the rigid disks and the one-dimensional disk we get a partial holomorphic disk building that contributes to the A_∞–operations in $\mathrm{CW}^*(L)$ followed by the map Ψ. Splittings of type (ii) correspond to the map Ψ followed by the differential μ_1 in $\mathrm{CE}^*(\Lambda)$. Finally splittings of type (iii) correspond to the map Ψ followed by the product μ_2 on $\mathrm{CE}^*(\Lambda)$. We conclude that the terms contributing to A_∞–relations express the codimension-one boundary of $[\mathcal{M}^{\mathrm{co}}(c)]$ in two different ways and hence Ψ is an A_∞–map. $\qquad\square$

Remark 87 In the boundary of the moduli space $\mathcal{M}^{\mathrm{co}}(c)$ considered in the proof of Theorem 86 there are also higher-dimensional curves splitting off in the positive symplectization end. Such splittings contribute neither to the codimension-one boundary of the chains of loops, nor to the operations in the wrapped Floer cohomology, and hence play no role in the A_∞ chain map equation.

We will use slight generalizations of the map Ψ. More precisely, if p_j for $j = 1, \ldots, m$ are points in $[0, \infty) \times \Lambda$ and if F_j is the cotangent fiber at p_j, then we have a similar surgery map

$$\Psi^{p_i p_j} : \mathrm{CW}^*(F_i, F_j) \to \mathrm{CE}^*_{p_i p_j}(\Lambda),$$

which counts holomorphic disks with a positive Reeb chord connecting F_i to F_j, two Lagrangian intersection punctures at p_i and at p_j, and a word of chains of loops in Λ and Reeb chords of Λ as output, and where CE^*_{ij} is directly analogous to CE^* but where the first chain of loops is in a word is replaced by a chain of paths from p_i to the basepoint and the last is replaced by a chain of paths from the basepoint to p_j. In this setup the counterpart of the second component Ψ_2 is

$$(46) \qquad \Psi^{p_i p_j p_k} : \mathrm{CW}^*(F_j, F_k) \otimes \mathrm{CW}^*(F_i, F_j) \to \mathrm{CE}^*_{p_i p_k}(\Lambda),$$

and counts disks with two positive punctures at Reeb chords and two Lagrangian intersection punctures at p_i and p_k. The counterpart of the A_∞–equations in this setup is then

$$(47) \quad d \circ \Psi^{p_i p_k} + \Psi^{p_j p_k} \cdot \Psi^{p_i p_j} + \Psi^{p_i p_j p_k} \circ (1 \otimes \mu_1 + \mu_1 \otimes 1) + \Psi^{p_i p_k} \circ \mu_2 = 0,$$

where d is the differential on CE^*_{ik} and $\cdot$ is the (Pontryagin) product

$$\mathrm{CE}^*_{p_j p_k} \otimes \mathrm{CE}^*_{p_i p_j} \to \mathrm{CE}^*_{p_i p_k}.$$

The proofs of these statements are word for word repetitions of the proof of Theorem 86.

We next sketch a proof that the map Ψ in Theorem 86 is in fact a quasi-isomorphism, or, in other words, that its first component Ψ_1 induces an isomorphism on homology. We filter $CW^*(C)$ by action of its Reeb chord generators. To get a corresponding filtration on $CE^*(\Lambda)$ we use action on Reeb chords in combination with the energy on the loops. We start with a discussion of the energy of loops, following [52].

Equip Λ with a Riemannian metric and let $\Omega = \Omega(\Lambda)$ denote the space of based loops in Λ with the supremum norm: for two loops $\gamma, \beta \colon [0, 1] \to \Lambda$,

$$d^*(\gamma, \beta) = \sup_{t \in [0,1]} \rho(\gamma(t), \beta(t)),$$

where ρ is the metric on Λ induced by the Riemannian structure. Then the metric topology on Ω agrees with the standard compact open topology.

Let $\Omega' = \Omega'(\Lambda)$ denote the space of piecewise smooth paths with metric

$$d(\gamma, \beta) = d^*(\gamma, \beta) + \int_0^1 (|\dot\gamma| - |\dot\beta|)^2 \, dt,$$

where $\dot\gamma$ denotes the derivative of γ. The natural inclusion $\Omega' \to \Omega$ is a homotopy equivalence [52, Theorem 17.1]. We will use finite-dimensional approximations to study Ω'. The energy of a piecewise smooth loop in Λ is

$$E(\gamma) = \int_0^1 |\dot\gamma|^2 \, dt.$$

For $c > 0$, let $\Omega^c \subset \Omega'$ denote the subset of loops of energy $E < c$. The space Ω^c can be approximated by piecewise geodesic loops. More precisely, fixing a subdivision $0 = t_0 < t_1 < \cdots < t_m = 1$ of $[0, 1]$ we consider the space B^c of loops of energy $E < c$ that are geodesic on each interval $[t_i, t_{i+1}]$. Then [52, Lemma 16.1] shows that for all sufficiently fine subdivisions, B^c is a finite-dimensional manifold (a submanifold of the product $\Lambda^{\times m}$ in a natural way). Moreover, by [52, Theorem 16.2], all critical points of $E|_{\Omega^c}$ lie in B^c, which is a deformation retract of Ω^c, and for a generic metric $E|_{B^c}$ is a Morse function.

With these preliminaries established we turn to the actual proof. The first step will be to describe the Reeb chords of C^τ. Let g be a Riemannian metric on Λ as above and let $f \colon [0, \infty) \to \mathbb{R}$ be a positive function with $f(0) = 1$, $f'(0) = -1$ and $f'(t) < 0$ monotone increasing. Define the metric h on $\Lambda \times \mathbb{R}$ by

$$h = dt^2 + f(t)g.$$

Then, if x and y are points in Λ and $\gamma \colon [0, s] \to \Lambda$ is a geodesic with $\gamma(0) = x$ and $\gamma(s) = y$, there is a unique geodesic $(\gamma(t), r(t)) \in \Lambda \times [0, \infty)$ such that

- $(\gamma(0), 0) = (x, 0)$ and $(\gamma(s), r(s)) = (y, 0)$;
- $r(t)$ is Morse and has a unique maximum at an interior point $t = t_0$.

Note that the Reeb flow in the unit disk bundle is the natural lift of the geodesic flow. Assume next as above that the metric g on Λ is generic in the sense that the length functional for curves connecting any two Reeb chord endpoints in Λ has only Morse critical points. Concretely, this means that the index form of any geodesic connecting two Reeb chord endpoints is nondegenerate. As in [Lemma 81](#), this allows us to control the Reeb chords of C^τ below a given action for all sufficiently thin handles. More precisely, let ϵ denote the size of the tubular neighborhood of Λ in Y where we attach $T^*(\Lambda \times [0, \infty))$. We introduce the following notion of a geodesic-Reeb chord word. A *geodesic-Reeb chord word* is a word

$$\gamma_1 c_1 \gamma_2 c_2 \cdots c_m \gamma_m,$$

where γ_1 is a geodesic from one of the basepoints p_v to the start point of the Reeb chord c_1, where γ_2 is a geodesic from the endpoint of c_1 to the start point of c_2, etc, until finally γ_m is a geodesic from the endpoint of the Reeb chord c_m to one of the basepoints p_w. We define the action of a geodesic-Reeb chord word to be the sum of actions of its Reeb chords and the energies of its geodesics.

Lemma 88 *For any $A > 0$ there exist $\epsilon_0 > 0$ and $\tau_0 > 0$ such that for any $\epsilon < \epsilon_0$ and any $\tau > \tau_0$ there is a natural one-to-one correspondence between Reeb chords of C^τ of action $< A$, and geodesic-Reeb chord words of Λ of action $< A$.*

Sketch of proof The proof uses the transversality of the Reeb chords and of the geodesics. The basic observation is that the point in the normal fiber of Λ where the Reeb flow hits determines the direction of the geodesic in $\Lambda \times [0, \infty)$. After introducing a concrete smoothing of corners the lemma then follows from the finite-dimensional inverse function theorem. $\square$

To show that Ψ_1 is a quasi-isomorphism we will show that it is represented by a triangular matrix with ones on the diagonal with respect to the action/energy filtration. To this end we will use the Morse-theoretic finite-dimensional model for the chain complex underlying the homology of the based loop space described above. In order to have $[\mathcal{M}^{co}(c)]$ defined as a chain in this model we need to ensure that the paths on

the boundary of the holomorphic disk are sufficiently well behaved. We only sketch the construction. On holomorphic disks with unstable domains, we fix gauge using small spheres surrounding a Reeb chord endpoint; compare [31, Section A.2]. As in [31, Section A.1] we use a configuration space for holomorphic curves consisting of maps with two derivatives in L^2. This means that the restriction to the boundary has $\frac{3}{2}$ derivatives in L^2 and in particular the projection to Λ has bounded energy. Since the action of the positive puncture in a holomorphic disk contributing to the differential controls the norm of the solution, it follows that we can use configuration spaces of bounded energy to study the disks in the differential: we approximate the boundary curves uniformly by a piecewise geodesic curve by introducing a uniformly bounded number of subdivision points and straight-line homotopies in small charts.

Conjecture 89 *The chain map $\Psi_1 : \mathrm{CW}^*(C) \to \mathrm{CE}^*(\Lambda)$ induces an isomorphism on homology.*

Sketch of proof Consider a word of the form

$$\gamma_0 c_1 \gamma_1 c_2 \cdots c_m \gamma_m,$$

where γ_j are geodesics in $\Lambda \times [0, \infty)$ and c_j are Reeb chords. We aim to construct algebraically one disk connecting the Reeb chord a of C, corresponding to this word (see Lemma 88), to the word itself. We use an inductive argument and energy filtration. To start the argument we pick additional fiber disks F_{c^+} and F_{c^-} in $T^*(\Lambda \times [0, \infty))$ at (c^+, ϵ) and (c^-, ϵ) for very small $\epsilon > 0$ near all Reeb chord endpoints c_+ and c_- in Λ. We use the natural counterparts of the correspondence between mixed words of geodesics and Reeb chords before surgery and Reeb chords after, for mixed wrapped Floer cohomologies. For example, there is a straightforward analogue of Lemma 88: Reeb chord generators of $\mathrm{CW}^*(F_{c^+}, C)$ correspond to before-surgery words of the form

$$\gamma_1 c_1 \gamma_2 \cdots c_m \gamma_m,$$

where γ_1 is a geodesic connecting the basepoint of F_{c^+} to the initial point of c_1, and γ_2 connects the endpoint of c_1 to the start point of c_2, etc. To start the argument we note that it is straightforward to construct holomorphic strips corresponding to the short geodesics starting at F_{c^-} followed by the chord c and then the short geodesic to F_{c^+} and to show that they are unique. This corresponds to a generator of $\mathrm{CW}^*(F_{c^-}, F_{c^+})$. Likewise, it is immediate to construct the holomorphic disk connecting a Reeb chord generator of $\mathrm{CW}^*(F_{c^+}, C)$ corresponding to a geodesic, and show that it is unique; compare Theorem 83.

We now use these two to construct algebraically one disk from the Reeb chord generator of $\mathrm{CW}^*(F_{c_-}, C)$ corresponding to the short geodesic, the chord c, and a geodesic connecting the endpoint of c to the basepoint of C. To this end we consider the natural map

$$\Psi^{p_v c^+ c^-} : \mathrm{CW}^*(F_{c_+}, C) \otimes \mathrm{CW}^*(F_{c_-}, F_{c_+}) \to \mathrm{CE}^*(\Lambda);$$

see (47). For the two Reeb chords, a connecting F_{c_-} to F_{c_+} corresponding to the chord c of Λ, and b connecting F_{c_+} to C corresponding to the geodesic, we then have, with p_v denoting the basepoint,

$$d(\Psi^{p_v c^+ c^-}(b, a)) + (\Psi^{p_v c^-}(b)) \cdot (\Psi^{c^+ c^-}(a)) + \Psi^{p_v c^-}(\mathfrak{m}_2(b, a))$$
$$+ (-1)^{|a|-1} \Psi^{p_v c^+ c^-}(\mathfrak{m}_1(b), a) + \Psi^{p_v c^+ c^-}(b, \mathfrak{m}_1(a)) = 0.$$

Here we know that the terms containing $\mathfrak{m}_1$ and d involve nontrivial holomorphic disks or Morse flows in the finite-dimensional approximation, and hence lower action/energy by an amount bounded below by some $\delta > 0$, which we assume is much larger than $\epsilon > 0$ above. Therefore, if we restrict attention to a small action window, we find

$$(\Psi^{p_v c^-}(b) \cdot \Psi^{c^+ c^-}(a)) + \Psi^{p_v c^-}(\mathfrak{m}_2(b, a)) = 0.$$

Here the first term is simply the Pontryagin product at the common endpoint of the curves, which is homologous to the word $\epsilon' c \gamma$ of the small geodesic, the Reeb chord and then the longer geodesic, by rounding the corner at c^+. It follows that $\mathfrak{m}_2(a, b) = r$, where r is a Reeb chord with action between the sum of the actions of a and b and the action of $\epsilon' c \gamma$, and that $\Psi^{c^- p_v}(r)$ contains this word with coefficient ± 1. Noting that there is only one Reeb chord in the action window studied, we find that the desired coefficient equals ± 1. It is now clear how to continue the induction: at each step we add one more geodesic or Reeb chord to any word. Using already constructed curves and (47) in a small action window, we find that the map Ψ_1 has a triangular action matrix with ± 1 on the diagonal, hence it is a quasi-isomorphism. □

References

[1] **M Abouzaid**, *Morse homology, tropical geometry, and homological mirror symmetry for toric varieties*, Selecta Math. 15 (2009) 189–270 MR Zbl

[2] **M Abouzaid**, *On the wrapped Fukaya category and based loops*, J. Symplectic Geom. 10 (2012) 27–79 MR Zbl

[3] **M Abouzaid**, **P Seidel**, *An open string analogue of Viterbo functoriality*, Geom. Topol. 14 (2010) 627–718 MR Zbl

[4]   **J F Adams**, *On the cobar construction*, Proc. Nat. Acad. Sci. U.S.A. 42 (1956) 409–412 MR Zbl

[5]   **J F Adams**, **P J Hilton**, *On the chain algebra of a loop space*, Comment. Math. Helv. 30 (1956) 305–330   MR Zbl

[6]   **M Aganagic**, **T Ekholm**, **L Ng**, **C Vafa**, *Topological strings, D–model, and knot contact homology*, Adv. Theor. Math. Phys. 18 (2014) 827–956   MR Zbl

[7]   **J-F Barraud**, **O Cornea**, *Lagrangian intersections and the Serre spectral sequence*, Ann. of Math. 166 (2007) 657–722   MR Zbl

[8]   **A Beilinson**, **V Ginzburg**, **W Soergel**, *Koszul duality patterns in representation theory*, J. Amer. Math. Soc. 9 (1996) 473–527   MR Zbl

[9]   **J M Boardman**, *Conditionally convergent spectral sequences*, from "Homotopy invariant algebraic structures" (J-P Meyer, J Morava, W S Wilson, editors), Contemp. Math. 239, Amer. Math. Soc., Providence, RI (1999) 49–84   MR Zbl

[10] **F Bourgeois**, **B Chantraine**, *Bilinearized Legendrian contact homology and the augmentation category*, J. Symplectic Geom. 12 (2014) 553–583   MR Zbl

[11] **F Bourgeois**, **T Ekholm**, **Y Eliashberg**, *Effect of Legendrian surgery*, Geom. Topol. 16 (2012) 301–389   MR Zbl

[12] **F Bourgeois**, **Y Eliashberg**, **H Hofer**, **K Wysocki**, **E Zehnder**, *Compactness results in symplectic field theory*, Geom. Topol. 7 (2003) 799–888   MR Zbl

[13] **E H Brown, Jr**, *Twisted tensor products, I*, Ann. of Math. 69 (1959) 223–246   MR Zbl

[14] **G Carlsson**, **R J Milgram**, *Stable homotopy and iterated loop spaces*, from "Handbook of algebraic topology" (I M James, editor), North-Holland, Amsterdam (1995) 505–583 MR Zbl

[15] **R Casals**, **E Murphy**, *Legendrian fronts for affine varieties*, Duke Math. J. 168 (2019) 225–323   MR Zbl

[16] **K T Chen**, *Iterated path integrals*, Bull. Amer. Math. Soc. 83 (1977) 831–879   MR Zbl

[17] **K Cieliebak**, **T Ekholm**, **J Latschev**, *Compactness for holomorphic curves with switching Lagrangian boundary conditions*, J. Symplectic Geom. 8 (2010) 267–298 MR Zbl

[18] **K Cieliebak**, **Y Eliashberg**, *Flexible Weinstein manifolds*, from "Symplectic, Poisson, and noncommutative geometry" (T Eguchi, Y Eliashberg, editors), Math. Sci. Res. Inst. Publ. 62, Cambridge Univ. Press (2014) 1–42   MR Zbl

[19] **K Cieliebak**, **J Latschev**, *The role of string topology in symplectic field theory*, from "New perspectives and challenges in symplectic field theory" (M Abreu, F Lalonde, L Polterovich, editors), CRM Proc. Lecture Notes 49, Amer. Math. Soc., Providence, RI (2009) 113–146   MR Zbl

[20] **G Civan**, **P Koprowski**, **J Etnyre**, **J M Sabloff**, **A Walker**, *Product structures for Legendrian contact homology*, Math. Proc. Cambridge Philos. Soc. 150 (2011) 291–311 MR Zbl

[21] **S Eilenberg**, **J C Moore**, *Limits and spectral sequences*, Topology 1 (1962) 1–23 MR Zbl

[22] **T Ekholm**, *Morse flow trees and Legendrian contact homology in 1–jet spaces*, Geom. Topol. 11 (2007) 1083–1224 MR Zbl

[23] **T Ekholm**, *Rational symplectic field theory over $\mathbb{Z}_2$ for exact Lagrangian cobordisms*, J. Eur. Math. Soc. 10 (2008) 641–704 MR Zbl

[24] **T Ekholm**, *A version of rational SFT for exact Lagrangian cobordisms in 1–jet spaces*, from "New perspectives and challenges in symplectic field theory" (M Abreu, F Lalonde, L Polterovich, editors), CRM Proc. Lecture Notes 49, Amer. Math. Soc., Providence, RI (2009) 173–199 MR Zbl

[25] **T Ekholm**, *Holomorphic curves for Legendrian surgery*, preprint (2019) arXiv

[26] **T Ekholm**, **J B Etnyre**, **L Ng**, **M G Sullivan**, *Knot contact homology*, Geom. Topol. 17 (2013) 975–1112 MR Zbl

[27] **T Ekholm**, **J Etnyre**, **M Sullivan**, *Orientations in Legendrian contact homology and exact Lagrangian immersions*, Internat. J. Math. 16 (2005) 453–532 MR Zbl

[28] **T Ekholm**, **J Etnyre**, **M Sullivan**, *Legendrian contact homology in $P \times \mathbb{R}$*, Trans. Amer. Math. Soc. 359 (2007) 3301–3335 MR Zbl

[29] **T Ekholm**, **K Honda**, **T Kálmán**, *Legendrian knots and exact Lagrangian cobordisms*, J. Eur. Math. Soc. (JEMS) 18 (2016) 2627–2689 MR Zbl

[30] **T Ekholm**, **A Oancea**, *Symplectic and contact differential graded algebras*, Geom. Topol. 21 (2017) 2161–2230 MR Zbl

[31] **T Ekholm**, **I Smith**, *Exact Lagrangian immersions with one double point revisited*, Math. Ann. 358 (2014) 195–240 MR Zbl

[32] **T Etgü**, **Y Lekili**, *Koszul duality patterns in Floer theory*, Geom. Topol. 21 (2017) 3313–3389 MR Zbl

[33] **Y Félix**, **J C Thomas**, *Extended Adams–Hilton's construction*, Pacific J. Math. 128 (1987) 251–263 MR Zbl

[34] **K Fukaya**, *Application of Floer homology of Langrangian submanifolds to symplectic topology*, from "Morse theoretic methods in nonlinear analysis and in symplectic topology" (P Biran, O Cornea, F Lalonde, editors), NATO Sci. Ser. II Math. Phys. Chem. 217, Springer (2006) 231–276 MR Zbl

[35] **K Fukaya**, **Y-G Oh**, *Zero-loop open strings in the cotangent bundle and Morse homotopy*, Asian J. Math. 1 (1997) 96–180 MR Zbl

[36] **K Fukaya**, **Y-G Oh**, **H Ohta**, **K Ono**, *Lagrangian intersection Floer theory: anomaly and obstruction, I*, AMS/IP Studies in Advanced Mathematics 46, Amer. Math. Soc., Providence, RI (2009) MR Zbl

[37] **P G Goerss**, **J F Jardine**, *Simplicial homotopy theory*, Progr. Math. 174, Birkhäuser, Basel (1999) MR Zbl

[38] **M Gromov**, *Pseudo holomorphic curves in symplectic manifolds*, Invent. Math. 82 (1985) 307–347 MR Zbl

[39] **E Herscovich**, *Hochschild* (co)*homology and Koszul duality*, preprint (2014) arXiv

[40] **J V S Holstein**, *Morita cohomology*, PhD thesis, University of Cambridge (2014) Available at https://api.repository.cam.ac.uk/server/api/core/bitstreams/38e69cc9-eb4d-435a-b49b-5a09764ceda5/content

[41] **J V S Holstein**, *Morita cohomology*, Math. Proc. Cambridge Philos. Soc. 158 (2015) 1–26 MR Zbl

[42] **J D S Jones**, *Cyclic homology and equivariant homology*, Invent. Math. 87 (1987) 403–423 MR Zbl

[43] **M Kalck**, **D Yang**, *Relative singularity categories, I: Auslander resolutions*, Adv. Math. 301 (2016) 973–1021 MR Zbl

[44] **D M Kan**, *Abstract homotopy, IV*, Proc. Nat. Acad. Sci. U.S.A. 42 (1956) 542–544 MR Zbl

[45] **B Keller**, *Deriving DG categories*, Ann. Sci. École Norm. Sup. 27 (1994) 63–102 MR Zbl

[46] **M Kontsevich**, *Symplectic geometry of homological algebra*, preprint (2009) Available at https://www.ihes.fr/~maxim/TEXTS/Symplectic_AT2009.pdf

[47] **K Lefèvre-Hasegawa**, *Sur les A_∞–catègories*, PhD thesis, Université Paris 7 (2003) arXiv

[48] **Y Lekili**, **A Polishchuk**, *Arithmetic mirror symmetry for genus 1 curves with n marked points*, Selecta Math. 23 (2017) 1851–1907 MR Zbl

[49] **J-L Loday**, **B Vallette**, *Algebraic operads*, Grundl. Math. Wissen. 346, Springer (2012) MR Zbl

[50] **D M Lu**, **J H Palmieri**, **Q S Wu**, **J J Zhang**, *Koszul equivalences in A_∞–algebras*, New York J. Math. 14 (2008) 325–378 MR Zbl

[51] **J McCleary**, *User's guide to spectral sequences*, Mathematics Lecture Series 12, Publish or Perish, Wilmington, DE (1985) MR Zbl

[52] **J Milnor**, *Morse theory*, Annals of Mathematics Studies 51, Princeton Univ. Press (1963) MR Zbl

[53] **D Nadler**, *Wrapped microlocal sheaves on pairs of pants*, preprint (2016) arXiv

[54] **J A Neisendorfer**, *What is loop multiplication anyhow?*, J. Homotopy Relat. Struct. 12 (2017) 659–690 MR Zbl

[55] **L Ng**, **D Rutherford**, **V Shende**, **S Sivek**, **E Zaslow**, *Augmentations are sheaves*, Geom. Topol. 24 (2020) 2149–2286 MR Zbl

[56] **L Positselski**, *Two kinds of derived categories, Koszul duality, and comodule–contramodule correspondence*, Mem. Amer. Math. Soc. 996, Amer. Math. Soc., Providence, RI (2011) MR Zbl

[57] **A Prouté**, *A_∞–structures: modèles minimaux de Baues–Lemaire et Kadeishvili et homologie des fibrations*, Repr. Theory Appl. Categ. (2011) 1–99 MR Zbl

[58] **S Schwede**, **B Shipley**, *Equivalences of monoidal model categories*, Algebr. Geom. Topol. 3 (2003) 287–334 MR Zbl

[59] **P Seidel**, *Fukaya categories and Picard–Lefschetz theory*, Eur. Math. Soc., Zürich (2008) MR Zbl

[60] **J-P Serre**, *Homologie singulière des espaces fibrés: applications*, Ann. of Math. 54 (1951) 425–505 MR Zbl

[61] **Z Sylvan**, *Partially wrapped Floer theory*, talk, IAS workshop "Homological mirror symmetry: methods and structures" (2016) Available at `https://www.ias.edu/video/whms/2016/1110-ZachSylvan`

[62] **Z Sylvan**, *On partially wrapped Fukaya categories*, J. Topol. 12 (2019) 372–441 MR Zbl

[63] **G Tabuada**, *Theorie homotopique des DG–categories*, preprint (2007) arXiv

[64] **C A Weibel**, *An introduction to homological algebra*, Cambridge Studies in Advanced Mathematics 38, Cambridge Univ. Press (1994) MR Zbl

TE: *Department of Mathematics, Uppsala University*
Uppsala, Sweden

Insitut Mittag-Leffler
Djursholm, Sweden

YL: *Department of Mathematics, Imperial College London*
South Kensington, London, United Kingdom

`tobias.ekholm@math.uu.se`, `y.lekili@imperial.ac.uk`

Filtering the Heegaard Floer contact invariant

Çağatay Kutluhan
Gordana Matić
Jeremy Van Horn-Morris
Andy Wand

We define an invariant of contact structures in dimension three from Heegaard Floer homology. This invariant takes values in the set $\mathbb{Z}_{\geq 0} \cup \{\infty\}$. It is zero for overtwisted contact structures, ∞ for Stein-fillable contact structures, nondecreasing under Legendrian surgery, and computable from any supporting open book decomposition. As an application, we give an easily computable obstruction to Stein-fillability on closed contact 3–manifolds with nonvanishing Ozsváth–Szabó contact class.

1 Introduction

Let M be a closed orientable 3–manifold and ξ be a contact structure on M. The goal of this article is to define an invariant of (M, ξ) as a refinement of the contact invariant in Heegaard Floer homology, the *Ozsváth–Szabó contact class* $\hat{c}(\xi)$ [50], and to study some of its properties. To define our invariant, we start from an open book decomposition of M supporting ξ and a collection of pairwise disjoint properly embedded arcs on a page of the open book decomposition. From this data we build a

filtered chain complex out of the corresponding Heegaard Floer chain complex, whose filtration captures in an algebraic sense the topological complexity of curves counted by the differential. We then consider how far the Ozsváth–Szabó contact class survives in the associated spectral sequence. The result is an invariant of the contact manifold, denoted by $o(M, \xi)$, and read the *spectral order*, or simply *order*, of (M, ξ), taking values in $\mathbb{Z}_{\geq 0} \cup \{\infty\}$.

Theorem 1.1 *The contact invariant o satisfies the following properties*:

- $o(M, \xi) = 0$ *if* (M, ξ) *is overtwisted.*
- $o(M, \xi) = \infty$ *if* (M, ξ) *is Stein-fillable.*
- $o(M, \xi)$ *can be detected on an arbitrary supporting open book decomposition of* (M, ξ).

The second bullet point property in Theorem 1.1 follows from the fact that the contact invariant o behaves well under Legendrian surgery, giving a map of partially ordered sets from contact manifolds ordered by Stein cobordisms to the set $\mathbb{Z}_{\geq 0} \cup \{\infty\}$ with the usual ordering:

Theorem 1.2 *The contact invariant o is nondecreasing under Legendrian surgery and in particular gives an obstruction to the existence of Stein cobordisms between contact 3–manifolds. Specifically, if (M_-, ξ_-) and (M_+, ξ_+) are respectively the concave and convex ends of a Stein cobordism, then $o(M_-, \xi_-) \leq o(M_+, \xi_+)$.*

Aside from the properties listed in Theorem 1.1, the contact invariant o behaves well under connected sums. To be more explicit:

Theorem 1.3 *Let (M_1, ξ_1) and (M_2, ξ_2) be closed contact 3–manifolds. Then their connected sum satisfies $o(M_1 \# M_2, \xi_1 \# \xi_2) = \min\{o(M_1, \xi_1), o(M_2, \xi_2)\}$.*

The above theorem fits into a broader pattern of similar contact connected sum results. Loosely, various measures of rigidity of $(M_1 \# M_2, \xi_1 \# \xi_2)$ — for example, Stein-fillability, having a nonvanishing Ozsváth–Szabó contact class, or tightness — is the weaker of that property for (M_1, ξ_1) or (M_2, ξ_2) (see Eliashberg [11], Cieliebak and Eliashberg [7], Ozsváth and Szabó [50] and Colin [8]). In addition, Theorem 1.3 leads to existence of a family of monoids $o^k(S)$ in the mapping class group $\mathrm{Mod}(S, \partial S)$: $\phi \in \mathrm{Mod}(S, \partial S)$ belongs to $o^k(S)$ if and only if $o \geq k$ for the contact 3–manifold specified by the open book decomposition (S, ϕ).

Our contact invariant is inspired by an analog of Latschev and Wendl's algebraic torsion introduced by Hutchings in the context of embedded contact homology (ECH) in [33, Appendix]. To a closed oriented 3–manifold M, a nondegenerate contact 1–form λ on M, and a generic almost complex structure J on $\mathbb{R} \times M$ as needed to define the ECH chain complex, Hutchings associates a number $f(M, \lambda, J)$ in $\mathbb{Z}_{\geq 0} \cup \{\infty\}$. The latter is shown to vanish for overtwisted contact structures for all choices of λ and J, and can be used to obstruct exact symplectic cobordisms. Our initial definitions follow the ideas of Hutchings' construction, ported to the setting of Heegaard Floer homology (see our work [32] for more on this). We choose to work with Heegaard Floer homology because of its computational advantages.

As an application, it follows from the second bullet point above that, even for closed contact 3–manifolds with nonvanishing Ozsváth–Szabó contact class, one can obstruct Stein-fillability by finding a finite upper bound on its spectral order, which is easier than computing the spectral order itself.

Theorem 1.4 *There is an infinite family of contact 3–manifolds* $\{(Y_p, \xi_p)\}_{p \in \mathbb{Z}_{>0}}$ *each with* $\hat{c}(\xi_p) \neq 0$ *but with* $o(Y_p, \xi_p) = 0$ *(see Figure 19, left, for a description of this family via open book decompositions).* In particular, *these contact 3–manifolds are not Stein-fillable.*

Remark During the course of this project we learned that John Baldwin and David Shea Vela-Vick have independently been working on a filtration in Heegaard Floer homology similar in spirit to the J_+–filtration defined in Section 2.2. This led to an interesting application in knot Floer homology [4].

Future considerations

In upcoming work in progress [31], we present an infinite family of contact structures with vanishing Ozsváth–Szabó contact class but with nonzero spectral order. Furthermore, we compute upper bounds on the spectral order of these contact structures and these upper bounds span the range of all positive integers. The next step will be to show that there is an increasing sequence of positive integers that provides lower bounds on the spectral order of our family of contact structures. These computations would resolve the following conjecture:

Conjecture 1.5 *An infinite sequence of distinct positive integers is realized by the spectral order of an infinite family of contact structures with vanishing Ozsváth–Szabó contact class.*

In addition, such a family of examples would provide a nested sequence of monoids

$$\cdots \subsetneq o^{k_{n+1}}(S) \subsetneq o^{k_n}(S) \subsetneq \cdots ,$$

where $o^{k_n}(S)$ is the set of orientation-preserving homeomorphisms ϕ in the mapping class group $\mathrm{Mod}(S, \partial S)$ such that the open book decomposition (S, ϕ) supports a contact structure with $o \geq k$, and S may have arbitrary genus. Note that this family of monoids would be contained in the monoid $\mathrm{Tight}(S, \partial S)$ and would contain the monoid $\mathrm{Stein}(S, \partial S)$ (see Etnyre and Van Horn-Morris [13], as well as Baldwin [3] and Baker, Etnyre and Van Horn-Morris [1]), and it would provide an answer to [13, Question 6.8].

A more conceptual question concerns the potential of a converse to the first bullet point of Theorem 1.1:

Question 1.6 *Suppose that (M, ξ) has vanishing Ozsváth–Szabó contact class. Does $o(M, \xi) = 0$ imply that ξ is overtwisted?*

An affirmative answer to Question 1.6 would imply that Heegaard Floer package detects tight contact structures. In this regard, spectral order gives a potential interpretation of *consistency* of an open book decomposition (see Wand [55]), a combinatorial condition equivalent to tightness of the supported contact structure, in the context of pseudoholomorphic curves. Furthermore, along with the nondecreasing behavior of spectral order under Legendrian surgery, an affirmative answer to Question 1.6 would provide an alternative and more conceptual proof of the following theorem, which has recently been proved by the last author in [56]:

Theorem 1.7 *Let ξ be a tight contact structure on M, and $K \subset M$ be a null-homologous Legendrian knot. Then contact (-1)–surgery on K produces a 3–manifold with a tight contact structure.*

Another question of interest is related to generalizing our invariant to compact contact 3–manifolds with convex boundary. In this regard, our construction of a filtered chain complex out of the Heegaard Floer chain complex readily generalizes to the case of partial open book decompositions introduced by Honda, Kazez and Matić [22]. This allows us to extend the definition of spectral order (Definition 2.2) to compact contact 3–manifolds with convex boundary. This was independently observed by Juhász and Kang [26], who used it to find an upper bound on the spectral order for a closed contact 3–manifold that contains a Giroux torsion domain. More generally, Juhász and Kang showed that the spectral order of a codimension zero contact submanifold with convex

boundary gives an upper bound on the spectral order of the ambient manifold. Among other things, we will compare o to Wendl's *planar torsion* [58]. As is stated by Latschev and Wendl [33, Theorem 6], planar torsion provides an upper bound to Latschev and Wendl's algebraic torsion. Moreover, planar torsion detects overtwistedness. One could expect a similar relationship between spectral order and Wendl's planar torsion. These are the content of another work in progress by the authors [30].

Question 1.8 *Suppose that the closed contact 3–manifold (M, ξ) has planar k–torsion. Does this imply $o(M, \xi) \leq k$?*

Organization

In Section 2, we provide the definitions required throughout the article, leading to the definition of spectral order. These include a preliminary version of the latter, denoted by o, which a priori depends on the choices made to define it.

Section 3 investigates the dependence of o on various choices made in its definition. Among these are a choice of the monodromy of an open book decomposition in its isotopy class and a choice of a collection of pairwise disjoint properly embedded arcs on a page of an open book decomposition.

In Section 4, we exhibit several properties of spectral order, and in doing so prove Theorems 1.1, 1.2 and 1.3.

In Section 5, we present an infinite family of contact structures with nonvanishing Ozsváth–Szabó contact class but with zero spectral order. This implies, by Theorem 1.1, that these contact structures are not Stein-fillable. We also compare our method to other known obstructions to fillability of closed contact 3–manifolds.

Acknowledgements

2 Definitions

2.1 Background

To set the stage, let M be a closed, connected and oriented 3–manifold endowed with a cooriented contact structure ξ. It is understood that the orientation on M is induced by ξ. A celebrated theorem of Giroux states that there is a one-to-one correspondence between contact structures up to isotopy and open book decompositions up to positive stabilization [17]. An abstract open book decomposition of M is a pair (S, ϕ), where S is a compact oriented surface of genus g with B boundary components, called the *page*, and ϕ is an orientation-preserving diffeomorphism of S which restricts to the identity in a neighborhood of the boundary, called the *monodromy*. The manifold M is homeomorphic to $S \times [0, 1]/\sim$, where $(p, 1) \sim (\phi(p), 0)$ for any $p \in S$ and $(p, t) \sim (p, t')$ for any $p \in \partial S$ and $t, t' \in [0, 1]$. The open book decomposition is said to support the contact structure ξ if there exists a 1–form λ such that $\xi = \ker(\lambda)$, $\lambda|_{\partial S} > 0$ and $d\lambda|_S > 0$.

Now fix an abstract open book decomposition (S, ϕ) of M supporting ξ and a collection of pairwise disjoint properly embedded arcs $\boldsymbol{a} = \{a_1, \ldots, a_N\}$ on S that contains a basis, that is, a subcollection of arcs cutting S into a polygon. This arc collection together with the monodromy ϕ defines a Heegaard diagram $(\Sigma, \{\beta_1, \ldots, \beta_N\}, \{\alpha_1, \ldots, \alpha_N\})$ for $-M$ as in [23, Section 3.1]. To be more explicit, let $\boldsymbol{b} = \{b_1, \ldots, b_N\}$ be a collection of arcs on S where b_i is isotopic to a_i via a small isotopy satisfying the following conditions:

- The endpoints of b_i are obtained from the endpoints of a_i by pushing along ∂S in the direction of the boundary orientation.
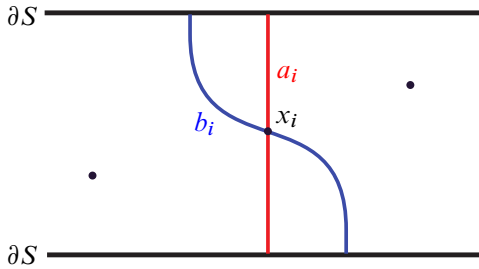
Figure 1: The arcs a_i and b_i on the surface S.

- a_i intersects b_i transversally at one point, x_i, in the interior of S.

- Having fixed an orientation of a_i, there is an induced orientation on b_i, and the sign of the oriented intersection $a_i \cap b_i$ is positive (see Figure 1).

Then $\Sigma = S \times \{\frac{1}{2}\} \cup_{\partial S} -S \times \{0\}$, $\alpha_i = a_i \times \{\frac{1}{2}\} \cup a_i \times \{0\}$ and $\beta_i = b_i \times \{\frac{1}{2}\} \cup \phi(b_i) \times \{0\}$. Note that the Heegaard diagram $(-\Sigma, \{\alpha_1, \ldots, \alpha_N\}, \{\beta_1, \ldots, \beta_N\})$ also describes the manifold $-M$, and we may sometimes prefer to use this diagram in figures.

With the preceding understood, we recall the definition of the Heegaard Floer chain complex $(\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}), \widehat{\partial}_{HF})$. In doing so, we adopt Lipshitz's cylindrical reformulation of Heegaard Floer homology [35]. The definition also requires the choice of basepoints $z \subset \Sigma \smallsetminus \bigcup_{i \in \{1, \ldots, N\}} (\alpha_i \cup \beta_i)$. In the present context, this is done according to the convention in [23, Section 3.1]. To be more explicit, place a single basepoint in every connected component of $S \smallsetminus \bigcup_{i \in \{1, \ldots, N\}} a_i$ outside the small strips between a_i and b_i (see Figure 1). Following Lipshitz, the chain group $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is freely generated over $\mathbb{F} := \mathbb{Z}/2\mathbb{Z}$ by I–*chord collections* $\vec{\boldsymbol{x}} := \boldsymbol{x} \times [0, 1]$ specified by unordered N–tuples of points in Σ of the form $\boldsymbol{x} = \{x_1, \ldots, x_N\}$, where $x_i \in \alpha_i \cap \beta_{\sigma(i)}$ for some element σ of the symmetric group S_N. Given a generic almost complex structure J_{HF} on $\Sigma \times [0, 1] \times \mathbb{R}$ satisfying conditions (J1)–(J5) in [35, Section 1, page 959], the differential $\widehat{\partial}_{HF}$ on $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is defined to be the endomorphism of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ sending a generator $\vec{\boldsymbol{x}}$ to

$$\sum_{\boldsymbol{y}} \sum_{\substack{A \in \widehat{\pi}_2(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) \\ \mathrm{ind}(A)=1}} n(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}; A) \vec{\boldsymbol{y}}.$$

Here $\widehat{\pi}_2(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}})$ denotes the set of relative homology classes of continuous maps from a Riemann surface with boundary and boundary punctures into $\Sigma \times [0, 1] \times \mathbb{R}$ such that it maps the boundary of the surface into $\boldsymbol{\alpha} \times \{0\} \times \mathbb{R} \cup \boldsymbol{\beta} \times \{1\} \times \mathbb{R}$, it converges to $\vec{\boldsymbol{x}}$ and $\vec{\boldsymbol{y}}$ at its punctures, and it has trivial homological intersection with $\{z\} \times [0, 1] \times \mathbb{R}$.

Meanwhile, ind(A) denotes the index of a class $A \in \hat{\pi}_2(\vec{x}, \vec{y})$ (see [35, Definition 4.4]), and $n(\vec{x}, \vec{y}; A)$ is a signed count, modulo $\mathbb{R}$–translation, of J_{HF}–holomorphic curves in $\Sigma \times [0,1] \times \mathbb{R}$ satisfying conditions (M0)–(M6) in [35, Section 1, page 960] and representing the class A. The latter is guaranteed to be finite if we choose the monodromy ϕ appropriately in its isotopy class so as to make the multipointed Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$ *admissible*. A multipointed Heegaard diagram is admissible if every nontrivial periodic domain has both positive and negative coefficients (see [35, Definition 5.1]).

Remark Even though Lipshitz carried out his construction of a cylindrical reformulation of Heegaard Floer homology in the case $N = 2g + B - 1$ (in other words, the case with one basepoint), the details of his construction and especially the results in [35, Sections 4 and 10] carry over to the multipointed case but for cosmetic changes.

2.2 The J_+ filtration

Next we build a filtered chain complex out of $(\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}), \hat{\partial}_{HF})$. To do this, we adopt Hutchings' recipe in [24, Section 6]. Given a pair of generators $\vec{x}$ and $\vec{y}$, define a function J_+ on $\hat{\pi}_2(\vec{x}, \vec{y})$ by[1]

$$(2\text{-}1) \qquad J_+(A) := \mu(\mathcal{D}(A)) - 2e(\mathcal{D}(A)) + |\boldsymbol{x}| - |\boldsymbol{y}|,$$

where $|\cdot|$ denotes the number of disjoint cycles in the element of the symmetric group S_N associated to a given generator following the convention described above in Section 2.1 (eg the generator x_ξ corresponding to the distinguished set of points $\{x_1, \ldots, x_N\}$ indicated in Figure 1 has $|\vec{x}_\xi| = N$), $\mathcal{D}(A)$ is the *domain* in the pointed Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$ representing a class $A \in \hat{\pi}_2(\vec{x}, \vec{y})$, $\mu(\mathcal{D}(A))$ is the *Maslov index* of $\mathcal{D}(A)$ as in the traditional setting of [48], and $e(\mathcal{D}(A))$ is the *Euler measure* of $\mathcal{D}(A)$ (see [35, Section 4.1, page 973] for the definition). Since the Maslov index and Euler measure are additive under concatenation of domains, so is J_+. More precisely, for any $A \in \hat{\pi}_2(\vec{x}, \vec{y})$ and $A' \in \hat{\pi}_2(\vec{y}, \vec{z})$, we have

$$J_+(A + A') = J_+(A) + J_+(A').$$

Now suppose that $A \in \hat{\pi}_2(\vec{x}, \vec{y})$ is represented by a J_{HF}–holomorphic curve C_L in $\Sigma \times [0,1] \times \mathbb{R}$ satisfying conditions (M0)–(M6) in [35, Section 1]. Then, by [35, Proposition 4.2 (see also Proposition 4.2′ in the correction)],

$$(2\text{-}2) \qquad \chi(C_L) = N - n_{\boldsymbol{x}}(\mathcal{D}(A)) - n_{\boldsymbol{y}}(\mathcal{D}(A)) + e(\mathcal{D}(A)).$$

---

[1]The interested reader may refer to [32] to see how the authors originally came up with this formula.

Here, $n_p(\mathcal{D}(A))$ denotes the *point measure*, namely, the average of the coefficients of $\mathcal{D}(A)$ for the four regions with corners at $p \in \alpha_i \cap \beta_j$. Meanwhile, Lipshitz's formula for the Maslov index of domains [35, Corollary 4.10 (see also Proposition 4.8′ in the correction)] asserts that

$$(2\text{-}3) \qquad \mu(\mathcal{D}(A)) = n_{\boldsymbol{x}}(\mathcal{D}(A)) + n_{\boldsymbol{y}}(\mathcal{D}(A)) + e(\mathcal{D}(A)).$$

Combining (2-2) and (2-3), we obtain

$$\mu(\mathcal{D}(A)) - 2e(\mathcal{D}(A)) = -\chi(C_L) + \mathrm{N},$$

and hence (2-1) can be rewritten as

$$(2\text{-}4) \qquad J_+(A) = -\chi(C_L) + \mathrm{N} + |\boldsymbol{x}| - |\boldsymbol{y}|.$$

With the preceding understood, consider the smooth compact oriented surface C obtained from the compactification of C_L by attaching 2–dimensional 1–handles along pairs of points in $\alpha_i \times \{0\} \times \mathbb{R} \cap C_L$ and $\beta_i \times \{1\} \times \mathbb{R} \cap C_L$ for each $i = 1, \ldots, \mathrm{N}$, and then smoothing. Then $\chi(C) = \chi(C_L) - \mathrm{N}$, and $|\boldsymbol{x}|$ (resp. $|\boldsymbol{y}|$) is equal to the number of boundary components of C arising from the I–chord $\vec{x}$ (resp. $\vec{y}$). Hence, we can further rewrite (2-4) as

$$(2\text{-}5) \qquad J_+(A) = \sum_{C_j \subset C} (2g_j - 2 + 2|\boldsymbol{x}_j|),$$

where each C_j denotes a connected component of C, g_j denotes the genus of C_j, and each $\boldsymbol{x}_j \subset \boldsymbol{x}$ denotes the maximal subcollection of points in $\boldsymbol{x}$ such that $\boldsymbol{x}_j \times [0, 1]$ lies on the boundary of the component C_j. Note that each connected component of C has nonempty intersections with the I–chord collections specified by $\boldsymbol{x}$ and $\boldsymbol{y}$ since each connected component of C_L has nonempty negative and positive ends. Therefore, it follows from (2-5) that $2 \mid J_+(A)$ and $J_+(A) \geq 0$.

Remark If there exists an embedded J_{HF}–holomorphic curve C_L representing the class A, then the Maslov index of $\mathcal{D}(A)$ agrees with the Fredholm index of C_L. For Maslov index-1 domains, we prefer to use the equivalent formula

$$(2\text{-}6) \qquad J_+(A) = 2[n_{\boldsymbol{x}}(\mathcal{D}(A)) + n_{\boldsymbol{y}}(\mathcal{D}(A))] - 1 + |\boldsymbol{x}| - |\boldsymbol{y}|.$$

2.3 The filtered chain complex

Following Hutchings, we decompose the Heegaard Floer differential as

$$\widehat{\partial}_{HF} = \partial_0 + \partial_1 + \cdots + \partial_l + \cdots,$$

where ∂_l counts J_{HF}–holomorphic curves with $J_+ = 2l$ and having empty intersection with $\{z\} \times [0,1] \times \mathbb{R}$. Since J_+ is additive under gluing of J–holomorphic curves, the above decomposition induces a spectral sequence with pages

$$E^k(S, \phi, \boldsymbol{a}; J_{HF}) = H_*(E^{k-1}(S, \phi, \boldsymbol{a}; J_{HF}), d_{k-1}).$$

To be more explicit, consider the $\mathbb{Z}$–graded module

$$\widehat{C\mathcal{F}}(S, \phi, \boldsymbol{a}) := \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}) \otimes_{\mathbb{F}} \mathbb{F}[t, t^{-1}]$$

endowed with the endomorphism $\widehat{\partial}$ defined by

$$\widehat{\partial}\left(\sum_{i \in \mathbb{Z}} c_i t^i\right) := \sum_{i \in \mathbb{Z}} \left(\sum_{l \in \mathbb{Z}} (\partial_l c_i) t^{i-l}\right).$$

Here $c_i \neq 0$ for only finitely many $i \in \mathbb{Z}$. Note that the additivity property of J_+ implies that

$$\sum_{i+j=l} \partial_i \circ \partial_j = 0$$

for any $l \geq 0$; hence, $\widehat{\partial} \circ \widehat{\partial} = 0$, making $(\widehat{C\mathcal{F}}(S, \phi, \boldsymbol{a}), \widehat{\partial})$ into a filtered chain complex, where the pth filtration level

$$\mathcal{F}^p(S, \phi, \boldsymbol{a}) = \left\{\sum_{i \leq p} c_i t^i \;\middle|\; c_i \in \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})\right\}.$$

Then $(E^k(S, \phi, \boldsymbol{a}; J_{HF}), d_k)$ is the spectral sequence associated to this filtered chain complex, where d_k is the restriction of $\widehat{\partial}$ to $E^k(S, \phi, \boldsymbol{a}; J_{HF})$. To be more explicit, let A_p^k denote the subcomplex defined by

$$A_p^k = \{c \in \mathcal{F}^p(S, \phi, \boldsymbol{a}) \mid \widehat{\partial}c \in \mathcal{F}^{p-k}(S, \phi, \boldsymbol{a})\},$$

ie

$$A_p^k = \left\{\sum_{i \leq p} c_i t^i \;\middle|\; c_i \in \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}) \text{ with } \sum_{i=0}^{j} \partial_i c_{p+i-j} = 0 \text{ for } 0 \leq j < k\right\}.$$

Then

$$E_p^k(S, \phi, \boldsymbol{a}; J_{HF}) = \frac{A_p^k}{\widehat{\partial} A_{p+k-1}^{k-1} + A_{p-1}^{k-1}}.$$

A straightforward calculation shows that $E_0^k(S, \phi, \boldsymbol{a}; J_{HF})$ is isomorphic to

(2-7)
$$\frac{\mathcal{Z}^k(S, \phi, \boldsymbol{a}; J_{HF})}{\mathcal{B}^k(S, \phi, \boldsymbol{a}; J_{HF})},$$

where

$$Z^k(S, \phi, \boldsymbol{a}; J_{HF})$$
$$:= \left\{ c_0 \in \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}) \;\middle|\; \exists c_i \in \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}) \text{ for } 1-k \leq i \leq -1 \text{ with } \partial_0 c_0 = 0 \right.$$
$$\left. \text{and } \partial_j c_0 = \sum_{i=0}^{j-1} \partial_i c_{i-j} \text{ for } 0 < j < k \right\}$$

and

$$\mathcal{B}^k(S, \phi, \boldsymbol{a}; J_{HF})$$
$$:= \left\{ \sum_{i=0}^{k-1} \partial_i b_i \;\middle|\; b_i \in \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}) \text{ and } \sum_{i=0}^{k-1-j} \partial_i b_{i+j} = 0 \text{ for } 0 < j < k \right\}.$$

(Note that, for an element $\sum_{i \leq 0} c_i t^i \in A_0^k$, the chains c_i for $1 - k \leq i \leq -1$ are uniquely determined by c_0 up to chains a_i for $1 - k \leq i \leq -1$ belonging to some $\sum_{i \leq -1} a_i t^i \in A_{-1}^{k-1}$, and that Z^k is isomorphic to A_0^k / A_{-1}^{k-1}.) Since $\mathcal{F}^p(S, \phi, \boldsymbol{a}) \cong \mathcal{F}^{p-1}(S, \phi, \boldsymbol{a})$ canonically as chain complexes, $E_p^k(S, \phi, \boldsymbol{a}; J_{HF})$ is canonically isomorphic to the quotient (2-7) for every p.

By [23, Theorem 3.1], the distinguished generator $\vec{x}_\xi$ represents the Ozsváth–Szabó contact class $\hat{c}(\xi) \in \widehat{HF}(-M)$, and it satisfies $\partial_i \vec{x}_\xi = 0$ for all $i \geq 0$. This is because there is no Fredholm index-1 J_{HF}–holomorphic curve in $\Sigma \times [0, 1] \times \mathbb{R}$ satisfying conditions (M0)–(M6) in [35, Section 1] with $\vec{x}_\xi$ at its negative punctures and having empty intersection with $\{z\} \times [0, 1] \times \mathbb{R}$. Hence, $\vec{x}_\xi$ represents a cycle in $E^k(S, \phi, \boldsymbol{a}; J_{HF})$ for all $k \geq 1$.

Definition 2.1 Define $o(S, \phi, \boldsymbol{a}; J_{HF})$ to be the smallest nonnegative integer k such that the generator $\vec{x}_\xi$ represents the trivial class in $E^{k+1}(S, \phi, \boldsymbol{a}; J_{HF})$.

Ideally, one would like to show that $o(S, \phi, \boldsymbol{a}; J_{HF})$ does not depend on choices of $(S, \phi, \boldsymbol{a})$ and J_{HF}. This is not true in general. For example, consider the closed contact 3–manifold where the contact structure is supported by the open book decomposition (S, ϕ), where S is a 4–holed sphere and ϕ is the product of Dehn twists depicted in Figure 2, left. Using the basis of arcs $\boldsymbol{a}$ shown in Figure 2, left, and a generic *split* almost complex structure J_{HF}, we observe that the shaded domain $\mathcal{D}$ in Figure 2, right, has a unique holomorphic representative up to translation (see [49, Lemma 3.4]), and this is sufficient for the vanishing of the Ozsváth–Szabó contact class. A simple computation shows that $J_+(\mathcal{D}) = 2$. Therefore, $\vec{x}_\xi$ represents the trivial class in $E^2(S, \phi, \boldsymbol{a}; J_{HF})$, and $o(S, \phi, \boldsymbol{a}; J_{HF}) \leq 1$. Furthermore, using the symmetry of the
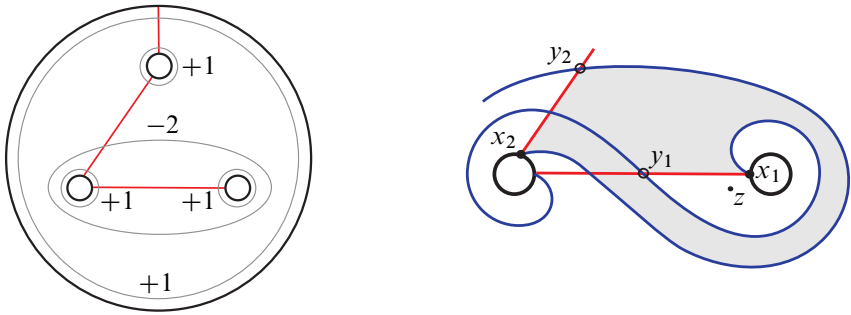
Figure 2: Left: an open book decomposition (S, ϕ) supporting an overtwisted contact structure and a basis of arcs depicted in red. Right: a Maslov index-1 holomorphic domain with $J_+ = 2$ in the $S \times \{0\}$ half of the Heegaard diagram $(-\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

open book decomposition and the choice of the arc basis, one can argue as in [31] that $o(S, \phi, \boldsymbol{a}; J_{HF}) = 1$. However, the contact structure supported by the open book decomposition (S, ϕ) is overtwisted, which can be seen after a sequence of positive stabilizations to reveal the overtwisted disk (see [55] for an explicit algorithm). Then there exists another open book decomposition (S', ϕ') and a basis of arcs $\boldsymbol{a}'$ on S' for which $o(S', \phi', \boldsymbol{a}'; J'_{HF}) = 0$ using a generic split almost complex structure J'_{HF} (see the proof of Theorem 2.3). As a result, o is not independent of these choices.

Definition 2.2 Let (M, ξ) be a closed contact 3–manifold. Then define the *spectral order*

$$\boldsymbol{o}(M, \xi) := \min\{o(S, \phi, \boldsymbol{a}; J_{HF})\},$$

where the minimum is taken over all data $(S, \phi, \boldsymbol{a}; J_{HF})$ such that (S, ϕ) is an open book decomposition of M supporting ξ, $\boldsymbol{a}$ is a collection of pairwise disjoint properly embedded arcs on S that contains a basis, and J_{HF} is a generic almost complex structure on $\Sigma \times [0, 1] \times \mathbb{R}$ satisfying conditions (J1)–(J5) in [35, Section 1].

It follows immediately that Definition 2.2 yields an invariant of contact structures. With the definition of our contact invariant in place, the first bullet point of Theorem 1.1 follows without much effort:

Theorem 2.3 Let ξ_{OT} be an overtwisted contact structure on a closed 3–manifold M. Then $\boldsymbol{o}(M, \xi_{OT}) = 0$.

Proof Note that an overtwisted contact structure is supported by an open book decomposition (S, ϕ) where the monodromy ϕ is not right-veering [21, Theorem 1.1].

One can find a basis of arcs $\boldsymbol{a}$ on S such that, in the corresponding Heegaard diagram, $\hat{\partial}_{\mathrm{HF}}\vec{\boldsymbol{y}} = \vec{\boldsymbol{x}}_{\xi_{\mathrm{OT}}}$, where $\boldsymbol{y} = \{y_1, x_2, \ldots, x_G\}$ and there is exactly one Maslov index-1 holomorphic domain $\mathcal{D}$, a bigon, that contributes to the differential [23, Lemma 3.2] as defined by a split complex structure on $\Sigma \times [0,1] \times \mathbb{R}$. Therefore, $n_{\boldsymbol{y}}(\mathcal{D}(A)) = \frac{1}{4}$, $n_{\boldsymbol{x}_{\xi_{\mathrm{OT}}}}(\mathcal{D}) = \frac{1}{4}$, $|\boldsymbol{y}| = G$ and $|\boldsymbol{x}_{\xi_{\mathrm{OT}}}| = G$. Applying (2-6), we find $J_+(\mathcal{D}) = 0$. As a result, $\boldsymbol{o}(M, \xi_{\mathrm{OT}}) = 0$. $\qquad\square$

3 Dependence on choices

This section investigates the question of dependence of $o(S, \phi, \boldsymbol{a}; J_{HF})$ on a choice of generic almost complex structure J_{HF} on $\Sigma \times [0,1] \times \mathbb{R}$, where $\Sigma = S \times \{\frac{1}{2}\} \cup_{\partial S} -S \times \{0\}$, a choice of the monodromy ϕ in its isotopy class, and how it changes under certain modifications of arc collections. We start with a priori dependence of o on a choice of generic almost complex structure.

3.1 Independence of almost complex structures

Proposition 3.1 *Fix an open book decomposition (S, ϕ) of M supporting ξ and a collection of pairwise disjoint properly embedded arcs $\boldsymbol{a}$ on S that contains a basis. Suppose that $(S, \phi, \boldsymbol{a})$ yields an admissible Heegaard diagram, and let J_{HF}^0 and J_{HF}^1 be two generic almost complex structures on $\Sigma \times [0,1] \times \mathbb{R}$ satisfying conditions (J1)–(J5) in [35, Section 1]. Then $o(S, \phi, \boldsymbol{a}; J_{HF}^0) = o(S, \phi, \boldsymbol{a}; J_{HF}^1)$.*

Proof There exists a smooth 1–parameter family of $\mathbb{R}$–invariant almost complex structures $\{J_{HF}^s\}_{s \in \mathbb{R}}$ on $\Sigma \times [0,1] \times \mathbb{R}$ that agrees with J_{HF}^0 if $s < \epsilon$ and with J_{HF}^1 if $s > 1 - \epsilon$ for some $\epsilon \ll 1$. As is explained in [35, Section 9], this family of almost complex structures can be chosen to satisfy conditions (J1), (J2) and (J4) in [35, Section 1] when considered as a non-$\mathbb{R}$–invariant almost complex structure on $\Sigma \times [0,1] \times \mathbb{R}$. Furthermore, this almost complex structure guarantees transversality for pseudoholomorphic curves with prescribed boundary conditions. It is used in [35, Section 9] to define a chain map

$$\Phi \colon (\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}), \hat{\partial}_{HF}^0) \to (\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}), \hat{\partial}_{HF}^1)$$

via a signed count of J_{HF}^s–holomorphic curves in $\Sigma \times [0,1] \times \mathbb{R}$ satisfying conditions (M0)–(M6) in [35, Section 1] and representing relative homology classes $A \in \hat{\pi}_2(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}})$ with $\mathrm{ind}(A) = 0$. If J_{HF}^s is generic, then the moduli space of such J_{HF}^s–holomorphic curves representing a class $A \in \hat{\pi}_2(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}})$ with $\mathrm{ind}(A) = 0$ (resp. $\mathrm{ind}(A) = 1$) is a smooth

orientable 0–dimensional (resp. 1–dimensional) manifold whose compactification in the 1–dimensional case is obtained by adding on pseudoholomorphic buildings of height 2 in which one level is J_{HF}^s–holomorphic and the other is either J_{HF}^0–holomorphic or J_{HF}^1–holomorphic as the case may be. The topology of the curves in each component of these moduli spaces is fixed.

Now we define an integer-valued function on moduli spaces of J_{HF}^s–holomorphic curves in $\Sigma \times [0,1] \times \mathbb{R}$ with ind ≤ 1 satisfying conditions (M0)–(M6) in [35, Section 1]. If C_L is such a curve representing a class in $\hat{\pi}_2(\vec{x}, \vec{y})$, then define

$$(3\text{-}1) \qquad J_+(C_L) := -\chi(C_L) + \mathrm{N} + |\boldsymbol{x}| - |\boldsymbol{y}|.$$

Note that (3-1) is additive in the sense that, if a pseudoholomorphic building of height 2 consists of a J_{HF}^0–holomorphic curve C_L^1 with ind $= 1$ representing a class in $\hat{\pi}_2(\vec{x}, \vec{x}')$ and a J_{HF}^s–holomorphic curve C_L^0 with ind $= 0$ representing a class in $\hat{\pi}_2(\vec{x}', \vec{y})$, then the J_{HF}^s–holomorphic curve C_L obtained from these by gluing (see [35, Appendix A]) satisfies

$$(3\text{-}2) \qquad J_+(C_L) = J_+(C_L^1) + J_+(C_L^0),$$

since $\chi(C_L) = \chi(C_L^1) + \chi(C_L^0) - \mathrm{N}$. The same holds for a pseudoholomorphic building of height 2 consisting of a J_{HF}^s–holomorphic curve C_L^0 with ind $= 0$ representing a class in $\hat{\pi}_2(\vec{x}, \vec{y}')$ and a J_{HF}^1–holomorphic curve C_L^1 with ind $= 1$ representing a class in $\hat{\pi}_2(\vec{y}', \vec{y})$. Note also that (3-1) coincides with (2-4), which allows us to deduce similarly that $J_+(C_L)$ is a nonnegative even integer. Hence, we may decompose Φ as

$$\Phi = \Phi^0 + \Phi^1 + \cdots + \Phi^l + \cdots,$$

where Φ^l counts J_{HF}^s–holomorphic curves with $J_+ = 2l$. Since Φ is a chain map and J_+ is additive under gluing, it follows that

$$\sum_{i+j=l} (\Phi^i \circ \partial_j^0 - \partial_i^1 \circ \Phi^j) = 0.$$

This identity implies that there is a filtered chain map $\hat{\Phi}$ from $(\widehat{\mathcal{CF}}(S, \phi, \boldsymbol{a}), \hat{\partial}^0)$ to $(\widehat{\mathcal{CF}}(S, \phi, \boldsymbol{a}), \hat{\partial}^1)$ defined by

$$\hat{\Phi}\left(\sum_{i \in \mathbb{Z}} c_i t^i \right) := \sum_{i \in \mathbb{Z}} \left(\sum_{l \in \mathbb{Z}} (\Phi^l c_i) t^{i-l} \right),$$

and hence a morphism of spectral sequences from $E^*(S, \phi, \boldsymbol{a}; J_{HF}^0)$ to $E^*(S, \phi, \boldsymbol{a}; J_{HF}^1)$. Moreover, $\Phi(\vec{x}_\xi) = \vec{x}_\xi$ since the only J_{HF}^s–holomorphic curve with negative ends at $\vec{x}_\xi$ satisfying conditions (M0)–(M6) in [35, Section 1] is $\vec{x}_\xi \times \mathbb{R}$. Therefore, we

have $o(S, \phi, \boldsymbol{a}; J_{HF}^0) \geq o(S, \phi, \boldsymbol{a}; J_{HF}^1)$. On the other hand, we may also consider the chain map induced by the smooth 1–parameter family of almost complex structures $\{J_{HF}^{1-s}\}_{s \in \mathbb{R}}$. Likewise, we obtain $o(S, \phi, \boldsymbol{a}; J_{HF}^0) \leq o(S, \phi, \boldsymbol{a}; J_{HF}^1)$. As a result, $o(S, \phi, \boldsymbol{a}; J_{HF}^0) = o(S, \phi, \boldsymbol{a}; J_{HF}^1)$. $\qquad\qquad\square$

3.2 Isotopy independence

Given [Proposition 3.1](#), we may drop a choice of generic almost complex structure from the notation and simply write $o(S, \phi, \boldsymbol{a})$. We proceed to discuss the dependence of o on the monodromy. In this regard, let ϕ and ϕ' be two orientation-preserving diffeomorphisms of S that restrict to the identity in a neighborhood of ∂S. Suppose that ϕ is isotopic to ϕ', and fix an isotopy $\{\phi_t\}_{t \in [0,1]}$ relative to ∂S such that $\phi_0 = \phi$ and $\phi_1 = \phi'$. Given a collection of pairwise disjoint properly embedded arcs $\boldsymbol{a}$ on S that contains a basis, the isotopy $\{\phi_t\}_{t \in [0,1]}$ yields an isotopy of arcs $\{\phi_t(\boldsymbol{b})\}_{t \in [0,1]}$, where $\boldsymbol{b}$ is the collection of arcs as in [Section 2.1](#). Of interest to us are two kinds of isotopies:

(1) For any $t \in [0, 1]$, $\boldsymbol{a}$ intersects $\phi_t(\boldsymbol{b})$ transversally in the interior of S.

(2) The isotopy creates/annihilates a pair of transverse intersections between $\boldsymbol{a}$ and $\phi(\boldsymbol{b})$.

Following [35], we refer to such isotopies as *basic isotopies*. In general, a pointed isotopy between two multipointed Heegaard diagrams, namely an isotopy supported in the complement of the basepoints, is called *admissible* if each intermediate multipointed Heegaard diagram is admissible. Any two admissible multipointed Heegaard diagrams that are pointed isotopic are in fact isotopic through a sequence of admissible basic isotopies (see [35, Proposition 5.6]). Note that isotopies of the monodromy of an open book decomposition yield pointed isotopies of the corresponding multipointed Heegaard diagram. Therefore, it suffices to investigate the behavior of o under admissible basic isotopies of the monodromy.

Proposition 3.2 *Let (S, ϕ) be an open book decomposition and $\boldsymbol{a}$ be a collection of pairwise disjoint properly embedded arcs $\boldsymbol{a}$ on S that contains a basis. Suppose that $(S, \phi, \boldsymbol{a})$ yields an admissible multipointed Heegaard diagram and that ϕ' is isotopic to ϕ via an admissible basic isotopy. Then $o(S, \phi', \boldsymbol{a}) = o(S, \phi, \boldsymbol{a})$.*

Proof As is explained in [35, Chapter 9] (see also [48, Section 7.3]), basic isotopies of the first kind above are equivalent to deformations of the complex structure on Σ.
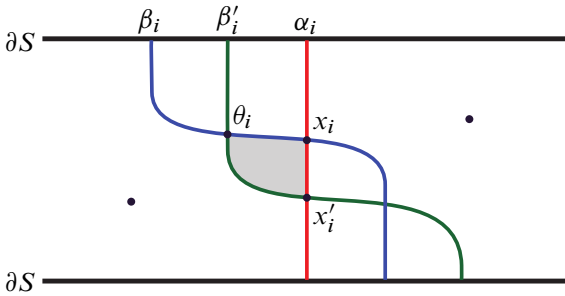
Figure 3: Part of the restriction of the multipointed Heegaard triple diagram $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$ to $S \times \{\frac{1}{2}\} \subset \Sigma$.

With this understood, o is unchanged under isotopies of this sort by Proposition 3.1. As for basic isotopies of the second kind above, we consider the chain maps induced by the multipointed Heegaard triple diagram $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$, where $\boldsymbol{\beta}' = \{\beta_1', \ldots, \beta_N'\}$ is such that each β_i' is obtained from a small Hamiltonian isotopy of $a_i \cup \phi'(b_i)$ so that it intersects β_i transversally in exactly two points near the point x_i, as shown in Figure 3, while it is disjoint from β_j for $j \neq i$. As a result, the Heegaard diagram $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta})$ represents the manifold $\#_G S^1 \times S^2$, where G is the genus of Σ; we may assume that the signed area of the region between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ is zero with respect to an area form on Σ which delivers the admissibility criteria for the multipointed Heegaard diagram $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, z)$ as stated in [35, Lemma 5.3]. Consequently, the multipointed Heegaard triple diagram $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$ is also admissible by [35, Lemma 10.14].

The Heegaard triple diagram $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, \boldsymbol{\alpha})$ describes a cobordism with one outgoing boundary component and two incoming boundary components, one of which is diffeomorphic to the manifold $\#_G S^1 \times S^2$. To be more specific, this cobordism is diffeomorphic to the complement of a tubular neighborhood of a bouquet of G embedded circles in the product cobordism $[0, 1] \times M$. It follows that there is a unique Spin^c structure $\mathfrak{t}_\xi$ on this cobordism which restricts to the trivial Spin^c structure $\mathfrak{s}_o$ on $\#_G S^1 \times S^2$ and to $\mathfrak{s}_\xi$ on M.

With the preceding understood, there exists a chain map

$$\hat{f}_{\boldsymbol{\beta}', \boldsymbol{\beta}, \boldsymbol{\alpha}; \mathfrak{t}_\xi} \colon \widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, \mathfrak{s}_o) \otimes_{\mathbb{F}} \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathfrak{s}_\xi) \to \widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}, \mathfrak{s}_\xi),$$

defined by counting embedded Fredholm index-0 pseudoholomorphic curves in $\Sigma \times T$ subject to appropriate boundary conditions. Here T denotes a disk with three marked points on its boundary and $\Sigma \times T$ is equipped with an almost complex structure satisfying conditions $(J'1)$–$(J'4)$ in [35, Section 10.2, page 1018].

No matter the almost complex structure, the differential on $\widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, \mathfrak{s}_\circ)$ vanishes identically. Therefore, restricting to the subcomplex $\mathbb{F} \cdot \vec{\boldsymbol{\theta}} \otimes_{\mathbb{F}} \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathfrak{s}_\xi)$, where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_N\}$ and $\vec{\boldsymbol{\theta}}$ is the top degree generator of $\widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, \mathfrak{s}_\circ)$, results in a chain map

$$\hat{f}_{\boldsymbol{\beta}', \boldsymbol{\beta}, \boldsymbol{\alpha}; \mathfrak{t}_\xi}(\vec{\boldsymbol{\theta}} \otimes \cdot) \colon \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathfrak{s}_\xi) \to \widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}, \mathfrak{s}_\xi).$$

The latter induces an isomorphism of homologies by [35, Proposition 11.4] (see also [48, Proposition 9.8]). In what follows, we work with a generic split complex structure on $\Sigma \times T$. We are allowed to do so since transversality of moduli spaces as defined by such almost complex structures can be guaranteed by slight perturbation of the $\boldsymbol{\alpha}-$, $\boldsymbol{\beta}-$ and $\boldsymbol{\beta}'-$curves. To be more precise, we may invoke the technique of [46]. This is because any class A in $\hat{\pi}_2(\vec{\boldsymbol{\theta}}, \cdot, \cdot)$ satisfies the boundary injectivity criterion in the sense of [35]. By way of a reminder, a class A in $\hat{\pi}_2(\vec{\boldsymbol{\theta}}, \cdot, \cdot)$ is said to satisfy the boundary injectivity criterion if any pseudoholomorphic curve u for some split complex structure on $\Sigma \times T$ representing the class A has $\pi_\Sigma \circ u$ somewhere injective in its boundary. This criterion is guaranteed as long as the domain representing the class has a region with multiplicity one adjacent to a region of multiplicity zero. Note that this is the case for any class in $\hat{\pi}_2(\vec{\boldsymbol{\theta}}, \cdot, \cdot)$ due to the placement of the basepoints, in that basepoints appear on both sides of every $\boldsymbol{\alpha}-$, $\boldsymbol{\beta}-$ and $\boldsymbol{\beta}'-$curve.

Next we show that the chain map $\hat{f}_{\boldsymbol{\beta}', \boldsymbol{\beta}, \boldsymbol{\alpha}; \mathfrak{t}_\xi}(\vec{\boldsymbol{\theta}} \otimes \cdot)$ induces a morphism of spectral sequences from $E^*(S, \phi, \boldsymbol{a}; J_{HF})$ to $E^*(S, \phi', \boldsymbol{a}; J'_{HF})$. First, define an analog of (2-1) for the cobordism described by the Heegaard triple diagram $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, \boldsymbol{\alpha})$ via

$$(3\text{-}3) \qquad J_+(A) = \tfrac{1}{2}N + \mu(\mathcal{D}(A)) - 2e(A) + |\boldsymbol{x}| - |\boldsymbol{y}|,$$

where $A \in \hat{\pi}_2(\vec{\boldsymbol{\theta}}, \vec{x}, \vec{y})$; $\mu(\mathcal{D}(A))$ denotes the Maslov index of the domain $\mathcal{D}(A)$ associated to A, which is the expected dimension of the moduli space of pseudoholomorphic curves representing the class A; and $e(A)$ is the Euler measure of the domain associated to the class A. If A can be represented by an embedded Fredholm index-0 pseudoholomorphic curve C_L, then (3-3) becomes

$$J_+(A) = \tfrac{1}{2}N - 2e(A) + |\boldsymbol{x}| - |\boldsymbol{y}| = \underbrace{-\chi(C_L) + N}_{\text{by } [35, \text{ Section } 10.2]} + |\boldsymbol{x}| - |\boldsymbol{y}|.$$

It follows from this formula that $J_+(A) = 2l$ for some $l \geq 0$. To see this, consider the smooth compact oriented surface C obtained from the compactification of C_L by first adding 2–dimensional 1–handles, one for each pair (β'_i, β_i) and one for each pair (β'_i, α_i), and then capping off the boundary components of the resulting surface

containing the I–chord collection $\vec{\boldsymbol{\theta}}$. Note that $\chi(C) = \chi(C_L) - \mathrm{N}$, and $|\boldsymbol{x}|$ (resp. $|\boldsymbol{y}|$) is equal to the number of boundary components of C arising from the I–chord collection $\vec{\boldsymbol{x}}$ (resp. $\vec{\boldsymbol{y}}$). The claim then follows in exactly the same way as in Section 2. Consequently, we can decompose the chain map $\hat{f}_{\boldsymbol{\beta}',\boldsymbol{\beta},\boldsymbol{\alpha};\mathfrak{t}_\xi}(\vec{\boldsymbol{\theta}} \otimes \cdot)$ as

$$\hat{f}_{\boldsymbol{\beta}',\boldsymbol{\beta},\boldsymbol{\alpha};\mathfrak{t}_\xi}(\vec{\boldsymbol{\theta}} \otimes \cdot) = f^0 + f^1 + \cdots + f^l + \cdots,$$

where f^l counts embedded Fredholm index-0 pseudoholomorphic curves with $J_+ = 2l$. Since the Maslov index and the Euler measure are additive under concatenation, it follows using (2-1) and (3-3) that J_+ is also additive. Therefore, we have

$$(3\text{-}4) \qquad \sum_{i+j=l} (f^i \circ \partial_j - \partial'_i \circ f^j) = 0$$

since $\hat{f}_{\boldsymbol{\beta}',\boldsymbol{\beta},\boldsymbol{\alpha};\mathfrak{t}_\xi}$ is a chain map and the J_+–filtered differential on $\widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, \mathfrak{s}_\circ)$ is identically zero. The latter is due to the fact that $\widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, \mathfrak{s}_\circ)$ is isomorphic to $(\mathbb{F}_{(0)} \oplus \mathbb{F}_{(1)})^{\otimes \mathrm{N}}$, where $\mathbb{F}_{(0)} \oplus \mathbb{F}_{(1)}$ is a graded module over $\mathbb{F}$ with vanishing differential and the domains corresponding to the pseudoholomorphic curves that contribute to the differential of the generator, $\theta_i \times [0, 1]$, of $\mathbb{F}_{(1)}$ are both bigons, which have $J_+ = 0$. In short, the restriction of the differential on $\widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\beta}, \mathfrak{s}_\circ) \otimes_\mathbb{F} \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathfrak{s}_\xi)$ to the subcomplex $\mathbb{F} \cdot \vec{\boldsymbol{\theta}} \otimes_\mathbb{F} \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathfrak{s}_\xi)$ is J_+–filtered.

The identity (3-4) implies that there is a filtered chain map from $(\widehat{\mathcal{CF}}(S, \phi, \boldsymbol{a}), \hat{\partial})$ to $(\widehat{\mathcal{CF}}(S, \phi', \boldsymbol{a}), \hat{\partial}')$ as before, and hence a morphism of spectral sequences from $E^*(S, \phi, \boldsymbol{a}; J_{HF})$ to $E^*(S, \phi', \boldsymbol{a}; J'_{HF})$. In addition,

$$\hat{f}_{\boldsymbol{\beta}',\boldsymbol{\beta},\boldsymbol{\alpha};\mathfrak{t}_\xi}(\vec{\boldsymbol{\theta}} \otimes \vec{\boldsymbol{x}}_\xi) = \vec{\boldsymbol{x}}'_\xi$$

since the shaded triangles in Figure 3 constitute the only holomorphic domain that contributes to this chain map due to the placement of the basepoints, and it is represented by a unique pseudoholomorphic curve by the Riemann mapping theorem. Hence, $o(S, \phi, \boldsymbol{a}; J_{HF}) \geq o(S, \phi', \boldsymbol{a}; J'_{HF})$. Likewise, the isotopy from ϕ' to ϕ yields $o(S, \phi, \boldsymbol{a}; J_{HF}) \leq o(S, \phi', \boldsymbol{a}; J'_{HF})$. As a result, $o(S, \phi, \boldsymbol{a}; J_{HF}) = o(S, \phi', \boldsymbol{a}; J'_{HF})$. $\square$

Remark Sarkar and Wang [54] and Plamenevskaya [52] proved that the Heegaard diagram resulting from an arbitrary choice of $(S, \phi, \boldsymbol{a})$, where $\boldsymbol{a}$ contains a basis, can be made *nice* by choosing ϕ appropriately in its isotopy class. On a nice Heegaard diagram, every Maslov index-1 holomorphic domain is represented by an *empty* embedded bigon or an *empty* embedded square [54, Theorem 3.3]. It is easy to see from (2-1) that such domains have either $J_+ = 0$ or $J_+ = 2$. This observation indicates that there should be a combinatorial description of o.
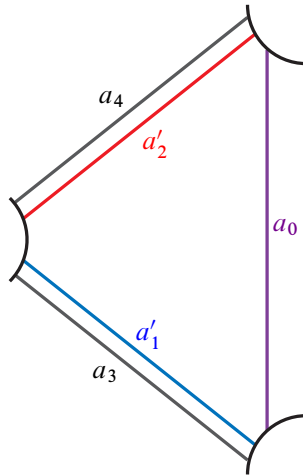
Figure 4: The configuration of arcs in the $S \times \{\frac{1}{2}\}$ page of the open book decomposition representing a triangle elimination operation.

3.3 Eliminating triangles

In this subsection, we investigate dependence of o on a choice of collection of pairwise disjoint properly embedded arcs containing a basis. More specifically, given an open book decomposition (S, ϕ) and such an arc collection $\boldsymbol{a}$ on S, we prove that o is nonincreasing under a *triangle elimination* operation on $\boldsymbol{a}$, which we will describe in a moment. As we shall see in Section 4, this operation gives us quite a bit of flexibility in our arguments that lead to the proofs of our main theorems. To set the stage, let (S, ϕ) be an open book decomposition supporting a contact structure ξ, and $\boldsymbol{a} = \{a_0, a_3, a_4, \dots, a_N\}$ be a collection of pairwise disjoint properly embedded arcs on S that contains a basis. Suppose that the three arcs $a_0, a_3, a_4 \in \boldsymbol{a}$ bound a connected component of $S \smallsetminus \bigcup \boldsymbol{a}$. Denote by $\boldsymbol{a}'$ the collection of pairwise disjoint properly embedded arcs on S obtained by discarding a_0 and "doubling" a_3 and a_4, ie $\boldsymbol{a}' = \{a_1', a_2', a_3, a_4, \dots, a_N\}$, where a_1' and a_2' are parallel and sufficiently close to a_3 and a_4, respectively (see Figure 4). Then:

Proposition 3.3 *Let (S, ϕ) be an open book decomposition, $\boldsymbol{a}$ be a collection of pairwise disjoint properly embedded arcs on S that contains a basis, and $\boldsymbol{a}'$ be obtained from $\boldsymbol{a}$ via triangle elimination. Then $o(S, \phi, \boldsymbol{a}') \le o(S, \phi, \boldsymbol{a})$.*

In preparation for the proof of the above proposition, we assume that the monodromy ϕ moves the arcs a_0, a_3 and a_4 to the right, since otherwise it would not move a_1'
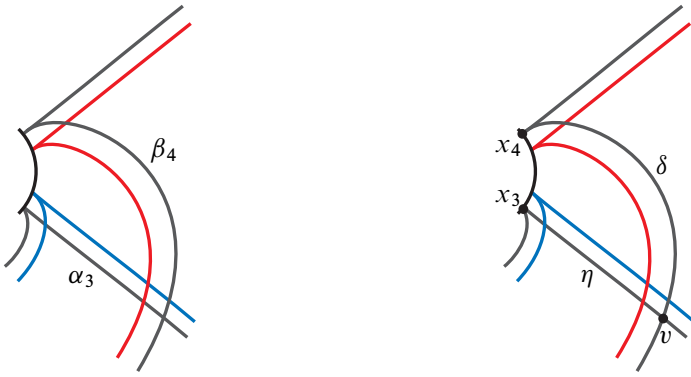
Figure 5: Left: the local behavior of the $\boldsymbol{\beta}$–curves as shown in the $S \times \{0\}$ half of the Heegaard diagram $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}')$. Right: the arc configuration in the desired nice Heegaard diagram with the arcs prohibited from forming bigons indicated.

and a_2' to the right either, resulting in both $o(S, \phi, \boldsymbol{a})$ and $o(S, \phi, \boldsymbol{a}')$ being zero as in the proof of Theorem 2.3. We further assume that β_4 stays parallel to the boundary of S immediately after turning right in the $S \times \{0\}$ half of the Heegaard diagram
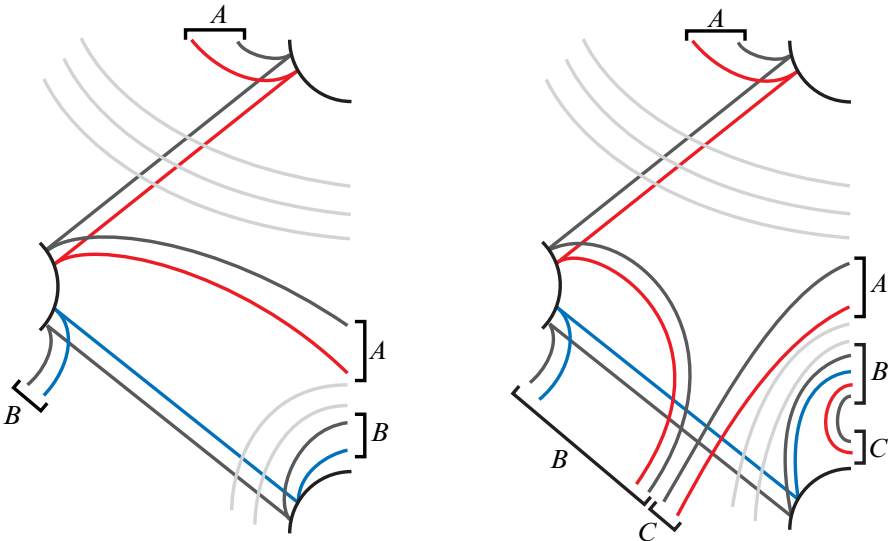

Figure 6: Left: configuration of arcs when β_4 doesn't stay parallel to the boundary of S immediately after turning right in the $S \times \{0\}$ half of the Heegaard diagram. Right: configuration of arcs after an isotopy to guarantee that β_4 intersects α_3 immediately after turning right. In both figures, brackets indicate the ends of arcs that are identified.

until it intersects α_3 as in Figure 5, left. Otherwise (see Figure 6, left), isotope the monodromy ϕ so as to guarantee that this is the case (see Figure 6, right). Note that, by Proposition 3.2, o is invariant under isotopies of the monodromy ϕ. With the preceding understood, we prove that we can work with a special kind of *nice* Heegaard diagram after a sequence of isotopies of the monodromy.

Lemma 3.4 *We may isotope the monodromy ϕ so that the multipointed Heegaard diagram $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}', \boldsymbol{z}')$ corresponding to $(S, \phi, \boldsymbol{a}')$ is nice while making sure that the intersection pattern as depicted in Figure 5, left, is preserved.*

Proof As is argued in [52], we may apply the algorithm of Sarkar and Wang [54, Section 4.1] to produce a nice Heegaard diagram by performing finger moves on $\boldsymbol{\beta}$–curves only in the $S \times \{0\}$ half of the Heegaard surface Σ. This is because, in a Heegaard diagram arising from an open book decomposition, there are regions with basepoints on either side of every $\boldsymbol{\beta}$–curve. In order to preserve the intersection pattern in Figure 5, left, we will show that these finger moves on $\boldsymbol{\beta}$–curves can be performed in such a way that the arc δ along β_4 between the points x_4 and v, shown in Figure 5, right, remains unchanged, and, in the resulting nice Heegaard diagram, no $\boldsymbol{\beta}$–curve forms a bigon with the arc η along α_3 between the points x_3 and v. It suffices to perform these finger moves in the Heegaard diagram resulting from the arc collection $\{a_3, a_4, \ldots, a_N\}$, which still contains a basis, since adding a'_1 and a'_2, parallel to a_3 and a_4, respectively, merely subdivides bigon and rectangle regions into smaller bigon and rectangle regions. With the preceding understood, we produce a nice diagram with the desired properties in the three steps that follow. Throughout, we change the definition of the *distance* of a region used in the Sarkar–Wang algorithm to be the minimum number of intersection points between the $\boldsymbol{\beta}$–curves and an arc connecting the interior of that region to a region with basepoint, with the arc taken to be in the complement of the $\boldsymbol{\alpha}$–curves *and* the arc δ.

Step 1 Note that, given a region, there is exactly one region with basepoint that can be connected to the interior of that region via an arc in the complement of the $\boldsymbol{\alpha}$–curves and the arc δ. Proceed as in the algorithm of Sarkar and Wang by first killing all nondisk regions without performing a finger move starting at δ and then performing finger moves as in the proof of [54, Lemma 4.1] to reduce the *distance d complexity* of the Heegaard diagram to (0) starting from *bad regions* with the largest distance. By way of reminder, the badness of a $2n$–gon is defined in [54, Section 4.1] to be $\max\{n-2, 0\}$,
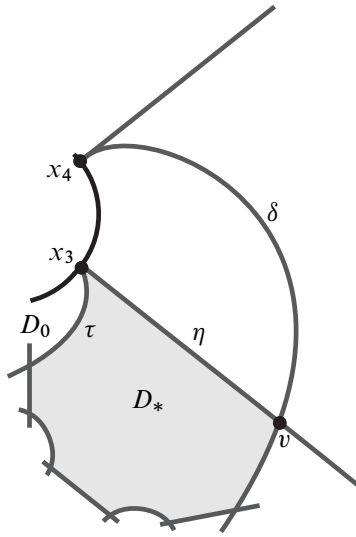
Figure 7: The domain D_*, the adjacent region with basepoint D_0, and the arc τ along β_3 which they both have on their boundaries.

and the distance d complexity of a multipointed Heegaard diagram is the tuple

$$\left(\sum_{i=1}^{m} b(D_i), -b(D_1), \dots, -b(D_m) \right),$$

where $D_1, \dots, D_m$ are all the distance d bad regions ordered in decreasing measure of badness $b(D_1) \geq \cdots \geq b(D_m)$. Given a distance d bad region, a finger move used to break up that region into regions of smaller badness as in the proof of [54, Lemma 4.1] starts from an arc along a $\boldsymbol{\beta}$–curve that is common to that bad region and another region of distance $d-1$. As a result of our definition of the distance of a region, the region without a basepoint that has the arc δ on its boundary is adjacent to a region with distance one less along an arc along a $\boldsymbol{\beta}$–curve other than δ. Therefore, at no point in the process do the finger moves needed to break up the former region into rectangles and bigons start at δ. Continue performing finger moves as in the proof of [54, Lemma 4.1] until the *distance* of the Heegaard diagram is reduced to 1; that is, until all bad regions are of distance at most 1.

Step 2 Having completed Step 1, all bad regions now have distance at most 1. The region with no basepoints and the arc η on its boundary has distance 1, and it is adjacent to a region with basepoint D_0 along β_3 (see Figure 7). Denote this region by D_*. The goal of this step is to break up every bad region except for D_*, if it is a bad region

at any point during the process, into rectangles and bigons, while avoiding crossing the arc η. Perform finger moves as in the proof of [54, Lemma 4.1], ignoring D_* in the measure of distance 1 complexity of the Heegaard diagram and stopping all finger moves once they enter D_*. Doing so breaks up every bad region other than D_* into rectangles and bigons and preserves the intersection pattern in Figure 5, left. We can do this because the Sarkar–Wang algorithm terminates after a finite number of finger moves, and we can stop those finger moves that enter D_* once they enter D_*. This modification of the algorithm does not increase the distance of any bad regions, and the modified algorithm eventually breaks up every bad region other than D_* into rectangles and bigons at the expense of possibly increasing the badness of D_*. The proof of the lemma is complete if $\mathcal{D}_*$ is not a bad region at the end of this step. Otherwise, we proceed to Step 3 in order to break up $\mathcal{D}_*$ into rectangles and bigons without changing badness of any other region without a basepoint.

Step 3 Finally, we break up the only remaining bad region, namely, D_*. We claim that we can perform a sequence of finger moves as in the proof of [54, Lemma 4.1] so that, in the resulting nice Heegaard diagram, no β–curve forms a bigon with η. We prove this claim by strong induction on the badness $b(D_*)$ of the region D_*. If $b(D_*) = 1$—that is, if D_* is a hexagon—then performing a finger move as in the proof of [54, Lemma 4.1] starting at the arc τ along β_3 with an end at x_3 on the boundary of D_* breaks D_* up into two rectangles. Moreover, since all other regions are either bigons or rectangles, this finger has to push through a "tunnel" of rectangular regions, forcing it to stay "parallel" to a β–curve. Therefore, it won't come back to D_*, since otherwise it would have to follow a full β–curve, which in turn would force our finger to cross a region with basepoint because there are regions with basepoint on either side of every β–curve. Next suppose that $b(D_*) > 1$ and perform a finger move as in the proof of [54, Lemma 4.1] starting at τ (see Figure 8, left). If the finger doesn't come back to D_*, then it will end up in a bigon region or a region with basepoint, and D_* will be broken up into a region $D_{*,1}$ with badness $b(D_*) - 1$ and a rectangle $D_{*,2}$. Note that both $D_{*,1}$ and $D_{*,2}$ are adjacent to D_0 along β_3, and that $D_{*,1}$ has the arc η on its boundary (see Figure 8, right). Then, by the induction hypothesis, the claim is true. Suppose instead that the finger comes back to D_*. Then, by the argument in [54, Subcase 4.2] and the fact that there are regions with basepoint on either side of every β–curve, there exists another finger move starting at τ that doesn't come back to D_*. This finger move would break D_* up into two regions, $D_{*,1}$ and $D_{*,2}$, both adjacent to D_0 along β_3 and $D_{*,1}$ having the arc η on its boundary. Then we have $b(D_{*,1}) + b(D_{*,2}) = b(D_*) - 1$
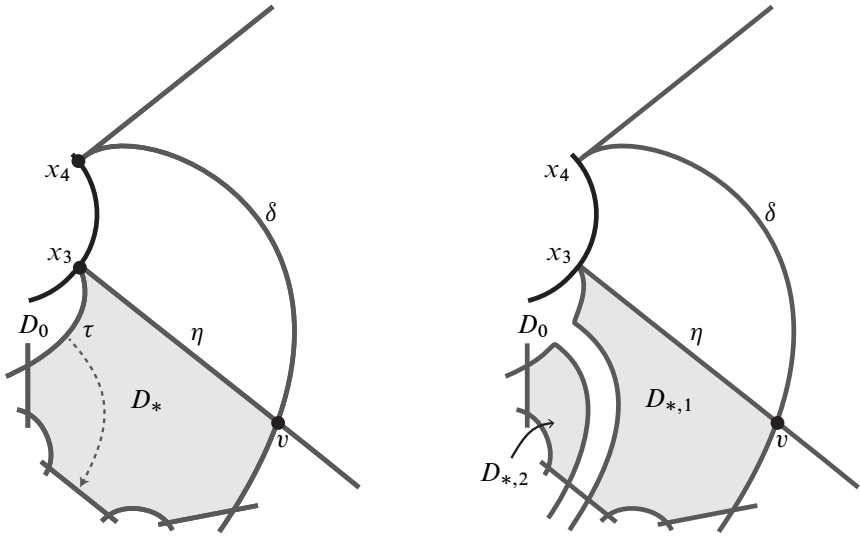
Figure 8: Left: the dashed line indicates the finger move to break up D_*.
Right: the regions $D_{*,1}$ and $D_{*,2}$ formed after the finger move.

and $b(D_{*,2}) \geq 1$. Once again, in contrast to the Sarkar–Wang algorithm, which requires ordering bad regions with increasing badness and then breaking up bad regions starting with the regions having the least positive badness, we first break up the region $D_{*,2}$ regardless of whether it is a bad region with the least positive badness. As we perform finger moves to break up $D_{*,2}$, as well as any subsequent new bad region that might emerge in that process, we stop a finger move that enters $D_{*,1}$ once it enters $D_{*,1}$, regardless of whether $b(D_{*,1}) > 0$ or not. In order to break up a bad region with badness b into rectangles, we need to perform exactly b finger moves, assuming no finger comes back to that region, and each finger pushed into a region would increase its badness by 1. Therefore, the process of breaking up $D_{*,2}$ into rectangles would increase the badness of $D_{*,1}$ by at most $b(D_{*,2})$. In the end, we have a Heegaard diagram with a single bad region of distance 1 adjacent to D_0 along τ having the arc η on its boundary. The badness of this region is at most $b(D_{*,1}) + b(D_{*,2}) = b(D_*) - 1$. Hence, by the induction hypothesis, our claim holds true, and a further sequence of finger moves as described above yields the desired nice Heegaard diagram.   □

With the above lemma understood, isotope a_0 and ϕ so that α_0 and β_0 intersect β_4 and α_3, respectively, to form bigons as in Figure 9, left. Then the multipointed Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{z})$ corresponding to $(S, \phi, \boldsymbol{a})$ is also nice. This is because outside the shaded areas in Figure 9, center and right, the multipointed Heegaard diagrams

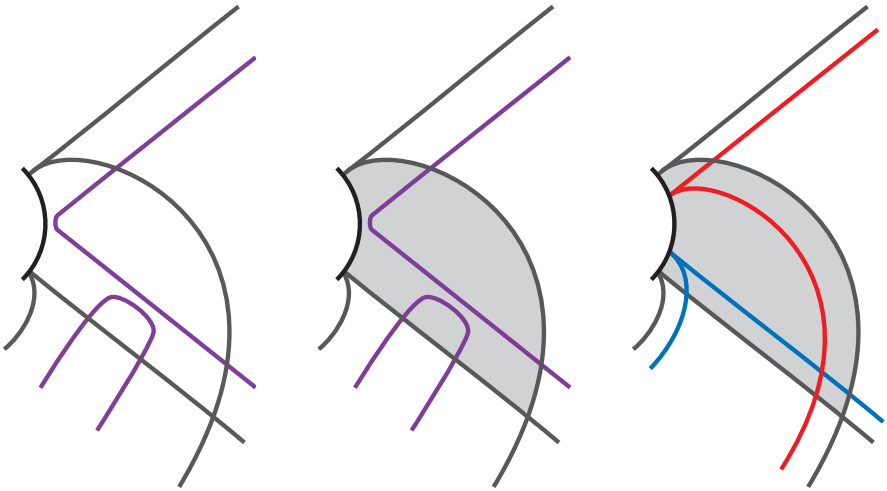Figure 9: Left: the $S \times \{0\}$ half of the multipointed Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{z})$. Center and right: the shaded areas in which the regions in the two multipointed Heegaard diagrams $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{z})$ and $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}', \boldsymbol{z}')$ essentially differ.

$(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{z})$ and $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}', \boldsymbol{z}')$, respectively, are isomorphic and, in the shaded area in Figure 9, center, all regions without a basepoint in $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{z})$ are bigons. Since o is invariant under isotopy of ϕ, we may assume without loss of generality that the monodromy ϕ is such that the multipointed Heegaard diagrams $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{z})$ and $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}', \boldsymbol{z}')$ are both nice and have the intersection patterns depicted in Figures 5, left, and 9, left. Note also that, in the multipointed Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{z})$, no $\boldsymbol{\alpha}$–curve other than α_0 forms a bigon with δ and no $\boldsymbol{\beta}$–curve other than β_0 forms a bigon with η.

Proof of Proposition 3.3 To start, associate to each $(\mathrm{N}-1)$–tuple of intersection points $\boldsymbol{y} = \{y_0, y_3, y_4, \dots, y_\mathrm{N}\}$ defining a generator $\vec{\boldsymbol{y}}$ of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ a unique $(\mathrm{N}-1)$–tuple of intersection points $\boldsymbol{y}'$ in $\boldsymbol{\alpha}' \cap \boldsymbol{\beta}'$ using the following recipe. For points belonging to $\boldsymbol{y}$ that lie on α_0 or β_0, associate a unique point in $\boldsymbol{\alpha}' \cap \boldsymbol{\beta}'$ according to the following rules:

- If $y_0 \in \alpha_0 \cap \beta_0$ and $y_0 \neq x_0$, then the associated point in $\boldsymbol{\alpha}' \cap \boldsymbol{\beta}'$ lies in $\alpha_i' \cap \beta_j'$, where $i, j \in \{1, 2\}$ (see Figure 10, left). If $y_0 = x_0$, we associate to it the point x_1'.
- If $y_0 \in \alpha_0 \cap \beta_j$ with $j \geq 3$, then the associated point in $\boldsymbol{\alpha}' \cap \boldsymbol{\beta}'$ lies in $\alpha_i' \cap \beta_j$, where $i \in \{1, 2\}$ (see Figure 10, center).
- If $y_i \in \alpha_i \cap \beta_0$ with $i \geq 3$, then the associated point in $\boldsymbol{\alpha}' \cap \boldsymbol{\beta}'$ lies in $\alpha_i \cap \beta_j'$, where $j \in \{1, 2\}$ (see Figure 10, right).
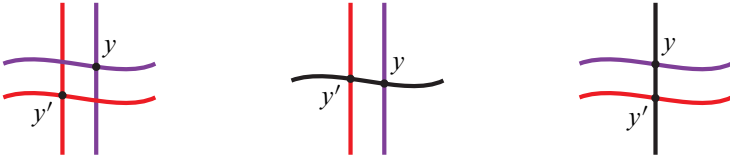
Figure 10: Assigning to an intersection point in $\boldsymbol{\alpha} \cap \boldsymbol{\beta}$ an intersection point in $\boldsymbol{\alpha}' \cap \boldsymbol{\beta}'$. Straight arcs indicate $\boldsymbol{\alpha}$–curves, while wavy arcs indicate $\boldsymbol{\beta}$–curves. Purple corresponds to α_0 or β_0, black corresponds to α_i or β_j for $i, j \geq 3$, and red corresponds to α_i' or β_j' for $i, j \in \{1, 2\}$.

In all other cases, the intersection points remain the same. Note that y' uses exactly one of α_1' or α_2', and exactly one of β_1' and β_2'. Depending on which pair of α_i' and β_j' that y' uses, we assign y the ordered pair $p_y := (i, j)$. Then, unless $p_y = (1, 2)$, we associate to y a unique N–tuple of intersection points $\tilde{y} := \{y_1', y_2', y_3', y_4', \ldots, y_N'\}$ defining a generator $\vec{\tilde{y}}$ of the chain complex $\widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}')$ by adding to y'

- the point x_2' if $p_y = (1, 1)$,
- the point $w \in \alpha_1' \cap \beta_2'$ indicated in Figure 11 if $p_y = (2, 1)$,
- the point x_1' if $p_y = (2, 2)$.

Note that this recipe associates to the distinguished (N−1)–tuple of intersection points $x_\xi = \{x_0, x_3, x_4, \ldots, x_N\}$ the distinguished N–tuple of intersection points

$$x_\xi' = \{x_1', x_2', x_3, x_4, \ldots, x_N\}.$$
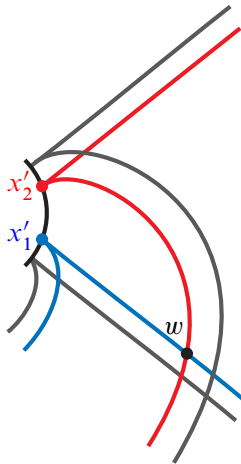

Figure 11: The intersection point w.

These two sets of intersection points define the distinguished generators that represent the Ozsváth–Szabó contact class in the homology of the chain complexes $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}')$, respectively.

Lemma 3.5 *Let $\mathcal{D} \in \hat{\pi}_2(\vec{\boldsymbol{y}}^1, \vec{\boldsymbol{y}}^2)$ be a Maslov index-1 holomorphic domain. Then $p_{\boldsymbol{y}^1} = p_{\boldsymbol{y}^2}$ unless $\mathcal{D}$ is a bigon. Furthermore, if $p_{\boldsymbol{y}^1} = (1, 2)$, then $p_{\boldsymbol{y}^2} = (1, 2)$.*

Proof Given $\boldsymbol{y}^1 = \{y_0^1, y_3^1, y_4^1, \dots, y_N^1\}$ defining a generator $\vec{\boldsymbol{y}}^1$ of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$, the first entry of the ordered pair $p_{\boldsymbol{y}^1}$ is determined by y_0^1, specifically by whether y_0^1 is near α_1' or α_2'. Similarly, the second entry of $p_{\boldsymbol{y}}$ is determined by $y_i^1 \in \alpha_i \cap \beta_0$, specifically by whether y_i^1 is near β_1' or β_2'. Let $\mathcal{D} \in \hat{\pi}_2(\vec{\boldsymbol{y}}^1, \vec{\boldsymbol{y}}^2)$ be a rectangular Maslov index-1 domain and $p_{\boldsymbol{y}^1} = (i, j)$. If $\mathcal{D}$ has neither an edge along α_0 nor an edge along β_0 on its boundary, then it follows at once from the definition of $p_{\boldsymbol{y}}$ that $p_{\boldsymbol{y}^2} = (i, j)$. If, on the other hand, $\mathcal{D}$ has an edge along α_0 and/or an edge along β_0 on its boundary, but it does not overlap the shaded area in Figure 9, center, then $\mathcal{D}$ has to have an edge parallel to α_3 or to α_4 depending on whether $i = 1$ or $i = 2$, and/or an edge parallel to β_3 or β_4 depending on whether $j = 1$ or $j = 2$ on its boundary; hence, $p_{\boldsymbol{y}^2} = (i, j)$. Finally, if $\mathcal{D}$ has an edge along α_0 and/or an edge along β_0 on its boundary, and it overlaps the shaded area in Figure 9, center, then it has an edge along α_0 or along β_0 running parallel to both α_3 and α_4 or to both β_3 and β_4, respectively, on its boundary. Such a rectangular domain would have to have an edge on its boundary along either another α_k or another β_k for some $k \geq 3$ running parallel to both α_3 and α_4 or to both β_3 and β_4, as the case may be. This would force either δ to form a bigon with α_k or η to form a bigon with β_k, since otherwise $\mathcal{D}$ would contain the bigon region between α_0 and δ or the bigon region between β_0 and η, and a Maslov index-1 rectangular domain in a nice Heegaard diagram can only be tiled by rectangular regions. But the nice Heegaard diagrams we produced in Lemma 3.4 do not allow any α_k to intersect δ or any β_k to intersect η for $k \geq 3$. Therefore, such a rectangular domain cannot exist. On the other hand, if $\mathcal{D}$ is a bigon and $p_{\boldsymbol{y}^1} \neq p_{\boldsymbol{y}^2}$, then we have either $p_{\boldsymbol{y}^1} = (2, j)$ or $p_{\boldsymbol{y}^1} = (i, 1)$ while $p_{\boldsymbol{y}^2} = (1, j)$ or $p_{\boldsymbol{y}^2} = (i, 2)$, respectively. (Think of the bigons formed between α_0 and δ, and between β_0 and η, as models.) It follows, in particular, that if $p_{\boldsymbol{y}^1} = (1, 2)$, then $p_{\boldsymbol{y}^2} = (1, 2)$. $\square$

Consequently, the submodule of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ generated by $\vec{\boldsymbol{y}}$ with $p_{\boldsymbol{y}} = (1, 2)$ is a subcomplex. We will denote this subcomplex by $\widehat{CF}_\circ(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ for future reference. Next we investigate the holomorphic domains contributing to the differential of a generator $\vec{\tilde{\boldsymbol{y}}}^1$ of $\widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}')$ corresponding to a generator $\vec{\boldsymbol{y}}^1$ of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$.
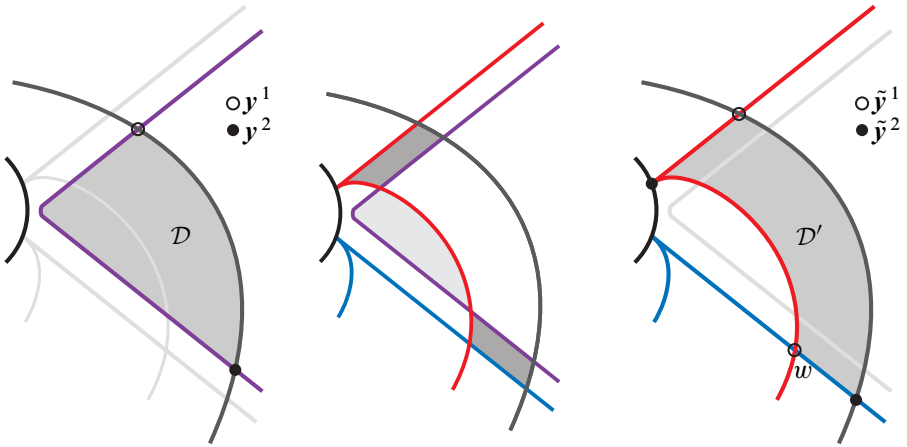
Figure 12: Constructing domains in $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}', z')$ from domains in $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$. Starting with a domain in the multipointed Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$ as on the left, add the darker shaded rectangular regions and subtract the lighter shaded bigon region in the center to get the domain in the multipointed Heegaard diagram $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}', z')$ shown on the right.

Lemma 3.6 *Given a generator $\boldsymbol{y}^1$ of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ and a generator $\vec{\tilde{y}}$ of $\widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}')$, if $p_{\boldsymbol{y}^1} \neq (1, 2)$, then there exists a Maslov index-1 holomorphic domain $\mathcal{D}' \in \hat{\pi}_2(\vec{\tilde{y}}^1, \vec{\tilde{y}})$ only if $\tilde{y} = \tilde{y}^2$ for some generator $\vec{\tilde{y}}^2$ of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ with $p_{\boldsymbol{y}^2} \neq (1, 2)$.*

Proof To see this, write $\tilde{y}^1 = \{y'^1_1, y'^1_2, \ldots, y'^1_N\}$ and $\tilde{y} = \{y'_1, y'_2, \ldots, y'_N\}$, and recall that either $y'^1_1 = x'_1$, $y'^1_1 = w$ or $y'^1_2 = x'_2$. If $y'^1_1 = x'_1$ or $y'^1_2 = x'_2$, then $y'_1 = x'_1$ or $y'_2 = x'_2$, respectively, since there are no nontrivial Maslov index-1 holomorphic domains with a corner at x'_1 or x'_2. If $y'^1_1 = w$, then either $y'_1 = x'_1$, $y'_1 = w$ or $y'_2 = x'_2$ since a Maslov index-1 holomorphic domain with a corner at w has to have a corner at x'_1 or x'_2. The latter is due to the fact that the multipointed Heegaard diagram $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}', z')$ is nice, so all Maslov index-1 holomorphic domains are empty embedded bigons or rectangles, and that starting at w and moving along α'_1 or β'_2 there is nowhere else to turn a corner other than at x'_1 or at x'_2. As a result, $\tilde{y} = \tilde{y}^2$ for some generator $\vec{\tilde{y}}^2$ of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ with $p_{\boldsymbol{y}^2} \neq (1, 2)$. □

Lemma 3.7 *Given generators $\vec{y}^1$ and $\vec{y}^2$ of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$, if $p_{\boldsymbol{y}^1} \neq (1, 2)$ and $p_{\boldsymbol{y}^2} \neq (1, 2)$, then there is a canonical one-to-one correspondence between Maslov index-1 holomorphic domains in $\hat{\pi}_2(\vec{y}^1, \vec{y}^2)$ and Maslov index-1 holomorphic domains in $\hat{\pi}_2(\vec{\tilde{y}}^1, \vec{\tilde{y}}^2)$.*

Proof Keep in mind that the Heegaard diagrams $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$ and $(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}', z')$ are both nice. In particular, a Maslov index-1 holomorphic domain has a unique holomorphic representative up to translation. If $\vec{y}^1$ and $\vec{y}^2$ are generators of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ with $p_{y^1} \neq (1, 2)$ and $p_{y^2} \neq (1, 2)$, then a Maslov index-1 holomorphic domain $\mathcal{D} \in \hat{\pi}_2(\vec{y}^1, \vec{y}^2)$ gives rise to a canonical Maslov index-1 holomorphic domain $\mathcal{D}' \in \hat{\pi}_2(\vec{\tilde{y}}^1, \vec{\tilde{y}}^2)$, and vice versa. If a domain $\mathcal{D}$ has neither α_0 nor β_0 on its boundary, then $\mathcal{D}' = \mathcal{D}$. Otherwise, to construct $\mathcal{D}'$ from $\mathcal{D}$ we add rectangular regions between α_0 and α_1', α_0 and α_2', β_0 and β_1' or β_0 and β_2', while removing the bigon regions between α_0 and β_2' or α_1' and β_0 as needed (see Figure 12). The former operation is reversible if $\mathcal{D}'$ has α_1' or α_2', and β_1' or β_2' on its boundary. $\qquad\square$

Lemma 3.8 If $\vec{y}^1$ and $\vec{y}^2$ are generators of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ with $p_{y^1} \neq (1, 2)$ and $p_{y^2} \neq (1, 2)$ and $\mathcal{D} \in \hat{\pi}_2(\vec{y}^1, \vec{y}^2)$ is a Maslov index-1 holomorphic domain, then the corresponding Maslov index-1 holomorphic domain $\mathcal{D}' \in \hat{\pi}_2(\vec{\tilde{y}}^1, \vec{\tilde{y}}^2)$ has $J_+(\mathcal{D}') = J_+(\mathcal{D})$.

Proof To see this, first note the following:

- If $p_y = (1, 1)$ or $p_y = (2, 2)$, then $|\tilde{y}| = |y| + 1$.
- If $p_y = (2, 1)$, then $|\tilde{y}| = |y|$.

As before, if $\mathcal{D}$ has neither α_0 nor β_0 on its boundary, then $\mathcal{D}' = \mathcal{D}$, and hence $J_+(\mathcal{D}') = J_+(\mathcal{D})$. Now suppose that $\mathcal{D}$ has either α_0 or β_0 on its boundary.

- If $\mathcal{D}$ is a rectangle, then $p_{y^1} = p_{y^2}$ (by Lemma 3.5) and $\mathcal{D}'$ is a rectangle. Hence, $|\tilde{y}^1| - |\tilde{y}^2| = |y^1| - |y^2|$ and $J_+(\mathcal{D}') = J_+(\mathcal{D})$.
- If $\mathcal{D}$ is a bigon, then $p_{y^1} = (2, 1)$ (otherwise $p_{y^2} = (1, 2)$) and either $p_{y^2} = (1, 1)$ or $p_{y^2} = (2, 2)$ (by Lemma 3.5), and $\mathcal{D}'$ is a rectangle. Hence, $|\tilde{y}^1| - |\tilde{y}^2| = |y^1| - |y^2| - 1$ and

$$J_+(\mathcal{D}') = 2 \cdot 1 - 1 + |\tilde{y}^1| - |\tilde{y}^2| = 2 \cdot \tfrac{1}{2} - 1 + |y^1| - |y^2| = J_+(\mathcal{D}),$$

  by (2-6). $\qquad\square$

By Lemma 3.5, the module $\widehat{CF}_\circ(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ generated by $\vec{y}$ with $p_y = (1, 2)$ is a subcomplex of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$. Therefore, we may construct the quotient complex $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})/\widehat{CF}_\circ(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$. Note that, since $p_{x_\xi} = (1, 1)$, it is sent under the quotient map $q \colon \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}) \to \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})/\widehat{CF}_\circ(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ to a nonzero class. The filtered extension of the quotient, $(\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})/\widehat{CF}_\circ(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})) \otimes_{\mathbb{F}} \mathbb{F}[t, t^{-1}]$, is canonically

isomorphic as a filtered chain complex to the quotient $\widehat{\mathcal{CF}}(S, \phi, \boldsymbol{a})/\widehat{\mathcal{CF}}_\circ(S, \phi, \boldsymbol{a})$. The quotient map

$$\widehat{\mathcal{CF}}(S, \phi, \boldsymbol{a}) \to \widehat{\mathcal{CF}}(S, \phi, \boldsymbol{a})/\widehat{\mathcal{CF}}_\circ(S, \phi, \boldsymbol{a})$$

is a filtered chain map and it induces a morphism of associated spectral sequences. Therefore, if we define $o_q(S, \phi, \boldsymbol{a})$ to be the spectral order as determined by the class $q(x_\xi)$ and the spectral sequence associated to the filtered quotient chain complex $(\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})/\widehat{CF}_\circ(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})) \otimes_{\mathbb{F}} \mathbb{F}[t, t^{-1}]$, then $o(S, \phi, \boldsymbol{a}) \geq o_q(S, \phi, \boldsymbol{a})$. Meanwhile, by Lemmas 3.6 and 3.7, there exists an injective map from $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})/\widehat{CF}_\circ(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ to $\widehat{CF}(\Sigma, \boldsymbol{\beta}', \boldsymbol{\alpha}')$ sending $\vec{x}_\xi$ to $\vec{x}'_\xi$, and hence an injective map of filtered chain complexes from $(\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})/\widehat{CF}_\circ(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})) \otimes_{\mathbb{F}} \mathbb{F}[t, t^{-1}]$ into $\widehat{\mathcal{CF}}(S, \phi, \boldsymbol{a}')$ by Lemma 3.8, which induces a morphism of associated spectral sequences. As a result, $o_q(S, \phi, \boldsymbol{a}) \geq o(S, \phi, \boldsymbol{a}')$, finishing the proof of Proposition 3.3.   □

Definition 3.9 It follows from Proposition 3.3 that, for the purpose of defining the contact invariant $\boldsymbol{o}$, it suffices to work with arc collections that are bases with multiple parallel copies of some arcs added, since one can always pass to such an arc collection, which we will refer to as a *multibasis*, via triangle elimination without increasing the value of o. In other words, we may define $\boldsymbol{o}(M, \xi)$ to be the minimum of $o(S, \phi, \boldsymbol{a})$ over all choices of open book decompositions (S, ϕ) of M supporting ξ and multibases $\boldsymbol{a}$.

4   Properties of o

The first bullet point of Theorem 1.1, that is, $\boldsymbol{o}$ vanishes for overtwisted contact structures, was proved at the end of Section 2. This section proves the remaining properties of the contact invariant $\boldsymbol{o}$ summarized in Theorems 1.1, 1.2 and 1.3.

To start, we establish a few basic properties of o. To do so, we work in a slightly more general context, where we consider arc collections that may not contain a basis. Let (S, ϕ) be an open book decomposition. Given an arc collection $\boldsymbol{a}$ on S that does not necessarily contain a basis, we can extend it to an arc collection $\tilde{\boldsymbol{a}}$ that contains a basis. Then we fix a generic almost complex structure J_{HF} for the multipointed Heegaard diagram $(\Sigma, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}, \tilde{z})$ associated to the arc collection $\tilde{\boldsymbol{a}}$. We may regard $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ as a submodule of $\widehat{CF}(\Sigma, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$ by identifying the generators of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ with the generators obtained from these via adding on the distinguished points lying in $S \times \{\frac{1}{2}\}$ for each of the arcs in $\tilde{\boldsymbol{a}} \smallsetminus \boldsymbol{a}$. Due to the placement of the basepoints, there can be no pseudo-holomorphic curves with negative punctures at the chords resulting from these points.

Therefore, the differentials on $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ and on the submodule of $\widehat{CF}(\Sigma, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$ that it is identified with coincide. As a result, we may consider $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ as a subcomplex of $\widehat{CF}(\Sigma, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$. With the preceding understood, the first basic property of o is that it is nonincreasing under enlargement of arc collections.

Lemma 4.1 *Suppose that $\boldsymbol{a}_1 \subset \boldsymbol{a}_2$ are two collections of pairwise disjoint properly embedded arcs on S. Then there exists a generic almost complex structure J_{HF} on $\Sigma \times [0, 1] \times \mathbb{R}$ and an inclusion of chain complexes*

$$I \colon \widehat{CF}(\Sigma, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_1) \to \widehat{CF}(\Sigma, \boldsymbol{\beta}_2, \boldsymbol{\alpha}_2), \quad \mathcal{I} \colon \widehat{\mathcal{CF}}(S, \phi, \boldsymbol{a}_1) \to \widehat{\mathcal{CF}}(S, \phi, \boldsymbol{a}_2)$$

such that the contact generator is mapped to the contact generator by the first inclusion while the latter inclusion induces a morphism of spectral sequences from $E^*(S, \phi, \boldsymbol{a}_1; J_{HF})$ to $E^*(S, \phi, \boldsymbol{a}_2; J_{HF})$; hence, $o(S, \phi, \boldsymbol{a}_1; J_{HF}) \geq o(S, \phi, \boldsymbol{a}_2; J_{HF})$.

Proof It suffices to find a generic almost complex structure J_{HF} on $\Sigma \times [0, 1] \times \mathbb{R}$ such that moduli spaces of J_{HF}–holomorphic curves associated to the Heegaard diagram $(\Sigma, \boldsymbol{\beta}_2, \boldsymbol{\alpha}_2)$ are cut out transversally, because this immediately implies transversality of moduli spaces of J_{HF}–holomorphic curves associated to the Heegaard diagram $(\Sigma, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_1)$. Having fixed such a generic almost complex structure, the inclusion map I is defined on the set of generators of $\widehat{CF}(\Sigma, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_1)$ by

$$I(\vec{y}) = \vec{y}',$$

where $\boldsymbol{y}' = \boldsymbol{y} \cup \{x_a\}_{a \in \boldsymbol{a}_2 \smallsetminus \boldsymbol{a}_1}$ and x_a is the unique intersection point of a and b for an arc $a \in \boldsymbol{a}_2 \smallsetminus \boldsymbol{a}_1$. It follows that $I(\vec{x}_\xi^1) = \vec{x}_\xi^2$. Meanwhile, the J_{HF}–holomorphic curves that define the differential acting on elements of the subgroup $I(\widehat{CF}(\Sigma, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_1))$ are the same as the J_{HF}–holomorphic curves that define the differential on $\widehat{CF}(\Sigma, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_1)$. Therefore, I is a chain map and the induced inclusion map $\mathcal{I}$ is a filtered chain map. The latter induces a morphism of spectral sequences from $E^*(S, \phi, \boldsymbol{a}_1; J_{HF})$ to $E^*(S, \phi, \boldsymbol{a}_2; J_{HF})$; hence, $o(S, \phi, \boldsymbol{a}_1; J_{HF}) \geq o(S, \phi, \boldsymbol{a}_2; J_{HF})$. □

The next lemma claims that o remains the same under suitable enlargement of the pages of an open book decomposition while keeping the arc collection untouched.

Lemma 4.2 *Let $\boldsymbol{a}$ be a collection of pairwise disjoint properly embedded arcs on S and S' be a compact oriented surface with boundary obtained from S by attaching 1–handles away from a neighborhood of $\partial \boldsymbol{a}$. Let $\phi' \colon S' \to S'$ be an orientation-preserving diffeomorphism whose restriction to $\boldsymbol{a}$ agrees with ϕ. Then there are generic*

almost complex structures J_{HF} and J'_{HF} to define the differentials on $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\widehat{CF}(\Sigma', \boldsymbol{\beta}, \boldsymbol{\alpha})$, respectively, such that $(\widehat{\mathcal{CF}}(S, \phi, \boldsymbol{a}), \widehat{\partial})$ and $(\widehat{\mathcal{CF}}(S', \phi', \boldsymbol{a}), \widehat{\partial}')$ are isomorphic as filtered chain complexes. As a result, $o(S, \phi, \boldsymbol{a}; J_{HF}) = o(S', \phi', \boldsymbol{a}; J'_{HF})$.

Proof　It follows from the description of the surface S' that $\boldsymbol{a}$ can also be seen as a pairwise disjoint collection of properly embedded arcs on S'. Moreover, there is a canonical one-to-one correspondence between unordered tuples of intersection points in the Heegaard diagrams $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ and $(\Sigma', \boldsymbol{\beta}, \boldsymbol{\alpha})$. Also note that Σ' is obtained from Σ by connect-summing with tori along regions in the Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ with basepoints. Therefore, having fixed a generic almost complex structure J_{HF} on $\Sigma \times [0, 1] \times \mathbb{R}$, we can "extend" it to a generic almost complex structure J'_{HF} on $\Sigma' \times [0, 1] \times \mathbb{R}$ so that the holomorphic domains in the pointed Heegaard diagrams $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$ and $(\Sigma', \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$ agree, and the claim follows. □

With the above understood, the proofs of Theorems 1.1, 1.2 and 1.3 require working with a more tractable version of $\boldsymbol{o}$:

Definition 4.3　Let (M, ξ) be a closed contact 3–manifold. Fix an open book decomposition $\mathcal{B} = (S, \phi)$ of M supporting ξ. Then define

$$\mathbb{o}(\mathcal{B}) := \min_{\boldsymbol{a}}\{o(S, \phi, \boldsymbol{a})\},$$

where the minimum is taken over all choices of multibasis $\boldsymbol{a}$ on S. Indeed,

$$\boldsymbol{o}(M, \xi) = \min_{\mathcal{B}}\{\mathbb{o}(\mathcal{B})\}.$$

The quantity $\mathbb{o}$ yields an invariant of open book decompositions. We would like to understand its behavior under positive stabilization. Recall that a positive stabilization of an open book decomposition (S, ϕ) is an open book decomposition (S', ϕ'), where S' is obtained from S by attaching a 1–handle H and ϕ' differs from ϕ by a right-handed Dehn twist around a simple closed curve $c \subset S'$ that intersects the cocore of H in exactly one point; in other words, $\phi' = \phi \circ \tau_c$. As we will show next, $\mathbb{o}$ is nonincreasing under positive stabilization. To prove this, we need the flexibility to move from one arc collection to another without increasing the value of o. Recall that one can pass from one basis on S to another via a sequence of *arc slides*. Given a basis $\{a_1, a_2, \ldots, a_G\}$ on S where a_1 and a_2 are adjacent — namely, there is an arc $\tau \subset \partial S$ with endpoints on a_1 and a_2 that intersects no other a_i — define $a_1 + a_2$ to be a properly embedded arc in S isotopic rel $\partial(a_1 \cup a_2) \smallsetminus \partial\tau$ to $a_1 \cup \tau \cup a_2$ and

disjoint from all other a_i. Then passing from $\{a_1, a_2, \ldots, a_G\}$ to $\{a_1 + a_2, a_2, \ldots, a_G\}$ is called an arc slide. Somewhat similarly, given a multibasis on S, one can pass to a multibasis containing an arbitrary arc basis on S via a sequence of *multiarc slide*s. Given a multibasis $\boldsymbol{a}$ containing a basis $\{a_1, a_2, \ldots, a_G\}$ on S where a_1 and a_2 are adjacent and $\boldsymbol{a}$ contains m parallel copies of the arc a_1, a multiarc slide removes all parallel copies of the arc a_1 and adds $m + 1$ parallel copies of the arc $a_1 + a_2$ as well as m additional parallel copies of the arc a_2. This modification is equivalent to adding a copy of the arc $a_1 + a_2$ and then removing each parallel copy of the arc a_1 one by one via triangle elimination, resulting in a new multibasis $\boldsymbol{a}'$. Note that a multiarc slide with $m = 1$ is *not* an arc slide.

Lemma 4.4 *Let $\boldsymbol{a}$ be a multibasis on S and $\boldsymbol{a}'$ be obtained from $\boldsymbol{a}$ by a multiarc slide. Then $o(S, \phi, \boldsymbol{a}') \leq o(S, \phi, \boldsymbol{a})$.*

Proof This follows readily from Lemma 4.1 and Proposition 3.3. □

Corollary 4.5 *Let $\mathcal{B} := (S, \phi)$ be an open book decomposition and $\mathcal{B}' := (S', \phi')$ be a positive stabilization of $\mathcal{B}$. Then $\mathbb{o}(\mathcal{B}') \leq \mathbb{o}(\mathcal{B})$.*

Proof Let $\boldsymbol{a}$ be a multibasis such that $\mathbb{o}(\mathcal{B}) = o(S, \phi, \boldsymbol{a})$. By a sequence of multiarc slides, pass to a multibasis $\boldsymbol{a}'$ on S that is disjoint from c. Then $o(S, \phi, \boldsymbol{a}') = o(S, \phi, \boldsymbol{a})$ by Lemma 4.4, since $\mathbb{o}(\mathcal{B}) = o(S, \phi, \boldsymbol{a})$, and $o(S', \phi', \boldsymbol{a}') = o(S, \phi, \boldsymbol{a}')$ by Lemma 4.2, since $\boldsymbol{a}'$ is disjoint from c. As a result,

$$\mathbb{o}(\mathcal{B}') \leq o(S', \phi', \boldsymbol{a}') = o(S, \phi, \boldsymbol{a}') = o(S, \phi, \boldsymbol{a}) = \mathbb{o}(\mathcal{B}). \qquad \square$$

Corollary 4.6 *Let $\mathcal{B} := (S, \phi)$ be an open book decomposition of M supporting ξ. Then we can apply a sequence of stabilizations to get to an open book decomposition $\mathcal{B}'$ that realizes $o(M, \xi)$.*

Proof This follows from Giroux correspondence together with Corollary 4.5. □

We move on to analyze the behavior of o under Legendrian surgery.

Proposition 4.7 *Let (S, ϕ) be an open book decomposition and $\boldsymbol{a}$ be any collection of pairwise disjoint properly embedded arcs on S that contains a basis. Suppose c is a homologically essential simple closed curve on S which meets each arc in $\phi(\boldsymbol{a})$ at most once. Then $o(S, \tau_c \circ \phi, \boldsymbol{a}) \geq o(S, \phi, \boldsymbol{a})$.*
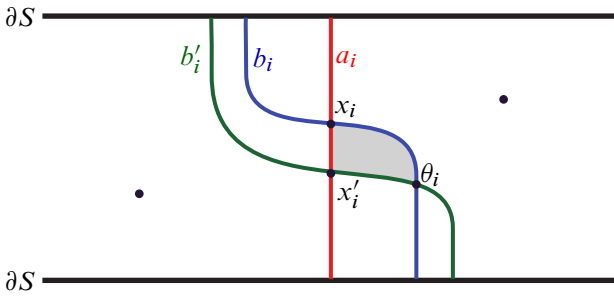
Figure 13: Part of the restriction of the multipointed Heegaard triple diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, z)$ to $S \times \{\frac{1}{2}\} \subset \Sigma$.

Proof To start, use $(S, \phi, \boldsymbol{a})$ and the curve c to form a multipointed triple Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, z)$, where $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_N\}$ with $\gamma_i = b_i' \times \{\frac{1}{2}\} \cup \tau_c \circ \phi(b_i') \times \{0\}$ such that b_i' is obtained from b_i by slightly pushing along ∂S in the direction of the boundary orientation as in Figure 13.

Notice that $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$ is the multipointed Heegaard diagram associated to $(S, \phi, \boldsymbol{a})$ and $(\Sigma, \boldsymbol{\gamma}, \boldsymbol{\alpha}, z)$ is the multipointed Heegaard diagram associated to $(S, \tau_c \circ \phi, \boldsymbol{a})$. Meanwhile, the multipointed Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma})$ describes a connected sum of some number of copies of the manifold $S^1 \times S^2$. Note also that the open book decomposition (S, τ_c) together with the collection of arcs $\{b_1, \ldots, b_N\}$ specifies the Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma})$, as in [23]. The chain complex $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma})$ has trivial
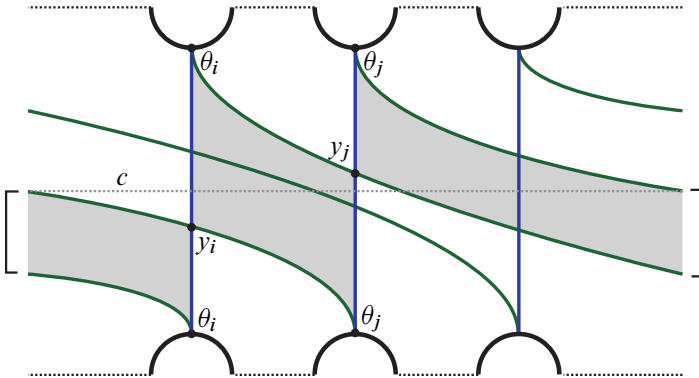
Figure 14: A local picture of the $-S \times \{0\} \subset \Sigma$ part of the Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma})$ near the surgery curve and all intersecting arcs. The shaded domains representing pseudoholomorphic curves with negative punctures at $\vec{\boldsymbol{\theta}}$ have the same J_+ value. The brackets indicate that the ends of the shaded region connect to one another.

differential and the generator $\vec{\theta}$ indicated in Figure 13 is the topmost generator. In fact, the J_+–filtered differential on $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is identically zero since all homology classes in $\hat{\pi}_2(\vec{\theta}, \cdot)$ have the same J_+ value (see Figure 14).

The placement of the basepoints guarantees, once again, that the multipointed triple Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, z)$ is admissible. Therefore, there is a chain map

$$(4\text{-}1) \qquad \hat{f}_{\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha}} \colon \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma}) \otimes_{\mathbb{F}} \widehat{CF}(\Sigma, \boldsymbol{\gamma}, \boldsymbol{\alpha}) \to \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$$

induced by the cobordism described by the triple Heegaard diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha})$. Since the differential on $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is identically zero, $\mathbb{F} \cdot \vec{\theta} \otimes_{\mathbb{F}} \widehat{CF}(\Sigma, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathfrak{s}_{\xi'})$ is a subcomplex of $\widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\gamma}) \otimes_{\mathbb{F}} \widehat{CF}(\Sigma, \boldsymbol{\gamma}, \boldsymbol{\alpha})$. Restricting (4-1) to this subcomplex, we obtain a chain map

$$\hat{f}_{\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha}}(\vec{\theta} \otimes \cdot) \colon \widehat{CF}(\Sigma, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathfrak{s}_{\xi'}) \to \widehat{CF}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathfrak{s}_{\xi}).$$

Therefore, having decomposed the above chain map as

$$\hat{f}_{\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha}}(\vec{\theta} \otimes \cdot) = f^0 + f^1 + \cdots + f^l + \cdots,$$

where f^l counts embedded Fredholm index-0 pseudoholomorphic curves with $J_+ = 2l$, we have

$$(4\text{-}2) \qquad \sum_{i+j=l} (f^i \circ \partial'_j - \partial_i \circ f^j) = 0$$

just as in Section 3. The identity (4-2) implies that there is a filtered chain map from $(\widehat{\mathcal{CF}}(S, \tau_c \circ \phi, \boldsymbol{a}), \hat{\partial}')$ to $(\widehat{\mathcal{CF}}(S, \phi, \boldsymbol{a}), \hat{\partial})$ and hence a morphism of spectral sequences from $E^*(S, \tau_c \circ \phi, \boldsymbol{a}; J'_{HF})$ to $E^*(S, \phi, \boldsymbol{a}; J_{HF})$. In addition, $\hat{f}_{\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha}}(\vec{\theta} \otimes \vec{x}'_\xi) = \vec{x}_\xi$ since the shaded triangle in Figure 13 is the only holomorphic domain that contributes to this chain map due to the placement of the basepoints, and it is represented by a unique pseudoholomorphic curve by the Riemann mapping theorem. Hence, $o(S, \tau_c \circ \phi, \boldsymbol{a}; J'_{HF}) \geq o(S, \phi, \boldsymbol{a}; J_{HF})$, as desired. $\qquad \square$

Corollary 4.8 *Let $\mathcal{B} := (S, \phi)$ be an open book decomposition and suppose $\mathcal{B}' := (S, \phi')$ is obtained from $\mathcal{B}$ by Legendrian surgery, ie $\phi' = \tau_{c_n} \circ \cdots \circ \tau_{c_1} \circ \phi$. Then*

$$(4\text{-}3) \qquad \mathbb{o}(\mathcal{B}) \leq \mathbb{o}(\mathcal{B}').$$

As a consequence, if $\mathcal{B} := (S, \phi)$ is an open book decomposition where ϕ can be written as a product of positive Dehn twists, then $\mathbb{o}(\mathcal{B}) = \infty$.

Proof We will apply Proposition 4.7 one Dehn twist at a time, noting that, for each Dehn twist curve c_i, we can find a multibasis $\boldsymbol{a}$ on S so that c_i intersects each arc in

the image of $\boldsymbol{a}$ under the monodromy at most once. With the preceding understood, for each $i \in \{0, 1, \ldots, n\}$ denote by $\mathcal{B}_i$ the open book decomposition (S, ϕ_i), where $\phi_0 = \phi$ and $\phi_i = \tau_{c_i} \circ \cdots \circ \tau_{c_1} \circ \phi$ for $i \in \{1, \ldots, n\}$. For each $i \in \{1, \ldots, n\}$, fix a multibasis $\boldsymbol{a}_i$ on S such that $\mathbb{o}(\mathcal{B}_i) = o(S, \phi_i, \boldsymbol{a}_i)$. Performing a sequence of multiarc slides, pass to a multibasis $\boldsymbol{a}'_i$ on S such that c_i intersects each arc in $\phi_{i-1}(\boldsymbol{a}'_i)$ at most once. It follows from [Lemma 4.4](#) and [Proposition 4.7](#) that

$$\mathbb{o}(\mathcal{B}_{i-1}) \leq o(S, \phi_{i-1}, \boldsymbol{a}'_i) \leq o(S, \phi_i, \boldsymbol{a}'_i) = \mathbb{o}(\mathcal{B}_i).$$

Concatenating these inequalities for $i \in \{1, \ldots, n\}$ while noting that $\mathcal{B}_0 = \mathcal{B}$ and $\mathcal{B}_n = \mathcal{B}'$, we achieve the first claim of the corollary.

The last claim of the corollary follows immediately from [(4-3)](#) once we note that $\mathbb{o}(S, \mathrm{id}_S) = \infty$. The latter is because the J_+–filtered differential in the corresponding Heegaard Floer chain complex is zero. □

With all the results needed in place, we are ready to prove [Theorem 1.2](#), and the second bullet point of [Theorem 1.1](#).

Proof of [Theorem 1.2](#) First note that, if (M', ξ') is obtained from (M, ξ) by attaching a Weinstein 1–handle, then an open book decomposition supporting ξ' can be built from an open book decomposition (S, ϕ) supporting ξ by attaching a 1–handle to S and extending the monodromy ϕ as the identity over this handle. It is easy to see that the latter operation does not change the value of $\mathbb{o}$, which then leads to the conclusion that $\boldsymbol{o}(M', \xi') = \boldsymbol{o}(M, \xi)$.

Next, for the case of a Weinstein 2–handle, let (M', ξ') be obtained from (M, ξ) by Legendrian surgery on a single curve c in (M, ξ). Let c' be the Legendrian curve in (M', ξ') that is the core of the surgery solid torus, which has the property that contact $+1$–surgery on it yields (M, ξ). The Legendrian c' lies on a page of an open book decomposition of M' supporting ξ', which by [Corollary 4.6](#) one can positively stabilize a number of times to get to an open book decomposition $\mathcal{B}'$ which realizes $\boldsymbol{o}(M', \xi')$; namely, $\mathbb{o}(\mathcal{B}') = \boldsymbol{o}(M', \xi')$. Now let $\mathcal{B}$ be the open book decomposition of M supporting ξ that is obtained by contact $+1$–surgery on c'. By [Corollary 4.8](#),

$$\boldsymbol{o}(M, \xi) \leq \mathbb{o}(\mathcal{B}) \leq \mathbb{o}(\mathcal{B}') = \boldsymbol{o}(M', \xi').$$ □

Corollary 4.9 Let (M, ξ) be Stein-fillable. Then $\boldsymbol{o}(M, \xi) = \infty$.

Proof A Stein-fillable contact 3–manifold admits a supporting open book decomposition (S, ϕ), where ϕ is a product of positive Dehn twists. To be more explicit, a Stein-fillable contact 3–manifold can be obtained via Legendrian surgery on some connected sum $\#_N S^1 \times S^2$ equipped with its standard contact structure ξ_{std} (see [18]). Therefore, by Theorem 1.2, it suffices to prove that $o(\#_N S^1 \times S^2, \xi_{\text{std}}) = \infty$. To see this, let $\mathcal{B}$ be an open book decomposition of $\#_N S^1 \times S^2$ supporting ξ_{std} which realizes $o(\#_N S^1 \times S^2, \xi_{\text{std}})$; in other words, $\circ(\mathcal{B}) = o(\#_N S^1 \times S^2, \xi_{\text{std}})$. As $(\#_N S^1 \times S^2, \xi_{\text{std}})$ is supported by an open book with trivial monodromy, a common stabilization, $\mathcal{B}'$, of that and $\mathcal{B}$ will have a monodromy which can be written as a product of positive Dehn twists and will also realize the minimal o. To see this, note that, by the second claim in Corollary 4.8, we have $\circ(\mathcal{B}') = \infty$. By Corollary 4.5, we also have $\circ(\mathcal{B}') \leq \circ(\mathcal{B}) = o(\#_N S^1 \times S^2, \xi_{\text{std}})$. Therefore, $o(\#_N S^1 \times S^2, \xi_{\text{std}}) = \infty$. $\qquad\square$

Next we prove the third bullet point of Theorem 1.1:

Theorem 4.10 *Given an open book decomposition $\mathcal{B} = (S, \phi)$ of M supporting ξ, and a basis $\boldsymbol{a}$ on S, there exists a multibasis $\boldsymbol{a}^m$ on S containing $\boldsymbol{a}$ such that*

$$o(S, \phi, \boldsymbol{a}^m) = o(M, \xi).$$

Proof By Corollary 4.6, we can positively stabilize $\mathcal{B}$ to pass to an open book decomposition $\mathcal{B}' = (S', \phi')$ with $\circ(\mathcal{B}') = o(M, \xi)$, where S' is built from S by adding 1–handles and $\phi' = \tau_{c_n} \circ \cdots \circ \tau_{c_1} \circ \phi$. Extending ϕ to S' as the identity on all the 1–handles, we form the open book decomposition $\widetilde{\mathcal{B}} = (S', \phi)$. Since ϕ' is obtained from ϕ by adding positive Dehn twists, $\circ(\widetilde{\mathcal{B}}) \leq \circ(\mathcal{B}')$ by Corollary 4.8.

Now fix a multibasis $\boldsymbol{a}'$ on S' such that

$$o(S', \phi', \boldsymbol{a}') = \circ(\mathcal{B}') = o(M, \xi).$$

Let $a_1, \ldots, a_n$ denote the cocores of the 1–handles added to S so as to build S' and perform a sequence of multiarc slides so as to pass to a multibasis $\boldsymbol{a}''$ that contains the arcs $a_1, \ldots, a_n$ and satisfies $o(S', \phi, \boldsymbol{a}'') = \circ(\widetilde{\mathcal{B}})$. We also have $o(S', \phi', \boldsymbol{a}') = o(S', \phi', \boldsymbol{a}'')$ by Lemma 4.4. Let $\boldsymbol{a}^\circ = \boldsymbol{a}'' \cap S$ and note that $\boldsymbol{a}^\circ$ is a multibasis on S. Furthermore, ϕ acts trivially on all arcs in $\boldsymbol{a}'' \smallsetminus \boldsymbol{a}^\circ$. Looking at the Heegaard diagram resulting from $(S', \phi, \boldsymbol{a}'')$, the α– and β–curves corresponding to arcs in $S' \smallsetminus S$ intersect each other exactly twice, forming two canceling bigons and thus contributing zero to $\widehat{\partial}_{HF}$. Furthermore, α_i and β_i intersect no other $\boldsymbol{\alpha}$–curves or $\boldsymbol{\beta}$–curves. Thus,

$$\widehat{\mathcal{CF}}(S', \phi, \boldsymbol{a}'') \equiv \widehat{\mathcal{CF}}(S', \phi, \boldsymbol{a}^\circ) \otimes_{\mathbb{F}} \left(\mathbb{F}_{(0)} \oplus \mathbb{F}_{(1)}\right)^{\otimes n},$$

where $\mathbb{F}_{(0)} \oplus \mathbb{F}_{(1)}$ is a graded module over $\mathbb{F}$ with vanishing differential and n is the number of arcs in $\boldsymbol{a}'' \smallsetminus \boldsymbol{a}°$. In particular,

$$o(S', \phi, \boldsymbol{a}'') = o(S', \phi, \boldsymbol{a}°).$$

By Lemma 4.2, we have $o(S, \phi, \boldsymbol{a}°) = o(S', \phi, \boldsymbol{a}°)$. Consequently,

$$o(S, \phi, \boldsymbol{a}°) = o(S', \phi, \boldsymbol{a}°) = o(S', \phi, \boldsymbol{a}'') \leq o(S', \phi', \boldsymbol{a}'') = o(S', \phi', \boldsymbol{a}') = \boldsymbol{o}(M, \xi).$$

Since, by definition, $o(S, \phi, \boldsymbol{a}°) \geq \boldsymbol{o}(M, \xi)$, we have $o(S, \phi, \boldsymbol{a}°) = \boldsymbol{o}(M, \xi)$. Finally, given a basis $\boldsymbol{a}$ on S, perform a sequence of multiarc slides to pass from $\boldsymbol{a}°$ to a multibasis $\boldsymbol{a}^m$ on S containing $\boldsymbol{a}$. Then, by Lemma 4.4,

$$o(S, \phi, \boldsymbol{a}^m) = o(S, \phi, \boldsymbol{a}°) = \boldsymbol{o}(M, \xi). \qquad \square$$

Remark Given an open book decomposition (S, ϕ) and a multibasis $\boldsymbol{a}$ on S, we can positively stabilize (S, ϕ) to pass to a new open book decomposition where $\boldsymbol{a}$ becomes a basis. Then it follows from Corollary 4.6 and Theorem 4.10 that $\boldsymbol{o}(M, \xi) = o(S, \phi, \boldsymbol{a})$ for some open book decomposition (S, ϕ) supporting the contact structure ξ and a basis $\boldsymbol{a}$ on S.

Another application of the Legendrian surgery statement in Theorem 1.2 is Theorem 1.3, namely that the spectral order of a contact connected sum is the minimum of the orders of the summands:

Proof of Theorem 1.3 Let $\mathcal{B}_1 = (S_1, \phi_1)$ and $\mathcal{B}_2 = (S_2, \phi_2)$ be open book decompositions which realize $\boldsymbol{o}(M_1, \xi_1)$ and $\boldsymbol{o}(M_2, \xi_2)$, respectively. Fix multibases $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ on S_1 and S_2, respectively, such that $\mathbb{o}(\mathcal{B}_i) = o(S_i, \phi_i, \boldsymbol{a}_i)$ for $i = 1, 2$. Then both $\widehat{\mathcal{CF}}(S_1, \phi_1, \boldsymbol{a}_1)$ and $\widehat{\mathcal{CF}}(S_2, \phi_2, \boldsymbol{a}_2)$ can be seen as filtered subcomplexes of $\widehat{\mathcal{CF}}(S_\#, \phi_\#, \boldsymbol{a}_\#)$, where $\mathcal{B}_1 \# \mathcal{B}_2 = (S_\#, \phi_\#)$ is the boundary connected sum open book decomposition with $\phi_\# = \phi_2 \circ \phi_1$, where we extend each by the identity across the complementary subsurface and $\boldsymbol{a}_\# = \boldsymbol{a}_1 \sqcup \boldsymbol{a}_2$. Hence, by Lemmas 4.1 and 4.2,

$$\boldsymbol{o}(M_1 \# M_2, \xi_1 \# \xi_2) \leq \mathbb{o}(\mathcal{B}_1 \# \mathcal{B}_2) \leq \mathbb{o}(\mathcal{B}_i) = \boldsymbol{o}(M_i, \xi_i)$$

for both $i = 1$ and $i = 2$, and $\boldsymbol{o}(M_1 \# M_2, \xi_1 \# \xi_2) \leq \min\{\boldsymbol{o}(M_1, \xi_1), \boldsymbol{o}(M_2, \xi_2)\}$.

For the reverse inequality, let $\mathcal{B} = (S, \phi)$ be a stabilization of $\mathcal{B}_1 \# \mathcal{B}_2$ realizing $\boldsymbol{o}(M_1 \# M_2, \xi_1 \# \xi_2)$. Ignore the extra positive Dehn twists on $\mathcal{B}$ which arise from its description as a positive stabilization of $\mathcal{B}_1 \# \mathcal{B}_2$. The resulting open book decomposition $\mathcal{B}' = (S, \phi')$ describes the 3–manifold $M_1 \# M_2 \# \#_k S^1 \times S^2$ for some k, the page S contains $S_1 \natural S_2$ as a subsurface due to $\mathcal{B}$ being a positive stabilization of $\mathcal{B}_1 \# \mathcal{B}_2$,

and the monodromy ϕ' extends $\phi_\#$ as the identity to the rest of S. In particular, $\mathcal{B}$ is obtained from $\mathcal{B}'$ by Legendrian surgery along curves contained in a page of $\mathcal{B}$; hence,

$$\boldsymbol{o}(M_1 \# M_2, \xi_1 \# \xi_2) = \mathbb{o}(\mathcal{B}) \geq \mathbb{o}(\mathcal{B}'),$$

by Theorem 1.2.

Fix a multibasis $\boldsymbol{a}'$ on S such that $\mathbb{o}(\mathcal{B}') = o(S, \phi', \boldsymbol{a}')$. After a sequence of multiarc slides, we can pass to a multibasis $\tilde{\boldsymbol{a}}$ on S which contains $\boldsymbol{a}_1 \sqcup \boldsymbol{a}_2$. By Lemma 4.4 $\mathbb{o}(\mathcal{B}') = o(S, \phi', \tilde{\boldsymbol{a}})$, and we have

$$\widehat{\mathcal{CF}}(S, \phi', \tilde{\boldsymbol{a}}) \cong \widehat{\mathcal{CF}}(S_1, \phi_1, \boldsymbol{a}_1) \otimes_{\mathbb{F}} \widehat{\mathcal{CF}}(S_2, \phi_2, \boldsymbol{a}_2) \otimes_{\mathbb{F}} (\mathbb{F}_{(0)} \oplus \mathbb{F}_{(1)})^{\otimes l}$$

as filtered chain complexes, where $\mathbb{F}_{(0)} \oplus \mathbb{F}_{(1)}$ is a graded module over $\mathbb{F}$ with vanishing differential and l is some nonnegative integer. As a result, $\mathbb{o}(\mathcal{B}') = o(S, \phi', \tilde{\boldsymbol{a}}) = \min\{o(S_1, \phi_1, \boldsymbol{a}_1), o(S_2, \phi_2, \boldsymbol{a}_2)\}$. On the other hand, since $o(S_1, \phi_1, \boldsymbol{a}_1) = \mathbb{o}(\mathcal{B}_1) = \boldsymbol{o}(M_1, \xi_1)$ and $o(S_2, \phi_2, \boldsymbol{a}_2) = \mathbb{o}(\mathcal{B}_2) = \boldsymbol{o}(M_2, \xi_2)$, by the above inequality we have

$$\min\{\boldsymbol{o}(M_1, \xi_1), \boldsymbol{o}(M_2, \xi_2)\} \leq \boldsymbol{o}(M_1 \# M_2, \xi_1 \# \xi_2). \qquad \square$$

Corollary 4.11 *For any surface S with boundary, the set of monodromies yielding open book decompositions supporting contact 3–manifolds (M, ξ) with $\boldsymbol{o}(M, \xi) \geq k$ forms a monoid in the mapping class group* $\mathrm{Mod}(S, \partial S)$.

We use $\boldsymbol{o}^k(S)$ to denote this monoid.

Proof By [1], for any two mapping classes ϕ_1 and ϕ_2, there is a Stein cobordism starting at the disconnected contact manifold $(M_{\phi_1}, \xi_{\phi_1}) \sqcup (M_{\phi_2}, \xi_{\phi_2})$ and ending at $(M_{\phi_2 \circ \phi_1}, \xi_{\phi_2 \circ \phi_1})$. By Theorems 1.2 and 1.3, this implies that

$$\boldsymbol{o}(M_{\phi_2 \circ \phi_1}, \xi_{\phi_2 \circ \phi_1}) \geq \boldsymbol{o}((M_{\phi_1}, \xi_{\phi_1}) \sqcup (M_{\phi_2}, \xi_{\phi_2}))$$
$$= \min\{\boldsymbol{o}(M_{\phi_1}, \xi_{\phi_1}), \boldsymbol{o}(M_{\phi_2}, \xi_{\phi_2})\}. \qquad \square$$

5 Obstructing Stein-fillability

In this section, we use spectral order to obstruct Stein-fillability by demonstrating a family of contact 3–manifolds with nonzero Ozsváth–Szabó contact class but with zero spectral order. In Section 5.1, we give a warm-up example of this application on a contact manifold which had previously been shown to be nonfillable in [39; 9]. In Section 5.2, we generalize this method to a previously unstudied family of contact 3–manifolds thereby proving Theorem 1.4. Finally, in Section 5.3, we compare this
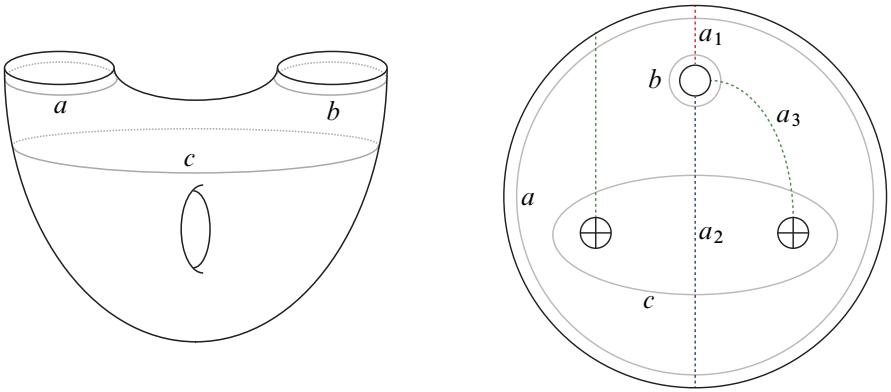
Figure 15: Left: the open book decomposition (S, ϕ) supporting the contact 3–manifold (Y, ξ), where $\phi = \tau_a \tau_b \tau_c^{-1}$. Right: the basis of arcs $\boldsymbol{a} = \{a_1, a_2, a_3\}$ on S, where the two middle circles intersecting a_3 decorated with "plus" are identified.

method to other techniques in the literature for obstructing symplectic-fillability (in its various forms) in the context of these examples.

5.1 A warm-up

We start with a warm-up example (Y, ξ), which is the base case of a family of contact 3–manifolds used by Conway in [9, Section 4]. The contact structure ξ is supported by the open book decomposition (S, ϕ), where S is a compact oriented genus-1 surface with two boundary components and $\phi = \tau_a \tau_b \tau_c^{-1}$, the product of positive Dehn twists around the curves a and b and a negative Dehn twist around the curve c indicated in Figure 15, left. This is an open book decomposition for inadmissible transverse 2–surgery on the binding of an open book decomposition $(S_{1,1}, \mathrm{id}_{S_{1,1}})$, where the page $S_{1,1}$ has genus 1 and one boundary component. The contact structure ξ has nonzero Ozsváth–Szabó contact class by [19, Corollary 4], as indicated by Conway.

Theorem 5.1 $o(Y, \xi) = 0$. Hence, (Y, ξ) is not Stein-fillable.

Proof To show that $o(Y, \xi) = 0$, we need to find a multibasis $\boldsymbol{a}$ on S such that $o(S, \phi, \boldsymbol{a}) = 0$; more explicitly, we will find a generator $\vec{y}$ of the resulting Heegaard Floer chain complex such that $\partial_0 \vec{y} = \vec{x}_\xi$. As we will show, it suffices to work with the basis of arcs $\{a_1, a_2, a_3\}$ depicted in Figure 15, right. The effect of the monodromy on this basis of arcs is shown in Figure 16. In what follows, a region without a basepoint will be denoted by R_i if it is numbered i in Figure 16.
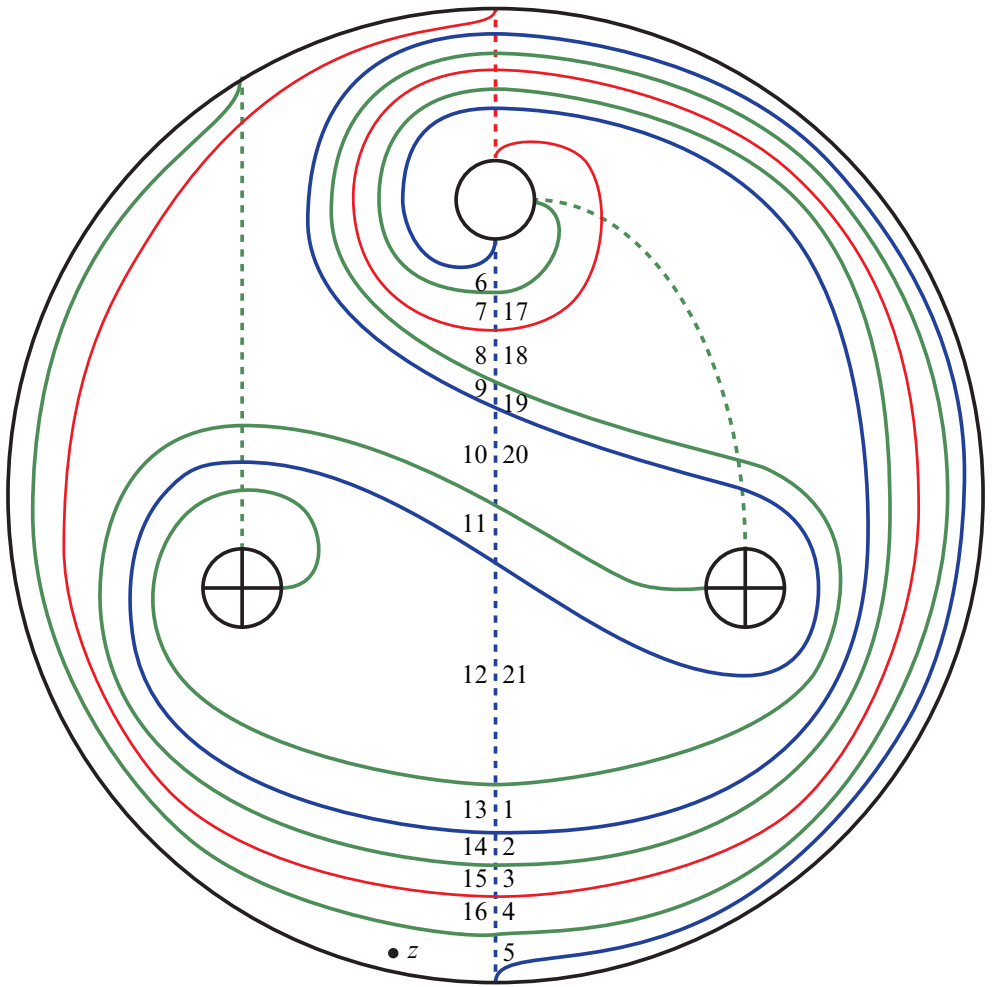
Figure 16: The effect of the monodromy applied to the basis of arcs. The resulting regions without basepoint are numbered $1, \ldots, 21$.

We claim that the generator $\vec{y}$ determined by the tuple of intersection points $y = (x_1, y_2, y_3)$ satisfies $\partial_0 \vec{y} = \vec{x}_\xi$ (see Figure 17). To show this, we need to know in general what kind of Maslov index-1 domains have $J_+ = 0$.

Lemma 5.2 *Let D be a domain from y to x. If $\mathcal{D}$ has Maslov index 1 and $J_+(\mathcal{D}) = 0$, then it is an immersed $2k$–gon with only acute corners and no corners in its interior. Moreover, if $\mathcal{D}$ is any immersed $2k$–gon with only acute corners and no corners in its interior, then it has Maslov index 1 and, furthermore, $J_+(\mathcal{D}) = 0$ if and only if $|y| - |x| = 1 - k$.*
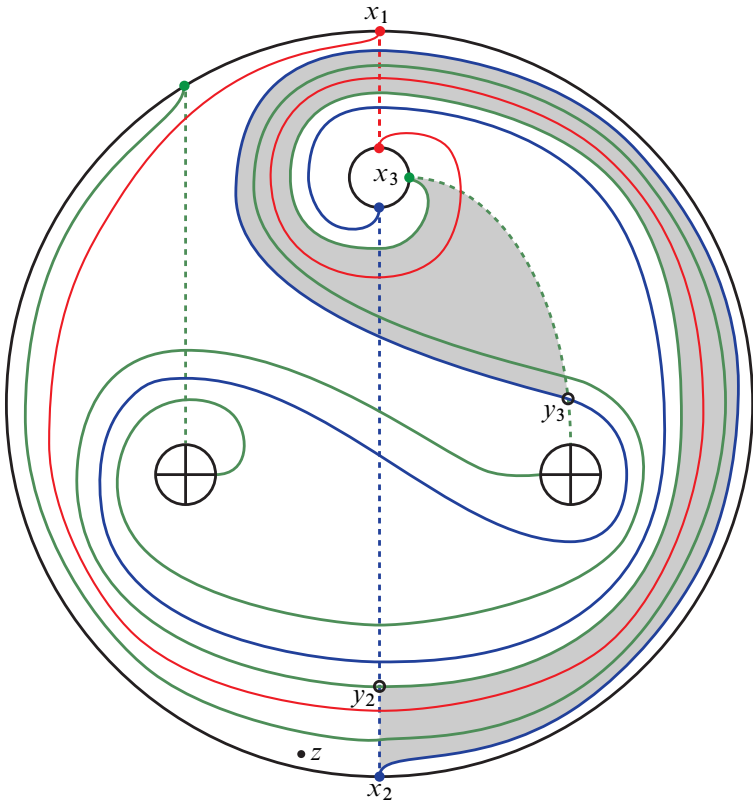
Figure 17: The domain $\mathcal{D}_0$ (shaded).

Proof Let $\mathcal{D}$ be a Maslov index-1 domain with $J_+ = 0$, and suppose that $\boldsymbol{y}$ and $\boldsymbol{x}$ differ on k $\boldsymbol{\alpha}$–curves; hence, $\mathcal{D}$ has $2k$ corners. By (2-6), we have

$$0 = J_+(\mathcal{D}) = 2(n_{\boldsymbol{y}}(\mathcal{D}) + n_{\boldsymbol{x}}(\mathcal{D})) - 1 + |\boldsymbol{y}| - |\boldsymbol{x}| \geq 2 \cdot \tfrac{2}{4}k - 1 + |\boldsymbol{y}| - |\boldsymbol{x}| = k - 1 + |\boldsymbol{y}| - |\boldsymbol{x}|.$$

In other words, $|\boldsymbol{x}| \geq |\boldsymbol{y}| + k - 1$. Conversely, let $\sigma_{\boldsymbol{y}}$ and $\sigma_{\boldsymbol{x}}$ denote the permutations associated to $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively, and denote by σ the composition $\sigma_{\boldsymbol{x}} \sigma_{\boldsymbol{y}}{}^{-1}$. Since $\boldsymbol{y}$ and $\boldsymbol{x}$ differ on k $\boldsymbol{\alpha}$–curves, the smallest number of transpositions that σ can be written as a composition of is bounded from above by $k - 1$, which is realized if and only if σ is a k–cycle. Next write σ as the composition of disjoint cycles. Note that composing a permutation with a transposition either merges two disjoint cycles, which reduces the number of disjoint cycles by 1, or breaks up a cycle into two disjoint cycles, which increases the number of disjoint cycles by 1. Therefore, $\sigma_{\boldsymbol{x}} = \sigma \sigma_{\boldsymbol{y}}$ can have at most $k - 1$ more disjoint cycles than $\sigma_{\boldsymbol{y}}$ has:

$$|\boldsymbol{x}| \leq |\boldsymbol{y}| + k - 1.$$

As a result, $|\boldsymbol{x}| = |\boldsymbol{y}| + k - 1$ and, in particular, σ is a k–cycle. We deduce from (2-6) that

$$n_{\boldsymbol{y}}(\mathcal{D}) + n_{\boldsymbol{x}}(\mathcal{D}) = \tfrac{1}{2}k,$$

implying that $\mathcal{D}$ has point measure $\tfrac{1}{4}$ at each corner and that $\mathcal{D}$ has connected boundary since σ is a k–cycle. Finally, by (2-3), we have $e(\mathcal{D}) = 1 - \tfrac{1}{2}k$, which is the Euler measure of a $2k$–gon with only acute corners, none of which is in the interior of $\mathcal{D}$.

For the second claim, note that, if $\mathcal{D}$ is a $2k$–gon from $\boldsymbol{y}$ to $\boldsymbol{x}$ with each corner having point measure $\tfrac{1}{4}$, then it has Euler measure $1 - \tfrac{1}{2}k$ and Maslov index 1 by (2-3). Therefore,

$$J_+(\mathcal{D}) = k - 1 + 1 - k = 0. \qquad \square$$

With regard to the second part of Lemma 5.2, note that, if $\boldsymbol{x} = \boldsymbol{x}_\xi$ and $\mathcal{D}$ is a $2k$–gon from $\boldsymbol{y}$ to $\boldsymbol{x}_\xi$, then $|\boldsymbol{y}| - |\boldsymbol{x}_\xi| = 1 - k$.

A positive Maslov index-1 $J_+ = 0$ domain $\mathcal{D}_0$ from $\boldsymbol{y}$ to $\boldsymbol{x}_\xi$ is shaded in Figure 17. As a formal sum of regions without basepoints in the Heegaard diagram, it is given by

$$\mathcal{D}_0 = R_3 + R_4 + R_5 + R_7 + R_8 + R_9 + R_{17} + R_{18} + R_{19}.$$

This domain is an embedded rectangle. Therefore, it has a unique holomorphic representative for a generic split almost complex structure by the Riemann mapping theorem. In fact, this is the only domain that represents a positive class in $\hat{\pi}_2(\vec{\boldsymbol{y}}, \vec{\boldsymbol{x}}_\xi)$. This is because any other domain from $\boldsymbol{y}$ to $\boldsymbol{x}_\xi$ has to differ from $\mathcal{D}_0$ by a *periodic domain* representing a *periodic class* in $\hat{\pi}_2(\vec{\boldsymbol{y}}, \vec{\boldsymbol{y}})$. The latter is isomorphic to $H_2(Y; \mathbb{Z})$, which is a free abelian group of rank 2. A basis for $\hat{\pi}_2(\vec{\boldsymbol{y}}, \vec{\boldsymbol{y}})$ is given by the periodic domains

$$P_1 = R_1 + R_4 + R_5 - R_6 - R_7 - R_{10} - R_{11} - R_{14} - R_{15} + R_{18} + R_{19} + R_{21},$$
$$P_2 = R_2 - R_5 + R_6 - R_9 + R_{11} + R_{13} + 2R_{14} + R_{15} + R_{16} - R_{17} - R_{18} - 2R_{19}$$
$$- R_{20} - R_{21}.$$

If $\mathcal{D}_0 + aP_1 + bP_2$ is a positive domain from $\boldsymbol{y}$ to $\boldsymbol{x}_\xi$, then, in particular,

$$0 + a \geq 0, \quad 0 + b \geq 0, \quad 0 - a \geq 0, \quad 0 - b \geq 0$$

via the multiplicities of the regions R_1, R_2, R_{10} and R_{20}, respectively. As a result, $a = 0 = b$.

Next we argue that there are no other positive Maslov index-1 $J_+ = 0$ domains from $\boldsymbol{y}$. To see this, let $\mathcal{D}$ be such a domain from $\boldsymbol{y}$ to some $\boldsymbol{v}$ defining a generator of the

Heegaard Floer chain complex, and move along the boundary of $\mathcal{D}$ in its boundary orientation. Note firstly that, due to the placement of the basepoint, $\mathcal{D}$ cannot have a corner at x_1. Therefore, $\mathcal{D}$ must be an immersed rectangle as none of the regions are bigons. As a result, $\boldsymbol{v} = (x_1, v_2, v_3)$ for some $v_i \in \alpha_i \cap \beta_i$ for $i = 2, 3$. Note further that the region R_{12} adjacent to y_3 is an immersed 8–gon. Being a positive Maslov index-1 $J_+ = 0$ domain with four corners, $\mathcal{D}$ must have Euler measure $e(\mathcal{D}) = 0$. As Euler measure is additive under unions and $\mathcal{D}$ is a positive domain, $\mathcal{D}$ cannot contain the region R_{12}, which has Euler measure -1. Hence, $\mathcal{D}$ contains only the region R_{19} among the four regions adjacent to y_3. Now $v_3 \neq x_3$, since otherwise $\mathcal{D} = \mathcal{D}_0$. Moreover, as $\mathcal{D}$ does not contain the region R_0 with basepoint, it contains R_{19} with multiplicity 1 and does not contain the region R_{18}. This forces $\mathcal{D}$ to be contained in the formal sum

$$R_5 + R_9 + R_{19},$$

as $\mathcal{D}$ cannot contain the 6–gon regions R_1 and R_{10}, which have Euler measure $-\frac{1}{2}$. But then, $\mathcal{D}$ cannot have a corner at y_2, which is a contradiction.

Consequently, we have $\partial_0 \vec{y} = \vec{x}_\xi$ which implies that $\boldsymbol{o}(Y, \xi) = o(S, \phi, \boldsymbol{a}) = 0$ as the spectral order is defined to be the minimum over all choices of open book decompositions (S, ϕ) supporting ξ and multibases $\boldsymbol{a}$ on S. Consequently, by the second bullet point of Theorem 1.1, (Y, ξ) is not Stein-fillable. $\square$

Remark   In fact, $\hat{\partial}_{HF} \vec{y} = \partial_0 \vec{y} + \partial_1 \vec{y} = \vec{x}_\xi + \vec{w}$ where $\vec{w}$ is determined by the tuple of intersection points $\boldsymbol{w} = (x_1, w_2, w_3)$ (see Figure 18). The domain $\mathcal{D}_1$ from $\boldsymbol{y}$ to $\boldsymbol{w}$ shown in Figure 18 is an embedded genus-1 surface with one boundary component and $J_+(\mathcal{D}_1) = 2$ given by the formal sum

$$\mathcal{D}_1 = R_{11} + R_{12} + R_{13} + R_{14}.$$

Arguing similarly to before, we see that, if $\mathcal{D}_1 + aP_1 + bP_2$ is another positive domain from $\boldsymbol{y}$ to $\boldsymbol{w}$, then, in particular,

$$0 + a \geq 0, \quad 0 + b \geq 0, \quad 0 - a \geq 0, \quad 0 - b \geq 0$$

via the multiplicities of the regions R_1, R_2, R_7 and R_9, respectively. As a result, $a = 0 = b$. Furthermore, a slightly more general version of the argument above proves that there are no other positive Maslov index-1 domains from $\boldsymbol{y}$. In particular, the domain $\mathcal{D}_1$ has a unique (up to a signed count) holomorphic representative, since otherwise $\hat{\partial}_{HF} \vec{y} = \vec{x}_\xi$, contradicting the nonvanishing of the Ozsváth–Szabó contact class.
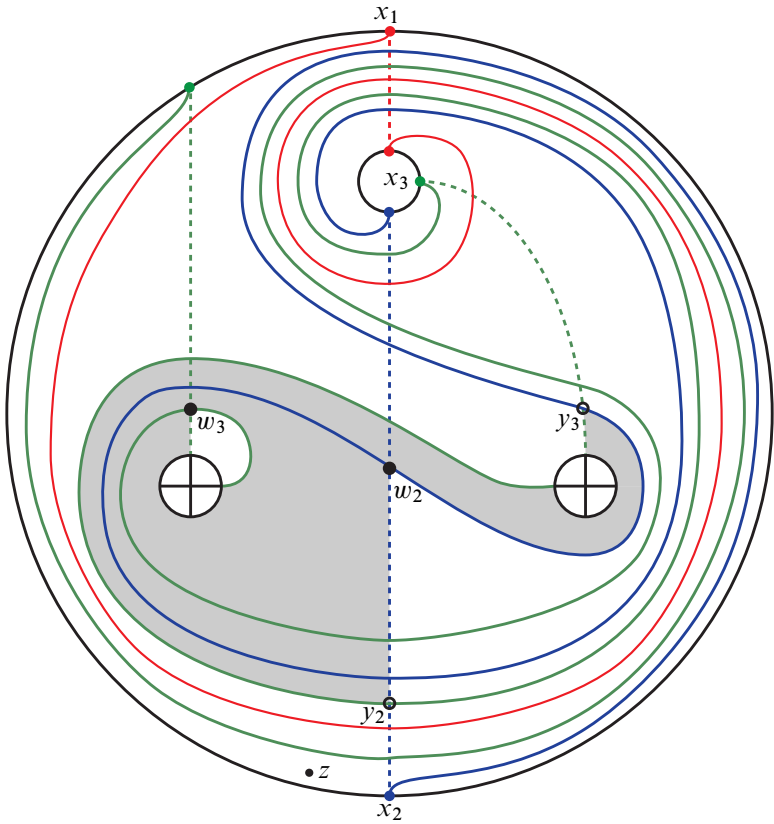
Figure 18: The domain $\mathcal{D}_1$ (shaded). Keep in mind that the middle two circles are identified.

5.2 A family of examples

In this section, we investigate an infinite family of contact 3–manifolds $\{(Y_p, \xi_p)\}_{p \in \mathbb{Z}_{>0}}$. For each $p \in \mathbb{Z}_{\geq 0}$, the contact structure ξ_p is supported by the open book decomposition $(S_{2,2}, \phi_p)$, where $S_{2,2}$ is a compact oriented genus-2 surface with two boundary components and $\phi_p = \tau_a^3 \tau_b \tau_c^{-1} \tau_d^p$, the product of positive Dehn twists around the curves a and b, a negative Dehn twist around the curve c, and p positive Dehn twists around the curve d indicated in Figure 19, left. The manifolds Y_p are obtained by $-1/p$–surgery on a horizontal curve in the circle bundle, Y_0, with Euler number $+4$ over a closed oriented surface of genus 2. Therefore, these manifolds are toroidal and have nontrivial JSJ decompositions. In the case $p = 0$, Honda [20] gave a complete a classification of tight contact structures (see also Giroux [16]). The contact structure ξ_0 is the unique virtually overtwisted contact structure on Y_0, and its nonfillability was established by Lisca and
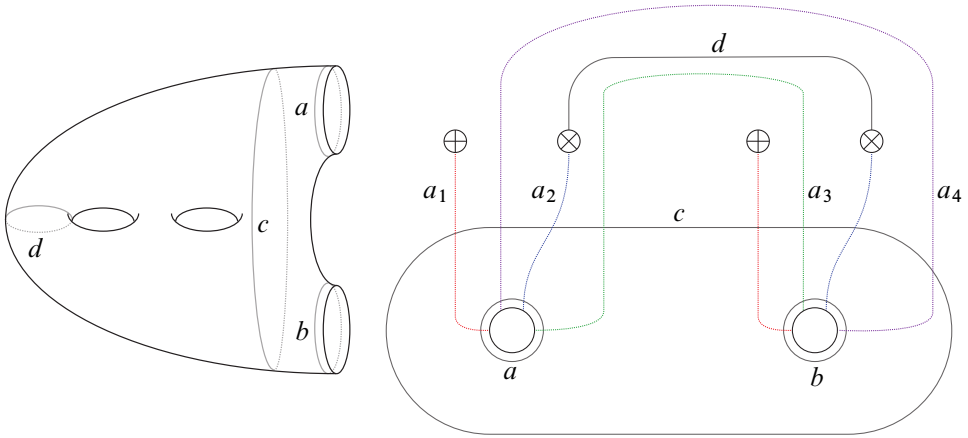
Figure 19: Left: a page of the open book decomposition $(S_{2,2}, \phi_p)$ supporting the contact structure ξ_p, where $\phi_p = \tau_a^3 \tau_b \tau_c^{-1} \tau_d{}^p$. Right: the same surface, where the two circles decorated with "plus" are identified, as are the two circles decorated with "cross".

Stipsicz [39] (see Section 5.3 below for a detailed discussion). The contact structure ξ_p for $p \geq 0$ can be constructed by first applying inadmissible transverse surgery (with framing $+4$) on the genus-2 Borromean knot K in $L(p, p-1) \# \#_3 S^1 \times S^2$, which is the binding of an open book decomposition that supports the unique tight contact structure on this manifold, then resolving the resulting rational open book decomposition into an integral one following Conway [9]. As with the genus-1 example in Section 5.1, the contact structures ξ_p have nonzero Ozsváth–Szabó contact class by [19, Corollary 4].

Theorem 5.3 $o(Y_p, \xi_p) = 0$ for $p \geq 1$. Hence, (Y_p, ξ_p) is not Stein-fillable for $p \geq 1$.

To put the above theorem in context, our examples fit somewhere in between the circle bundle example of Lisca and Stipsicz and positive-integer surgeries on the $(2, 5)$–torus knot. In the former case, the monodromy is trivial away from the pair of pants at the boundary. In the latter, the monodromy has four positive Dehn twists: those that fit along the standard length-four chain. These two examples — Lisca and Stipsicz's and positive-integer surgeries on the $(2, 5)$–torus knot — behave differently as one increases the surgery coefficient. Increasing the surgery coefficient on K by 1 corresponds to adding a single positive Dehn twist along the curve a to the monodromy. The Lisca–Stipsicz examples remain nonfillable for all higher-integer surgeries, whereas $+9$–surgery on the $(2, 5)$–torus knot yields a tight contact structure on a lens space; hence, it is Stein-fillable. All higher-integer surgeries on the $(2, 5)$–torus knot then
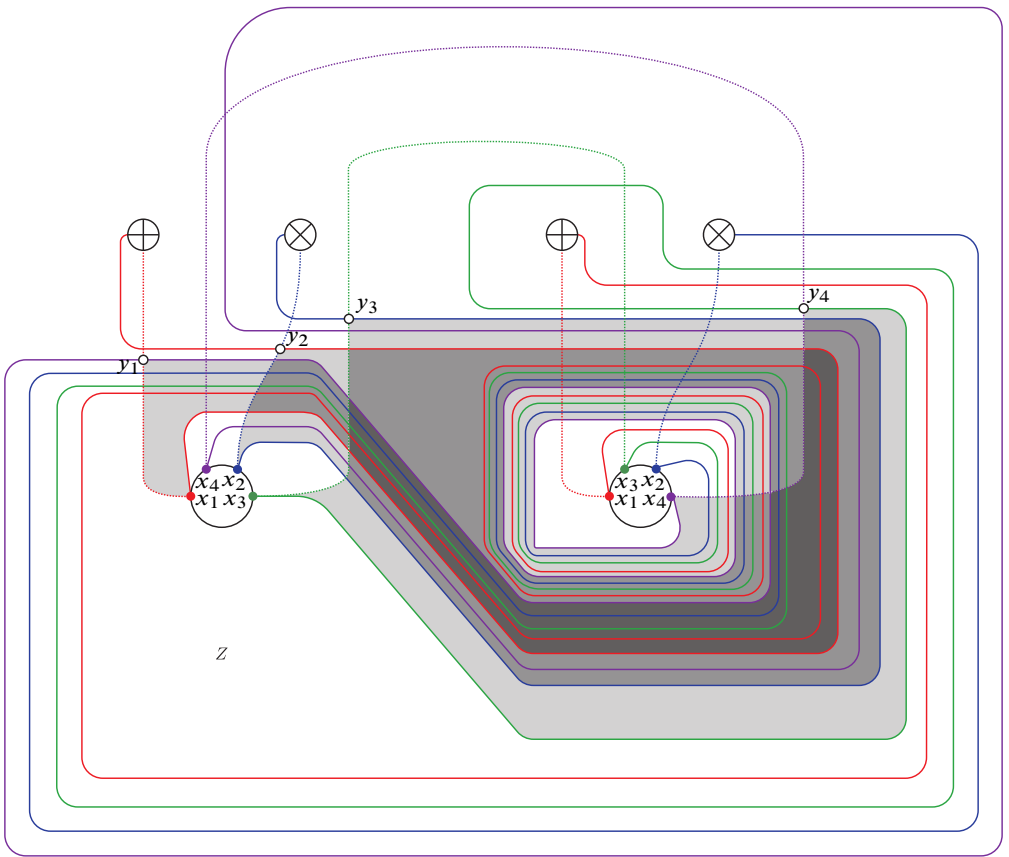
Figure 20: The effect of the monodromy applied to the arcs a_1, a_2, a_3 and a_4 (dotted) on $S_{2,2}$. Also shown is the domain $\mathcal{D}$ (shaded) with darker shading indicating 1 higher multiplicity.

remain Stein-fillable. Our initial calculations for $+4$–surgery on the $(2, 5)$–torus knot suggest that in this case $\boldsymbol{o} = \infty$. For the examples ξ_p when $p > 0$, we expect $\boldsymbol{o}$ to remain finite (though possibly nonzero) for all integer surgeries higher than $+4$, and therefore that all resulting contact structures remain non-Stein-fillable. It would be interesting to know whether these contact structures are weakly or strongly fillable.

Proof of Theorem 5.3   A basis of arcs on $S_{2,2}$ consists of five pairwise disjoint properly embedded arcs. In what follows, we work with a collection of four pairwise disjoint properly embedded arcs $\boldsymbol{a} = \{a_1, a_2, a_3, a_4\}$ to show that $\boldsymbol{o}(Y_p, \xi_p) = 0$. Adding extra arcs would not change the result in light of Lemma 4.1. These arcs are shown in Figure 19, right, while their respective images under the monodromy are shown
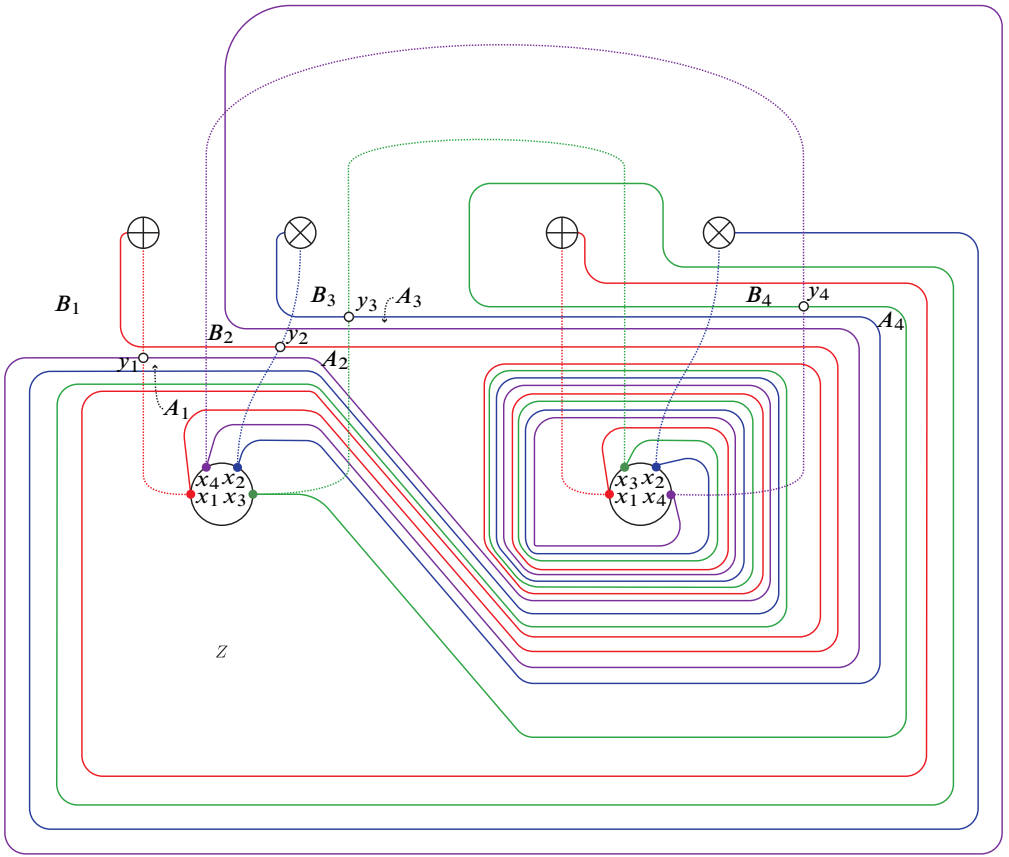
Figure 21: The regions A_i and B_i for $i = 1, \ldots, 4$ and the region Z with basepoint.

in Figure 20. Also shown in the latter figure is a positive Maslov index-1, $J_+ = 0$ domain $\mathcal{D}$ from $\boldsymbol{y} = (y_1, y_2, y_3, y_4)$ to $\boldsymbol{x}_{\xi_p} = (x_1, x_2, x_3, x_4)$, which is an immersed octagon and therefore has a unique holomorphic representative for a generic split almost complex structure. Our goal is to show that $\mathcal{D}$ is the only positive Maslov index-1, $J_+ = 0$ domain from $\boldsymbol{y}$.

Suppose that $\mathcal{D}'$ is a positive Maslov index-1, $J_+ = 0$ domain from $\boldsymbol{y}$ to some $\boldsymbol{w} = (w_1, w_2, w_3, w_4)$. We will show that $\mathcal{D}' = \mathcal{D}$. To begin, for each y_i, label the region with corner at y_i having nonzero coefficient in $\mathcal{D}$ as A_i, and let B_i denote the region whose intersection with A_i in a neighborhood of y_i consists only of the point y_i (see Figure 21). We label one of the regions with basepoint as Z and denote the multiplicity of a region R in $\mathcal{D}'$ by $|R|$. Since $J_+(\mathcal{D}') = 0$, $\mathcal{D}'$ has only acute corners and a connected boundary by Lemma 5.2; hence, if y_i is a corner of $\mathcal{D}'$, then $\{|A_i|, |B_i|\} = \{0, 1\}$. Note also

that, as B_3 is an annulus with eight acute corners, it has Euler measure -2. Therefore, $|B_3| = 0$, since $e(\mathcal{D}') \geq -1$ and Euler measure is additive.

Next suppose that y_4 is a corner of $\mathcal{D}'$ and that $|A_4| = 1$. Following β_3 along the boundary of $\mathcal{D}'$ starting from y_4, we deduce that, in order to avoid Z, we must have $w_3 = x_3$, which forces y_3 to be a corner of $\mathcal{D}'$ and $|A_3| = 1$. Similarly, for $i = 2$ and $i = 3$, if y_i is a corner of $\mathcal{D}'$ and $|A_i| = 1$, we may follow along β_{i-1} to conclude that, so as to avoid Z, we must have $w_{i-1} = x_{i-1}$; hence, y_{i-1} is a corner of $\mathcal{D}'$ and $|A_{i-1}| = 1$. Finally, if y_1 is a corner of $\mathcal{D}'$ and $|A_1| = 1$, following β_4 along the boundary of $\mathcal{D}'$, we conclude that, so as to avoid B_3, we must have $w_4 \in \beta_4 \cap \alpha_4$; therefore, y_4 should be a corner of $\mathcal{D}'$ and $|A_4| = 1$. In conclusion, if $|A_i| \neq 0$ for any $i = 1, \ldots, 4$, then $|A_i| \neq 0$ and $w_i = x_i$ for all $i = 1, \ldots, 4$. Checking the coefficients this forces in the remaining regions, we conclude that the only such domain is $\mathcal{D}$ itself.

It remains to consider the case that $|A_i| = 0$ for all $i = 1, \ldots, 4$. As noted above, $|B_3| = 0$, from which we conclude that y_3 is not a corner of $\mathcal{D}'$; hence, $\partial \mathcal{D}'$ contains no segment of α_3 or β_2. Supposing then that y_2 is a corner of $\mathcal{D}'$ with $|B_2| = 1$, we may follow α_2 along the boundary of $\mathcal{D}'$ to see that, in order to avoid B_3, w_2 should be the corner of B_2 along β_4, forcing y_1 to be a corner of $\mathcal{D}'$. Similarly, if y_1 is a corner of $\mathcal{D}'$ with $|B_1| = 1$, then, following α_1 along the boundary of $\mathcal{D}'$, we deduce that there is a unique candidate for w_1 (ie a unique intersection point at which turning left leads to a y_i without creating a self-intersection in $\partial \mathcal{D}'$), which is along β_3, forcing a corner at y_4. Finally, if y_4 is a corner of $\mathcal{D}'$ with $|B_4| = 1$, following α_4 along the boundary of $\mathcal{D}'$, we conclude that there are three possibilities for w_4 to avoid B_3. Of these, one is along β_2, which would force y_3 to be a corner of $\mathcal{D}'$, and another is along β_3, which would lead us back to y_4 along a homotopically nontrivial path, a contradiction as $\mathcal{D}'$ can have only one boundary component. We conclude that w_4 should be the corner of B_4 along β_1; hence, y_2 is a corner of $\mathcal{D}'$. It follows that $\partial \mathcal{D}'$ has a single self-intersection, at the corner of B_1 in $\alpha_1 \cap \beta_1$, giving a contradiction. $\square$

5.3 Comparison with other known obstructions

Our goal in this section is to put our calculations of o for the contact 3–manifolds $\{(Y_p, \xi_p)\}_{p \in \mathbb{Z}_{>0}}$ into the broader context of obstructions to existence of weak, strong, exact and Stein fillings. Note that:

(1) The contact structure ξ_p results from inadmissible transverse surgery with framing $+4$ on the genus-2 Borromean knot in $L(p, p-1) \# \#_3 S^1 \times S^2$ [9].

(2) $\hat{c}(\xi_p) \neq 0$ by [19], since the surgery coefficient is $2g = 4$.

(3) Capping the boundary along the curve a gives a weak symplectic 2–handle cobordism from Y_p to $L(p, p-1) \# \#_3 S^1 \times S^2$ [14].

(4) $c_1(\xi_p)$ is torsion.

(5) $d_3(\xi_p) = \frac{1}{4}(p-3)$ [39].

To see (4), note that $b_2(L(p, p-1) \# \#_3 S^1 \times S^2) = 3$ (or $= 4$ if $p = 0$) and that every homology class can be represented by embedded tori. As a result, the Benneqiun–Eliashberg bound implies that $c_1(\xi_p)$ evaluates trivially on $H_2(L(p, p-1) \# \#_3 S^1 \times S^2)$; therefore, it must be torsion.

Coarsely, there are two methods to obstruct symplectic-fillability: via the vanishing of contact invariants from Floer homology and gauge theory — such as monopole Floer homology, Heegaard Floer homology and embedded contact homology — via structural algebraic properties of symplectic field theory (SFT) or contact homology, or by applying more context specific ad hoc methods. The vanishing of contact invariants in Floer homology obstructs strong-fillability and can be used to obstruct weak-fillability with a suitable coefficient system. For our examples, because $\hat{c}(\xi_p) \neq 0$, any obstruction to symplectic-fillability would fall into the ad hoc category. It is possible that weak/strong-fillability of these contact 3–manifolds could be obstructed by SFT, for example were the algebraic torsion to be nonzero [33]. It is also possible that one could obstruct strong-fillability using contact homology [51; 5; 25], again assuming one could both calculate it and show that there are no augmentations of the algebra. Neither of these methods seems particularly practical for these examples, but we don't know.

In situations where contact invariants fail to obstruct symplectic-fillability or they are too difficult to calculate, other information can sometimes be utilized. Interesting families of contact 3–manifolds have been shown to be nonfillable by symplectic caps or other cobordisms. Prior to the introduction of contact invariants from Floer homology, all methods of obstructing fillability were ad hoc and relied on Gromov's theory of pseudoholomorphic curves (eg [11; 12]), but they only apply to obstructing existence of strong fillings. The introduction of the contact invariant in Seiberg–Witten theory [29] provided a more universal tool to obstruct symplectic-fillability, but it was notoriously difficult to calculate. In [39], Lisca and Stipsicz studied a family of contact 3–manifolds $\{(Y_{g,n}, \xi_i)\}_{n \geq 2g \geq 0, i=0,1}$ described by Honda [20] and Giroux [16] to show that they are not symplectically fillable. Rather than directly showing that the monopole contact

invariant is zero for this family, Lisca and Stipsicz first calculated the d_3–invariants of these contact structures using descriptions of $(Y_{g,n}, \xi_i)$ as Legendrian surgeries on some Stein-fillable contact 3–manifolds. Building on earlier work of Lisca [36] and Mrowka, Ozsváth and Yu [43], they then conclude that in a given Spin^c structure, there is a unique homotopy class of 2–plane fields ξ containing a symplectically fillable contact structure. Finally, they use calculations of an η–invariant by Nicolaescu [44] to calculate the d_3–invariant of ξ and see that it does not agree with the d_3–invariants of ξ_i.

Often filling obstructions follow by finding a symplectic cobordism to either the empty set or some target contact 3–manifold (M, ξ) whose symplectic fillings are classified — such as certain contact structures on the lens spaces $L(p, q)$ (eg [42; 37]) — or are obstructed entirely [14]. One then attempts to obstruct this cobordism from embedding in any filling of (M, ξ). These methods build on the foundational examples and methods of Lisca in [36] (see [40; 28; 34]). There, strong-fillability is obstructed by finding a smooth 4–manifold cap which cannot be embedded into a diagonal lattice, noticing that the projection $c_{\mathrm{red}}(\xi)$ of $c^+(\xi)$ onto the reduced Heegaard Floer homology is zero, so that all strong fillings must be negative-definite, and then invoking Donaldson's theorem to obstruct the existence of resulting closed smooth manifold. One can also invoke a relative version of this obstruction by Owens and Strle [47], using the Heegaard Floer d–invariants of M. This method also often obstructs existence of weak fillings as well, as at least some of the manifolds involved are rational homology spheres, where the two notions of weak and strong filling are equivalent. Similar obstructions are possible by finding symplectic caps which contain symplectic spheres of nonnegative square and then analyzing the resulting embedding into a ruled surface (eg [41]).

There are other methods of obstructing existence of even weak fillings in situations where property (2) and some version of properties (1) [10], (4) [28; 34] and (3) [40; 41] above hold. In [10], Conway, Etnyre and Tosun study a particular case. They investigate contact 3–manifolds Y_K obtained by inadmissible contact surgery on a transverse knot K in S^3 and obstruct existence of weak fillings in a very interesting range of surgery slopes determined by $\tau(K)$. Their obstruction is obtained by the relative adjunction inequality of Raoux [53] for knots in rational homology spheres, noting that any weak filling of Y_K embeds into a strong filling of S^3 in which K bounds a symplectic disk. There are generalizations of this method that work in the exact setting where Y_K is built by inadmissible surgery on a transverse knot K in a 3–manifold Y all of whose weak fillings are classified. We note that neither the Conway–Etnyre–Tosun nor Lisca–Stipsicz methods appear to be applicable to our family of contact

3–manifolds. The sphere of square $-p$ easily embeds into the diagonal lattice and there does not appear to be an obstruction to K bounding a symplectic disk of the appropriate type in a symplectic filling of $L(p, p-1)$. We also note that, if our family of contact 3–manifolds were supported by planar open book decompositions, one could conceivably invoke Niederkrüger and Wendl [45; 57], as used by [27], to obstruct weak and strong fillings.

The obstructions particular to Stein-fillability generally require both that $c_{\text{red}}(\xi) = 0$ and that $d_3(\xi)$ be small (either less than 1 [28] or less than 0 [2]). To date, there is exactly one method to obstruct Stein-fillability of an exactly fillable contact 3–manifold [6]. This uses Eliashberg's theorem on decomposing spheres [11] and requires the 3–manifold in question to be reducible. This method is entirely dependent on Ghiggini's obstruction to existence of exact fillings of certain strongly fillable contact 3–manifolds. In [15], Ghiggini used properties of Stein fillings [38] and the behavior of Heegaard Floer homology under Spin^c conjugation to obstruct Stein-fillability on a number of Brieskorn spheres. Ghiggini's method requires, among much else, that ξ be homotopic to its coorientation reversal, $\overline{\xi}$, which implies that $c_1(\xi) = 0$.

We note that the four simplifying properties (1)–(4) hold only because we have chosen a particularly simple family of contact 3–manifolds. One can tweak this family to construct examples where $\hat{c}(\xi) \neq 0$ but $o(\xi) = 0$ and where none of the properties (1)–(4) hold. In general, we expect that there are examples of contact 3–manifolds where both $c^+(\xi) \neq 0$ and $c_{\text{red}}(\xi) \neq 0$, but $o(\xi) < \infty$, and which have no reasonable cobordism to a contact 3–manifold whose fillings are classified or whose symplectic caps are constrained. For such contact 3–manifolds, the spectral order obstructs Stein-fillability but it is likely that no other current method could be applied to show this.

Finally, one major practical advantage of working with Heegaard Floer homology is its computability. Finding an upper bound for o is a direct calculation that can be done easily on any fixed open book decomposition. We carried out this task using a computer program that we wrote, building on a program of Sucharit Sarkar that analyzes Heegaard Floer chain complexes. The proofs given in Sections 5.1 and 5.2 were done by hand and verified by this computer program, which also gives us the capability to do calculations on much larger chain complexes. In conclusion, if o is finite, finding an upper bound for the explicit value is a relatively simple endeavor even if calculating the exact value is difficult. Hence, as a fillability obstruction when $\hat{c}(\xi) \neq 0$, o is both a robust contact invariant with its fundamental properties, and it is computable.

References

[1] **K L Baker**, **J B Etnyre**, **J Van Horn-Morris**, *Cabling, contact structures and mapping class monoids*, J. Differential Geom. 90 (2012) 1–80 MR Zbl

[2] **J A Baldwin**, *Heegaard Floer homology and genus one, one-boundary component open books*, J. Topol. 1 (2008) 963–992 MR Zbl

[3] **J A Baldwin**, *Contact monoids and Stein cobordisms*, Math. Res. Lett. 19 (2012) 31–40 MR Zbl

[4] **J Baldwin**, **D S Vela-Vick**, *A note on the knot Floer homology of fibered knots*, Algebr. Geom. Topol. 18 (2018) 3669–3690 MR Zbl

[5] **E Bao**, **K Honda**, *Definition of cylindrical contact homology in dimension three*, J. Topol. 11 (2018) 1002–1053 MR Zbl

[6] **J Bowden**, *Exactly fillable contact structures without Stein fillings*, Algebr. Geom. Topol. 12 (2012) 1803–1810 MR Zbl

[7] **K Cieliebak**, **Y Eliashberg**, *From Stein to Weinstein and back: symplectic geometry of affine complex manifolds*, American Mathematical Society Colloquium Publications 59, Amer. Math. Soc., Providence, RI (2012) MR Zbl

[8] **V Colin**, *Chirurgies d'indice un et isotopies de sphères dans les variétés de contact tendues*, C. R. Acad. Sci. Paris Sér. I Math. 324 (1997) 659–663 MR Zbl

[9] **J Conway**, *Transverse surgery on knots in contact 3–manifolds*, Trans. Amer. Math. Soc. 372 (2019) 1671–1707 MR Zbl

[10] **J Conway**, **J B Etnyre**, **B Tosun**, *Symplectic fillings, contact surgeries, and Lagrangian disks*, Int. Math. Res. Not. 2021 (2021) 6020–6050 MR Zbl

[11] **Y Eliashberg**, *Filling by holomorphic discs and its applications*, from "Geometry of low-dimensional manifolds, II" (S K Donaldson, C B Thomas, editors), London Math. Soc. Lecture Note Ser. 151, Cambridge Univ. Press (1990) 45–67 MR Zbl

[12] **Y Eliashberg**, *Unique holomorphically fillable contact structure on the 3–torus*, Int. Math. Res. Not. 1996 (1996) 77–82 MR Zbl

[13] **J B Etnyre**, **J Van Horn-Morris**, *Monoids in the mapping class group*, from "Interactions between low-dimensional topology and mapping class groups" (R I Baykur, J Etnyre, U Hamenstädt, editors), Geom. Topol. Monogr. 19, Geom. Topol. Publ., Coventry (2015) 319–365 MR Zbl

[14] **D T Gay**, *Four-dimensional symplectic cobordisms containing three-handles*, Geom. Topol. 10 (2006) 1749–1759 MR Zbl

[15] **P Ghiggini**, *Strongly fillable contact 3–manifolds without Stein fillings*, Geom. Topol. 9 (2005) 1677–1687 MR Zbl

[16] **E Giroux**, *Structures de contact en dimension trois et bifurcations des feuilletages de surfaces*, Invent. Math. 141 (2000) 615–689 MR Zbl

[17] **E Giroux**, *Géométrie de contact: de la dimension trois vers les dimensions supérieures*, from "Proceedings of the International Congress of Mathematicians, II" (T Li, editor), Higher Ed., Beijing (2002) 405–414   MR   Zbl

[18] **R E Gompf**, *Handlebody construction of Stein surfaces*, Ann. of Math. 148 (1998) 619–693   MR   Zbl

[19] **M Hedden**, **O Plamenevskaya**, *Dehn surgery, rational open books and knot Floer homology*, Algebr. Geom. Topol. 13 (2013) 1815–1856   MR   Zbl

[20] **K Honda**, *On the classification of tight contact structures, II*, J. Differential Geom. 55 (2000) 83–143   MR   Zbl

[21] **K Honda**, **W H Kazez**, **G Matić**, *Right-veering diffeomorphisms of compact surfaces with boundary*, Invent. Math. 169 (2007) 427–449   MR   Zbl

[22] **K Honda**, **W H Kazez**, **G Matić**, *The contact invariant in sutured Floer homology*, Invent. Math. 176 (2009) 637–676   MR   Zbl

[23] **K Honda**, **W H Kazez**, **G Matić**, *On the contact class in Heegaard Floer homology*, J. Differential Geom. 83 (2009) 289–311   MR   Zbl

[24] **M Hutchings**, *The embedded contact homology index revisited*, from "New perspectives and challenges in symplectic field theory" (M Abreu, F Lalonde, L Polterovich, editors), CRM Proc. Lecture Notes 49, Amer. Math. Soc., Providence, RI (2009) 263–297   MR   Zbl

[25] **M Hutchings**, **J Nelson**, *Cylindrical contact homology for dynamically convex contact forms in three dimensions*, J. Symplectic Geom. 14 (2016) 983–1012   MR   Zbl

[26] **A Juhász**, **S Kang**, *Spectral order for contact manifolds with convex boundary*, Algebr. Geom. Topol. 18 (2018) 3315–3338   MR   Zbl

[27] **A Kaloti**, *Stein fillings of planar open books*, preprint (2013)   arXiv 1311.0208

[28] **A Kaloti**, **B Tosun**, *Hyperbolic rational homology spheres not admitting fillable contact structures*, Math. Res. Lett. 24 (2017) 1693–1705   MR   Zbl

[29] **P B Kronheimer**, **T S Mrowka**, *Monopoles and contact structures*, Invent. Math. 130 (1997) 209–255   MR   Zbl

[30] **Ç Kutluhan**, **G Matić**, **J V Horn-Morris**, **A Wand**, *Spectral order for sutured contact manifolds*, in preparation

[31] **Ç Kutluhan**, **G Matić**, **J V Horn-Morris**, **A Wand**, *Spectral order is a non-trivial contact invariant*, in preparation

[32] **Ç Kutluhan**, **G Matić**, **J Van Horn-Morris**, **A Wand**, *A Heegaard Floer analog of algebraic torsion*, from "Breadth in contemporary topology" (D T Gay, W Wu, editors), Proc. Sympos. Pure Math. 102, Amer. Math. Soc., Providence, RI (2019) 119–130   MR   Zbl

[33] **J Latschev**, **C Wendl**, *Algebraic torsion in contact manifolds*, Geom. Funct. Anal. 21 (2011) 1144–1195   MR   Zbl   With an appendix by Michael Hutchings

[34] **Y Li**, **Y Liu**, *Hyperbolic 3–manifolds admitting no fillable contact structures*, Proc. Amer. Math. Soc. 147 (2019) 351–360   MR   Zbl

[35] **R Lipshitz**, *A cylindrical reformulation of Heegaard Floer homology*, Geom. Topol. 10 (2006) 955–1096   MR   Zbl   Correction in 18 (2014) 17–30

[36] **P Lisca**, *Symplectic fillings and positive scalar curvature*, Geom. Topol. 2 (1998) 103–116   MR   Zbl

[37] **P Lisca**, *On lens spaces and their symplectic fillings*, Math. Res. Lett. 11 (2004) 13–22   MR   Zbl

[38] **P Lisca**, **G Matić**, *Tight contact structures and Seiberg–Witten invariants*, Invent. Math. 129 (1997) 509–525   MR   Zbl

[39] **P Lisca**, **A I Stipsicz**, *Tight, not semi-fillable contact circle bundles*, Math. Ann. 328 (2004) 285–298   MR   Zbl

[40] **P Lisca**, **A I Stipsicz**, *Ozsváth–Szabó invariants and tight contact 3–manifolds, III*, J. Symplectic Geom. 5 (2007) 357–384   MR   Zbl

[41] **S Lisi**, **C Wendl**, *Spine removal surgery and the geography of symplectic fillings*, Michigan Math. J. 70 (2021) 403–422   MR

[42] **D McDuff**, *The structure of rational and ruled symplectic 4–manifolds*, J. Amer. Math. Soc. 3 (1990) 679–712   MR   Zbl

[43] **T Mrowka**, **P Ozsváth**, **B Yu**, *Seiberg–Witten monopoles on Seifert fibered spaces*, Comm. Anal. Geom. 5 (1997) 685–791   MR   Zbl

[44] **L I Nicolaescu**, *Eta invariants of Dirac operators on circle bundles over Riemann surfaces and virtual dimensions of finite energy Seiberg–Witten moduli spaces*, Israel J. Math. 114 (1999) 61–123   MR   Zbl

[45] **K Niederkrüger**, **C Wendl**, *Weak symplectic fillings and holomorphic curves*, Ann. Sci. Éc. Norm. Supér. 44 (2011) 801–853   MR   Zbl

[46] **Y-G Oh**, *Fredholm theory of holomorphic discs under the perturbation of boundary conditions*, Math. Z. 222 (1996) 505–520   MR   Zbl

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

[48] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. 159 (2004) 1027–1158   MR   Zbl

[49] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and three-manifold invariants: properties and applications*, Ann. of Math. 159 (2004) 1159–1245   MR   Zbl

[50] **P Ozsváth**, **Z Szabó**, *Heegaard Floer homology and contact structures*, Duke Math. J. 129 (2005) 39–61   MR   Zbl

[51] **J Pardon**, *Contact homology and virtual fundamental cycles*, J. Amer. Math. Soc. 32 (2019) 825–919 MR Zbl

[52] **O Plamenevskaya**, *A combinatorial description of the Heegaard Floer contact invariant*, Algebr. Geom. Topol. 7 (2007) 1201–1209 MR Zbl

[53] **K Raoux**, *tau-invariants for knots in rational homology spheres*, PhD thesis, Brandeis University (2017) MR Available at https://www.proquest.com/docview/1873863961

[54] **S Sarkar**, **J Wang**, *An algorithm for computing some Heegaard Floer homologies*, Ann. of Math. 171 (2010) 1213–1236 MR Zbl

[55] **A Wand**, *Detecting tightness via open book decompositions*, from "Interactions between low-dimensional topology and mapping class groups" (R I Baykur, J Etnyre, U Hamenstädt, editors), Geom. Topol. Monogr. 19, Geom. Topol. Publ., Coventry (2015) 291–317 MR Zbl

[56] **A Wand**, *Tightness is preserved by Legendrian surgery*, Ann. of Math. 182 (2015) 723–738 MR Zbl

[57] **C Wendl**, *Strongly fillable contact manifolds and J–holomorphic foliations*, Duke Math. J. 151 (2010) 337–384 MR Zbl

[58] **C Wendl**, *A hierarchy of local symplectic filling obstructions for contact 3–manifolds*, Duke Math. J. 162 (2013) 2197–2283 MR Zbl

Department of Mathematics, University at Buffalo
Buffalo, NY, United States

Department of Mathematics, University of Georgia
Athens, GA, United States

Department of Mathematics, University of Arkansas
Fayetteville, AR, United States

School of Mathematics and Statistics, University of Glasgow
Glasgow, United Kingdom

kutluhan@buffalo.edu, gordanam@uga.edu, jvhm@uark.edu, andy.wand@glasgow.ac.uk

Large-scale geometry of big mapping class groups

KATHRYN MANN

KASRA RAFI

We study the large-scale geometry of mapping class groups of surfaces of infinite type, using the framework of Rosendal for coarse geometry of non-locally-compact groups. We give a complete classification of those surfaces whose mapping class groups have local *coarse boundedness* (the analog of local compactness). When the end space of the surface is countable or *tame*, we also give a classification of those surfaces where there exists a coarsely bounded generating set (the analog of finite or compact generation, giving the group a well-defined quasi-isometry type) and those surfaces with mapping class groups of bounded diameter (the analog of compactness).

We also show several relationships between the topology of a surface and the geometry of its mapping class groups. For instance, we show that *nondisplaceable subsurfaces* are responsible for nontrivial geometry and can be used to produce unbounded length functions on mapping class groups using a version of subsurface projection; while *self-similarity* of the space of ends of a surface is responsible for boundedness of the mapping class group.

57K20, 57M07

1 Introduction

Mapping class groups of surfaces of infinite type (with infinite genus or infinitely many ends) form a rich class of examples of non-locally-compact Polish topological groups. These "big" mapping class groups can be seen as natural generalizations of, or limit objects of, the mapping class groups of finite type surfaces, and also arise naturally in the study of laminations and foliations, and the dynamics of group actions on finite type surfaces.

Several recent papers have studied big mapping class groups through their actions on associated combinatorial structures such as curve or arc complexes; see for instance Aramayona, Fossas and Parlier [1], Bavard, Dowdall and Rafi [4] and Durham, Fanoni and Vlamis [7]. From this perspective, an important problem is to understand whether a given mapping class group admits a *metrically nontrivial* action on such a space, namely, an action with unbounded orbits. It is our observation that this should be framed as part of a larger question, one of the *coarse* or *large-scale geometry* of big mapping class groups. This is the goal of the present work.

However, describing the large-scale structure of big mapping class groups — or even determining whether this notion makes sense — is a nontrivial problem, as standard tools of geometric group theory apply only to locally compact, compactly generated groups, and big mapping class groups do not fall in this category. Instead, we use recent work of Rosendal [18] that extends the framework of geometric group theory to a broader class of topological groups, using the notion of *coarse boundedness*.

Definition 1.1 Let G be a topological group. A subset $A \subset G$ is *coarsely bounded*, abbreviated *CB*, if every compatible left-invariant metric on G gives A finite diameter. A group is *locally CB* if it admits a CB neighborhood of the identity, and *CB generated* if it admits a CB generating set.[1]

To give an example, in a locally compact group, the CB sets are precisely the compact ones. As is well known, among locally compact groups, those who admit a CB (ie compact) generating set have a well-defined quasi-isometry type, namely that given by the word metric with respect to any compact generating set (the discrete, finitely

---

[1]In Rosendal [17] and much earlier work, this condition is called (OB), for *orbites bornées*, as it is equivalent to the condition that for any continuous action of G on a metric space X by isometries, the diameter of every orbit $A \cdot x$ is bounded. *Coarsely bounded* appears in [18]; we prefer this terminology as it is more suggestive of the large-scale geometric context.

generated groups are a special case of this). Extending this notion, one says that a left-invariant metric d on a group G is said to be *maximal* if for any other left-invariant metric d' there exist constants C and K such that

$$d'(f,g) \leq Kd(f,g) + C$$

holds for all $f, g \in G$. If G admits a maximal metric, then the coarse equivalence class of this metric gives G a well-defined quasi-isometry type. Rosendal shows the following.

Theorem 1.2 (Rosendal [18, Theorem 1.2]) *Let G be a Polish group. The following are equivalent:*

  (i) *G is generated by a CB subset.*

 (ii) *G admits a maximal left-invariant metric, among the left-invariant metrics which generate its topology.*

(iii) *G has a CB neighborhood of the identity and cannot be expressed as the union of a countable chain of proper open subgroups.*

Furthermore, the word metric from any CB generating set is in the quasi-isometry class of the maximal metric, giving a concrete description of the geometry of the group [18, Proposition 2.5].

In this work, we show that among the big mapping class groups there is a rich family of examples to which Rosendal's theory applies, and give the first steps towards a classification of such groups up to quasi-isometry.

1.1 Main results

For simplicity, we assume all surfaces are oriented and have empty boundary, and all homeomorphisms are orientation-preserving. (The cases of nonorientable surfaces, and those with finitely many boundary components can be approached using essentially the same tools.)

Summary We give a complete classification of surfaces Σ for which $\mathrm{Map}(\Sigma)$ is *locally CB* (Theorem 1.4). By Theorem 1.2, this is necessary for the group to be generated by a CB subset, but these are not equivalent. Under mild hypotheses, we give a full classification of those surfaces which are *CB generated* and therefore have a well-defined quasi-isometry type (Theorem 1.6), as well as those which are *globally CB*, ie have trivial QI type (Theorem 1.7).

To give the precise statements, we need to recall the classification of surfaces and state two key definitions.

End spaces Recall that topological spaces admit a standard compactification by a *space of ends*. By a theorem of Richards [16] orientable, boundaryless, infinite-type surfaces are completely classified by the following data: the genus (possibly infinite), the *space of ends* E, which is a totally disconnected, separable, metrizable topological space, and the subset of ends E^G that are accumulated by genus, which is a closed subset of E. Every such pair (E, E^G) occurs as the space of ends of some surface, with $E^G = \varnothing$ if and only if the surface has finite genus. We call a pair (E, E^G) *self-similar* if for any decomposition $E = E_1 \sqcup E_2 \sqcup \cdots \sqcup E_n$ of E into pairwise disjoint clopen sets, there exists a clopen set D contained in some E_i such that the pair $(D, D \cap E^G)$ is homeomorphic to (E, E^G).

Complexity A key tool in our classification is the following ranking of the "local complexity" of an end, which (as we show) gives a partial order on equivalence classes of ends.

Definition 1.3 For $x, y \in E$, we say $x \preccurlyeq y$ if every neighborhood of y contains a homeomorphic copy of a neighborhood of x. We say x and y are *equivalent* if $x \preccurlyeq y$ and $y \preccurlyeq x$.

We show that this order has maximal elements (Proposition 4.7), and for A a clopen subset of E, we denote the maximal ends of A by $\mathcal{M}(A)$.

The following theorem gives the classification of locally CB mapping class groups. While the statement is technical, it is easy to apply in specific examples. For instance, the surfaces in Figure 1, left, satisfy the conditions, while those on the right fail to have CB mapping class group.

Theorem 1.4 (classification of locally CB mapping class groups) $\mathrm{Map}(\Sigma)$ *is locally CB if and only if there is a finite-type surface $K \subset \Sigma$ with the following properties:*

(i) *Each complementary region of K has one or infinitely many ends and infinite or zero genus.*

(ii) *The complementary regions of K partition E into clopen sets, indexed by finite sets $\mathcal{A}$ and $\mathcal{P}$ such that*

- *each $A \in \mathcal{A}$ is self-similar, with $\mathcal{M}(A) \subset \mathcal{M}(E)$ and $\mathcal{M}(E) \subset \bigsqcup_{A \in \mathcal{A}} \mathcal{M}(A)$,*
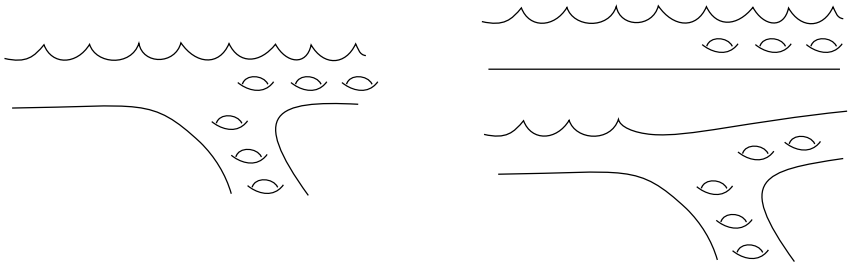
Figure 1: By Theorem 1.4, the surface on the left has a locally CB mapping class group and those on the right do not. All have $\mathcal{P} = \varnothing$.

- each $P \in \mathcal{P}$ is homeomorphic to a clopen subset of some $A \in \mathcal{A}$, and
- for any $x_A \in \mathcal{M}(A)$ and any neighborhood V of the end x_A in Σ, there is $f_V \in \mathrm{Homeo}(\Sigma)$ such that $f_V(V)$ contains the complementary region to K with end set A.

Moreover, in this case the set $\mathcal{V}_K := \{g \in \mathrm{Homeo}(\Sigma) : g|_K = \mathrm{id}\}$ *is a CB neighborhood of the identity.*

In order to illustrate Theorem 1.4 and motivate the conditions in the next two classification theorems, we now state results in the much simpler special case when Σ has genus zero and countable end space.

Special case: E countable, genus zero If E is a countable set and $E^G = \varnothing$, a classical result of Mazurkiewicz and Sierpinski [13] states that there exists a countable ordinal α such that E is homeomorphic to the ordinal $\omega^\alpha n + 1$ equipped with the order topology. Thus, any $x \in E$ is locally homeomorphic to $\omega^\beta + 1$ for some $\beta \leq \alpha$ (here β is the Cantor–Bendixon rank of the point x). In this case, our partial order $\prec$ agrees with the usual ordering of the ordinal numbers, points are equivalent if and only if they are locally homeomorphic, and we have the following.

Theorem 1.5 (special case of Theorems 1.4, 1.6 and 1.7) *Suppose Σ is an infinite-type surface of genus zero with $E \cong \omega^\alpha n + 1$. Then:*

(i) $\mathrm{Map}(\Sigma)$ *is CB if and only if* $n = 1$; *in this case E is self-similar.*

(ii) *If $n \geq 2$ and α is a successor ordinal, then $\mathrm{Map}(\Sigma)$ is locally CB and generated by a CB set, but admits a surjective homomorphism to $\mathbb{Z}$, so is not globally CB.*

(iii) *If $n \geq 2$ and α is a limit ordinal, then $\mathrm{Map}(\Sigma)$ is locally CB, but not generated by any CB set.*

Classification: general case   One cannot hope for such a clean statement as that of Theorem 1.5 to hold in general, since there is no similarly clean classification of end spaces. In fact, even in the genus-zero case, classifying possible end spaces E (ie closed subsets of Cantor sets) up to homeomorphism is a difficult and well-studied problem, equivalent to the classification problem for countable Boolean algebras.[2] Ketonen [10] gives some description and isomorphism invariants. In practice these invariants are difficult to use, and yet they are in some sense an optimal classification, as Carmelo and Gao show in [5] that the isomorphism relation is Borel complete. Our definition of the partial order $\preccurlyeq$ allows us to sidestep the worst of these issues.

For technical reasons, the order is better behaved under a weak hypothesis on the topology of the end space, which we call "tameness". See Section 6 for motivation and the definition. To our knowledge, tame surfaces include all concrete examples studied thus far in the literature, including the mapping class groups of some specific infinite-type surfaces in Aramayona, Patel and Vlamis [2], Bavard [3] and Fanoni, Hensel and Vlamis [8], and the discussion of geometric or dynamical properties of various translation surfaces of infinite type in Chamanara [6], Hooper [9] and Randecker [15]. Although nontame examples do exist (see Example 6.13), there are no known nontame surface that have a well-defined quasi-isometry type (Problem 6.12). Under this hypothesis, we can give a complete classification of surfaces with a well-defined QI type, and those with a trivial QI type, as follows.

Theorem 1.6   (classification of CB generated mapping class groups)   *For a tame surface Σ with locally (but not globally) CB mapping class group, $\mathrm{Map}(\Sigma)$ is CB generated if and only if E is **finite rank** and not of **limit type**.*

Theorem 1.7   (classification of globally CB mapping class groups)   *Suppose Σ is either tame or has countable end space. Then $\mathrm{Map}(\Sigma)$ is CB if and only if Σ has infinite or zero genus and E is self-similar or a variant of this called "telescoping". The telescoping case occurs only when E is uncountable.*

Finite rank, loosely speaking, means that finite-index subgroups of $\mathrm{Map}(\Sigma)$ do not admit surjective homomorphisms to $\mathbb{Z}^n$ for arbitrarily large n. *Limit type* refers to behavior of equivalence classes for the partial order that mimics the behavior of limit ordinals in the special countable case stated above; see Section 6.2. *Telescoping* is a slightly broader notion of homogeneity or local similarity of an end space. Informally

---

[2]By Stone duality, totally disconnected, separable, compact sets are in one-to-one correspondence with countable Boolean algebras.

speaking, *self-similar* sets either appear very homogeneous (eg a Cantor set) or may have one "special" point, any neighborhood of which contains a copy of the whole set — for instance, a countable set with a single accumulation point is self-similar. Telescoping is a generalization that allows for two special points. Further motivation and a precise definition are given in Section 3.2.

Key tool: nondisplaceable subsurfaces The following tool is of independent interest and provides an easily employable criterion to certify that a surface has non-CB mapping class group (or, equivalently, admits a continuous isometric action on a metric space with unbounded orbits).

Definition 1.8 A connected, finite-type subsurface S of a surface Σ is said to be *nondisplaceable* if $f(S) \cap S \neq \varnothing$ for each $f \in \mathrm{Homeo}(\Sigma)$. A nonconnected surface is nondisplaceable if, for every $f \in \mathrm{Homeo}(\Sigma)$, there are connected components S_i and S_j of S such that $f(S_i) \cap S_j \neq \varnothing$.

Theorem 1.9 *If Σ is a surface that contains a nondisplaceable finite-type subsurface, then* $\mathrm{Map}(\Sigma)$ *is not globally CB.*

A key ingredient of the proof is *subsurface projection*, a familiar tool from the study of mapping class groups of finite-type surfaces, introduced by Masur and Minsky [12].

Theorem 1.9 immediately gives many examples of surfaces whose mapping class groups are not CB, and hence admit unbounded orbits on combinatorial complexes. For instance, any surface with finite but nonzero genus has this property. (See Theorem 1.5 below for a number of other easily described examples.) Theorem 1.9 also recovers, with a new proof, some of the work of Bavard in [3] and Durham, Fanoni and Vlamis in [7].

Outline

- Section 2 contains background information on standard mapping class group techniques, and the proof of Theorem 1.9.
- Section 3 gives two criteria for CB mapping class groups: self-similarity and telescoping end spaces. This is used later in the proof of the local and global CB classification theorems.
- Section 4 introduces the partial order on the end space and proves key properties of this relation, and a characterization of self-similar end spaces in terms of the partial order.

- Section 5 contains the proof of Theorem 1.4. This and the following section form the technical core of this work.
- Section 6 contains the proof of Theorem 1.6.
- Section 7 gives the proof of Theorem 1.7.

2 Proof of Theorem 1.9

In this section we prove that nondisplaceable finite-type subsurfaces of a surface Σ are responsible for nontrivial geometry in $\mathrm{Map}(\Sigma)$. We begin by introducing some notions from large-scale geometry and setting some conventions that will be useful throughout.

A criterion for coarse boundedness Recall that a subset $A \subset G$ of a metrizable, topological group is said to be *coarsely bounded* or *CB* if it has finite diameter in every compatible left-invariant metric on G. The following result gives an equivalent condition that is often easier to use in practice.

Theorem 2.1 (Rosendal [18, Proposition 2.7(5)]) *Let A be a subset of a Polish group G. The following are equivalent*:

(i) *A is coarsely bounded.*

(ii) *For every neighborhood $\mathcal{V}$ of the identity in G, there is a finite subset $\mathcal{F}$ and some $k \geq 1$ such that $A \subset (\mathcal{F}\mathcal{V})^k$.*

While Rosendal's theory is quite broadly applicable, mapping class groups (of any manifold) fall into the nicest family to which it applies, namely the completely metrizable or *Polish* groups. For any manifold M, the homeomorphism group $\mathrm{Homeo}(M)$ endowed with the compact-open topology is Polish, and hence also for any closed subset of M, the closed subgroups $\mathrm{Homeo}(M, X)$ and $\mathrm{Homeo}(M \operatorname{rel} X)$ of homeomorphisms, respectively preserving and pointwise fixing X. (In the mapping class groups context, X is typically taken to be the boundary of M or a set of marked points.) Thus, since

the identity component $\mathrm{Homeo}_0(M, X)$ is a closed, normal subgroup, the quotient $\mathrm{Homeo}(M, X)/\mathrm{Homeo}_0(M, X)$ is also a Polish group.[3]

One useful tool for probing the geometry of a topological group is the following concept of a length function.

Definition 2.2 A *length function* on a topological group G is a continuous function $\ell : G \to [0, \infty)$ satisfying $\ell(g) = \ell(g^{-1})$, $\ell(\mathrm{id}) = 0$ and $\ell(gh) \le \ell(g) + \ell(h)$ for all $g, h \in G$.

If ℓ is any length function, then for any $\epsilon > 0$ the set $\ell^{-1}([0, \epsilon))$ is a neighborhood of the identity in G. It follows from the criterion in Theorem 2.1 that ℓ is bounded on any CB subset.

Our strategy for the proof of Theorem 1.9 is to use the presence of a nondisplaceable subsurface to construct an unbounded length function. In order to do this, we introduce some notation and conventions which will also be used in later sections.

Surfaces: conventions The following conventions will be used throughout this work. Infinite-type surfaces, typically denoted by Σ, are assumed to be connected and orientable, and unless otherwise specified will be assumed to have empty boundary. By a *curve* in Σ we mean a free homotopy class of a nontrivial, nonperipheral, simple closed curve. In the first part of this section, when we talk about a subsurface $S \subset \Sigma$, we always assume that S is connected, has finite type and is *essential*, meaning that every curve in ∂S is nontrivial and nonperipheral in Σ. (Later we will broaden our discussion to include nonconnected subsurfaces.) As is standard, the *complexity* of a finite-type surface S is defined to be $\xi(S) = 3g_S + b_S + p_S$, where g_S is the genus, p_S is the number of punctures and b_S is the number of boundary components of S. *Finite type* simply means that all these quantities are finite.

The *intersection number* between two curves γ_1 and γ_2 is the usual geometric intersection number $\mathrm{i}(\gamma_1, \gamma_2)$, defined to be the minimal intersection number between representatives in the free homotopy classes of γ_1 and γ_2. To simplify the exposition going forward, we will fix a complete hyperbolic structure on Σ. Then every curve has a unique geodesic representative and the homotopy class of every subsurface has a unique representative that has geodesic boundary. A pair of curves γ_1 and γ_2 have

---

[3]For the case where M is a surface, that mapping class groups are Polish was also observed in [2] using the property that these groups are the automorphism groups of the curve complex of the surface.

disjoint representatives if and only if their geodesic representatives are disjoint. In this case, we say that $i(\gamma_1, \gamma_2) = 0$. Otherwise, we say γ_1 intersects γ_2 and in this case, the intersection number $i(\gamma_1, \gamma_2)$ is the cardinality of the intersection of their geodesic representatives.

Similarly, two subsurfaces R and S (or a subsurface R and geodesic γ) intersect if every subsurface homotopic to R intersects every subsurface homotopic to S (or analogously for γ), and this is equivalent to saying that the representatives of R and S with geodesic boundaries intersect each other. Hence, from now on, every time we consider a curve we assume it is a geodesic and every time we consider a subsurface we assume it has geodesic boundary. This allows us to unambiguously speak of intersections.

Definition 2.3 A finite-type, connected subsurface $S \subset \Sigma$ is *nondisplaceable* if $S \cap f(S) \neq \varnothing$ for all $f \in \mathrm{Map}(\Sigma)$.

Example 2.4 When Σ has positive, finite genus, any subsurface S whose genus matches that of Σ is nondisplaceable. This is because S contains nonseparating curves but $\Sigma - S$ does not. Since every image of S under a homeomorphism of Σ will also contain a nonseparating curve, it must intersect S.

Example 2.5 (nondisplaceable subsurfaces) It is also easy to construct examples of nondisplaceable surfaces using the topology of the end space. Suppose Σ has infinite end space, and Z is an invariant, finite set of ends of cardinality at least 3. Then any surface S which separates all the points of Z into different complementary regions will be nondisplaceable.

To give another prototypical example, if X and Y are disjoint, closed invariant sets of ends, with X homeomorphic to a Cantor set, then a subsurface homeomorphic to a pair of pants which contains points of X in two complementary regions, and all of Y in the third complementary region, will also be nondisplaceable.

Curve graphs and subsurface projections We recall some basic material on curve graphs. A reader unfamiliar with this machinery may wish to consult the introductory notes [19] or paper [11] for more details. As in the previous paragraph, we continue to assume here that surfaces are connected.

The *curve graph* $\mathcal{C}(S)$ of a finite-type surface S is a graph whose vertices are curves in S and whose edges are pairs of disjoint curves. We give each edge length one and denote the induced metric on $\mathcal{C}(S)$ by d_S. With this metric, as soon as $\xi(S) \geq 5$,

$(\mathcal{C}(S), d_S)$ has infinite diameter and is Gromov hyperbolic [11]. One can define curve graphs analogously for infinite-type surfaces, but these no longer have infinite diameter and we will use only the classical finite-type setting.

If Σ is any surface and $S \subset \Sigma$ a subsurface, there is a *projection map* π_S from the set of curves in Σ that intersect S to the set of subsets of $\mathcal{C}(S)$, defined as follows: for a curve γ, the intersection $\gamma \cap S$ of the geodesic γ with the subsurface S is either equal to γ (if $\gamma \subset S$) or is a union of arcs with endpoints in ∂S. For every such arc ω, one may perform a surgery between γ and ∂S to obtain in curve in S disjoint from ω, possibly in two different ways (the curve is a concatenation of one or two copies of ω and one or two arcs in ∂S). We define the projection $\pi_S(\gamma)$ to be γ if $\gamma \subset S$ and otherwise to be the *union* of curves associated to each arc on $\gamma \cap S$ obtained by surgery as above. When $\xi(S) \geq 5$, the set $\pi_S(\gamma)$ has diameter at most 2 in $\mathcal{C}(S)$; in fact, we have

$$(1) \qquad \qquad \mathrm{i}(\gamma_1, \gamma_2) = 0 \implies \mathrm{diam}_S \, \pi_S(\gamma_1 \cup \gamma_2) \leq 2.$$

See [12, Lemma 2.2] for more details. In general, if μ is a subset of $\mathcal{C}(S)$, we define

$$\pi_S(\mu) = \bigcup_{\gamma \in \mu} \pi_S(\gamma).$$

The natural distance d_S on $\mathcal{C}(S)$ can be extended to a distance function on curves in Σ that intersect S via

$$d_S(\gamma_1, \gamma_2) = \max_{\alpha_i \in \pi_S(\gamma_i)} d_S(\alpha_1, \alpha_2).$$

The following result states that a bound on the intersection number between two curves gives a bound on their projection distance in any subsurface. This principle is well known and there are many similar results in the literature. We give a short proof with a suboptimal bound.

Lemma 2.6 Let γ_1 and γ_2 be curves in Σ that intersect S. Then

$$(2) \qquad \qquad d_S(\gamma_1, \gamma_2) \leq 2 \log_2(\mathrm{i}(\gamma_1, \gamma_2) + 1) + 6.$$

Proof Let ω_1 be an arc in S that is a component of the restriction of γ_1 and let $\alpha_1 \in \pi_S(\gamma_1)$ be the curve in $\mathcal{C}(S)$ that is obtained by doing a surgery between ω_1 and the boundary of S. Then α_1 is a concatenation of one or two copies of ω_1 (depending on whether the endpoints of ω_1 are on the same boundary or different boundary components of S) and some arcs in ∂S. Similarly, let ω_2 be an arc in S that is a restriction of γ_2

and let α_2 be the associated curve in $\pi_S(\gamma_2)$. Then every intersection point between ω_1 and ω_2 results in 1, 2 or 4 intersection points between α_1 and α_2. Also, applying surgery between ω_2 and ∂S can result in two intersection points between α_2 and α_1 at each end of ω_2. Therefore,

$$\mathrm{i}(\alpha_1, \alpha_2) \leq 4\,\mathrm{i}(\omega_1, \omega_2) + 4.$$

On the other hand, from [19, Lemma 1.21], we have

$$d_S(\alpha_1, \alpha_2) \leq 2\log_2(\mathrm{i}(\alpha_1, \alpha_2)) + 2.$$

Therefore,

$$d_S(\alpha_1, \alpha_2) \leq 2\log_2(4\,\mathrm{i}(\omega_1, \omega_2) + 4) + 2 \leq 2\log_2(\mathrm{i}(\gamma_1, \gamma_2) + 1) + 6,$$

which is as we claimed. $\square$

The notions of distance d_S and intersection number can also be extended further to take finite sets of curves as arguments. If μ_i are finite sets of curves, we define

$$d_S(\mu_1, \mu_2) = \max_{\gamma_1 \in \mu_1, \gamma_2 \in \mu_s} d_S(\gamma_1, \gamma_2) \quad \text{and} \quad \mathrm{i}(\mu_1, \mu_2) = \max_{\gamma_1 \in \mu_1, \gamma_2 \in \mu_s} \mathrm{i}(\gamma_1, \gamma_2).$$

Using equation (2), for any finite subsets μ_1 and μ_1 of $\mathcal{C}(S)$, we have

$$(3) \qquad\qquad d_S(\mu_1, \mu_2) \leq 2\log_2(\mathrm{i}(\mu_1, \mu_2) + 1) + 6.$$

Note that the triangle inequality still holds for this generalized distance d_S.

Construction of an unbounded length function We now proceed with the proof of Theorem 1.9. Let Σ be any surface, and let S be a nondisplaceable subsurface. Enlarge S if needed so that $\xi(S) \geq 5$ and so that S is connected. (In Section 2.1, we give an alternative modification for nonconnected subsurfaces that will be useful in later work.)

Let $\mathcal{I}$ denote the set of (isotopy classes of) subsurfaces of the same type as S, ie

$$\mathcal{I} = \{f(S) \mid f \in \mathrm{Map}(\Sigma)\}.$$

As before, while $f(S)$ denotes only an isotopy class of a surface when $f \in \mathrm{Map}(\Sigma)$, the reader may identify it with an honest subsurface by taking the representative with geodesic boundary. Let μ_S be a filling set of curves in $\mathcal{C}(S)$, ie a set of curves with the property that every curve in S intersects some curve in μ.

For $R \in \mathcal{I}$ let $\mu_R = \pi_R(\mu_S)$. Note that this is always defined since μ_S fills S, and R intersects S because S was assumed nondisplaceable.

Now, define

$$\ell \colon \mathrm{Map}(\Sigma) \to \mathbb{Z} \quad \text{by} \quad \ell(\phi) = \max_{R \in \mathcal{I}} d_{\phi(R)}(\phi(\mu_R), \mu_{\phi(R)}).$$

Equivalently, we have

$$(4) \qquad\qquad \ell(\phi) = \max_{T \in \mathcal{I}} d_T(\phi(\mu_{\phi^{-1}(T)}), \mu_T).$$

Note that ℓ is finite because, for every ϕ, the intersection number $i(\mu_S, \phi(\mu_S))$ is a finite number. Hence, by equation (3), their projections to $\phi(R)$ lie at a bounded distance in $\mathcal{C}(R)$, with a bound that depends on ϕ alone, not on R.

The latter definition also makes it clear that $\ell(\phi) = \ell(\phi^{-1})$, since

$$\begin{aligned}
\ell(\phi^{-1}) &= \max_{T \in \mathcal{I}} d_T(\phi^{-1}(\mu_{\phi(T)}), \mu_T) \\
&= \max_{T \in \mathcal{I}} d_{\phi(T)}(\mu_{\phi(T)}, \phi(\mu_T)) \\
&= \max_{R = \phi(T) \in \mathcal{I}} d_R(\mu_R, \phi(\mu_{\phi^{-1}(R)})) = \ell(\phi).
\end{aligned}$$

We now check the triangle inequality. Let ψ and ϕ be given, and let $R \in \mathcal{I}$ be a surface such that $\ell(\psi\phi) = d_{\psi\phi(R)}(\psi\phi(\mu_R), \mu_{\psi\phi(R)})$. Then we have

$$\begin{aligned}
\ell(\psi\phi) &= d_{\psi\phi(R)}(\psi\phi(\mu_R), \mu_{\psi\phi(R)}) \\
&\leq d_{\psi\phi(R)}(\psi\phi(\mu_R), \psi(\mu_{\phi(R)})) + d_{\psi\phi(R)}(\psi(\mu_{\phi(R)}), \mu_{\psi\phi(R)}) \\
&= d_{\phi(R)}(\phi(\mu_R), \mu_{\phi(R)}) + d_{\psi(Q)}(\psi(\mu_Q), \mu_{\psi(Q)}) \quad (\text{where } Q = \phi(R)) \\
&\leq \ell(\phi) + \ell(\psi).
\end{aligned}$$

Continuity of ℓ as a function on $\mathrm{Map}(\Sigma)$ is a consequence of the following observation.

Observation *If ϕ and ϕ' agree on S, then $\ell(\phi) = \ell(\phi')$.*

Proof First note that for any $T \in \mathcal{I}$, we have $\mu_{\phi^{-1}(T)} \subset S \cap \phi^{-1}(T)$, hence

$$\phi(\mu_{\phi^{-1}(T)}) \subset \phi(S) \cap T.$$

Similarly,

$$\phi'(\mu_{\phi'^{-1}(T)}) \subset \phi'(S) \cap T.$$

But $\phi(S) \cap T = \phi'(S) \cap T$. In fact, $\phi(\mu_{\phi^{-1}(T)})$ is the projection of $\phi(\mu_S)$ to T and $\phi'(\mu_{\phi'^{-1}(T)})$ is the projection of $\phi'(\mu_S)$ to T. Since ϕ and ϕ' agree on S,

$$\phi(\mu_{\phi^{-1}(T)}) = \phi'(\mu_{\phi^{-1}(T)}),$$

from which it follows from (4) that $\ell(\phi) = \ell(\phi')$. $\qquad\square$

Thus, the preimage of $\ell(\phi)$ under ℓ contains the open set consisting of mapping classes agreeing with ϕ on S. The remaining condition on a length function is that the length of identity should be zero. This is not a consequence of our definition, however we may simply redefine $\ell(\phi) = 0$ for all ϕ which restrict to the identity on S, without affecting the validity of the triangle-inequality computation above, as can be checked easily by hand.

To see that ℓ is unbounded, let $\phi \in \mathrm{Map}(\Sigma)$ be a homeomorphism that preserves S and such that the restriction $\phi|_S$ of ϕ to S is a pseudo-Anosov homeomorphism of S. Then, for any curve γ in S,

$$(5) \qquad\qquad d_S(\gamma, \phi^n(\gamma)) \to \infty \quad \text{as } n \to \infty.$$

See eg [11] for details. Thus, ℓ is an unbounded length function, and so $\mathrm{Map}(\Sigma)$ is not coarsely bounded. $\square$

2.1 Disconnected subsurfaces

While we have so far worked only with connected nondisplaceable subsurfaces, there is a natural generalization of the work above to nonconnected subsurfaces. This will be useful when we need to find a nondisplaceable subsurface that is disjoint from a given compact subset of Σ to determine if $\mathrm{Map}(\Sigma)$ is locally CB. The extension to this broader framework requires a little care since, if we simply take the definitions above verbatim, then the diameter of the curve graph $C(S)$ is finite as soon as S is not connected. However, the following minor adaptations allow our work above to carry through in this case.

Definition 2.7 A *disconnected finite-type subsurface* is a finite union of pairwise disjoint finite-type surfaces. We say such a subsurface $\bar{S}$ is *nondisplaceable* if, for any $f \in \mathrm{Map}(\Sigma)$ and any connected component S_i of $\bar{S}$, there is a connected component S_j of $\bar{S}$ such that $S_j \cap f(S_i) \neq \varnothing$.

We now show how to use such a disconnected surface $\bar{S}$ to construct a length function on $\mathrm{Map}(\Sigma)$. As before, let $\mathcal{I}$ denote the set of images of $\bar{S}$ under mapping classes, ie

$$\mathcal{I} = \{f(\bar{S}) \mid f \in \mathrm{Map}(\Sigma)\}.$$

If $\bar{S} = \bigsqcup_{i=1}^{k} S_i$, where S_i are the connected components, then an element $\bar{R}$ of $\mathcal{I}$ is simply the disjoint union of a set $\{R_1, \dots R_k\}$, where $R_i = f(S_i)$. Let $\mu_{\bar{S}}$ be a set

of curves in $\bigcup_i \mathcal{C}(S_i)$ that fill every S_i. Keeping the notation from before, note that $\pi_{R_i}(\mu_{\bar{S}})$ is always defined since R_i intersects some S_j, and curves in $\mu_{\bar{S}}$ fill S_j. Now, define $\ell_{\bar{S}} \colon \mathrm{Map}(\Sigma) \to \mathbb{Z}$ by

$$\ell_{\bar{S}}(\phi) = \max_{\bar{R} \in \mathcal{I}} \max\{d_{\phi(R_i)}(\phi(\mu_{R_i}), \mu_{\phi(R)}) \mid R_i \text{ a component of } \bar{R}\}.$$

The same computation as in the connected case shows that $\ell_{\bar{S}}$ is finite, is continuous as a function on $\mathrm{Map}(\Sigma)$, and satisfies the triangle inequality with the same adjustment that $\ell_{\bar{S}}(\phi) = 0$ when ϕ is identity on $\bar{S}$. To see that $\ell_{\bar{S}}$ is unbounded, let $\phi \in \mathrm{Map}(\Sigma)$ be a homeomorphism that preserves $\bar{S}$ and such that the restriction $\phi|_{S_1}$ of ϕ to S_1 is a pseudo-Anosov homeomorphism of S_1. Since $\ell_{\bar{S}}$ is defined as a maximum of distances in various curve graphs, if ϕ has a positive translation length in $\mathcal{C}(S_1)$ (or in any $\mathcal{C}(S_i)$) then $\ell_{\bar{S}}(\phi^n) \to \infty$ as $n \to \infty$. This gives an alternative proof of Theorem 1.9 in the disconnected case, and the following more general statement:

Proposition 2.8 *If Σ contains a connected or disconnected, nondisplaceable, finite-type subsurface S such that each connected component of S has complexity at least 5, then there exists a length function ℓ defined on $\mathrm{Map}(\Sigma)$ such that the restriction of ℓ to mapping classes supported on S is unbounded.*

3 Self-similar and telescoping end spaces

In this section we give two topological conditions (in Propositions 3.1 and 3.5) that imply coarse boundedness of the mapping class group: *self-similarity* and *telescoping*.

3.1 Self-similar end spaces

Recall that a space of ends (E, E^G) is said to be *self-similar* if for any decomposition $E = E_1 \sqcup E_2 \sqcup \cdots \sqcup E_n$ of E into pairwise disjoint clopen sets, there exists a clopen set D in some E_i such that $(D, D \cap E^G)$ is homeomorphic to (E, E^G). There are many examples of such sets; a few basic ones are:

- E equal to a Cantor set, and E^G either empty, equal to E, or a singleton.

- E a countable set homeomorphic to $\omega^\alpha + 1$ with the order topology, for some countable ordinal α, and E^G the set of points of type $\omega^\beta + 1$ for all ordinals $\beta \geq \beta_0$, where β_0 is a some fixed ordinal.

- E the union of a countable set Q and a Cantor set where the sole accumulation point of Q is a point in the Cantor set, and $E^G = \bar{Q}$.

Convention Going forward, we drop the notation E^G, assuming that E comes with a designated closed subset of ends accumulated by genus, empty if the genus of Σ is finite, and that all homeomorphisms between sets or subsets of end spaces preserve (setwise) the ends accumulated by genus.

As E and E^G are totally disconnected spaces, we also make the following convention.

Convention For the remainder of this work, when we speak of a *neighborhood* in an end space E, we always mean a *clopen neighborhood*.

Proposition 3.1 (self-similar implies CB) *Let Σ be a surface of infinite or zero genus. If the space of ends of Σ is self-similar, then* $\mathrm{Map}(\Sigma)$ *is CB.*

Note that finite, nonzero-genus surfaces cannot have CB mapping class groups by Example 2.4, so Proposition 3.1 is optimal in this sense. Note also that the proposition holds for finite-type surfaces as well, but the only applicable example is the once-punctured sphere, which has trivial mapping class group.

Proof of Proposition 3.1 Let Σ be an infinite-type surface satisfying the hypotheses of the proposition, and let $\mathcal{V}$ be a neighborhood of the identity in $\mathrm{Map}(\Sigma)$. Then there exists some finite-type subsurface S such that $\mathcal{V}$ contains the open set $\mathcal{V}_S$ consisting of mapping classes of homeomorphisms that restrict to the identity on S. By Theorem 2.1, it suffices to find a finite set $\mathcal{F} \subset \mathrm{Map}(\Sigma)$ and $k \in \mathbb{N}$ (which are allowed to depend on $\mathcal{V}_S$, hence on S) such that $\mathrm{Map}(\Sigma) = (\mathcal{F}\mathcal{V}_S)^k$. Enlarging S (and therefore shrinking $\mathcal{V}_S$) if needed, we may assume that each connected component of $\Sigma - S$ is of infinite type.

Since the proof is somewhat technical, we begin with an outline. The first step is to find a suitable homeomorphism f of Σ so that $f(S) \cap S = \varnothing$, and declare $\mathcal{F}$ to be the finite set consisting of f and f^{-1}. Now suppose one is given $g \in \mathrm{Map}(\Sigma)$. Obviously if g restricts to the identity on S, then $g \in \mathcal{V}_S$ and we are done (in fact $k = 1$ would work). If instead g restricted to the identity on $f(S)$, then we would have $g \in f\mathcal{V}_S f^{-1}$, and again are done, and could have taken $k = 2$. The general philosophy of the proof is to cleverly choose f so that *every* mapping class g can be written as a product of at most three elements which are either the identity on S or on $f(S)$, and use this to get the desired bound on k. In practice, we do this by finding an additional homeomorphic copy of S in Σ. Now we provide the details.
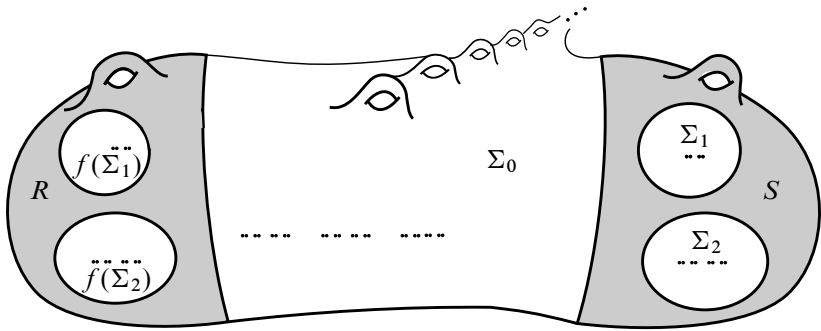
Figure 2: A homeomorphic copy R of S contained in the complementary region Σ_0.

The connected components of $\Sigma - S$, together with the finite set P of punctures of S, partition E into clopen sets. Let

$$E = E_0 \sqcup E_1 \sqcup \cdots \sqcup E_n \sqcup P$$

denote this decomposition, and let Σ_i denote the connected component of $\Sigma - S$ containing E_i. Since S is of finite type, $E^G \cap P = \varnothing$. Since E is self-similar, one of the E_i contains a copy of E. Without loss of generality, we assume this is E_0, the set of ends of Σ_0; thus we may write $E_0 = E' \sqcup D$, where E' is homeomorphic to E. The next lemma asserts that we may find a surface $R = f(S)$ as depicted in Figure 2.

Lemma 3.2 *With the notation above, there exists $f \in \mathrm{Homeo}(\Sigma)$ such that*

(i) $R = f(S) \subset \Sigma_0$,

(ii) $S \subset f(\Sigma_0)$, *and*

(iii) *the end set of $f(\Sigma_0) \cap \Sigma_0$ contains a homeomorphic copy of E.*

Proof Since E' is homeomorphic to E we can write E' as the disjoint union of sets E_i', with $i = 0, 1, \ldots, n$, and P', where $E_i' \cong E_i$ and $P' \cong P$. (Of course, by $\cong$ we mean homeomorphic via a homeomorphism which respects E^G.) We can further write $E_0' = E'' \sqcup D'$, where $E'' \cong E$ and $D' \cong D$.

Consider a subsurface R disjoint from S with puncture set P' and $n+1$ complementary regions, one with end space E_i' for each $i = 1, 2, \ldots, n$, and the final one containing the remaining ends, namely $D' \sqcup E'' \sqcup D \sqcup \left(\bigsqcup_{1 \le i \le n} E_i \right)$. Now we have

$$E'' \sqcup D \sqcup \left(\bigsqcup_{1 \le i \le n} E_i \right) \cong E,$$
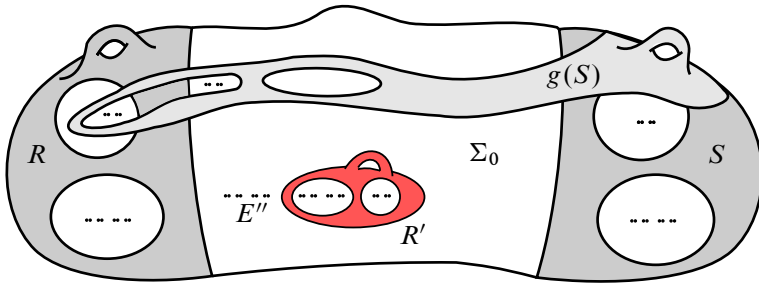
Figure 3: The surface $g(S)$ may intersect R and S in a complicated way, but R' lies in the "big" complementary region of at least one of them (in this illustration, it is in their intersection Z).

therefore,

$$D' \sqcup E'' \sqcup D \sqcup \left(\bigsqcup_{1 \leq i \leq n} E_i \right) \cong D' \sqcup E \cong E_0.$$

Thus, we may apply Richards' classification of surfaces and conclude that there is a homeomorphism f of Σ such that $f(S) = R$ and for $i \geq 1$ we have $f(E_i) = E'_i$, and $f(E_0) = D' \sqcup E'' \sqcup D \sqcup \left(\bigsqcup_{1 \leq i \leq n} E_i \right)$. □

Now fix R and f as in Lemma 3.2 and let $\mathcal{F} = \{f, f^{-1}\}$. We will show

$$\mathrm{Map}(\Sigma) = (\mathcal{F}\mathcal{V}_S)^5.$$

Let $g \in \mathrm{Map}(\Sigma)$. Let E' be a homeomorphic copy of E in the end space of $\Sigma_0 \cap f(\Sigma_0)$, and consider the set $g(E')$.

Since the clopen sets $Z := (f(E_0) \cap E_0)$, $(E_0 - Z)$ and $(E - E_0)$ partition E, their intersections with $g(E')$ partition $g(E')$. Since $g(E') \cong E$ is a self-similar set, one of these three sets in the partition contains a homeomorphic copy of E; call this E''. Thus, E'' lies either in $g(E') \cap f(E_0)$ or in $g(E') \cap E_0$ (or both). If the first case occurs, then we have $f^{-1}g(E') \cap f^{-1}(f(E_0))$. This means that, at the cost of replacing g by $f^{-1}g$, and therefore using one more letter from $\mathcal{F}$, we can assume that we are in the second case, ie where $E'' \subset g(E') \cap E_0$. So it suffices to show that in this case, we have $g \in (\mathcal{F}\mathcal{V}_S)^4$. This situation is illustrated in Figure 3. (For simplicity, we did not draw infinite genus on this image.)

Assuming that $E'' \subset g(E') \cap E_0$, the next step is to find another copy of S in a small neighborhood of E'', and hence in $g(\Sigma_0) \cap \Sigma_0$. In detail, just as in Lemma 3.2, but using E'' instead of E', and working with the subsurface R of the surface Σ_0 instead of the subsurface S of Σ, we may find a surface $R' \subset \Sigma_0 \cap g(\Sigma_0)$ homeomorphic to R,

and a homeomorphism v of Σ_0 mapping R to R' that satisfies $R \subset v(f(\Sigma_0) \cap \Sigma_0)$. Extend v to a homeomorphism of Σ by declaring it to be the identity on $\Sigma - \Sigma_0$; abusing notation slightly, denote this homeomorphism also by v, and so we have $v \in \mathcal{V}_S$. Then R, S and $g(S)$ are all contained in $v(f(\Sigma_0))$. See Figure 3 for a schematic.

The same argument as that in Lemma 3.2 using the classification of surfaces now shows that we may find u restricting to the identity on R', with $ug(S) = S$ and ug equal to identity on S. (The details are a straightforward exercise.) Since u is the identity on R', it follows that $(vf)^{-1}u(vf)$ is the identity on $(vf)^{-1}(R') = S$, which implies that $u \in (\mathcal{F}\mathcal{V}_S)^3$, hence $g \in (\mathcal{F}\mathcal{V}_S)^4$. This concludes the proof of Proposition 3.1. □

3.2 Telescoping end spaces

Motivation Recall from Example 2.5 that, if Σ is a surface such that there exists a finite, $\mathrm{Map}(\Sigma)$–invariant set $F \subset E$ of cardinality at least three, then $\mathrm{Map}(\Sigma)$ is not CB: any finite-type subsurface S such that the elements of F each lie in different complementary regions of S is easily seen to be nondisplaceable. The definition of *telescoping* below was motivated by the question: *Under what conditions is a two-element* $\mathrm{Map}(\Sigma)$*–invariant subset of E compatible with global coarse boundedness?* As will follow from our work in Section 7, this never happens if E is countable: every surface with countable end space and coarsely bounded mapping class group is self-similar. However, in the uncountable case, surfaces with telescoping end spaces provide additional examples (and are the only additional examples among tame surfaces). Informally speaking, telescoping spaces of ends have two "special" points with the property that neighborhoods of each point can be expanded an arbitrary amount, and can also be expanded a fixed amount relative to a neighborhood of the other point.

Convention In the following definition, and for the remainder of this work, we wish to work only with specific neighborhoods of ends in Σ, not every open subset of the surface containing this end. Thus, going forward, a *neighborhood of an end x in Σ* means a connected subsurface with a single boundary component that has x as an end.

Definition 3.3 A surface Σ is *telescoping* if there are ends $x_1, x_2 \in E$ and disjoint clopen neighborhoods V_i of x_i in Σ such that for all clopen neighborhoods $W_i \subset V_i$ of x_i, there exist homeomorphisms f_i and h_i of Σ, both pointwise fixing $\{x_1, x_2\}$, with

$$f_i(W_i) \supset (\Sigma - V_{3-i}), \quad h_i(W_i) = V_i, \quad h_i(V_{3-i}) = V_{3-i}.$$

When we wish to make the points x_1, x_2 explicit, we say also *telescoping with respect to $\{x_1, x_2\}$*. We may equivalently require h_i to restrict to the identity on V_{3-i}.

Note that this definition implies that Σ has infinite or zero genus, as does $\Sigma - (V_1 \cup V_2)$.

While the complement of a Cantor set in S^2 is both self-similar and telescoping with respect to any pair of points, there are many examples of telescoping sets that are *not* self-similar, for instance:

- E^G a Cantor set, and E the union of E^G and another Cantor set which intersects E^G at exactly two points.

- E the union of two copies of the Cantor set, C_1 and C_2, which intersect at exactly two points, and a countable set Q such that the accumulation points of Q are exactly C_1. E^G could be empty, equal to the closure of Q, or equal to E.

Note that, in Definition 3.3, f_i and h_i are required to be homeomorphisms of the surface, not merely the end space.

Remark 3.4 An equivalent definition of telescoping may be given by replacing "there exist disjoint neighborhoods V_i of x_i" with "for all sufficiently small neighborhoods V_i of x_i". The proof is an immediate consequence of the definition.

The telescoping condition also implies that all neighborhoods of x_i in $\Sigma - \{x_{3-i}\}$ are homeomorphic. With this fact, one can use a standard back-and-forth argument to show that there is a homeomorphism of Σ taking x_i to x_{3-i}. We omit the proof as it is not needed for what follows.

Proposition 3.5 (telescoping implies CB) *Let Σ be a surface that is telescoping with respect to $\{x_1, x_2\}$. Then the pointwise stabilizer of $\{x_1, x_2\}$ in $\mathrm{Map}(\Sigma)$ is CB.*

In particular, if $\{x_1, x_2\}$ is a $\mathrm{Map}(\Sigma)$–invariant set, then $\mathrm{Map}(\Sigma)$ is itself CB.

Remark 3.6 In fact, it will follow easily from the tools developed in the next section (see Proposition 4.8) that if $\{x_1, x_2\}$ is not invariant, then the end space of Σ is self-similar and so $\mathrm{Map}(\Sigma)$ is CB in this case as well.

Proof of Proposition 3.5 Suppose that Σ is telescoping and let x_i and V_i be as in the definition. To simplify notation, let G denote the pointwise stabilizer of $\{x_1, x_2\}$ in $\mathrm{Map}(\Sigma)$. Fix a neighborhood of the identity in $\mathrm{Map}(\Sigma)$; shrinking this if needed we may take it to be the set $\mathcal{V}_S$ of mapping classes that restrict to the identity on some finite-type subsurfaces S. By Remark 3.4, we may assume that $S \subset \Sigma - (V_1 \cup V_2)$. Let

$V \subset \mathcal{V}_S$ be the set of mapping classes that restrict to the identity on $\Sigma' := \Sigma - (V_1 \cup V_2)$. We will exhibit a finite set $\mathcal{F}$ such that $G \subset (\mathcal{F}V)^{10} \subset (\mathcal{F}\mathcal{V}_S)^{10}$. This is sufficient to show that G is CB, by Theorem 2.1.

Fix neighborhoods $W_i \subset V_i$ of x_i in Σ and homeomorphisms f_i with $f_i(W_i) \supset (\Sigma - V_{3-i})$, as given by the definition of telescoping. Let $\mathcal{F} = \{f_1^{\pm 1}, f_2^{\pm 1}\}$. This is our finite set. Note that any homeomorphism which restricts to the identity on V_i lies in $f_{3-i}V f_{3-i}^{-1}$.

Given $g \in G$, let W_i' be a neighborhood of x_i small enough that $W_i' \subset g^{-1}(V_i) \cap g(V_i)$. By definition of telescoping, there exist homeomorphisms h_1 and j_1, both restricting to the identity on V_2, with $h_1(g(W_1')) = V_1$ and $j_1(W_1) = V_1$. Then $g_1 := j_1^{-1}h_1$ is the identity on V_2, hence lies in $f_1 V f_1^{-1}$, and satisfies $g_1 g(W_1') = W_1$.

Similarly, we can find $g_2 \in f_2 V f_2^{-1}$ restricting to the identity on V_1, and satisfying $g_2 g(W_2') = W_2$. Thus,

$$g_2 g_1 g(W_i') = W_i \quad \text{for } i = 1, 2.$$

It follows that $g_2 g_1 g(\Sigma') \subset (\Sigma - W_1 \cup W_2)$, so $f_1 g_2 g_1 g(\Sigma') \subset V_2$ and

$$f_2^{-1} f_1 g_2 g_1 g(\Sigma') \subset W_2.$$

For notational convenience, let $\phi = f_2^{-1} f_1 g_2 g_1 g$. Since $\phi(\Sigma')$ and $f_2^{-1} f_1 \Sigma'$ both lie in W_2, as a consequence of the definition of telescoping there exists a homeomorphism ψ restricting to the identity on V_1, with $\psi\phi(\Sigma') = f_2^{-1} f_1(\Sigma')$. Precomposing ψ with a homeomorphism that is also the identity on V_1, we can also ensure that $(f_2^{-1} f_1)^{-1} \psi\phi$ restricts to the identity on Σ'. Thus, we have shown that $(f_2^{-1} f_1)^{-1} \psi\phi = (f_2^{-1} f_1)^{-1} \psi f_2^{-1} f_1 g_2 g_1 g \in V$. Since $\psi^{-1} \in (\mathcal{F}V)^2$, and $g_i^{-1} \in (\mathcal{F}V)^2$, we conclude that $g \in (\mathcal{F}V)^{10}$. Since $\mathcal{F}$ and the exponent are independent of g, we have proved the desired result. $\qquad\square$

We conclude this section with a result whose proof serves as a good warm-up for the technical work to come.

Proposition 3.7 *No telescoping surface has countable end space.*

Proof Suppose that Σ has countable end space E. Recall in this case $E \cong \omega^\alpha n + 1$ by [13], and $E^G \subset E$ is some closed subset. Assume for contradiction that E is telescoping with respect to some pair of ends x_1, x_2. For each point $x \in E$, there exists

$\beta = \beta(x) \leq \alpha$ such that every sufficiently small neighborhood of x is homeomorphic to $\omega^\beta + 1$ (this ordinal $\beta(x)$ is simply the Cantor–Bendixon rank of x). It follows from the definition of telescoping that every clopen neighborhood U of x_i disjoint from x_{3-i} is homeomorphic to every other such neighborhood. In particular, necessarily $n = 2$ and x_1 and x_2 are points of equal and maximal rank α. Suppose as a first case that α is a successor ordinal and let η denote its immediate predecessor. Then the set of points of rank η accumulates only at x_1 and x_2. If V_i is any neighborhood of x_i, then $\Sigma - (V_1 \cup V_2)$ contains finitely many points of rank η. Thus, if $W_1 \subset V_1$ satisfies that $V_1 - W_1$ contains exactly one point of rank η, then no homeomorphism fixing V_2 can send W_1 to V_1, and the definition of telescoping fails.

The case where α has limit type is similar. Given neighborhoods V_i of x_i, let $\eta < \alpha$ be the supremum of the ranks of points in $E - (V_1 \cup V_2)$. Let $W_1 \subset V_1$ be a set such that $V_1 - W_1$ contains a point of rank α where $\eta < \alpha$. Then no homeomorphism fixing V_2 can send W_1 to V_1, and the definition of telescoping fails. $\square$

As we will see in the next sections, this limit type phenomenon is closely related to the failure of the mapping class group to be generated by a CB set. However, to treat this in the case where E is uncountable, we will need to develop a more refined ordering on the space of ends.

4 A partial order on the space of ends

Let Σ be an infinite-type surface with set of ends (E, E^G). As in the previous section, we drop the notation E^G and, by convention, all homeomorphisms of an end space E of a surface Σ are required to preserve E^G, so to say that $A \subset E$ is homeomorphic to $B \subset E$ means that there is a homeomorphism from $(A, A \cap E^G)$ to $(B, B \cap E^G)$. It follows from Richards' classification of surfaces in [16, Theorem 1] that each homeomorphism of (E, E^G) is induced by a homeomorphism of Σ.[4] Thus, we will pass freely between speaking of homeomorphisms of the end space and the underlying surface.

Observe also that, if U and V are two disjoint, clopen subsets of E, then any homeomorphism f from U onto V can be extended to a globally defined homeomorphism $\bar{f}$ of E by declaring $\bar{f}$ to agree with f^{-1} on V and to pointwise fix the complement

---

[4]While this is not in the statement of [16, Theorem 1], the proof gives such a construction. This was originally explained to the authors by J Lanier following work of S Afton.

of $U \cup V$. Thus, to say points x and y are locally homeomorphic is equivalent to the condition that there exists $\bar{f} \in \text{Map}(\Sigma)$ with $\bar{f}(x) = y$. We will use this fact frequently. In particular, we have the following equivalent rephrasing of Definition 1.3:

Definition 4.1 Let $\preccurlyeq$ be the binary relation on E given by $y \preccurlyeq x$ if, for every neighborhood U of x, there exists a neighborhood V of y and $f \in \text{Map}(\Sigma)$ such that $f(V) \subset U$.

Note that this relation is transitive.

Notation 4.2 For $x, y \in E$ we say that $x \sim y$ or "x and y are of the same type" if $x \preccurlyeq y$ and $y \preccurlyeq x$, and write $E(x)$ for the set $\{y \mid y \sim x\}$ of "ends of type x".

It is easily verified that $\sim$ defines an equivalence relation: symmetry and reflexivity are immediate from the definition, while transitivity follows from the transitivity of $\preccurlyeq$. From this it follows that the relation $\prec$, defined by $x \prec y$ if $x \preccurlyeq y$ and $x \nsim y$, gives a partial order on the set of equivalence classes under $\sim$. For any homeomorphism f of Σ, we have $x \succ y$ (resp. $x \succcurlyeq y$) if and only if $f(x) \succ f(y)$ (resp. $f(x) \succcurlyeq f(y)$).

Proposition 4.3 *If E is countable, then $x \sim y$ if and only if x and y are locally homeomorphic. If additionally $E^G = \varnothing$, then the Cantor–Bendixon rank gives an order isomorphism between equivalence classes of points and countable ordinals.*

Proof Suppose that E is countable. Consider first the case where $E^G = \varnothing$. Then every point $x \in E$ has a neighborhood U_x homeomorphic to the set $\omega^{\alpha(x)} + 1$, where $\alpha(x)$ is the Cantor–Bendixon rank of x. If $x \preccurlyeq y$ and $y \preccurlyeq x$ both hold, it follows that $\alpha(x) = \alpha(y)$, and so any homeomorphism from a neighborhood of x into a neighborhood of y necessarily takes x to y. Thus, x and y are locally homeomorphic. In particular, these points also have the same rank.

In the general case where $E^G \neq \varnothing$, let $\bar{E}$ denote the topological space of ends (with no distinction between those accumulated by genus or not). Any homeomorphism of E induces one of $\bar{E}$ by simply forgetting that E^G is preserved. Thus, the argument above shows that if $x \sim y$ in E, then they admit neighborhoods U_x and U_y in $\bar{E}$ which are homeomorphic. Moreover, such a homeomorphism necessarily takes x to y, and in fact no homeomorphism of E can take x to another point of U_y. Thus, $x \preccurlyeq y$ implies that there is a homeomorphism of E taking a neighborhood of x in E to one of y. The converse statement is immediate. $\qquad\square$

Remark 4.4 We do not know if Proposition 4.3 holds in the uncountable case. This appears to be an interesting question. However, it is quite easy to construct large families of examples for which it does hold.

Remark 4.5 Despite the above remark, there are indeed some marked differences between the behavior of $\prec$ when E is countable and when E is uncountable. In the countable case, it follows from Proposition 4.3 that $x \prec y$ if and only if y is an accumulation point of $E(x)$, giving a convenient alternative description of $\prec$. In general, a weaker statement holds: we show below that if y is an accumulation point of $E(x)$, then $x \preccurlyeq y$. However, if E is a Cantor set and $E^G = \varnothing$, for example, then all points are equivalent and all are accumulation points of their equivalence class.

We now prove some general results on the structure of $\preccurlyeq$.

Lemma 4.6 *For every $y \in E$, the set $\{x \mid x \succcurlyeq y\}$ is closed.*

Proof Consider a sequence $x_n \to x$ where $x_n \succcurlyeq y$ holds for all n. Let U be a neighborhood of x. Then, for large n, U is also a neighborhood of x_n and hence contains homeomorphic copies of some neighborhood of y. $\square$

Proposition 4.7 *The partial order $\succ$ has maximal elements. Furthermore, for every maximal element x, the equivalence class $E(x)$ is either finite or a Cantor set.*

Proof To show that E has maximal elements, by Zorn's lemma it suffices to show that every chain has an upper bound. Suppose that $\mathcal{C}$ is a totally ordered chain. Consider the family of sets $\{x \mid x \succcurlyeq y\}$, for $y \in \mathcal{C}$. Then, by Lemma 4.6, this is a family of nested, closed, nonempty sets and hence

$$\mathcal{C}_M = \bigcap_{y \in \mathcal{C}} \{x \mid x \succcurlyeq y\}$$

is nonempty. By definition, any point of this intersection is an upper bound for $\mathcal{C}$.

To see the second assertion, consider a maximal element x. If $E(x)$ is an infinite set, then it has an accumulation point, say z. Then $z \succcurlyeq x$, but since x is maximal, we have $z \sim x$. Since any neighborhood of any other point in $E(x)$ contains a homeomorphic copy of a neighborhood of z, it follows that all points of $E(x)$ are accumulation points and hence $E(x)$ is a Cantor set. $\square$

Going forward, we let $\mathcal{M} = \mathcal{M}(E)$ denote the set of maximal elements for $\succ$.

4.1 Characterizing self-similar end sets

The remainder of this section consists of a detailed study of the behavior of end sets using the partial order. We will develop a number of tools for the classification of locally CB and CB generated mapping class groups that will be carried out in the next sections.

Proposition 4.8 *Let Σ be a surface with end space E and no nondisplaceable subsurfaces. Then E is self-similar if and only if $\mathcal{M}$ is either a singleton or a Cantor set of points of the same type.*

One direction is easy and does not require the assumption that Σ has no nondisplaceable subsurfaces: if $\mathcal{M}$ contains two distinct maximal types x_1 and x_2, then a partition $E = E_1 \sqcup E_2$, where $E(x_i) \subset E_i$, fails the condition of self-similarity. Similarly, if $\mathcal{M}$ is a finite set of cardinality at least two, then any partition separating points of $\mathcal{M}$ similarly fails the condition. By Proposition 4.7, the only remaining possibility is that $\mathcal{M}$ is a Cantor set of points of the same type. This proves the first direction. The converse is more involved, so we treat the singleton and Cantor set case separately. We will need the following easy observation.

Observation 4.9 ("shift maps") *Suppose $U_1, U_2, \ldots$ are disjoint, pairwise homeomorphic clopen sets which Hausdorff converge to a point x. Then $\bigcup_{i=1}^{\infty} U_i \cup \{x\}$ is homeomorphic to $\bigcup_{i=2}^{\infty} U_i \cup \{x\}$.*

Proof For each i, fix a homeomorphism $f_i \colon U_i \to U_{i+1}$. Since the U_i Hausdorff converge to a point, the union of these defines a global homeomorphism $\bigcup_{i=1}^{\infty} U_i \to \bigcup_{i=2}^{\infty} U_i$ that extends continuously to x. □

We will also use the following alternative characterization of self-similarity:

Lemma 4.10 *Self-similarity is equivalent to the following condition: if $E = E_1 \sqcup E_2$ is a decomposition into clopen sets, then some E_i contains a clopen set homeomorphic to E.*

Proof Self-similarity implies the condition by taking $n = 2$. For the converse, suppose the condition holds and let $E = E_1 \sqcup E_2 \sqcup \cdots \sqcup E_n$ be a decomposition into clopen sets. Grouping these as $E_1 \sqcup (E_2 \sqcup \cdots \sqcup E_n)$, by assumption one of these subsets contains

a clopen set E' homeomorphic to E. If it is E_1, we are done. Else, the sets $E' \cap E_i$ with $i = 2, 3, \ldots, n$ form a decomposition of $E' \cong E$ into clopen sets; so by the same reasoning either $E_2 \cap E'$ contains a clopen set homeomorphic to E, or the union of the sets $E' \cap E_i$, for $i \geq 3$, does. Iterating this argument eventually produces a set homeomorphic to E in one of the E_i. □

The next three lemmas give the proof of Proposition 4.8.

Lemma 4.11 *Suppose Σ has no nondisplaceable subsurfaces and $\mathcal{M}$ is a singleton. Let $E = A \sqcup B$ be a decomposition into clopen sets. If $\mathcal{M} \subset A$, then A contains a homeomorphic copy of B.*

Proof Let $E = A \sqcup B$ be a decomposition of E into clopen sets with $\mathcal{M} = \{x\} \subset A$. Since A is a neighborhood of x, every point $y \in B$ has a neighborhood homeomorphic to a subset of A. Since B is compact, finitely many of these cover B, say $U_1, U_2, \ldots, U_k$. Without loss of generality, we may assume all the U_i are disjoint and their union is equal to B. For each i, let V_i be a homeomorphic copy of U_i in A; note that $x \notin \bigcup_i V_i$. Let S be a three-holed sphere subsurface such that the disjoint sets $\{x\}$, $\bigcup_i V_i$ and B all lie in different connected components of the complement of S. Let f be a homeomorphism displacing S. Since $f(x) = x$, up to replacing f with its inverse, we have either $f(B) \subset A$, in which case we are done, or A contains a homeomorphic copy of $A \sqcup (\bigcup_i V_i)$. In this latter case, by iterating f we can find k disjoint copies of $\bigcup_i V_i$ inside A. Since each contains a copy of U_i, this gives a subset of A homeomorphic to $\bigsqcup U_i = B$. □

As a consequence, we can prove the first case of Proposition 4.8.

Lemma 4.12 *Suppose Σ has no nondisplaceable subsurfaces and $\mathcal{M}$ is a singleton. Then E is self-similar.*

Proof Let $E = E_1 \sqcup E_2$ be a decomposition of E into clopen sets. Without loss of generality, suppose $\mathcal{M} = \{x\} \subset E_1$. Lemma 4.11 says that there is a homeomorphic copy U_2 of E_2 inside E_1, necessarily this is disjoint from $\{x\}$. Let A be a small neighborhood of x, disjoint from U_2. Lemma 4.11 again gives a homeomorphic copy U_3 of E_2 inside A. Proceeding in this way, we may find $E_2 = U_1, U_2, U_3, \ldots$, each homeomorphic to E_2 and Hausdorff converging to x. Define $f : E_1 \sqcup E_2 \to E_1$ to be the homeomorphism where the restriction of f to $\bigcup_{i=1}^{\infty} U_i \cup \{x\}$ is constructed as in Observation 4.9, and the restriction of f to the rest of E is the identity. □

The second case is covered by the following:

Lemma 4.13 *Suppose Σ has no nondisplaceable subsurfaces and $\mathcal{M}$ is a Cantor set of points all of the same type. Then E is self-similar.*

Proof Let $E = E_1 \sqcup E_2$ be a decomposition of E into clopen sets. If $\mathcal{M}$ is contained in only one of the E_i, then one may apply the argument from Lemma 4.12, by letting x be any point of $\mathcal{M}$. Thus, we assume that both E_1 and E_2 contain points of $\mathcal{M}$.

For concreteness, fix a metric on E. For each $n \in N$, fix a decomposition $A_1^{(n)}, \ldots, A_{j_n}^{(n)}$ of E into clopen sets of diameter at most 2^{-n}, such that E_1 and E_2 are each the union of some number of these sets. Let S_n be a subsurface homeomorphic to a j_n–holed sphere, with complementary regions containing the sets $A_k^{(n)}$. Since S_n is displaceable, there exists some k such that $A_k^{(n)}$ contains a copy of all but one of the sets $A_j^{(n)}$; in particular, it contains a copy of either E_1 or E_2. Passing to a subsequence, we conclude that for either $i = 1$ or $i = 2$ there exist homeomorphic copies of E_i of diameter less than 2^{-n}, for each n. Without loss of generality, say that this holds for E_1. Passing to a further subsequence, we can assume these copies of E_1 Hausdorff converge to a point x, so in particular every neighborhood of x contains a copy of E_1.

It follows from the definition of $\preccurlyeq$ that each $y \in \mathcal{M}$ therefore also has this property: every neighborhood of y contains a homeomorphic copy of E_1. Let $y_2, y_3, y_4, \ldots$ be a sequence of points in E_2 converging to $y \in E_2$, and let $U_1 = E_1$. Fix disjoint neighborhoods N_i of y_i converging to y, and let U_i be a homeomorphic copy of E_1 in N_i. Now apply Observation 4.9. $\qquad\square$

This completes the proof of Proposition 4.8.

4.2 Stable neighborhoods

Motivated by the behavior of maximal points in the proposition above, we make the following definition:

Definition 4.14 For $x \in E$, call a neighborhood U of x *stable* if for any smaller neighborhood $U' \subset U$ of x, there is a homeomorphic copy of U contained in U'.

Our use of the terminology "stable" is justified by Lemma 4.17 below, which says that all such neighborhoods of a point are homeomorphic. (Recall that, by convention, neighborhood always means clopen neighborhood.)

Remark 4.15 Stable neighborhoods are automatically self-similar sets, and if U is a stable neighborhood of x, then $x \in \mathcal{M}(U)$. Our work in the previous section shows that when $\prec$ has a unique maximal type and all subsurfaces are displaceable, each maximal point has a stable neighborhood.

It follows immediately from the definition that if x has one stable neighborhood, then *every* sufficiently small neighborhood of x is also stable. More generally, we have the following.

Lemma 4.16 *If x has a stable neighborhood, and $y \sim x$, then y has a stable neighborhood.*

Proof Let U be a stable neighborhood of x. Since $y \prec x$, there is a neighborhood V of y such that U contains a homeomorphic copy of V. Suppose $V' \subset V$ is a smaller neighborhood of y. Since $x \prec y$, there is some neighborhood U' of x (without loss of generality, we may assume that $U' \subset U$) such that V' contains a homeomorphic copy of U'. By definition of stable neighborhoods, U' contains a homeomorphic copy of U, thus V' contains a homeomorphic copy of U and hence a homeomorphic copy of V. □

Lemma 4.17 *If x has a stable neighborhood U, then for any $y \sim x$, all sufficiently small neighborhoods of y are homeomorphic to U via a homeomorphism taking x to y.*

Proof The proof is a standard back-and-forth argument. Suppose $x \prec y$ and $y \prec x$. Let V_x be a stable neighborhood of x and V_y a stable neighborhood of y. Take a neighborhood basis $V_x = V_0 \supset V_1 \supset V_2 \supset \cdots$ of x consisting of nested neighborhoods, and take a neighborhood basis $V_y = V_0' \supset V_1' \supset V_2' \supset \cdots$ of y. Since $y \prec x$ and $x \prec y$, each V_i contains a homeomorphic copy of V_0' and each V_i' a copy of V_0.

Let f_1 be a homeomorphism from $V_0 - V_1$ into V_0'. Note that we may assume the image of f_1 avoids y: if y is the unique maximal point of V_0', then this is automatic, otherwise, $E(y)$ is a Cantor set of points, each of which contains copies of V_0 in every small neighborhood. Let g_1 be a homeomorphism from the complement of the image of f_1 in $V_0' - V_1'$ onto a subset of $V_1 - \{x\}$. Iteratively, define f_i to be a homeomorphism from the complement of the image of g_{i-1} in $V_{i-1} - V_i$ onto a subset of $V_{i-1}' - \{y\}$, and g_i a homeomorphism from the complement of the image of f_i in $V_{i-1}' - V_i'$ onto a subset of $V_i - \{x\}$. Then the union of all f_i and g_i^{-1} is a homeomorphism from $V_0 - \{x\}$ to $V_0' - \{y\}$ that extends to a homeomorphism from V_0 to V_0' taking x to y. □

The following variation on Lemma 4.11 uses stable neighborhoods as a replacement for displaceable subsurfaces.

Lemma 4.18 *Let $x, y \in E$, and assume x has a stable neighborhood V_x and that x is an accumulation point of $E(y)$. Then for any sufficiently small neighborhood U of y, $U \cup V_x$ is homeomorphic to V_x.*

Proof If $x \sim y$, then let U be a stable neighborhood of y disjoint from V_x. Let $V_1 \supset V_2 \supset V_3 \supset \cdots$ be a neighborhood basis for x consisting of stable neighborhoods. Since x is an accumulation point of $E(y)$, for any sufficiently small neighborhood U_0 of y (and hence for any stable neighborhood U), there is a homeomorphic copy U_1 of U_0 in $V_1 - \{x\}$. Shrinking neighborhoods if needed, we may take U_1 to be disjoint from V_{i_1} for some $i_1 \in \mathbb{N}$. Since V_{i_1} is homeomorphic to V_1, there is also a homeomorphic copy of U_2 of U_0 in V_{i_1}, disjoint from some V_{i_2}. Iterating this process we can find disjoint sets $U_n \subset V_1$, each homeomorphic to U, and Hausdorff converging to x. Define $f : V_1 \cup U_0 \to V_1$ to be the identity on the complement of $\bigcup_n U_n$ and send U_i to U_{i+1} by a homeomorphism as in Observation 4.9.

If instead $y \prec x$, then take any neighborhood U of y disjoint from V_x and small enough that V_x contains a homeomorphic copy of U. Since $y \prec x$, this copy lies in $V_x - \{x\}$, and we may repeat the same line of argument above. □

5 Classification of locally CB mapping class groups

We now prove properties of locally CB mapping class groups, building towards our general classification theorem. Recall that we have the following notational convention.

Notation 5.1 If $K \subset \Sigma$ is a finite-type subsurface, we denote by $\mathcal{V}_K$ the identity neighborhood consisting of mapping classes of homeomorphisms that restrict to the identity on K.

Lemma 5.2 *Let $K \subset \Sigma$ be a finite-type subsurface such that each component of $\Sigma - K$ has infinite type. If there exists a finite-type, nondisplaceable (possibly disconnected) subsurface S in $\Sigma - K$, then $\mathcal{V}_K$ is not CB. If this holds for every such finite-type $K \subset \Sigma$, then $\mathrm{Map}(\Sigma)$ is not locally CB.*

Proof Let K be a surface as in the statement of the proposition, with a nondisplaceable subsurface $S \subset \Sigma - K$. Since each complementary region to K was assumed to have infinite type, by enlarging S if needed we may assume that S still remains in the complement of K, but is such that each component of S has high enough complexity

that the length function ℓ_S defined in Section 2 will be unbounded. As in Proposition 2.8, this gives a length function which is unbounded on $\mathcal{V}_K$, hence on $\mathcal{V}$, so $\mathrm{Map}(\Sigma)$ is not locally CB.

As remarked above, the sets $\mathcal{V}_L$, where L ranges over finite-type subsurfaces, form a neighborhood basis of the identity in $\mathrm{Map}(\Sigma)$. But one may in fact restrict this to range over finite-type surfaces whose complementary regions are all of infinite type, since if L is finite type, then the union of L and its finite-type complementary regions is again a compact surface, say K, and $\mathcal{V}_K \subset \mathcal{V}_L$. Thus, $\mathrm{Map}(\Sigma)$ is locally CB if and only if some such set $\mathcal{V}_K$ is CB.   □

Going forward, we reference the partial order $\prec$ defined in Section 4.

Lemma 5.3 *If* $\mathrm{Map}(\Sigma)$ *is locally CB, then the number of distinct maximal types under $\prec$ is finite.*

Proof We prove the contrapositive. Suppose that there are infinitely many distinct maximal types. Let K be any subsurface of finite type. By Lemma 5.2, it suffices to find a nondisplaceable subsurface contained in $\Sigma - K$, which we do now.

To every end $x \in E$ of maximal type, let $\sigma(x)$ denote the set of connected components of $\Sigma - K$ which contain ends from $E(x)$. Since $\Sigma - K$ has finitely many connected components, by the pigeonhole principle there are two ends x and y with $x \sim y$ but $\sigma(x) = \sigma(y)$. That is, each complementary region of $\Sigma - K$ that has an end from $E(x)$ also contains ends from $E(y)$, and vice versa. Fix any $z \in E$ with $z \nsim x$ and $z \nsim y$.

Construct a surface S as follows. For each component τ of $\sigma(x)$, take a three-holed sphere subsurface contained in τ so that the complementary regions of the three-holed sphere separate $E(x)$ from $E(y)$ and $E(z)$ in τ. That is to say, one complementary region contains only ends from $E(x)$ and none from $E(y)$ or $E(z)$, while another contains only ends from $E(y)$ and none from $E(x)$ or $E(z)$, and the third contains at least some points of $E(z)$ (possibly those from another complementary region of K). Let S be the union of these three holed spheres, one in each component of $\sigma(x)$. Thus, each end from $E(x)$ is the end of some complementary region of S which has no ends of type y, and vice versa.

We claim that S is nondisplaceable. For if S_i is a connected component of S, then one complementary region of S_i contains ends from $E(x)$, but none from $E(y)$. By

invariance of $E(x)$ and $E(y)$, if some homeomorphic image $f(S_i)$ were disjoint from S, then we would have to have $f(S_i)$ contained in one of the complementary regions of S containing points of $E(x)$. However, this region contains no points of $E(y)$ or $E(z)$, contradicting our construction of S_i. Hence, S is nondisplaceable and, by Lemma 5.2, $\mathrm{Map}(\Sigma)$ is not locally CB. □

We now state the first structure theorem for end spaces of surfaces with locally CB mapping class groups.

Proposition 5.4 *If* $\mathrm{Map}(\Sigma)$ *is locally CB, then there is a partition*

$$E = \bigsqcup_{A \in \mathcal{A}} A,$$

where $\mathcal{A}$ *is finite, each* $A \in \mathcal{A}$ *is clopen and self-similar, and* $\mathcal{M}(A) \subset \mathcal{M}(E)$. *Moreover, this decomposition can be realized by the complementary regions to a finite-type surface* $L \subset \Sigma$ *with* $|\mathcal{A}|$ *boundary components, either of zero genus or of finite genus equal to the genus of* Σ.

This will be a quick consequence of the following stronger result:

Proposition 5.5 *Suppose that* $\mathrm{Map}(\Sigma)$ *is locally CB. Then there exists a CB neighborhood* $\mathcal{V}_K$ *of the identity, where* K *is a finite-type surface with the following properties:*

(i) *Each connected component of* $\Sigma - K$ *has one or infinitely many ends and zero or infinite genus.*

(ii) *The connected components of* $\Sigma - K$ *partition* E *as*

$$E = \bigsqcup_{\widehat{A} \in \mathcal{A}} \widehat{A} \sqcup \bigsqcup_{P \in \mathcal{P}} P,$$

where each $\widehat{A} \in \mathcal{A}$ *is self-similar, and for each* $P \in \mathcal{P}$, *there exists some* $\widehat{A} \in \mathcal{A}$ *such that* P *is homeomorphic to a clopen subset of* $\widehat{A}$.

(iii) *For all* $\widehat{A} \in \mathcal{A}$, *the maximal points* $\mathcal{M}(\widehat{A})$ *are maximal in* E, *and* $\mathcal{M}(E) = \bigsqcup_{\widehat{A} \in \mathcal{A}} \mathcal{M}(\widehat{A})$.

Our choice of $\mathcal{A}$ as the notation for the index set in both propositions is because they may be canonically identified. In fact, the proof of Proposition 5.4 consists of showing that each of the sets A is a union of one set $\widehat{A}$ from Proposition 5.5 and some number of the sets in $\mathcal{P}$, and that A is homeomorphic to $\widehat{A}$.

Proof of Proposition 5.5   Suppose that $\mathcal{V}$ is a CB neighborhood of the identity in $\mathrm{Map}(\Sigma)$. Let K be a finite-type surface such that $\mathcal{V}_K \subset \mathcal{V}$, so $\mathcal{V}_K$ is also CB. Enlarging K if needed (and hence shrinking $\mathcal{V}_K$), we may assume that each complementary region to K has either zero or infinite genus. Since $\prec$ has only finitely many maximal types, enlarging K further, we may assume that its complementary regions separate the different maximal types, and moreover, if for some maximal x the set $E(x)$ is finite, then all the ends from $E(x)$ are separated by K. Thus, complementary regions to K have either no end from $\mathcal{M}(E)$, a single end from $\mathcal{M}(E)$ or a Cantor set of ends of a single type from $\mathcal{M}(E)$.

Our goal is to show that the complementary regions containing ends from $\mathcal{M}(E)$ are all self-similar sets, and the end sets of the remaining regions have the property desired of the sets $P \in \mathcal{P}$ described above. It will be convenient to introduce some terminology for the set of ends of a complementary region to K, so call such a subset of E a *complementary end set*.

For simplicity, assume as a first case that for each maximal type x, the set $E(x)$ is *finite*. Fix a maximal type point $x \in E$, and let $B_1, B_2, \ldots, B_k \subset E$ be the complementary end sets whose maximal points lie in $E(x)$. We start by showing that at least one of the sets B_i is self-similar. Let x_i denote the maximal point in B_i. Let U_i be any clopen neighborhood of x_i in B_i. Since $x_i \in E(x)$, we may find smaller neighborhoods $V_i \subset U_i$ such that each U_i contains a homeomorphic copy of V_j for all $j = 1, 2, \ldots, k$. Let $S \subset \Sigma - K$ be a subsurface, homeomorphic to the disjoint union of k pairs of pants, such that the complementary regions of the ith pair of pants partitions the ends of Σ into V_i, $B_i - V_i$ and $E - B_i$.

Since $\mathcal{V}_K$ is assumed CB, the surface S is displaceable by Lemma 5.2, so at least one of the connected components of S can be moved to be disjoint from S by a homeomorphism. Since $E(x)$ is homeomorphism invariant, we conclude that there is a copy of B_j in some V_i, possibly with $i \neq j$. Our choice of V_i now implies that there is in fact a homeomorphic copy of B_j in U_j. Thus, we have shown that, for any neighborhoods U_i of x_i, there exists j such that U_j contains a copy of B_j. Applying this conclusion to each of a nested sequence of neighborhoods of the x_i which give a neighborhood basis, we conclude that some j must satisfy this conclusion infinitely often (ie has a homeomorphic copy contained in every neighborhood of x_j), giving us some B_j which is self-similar.

Since x_i are the unique maximal points of B_i, this implies that each x_i has a neighborhood M_i homeomorphic to B_j, ie a self-similar set. Repeating this process for all

of the distinct maximal types, we conclude that each maximal point has a self-similar neighborhood. Fix a collection of such neighborhoods. Since this collection is finite we may enumerate them $A_1, A_2, \ldots, A_n$.

For each nonmaximal point y, Lemma 4.18 implies that there exists a neighborhood P_y of y such that $P_y \cup A_i$ is homeomorphic to some A_i, a neighborhood of a maximal point that is a successor (though not necessarily an immediate successor) of y. Since $E - \bigsqcup_i A_i$ is compact, finitely many such neighborhoods P_y cover it. Enlarging K, we may assume that it partitions the end sets into the disjoint union of such sets of the form P_y and A_i. This concludes the proof in the case where $\mathcal{M}$ is finite.

Now we treat the general case where, for some maximal types, the set $E(x)$ is a Cantor set. The strategy is essentially the same. We use the following lemma, which parallels the argument just given above.

Lemma 5.6 *Keeping the hypotheses of the proposition, let x be a maximal type with $E(x)$ a Cantor set. Then x has a neighborhood which is self-similar.*

Proof Let $A_1, \ldots, A_k$ be the complementary end sets which contain points of $E(x)$, and fix a maximal end x_i in each A_i. As before, we start by showing that, for some j, every neighborhood of x_j contains a homeomorphic copy of A_j, so in particular A_j is self-similar. Let U_i be a neighborhood of x_i. For each $z \in E(x)$, let V_z be a neighborhood of z such that each of the sets U_i contains a homeomorphic copy of V_z. Since $E(x)$ is compact, finitely many such V_z cover $E(x)$, so from now on we consider only a finite subcollection that covers. Let $S \subset \Sigma - K$ be a subsurface homeomorphic to the union of k disjoint n–holed spheres, where n is chosen large enough that each complementary region of S has its set of ends either contained in one of the finitely many V_z, or containing all but one of the sets A_i.

Again, since $E(x)$ is invariant, and S is displaceable, this means that there is some V_z and some A_j such that V_z contains a homeomorphic copy of A_j. Thus, by definition of V_z, we have that U_j contains a homeomorphic copy of A_j. Repeating this for a nested sequence of neighborhoods of the x_i, we conclude that some x_j satisfies this infinitely often. This means that A_j is a stable neighborhood of x_j, hence by Lemma 4.17, each point of $E(x)$ has a stable neighborhood, which is necessarily a self-similar set. $\qquad\square$

Now we can finish the proof as in the case where all $E(x)$ are finite, by fixing a finite cover of $\bigcup_{x \in \mathcal{M}(E)} E(x)$ by stable neighborhoods, and using Lemma 4.18 as before. $\quad\square$

Proof of Proposition 5.4 Let $E = \bigsqcup_{\widehat{A} \in \mathcal{A}} \widehat{A} \sqcup \bigsqcup_{P \in \mathcal{P}} P$ be the decomposition given by Proposition 5.5. By construction of the sets P and Lemma 4.18, for each $P \in \mathcal{P}$, there exists $\widehat{A} \in \mathcal{A}$ such that $P \sqcup \widehat{A} \cong \widehat{A}$. Applying this to each P iteratively, we conclude that E is homeomorphic to the disjoint union $\bigsqcup_{A \in \mathcal{A}} \widehat{A}$. Relabeling $\widehat{A}$ as A gives the desired result, and we may take L to be a subset of K.   □

With this groundwork in place, we can prove Theorem 1.4. We restate it in slightly different form, for convenience.

Theorem 5.7 *Map*(Σ) *is locally CB if and only if there is a finite-type surface K such that the complementary regions of K each have one or infinitely many ends and zero or infinite genus, and partition of E into finitely many clopen sets*

$$E = \left(\bigsqcup_{\widehat{A} \in \mathcal{A}} \widehat{A} \right) \sqcup \left(\bigsqcup_{P \in \mathcal{P}} P \right)$$

with the following properties:

(i) *Each $\widehat{A} \in \mathcal{A}$ is self-similar, $\mathcal{M}(\widehat{A}) \subset \mathcal{M}(E)$ and $\mathcal{M}(E) = \bigsqcup_{\widehat{A} \in \mathcal{A}} \mathcal{M}(\widehat{A})$.*

(ii) *Each $P \in \mathcal{P}$ is homeomorphic to a clopen subset of some $\widehat{A} \in \mathcal{A}$.*

(iii) *For any $x_A \in \mathcal{M}(\widehat{A})$ and any neighborhood V of the end x_A in Σ, there is $f_V \in \mathrm{Homeo}(\Sigma)$ such that $f_V(V)$ contains the complementary region to K with end set $\widehat{A}$.*

Moreover, in this case $\mathcal{V}_K := \{g \in \mathrm{Homeo}(\Sigma) : g|_K = \mathrm{id}\}$ is a CB neighborhood of the identity, and K may always be taken to have genus zero if Σ has infinite genus, and genus equal to that of Σ otherwise, and if the number of isolated planar ends of Σ is finite, we may additionally take all of these ends to be punctures of K.

Note that the case where $K = \varnothing$ implies that Σ has zero or infinite genus and self-similar end space, in which case we already showed that $\mathcal{V}_\varnothing = \mathrm{Map}(\Sigma)$ is CB. In this case, conditions (ii) and (iii) are vacuously satisfied. The reader may find it helpful to refer to Figure 1 for some very basic examples, all with $\mathcal{P} = \varnothing$, and keep this in mind during the proof.

Proof of Theorem 5.7   ($\Longrightarrow$)   The forward direction is obtained by a minor improvement of Proposition 5.5. Assume Map(Σ) is locally CB. Let $K \subset \Sigma$ be a finite-type surface with $\mathcal{V}_K$ a CB neighborhood of the identity and the properties given

in Proposition 5.5. We may enlarge K if needed so that each of its boundary curves are separating, and so that whenever some maximal type x has $E(x)$ homeomorphic to a Cantor set, then $E(x)$ is contained in at least two complementary regions to K. This latter step can be done as follows: if $\hat{A}$ is the unique complementary region of K containing the Cantor set $E(x)$, then glue a strip to K that separates $\hat{A}$ into two clopen sets, each containing points of $E(x)$. Since $\hat{A}$ is self-similar, each point of $\mathcal{M}(\hat{A})$ has a stable neighborhood by Lemma 4.16, and so the two clopen sets of our partition are again each self-similar and each homeomorphic to $\hat{A}$. Enlarging K further if needed, we may assume it also contains all isolated punctures if this number is finite.

Thus, we assume K now has these properties, and let $E = \left(\bigsqcup_{\hat{A} \in \mathcal{A}} \hat{A} \right) \sqcup \left(\bigsqcup_{P \in \mathcal{P}} P \right)$ be the resulting decomposition of E, with $\Sigma_{\hat{A}}$ and Σ_P denoting the connected component of K with end space $\hat{A}$ or P, respectively. We need to establish that the third condition holds. Fix $\hat{A}$, let $x_A \in \mathcal{M}(\hat{A})$, let $V \subset \Sigma$ be a neighborhood of the end x_A, and let $E(V) \subset \hat{A}$ denote the end space of V. We may without loss of generality assume that V has a single boundary component. Recall that our goal is to show that the pair $V, (\Sigma - V)$ is homeomorphic to the pair $\Sigma_{\hat{A}}, (\Sigma - \Sigma_{\hat{A}})$.

First consider the case where $|E(x_A)| > 1$. By construction there exists $\hat{B} \neq \hat{A} \in \mathcal{A}$ with $E(x) \cap \hat{B} \neq \varnothing$. Since points of $E(x_A)$ have stable neighborhoods, $\hat{B} \cup (\hat{A} - E(V))$ is homeomorphic to $\hat{B}$. Moreover, if $\Sigma_{\hat{A}}$ has infinite genus, then V and $\Sigma - \Sigma_{\hat{A}}$ and $\Sigma - V$ all do as well, while if $\Sigma_{\hat{A}}$ has genus 0, then so does V, and both complementary regions are of the same genus as well (equal to the genus of Σ). Thus, by the classification of surfaces, the pair $V, \Sigma - V$ is homeomorphic to $\Sigma_{\hat{A}}, \Sigma - \Sigma_{\hat{A}}$ and so there is some $f_V \in \text{Map}(\Sigma)$ taking V to $\Sigma_{\hat{A}}$. This is what we needed to show.

Now suppose instead $|E(x_A)| = 1$. Here we will use the displaceable subsurfaces condition to find the desired f_V. Let S be a pair of pants in the complement of K, with one boundary component equal to ∂V and another homotopic to $\partial \Sigma_{\hat{A}}$. Since $S \subset (\Sigma - K)$, it is displaceable, so let f be a homeomorphism displacing S. Since $E(x_A) = x_A$ is an invariant set, up to replacing f with its inverse, we have $f(S) \subset V$. If, as a first case, there exists a maximal end $y \sim x$, then $E(y)$ is also an invariant set. Thus, $f(\Sigma_{\hat{A}}) \subset V$, hence we may take $f_V = f^{-1}$ and have $f_V(V) \supset \hat{A}$.

If, as a second case, Σ has finite genus, then $f(\Sigma - \Sigma_A)$ necessarily contains all the genus of Σ, hence again we have $f(\Sigma_{\hat{A}}) \subset V$. Finally, if neither of these two cases holds, then Σ has infinite or zero genus, and a unique maximal end, so $|\mathcal{A}| = 1$ and E is self-similar. Thus, $\text{Map}(\Sigma)$ is CB by Proposition 3.1.

($\Longleftarrow$) For the converse, the case where $K = \varnothing$, we have that Σ has zero or infinite genus and a self-similar end space is covered by Proposition 3.1.

So suppose Σ is not zero or infinite genus with a self-similar end space, but instead we have a finite-type surface K with the properties listed. We wish to show that $\mathcal{V}_K$ is CB. Let $T \subset \Sigma$ be a finite-type surface with $\mathcal{V}_T \subset \mathcal{V}_K$, ie $T \supset K$. We need to find a finite set F and some n such that $(F\mathcal{V}_T)^n$ contains $\mathcal{V}_K$.

For each $\widehat{A} \in \mathcal{A}$, fix $x_A \in \mathcal{M}(\widehat{A})$ and let V_A be the connected component of T containing x_A. Let f_V be the homeomorphism provided by our assumption. Also, for each $P \in \mathcal{P}$, choose a homeomorphism f_P of Σ that exchanges P with a clopen subset of some $\widehat{A} \in \mathcal{A}$ which is homeomorphic to P. Let F be the set of all such $f_V^{\pm 1}$ and $f_P^{\pm 1}$.

Now suppose $g \in \mathcal{V}_K$. We can write g as a product of $|\mathcal{A}| + |\mathcal{P}|$ homeomorphisms, where each one is supported on a surface of the form Σ_A for $\widehat{A} \in \mathcal{A}$ or Σ_P for $P \in \mathcal{P}$ (adopting our notation from the previous direction of the proof).

If some such homeomorphism g_A is supported on Σ_A, then $f_V^{-1} g_A f_V$ restricts to the identity on T, so $g_A \in F\mathcal{V}_T F$. For a homeomorphism g_P supported on Σ_P, we have that $f_P^{-1} g_P f_P$ is supported in Σ_A, so $g_P \in F^2 \mathcal{V}_T F^2$. This shows that $g \in (F^2 \mathcal{V}_T F^2)^{|\mathcal{A}| + |\mathcal{P}|}$, which is what we needed to show. $\square$

5.1 Examples

While the statement of Theorem 5.7 is somewhat involved, it is practical to apply in specific situations. Below are a few examples illustrating some of the subtlety of the phenomena at play. The first is an immediate consequence:

Corollary 5.8 *If Σ has finite nonzero genus and countable self-similar end space, then Σ is not locally CB.*

As another example, one could take Σ to have finite nonzero genus, and end space equal to the union of cantor set and a countable set of isolated points, accumulating on the Cantor set at exactly one point. Many other variations are possible. As a more involved example, we have the following:

Corollary 5.9 *Suppose that Σ has finite nonzero genus and self-similar end space, with a single maximal end x, but infinitely many distinct immediate precursors to x. Then $\mathrm{Map}(\Sigma)$ is not locally CB.*
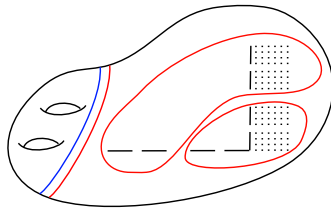
Figure 4: The subsurface with red boundary defines a CB neighborhood, while the smaller subsurface with blue boundary does not.

As a concrete example, one could construct E by taking countably many copies of a Cantor set indexed by $\mathbb{N}$, all sharing a single point in common and Hausdorff converging to that point, with the nth copy accumulated everywhere by points locally homeomorphic to $\omega^n + 1$.

Proof If $\mathrm{Map}(\Sigma)$ were locally CB, then we would have a finite-type surface K as in Theorem 5.7. Since $\mathcal{M}$ is a singleton, $\mathcal{A} = \{\widehat{A}\}$, $x_A = x$ and Σ_A is some neighborhood of the end x. However, by construction, $E - \widehat{A}$ contains ends of only finitely many types of immediate precursors. Thus, we may choose a smaller neighborhood V of x so that $\Sigma - V$ has more distinct types of ends. Then $\Sigma - V$ cannot possibly be homeomorphic to $\Sigma - \Sigma_A$, so no such f_V exists. $\qquad\square$

By contrast, if Σ is finite genus with end space equal to a Cantor set, or attained by the construction in Corollary 5.9 but replacing $\mathbb{N}$ with a finite number, then $\mathrm{Map}(\Sigma)$ is locally CB. We draw attention to a specific case of this to highlight the role played by $\mathcal{A}$ and $\mathcal{P}$.

Example 5.10 Let Σ be a surface of finite nonzero genus g, with E homeomorphic to the union of a Cantor set C and a Cantor set D, and a countable set Q, with $C \cap D = \{x\}$ and the accumulation points of Q equal to D, as illustrated in Figure 4. Then by Theorem 5.7, a CB neighborhood of the identity in $\mathrm{Map}(\Sigma)$ can be taken to be $\mathcal{V}_K$ where K is a finite-type subsurface of genus g with two boundary components, with one complementary region to K having x as an end, and the other containing points of both C and D. In this case, $\mathcal{A}$ and $\mathcal{P}$ are both singletons, with one complementary region in each.

The set E itself is self-similar, and the decomposition into self-similar sets given by Proposition 5.4 is trivial. However, if K' is a finite-type subsurface realizing this

decomposition (with a single complementary region), then $\mathcal{V}_{K'}$ is not a CB set. Indeed, one may find a nondisplaceable subsurface in the complement homeomorphic to a three-holed sphere, where one complementary region has x as an end, one contains all the genus of Σ but no ends, and the third contains points of C, for example.

6   CB generated mapping class groups

In this section we give general criteria for when mapping class groups are CB generated, building towards the proof of Theorem 1.6.

6.1   Two criteria for CB generation

Notation 6.1   For a subset $X \subset E$, we say a family of neighborhoods U_n in E *descends to* X if U_n are nested, meaning $U_{n+1} \subseteq U_n$, and if $\bigcap_{n \in \mathbb{N}} U_n = X$. As a shorthand, we write $U_n \searrow X$. If $X = \{x\}$ is a singleton, we abuse notation slightly and write $U_n \searrow x$ and say U_n descends to x.

Definition 6.2   (limit type)   We say that an end set E is *limit type* if there is a finite-index subgroup G of $\mathrm{Map}(\Sigma)$, a G–invariant set $X \subset E$, points $z_n \in E$, indexed by $n \in \mathbb{N}$ which are pairwise inequivalent, and a family of neighborhoods $U_n \searrow X$ such that

$$E(z_n) \cap U_n \neq \varnothing, \quad E(z_n) \cap U_0^c \neq \varnothing, \quad E(z_n) \subset (U_n \cup U_0^c).$$

Here $U_0^c = E - U_0$ denotes the complement of U_0 in E.

The following example explains our choice of the terminology "limit type":

Example 6.3   Suppose that α is a countable limit ordinal, and $E \cong \omega^\alpha \cdot n + 1$, with $n \geq 2$ and $E^G = \varnothing$. To see that this end space is limit type, take G to be the finite-index subgroup pointwise fixing the n maximal ends. Fix a maximal end x and a clopen neighborhood U_0 of x disjoint from the other maximal ends, and let $U_n \subset U_0$ be nested clopen sets forming a neighborhood basis of x. Since $U_n - U_{n+1}$ is closed, there is a maximal ordinal β_n such that U_n contains points locally homeomorphic to $\omega^{\beta_n} + 1$. Passing to a subsequence we may assume that all of these are distinct, and one may choose $z_n \in U_n$ to be a point locally homeomorphic to $\omega^{\beta_n} + 1$. Note that necessarily the sequence β_n converges to the limit ordinal α. The assumption that $n \geq 2$ ensures that the sets $E(z_n)$ contain points outside of U_0, and we require this in the definition to ensure that E is not self-similar.

Lemma 6.4 (limit-type criterion) *If an end set E has limit type, then $\mathrm{Map}(\Sigma)$ is not CB generated.*

Proof Let G, X, U_n and z_n be as in the definition of limit type. We will show G is not CB generated. Since G is finite index, this is enough to show that $\mathrm{Map}(\Sigma)$ is not CB generated. Furthermore, since $\mathrm{Map}(\Sigma)$ (and hence G) is assumed to be locally CB, it suffices to show that there is some neighborhood $\mathcal{V}_G$ of the identity in G such that for any finite set F, the set $F\mathcal{V}_G$ does not generate G.

Let $\mathcal{V}_G$ be a neighborhood of the identity in G, chosen small enough that, for every $g \in \mathcal{V}_G$ and all $n > 0$, we have $g(U_n) \subset U_0$ and $g(U_0^c) \cap U_n = \varnothing$.

Let F be any finite subset of G. Since G preserves both the set X and the set $E(z_n) \subset U_n \sqcup U_0^c$, there exists $N \in \mathbb{N}$ such that for all $n > N$ and all $f \in F$, we have

$$f(E(z_n) \cap U_n) \subset U_n.$$

The same holds for elements of $\mathcal{V}_G$.

Fix such an $n > N$, and let $x_n \in E(z_n) \cap U_n$ and $y_n \in E(z_n) \cap U_0^c$. Since $x_n \sim y_n$, there is a homeomorphism h with $h(x_n)$ lying in a small neighborhood of y_n contained in U_0^c. By our observation above, h is not in the subgroup generated by $F\mathcal{V}_G$, which shows that G is not CB generated, as desired. $\qquad\square$

A second obstruction to CB generation is the following "rank" condition:

Definition 6.5 (infinite rank) We say $\mathrm{Map}(\Sigma)$ has *infinite rank* if there is a finite-index subgroup G of $\mathrm{Map}(\Sigma)$, a closed G–invariant set X, neighborhood U of X and points z_n, for $n \in \mathbb{N}$, each with a *stable neighborhood* (see Definition 4.14) such that

- $z_n \notin E(z_m)$ if $m \neq n$,
- for all n, $E(z_n)$ is countably infinite and has at least one accumulation point in both X and in $E - U$, and
- the set of accumulation points of $E(z_n)$ in U is a subset of X.

If the above does not hold, we say instead that $\mathrm{Map}(\Sigma)$ has *finite rank*.

Example 6.6 A simple example of such a set is as follows. Let C_n be the union of a countable set and a Cantor set, with Cantor–Bendixson rank n, and n^{th} derived set equal to the Cantor set. For each C_n, select a single point z_n of the Cantor set to be

an end accumulated by genus. Now create an end space E by taking $\mathbb{N}$ copies of each C_n, arranged so that they have exactly two accumulation points x and y (and these accumulation points are independent of n). Then $X = \{x\}$ and the points z_n satisfy the definition.

Examples of surfaces with *countable* end spaces and infinite-rank mapping class groups are much more involved to describe. (Note that these necessarily must have infinite genus.) It would be nice to see a general procedure for producing families of examples.

Lemma 6.7 (infinite-rank criterion) *If* $\mathrm{Map}(\Sigma)$ *has infinite rank, then it is not CB generated.*

Proof Let G, X, U and z_n be as the definition of infinite rank. For every z_n, we define a function $\ell_n \colon G \to \mathbb{Z}$ as follows. For $\phi \in G$, define

$$\ell_n(\phi) = |(E(z_n) \cap U) - \phi^{-1}(U)| - |(E(z_n) \cap \phi^{-1}(U)) - U|.$$

That is, $\ell_n(\phi)$ is the difference between the number of points in $E(z_n)$ that ϕ maps out of U, and the number of points in $E(z_n)$ that ϕ maps into U.

Since X is G–invariant and contains all of the accumulation points of $E(z_n)$ in U, the value of ℓ_n is always finite. It is also easily verified that ℓ_n is a homomorphism. Moreover, as each z_n has a stable neighborhood (all of which are pairwise homeomorphic), for any finite collection $n_1, \ldots, n_k$ one may construct, for each i, a "shift" homomorphism ϕ_i supported on a union of disjoint stable neighborhoods of $E(z_{n_i})$, taking one stable neighborhood to the next, so that $\ell_{n_i}(\phi_i) = 1$ and $\ell_{n_j}(\phi_j) = 0$ for $j \neq i$. Finally, ℓ_n is continuous; in fact for any neighborhood $\mathcal{V}$ of the identity in G which is small enough that elements of $\mathcal{V}$ fix the isotopy class of a curve separating U from $E - U$, we will have $\ell_n(\mathcal{V}) = 0$.

Thus, we have for each $k \in \mathbb{N}$ a surjective, continuous homomorphism

$$(\ell_{n_1}, \ldots, \ell_{n_k}) \colon G \to \mathbb{Z}^k,$$

which restricts to the trivial homomorphism on the neighborhood $\mathcal{V}$ of the identity described above.

By Theorem 2.1, any CB set is contained in a set of the form $(F\mathcal{V})^k$ for some finite set F and $k \in \mathbb{N}$. Given any such F, choose $j > |F|$. Then restriction of $(\ell_{n_1}, \ldots, \ell_{n_j})$ to the subgroup generated by $(F\mathcal{V})^k$ cannot be surjective, as $\mathcal{V}$ lies in its kernel. It follows that no CB set can generate G. Since G is finite index in $\mathrm{Map}(\Sigma)$, the same is also true for $\mathrm{Map}(\Sigma)$. $\qquad\square$

6.2 End spaces of locally CB mapping class groups

For the remainder of this section, we assume that $\mathrm{Map}(\Sigma)$ is locally CB, our ultimate goal being to understand which such groups are CB generated. Recall that Proposition 5.4 gave a decomposition of E into a disjoint union of self-similar sets homeomorphic to $A \in \mathcal{A}$, realized by a finite-type subsurface $L \subset K$. However, as shown in Example 5.10, the neighborhood $\mathcal{V}_L$ might not be CB. We now show that $\mathcal{V}_L$ is CB generated.

Lemma 6.8 *Assume that* $\mathrm{Map}(\Sigma)$ *is locally CB. Let* L *be a finite-type surface whose complementary regions realize the decomposition* $E = \bigsqcup_{A \in \mathcal{A}} A$ *given by* Proposition 5.4. *Then* $\mathcal{V}_L$ *is CB generated.*

Furthermore, we may take L *to have genus zero if* Σ *has infinite genus, and genus equal to that of* Σ *otherwise; and a number of punctures equal to the number of isolated planar (not accumulated by genus) ends of* Σ *if that number is finite, and zero otherwise.*

For the proof, we need the following observation, which follows from well-known results on standard generators for mapping class groups of finite-type surfaces.

Observation 6.9 *Let* Σ *be an infinite-type surface, possibly with finitely many boundary components, and* $S \subset \Sigma$ *a finite-type subsurface. Then there is a finite set of Dehn twists* D *such that for any finite-type surface* S', $\mathrm{Map}(S')$ *is generated by* D *and* $\mathcal{V}_S$.

In fact, akin to Lickorish's Dehn twist generators for the mapping class group of a surface of finite type, one can find a set $\mathcal{D}$ of simple closed curves in Σ such that every curve in $\mathcal{D}$ intersects only finitely many other curves in $\mathcal{D}$, and such that the set of Dehn twists around curves in $\mathcal{D}$ generates the subgroup of $\mathrm{Map}(\Sigma)$ consisting of mapping classes supported on finite-type subsurfaces of Σ; see [14]. One can then take the set D of Observation 6.9 to be the set of Dehn twists around the curves in $\mathcal{D}$ that intersect S.

Proof of Lemma 6.8 Let K be the surface given by Theorem 5.7. For each $P \in \mathcal{P}$, there exists $\widehat{A} \in \mathcal{A}$ such that $\widehat{A} - \{x_A\}$ contains a homeomorphic copy of P. Choose one such $\widehat{A}$ for each $P \in \mathcal{P}$, and for $\widehat{A} \in \mathcal{A}$ let P_A denote the union of the elements of $\mathcal{P}$ assigned to $\widehat{A}$. Let $L \subset K$ be a connected, finite-type surface with $|\mathcal{A}|$ boundary components, and such that the complementary regions of L partition E into the sets $\widehat{A} \cup P_A$ as $\widehat{A}$ ranges over $\mathcal{A}$. We take L to have the same number of punctures and genus as K. For each $\widehat{A}$, let Σ_A denote the complementary region to L with end space $\widehat{A} \cup P_A$.

If $f \in \mathcal{V}_L$, then f can be written as a product of $|\mathcal{A}|$ homeomorphisms, one supported on each surface Σ_A (and hence identifiable with an element of $\mathrm{Map}(\Sigma_A)$). So it suffices to show, for each $\widehat{A} \in \mathcal{A}$, that $\mathrm{Map}(\Sigma_A)$ is generated by $\mathcal{V}_K \cap \mathrm{Map}(\Sigma_A)$, which is a CB subset of $\mathrm{Map}(\Sigma)$, together with a finite set.

Fix $\widehat{A}$, let K' denote $K \cap \Sigma_A$, let $\Sigma_1, \Sigma_2, \ldots, \Sigma_n$ denote the connected components of $\Sigma_A - K'$ with end spaces elements of $\mathcal{P}$, and let Σ' be the connected component with end space $\widehat{A}$. Let

$$\mathcal{G} = \mathcal{V}_K \cap \mathrm{Map}(\Sigma_A) = \mathrm{Map}(\Sigma') \times \mathrm{Map}(\Sigma_1) \times \cdots \times \mathrm{Map}(\Sigma_n).$$

In view of Observation 6.9, we can find a finite set of Dehn twists D_A whose support is contained in Σ_A such that, for any finite-type surface $S' \subset \Sigma_A$, $\mathrm{Map}(S')$ is contained in the group generated by D_A and $\mathcal{G}$.

Recall from Proposition 5.5 that P_A contains no maximal points, that $A = \widehat{A} \cup P_A$ is a self-similar set (and homeomorphic to $\widehat{A}$), and in particular we can find a copy of P_A in any neighborhood of x_A. This implies there is some homeomorphism g_A of Σ_A with $g_A(P_A) \subset \mathrm{End}(\Sigma')$, where $\mathrm{End}(\Sigma)$ denotes the space of ends of the surface Σ'. We now set our desired finite set to be

$$\mathcal{F} = D_A \cup \{g_A\}.$$

We now show that $\mathrm{Map}(\Sigma_A)$ is generated by

$$\mathcal{G}' = \mathcal{G} \cup \mathcal{F}.$$

Let $f \in \mathrm{Map}(\Sigma_A)$. Since $\mathcal{M}(\widehat{A})$ is an invariant set, we may find a neighborhood U of $\mathcal{M}(\widehat{A})$ in Σ_A, which we may take to be a (infinite-type) subsurface of Σ' with a single boundary component, such that $f(U) \subset \Sigma'$. Let P'_A be a homeomorphic copy of P_A contained in $\mathrm{End}(U)$. Thus, $f(P'_A) \subset \mathrm{End}(\Sigma')$, and so there exists $h \in \mathrm{Map}(\Sigma')$ with $hf(P'_A) = g_A(P_A)$. This means $g_A^{-1}hf(P'_A) = P_A$ and therefore,

$$g_A^{-1}hf(P_A) \subset \mathrm{End}(\Sigma').$$

Thus, there exists $h' \in \mathrm{Map}(\Sigma')$ interchanging $g_A^{-1}hf(P_A)$ with $g_A(P_A)$, such that the map $h' \circ (g_A^{-1}hf)$ agrees with g_A on P_A. It follows that

$$g_A^{-1} \circ h' \circ g_A^{-1}hf|_{P_A} = \mathrm{id}.$$

Applying another element $h'' \in \mathrm{Map}(\Sigma')$, we can ensure that

$$f' = h'' \circ g_A^{-1} \circ h' \circ g_A^{-1}hf$$

is the identity on $\text{End}(\Sigma_A)$—that is, it is an element of the pure mapping class group of Σ_A. Since h, h', h'' and g_A are in $\mathcal{G}'$, it is sufficient to show that f' is also contained in the group generated by $\mathcal{G}'$.

Let S' be a genus-zero surface of finite type that contains $K' \cup f'(K')$. Since f' is a pure mapping class, for each boundary curve α of K' the curves α and $f'(\alpha)$ cut out the same subset of $\text{End}(\Sigma_A)$. Hence they also cut out the same set of boundary curves of S'. But S' has genus zero, therefore the component of $S' - K'$ associated to α is homeomorphic to the component of $S' - f'(K')$ associated to $f'(\alpha)$. That is, there is a homeomorphism $g' \in \text{Map}(S')$ such that

$$g' f'(K') = K'.$$

But, as mentioned above, g' (which has finite support) is in the group generated by $\mathcal{G}'$. Also, $g' f'$ fixes Σ' and hence is contained in $\mathcal{V}_K \cup \text{Map}(K')$. But $\mathcal{V}_K \subset \mathcal{G}'$ and K' has finite type, so $\text{Map}(K')$ is also contained in the group generated by $\mathcal{G}'$. This finishes the proof. $\qquad\square$

Going forward, we will ignore the surface K produced earlier that defined the CB neighborhood $\mathcal{V}_K$, and instead use the surface L, which gives a simpler decomposition of the end space. The sets $P \in \mathcal{P}$ play no further role, and we focus on the decomposition $E = \bigsqcup_{A \in \mathcal{A}} A$ given by the end spaces of complementary regions to the surface L. This is the reason for our choice of notation $\hat{A}$ for the smaller sets of the finer partition of E, for we may now abandon the cumbersome hats.

Further decompositions of end sets Now we begin the technical work of the classification of CB generated mapping class groups. As motivation for our next lemmas, consider the surface depicted in Figure 1 on the left. This surface has a mapping class group which is both locally CB and CB generated — we have not proved CB generation yet, but the reader may find it an illustrative exercise to attempt this case by hand. Here, the decomposition of E given by the surface L is $E = A \sqcup B \sqcup C$, where A and C are accumulated by genus, A and B are homeomorphic to $\omega + 1$, and C is a singleton. As well as a neighborhood of the identity of the form $\mathcal{V}_L$, any generating set must include a "handle shift" moving genus from A into C (see Definition 6.20 below), as well as a "puncture shift" that moves isolated punctures out of A and into B. If each handle was replaced by, say, a puncture accumulated by genus, one would need a shift moving these end types in and out of neighborhoods of A and C instead.

To generalize this observation to other surfaces with more complicated topology, we need to identify types of ends of Σ that accumulate at the maximal ends of the various sets in the decomposition. The sets $W_{A,B}$ defined in Lemma 6.10 and refined in Lemma 6.17 below pick out blocks of ends that can be shifted between elements A and B in $\mathcal{A}$. Ultimately, we will have to further subdivide these blocks to distinguish different ends that can be independently shifted; this is carried out in Section 6.4.

Lemma 6.10 *Assume that* $\mathrm{Map}(\Sigma)$ *is locally CB and that* E *does not have limit type. Then*:

- *For every* $A \in \mathcal{A}$, *there is a neighborhood* $N(x_A) \subset A$ *containing* x_A *such that* $A - N(x_A)$ *contains a representative of every type in* $A - \{x_A\}$.

- *For every pair* $A, B \in \mathcal{A}$, *there is a clopen set* $W_{A,B} \subset (A - N(x_A))$ *with the property that* $E(z) \cap W_{A,B} \neq \varnothing$ *if and only if*

$$E(z) \cap (A - \{x_A\}) \neq \varnothing \quad \text{and} \quad E(z) \cap (B - \{x_B\}) \neq \varnothing.$$

- *For every* $A \in \mathcal{A}$, *there is a clopen set* $W_A \subset (A - N(x_A))$ *with the property that if* $E(z) \cap (A - \{x_A\}) \neq \varnothing$ *and, for all* $B \neq A$, $E(z) \cap (B - \{x_B\}) = \varnothing$ *then* $E(z) \cap W_A \neq \varnothing$.

In other words, $W_{A,B}$ contains representatives of every type of end that appears in both $A - \{x_A\}$ and $B - \{x_B\}$, and W_A contains representatives of every type that appears only in A.

We declare $W_{A,B} = \varnothing$ if $A - \{x_A\}$ and $B - \{x_B\}$ have no common types of ends, and similarly take $W_A = \varnothing$ if each type of end in A appears also in some $B \neq A$.

Proof We start with the first assertion. If $\mathcal{M}(A)$ is a Cantor set then we can take $N(x_A)$ to be any neighborhood of x_A that does not contain all of $\mathcal{M}(A)$, and the first assertion follows since $\mathcal{M}(A)$ is the set of maximal points. Otherwise, $\mathcal{M}(A) = \{x_A\}$. Let G be the finite-index subgroup of $\mathrm{Map}(\Sigma)$ that fixes $E(x_A)$ (which we know is finite). Also recall that $A = \widehat{A} \cup P_A$. If such a neighborhood $N(x_A)$ does not exist, then there is a nested family of neighborhoods $U_n \subset \widehat{A}$ descending to x_A and points $z_n \in U_n$ where $(E(z_n) \cap \widehat{A}) \subset U_n$. We also have that $E(z_n)$ has nontrivial intersection with the complement of $\widehat{A}$, in fact if we choose V to be a neighborhood of x_A excluding z_n, then for f_V as in part (iii) of Theorem 5.7, $f_V(z_n)$ is not in A. Then, letting $X = \{x_A\}$ and assuming $U_0 = \widehat{A}$, we see that E has limit type. The contradiction proves the first assertion.

For the second assertion, fix A and $B \in \mathcal{A}$ and let

$$X = \{x \in E \mid E(x) \cap A \neq \varnothing \text{ and } E(x) \cap B = \varnothing\}.$$

Then $X \cap A$ is closed: this follows since A is closed, and if x_n is a sequence of points in $X \cap A$ converging to x_∞ but there is some point $z \in E(x_\infty) \cap B$, then any neighborhood of z would contain homeomorphic copies of neighborhoods of x_n, for sufficiently large n, contradicting the fact that $E(x_n) \cap B = \varnothing$.

Now consider a family of neighborhoods U_n of $X \cap A$ with $U_n \searrow X$ and $U_0 \cap B = \varnothing$. Let $W_n = A - (U_n \cup N(x_A))$. Since we have removed the neighborhood U_n of X, every point in W_n has a representative in B. We claim that, for some $N \in \mathbb{N}$, W_N contains a representative of all points that appear in both A and B, that is to say, $W_{A,B}$ can be taken to be W_N. To prove the claim, suppose for contradiction that it fails. Then after passing to a subsequence, we may find points z_n, all of distinct types, such that $z_n \in U_n$, $E(z_n) \cap A \neq \varnothing$ and $E(z_n) \cap B \neq \varnothing$. Since $E(z_n)$ intersects $U_0^c \supset B$, this implies that E has limit type. The contradiction proves the second assertion.

For the third assertion, consider the closed set

$$X = \{x \in E \mid E(x) \cap A \neq \varnothing \text{ and } E(x) \cap B = \varnothing \text{ for all } B \neq A\}.$$

Let U be any clopen neighborhood of $X \cap A$ in A, and let $W_A = U - N(x_A)$. Then by definition of $N(x_A)$, $(X \cap A) - N(x_A)$ contains a representative of every type appearing only in A, so this remains true of its clopen neighborhood W_A. □

6.3 Tame end spaces

Definition 6.11 An end space E is *tame* if, for every $A \in \mathcal{A}$, the point x_A has a stable neighborhood (as in Definition 4.14), and for any $A, B \in \mathcal{A}$, every maximal point in $W_{A,B}$ has a stable neighborhood.

If Σ has locally CB mapping class group, then Theorem 1.4 implies that maximal points have stable neighborhoods, so half of the tameness condition is satisfied. The other half is an assumption that will be used in the next two sections. While this seems like a restrictive hypothesis, the class of tame surfaces is very large. In fact, the following problem seems to be challenging, as the examples of nontame surfaces (excluding those which are self-similar; see Example 6.13 below) which we can easily construct all seem to have infinite-rank or limit-type like behavior.

Problem 6.12   Does there exist an example of a nontame surface whose mapping class group has nontrivial, well-defined quasi-isometry type (ie is locally, but not globally, CB and CB generated)?

Example 6.13   (nontame surfaces)   Suppose $\{z_n\}_{n\in\mathbb{N}}$ is a sequence of points in an end space which are not comparable, ie for all $i \neq j$ we have neither $z_i \preccurlyeq z_j$ nor $z_j \preccurlyeq z_i$. An end space containing such a sequence may be constructed, for instance, as in Example 6.6, and even (as in that example) have the property that each z_n admits a stable neighborhood V_n. Let D denote a set consisting of the disjoint union of one copy of each stable neighborhood V_n and a singleton x, so that the sets V_n Hausdorff converge to x. Then x is a maximal point in D, but fails the stable neighborhood condition in the definition of tame, since the homeomorphism types of small neighborhoods of x do not eventually stabilize.

A surface with end space D fails the condition of Theorem 1.4 so is not locally CB, but one can easily modify this construction to provide locally, and even globally, CB examples. For instance, let E be the disjoint union of countably many copies of D, arranged to have exactly k accumulation points. If $k = 1$, the end space constructed is self-similar, with the sole accumulation point the unique maximal point. If $k > 1$, the end space may be partitioned into finitely many self-similar sets satisfying the condition of Theorem 1.4, but has immediate predecessors to the maximal points with no tame neighborhood. (However, we note that this example is infinite rank, so the mapping class group of a surface with this end type is not CB generated.)

The main application of the tameness condition is that it allows us to give a standard form to other subsets of E. We begin with a definition and some preliminary lemmas.

Definition 6.14   When $E(z)$ is countable, we will say that z is a point *of countable type*. Define $E_{\mathrm{cp}}(A, B)$ (the *countable predecessor set*) to be the subset of $W_{A,B}$ consisting of points z where z is maximal in $W_{A,B}$ and of countable type. Since $W_{A,B}$ is clopen, it has maximal points as in Proposition 4.7.

Observation 6.15   If z is any point of countable type, then any accumulation point p of $E(z)$ satisfies $z \prec p$. Thus, if $z \in E_{\mathrm{cp}}(A, B)$, then $E(z)$ does not have any accumulation points in $W_{A,B}$ and hence $E(z) \cap W_{A,B}$ is a finite set.

Lemma 6.16   Suppose E is tame and $\mathrm{Map}(\Sigma)$ has neither limit type nor infinite rank. Then, for any $A, B \in \mathcal{A}$, the set $E_{\mathrm{cp}}(A, B)$ contains only finitely many different types.

Proof As a first case, suppose that $\mathcal{M}(A)$ is a single point. Let G be the finite-index subgroup of $\text{Map}(\Sigma)$ that fixes x_A; recall that $E(x_A)$ is finite. Now $X = \{x_A\}$ is G–invariant and since $\text{Map}(\Sigma)$ does not have infinite rank, we can take $U = A$ and conclude that $E_{\text{cp}}(A, B)$ has finitely many different types.

Otherwise, $\mathcal{M}(A)$ is a Cantor set. If $E(x_A)$ does not intersect B, we can take $X = E(x_A)$ and $U = B^c$. Then X is $\text{Map}(\Sigma)$–invariant and again the fact that $\text{Map}(\Sigma)$ does not have infinite rank implies that $E_{\text{cp}}(A, B)$ has finitely many different types.

If $\mathcal{M}(A)$ is a Cantor set and $E(x_A)$ intersects B, then $E(x_A)$ intersects $W_{A,B}$ and thus $E_{\text{cp}}(A, B)$ is empty. $\qquad\square$

Lemma 6.17 *Suppose that Σ has tame end space. Then, under the hypotheses of Lemma 6.16, the sets $W_{A,B}$ from Lemma 6.10 can be chosen so that for any z in $E_{\text{cp}}(A, B)$, the set $E(z) \cap W_{A,B}$ is a singleton. Such a choice specifies a set which is unique up to homeomorphism, and in this case $W_{A,B}$ is homeomorphic to $W_{B,A}$.*

Proof Fix a choice of set $W_{A,B}$ as given by Lemma 6.10. For each $z \in E_{\text{cp}}(A, B)$, choose disjoint stable neighborhoods around every point in the finite set $E(z) \cap W_{A,B}$ (this set is finite by Observation 6.15) and remove all but one neighborhood, leaving the rest of $W_{A,B}$ unchanged. Denote this new set by $W'_{A,B}$. Since one such neighborhood remains, any type that was represented in $W_{A,B}$ is still represented there, so it satisfies the conditions of Lemma 6.10. We wish to show that the homeomorphism type of $W'_{A,B}$ is independent of our choices of stable neighborhoods, and that $W'_{A,B}$ is homeomorphic to $W'_{B,A}$. We prove both assertions simultaneously, by showing that $W'_{A,B}$ is homeomorphic to any choice of set $W'_{B,A}$ as defined by the same procedure.

Let $z_1, \ldots, z_k \in W'_{A,B}$ be the points of $E_{\text{cp}}(A, B)$; recall there is one of each type. Let $V_1, \ldots, V_k$ be the chosen disjoint stable neighborhoods of these points in $W'_{A,B}$, which exist by the tameness assumption. Let $W = W'_{A,B} - \bigcup_i V_i$. Similarly, choose $V'_1, \ldots, V'_k$ to be disjoint stable neighborhoods of points of countable predecessor type in $W_{B,A}$ so that V_i is homeomorphic to V'_i, and let $W' = W'_{B,A} - \bigcup_i V'_i$. We start by showing that

$$W \cup W'_{B,A} \cong W'_{B,A}.$$

This is because, for any point in $x \in W$, there is a point $y \in W'_{B,A}$ that is maximal in $W'_{A,B}$, where y is an accumulation point of $E(x)$. Hence, by Lemma 4.18, there is a neighborhood U_x of x and stable neighborhood V_y of y such that $U_x \cup V_y$ is

homeomorphic to V_y. Since W is compact, finitely many such neighborhoods are enough to cover W and, shrinking these neighborhoods if needed, we can write W as the disjoint union of finitely many such neighborhoods. Thus, W can be absorbed into $W'_{B,A}$.

Similarly we have that $W' \cup W'_{A,B}$ is homeomorphic to $W'_{A,B}$. That is,

$$W'_{A,B} \cong W'_{A,B} \cup W' \cong W \cup W' \cup \left(\bigcup_i V_i \right) \cong W \cup W' \cup \left(\bigcup_i V'_i \right)$$
$$\cong W \cup W'_{B,A} \cong W'_{B,A}. \qquad \square$$

Going forward, we will use $W_{A,B}$ to denote the (well-defined up to homeomorphism) sets constructed in the lemma, each containing a single representative of each of its countable predecessor types.

6.4 Classification of CB generated mapping class groups

The purpose of this section is to prove Theorem 1.6, namely, the statement that the necessary conditions for CB generation introduced in Section 6.1 are also sufficient for tame surfaces.

We continue with the notation and conventions introduced in the previous section, in particular the following.

Convention Going forward, we let L denote the finite-type surface furnished by Proposition 5.4, so that the complementary regions to L produce a decomposition $E = \bigsqcup_{A \in \mathcal{A}} A$, where each A is self-similar, and we have $\bigsqcup \mathcal{M}(A) = \mathcal{M}(E)$.

The next proposition is the main technical ingredient in the proof of Theorem 1.6. It says that, by using elements from a CB set, one may map any neighborhood U of x_A in E homeomorphically onto A while pointwise fixing any set $B \in \mathcal{A}$ which shares no end types with $A - U$.

Proposition 6.18 *Assume that E is tame and not of limit type, and that $\mathrm{Map}(\Sigma)$ does not have infinite rank. Then there is a finite set $F \subset \mathrm{Map}(\Sigma)$ such that the following holds*:

Let $A \in \mathcal{A}$, and let $U \subset A$ be a neighborhood of x_A. If $\mathcal{B}_U \subset \mathcal{A}$ is a subset that satisfies $E(y) \cap \left(\bigcup_{B \in \mathcal{B}_U} B \right) \neq \varnothing$ for all $y \in A - U$, then there is an element f in the group generated by F and $\mathcal{V}_L$ with $f(U) = A$, and $f|_C = \mathrm{id}$ for all $C \in (\mathcal{A} - \mathcal{B}_U)$.

Proof The proof consists of several preliminary structural results on end spaces, carried out in Steps 1–4; the set U and $\mathcal{B}_U$ are introduced in the final step.

Step 1: decomposition of the sets $A \in \mathcal{A}$ Fix $A \in \mathcal{A}$. For every $B \in \mathcal{A}$, consider a copy of $W_{A,B} \subset A$ as in Lemma 6.17, as well as a homeomorphic copy of W_A. A short argument shows that we may choose these sets to be pairwise disjoint, so that we have $W_{A,B} \cap W_{A,B'} = \varnothing$ whenever $B \neq B'$ and $W_{A,B} \cap W_A = \varnothing$ for all B. This is as follows: enumerate the sets $B_1, B_2, \ldots, B_k$ of $\mathcal{A} - \{A\}$ and perform our original construction to obtain W_{A,B_1}. This set is disjoint from $N(x_A)$. By self-similarity, there is a homeomorphic copy of A inside $N(x_A)$, hence we may find a set W_{A,B_2} disjoint from W_{A,B_1} and also disjoint from a smaller copy of $N(x_A)$. Continuing in this manner, we may produce the desired sets. Doing this one more time, we also find a disjoint copy of W_A. We keep these sets (and refer to them to by this notation, $W_{A,B}$ and W_A) for the remainder of the proof.

Let
$$T_0 = W_A \sqcup \left(\bigsqcup_{B \in \mathcal{A} - \{A\}} W_{A,B} \right) \subset A.$$

By construction, for every $y \in A - \{x_A\}$, $E(y)$ intersects T_0 by Theorem 1.4.

Let $V_1 = A - T_0$ and consider a family of neighborhoods $V_k \searrow x_A$. Each V_k contains a copy of A and hence a copy T_k of T_0. After dropping some of the sets V_k from the nested sequence and reindexing, we can assume $T_1 \subset (V_1 - V_2)$. Continuing in this way, we find a new nested sequence of neighborhoods, which we again denote by V_k, so that $(V_k - V_{k+1})$ contains a copy T_k of T_0. In particular, the sets T_k are disjoint.

Our next goal is to modify this construction so that we in fact have $(V_k - V_{k+1}) \cong T_k$, ie we obtain a nested family of neighborhoods such that the annular regions between them are homeomorphic to the sets T_k above. To do this, we first show that we can distribute the set
$$Q = (V_1 - V_2) - T_1$$

among finitely many of the other sets T_k, with $k > 1$, while preserving the homeomorphism class of the T_k; and then proceed iteratively.

For each point $y \in Q$, $E(y)$ intersects T_0 and hence y has a neighborhood $V_y \subset Q$ that has a homeomorphic copy inside T_0. Since Q is compact, finitely many such neighborhoods are sufficient to cover Q. Making some of these neighborhoods smaller, we can write $Q = Q_1 \sqcup \cdots \sqcup Q_m$, where every Q_i has a copy in T_0 and hence in

every T_k. For $j = 1, \ldots, m$ and $k \equiv j \mod m$ let Q'_k be the copy of Q_j in V_k. For $k = 1, \ldots, m$ define

$$T'_k = (T_k - Q'_k) \cup Q_k,$$

and for $k > m$ define

$$T'_k = (T_k - Q'_k) \cup Q'_{k-m}.$$

Each T'_k is still homeomorphic to T_0, the sets T'_k are disjoint and every point in $(V_1 - V_2)$ is contained in some T'_k. Note that T_0 is not modified.

Similarly to the above, we can distribute the points in

$$Q' = (V_2 - V_3) - \bigcup_{k \geq 1} T'_k$$

among the sets T'_k, with $k = 2, 3, \ldots$, without changing their topology. That is, we obtain a family T''_k of disjoint sets homeomorphic to T_0 whose union covers $A - V_3$, without modifying T_0 or T'_1. Continuing in this way, every T_k is modified finitely many times and stabilizes after k steps. Thus, $\{T_k^{(k)} \mid k \in \mathbb{N}\}$ is a family of disjoint copies of T_0 that covers $A - \{x_A\}$. To simplify notation, denote $T_k^{(k)}$ by $T_k(A)$. To summarize,

$$A - \{x_A\} = \bigsqcup_{k \geq 0} T_k(A),$$

and, defining

$$U_n := \bigsqcup_{k \geq n} T_k(A),$$

we have

$$U_n \searrow \{x_A\}.$$

Since $T_0 = W_A \sqcup \left(\bigsqcup_{B \neq A} W_{A,B} \right)$, we have a similar decomposition of each homeomorphic set $T_k(A)$ into sets homeomorphic to W_A and $W_{A,B}$, which we notate by

$$T_k(A) = W_A^k \sqcup \left(\bigsqcup_{B \in \mathcal{A} - \{A\}} W_{A,B}^k \right),$$

where, for $k \in \mathbb{N}$, W_A^k is a set homeomorphic to W_A and $W_{A,B}^k$ is a set homeomorphic to $W_{A,B}$.

We also have the above decomposition for every $B \in \mathcal{A} - \{A\}$. For notational convenience, when $k < 0$, we define

$$W_{A,B}^k := W_{B,A}^{-k-1}.$$

Step 2: a first shift map Using the decomposition above, we define the first homeo-morphism (of several) that shifts points between A and B. Since the sets $W_{A,B}^k$ for $k \in \mathbb{Z}$ are disjoint and homeomorphic and Hausdorff converge to the points x_A and x_B as k approaches ∞ and $-\infty$, respectively, there exists a homeomorphism $\eta_{A,B}$ such that

$$\eta_{A,B}(W_{A,B}^k) = W_{A,B}^{k-1} \quad \text{for all } k \in \mathbb{Z},$$

and restricts to the identity elsewhere in E. Fix one such map for each (unordered) pair $A, B \in \mathcal{A}$. Visually, the map $\eta_{A,B}$ pushes a copy of $W_{A,B}$ out of A and into B.

Step 3: shifting countable predecessor ends independently Now we define homeo-morphisms allowing one to shift the countable predecessor ends one by one. As motivation, consider, for instance, a surface with $E \cong \omega \cdot 2 + 1$, such that E^G and the closure of $E - E^G$ are both homeomorphic to $\omega \cdot 2 + 1$, as shown in Figure 5. There are two maximal ends, $\mathcal{A} = \{A, B\}$, and we have the simple situation where $W_{A,B} = T_0$ consists of one of each type of nonmaximal end. The map $\eta_{A,B}$ shifts ends of both types towards B, simultaneously. However, there is evidently a homeomorphism of Σ which pointwise fixes $E - E^G$ and shifts the nonmaximal ends of E^G.

For $z \in E_{\mathrm{cp}}(A, B)$, let $W_{A,B}^k(z) \subset W_{A,B}^k$ be a stable neighborhood of the unique intersection point of $E(z)$ with $W_{A,B}^k$. By making these neighborhoods smaller, we can assume that the $W_{A,B}^k(z)$ for different $z \in E_{\mathrm{cp}}(A, B)$ are disjoint. (This is a very slight abuse of notation since $W_{A,B}^k(z)$ depends only on the equivalence class of z under $\sim$, not the point itself.) Define $\eta_{A,B,z}$ to be a homeomorphism of Σ such that

$$\eta_{A,B,z}(W_{A,B}^k(z)) = W_{A,B}^{k-1}(z) \quad \text{for } k \in \mathbb{Z}$$

and acts by the identity elsewhere in E. Note that the actions of $\eta_{A,B,z}$ on E commute with each other and have support in $A \cup B$.
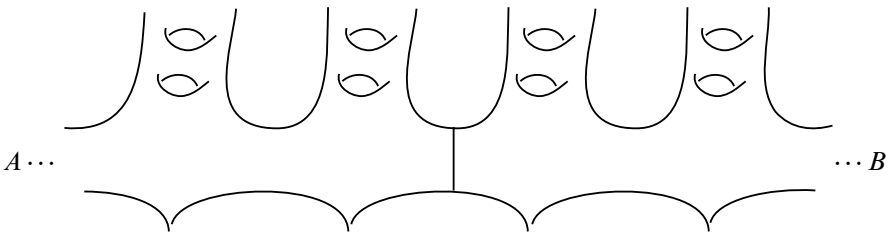
Figure 5: $E - E^G$ and the nonmaximal ends of E^G can be shifted independently.

Step 4: standard decomposition for sets of shared ends   Define

$$E(A, B) = \bigsqcup_{k=0}^{\infty} W_{A,B}^k.$$

The following claim shows that clopen subsets of $E(A, B)$ have a standard form:

Claim   Let $W \subset E(A, B)$ be any clopen set in $E(A, B)$ containing $W_{A,B}$ and disjoint from x_A. For $z \in E_{\mathrm{cp}}(A, B)$, let $p_z(W) = |E(z) \cap W|$. Then W is homeomorphic to the set

$$W_{A,B} \sqcup \left(\bigsqcup_{z \in E_{\mathrm{cp}}(A,B)} \bigsqcup_{k=1}^{p_z(W)-1} W_{A,B}^k(z) \right).$$

Recall that $W_{A,B} \subset T_0$ was a fixed set, chosen in Step 1. However, note that this structure theorem also applies to any clopen subset of $E(A, B)$ which contains a *homeomorphic copy* of $W_{A,B}$.

Proof of claim   For $z \in E_{\mathrm{cp}}(A, B)$ and $y \in E(z) \cap (W - W_{A,B})$, choose a stable neighborhood V_y of y in W. Making the neighborhoods small enough, we can assume they are disjoint from each other and from $W_{A,B}$. Since stable neighborhoods are canonical, we can map the union of these neighborhoods homeomorphically to

$$\bigsqcup_{z \in E_{\mathrm{cp}}(A,B)} \bigsqcup_{k=1}^{p_z(W)-1} W_{A,B}^k(z).$$

It remains to show that if $p_z(W) = 1$ for every $z \in E_{\mathrm{cp}}(A, B)$, then W is homeomorphic to $W_{A,B}$.

For every point in $y \in (W - W_{A,B})$, there is a point $x \in W_{A,B}$ that is maximal in $W_{A,B}$ where x is an accumulation point of $E(y)$. By the tameness assumption, x has a stable neighborhood and by Lemma 4.18, for any stable neighborhoods V_x of x and any neighborhood V_y of y, $V_x \cup V_y$ is homeomorphic to V_x. Taking a cover of $W - W_{A,B}$ by such neighborhoods, we conclude that

$$W = (W - W_{A,B}) \cup W_{A,B} \cong W_{A,B}. \qquad \square$$

Step 5: finishing the proof of Proposition 6.18   Let

$$F = \{\eta_{A,B}^{\pm 1}, \eta_{A,B,z}^{\pm 1} \mid B \in \mathcal{A} - \{A\} \text{ and } z \in E_{\mathrm{cp}}(A, B)\}.$$

Let $U \subset A$ be a neighborhood of x_A and let $\mathcal{B}_U \subset \mathcal{A} - \{A\}$ be as in the statement of the proposition. The homeomorphism $\prod_{B \in \mathcal{B}_U} \eta_{A,B}^{-1}$ shifts the sets $W_{B,A}$ from $\bigsqcup_{B \in \mathcal{B}_U} B$

into A, and, in particular,

$$\bigsqcup_{B\in\mathcal{B}_U} W_{A,B} \subset A - \left(\prod_{B\in\mathcal{B}_U} \eta_{A,B}^{-1}\right)(U).$$

Thus, up to applying this homomorphism, we may assume that U is sufficiently small that its complement contains $\bigsqcup_{B\in\mathcal{B}_U} W_{A,B}$, the subset of T_0.

Fix $B_1 \in \mathcal{B}_U$. Since $(A-U)\cap E(A,B_1)$ contains W_{A,B_1}, the claim proved in Step 4 implies that $(A-U)\cap E(A,B_1)$, it is homeomorphic to the standard set

$$W_{A,B_1} \sqcup \left(\bigsqcup_{z\in E_{\mathrm{cp}}(A,B_1)} \bigsqcup_{k=1}^{p_z(W)-1} W_{A,B_1}^k(z)\right)$$

in A, and the complements of both this standard set and of $(A-U)\cap E(A,B_1)$ in A are homeomorphic (each is easily seen to be homeomorphic to A). Thus, by the classification of surfaces there is a homeomorphism v_1 supported on the complementary region to L with end space A, hence in $\mathcal{V}_L$, taking $(A-U)\cap E(A,B)$ to this standard set. However, by construction, the image of this standard set under

$$\eta_{A,B_1} \circ \prod_{z\in E_{\mathrm{cp}}(A,B_1)} \eta_{A,B_1,z}^{p_z(W)-1}$$

is disjoint from A, and the image of its complement in A is equal to A. Let

$$U' = \eta_{A,B_1} \circ \prod_{z\in E_{\mathrm{cp}}(A,B_1)} \eta_{A,B_1,z}^{p_z(W)-1} \circ v_1(U).$$

Note that $\mathcal{B}_{U'} = \mathcal{B}_U - \{B_1\}$. We now repeat the process above using $B_2 \in \mathcal{B}_{U'}$ and U' and produce an element of the subgroup generated by F and $\mathcal{V}_L$ which takes U' to a subset of A containing $E(A,B_2)$. Iterating this process for each $B \in \mathcal{B}_U$ achieves the desired result. $\qquad\square$

We are almost ready to prove the main result of this section. In order to do so, we need another finite set of mapping classes, the *handle shifts*, which we define now. See also [14, Section 6] for earlier use of this class of maps.

Definition 6.19 An *infinite strip with genus* is the surface $\mathbb{R}\times[-1,1]$ with a handle attached to the interior of each set $[m,m+1]\times[0,1]$ so that $(x,y)\mapsto(x+1,y)$ is a homeomorphism of the surface.

A *handle shift* on the infinite strip with genus is the mapping class of the homeomorphism h which pointwise fixes the boundary, agrees with $(x, y) \mapsto (x + 1, y)$ outside an ϵ–neighborhood of the boundary, and on the ϵ–neighborhood agrees with $(x, y) \mapsto (x + (1 - |y|)/\epsilon, y)$.

Definition 6.20 Suppose that Σ has locally CB mapping class group and L is a surface as in Lemma 6.8 We call a (infinite-type) subsurface $R \subset \Sigma$ an *infinite strip with genus* in Σ if it is homeomorphic to an infinite strip with genus, and has the property that the complement of R in each complementary region to L has infinite genus.

A *handle shift* on R is the mapping class of the map h above (under our identification), extended to agree with the identity on the complement of R.

Recall that the *pure mapping class group*, denoted by $\mathrm{PMap}(\Sigma)$, is the subgroup of $\mathrm{Map}(\Sigma)$ which pointwise fixes E. We now prove a lemma on generating pure mapping classes.

For each pair (A, B) such that x_A and x_B are both accumulated by genus, let $R_{AB} \subset \Sigma$ be an infinite strip with genus, with one end in A and one end in B. We may choose these (one at a time) so that they are disjoint subsurfaces of Σ. Fix also a handle shift $h_{AB} \in \mathrm{Homeo}(\Sigma)$ on R_{AB}.

Lemma 6.21 (generating $\mathrm{PMap}(\Sigma)$) *Let G be a subgroup of $\mathrm{Map}(\Sigma)$ containing all mapping classes supported on finite-type subsurfaces, all mapping classes that fix each of the boundary components of L and the handle shifts h_{AB} defined above. Then G contains $\mathrm{PMap}(\Sigma)$.*

Proof For $A \in \mathcal{A}$, let Σ_A denote the connected component of $\Sigma - L$ with end space A, and let ∂_A denote its boundary component. Let $g \in \mathrm{PMap}(\Sigma)$. Then $g(\Sigma_A)$ also has end space A, and a single boundary component $g(\partial A)$. Let $T \subset \Sigma$ be a connected, finite-type subsurface large enough to contain $L \cup g(L)$. If, for each $A \in \mathcal{A}$, the surface $\Sigma_A \cap T$ is homeomorphic rel ∂T to $g(\Sigma_A) \cap T$, then there is a mapping class ϕ supported on T such that $\phi g(L) = L$, preserving each of its boundary components, which proves what we needed to show.

So we are reduced to the case where, for some A, the surface $\Sigma_A \cap T$ is not homeomorphic to $g(\Sigma_A) \cap T$. Both are connected surfaces with the same number of boundary components, so we conclude that they must have different genus. In particular, this
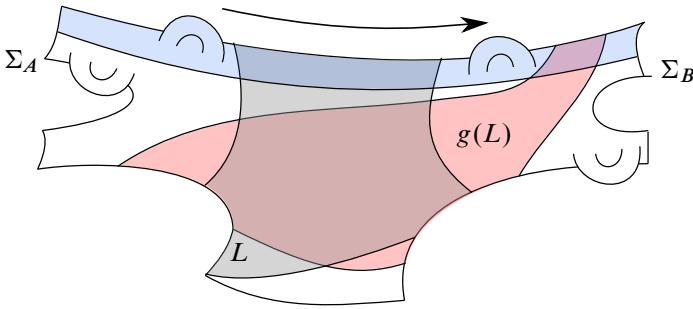
Figure 6: T containing L and $g(L)$, and the domain $\phi(R_{AB})$ of the handle shift.

only occurs if Σ is itself of infinite genus, for otherwise we choose K by convention to contain all the genus of Σ.

Without loss of generality, assume that the genus of $g(\Sigma_A) \cap T$ is larger than that of $\Sigma_A \cap T$. Since T is finite genus, there must also be another $B \in \mathcal{A}$ such that the genus of $g(\Sigma_B) \cap T$ is smaller than that of $\Sigma_B \cap T$. Since L is chosen so that complementary regions have either zero or infinite genus, we conclude that $\mathcal{M}(A)$ and $\mathcal{M}(B)$ must be accumulated by genus.

Consider the handle shift h_{AB} supported on R_{AB}, which has one end in A and one end in B. Let ϕ be a homeomorphism preserving the ends of Σ, preserving each of the boundary components of L, and such that the intersection of $\phi(R_{AB})$ with $T \cap (g(\Sigma_A) - \Sigma_A)$ and with $T \cap (\Sigma_B \cap g(\Sigma_B))$ each have genus one, and $\phi(R_{AB}) \cap T$ has genus two (so there is no genus elsewhere in T), and so that, up to replacing h_{AB} with its inverse, $\phi h_{AB} \phi^{-1}$ shifts the genus from $T \cap g(\Sigma_A)$ into $T \cap \Sigma_B$. See Figure 6 for an illustration in a simple setting. Such a homeomorphism ϕ exists by the classification of surfaces, and our stipulation that the complement of R_{AB} have infinite genus in complementary regions of L.

Then the genus of $\phi h_{AB} \phi^{-1} g(\Sigma_A) \cap T$ is one less than that of $g(\Sigma_A) \cap T$, and the genus of $\phi h_{AB}^{-1} \phi^{-1} g(\Sigma_A) \cap T$ is one more, and there is no change otherwise in the genus of complementary regions. Continuing in this fashion, one may iteratively modify g by composing by elements of G so as to arrive at a homeomorphism g' with the property that $\Sigma_A \cap T$ is homeomorphic to $g'(\Sigma_A) \cap T$ for all $A \in \mathcal{A}$, which is what we needed to show. □

A CB generating set We are now in a position to prove the main theorem on CB generation. Our CB generating set will consist of $\mathcal{V}_K$, together with the finite set consisting

of the Dehn twists D from Observation 6.15, the finite set F from Proposition 6.18, the handle shifts h_{AB}, and a finite collection of homeomorphisms g_{AB} (to be specified), one for each pair $A, B \in \mathcal{A}$ such that x_A and x_B are of the same type.

Proof of Theorem 1.6   One direction follows from Lemmas 6.4 and 6.7. We prove the other direction. For this, we show that the generating set described in the paragraph above (after giving precise definitions of g_{AB}) is in fact CB.

Let $\mathcal{V}_K \cup D$ be the CB set given by Observation 6.15 (recall that D is a finite collection of Dehn twists). Let F be the finite set from Proposition 6.18. For each pair of maximal points x_A, x_B in E^G, let h_{AB} be the handle shift defined above Lemma 6.21. Let χ be the CB set consisting of $\mathcal{V}_K \cup D$ together with the homeomorphisms from F and all the h_{AB}. By Lemma 6.21, we already know that this set generates the pure mapping class group, so we start by considering only the action on the end space.

We show first that χ generates the pointwise stabilizer of $\{x_A : A \in \mathcal{A}\}$. After this, we will add finitely many more homeomorphisms g_{AB} to generate $\mathrm{Map}(\Sigma)$.

Suppose that ϕ fixes each of the points x_A. We proceed inductively on the number of elements of $\mathcal{A}$ which are pointwise fixed by the action of ϕ on E. Let $\mathcal{A}_{\mathrm{id}}$ denote the subset (possibly empty) of $\mathcal{A}$ such that, for each $A \in \mathcal{A}_{\mathrm{id}}$, the ends of A are pointwise fixed by ϕ. Let $\mathcal{A}^c = \mathcal{A} - \mathcal{A}_{\mathrm{id}}$. Choose a set $A \in \mathcal{A}^c$. For every $B \neq A \in \mathcal{A}^c$, let $U_B = B - \phi(A)$. Then for every end $z \in (B - U_B) \subset \phi(A)$, there is some end $y \sim z$ which lies in A. Hence, by Proposition 6.18 setting $\mathcal{B}_{U_B} = \mathcal{A} - \{A, B\}$, there is an element g in the group generated by χ with support in $A \cup B$ that sends U_B to B. In particular, $g\,\phi(A) \cap B = \varnothing$ and the restriction of $g\,\phi$ to sets in $\mathcal{A}_{\mathrm{id}}$ is still the identity.

Repeating this for each element of $\mathcal{A}^c$, we may modify ϕ by elements of χ to obtain a map ϕ' such that $\phi'(A)$ is disjoint from every $C \in \mathcal{A} - \{A\}$, ie $\phi'(A) \subset A$, and so that ϕ' restricts to identity on each element of $\mathcal{A}_{\mathrm{id}}$. Letting $U = \phi'(A)$, we see that the conditions of Proposition 6.18 are again satisfied taking $\mathcal{B}_U = \mathcal{A}^c$. Hence, there is a $g' \in \langle \chi \rangle$ that is also the identity on every set in $\mathcal{A}_{\mathrm{id}}$, and sends U to A. Thus, $g'\phi'(A) = A$ and we may take some $\psi \in \mathcal{V}_L$ such that the restriction of $\psi g'\phi'$ to A is the identity.

Continuing in this way, at every step, we increase the number of sets in $\mathcal{A}_{\mathrm{id}}$, and eventually obtain a homeomorphism which pointwise fixes all ends. Since χ generates $\mathrm{PMap}(\Sigma)$, we conclude that $\phi \in \langle \chi \rangle$.

Now we show that there is a finite set F' such that $\chi \cup F'$ generates $\mathrm{Map}(\Sigma)$. Construct F' as follows. For any $A, B \in \mathcal{A}$ such that points in $\mathcal{M}(A)$ and $\mathcal{M}(B)$ are of the same

type, choose one element $g_{A,B}$ sending $N(x_A)$ to $N(x_B)$ (recall that these are stable neighborhoods) and restricting to the identity on every set in $\mathcal{A} - \{A, B\}$. Let F' be the set of all such chosen $g_{A,B}$. To see that $\chi \cup F'$ generates, let $\phi \in \mathrm{Map}(\Sigma)$. Suppose $\phi(x_A) \in B$. We modify ϕ to a map ϕ' in one of the following ways.

Case 1 Assume $\phi(x_B) \neq x_B$. There is a $\psi \in \mathcal{V}_L$ with support in B that sends $\phi(x_A)$ to x_B and hence

$$\phi' = g_{A,B}\psi\phi$$

fixes x_A.

Case 2 Assume $\phi(x_B) = x_B$. Then $\mathcal{M}(B)$ has more than one point and hence it is a Cantor set. Take a map $\psi \in \mathcal{V}_L$ with support in B that sends $\phi(x_A)$ to x_B and sends x_B to a point in $B - N(x_B)$. Then

$$\phi' = \psi^{-1} g_{A,B}\psi\phi$$

sends x_A to x_A and still fixes x_B.

The number of points x_A that are fixed by ϕ' is one more than that for ϕ. Hence, after repeating this process finitely many times, we arrive at an element fixing each maximal point, hence generated by χ. This finishes the proof. $\qquad\square$

7 Classification of CB mapping class groups

In this section we prove Theorem 1.7 classifying the surfaces Σ for which the group $\mathrm{Map}(\Sigma)$ is CB. In the case where E is uncountable, we will add the hypothesis that Σ is tame. However, we expect the classification theorem to hold without this additional hypothesis, since it is only used in the very last portion of the proof.

Note that the telescoping case occurs only when E is uncountable, by Proposition 3.7.

Proof of Theorem 1.7 If Σ has zero or infinite genus and is either telescoping or has self-similar end space, then it was shown in Propositions 3.1 and 3.5 that $\mathrm{Map}(\Sigma)$ is CB, with no hypothesis on tameness. We prove the other direction. Assume that Σ has a CB mapping class group. By Example 2.4, this implies that Σ has zero or infinite genus. Also, being globally CB, $\mathrm{Map}(\Sigma)$ is in particular locally CB so the end space admits a decomposition $E = \bigsqcup_{A \in \mathcal{A}} A$ into finitely many self-similar sets as in Theorem 1.4. Then Example 2.5 implies that, if we take such a decomposition with $\mathcal{A}$ of minimal cardinality, then $\mathcal{A}$ has either one or two elements. Finally, if $\mathcal{A}$ is a singleton, then E is self-similar. Thus, we only need to take care of the case where $\mathcal{A}$ has exactly two elements.

Example 2.5 also shows that, if $\mathcal{A} = \{A, B\}$, then $\mathcal{M}(A)$ and $\mathcal{M}(B)$ are either both singletons or Cantor sets. A slight variation on the argument there also allows us to eliminate the case where they are both Cantor sets: if points of $\mathcal{M}(A)$ are not of the same type as those in $\mathcal{M}(B)$, then one may construct a nondisplaceable subsurface just as in the example by having $\mathcal{M}(A)$ play the role of the singleton. Otherwise, points of $\mathcal{M}(A)$ and $\mathcal{M}(B)$ are all of the same type and hence

$$\mathcal{M}(E) = \mathcal{M}(A) \cup \mathcal{M}(B) = E(x_A),$$

and Lemmas 5.6 and 4.18 together imply that E is self-similar.

Thus, we can assume that $\mathcal{M}(A) = \{x_A\}$ and $\mathcal{M}(B) = \{x_B\}$. We start by showing in this case that $E_{\mathrm{mc}}(A, B) = \varnothing$. To show this, suppose for contradiction that we have some $z \in E_{\mathrm{mc}}(A, B)$. Then $E(z)$ accumulates to both x_A and x_B and since z is maximal in $E - \{x_A, x_B\}$, the set $E(z)$ has no other accumulation points. As in Lemma 6.7, we can define a continuous homomorphism to $\mathbb{Z}$ on the subgroup that pointwise fixes $\{x_A, x_B\}$ (which is of index at most two in $\mathrm{Map}(\Sigma)$), via

$$\ell(\phi) = |\{x \in E(z) : x \in A, \phi(x) \in B\}| - |\{x \in E(z) : x \in B, \phi(x) \in A\}|.$$

Let $U_0 \subset A$ be a neighborhood of z not containing x_A. Since $z \in E_{\mathrm{cp}}(A, B)$, we can find a homeomorphic copy $U_1 \subset B$ of U_0 in B. Since A and B are self-similar, we may find disjoint homeomorphic copies $U_2, U_3, \ldots$ of U_0 in A descending to x_A, and homeomorphic copies $U_{-1}, U_{-2}, \ldots$ of U_0 in B descending to x_B. Let η be a homeomorphism that sends U_i to U_{i+1} and restricts to the identity everywhere else. Then $\ell(\eta^n) = n$, so the homomorphism ℓ is unbounded and $\mathrm{Map}(\Sigma)$ is not CB. This gives the desired contradiction, so we conclude that $E_{\mathrm{mc}}(A, B) = \varnothing$. Note that, in particular, this implies E is not countable.

We now show that E is telescoping. Let $N(x_A)$ and $N(x_B)$ be as in Lemma 6.10. Let V_1 and V_2 be subsurfaces with a single boundary component, such that the end space of V_1 is $N(x_A)$ and that of V_2 is $N(x_B)$. We will check the definition of telescoping by using these neighborhoods of $x_1 = x_A$ and $x_2 = x_B$.

Let $W_1 \subset V_1$ and $W_2 \subset V_2$ be neighborhoods of x_A and x_B respectively. Let S be a finite-type subsurface, homeomorphic to a pair of pants, whose complementary regions partition E into W_1, V_2 and the remaining ends. Provided $N(x_A)$ and $N(x_B)$ are chosen small enough, condition (ii) of Theorem 1.4 ensures that either Σ has genus zero, or $\Sigma - (V_1 \cup V_2)$ has infinite genus.

Let f_1 be a homeomorphism displacing S. We may also assume that f_1 fixes x_A and x_B, since existence of a nondisplaceable subsurface in the finite-index subgroup

of $\mathrm{Map}(\Sigma)$ stabilizing x_A and x_B is sufficient to show that $\mathrm{Map}(\Sigma)$ is not CB. Then, up to replacing f_1 with its inverse, we have $f_1(\Sigma - W_1) \subset V_2$. A similar argument gives a homeomorphism f_2 with $f_2(\Sigma - W_2) \subset V_1$ and so the second condition in the definition of telescoping is satisfied.

For the first condition, we need to find a homeomorphism of the subsurface $\Sigma - V_2$ that maps W_1 to V_1. By Lemma 4.18, we know that V_1 and W_1 are homeomorphic — their end sets are homeomorphic, and they each have zero or infinite genus and one boundary component — so we need only show that their complements are homeomorphic and apply the classification of surfaces. Since, as remarked above, Σ either has genus zero or $\Sigma - (V_1 \cup V_2)$ has infinite genus, we need only produce such a homeomorphism on the level of end spaces. Here we will finally invoke tameness. Let

$$\Sigma' = \Sigma - (V_1 \cup V_2).$$

By definition of $N(x_A)$, for any end z of $V_1 - W_1$ there exists a maximal point $x \in W_1$ with $z \preceq x$. Tameness means that x has a stable neighborhood. Since x is not of countable type, it is necessarily an accumulation point of $E(z)$ (even if z and x are of the same type), and hence Lemma 4.18 implies that z has a neighborhood U_z such that $U_z \cup V_x$ is homeomorphic to V_x. Thus, on the level of ends, the end space of Σ' is homeomorphic to that of its union with U_z.

Since the end space of $V_1 - W_1$ is compact, it may be covered by finitely many such neighborhoods U_z (varying z); applying the procedure above to each of them in turn produces the desired homeomorphism on the level of end spaces, showing the two subsurfaces are homeomorphic. $\qquad\square$

References

[1] **J Aramayona**, **A Fossas**, **H Parlier**, *Arc and curve graphs for infinite-type surfaces*, Proc. Amer. Math. Soc. 145 (2017) 4995–5006 MR Zbl

[2] **J Aramayona**, **P Patel**, **N G Vlamis**, *The first integral cohomology of pure mapping class groups*, Int. Math. Res. Not. 2020 (2020) 8973–8996 MR Zbl

[3] **J Bavard**, *Hyperbolicité du graphe des rayons et quasi-morphismes sur un gros groupe modulaire*, Geom. Topol. 20 (2016) 491–535 MR Zbl

[4] **J Bavard**, **S Dowdall**, **K Rafi**, *Isomorphisms between big mapping class groups*, Int. Math. Res. Not. 2020 (2020) 3084–3099 MR Zbl

[5] **R Camerlo**, **S Gao**, *The completeness of the isomorphism relation for countable Boolean algebras*, Trans. Amer. Math. Soc. 353 (2001) 491–518 MR Zbl

[6] **R Chamanara**, *Affine automorphism groups of surfaces of infinite type*, from "In the tradition of Ahlfors and Bers, III" (W Abikoff, A Haas, editors), Contemp. Math. 355, Amer. Math. Soc., Providence, RI (2004) 123–145 MR Zbl

[7] **M G Durham**, **F Fanoni**, **N G Vlamis**, *Graphs of curves on infinite-type surfaces with mapping class group actions*, Ann. Inst. Fourier (Grenoble) 68 (2018) 2581–2612 MR Zbl

[8] **F Fanoni**, **S Hensel**, **N G Vlamis**, *Big mapping class groups acting on homology*, Indiana Univ. Math. J. 70 (2021) 2261–2294 MR Zbl

[9] **W P Hooper**, *Grid graphs and lattice surfaces*, Int. Math. Res. Not. 2013 (2013) 2657–2698 MR Zbl

[10] **J Ketonen**, *The structure of countable Boolean algebras*, Ann. of Math. 108 (1978) 41–89 MR Zbl

[11] **H A Masur**, **Y N Minsky**, *Geometry of the complex of curves, I: Hyperbolicity*, Invent. Math. 138 (1999) 103–149 MR Zbl

[12] **H A Masur**, **Y N Minsky**, *Geometry of the complex of curves, II: Hierarchical structure*, Geom. Funct. Anal. 10 (2000) 902–974 MR Zbl

[13] **S Mazurkiewicz**, **W Sierpiński**, *Contribution à la topologie des ensembles dénombrables*, Fund. Math. 1 (1920) 17–27 Zbl

[14] **P Patel**, **N G Vlamis**, *Algebraic and topological properties of big mapping class groups*, Algebr. Geom. Topol. 18 (2018) 4109–4142 MR Zbl

[15] **A Randecker**, *Wild translation surfaces and infinite genus*, Algebr. Geom. Topol. 18 (2018) 2661–2699 MR Zbl

[16] **I Richards**, *On the classification of noncompact surfaces*, Trans. Amer. Math. Soc. 106 (1963) 259–269 MR Zbl

[17] **C Rosendal**, *Global and local boundedness of Polish groups*, Indiana Univ. Math. J. 62 (2013) 1621–1678 MR Zbl

[18] **C Rosendal**, *Coarse geometry of topological groups*, Cambridge Tracts in Mathematics 223, Cambridge Univ. Press (2021) MR Zbl

[19] **S Schleimer**, *Notes on the complex of curves*, lecture notes (2020) Available at http://homepages.warwick.ac.uk/~masgar/Maths/notes2.pdf

Department of Mathematics, Cornell University
Ithaca, NY, United States

Department of Mathematics, University of Toronto
Toronto, ON, Canada

k.mann@cornell.edu, rafi@math.toronto.edu

On dense totipotent free subgroups in full groups

Alessandro Carderi

Damien Gaboriau

François Le Maître

We study probability measure preserving (p.m.p.) nonfree actions of free groups and the associated IRSs. The perfect kernel of a countable group Γ is the largest closed subspace of the space of subgroups of Γ without isolated points. We introduce the class of totipotent ergodic p.m.p. actions of Γ: those for which almost every point-stabilizer has dense conjugacy class in the perfect kernel. Equivalently, the support of the associated IRS is as large as possible, namely it is equal to the whole perfect kernel. We prove that every ergodic p.m.p. equivalence relation $\mathscr{R}$ of cost $< r$ can be realized by the orbits of an action of the free group F_r on r generators that is totipotent and such that the image in the full group $[\mathscr{R}]$ is dense. We explain why these actions have no minimal models. This also provides a continuum of pairwise orbit inequivalent invariant random subgroups of F_r, all of whose supports are equal to the whole space of infinite-index subgroups. We are led to introduce a property of topologically generating pairs for full groups (which we call evanescence) and establish a genericity result about their existence. We show that their existence characterizes cost 1.

1 Introduction

> In this context, clarifying precisely what is meant by "totipotency" and how it is experimentally determined will both avoid unnecessary controversy and potentially reduce inappropriate barriers to research.
>
> — M Condic [5]

Let Γ be a countable discrete group. Denote by $\mathrm{Sub}(\Gamma)$ the space of subgroups of Γ. It is equipped with the compact totally disconnected topology of pointwise convergence and with the continuous Γ–action by conjugation. Let β be a Borel Γ–action on the standard Borel space $X \simeq [0, 1]$. Its *stabilizer map*

$$\mathrm{Stab}^\beta \colon X \to \mathrm{Sub}(\Gamma), \quad x \mapsto \{\gamma \in \Gamma : \beta(\gamma)x = x\},$$

is Γ–equivariant. If μ is a probability measure on X which is preserved by β, then the pushforward measure $\mathrm{Stab}^\beta_* \mu$ is invariant under conjugation. It is the prototype of an invariant random subgroup (IRS). When μ is atomless and the stabilizer map is essentially injective (a.k.a. the action β is *totally nonfree*), the support of the associated IRS $\mathrm{Stab}^\beta_*(\mu)$ has no isolated points: it is a perfect set. The largest closed subspace of $\mathrm{Sub}(\Gamma)$ with no isolated points is called the *perfect kernel* of $\mathrm{Sub}(\Gamma)$. We say that an ergodic probability measure preserving (p.m.p.) action is *totipotent* when the support of its IRS is equal to the perfect kernel of $\mathrm{Sub}(\Gamma)$. By ergodicity, the following stronger property holds: *almost every element of the associated IRS has dense orbit in the perfect kernel*; see Proposition 2.3. We call such an IRS *totipotent*.

Given a p.m.p. action $\Gamma \curvearrowright^\beta (X, \mu)$, we consider the associated *p.m.p. equivalence relation*

$$\mathscr{R}^\beta := \{(x, y) \in X \times X : \beta(\Gamma)x = \beta(\Gamma)y\},$$

and its *full group* $[\mathscr{R}^\beta]$ as the group of all measure-preserving transformations whose graph is contained in $\mathscr{R}^\beta$. The (bi-invariant) *uniform distance* between two measure-preserving transformations S and T is defined by

$$d_u(T, S) := \mu(\{x \in X : S(x) \neq T(x)\}).$$

It endows the full group $[\mathscr{R}^\beta]$ with a Polish group structure. The *cost* is a numerical invariant attached to the equivalence relation $\mathscr{R}^\beta$. If β is a p.m.p. action of the free group $\boldsymbol{F}_r$ on r generators, then *the cost of $\mathscr{R}^\beta$ is exactly r when β is free, and the cost of $\mathscr{R}^\beta$ is $< r$ when β is nonfree*; see Gaboriau [8].

The main result of Le Maître [15] is that for any ergodic p.m.p. equivalence relation $\mathscr{R}$, *if $\mathscr{R}$ has cost $< r$ for some integer $r \geq 2$, then there exists a homomorphism $\tau : F_r \to [\mathscr{R}]$ with dense image.*

This result has been sharpened in order to ensure that *the homomorphism τ is injective.* Actually, the associated (almost everywhere defined) p.m.p. action α_τ can be made to satisfy the following two opposite conditions: *high faithfulness* and *amenability on μ–almost every orbit*; see Le Maître [18].

These two conditions can be phrased in terms of the support of the IRS associated to the action: the first one means that the support contains the trivial subgroup, and one can show that the second one is equivalent to the support containing a coamenable subgroup (which in the construction of [18] is the kernel of a certain surjective homomorphism $F_r \to \mathbb{Z}$).

The purpose of the present paper is to show that the homomorphism can be chosen so that the support of the associated IRS is actually the largest perfect subspace of $\mathrm{Sub}(F_r)$, which consists of all its infinite-index subgroups; see Proposition 2.1.

Theorem *Let $\mathscr{R}$ be an ergodic p.m.p. equivalence relation whose cost is $< r$ for some integer $r \geq 2$. Then there exists a homomorphism $\tau : F_r \to [\mathscr{R}]$ whose image is dense and whose associated p.m.p. action α_τ is totipotent.*

The density in $[\mathscr{R}]$ of the image of τ implies that $\mathscr{R}^{\alpha_\tau} \simeq \mathscr{R}$ and that the stabilizer map $\mathrm{Stab}^{\alpha_\tau}$ is essentially injective [18, Proposition 2.4]. In particular, the actions $F_r \curvearrowright (\mathrm{Sub}(F_r), \mathrm{Stab}^{\alpha_\tau}_* \mu)$ and $F_r \curvearrowright^{\alpha_\tau} (X, \mu)$ are conjugate (thus produce the same equivalence relation) and almost every subgroup for the IRS $\mathrm{Stab}^{\alpha_\tau}_* \mu$ equals its own normalizer. It follows that, up to isomorphism, *every p.m.p. ergodic equivalence relation of cost $< r$ comes from a totipotent IRS of F_r* (actually, from continuum many different totipotent IRSs of F_r; see Remark 5.1).

Such a statement is optimal since p.m.p. equivalence relations of cost $\geq r$ cannot come from a nonfree F_r action. To our knowledge, it was not even clear until now whether F_r admits ergodic totipotent IRSs. Since there are continuum many pairwise nonisomorphic ergodic p.m.p. equivalence relations of cost $< r$, our approach provides continuum many pairwise distinct ergodic totipotent IRSs of the free group on r generators, whose associated equivalence relations are even nonisomorphic.

Another interesting fact about totipotent p.m.p. F_r–actions is that they have no minimal model, ie they cannot be realized as minimal actions on a compact space. Indeed, it

follows from a result of Glasner and Weiss [10, Corollary 4.3] that as soon as the support of the IRS of a given p.m.p. action contains two distinct minimal subsets (eg when it contains two distinct fixed points), the action does not admit a minimal model; see Theorem 2.5. In our case the perfect kernel of $\mathrm{Sub}(F_r)$ contains a continuum of fixed points (namely, all infinite-index normal subgroups), so that totipotent p.m.p. actions of F_r are actually very far from admitting a minimal model.

Let us now recall the context around our construction. The term IRS was coined by Abert, Glasner and Virag [1] and has become an important subject on its own at the intersection of group theory, probability theory and dynamical systems. The notion of IRS is a natural generalization of a normal subgroup, especially in the direction of superrigidity type results. It has thus been present implicitly in the work of many authors, a famous landmark being the Stuck–Zimmer theorem [21], which gives examples of groups admitting very few IRSs. On the contrary, some groups admit a "zoo" of IRSs, starting with free groups; see Bowen [2] and, for other examples, Bowen, Grigorchuk and Kravchenko [3; 4] and Kechris and Quorning [13].

In particular, Bowen proved that every p.m.p. ergodic equivalence relation of cost $< r$ comes from some IRS of F_r. He obtained this result through a Baire category argument which required that the first generator act freely. In particular, such IRSs can never be totipotent.

Eisenmann and Glasner [7] then used homomorphisms $F_r \to [\mathscr{R}]$ with dense image so as to obtain interesting IRSs of F_r. They proved that given a homomorphism $\Gamma \to [\mathscr{R}]$ with dense image, the associated IRS is always cohighly transitive almost surely, which means that for almost every $\Lambda \leq \Gamma$, the Γ–action on Γ/Λ is n–transitive for every $n \in \mathbb{N}$. They also showed that the IRSs of F_r obtained by Bowen for cost 1 equivalence relations are faithful and, moreover, almost surely coamenable.

The third author [18] then used a modified version of his result on the topological rank of full groups to show that every p.m.p. ergodic equivalence relation of cost $< r$ comes from a coamenable, cohighly transitive and faithful IRS of F_r. Also in this construction, the first generator continues to act freely, thus preventing totipotency. Let us now briefly explain how our new construction (Section 5) allows us to circumvent this.

The main idea is to use a smaller set $Y \subsetneq X$ such that the restriction of $\mathscr{R}$ to Y still has cost $< r$, so that we can find some homomorphism $F_r \to [\mathscr{R}_{\restriction Y}]$ with dense image.

This provides us with some extra space in order to obtain totipotency via a well-chosen perturbation of the above homomorphism.

This perturbation is obtained by mimicking all Schreier balls on $X \setminus Y$ and then merging these amplifications with the action on Y so as to obtain both density in $[\mathscr{R}]$ and totipotency. The use of evanescent pairs of topological generators (see Definition 4.1) with Theorem 4.5 and Proposition 3.8 will grant us that this perturbation maintains the density. We establish in Theorem 4.6 that the existence of an evanescent pair of topological generators is equivalent to $\mathscr{R}$ having cost 1.

Finally, let us mention the case of the free group on infinitely many generators $\boldsymbol{F}_\infty$. Here, the space of subgroups is already perfect (see Proposition 2.1), and one can easily adapt our arguments to show that: *For every ergodic p.m.p. equivalence relation $\mathscr{R}$, there exists a homomorphism $\tau : \boldsymbol{F}_\infty \to [\mathscr{R}]$ whose image is dense and whose associated p.m.p. action α_τ is totipotent.*

This result could, however, also be obtained by a purely Baire-categorical argument: it is not hard to see that the space of such homomorphisms is dense G_δ in the Polish space of all homomorphisms $\tau : \boldsymbol{F}_\infty \to [\mathscr{R}]$.

Going back to the case of finite rank, it is not even true that a generic homomorphism $\tau : \boldsymbol{F}_r \to [\mathscr{R}]$ generates the equivalence relation $\mathscr{R}$. In order to hope for a similar genericity statement, one should first answer the following question.

Question Consider a p.m.p. ergodic equivalence relation $\mathscr{R}$ of cost $< r$. Is it true that, in the space of homomorphisms $\tau : \boldsymbol{F}_r \to [\mathscr{R}]$ whose image generates $\mathscr{R}$, those with dense image are dense?

The fact that Bowen and then Eisenmann and Glasner had to work in the even smaller space where the first generator acts freely, indicates that a Baire-categorical approach to our main result is out of reach at the moment, if not impossible.

2 Perfect kernel for groups and minimal models

Let Γ be a countable discrete group. The topology on its space of subgroups $\mathrm{Sub}(\Gamma)$ admits $V(\mathcal{I}, \mathcal{O}) := \{\Lambda \in \mathrm{Sub}(\Gamma) : \mathcal{I} \subseteq \Lambda \text{ and } \mathcal{O} \cap \Lambda = \varnothing\}$ as a basis of open sets, where $\mathcal{I}$ and $\mathcal{O}$ are finite subsets of Γ. By the Cantor–Bendixson theorem, $\mathrm{Sub}(\Gamma)$ decomposes in a unique way as the disjoint union of a perfect set, called the *perfect kernel* $\mathcal{K}(\Gamma)$ of $\mathrm{Sub}(\Gamma)$, and of a countable set. We indicate some isolation properties of subgroups:

(1) If $\Lambda \in \mathrm{Sub}(\Gamma)$ is not finitely generated, then writing $\Lambda = (\lambda_j)_{j \in \mathbb{N}}$ we obtain Λ as the nontrivial limit of the infinite index (both in Λ and in Γ) of the finitely generated subgroups $\Lambda_n := \langle \lambda_0, \lambda_1, \ldots, \lambda_n \rangle$.

(2) If Γ is finitely generated, then its finite-index subgroups are isolated. Indeed, a finite-index subgroup Λ is finitely generated as well and it is alone in the open subset defined by a finite family $\mathcal{I}$ of generators and a finite family $\mathcal{O}$ of representatives of its cosets Γ/Λ except $\{\Lambda\}$.

(3) If Γ is not finitely generated, then its finite-index subgroups are also not finitely generated and thus are not isolated by property (1).

Let us denote by $\mathrm{Sub}_{\infty i}(\Gamma)$ the subspace of infinite-index subgroups of Γ. The following is probably well known, but we were not able to locate a proof in the literature.

Proposition 2.1 *For the free group F_r on r generators, with $2 \leq r \leq \infty$:*

(i) *For finite $r \geq 2$, $\mathcal{K}(F_r) = \mathrm{Sub}_{\infty i}(F_r)$.*

(ii) *For r infinite, $\mathcal{K}(F_\infty) = \mathrm{Sub}(F_\infty)$.*

Proof We first show that if $\Lambda \in \mathrm{Sub}_{\infty i}(F_r)$, with $2 \leq r \leq \infty$, then it is a nontrivial limit of finitely generated infinite-index subgroups of F_r. If Λ is not finitely generated, then property (1) above applies. Thus, assume Λ is finitely generated. If r is infinite, then Λ has infinite index in some finitely generated noncyclic free subgroup $\Lambda \leq \Lambda * F_2 \leq F_\infty$. We can thus assume that the rank $r \geq 2$ is finite. By the Hall theorem, Λ is a free factor of a finite-index subgroup $\Lambda * \Delta$ of the free group F_r (we include the case $\Lambda = \{1\}$). Since Λ has infinite index, Δ is nontrivial. If $g \in \Delta$ is a nontrivial element, then Λ is the nontrivial limit of the sequence of finitely generated infinite-index subgroups $(\Lambda * \langle g^n \rangle)_{n \geq 2}$ of F_r.

This (with property (2), and property (3) above, respectively) shows that $\mathcal{K}(F_r) = \mathrm{Sub}_{\infty i}(F_r)$ for $r < \infty$ and $\mathcal{K}(F_\infty) = \mathrm{Sub}(F_\infty)$. $\qquad \square$

Remark 2.2 This also shows that the Cantor–Bendixson rank of $\text{Sub}(\boldsymbol{F_r})$ equals 1 when r is finite and equals 0 when $r = \infty$.

Computations of the perfect kernel for some other groups have been performed in [3; 20].

The following is a classical result:

Assume Γ acts by homeomorphisms on a Polish space Z and ν is an ergodic Γ–invariant probability measure on Z. Then the orbit of ν–almost every point $z \in Z$ is dense in the support of ν.

In particular:

Proposition 2.3 *If $\Gamma \curvearrowright (X, \mu)$ is a p.m.p. ergodic action on a standard probability space, then the stabilizer $\text{Stab}(x)$ of almost every point $x \in X$ has dense Γ–orbit in the support of the associated IRS $\nu = \text{Stab}_* \mu$ of $\text{Sub}(\Gamma)$.*

Thus, our main theorem produces IRSs on $\text{Sub}(\boldsymbol{F_r})$ for which almost every $\boldsymbol{F_r}$–orbit (under conjugation) is dense in $\mathcal{K}(\boldsymbol{F_r}) = \text{Sub}_{\infty i}(\boldsymbol{F_r})$. In other words, for almost every subgroup Λ, the Schreier graph of the action $\boldsymbol{F_r} \curvearrowright \boldsymbol{F_r}/\Lambda$ contains arbitrarily large copies of Schreier balls of every infinite transitive $\boldsymbol{F_r}$–action.

Remark 2.4 In the introduction, we defined an IRS to be totipotent when almost every subgroup has dense orbit in the perfect kernel. But an IRS can also be considered as a p.m.p. dynamical system whose associated IRS can be different. The connections between the two notions of totipotency are unclear to us. However, since the actions that we construct are totally nonfree, this situation does not happen and our IRSs are totipotent in both senses.

Moreover, this proposition can be combined with [10, Corollary 4.3] to give the following result.

Theorem 2.5 *Let $\Gamma \curvearrowright (X, \mu)$ be a p.m.p. ergodic action on a standard probability space. Suppose that the support of the associated IRS contains at least two distinct minimal subsets. Then the action has no minimal model.*

This is in wide contrast with free actions of countable groups: they always admit minimal models [22].

Proof By the previous proposition, the orbit closure of the stabilizer of μ–almost every point is equal to the support of the IRS, and hence contains two distinct minimal subsets. Admitting a minimal model would thus be incompatible with [10, Corollary 4.3]. □

3 Full groups and density

We fix once and for all a standard probability space (X, μ) and denote by $\mathrm{Aut}(X, \mu)$ the group of all its measure-preserving transformations, two such transformations being identified if they coincide on a full measure set. In order to ease notation, we will always neglect what happens on null sets. Given an element $T \in \mathrm{Aut}(X, \mu)$, its set of fixed points is denoted by

$$\mathrm{Fix}(T) := \{x \in X : T(x) = x\}.$$

A *partial isomorphism* of (X, μ) is a partially defined Borel bijection $\varphi \colon \mathrm{dom}\,\varphi \to \mathrm{rng}\,\varphi$, with $\mathrm{dom}\,\varphi$ and $\mathrm{rng}\,\varphi$ Borel subsets of X, such that φ is measure-preserving for the measures induced by μ on its domain $\mathrm{dom}\,\varphi$ and its range $\mathrm{rng}\,\varphi$. In particular, we have $\mu(\mathrm{dom}\,\varphi) = \mu(\mathrm{rng}\,\varphi)$. The *support* of φ is the set

$$\mathrm{supp}\,\varphi := \{x \in \mathrm{dom}\,\varphi : \varphi(x) \neq x\} \cup \{x \in \mathrm{rng}\,\varphi : \varphi^{-1}(x) \neq x\}.$$

Given two partial isomorphisms with φ, ψ disjoint domains and ranges, one can form their *union*, which is the partial isomorphism

$$\varphi \sqcup \psi \colon \mathrm{dom}\,\varphi \sqcup \mathrm{dom}\,\psi \to \mathrm{rng}\,\varphi \sqcup \mathrm{rng}\,\psi, \quad x \mapsto \begin{cases} \varphi(x) & \text{if } x \in \mathrm{dom}\,\varphi, \\ \psi(x) & \text{if } x \in \mathrm{dom}\,\psi. \end{cases}$$

A *graphing* is a countable set of partial isomorphisms Φ. Its *cost* $\mathcal{C}(\Phi)$ is the sum of the measures of the domains of its elements, which is also equal to the sum of the measures of their ranges since they preserve the measure.

Given a graphing Φ, the smallest equivalence relation which contains all the graphs of the elements of Φ is denoted by $\mathscr{R}_\Phi$ and called the *equivalence relation generated* by Φ. When $\Phi = \{\varphi\}$, we also write it as $\mathscr{R}_\varphi$ and call it the equivalence relation generated by φ.

The equivalence relations that can be generated by graphings are called *p.m.p. equivalence relations*; they are Borel as subsets of $X \times X$ and have countable classes. The *cost* $\mathcal{C}(\mathscr{R})$ of a p.m.p. equivalence relation $\mathscr{R}$ is the infimum of the costs of the graphings which generate it.

Whenever $\alpha\colon \Gamma \to \mathrm{Aut}(X, \mu)$ is a p.m.p. action, we denote by $\mathscr{R}^{\alpha}$ the equivalence relation generated by $\alpha(\Gamma)$.

Given a p.m.p. equivalence relation $\mathscr{R}$, the set of partial isomorphisms whose graph is contained in $\mathscr{R}$ is denoted by $[\![\mathscr{R}]\!]$ and called the *pseudo full group* of $\mathscr{R}$. Here is a useful way of obtaining elements of the pseudo full group that we will use implicitly. Say that $\mathscr{R}$ is *ergodic* when every Borel $\mathscr{R}$–saturated set has measure 0 or 1. Under this assumption, given any two Borel subsets $A, B \subseteq X$ of equal measure, there is $\varphi \in [\![\mathscr{R}]\!]$ such that $\mathrm{dom}\,\varphi = A$ and $\mathrm{rng}\,\varphi = B$ [12, Lemma 7.10].

The *full group* of $\mathscr{R}$ is the subgroup $[\mathscr{R}]$ of $\mathrm{Aut}(X, \mu)$ consisting of almost everywhere defined elements of the pseudo full group. Endowed with the *uniform metric* given by $d_u(S, T) = \mu(\{x \in X : S(x) \neq T(x)\})$, it becomes a Polish group. Observe that $d_u(T, \mathrm{id}_X) = \mu(\mathrm{supp}\,T)$.

For more material about this section, we refer for instance to [12; 9].

3.1 Around a theorem of Kittrell and Tsankov

In this paper, we will be interested in p.m.p. actions $\tau\colon \boldsymbol{F}_r \to [\mathscr{R}]$ with dense image in $[\mathscr{R}]$. To that end, the following result of Kittrell and Tsankov is very useful. Given a family $(\mathscr{R}_i)$ of equivalence relations on the same set X, we define $\bigvee_{i \in I} \mathscr{R}_i$ as the smallest equivalence relations which contains each $\mathscr{R}_i$.

Theorem 3.1 [14, Theorem 4.7] *Let $\mathscr{R}$ be a p.m.p. equivalence relation on (X, μ), and suppose that $(\mathscr{R}_i)_{i \in I}$ is a family of Borel subequivalence relations such that $\mathscr{R} = \bigvee_{i \in I} \mathscr{R}_i$. Then $[\mathscr{R}] = \langle \bigcup_{i \in I} [\mathscr{R}_i] \rangle$.*

We will also use two easy corollaries of their result, which require us to set up a bit of notation.

Definition 3.2 Given an equivalence relation $\mathscr{R}$ on a set X and $Y \subseteq X$, we define the equivalence relation $\mathscr{R}_{\restriction Y}$ *restricted* to Y and the equivalence relation $\mathscr{R}_{\updownarrow Y}$ *induced* on Y by

$$\mathscr{R}_{\restriction Y} := \mathscr{R} \cap Y \times Y = \{(x, y) \in \mathscr{R} : x, y \in Y\} \subseteq Y \times Y,$$

$$\mathscr{R}_{\updownarrow Y} := \mathscr{R}_{\restriction Y} \cup \{(x, x) : x \in X\} \subseteq X \times X.$$

Observe that given a p.m.p. equivalence relation $\mathscr{R}$, we have a natural way of identifying the full group of the restriction $\mathscr{R}_{\restriction Y}$ with the full group of the induced equivalence relation $\mathscr{R}_{\updownarrow Y}$ by making its elements act trivially outside of Y.

Corollary 3.3 *Let $\mathscr{R}$ be an ergodic p.m.p. equivalence relation on (X, μ). Let $T \in [\mathscr{R}]$ and $Y \subseteq X$ be measurable and such that $\mu(Y \cap TY) > 0$. Put $Y_T := \bigcup_{n \in \mathbb{Z}} T^n Y$. Then $\overline{\langle T, [\mathscr{R}_{\updownarrow Y}] \rangle} \geq [\mathscr{R}_{\updownarrow Y_T}]$.*

Proof Since $\mu(Y \cap TY) > 0$ and $\mathscr{R}$ is ergodic, we have that $\mathscr{R}_{\updownarrow Y \cup TY} = \mathscr{R}_{\updownarrow Y} \vee \mathscr{R}_{\updownarrow TY}$. Therefore Theorem 3.1 implies that

$$\overline{\langle [\mathscr{R}_{\updownarrow Y}], T[\mathscr{R}_{\updownarrow Y}]T^{-1} \rangle} = [\mathscr{R}_{Y \cup TY}].$$

Now observe that $(Y \cup TY) \cap T(Y \cup TY) \supseteq TY$ has positive measure. Therefore Theorem 3.1 implies that $\overline{\langle T, [\mathscr{R}_{\updownarrow Y}] \rangle}$ contains $[\mathscr{R}_{\updownarrow(Y \cup TY \cup T^2 Y)}]$, and the corollary follows by induction. $\square$

Corollary 3.4 *Consider an ergodic p.m.p. equivalence relation $\mathscr{R}$ on (X, μ) and let $Y \subseteq X$ be a positive-measure subset. Let α be a p.m.p. action of Γ on (X, μ) such that $\alpha(\Gamma) \leq [\mathscr{R}]$, $\mu(\alpha(\Gamma)Y) = 1$ and $[\mathscr{R}_{\updownarrow Y}] \leq \overline{\alpha(\Gamma)}$. Then either $\overline{\alpha(\Gamma)} = [\mathscr{R}]$, or Γ preserves a finite partition $\{Y_i\}_{i=1}^k$ of X, with $Y \subseteq Y_1$ and $[\mathscr{R}_{\updownarrow Y_i}] \leq \overline{\alpha(\Gamma)}$ for each $i \leq k$.*

In particular, if $\mu(Y) > \frac{1}{2}$, then $k = 1$ and hence $\overline{\alpha(\Gamma)} = [\mathscr{R}]$.

Proof Let $B \supset Y$ be a subset of maximal measure such that $\overline{\alpha(\Gamma)} \geq [\mathscr{R}_{\updownarrow B}]$. Then by the above corollary, for every $\gamma \in \Gamma$ such that $\alpha(\gamma)B \neq B$ we must have $\mu(B \cap \alpha(\gamma)B) = 0$; hence B is an atom of a finite partition preserved by the Γ–action α. $\square$

3.2 From graphings to density

The following is a slight variation of [15, Definition 8].

Definition 3.5 Let $n \geq 2$. A *precycle* of *length* n is a partial isomorphism φ such that if we set $B := \text{dom}\,\varphi \setminus \text{rng}\,\varphi$ (the *basis* of the precycle), then $\{\varphi^i(B)\}_{i=0,\ldots,n-2}$ is a partition 0 of dom φ, and $\{\varphi^i(B)\}_{i=1,\ldots,n-1}$ is a partition of rng φ.

We say that $T \in \text{Aut}(X, \mu)$ *extends* φ if $Tx = \varphi x$ for every $x \in \text{dom}(\varphi)$.

Observe that a precycle of length 2 is an element $\varphi \in [\![\mathscr{R}]\!]$ such that $\text{dom}(\varphi) \cap \text{rng}(\varphi) = \varnothing$. If φ is a precycle of length n, then $\mu(\text{supp}\,\varphi) = n\mu(B)$ and $\mu(\text{dom}\,\varphi) = (n-1)\mu(B)$.

An n–*cycle* is a measure-preserving transformation all of whose orbits have cardinality either 1 or n. Given a precycle φ of length n, we can extend it to an n–cycle $U_\varphi \in [\mathscr{R}_\varphi]$ as follows:

$$U_\varphi(x) := \begin{cases} \varphi(x) & \text{if } x \in \operatorname{dom} \varphi, \\ \varphi^{-(n-1)}(x) & \text{if } x \in \operatorname{rng} \varphi \setminus \operatorname{dom} \varphi, \\ x & \text{otherwise.} \end{cases}$$

This n–cycle U_φ is called the *closing cycle* of φ and $\operatorname{supp} U_\varphi = \operatorname{supp} \varphi$.

Remark 3.6 If $\{\varphi_1, \ldots, \varphi_{n-1}\}$ is a pre-n–cycle in the sense of [15, Definition 8], then $\varphi_1 \sqcup \cdots \sqcup \varphi_{n-1}$ is a precycle of length n in our sense; and if φ is a precycle of length n in our sense, then $\{\varphi_{\restriction \varphi^i(B)} : i = 0, \ldots, n-2\}$ is a pre-n–cycle in the sense of [15, Definition 8]. The reason for this change of terminology will become apparent in the statement of the next lemma, which was proved for $U = U_\varphi$ in [15, Proposition 10].

Lemma 3.7 *Suppose that φ is a precycle of basis B, let $\psi := \varphi_{\restriction B}$, and suppose that $U \in \operatorname{Aut}(X, \mu)$ extends φ. Then $[\mathscr{R}_\varphi]$ is contained in the closure of the group generated by $[\mathscr{R}_\psi] \cup \{U\}$.*

Proof Let n be the length of φ. For $i = 0, \ldots, n-2$, let $\psi_i = \varphi_{\restriction \varphi^i(B)}$. Then we have $\mathscr{R}_\varphi = \bigvee_{i=0}^{n-2} \mathscr{R}_{\psi_i}$. Since U extends φ, we have $U \psi_i U^{-1} = \psi_{i+1}$ for all $i = 0, \ldots, n-3$, and hence $U[\mathscr{R}_{\psi_i}]U^{-1} = [\mathscr{R}_{\psi_{i+1}}]$. Since $\psi_0 = \psi$, the group generated by $U \cup [\mathscr{R}_\psi]$ contains $[\mathscr{R}_{\psi_i}]$ for all $i = 0, \ldots, n-2$. Theorem 3.1 finishes the proof. □

The following proposition is obtained by a slight modification of the proof of the main theorem of [15].

Proposition 3.8 *Let $\mathscr{R}$ be a p.m.p. ergodic equivalence relation on X and let $Y \subseteq X$ be a positive measure subset. Let $\mathscr{R}_0 \leq \mathscr{R}_{\updownarrow Y}$ be a hyperfinite equivalence relation whose restriction to Y is ergodic (and trivial on $X \setminus Y$). Suppose that $\mathcal{C}(\mathscr{R}_{\updownarrow Y}) < r\mu(Y)$ for some integer $r \geq 2$. Then there are $r - 1$ precycles $\varphi_2, \varphi_3, \ldots, \varphi_r \in [\![\mathscr{R}_{\updownarrow Y}]\!]$ such that $\mu(\operatorname{supp}(\varphi_i)) < \mu(Y)$ and such that whenever $U_2, U_3, \ldots, U_r \in [\mathscr{R}]$ extend $\varphi_2, \varphi_3, \ldots, \varphi_r$, we have $\overline{\langle [\mathscr{R}_0], U_2, U_3, \ldots, U_r \rangle} \geq [\mathscr{R}_{\updownarrow Y}]$.*

For instance, one can take $U_2, U_3, \ldots, U_r$ to be the closing cycles of $\varphi_2, \varphi_3, \ldots, \varphi_r$.

Proof Let $T \in [\mathscr{R}_0]$ be such that its restriction to Y is ergodic. Our assumption $\mathcal{C}(\mathscr{R}_{\updownarrow Y}) < r\mu(Y)$ means that the normalized cost of the restriction $\mathscr{R}_{\restriction Y}$ is less than r.

Lemma III.5 from [8] then provides a graphing Φ on Y of normalized cost $< (r-1)$ such that $\{T_{\restriction Y}\} \cup \Phi$ generates the restriction $\mathscr{R}_{\restriction Y}$. We now view Φ as a graphing on X, so that $\{T\} \cup \Phi$ generates $\mathscr{R}_{\updownarrow Y}$, and $\mathcal{C}(\Phi) < (r-1)\mu(Y)$. Let $c := \mathcal{C}(\Phi)/(r-1) < \mu(Y)$. We take $p \in \mathbb{N}$ so large that $c(p+2)/p < \mu(Y)$.

Pick $\psi \in [\![\mathscr{R}_0]\!]$ a precycle of length 2 whose domain B has measure c/p. By cutting and pasting the elements of Φ and by conjugating them by elements of $[\mathscr{R}_0]$, we may as well assume that $\Phi = \{\varphi_2, \ldots, \varphi_r\}$, where each φ_i is a precycle of length $p+2$ extending ψ of basis B, whose support is a strict subset of Y. Assume that $U_i \in [\mathscr{R}]$ extends φ_i for every $i = 2, 3, \ldots, r$. Since $\psi \in [\![\mathscr{R}_0]\!]$, then $[\mathscr{R}_\psi] \le [\mathscr{R}_0]$. We can apply Lemma 3.7 and obtain that the closure of the group generated by $[\mathscr{R}_0]$ and U_i contains $[\mathscr{R}_{\varphi_i}]$. Since $\mathscr{R}_{\updownarrow Y} = \mathscr{R}_0 \vee \mathscr{R}_{\varphi_2} \vee \cdots \vee \mathscr{R}_{\varphi_r}$, we can conclude the proof of the theorem using Theorem 3.1. $\qquad\square$

Remark 3.9 We have a lot of freedom in constructing the precycles $\varphi_2, \varphi_3, \ldots, \varphi_r$ of Proposition 3.8. To start with, their length can be chosen to be any integer $n = p+2$ large enough that $c/\mu(Y) < (n-2)/n$. Actually, they could even have been chosen with any (possibly different) lengths $n_2, n_3, \ldots, n_r$, integers large enough that $c/\mu(Y) < (n_j - 2)/n_j$: simply pick $r-1$ precycles $\psi_j \in [\![\mathscr{R}_0]\!]$ of length 2 whose domain B_j has measure $c/(n_j - 2)$, and proceed as in the proof above.

In particular, the periodic closing cycles $U_2, U_3, \ldots, U_r$ can be assumed to have any large enough period $n_2, n_3, \ldots, n_r$ and domains contained in Y of measure $< \mu(Y)$. Up to conjugating by elements of $[\mathscr{R}_0]$, one can further assume that the closing cycles have a nonnull common subset of fixed points in Y:

$$\mu(\mathrm{Fix}(U_2) \cap \mathrm{Fix}(U_3) \cap \cdots \cap \mathrm{Fix}(U_r) \cap Y) > 0.$$

4 Evanescent pairs and topological generators

In this section our main goal is to obtain two topological generators of the full group of a hyperfinite ergodic equivalence relation with new flexibility properties relying on the following definition.

Definition 4.1 A pair (T, V) of elements of the full group $[\mathscr{R}]$ of the p.m.p. equivalence relation $\mathscr{R}$ is called an *evanescent pair of topological generators* of $\mathscr{R}$ if

(1) V is periodic, and

(2) for every $n \in \mathbb{N}$, the full group $[\mathscr{R}]$ is topologically generated by the conjugates of V^n by the powers of T, ie $\overline{\langle T^j V^n T^{-j} : j \in \mathbb{Z} \rangle} = [\mathscr{R}]$.

In particular, if (T, V) is an evanescent pair of topological generators, then

- the pair (T, V) topologically generates $[\mathcal{R}]$,
- (T, V^n) is an evanescent pair of topological generators for any $n \in \mathbb{N}$,
- $d_u(V^{n!}, \mathrm{id}_X)$ tends to 0 when n tends to ∞.

We will show in Theorem 4.5 that the odometer T_0 can be completed to form an evanescent pair (T_0, V) of topological generators for $\mathcal{R}_{T_0}$, and that the set of possible V is actually a dense G_δ.

In this section, we set $X = \{0, 1\}^{\mathbb{N}}$, and endow it with the Bernoulli $\frac{1}{2}$ measure $\mu = \left(\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1\right)^{\otimes \mathbb{N}}$. Given $s \in \{0, 1\}^n$, we define the basic clopen set

$$N_s := \{x \in \{0, 1\}^{\mathbb{N}} : x_i = s_i \text{ for } 1 \le i \le n\}.$$

The *odometer* T_0 is the measure-preserving transformation of this space defined as adding the binary sequence $(1, 0, 0, \dots)$ with carry to the right. More precisely, for each sequence $x \in \{0, 1\}^{\mathbb{N}}$, if k is the (possibly infinite) first integer such that $x_k = 0$, then $y = T_0(x)$ is defined by

$$y_n := \begin{cases} 0 & \text{if } n < k, \\ 1 & \text{if } n = k, \\ x_n & \text{if } n > k. \end{cases}$$

For each $n \in \mathbb{N}$, the permutation group $\mathrm{Sym}(\{0, 1\}^n)$ has a natural action α_n on $\{0, 1\}^{\mathbb{N}} \simeq \{0, 1\}^n \times \{0, 1\}^{\mathbb{N}}$ given for $x \in \{0, 1\}^{\mathbb{N}}$ and $\sigma \in \mathrm{Sym}(\{0, 1\}^n)$ by

$$\alpha_n(\sigma)(x_1, \dots, x_n, x_{n+1}, \dots) := (\sigma(x_1, \dots, x_n), x_{n+1}, \dots).$$

The sequence $(\alpha_n(\mathrm{Sym}(\{0, 1\}^n)))_{n \in \mathbb{N}}$ is an increasing sequence of subgroups of the full group $[\mathcal{R}_{T_0}]$ whose reunion is dense in $[\mathcal{R}_{T_0}]$; see [11, Proposition 3.8].

We now define a sequence of involutions $U_n \in [\mathcal{R}_{T_0}]$ with disjoint supports as in [17, Section 4.2]: $U_n := \alpha_n(\upsilon_n)$, where $\upsilon_n \in \mathrm{Sym}(\{0, 1\}^n)$ is the 2–point support transposition that exchanges $0^{n-1}1$ and $1^{n-1}0$. Observe that U_n is the involution with support $N_{1^{n-1}0} \sqcup N_{0^{n-1}1}$ (of measure 2^{-n+1}) which is equal to T_0 on $N_{1^{n-1}0}$ and T_0^{-1} on $N_{0^{n-1}1}$.

Recall that if $\tau_n \in \mathrm{Sym}(\{0, 1\}^n)$ is 2^n–cycle and w_n is a transposition which exchanges two τ_n–consecutive elements, then the group $\mathrm{Sym}(\{0, 1\}^n)$ is generated by the conjugates of w_n by powers of τ_n (actually $2^n - 1$ of them are enough). A straightforward modification gives the following lemma; see [17, Lemma 4.3] for a detailed proof.

Lemma 4.2 *For every $n \in \mathbb{N}$, the group $\alpha_n(\mathrm{Sym}(\{0, 1\}^n))$ is contained in the group generated by the conjugates of U_n by powers of T_0.*

Given a periodic p.m.p. transformation U and $k \in \mathbb{N}$, we say that V is a kth *root* of U when $\mathrm{supp}\, U = \mathrm{supp}\, V$ and $V^k = U$. The following lemma is well-known.

Lemma 4.3 *Whenever $\mathscr{R}$ is an ergodic equivalence relation, every periodic element in $[\mathscr{R}]$ admits a kth root in $[\mathscr{R}]$.*

Proof Let us first prove that every n–cycle $U \in [\mathscr{R}]$ admits a kth root. To this end, pick a fundamental domain A for the restriction of U to its support. Since $\mathscr{R}$ is ergodic, we can pick a k–cycle $V \in [\mathscr{R}]$ supported on A. Let B be a fundamental domain for V, and put $C := A \setminus B$. Then it is straightforward to check that $W \in [\mathscr{R}]$, defined as follows, is a kth root of U:

$$W(x) := \begin{cases} UU^iVU^{-i}(x) & \text{if } x \in U^i(B), \\ U^iVU^{-i}(x) & \text{if } x \in U^i(C), \\ x & \text{otherwise.} \end{cases}$$

In the general case, one glues together the kth roots obtained for every $n \in \mathbb{N}$ by considering the restrictions of U to U–orbits of cardinality n. $\qquad\square$

Remark 4.4 The same proof works more generally for *aperiodic* p.m.p. equivalence relations.

Theorem 4.5 *The set of $V \in [\mathscr{R}_{T_0}]$ such that (T_0, V) is an evanescent pair of topological generators of $\mathscr{R}_{T_0}$ is a dense G_δ subset of $[\mathscr{R}_{T_0}]$.*

Proof Denote by $\mathcal{P}$ the set of periodic elements of $[\mathscr{R}_{T_0}]$. It is a direct consequence of Rokhlin's lemma that $\mathcal{P}$ is dense in $[\mathscr{R}_{T_0}]$. And similarly the subset $\mathcal{P}' \subseteq \mathcal{P}$ of $V \in [\mathscr{R}_{T_0}]$ with finite order (or equivalently, with bounded orbit size) is dense in $[\mathscr{R}_{T_0}]$.

Writing $\mathcal{P}$ as the intersection (over the positive integers q) of the open sets

$$\left\{ V \in [\mathscr{R}_{T_0}] : d(V^{p!}, \mathrm{id}_X) < \frac{1}{q} \text{ for some } p \in \mathbb{N} \right\}$$

shows that $\mathcal{P}$ is a G_δ subset of $[\mathscr{R}_{T_0}]$.

Denote by $\mathcal{E}$ the set of $V \in [\mathscr{R}_{T_0}]$ such that for every n, the group $[\mathscr{R}_{T_0}]$ is topologically generated by conjugates of V^n by powers of T_0. We want to show that $\mathcal{P} \cap \mathcal{E}$ is

dense G_δ, and since $\mathcal{P}$ is dense G_δ it suffices (by the Baire category theorem in the Polish group $[\mathscr{R}_{T_0}]$) to show that $\mathcal{E}$ is dense G_δ.

For every $m, n \in \mathbb{N}$, set

$$\mathcal{E}_{m,n} := \{V \in [\mathscr{R}_{T_0}] : \alpha_n(\mathrm{Sym}(\{0,1\}^n)) \leq \overline{\langle T_0^k V^m T_0^{-k} : k \in \mathbb{Z} \rangle}\}.$$

The density of the union of the $\alpha_n(\mathrm{Sym}(\{0,1\}^n))$ in $[\mathscr{R}_{T_0}]$ recalled above implies that $\mathcal{E} = \bigcap_{m,n \in \mathbb{N}} \mathcal{E}_{m,n}$. So it suffices to show that each $\mathcal{E}_{m,n}$ is dense G_δ.

Let us first check that each $\mathcal{E}_{m,n}$ is G_δ. Denote by W the subgroup of $F_2 = \langle a_1, a_2 \rangle$ generated by the conjugates of a_2 by powers of a_1. So for $w = w(a_1, a_2) \in W$ and $V \in [\mathscr{R}_{T_0}]$, the element $w(T_0, V^m)$ is a product of conjugates of V^m by powers of T_0. By the definition of the closure we can write $\mathcal{E}_{m,n}$ as

$$\mathcal{E}_{m,n} = \bigcap_{p \in \mathbb{N}} \bigcap_{\sigma \in \mathrm{Sym}(\{0,1\}^n)} \bigcup_{w \in W} \left\{ V \in [\mathscr{R}_{T_0}] : d_u(w(T_0, V^m), \sigma) < \frac{1}{p} \right\}.$$

Since the map $V \mapsto w(T_0, V)$ is continuous, each of the above right-hand sets is open, so their union over $w \in W$ is also open, and we conclude that $\mathcal{E}_{m,n}$ is G_δ.

To check the density, it suffices to show that, for each m, n, one can approximate arbitrary elements of $\mathcal{P}'$ by elements of $\mathcal{E}_{m,n}$. So let $U \in \mathcal{P}'$ and let $\epsilon > 0$. Denote by K the order of U. Pick $p \geq n$ such that $2^{-p} K < \frac{1}{2}\epsilon$. Let A be the U–saturation of the support of $U_p = \alpha_p(\upsilon_p)$ (defined at the beginning of the section). The measure of A is at most ϵ. Finally, let V be a $(Km)^{\text{th}}$ root of U_p and define

$$\widetilde{U}(x) := \begin{cases} U(x) & \text{if } x \in X \setminus A, \\ V(x) & \text{if } x \in A. \end{cases}$$

By construction $d_u(U, \widetilde{U}) \leq \mu(A) < \epsilon$. Observe that $\widetilde{U}^{Km} = (\widetilde{U}^m)^K = U_p$; thus, [Lemma 4.2](#) yields that $\widetilde{U} \in \mathcal{E}_{Km,p} \subseteq \mathcal{E}_{m,p}$. Since $p \geq n$, $\widetilde{U} \in \mathcal{E}_{m,p} \subseteq \mathcal{E}_{m,n}$, so we are done. $\square$

Let us make a few comments on the above result. First, one can check that the pair (T_0, V) produced in the construction of [17, Theorem 4.2] provides an explicit example of an evanescent pair of topological generators of $\mathscr{R}_{T_0}$. Also, the above proof can be adapted to show that any *rank one* p.m.p. ergodic transformation [19, Section 8] can be completed to form an evanescent pair of topological generators; see [16, Theorem 5.28] for an explicit example of a pair which is evanescent. Proving these results is beyond the scope of this paper, so we leave it as an exercise for the interested reader.

It is unclear whether every p.m.p. ergodic transformations can be completed to form an evanescent pair of topological generators for its full group. Nevertheless, we can characterize the existence of an evanescent pair as follows.

Theorem 4.6 *Let $\mathscr{R}$ be an ergodic p.m.p. equivalence relation. Then $\mathscr{R}$ admits an evanescent pair of topological generators if and only if $\mathscr{R}$ has cost 1.*

Proof If $\mathscr{R}$ admits an evanescent pair (T, V), then since V is periodic we have $\mu(\text{supp } V^{n!}) \to 0$. Since any set of topological generators for $[\mathscr{R}]$ generates the equivalence relation $\mathscr{R}$, we conclude that $\mathscr{R}$ has cost 1.

As for the converse, Theorems 4 and 5 from [6] provide an ergodic hyperfinite subequivalence relation which is isomorphic to that of the odometer. So we can pick a conjugate of the odometer $T \in [\mathscr{R}]$. Repeating the proof of Theorem 4.5, we see that the set $\mathcal{E}_T$ of $V \in [\mathscr{R}]$ such that for every $n \in \mathbb{N}$, $[\mathscr{R}_T]$ is *contained* in $\overline{\langle T^j V^n T^{-j} : j \in \mathbb{Z} \rangle}$, is dense G_δ in $[\mathscr{R}]$.

Let us now consider the set $\mathcal{E}_{\mathscr{R}}$ of $V \in [\mathscr{R}]$ such that (T, V) is an evanescent pair of topological generators of $\mathscr{R}$, and for $n \in \mathbb{N}$ the set $\mathcal{E}_n$ of $V \in [\mathscr{R}]$ such that V is periodic and $\overline{\langle T^j V^n T^{-j} : j \in \mathbb{Z} \rangle} = [\mathscr{R}]$. Each $\mathcal{E}_n$ is G_δ by the same argument as in the proof of Theorem 4.5. Since $\mathcal{E}_{\mathscr{R}} = \bigcap_n \mathcal{E}_n$, it suffices to show that each $\mathcal{E}_n$ is dense in order to apply the Baire category theorem and finish the proof.

Let us fix $n \in \mathbb{N}$. Since $\mathcal{E}_T$ is dense in $[\mathscr{R}]$, we only need to approximate elements of $\mathcal{E}_T$ by elements of $\mathcal{E}_n$. Moreover, the set of $V \in [\mathscr{R}]$ such that $\mu(\text{supp } V) < 1$ is open and dense, so we only need to approximate every $V \in \mathcal{E}_T$ with $\mu(\text{supp } V) < 1$ by elements of $\mathcal{E}_n$.

So let $V \in \mathcal{E}_T$ with $\mu(\text{supp } V) < 1$, and take $\epsilon > 0$.

Lemma III.5 from [8] yields a graphing Φ of cost $< \frac{1}{3} \min(\epsilon, \mu(X \setminus \text{supp } V))$ such that $\{T\} \cup \Phi$ generates $\mathscr{R}$, since $\mathscr{R}$ has cost 1. Conjugating by elements of $[\mathscr{R}_T]$ and pasting the elements of Φ, we may as well assume that $\Phi = \{\varphi\}$, where $\mu(\text{dom } \varphi) < \frac{1}{3}\epsilon$, and φ is a precycle of length 2 whose support is disjoint from $\text{supp } V$. We then pick $\psi \in [\![\mathscr{R}_T]\!]$ such that $\varphi \sqcup \psi$ is a precycle of length 3 of support disjoint from $\text{supp } V$, and denote by U_1 the associated 3–cycle.

Now let U_2 be an nth root of U_1 and let $V_2 := V U_2$. Then $d_u(V_2, V) < \epsilon$, and we claim that V_2 belongs to $\mathcal{E}_n$. In order to prove this, let us denote by G the closed group generated by the conjugates of V_2^n by powers of T.

Since U_2 and V have disjoint support, they commute, and so $V_2^n = U_2^n V^n = U_1 V^n$. So $(V_2^n)^3 = V^{3n}$, and since $V \in \mathcal{E}_T$, we have that $[\mathcal{R}_T] \leq G$. In particular $[\mathcal{R}_\psi] \leq G$, and conjugating by V_2^n (which acts as U_1 on supp U_1), we get that $[\mathcal{R}_\varphi] \leq G$; see also Lemma 3.7. Since $\mathcal{R} = \mathcal{R}_T \vee \mathcal{R}_\psi$, we conclude by Theorem 3.1 that G contains $[\mathcal{R}]$, as wanted. □

5 Proof of the main theorem

As shown in Proposition 2.1, the perfect kernel of $\mathrm{Sub}(F_r)$ for $1 < r < \infty$ is the space of infinite-index subgroups. We will construct a p.m.p. action of F_r for which almost every Schreier graph contains all possible balls of Schreier graphs of transitive F_r–actions on infinite sets.

Step 1 (using a smaller subset) We start with a p.m.p. ergodic equivalence relation $\mathcal{R}$ on (X, μ) of cost $< r$. By the induction formula [8, Proposition II.6], there is a subset $Y \subseteq X$ such that $\frac{1}{2} < \mu(Y) < 1$ and such that the (normalized) cost of the restriction $\mathcal{R}_{\upharpoonright Y}$ is still $< r$. Thus the cost of the induced equivalence relation $\mathcal{R}_{\updownarrow Y}$ is $< r\mu(Y)$.

Using results of Dye [6, Theorems 4 and 5] as in the proof of Theorem 4.6, one can pick a conjugate of the odometer $T \in [\mathcal{R}_{\upharpoonright Y}]$. We view T as an element of $[\mathcal{R}_{\updownarrow Y}]$.

Now we apply Proposition 3.8 (where $\mathcal{R}_0 = \mathcal{R}_T$) to obtain precycles $\varphi_2, \ldots, \varphi_r \in [\![\mathcal{R}_{\upharpoonright Y}]\!]$ whose supports have measure $< \mu(Y)$. For $i \leq r$, we let U_i be the closing cycle of φ_i as defined after Definition 3.5. Set $\eta := \mu(Y \setminus \mathrm{supp}\, U_2) > 0$. Let m_0 be a positive integer such that $\mu(X \setminus Y)/m_0 < \frac{1}{2}\eta$.

Step 2 (preparing the finite actions) Let $(G_n)_{n \geq 1}$ be an enumeration of the (finite radius) balls of the Schreier graphs of all the transitive F_r–actions over an infinite set, up to labeled graph isomorphism, and for which the number of vertices satisfies $|G_n| \geq m_0$.

Since G_n comes from a transitive action over an infinite set, we can choose some $\ell \in \{1, \ldots, r\}$ and some $\zeta_n \in G_n$ so that there is no a_ℓ–labeled edge whose source is equal to ζ_n.

Pick $\delta_n, \xi_n \notin G_n$, set $G'_n := G_n \sqcup \{\delta_n, \xi_n\}$ and add an a_ℓ–edge from ζ_n to δ_n, an a_1–edge from δ_n to ξ_n and an a_2–edge from ξ_n to itself.

In this way we obtain a finite *partial Schreier graph*, and this can be extended to a genuine Schreier graph of an $F_r = \langle a_1, \ldots, a_r \rangle$–action on the same set as follows: for each $i \in \{1, 2, \ldots, r\}$, we consider the connected components of the subgraph obtained

by keeping only the edges labeled a_i. These are either cycles (we don't modify them) or oriented segments (possibly reduced to a single vertex), in which case we add one edge labeled a_i from the end to the beginning of the segment.

Therefore we obtain an *action ρ_n of $\boldsymbol{F}_r$ on the finite set G'_n and a special point $\xi_n \in G'_n \setminus G_n$ such that $\rho_n(a_2)\xi_n = \xi_n$.*

Step 3 (defining the action) Set $C := X \setminus Y$. Consider a partition $C = \bigsqcup_{n \geq 1} C_n$, where $\mu(C_n) > 0$ for every n. We are going to define an amplified version of the action ρ_n on C_n as follows.

For each $n \geq 1$, we take a measurable partition $C_n = \bigsqcup_{g \in G'_n} B_n^g$ such that $\mu(B_n^g)|G'_n| = \mu(C_n)$ for every $g \in G'_n$. Set $B_n := B_n^{\xi_n}$. Using ergodicity of $\mathscr{R}$, for every $g \in G'_n \setminus \{\xi_n\}$ we choose $\psi_g \colon B_n \to B_n^g$ in the pseudo full group $[\![\mathscr{R}]\!]$ of $\mathscr{R}$. In this way we obtain an action α_n of $\boldsymbol{F}_r$, defined on C_n by the formula

$$\text{if} \quad x \in B_n^{g_0} \text{ and } \rho_n(\gamma)g_0 = g_1, \quad \text{then} \quad \alpha_n(\gamma)x := \psi_{g_1}\psi_{g_0}^{-1}(x),$$

and trivial on $X \setminus C_n$. Thus, $\alpha_n(\boldsymbol{F}_r) \leq [\mathscr{R}_{\updownarrow C_n}]$.

Gluing all the α_n together, we obtain an action α_∞ of $\boldsymbol{F}_r$ on X with the properties that $\alpha_\infty(\boldsymbol{F}_r) \leq [\mathscr{R}_{\updownarrow C}]$ and α_∞ restricted to C_n is α_n.

Let $T \in [\mathscr{R}_{\updownarrow Y}]$ be the conjugate of the odometer introduced in Step 1. Theorem 4.5 states that the set of $V \in [\mathscr{R}_T]$ such that (T, V) is an evanescent pair of generators for $\mathscr{R}_T$ is dense so we can choose such a V with $\mu(\operatorname{supp} V) < \frac{1}{2}\eta$. Let $W \in [\mathscr{R}_T]$ be such that $\mu\big(\operatorname{supp}(WU_2W^{-1}) \cap \operatorname{supp} V\big) = 0$. Set

- $B := \bigcup_n B_n$, and note that $\mu(B) \leq \mu(C)/m_0 < \frac{1}{2}\eta$;
- $D := \operatorname{supp}(WU_2W^{-1}) \cup \operatorname{supp} V$, and observe that $\mu(Y \setminus D) > \frac{1}{2}\eta$.

Therefore there exists a subset $A \subseteq Y \setminus D$ of measure $\mu(A) = \mu(B)$. Let $I \in [\mathscr{R}]$ be an involution with support $A \cup B$ and which exchanges A and B.

We finally define the desired action α of $\boldsymbol{F}_r$ by setting

$$\begin{aligned}
\alpha(a_1) &:= T\alpha_\infty(a_1), \\
\alpha(a_2) &:= V(WU_2W^{-1})(I\alpha_\infty(a_2)), \\
\alpha(a_i) &:= U_i\alpha_\infty(a_i) \quad \text{for } i \geq 3.
\end{aligned}$$

See Figure 1 for the action of $\alpha(a_2)$. Note that a_2 is the only generator of $\boldsymbol{F}_r$ which does not leave the set Y invariant, because of the presence of the involution I in the definition of its action.
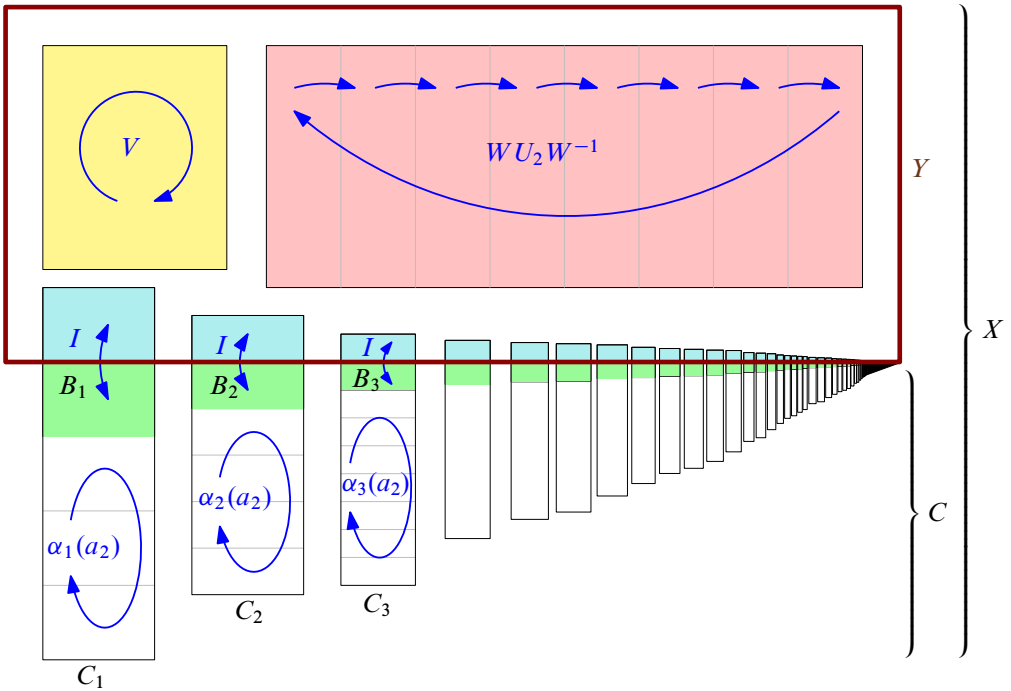
Figure 1: The action of $\alpha(a_2)$ on X.

Step 4 (density) (a) We claim that $\overline{\alpha(F_r)} \geq [\mathscr{R}_T]$.

Indeed let $S \in [\mathscr{R}_T]$ and fix $\epsilon > 0$. There exists n_0 such that if we set $C_{>n_0} := \bigcup_{n>n_0} C_n$, then $\mu(C_{>n_0}) < \frac{1}{2}\epsilon$. The elements $U_2, I, \alpha_1(a_2), \ldots, \alpha_{n_0}(a_2)$ have uniformly bounded orbits. So we can pick $k \in \mathbb{N}$ so that U_2^k, I^k and $\alpha_1(a_2)^k, \ldots, \alpha_{n_0}(a_2)^k$ are the identity.

By construction V, WU_2W^{-1}, I and $\alpha_\infty(a_2)$ have mutually disjoint supports and hence commute. Therefore $\alpha(a_2)^k = V^k\alpha_\infty(a_2)^k$.

The crucial assumption that (T, V) is an evanescent pair of generators now comes into play: there is a word $w(T, V^k)$, which is a product of conjugates of V^k by powers of T, such that $d_u(w(T, V^k), S) < \frac{1}{2}\epsilon$. We remark that $\alpha(a_1)$ acts on Y the same way as T, and that $\alpha(a_2)^k$ acts on Y the same way as V^k. Also note that $\alpha(a_1)$ preserves each C_j while $\alpha(a_2)^k$ is the identity on each C_j for $j = 1, 2, \ldots, n_0$, so that for all $m \in \mathbb{Z}$, the transformation $\alpha(a_1)^m\alpha(a_2)^k\alpha(a_1)^{-m}$ acts on $C_1 \cup C_2 \cup \cdots \cup C_{n_0}$ as the identity.

It now follows from the fact that w is a product of conjugates of V^k by powers of T that $w(T, V^k)$ and $w(\alpha(a_1), \alpha(a_2)^k)$ coincide on Y and can only differ on $C_{>n_0}$, which

has measure less than $\frac{1}{2}\epsilon$. Hence $d_u(w(\alpha(a_1), \alpha(a_2)^k), S) < \epsilon$, which implies that $\overline{\alpha(F_r)} \geq [\mathscr{R}_T]$.

(b) We claim that $\overline{\alpha(F_r)} \geq [\mathscr{R}_{\updownarrow Y}]$.

Recall that $\alpha(a_2) = V(WU_2W^{-1})(I\alpha_\infty(a_2))$, where V, WU_2W^{-1} and $I\alpha_\infty(a_2)$ have pairwise disjoint support. Since U_2 extends φ_2, we get that $W^{-1}\alpha(a_2)W$ extends φ_2. By assumption $\alpha(a_3), \ldots, \alpha(a_r)$ extend $\varphi_3, \ldots, \varphi_r$, respectively. Moreover, $W \in [\mathscr{R}_T]$, so the claim follows from [Proposition 3.8](#):

$$\overline{\alpha(F_r)} \geq \overline{\langle [\mathscr{R}_T], W^{-1}\alpha(a_2)W, \alpha(a_3), \ldots, \alpha(a_r) \rangle} \geq [\mathscr{R}_{\updownarrow Y}].$$

(c) We claim that $\overline{\alpha(F_r)} \geq [\mathscr{R}]$.

This is a direct consequence of [Corollary 3.4](#) granted that $\alpha(\Gamma)Y = X$, which we will now show.

Clearly $\alpha(a_2)Y \supset \alpha(a_2)A = B = \bigcup_n B_n$. For every n and $g \in G'_n \setminus \{\xi_n\}$, there exists $\gamma \in \Gamma$ of minimal length such that $\rho_n(\gamma)\xi_n = g$. Since $\rho_n(a_2)\xi_n = \xi_n$ and since $\alpha(a_2)_{\upharpoonright C_n \setminus B_n} = \alpha_n(a_2)$ mimics the action of $\rho_n(a_2)$ on $G'_n \setminus \{\xi_n\}$, the minimality of the length of γ implies that $\alpha(\gamma)B_n = B_n^g$. Since this is true for every $g \in G'_n$ we get $\alpha(\Gamma)Y \supset C_n$; and this holds for every n. We thus have $\alpha(\Gamma)Y = X$ as wanted.

Step 5 (totipotency) Consider a transitive action ρ of F_r on some infinite set. Let H be a Schreier ball such that $|H| \geq m_0$. Then by construction there exists n such that $H = G_n \subseteq G'_n$. We also remark that the restriction of the Schreier graph of the action α to $\bigcup_{g \in G_n} B_n^g \subseteq C_n$ mimics the partial Schreier graph H. Since ρ and H are arbitrary, the Schreier graph of α contains every sufficiently large Schreier ball of every transitive action of F_r, and this finishes the proof of the main theorem. $\square$

Remark 5.1 The subspace $Y \subseteq X$ chosen in [Step 1](#) of the above proof coincides with the subset where $\alpha(a_1)$ is aperiodic. So the event "no power of a_1 belongs to Λ" has measure $\mu(Y)$ in the IRS $\text{Stab}_*^\alpha \mu$ associated with α. The measure $\mu(Y)$ can be chosen to take any value from the nonempty interval

$$\left(\max\left\{ \frac{\mathcal{C}(R) - 1}{r - 1}, \frac{1}{2} \right\}, 1 \right).$$

Recalling that the density of $\alpha(\Gamma)$ implies Stab^α is essentially injective, then the following holds: *Every ergodic p.m.p. equivalence relation $\mathscr{R}$ of cost $< r$ can be realized (up to a null set) by the action of $F_r \curvearrowright \text{Sub}(F_r)$ for continuum many different totipotent IRSs of F_r.*

References

[1] **M Abért**, **Y Glasner**, **B Virág**, *Kesten's theorem for invariant random subgroups*, Duke Math. J. 163 (2014) 465–488 MR Zbl

[2] **L Bowen**, *Invariant random subgroups of the free group*, Groups Geom. Dyn. 9 (2015) 891–916 MR Zbl

[3] **L Bowen**, **R Grigorchuk**, **R Kravchenko**, *Invariant random subgroups of lamplighter groups*, Israel J. Math. 207 (2015) 763–782 MR Zbl

[4] **L Bowen**, **R Grigorchuk**, **R Kravchenko**, *Characteristic random subgroups of geometric groups and free abelian groups of infinite rank*, Trans. Amer. Math. Soc. 369 (2017) 755–781 MR Zbl

[5] **M L Condic**, *Totipotency: what it is and what it is not*, Stem Cells and Development 23 (2014) 796–812

[6] **H A Dye**, *On groups of measure preserving transformations, I*, Amer. J. Math. 81 (1959) 119–159 MR Zbl

[7] **A Eisenmann**, **Y Glasner**, *Generic IRS in free groups, after Bowen*, Proc. Amer. Math. Soc. 144 (2016) 4231–4246 MR Zbl

[8] **D Gaboriau**, *Coût des relations d'équivalence et des groupes*, Invent. Math. 139 (2000) 41–98 MR Zbl

[9] **D Gaboriau**, *Orbit equivalence and measured group theory*, from "Proceedings of the International Congress of Mathematicians, III" (R Bhatia, A Pal, G Rangarajan, V Srinivas, M Vanninathan, editors), Hindustan, New Delhi (2010) 1501–1527 MR Zbl

[10] **E Glasner**, **B Weiss**, *Uniformly recurrent subgroups*, from "Recent trends in ergodic theory and dynamical systems" (S Bhattacharya, T Das, A Ghosh, R Shah, editors), Contemp. Math. 631, Amer. Math. Soc., Providence, RI (2015) 63–75 MR Zbl

[11] **A S Kechris**, *Global aspects of ergodic group actions*, Mathematical Surveys and Monographs 160, Amer. Math. Soc., Providence, RI (2010) MR Zbl

[12] **A S Kechris**, **B D Miller**, *Topics in orbit equivalence*, Lecture Notes in Math. 1852, Springer (2004) MR Zbl

[13] **A S Kechris**, **V Quorning**, *Co-induction and invariant random subgroups*, Groups Geom. Dyn. 13 (2019) 1151–1193 MR Zbl

[14] **J Kittrell**, **T Tsankov**, *Topological properties of full groups*, Ergodic Theory Dynam. Systems 30 (2010) 525–545 MR Zbl

[15] **F Le Maître**, *The number of topological generators for full groups of ergodic equivalence relations*, Invent. Math. 198 (2014) 261–268 MR Zbl

[16] **F Le Maître**, *Sur les groupes pleins préservant une mesure de probabilité*, PhD thesis, ENS Lyon (2014) Available at http://math.univ-lyon1.fr/~melleray/these_FLM.pdf

[17] **F Le Maître**, *On full groups of non-ergodic probability-measure-preserving equivalence relations*, Ergodic Theory Dynam. Systems 36 (2016) 2218–2245 MR Zbl

[18] **F Le Maître**, *Highly faithful actions and dense free subgroups in full groups*, Groups Geom. Dyn. 12 (2018) 207–230 MR Zbl

[19] **D S Ornstein**, **D J Rudolph**, **B Weiss**, *Equivalence of measure preserving transformations*, Mem. Amer. Math. Soc. 262, Amer. Math. Soc., Providence, RI (1982) MR Zbl

[20] **R Skipper**, **P Wesolek**, *On the Cantor–Bendixson rank of the Grigorchuk group and the Gupta–Sidki 3 group*, J. Algebra 555 (2020) 386–405 MR Zbl

[21] **G Stuck**, **R J Zimmer**, *Stabilizers for ergodic actions of higher rank semisimple groups*, Ann. of Math. 139 (1994) 723–747 MR Zbl

[22] **B Weiss**, *Minimal models for free actions*, from "Dynamical systems and group actions", Contemp. Math. 567, Amer. Math. Soc., Providence, RI (2012) 249–264 MR Zbl

Fakultät für Mathematik, Institut für Algebra und Geometrie, Karlsruhe Institute of Technology
Karlsruhe, Germany

Unité de Mathématiques Pures et Appliquées, École Normale Supérieure de Lyon
Lyon, France

Institut de Mathématiques de Jussieu-PRG, Université de Paris
Paris, France

alessandro.carderi@kit.edu,   damien.gaboriau@ens-lyon.fr,
francois.le-maitre@imj-prg.fr

The infimum of the dual volume of
convex cocompact hyperbolic 3–manifolds

Filippo Mazzoli

We show that the infimum of the dual volume of the convex core of a convex co-compact hyperbolic 3–manifold with incompressible boundary coincides with the infimum of the Riemannian volume of its convex core, as we vary the geometry by quasi-isometric deformations. We deduce a linear lower bound of the volume of the convex core of a quasi-Fuchsian manifold in terms of the length of its bending measured lamination, with optimal multiplicative constant.

Introduction

Let M be a complete hyperbolic 3–manifold and let CM be its convex core, namely the smallest nonempty convex subset of M. Then M is said to be *convex cocompact* if CM is a compact subset. The notion of dual volume of the convex core $V_C^*(M)$ arises from the polarity correspondence between the hyperbolic and the de Sitter spaces; see Schlenker [36, Section 1] and Mazzoli [28]. If M is a convex cocompact hyperbolic 3–manifold, then $V_C^*(M)$ coincides with $V_C(M) - \frac{1}{2}\ell_m(\mu)$, where $V_C(M)$ stands for the usual Riemannian volume of the convex core and $\ell_m(\mu)$ denotes the length of the bending measured lamination μ with respect to the hyperbolic metric m of the boundary of the convex core of M. Our aim is to study the infimum of V_C^*, considered as a function over the space $\mathcal{QD}(M)$ of quasi-isometric deformations of a given convex cocompact hyperbolic 3–manifold M with incompressible boundary. In particular, we will prove:

Theorem A *For a convex cocompact hyperbolic* 3*–manifold* M *with incompressible boundary*,

$$\inf_{M' \in \mathcal{QD}(M)} V_C^*(M') = \inf_{M' \in \mathcal{QD}(M')} V_C(M').$$

Moreover, $V_C^*(M') = V_C(M')$ *if and only if the boundary of the convex core of* M' *is totally geodesic.*

When M is a quasi-Fuchsian manifold, Theorem A can be equivalently stated as

$$(1) \qquad V_C(M') \geq \tfrac{1}{2}\ell_{m'}(\mu')$$

for every $M' \in \mathcal{QD}(M)$, where $\ell_{m'}(\mu')$ is the length of the bending measure of $\partial CM'$. As a consequence of the variation formulae of V_C (see Bonahon [4]) and of V_C^* (see Mazzoli [29] and see also Krasnov and Schlenker [20]), we will see in Corollary 4.1 that the multiplicative constant $\tfrac{1}{2}$ is optimal, and is realized near the Fuchsian locus.

Theorem A is to the dual volume as the following result of Bridgeman, Brock and Bromberg is to the renormalized volume:

Theorem [9, Theorem 3.11] *For every convex cocompact hyperbolic 3–manifold M with incompressible boundary,*

$$\inf_{M' \in \mathcal{QD}(M)} V_R(M') = \inf_{M' \in \mathcal{QD}(M)} V_C(M').$$

Moreover, $V_R(M') = V_C(M')$ if and only if the boundary of the convex core of M is totally geodesic.

By work of W P Thurston, if the compact 3–manifold with boundary $N := M \cup \partial_\infty M$ is acylindrical, then there exists a unique convex cocompact structure $M_0 \in \mathcal{QD}(M)$ whose convex core has totally geodesic boundary. In [41] (see also [40]), Storm proved that the infimum of the volume of the convex core function $V_C \colon \mathcal{QD}(M) \to \mathbb{R}$ is equal to half the simplicial volume of the doubled manifold $D(N)$. Moreover, the infimum is realized exactly when N is acylindrical, and it is achieved at M_0. Theorem A and [9, Theorem 3.11] then imply that the same characterization holds true for the infimum of the dual volume and the renormalized volume, respectively. In the case of the renormalized volume V_R, such description of $\inf V_R$ was first established by Pallete [31], without making use of Storm's result. Bridgeman, Brock and Bromberg [10] recently introduced a notion of surgered gradient flow of the renormalized volume in the relatively acylindrical case, which allowed them to obtain new comparisons between the renormalized volume and the Weil–Petersson geometry of the deformation spaces of convex cocompact 3–manifolds, generalizing in particular the works of Brock [12] and Schlenker [38]. In the same work, a new proof of Storm's result in the acylindrical case is obtained as a byproduct of their analysis; see in particular [10, Corollary 6.5].

Dual volume, renormalized volume and Riemannian volume of the convex core are related by the chain of inequalities

$$V_C^*(M) := V_C(M) - \tfrac{1}{2}\ell_m(\mu) \leq V_R(M) \leq V_C(M) - \tfrac{1}{4}\ell_m(\mu) \leq V_C(M).$$

Here the second inequality is due to Schlenker [38], and the lower bound of V_R is proved in [9, Theorem 3.7]. Observe in particular that Theorem A implies [9, Theorem 3.11], which concerns the infimum of the renormalized volume. The requirement on M to have incompressible boundary is necessary; indeed, it has been shown by Pallete [32] that there exist Schottky groups with negative renormalized volume.

Our proof of Theorem A broadly follows the same strategy as the work of Bridgeman, Brock and Bromberg [9], with some necessary differences. The authors of [9] interpret the renormalized volume as a function V_R over the Teichmüller space $\mathcal{T}(\partial_\infty M)$ of the *conformal boundary at infinity* of M (by the works of Bers [1], Kra [18] and Maskit [27]), and they estimate the difference $|V_R - V_C|$ as one follows the (opposite of the) Weil–Petersson gradient flow of V_R on $\mathcal{T}(\partial_\infty M)$. In order to study the dual volume function, the analogy between the variation formula of the renormalized volume (see the work of Krasnov and Schlenker [19, Lemma 5.8], or Section 1.6) and the dual Bonahon–Schläfli formula [29] would tempt us to consider V_C^* as a function of the Teichmüller space $\mathcal{T}(\partial CM)$, seen as the deformation space of *hyperbolic structures* on the boundary of the convex core of M. However, the hyperbolic structure on ∂CM is only conjecturally thought to provide a parametrization of the quasi-isometric deformation space of M. To avoid this difficulty, we rather focus our attention on a family of functions V_k^* approximating V_C^*, for which a similar procedure is possible.

Given k, a real number in the interval $(-1, 0)$, we say that an embedded surface $\Sigma_k \subset M$ is a k*–surface* if its first fundamental form (namely the restriction of the metric of M on the tangent space to Σ_k) is a Riemannian metric with constant Gaussian curvature equal to k. Then, by the work of Labourie [21], the complementary region of the convex core of M is foliated by k–surfaces, which converge to ∂CM as k goes to -1, and tend towards the conformal boundary at infinity $\partial_\infty M$ as k goes to 0. The function $V_k^*(M)$ is then defined to be the dual volume of the region M_k of M enclosed by its k–surfaces, one for each geometrically finite end of M. By the works of Labourie [22] and Schlenker [37], the hyperbolic structures of the k–surfaces do provide a parametrization of $\mathcal{QD}(M)$, a fact that allows us to study V_k^* as a function over the Teichmüller space of ∂M_k. At this point, studying the Weil–Petersson gradient of V_k^* on $\mathcal{T}(\partial M_k)$, we prove that the difference between the dual volume and the standard volume of the regions M_k is well-behaved as one follows the lines of the flow backwards, and finally we deduce the statement of Theorem A by taking a limit for k that goes to -1. While the methods of [9] for the study of the renormalized volume heavily rely on the relations between the geometry of the boundary of the convex

core and the properties of the *Schwarzian at infinity* of $\partial_\infty M$, here we use a more analytical approach to determine the necessary bounds on the geometric quantities of the k–surfaces $\partial_k M$ of M, which will guarantee us the existence and the good behavior of the flow of the Weil–Petersson gradient vector fields of V_k^*.

Outline of the paper

After the first section of background, we suggest the reader move backwards (as with the flow of the gradient of the functions V_k^*) while going through this exposition. In Section 4 the proof of Theorem A is described. Here the analogy with the work of Bridgeman, Brock and Bromberg [9] is manifest; the required technical ingredients (Lemmas 3.4 and 3.7 and Corollary 3.6) are formally very similar to the ones developed for the renormalized volume.

Section 3 focuses on the study of the Weil–Petersson gradient $\mathrm{grad}_{\mathrm{WP}} V_k^*$ of the dual volume functions V_k^* and the proofs of the ingredients mentioned above. In Lemma 3.4 we determine a lower bound of the norm of $\mathrm{grad}_{\mathrm{WP}} V_k^*$ in terms of the integral of the mean curvature of ∂M_k (which replaces the role of the length $\ell_m(\mu)$ in the definition of the dual volume of the regions M_k). In Corollary 3.6 we show that the flow of the vector field $\mathrm{grad}_{\mathrm{WP}} V_k^*$ is defined for all times, and in Lemma 3.7 we prove the existence of a global lower bound of the dual volumes V_k^* over $\mathcal{QD}(M)$. All the proofs in this section rely on differential-geometric methods and are consequences of an explicit description of the Weil–Petersson gradient of V_k^* developed in Proposition 3.2. This presentation of the vector field $\mathrm{grad}_{\mathrm{WP}} V_k^*$ is inspired by an orthogonal decomposition of the space of symmetric tensors due to Fischer and Marsden [15], and it involves the solution u_k of a simple PDE (4) over the k–surface ∂M_k. In particular, the proof of Corollary 3.6 will require us to have uniform control of the $\mathscr{C}^2$–norm of the function u_k. Section 2 (and in particular Lemma 2.3) provides us this last ingredient, and it is essentially based on the classical regularity theory for linear elliptic differential operators (see eg Evans [14]), and on the following property of k–surfaces:

Proposition (see Proposition 2.1) *For any $k \in (0, 1)$ and $n \in \mathbb{N}$ there exists a positive constant $N_{k,n}$ such that, for every convex cocompact hyperbolic 3–manifold M and for every incompressible k–surface Σ_k in M, the $\mathscr{C}^n$–norm of the mean curvature of Σ_k is bounded above by $N_{n,k}$.*

The existence of such a universal upper bound was proved (with weaker assumptions than the ones appearing above) by Bonsante, Danciger, Maloni and Schlenker in

[6, Proposition 3.8] for $n = 0$ (and the same strategy actually shows that the statement holds for any n), and its proof heavily relies on a compactness criterion for isometric immersions of surfaces established by Labourie [21]; see also Bonsante, Danciger, Maloni and Schlenker [6, Proposition 3.6]. As will be manifest in the proof of Proposition 2.1, the constants $N_{n,k}$ that we will produce are unfortunately not explicit.

Acknowledgments

1 Preliminaries

1.1 Hyperbolic 3–manifolds

Let M be an orientable complete hyperbolic 3–manifold, namely a complete Riemannian 3–manifold with constant sectional curvature equal to -1, and let Γ be a discrete and torsion-free group of orientation-preserving isometries of the hyperbolic 3–space $\mathbb{H}^3$, such that M is isometric to $\mathbb{H}^3 / \Gamma$. We define the *limit set* of Γ to be

$$\Lambda_\Gamma := \overline{\Gamma \cdot x_0} \cap \partial_\infty \mathbb{H}^3,$$

where $\overline{\Gamma \cdot x_0}$ denotes the closure of the Γ–orbit of x_0 in $\overline{\mathbb{H}}^3 := \mathbb{H}^3 \cup \partial_\infty \mathbb{H}^3$. It is simple to see that the definition of Λ_Γ does not depend on the choice of basepoint $x_0 \in \mathbb{H}^3$. If Γ is nonelementary (it does not have any finite orbit in $\overline{\mathbb{H}}^3$), then Λ_Γ can be characterized as the smallest closed Γ–invariant subset of $\partial_\infty \mathbb{H}^3$; see eg [33, Chapter 12]. The complementary region Ω_Γ of the limit set in $\partial_\infty \mathbb{H}^3$ is called the *domain of discontinuity* of Γ.

1.2   The convex core

If $\pi: \mathbb{H}^3 \to \mathbb{H}^3/\Gamma \cong M$ denotes the universal cover of M, then a subset C of M is *convex* if and only if $\pi^{-1}(C)$ is convex in $\mathbb{H}^3$. If Γ is nonelementary, then every nonempty Γ–invariant convex subset of $\mathbb{H}^3$ contains the *convex hull* C_Γ of Γ, which consists of the intersection of all half-spaces H of $\mathbb{H}^3$ satisfying $\overline{H} \supseteq \Lambda_\Gamma$ ($\overline{H}$ stands for the closure of H in $\overline{\mathbb{H}}^3$). The image $CM := \pi(C_\Gamma)$ describes a convex subset of M, called the *convex core* of M, which is minimal among the family of nonempty convex subsets of M.

Now let M be a *convex cocompact* hyperbolic 3–manifold, namely a noncompact complete hyperbolic 3–manifold whose convex core is compact. The boundary of the convex core ∂CM of M is the union of a finite collection of connected surfaces, each of which is totally geodesic outside a subset of Hausdorff dimension 1. As described in [13], the hyperbolic metrics on the totally geodesic pieces "merge" together, defining a complete hyperbolic metric m on ∂CM. The locus where the boundary of the convex core is not flat is a geodesic lamination λ, ie a closed subset that is union of disjoint simple geodesics. The surface ∂CM is bent along λ, and the amount of bending can be described by a measured lamination μ called the *bending measure* of ∂CM. The μ–measure along an arc k transverse to λ consists of an integral sum of the exterior dihedral angles along the leaves that k meets. By locally integrating the lengths of the leaves of the lamination in $\mathrm{d}\mu$, we obtain the notion of length of the bending measure with respect to the hyperbolic structure m, which will be denoted by $\ell_m(\mu)$. For a more detailed description we refer to [13, Section II.1.11] or [2].

1.3   Incompressible boundary

When M is convex cocompact and Γ is a discrete and torsion-free subgroup of isometries of $\mathbb{H}^3$ such that $M \cong \mathbb{H}^3/\Gamma$, Γ acts freely and properly discontinuously on the domain of discontinuity Ω_Γ, and the quotient of $\mathbb{H}^3 \cup \Omega_\Gamma$ by Γ determines a natural compactification of M, which will be denoted by $\overline{M} = M \cup \partial_\infty M$. Then M is said to have *incompressible boundary* if the inclusion $S \to \overline{M}$ of each connected component S of $\partial_\infty M$ induces an injection at the level of the fundamental groups. This implies in particular that any lift of the inclusion $S \to \overline{M}$ to the universal covers $\widetilde{S} \to \widetilde{\overline{M}}$ is a homeomorphism onto its image.

1.4   Constant Gaussian curvature surfaces

Definition 1.1   Let S be an immersed surface inside a Riemannian 3–manifold N. The *first fundamental form* I of S is the Riemannian metric of S given by the restriction of

the metric of N to the tangent spaces of S. If S admits a unitary normal vector field $v: S \to T^1 N$, we define its *shape operator* B to be the endomorphism of TS given by $BU := -\mathcal{D}_U v$, for every tangent vector field U of S (here $\mathcal{D}$ denotes the Levi-Civita connection of N). The trace of the shape operator will be called the *mean curvature* of S, and the tensor $I\!I := I(B \cdot, \cdot)$ the *second fundamental form* of S.

Let Σ be a surface immersed in a hyperbolic 3–manifold M, with first and second fundamental forms I and $I\!I$, and shape operator B. We denote by K_e its *extrinsic curvature*, $K_e = \det B$, and by K_i its *intrinsic curvature*, the Gaussian curvature of the Riemannian metric I. Then the Gauss–Codazzi equations of $(\Sigma, I, I\!I)$ can be expressed as

$$K_i = K_e - 1, \qquad (\nabla_U B)V = (\nabla_V B)U \quad \text{for all } U, V,$$

where U and V are tangent vector fields to Σ, and ∇ is the Levi-Civita connection of I.

Definition 1.2 Let Σ be an immersed surface inside a hyperbolic 3–manifold, and let $k \in (-1, 0)$. If the intrinsic curvature of Σ is constantly equal to k, it is a *k–surface*.

If Σ is a k–surface, then its extrinsic curvature $K_e = k + 1$ is positive, since $k \in (-1, 0)$. In particular, Σ is a (locally) strictly convex surface.

In every convex cocompact 3–manifold M, the subset $M \setminus CM$ is the disjoint union of a finite number of geometrically finite hyperbolic ends $(E_i)_i$, each of which is homeomorphic to $\Sigma_i \times (0, \infty)$ for some compact orientable surface Σ_i of genus larger than or equal to 2. By the work of Labourie [21], the sets E_i are foliated by embedded k–surfaces $(\Sigma_{i,k})_k$, with k that varies in $(-1, 0)$. The surfaces $\Sigma_{i,k}$ approach the components of the pleated boundary ∂CM of the convex core of M as k goes to -1, and the components of conformal boundary at infinity $\partial_\infty M$ as k goes to 0.

We will denote by M_k the compact region of M whose boundary ∂M_k consists of the union of the surfaces $\bigcup_i \Sigma_{i,k}$, and we will endow ∂M_k with the second fundamental form $I\!I_k$ defined by the normal vector field pointing towards the interior of M_k, so that $I\!I_k$ is positive definite, and H_k is a positive function (observe that the eigenvalues of the shape operator have the same sign since $K_e = \det B > 0$).

1.5 Deformation spaces

Let Σ be a compact orientable surface of genus larger than or equal to 2. The *Teichmüller space* of Σ, denoted by $\mathcal{T}(\Sigma)$, is the space of isotopy classes of hyperbolic metrics

on Σ. Equivalently, in light of the uniformization theorem, $\mathcal{T}(\Sigma)$ can be described as the space of isotopy classes of conformal structures over Σ (compatible with the choice of a fixed orientation on Σ).

Since convex cocompact hyperbolic 3–manifolds are not closed, several different notions of deformation spaces can be introduced. In this exposition we will consider the *quasi-isometric* (or quasiconformal) deformation space.

Definition 1.3 Given M and M' hyperbolic manifolds, a diffeomorphism $M \to M'$ is a *quasi-isometric deformation* of M if it is globally bi-Lipschitz. We denote by $\mathcal{QD}(M)$ the space of quasi-isometric deformations of M, where we identify two deformations $M \to M'$ and $M \to M''$ if their pullback metrics are isotopic to each other.

Remark 1.4 By a theorem of Thurston [42, Proposition 8.3.4], two hyperbolic n–manifolds M and M' are quasi-isometric if and only if their fundamental groups Γ and Γ' (as subgroups of the isometry group of $\mathbb{H}^n$) are quasiconformally conjugated, ie there exists a quasiconformal self-homeomorphism φ of $\partial_\infty \mathbb{H}^n$ such that $\varphi \Gamma \varphi^{-1} = \Gamma'$.

We denote by $m_k(M) \in \mathcal{T}(\partial M_k) = \prod_i \mathcal{T}(\Sigma_i)$ the isotopy class of the hyperbolic metric $(-k)I_k$, where I_k is the first fundamental form of the k–surface $\partial_k M$ of M. Then for every $k \in (-1, 0)$ we have maps

$$T_k \colon \mathcal{QD}(M) \to \mathcal{T}(\partial M_k), \quad M \mapsto m_k(M).$$

The convenience in considering foliations by k–surfaces relies in the following result, based on the works of Labourie [22] and Schlenker [37]:

Theorem 1.5 *If M has incompressible boundary, the map T_k is a $\mathscr{C}^1$–diffeomorphism for every $k \in (-1, 0)$.*

In the compressible case a similar statement can be recovered, replacing the role of the Teichmüller space $\mathcal{T}(\partial M_k)$ with its quotient by the action of a suitable subgroup of the mapping class group of ∂M_k; see eg [26, Theorem 5.1.3] for the corresponding statement concerning the conformal structure of the boundary at infinity.

As mentioned in the introduction it is an open question, asked by Thurston, whether the same statement is true for the hyperbolic structures on the boundary of the convex core, which could be considered as the case $k = -1$ in Theorem 1.5. More precisely, the

map T_{-1} is known to be continuously differentiable by [5] and surjective by the work of Sullivan (described in [13]), but there are no results concerning its global injectivity.

1.6 Dual volume

Let M be a convex cocompact hyperbolic 3–manifold. If N is a compact convex subset of M with smooth boundary, we define the *dual volume* of N to be

$$V^*(N) := V(N) - \frac{1}{2} \int_{\partial N} H \, \mathrm{d}a,$$

where H stands for the mean curvature of ∂N defined using the inner normal vector field, and $V(N)$ is the Riemannian volume of N. We refer to [28] for a description of the relation between the notion of dual volume and the polarity correspondence between the hyperbolic and de Sitter spaces.

For every $k \in (-1, 0)$ we let $V_k^* : \mathcal{T}(\partial M_k) \to \mathbb{R}$ denote the function that associates, with a hyperbolic structure $m_k \in \mathcal{T}(\partial M_k)$, the dual volume of the region $\partial M_k'$ enclosed by the k–surfaces of the unique convex cocompact hyperbolic 3–manifold $M' = T_k^{-1}(m_k)$ whose k–surfaces have hyperbolic structure m_k.

If $(N_h)_h$ is a sequence of convex compact subsets approaching CM, then the integral of the mean curvature over ∂N_h approaches $\ell_m(\mu)$, the length of the bending measure μ with respect to the hyperbolic structure of ∂CM; see eg [9, Proposition 3.4]. This suggests we should set the *dual volume of the convex core* of M as

$$V_C^*(M) := V(CM) - \tfrac{1}{2}\ell_m(\mu).$$

In [29], a first-order variation formula for the function V_C^* over $\mathcal{QD}(M)$ is studied, called the *dual Bonahon–Schläfli formula*,

$$\mathrm{d}V_C^*(\dot{M}) = -\tfrac{1}{2} \, \mathrm{d}L_\mu(\dot{m}).$$

Here $\dot{m}$ denotes the first-order variation of the hyperbolic metric on ∂CM along $\dot{M}$, and $L_\mu : \mathcal{T}(\partial CM) \to \mathbb{R}$ is the function that associates with every hyperbolic structure m the length of the m–geodesic realization of μ.

A strong similarity between dual and renormalized volumes is displayed by their variation formulae. The renormalized volume satisfies

$$\mathrm{d}V_R(\dot{M}) = -\tfrac{1}{2} \, \mathrm{dext}_{\mathcal{F}_\infty}(\dot{c}_\infty),$$

where $\dot{c}_\infty$ denotes the first-order variation of the conformal structure on $\partial_\infty M$ along $\dot{M}$, and $\mathrm{ext}_{\mathcal{F}_\infty} : \mathcal{T}(\partial_\infty M) \to \mathbb{R}$ is the function that associates with every conformal structure

c the extremal length of the horizontal measured foliation of the Schwarzian at infinity of M with respect to c; see Schlenker [39] for a proof of this relation.

1.7 Norms on $T\mathcal{T}(\Sigma)$

First we introduce the necessary notation for the "Riemannian geometric tools" that will be used in the rest of the paper. Let (N, g) be a Riemannian manifold with Levi-Civita connection ${}^g\nabla$, and consider $(e_i)_i$, a local g–orthonormal frame. Given T, a symmetric 2–tensor on N, we define its g–divergence as the 1–form

$$(\mathrm{div}_g T)(X) := \sum_i ({}^g\nabla_{e_i} T)(e_i, X)$$

for every tangent vector field X. Similarly, the g–divergence of a vector field X is the function

$$\mathrm{div}_g X = \sum_i g({}^g\nabla_{e_i} X, e_i).$$

The Laplace–Beltrami operator can be expressed as $\Delta_g f = \mathrm{div}_g \mathrm{grad}_g f$. Given two symmetric 2–tensors T and T', their scalar product is defined as

$$(T, T')_g := g^{ij} g^{hk} T_{ih} T'_{jk} = \mathrm{tr}(g^{-1} T g^{-1} T').$$

In particular, we set $\mathrm{tr}_g T := (g, T)_g = \mathrm{tr}(g^{-1} T)$. In the next sections it will also be useful to keep in mind the way that these operators change if we replace g with λg, for some positive constant λ. If $\dim N = n$,

$$(2) \qquad \mathrm{div}_{\lambda g} T = \lambda^{-1} \mathrm{div}_g T, \quad \Delta_{\lambda g} f = \lambda^{-1} \Delta_g f, \quad \mathrm{d}a_{\lambda g} = \lambda^{n/2} \mathrm{d}a_g,$$

$$(3) \qquad (T, T')_{\lambda g} = \lambda^{-2} (T, T')_g, \quad \mathrm{tr}_{\lambda g} T = \lambda^{-1} \mathrm{tr}_g T.$$

Now let $\mathcal{M}$ be the set of Riemannian metrics on Σ, and let $\mathcal{H}$ be the subset of hyperbolic ones. The first-order variations $\dot{g}$ of elements of $\mathcal{M}$ identify with smooth symmetric 2–tensors on Σ. The choice of a metric $g \in \mathcal{M}$ determines a scalar product on $T_g\mathcal{M}$, which can be expressed as

$$(\sigma, \tau)_{\mathrm{FT}, g} := \int_\Sigma (\sigma, \tau)_g \, \mathrm{d}a_g,$$

where FT stands for Fischer–Tromba. We define $S_2^{\mathrm{tt}}(\Sigma, g)$ to be the space of those symmetric tensors σ that are traceless with respect to g (meaning $(\sigma, g)_g = 0$) and g–divergence-free (meaning $\mathrm{div}_g \sigma = 0$, as defined above). Such tensors are also called *transverse traceless*. A simple way to characterize the space $S_2^{\mathrm{tt}}(\Sigma, g)$ is through

holomorphic quadratic differentials. A holomorphic quadratic differential ϕ on (Σ, g) is a $\mathbb{C}$–valued symmetric tensor that can be locally written as $\phi = f \, dz^2$, where z is a local coordinate conformal to the metric g (and compatible with a given orientation), and $f = f(z)$ is a holomorphic function. Transverse traceless tensors are exactly those 2–tensors that can be written as $\Re\phi$ for some ϕ holomorphic quadratic differential on (Σ, h).

It is shown in [43] that, for every hyperbolic metric h, $S_2^{\mathrm{tt}}(\Sigma, h)$ coincides with

$$T_h \mathcal{H} \cap \left(T_h (\mathrm{Diff}_0(\Sigma) \cdot h) \right)^{\perp},$$

where $T_h(\mathrm{Diff}_0(\Sigma) \cdot h)$ is the tangent space to the orbit of h by the action of the group of diffeomorphisms of Σ isotopic to the identity, and $(\cdot)^{\perp}$ is taken with respect to the scalar product $(\cdot, \cdot)_{\mathrm{FT}, h}$ on $T_h \mathcal{M}$. Therefore, if $m = [h]$ denotes the isotopy class of a hyperbolic metric on Σ, we can identify $S_2^{\mathrm{tt}}(\Sigma, h)$ with $T_m \mathcal{T}(\Sigma)$, the tangent space at m to the Teichmüller space $\mathcal{T}(\Sigma) = \mathcal{H}/\mathrm{Diff}_0(\Sigma)$, seen as the space of isotopy classes of hyperbolic metrics on Σ. Moreover, the restriction of the scalar product $(\cdot, \cdot)_{\mathrm{FT}, h}$ to $S_2^{\mathrm{tt}}(\Sigma, h)$ coincides with (a multiple of) the *Weil–Petersson metric* $\langle \cdot, \cdot \rangle_{\mathrm{WP}}$ (see Lemma 1.6 for the explicit multiplicative constant).

The Teichmüller space can also be endowed with another Finsler norm that arises from its conformal (or quasiconformal) interpretation, namely the *Teichmüller norm*. The Teichmüller norm $\| \cdot \|_{\mathcal{T}}$ of a tangent vector $\dot{m} \in T_m \mathcal{T}(\Sigma)$ is the infimum of the L^∞–norms of the Beltrami differentials representing $\dot{m}$. It is not difficult to see that the Beltrami differential associated to the tangent direction $2\Re\phi$ coincides with ν_ϕ, the *harmonic Beltrami differential* associated to ϕ (see eg [16] for a detailed description of these notions, and [30, Lemma 1.2] for a direct computation of this relation). Moreover, the L^∞–norm of ν_ϕ can be computed as

$$\|\nu_\phi\|_\infty = \frac{1}{\sqrt{2}} \sup_{\Sigma} \|\Re\phi\|_h.$$

We summarize what we observed:

Lemma 1.6 *For every hyperbolic metric h representing the isotopy class $m \in \mathcal{T}(\Sigma)$, the tangent space $T_m \mathcal{T}(\Sigma)$ identifies with $S_2^{\mathrm{tt}}(\Sigma, h)$. For every $\dot{m} \in T_m \mathcal{T}(\Sigma)$,*

$$\|\dot{m}\|_{\mathrm{WP}} = \frac{1}{\sqrt{2}} \|\Re\phi\|_{\mathrm{FT}, h}, \quad \|\dot{m}\|_{\mathcal{T}} \leq \frac{1}{\sqrt{2}} \sup_{\Sigma} \|\Re\phi\|_h,$$

where ϕ is a holomorphic quadratic differential such that $2\Re\phi$ represents $\dot{m}$ inside $S_2^{\mathrm{tt}}(\Sigma, h)$.

2 Some useful estimates

In this section we determine estimates for the solution u_k of a certain linear PDEs, defined over a k–surface lying inside an end of a convex cocompact hyperbolic 3–manifold with incompressible boundary. The function u_k will be later used to describe the Weil–Petersson gradient of the dual volume functions V_k^*, and the bounds produced in this section will play an important role in the study of its flow.

Given (N, g) a Riemannian manifold with Levi-Civita connection ${}^g\nabla$ and area form $\mathrm{d}a_g$, we denote by $H^n(N, \mathrm{d}a_g)$ the Sobolev space of real-valued functions f on N with $L^2(N, \mathrm{d}a_g)$–integrable weak derivatives $({}^g\nabla)^i f$ for all $i \leq n$. The space $H^n(N, \mathrm{d}a_g)$ is Hilbert if endowed with the scalar product

$$(f, f') := \sum_{i=0}^n \int_N (({}^g\nabla)^i f, ({}^g\nabla)^i f')_g \, \mathrm{d}a_g \quad \text{for } f, f' \in H^n(N, \mathrm{d}a_g),$$

where $(\cdot, \cdot)_g$ denotes the scalar product induced by g on the space of i–tensors over N. Given $f : N \to \mathbb{R}$ a $\mathscr{C}^n$–function, we define its $\mathscr{C}^n(N, g)$–norm as

$$\|f\|_{\mathscr{C}^n(N,g)} := \sum_{i=0}^n \sup_{p \in N} \|({}^g\nabla)^i f|_p\|_g,$$

where $\|T\|_g = \sqrt{(T, T)_g}$.

Now let h_k denote the hyperbolic metric $(-k)I_k$ on the k–surface ∂M_k, with Levi-Civita connection ${}^k\nabla$ and Laplace–Beltrami operator Δ_k (here we consider $\Delta_k u$ to be the trace of the Hessian of u). We define the linear differential operator L_k to be

$$L_k u := (\Delta_k - 2\mathbb{1})u = \Delta_k u - 2u.$$

Let A be the symmetric bilinear form on $H^1(\partial M_k, \mathrm{d}a_k)$ with quadratic form

$$A(u, u) := -(L_k u, u) = \int_\Sigma (\|\mathrm{d}u\|_k^2 + 2u^2) \, \mathrm{d}a_k,$$

where $\|\cdot\|_k$ and $\mathrm{d}a_k$ denote the norm and the area form of h_k, respectively. By the Lax–Milgram theorem (see eg [7, Corollary 5.8]) applied to the Sobolev space $H^1(\partial M_k, \mathrm{d}a_k)$ and to the coercive symmetric bilinear form A, we have that, for every $f \in L^2(\partial M_k, \mathrm{d}a_k)$, there exists a unique weak solution $u \in H^1(\partial M_k, \mathrm{d}a_k)$ of the equation $L_k u = f$. We will in particular denote by u_k the function satisfying

$$(4) \qquad L_k u_k = -k^{-1} H_k \iff \Delta_{I_k} u_k + 2k u_k = H_k,$$

where H_k denotes the mean curvature of the k–surface ∂M_k. By the classical regularity theory for linear elliptic PDEs (see eg [14, Section 6.3]), the smoothness of the mean curvature H_k and the compactness of ∂M_k imply that the function u_k is smooth and it is a strong solution of (4).

By the work of Rosenberg and Spruck [35, Theorem 4], for every Jordan curve c in $\partial_\infty \mathbb{H}^3$ there exist exactly two k–surfaces $\widetilde{\Sigma}_k^{\pm}(c)$ asymptotic to c. A fundamental property of k–surfaces, which will be crucial in Lemma 2.3, is:

Proposition 2.1 [6, Proposition 3.8] *Let $k \in (-1, 0)$ and $n \in \mathbb{N}$. Then there exists a constant $N_{k,n} > 0$ such that, for every Jordan curve c in $\partial_\infty \mathbb{H}^3$, the mean curvature $H_{c,k}$ of the k–surface $\widetilde{\Sigma}_k(c) = \widetilde{\Sigma}_k^+(c) \sqcup \widetilde{\Sigma}_k^-(c)$ asymptotic to c satisfies*

$$\|H_{c,k}\|_{\mathscr{C}^n(\widetilde{\Sigma}_k(c))} \le N_{n,k}.$$

Proof We briefly recall here the proof of this statement (which was stated in [6] for $n = 0$). First, recall that k–surfaces satisfy the following compactness criterion:

Proposition 2.2 [6, Proposition 3.6] *Let $k \in (-1, 0)$, and consider $f_n \colon \mathbb{H}_k^2 \to \mathbb{H}^3$ a sequence of proper isometric embeddings of the hyperbolic plane $\mathbb{H}_k^2$ with constant Gaussian curvature k. If there exists a point $p \in \mathbb{H}^2$ such that $(f_n(p))_n$ is precompact, then there exists a subsequence of $(f_n)_n$ that converges $\mathscr{C}^\infty$–uniformly on compact sets to an isometric immersion $f \colon \mathbb{H}_k^2 \to \mathbb{H}^3$.*

Fixing $k \in (-1, 0)$ and $n \in \mathbb{N}$, assume by contradiction that there exists a sequence of Jordan curves $(c_m)_m$ such that the mean curvatures $H_m = H_{c_m,k}$ of the k–surfaces $\widetilde{\Sigma}_k(c_m)$ satisfy $\|H_m\|_{\mathscr{C}^n(\widetilde{\Sigma}_k(c_m))} > m$. Up to extracting a subsequence, there exists an $i \le n$ such that for every $m \in \mathbb{N}$

$$\sup_{\widetilde{\Sigma}_k(c_m)} \|({}^k\nabla)^i H_m\| > \frac{m}{n+1} = C_n m.$$

Now choose $q_m \in \widetilde{\Sigma}_k(c_m)$ for which the norm of $({}^k\nabla)^i H_m$ at q_m is at least $C_n m$. Since each component of $\widetilde{\Sigma}_k(c_m)$ is embedded and isometric to the hyperbolic plane $\mathbb{H}_k^2$ (which is homogeneous), we can find a sequence of proper isometric embeddings $f_m \colon \mathbb{H}_k^2 \to \mathbb{H}^3$, parametrizing a component of $\widetilde{\Sigma}_k(c_m)$, such that $f_m(\bar{p}) = q_m$ for some fixed basepoint $\bar{p} \in \mathbb{H}_k^2$. Up to postcomposing f_m by an isometry of $\mathbb{H}^3$, we can also assume that $f_m(\bar{p}) = \bar{q}$ is fixed. In this way, we have found a sequence of proper isometric embeddings $f_m \colon \mathbb{H}_k^2 \to \mathbb{H}^3$ satisfying

- $f_m(\bar{p}) = \bar{q} \in \mathbb{H}^3$ is independent of $m \in \mathbb{N}$,

- the mean curvature of the surfaces $f_m(\mathbb{H}_k^2)$ at $\bar{q}$ has some ith–order derivative that is unbounded as m goes to ∞.

This clearly contradicts the compactness criterion mentioned above.   □

From this result we can now obtain a uniform control on u_k:

Lemma 2.3 *Let M be a convex cocompact hyperbolic 3–manifold. Then the function* $u_k \colon \partial M_k \to \mathbb{R}$, *the solution of* (4), *satisfies*

$$\frac{\max_{\partial M_k} H_k}{2k} \le u_k \le \frac{\min_{\partial M_k} H_k}{2k} = \frac{\sqrt{k+1}}{k} < 0.$$

Moreover, if M has incompressible boundary, then there exists a constant $C_k > 0$ depending only on the intrinsic curvature $k \in (-1, 0)$, and in particular not on the hyperbolic structure of M, such that

$$\max_{\partial M_k} \|{}^k \nabla^2 u_k\|_k \le C_k.$$

Proof The first assertion is an immediate consequence of the maximum principle applied to u_k as a solution of the PDE (4). Moreover, since the product of the principal curvatures (the eigenvalues of the shape operator) of a k–surface is everywhere equal to $k + 1$, the trace of the shape operator is bounded from below by $2\sqrt{k+1}$ (see Remark 2.5 for an explanation of the equality $\min_{\partial M_k} H_k = 2\sqrt{k+1}$).

The proof of the second assertion requires more care. Let Σ_k be a connected component of the k–surface ∂M_k, and let $\widetilde{M} \cong \mathbb{H}^3$ denote the universal cover of M. Since M is a convex cocompact hyperbolic 3–manifold with incompressible boundary, every component $\widetilde{\Sigma}_k$ of the preimage of Σ_k in $\widetilde{M}$ is stabilized by a subgroup $\Gamma \cong \pi_1(\Sigma_k)$ of the fundamental group of M, acting by isometries on $\widetilde{M}$. Each of these subgroups Γ is quasi-Fuchsian (see eg [17, Corollary 4.112 and Theorem 8.17] for a proof of this assertion), and the surface $\widetilde{\Sigma}_k$ is a k–surface asymptotic to some Jordan curve in $\partial_\infty \widetilde{M} \cong \partial_\infty \mathbb{H}^3$. In particular, by Proposition 2.1, we can find a universal constant $N_k = N_{2,k} > 0$ that satisfies

$$(5) \qquad\qquad \|\widetilde{H}_k\|_{\mathscr{C}^2(\widetilde{\Sigma}_k)} \le N_k.$$

Here we stress that the constant N_k does not depend on the hyperbolic structure of M or Σ_k, but only on the value of $k \in (-1, 0)$.

Our goal is now to make use of this control to obtain a uniform bound of the norm of the Hessian of u_k. For this purpose, we will need a classical result of regularity for linear elliptic differential equations:

Theorem 2.4 [14, Theorem 2, page 314] *Let $m, n \in \mathbb{N}$ and $U \subset \mathbb{R}^n$ be a bounded open set. We consider a differential operator L of the form*

$$Lf = -\sum_{i,j=1}^{n} a^{ij}(x)\partial^2_{x_i, x_j} f + \sum_{i=0}^{n} b^i(x)\partial_{x_i} f + c(x)f,$$

where $a^{ij} = a^{ji}, b^i, c \in \mathscr{C}^{m+1}(U, \mathbb{R})$. Assume that L is uniformly elliptic, ie there exists a constant $\varepsilon > 0$ such that $\sum_{i,j} a^{ij}(x)v_i v_j \geq \varepsilon \|v\|^2$ for all $v \in \mathbb{R}^n$ and $x \in U$. If $f \in H^1(U)$ is a weak solution of the equation $Lf = \lambda$ for some $\lambda \in H^m(U)$, then for every bounded open set V with closure contained in U there exists a constant C, depending only on m, U and V and the functions a^{ij}, b^i and c, such that

$$\|f\|_{H^{m+2}(V)} \leq C(\|\lambda\|_{H^m(U)} + \|f\|_{L^2(U)}),$$

where the Sobolev spaces $H^{m+2}(V), H^m(U)$ and $L^2(U)$ are defined with respect to the Euclidean metric of $U \subset \mathbb{R}^n$.

The surface $\widetilde{\Sigma}_k$ endowed with the lift of the hyperbolic metric h_k of Σ_k is isometric to the hyperbolic plane $\mathbb{H}^2$. In the rest of the proof we will identify $\widetilde{\Sigma}_k$ with the Poincaré disk model $\mathbb{H}^2 := (B_1, g)$, where B_1 is the Euclidean ball of radius 1 and center 0 in $\mathbb{C}$, and g is the Riemannian metric

$$g = \left(\frac{2}{1 - |z|^2}\right)^2 |\mathrm{d}z|^2.$$

Now we choose U and V to be the g–geodesic balls of center $0 \in B_1$ and hyperbolic radius equal to 2 and 1, respectively. The lift of the operator $-L_k$ over U is clearly uniformly elliptic because of the compactness of $\overline{U}$ and its expression in coordinates

$$-L_k f = -g^{ij}(\partial^2_{ij} f - \Gamma^h_{ij}(g)\partial_h f) + 2f,$$

where the $\Gamma^h_{ij}(g)$ denote the Christoffel symbols of g. Again by the compactness of $\overline{U}$ and $\overline{V}$, the norms of the Sobolev spaces $\|\cdot\|_{H^j(U)}$ and $\|\cdot\|_{H^j(V)}$, computed with respect to the flat connection of $B_1 \subset \mathbb{R}^2$ and the Euclidean volume form, are equivalent to the norms of the corresponding Sobolev spaces defined using the Levi-Civita connection of g and the g–volume form. Moreover, the bi-Lipschitz constants involved in the equivalence only depend on a bound of the $\mathscr{C}^{j+1}$–norm of g over U, so they can be chosen to depend only on $j \in \mathbb{N}$. From now on, we will consider the norms on the spaces $H^j(U)$ and $H^j(V)$ to be defined using the metric g and its connection.

Now we apply Theorem 2.4 to $m = n = 2$, the operator $-L_k$ and the functions $f = \tilde{u}_k$ and $\lambda = -k^{-1}\widetilde{H}_k$, where $\widetilde{F}$ denotes the lift of the function F over $\widetilde{\Sigma}_k$. We can find

a universal constant $C > 0$ (depending only on the open sets U and V, and on the metric $g|_U$) such that

$$\|\tilde{u}_k\|_{H^4(V)} \leq C(-k^{-1}\|\tilde{H}_k\|_{H^2(U)} + \|\tilde{u}_k\|_{L^2(U)}).$$

By the first part of Lemma 2.3, $\|\tilde{u}_k\|_{\mathscr{C}^0(U)} \leq -(2k)^{-1}\|\tilde{H}_k\|_{\mathscr{C}^0(\mathbb{H}^2)}$. In addition,

$$\|\tilde{u}_k\|_{L^2(U)} \leq \mathrm{Area}(U, g)^{1/2}\|\tilde{u}_k\|_{\mathscr{C}^0(U)} \leq -(2k)^{-1}\,\mathrm{Area}(U, g)^{1/2}\|\tilde{H}_k\|_{\mathscr{C}^0(\mathbb{H}^2)},$$

and

$$\|\tilde{H}_k\|_{H^2(U)} \leq \mathrm{Area}(U, g)^{1/2}\|\tilde{H}_k\|_{\mathscr{C}^2(\mathbb{H}^2)}.$$

In conclusion, we deduce that

$$\|\tilde{u}_k\|_{H^4(V)} \leq -2k^{-1}C\,\mathrm{Area}(U, g)^{1/2}\|\tilde{H}_k\|_{\mathscr{C}^2(\mathbb{H}^2)}.$$

By the Sobolev embedding theorem (see eg [7, Corollary 9.13, page 283]), given W an open set satisfying $0 \in W \subset \overline{W} \subset V$, the $\mathscr{C}^2(W)$–norm of $\tilde{u}_k$ (again, computed with respect to the Levi-Civita connection of g) is controlled by a multiple of its H^4–norm over V, and the multiplicative factor depends only on W and V. Therefore, if we choose for instance $W = B_{\mathbb{H}^2}(0, \frac{1}{2})$,

$$\|^k\nabla^2\tilde{u}_k\|_{\mathscr{C}^0(W)} \leq C'(k)\|\tilde{H}_k\|_{\mathscr{C}^2(\mathbb{H}^2)}.$$

Now the desired statement easily follows. From relation (5) and the last inequality, we obtain a uniform bound of the Hessian of $\tilde{u}_k$ over $W \ni 0$. Now let q be any other point of $\mathbb{H}^2$, and choose a g–isometry $\varphi_q \colon B_1 \to B_1$ such that $\varphi_q(0) = q$. If we replace $\tilde{u}_k$ and $\tilde{H}_k$ with $\tilde{u}_k \circ \varphi_q$ and $\tilde{H}_k \circ \varphi_q$, respectively, the exact same argument above applies, since the operator L_k and the norms $\|\cdot\|_{H^j}$ and $\|\cdot\|_{\mathscr{C}^l}$ are invariant under the action of the isometry group of $\mathbb{H}^2$ (and since $\|\tilde{H}_k\|_{\mathscr{C}^2(\mathbb{H}^2)} = \|\tilde{H}_k \circ \varphi_q\|_{\mathscr{C}^2(\mathbb{H}^2)}$). In particular, this gives us a control on the norm of $^k\nabla^2\tilde{u}_k$ over $\varphi_q(W)$ for any point $q \in \mathbb{H}^2$, and the last part of our assertion follows. $\qquad\square$

Remark 2.5 The minimum of the mean curvature $2\sqrt{k+1}$ is always realized. As described by Labourie in [23], whenever we have a k–surface Σ_k with first and second fundamental forms I_k and II_k, respectively, the identity map $\mathrm{id}\colon (\Sigma_k, II_k) \to (\Sigma_k, I_k)$ is harmonic, with Hopf differential ψ_k satisfying

$$2\Re\psi_k = I_k - \frac{H_k}{2(k+1)}II_k.$$

Its squared norm with respect to II_k can be expressed as

$$\|2\Re\psi_k\|^2_{II_k} = \frac{H_k^2 - 4(k+1)}{(k+1)^2}.$$

In particular, at each zero of ψ_k (which necessarily exist because $\chi(\Sigma_k) < 0$), we have $H_k = 2\sqrt{k+1}$.

We stress that, even if the maximum of the mean curvature H_k will clearly depend on the hyperbolic structure of M, Proposition 2.1 guarantees that max H_k is controlled by a function of k independent of the geometry of M, as long as ∂M is incompressible.

We will make use of the upper bound $u_k \le \sqrt{k+1}/k$ in Lemma 3.4, where we will determine a lower bound of the Weil–Petersson norm of the differential of V_k^* in terms of the integral of the mean curvature.

3 The gradient of the dual volume

The aim of this section is to describe the gradient of the dual volume function V_k^* with respect to the Weil–Petersson metric on the Teichmüller space of ∂M_k in terms of the function u_k studied in the previous section.

The first-order variation of the dual volume of M_k as we vary the convex cocompact hyperbolic structure of M can be computed by applying the *differential Schläfli formula* due to Rivin and Schlenker [34]. In particular:

Proposition 3.1 *We have*

$$
\mathrm{d}(V_k^* \circ T_k)(\dot{M}) = \frac{1}{4} \int_{\partial M_k} (\dot{I}_k, \mathit{II}_k - H_k I_k)_{I_k} \, \mathrm{d}a_{I_k}
$$
$$
= \frac{1}{4} \int_{\partial M_k} (\dot{h}_k, \mathit{II}_k + k^{-1} H_k h_k)_{h_k} \, \mathrm{d}a_{h_k},
$$

where $\dot{I}_k = -k^{-1}\dot{h}_k$ is the first-order variation of the first fundamental form on ∂M_k along the variation $\dot{M}$, and $T_k \colon \mathcal{QD}(M) \to \mathcal{T}(\partial M_k)$ is the diffeomorphism introduced in Section 1.5.

A proof of this relation based on the results of Rivin and Schlenker can be found in [29, Proposition 2.5]. From its variation formula, we can give an explicit description of the Weil–Petersson gradient of the dual volume function V_k^*, which will turn out to be useful for the study of its flow.

Proposition 3.2 *The vector field* $\mathrm{grad}_{\mathrm{WP}} V_k^*$ *is represented by the symmetric 2–tensor* $2\Re\phi_k$, *where ϕ_k is the (unique) holomorphic quadratic differential satisfying*

$$
\Re\phi_k = \mathit{II}_k - {}^k\nabla^2 u_k + u_k h_k,
$$

where u_k denotes the solution of (4).

Proof Let $\dot{m}_k$ denote a tangent vector to the Teichmüller space of ∂M_k at m_k. As described in Section 1.5, given any hyperbolic metric h_k representing the isotopy class $m_k \in \mathcal{T}(\partial M_k)$, we can find a unique transverse traceless tensor $\dot{h}_k \in S_2^{tt}(\Sigma, h_k)$ representing $\dot{m}_k$. Assume for a moment that we can find a decomposition of the symmetric tensor $I\!I_k + k^{-1} H_k h_k$ of the form

$$I\!I_k + k^{-1} H_k h_k = S_{tt} + \mathcal{L}_X h_k + \lambda h_k,$$

where S_{tt} is a transverse traceless tensor with respect to h_k, $\mathcal{L}_X h_k$ is the Lie derivative of h_k with respect to a vector field X, and λ is a smooth function on ∂M_k. Then, by Proposition 3.1, we can express the variation of the dual volume V_k^* along a transverse traceless variation $\dot{h}_k$:

$$dV_k^*(\dot{h}_k) = \frac{1}{4} \int_{\partial M_k} (\dot{h}_k, S_{tt} + \mathcal{L}_X h_k + \lambda h_k)_{h_k}\, da_{h_k}.$$

Since $\dot{h}_k$ is traceless, the scalar product $(\dot{h}_k, h_k)_{h_k} = \mathrm{tr}_{h_k}(\dot{h}_k)$ vanishes identically. The L^2–scalar product between $\dot{h}_k$ and $\mathcal{L}_X h_k$ vanishes too, because $\mathcal{L}_X h_k$ is tangent to the orbit of h_k by the action of $\mathrm{Diff}_0(\Sigma)$; see Section 1.7. In particular,

$$dV_k^*(\dot{h}_k) = \frac{1}{4} \int_{\partial M_k} (\dot{h}_k, S_{tt})_{h_k}\, da_{h_k} = \tfrac{1}{8}(\dot{h}_k, 2S_{tt})_{\mathrm{FT}, h_k}.$$

In light of Lemma 1.6, by varying the tangent vector $\dot{m}_k \in T_{m_k} \mathcal{T}(\partial M_k)$, we deduce that the tensor $2S_{tt}$ is the element of $S_2^{tt}(\Sigma, h_k)$ that represents $\mathrm{grad}_{\mathrm{WP}} V_k^*$.

In conclusion, this argument shows us that, in order to prove our assertion, we need to determine a decomposition of the tensor $I\!I_k + k^{-1} H_k h_k$ of the form we described above, with $S_{tt} = I\!I_k - {}^k\nabla^2 u_k + u_k h_k$. For this purpose, we consider the expression

$$\begin{aligned}
I\!I_k + k^{-1} H_k h_k &= (I\!I_k - {}^k\nabla^2 u_k + u_k h_k) + {}^k\nabla^2 u_k + (k^{-1} H_k - u_k) h_k \\
&= (I\!I_k - {}^k\nabla^2 u_k + u_k h_k) + \tfrac{1}{2}\mathcal{L}_{\mathrm{grad}_{h_k} u_k} h_k + (k^{-1} H_k - u_k) h_k,
\end{aligned}$$

where we used the relation $\mathcal{L}_{\mathrm{grad}_{h_k} u_k} h_k = 2({}^k\nabla^2 u_k)$. In this expression, the second term of the sum is of the type $\mathcal{L}_X h_k$, while the third term has the form λh_k. Therefore, by the argument above, it is enough to show that the first term is h_k–traceless and h_k–divergence-free. The trace of $I\!I_k - {}^k\nabla^2 u_k + u_k h_k$ satisfies

$$\mathrm{tr}_{h_k}(I\!I_k - {}^k\nabla^2 u_k + u_k h_k) = -k^{-1} H_k - \Delta_k u_k + 2u_k.$$

This expression vanishes because u_k is a solution of (4). In order to compute the divergence of our tensor, we will need the relations

$$\mathrm{div}_{h_k} I\!I_k = -k^{-1} dH_k, \quad \mathrm{div}_g({}^g\nabla^2 f) = d(\Delta_g f) + \mathrm{Ric}_g(\mathrm{grad}_g f, \cdot).$$

The first equality follows from the Codazzi equation $(^k\nabla_X B_k)Y = (^k\nabla_Y B_k)X$ satisfied by the shape operator B_k of ∂M_k (the Levi-Civita connections of h_k and the first fundamental form I_k are the same, since they differ by a multiplicative constant). The second relation is true for any Riemannian metric g, and we will apply it in the case $g = h_k$ and $f = u_k$. Since h_k is a hyperbolic metric on a 2–manifold, we have $\text{Ric}_{h_k} = -h_k$. Therefore

$$\text{div}_{h_k}(I\!I_k - \nabla_k^2 u_k + u_k h_k) = -k^{-1}dH_k - d(\Delta_k u_k) + du_k + du_k$$
$$= d(-k^{-1}H_k - \Delta_k u_k + 2u_k),$$

where we used the relation $\text{div}_g(fg) = df$. Again, the expression above vanishes because u_k solves (4). Then we have shown that $I\!I_k - {}^k\nabla^2 u_k + u_k h_k$ is a transverse traceless tensor, as desired. □

Remark 3.3 In fact, the decomposition we presented for the tensor $I\!I_k + k^{-1}H_k h_k$ is related to the orthogonal decomposition of the space of symmetric tensors due to Fischer and Marsden [15]. Given g, a hyperbolic metric, every symmetric tensor S admits an orthogonal decomposition of the form

$$S = S_{tt} + \mathcal{L}_X g + ((-\Delta_g f + f)g + {}^g\nabla^2 f),$$

where:

- S_{tt} is transverse traceless with respect to g.
- $S_{tt} + \mathcal{L}_X g$ is tangent to the space of Riemannian metrics with constant Gaussian curvature equal to -1. That is, if $g' \mapsto K(g')$ denotes the operator that associates to the Riemannian metric g' its Gaussian curvature, then $S_{tt} + \mathcal{L}_X g \in \ker dK_g$.
- $(-\Delta_g f + f)g + {}^g\nabla^2 f$ is L^2–orthogonal to $\ker dK_g$.

Then, the expression

$$I\!I_k + k^{-1}H_k h_k = (I\!I_k - {}^k\nabla^2 u_k + u_k h_k) + 0 + ((k^{-1}H_k - u_k)h_k + {}^k\nabla^2 u_k)$$
$$= (I\!I_k - {}^k\nabla^2 u_k + u_k h_k) + 0 + ((-\Delta_k u_k + u_k)h_k + {}^k\nabla^2 u_k)$$

is the Fischer–Marsden decomposition of $I\!I_k + k^{-1}H_k h_k$, where $f = u_k$, $X = 0$ and $S_{tt} = (I\!I_k - {}^k\nabla^2 u_k + u_k h_k)$.

Using this explicit description of the Weil–Petersson gradient of the dual volume function V_k^*, we can determine a lower bound of its norm in terms of the integral of the mean curvature:

Lemma 3.4 *For every $k \in (-1, 0)$,*

$$\|\mathrm{d}V_k^*\|_{\mathrm{WP}}^2 \geq -\frac{\sqrt{k+1}}{2k} \int_{\partial M_k} H_k \, \mathrm{d}a_{I_k} - \frac{2\pi(k+1)}{k^2} |\chi(\partial M)|.$$

Proof In what follows, we will prove the expression

$$(6) \qquad \|I\!I_k - \nabla_k^2 u_k + u_k h_k\|_{I_k}^2 = k u_k H_k - 2(k+1) + \mathrm{div}_{I_k} W$$

for some tangent vector field W on ∂M_k. Assuming for the moment this relation,

$$\begin{aligned}
\|\mathrm{d}V_k^*\|_{\mathrm{WP}}^2 &= \frac{1}{2} \int_{\partial M_k} \|\Re\phi_k\|_{h_k}^2 \, \mathrm{d}a_{h_k} \qquad \text{(Proposition 3.2 and Lemma 1.6)} \\
&= \frac{1}{2} \int_{\partial M_k} (-k)^{-2} \|\Re\phi_k\|_{I_k}^2 (-k) \, \mathrm{d}a_{I_k} \\
&= -\frac{1}{2k} \int_{\partial M_k} (k u_k H_k - 2(k+1)) \, \mathrm{d}a_{I_k} \qquad \text{(relation (6))},
\end{aligned}$$

where we used that $h_k = (-k)I_k$, relations (2) and (3), and that the integral of the term $\mathrm{div}_{I_k} W$ vanishes by the divergence theorem. By Lemma 2.3 $u_k \leq \sqrt{k+1}/k$, so

$$\|\mathrm{d}V_k^*\|_{\mathrm{WP}}^2 \geq -\frac{\sqrt{k+1}}{2k} \int_{\partial M_k} H_k \, \mathrm{d}a_{I_k} - \frac{2\pi(k+1)}{k^2} |\chi(\partial M)|,$$

where we applied the Gauss–Bonnet theorem to say that the area of ∂M_k with respect to I_k is equal to $-2\pi k^{-1} |\chi(\partial M)|$.

The only ingredient left to prove is relation (6). For this computation, we will use *Bochner's formula* (see eg [25, page 223]),

$$(7) \quad \tfrac{1}{2}\Delta_g \|\mathrm{d}f\|_g^2 = \|{}^g\nabla^2 f\|_g^2 + g(\mathrm{grad}_g \, f, \mathrm{grad}_g \, \Delta_g f) + \mathrm{Ric}_g(\mathrm{grad}_g \, f, \mathrm{grad}_g \, f),$$

and the expressions

$$(8) \qquad\qquad \mathrm{div}_g(fX) = g(\mathrm{grad}_g \, f, X) + f \, \mathrm{div}_g X,$$

$$(9) \qquad\qquad \tfrac{1}{2}(\mathcal{L}_X g, T)_g = -(\mathrm{div}_g T)(X) + \mathrm{div}_g Y,$$

where X is a tangent vector field, f is a smooth function, T is a symmetric 2–tensor, and $Y = T(X, \cdot)^{\sharp}$ is the vector field defined by requiring that $g(Y, Z) = T(X, Z)$ for all vector fields Z. From now on, we will omit everywhere the dependence of the connections, norms, gradients, and the Laplace–Beltrami operator on the Riemannian

metric g, and everything has to be interpreted as associated to $g = I_k$. Observe also that the Levi-Civita connections of I_k and h_k are equal, since these metrics differ by multiplication by a constant and, in particular, the h_k– and I_k–Hessians coincide. Then

$$(10) \quad \|I\!I_k - \nabla^2 u_k + u_k h_k\|^2 = \|I\!I_k - \nabla^2 u_k - k u_k I_k\|^2$$
$$= \|I\!I_k\|^2 + \|\nabla^2 u_k\|^2 + k^2 u_k^2 \|I_k\|^2 - 2(I\!I_k, \nabla^2 u_k)$$
$$- 2k u_k(I\!I_k, I_k) + 2k u_k(\nabla^2 u_k, I_k).$$

First, we focus our attention on the terms $\|\nabla^2 u_k\|^2$ and $(I\!I_k, \nabla^2 u_k)$. In order to simplify the notation, we say that two functions a and b on ∂M_k are equal "modulo divergence", and we write $a \equiv_{\mathrm{div}} b$, if their difference coincides with the divergence of some smooth vector field. Then

$$\|\nabla^2 u_k\|^2 = \tfrac{1}{2}\Delta\|du_k\|^2 - \langle \mathrm{grad}\, u_k, \mathrm{grad}\, \Delta u_k\rangle - k\|du_k\|^2 \quad \text{(relation (7))}$$
$$\equiv_{\mathrm{div}} -\langle \mathrm{grad}\, u_k, \mathrm{grad}\, \Delta u_k\rangle - k\|du_k\|^2 \qquad (\Delta_g f = \mathrm{div}_g\, \mathrm{grad}_g\, f)$$
$$= -\mathrm{div}(\Delta u_k\, \mathrm{grad}\, u_k) + (\Delta u_k)^2 - k\|du_k\|^2 \qquad \text{(relation (8))}$$
$$\equiv_{\mathrm{div}} (\Delta u_k)^2 - k\, \mathrm{div}(u_k\, \mathrm{grad}\, u_k) + k u_k \Delta u_k \qquad \text{(relation (8))}$$
$$\equiv_{\mathrm{div}} \Delta u_k(\Delta u_k + k u_k),$$

and

$$(I\!I_k, \nabla^2 u_k) = \tfrac{1}{2}(I\!I_k, \mathcal{L}_{\mathrm{grad}\, u_k} I_k) \qquad\qquad (\mathcal{L}_{\mathrm{grad}_g\, f}\, g = 2^g \nabla^2 f)$$
$$\equiv_{\mathrm{div}} -(\mathrm{div}\, I\!I_k)(\mathrm{grad}\, u_k) \qquad\qquad \text{(relation (9))}$$
$$= -\langle \mathrm{grad}\, H_k, \mathrm{grad}\, u_k\rangle \qquad\qquad (\mathrm{div}\, I\!I_k = \mathrm{d}H_k)$$
$$= -\mathrm{div}(H_k\, \mathrm{grad}\, u_k) + H_k \Delta u_k \quad \text{(relation (8))}$$
$$\equiv_{\mathrm{div}} H_k \Delta u_k.$$

The other terms in (10) are simpler to handle. In particular,

$$\|I\!I_k\|^2 = H_k^2 - 2(k+1), \qquad \|I_k\|^2 = 2,$$
$$(I\!I_k, I_k) = H_k, \qquad\qquad (\nabla^2 u_k, I_k) = \Delta u_k.$$

Replacing all the relations we found in (10), we obtain

$$\|I\!I_k - \nabla^2 u_k + u_k h_k\|^2$$
$$\equiv_{\mathrm{div}} H_k^2 - 2(k+1) + \Delta u_k(\Delta u_k + k u_k) + 2k^2 u_k^2$$
$$- 2H_k \Delta u_k - 2k u_k H_k + 2k u_k \Delta u_k$$
$$= H_k^2 - 2(k+1) + 2k^2 u_k^2 - 2k u_k H_k + \Delta u_k(\Delta u_k + 3k u_k - 2H_k).$$

Finally, by replacing $\Delta u_k = \Delta_{I_k} u_k$ using (4) in the equality above, we find that

$$\| I\!I_k - \nabla^2 u_k + u_k h_k \|^2 \equiv_{\mathrm{div}} k u_k H_k - 2(k+1),$$

which is equivalent to relation (6). □

Since the Weil–Petersson metric of the Teichmüller space is noncomplete, a control from above of the quantity $\| dV_k^* \|_{\mathrm{WP}}$ would not suffice to guarantee the existence of the flow for every time. For this purpose, we rather study the L^∞–norm of the Beltrami differentials equivalent to $\mathrm{grad}_{\mathrm{WP}} V_k^*$, which gives a control with respect to the Teichmüller metric (that is complete). At this point, the estimates determined in Lemma 2.3 will play an essential role.

Proposition 3.5 *There exists a constant $D_k > 0$ depending only on the intrinsic curvature $k \in (-1, 0)$ such that*

$$\| \mathrm{grad}_{\mathrm{WP}} V_k^* \|_{\mathcal{T}} \le D_k,$$

where $\| \cdot \|_{\mathcal{T}}$ denotes the Teichmüller norm on $T\mathcal{T}(\partial M_k)$.

Proof Let m_k be a point of the Teichmüller space $\mathcal{T}(\partial M_k)$ and let h_k be a hyperbolic metric in the isotopy class m_k. In Proposition 3.2, we showed that the vector field $\mathrm{grad}_{\mathrm{WP}} V_k^*$ at a point $m_k \in \mathcal{T}(\partial M_k)$ is represented by the transverse traceless tensor $2\Re \phi_k \in S_2^{\mathrm{tt}}(\partial M_k, h_k)$. Therefore by Lemma 1.6,

$$\| \mathrm{grad}_{\mathrm{WP}} V_k^* \|_{\mathcal{T}} \le \frac{1}{\sqrt{2}} \sup_{\partial M_k} \| \Re \phi_k \|_{h_k}.$$

So it is enough to show that the norm $\| I\!I_k - {}^k\nabla^2 u_k + u_k h_k \|_{h_k}$ is uniformly bounded by a constant depending only on k. The norm of $I\!I_k$ is equal to $-k^{-1}\sqrt{H_k^2 - 2(k+1)}$, and $\| u_k h_k \|_{h_k} = \sqrt{2} |u_k|$. Therefore

$$\| I\!I_k - {}^k\nabla^2 u_k + u_k h_k \|_{h_k} \le -k^{-1}\sqrt{\| H_k \|_{\mathscr{C}^0}^2 - 2(k+1)} + \| {}^k\nabla^2 u_k \|_{h_k} + \sqrt{2}\| u_k \|_{\mathscr{C}^0}.$$

Our assertion is now an immediate consequence of Proposition 2.1 and Lemma 2.3. □

Corollary 3.6 *The flow Θ_t of the vector field $-\mathrm{grad}_{\mathrm{WP}} V_k^*$ over $\mathcal{T}(\partial M_k)$ is defined for all times $t \in \mathbb{R}$.*

Proof The assertion follows from the fact that the Teichmüller distance is complete, and the bound shown in Proposition 3.5. □

The last ingredient that we will need for the proof of Theorem A is the existence of some lower bound for the dual volume function V_k^*. To find one, we will make use of the properties of the dual volume proved in [29], and of an upper bound for the length of the bending measure of the boundary of the convex core of a convex cocompact manifold with incompressible boundary, whose existence was first proved by Bridgeman [8] and improved in later works; see [11]. We will make use of the best result currently known in this direction for convex cocompact manifolds with incompressible boundary, which is due to Bridgeman, Brock and Bromberg [9].

Lemma 3.7 *For every $k \in (-1, 0)$ and for every convex cocompact hyperbolic 3– manifold M with incompressible boundary,*

$$V_k^*(M) \geq F(k, \chi(\partial M)),$$

where F is an explicit function of the curvature $k \in (-1, 0)$ and the Euler characteristic of ∂M.

Proof Since the k–surfaces foliate the complement of the convex core CM, a simple application of the geometric maximum principle (see for instance [24, lemme 2.5.1]) shows that the k–surface ∂M_k is contained in $N_{\varepsilon_k} CM$, the ε_k–neighborhood of the convex core CM, for $\varepsilon_k = \operatorname{arctanh} \sqrt{k+1}$. The dual volume of a convex set is a decreasing function with respect to inclusion (see [29, Proposition 2.6] for a proof of this assertion), therefore the quantity $V_k^*(M)$ is bounded from below by the dual volume of the ε_k–neighborhood of the convex core. It is not difficult to show that, for every $\varepsilon > 0$,

$$V^*(N_\varepsilon CM) = V(CM) - \tfrac{1}{4}\ell_m(\mu)(\cosh 2\varepsilon + 1) - \tfrac{1}{2}\pi |\chi(\partial CM)|(\sinh 2\varepsilon - 2\varepsilon),$$

where $\ell_m(\mu)$ denotes the length of the bending measured lamination on the boundary of the convex core of M; see eg [29, Proposition 2.4]. By [9, Theorem 2.16], the term $\ell_m(\mu)$ is less than or equal to $6\pi |\chi(\partial M)|$. Combining these observations,

$$\begin{aligned}
V_k^*(M) &\geq V^*(N_{\varepsilon_k} CM) \\
&= V(CM) - \tfrac{1}{4}\ell_m(\mu)(\cosh 2\varepsilon_k + 1) - \tfrac{1}{2}\pi |\chi(\partial CM)|(\sinh 2\varepsilon_k - 2\varepsilon_k) \\
&\geq -\tfrac{1}{4}\ell_m(\mu)(\cosh 2\varepsilon_k + 1) - \tfrac{1}{2}\pi |\chi(\partial CM)|(\sinh 2\varepsilon_k - 2\varepsilon_k) \\
&\geq -\tfrac{1}{2}\pi |\chi(\partial M)|(3\cosh \varepsilon_k + 3 + \sinh 2\varepsilon_k - 2\varepsilon_k),
\end{aligned}$$

which proves the desired inequality. □

4 The proof of Theorem A

This section is dedicated to the proof of our main theorem, and to the proof of the optimality of the multiplicative constant appearing in (1).

Proof of Theorem A   Let M be a convex cocompact hyperbolic 3–manifold with incompressible boundary. We denote by $M_t := \Theta_t(M)$ the hyperbolic 3–manifold obtained by following the flow of the vector field $-\operatorname{grad}_{\mathrm{WP}} V_k^*$, which is defined for every $t \in \mathbb{R}$ in light of Corollary 3.6. In order to simplify the notation, we will continue to denote by V_k^* the k–dual volume as a function over the space of quasi-isometric deformations of M. This abuse is justified by the fact that, for every $k \in (-1,0)$, a convex cocompact manifold is uniquely determined by the hyperbolic structures on its k–surfaces (see Theorem 1.5). We have

$$V_k^*(M) - V_k^*(M_t) = \int_0^t \|dV_k^*\|_{M_s}^2 \, ds.$$

By Lemma 3.7, the left side of the relation is bounded from above with respect to t. In particular, the integral on the right side has to converge as t goes to $+\infty$. Therefore we can find an unbounded increasing sequence $(t_n)_n$ for which the Weil–Petersson norm $\|dV_k^*\|^2$ evaluated at M_{t_n} goes to 0 as n goes to ∞. Then, by Lemma 3.4,

$$\limsup_{n\to\infty} \int_{\partial M_{t_n,k}} H_k \, da_{I_k} \le -4\pi k^{-1} \sqrt{k+1} |\chi(\partial M)|,$$

where $M_{t_n,k}$ stands for $(M_{t_n})_k$, the region of M_{t_n} enclosed by its k–surfaces. Therefore

$$
\begin{aligned}
V_k^*(M) &\ge \lim_{n\to\infty} V_k^*(M_{t_n}) = \lim_{n\to\infty} \left(V_k(M_{t_n}) - \frac{1}{2} \int_{\partial M_{t_n,k}} H_k \, da_{I_k} \right) \\
&\ge \inf_{M' \in \mathcal{QD}(M)} V_k(M') - \frac{1}{2} \limsup_{n\to\infty} \int_{\partial M_{t_n,k}} H_k \, da_{I_k} \\
&\ge \inf_{M' \in \mathcal{QD}(M)} V_k(M') + 2\pi k^{-1} \sqrt{k+1} |\chi(\partial M)|,
\end{aligned}
$$

where $V_k(M')$ denotes the Riemannian volume of the region M_k' of M' enclosed by its k–surface. Observe that the term $2\pi k^{-1} \sqrt{k+1} |\chi(\partial M)|$ is equal to $-\frac{1}{2} \int_{\partial M_k'} H_k \, da_{I_k}$ when the boundary of the convex core of M' is totally geodesic.

Finally, by taking the limit as k goes to $(-1)^+$, we obtain that $V_C^*(M) \ge \inf_{M'} V_C(M')$ for every convex cocompact structure M. This proves that

$$\inf_{\mathcal{QD}(M)} V_C^* \ge \inf_{\mathcal{QD}(M)} V_C.$$

On the other hand, the dual volume $V_C^*(M) := V_C(M) - \frac{1}{2}\ell_m(\mu)$ is always smaller than or equal to $V_C(M)$, so the other inequality between the infima is clearly satisfied.

If $V_C^*(M) = V_C(M)$, then the length of the bending measured lamination μ of the convex core of M has to vanish. Therefore $\mu = 0$ or, in other words, ∂CM is totally geodesic. $\square$

Corollary 4.1 *For every quasi-Fuchsian manifold M, $V_C(M) \geq \frac{1}{2}\ell_m(\mu)$, where $m = m(M)$ and $\mu = \mu(M)$ denote the hyperbolic metric and the bending measure of the boundary of the convex core of M, respectively. Moreover, for every positive ε and for every neighborhood U of a Fuchsian manifold M_0 inside $\mathcal{QD}(M_0) = \mathcal{QD}(M)$, there exists a quasi-Fuchsian manifold M_ε in U that satisfies $V_C(M_\varepsilon) < \left(\frac{1}{2} + \varepsilon\right)\ell_{m_\varepsilon}(\mu_\varepsilon)$, where $m_\varepsilon = m(M_\varepsilon)$ and $\mu_\varepsilon = \mu(M_\varepsilon)$.*

Proof If M is quasi-Fuchsian, the infimum of the volume of the convex core over the space of quasi-isometric deformations $\mathcal{QD}(M)$ is equal to 0, and it is realized on the Fuchsian locus.

For the second part of the statement, consider M_0 a Fuchsian manifold whose convex core is a totally geodesic surface homeomorphic to Σ with hyperbolic metric m_0. Let $\alpha \colon [0,1] \to \mathcal{QD}(M)$ be a path starting at $\alpha(0) = M_0$ and for which the right derivative of the bending measure $\dot{\mu}_0^+$ exists and it is equal to a nonzero measured lamination on $\Sigma \sqcup \Sigma$. A fairly explicit way to produce such a path is to choose a measured lamination $\lambda \in \mathcal{ML}(\Sigma)$ and consider the deformation of M_0 given by the holonomies of pleated surfaces with bending Hölder cocycle equal to $t\lambda$ and hyperbolic metric m_0, as t varies in $[0,1]$; compare with [3]. Then, for every $\varepsilon > 0$, we define

$$f_\varepsilon(t) := V_C(\alpha(t)) - \left(\tfrac{1}{2} + \varepsilon\right)\ell_{m_t}(\mu_t) = V_C^*(\alpha(t)) - \varepsilon\ell_{m_t}(\mu_t) \quad \text{for } t \in [0,1],$$

where $m_t = m(M_t)$ and $\mu_t = \mu(M_t)$ denote the hyperbolic metric and the bending measure of the boundary of the convex core of $M_t = \alpha(t)$. As shown in [20, (4)],

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_{m_t}(\mu_t)\Big|_{t=0^+} = \mathrm{d}(L_{\mu_0})(\dot{m}_0) + \ell_{m_0}(\dot{\mu}_0^+) = \ell_{m_0}(\dot{\mu}_0^+),$$

where we use that $\mu_0 = 0$ (here $L_{\mu_0} \colon \mathcal{T}(\partial CM) \to \mathbb{R}$ is the function that associates with every hyperbolic structure m the length of the m–geodesic realization of μ_0). Then

$$\begin{aligned}
f_\varepsilon(t) &= f_\varepsilon(0) + f_\varepsilon'(0)t + o(t; \varepsilon) \\
&= 0 + (\mathrm{d}(V_C{}^*)_{M_0}(v) - \varepsilon\ell_{m_0}(\dot{\mu}_0^+))t + o(t; \varepsilon) \\
&= -\varepsilon\ell_{m_0}(\dot{\mu}_0^+)t + o(t; \varepsilon) \qquad\qquad (V_C^* \in \mathscr{C}^1 \text{ and } M_0 \text{ minimum}).
\end{aligned}$$

This proves that $f_\varepsilon(t) < 0$ for t sufficiently small (depending on ε), and therefore the existence of a quasi-Fuchsian manifold M_ε satisfying the desired properties. □

References

[1] **L Bers**, *Spaces of Kleinian groups*, from "Several complex variables, I" (J Horváth, editor), Springer (1970) 9–34 MR Zbl

[2] **F Bonahon**, *The geometry of Teichmüller space via geodesic currents*, Invent. Math. 92 (1988) 139–162 MR Zbl

[3] **F Bonahon**, *Shearing hyperbolic surfaces, bending pleated surfaces and Thurston's symplectic form*, Ann. Fac. Sci. Toulouse Math. 5 (1996) 233–297 MR Zbl

[4] **F Bonahon**, *A Schläfli-type formula for convex cores of hyperbolic* 3*–manifolds*, J. Differential Geom. 50 (1998) 25–58 MR Zbl

[5] **F Bonahon**, *Variations of the boundary geometry of* 3*–dimensional hyperbolic convex cores*, J. Differential Geom. 50 (1998) 1–24 MR Zbl

[6] **F Bonsante**, **J Danciger**, **S Maloni**, **J-M Schlenker**, *The induced metric on the boundary of the convex hull of a quasicircle in hyperbolic and anti–de Sitter geometry*, Geom. Topol. 25 (2021) 2827–2911 MR Zbl

[7] **H Brezis**, *Functional analysis, Sobolev spaces and partial differential equations*, Springer (2011) MR Zbl

[8] **M Bridgeman**, *Average bending of convex pleated planes in hyperbolic three-space*, Invent. Math. 132 (1998) 381–391 MR Zbl

[9] **M Bridgeman**, **J Brock**, **K Bromberg**, *Schwarzian derivatives, projective structures, and the Weil–Petersson gradient flow for renormalized volume*, Duke Math. J. 168 (2019) 867–896 MR Zbl

[10] **M Bridgeman**, **J Brock**, **K Bromberg**, *The Weil–Petersson gradient flow of renormalized volume and* 3*–dimensional convex cores*, preprint (2021) arXiv 2003.00337 To appear in Geom. Topol.

[11] **M Bridgeman**, **R D Canary**, *Bounding the bending of a hyperbolic* 3*–manifold*, Pacific J. Math. 218 (2005) 299–314 MR Zbl

[12] **J F Brock**, *The Weil–Petersson metric and volumes of* 3*–dimensional hyperbolic convex cores*, J. Amer. Math. Soc. 16 (2003) 495–535 MR Zbl

[13] **R D Canary**, **D Epstein**, **A Marden** (editors), *Fundamentals of hyperbolic geometry: selected expositions*, Lond. Math. Soc. Lect. Note Ser. 328, Cambridge Univ. Press (2006) MR Zbl

[14] **L C Evans**, *Partial differential equations*, Graduate Studies in Math. 19, Amer. Math. Soc., Providence, RI (1998) MR Zbl

[15] **A E Fischer**, **J E Marsden**, *Deformations of the scalar curvature*, Duke Math. J. 42 (1975) 519–547 MR Zbl

[16] **F P Gardiner**, **N Lakic**, *Quasiconformal Teichmüller theory*, Math. Surv. Monogr. 76, Amer. Math. Soc., Providence, RI (2000) MR Zbl

[17] **M Kapovich**, *Hyperbolic manifolds and discrete groups*, Progr. Math. 183, Birkhäuser, Boston, MA (2001) MR Zbl

[18] **I Kra**, *On spaces of Kleinian groups*, Comment. Math. Helv. 47 (1972) 53–69 MR Zbl

[19] **K Krasnov**, **J-M Schlenker**, *On the renormalized volume of hyperbolic* 3–*manifolds*, Comm. Math. Phys. 279 (2008) 637–668 MR Zbl

[20] **K Krasnov**, **J-M Schlenker**, *A symplectic map between hyperbolic and complex Teichmüller theory*, Duke Math. J. 150 (2009) 331–356 MR Zbl

[21] **F Labourie**, *Problème de Minkowski et surfaces à courbure constante dans les variétés hyperboliques*, Bull. Soc. Math. France 119 (1991) 307–325 MR Zbl

[22] **F Labourie**, *Métriques prescrites sur le bord des variétés hyperboliques de dimension* 3, J. Differential Geom. 35 (1992) 609–626 MR Zbl

[23] **F Labourie**, *Surfaces convexes dans l'espace hyperbolique et* $\mathbb{C}P^1$–*structures*, J. London Math. Soc. 45 (1992) 549–565 MR Zbl

[24] **F Labourie**, *Un lemme de Morse pour les surfaces convexes*, Invent. Math. 141 (2000) 239–297 MR Zbl

[25] **J M Lee**, *Introduction to Riemannian manifolds*, Graduate Texts in Math. 176, Springer (2018) MR Zbl

[26] **A Marden**, *Outer circles: an introduction to hyperbolic* 3–*manifolds*, Cambridge Univ. Press (2007) MR Zbl

[27] **B Maskit**, *Self-maps on Kleinian groups*, Amer. J. Math. 93 (1971) 840–856 MR Zbl

[28] **F Mazzoli**, *Constant curvature surfaces and volumes of convex co-compact hyperbolic manifolds*, PhD thesis, Université du Luxembourg (2020) Available at http://hdl.handle.net/10993/43901

[29] **F Mazzoli**, *The dual Bonahon–Schläfli formula*, Algebr. Geom. Topol. 21 (2021) 279–315 MR Zbl

[30] **F Mazzoli**, *The dual volume of quasi-Fuchsian manifolds and the Weil–Petersson distance*, Trans. Amer. Math. Soc. 375 (2022) 695–723 MR Zbl

[31] **F V Pallete**, *Continuity of the renormalized volume under geometric limits*, preprint (2016) arXiv 1605.07986

[32] **F V Pallete**, *Upper bounds on renormalized volume for Schottky groups*, preprint (2019) arXiv 1905.03303

[33] **J G Ratcliffe**, *Foundations of hyperbolic manifolds*, 2nd edition, Graduate Texts in Math. 149, Springer (2006) MR Zbl

[34] **I Rivin**, **J-M Schlenker**, *On the Schlafli differential formula*, preprint (2000) arXiv math/0001176

[35] **H Rosenberg**, **J Spruck**, *On the existence of convex hypersurfaces of constant Gauss curvature in hyperbolic space*, J. Differential Geom. 40 (1994) 379–409 MR Zbl

[36] **J-M Schlenker**, *Hypersurfaces in H^n and the space of its horospheres*, Geom. Funct. Anal. 12 (2002) 395–435 MR Zbl

[37] **J-M Schlenker**, *Hyperbolic manifolds with convex boundary*, Invent. Math. 163 (2006) 109–169 MR Zbl

[38] **J-M Schlenker**, *The renormalized volume and the volume of the convex core of quasifuchsian manifolds*, Math. Res. Lett. 20 (2013) 773–786 MR Zbl

[39] **J-M Schlenker**, *Notes on the Schwarzian tensor and measured foliations at infinity of quasifuchsian manifolds*, preprint (2017) arXiv 1708.01852

[40] **P A Storm**, *Minimal volume Alexandrov spaces*, J. Differential Geom. 61 (2002) 195–225 MR Zbl

[41] **P A Storm**, *Hyperbolic convex cores and simplicial volume*, Duke Math. J. 140 (2007) 281–319 MR Zbl

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

[43] **A J Tromba**, *Teichmüller theory in Riemannian geometry*, Birkhäuser, Basel (1992) MR Zbl

Department of Mathematics, University of Virginia
Charlotteville, VA, United States

filippomazzoli@me.com

Discrete subgroups of small critical exponent

Beibei Liu

Shi Wang

We prove that finitely generated Kleinian groups $\Gamma < \mathrm{Isom}(\mathbb{H}^n)$ with small critical exponent are always convex cocompact. We also prove some geometric properties for any complete pinched negatively curved manifold with critical exponent less than 1.

1 Introduction

A *Kleinian* group is a discrete isometry subgroup of $\mathrm{Isom}(\mathbb{H}^n)$. The study of 3–dimensional finitely generated Kleinian groups dates back to Schottky, Poincaré and Klein. It is only recently that the geometric picture of the associated hyperbolic manifold has been much better understood, after the celebrated work of Ahlfors' finiteness theorem [2], the proof of the tameness conjecture (see Agol [1], Bonahon [10] and Calegari and Gabai [18]), and the unraveling of the ending lamination conjecture; see Bowditch [13], Brock, Canary and Minsky [14], Minsky [36] and Soma [42]. However, such geometric descriptions fail in higher dimensions; see Kapovich [29; 30], Kapovich and Potyagailo [33; 34] and Potyagailo [41; 40].

One way to study higher-dimensional Kleinian groups is to consider the interplay between the group-theoretic properties, the geometry of the quotient manifolds, and the *measure-theoretic size* of the limit set. It was shown by Gusevskii [23] that if the Hausdorff dimension of the entire limit set $\dim_{\mathcal{H}}(\Lambda(\Gamma))$ is less than 1, then Γ is geometrically finite. In this case, the Hausdorff dimension of the entire limit set equals the Hausdorff dimension of the conical limit set (see Bowditch [12]), which is smaller than 1. However, when Γ is geometrically infinite, the size of the entire limit set could a priori be much larger, so $\dim_{\mathcal{H}} \Lambda(\Gamma) > \dim_{\mathcal{H}} \Lambda_c(\Gamma)$. Thus, it is interesting to ask what the relative size of $\Lambda_c(\Gamma)$ is compared to the entire $\Lambda(\Gamma)$, or rather, to what extent is the size of $\Lambda_c(\Gamma)$ able to determine the geometric finiteness of the group. By the

work of Bishop and Jones [9], the Hausdorff dimension of the conical limit set $\Lambda_c(\Gamma)$ equals the critical exponent $\delta(\Gamma)$. Hence, Kapovich [31, Problem 1.6] asked:

Question 1.1 *Is every finitely generated Kleinian group $\Gamma < \mathrm{Isom}(\mathbb{H}^n)$ with $\delta(\Gamma) < 1$ geometrically finite?*

We partly answer this in the affirmative in a slightly more general context.

Theorem 1.2 *For each n and κ there exists a positive constant $D(n,\kappa) < \frac{1}{2}$ with the property that, for every n–dimensional Hadamard manifold with pinched sectional curvature $-\kappa^2 \le K \le -1$ and any finitely generated torsion-free discrete isometry subgroup $\Gamma < \mathrm{Isom}(X)$, Γ is convex cocompact if $\delta(\Gamma) < D(n,\kappa)$.*

Remark 1.3 The constant $D(n,\kappa)$ can be obtained from the quantitative version of the Tits alternative for pinched negatively curved manifolds; see Dey, Kapovich and Liu [20].

Remark 1.4 For 3–dimensional finitely generated Kleinian groups Γ of *second kind*, ie $\Lambda(\Gamma) \ne S^2$, Bishop and Jones [9] showed that Γ is geometrically finite if $\delta(\Gamma) < 2$. Hou [25; 26; 27] proved that a 3–dimensional Kleinian group Γ is a classical Schottky group if $\dim_{\mathcal{H}}(\Lambda(\Gamma)) < 1$.

In [31], Kapovich established a relation between the homological dimension and the critical exponent of a Kleinian group. A similar homological vanishing feature has been extended to other rank-one symmetric spaces by Connell, Farb and McReynolds [19]. It is conjectured [31, Conjecture 1.4] that the virtual cohomological dimension $\mathrm{vcd}(\Gamma)$ is bounded above by $\delta(\Gamma) + 1$ (assuming Γ has no higher-rank cusps). Under the condition $\delta(\Gamma) < 1$, it is equivalent to ask (see Stallings [43] and also a weaker form by Bestvina [8, Question 5.6]):

Question 1.5 *Is every finitely generated Kleinian group $\Gamma < \mathrm{Isom}(\mathbb{H}^n)$ with $\delta(\Gamma) < 1$ virtually free?*

In the same paper, Kapovich gave a positive answer to this question under the stronger assumption that Γ is finitely presented. On the other hand, when $\delta(\Gamma)$ is sufficiently small, our Theorem 1.2 automatically implies $\dim_{\mathcal{H}}(\Lambda(\Gamma)) = \delta(\Gamma) < D(n,\kappa) < 1$. This implies that the limit set $\Lambda(\Gamma)$ is a Cantor set since it is perfect. Following the classical result of Kulkarni [35, Theorem 6.11]:

Corollary 1.6 *For each n there is a positive constant $D(n) < \frac{1}{2}$ such that any finitely generated discrete isometry subgroup $\Gamma < \mathrm{Isom}(\mathbb{H}^n)$ is virtually free if $\delta(\Gamma) < D(n)$.*

Remark 1.7 Under the assumption that $\dim_{\mathcal{H}}(\Lambda(\Gamma)) < 1$, Pankka and Souto [39] proved that any torsion-free Kleinian group (not necessarily finitely generated) is free.

The method in [31] also works for discrete isometry subgroups of Hadamard manifolds with negatively pinched sectional curvature $-\kappa^2 \leq K \leq -1$, and Question 1.5 can be asked for this family of groups. If in addition we know Γ is free in Theorem 1.2, then the constant $D(n, \kappa)$ can actually be made effective, and independent of n and κ.

Theorem 1.8 *Let* $\Gamma < \mathrm{Isom}(X)$ *be a finitely generated virtually free discrete isometry subgroup of an* n*–dimensional Hadamard manifold with pinched negative curvature* $-\kappa^2 \leq K \leq -1$. *If* $\delta(\Gamma) < \frac{1}{16}$, *then* Γ *is convex cocompact.*

Thus, in view of Kapovich's result [31, Corollary 1.5], we obtain:

Corollary 1.9 *A finitely presented Kleinian group with* $\delta(\Gamma) < \frac{1}{16}$ *is convex cocompact.*

One of the main efforts in our proofs is investigating the geometric properties of the quotient manifold $M = X/\Gamma$ under the condition that δ is small. While these results are only restricted to $\delta < 1$, we still find that they might be of independent interest and worth highlighting. The following theorem is closely related to the classical Plateau's problem, where we obtain a certain type of linear isoperimetric inequality for the quotient manifold $M = X/\Gamma$.

Theorem 1.10 *Suppose that* $\mathcal{C}$ *is a union of smooth loops in* $M = X/\Gamma$ *which represents a trivial homology class in* $H_1(M, \mathbb{Z})$. *If* $\delta(\Gamma) = \delta < 1$, *then* $\mathcal{C}$ *bounds a smooth surface* $i : \Sigma \to M$ *(see Definition 2.6) whose area satisfies*

$$A(i) \leq \frac{4}{1-\delta} \ell(\mathcal{C}),$$

where $\ell(\mathcal{C})$ *denotes the total length of the smooth loops in* $\mathcal{C}$.

Finitely generated Kleinian groups in dimension 3 have only finitely many cusps (see Sullivan [44]), but the same result does not hold in higher dimensions; see Kapovich [29]. As an application of Theorem 1.10, we show that, under the assumption $\delta < 1$, the ϵ–thin part of M has only finitely many connected components when ϵ is small enough. In particular, M has only finitely many cusps.

Theorem 1.11 *Let* $\Gamma < \mathrm{Isom}(X)$ *be a finitely generated torsion-free discrete isometry subgroup of an* n*–dimensional Hadamard manifold with pinched negative curvature* $-\kappa^2 \leq K \leq -1$. *Suppose that* $\delta(\Gamma) < 1$. *Then:*

(1) *The number of cusps in $M = X/\Gamma$ is at most the first Betti number of M.*

(2) *M has bounded geometry. That is, the noncuspidal part of M has a uniform lower bound on its injectivity radius.*

(3) *Γ is convex cocompact if and only if the injectivity radius function* $\mathrm{inj}: M \to \mathbb{R}$ *is proper.*

Remark 1.12 Without the assumption on the critical exponent, Benoist and Hulin [5, Proposition 2.6] showed that Γ is convex cocompact if and only if M is Gromov hyperbolic and the injectivity radius function is proper.

Outline of the proof of Theorem 1.2

We first observe that whenever $\delta < 1$ there is an area-decreasing self-map (the Besson–Courtois–Gallot map) on M. This allows us to prove the linear isoperimetric type inequality as in Theorem 1.10, from which we deduce further that closed geodesics on M asymptotically have uniformly bounded normal injectivity radii. This means that if there is an escaping sequence of closed geodesics on M, then there exists a subsequence on which the normal injectivity radii are uniformly bounded. Next we observe that, given a long closed geodesic with small normal injectivity radius, one can always separate along the normal direction to replace it by a shorter closed geodesic nearby. Then, we use the result by Kapovich and Liu [32] which states that Γ is geometrically infinite if and only if there exists an escaping sequence of closed geodesics. The assumption that $D(n,\kappa)$ is smaller than $\frac{1}{2}$ excludes parabolic elements, so assume for the sake of contradiction that there is one such escaping sequence. Using the idea of infinite descent we can reduce the length of the closed geodesics and find another escaping sequence whose lengths and normal injectivity radii are both uniformly bounded, from which we can find two loxodromic isometries that move a common point within a uniformly bounded distance. This means the nonelementary subgroup generated by the two isometries will have large critical exponent, thus leading to a contradiction if we assume δ is small enough.

Organization of the paper

In Section 2 we review some elementary results of negatively pinched Hadamard manifolds and the Besson–Courtois–Gallot map. In Section 3 we give the proofs of Theorems 1.10 and 1.11. In Section 4 we prove Theorem 4.1, which together with Theorem 1.11 implies Theorems 1.2 and 1.8.

Acknowledgments

2 Preliminaries

2.1 Discrete isometry groups

Let X be a complete simply connected n–dimensional Riemannian manifold of pinched negative curvature $-\kappa^2 \le K \le -1$ where $\kappa \ge 1$. The Riemannian metric on X induces the distance function d_X, and (X, d_X) is a uniquely geodesic space. With the curvature assumption, the metric space (X, d_X) is Gromov hyperbolic, where the hyperbolicity constant δ_0 can be chosen as $\cosh^{-1}(\sqrt{2})$, ie every geodesic triangle in X is δ_0–slim.

By the Cartan–Hadamard theorem, X is diffeomorphic to the Euclidean space $\mathbb{R}^n$ via the exponential map at any point in X. We can naturally compactify X by adding the ideal boundary $\partial_\infty X$, thus the compactified space $\overline{X} = X \cup \partial_\infty X$ is homeomorphic to the unit n–ball B^n.

Every isometry $\gamma \in \mathrm{Isom}(X)$ extends the action to the ideal boundary, so it induces a diffeomorphism on $\overline{X}$. Based on its fixed-point set $\mathrm{Fix}(\gamma)$, the isometry γ on X can be classified:

(1) γ is *parabolic* if $\mathrm{Fix}(\gamma)$ is a singleton $\{p\} \subset \partial_\infty X$.

(2) γ is *elliptic* if it has a fixed point in X. In this case, the fixed-point set $\mathrm{Fix}(\gamma)$ is a totally geodesic subspace of X invariant under γ. In particular, the identity map is elliptic.

(3) γ is *loxodromic* if $\mathrm{Fix}(\gamma)$ consists of two distinct points $p, q \in \partial_\infty X$. In this case, γ stabilizes and translates along the geodesic pq, and we call the geodesic pq the *axis* of γ.

One can also use the translation length to classify the isometries on X. For each isometry $\gamma \in \mathrm{Isom}(X)$, we define its *translation length* $\tau(\gamma)$ as

$$\tau(\gamma) := \inf_{x \in X} d_X(x, \gamma(x)).$$

The isometry γ is loxodromic if and only if $\tau(\gamma) > 0$. In this case, the infimum is attained exactly when the points are on the axis of γ. The isometry γ is parabolic if and only if $\tau(\gamma) = 0$ and the infimum is not attained. The isometry γ is elliptic if and only if $\tau(\gamma) = 0$ and the infimum is attained.

Let $\Gamma < \mathrm{Isom}(X)$ be a discrete subgroup which acts on X properly discontinuously. If Γ is torsion-free, then any nontrivial element in Γ is either loxodromic or parabolic. We denote the quotient manifold X/Γ by M, and let $\pi: X \to M$ denote the canonical projection. The geodesic loops $c:[a,b] \to M$ at $p = c(a) = c(b) \in M$ are in one-to-one correspondence with geodesic segments from x to $\gamma(x)$, where $x \in X$ with $\pi(x) = p$ and $\gamma \in \Gamma$. Recall that the injectivity radius at a point $p \in M$ is the largest radius for which the exponential map at p is a diffeomorphism. The injectivity radius at a point $p \in M$ is half the length of shortest geodesic loop at p since there are no conjugate points in M. We use $\mathrm{inj}(p)$ to denote the injectivity radius at p and define

$$d_\Gamma(x) := \min_{\gamma \in \Gamma \setminus \{id\}} d_X(x, \gamma(x))$$

for $x \in X$. Then $d_\Gamma(x) = 2 \, \mathrm{inj}(\pi(x))$. We say the injectivity radius function $\mathrm{inj}: M \to \mathbb{R}$ is *proper* if the preimage of a compact set is compact. The injectivity radius function is 1–Lipschitz. To see this, given any two points $p, q \in M$, let $\tilde{p}$ and $\tilde{q}$ be lifts of p and q in X whose distance is the same as the distance $d(p,q)$ of $p, q \in M$. There exists an isometry $\gamma \in \Gamma$ such that $d_X(\tilde{p}, \gamma\tilde{p}) = d_\Gamma(\tilde{p})$, and

$$2 \, \mathrm{inj}(q) \leq d_X(\tilde{q}, \gamma(\tilde{q})) \leq d_X(\tilde{q}, \tilde{p}) + d_X(\tilde{p}, \gamma(\tilde{p})) + d_X(\gamma(\tilde{p}), \gamma(\tilde{q}))$$
$$= 2d(p,q) + 2 \, \mathrm{inj}(p).$$

Hence, $\mathrm{inj}(q) - \mathrm{inj}(p) \leq d(p,q)$.

Now recall that the *critical exponent* $\delta(\Gamma)$ of a torsion-free discrete isometry group $\Gamma < \mathrm{Isom}(X)$ is defined to be

$$\delta(\Gamma) := \inf\left\{ s \;\middle|\; \sum_{\gamma \in \Gamma} \exp(-s d_X(p, \gamma(p))) < \infty \right\},$$

where p is a given point in X. Note that $\delta(\Gamma)$ is independent of the choice of p. Alternatively, one can also define the critical exponent $\delta(\Gamma)$ [38] as

$$(2\text{-}1) \qquad\qquad \delta(\Gamma) = \limsup_{R \to \infty} \frac{\log(N(R))}{R},$$

where $N(R) = \#\{\gamma \in \Gamma \mid d_X(x, \gamma(x)) \leq R\}$ for any given point $x \in X$.

We will need to use the following proposition later in the proofs:

Proposition 2.1 [32, Corollary 6.12] *Let $w \in M = X/\Gamma$ be a piecewise geodesic loop which consists of r geodesic segments, and let α be the closed geodesic freely homotopic to w such that $\ell(\alpha) \geq \epsilon > 0$. Then α is contained in the D–neighborhood of the loop w, where*

$$D = \cosh^{-1}(\sqrt{2})\lceil \log_2 r \rceil + \sinh^{-1}\left(\frac{2}{\epsilon}\right) + 2\delta_0.$$

Remark 2.2 The original corollary was stated under the extra assumption that α is simple. However, the proof of [32, Corollary 6.12] does not rely on this fact so we have removed the assumption here.

2.2 Thick–thin decomposition

Given an isometry $\gamma \in \mathrm{Isom}(X)$ and a constant $\epsilon > 0$, we define the *Margulis region* $\mathrm{Mar}(\gamma, \epsilon)$ of γ as

$$\mathrm{Mar}(\gamma, \epsilon) := \{x \in X \mid d_X(x, \gamma(x)) \leq \epsilon\}.$$

It is a convex subset by the convexity of the distance function. Given a point $x \in X$ and a constant $\epsilon > 0$, the set

$$\mathcal{F}_\epsilon(x) := \{\gamma \in \mathrm{Isom}(X) \mid d_X(x, \gamma(x)) \leq \epsilon\}$$

consists of all isometries that translate x by at most ϵ. For any discrete subgroup $\Gamma < \mathrm{Isom}(X)$, we denote by $\Gamma_\epsilon(x)$ the group generated by $\mathcal{F}_\epsilon(x) \cap \Gamma$. The Margulis lemma [3, Theorem 9.5] states that $\Gamma_\epsilon(x)$ is a finitely generated virtually nilpotent group for any $0 < \epsilon < \epsilon(n, \kappa)$, where $\epsilon(n, \kappa)$ is the Margulis constant depending on the dimension n of X and the sectional curvature bound κ.

We define the Γ–invariant set

$$\mathcal{T}_\epsilon(\Gamma) := \{p \in X \mid \Gamma_\epsilon(p) \text{ is infinite}\}.$$

The *thin part* (more precisely, the ϵ–thin part) of the quotient orbifold $M = X/\Gamma$, which we denote by $\mathrm{thin}_\epsilon(M)$, is defined to be $\mathcal{T}_\epsilon(\Gamma)/\Gamma$. The closure of the complement $M \setminus \mathrm{thin}_\epsilon(\Gamma)$ is called the *thick part* of M and is denoted by $\mathrm{thick}_\epsilon(M)$. The thin part consists of bounded and unbounded components. The bounded components are called the *Margulis tubes*, and are neighborhoods of short closed geodesics of length no greater than ϵ. More precisely, for every point x in the closed geodesic and every tangent vector v at x perpendicular to the geodesic, we consider a unit-speed ray ρ emanating from x in the direction of v. There exists R, depending on x and v, such that

$$d_\Gamma(\rho(R)) = \epsilon \quad \text{and} \quad d_\Gamma(\rho(t)) < \epsilon$$

for all $t < R$. We call the arc $\rho([0, R])$ a *maximal radial arc*, and a Margulis tube is the union of all radial arcs emanating from a short closed geodesic. For details, see for example [16].

The unbounded components are called the *Margulis cusps*, and can be described more precisely as follows. Denote the fixed-point set of Γ by

$$\operatorname{Fix}(\Gamma) := \bigcap_{\gamma \in \Gamma} \operatorname{Fix}(\gamma).$$

A discrete subgroup $P < \Gamma$ is called a *parabolic subgroup* if $\operatorname{Fix}(P)$ consists of a single point $\xi \in \partial_\infty X$. Given a constant $0 < \epsilon < \epsilon(n, \kappa)$ and a maximal parabolic subgroup $P < \Gamma$, the set $\mathcal{T}_\epsilon(P) \subset X$ is precisely invariant under P, and we have $\operatorname{stab}_\Gamma(\mathcal{T}_\epsilon(P)) = P$; see [12, Corollary 3.5.6]. In this case, $\mathcal{T}_\epsilon(P)/P$ can be regarded as a subset of M, called a Margulis cusp. The *cuspidal* part of M is the union of all Margulis cusps, denoted by $\operatorname{cusp}_\epsilon(M)$. Note that $\operatorname{cusp}_\epsilon(M) \subset \operatorname{thin}_\epsilon(M)$.

In our context, the parabolic subgroups in Γ (hence also the cuspidal part of M) turn out to be very simple due to the following proposition:

Proposition 2.3 Let $\Gamma < \operatorname{Isom}(X)$ be a torsion-free discrete isometry group, and $P < \Gamma$ be any parabolic subgroup. If δ is the critical exponent of Γ and P has polynomial growth rate r, then we have $r \leq 2\delta$. Thus:

(1) If $\delta < 1$, then all parabolic subgroups (if they exist) are isomorphic to $\mathbb{Z}$.

(2) If $\delta < \frac{1}{2}$, then all nontrivial isometries in Γ are loxodromic.

Proof Let $\mathcal{H}$ be a horosphere that P acts on and choose any basepoint $O \in \mathcal{H}$. Denote by $d_\mathcal{H}$ the horospherical distance and by d_P the Cayley metric with respect to some fixed finite generating set of P. Then there exists a constant $C > 0$ such that

$$(2\text{-}2) \qquad\qquad d_\mathcal{H}(O, \gamma(O)) \leq C d_P(1, \gamma)$$

holds for all $\gamma \in P$. By [24, Theorem 4.6] there exists a constant $C' > 0$ such that, for any $p, q \in \mathcal{H}$ with $d_X(p, q) > C'$, we have

$$(2\text{-}3) \qquad\qquad d_X(p, q) \leq 2\ln(C' d_\mathcal{H}(p, q)).$$

By possibly replacing C or C' by a larger constant, we may assume $C' = C$. Therefore we obtain, from the above the asymptotic inequalities (for R large),

$$|\{\gamma \in P : d_P(1, \gamma) \leq R\}| \leq |\{\gamma \in P : d_\mathcal{H}(O, \gamma(O)) \leq CR\}| \qquad \text{(by (2-2))}$$
$$\lesssim |\{\gamma \in P : d_X(O, \gamma(O)) \leq 2\ln(C^2 R)\}| \quad \text{(by (2-3))}$$

$$\simeq e^{2\ln(C^2 R)\delta(P)} \qquad \qquad \text{(by (2-1))}$$
$$\simeq R^{2\delta(P)},$$

where $\delta(P)$ is the critical exponent of P. Since $\delta(P) \leq \delta$, it follows that $r \leq 2\delta$.

In particular, if $\delta < 1$, then $r < 2$ and by the Bass–Guivarc'h formula [4; 22], P must be virtually $\mathbb{Z}$. But since P is torsion-free, it must be $\mathbb{Z}$ [43]. If $\delta < \frac{1}{2}$, then $r < 1$ and P cannot exist. Thus all nontrivial elements in Γ are loxodromic. $\qquad \square$

2.3 Geometric finiteness

Recall that the *limit set* $\Lambda(\Gamma)$ of a discrete subgroup $\Gamma < \mathrm{Isom}(X)$ is defined to be the set of accumulation points of the Γ–orbit $\Gamma(p)$ in $\partial_\infty X$, where p is an arbitrary given point in X, and that the definition is independent of the choice of p. If $\Lambda(\Gamma)$ is finite, then Γ is called *elementary*. Otherwise, it is called *nonelementary*. A point $\xi \in \Lambda(\Gamma)$ is called *a conical limit point* if every geodesic ray $\rho: \mathbb{R}_+ \to X$ asymptotic to ξ projects to a nonproper map $\pi \circ \rho: \mathbb{R}_+ \to M = X/\Gamma$. We denote by $\Lambda_c(\Gamma)$ the set of all conical limit points.

We denote by $\mathrm{Hull}(\Lambda) \subset X$ the closed convex hull of $\Lambda \subset \partial_\infty X$, which is the smallest closed convex subset in X whose accumulation set in $\partial_\infty X$ is Λ, and by $C(\Gamma) = \mathrm{Hull}(\Lambda)/\Gamma$ the *convex core* of Γ.

A discrete isometry subgroup $\Gamma < \mathrm{Isom}\, X$ is *geometrically finite* if the noncuspidal part of the convex core $C(\Gamma)$ in $M = X/\Gamma$ is compact. Otherwise, it is called *geometrically infinite*. If $C(\Gamma)$ is compact, then the discrete subgroup Γ is called *convex cocompact*.

There are various equivalent definitions of geometric finiteness, but we will only mention one of them, proved by Kapovich and the first author. For the other equivalent definitions we refer the readers to [12]. The following theorem is a generalization of a previous result of Bonahon [10]:

Theorem 2.4 [32, Theorem 1.5] *A discrete subgroup $\Gamma < \mathrm{Isom}(X)$ is geometrically infinite if and only if there exists a sequence of closed geodesics $\alpha_i \subset M = X/\Gamma$ which escapes every compact subset of M.*

2.4 Admissible surfaces

In this section, we give a sketch of the existence of smooth admissible surfaces. This can be treated as a smooth version of [17, Section 1.1.5]. In our case, we will need a slightly broader category of admissible surfaces than smooth maps in order to include

the gluing of two maps along a smooth boundary. In general the notion of a piecewise smooth map is rather technical (using Whitney stratification), but we only consider maps from a smooth surface with boundary to a smooth manifold. Thus we simplify the notion:

Definition 2.5 Given a smooth surface Σ (possibly with boundary) and a smooth manifold M, we say a map $f : \Sigma \to M$ is a *piecewise smooth* map if there is a smooth triangulation $\Delta = \{\sigma_1, \ldots, \sigma_m\}$ on Σ (ie edges are all smooth paths) such that:

(1) f is continuous.

(2) f is smooth on the interior of each face σ_i.

(3) If $e = \sigma_i \cap \sigma_j$ is a common edge, then the restriction $f|_p$ is smooth.

Roughly speaking, a piecewise smooth map is just a finite concatenation of smooth maps, possibly pleating along the gluing edges. The singular set forms a piecewise smooth 1–skeleton on Σ. Now we return to our context, where $M = X/\Gamma$ is a complete pinched negatively curved manifold. Suppose $\{\eta_1, \ldots, \eta_k\}$ is a collection of k smooth loops in M. If there exists a set of integers $c_1, \ldots, c_k$ such that $\sum_{i=1}^{k} c_i[\eta_i] = 0$ in $H_1(M, \mathbb{Z})$, then we claim that $\bigcup_i c_i \eta_i$ will bound a piecewise smooth surface in the sense explained below.

Choose a basepoint $x_0 \in M$ and connect x_0 to each of the loops η_i by a smooth path p_i. Then the loop $q_i := p_i * (c_i \eta_i) * p_i^{-1}$ is free homotopic to $c_i \eta_i$, which also represents an element $\gamma_i \in \Gamma \cong \pi_1(M, x_0)$. Since $\sum_{i=1}^{k} c_i[\eta_i] = 0$ in $H_1(M, \mathbb{Z}) \cong \Gamma/[\Gamma, \Gamma]$, it follows that the product $\gamma = \gamma_1 \cdots \gamma_k$ is an element in the commutator subgroup $[\pi_1(M, x_0), \pi_1(M, x_0)]$. Thus we can write

$$\gamma = [a_1, b_1] \cdots [a_g, b_g]$$

for some $a_i, b_i \in \Gamma$. We choose smooth loops α_i and β_i from x_0 that represent a_i and b_i, respectively. Fix a preimage $\tilde{x}_0 \in X$ of x_0 under the projection map $\pi : X \to M$. The loop $\sigma = \alpha_1 * \beta_1 * \alpha_1^{-1} * \beta_1^{-1} * \cdots * \alpha_g * \beta_g * \alpha_g^{-1} * \beta_g^{-1} * (q_1 * \cdots * q_k)^{-1}$ is nullhomotopic, thus lifts to a piecewise smooth loop on X. Therefore it bounds a smooth disk on X, that is, there exists a disk $D \subset \mathbb{R}^2$ and a piecewise smooth map $f : D \to X$ with $f(\partial D) = \sigma$. Moreover, by identifying D with a $(4g+3k)$–polygon with the label of $\prod_{i=1}^{g}[\bar{a}_i, \bar{b}_i]\bar{p}_1 \ell_1 \bar{p}_1^{-1} \cdots \bar{p}_k \ell_k \bar{p}_k^{-1}$, we can make the map f explicit by sending the edge labels $\bar{a}_i, \bar{b}_i, \bar{a}_i^{-1}, \bar{b}_i^{-1}, \bar{p}_i, \ell_i$ and $\bar{p}_i^{-1}$ to $\alpha_i, \beta_i, \alpha_i^{-1}, \beta_i^{-1}, p_i, c_i \eta_i$ and p_i^{-1}, respectively. Therefore, after gluing along the edge labels, f descends to

a piecewise smooth map from $\Sigma_{g,k}$ (a genus g surface with k boundary components) to M, which sends the boundary components (corresponding to ℓ_i) to $c_i \eta_i$.

In general:

Definition 2.6 Let Σ be a compact oriented (not necessarily connected) surface with k boundary components. Given a collection of k loops $\{\alpha_1, \ldots, \alpha_k\}$ on M, we say a map $f : \Sigma \to M$ is *admissible* with respect to $\{\alpha_1, \ldots, \alpha_k\}$ if the following diagram commutes:

$$
\begin{array}{ccc}
\partial\Sigma & \xrightarrow{\ i\ } & \Sigma \\
{\scriptstyle \partial f}\big\downarrow & & \big\downarrow{\scriptstyle f} \\
\bigcup_{i=1}^{k} \alpha_i & \xrightarrow{\ i\ } & M
\end{array}
$$

Note that α_i could carry multiplicities, and the orientation of the surface Σ induces an orientation on $\partial\Sigma$. In the above commutative diagram we also require ∂f to preserve the orientations. If there exist such Σ and f, then we simply say $\bigcup_{i=1}^{k} \alpha_i$ bounds a surface f.

By the above discussion:

Proposition 2.7 *Suppose $\{\alpha_1, \ldots, \alpha_k\}$ is a collection of k smooth loops in M. If there exists a set of integers $c_1, \ldots, c_k$ such that $\sum_{i=1}^{k} c_i[\alpha_i] = 0$ in $H_1(M, \mathbb{Z})$, then there exists a piecewise smooth admissible map with respect to $\{c_1\alpha_1, \ldots, c_k\alpha_k\}$, that is, $\bigcup_{i=1}^{k} c_i\alpha_i$ bounds a piecewise smooth surface $f : \Sigma \to M$.*

Given two Riemannian manifolds N and M, a smooth map $F : N \to M$ and a positive integer $\mathfrak{p} \leq \min\{\dim(N), \dim M\}$, the $\mathfrak{p}$–Jacobian of F at a point $x \in N$ is defined to be

$$
\mathrm{Jac}_{\mathfrak{p}}(F)(x) = \sup \|dF_x(e_1) \wedge dF_x(e_2) \wedge \cdots \wedge dF_x(e_{\mathfrak{p}})\|,
$$

where the supremum is taken over all orthonormal $\mathfrak{p}$–frames $\{e_1, \ldots, e_{\mathfrak{p}}\}$ on $T_x N$, and the norm is induced by the Riemannian inner product at $T_{F(x)} M$. Note that when $\mathfrak{p} = \dim N \leq \dim M$, the $\mathfrak{p}$–Jacobian of F coincides with $\sqrt{\det_{g_N} F^* g_M}$.

Definition 2.8 Given a Riemannian manifold M, a smooth map $f : \Sigma \to M$ and a smooth region $U \subset \Sigma$, we define the area of the map on U to be

$$
A(f|_U) := \int_U |\mathrm{Jac}_2\, f|(x)\, dV_\Sigma,
$$

where dV_Σ is the volume form on Σ with respect to some chosen Riemannian metric g_Σ, and it is clear the definition of area is independent of the choice of g_Σ. When $U = \Sigma$,

we simply denote it by $A(f)$. The definition naturally extends to a piecewise smooth map. Note that, at the region where df is degenerate, $(\text{Jac}_2\, f)$ vanishes, so it does not contribute to the area.

2.5 Besson–Courtois–Gallot map

In this section, we give a brief introduction to the Besson–Courtois–Gallot map and we refer the readers to [6] for a more detailed exposition. First we recall that, given any discrete subgroup $\Gamma < \text{Isom}(X)$, there exists a family of positive finite Borel measures called the Patterson–Sullivan measures, which satisfy:

(1) μ_x is Γ–equivariant for all $x \in X$.

(2) $d\mu_x(\theta) = e^{-\delta B(x,\theta)} d\mu_o(\theta)$ for all $x \in X$ and $\theta \in \partial_\infty X$.

Here δ is the critical exponent of Γ, o is a basepoint on X, and $B(x, \theta)$ is the Busemann function on X with respect to o. Recall that the Busemann function B is defined by

$$B(x,\theta) = \lim_{t\to\infty} \big(d(x, \alpha_\theta(t)) - t\big),$$

where $\alpha_\theta(t)$ is the unique geodesic ray from o to θ.

We note that the Busemann function $B(x, \theta)$ is convex on X. If μ is any finite Borel measure supported on at least two points on $\partial_\infty X$, then the function

$$x \mapsto \mathcal{B}_\mu(x) := \int_{\partial_\infty X} e^{B(x,\theta)}\, d\mu(\theta)$$

is strictly convex, and one can check it tends to $+\infty$ as $x \to \partial_\infty X$. Hence we can define the barycenter $\text{bar}(\mu)$ of μ to be the unique point in X where the function attains its minimum.

Now we construct the map $\widetilde{F}\colon X \to X$ given by

$$x \mapsto \text{bar}(e^{-B(x,\theta)}\mu_x),$$

where $e^{-B(x,\theta)}\mu_x$ denotes the unique (up to measure zero) Borel measure which is absolutely continuous with respect to μ_x, with the corresponding Radon–Nikodym derivative $e^{-B(x,\theta)}$.

Theorem 2.9 (Besson–Courtois–Gallot [6])   *The map $\widetilde{F}\colon X \to X$ constructed above satisfies*:

(1) $\widetilde{F}$ *is Γ–equivariant, and thus descends to a map $F\colon M \to M$.*

(2) F *is smooth and homotopic to the identity.*

(3) $|\text{Jac}_\mathfrak{p}(F)(x)| \le ((1+\delta)/\mathfrak{p})^\mathfrak{p}$ *for any integer $\mathfrak{p} \in [1, \dim M]$ and any $x \in M$.*

Remark 2.10 The case of $\mathfrak{p} = 1$ in (3) is not directly stated in the paper, however it is clear from the 2–form equation [6, (4.11)] that $\|dF\| \leq (1 + \delta)$. According to the theorem, if $\delta \leq \mathfrak{p} - 1$, then $|\mathrm{Jac}_\mathfrak{p}(F)| \leq 1$ hence F is a $\mathfrak{p}$–dimensional volume-decreasing map. However, in order to obtain the linear isoperimetric inequality in Section 3.1, we will need an area-decreasing map, which is assured only in the case $\delta < 1$. Thus, we will only apply the theorem to the cases $\mathfrak{p} = 1, 2$.

Notation

Henceforth X always denotes a negatively pinched Hadamard manifold with sectional curvature $-\kappa^2 \leq K \leq -1$, and $\Gamma < \mathrm{Isom}(X)$ denotes a torsion-free discrete isometry subgroup. Let $M = X/\Gamma$ be the quotient manifold, $\pi \colon X \to M$ be the quotient map, and d be the distance on M. Let δ denote the critical exponent of Γ and $C(\delta) = 4/(1 - \delta)$. We use ℓ and A to denote the length and area functions, respectively. We let $\mathrm{inj}(x)$ denote the injectivity radius at a point $x \in M$, and let $\mathrm{NJ}(S)$ denote the normal injectivity radius of a submanifold $S \subset M$; see Section 3.2.

3 Geometry with small critical exponent

In this section, we investigate the geometry of the quotient manifold M under the assumption $\delta < 1$.

3.1 Linear isoperimetric type inequality

The study of the isoperimetric problem has a long and significant history. In the classical context, given a region $\Omega \subset \mathbb{R}^2$, it is natural to ask what the optimal relation between its area $A(\Omega)$ and the length of its bounding curve $\ell(\partial\Omega)$ is. It is proved that there is a quadratic relation $A(\Omega) \leq \ell(\partial\Omega)^2/4\pi$, and that equality holds if and only if Ω has a circular boundary. However, our main interest has driven us to work in a slightly different context. Let $M = X/\Gamma$ be a complete quotient manifold and $\mathcal{C} \subset M$ be a union of smooth loops which represents a trivial homology class in M. By the discussion in Section 2.4, $\mathcal{C}$ bounds an admissible surface. Among all admissible surfaces, we find one surface Σ such that $A(\Sigma)$ and $\ell(\partial\Sigma)$ satisfy a linear isoperimetric type inequality.

Definition 3.1 A family of loops $\mathcal{F} = \{\alpha_1, \ldots, \alpha_k\}$ in M is *irreducible* if either

(1) $k = 1$ and α_1 represents a trivial or torsion homology class, or

(2) $\mathcal{F}$ consists of linearly dependent loops, and any nontrivial subfamily of $\mathcal{F}$ is linearly independent.

Suppose $\mathcal{F} = \{\alpha_1, \ldots, \alpha_k\}$ is an irreducible family of loops. In case (1), $\mathcal{F}$ consists of one homology class $[\alpha]$, so there is a minimal positive integer c such that $c[\alpha] = 0$. In case (2), there exists a unique (up to a sign) set of integers $c_1, \ldots, c_k$ such that $\gcd(c_1, \ldots, c_k) = 1$ and $\sum_{i=1}^{k} c_i[\alpha_i] = 0$ in $H_1(M)$. Thus, there exist admissible surfaces in M with respect to $c[\alpha]$ (or $\bigcup_{i=1}^{k} c_i \alpha_i$) and by irreducibility they are necessarily connected. Note that $c_i \alpha_i$ denotes the c_i multiple of α_i, and c_i being negative corresponds to reversing the orientation of α_i. We call the set of integers $c_1, \ldots, c_k$ (or, in case 1, c) the *associated integers* of the irreducible family.

Theorem 3.2 *Let $\mathcal{F} = \{\alpha_1, \ldots, \alpha_k\}$ be any family of smooth loops in M which are linearly dependent in $H_1(M, \mathbb{Z})$ such that there are integers $c_1, \ldots, c_k$ satisfying $\sum_{i=1}^{k} c_i[\alpha_i] = 0$ in $H_1(M)$. Suppose the critical exponent δ is less than 1. Then $\bigcup_{i=1}^{k} c_i \alpha_i$ bounds a smooth surface $f_0 \colon \Sigma \to M$ whose area satisfies*

$$A(f_0) \leq \frac{4}{1-\delta} \ell(f_0(\partial \Sigma)) = \frac{4}{1-\delta}\left(\sum_{i=1}^{k} |c_i| \ell(\alpha_i) \right).$$

Proof We may assume $\mathcal{F}$ is irreducible. Otherwise, we decompose $\mathcal{F}$ into irreducible subfamilies and use the additivity of area and length functions on disjoint unions. We consider the set $\mathfrak{S}$ which consists of all piecewise smooth surfaces bounded by $\bigcup_{i=1}^{k} c_i \alpha_i$, or more precisely, we set $\mathfrak{S}$ equal to

$$\{ f \colon \Sigma \to M \mid f \text{ is piecewise smooth admissible with respect to } \{c_1\alpha_1, \ldots, c_k\alpha_k\} \}.$$

By Proposition 2.7 it is nonempty. Let $A_0 = \inf\{A(f) : f \in \mathfrak{S}\}$. To avoid possible existence and regularity issues (see the following remark) of minimal surfaces in M, we can choose a piecewise smooth admissible map $f_\epsilon \in \mathfrak{S}$ such that $A(f_\epsilon) \leq (1+\epsilon)A_0$ for any $\epsilon > 0$. Composing with the Besson–Courtois–Gallot map F as described in Section 2.5, we obtain a piecewise smooth admissible map $F \circ f_\epsilon$ with respect to $\bigcup_{i=1}^{k} c_i F(\alpha_i)$. By Theorem 2.9 we have the area estimate

$$A(F \circ f_\epsilon) = \int_\Sigma |\mathrm{Jac}_2(F \circ f_\epsilon)| \, dV_\Sigma \leq \int_\Sigma |\mathrm{Jac}_2 F| \cdot |\mathrm{Jac}_2 f_\epsilon| \, dV_\Sigma$$
$$\leq \left(\tfrac{1}{2}(1+\delta)\right)^2 A(f_\epsilon) \leq \left(\tfrac{1}{2}(1+\delta)\right)^2 (1+\epsilon)A_0,$$

and the length estimate $\ell(F(\alpha_i)) \leq (1+\delta)\ell(\alpha_i)$. For each α_i, since $F(\alpha_i)$ is free homotopic to α_i, we can build an (immersed) cylindrical homotopy $\Sigma_i \subset M$ between them by taking the image of the union of two geodesic cones $\mathrm{Cone}_p(\widetilde{F}(\tilde{\alpha}))$ and $\mathrm{Cone}_{\gamma(q)}(\tilde{\alpha})$ under the projection $\pi \colon X \to M$; see Figure 1. Here $\gamma \in \Gamma$ is an element
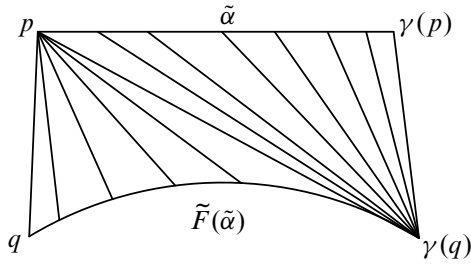
Figure 1

represented by α, $\tilde{\alpha}$ is a lift of α, and p and q as well as $\gamma(p)$ and $\gamma(q)$ are connected by geodesics. To estimate the area of Σ_i, we will need:

Lemma 3.3 *For any $p \in X$ and any smooth curve $\alpha \subset X$, the geodesic cone $\mathrm{Cone}_p(\alpha)$ has the area bound*

$$A(\mathrm{Cone}_p(\alpha)) \leq \ell(\alpha).$$

Proof We parametrize the smooth curve by $\alpha : [0, 1] \to X$, and write $D(s) = d(p, \alpha(s))$. The geodesic cone $\mathrm{Cone}_p(\alpha)$ can be parametrized by the smooth map

$$\Phi : [0, 1] \times [0, D(s)] \to X, \quad (s, t) \mapsto \exp_p(t\beta(s)),$$

where $\beta(s)$ is the unit vector in the direction of the preimage of α under the exponential map, that is, the unique curve in $T_p X$ satisfying $\exp_p(D(s)\beta(s)) = \alpha(s)$. Since $\alpha(s) = \Phi(s, D(s))$, we have

$$\alpha'(s) = \left[\frac{\partial \Phi}{\partial s} + \frac{\partial \Phi}{\partial t} D'(s) \right](s, D(s)).$$

Let $\gamma_s(t) = \Phi(s, t)$. For each s, $\gamma_s(t)$ is a unit-speed geodesic connecting p to $\alpha(s)$, so, at any point $(s, t) \in [0, 1] \times [0, D(s)]$,

$$\frac{\partial \Phi}{\partial t} = \gamma_s'(t), \quad \frac{\partial \Phi}{\partial s} = J_s(t),$$

where $J_s(t)$ is the unique Jacobi field along γ_s satisfying $J_s(0) = 0$ and

$$J_s(D(s)) = \frac{\partial \Phi}{\partial s}(s, D(s)) = \alpha'(s) - \gamma_s'(D(s))D'(s),$$

which is the projection of $\alpha'(s)$ orthogonal to $\gamma_s'(D(s))$. This implies that $J_s(t)$ is a normal Jacobi field and that $\partial \Phi / \partial t \perp \partial \Phi / \partial s$. Therefore

$$|\mathrm{Jac}(\Phi)| = \left\| \frac{\partial \Phi}{\partial s} \wedge \frac{\partial \Phi}{\partial t} \right\| = \left\| \frac{\partial \Phi}{\partial s} \right\| \cdot \left\| \frac{\partial \Phi}{\partial t} \right\| = \| J_s(t) \|.$$

Using [24, Proposition 2.3] and the curvature assumption $K \leq -1$, we can estimate the norm of the Jacobi fields by

(3-1) $$\| J_s(t) \| \leq \frac{\sinh t}{\sinh(D(s))} \| J_s(D(s)) \| \leq \frac{\sinh t}{\sinh(D(s))} \| \alpha'(s) \|.$$

Finally we obtain the area estimate of the geodesic cone:

$$
(3\text{-}2) \qquad A(\mathrm{Cone}_p(\alpha)) \leq \int_0^1 \int_0^{D(s)} |\mathrm{Jac}(\Phi)|\, dt\, ds
$$

$$
\leq \int_0^1 \int_0^{D(s)} \frac{\sinh t}{\sinh(D(s))} \|\alpha'(s)\|\, dt\, ds \quad \text{(by (3-1))}
$$

$$
\leq \int_0^1 \|\alpha'(s)\|\, ds \leq \ell(\alpha). \qquad \square
$$

Now we continue with the proof. By the lemma above,

$$
(3\text{-}3) \qquad\qquad A(\Sigma_i) \leq \ell(\alpha_i) + \ell(F(\alpha_i)) \leq (2+\delta)\ell(\alpha_i).
$$

Here Σ_i is a piecewise immersed surface in M and we can choose any piecewise smooth parametrization $\sigma_i \colon S^1 \times [0,1] \to M$ to represent Σ_i. If we concatenate each σ_i with $F \circ f_\epsilon$ (glue $\bigcup_{i=1}^k c_i \Sigma_i$ onto $F \circ f_\epsilon(\Sigma)$ on M), we get a new piecewise smooth admissible surface f'_ϵ with respect to $\bigcup_{i=1}^k c_i \alpha_i$, and by assumption $A(f'_\epsilon) \geq A_0$. On the other hand, combining the above inequalities,

$$
A_0 \leq A(f'_\epsilon) = A(F \circ f_\epsilon) + \sum_{i=1}^k |c_i| A(\Sigma_i)
$$

$$
\leq \left(\tfrac{1}{2}(1+\delta)\right)^2 (1+\epsilon) A_0 + (2+\delta)\left(\sum_{i=1}^k |c_i|\ell(\alpha_i)\right) \quad \text{(by (3-2) and (3-3))}.
$$

Thus, by letting ϵ tend to zero, we obtain

$$
A_0 \leq \frac{4(2+\delta)}{(1-\delta)(3+\delta)}\left(\sum_{i=1}^k |c_i|\ell(\alpha_i)\right) < \frac{4}{1-\delta}\left(\sum_{i=1}^k |c_i|\ell(\alpha_i)\right).
$$

Therefore we can always choose a piecewise smooth map in $\mathfrak{S}$ whose area is arbitrarily close to A_0, and finally we can always smoothen it with an arbitrarily small increase on the area. In particular, there is a smooth admissible map f_0 with area

$$
A(f_0) \leq \frac{4}{1-\delta}\left(\sum_{i=1}^k |c_i|\ell(\alpha_i)\right). \qquad \square
$$

Remark 3.4 The existence and regularity of minimal surfaces for a general complete manifold relate to the generalized Plateau problem, which has been studied in [37]. If there is a uniform lower bound on the injectivity radius on M, then the condition of "homogeneously regular" in [37] is satisfied; hence, the existence and regularity

of the area minimizer hold. Although in Theorem 3.7 we manage to show M has bounded geometry, the proof relies on this theorem; hence, using this would fall into circular reasoning.

We do not pursue the optimal bound in the theorem above. Indeed, the linear isoperimetric constant we produce via this method will always tend to infinity as $\delta \to 1$. This stands as an obstacle in improving our main theorems as δ approaches 1.

3.2 Asymptotically uniformly bounded tubular neighborhood

Let S be a closed submanifold of M, $N(S, M) = \{(x, v) \in TM : x \in S \text{ and } v \perp T_x S\}$ be the *normal bundle* of S in M, and $N_r(S, M) = \{(x, v) \in N(S, M) : |v| < r\}$ be the *r–normal bundle* of S in M. The *normal exponential map* $\exp_S$ is defined to be the restriction of the exponential map $\exp: TM \to M$ to the normal bundle $N(S, M)$ of S in M. The *normal injectivity radius* $\mathrm{NJ}(S)$ is defined to be the supremum of r such that $\exp_S$ is an embedding on $N_r(S, M)$. In the case where $r \leq \mathrm{NJ}(S)$, we say $\exp_S(N_r(S, M)) = \{x \in M \mid d(x, S) < r\}$ is the *r–tubular neighborhood* of S in M, and we denote it by $T_r(S)$. By convention, if the submanifold has a self-intersection, we declare that it has normal injectivity radius zero.

Lemma 3.5 *Let α be a closed geodesic in M with $\mathrm{NJ}(\alpha) = R > 0$, and let $T_R(\alpha)$ be its R–tubular neighborhood in M. If $i : \Sigma \to M$ is any smooth admissible map with respect to $\{k\alpha, \alpha'\}$ such that either α' is empty or α' consists of a union of smooth loops outside of $T_R(\alpha)$ (ie $d_M(\alpha', \alpha) > R$), then*

$$A(i|_{i^{-1}(T_R(\alpha))}) \geq kR\ell(\alpha).$$

Proof We choose a Riemannian metric g_0 on Σ, and let ϵ_1 and ϵ_2 be two positive real numbers recognized to be small and to be determined later. First, we perturb the pullback metric $i^* g_M$ to be Riemannian on Σ by setting $g = i^* g_M + \epsilon_1 g_0$ and use this to estimate the area of i. It follows that, for any $\epsilon > 0$ and any region $U \subset \Sigma$,

$$
\begin{aligned}
(3\text{-}4) \qquad |\mathrm{vol}_g(U) - A(i|_U)| &= \left| \int_U 1 \, dV_g - \int_U |\mathrm{Jac}_2\, i| \, dV_{g_0} \right| \\
&= \int_U \left(\sqrt{\det_{g_0}(g)} - \sqrt{\det_{g_0}(i^* g_M)} \right) dV_{g_0} \\
&\leq \int_\Sigma \left(\sqrt{\det_{g_0}(g)} - \sqrt{\det_{g_0}(i^* g_M)} \right) dV_{g_0} < \epsilon,
\end{aligned}
$$

after choosing ϵ_1 small enough. Note that this follows from the continuity of the determinant function, and that the estimate is uniform on U.

Next, we choose a suitable function on Σ and use the coarea formula to estimate $\mathrm{vol}_g(U)$. Denote by $\sigma \subset \partial\Sigma$ the boundary component which sends to $k\alpha$ under i, and by $\rho_\alpha \colon M \to \mathbb{R}$ the distance function to α on M. Now we construct a function $f \colon \Sigma \to \mathbb{R}$ by setting

$$f = \rho_\alpha \circ i + \epsilon_2\varphi,$$

where φ is a smooth function on Σ chosen so that:

(1) $\varphi(x) = 0$ on σ and $\varphi(x) > 0$ on $\Sigma \setminus \sigma$.

(2) There exists a collar neighborhood V of σ such that $d\varphi(x) \neq 0$ when $x \in V \setminus \sigma$.

For example, one can choose φ to be the distance function to σ on its local neighborhood and then extend smoothly to any positive function outside. For this choice, it is clear that $f(x) \geq 0$ and $f^{-1}(0) = \sigma$. Since M is negatively curved, there is no conjugate point for M. Thus, for any $y \in T_R(\alpha)$, there is a unique geodesic projection onto α, so ρ_α is smooth on $T_R(\alpha) \setminus \alpha$. It follows that f is smooth on $i^{-1}(T_R(\alpha)) \setminus \sigma \subset \Sigma$. We can estimate the norm of its differential with respect to the metric g by

$$(3\text{-}5) \qquad \|df\| = \|d\rho_\alpha \circ di + \epsilon_2 d\varphi\|$$
$$\leq \|d\rho_\alpha\| \cdot \|di\| + \epsilon_2\|d\varphi\| \quad \text{(note that } i \text{ is 1–Lipschitz)}$$
$$< (1 + \epsilon),$$

after choosing ϵ_2 small enough. This uses the compactness of Σ.

Finally we estimate the area of i on $i^{-1}(T_R(\alpha))$. By the construction of f, we have $f^{-1}([0, R)) \subset i^{-1}(T_R(\alpha))$. Thus, if we set $U = f^{-1}([0, R))$, then

$$\mathrm{vol}_g(U) \leq \mathrm{vol}_g\big(i^{-1}(T_R(\alpha))\big).$$

On the other hand, by the coarea formula [15, Section 13.4], we obtain from (3-5) that

$$(3\text{-}6) \qquad \mathrm{vol}_g(U) > \frac{1}{1+\epsilon} \int_U \|df\| \, dV_g = \frac{1}{1+\epsilon} \int_0^R \ell_g(f^{-1}(t)) \, dt.$$

Note that in the above formula, $f^{-1}(t)$ might not be a smooth curve if t is a singular value. But by Sard's theorem, almost all values $r \in (0, R)$ are regular, in which case the level sets are unions of smooth circles on Σ, and ℓ_g denotes the total length of the circles. In particular, the above integral makes sense. Other boundary components (if any) of Σ do not intersect with $i^{-1}(T_R(\alpha))$ by assumption, so, given any regular value $t \in [0, R)$, $f^{-1}(t)$ (up to orientation) is homologous to $f^{-1}(0) = \sigma$ on Σ. Hence, taking their images in M, we obtain that $i(f^{-1}(t))$, which is also a union of smooth loops, is

homologous to $k\alpha$ on M. Since they are entirely contained in $T_R(\alpha)$, $i(f^{-1}(t))$ is in fact free homotopic to $k\alpha$. More precisely, for almost all $t \in (0, R)$, if we write $i(f^{-1}(t))$ as a disjoint union of circles $\bigcup_{i=1}^m \alpha_i$, then each α_i is a smooth loop free homotopic to $k_i\alpha$ for $k_i \in \mathbb{Z}$, since the fundamental group of the R–neighborhood of α is a cyclic group generated by the loop α. (Some k_i could be zero, in which case α_i is homotopically trivial in M.) Moreover, $\sum_{i=1}^m k_i = k$. Since α is a closed geodesic, we have that $\ell(i(f^{-1}(t))) = \sum_{i=1}^m \ell(\alpha_i) \geq \sum_{i=1}^m |k_i|\ell(\alpha) \geq k\ell(\alpha)$. Note that i is 1–Lipschitz, so $\ell_g(f^{-1}(t)) \geq \ell(i(f^{-1}(t)))$. Combining the above inequality with (3-4) and (3-6),

$$A(i|_{i^{-1}(T_R(\alpha))}) > \frac{1}{1+\epsilon} k R \ell(\alpha) - \epsilon.$$

Since $\epsilon > 0$ is arbitrary, the lemma follows. $\qquad\square$

Lemma 3.6 *Assume we have N cusps in M and a constant $\epsilon > 0$ small enough that $\{M_{12\epsilon}^{(i)} : 1 \leq i \leq N\}$ are disjoint components of the cuspidal part $\mathrm{cusp}_{12\epsilon}(M)$. Suppose $\iota: \Sigma \to M$ bounds an irreducible collection of smooth loops $\bigcup_{i=1}^N c_i\alpha_i$, where each α_i is contained in the 2ϵ–thinner part $M_{2\epsilon}^{(i)} \subset M_{12\epsilon}^{(i)}$ in each cusp component and is homologically nontrivial. Then*

$$A(\iota) \geq 4\epsilon^2.$$

Proof Since the collection is irreducible and α_1 is homologically nontrivial in its cusp component (which might be homologically trivial in M), $\iota(\Sigma)$ has to leave $M_{12\epsilon}^{(1)}$. We will only focus on the region $U_0 := \iota^{-1}(M_{12\epsilon}^{(1)})$ as shown in Figure 2. If we let $M_{4\epsilon}^{(1)} \subset M_{12\epsilon}^{(1)}$ be the 4ϵ–thinner part and set $T_1 = M_{12\epsilon}^{(1)} \setminus M_{4\epsilon}^{(1)}$, then certainly

$$A(\iota) \geq A(\iota|_{i^{-1}(T_1)}).$$

So it suffices to give a lower bound on the area restricted to the T_1 region.

Similar to the proof of Lemma 3.5, we first choose the same perturbed Riemannian metric on Σ as $g = \iota^* g_M + \epsilon_1 g_0$, and for any $\epsilon' > 0$ the estimate of (3-4) still works after choosing ϵ_1 small enough. Thus, for any $U \subset \Sigma$, we have

$$(3\text{-}7) \qquad\qquad |\mathrm{vol}_g(U) - A(\iota|_U)| < \epsilon'.$$

Denote by $\sigma \subset \partial\Sigma$ the boundary component which maps to $c_1\alpha_1$ under ι, and let φ be, as before, the smooth function on Σ such that:

(1) $\varphi(x) = 0$ on σ and $\varphi(x) > 0$ on $\Sigma \setminus \sigma$.

(2) There exists a collar neighborhood V of σ such that $d\varphi(x) \neq 0$ when $x \in V \setminus \sigma$.
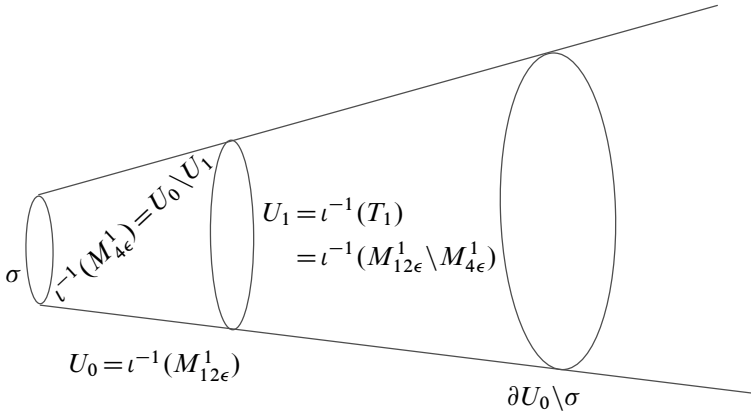
Figure 2

We choose a smooth approximation [21, Proposition 2.1] of the injectivity radius function on a neighborhood of $\iota(\Sigma)$, denoted by j, such that

(1) $j > 0$ on $\iota(\Sigma)$,

(2) j is $(1+\epsilon')$–Lipschitz, and

(3) $|j(y) - \mathrm{inj}(y)| < \epsilon$ on $\iota(\Sigma)$.

Choose a smooth bump function $0 \leq \psi \leq 1$ on Σ such that $\psi = 1$ on $\iota^{-1}(T_1)$ and $\psi = 0$ on σ. Since Σ is compact, there exists $\mathcal{K} > 0$ such that $\|\varphi\| < \mathcal{K}$ and $\|d\varphi\| < \mathcal{K}$. Choose a positive constant $\epsilon_2 < \min\{\epsilon, \epsilon'\}/\mathcal{K}$. Now define the smooth function $f: \Sigma \to \mathbb{R}$ by

$$f = \epsilon_2 \varphi + \psi(j \circ \iota).$$

By the construction of f, we see that $f(x) \geq 0$ on U_0 and $f^{-1}(0) = \sigma$. When restricting to $U_1 := \iota^{-1}(T_1) = \iota^{-1}(M^{(1)}_{12\epsilon} \setminus M^{(1)}_{4\epsilon})$, the norm of its differential under the metric g can be estimated by

$$(3\text{-}8) \qquad \|df\|_{U_1} = \|\epsilon_2 d\varphi + dj \circ d\iota\| \leq \epsilon_2 \|d\varphi\| + \|dj\| \cdot \|d\iota\| < 1 + 2\epsilon'.$$

The first inequality follows from the fact that $\psi = 1$ on $\iota^{-1}(T_1)$, and the last inequality uses that ι is 1–Lipschitz and also the choice of j and ϵ_2. Now we investigate the value of f on U_0, and apply the coarea formula to give a lower bound for the area of $\iota|_{f^{-1}([4\epsilon, 5\epsilon]) \cap U_0}$.

Claim   The subset $f^{-1}([4\epsilon, 5\epsilon]) \cap U_0$ is contained in U_1, and $f^{-1}([0, 5\epsilon]) \cap U_0$ is disjoint from $\partial U_0 \setminus \sigma$.

Proof For any $x \in U_0 \setminus U_1 = \iota^{-1}(M_{4\epsilon}^{(1)})$,

$$f(x) = \epsilon_2 \varphi(x) + \psi(x) j(\iota(x)) < \epsilon + j(\iota(x)) < \epsilon + \mathrm{inj}(\iota(x)) + \epsilon < 4\epsilon.$$

This implies that $f^{-1}([4\epsilon, 5\epsilon]) \cap U_0$ is contained in U_1. Next, we notice that ∂U_0 consists of σ and other boundary components on which $\mathrm{inj} = 6\epsilon$. For any $x \in \partial U_0 \setminus \sigma$,

$$f(x) = \epsilon_2 \varphi(x) + \psi(x) j(\iota(x)) > j(\iota(x)) > \mathrm{inj}(\iota(x)) - \epsilon > 5\epsilon.$$

So, for any $t \in [0, 5\epsilon]$, $f^{-1}(t)$, restricted on U_0, does not intersect with ∂U_0. □

As a consequence, for any regular values $t \in [0, 5\epsilon]$, $f^{-1}(t)$ is a union of smooth loops that cobounds with $f^{-1}(0) = \sigma$, and in particular is homologous to σ. Under the image of ι, it shows that $\iota(f^{-1}(t) \cap U_0)$ is homologous to $\iota(\sigma) = c_1[\alpha_1] \neq 0$. Moreover, for regular values $t \in (4\epsilon, 5\epsilon)$ and any point $y \in \iota(f^{-1}(t) \cap U_0)$, we let $x \in f^{-1}(t) \cap U_0 \subset U_1$ be any preimage of y. Then

$$\mathrm{inj}(y) = \mathrm{inj}(\iota(x)) \geq j(\iota(x)) - \epsilon = f(x) - \epsilon_2 \varphi(x) - \epsilon \quad (\psi(x) = 1 \text{ since } x \in U_1)$$

$$\geq t - 2\epsilon > 2\epsilon.$$

In particular, $\ell\big(\iota(f^{-1}(t) \cap U_0)\big) \geq 2\,\mathrm{inj}(y) \geq 4\epsilon$. Since ι is 1–Lipschitz, we obtain $\ell_g(f^{-1}(t) \cap U_0) \geq 4\epsilon$ for any regular values $t \in (4\epsilon, 5\epsilon)$. Finally, we apply the coarea formula together with (3-7) and (3-8), and obtain

$$\begin{aligned}
A(\iota) &\geq A(\iota|_{f^{-1}([4\epsilon,5\epsilon]) \cap U_0}) > \mathrm{vol}_g(f^{-1}([4\epsilon, 5\epsilon]) \cap U_0) - \epsilon' \\
&> \frac{1}{1+2\epsilon'} \int_{f^{-1}([4\epsilon,5\epsilon]) \cap U_0} \|df\| \, dV_g - \epsilon' \\
&= \frac{1}{1+2\epsilon'} \int_{4\epsilon}^{5\epsilon} \ell_g(f^{-1}(t) \cap U_0) \, dt - \epsilon' \geq \frac{1}{1+2\epsilon'} 4\epsilon^2 - \epsilon'.
\end{aligned}$$

Since $\epsilon' > 0$ is arbitrary, the lemma follows. □

Now we are ready to prove (1) and (2) of Theorem 1.11.

Theorem 3.7 *Let $\Gamma < \mathrm{Isom}(X)$ be a finitely generated torsion-free discrete isometry subgroup of a negatively pinched (normalized to $K \leq -1$) Hadamard manifold X. Let $N(\Gamma)$ be the number of cusps in M, and $\beta_1(\Gamma)$ be the first Betti number of M. If $\delta < 1$, then:*

(1) $N(\Gamma) \leq \beta_1(\Gamma)$.

(2)   *For an integer $k > \beta_1(\Gamma) - N(\Gamma)$ and any family of closed geodesics $\{\alpha_1, \ldots, \alpha_k\}$ that are mutually $2C(\delta) + 1$ apart, there exists at least one closed geodesic whose normal injectivity radius is $\leq C(\delta)$, where $C(\delta) = 4/(1 - \delta)$.*

(3)   *M has bounded geometry.*

Proof   For (1), suppose to the contrary $N(\Gamma) > \beta_1(\Gamma)$, where $N(\Gamma)$ could be infinite. Choose ϵ small enough such that the cuspidal part $\mathrm{cusp}_{12\epsilon}(M)$ consists of $N(\Gamma)$ disjoint components $\bigcup_{i=1}^{N} M_{12\epsilon}^{(i)}$. For each component $M_{12\epsilon}^{(i)}$, the corresponding parabolic subgroup P_i is infinite cyclic by [Proposition 2.3](#), so we can choose $\gamma_i \in P_i < \Gamma$ which represents a nontrivial torsion-free homology class in X/P_i (not necessarily in M). Since $N(\Gamma) > \beta_1(\Gamma)$, we have that $\{[\gamma_1], \ldots, [\gamma_{N(\Gamma)}]\}$ is linearly dependent in $H_1(M)$. We can choose an irreducible subfamily containing $[\gamma_1]$ and without loss of generality we assume this to be $\{\gamma_1, \ldots, \gamma_k\}$, where $k \leq \beta_1(\Gamma) + 1 < \infty$. Let $c_1, \ldots, c_k$ be the associated integers such that $\sum_{i=1}^{k} c_i[\gamma_i] = 0$ (with $c_1 \neq 0$). On each component $M_{12\epsilon}^{(i)}$ choose a thinner part $M_{4\epsilon}^{(i)} \subset M_{12\epsilon}^{(i)}$ and let $T_i = M_{12\epsilon}^{(i)} \setminus M_{4\epsilon}^{(i)}$. In particular, the T_i are disjoint and, for any $x \in T_i$, we have $2\epsilon \leq \mathrm{inj}(x) \leq 6\epsilon$. We choose a loop $\alpha_i \subset M_{2\epsilon}^{(i)}$ representing $[\gamma_i]$ such that $\ell(\alpha_i)$ is small enough that $\sum_{i=1}^{k} |c_i| \ell(\alpha_i) < \epsilon^2 / C(\delta)$; see [12, Proposition 1.1.11]. By [Theorem 3.2](#), $\bigcup_{i=1}^{k} c_i \alpha_i$ bounds a smooth surface $\iota : \Sigma \to M$ whose area satisfies

$$(3\text{-}9) \qquad A(\iota) \leq C(\delta) \left(\sum_{i=1}^{k} |c_i| \ell(\alpha_i) \right) < \epsilon^2.$$

However, by [Lemma 3.6](#), $A(\iota) \geq 4\epsilon^2$, which contradicts to [(3-9)](#). Hence, $N(\Gamma) \leq \beta_1(\Gamma)$.

For (2), suppose there are $k = \beta_1(\Gamma) - N(\Gamma) + 1$ mutually $2C(\delta) + 1$ apart simple closed geodesics $\alpha_1, \ldots, \alpha_k$ whose normal injectivity radii are greater than $C(\delta)$. To illustrate the idea, we first assume M has no cusps. Then $[\alpha_1], \ldots, [\alpha_k]$ are linearly dependent on $H_1(M)$. By [Theorem 3.2](#), there exist integers $c_1, \ldots, c_k$ such that $\bigcup_{i=1}^{k} c_i \alpha_i$ bounds a smooth surface $f : \Sigma \to M$ whose area satisfies

$$(3\text{-}10) \qquad A(f) \leq C(\delta) \left(\sum_{i=1}^{k} |c_i| \ell(\alpha_i) \right).$$

Let $R_i = \mathrm{NJ}(\alpha_i)$ and, by the assumption $R_i > C(\delta)$, we can pick $\epsilon > 0$ small enough that $\epsilon < \frac{1}{2}$ and $C(\delta) + \epsilon < R_i$ for all i. Denote by T_i the $(C(\delta) + \epsilon)$–tubular neighborhood of α_i, and, since $\{\alpha_i\}$ are mutually $2C(\delta) + 1$ apart, $\{T_i\}$ are disjoint, and so are

$\{f^{-1}(T_i)\}$. Therefore, by Lemma 3.5,

$$(3\text{-}11) \qquad A(f) \geq \sum_{i=1}^{k} A(f|_{f^{-1}(T_i)}) \geq (C(\delta) + \epsilon)\left(\sum_{i=1}^{k} |c_i|\ell(\alpha_i)\right).$$

This contradicts (3-10).

For the general case, pick nontrivial torsion-free homology classes $\{[\gamma_1], \ldots, [\gamma_{N(\Gamma)}]\}$ on each cusp component as in (1). This together with $[\alpha_1], \ldots, [\alpha_k]$ forms a linearly dependent system on $H_1(M)$. Choose an irreducible system containing $[\alpha_1]$, and without loss of generality assume it to be $\{[\gamma_1], \ldots, [\gamma_{N(\Gamma)}], [\alpha_1], \ldots, [\alpha_k]\}$. Thus there are integers $b_1, \ldots, b_{N(\Gamma)}$ and $c_1, \ldots, c_k$ such that $\sum_{i=1}^{N(\Gamma)} b_i[\gamma_i] + \sum_{j=1}^{k} c_j[\alpha_j] = 0$. Now choose a loop η_i on each cusp component representing γ_i such that $\ell(\eta_i)$ is small enough that $\sum_{i=1}^{N(\Gamma)} |b_i|\ell(\eta_i) < \epsilon\left(\sum_{j=1}^{k} |c_j|\ell(\alpha_j)\right)/C(\delta)$, where ϵ is the same constant as above in the noncusp case. By Theorem 3.2, $\left(\bigcup_{i=1}^{N(\Gamma)} b_i \eta_i\right) \cup \left(\bigcup_{j=1}^{k} c_j \alpha_j\right)$ bounds a smooth surface $f: \Sigma \to M$ whose area satisfies

$$A(f) \leq C(\delta)\left(\sum_{i=1}^{N(\Gamma)} |b_i|\ell(\eta_i) + \sum_{j=1}^{k} |c_j|\ell(\alpha_j)\right).$$

Thus we have

$$A(f) < C(\delta)\left(1 + \frac{\epsilon}{C(\delta)}\right)\left(\sum_{j=1}^{k} |c_j|\ell(\alpha_j)\right) = (C(\delta) + \epsilon)\left(\sum_{j=1}^{k} |c_j|\ell(\alpha_j)\right).$$

However, the area lower bound estimate in (3-11) still holds, which is a contradiction.

For (3), suppose M has unbounded geometry, that is, there exists a sequence of closed geodesics $\{\alpha_i\}$ with $\ell(\alpha_i) \to 0$. When $\ell(\alpha_i)$ is smaller than the Margulis constant, α_i determines a Margulis tube such that the length of every maximal radial arc tends to ∞ as $\ell(\alpha_i) \to 0$; see for example [16, Lemma 2.4]. In particular, the normal injectivity radius $\mathrm{NJ}(\alpha_i)$ goes to ∞. By passing to a subsequence, we can assume that the geodesics α_i are arbitrarily far apart and their normal injectivity radii are all greater than $C(\delta)$, which contradicts (2). $\qquad\square$

Remark 3.8 The assumption $\delta < 1$ is crucial in Theorem 3.7 (which also traces back to Theorem 3.2). Indeed, the main strategy of the proof is to apply an area-decreasing map on the (approximated) area-minimizing surfaces, which are bounded either by tiny loops in different cusps or by far apart closed geodesics. The existence of such a map follows from a construction of Besson, Courtois and Gallot (Theorem 2.9), where $\delta < 1$ has been used to obtain that the area is decreasing.
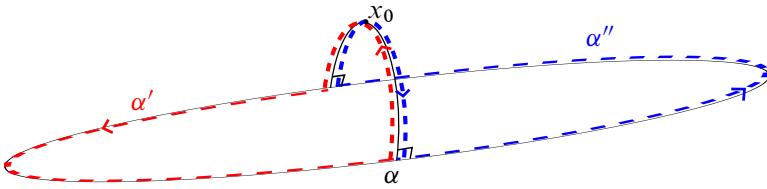
Figure 3

In general, there are examples [29] of finitely generated Kleinian groups $\Gamma < \mathrm{Isom}(\mathbb{H}^4)$ with infinitely many (rank-one) cusps, and by construction it is clear that $\delta \in [2, 3]$. Thus, for every $n \geq 4$, one can construct, via the totally geodesic embedding $\mathbb{H}^4 \to \mathbb{H}^n$, a Kleinian group $\Gamma < \mathrm{Isom}(\mathbb{H}^n)$ of the same critical exponent which contains infinitely many cusps. Italiano, Martelli and Migliorini [28] constructed new examples of finitely generated Kleinian groups $\Gamma \lhd G < \mathrm{Isom}(\mathbb{H}^n)$ for $5 \leq n \leq 8$ with infinitely many cusps, where G is a lattice and $G/\Gamma \cong \mathbb{Z}$. Hence it follows that $\delta(\Gamma) = \delta(G) = n - 1$. We believe that finitely generated Kleinian groups must have finitely many cusps if $\delta < 2$.

We end this section with a corollary which turns out to be essential to our proofs of the main theorems. It is a direct consequence of Theorem 3.7(2). Roughly speaking, if $\delta < 1$ then closed geodesics asymptotically have uniformly bounded tubular neighborhoods.

Corollary 3.9 *Suppose $\delta < 1$ and M has a sequence of escaping closed geodesics. Then there exists a subsequence of escaping closed geodesics whose normal injectivity radii are $\leq C(\delta)$.*

3.3 Decomposing a closed geodesic

Suppose α is a closed geodesic in M with $\mathrm{NJ}(\alpha) \leq C(\delta)$. By definition, there exists $x_0 \in M$ achieving the normal injectivity radius such that it projects to α in two different geodesic minimizing paths. The two geodesic paths have an angle of π. Thus we can decompose α into two piecewise geodesic loops α' and α'' as shown in Figure 3. It is clear that their lengths satisfy $\ell(\alpha') + \ell(\alpha'') \leq \ell(\alpha) + 4C(\delta)$.

Equivalently, in the universal cover (as shown in Figure 4), there exists an isometry $g \in \Gamma$ and $\tilde{x}_0 \in X$ such that

$$d(\tilde{x}_0, A_\gamma) \leq C(\delta), \quad d(\tilde{x}_0, g^{-1}(A_\gamma)) \leq C(\delta),$$

where A_γ is a lift of α in X. Let $\tilde{x}$ and $\tilde{y}$ be the projections of $\tilde{x}_0$ onto $g^{-1}(A_\gamma)$ and A_γ, respectively, which will realize the shortest distance between $g^{-1}(A_\gamma)$ and A_γ

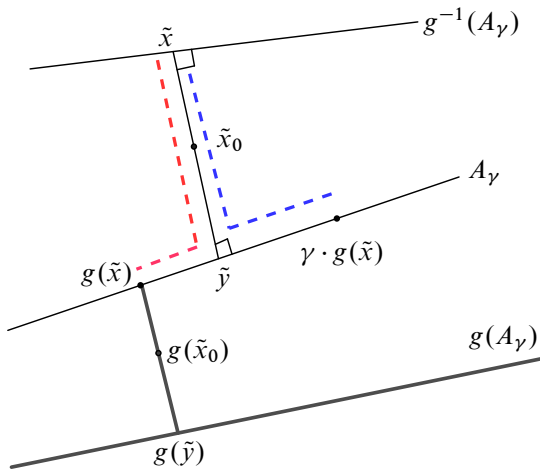Figure 4

(so $\ell(\tilde{x}\tilde{y}) \leq 2C(\delta)$). Under the projection map $\pi \colon X \to M$, the consecutive geodesic segments connecting $g(\tilde{x})$, $\tilde{y}$ and $\tilde{x}$ maps to α' and the one connecting $\tilde{x}$, $\tilde{y}$ and $\gamma \cdot g(\tilde{x})$ maps to α'', where γ translates along A_γ and corresponds to α. From Figure 3, we see that α' represents the isometry g and α'' represents the isometry $\gamma \cdot g$; these are nontrivial elements in Γ. We claim that the group $\langle g, \gamma \cdot g \rangle$ is nonelementary. Otherwise, $\langle g, \gamma \cdot g \rangle$ is parabolic or loxodromic. If $\langle g, \gamma \cdot g \rangle$ is parabolic, then both g and $\gamma \cdot g$ are parabolic and they have the same fixed point, which implies that γ has the same fixed point as the one of the parabolic isometry g, which contradicts the assumption that Γ is discrete by [11, Lemma 3.1.2]. (The proof of Lemma 3.1.2 can be applied to the case of negatively pinched Hadamard manifolds directly.) If $\langle g, \gamma \cdot g \rangle$ is loxodromic, then g and $\gamma \cdot g$ are both loxodromic and they preserve an axis setwise, which means that γ will preserve the same axis as g. However, note that γ preserves the axis A_γ, which is not preserved by g.

It is possible that x_0 projects to the same point on α, in which case α' is the entire transverse geodesic loop, and α'' is the concatenation $\alpha'^{-1} * \alpha$. It is also possible that α may have a transverse self-intersection, in which case the above decomposition coincides with the obvious separation at the self-intersection. Note that nontransverse self-intersection of a closed geodesic α can only occur when α is a multiple of some primitive closed geodesic $\bar{\alpha}$, in which case the above decomposition on α can essentially be treated on $\bar{\alpha}$. We remark that in all the abovementioned "exceptional" cases, the decomposition as described always exists.
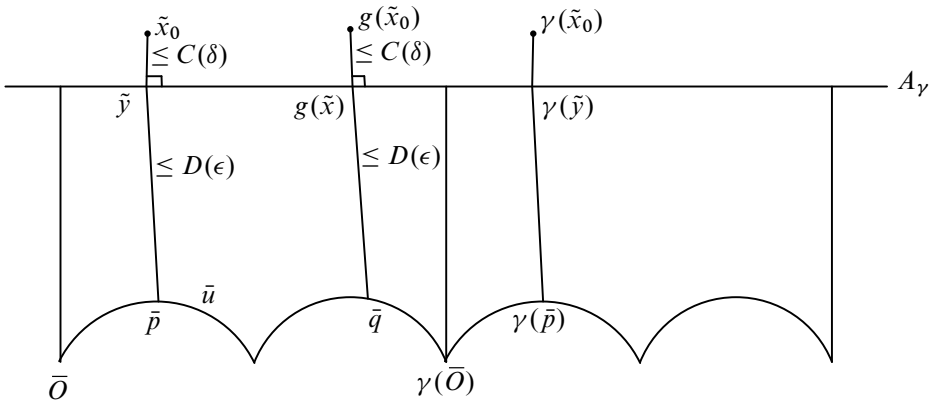
Figure 5

We can extend the above decomposition to a piecewise geodesic loop:

Lemma 3.10 Let $u \subset M$ be a piecewise geodesic loop consisting of at most two geodesics, and let $\alpha \subset M$ be the closed geodesic free homotopic to u with $\mathrm{NJ}(\alpha) \leq C(\delta)$ and $\ell(\alpha) \geq \epsilon$. Then there exist points $p, q \in u$ (which could be the same) and a geodesic segment ω connecting p and q whose length is bounded above by $C_0 = 2C(\delta) + 2D(\epsilon)$. Here $D(\epsilon)$ is the constant in Proposition 2.1. Moreover, the two piecewise geodesic loops under the decomposition shown in Figure 3 are homotopically nontrivial.

Proof Write u as the union of two geodesic segments in M which start and end at O. Let $\bar{u}$ be a lift of u in X consisting of two geodesic segments from the lift $\overline{O}$ to $\gamma(\overline{O})$ as in Figure 5, where $\gamma \in \Gamma$ is represented by u. We denote the axis of γ by A_γ, which is a lift of α. Since $\mathrm{NJ}(\alpha) \leq C(\delta)$, by the discussion above there exists a point $\tilde{x}_0 \in X$ and a nontrivial element $g \in \Gamma$ with $g \neq \gamma$ such that $\tilde{x}_0$ and $g(\tilde{x}_0)$ project onto A_γ at two points $\tilde{y}$ and $g(\tilde{x})$ (which could be the same point) satisfying $d(\tilde{x}_0, \tilde{y}) \leq C(\delta)$ and $d(g(\tilde{x}_0), g(\tilde{x})) \leq C(\delta)$; see Figure 4.

By Proposition 2.1 there exist $\bar{p}, \bar{q} \in \bar{u}$ such that $d(\tilde{y}, \bar{p}) \leq D(\epsilon)$ and $d(g(\tilde{x}), \bar{q}) \leq D(\epsilon)$. Thus, the piecewise geodesic consecutively connecting $\bar{p}$, $\tilde{y}$ and $\tilde{x}_0$ together with the one connecting $g(\tilde{x}_0)$, $g(\tilde{x})$ and $\bar{q}$ projects to a piecewise geodesic path connecting $\pi(\bar{p}) = p$ and $\pi(\bar{q}) = q \in M$ with total length $\leq 2C(\delta) + 2D(\epsilon)$. Finally, there is a unique geodesic segment ω connecting p and q which is homotopic to this piecewise geodesic path and it is clear that $\ell(\omega) \leq 2C(\delta) + 2D(\epsilon)$.

The geodesic segment ω divides the piecewise geodesic loop u into two parts, u_1 and u_2. The concatenation of u_i with the geodesic segment ω gives two piecewise
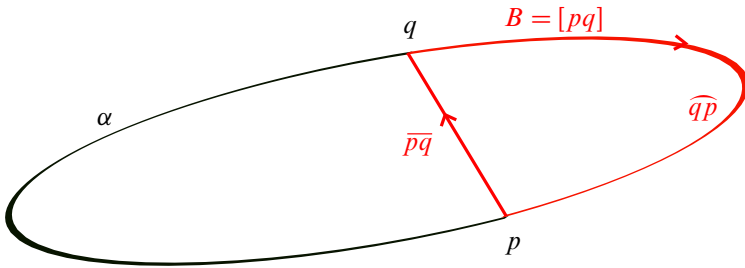
Figure 6

geodesic loops under this decomposition, where $i = 1, 2$. If the two piecewise geodesic loops are homotopically trivial, then $\tilde{x}_0 = g(\tilde{x}_0) = \gamma(\tilde{x}_0)$. By our construction, $g \neq \gamma$ and $g \neq 1$. Hence, they are homotopically nontrivial. $\qquad \square$

3.4 Injectivity radius and convex cocompactness

In this section, we prove (3) of Theorem 1.11. We start by introducing the definition of a *bow* which will be used later in the proof.

Definition 3.11 Given a closed geodesic α, we say $B = \overline{pq} * \widehat{qp}$ is a *bow* on α if:

(1) B consists of two edges $\overline{pq}$ and $\widehat{qp}$, where p and q are two distinct points on α.

(2) $\overline{pq}$ is a minimizing geodesic connecting p to q on M, which might not lie on α.

(3) $\widehat{qp}$ is a geodesic segment on α connecting q to p, which might not be length minimizing; see Figure 6.

We say a bow $B = \overline{pq} * \widehat{qp}$ is C–*thin* if $d(p, q) \leq C$, and we say B is *nontrivial* if the loop $\overline{pq} * \widehat{qp}$ of B is homotopically nontrivial in M. The *length* of a bow $B = \overline{pq} * \widehat{qp}$ is the length of the loop $\overline{pq} * \widehat{qp}$.

Lemma 3.12 *Suppose that* $\delta < 1$ *and the injectivity radius on* M *is bounded by some constant* $\frac{1}{2}\epsilon_0 > 0$ *from below. Then there are no closed geodesics* α *in* M *satisfying*:

(1) α *has normal injectivity radius at most* $C(\delta)$.

(2) *All points of* α *have injectivity radii greater than* $4C_0 + 1$, *where* C_0 *is the constant in Lemma 3.10*.

Proof Suppose that there exists such a closed geodesic α in M. We consider the set $\mathcal{B} = \mathcal{B}(\alpha, 2C_0)$ that consists of all nontrivial $2C_0$–thin bows on α. The set is never
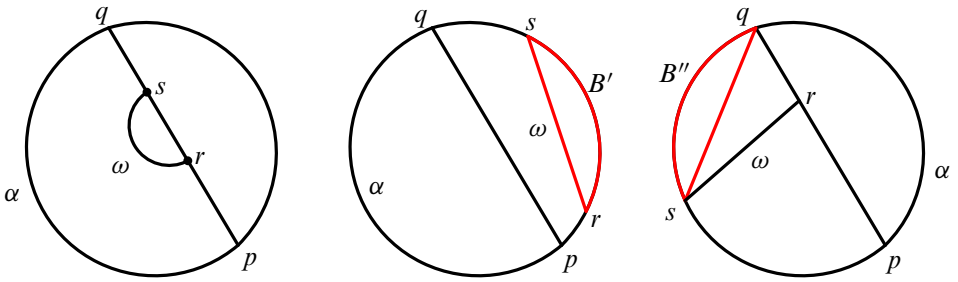
Figure 7

empty. Indeed, choose $p, q \in \alpha$ sufficiently close and choose $\widehat{qp}$ the longer segment on α connecting q to p such that $\ell(\overline{pq}) < \ell(\widehat{qp})$ and $\ell(\overline{pq}) \leq 2C_0$. This gives a nontrivial $2C_0$–thin bow on α. Let $t = \inf\{\ell(B) : B \in \mathcal{B}\}$. We choose $B = \overline{pq} * \widehat{qp} \in \mathcal{B}$ to be a bow with length $\leq t + 1$. Since B is a 2–piecewise geodesic path, by Lemma 3.10 there exist $r, s \in B$ and a geodesic segment $\omega \subset M$ connecting r and s such that

$$(3\text{-}12) \qquad\qquad \ell(\overline{rs}) = \ell(\omega) \leq C_0$$

and that ω splits B nontrivially. Although Lemma 3.10 by itself does not assure that ω is length minimizing, and r and s might even be the same point, we claim this is not the case. Indeed, since $\ell(\overline{pq}) \leq 2C_0$, r must be contained in the C_0–neighborhood of α. By the assumption on the injectivity radius, all the points on α have injectivity radius $> 4C_0 + 1$. Since the injectivity radius function is 1–Lipschitz, $\mathrm{inj}(r) > 3C_0 + 1$. This implies that any geodesic segment emanating from r whose length is at most $3C_0 + 1$ must be uniquely length minimizing. In particular, ω is uniquely length minimizing and $r \neq s$.

Based on the positions of r and s, we discuss three cases:

(1) r and s are both on $\overline{pq}$.

(2) r and s are both on $\widehat{qp}$.

(3) $r \in \overline{pq}$ and $s \in \widehat{qp}$.

Observe that (1) is impossible since both ω and $\overline{pq}$ are uniquely length minimizing, so ω has to be entirely contained in $\overline{pq}$, which contradicts the fact that ω splits B nontrivially. Case (2) is also impossible. To see this, we assume without loss of generality that q, s, r and p are in cyclic order in $\widehat{qp}$, as in Figure 7, and r and s cut $\widehat{qp}$ into three geodesic segments, denoted by $\widehat{qs}, \widehat{sr}$ and $\widehat{rp}$. By assumption, the bow

$B' = \overline{rs} * \widehat{sr}$ is a nontrivial C_0–thin (of course also $2C_0$–thin) bow on α. So by the choice of B we have $\ell(B') + 1 \geq t + 1 \geq \ell(B)$, hence

$$(3\text{-}13) \qquad\qquad \ell(\overline{rs}) + 1 \geq \ell(\widehat{rp}) + \ell(\overline{pq}) + \ell(\widehat{qs}).$$

Since ω splits B nontrivially, we have obtained a homotopically nontrivial piecewise geodesic loop $\eta = \overline{rs} * \widehat{sq} * \overline{qp} * \widehat{pr}$ whose total length can be estimated as

$$\ell(\eta) = \ell(\overline{rs}) + \ell(\widehat{sq}) + \ell(\overline{qp}) + \ell(\widehat{pr}) \leq 2\ell(\overline{rs}) + 1 \quad \text{(by (3-13))}$$
$$\leq 2C_0 + 1 \qquad\qquad\qquad\qquad\qquad\qquad \text{(by (3-12))}.$$

This contradicts the assumption on injectivity radius.

For case (3), note that $\ell(\overline{pq}) \leq 2C_0$, so r is C_0 close to either p or q, and without loss of generality we assume it is closer to q. Therefore by the triangle inequality, $d(q, s) \leq \ell(\overline{rq}) + \ell(\omega) \leq 2C_0$. Now we consider the bow $B'' = \overline{sq} * \widehat{qs}$, where $\widehat{qs}$ is the geodesic segment on α. The bow is nontrivial. Otherwise, $\overline{sq}$ coincides with $\widehat{qs}$, which indicates that $\ell(\widehat{qs}) \leq 2C_0$. Then we have a piecewise geodesic loop $\overline{sr} * \overline{rq} * \widehat{qs}$ with length $\leq 4C_0$. By the injectivity radius assumption it must represent a trivial element, which contradicts the fact that ω cuts B_i nontrivially. Hence, $B'' \in \mathcal{B}$. By the choice of B, we have $\ell(B'') + 1 \geq t + 1 \geq \ell(B)$, hence $\ell(\overline{sq}) + 1 \geq \ell(\widehat{sp}) + \ell(\overline{pq})$. So we have obtained a piecewise geodesic loop $\eta' = \overline{qs} * \widehat{sp} * \overline{pq}$ whose total length satisfies

$$\ell(\eta') = \ell(\overline{qs}) + \ell(\widehat{sp}) + \ell(\overline{pq}) \leq 2\ell(\overline{qs}) + 1 \leq 4C_0 + 1.$$

So η' must be homotopically trivial according to the injectivity radius assumption. Since ω splits B_i nontrivially, the piecewise geodesic loop $\overline{rs} * \widehat{sp} * \widehat{pr}$ is homotopically nontrivial, and therefore, differing by an η', the geodesic triangle $\eta'' = \overline{rs} * \overline{sq} * \overline{qr}$ is also homotopically nontrivial. On the other hand

$$\ell(\eta'') = \ell(\overline{rs}) + \ell(\overline{sq}) + \ell(\overline{qr}) \leq 4C_0,$$

which contradicts the injectivity radius assumption. $\qquad\square$

The following is a restatement of Theorem 1.11(3), which gives an alternative geometric characterization of convex compactness under the assumption that $\delta < 1$.

Theorem 3.13 *If $\delta < 1$, then Γ is convex cocompact if and only if the injectivity radius function* $\mathrm{inj}\colon M \to \mathbb{R}$ *is proper.*

Proof We start with the "only if" part, which does not need the condition $\delta < 1$. Since Γ is convex cocompact, it consists of only loxodromic isometries. Note that all the

closed geodesics are in the compact convex core since their lifts in X are in $\mathrm{Hull}(\Lambda(\Gamma))$. Therefore, the length of all closed geodesics in M is uniformly bounded from below. Otherwise, there is an escaping sequence of closed geodesics (whose length tends to 0) inside the convex core, contradicting compactness. Suppose the injectivity radius function is not proper. Then there exists an escaping sequence of points $x_i \in M$ whose injectivity radii are uniformly bounded by some constant R. At each point x_i, we choose a geodesic loop w_i whose length satisfies $\ell(w_i) = 2\,\mathrm{inj}(x_i) \leq 2R$. By Proposition 2.1, the closed geodesic free homotopic to w_i is within a D–neighborhood of w_i for some constant D. Hence we get an escaping sequence of closed geodesics in the convex core of M, which contradicts compactness.

To show the "if" part, we first note that properness of the injectivity radius function automatically implies that M has no cusps, and there is a uniform lower bound ϵ_0 on the length of closed geodesics in M. Suppose that Γ is not convex cocompact, ie geometrically infinite. By Theorem 2.4 there is an escaping sequence of closed geodesics $\{\alpha_i\} \subset M$. By Corollary 3.9, there is a subsequence of closed geodesics whose normal injectivity radii are all at most $C(\delta)$. For convenience, we still denote it by $\{\alpha_i\}$. Now we fix a constant $C_0 = 2C(\delta) + 2D(\epsilon_0)$ as in Lemma 3.10. Since the injectivity radius function is proper and the sequence $\{\alpha_i\}$ is escaping, all points on α_i have injectivity radii greater than $4C_0 + 1$ when i is sufficiently large. Hence, there exists a closed geodesic in M whose normal injectivity radius is at most $C(\delta)$, and where all points on the geodesic have injectivity radii greater than $4C_0+1$, contradicting Lemma 3.12. Therefore, Γ is convex cocompact. $\qquad\square$

4   Proofs of the main theorems

Theorem 4.1 *For each n and κ there exists a positive constant $D(n,\kappa) < \frac{1}{2}$ such that, for any finitely generated torsion-free discrete isometry subgroup $\Gamma < \mathrm{Isom}\,X$, if either*

(1) $\delta < D(n,\kappa)$, or

(2) Γ is free and $\delta < \frac{1}{16}$,

then the injectivity radius function on M is proper.

Proof   Since $D(n,\kappa) < \frac{1}{2}$, there are no parabolic isometries in Γ by Proposition 2.3. Suppose that the injectivity radius function is not proper. By the same argument as in the first paragraph of the proof of Theorem 3.13, there exists an escaping sequence of closed geodesics $\{\alpha_i\}$ of uniformly bounded length in M. Let $\mathcal{G}^\infty$ be the set of all escaping

sequences of closed geodesics in M, and let $t = \inf\{\liminf_{i\to\infty} \ell(\alpha_i) : \{\alpha_i\} \in \mathcal{G}^\infty\}$. From the previous discussion, we see that $t < \infty$. On the other hand, M has bounded geometry according to Theorem 3.7, so $t > 0$.

We claim that $t \leq 4C(\delta)$. Suppose $t > 4C(\delta)$. Then there exists an escaping sequence of closed geodesics α_i with $\liminf_{i\to\infty} \ell(\alpha_i) = s \in (t, t + \epsilon_0)$, where ϵ_0 is a fixed positive number smaller than $\frac{1}{2}(t - 4C(\delta))$. By Corollary 3.9 there exists a subsequence, which by abuse of notation we still denote by $\{\alpha_i\}$, such that $\lim_{i\to\infty} \ell(\alpha_i) = s$ and $\mathrm{NJ}(\alpha_i) \leq C(\delta)$ for all i. Without loss of generality, we assume $\ell(\alpha_i) \in (t, t + \epsilon_0)$ for all i. By Section 3.3, each α_i can be decomposed into two nontrivial loops α_i' and α_i'' such that $\ell(\alpha_i') + \ell(\alpha_i'') \leq \ell(\alpha_i) + 4C(\delta)$. So the shorter one, which we assume to be α_i', has length $\leq \frac{1}{2}\ell(\alpha_i) + 2C(\delta)$, and it represents a nontrivial isometry in Γ. There is a closed geodesic ν_i free homotopic to α_i' with length $\leq \frac{1}{2}\ell(\alpha_i) + 2C(\delta)$. Since M has bounded geometry, ν_i is inside a uniformly bounded neighborhood of α_i' by Proposition 2.1. Thus we have found another escaping sequence of closed geodesics ν_i which satisfies

$$\ell(\nu_i) \leq \ell(\alpha_i') \leq \tfrac{1}{2}\ell(\alpha_i) + 2C(\delta) \leq \tfrac{1}{2}(t + \epsilon_0) + 2C(\delta) < \tfrac{1}{2}\left(t + \tfrac{1}{2}(t - 4C(\delta))\right) + 2C(\delta)$$
$$= \tfrac{3}{4}t + C(\delta).$$

The last two inequalities follow from the choices of $\{\alpha_i\}$ and ϵ_0. Hence

$$\liminf_{i\to\infty} \ell(\nu_i) \leq \tfrac{3}{4}t + C(\delta) < t.$$

This contradicts the choice of t, therefore $t \leq 4C(\delta)$.

This means that, for any $\epsilon > 0$, there exists a primitive closed geodesic, denoted by α_0, such that $\ell(\alpha_0) \leq t + \epsilon \leq 4C(\delta) + \epsilon$ and $\mathrm{NJ}(\alpha_0) \leq C(\delta)$. By Section 3.3, α_0 can be decomposed to two nontrivial loops α_0' and α_0'', and again we assume α_0' is the shorter one. So $\ell(\alpha_0') < 4C(\delta) + \epsilon$. Let x_0 be a common point of α_0 and α_0'. Note that α_0 and α_0' represent two loxodromic elements $\gamma_0, \gamma_0' \in \pi_1(M, x_0) \cong \Gamma$, which generate a nonelementary subgroup $\langle \gamma_0, \gamma_0' \rangle = \Gamma_0 < \Gamma$.

Recall that, for any group G with finite generating set S, its entropy is defined as

$$h(G, S) = \lim_{N\to\infty} \frac{\ln |\{g \in G : d_S(1, g) \leq N\}|}{N},$$

where d_S is the Cayley graph metric determined by S.

Since Γ is free in (2), Γ_0 must be a free subgroup isomorphic to F_2. So $h(\Gamma_0, S) = \ln 3$ for $S = \{\gamma_0, \gamma_0'\}$. Note that the lengths of geodesic loops from x_0 representing γ_0 and

γ_0' are both bounded by $4C(\delta) + \epsilon$. We conclude that the orbit map $\gamma \mapsto \gamma \cdot x_0$ gives a $(4C(\delta)+\epsilon)$–Lipschitz injection from (Γ_0, d_S) to (X, d). This implies

$$\delta = \delta(\Gamma) \geq \delta(\Gamma_0) \geq \frac{1}{4C(\delta) + \epsilon} h(\Gamma_0, S) = \frac{\ln 3}{4C(\delta) + \epsilon},$$

where the last inequality follows from (2-1). By choosing ϵ small enough and assuming $\delta < \frac{1}{16}$, one can check that the above inequality cannot hold. The contradiction implies that the injectivity radius is proper.

If we are in case (1), then according to [20, Theorem 1.1] there is a free subgroup $\Gamma_0' < \Gamma_0$ generated by two elements g_0 and g_0' whose word lengths measured in (Γ_0, S) are bounded above by some universal constant $C(n, \kappa)$ depending only on the dimension and lower sectional curvature of X. Write $S_0 = \{g_0, g_0'\}$. Therefore, the orbit map $(\Gamma_0', d_{S_0}) \to (X, d)$ through the inclusion $\Gamma_0' \to \Gamma_0$ is a $(4C(\delta)+\epsilon)C(n, \kappa)$–Lipschitz injection. This implies

$$\delta \geq \delta(\Gamma_0) \geq \frac{1}{(4C(\delta) + \epsilon)C(n, \kappa)} h(\Gamma_0', S_0) = \frac{\ln 3}{(4C(\delta) + \epsilon)C(n, \kappa)}.$$

Thus, there exists a constant $D(n, \kappa)$ which is smaller than $\frac{1}{2}$ such that, by choosing ϵ small enough and assuming $\delta < D(n, \kappa)$, the above inequality fails. The contradiction again implies that the injectivity radius is proper. □

Remark 4.2 For case (1), instead of passing to a rank-2 free subgroup, one can also apply the result of [7] to give a uniform lower bound on the entropy of Γ_0.

Now we can finish the proofs of our main results from the introduction.

Proof of Theorems 1.2 and 1.8 Theorem 1.2 follows from Theorems 3.13 and 4.1. For the proof of Theorem 1.8, there exists a finite-index free subgroup $\Gamma' < \Gamma$ such that $\delta(\Gamma') = \delta(\Gamma) < \frac{1}{16}$. Then Γ' is convex cocompact by Theorems 3.13 and 4.1, which implies that Γ is also convex cocompact. □

Proof of Corollary 1.6 Let $D(n)$ be the constant $D(n, \kappa)$ in Theorem 1.2 with $\kappa = 1$. Suppose that $\Gamma < \mathrm{Isom}(\mathbb{H}^n)$ is a finitely generated discrete isometry subgroup with $\delta(\Gamma) < D(n) < \frac{1}{2}$. By the Selberg lemma, there exists a finite-index torsion-free subgroup $\Gamma' < \Gamma$ with $\delta(\Gamma') = \delta(\Gamma) < D(n) < \frac{1}{2}$. By Theorem 1.2, Γ' is convex cocompact. Hence, the Hausdorff dimension of the limit set equals $\delta(\Gamma')$ [9], which is smaller than 1. Note that since the limit set is a second-countable compact metric space (hence also locally compact and Hausdorff) its topological dimension equals the

small inductive dimension, which is bounded above by its Hausdorff dimension, which hence must be zero. This implies that the limit set is totally disconnected (and is in fact a Cantor set). Then we apply a result of Kulkarni [35, Theorem 6.11], which states that if the limit set of a finitely generated Kleinian group is totally disconnected, then the group splits as a free amalgamation of a free group with virtually abelian groups corresponding to the parabolic subgroups. Since the condition $\delta(\Gamma') < 1$ excludes all free abelian factors of higher rank, we conclude Γ' must be free. Therefore, Γ is virtually free. $\square$

References

[1] **I Agol**, *Tameness of hyperbolic* 3*–manifolds*, preprint (2004) arXiv math/0405568

[2] **L V Ahlfors**, *Finitely generated Kleinian groups*, Amer. J. Math. 86 (1964) 413–429 MR Zbl

[3] **W Ballmann**, **M Gromov**, **V Schroeder**, *Manifolds of nonpositive curvature*, Progr. Math. 61, Birkhäuser, Boston (1985) MR Zbl

[4] **H Bass**, *The degree of polynomial growth of finitely generated nilpotent groups*, Proc. London Math. Soc. 25 (1972) 603–614 MR Zbl

[5] **Y Benoist**, **D Hulin**, *Harmonic quasi-isometric maps III: Quotients of Hadamard manifolds*, Geom. Dedicata 217 (2023) art. id. 52 MR Zbl

[6] **G Besson**, **G Courtois**, **S Gallot**, *Rigidity of amalgamated products in negative curvature*, J. Differential Geom. 79 (2008) 335–387 MR Zbl

[7] **G Besson**, **G Courtois**, **S Gallot**, *Uniform growth of groups acting on Cartan–Hadamard spaces*, J. Eur. Math. Soc. 13 (2011) 1343–1371 MR Zbl

[8] **M Bestvina**, *Questions in geometric group theory* (2000) Available at http://www.math.utah.edu/~bestvina/eprints/questions-updated.pdf

[9] **C J Bishop**, **P W Jones**, *Hausdorff dimension and Kleinian groups*, Acta Math. 179 (1997) 1–39 MR Zbl

[10] **F Bonahon**, *Bouts des variétés hyperboliques de dimension* 3, Ann. of Math. 124 (1986) 71–158 MR Zbl

[11] **B H Bowditch**, *Geometrical finiteness for hyperbolic groups*, J. Funct. Anal. 113 (1993) 245–317 MR Zbl

[12] **B H Bowditch**, *Geometrical finiteness with variable negative curvature*, Duke Math. J. 77 (1995) 229–274 MR Zbl

[13] **B H Bowditch**, *The ending lamination theorem*, preprint (2011) Available at http://homepages.warwick.ac.uk/~masgak/papers/elt.pdf

[14] **J F Brock**, **R D Canary**, **Y N Minsky**, *The classification of Kleinian surface groups, II: The ending lamination conjecture*, Ann. of Math. 176 (2012) 1–149 MR Zbl

[15] **Y D Burago**, **V A Zalgaller**, *Geometric inequalities*, Grundl. Math. Wissen. 285, Springer (1988) MR Zbl

[16] **P Buser**, **B Colbois**, **J Dodziuk**, *Tubes and eigenvalues for negatively curved manifolds*, J. Geom. Anal. 3 (1993) 1–26 MR Zbl

[17] **D Calegari**, *scl*, MSJ Memoirs 20, Math. Soc. Japan, Tokyo (2009) MR Zbl

[18] **D Calegari**, **D Gabai**, *Shrinkwrapping and the taming of hyperbolic* 3*–manifolds*, J. Amer. Math. Soc. 19 (2006) 385–446 MR Zbl

[19] **C Connell**, **B Farb**, **D B McReynolds**, *A vanishing theorem for the homology of discrete subgroups of* $\mathrm{Sp}(n, 1)$ *and* F_4^{-20}, J. Lond. Math. Soc. 94 (2016) 357–376 MR Zbl

[20] **S Dey**, **M Kapovich**, **B Liu**, *Ping-pong in Hadamard manifolds*, Münster J. Math. 12 (2019) 453–471 MR Zbl

[21] **R E Greene**, **H Wu**, C^∞ *approximations of convex, subharmonic, and plurisubharmonic functions*, Ann. Sci. École Norm. Sup. 12 (1979) 47–84 MR Zbl

[22] **Y Guivarc'h**, *Croissance polynomiale et périodes des fonctions harmoniques*, Bull. Soc. Math. France 101 (1973) 333–379 MR Zbl

[23] **N A Gusevskii**, *Geometric decomposition of Kleinian groups in space*, Dokl. Akad. Nauk SSSR 301 (1988) 529–532 MR Zbl In Russian; translated in Dokl. Math. 38 (1989) 89–101

[24] **E Heintze**, **H-C Im Hof**, *Geometry of horospheres*, J. Differential Geometry 12 (1977) 481–491 MR Zbl

[25] **Y Hou**, *Kleinian groups of small Hausdorff dimension are classical Schottky groups, I*, Geom. Topol. 14 (2010) 473–519 MR Zbl

[26] **Y Hou**, *All finitely generated Kleinian groups of small Hausdorff dimension are classical Schottky groups*, Math. Z. 294 (2020) 901–950 MR Zbl

[27] **Y Hou**, *The classification of kleinian groups of hausdorff dimension at most one*, Q. J. Math. 74 (2023) 607–625 MR Zbl

[28] **G Italiano**, **B Martelli**, **M Migliorini**, *Hyperbolic manifolds that fiber algebraically up to dimension* 8, J. Inst. Math. Jussieu (online publication November 2022)

[29] **M Kapovich**, *On the absence of Sullivan's cusp finiteness theorem in higher dimensions*, from "Algebra and analysis" (L A Bokut', M Hazewinkel, Y G Reshetnyak, S Ivanov, editors), Amer. Math. Soc. Transl. Ser. 2 163, Amer. Math. Soc., Providence, RI (1995) 77–89 MR Zbl

[30] **M Kapovich**, *Kleinian groups in higher dimensions*, from "Geometry and dynamics of groups and spaces" (M Kapranov, S Kolyada, Y I Manin, P Moree, L Potyagailo, editors), Progr. Math. 265, Birkhäuser, Basel (2008) 487–564 MR Zbl

[31]   **M Kapovich**, *Homological dimension and critical exponent of Kleinian groups*, Geom. Funct. Anal. 18 (2009) 2017–2054   MR Zbl

[32]   **M Kapovich**, **B Liu**, *Geometric finiteness in negatively pinched Hadamard manifolds*, Ann. Acad. Sci. Fenn. Math. 44 (2019) 841–875   MR Zbl

[33]   **M Kapovich**, **L Potyagailo**, *On the absence of Ahlfors' finiteness theorem for Kleinian groups in dimension three*, Topology Appl. 40 (1991) 83–91   MR Zbl

[34]   **M E Kapovich**, **L D Potyagailo**, *On the absence of finiteness theorems of Ahlfors and Sullivan for Kleinian groups in higher dimensions*, Sibirsk. Mat. Zh. 32 (1991) 61–73   MR Zbl   In Russian; translated in Siberian Math. J. 32 (1991), 227–237

[35]   **R S Kulkarni**, *Groups with domains of discontinuity*, Math. Ann. 237 (1978) 253–272   MR Zbl

[36]   **Y Minsky**, *The classification of Kleinian surface groups, I: Models and bounds*, Ann. of Math. 171 (2010) 1–107   MR Zbl

[37]   **C B Morrey, Jr**, *The problem of Plateau on a Riemannian manifold*, Ann. of Math. 49 (1948) 807–851   MR Zbl

[38]   **P J Nicholls**, *The ergodic theory of discrete groups*, London Mathematical Society Lecture Note Series 143, Cambridge Univ. Press (1989)   MR Zbl

[39]   **P Pankka**, **J Souto**, *Free vs. locally free Kleinian groups*, J. Reine Angew. Math. 746 (2019) 149–170   MR Zbl

[40]   **L Potyagailo**, *The problem of finiteness for Kleinian groups in 3–space*, from "Knots 90" (A Kawauchi, editor), de Gruyter, Berlin (1992) 619–623   MR Zbl

[41]   **L Potyagailo**, *Finitely generated Kleinian groups in 3–space and 3–manifolds of infinite homotopy type*, Trans. Amer. Math. Soc. 344 (1994) 57–77   MR Zbl

[42]   **T Soma**, *Geometric approach to ending lamination conjecture*, preprint (2008)   arXiv 0801.4236

[43]   **J R Stallings**, *On torsion-free groups with infinitely many ends*, Ann. of Math. 88 (1968) 312–334   MR Zbl

[44]   **D Sullivan**, *A finiteness theorem for cusps*, Acta Math. 147 (1981) 289–299   MR Zbl

Department of Mathematics, The Ohio State University
Columbus, OH, United States

Institute of Mathematical Sciences, ShanghaiTech University
Shanghai, China

bbliumath@gmail.com,   shiwang.math@gmail.com

Stable cubulations, bicombings, and barycenters

Matthew G Durham

Yair N Minsky

Alessandro Sisto

We prove that the hierarchical hulls of finite sets of points in mapping class groups and Teichmüller spaces are stably approximated by CAT(0) cube complexes, strengthening a result of Behrstock, Hagen and Sisto. As applications, we prove that mapping class groups are semihyperbolic and Teichmüller spaces are coarsely equivariantly bicombable, and both admit stable coarse barycenters. Our results apply to the broader class of "colorable" hierarchically hyperbolic spaces and groups.

1 Introduction

Much of the coarse structure of mapping class groups has the flavor of CAT(0) geometry, in spite of the fact that mapping class groups have no geometric actions on CAT(0) spaces; see Bridson [17]. Manifestations of this include the weakly relatively hyperbolic structure associated to curve complexes — see Masur and Minsky [42] — and the equivariant embedding into finite products of quasitrees found by Bestvina, Bromberg and Fujiwara [13].

A notion of "hulls" of finite sets in mapping class groups was introduced by Behrstock, Kleiner, Minsky and Mosher in [9], and these were more recently shown by Behrstock, Hagen and Sisto [8] to be approximated in a uniform way by finite CAT(0) cube complexes — see also the alternative proof given by Bowditch in [16]. Our goal in this paper is to refine this construction to make it *stable*, in the sense that perturbation of the input data gives rise to bounded change in the cubical structure. As initial applications, we give a construction for equivariant barycenters and a proof that mapping class groups are bicombable.

As in [8], the proof works in a more general context of *hierarchically hyperbolic groups*, a class of groups (and spaces) introduced by Behrstock, Hagen and Sisto [6; 7] which are endowed with a structure similar to the hierarchical family of curve complexes associated to a surface; see Masur and Minsky [43]. See Section 2.2 for the definition of a hierarchically hyperbolic space (HHS).

Our main result, stated informally, is the following:

Theorem A *In a colorable HHS $(\mathcal{X}, \mathfrak{S})$, the coarse hull $H_\theta(F)$ of any finite set F can be approximated in a coarsely equivariant way by a finite CAT(0) cube complex whose dimension is bounded by the complexity of $(\mathcal{X}, \mathfrak{S})$, in such a way that a bounded change in F corresponds to a change of the cubical structure by a bounded number of hyperplane deletions and insertions.*

The colorability assumption—see Definition 2.8—in Theorem A is apparently quite weak and excludes none of the key examples of HHSs, though there are noncolorable HHGs; see Hagen [36].

For the general context of this result, see the discussion in Section 1.2, where we also give a more precise statement in Theorem 1.4. See Theorem 4.1 for the strongest version. Besides mapping class groups, there are several other classes of spaces and groups that are colorably hierarchically hyperbolic, including

- many cubical groups, including all right-angled Artin and Coxeter groups; see [6] and Hagen and Susse [38];
- Teichmüller spaces with either the Teichmüller or the Weil–Petersson metric; see Durham [27], Eskin, Masur and Rafi [32], and Rafi [48];
- fundamental groups of closed 3–manifolds without Nil or Sol summands [7];
- groups resulting from various combination and small-cancellation-type theorems; see Behrstock, Hagen and Sisto [5; 7], Berlai and Robbio [11], and Robbio and Spriano [51; 56];
- quotients of mapping class groups by suitable large powers of Dehn twists, and other related quotients; see Behrstock, Hagen, Martin and Sisto [4];
- extensions of lattice Veech subgroups of mapping class groups; see Dowdall, Durham, Leininger and Sisto [25; 26];
- extensions of multicurve stabilizer subgroups of mapping class groups; see Russell [52];

- the genus-2 handlebody group; see Chesser [23];
- Artin groups of extra large type; see Hagen, Martin and Sisto [37, Remark 6.16].

With the exception of any hyperbolic and cubical examples from above, our main results and its applications are novel for this wide class of objects.

1.1 Applications

We now discuss our two main applications of Theorem A, namely that mapping class groups and Teichmüller spaces are bicombable (Corollary D) and admit stable barycenters (Corollary F).

Bicombings and semihyperbolicity In CAT(0) spaces, geodesics are unique. In geodesic Gromov hyperbolic spaces, all geodesics between any pair of points fellow-travel. In fact, in both of these classes of spaces geodesics are stable under perturbation of their endpoints in the following sense: given points x, x', y and y' with $d(x, y), d(x', y') \leq 1$, all geodesics from x to y fellow-travel those from x' to y'.

The notion of a *bicombing* of a metric space X, introduced by Thurston, generalizes this stability property. Roughly speaking, a bicombing is a transitive family of uniform quasigeodesics with the above parametrized fellow-traveling property under perturbation of endpoints. See Section 6.2 for a precise definition.

Bicombability is a quasi-isometry invariant which imposes strong constraints on groups, such as property FP_∞, a quadratic isoperimetric inequality, and the Novikov conjecture; see Alonso and Bridson [1], Baumslag, Gersten, Shapiro and Short [2], Epstein, Cannon, Holt, Levy, Paterson and Thurston [31], Gersten and Short [33], and Storm [57]. Moreover, bicombings are the key geometric feature of biautomatic structures on groups (where one requires that the bicombing is constructible by a finite state automaton), thereby playing an important role in computational group theory. It is worth noting that bicombability is decidedly a feature of nonpositive curvature, with the 3–dimensional Heisenberg group not being bicombable because it does not satisfy a quadratic isoperimetric inequality [31].

The power of our stable cubical models is that they allow us to stably and hierarchically import geometric features of cube complexes into HHSs. In particular, ℓ^1–geodesics in the cubical models map to *hierarchy paths* (Definition 6.5), which are quasigeodesics that are finely attuned to the HHS structure, in that they project to

uniform, unparametrized quasigeodesics in every hyperbolic space in the hierarchical structure. The stability property of the cubulation then implies that carefully chosen ℓ^1–geodesics give a bicombing:

Theorem B *Any colorable HHS* $(\mathcal{X}, \mathfrak{S})$ *admits a coarsely* $\mathrm{Aut}(\mathcal{X}, \mathfrak{S})$*–equivariant, discrete, bounded, quasigeodesic bicombing by hierarchy paths with uniform constants.*

If the action by automorphisms is free, then coarse equivariance can be upgraded to equivariance. By the definition of semihyperbolicity [1], we obtain:

Corollary C *Colorable hierarchically hyperbolic groups are semihyperbolic.*

Note that semihyperbolicity has several novel consequences for HHGs. Besides novel consequences of bicombability, these include solvability of the conjugacy problem and the fact that abelian subgroups are undistorted [1].

While many HHSs were known to be bicombable for other reasons, eg many are CAT(0), this produces bicombings for many new examples, such as extensions of Veech subgroups of mapping class groups.

Our main application is:

Corollary D *For any finite type surface* Σ*, its mapping class group* $\mathrm{MCG}(\Sigma)$ *is semihyperbolic and its Teichmüller space* $\mathrm{Teich}(\Sigma)$ *with either the Teichmüller metric or the Weil–Petersson metric is coarsely* $\mathrm{MCG}(\Sigma)$*–equivariantly bicombable by hierarchy paths with uniform constants.*

Note that the HHS notion of hierarchy path that we are using here is more general than the hierarchy paths produced in [27; 43], which are explicitly constructed from hierarchies of tight geodesics in curve graphs.

We remark that semihyperbolicity of $\mathrm{MCG}(\Sigma)$ follows from work in a preprint of Hamenstädt [40]. The result for $\mathcal{T}(\Sigma)$ is new, though we were informed by M Kapovich and K Rafi that they know of a different construction for bicombing $\mathcal{T}(\Sigma)$. Note that $\mathcal{T}(\Sigma)$ with the Weil–Petersson metric is bicombable since its completion is CAT(0) — see Bridson and Haefliger [18], Tromba [59], and Wolpert [60] — though we note that it is unknown whether Weil–Petersson geodesics are hierarchy paths. Combability of $\mathrm{MCG}(\Sigma)$ follows from work of Mosher [45].

Notably, our bicombing construction applies to both mapping class groups and Teichmüller spaces simultaneously. Moreover, our bicombings are relatively straightforward applications of our more powerful stable cubulation construction. See Section 1.4 for a discussion.

Stable barycenters Another key feature of nonpositively curved spaces is that bounded sets admit (coarse) *barycenters*. Here, we think of barycenters simply as maps assigning a point to any finite subset. Some more properties are required to make this notion meaningful, such as *stability*, which requires the barycenter to vary a bounded amount when the finite set varies a bounded amount, and *coarse equivariance* when a group action is present; see Section 6.1.

In CAT(0) spaces there are a number of useful notions of barycenter which are equivariant and stable, for example center-of-mass constructions and circumcenters. Coarse barycenters are useful in the context of groups for understanding centralizers and solving the conjugacy problem for torsion elements and subgroups. Notably, Gromov hyperbolic spaces admit (coarse) barycenters: a coarse barycenter of a finite set F in a hyperbolic space X can be taken to be one of the standard CAT(0) barycenters in the CAT(0) space which models the hull of F in X, ie a simplicial tree. See Section 1.2 for a discussion of these ideas in the context of the this paper.

We should mention that coarse barycenters for triples of points are used to define coarse medians in the sense of Bowditch [15]; thus playing a central role in the theory of coarse median spaces and its many applications. However it is unclear how to construct barycenters even for pairs of points in a coarse median space, and stability properties appear just as difficult to obtain.

Barycenters in CAT(0) spaces are not in general well behaved under quasi-isometries. Using Theorem A and a construction reminiscent of Niblo and Reeves' normal paths [46], we are able to prove that most HHSs admit equivariant coarse barycenters, which are coarsely invariant under HHS automorphisms:

Theorem E Let $(\mathcal{X}, \mathfrak{S})$ be a colorable HHS. Then $\mathcal{X}$ admits coarsely $\mathrm{Aut}(\mathcal{X}, \mathfrak{S})$–equivariant stable barycenters for k points, for any $k \geq 1$.

We remark that the coarse barycenter we produce for a set F is contained in the hull of F.

As with Theorem B, Theorem E can be applied to mapping class groups and Teichmüller spaces:

Corollary F *For any finite type surface* Σ, *its mapping class group* $\mathrm{MCG}(\Sigma)$ *and Teichmüller space* $\mathcal{T}(\Sigma)$ *admit coarsely* $\mathrm{MCG}(\Sigma)$–*equivariant stable barycenters for* k *points, for any* $k \geq 1$.

Corollary F is new for arbitrary finite sets of points in $\mathrm{MCG}(\Sigma)$ and $\mathcal{T}(\Sigma)$ with the Teichmüller metric, even without the stability property. The corresponding statement for $\mathcal{T}(\Sigma)$ with the Weil–Petersson metric is an easy consequence of the fact that its completion is CAT(0). Corollary F, without the stability property, was proven for triples of points in $\mathrm{MCG}(\Sigma)$ by Behrstock and Minsky [10], for orbits of finite order elements of $\mathrm{MCG}(\Sigma)$ in $\mathrm{MCG}(\Sigma)$ by Tao [58], and more generally for orbits of finite subgroups of $\mathrm{MCG}(\Sigma)$ in both $\mathrm{MCG}(\Sigma)$ and $\mathcal{T}(\Sigma)$ with the Teichmüller metric by Durham [28].

In work that appeared simultaneously to ours, Haettel, Hoda and Petyt [35] proved that HHSs are coarse Helly spaces, in the sense of Chalopin, Chepoi, Genevois, Hirai and Osajda [20]. This property has a number of strong consequences, many of which overlap with the results in this paper. In particular, they obtain versions of Theorems B and E along with their corollaries, without the colorability assumption and the hierarchy path conclusion.

Their approach and constructions are very different from ours, using results from the theory of coarse Helly and injective metric spaces, whereas our work relies mostly on hyperbolic and cubical geometry.

1.2 Coarse hulls and their models

Given the technical nature of many of the proofs in this paper, we include here an extended but simplified discussion of the ideas that go into our constructions. The propositions stated in this section will not, however, be used elsewhere in the paper.

Consider first the notion of a *convex hull* in a CAT(0) space. The convex hull of a finite set F has the following nice property: the map $F \mapsto \mathrm{hull}(F)$ is 1–Lipschitz with respect to the Hausdorff metric on sets. We are interested in generalizing this notion to more coarse hulls (which we will just denote by $\mathrm{hull}(F)$ in each case) in more general spaces.

As a first motivating example, consider the Euclidean plane, $X = \mathbb{R}^2$ with the ℓ^2 metric. The convex hull of two points, $\mathrm{hull}(\{x, y\})$, is just the unique geodesic between them. If, on the other hand, we endow $\mathbb{R}^2$ with the ℓ^1 metric then the convex hull is the axis-parallel rectangle spanned by x and y. Note that $(\mathbb{R}^2, \ell^1)$ is not CAT(0) but is a

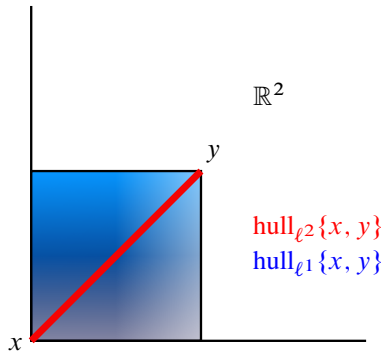Figure 1: A cartoon of the ℓ^2–hull (red) and ℓ^1–hull (blue) of two points in $\mathbb{R}^2$. The ℓ^1–hull reflects the intrinsic product structure of the space.

product of CAT(0) spaces, and this hull is a product of hulls in the CAT(0) factors. See Figure 1. This simple idea is a model for a useful construction in the HHS context.

We can think of an HHS as (coarsely) embedded in a product of hyperbolic spaces, in such a way that it is composed of products of certain factors, intersecting and nesting in a complicated fashion. The reader familiar with the foundational example, namely Masur and Minsky's hierarchy of curve graphs for mapping class groups [42; 43], will lose nothing by keeping it in mind during the ensuing discussion. In that setting, Behrstock, Kleiner, Minsky and Mosher [9] introduced a notion of hull which is essentially a coarse pullback of convex hulls in each hyperbolic factor; see Section 2.2. Behrstock, Hagen and Sisto [8] proved, in the general HHS setting, that these hulls are quasi-isometrically modeled by finite CAT(0) cubical complexes.

Their result is a partial generalization of the situation in Gromov hyperbolic spaces, where Gromov proved that hulls of finite sets of points are quasi-isometrically modeled by finite simplicial trees [34]. However, in the setting of hyperbolic spaces, the modeling trees satisfy additional strong stability properties under perturbation of the set of input points; see Proposition 1.3 below.

Our main theorem — in increasing specificity, Theorems A, 1.4 and 4.1 — endows the modeling cube complexes from [8] with a generalization of the stability properties that Gromov's modeling trees enjoy.

Before giving a full account of our results and an overview of their proofs, it will be beneficial to discuss the situation in hyperbolic spaces and cubical complexes. We will see that our results are a common generalization of the situations from these motivating examples.
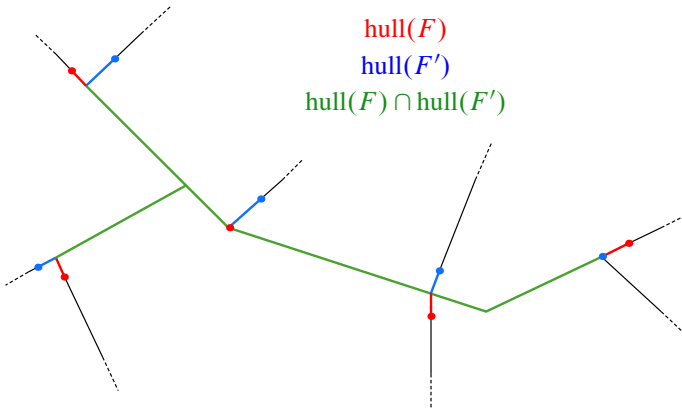
Figure 2: Stability of hulls in a tree: the intersection of $\text{hull}(F)$ (red) and $\text{hull}(F')$ (blue) subtrees is a (green) subtree which can be obtained by deleting the boundedly many complementary red and blue subtrees.

Hulls in trees and cube complexes Let X be a simplicial tree. Then the convex hull of any finite set of vertices $F \subset X^0$ is the subtree T_F of X spanned by F. Moreover, the subtree T_F is stable under small perturbations of F, in the following sense; see Figure 2:

Proposition 1.1 *Let X be a simplicial tree. If $F, F' \subset X^0$ satisfy $\#F' = \#F = k$ and $d_{\text{Haus}}(F, F') \leq 1$, then the intersection of their hulls, $T_0 = T_F \cap T_{F'}$, is itself a subtree with both $T_F \setminus T_0$ and $T_{F'} \setminus T_0$ a union of at most k subtrees each of diameter at most* 1.

We will not use this fact, so we leave its proof to the interested reader.

This situation generalizes to when X is a CAT(0) cube complex endowed with the ℓ^1 metric; see Section 2.1 for the relevant definitions. Recall that the ℓ^1 metric on X is completely determined by a special collection $\mathcal{H}_X$ of codimension-1 subspaces called *hyperplanes* (Section 2.1), in the sense that X is precisely the dual cube complex arising from Sageev's cubulation construction [54] applied to $\mathcal{H}_X$ as a wallspace; see Section 2.1.1.

In the cubical context, the ℓ^1 convex hull of any finite set of vertices $F \subset X^0$ is the cubical subcomplex $\mathcal{Q}_F \subset X$ realized as the dual to the hyperplanes $\mathcal{H}_F$ separating the points in F. In addition, these cubical hulls satisfy the following strong stability property; see Figure 3:
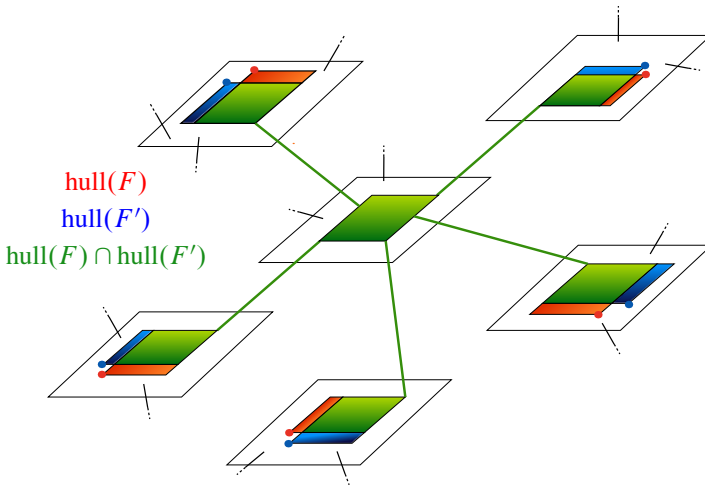
Figure 3: Stability of hulls in the universal cover of $S^1 \vee T^2$: the subcomplex dual to all hyperplanes common to both hull(F) (red) and hull(F') (blue) is here realized as the intersection (green) of the ℓ^1–hulls.

Proposition 1.2 *Let X be a CAT(0) cube complex endowed with the ℓ^1 metric. If $F, F' \subset X^0$ satisfy $\#F, \#F' \leq k$ and $d_{\mathrm{Haus}}(F, F') \leq 1$, then there are convex subcomplexes $X_F \subseteq \mathcal{Q}_F$ and $X_{F'} \subseteq \mathcal{Q}_{F'}$, both dual to the hyperplanes in $\mathcal{H}_0 = \mathcal{H}_F \cap \mathcal{H}_{F'}$, such that $d_{\mathrm{Haus}}(X_F, X_{F'}) \leq 1$. Moreover, both $\mathcal{H}_F \setminus \mathcal{H}_0$ and $\mathcal{H}_{F'} \setminus \mathcal{H}_0$ contain at most k hyperplanes.*

Again, we will not use this proposition, so we omit its proof.

In the cubical structure on a simplicial tree, the hyperplanes correspond to midpoints of edges. Hence Proposition 1.2 generalizes Proposition 1.1. Note that now the diameters of $\mathcal{Q}_F \setminus X_F$ and $\mathcal{Q}_{F'} \setminus X_{F'}$ can be arbitrarily large. However, since the ℓ^1 metric on X is completely determined by its defining hyperplanes, Proposition 1.2 says that $\mathcal{Q}_F$ and $\mathcal{Q}_{F'}$ are metrically and combinatorially related, depending only on k and X — and not on $\mathrm{diam}(F)$. In particular, one can delete boundedly many hyperplanes from the collections $\mathcal{H}_F$ and $\mathcal{H}_{F'}$ to generate a common model; see Section 2.1.2 for a discussion on hyperplane deletions.

Modeling hulls in hyperbolic spaces   In coarse geometry, eg when X is the Cayley graph of a finitely generated group, the notion of geodesic is often wobbly, and so our notion of hull needs to be more flexible. Moreover, it will often be more fruitful to construct quasi-isometric *models* of hulls, which we should think of as nice combi-
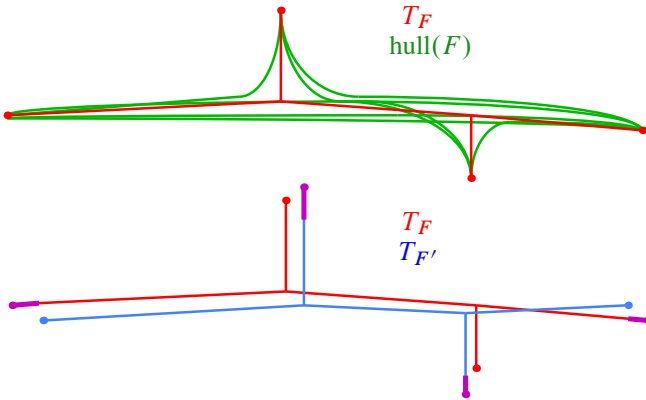
Figure 4: Stability of hulls in a hyperbolic space: top, the modeling tree (red) for the coarse hull (green) of a finite set F; bottom, the modeling trees T_F and $T_{F'}$ for hull(F) and hull(F') are $(1, K)$–quasi-isometric after deleting small subtrees (purple).

natorial objects which coarsely encode the key geometric features of hulls into their combinatorial structure. The main motivating examples here are hyperbolic spaces, where hulls are modeled by finite simplicial trees.

When X is δ–hyperbolic and $F \subset X$ with $\#F = k$, the right notion of hull(F) is the *weak hull*, namely the set of all geodesics between points in F. Notice then that the tripod-like δ–slim-triangles condition generalizes to a tree-like slimness for hull(F). The following proposition is an easy consequence of Gromov's original arguments [34]; see Figure 4:

Proposition 1.3 *For any $k \in \mathbb{N}$ and $\delta > 0$, there exists $L = L(k, \delta) > 0$ such that the following holds.*

Let X be δ–hyperbolic and $F \subset X$ with $\#F = k$. Then there is a simplicial tree T_F and a $(1, L)$–quasi-isometric embedding $\phi_F : T_F \to X$ with $d_{\mathrm{Haus}}(\phi_F(T_F), \mathrm{hull}(F)) < L$.

Moreover, if $F' \subset X$ with $\#F' \leq k$ and $d_{\mathrm{Haus}}(F, F') \leq 1$, then there exists a simplicial tree T_0 and a $(1, L)$–quasi-isometric embedding $\phi_0 : T_0 \to X$ such that the diagram

$$
\begin{array}{ccc}
T_F & & \\
h_F \downarrow & \searrow \phi_F & \\
T_0 & \xrightarrow{\phi_0} & X \\
h_{F'} \uparrow & \nearrow \phi_{F'} & \\
T_{F'} & &
\end{array}
$$

(1-1)

commutes up to error at most L, where h_F and $h_{F'}$ are quotient maps which collapse at most L subtrees each of diameter at most L.

Observe that Proposition 1.3 is a generalization of Proposition 1.1, where X is a tree and we can take the trees T_F, $T_{F'}$, and T_0 as before and the maps ϕ_F, $\phi_{F'}$, and ϕ_0 to be inclusions. The main difference here is that a general hyperbolic space is stably *locally* tree-like, and not a tree itself. Hence the need for a model for the hulls.

Our main theorem is a common generalization of the stability properties in Propositions 1.2 and 1.3.

1.3 Stable cubical models for hulls in HHSs

We will deal with *colorable* hierarchically hyperbolic spaces $(\mathcal{X}, \mathfrak{S})$, which means, for the reader familiar with HHSs, that there exists a decomposition of $\mathfrak{S}$ into finitely many families $\mathfrak{S}_i$ such that each $\mathfrak{S}_i$ is pairwise transverse. Colorable HHSs include mapping class groups and Teichmüller spaces of finite-type surfaces.

In fact, colorability is a rather mild condition which is satisfied by all of the main motivating examples. Its definition is inspired by Bestvina, Bromberg and Fujiwara's proof that curve graphs are finitely colorable [12]; see Section 2.2 for a discussion.

Given a finite set of points $F \subset \mathcal{X}$ in an HHS, the standard notions of a hull for F are very difficult to analyze. For example, while little is known about geodesics in the mapping class group, Rafi and Verberne [50] proved that geodesics do not always interact well with the curve graph machinery. In Teichmüller space with the Teichmüller metric, geodesics are unique, but it is an open question of Masur whether the classical convex hull of a set of three points can be the whole space. Moreover, it is a result of Rafi that hulls of two points, ie geodesics, do not behave stably under perturbation [49, Theorem D]. These complications motivate a more flexible definition of hull in this setting.

The *hierarchical hull* of a finite set $F \subset \mathcal{X}$, which we also denote by hull(F), was introduced in [9] to study subspaces of the asymptotic cones of the mapping class group, on the way to proving that these groups are quasi-isometrically rigid. In hyperbolic spaces and cube complexes, the hierarchical hull coincides with the notions of hull discussed above. In the hierarchical setting, one instead has a notion of projecting F to a family of hyperbolic spaces (eg curve graphs of subsurfaces). In each of these hyperbolic spaces, one then takes the weak hull of the projection — which is coarsely

a tree, as above — and the uses certain hierarchical consistency conditions [7; 9] to fashion these weak hulls in the various spaces into a hull in the ambient HHS which satisfies certain convexity properties [7; 9]. In particular, the hierarchical hull of F is *hierarchically quasiconvex* [7] and contains all of the hierarchy paths between points in F [9].

In [8], Behrstock, Hagen and Sisto proved that the hierarchical hull of a finite set of points is quasi-isometric to a finite CAT(0) cube complex. Their main observation was that the hierarchical consistency conditions are closely related to the consistency conditions on a wallspace from Sageev's construction of cubical complexes (Section 2.1.1). Their idea was to look at points on the modeling trees in the hyperbolic spaces which are unseen by the other projection data. The preimages of these points under the projection maps turn out to behave like walls in the hull. See Section 1.4 for a sketch of these ideas, and Section 4.2 for a full discussion.

Our main theorem stabilizes their construction, simultaneously generalizing the stability properties from Proposition 1.3 for any hyperbolic space and Proposition 1.2 for cube complexes admitting an HHS structure. The following is a more detailed version of Theorem A.

Theorem 1.4 Let $(\mathcal{X}, \mathfrak{S})$ be a colorable HHS. Then for each k there exist K and N with the following properties. For any $F \subset \mathcal{X}$ with $\#F \leq k$, there exists a finite CAT(0) cube complex $\mathcal{Q}_F$ and a K–quasimedian, (K, K)–quasi-isometric embedding $\Phi_F : \mathcal{Q}_F \to \mathcal{X}$ with $d_{\mathrm{Haus}}(\Phi_F(\mathcal{Q}_F), \mathrm{hull}(F)) \leq K$.

Moreover, if $F' \subseteq \mathcal{X}$ is another subset with $\#F' \leq k$ and $d_{\mathrm{Haus}}(F, F') \leq 1$, there is a finite CAT(0) cube complex $\mathcal{Q}_0$ and a K–quasimedian, (K, K)–quasi-isometric embedding $\Phi_0 : \mathcal{Q}_0 \to \mathcal{X}$ such that the diagram

$$(1\text{-}2)\qquad
\begin{array}{ccc}
\mathcal{Q}_F & & \\
\eta_F \downarrow & \searrow \Phi_F & \\
\mathcal{Q}_0 & \xrightarrow{\Phi_0} & \mathcal{X} \\
\eta_{F'} \uparrow & \nearrow \Phi_{F'} & \\
\mathcal{Q}_{F'} & &
\end{array}$$

commutes up to error at most K, where η_F and $\eta_{F'}$ are hyperplane deletion maps which delete at most N hyperplanes.

See Theorem 4.1 for the full version of the theorem, the details of which are necessary for our applications.

We note that CAT(0) cubical complexes are *median* spaces and HHSs are *coarse median* [7] in the sense of Bowditch [15]. As with the cubical models in [8], our stable cubical models also coarsely preserve the medians, meaning that the maps Φ_F do (as stated in Theorem 4.1). It is worth noting that in view of Russell, Spriano and Tran [53, Corollary 5.12], our cube complexes also approximate coarse median hulls.

1.4 Sketch of proofs

The proof of Theorem 4.1, of which Theorem A is an informal version, is contained in Section 4 and depends crucially on our work in Section 3. Theorems B and E are a consequence of Theorem A and our work in Section 5. We now explain the various parts and how they fit together.

In what follows, we will keep our discussion within the context of mapping class groups and hierarchies of curve graphs [42; 43], though we work in the more general context of HHSs.

Let $F \subset \mathrm{MCG}(\Sigma)$ be a finite subset and consider essential subsurfaces $V \subset \Sigma$ which are not 3–holed spheres. Roughly, the *hierarchical hull* of F, hull(F), is the set of points of $\mathrm{MCG}(\Sigma)$ whose subsurface projections in each curve graph $\mathcal{C}(V)$ lie close to the weak hull of the subsurface projection $\pi_V(F)$ of F.

In the cubulation construction of [8], the authors build a wallspace for hull(F).

To do this, they first consider the collection $\mathcal{U}_F$ of *relevant* subsurfaces $V \subset \Sigma$ for which diam$_V \pi_V(F) > K$ for some fixed threshold $K > 0$. In each of these subsurfaces, they take a tree T_F^V which coarsely models the hull of $\pi_V(F)$ in $\mathcal{C}(V)$, as discussed in Section 1.2. For each such $V \in \mathcal{U}_F$, they then consider the collection of relative projections ρ_V^W of $W \in \mathcal{U}_F$ to $\mathcal{C}(V)$, which correspond to the projection of ∂W to $\mathcal{C}(V)$ and thus are nonempty if V is neither disjoint from nor contained in W. The bounded geodesic image theorem [43] and certain consistency properties of projections — see Behrstock [3] and [9] — imply that each ρ_V^W for such W lies uniformly close to the tree T_F^V.

They then consider, roughly, the complement P_F^V in T_F^V of a regular neighborhood of these projections, which consists of a number of subtrees of T_F^V which are "unseen" by the other subsurfaces in $\mathcal{U}_F$ which interact with V. Any point in $T_F^V \setminus P_F^V$ cuts T_F^V into two subtrees. The partitions of hull(F) that define the wallspace on hull(F) come from these subdivision points in the T_F^V, namely one considers the subspaces of hull(F) whose subsurface projections to $\mathcal{C}(V)$ lie close to either of the subtrees.

While this construction is useful for studying top-dimensional quasiflats, it is unstable under perturbation of F, in that given some other F' with $d_{\text{Haus}}(F, F') \leq 1$, then the cubical models $\mathcal{Q}_F$ and $\mathcal{Q}_{F'}$ might differ by a number of hyperplanes on the order of $\text{diam}(F)$, which is not bounded.

The proof of Theorem A involves stabilizing this process in a number of places. The first step is to robustly stabilize the collection of relevant subsurfaces $\mathcal{U}_F$ (Proposition 2.14), eg so that $|\mathcal{U}_F \bigtriangleup \mathcal{U}_{F'}|$ is bounded in terms of the topology of the surface S. We do this by applying work of Bestvina, Bromberg, Fujiwara and Sisto [14], which allows us to stabilize subsurface projections (Theorem 2.9), and then use standard projection complex type arguments.

In Section 3, we stabilize the modeling trees T_F^V for each $V \in \mathcal{U}_F$. Unlike before, it will not do to simply take any Gromov modeling tree, since unboundedly many pieces of it might change in the transition from F to F' when we cut it up using the relative projection data (the P_F^V above). Instead, we use the newly stabilized relative projection data to build a new stable tree. We do this by taking a regular neighborhood of the relative projections, which then group into connected components we call *clusters*. As before, these clusters lie close to any Gromov modeling tree, but we cannot use these trees. Instead, we define a separation graph for these clusters (Definition 3.5), and then prove that the combinatorics of this graph encode how these domain clusters are arranged on *any* Gromov modeling tree. We then construct our stable tree by connecting clusters both internally and externally via minimal spanning networks in $\mathcal{C}(V)$. The stability of the cluster data then is converted into stability of the tree construction in Theorem 3.3, which, in particular, says that the set of long edges of two related trees are in bijection and within bounded Hausdorff distance, with most long edges exactly the same. See Figures 5 and 9 below.

In Section 4, we then plug these stable trees into the cubulation machine from [8]. We must be mindful of how subdivision points change when transitioning from F to F'. In particular, we construct a common refinement of the sets of subdvision points for our two sets F and F' (Proposition 4.12), with the delicate nature of this process necessitating the intricacies in the statement and proof of the stable tree theorem (Theorem 3.3). With this in hand, we prove that this common refinement induces an isomorphism between the resulting cubical models for the hulls of both sets (Proposition 4.13); see Figure 18. This isomorphism depends on a careful hierarchical analysis of when two halfspaces corresponding to two subdivision points intersect (Lemma 4.10). The full

version of the stable cubulation theorem is achieved in Theorem 4.1, which says that the two modeling cube complexes $\mathcal{Q}_F$ and $\mathcal{Q}_{F'}$ become isomorphic when we delete a bounded number of hyperplanes from each, with the bound depending only on $|F|$ and $|F'|$.

In Section 5, we adapt the normal path construction of Niblo and Reeves [46] and analyze how it changes under hyperplane deletion. In particular, for any finite CAT(0) cube complex $\mathcal{Q}$, we develop a sequence of contractions which take the extremal vertices of $\mathcal{Q}$ (ie its corners) into a "barycentric" cube at the "center" of $\mathcal{Q}$, and we prove that this contraction sequence is only boundedly perturbed by hyperplane deletions (Theorem 5.1).

Stability of the cubical model and the contraction sequence easily give the barycenter theorem (Theorem E). In the context of a bicombing (Theorem B) when $F = \{x, y\}$, we take the bicombing path from x to y to be the image in MCG(Σ) of the path obtained by following the contraction sequence of x to the barycentric cube, and then traversing the contraction sequence from the barycentric cube to y in reverse order. Once again, stability of the contraction sequence and the cubical models implies that these are uniform quasigeodesics which fellow-travel in a parametrized fashion; see Figure 25. Theorems B and E are proved in Section 6.

1.5 Outline

In Section 2 we collect some background material.

Section 3 takes place entirely in a fixed hyperbolic space, using methods from coarse hyperbolic geometry but with HHS ends in mind. The main result there is Theorem 3.3, and no other result from that section will be used elsewhere.

In Section 4, we prove the precise version of Theorem A, which is Theorem 4.1. Again, no other statement from this section will be used elsewhere. In this section, we use the combinatorial geometry of HHSs.

Section 5 uses the tools of cubical geometry, and it is independent from the previous sections. Its main result is Theorem 5.1, which once again is the only result from here needed in the rest of the paper.

Finally, in Section 6 we put all the pieces together, and we prove Theorems B and E.

Acknowledgements

2 Background

In this section, we will collect and record the various facts about cube complexes and hierarchically hyperbolic spaces that we need.

2.1 CAT(0) cube complexes

We will briefly discuss some basic aspects of CAT(0) cubical geometry. We direct the reader to Sageev's lecture notes [55] for details.

A *cube complex* is a simplicial complex X obtained from a disjoint collection of Euclidean cubes which are glued along their faces by a collection of Euclidean isometries. A cube complex is *nonpositively curved* (NPC) if its vertex links are simplicial flag complexes. An NPC cube complex is *CAT*(0) if it is a 1–connected NPC complex.

A *midcube* of an n–cube $C \subset X$ is an $(n-1)$–dimensional cube $H' \subset C$ running through the barycenter of C and parallel to one of the faces of C. A *hyperplane $H \subset X$* is a connected subspace of X such that for all closed cubes C, the intersection $H \cap C$ is either empty or a midcube of C. The *carrier* of H is the union of all of the cubes in X whose intersection with H is a midcube, and it is naturally isomorphic to $H \times [0, 1]$.

Equivalently, there is a natural equivalence relation on the set of edges in the 1–skeleton of X generated by relating two edges if they are opposite edges of some square in X. Any hyperplane can be obtained as the collection of midcubes which intersect the edges in a given equivalence class.

In this paper, we will be considering *finite* cube complexes, namely those with finitely many cubes.

Metrics on cube complexes There are many interesting metrics one can put on a CAT(0) cube complex X. We will be interested in both

- the ℓ^1 or *combinatorial* metric, d_1, which is generated by the ℓ^1 norm in each cube of X, and can be equivalently defined on the 1–skeleton $X^{(1)}$ as the path metric thereon;

- the *cubical sup metric*, d_∞, which is the metric generated by the ℓ^∞ or sup norm in each cube in X.

The following is an easy consequence of the observation that, given n, the ℓ^1 and ℓ^∞ norm on an n–cube are bi-Lipschitz equivalent.

Lemma 2.1 *For any $n > 0$, there exists $K = K(n) > 0$ such that if X is an n–dimensional cube complex, then the identity* id$\colon (X, d_1) \to (X, d_\infty)$ *is a (K, K)–quasi-isometry.*

The differences between these metrics will come up in Sections 5 and 6. See [44] for a detailed discussion of these metrics.

2.1.1 Wallspaces and Sageev's construction In Section 4, we will adopt the perspective of obtaining cube complexes as duals to wallspaces. Wallspaces were first defined by Haglund and Paulin [39]; see Hruska and Wise [41] for a broader discussion.

Let Y be a nonempty set. A *wall* in Y is a pair of subsets $W = \{\overleftarrow{W}, \overrightarrow{W}\}$ where $Y = \overleftarrow{W} \sqcup \overrightarrow{W}$. In this case, $\overleftarrow{W}$ and $\overrightarrow{W}$ are called *halfspaces*.

Two points $x, y \in Y$ are *separated* by a wall W if x is contained in a different halfspace from y.

A *wallspace* is a set Y with a collection of walls $\mathcal{W}$ on Y such that the number of walls separating any pair of points is finite.

An *orientation* on a wallspace $(Y, \mathcal{W})$ is an assignment σ such that, for each $W \in \mathcal{W}$, we have $\sigma(W) \in \{\overleftarrow{W}, \overrightarrow{W}\}$. The orientation σ is called *coherent* if, for all $W, W' \in \mathcal{W}$, we have $\sigma(W) \cap \sigma(W') \neq \varnothing$. We call σ *canonical* if there exists $x \in X$ such that $x \in \sigma(W)$ for all but finitely many $W \in \mathcal{W}$.

Given a wallspace $(Y, \mathcal{W})$, we can consider the cube complex $X(Y, \mathcal{W})$ constructed as follows. The 0–cubes of $X(Y, \mathcal{W})$ are coherent, canonical orientations of $(Y, \mathcal{W})$. Two 0–simplices are connected by a 1–cube if, seen as orientations, they differ on only one

wall. Finally, all subcomplexes of the 1–skeleton isomorphic to the 1–skeleton of an n–cube cube get filled by an n–cube.

Work of Chatterji and Niblo [21], Chepoi [22], and Nica [47] — building off of work of Sageev [54] — gives that $X(Y, \mathcal{W})$ is a CAT(0) cube complex. We call $X(Y, \mathcal{W})$ the *dual cube complex* associated to the wallspace $(Y, \mathcal{W})$.

2.1.2 Hyperplane deletions In Section 5, we will be interested in understanding how cube complexes change under deletions of hyperplanes, so we will use the alternative perspective of obtaining cube complexes from sets of hyperplanes. We briefly explain how this works.

Let X be a CAT(0) cube complex and $\mathcal{H}_X$ its (finite) set of hyperplanes. Then we can identify each hyperplane $H \in \mathcal{H}_X$ with the two halfspaces into which it partitions $X^{(0)}$. As such, we can and will think of $(X^{(0)}, \mathcal{H}_X)$ as a wallspace, and one can show that X is the dual cube complex associated to $(X, \mathcal{H}_X)$.

Given any subset $\mathcal{H} \subset \mathcal{H}_X$ of hyperplanes in a cube complex X, there is a natural cube complex $X(\mathcal{H})$ defined as the dual cube complex associated to the wallspace defined by $\mathcal{H}$ in X. In particular, each point in $X(\mathcal{H})$ is a choice of coherent, canonical orientations of the half-spaces defined by $\mathcal{H}$.

With this notation, we can now define hyperplane deletions:

Definition 2.2 Let X be a CAT(0) cube complex obtained with hyperplanes $\mathcal{H}_X$. For a finite collection of hyperplanes $\mathcal{G} \subset \mathcal{H}_X$, the *hyperplane deletion map* for H is the map
$$\mathrm{Res}_{\mathcal{H}_X \setminus \mathcal{G}} \colon X \to X(\mathcal{H}_X \setminus \mathcal{G})$$
obtained by restriction of orientations, where $X(\mathcal{H}_X \setminus \mathcal{G})$ is the dual cube complex associated to the wallspace $(X, \mathcal{H}_X \setminus \mathcal{G})$.

Equivalently, the map $\mathrm{Res}_{\mathcal{H}_X \setminus \mathcal{G}}$ is the quotient map which collapses the $[0, 1]$ factor of each of the carriers of the hyperplanes in $\mathcal{G}$ (recall that the carrier of the hyperplane H is naturally isomorphic to $H \times [0, 1]$). We note that these maps appear elsewhere in the literature, eg [19].

We also record the following fact, which indicates that the isomorphism type of the cube complex coming from a wallspace is determined by the intersection pattern of halfspaces. The proof is elementary.

Lemma 2.3 Let $\mathcal{W}$ and $\mathcal{W}'$ be wallspaces, and let $\iota\colon \mathcal{H}_\mathcal{W} \to \mathcal{H}_{\mathcal{W}'}$ be a bijection of their halfspaces which preserves complements and disjointness.

Denote by j the induced map on walls

$$\{H, H^c\} \mapsto \{\iota(H), \iota(H)^c\}$$

and by $h(x) = \iota \circ x \circ j^{-1}$ the induced map on orientations. Then h, viewed as a map on 0–cubes, induces an isomorphism $h\colon \mathcal{Y}_\mathcal{W} \to \mathcal{Y}_{\mathcal{W}'}$ between the corresponding CAT(0) cube complexes.

2.2 HHS axioms

We recall from [7] the definition of a hierarchically hyperbolic space.

Definition 2.4 (HHS) The q–quasigeodesic space $(\mathcal{X}, \mathrm{dist}_\mathcal{X})$ is a *hierarchically hyperbolic space* if there exists $\delta \geq 0$, an index set $\mathfrak{S}$, and a set $\{\mathcal{C}W \mid W \in \mathfrak{S}\}$ of δ–hyperbolic spaces $(\mathcal{C}U, \mathrm{dist}_U)$, such that the following conditions are satisfied.

(1) **Projections** There is a set $\{\pi_W\colon \mathcal{X} \to 2^{\mathcal{C}W} \mid W \in \mathfrak{S}\}$ of *projections* sending points in $\mathcal{X}$ to sets of diameter bounded by some $\xi \geq 0$ in the various $\mathcal{C}W \in \mathfrak{S}$. Moreover, there exists K such that, for all $W \in \mathfrak{S}$, the coarse map π_W is (K, K)–coarsely Lipschitz and $\pi_W(\mathcal{X})$ is K–quasiconvex in $\mathcal{C}W$.

(2) **Nesting** $\mathfrak{S}$ is equipped with a partial order $\sqsubseteq$, and either $\mathfrak{S} = \varnothing$ or $\mathfrak{S}$ contains a unique $\sqsubseteq$–maximal element; when $V \sqsubseteq W$, we say V is *nested* in W. (We emphasize that $W \sqsubseteq W$ for all $W \in \mathfrak{S}$.) For each $W \in \mathfrak{S}$, we denote by $\mathfrak{S}_W$ the set of $V \in \mathfrak{S}$ such that $V \sqsubseteq W$. Moreover, for all $V, W \in \mathfrak{S}$ with $V \subsetneq W$, there is a specified subset $\rho_W^V \subset \mathcal{C}W$ with $\mathrm{diam}_{\mathcal{C}W}(\rho_W^V) \leq \xi$. There is also a *projection* $\rho_V^W\colon \mathcal{C}W \to 2^{\mathcal{C}V}$. (The similarity in notation is justified by viewing ρ_W^V as a coarsely constant map $\mathcal{C}V \to 2^{\mathcal{C}W}$.)

(3) **Orthogonality** $\mathfrak{S}$ has a symmetric and antireflexive relation called *orthogonality*, and we write $V \perp W$ when V and W are orthogonal. Also, whenever $V \sqsubseteq W$ and $W \perp U$, we require that $V \perp U$. We require that for each $T \in \mathfrak{S}$ and each $U \in \mathfrak{S}_T$ for which $\{V \in \mathfrak{S}_T \mid V \perp U\} \neq \varnothing$, there exists $W \in \mathfrak{S}_T \setminus \{T\}$ such that, whenever $V \perp U$ and $V \sqsubseteq T$, we have $V \sqsubseteq W$. Finally, if $V \perp W$, then V and W are not $\sqsubseteq$–comparable.

(4) **Transversality and consistency** If $V, W \in \mathfrak{S}$ are not orthogonal and neither is nested in the other, then we say V and W are *transverse*, denoted by $V \pitchfork W$. There

exists $\kappa_0 \geq 0$ such that if $V \pitchfork W$, then there are sets $\rho_W^V \subseteq \mathcal{C}W$ and $\rho_V^W \subseteq \mathcal{C}V$, each of diameter at most ξ and satisfying

$$(2\text{-}1) \qquad \min\{\operatorname{dist}_W(\pi_W(x), \rho_W^V), \operatorname{dist}_V(\pi_V(x), \rho_V^W)\} \leq \kappa_0$$

for all $x \in \mathcal{X}$.

For $V, W \in \mathfrak{S}$ satisfying $V \sqsubset W$ and for all $x \in \mathcal{X}$,

$$(2\text{-}2) \qquad \min\{\operatorname{dist}_W(\pi_W(x), \rho_W^V), \operatorname{diam}_{\mathcal{C}V}(\pi_V(x) \cup \rho_V^W(\pi_W(x)))\} \leq \kappa_0.$$

The preceding two inequalities are the *consistency inequalities* for points in $\mathcal{X}$.

Finally, if $U \sqsubset V$, then $\operatorname{dist}_W(\rho_W^U, \rho_W^V) \leq \kappa_0$ whenever $W \in \mathfrak{S}$ satisfies either $V \subsetneq W$ or $V \pitchfork W$ and $W \not\sqsubset U$.

(5) Finite complexity There exists $n \geq 0$, the *complexity* of $\mathcal{X}$ (with respect to $\mathfrak{S}$), such that any set of pairwise $\sqsubset$–comparable elements has cardinality at most n.

(6) Large links There exist $\lambda \geq 1$ and $E \geq \max\{\xi, \kappa_0\}$ such that the following holds. Let $W \in \mathfrak{S}$ and let $x, x' \in \mathcal{X}$. Let $N = \lambda \operatorname{dist}_W(\pi_W(x), \pi_W(x')) + \lambda$. Then there exists $\{T_i\}_{i=1,\dots,\lfloor N \rfloor} \subseteq \mathfrak{S}_W \setminus \{W\}$ such that for all $T \in \mathfrak{S}_W \setminus \{W\}$, either $T \in \mathfrak{S}_{T_i}$ for some i, or $\operatorname{dist}_T(\pi_T(x), \pi_T(x')) < E$. Also, $\operatorname{dist}_W(\pi_W(x), \rho_W^{T_i}) \leq N$ for each i.

(7) Bounded geodesic image There exists $\kappa_0 > 0$ such that for all $W \in \mathfrak{S}$, all $V \in \mathfrak{S}_W \setminus \{W\}$, and all geodesics γ of $\mathcal{C}W$, either

$$\operatorname{diam}_{\mathcal{C}V}(\rho_V^W(\gamma)) \leq \kappa_0 \quad \text{or} \quad \gamma \cap \mathcal{N}_{\kappa_0}(\rho_W^V) \neq \varnothing.$$

(8) Partial realization There exists a constant α with the following property. Let $\{V_j\}$ be a family of pairwise orthogonal elements of $\mathfrak{S}$, and let $p_j \in \pi_{V_j}(\mathcal{X}) \subseteq \mathcal{C}V_j$. Then there exists $x \in \mathcal{X}$ such that

- $\operatorname{dist}_{V_j}(x, p_j) \leq \alpha$ for all j;
- for each j and each $V \in \mathfrak{S}$ with $V_j \sqsubset V$, we have $\operatorname{dist}_V(x, \rho_V^{V_j}) \leq \alpha$; and
- if $W \pitchfork V_j$ for some j, then $\operatorname{dist}_W(x, \rho_W^{V_j}) \leq \alpha$.

(9) Uniqueness For each $\kappa \geq 0$, there exists $\theta_u = \theta_u(\kappa)$ such that if $x, y \in \mathcal{X}$ and $\operatorname{dist}_{\mathcal{X}}(x, y) \geq \theta_u$, then there exists $V \in \mathfrak{S}$ such that $\operatorname{dist}_V(x, y) \geq \kappa$.

We often refer to $\mathfrak{S}$, together with the nesting and orthogonality relations, and the projections as a *hierarchically hyperbolic structure* for the space $\mathcal{X}$.

Where it will not cause confusion, given $U \in \mathfrak{S}$, we will often suppress the projection map π_U when writing distances in $\mathcal{C}U$; ie given $x, y \in \mathcal{X}$ and $p \in \mathcal{C}U$, we write $\mathrm{dist}_U(x, y)$ for $\mathrm{diam}_{\mathcal{C}U}(\pi_U(x) \cup \pi_U(y))$ and $\mathrm{dist}_U(x, p)$ for $\mathrm{diam}_{\mathcal{C}U}(\pi_U(x) \cup \{p\})$. Given $A \subset \mathcal{X}$ and $U \in \mathfrak{S}$ we let $\pi_U(A)$ denote $\bigcup_{a \in A} \pi_U(a)$.

There is a natural notion of automorphism of an HHS, which we now briefly explain. These were originally defined in [7], but we give a more restrictive definition which is essentially equivalent, as explained in [30, Section 2.1]. An automorphism g of an HHS $(\mathcal{X}, \mathfrak{S})$ is an isometry of $\mathcal{X}$ together with a bijection $\mathfrak{S} \to \mathfrak{S}$, also denoted by $U \mapsto gU$, which preserves nesting and orthogonality, and isometries between corresponding hyperbolic spaces, again still denoted by $g \colon \mathcal{C}(U) \to \mathcal{C}(gU)$. We require that $g\pi_U(x) = \pi_{gU}(gx)$ for all $x \in \mathcal{X}$ and $U \in \mathfrak{S}$, and $g\rho_V^U = \rho_{gV}^{gU}$ for all $U, V \in \mathfrak{S}$ where this is defined.

We let $\mathrm{Aut}(\mathcal{X}, \mathfrak{S})$ denote the group of HHS automorphisms of $(\mathcal{X}, \mathfrak{S})$.

We say that a group G is a *hierarchically hyperbolic group* if it acts properly and coboundedly by HHS automorphisms on some HHS $(\mathcal{X}, \mathfrak{S})$.

2.2.1 Some useful facts We now recall results from [7] that will be useful later on.

Definition 2.5 Let $\kappa \geq 0$ and let $\vec{b} \in \prod_{U \in \mathfrak{S}} 2^{\mathcal{C}U}$ be a tuple such that for each $U \in \mathfrak{S}$, the U–coordinate b_U has diameter $\leq \kappa$. Then $\vec{b}$ is κ–*consistent* if for all $V, W \in \mathfrak{S}$,

$$\min\{\mathrm{dist}_V(b_V, \rho_V^W), \mathrm{dist}_W(b_W, \rho_W^V)\} \leq \kappa$$

whenever $V \pitchfork W$ and

$$\min\{\mathrm{dist}_W(x, \rho_W^V), \mathrm{diam}_V(b_V \cup \rho_V^W)\} \leq \kappa$$

whenever $V \subsetneq W$.

The following is [7, Theorem 4.5].

Theorem 2.6 (distance formula) *Let* $(\mathcal{X}, \mathfrak{S})$ *be a hierarchically hyperbolic space. Then there exists* s_0 *such that for all* $s \geq s_0$, *there exist* C *and* K *such that for all* $x, y \in \mathcal{X}$,

$$\mathrm{dist}(x, y) \asymp_{K,C} \sum_{U \in \mathfrak{S}} \{\!\!\{\mathrm{dist}_U(x, y)\}\!\!\}_s,$$

where $\{\!\!\{A\}\!\!\}_B$ *denotes the quantity which is* A *if* $A \geq B$ *and* 0 *otherwise.*

We recall the notion of a *hierarchical hull*, which originates in [9] for the setting of mapping class groups, and extends to the HHS setting in [7]. Given a constant θ, for any $F \subset \mathcal{X}$ we define

$$(2\text{-}3) \qquad H_\theta(F) = \{x \in \mathcal{X} \mid \forall V \in \mathfrak{S}, \ \pi_V(x) \in \mathcal{N}_\theta(\text{hull}(\pi_V(F)))\},$$

where hull(A) denotes the union of all geodesics connecting points of A. In words, H_θ is the set of points whose projections in every hyperbolic factor space land in a specified neighborhood of the hull of the image of F. That this is sufficiently nonvacuous is indicated by the following result which, as we will see, is an easy consequence of [7, Theorem 4.4].

Theorem 2.7 *Let $(\mathcal{X}, \mathfrak{S})$ be a hierarchically hyperbolic space. Given k, there exist θ and κ such that, if $F \subset \mathcal{X}$ is a set of cardinality k then for every $V \in \mathfrak{S}$ the image $\pi_V(H_\theta(F))$ and the hull of $\pi_V(F)$ lie within Hausdorff distance κ of each other.*

Proof By definition, $\pi_V(H_\theta(F))$ lies in a controlled neighborhood of hull($\pi_V(F)$), so we are left to show that any point on a geodesic connecting points of $\pi_V(F)$ lies close to $\pi_V(H_\theta(F))$. This follows from [7, Theorem 4.4], which says that any two points of F are connected by a hierarchy path (with uniform constant). These are defined in Definition 6.5 below, but here we only need that said path has projection to any given $\mathcal{C}(W)$ which, as a set, uniformly coarsely coincides with a geodesic between the projections of the endpoints. In particular, for any θ large enough, the path will be contained in $H_\theta(F)$. $\qquad\square$

2.3 Refined projections and stable subsurface collections

We will be working in a broad but restricted class of HHSs:

Definition 2.8 Let $(\mathcal{X}, \mathfrak{S})$ be an HHS and let $G < \text{Aut}(\mathfrak{S})$. We say that $(\mathcal{X}, \mathfrak{S})$ is *G–colorable* if there exists a decomposition of $\mathfrak{S}$ into finitely many families $\mathfrak{S}_i$ such that each $\mathfrak{S}_i$ is pairwise transverse and G acts on $\{\mathfrak{S}_i\}_i$ by permutations. We say that $(\mathcal{X}, \mathfrak{S})$ is *colorable* if it is $\text{Aut}(\mathfrak{S})$–colorable.

The notion of colorability is inspired by Bestvina, Bromberg and Fujiwara [12], who, essentially, proved that $\text{MCG}(\Sigma)$ and $\mathcal{T}(\Sigma)$ are finitely $\text{MCG}(\Sigma)$–colorable HHSs; we now explain how their work proves this fact. First of all, the standard HHS structures of $\text{MCG}(\Sigma)$ and $\mathcal{T}(\Sigma)$ both have as index set $\mathfrak{S}$ the set of all essential subsurfaces (including the disconnected ones) that do not have pairs of pants as connected com-

ponents. Except for allowing disconnected subsurfaces, this is the same as the set Y considered in [12, Proposition 5.8]. The proposition yields a certain decomposition $Y = Y^1 \sqcup \cdots \sqcup Y^k$, and from an inspection of the first paragraph of the proof one sees that the Y^i are the orbits of a certain finite-index subgroup G of MCG(Σ). We can define our $\mathfrak{S}_i$ to be the orbits in $\mathfrak{S}$ (rather than Y) of this same finite-index subgroup. The fact that distinct elements of each Y^i have intersecting boundaries — as given by [12, Proposition 5.8] — implies that for any two distinct subsurfaces in the same $\mathfrak{S}_i$ have connected components with intersecting boundaries. Since transversality in MCG(Σ) and $\mathcal{T}(\Sigma)$ is defined via intersecting boundaries, we are done.

We sometimes refer to the $\mathfrak{S}_i$ as *BBF families*.

For $A, B \subseteq \mathcal{C}(Y)$, we define $d_Y(A, B) := \mathrm{diam}_{\mathcal{C}(Y)}(A \cup B)$.

Theorem 2.9 [14] *Let $(\mathcal{X}, \mathfrak{S})$ be a G–colorable HHS for $G < \mathrm{Aut}(\mathfrak{S})$ with standard projections $\hat{\pi}_-$ and $\hat{\rho}_-^-$. There exist $\theta > 0$ and refined projections π_- and ρ_-^- with the same domains and ranges, respectively, such that:*

(1) *If X and Y lie in different $\mathfrak{S}_j$, and $\hat{\rho}_Y^X$ is defined, then $\rho_Y^X = \hat{\rho}_Y^X$.*

(2) *If $X, Y \in \mathfrak{S}_j$ are distinct, then the Hausdorff distance between ρ_Y^X and $\hat{\rho}_Y^X$ is at most θ.*

(3) *If $x \in \mathcal{X}$ and $Y \in \mathfrak{S}$, then the Hausdorff distance between $\pi_Y(x)$ and $\hat{\pi}_Y(x)$ is at most θ.*

(4) *If $X, Y, Z \in \mathfrak{S}_j$ for some j are pairwise distinct and $d_Y(\rho_Y^X, \rho_Y^Z) > \theta$, then $\rho_Z^X = \rho_Z^Y$.*

(5) *Let $x \in \mathcal{X}$, and $Y, Z \in \mathfrak{S}_j$ for some j be pairwise distinct. If $d_Y(\pi_Y(x), \rho_Y^Z) > \theta$ then $\pi_Z(x) = \rho_Z^Y$.*

Moreover, $(\mathcal{X}, \mathfrak{S})$ equipped with π_- and ρ_-^- is an HHS, $G < \mathrm{Aut}(\mathfrak{S})$, and it is G–colorable.

Proof The idea is to apply the construction from [14] to the standard projections $\hat{\pi}_-$ and $\hat{\rho}_-^-$ and distances $\hat{d}_-$ for the sets $\mathfrak{S}_i \cup \mathcal{X}$ for each i, where we think of $\mathcal{X}$ as a collection of single point spaces $x = \{x\}$ for each $x \in \mathcal{X}$.

Given a point $x \in \mathcal{X}$, we define projections ρ_x^- from domains in $\mathfrak{S}_i$ and $\mathcal{X}$ to x as the constant map $\rho_x^- \equiv x$. It is easily checked that $\mathfrak{S}_i \cup \mathcal{X}$, once equipped with the original projections $\hat{\pi}_-$ and $\hat{\rho}_-^-$ and these additional projections, satisfies the projection

axioms from [12]. The existence of projections and distances π_-, ρ_-^- and d_-, and that all properties hold for them, is then an immediate consequence of [14, Theorem 4.1].

Finally, the fact that $(\mathcal{X}, \mathfrak{S})$ equipped with these projections is an HHS follows from the fact that the new projections are bounded distance away from the old ones, by items (1), (2) and (3).

The fact that G still acts by automorphisms on the new structure follows from equivariance of the construction of the new projections, meaning [14, Theorem 4.1(3)]. □

Definition 2.10 We say that a G–colorable HHS $(\mathcal{X}, \mathfrak{S})$ with $G < \mathrm{Aut}(\mathfrak{S})$ has *stable projections* if it is equipped with the projections provided by Theorem 2.9.

For the rest of this section, fix a G–colorable HHS $(\mathcal{X}, \mathfrak{S})$ with $G < \mathrm{Aut}(\mathfrak{S})$ and with stable projections. In particular, we assume that the standard projections for $(\mathcal{X}, \mathfrak{S})$ satisfy the stability properties in Theorem 2.9.

As usual, $d_Y(x_1, x_2)$ denotes $\mathrm{diam}_{\mathcal{C}(Y)}(A_1 \cup A_2)$, where

- $A_i = \pi_Y(x_i)$ if $x_i \in \mathcal{X}$,
- $A_i = \rho_Y^{x_i}$ if $x_i \in \mathfrak{S}$ and either $x_i \subsetneq Y$ or $x_i \pitchfork Y$.

For any pair of points $x, y \in \mathcal{X}$ and constant $K > 0$, we let $\mathrm{Rel}_K(x, y) \subset \mathfrak{S}$ denote the collection of $Y \in \mathfrak{S}$ such that $d_Y(x, y) > K$; we also set $\mathrm{Rel}_K^i(x, y) = \mathrm{Rel}_K(x, y) \cap \mathfrak{S}_i$.

Let θ satisfy Theorem 2.9(5). Following eg [9; 12; 24], we now consider a relation on $\mathrm{Rel}_K(x, y)$ — the properties claimed below follow from [7, Proposition 2.8] and Definition 2.4(4). For any $K > 10\theta$, $\mathrm{Rel}_K(x, y)$ is a partially ordered set with order $\prec$ such that $X \prec Y$ whenever $X \pitchfork Y$ and one of the following equivalent conditions hold:

$$d_Y(x, \rho_Y^X) \le \theta, \quad d_X(\rho_X^Y, y) \le \theta, \quad d_Y(\rho_Y^X, y) \ge K - \theta, \quad d_X(x, \rho_X^Y) \ge K - \theta.$$

When restricted to $\mathrm{Rel}_K^i(x, y)$, the relation $\prec$ becomes a total order.

For a finite set $F \subset \mathcal{X}$, we define

$$\mathrm{Rel}_K(F) = \bigcup_{x, y \in F} \mathrm{Rel}_K(x, y) \quad \text{and} \quad \mathrm{Rel}_K^i(F) = \mathrm{Rel}_K(F) \cap \mathfrak{S}_i.$$

The following stability lemma follows directly from the construction in [14].

Lemma 2.11 *There exists $K \gg 2\theta$ such that if $x, y, y' \in \mathcal{X}$ satisfy $d_{\mathcal{X}}(y, y') \le 1$, then for each i,*

$$|\mathrm{Rel}_K^i(x, y) \bigtriangleup \mathrm{Rel}_K^i(x, y')| \le 2.$$

Proof By contradiction, suppose we have distinct elements

$$Y_0, Y_1, Y_2 \in \mathrm{Rel}_K^i(x, y) \setminus \mathrm{Rel}_K^i(x, y')$$

with $Y_0 \prec Y_1 \prec Y_2$. If $K > 10\theta$, applying the definition of $\prec$ and Theorem 2.9(5), we see that $\pi_{Y_1}(x) = \rho_{Y_1}^{Y_0}$ and $\pi_{Y_1}(y) = \rho_{Y_1}^{Y_2}$. Also, since $\pi_{Y_2}(y)$ and $\pi_{Y_2}(y')$ are uniformly close to each other, if K is sufficiently large then we have $d_{Y_2}(Y_1, y) > \theta$ and hence, one again, $\pi_{Y_1}(y') = \rho_{Y_1}^{Y_2}$. But then $d_Y(x, y') = d_Y(x, y) \geq K$, which contradicts $Y_1 \notin \mathrm{Rel}_K^i(x, y')$. $\qquad\square$

Proposition 2.12 *Let $K \gg 2\theta$ and $F \subset \mathcal{X}$ be any finite set. There exists*

$$M = M(K, \mathfrak{S}, |F|) > 0$$

such that, for any $F' \subset \mathcal{X}$ with $d_{\mathrm{Haus}}(F, F') \leq 1$ and $|F'| \leq |F|$,

$$|\mathrm{Rel}_K(F) \triangle \mathrm{Rel}_K(F')| < M.$$

Proof Assume throughout the proof that K is sufficiently large.

Since there are finitely many colors, it suffices to prove the analogous statement for $\mathrm{Rel}_K^i(F) \triangle \mathrm{Rel}_K^i(F')$ for any given i.

Applying Lemma 2.11 twice, we see that if $d_{\mathcal{X}}(x, x'), d_{\mathcal{X}}(y, y') \leq 1$, then

$$|\mathrm{Rel}_K^i(x, y) \setminus \mathrm{Rel}_K^i(x', y')| \leq 4.$$

For each of the $|F|^2$ pairs $x, y \in F$, we can pick any x', y' with $d_{\mathcal{X}}(x, x'), d_{\mathcal{X}}(y, y') \leq 1$ such that there are at most $4|F|^2$ elements of $\mathrm{Rel}_K^i(F) = \bigcup_{x,y \in F} \mathrm{Rel}_K^i(x, y)$ that are not in $\mathrm{Rel}_K^i(F')$; ie $|\mathrm{Rel}_K^i(F) \setminus \mathrm{Rel}_K^i(F')| \leq 4|F|^2$. Symmetrically, we have $|\mathrm{Rel}_K^i(F') \setminus \mathrm{Rel}_K^i(F)| \leq 4|F'|^2$, and since $|F'| \leq |F|$ by assumption, we finally get $|\mathrm{Rel}_K^i(F) \triangle \mathrm{Rel}_K^i(F')| < 8|F|^2$, as required. $\qquad\square$

2.4 Bounding involved domains

Let $(\mathcal{X}, \mathfrak{S})$ be a G–colorable HHS with stable projections for $G < \mathrm{Aut}(\mathfrak{S})$, as provided by Theorem 2.9.

Let $F, F' \subset \mathcal{X}$ with $|F|, |F'| \leq k$ and $d_{\mathrm{Haus}}(F, F') \leq 1$. We will now prove some stronger stability results about how the set of relevant domains (and their subdomains) changes between F and F'.

For any $K \gg 2\theta$ as above, let $\mathcal{U}(F) = \operatorname{Rel}_K(F)$ and $\mathcal{U}(F') = \operatorname{Rel}_K(F')$. Given $V \in \mathfrak{S}$, let $\mathcal{U}^V(F) = \{W \in \mathcal{U}(F) \mid W \subsetneq V\}$ and define $\mathcal{U}^V(F')$ similarly.

In many of our stability results, we will need to know how domains in $\mathcal{U}(F)$ may differ from those in $\mathcal{U}(F')$. We call such domains *involved*, and they come in two flavors:

Definition 2.13 We say that $V \in \mathcal{U}(F) \cup \mathcal{U}(F')$ is *involved* in the transition between F and F' if

(1) $\pi_V(F) \neq \pi_V(F')$, or

(2) $\mathcal{U}^V(F) \neq \mathcal{U}^V(F')$.

Proposition 2.14 *If K is sufficiently large then the following holds. Given $k > 0$ there exists $N_1 = N_1(k, \mathfrak{S}) > 0$ such that, if $|F|, |F'| \leq k$ and $d_{\mathrm{Haus}}(F, F') \leq 1$, then there are at most N_1 domains $V \in \mathcal{U}(F) \cup \mathcal{U}(F')$ involved in the transition between F and F'.*

Proof By Proposition 2.12, it suffices to bound the number of involved domains in $\mathcal{U}(F) \cap \mathcal{U}(F')$. However, we will still have to bound the number of involved domains of type (1) in $\mathcal{U}(F) \cup \mathcal{U}(F')$. We note that, since F and F' lie within Hausdorff distance 1, up to increasing K we can assume that for each $V \in \mathcal{U}(F)$ we have $\operatorname{diam}_{CV}(\pi_V(F')) \geq K/2$, and similarly for $V \in \mathcal{U}(F')$.

Involved of type (1) Let $x \in F$. We say that $V \in \mathcal{U}(F) \cup \mathcal{U}(F')$ is *exposed to x* if $\pi_V(x)$ is not contained in $\pi_V(F')$. We define exposure for $x \in F'$ similarly (with an abuse, here we are considering F and F' as disjoint, so we should actually define exposure for $x \in F \sqcup F'$).

Observe that $V \in \mathcal{U}(F) \cup \mathcal{U}(F')$ satisfies $\pi_V(F) \neq \pi_V(F')$ if and only if V is exposed to some x in either F or F'. Hence it suffices to bound the number of exposed domains.

Since $|F|, |F'| \leq k$, we may fix a point $x \in F$ and consider domains V which are exposed to x. The case of domains exposed to points in F' follows from a symmetric argument.

Given x and V as above, there is a $y \in F$ such that $d_V(x, y) \geq K/4$ (this is because $\operatorname{diam}_{C(V)}(\pi_V(F)) \geq K/2$). Since F has at most k elements, we can further fix y with said property.

Suppose for a contradiction that there exist domains $V_1, V_2, V_3 \in (\mathcal{U}(F) \cup \mathcal{U}(F')) \cap \mathfrak{S}_i$ which are exposed to x, where $\mathfrak{S}_i$ is the ith BBF family, making the V_i necessarily

pairwise transverse (this suffices since there are finitely many BBF families). Up to reordering, we have $V_1 \prec V_2 \prec V_3$ in $\mathrm{Rel}_{K/2}(x, y)$.

Since F and F' lie at Hausdorff distance at most 1, there is a pair $x_2, y_2 \in F'$ such that $d(x, x_2), d(y, y_2) \leq 1$, and necessarily we have $\pi_{V_2}(x) \neq \pi_{V_2}(x_2)$ (as we cannot have the containment "$\subseteq$").

Since $d(x, x_2), d(y, y_2) \leq 1$, by taking $K \gg 2\theta$ sufficiently large, we can ensure that $d_{V_i}(x_2, y_2) > 2\theta$ for $i = 1, 2, 3$. Since $V_1 \prec V_2 \prec V_3 \in \mathrm{Rel}_{K/2}(x, y)$, we also must have the same order $V_1 \prec V_2 \prec V_3$ in $\mathrm{Rel}_{2\theta}(x_2, y_2)$.

Thus by Theorem 2.9, it follows that $\pi_{V_2}(x_2) = \rho_{V_2}^{V_1}$. However, Theorem 2.9 also implies that $\pi_{V_2}(x) = \rho_{V_2}^{V_1}$. This contradicts the fact that $\pi_{V_2}(x) \neq \pi_{V_2}(x_2)$, and completes the proof that there is a bound on domains of type (1).

Involved of type (2) Notice that if $W \in \mathcal{U}(F) \cap \mathcal{U}(F')$ is of type (2), then there necessarily exists an exposed domain $V \in \mathcal{U}(F) \cup \mathcal{U}(F')$ of type (1) with $V \sqsubset W$ (by a $\sqsubset$–minimality argument: there must exist some $V \subsetneq W$ with $\pi_V(F) \neq \pi_V(F')$, by virtue of V lying in the symmetric difference of $\mathcal{U}(F)$ and $\mathcal{U}(F')$, and hence a $\sqsubset$–minimal such V exists by Definition 2.4(5)). We therefore bound the number of such containers W for a fixed exposed domain V, of which there is a bounded number by the first part of the proof.

Since $|F|, |F'| \leq k$, it suffices to fix $x, y \in F$ and provide a bound on the number of elements $W \in \mathrm{Rel}_K(x, y)$ which contain a fixed domain V.

In fact, we can conclude with an argument that does not rely on colorability, which we record here as a separate lemma since it might be of independent interest.

Lemma 2.15 *Let $(\mathcal{X}, \mathfrak{S})$ be an HHS. Then there exists N such that for any sufficiently large K the following holds. Let $x, y \in \mathcal{X}$ and $V \in \mathfrak{S}$. Then there are at most N elements $W \in \mathfrak{S}$ with $V \subsetneq W$ such that $W \in \mathrm{Rel}_K(x, y)$.*

Proof We fix K large enough that $\mathrm{Rel}_K(x, y)$ is partially ordered for all $x, y \in \mathcal{X}$, and larger than $10\kappa_0$ for κ_0 as in Definition 2.4(4).

First of all, there is a bound on the maximal number of pairwise $\sqsubset$–comparable domains by Definition 2.4(5), and similarly there is a bound on the maximal number of pairwise orthogonal domains by [29, Lemma 1.4]. Hence, in view of Ramsey's theorem, there exists N such that any collection of more than N domains contains three pairwise transverse elements.

Suppose, by way of contradiction, that there exist more than N domains W as in the statement, for given x, y and V, and consider $W_1 \prec W_2 \prec W_3$ in $\mathrm{Rel}_K(x, y)$. By Definition 2.4(4), we have that $d_{W_2}(\pi_{W_2}(x), \rho_{W_2}^{W_1}) \leq \kappa_0$ and $d_{W_2}(\pi_{W_2}(y), \rho_{W_2}^{W_3}) \leq \kappa_0$. However, since $V \sqsubset W_i$ for each $i = 1, 2, 3$, we have $d_{W_2}(\rho_{W_2}^{W_1}, \rho_{W_2}^{V}) < \kappa_0$ and $d_{W_2}(\rho_{W_2}^{W_3}, \rho_{W_2}^{V}) < \kappa_0$, and so $d_{W_2}(\rho_{W_2}^{W_1}, \rho_{W_2}^{W_3}) < 2\kappa_0$ by the triangle inequality. But since $d_{W_2}(x, y) \geq d_{W_2}(\rho_{W_2}^{W_1}, \rho_{W_2}^{W_3}) - 2\kappa_0 > K - 2\kappa_0 > 2\kappa_0$ by assumption, this is a contradiction. $\qquad\square$

As explained above, the lemma concludes the proof of the proposition. $\qquad\square$

Remark 2.16 While not strictly necessary, we can simplify the setup that we deal with in Section 3 thanks to the following: Given an HHS, we can $\mathrm{Aut}(\mathcal{X}, \mathfrak{S})$–equivariantly change the structure in a way that all $\pi_V(x)$ and ρ_V^U for $U \subsetneq V$ are points, rather than bounded sets, and that moreover the new structure has stable projections if the old one did. This can be achieved, for example, by replacing each $\mathcal{C}(V)$ by the nerve of the covering given by subsets of sufficiently large diameter — which is quasi-isometric to $\mathcal{C}(V)$. In particular, the vertices of the new $\mathcal{C}(V)$ are labeled by bounded sets, and we can redefine $\pi_V(x)$ to be the vertex labeled by $\pi_V(x)$, and similarly for ρ_V^U; all properties required are straightforward to check.

In Section 3, we will deal with finite subsets of a hyperbolic space. If in Section 4 we did not modify the HHS structure as outlined above, we would instead have to deal with finite collections of bounded subsets. This is possible, but would make the arguments more opaque.

3   Stable trees

In this section we will consider the geometry of trees in a δ–hyperbolic space, in preparation for arguments that will take place in the individual hyperbolic spaces of our hierarchical structure. Our main result will be Theorem 3.3, stated below after some preliminary definitions. This is the only result from this section that will get used later (namely, in Section 4).

Fix a geodesic δ–hyperbolic space $\mathcal{Z}$. For a finite subset $F \subset \mathcal{Z}$ let $\mathrm{hull}(F) \subset \mathcal{Z}$ be the set of geodesics connecting points of F. Hyperbolicity tells us that $\mathrm{hull}(F)$ can be approximated by a finite tree with accuracy depending only on δ and the cardinality $\#F$. To systematize this for the purposes of this section, we make the following definitions.

Let us fix a function λ which assigns, to any finite subset F of $\mathcal{Z}$, a minimal network spanning F. That is, $\lambda(F)$ is a 1–complex embedded in $\mathcal{Z}$ with the property that $\lambda(F) \cup F$ is connected, and has minimal length among all such 1–complexes (where the length of a 1–complex embedded in $\mathcal{Z}$ is the sum of the lengths of all edges). Minimality implies $\lambda(F)$ is a tree. Let us similarly define λ' which assigns, to any finite collection $A_1, \ldots, A_k$ of subsets of $\mathcal{Z}$, a minimal network that spans them. That is, $\lambda'(A_1, \ldots, A_k)$ is a 1–complex in $\mathcal{Z}$ of minimal length with the property that the quotient of $\lambda' \cup A_1 \cup \cdots \cup A_k$ obtained by collapsing each A_i to a point is connected. Minimality again implies that this collapsed graph is a tree. For convenience we assume that $\lambda(\{x_1, \ldots, x_k\}) = \lambda'(\{x_1\}, \ldots, \{x_k\})$.

The following lemma illustrates a basic property of hyperbolic spaces, and we omit its proof.

Lemma 3.1 *Let $\mathcal{Z}$ be a geodesic δ–hyperbolic space and λ a minimal network function as above. Then there exists $\epsilon_0 = \epsilon_0(k, \delta)$ such that, for all $\epsilon \geq \epsilon_0$, there exists $\epsilon' > \epsilon$ such that if $F \subset \mathcal{Z}$ has cardinality k then*

- *there is a $(1, \epsilon/2)$–quasi-isometry $\lambda(F) \to \mathrm{hull}(F)$ which is $\epsilon/2$–far from the identity;*

- *for any two points $x, y \in \mathcal{N}_\epsilon(\lambda(F))$, any geodesic joining them is in $\mathcal{N}_{\epsilon'}(\lambda(F))$.*

In the rest of this section we consider the following situation. Let a (large but) finite set $\mathcal{Y} \subset \mathcal{N}_{\epsilon/2}(\mathrm{hull}(F))$ be given (see Section 4 for what $\mathcal{Y}$ will be in our setting). It is possible to divide $\lambda(F)$ up into a union of subtrees some of which are close approximations to "clusters" in $\mathcal{Y}$ and the rest interconnect the clusters, but such a construction is not unique, depending on many choices (including the choice of $\lambda(F)$ itself). Our goal in this section is to describe a version of this which is stable, in the sense that small changes in the sets F and $\mathcal{Y}$ only alter the tree and its subtrees in a controlled way — independently of the diameter of F or the cardinality of $\mathcal{Y}$.

Remark 3.2 For convenience in our discussion we allow ourselves to assume that the points of F are all leaves of $\lambda(F)$. This can be arranged by a slight perturbation, or by considering each point of F as the endpoint of an additional edge of length 0.

Given $E \gg \epsilon$, let $\mathcal{C}_E(\mathcal{Y} \cup F)$ be the graph whose edges connect points of $\mathcal{Y} \cup F$ that are at most E apart. Vertex sets of connected components of $\mathcal{C}_E$ are called *clusters*.

We will choose E to be a suitably large multiple of ϵ'. We note that the relation of E to ϵ' and ϵ is the one sensitive part of the argument, and elsewhere we can be content with order-of-magnitude arguments.

For a simplicial tree T, let $d(v)$ be the valence of each vertex v and let $k(T)$ be the number of leaves, ie vertices of valence 1. We have $\sum_{d(v)>2}(d(v)-2)=k(T)-2$, for example by an Euler characteristic argument. We call this quantity the *total branching* of T.

The following theorem is the main result of this section.

Theorem 3.3 *Given k, N, δ, and $\epsilon \geq \epsilon_0(k,\delta)$ as in [Lemma 3.1](#) there exists $K > 0$ such that the following holds. Let $\mathcal{Z}$ be a geodesic δ–hyperbolic space and let $F, \mathcal{Y} \subset \mathcal{Z}$ be finite subsets, where $|F| \leq k$ and $\mathcal{Y} \subset \mathcal{N}_{\epsilon/2}(\mathrm{hull}(F))$.*

There exists a metric tree $T = T(F, \mathcal{Y})$ with a decomposition into two forests $T = T_c \cup T_e$ intersecting along a finite set of points and a map $\Xi = \Xi_{F,\mathcal{Y}} \colon T(F, \mathcal{Y}) \to \mathcal{Z}$ such that:

(a) *The total branching of T is bounded by $2k-4$.*

(b) *Ξ is a (K, K)–quasi-isometric embedding with image K–Hausdorff close to $\mathrm{hull}(F)$.*

(c) *For each component τ of T_e we have that $\Xi|_\tau$ is a $(1, K)$–quasi-isometric embedding, and an isometry onto $\Xi(\tau)$ endowed with its path metric.*

(d) *There is a bijection b between components of T_c and clusters in $\mathcal{C}_E(\mathcal{Y} \cup F)$ such that $\Xi(\tau)$ is K–Hausdorff close to $b(\tau)$ for each component τ of T_c.*

Furthermore, if $F', \mathcal{Y}' \subset \mathcal{Z}$ and $g \in \mathrm{Isom}(\mathcal{Z})$ are such that $|F'| \leq k$, $\mathcal{Y}'$ is finite, $d_{\mathrm{Haus}}(gF, F') \leq 1$, and $|g\mathcal{Y} \triangle \mathcal{Y}'| < N$, then there exists a constant $L = L(N, k, \delta) > 0$ and subsets $T_s \subset T_e(F, \mathcal{Y})$ and $T'_s \subset T_e(F', \mathcal{Y}')$ such that, identifying components of $T_e(F, \mathcal{Y})$ and $T_e(F', \mathcal{Y}')$ with their images in $\mathcal{Z}$, we have:

(1) *The components of T_s and T'_s are contained in the edges of $T_e(F, \mathcal{Y})$ and $T_e(F', \mathcal{Y}')$, respectively.*

(2) *The complements $T_e(F, \mathcal{Y}) \setminus T_s$ and $T_e(F', \mathcal{Y}') \setminus T'_s$ have at most L components, each of diameter at most L.*

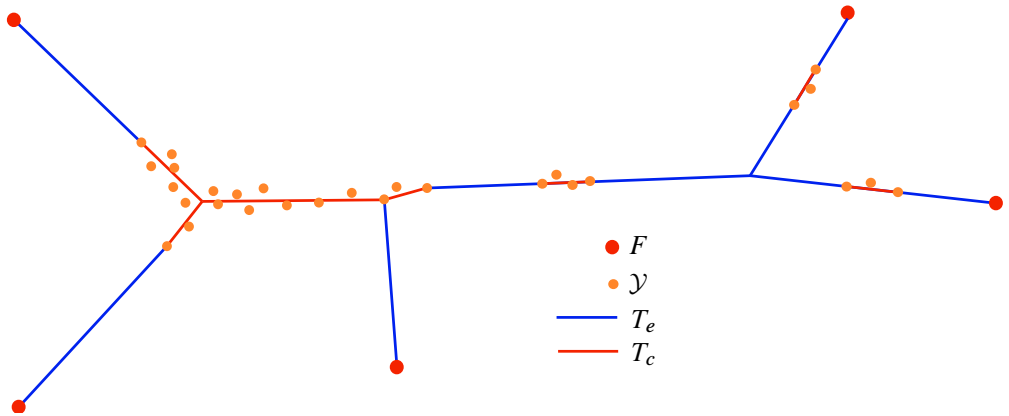(3) *There is a bijective correspondence between the sets of the components of gT_s and T'_s.*

Figure 5: An example of the stable tree $T = T_c \cup T_e$ provided by Theorem 3.3.

(4) *Under this correspondence, all but L components are exactly the same, and the identical components of T_s and T'_s come from the identical components of $T_e(F, \mathcal{Y})$ and $T_e(F', \mathcal{Y}')$.*

(5) *The remaining L components of gT_s are each at Hausdorff distance L to the corresponding component in T'_s.*

We call the trees $T(F, \mathcal{Y})$ *stable trees*.

Remark 3.4 (coarse equivariance and its proof) The "furthermore" part of Theorem 3.3 can be interpreted as simultaneously stating two facts. For g the identity, it says that the trees are stable under perturbations of F and $\mathcal{Y}$. Alternatively, for $F' = gF$ and $\mathcal{Y}' = g\mathcal{Y}$, it says that the construction is coarsely equivariant. In either case, what we have to prove is essentially the following. The construction relies on certain choices, namely the choices of functions λ and λ' as above, and we have to show that these only cause the kinds of perturbations described in the statement of the theorem. From this perspective, it is clear that the proof for a general g is the same as that for $g = 1$, as $gT(F, \mathcal{Y})$ coincides with the tree $T(gF, g\mathcal{Y})$ constructed based on different choices. To save notation and make the proof more readable, we will prove only the case where g is the identity.

3.1 Cluster separation graph

Let $\mathcal{C}_E = \mathcal{C}_E(F \cup \mathcal{Y})$ be as above and let $C_1, C_2, C_3 \subset \mathcal{C}_E^0$ be clusters (ie vertex sets of connected components). We say that C_2 *separates* C_1 *from* C_3 *in* $\mathcal{Z}$ if there

exists a minimal $\mathcal{Z}$–geodesic segment σ with endpoints on C_1 and C_3 which meets the $2\epsilon'$–neighborhood of C_2.

Definition 3.5 Let $\mathcal{G}_E = \mathcal{G}_E(F \cup \mathcal{Y})$ be a graph whose vertex set $\mathcal{G}_E^0$ is the set of clusters of $\mathcal{C}_E$, and where $[C_1, C_2]$ is an edge whenever there is no cluster separating C_1 from C_2 in $\mathcal{Z}$. We call $\mathcal{G}_E$ the *separation graph* for $\mathcal{C}_E$.

Lemma 3.6 *If $E > 4\epsilon'$, then $\mathcal{G}_E$ is connected.*

Proof Let $C, C' \in \mathcal{G}_E^0$. If $C \neq C'$, then $d_{\mathcal{Z}}(C, C') > E$ (here $d_{\mathcal{Z}}$ means the minimal distance, not the diameter of the union). If they are not adjacent in $\mathcal{G}$, then there is a third cluster B separating them in $\mathcal{Z}$. Let σ be a minimal geodesic connecting C to C' with $p \in \sigma$ within $2\epsilon'$ of B. Then p is distance at least $E - 2\epsilon'$ from each end of σ since B is at least E from both C and C'. It follows that $d_{\mathcal{Z}}(B, C) \leq d_{\mathcal{Z}}(C, C') - E + 4\epsilon'$, and similarly for $d_{\mathcal{Z}}(B, C')$.

If $d_{\mathcal{Z}}(C, C') \leq 2E - 4\epsilon'$ this gives $d_{\mathcal{Z}}(B, C) \leq E$ which is a contradiction, so C and C' must be connected by an edge in $\mathcal{G}_E$. For $d_{\mathcal{Z}}(C, C') > 2E - 4\epsilon'$, we have that $d_{\mathcal{Z}}(B, C)$ and $d_{\mathcal{Z}}(B, C')$ are smaller than $d_{\mathcal{Z}}(C, C')$ by at least $E - 4\epsilon'$, so we can proceed inductively. $\square$

For ease of notation, set $\mathcal{C} = \mathcal{C}_E$ and $\mathcal{G} = \mathcal{G}_E$.

Definition 3.7 For any subset A of $\mathcal{Z}$, let its *shadow* $s(A)$ be the subtree of $\lambda(F)$ obtained by taking the convex hull (in $\lambda(F)$) of all the points in $\lambda(F)$ within distance ϵ from points of A. For a singleton $\{x\}$ we also write $s(x) := s(\{x\})$.

Note that, since $\mathcal{Y} \cup F$ is in $\mathcal{N}_\epsilon(\lambda(F))$ by hypothesis, $s(C) \neq \varnothing$ for any nonempty subset $C \subset \mathcal{Y} \cup F$.

The rest of this subsection is devoted to establishing several properties of shadows which will connect the separation properties of clusters in $\mathcal{G}$ to separation properties of their shadows in $\lambda(F)$, thereby allowing us to work with $\mathcal{G}$ and independently of $\lambda(F)$.

The next lemma controls how and when shadows of clusters can intersect.

Lemma 3.8 *Let $E > 7\epsilon$ and let $C, C' \in \mathcal{G}^0$ be distinct clusters. Then*

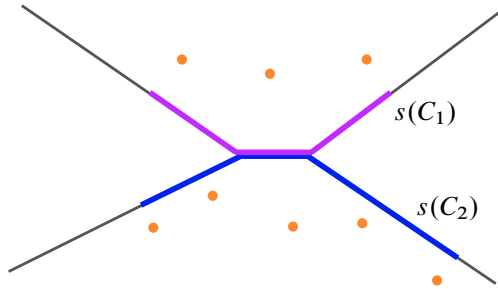(1) *$s(C) \cap s(C')$ can contain no leaf of $s(C)$ or $s(C')$;*

Figure 6: When there is branching, shadows can overlap in their interiors, but never at their leaves.

(2) the diameter of $s(C) \cap s(C')$ is bounded by a constant depending on $\#F$, E, and ϵ;

(3) if at least one of $s(C)$ and $s(C')$ is an interval along an edge of $\lambda(F)$, then $s(C) \cap s(C') = \varnothing$.

Proof Note first that for any $x \in C$, $s(x)$ is a subtree of diameter at most 3ϵ. This is because any two extreme points of $s(x)$ are within ϵ of x, and $\lambda(F)$ is $(1, \epsilon)$–quasi-isometrically embedded. Similarly, for any $x, y \in C$,

$$\mathrm{diam}_{\lambda(F)}(s(\{x, y\})) \leq d(x, y) + 3\epsilon.$$

Claim 1 For every $p \in s(C)$, there exists $q \in s(C)$ at distance (in $\lambda(F)$) at most $(E + 3\epsilon)/2$ such that $d(q, C) \leq \epsilon$.

Proof Either $p \in s(x)$ for some $s(x)$ containing an extreme point of $s(C)$, or p separates some $s(x)$ from $s(y)$, for $x, y \in C$. In the first case p is within $3\epsilon/2$ of a point q for which $d(q, x) \leq \epsilon$ and we are done. In the second case, a path in $\mathcal{C}_E$ from x to y then yields a sequence of points $x_i \in C$ such that $d(x_i, x_{i+1}) \leq E$ and p is contained in one of the shadows $s(\{x_i, x_{i+1}\})$. Since $\mathrm{diam}_{\lambda(F)}(s(\{x_i, x_{i+1}\})) \leq E + 3\epsilon$, we find that p is within $(E + 3\epsilon)/2$ of an extreme point q of $s(\{x_i, x_{i+1}\})$, so $d(q, x_i) \leq \epsilon$ or $d(q, x_{i+1}) \leq \epsilon$. The claim follows. $\square$

For (1), suppose that a leaf p of $s(C)$ is in $s(C')$. Note that the leaves of $s(C)$ and $s(C')$ are within ϵ of C and C', respectively. By the previous paragraph, there is a point q of $s(C')$ within $(E + 3\epsilon)/2$ of p which is ϵ close to C' Thus we obtain $d(C, C') \leq 2\epsilon + (E + 3\epsilon)/2 < E$, so $C = C'$.

For (2), suppose that $s(C) \cap s(C')$ contains an edge e of length greater than $2(E + 3\epsilon)$. Claim 1 implies that there is a set R in $s(C)$ consisting of points at distance ϵ from
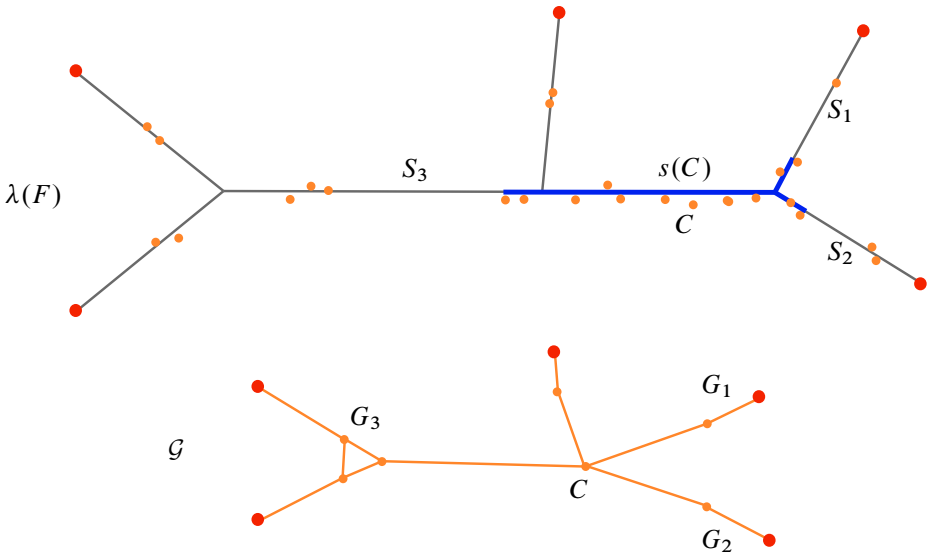
Figure 7: The cluster C is guaranteed valence at least 3 (via S_1, S_2 and S_3) by Lemma 3.10. In this case it has valence 4.

C and whose $(E + 3\epsilon)/2$–neighborhood covers $s(C)$; there is also a similar set R' in $s(C')$. Since e is in both shadows, it must be that $e \cap R$ and $e \cap R'$ both cut e into intervals of length at most $E + 3\epsilon$. Thus it must be that there is a point $r \in R \cap e$ and $r' \in R' \cap e$ that are distance $(E + 3\epsilon)/2$ apart. Then just as before we obtain $d(C, C') < E$ so $C = C'$. Now the number of edges in $s(C) \cap s(C')$ is bounded by the total branching of the tree, which depends on $\#F$. This gives (2).

Finally, for (3), if one of $s(C)$ and $s(C')$ is an interval contained in an edge of $\lambda(F)$ then it is easy to see that, if they overlap, then one must contain a leaf of the other, thereby violating (1). □

Definition 3.9 From now on we set $E = 8\epsilon'$, so that the conclusions of both Lemmas 3.6 and 3.8 hold.

The following lemma connects the separation properties in $\mathcal{G}$ of a cluster to the separation properties in $\lambda(F)$ of its shadow.

Lemma 3.10 *Let C be a cluster and $S_1, \ldots, S_m$ be the components of $\lambda(F) \backslash \mathrm{int}(s(C))$ which meet $s(C)$ at a leaf of $s(C)$. Let G_i be the set of clusters $B \in \mathcal{G}^0 \backslash \{C\}$ such that $s(B) \cap S_i \neq \varnothing$. Then each G_i is in a distinct component of $\mathcal{G} \backslash C$, and moreover the valence of C in $\mathcal{G}$ at least m.*
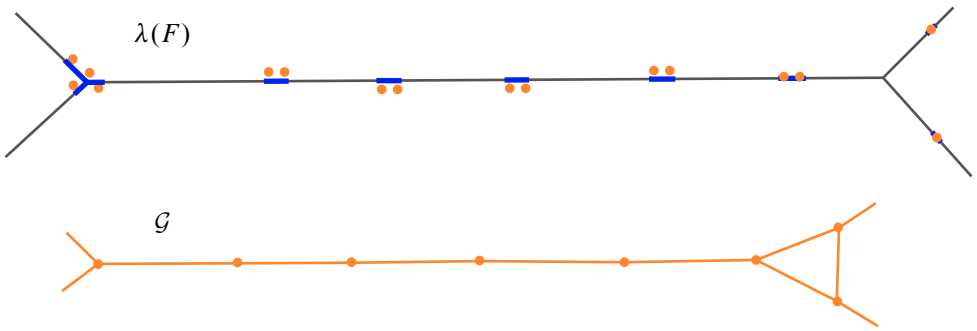
Figure 8: Clusters with shadows on an edge in $\lambda(F)$ give rise to a path of bivalent vertices in $\mathcal{G}$.

Proof Note that if B is a cluster in G_i then $s(B)$ is actually disjoint from $s(C)$, since the leaves of $s(C)$ cannot meet $s(B)$ by Lemma 3.8. Moreover, there may be clusters $C' \neq C$ that are not in any G_i; their shadows meet components of $\lambda(F) \setminus \text{int}(s(C))$ that do not meet leaves of $s(C)$.

Let $A \in G_i$ and $B \notin G_i \cup \{C\}$. A minimal geodesic σ in $\mathcal{Z}$ connecting A to B must be ϵ'–close to the path in $\lambda(F)$ connecting $s(A)$ to $s(B)$, and this path passes through a leaf p of $s(C)$ (namely $s(C) \cap S_i$). Thus there is a point of C within $\epsilon + \epsilon' < 2\epsilon'$ of σ, so C separates A from B in $\mathcal{Z}$. In particular A and B cannot be adjacent in $\mathcal{G}$.

Thus G_i cannot be connected to any vertex in $\mathcal{G}^0 \setminus (G_i \cup \{C\})$, which implies distinct G_i are in distinct components of $\mathcal{G} \setminus C$.

To see that the valence is at least m, we must check that each G_i is nonempty. But each S_i must contain a leaf of $\lambda(F)$, which is a point of F, so there must be a cluster whose shadow is in S_i. $\square$

Lemma 3.11 *If e is an edge of $\lambda(F)$, the clusters C whose shadows $s(C)$ are sub-intervals of e form a path in $\mathcal{G}$ whose interior vertices are bivalent. The ordering of this path matches the ordering of the shadows in e.*

Proof Let $\{C_1, \ldots, C_l\}$ be the set of clusters whose shadows are subintervals of e. By Lemma 3.8, $s(C_i) \cap s(C_j) = \varnothing$ for all i and j. We may therefore assume that their indices correspond to the order they appear along e in $\lambda(F)$.

The complement $\lambda(F) \setminus \text{int}(s(C_i))$ has two components for each i, labeled A_i^- and A_i^+, such that A_i^- contains $s(C_{i-1})$ when $i > 0$ and A_i^+ contains $s(C_{i+1})$ when $i < l$.

By our ordering, no shadows lie between $s(C_i)$ and $s(C_{i+1})$. Lemma 3.10 implies that C_i separates (in $\mathcal{G}$) the clusters whose shadows lie in A_i^- from those in A_i^+. In particular, no B can separate C_i from C_{i+1} in $\mathcal{G}$, so they are adjacent and we obtain a path $C_1, \ldots, C_l$ in $\mathcal{G}$. Moreover for $1 < i < l$ we can see that C_i is bivalent as follows: if $D \in \mathcal{C} \setminus \{C_{i-1}, C_i, C_{i+1}\}$, then one of C_{i-1} or C_{i+1} separates D from C_i in $\mathcal{G}$, again by Lemma 3.10, and so there can be no edge $[C_i, D]$ and the valence of C_i is exactly 2. $\qquad\square$

Lemma 3.12 *If C has valence 2 in $\mathcal{G}$ but $s(C)$ is not an interval inside an edge of $\lambda(F)$, then C contains a point of F.*

Proof If $s(C)$ is not an interval in an edge of $\lambda(F)$, it has a branch point and hence at least three leaves. At most two of these can be interior to $\lambda(F)$, because otherwise C would have valence at least 3 in $\mathcal{G}$ by Lemma 3.10.

Thus $s(C)$ contains a leaf q of $\lambda(F)$, which is a point of F. This means $d(q, C) \leq \epsilon < E$ (notice that, since q is a leaf, it lies in the convex hull of a subset of $\lambda(F)$ only if it lies in the subset). Hence, we have $q \in C$. $\qquad\square$

Structure of bivalent clusters Let $\mathcal{E}^0$ denote the set of clusters $C \in \mathcal{G}^0$ which have valence 2 in $\mathcal{G}$ and do not contain a point of F. Lemma 3.12 implies that each $C \in \mathcal{E}^0$ has shadow inside an edge of $\lambda(F)$.

The next lemma gives that almost all clusters are bivalent.

Lemma 3.13 $$\#(\mathcal{G}^0 \setminus \mathcal{E}^0) \leq 2k - 2.$$

Proof For a cluster $C \in \mathcal{G}^0 \setminus \mathcal{E}^0$, either C contains a point of F, or $s(C)$ contains a branch point of $\lambda(F)$. There are at most k clusters of the former type. The number of clusters of the latter type is bounded by the total branching $b(\lambda(F))$, but to show this we must contend with the fact that shadows can overlap.

Let $W \subset \lambda(F)$ be a connected union of shadows $s(C_1), \ldots, s(C_m)$, each of which contains a branch point. By Lemma 3.8, no leaf of $s(C_i)$ can be in $s(C_j)$ for $i \neq j$. Hence all leaves of $s(C_i)$ must be leaves of W and disjoint from each other. Since each $s(C_i)$ has at least two leaves,

$$m \leq \left\lfloor \frac{k(W)}{2} \right\rfloor = \left\lfloor \frac{b(W) + 2}{2} \right\rfloor,$$

where $k(W)$ is the number of leaves and $b(W)$ is the total branching of W. Since $b(W) \geq 1$, this implies $m \leq b(W)$. Summing over all such W we find that the number of clusters with branch points in their shadows is bounded by $b(\lambda(F))$, or $k - 2$. The desired inequality follows. □

Let $\mathcal{E}$ be the subgraph of $\mathcal{G}$ induced on the vertices $\mathcal{E}^0$.

Lemma 3.14 *Let $\mathcal{E}_1, \ldots, \mathcal{E}_m$ be the components of $\mathcal{E}$. For each $\mathcal{E}_i$ there is an edge e_i of $\lambda(F)$ such that $\mathcal{E}_i$ is a path $C_1, \ldots, C_{r_i}$ in $\mathcal{G}$ consisting of all elements of $\mathcal{E}$ whose shadows lie in the interior of e_i; the edges e_i are distinct.*

Proof Since each cluster $D \in \mathcal{E}_i$ is a bivalent vertex of $\mathcal{G}$ with shadow in an edge of $\lambda(F)$ by [Lemma 3.12](), and [Lemma 3.11]() implies that all such clusters with shadows on a given edge $e \in \lambda(F)$ form a path in $\mathcal{G}$, it suffices to prove that no two such edge paths of bivalent clusters in $\mathcal{G}$ are directly connected by an edge.

Suppose $C, D \in \mathcal{E}_i$ are connected by an edge in $\mathcal{G}$ but $s(C)$ and $s(D)$ are not contained in a single edge of $\lambda(F)$. Since $s(C) \cap s(D) = \varnothing$, we may label the components of $\lambda(F) \setminus s(C)$ and $\lambda(F) \setminus s(D)$ by $\gamma_\pm$ and $\delta_\pm$, respectively, so that $s(C) \subset \delta_-$ and $s(D) \subset \gamma_-$. Then the intersection $\gamma_- \cap \delta_-$ contains a vertex v of $\lambda(F)$ of valence at least 3.

By [Lemma 3.10](), $\mathcal{G} \setminus C$ is divided into subgraphs $\mathcal{G}(\gamma_\pm)$ spanned by clusters whose shadows are in $\gamma_\pm$, respectively, and are separated by C, and similarly $\mathcal{G}(\delta_\pm)$ are separated by D, respectively. In particular note $C \in \mathcal{G}(\delta_-)$ and $D \in \mathcal{G}(\gamma_-)$.

Since v has valence at least 3, there is a component of $\lambda(F) \setminus \{v\}$ that meets neither $s(C)$ or $s(D)$. A leaf of this component is in the shadow of a cluster B which is therefore in $\mathcal{G}(\gamma_-) \cap \mathcal{G}(\delta_-)$.

Since $\mathcal{G}$ is connected, B is connected to C within $\mathcal{G}(\gamma_-)$ and to D within $\mathcal{G}(\delta_-)$. Since C and D are bivalent and by hypothesis adjacent in $\mathcal{G}$, the edge between them is the only edge connecting C to $\mathcal{G}(\gamma_-)$, and the only edge connecting D to $\mathcal{G}(\delta_-)$. Hence any path from B to C must pass through this edge and must therefore meet D first. Reversing the roles of C and D we obtain a contradiction. □

3.2 Constructing the stable tree

In this section, we construct our stable tree $T(F, \mathcal{Y})$ from the structure of $\mathcal{G}$ without referring to $\lambda(F)$ directly. In [Proposition 3.17]() below, we prove it is quasi-isometric to $\lambda(F)$.
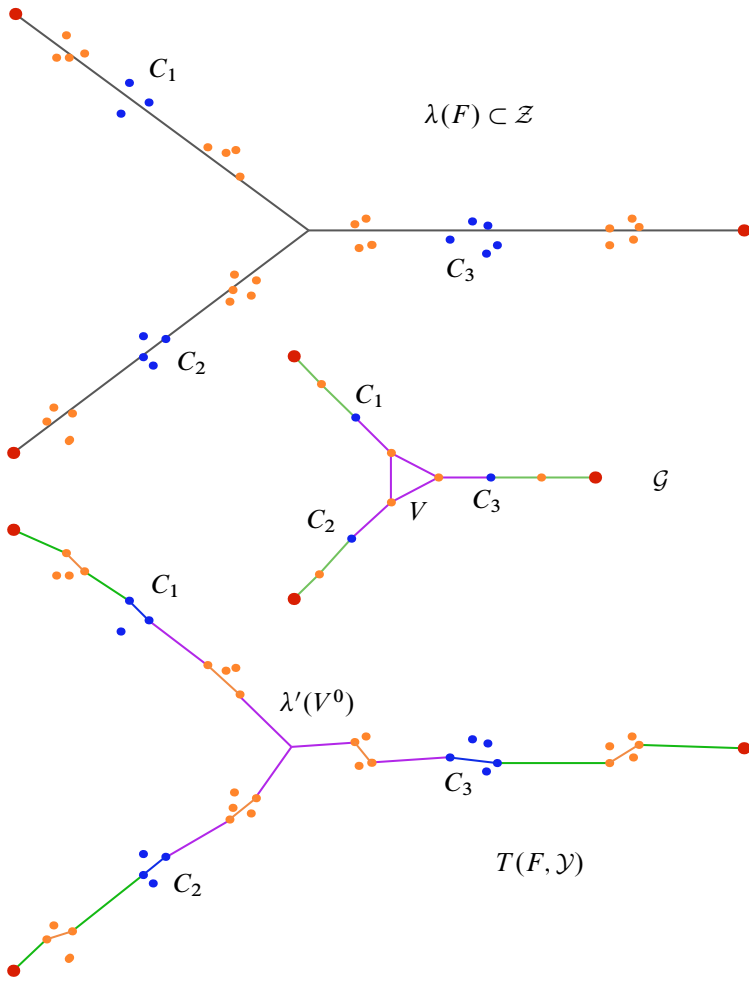
Figure 9: The construction of a stable tree. Note that each complementary component of $\mathcal{G} \setminus \mathcal{E}^0$ determines multiple components of T_e, eg the lavender forest $\lambda'(V^0)$ determined by the component V whose boundary is the bivalent clusters C_1, C_2, and C_3. Each cluster C then determines a single component $\mu(C)$ of T_c by connecting the points $r(C) = C \cap (T_e \cup F)$.

The two forests   Now let us proceed to define the forests $T_c(F, \mathcal{Y})$ and $T_e(F, \mathcal{Y})$. We let $\mathcal{G} = \mathcal{G}_E(F \cup \mathcal{Y})$ be as above.

Let $\mathcal{V}$ denote the set of closures of connected components of $\mathcal{G} \setminus \mathcal{E}^0$. Thus each element of $\mathcal{V}$ is a subgraph connected to the rest of $\mathcal{G}$ along vertices in $\mathcal{E}^0$. For each $V \in \mathcal{V}$ let V^0 denote its vertex set, which is a collection of clusters. Some elements of $\mathcal{V}$ are

single edges $[C, D]$, where $C, D \in \mathcal{E}^0$, and others are subgraphs containing vertices in $\mathcal{G}^0 \setminus \mathcal{E}^0$, and we note they are not necessarily trees, though Lemma 3.13 bounds their size.

For each $V \in \mathcal{V}$, let $\lambda'(V^0)$ be the minimal network defined in the beginning of this section, where V^0 is interpreted as a collection of subsets (clusters) in $\mathcal{Z}$.

Now define
$$T_e = T_e(F, \mathcal{Y}) = \bigsqcup_{V \in \mathcal{V}} \lambda'(V^0).$$

Remark 3.15 The forest T_e is a disjoint union of copies of forests each contained in $\mathcal{Z}$. It is important to note, however, that these trees might in fact intersect in $\mathcal{Z}$. With a slight abuse, we will conflate the abstract copies of the $\lambda'(V^0)$ that constitute T_e and their "concrete" counterparts in $\mathcal{Z}$. Similar comments apply to T_c below. Since the map $\Xi \colon T \to \mathcal{Z}$ is just going to be the identity on all the components of T_e and T_c, we will allow ourselves to regard T as a subset of $\mathcal{Z}$ for purposes that do not require understanding the metric of T, eg when measuring the Hausdorff distance between (the image in $\mathcal{Z}$ of) a subset of T and a subset of $\mathcal{Z}$.

Note that T_e is a forest whose leaves are points of clusters.

Collapsing clusters to points, T_e becomes a connected network N, by the definition of $\mathcal{V}$. This connected network is a union of trees joined at points that correspond to vertices of $\mathcal{E}^0$. Since any vertex in $\mathcal{E}^0$ disconnects $\mathcal{G}$, each of these join points disconnects N, so we see that N is a tree.

Now for each cluster $C \in \mathcal{G}^0$, we consider the set of points $r(C) = C \cap (T_e \cup F)$. We let $\mu(C)$ denote the tree $\lambda(r(C))$, and define
$$T_c = T_c(F, \mathcal{Y}) = \bigsqcup_{C \in \mathcal{G}^0} \mu(C).$$

The tree We now define $T(F, \mathcal{Y}) = T_c(F, \mathcal{Y}) \cup T_e(F, \mathcal{Y})$, or $T = T_c \cup T_e$ for short. Note that T is a tree because as above collapsing the subtrees of T_c to points yields a tree; see Figure 9.

Lemma 3.16 Let $T = T(F, \mathcal{Y}) = T_c \cup T_e$.

(1) The total branching $b = b(T)$ is bounded by $2k - 4$, and the leaves of T are contained in $F \cup \mathcal{Y}$.

(2) $\mu(C) \subseteq \mathcal{N}_{O(\epsilon)}(C)$, so $T_c \subset \mathcal{N}_{O(\epsilon)}(\mathcal{Y} \cup F)$.

(3) For all $p \in T_e$, we have $d_{\mathcal{Z}}(p, \mathcal{Y} \cup F) \geq \frac{1}{b} d_T(p, \partial T_e) - O(\epsilon)$.

Proof To bound $b(T)$, we bound the number of leaves T can have. Leaves of T are leaves of the various components of T_e and T_c, and thus can arise in two ways:

(a) If a cluster C contains points of F, the points in $C \cap F$ can be leaves of $\mu(C)$ which are also leaves of T. There are at most k such points.

(b) If a cluster C contains no points of F and a single point q of ∂T_e, which is connected to only one subtree of T_e, then q is a leaf of T; see Figure 10.

All other vertices of $\partial T_e \cup \partial T_c$ have valence at least 2. Notice that we already showed that all leaves of T are contained in $F \cup \mathcal{Y}$.

Clusters of type (b) must be in $\mathcal{G}^0 \setminus \mathcal{E}^0$ since every cluster in $\mathcal{E}^0$ belongs to two subgraphs in $\mathcal{V}$, and hence either has two points in ∂T_e or two subtrees of T_e meeting at a single point. The number of clusters in $\mathcal{G}^0 \setminus \mathcal{E}^0$ that don't contain points of F was bounded in Lemma 3.13 by $k - 2$.

This gives us a bound of $2k - 2$ on the total number of leaves in T, which bounds the total branching by $2k - 4$. This proves part (1).

Now for part (2), consider the minimal network $\mu(C)$ for the cluster C. By Lemma 3.1 and the definition of shadows, $\mu(C)$ is within $O(\epsilon)$ of the shadow of C in $\lambda(F)$, and it follows from Claim 1 (in the proof of Lemma 3.8) that every point of $s(C)$ is within $O(E)$ of C. This proves part (2).

For part (3), let $p \in \lambda'(V^0) \subset T_e$, where $V \in \mathcal{V}$, and let the distance $d_{\mathcal{Z}}(p, \mathcal{Y} \cup F)$ be realized on a point in a cluster C_1. Write $d_{\mathcal{Z}}(p, C_1) = t$.

Suppose first that $C_1 \in V^0$. The quotient of $\lambda'(V^0)$ obtained by collapsing the clusters of V^0 to points is a tree by minimality of the network, so there is some sequence of components of $\lambda'(V^0)$ which connects p to C_1, possibly through clusters $C_2, \ldots, C_l \in V^0$.

Consider the unique path α in $\lambda'(V^0)$ from p to C_2. The path α branches at no more than $b = b(T)$ points, so let $\alpha' \subset \alpha$ be the longest unbranched subsegment of α. We thus have $|\alpha'| \geq \frac{1}{b} d_T(p, \partial \lambda'(V^0))$. If $|\alpha'| > t$, we may remove α' from $\lambda'(V^0)$, attach a minimal length path in $\mathcal{Z}$ from p to C_1 (of length t), and obtain a network with smaller total length than $\lambda'(V^0)$ that still connects the clusters in V. This would violate the
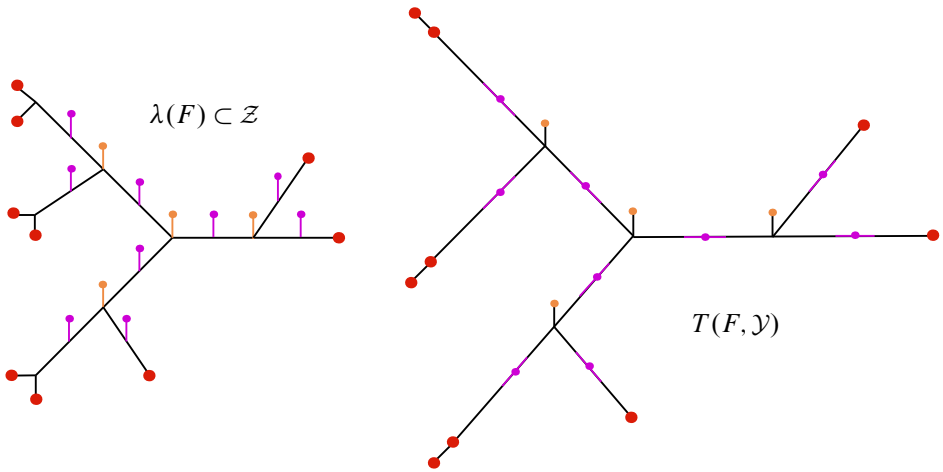
Figure 10: The stable tree $T = T(F, \mathcal{Y})$ may have leaves which are not points in F, and some points of F may not be leaves of T. In this example, the ambient space $\mathcal{Z}$ is the whole graph on the left, and T is realized geometrically on the right. The orange cluster points atop spikes at the branch points of the underlying tree create leaves in T. New leaves in T always arise from clusters near branch points of $\lambda(F)$. The pink bivalent cluster points determine bivalent vertices in T, with small neighborhoods thereof folding into the spikes upon inclusion of $T \to \mathcal{Z}$. By contrast, $\lambda(F)$ contains none of the spikes. Finally, the pairs of nearby points of F in $\lambda(F)$ on the left side of $\mathcal{Z}$ form clusters. The components of T_e connect one point from each pair to a pink cluster, while a component of T_c connects the pair. As a result, some points of F are not leaves of T.

minimality of $\lambda'(V^0)$, so we must have $t \geq |\alpha'|$, and therefore $t \geq \frac{1}{b} d_T(p, \partial \lambda'(V^0))$, as required.

Now consider the possibility that C_1 is a cluster outside of V. Let $s(V)$ denote the shadow of the union of clusters $s\left(\bigcup_{A \in V^0} A\right)$, which is the same as the hull in $\lambda(F)$ of the shadows $\{s(A) \mid A \in V^0\}$. We claim that $s(V) \cap s(C) = \varnothing$ for every $C \in \mathcal{G}^0 \setminus V^0$.

Recall from Lemmas 3.8 and 3.10 that the shadow $s(C)$ for each $C \in \mathcal{E}^0$ is disjoint from all other cluster shadows, and that the separation of shadows by $s(C)$ in $\lambda(F)$ is the same as the separation of the corresponding vertices in $\mathcal{G}$ by $C \in \mathcal{G}^0$. In particular, if $V \in \mathcal{V}$ and $C \in \mathcal{E}^0$ then all vertices $D \in V^0$ (other than C itself if C happens to lie in V^0) have shadows $s(D)$ on one side of $s(C)$. Any $V_1 \neq V_2$ in $\mathcal{V}$ are separated in $\mathcal{G}$ by some $C \in \mathcal{E}^0$, including the case when C is the common vertex of V_1 and V_2.
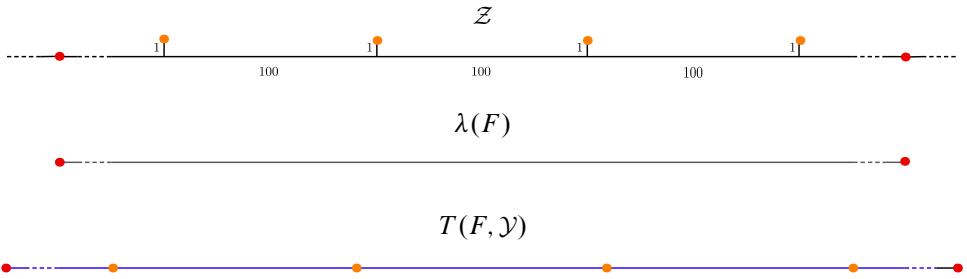
Figure 11: A basic, complicating example. The ambient space $\mathcal{Z}$ is a bi-infinite line with spikes of height 1 at distance 100 from each other, which we think of as 10–hyperbolic. The cluster points $\mathcal{Y}$ (in orange) sit on the ends of the spikes. The two points in F are very far apart. The spanning tree $\lambda(F)$ for F is a long segment. The stable tree $T(F, \mathcal{Y})$ is also abstractly a long concatenation of segments of length 102. The natural map $T(F, \mathcal{Y}) \to \mathcal{Z}$ folds the ends of these segments onto the spikes and is therefore not an embedding. It is a quasi-isometric embedding, but the multiplicative constant is at least $102/100$.

Thus the shadows $s(V_1)$ and $s(V_2)$ are either disjoint or overlap exactly on $s(C)$ for this common vertex C. The claim follows.

Now applying Lemma 3.1 again we find that $\lambda'(V^0)$ is in an $O(\epsilon)$ neighborhood of $s(V)$. Thus if C_1 is a cluster in $\mathcal{G}^0 \setminus V^0$ then any $\mathcal{Z}$–geodesic from p to C_1 has an $O(\epsilon)$ fellow-traveling path in $\lambda(F)$ which must exit $s(V)$ before it arrives at $s(C_1)$. It follows that $d(p, C_1) \geq d(p, C_2) - O(\epsilon)$, for some $C_2 \in V^0$. This reduces to the previous case. □

We are now ready to prove that our stable tree $T(F, \mathcal{Y})$ coarsely behaves like $\lambda(F)$. Unlike Gromov's trees, stable trees quasi-isometrically embed with multiplicative constants possibly larger than 1; see Figure 11. This is an inconvenient fact for what follows and later in Section 4.

Proposition 3.17 *The natural map $T(F, \mathcal{Y}) \to \mathcal{Z}$ is a (K_1, K_1)–quasi-isometric embedding, and $T(F, \mathcal{Y})$ lies within Hausdorff distance K_1 of $\lambda(F)$, where $K_1 = K_1(k, \delta, \epsilon)$.*

Proof It follows from Lemma 3.1 that each component of T_c and T_e is $(1, O(\epsilon))$–quasi-isometric to its shadow in $\lambda(F)$, and moreover is within Hausdorff distance $O(\epsilon)$ of its shadow in $\lambda(F)$.

Consider two distinct clusters $C_1, C_2 \in \mathcal{G}^0$, and their shadows. By Lemma 3.8(2), the shadow intersection $s(C_1) \cap s(C_2)$ has uniformly bounded diameter. If C_1 and C_2 belong to different pieces $V_1, V_2 \in \mathcal{V}$, then there is a cluster $D \in \mathcal{E}^0$ separating V_1 from V_2 in $\mathcal{G}$. If D is not equal to either C_i then its shadow separates $s(C_1)$ from $s(C_2)$ by Lemma 3.10, and hence the shadows are disjoint. If $C_1 = D$, say, then again the shadows are disjoint, by Lemma 3.8(3).

In particular, the clusters in $\mathcal{E}^0$ have pairwise disjoint shadows, and moreover by Lemma 3.10 their separation properties in the graph $\mathcal{G}$ are preserved in $\lambda(F)$ — that is, if C_2 separates C_1 from C_3 in $\mathcal{G}$ then $s(C_2)$ separates $s(C_1)$ from $s(C_3)$ in $\lambda(F)$. This means that any $V \in \mathcal{V}$ is associated to a complementary component $c(V)$ of the shadows of $\mathcal{E}^0$ in $\lambda(F)$, in the following way. For every $C \in \mathcal{E}_0$, any cluster in $V^0 \setminus \{C\}$ has shadow contained in one of the two components of $\lambda(F) \setminus s(C)$, by Lemma 3.10. We let $c(V)$ be the intersection of all these components. Notice that if $V \neq V'$ then $c(V) \neq c(V')$, since in that case some $C \in \mathcal{E}_0$ will separate V from V' in $\mathcal{G}$.

We now study overlaps of the shadows of the various relevant subtrees of T, showing that said overlaps are bounded.

Let $V \in \mathcal{V}$ be the closure of a component of $\mathcal{G} \setminus \mathcal{E}^0$. If λ_1 and λ_2 are distinct components of $\lambda'(V^0)$, we claim that their shadows in $\lambda(F)$ have an intersection of bounded diameter.

Indeed, if the shadows of λ_1 and λ_2 had overlap of size $\gg \epsilon$, then λ_1 would contain points within $O(\epsilon)$ of λ_2, at distance $\gg \epsilon$ from each other, and with no branch point of either λ_1 or λ_2 within $O(\epsilon)$ of the geodesic in λ_1 connecting the two points (this uses the bound on the branching of T). A simple surgery would then reduce the total length of $\lambda'(V^0)$, contradicting its minimality.

Now consider a component λ_1 of $\lambda'(V^0)$ and one of the clusters C in V. We claim their shadows also have bounded-diameter intersection. Lemma 3.16 tells us that any point of λ_1 within d of $\mu(C)$ is within $O(d)$ of the boundary of λ_1. This proves the claim.

Note that the number of clusters in V, and therefore the number of components of $\lambda'(V^0)$, is bounded via Lemma 3.13. Thus the subtree T^V comprising $\lambda'(V^0)$ together with all the components of T_c associated to clusters in V has a decomposition into a bounded number of subtrees, and a map to $\lambda(F)$ (using the shadows) which is a $(1, O(\epsilon))$–quasi-isometric embedding on each subtree and such that the images of

distinct subtrees have bounded overlap. Under these circumstances it follows that the map $T^V \to \lambda(F)$ is a (k, k)–quasi-isometric embedding, where k depends on these bounds. Moreover, the image of this map must, up to bounded error, lie in the component of $\lambda(F)$ minus the shadows of those clusters in $\mathcal{E}^0$ that separate V from the rest of $\mathcal{G}$ (by the preservation of separation properties noted above).

It follows that these maps piece together to give a (K_1, K_1)–quasi-isometry, where $K_1 = K_1(k, \delta, \epsilon)$. $\square$

3.3 Proof of Theorem 3.3

Property (a), the branching bound on $T(F, \mathcal{Y})$, was proved in Lemma 3.16.

Property (b), the quasi-isometry, is given by Proposition 3.17.

Property (c) follows from the construction of T_e and Lemma 3.1.

Regarding property (d), we have a natural bijection between components of T_c and clusters by construction, and each component $\mu(C)$ is contained in a controlled neighborhood of the corresponding cluster C by Lemma 3.16, where by controlled we mean that the corresponding constant depends on δ, ϵ, and E. We are left to argue that C lies in a controlled neighborhood of $\mu(C)$. This is equivalent to showing that $s(C)$ lies in a controlled neighborhood of $\mu(C)$. If this was not true then, in view of the bound on the total branching of $\lambda(F)$, we would have that $s(C)$ contains an interval I in an edge of $\lambda(F)$ of length $\gg E$ not contained in a controlled neighborhood of $\mu(C)$. From Proposition 3.17 we know that T lies within controlled Hausdorff distance of $\lambda(F)$, so I is contained in a union of controlled neighborhoods of the $\mu(C')$ for $C' \neq C$, and controlled neighborhoods of the components of T_e. Neighborhoods of the latter type cannot contain points in I far from its endpoints by Lemma 3.16(3), and the same holds true for neighborhoods of the former type in view of Lemma 3.8(1), a contradiction. Therefore, $s(C)$ and C are contained in a controlled neighborhood of $\mu(C)$, as required.

We now prove the "furthermore" part of the statement. Recall that, for the reasons explained in Remark 3.4, we only treat the case that g is the identity. Let $(F', \mathcal{Y}')$ be a second configuration differing from $(F, \mathcal{Y})$ as in the statement. We name the constructions arising from $(F', \mathcal{Y}')$ by $\mathcal{C}', \mathcal{G}', \mathcal{E}', \mathcal{V}'$, etc. Also, we write $T = T(F, \mathcal{Y})$ and $T' = T(F', \mathcal{Y}')$, and similarly for T_c, T_e, T'_c, and T'_e. Set $N = \#(\mathcal{Y} \triangle \mathcal{Y}')$.

Claim 1 *The cardinality $|\mathcal{G}^0 \triangle (\mathcal{G}')^0|$ is bounded in terms of k, δ, and N.*

Proof A cluster C is in the symmetric difference $\mathcal{G}^0 \triangle (\mathcal{G}')^0$ only if it is within E of a point of $F \cup F' \cup (\mathcal{Y} \triangle \mathcal{Y}')$, of which there are at most $2k + 2 + N$. Now each point of a cluster in $\mathcal{C}$ is within ϵ of some point in $\lambda(F)$, and there is a number R depending only on the total branching of $\lambda(F)$ such that among any R points in $\lambda(F)$ within a ball of radius $E + \epsilon$ (in $\lambda(F)$), there must be two which are less than ϵ apart (and the same is true for $\mathcal{C}'$ and $\lambda(F')$). Thus, if there are more than $(2k + 2 + N)R$ elements in $\mathcal{G}^0 \triangle (\mathcal{G}')^0$, then two are closer than E apart, which is a contradiction. $\qquad\square$

Claim 2 *The symmetric difference of the edge sets of $\mathcal{G}$ and $\mathcal{G}'$ has cardinality bounded in terms of k, δ, and N.*

Proof By [Lemma 3.13](#), the maximal valence of any vertex of $\mathcal{G}$ is bounded, and so the number of edges incident to elements of $\mathcal{G}^0 \triangle (\mathcal{G}')^0$ is bounded. Therefore it suffices to consider the case where $C, D \in \mathcal{G}^0 \cap (\mathcal{G}')^0$ with $[C, D]$ an edge in $\mathcal{G}$ but not in $\mathcal{G}'$. This implies there is a $B' \in (\mathcal{G}')^0 \setminus \mathcal{G}^0$ separating C from D when no such cluster in $\mathcal{G}^0$ did so before.

Since B' separates C from D, there is a point $q \in B' \cap ((\mathcal{Y}' \triangle \mathcal{Y}) \cup (F' \triangle F))$ which lies at distance at most ϵ from a $\mathcal{Z}$–geodesic γ joining C and D. The shadow $s(q)$ on $\lambda(F)$ must therefore be in a 10ϵ–neighborhood of the interval in $\lambda(F)$ between $s(C)$ and $s(D)$.

Each such q can only affect a bounded number of such edges (C, D) in this way, because the shadows of edges in each component of $\mathcal{E}$ are arranged sequentially and disjointly along edges of $\lambda(F)$ by [Lemma 3.14](#), and $\mathcal{G}^0 \setminus \mathcal{E}^0$ is bounded (again by [Lemma 3.13](#)). Since there are only boundedly many such q, this bounds the number of edges in the symmetric difference. $\qquad\square$

Let $\pi_0(\mathcal{T})$ denote the set of components of a forest $\mathcal{T}$. By [Claim 2](#), there is a bound on the number of collections of clusters in $\mathcal{V} \triangle \mathcal{V}'$, and this gives us a bound, say K, on $|\pi_0(T_e) \triangle \pi_0(T_e')|$.

Now the components of T_c and T_c' correspond to the elements of $\mathcal{G}^0$ and $(\mathcal{G}')^0$ respectively, whose symmetric difference is bounded by [Claim 1](#). Moreover for each element C of $\mathcal{G}^0 \cap (\mathcal{G}')^0$, the points $r(C)$ used to determine the component of T_c (resp. T_c') associated to C are determined by the components of T_e (resp. T_e') adjacent to C. Thus, together with the bound on $|\pi_0(T_e) \triangle \pi_0(T_e')|$, we obtain a bound on $|\pi_0(T_c) \triangle \pi_0(T_c')|$.

We now increase K in a controlled way a few times, with the result of each step depending only on k, δ, and ϵ.

By item (b), we can increase K to ensure that $d_{\text{Haus}}(T, T') < K$. By Lemma 3.18 (below) we can further assume that $d_{\text{Haus}}(\mathcal{B} \cup T_c \cup F, \mathcal{B}' \cup T'_c \cup F') \leq K$, where $\mathcal{B}$ and $\mathcal{B}'$ are the sets of branch points of T and T'. We have to be careful in using Lemma 3.18 because the sets of leaves of T and T' need not be within bounded Hausdorff distance of each other, since they might contain more than F and F'; see Figure 10. However, we can apply the lemma after slightly modifying T and T' by adding "spikes" (meaning edges attached at T, or T', at one side and having a leaf on the other side) of length, say, 1 to ensure that the sets of leaves of the new trees that we obtain do lie within controlled Hausdorff distance. Such spikes only need to be added close to T_c and T'_c by Lemma 3.16(1), yielding the required Hausdorff distance estimate (indeed, $\mathcal{B} \cup T_c \cup F$ is Hausdorff close to the union of T_c and the set of branch points and leaves of the modification of T, with T_c being Hausdorff close to T'_c in view of property (d), and the latter set being close to the corresponding one for T' by Lemma 3.18).

We can then increase K once more to ensure that $T_c \cup (\mathcal{Y}' - \mathcal{Y})$ and $T'_c \cup (\mathcal{Y} - \mathcal{Y}')$ also lie at Hausdorff distance bounded by K. This can be done since T_c, $\mathcal{Y}$ and T'_c, $\mathcal{Y}'$ are at bounded Hausdorff distance by Lemma 3.16(2). Finally, we also require that $K > K_1$ as in Proposition 3.17.

Now let $\sigma = \sigma(K, \delta)$ be the fellow-traveling constant for $(1, K)$–quasigeodesics with endpoints at distance at most K in a δ–hyperbolic space. This constant will be relevant later because geodesics in our trees T and T' are $(1, K)$–quasigeodesics in $\mathcal{Z}$ by item (b) and our choice of K.

For ease of notation, we will refer to components in $\mathcal{U} = \pi_0(T_e) \cap \pi_0(T'_e)$ as "unchanged" components, and the remaining components as "changed". We note that there are at most K changed components in each of T_e and T'_e.

For each component $E \in \pi_0(T_e) \setminus \pi_0(T'_e)$, let $E_{\mathcal{Y}'} = \text{hull}_E(E \cap \mathcal{N}_K(\mathcal{Y}'))$, where the hull of this intersection is taken in the tree E, while the neighborhood is taken in $\mathcal{Z}$. Now define $\mathcal{R} = \bigcup_{E \in \pi_0(T_e) \setminus \pi_0(T'_e)} E_{\mathcal{Y}'}$, and define $\mathcal{R}'$ similarly. We now collect the "unstable parts" of the trees along with the unchanged parts. Set

$$U = T_c \cup \mathcal{R} \cup \mathcal{U} \cup \mathcal{B} \cup F \quad \text{and} \quad U' = T'_c \cup \mathcal{R}' \cup \mathcal{U} \cup \mathcal{B}' \cup F'.$$

Note that we included F and F' in these sets to ensure that they lie within bounded Hausdorff distance of each other, but this is inconsequential for the purposes of considering the complementary forests, which is what we want to do next.

Let $K' = K'(K, N, k) \geq 10K$ be such that

  (i) $d_{\mathrm{Haus}}(T, T') < K$ and $d_{\mathrm{Haus}}(U, U') < 10K$,

  (ii) $|\#\pi_0(T \setminus U)| < K'$ and $|\#\pi_0(T' \setminus U')| < K'$,

  (iii) $(T \setminus U) \subset T_e$ and $(T' \setminus U') \subset T'_e$,

  (iv) $|\#\pi_0(T_e \cap U) \setminus \mathcal{U}| < K'$ and $|\#\pi_0(T'_e \cap U')) \setminus \mathcal{U}'| < K'$, and

  (v) for any $C \in (\pi_0(T_e \cap U) \cup \pi_0(T'_e \cap U')) \setminus \mathcal{U}$, we have $\mathrm{diam}\, C < K'$.

Regarding property (i), the "10" in "$10K$" is there to keep into account the fact that we took hulls.

Property (ii) can be shown observing that the intersection of U with each changed component can only have a bounded number of components because of the bound N on $|\mathcal{Y} \bigtriangleup \mathcal{Y}'|$ and the bound on $|\mathcal{B}|$ given by Lemma 3.16. This same observation shows item (iv). Property (iii) holds by construction.

Property (v) is nontrivial for $\mathcal{R}$ and $\mathcal{R}'$, in which case it holds since C is a union of boundedly many components, each of bounded diameter. In particular, any set $E_{\mathcal{Y}'}$ is a union of hulls in E of intersections with balls centered at an element of $\mathcal{Y}'$, which have bounded diameter, and the union consists of at most N elements.

Now let $L_1 = L_1(K', \delta), L_2 = L_2(K', \delta) > 0$ be the constants given by Lemma 3.19 (below) with T, U and T', U' satisfying the conditions of that lemma via (i) and (ii).

Let $\mathcal{L} \subset \pi_0(T \setminus U)$ be the set of all components of $T \setminus U$ of diameter greater than L_1.

Define $\mathcal{L}' \subset \pi_0(T' \setminus U')$ to be the set of components of $T' \setminus U'$ that lie within Hausdorff distance L_2 of an element of $\mathcal{L}$. Since there are at most K' components of $T \setminus U$, this bounds the cardinality of $\pi_0(T \setminus U) \setminus \mathcal{L}$, and any component in this set has diameter at most L_1. Similarly, the numbers and diameters of the components of $(T' \setminus U')$ not in $\mathcal{L}'$ are also bounded, this time the bound on the diameter being L_3 by the "moreover" part of Lemma 3.19.

Lemma 3.19 provides a bijection $\rho \colon \mathcal{L} \to \mathcal{L}'$ which sends any component in $\mathcal{L}$ to the unique component in $\mathcal{L}'$ which is within Hausdorff distance L_2. That is, for any $C \in \pi_0(\mathcal{L})$, we have $d_{\mathrm{Haus}}(C, \rho(C)) < L_2$.

Now let $T_{s,0}$ be the union of all elements of $\mathcal{L} \cup \mathcal{U}$, and define $T'_{s,0}$ similarly. We observe that we have by construction and (iii) above that $T_{s,0} \subset T_e$ and $T'_{s,0} \subset T'_e$. Moreover, the number of components of $T_e \setminus T_{s,0}$ and their diameters are bounded by

$2K'(L_1 + 10L_2 + K' + 1)$, and similarly for $T'_e \setminus T'_{s,0}$. In fact, the number of such components is bounded by $2K'$, since each is a union of

- changed components of $T_e \cap U$, and there are at most K' of those by (iv), and
- components of $T \setminus U$ of diameter at most L_1, and again there are at most K' of those.

The bound on the diameter also follows from this description.

To obtain the sets T_s and T'_s required by the theorem, it suffices now to remove the branch points from the unchanged components contained in $T_{s,0}$ and $T'_{s,0}$ to ensure (1) (which at this point is not satisfied only because of the unchanged components, since we included the branch points in U and U'), while all other properties have been checked above. $\qquad \square$

Two supporting lemmas The following two lemmas were used in the proof of Theorem 3.3 above. To simplify notation, we will not distinguish between a tree quasi-isometrically embedded in a metric space, and the image of said tree.

Lemma 3.18 *For each K and δ there exists L_0 such that the following holds. Let T and T' be trees (K, K)–quasi-isometrically embedded in the δ–hyperbolic metric space $\mathcal{Z}$, with $d_{\mathrm{Haus}}(F_0, F'_0) \leq K$, where F_0 and F'_0 are the sets of leaves of T and T' respectively. Then the sets of branch points $\mathcal{B}$ and $\mathcal{B}'$ of T and T' satisfy $d_{\mathrm{Haus}}(\mathcal{B} \cup F_0, \mathcal{B}' \cup F'_0) \leq L_0$.*

Proof The set $\mathcal{B} \cup F_0$ can be coarsely characterized as the set of points x of T such that there are f_1, f_2 and f_3 in F_0 (not necessarily distinct) with the property that the Gromov product at x between any f_i and f_j is small, and similarly for T'. We leave the details to the reader. $\qquad \square$

Lemma 3.19 *For each K there exist L_1, L_2 and L_3 such that the following holds. Let T and T' be trees (K, K)–quasi-isometrically embedded in the metric space $\mathcal{Z}$, with $d_{\mathrm{Haus}}(T, T') \leq K$. Also, let $U \subseteq T$ and $U' \subseteq T'$ be subforests such that $d_{\mathrm{Haus}}(U, U') \leq K$ and all branch points of T (resp. T') are contained in U (resp. U').*

Then for each component C of $T \setminus U$ of diameter at least L_1, there exists a unique component C' of $T' \setminus U'$ within Hausdorff distance L_2 of C. Moreover, every component C' of $T' \setminus U'$ of diameter at least L_3 arises in this way.

Proof We will conflate components of $T \setminus U$ with their closures, so we can talk about their leaves, and similarly for $T' \setminus U'$.

The main observation is that there exists $K_2 = K_2(K)$ such that the following holds. Let C be a component of $T \setminus U$ and let x and y be (not necessarily distinct) points in C that, in the metric of C, are at least K_2 from all the leaves of C. Then there exists a unique component C' of $T' \setminus U'$ which is within K of both x and y.

To prove this, suppose by contradiction that $K_2 \gg K$ and that there are distinct components of $T' \setminus U'$ that contain points x' and y' that are within K of x and y, respectively. Then there exists some $p' \in U'$ on the geodesic $[x', y']$ in T' from x' to y'. Let $p \in U$ be such that $d(p, p') < K$. Then p lies within $\sigma = \sigma(K) > 0$ from the geodesic from x to y, since considering points in T that are within K of those along $[x', y']$ yields a quasigeodesic in T. Since x and y are at least K_2 from the leaves of C, they cannot lie close to any point of U, in particular p. We can then deduce that either p lies along the geodesic $[x, y]$ in T from x to y, or there is a branch point of T along $[x, y]$. In either case, x and y do not lie in the same component of $T \setminus U$, a contradiction. (Recall that U contains all branch points of T by hypothesis.)

Consider now a component C of $T \setminus U$ of diameter sufficiently large that it contains a point which is at least K_2 from all the leaves of C. By the observation above, all such points are close to a unique component C' of T', and since the set of all such points has bounded Hausdorff distance from C, we have that C is contained in a uniform neighborhood of C'. Moreover, if C has sufficiently large diameter, then we can apply the same reasoning to C' and deduce that C' contains points that are within K of a unique component C'' of $T \setminus U$, and that C' is contained in a uniform neighborhood of C''. But the above observation implies that $C'' = C$, and it follows that C and C' lie within uniformly bounded Hausdorff distance.

Finally, the "moreover" part follows from a similar back-and-forth using the previous part of the statement. Namely C' has sufficiently large diameter and is within bounded Hausdorff distance of a component C of $T \setminus U$, then C also has large diameter and is thus in turn within bounded Hausdorff distance of some component of $T' \setminus U'$, which needs to be C'. $\qquad\square$

4 Stable cubulations

Fix a G–colorable HHS $(\mathcal{X}, \mathfrak{S})$ for $G < \mathrm{Aut}(\mathfrak{S})$, and let $F \subset \mathcal{X}$ be a finite set.

In this section, we use the stable trees constructed in Section 3 for the projections of F to the relevant domains to define a wallspace on the hull $H_\theta(F)$. This wallspace

can then be plugged into Sageev's machine to produce a cube complex which, by an argument from [8], coarsely models the hull of F in $\mathcal{X}$. Stability of the tree construction then induces stability in the cubulations under perturbations of F. We refer the reader to Section 2.1 for some background and references on cube complexes, wallspaces, and hyperplane deletions, as well as to Section 2.2 for background on HHSs.

The main result of this section, and the only statement from this section that we will use in the rest of the paper, is the following precise version of Theorem A.

Theorem 4.1 Let $(\mathcal{X}, \mathfrak{S})$ be a G–colorable HHS for $G < \mathrm{Aut}(\mathfrak{S})$. Then for each k there exist K and N with the following properties. To each subset $F \subseteq \mathcal{X}$ of cardinality at most k one can assign a triple $(\mathcal{Q}_F, \Phi_F, \psi_F)$ satisfying

(1) $\mathcal{Q}_F$ is a CAT(0) cube complex of dimension at most the maximal number of pairwise orthogonal domains of $(\mathcal{X}, \mathfrak{S})$;

(2) $\Phi_F : \mathcal{Q}_F \to H_\theta(F)$ is a K–median (K, K)–quasi-isometry;

(3) $\psi_F : F \to (\mathcal{Q}_F)^{(0)}$ satisfies $d_{\mathcal{X}}(\Phi_F \circ \psi_F(f), f) \le K$ for each $f \in F$.

Moreover, suppose that $F' \subseteq \mathcal{X}$ is another subset of cardinality at most k, $g \in G$, and $d_{\mathrm{Haus}}(gF, F') \le 1$. Choose any map $\iota_F : F \sqcup F' \to F$ such that $\iota_F(f) = f$ if $f \in F$ and $d_{\mathcal{X}}(g(\iota_F(f)), f) \le 1$ if $f \in F'$. Also, choose a map $\iota_{F'} : F \sqcup F' \to F'$ such that $\iota_{F'}(f) = f$ if $f \in F'$ and $d_{\mathcal{X}}(g(f), \iota_{F'}(f)) \le 1$ if $f \in F$. Then the following holds:

There is a third CAT(0) cube complex $\mathcal{Q}_0$ and K–median (K, K)–quasi-isometric embedding Φ_0 such that the diagram

(4-1)

$$
\begin{array}{ccc}
 & F \xrightarrow{\ \psi_F\ } \mathcal{Q}_F & \\
 \nearrow^{\iota_F} & \downarrow_\eta \quad \searrow^{g \circ \Phi_F} & \\
F \sqcup F' & \mathcal{Q}_0 \xrightarrow{\ \Phi_0\ } \mathcal{X} & \\
 \searrow_{\iota_{F'}} & \uparrow^{\eta'} \quad \nearrow_{\Phi_{F'}} & \\
 & F' \xrightarrow{\ \psi_{F'}\ } \mathcal{Q}_{F'} &
\end{array}
$$

commutes up to error at most K, where η and η' are hyperplane deletion maps that delete at most N hyperplanes. The left side commutes exactly, ie $\eta \circ \psi_F \circ \iota_F = \eta' \circ \psi_{F'} \circ \iota_{F'}$.

The notion of K–median refers to the coarse median structure on $\mathcal{X}$ in the sense of [15], and we only define it later where it is needed, since we obtain it directly from [8].

The "1" in "$d_{\text{Haus}}(gF, F') \le 1$" could be changed with any other constant up to changing K in the "moreover" part. We decided to keep the statement slightly simpler and not introduce further quantifiers since one can always replace an HHS with a graph, or rescale the metric.

The remainder of this section is devoted to the proof of this theorem.

Standing assumptions For this section, we fix an HHS $(\mathcal{X}, \mathfrak{S})$ as in Theorem 4.1, which, in light of Theorem 2.9, we can assume to have stable projections. Furthermore, we can assume that it has the property that all $\pi_V(x)$ and all ρ_V^U for $U \subsetneq V$ are single points; see Remark 2.16.

4.1 Subdivision sets for stable trees

In this section, we establish a formalism for subdividing trees. In Section 4.2 these subdivision points will give the walls in our cubulation. This mostly follows the strategy of [8], except that we need to take greater care in making choices for the subdivision.

Definition 4.2 Let $M' > M > 0$. An (M, M')–*subdivision* of a tree T is a collection of points $\boldsymbol{p}(T) \subset T$ satisfying:

(1) The points $\boldsymbol{p}(T)$ are contained in the interiors of edges of T. We set

$$\boldsymbol{p}(e) = \boldsymbol{p}(T) \cap \text{int}(e)$$

for each edge e of T.

(2) The $M'/2$–neighborhood of $\boldsymbol{p}(e) \cup \partial e$ covers e.

(3) All points of $\boldsymbol{p}(e) \cup \partial e$ are at least M apart in e.

In other words, the spacing between points of $\boldsymbol{p}(e) \cup \partial e$ along e is at least M and strictly less than M'.

We additionally say that $\boldsymbol{p}(T)$ is (M, M')–*evenly spaced* if $M' \ge 8M$ and the spacing between successive points of $\boldsymbol{p}(e)$ is exactly M for each edge e.

We will specify M and M' later, though for now it suffices to assume they are large relative to the various HHS constants.

We fix, once and for all, a *subdivision operator* $\wp_{M,M'}$ which to any tree T associates a fixed (M, M')–evenly spaced subdivision $\wp_{M,M'}(T)$ of (the edges of) T. Often the constants M and M' are fixed, and we simply write $\wp(T)$. Similarly we can define $\wp_{M,M'}$ on a forest as the union of subdivisions $\wp_{M,M'}$ on its components.

We now explain how to associate a collection of subdivisions on stable trees to a set of points $F \subset \mathcal{X}$.

Fix some large K (depending on M, to be specified later) and write $\mathcal{U}(F) = \mathrm{Rel}_K(F)$ for any finite set $F \subset \mathcal{X}$.

For any $V \in \mathfrak{S}$, define
$$\mathcal{Y}^V = \{\rho_V^W \mid W \in \mathcal{U}(F), W \subsetneq V\} \quad \text{and} \quad F^V = \pi_V(F).$$

Note that whenever K is larger than the bounded geodesic image constant κ_0, we have that $\mathcal{Y}^V$ is contained in the κ_0–neighborhood of the hull of F^V in $\mathcal{C}(V)$. This ensures that $\mathcal{Y}^V$ satisfies the requirements of Theorem 3.3 for any $\epsilon \geq 2\kappa_0$. We fix such ϵ as in Theorem 3.3, and we will always apply that theorem with this ϵ.

Note that if $V \notin \mathcal{U}(F)$ then $\mathcal{Y}^V \cup F^V$ has uniformly bounded diameter.

Let $V \in \mathcal{U}(F)$. From V, we get corresponding sets of projections F^V and $\mathcal{Y}^V$ in $\mathcal{C}(V)$, which we may consider independently of the set F. Doing so, we obtain from Theorem 3.3 a fixed stable tree
$$T_F^V := T(F^V, \mathcal{Y}^V)$$
and we denote its decomposition by $T_c^V \cup T_e^V$.

Let $M' > 8M > 0$ be subdivision constants (to be specified later). Applying $\wp_{M,M'}$ to each forest T_e^V in T_F^V, we call the resulting subdivision $\wp^V(F)$.

Remark 4.3 We emphasize that the distance between the subdivision points is measured in the various trees themselves, not in the corresponding $\mathcal{C}(V)$ where the trees quasi-isometrically embed. This is something that will require care throughout this section, but the fact that the trees are quasi-isometrically embedded in the corresponding $\mathcal{C}(V)$ will ensure that no real issues arise.

The disjoint union over all $V \in \mathcal{U}(F)$ gives us
$$\wp \equiv \wp(F) = \bigsqcup_{V \in \mathcal{U}(F)} \wp^V(F).$$

More generally, we will consider (M, M') subdivisions of the forests T_e^V which are not necessarily obtained from our subdivision operator $\wp_{M,M'}$, and in particular may not be evenly spaced (they will arise by taking subsets). In this case if we name the full configuration $\boldsymbol{p}$, we will again denote the restriction to T_e^V as $\boldsymbol{p}_V(F)$.
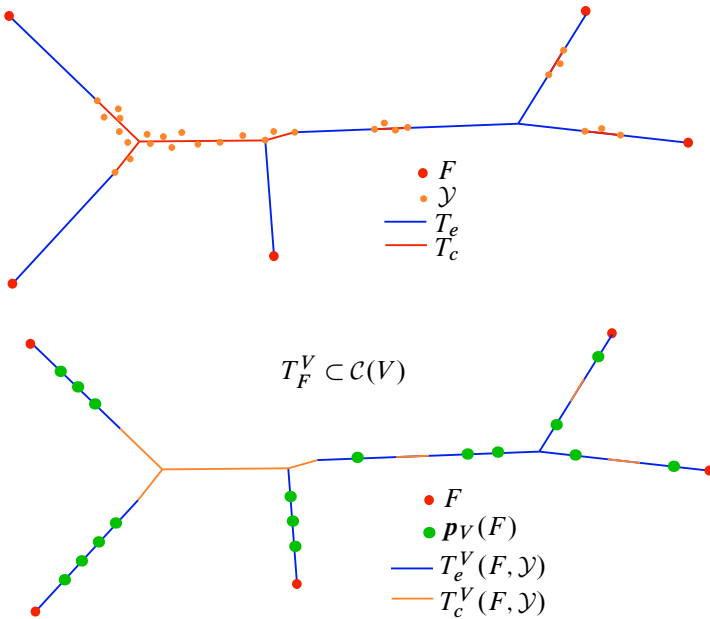
Figure 12: An evenly spaced subdivision of the stable tree T_F^V. Note that subdivision points are far away from leaves, components of T_c, and branch points.

Remark 4.4 Our choice of constants $M' > 8M$ gives that any point $p \in \wp^V(F)$ is at least distance $4M$ from any point $\rho_V^W \in \mathcal{Y}^V$ for $W \sqsubset V$ or leaf $f \in F^V$.

4.2 The cube complex

Following the scheme of [8], given a union p of (M', M) subdivisions as above, we now describe a cube complex $\mathcal{Q}_{F,p}$ and a map

$$\Phi_{F,p} \colon \mathcal{Q}_{F,p} \to \mathcal{X}$$

which turns out to be a quasi-isometric embedding whose image is within bounded Hausdorff distance of $H_\theta(F)$, when the constants M, M' and K are chosen suitably (see below).

To build the cube complex $\mathcal{Q}_{F,p}$, we build a wallspace structure on the hull $H_\theta(F)$ in which each $p \in p$ corresponds to a wall, ie a partition of $H_\theta(F)$ into two sets.

For each $V \in \mathfrak{S}$, we let $\beta_F^V \colon \mathcal{C}(V) \to T_F^V$ be, roughly, a closest-point projection map to T_F^V, and more precisely any fixed map such that $d_V(x, \Xi(\beta_F^V(x)))$ is minimal for

all $x \in \mathcal{C}(V)$, where we recall that Ξ is the quasi-isometric embedding of T_F^V in $\mathcal{C}(V)$. With slight abuse of notation, for $x \in \mathcal{X}$ we write $\beta_F^V(x)$ for $\beta_F^V(\pi_V(x))$.

Remark 4.5 To save notation, we will often omit the map Ξ, thereby identifying points and subsets of some T_F^V with their image in $\mathcal{C}(V)$ (as one often does with quasi-geodesics). For example, for $p \in \beta^V(F)$ and $x \in \mathcal{C}(V)$ we will write $d_V(\beta_F^V(p), x)$ rather than $d_V(\Xi(\beta_F^V(p)), x)$.

Given $p \in \boldsymbol{p}_V(F)$, let $T_{p,+}^V$ denote one of the components of $T_F^V \setminus \{p\}$, and let $T_{p,-}^V$ be the union of the other component and $\{p\}$ (we arbitrarily choose the first component). Let
$$W_{p,\pm}^V = (\beta_F^V)^{-1}(T_{p,\pm}^V) \cap H_\theta(F).$$
Note that $W_{p,+}^V$ and $W_{p,-}^V$ form a partition of $H_\theta(F)$. Let $\mathcal{L}_p^V = \{W_{p,+}^V, W_{p,-}^V\}$ be the wall associated to p. We call $T_{p,\pm}^V$ the *half-trees* associated to the wall $\mathcal{L}_p^V$.

Let $\mathcal{Q} = \mathcal{Q}_{F,\boldsymbol{p}}$ be the CAT(0) cube complex dual to the wallspace $\{\mathcal{L}_p^V\}$.

To define $\Phi_F = \Phi_{F,\boldsymbol{p}} \colon \mathcal{Y} \to \mathcal{X}$, note that it suffices to define Φ_F on the 0–skeleton of $\mathcal{Y}$. Let $x \in \mathcal{Y}^{(0)}$; we view x as a coherent orientation of the walls $\mathcal{L}_p^V$; see Section 2.1.1. That is, for each $p \in \boldsymbol{p}$, we have $x(p)$ equal to one of $W_{p,+}^V$ or $W_{p,-}^V$.

Coarsely, we would like to define $\Phi(x)$ by
$$\Phi(x) \approx \bigcap_{p \in \boldsymbol{p}} x(p).$$

This is done in [8] by considering the projections of $x(p)$ to the factor trees T_F^Y. That is, we set
$$S_{p,Y}(x) = \mathrm{hull}(\beta_F^Y(x(p))),$$
where $Y \in \mathfrak{S}$ and hull denotes convex hull in the tree T_F^Y. Note that $p \in \boldsymbol{p}_V(F)$ for some V, which is typically different from Y.

We now define the intersection
$$(4\text{-}2) \qquad b_Y = b_Y(x) = \bigcap_{p \in \boldsymbol{p}} S_{p,Y}(x).$$

These b_Y will serve as (coarse) coordinates for the map Φ, as given by the theorem that we state below, after we explain how it can be extracted from [8].

Remarks on the construction in [8] We now summarize various results proven in [8] regarding cubulations of hulls in HHSs. We note that the construction in [8] of the CAT(0) cube complexes approximating hulls is the same as the one we just described,

with one difference. The difference is that, rather than using T_F^V to approximate the hull of $\pi_V(F)$ in $\mathcal{C}(V)$, in [8] the authors use any choice of tree contained in $\mathcal{C}(V)$ that uniformly approximates the hull and is quasi-isometrically embedded with multiplicative constant 1. This choice is not due to the fact that the multiplicative constant being 1 is needed for the arguments, but more simply due to the fact that one such tree exists, and so it is more convenient to use it. For our purposes, we have to be more careful in the choice of the tree, and as a result we cannot guarantee that the multiplicative constant is 1 with our construction (recall Figure 11). However, this does not affect the arguments of [8], except that the subdivision constant M has to be chosen large compared to the quasi-isometric embedding constants, so that the subdivision points are sufficiently far apart in the various hyperbolic spaces $\mathcal{C}(V)$.

Another remark about the statement below is that the constant dependencies that we give below are not explicitly stated in [8], but can be recovered as follows. In [8, Section 2], the constant M is chosen large compared to various HHS constants and $k = |F|$, so that the construction has all the stated properties for any sufficiently large M. Regarding M', in [8, Section 2.1] it is taken to be $10Mk$, as can be seen from point (4) of the construction of the walls. The reason for the constant $10Mk$ is that one can choose subdivision points that make the diameter of the complementary components at most that quantity, but with any other bound one would obtain the same properties (eg that the CAT(0) cube complex quasi-isometrically embeds in the HHS), with different constants. Regarding K, in [8, Section 2.1] it is chosen to be $100Mk$, and similar remarks to those regarding M' apply.

Properties of the cubulation With this in mind, we now state various results about the construction we explained above, and point out where the arguments for those can be found in [8].

Theorem 4.6 *Given an HHS $(\mathcal{X}, \mathfrak{S})$ and an integer k, there exist $M_0 \geq 1$ and functions $M_0' \colon \mathbb{R} \to \mathbb{R}$ and $K_0 \colon \mathbb{R} \to \mathbb{R}$ with the following property. Whenever $M \geq M_0$, $M' \geq M_0'(M)$, and $K \geq K_0(M)$, there exists ξ such that for every $F \subseteq \mathcal{X}$ with $|F| \leq k$, the following hold:*

(1) [8, Lemma 2.6, paragraph "Definition of $\mathfrak{p}_A$" in proof of Theorem 2.1] *For every $x \in \mathcal{Q}_{F,\boldsymbol{p}}^{(0)}$, $b_Y(x)$ is nonempty and $\mathrm{diam}_{\mathcal{C}(Y)}(b_Y(x)) \leq \xi$.*

(2) [8, Lemma 2.7, paragraph "Definition of $\mathfrak{p}_A$" in proof of Theorem 2.1] *For every $x \in \mathcal{Q}_{F,\boldsymbol{p}}^{(0)}$, there exists a point in $H_\theta(F)$, denoted by $\Phi(x)$, whose projections to all $\mathcal{C}(Y)$ are within distance ξ of $b_Y(x)$.*

(3) [8, Theorem 2.1] Φ is a ξ–median (ξ, ξ)–quasi-isometry to $H_\theta(F)$.

(4) [8, Theorem 2.1] *The dimension of $\mathcal{Q}_{F,\boldsymbol{p}}$ is bounded by the maximal number of pairwise orthogonal domains in $\mathfrak{S}$.*

We will also need some more technical properties of the trees T_F^V and projections ρ_V^U related to the HHS consistency axioms.

Proposition 4.7 *Given an HHS $(\mathcal{X}, \mathfrak{S})$ and an integer k, there exists κ such that given K sufficiently large (depending only on $(\mathcal{X}, \mathfrak{S})$):*

(1) [8, Lemma 2.3] *If $U, V \in \mathcal{U}(F)$ and $U \pitchfork V$, then ρ_V^U lies κ–close in $\mathcal{C}(V)$ to a point of $\pi_V(F)$.*

(2) [8, Lemma 2.5] *If $U, V \in \mathcal{U}(F)$, $V \subsetneqq U$, and $q \in \boldsymbol{p}(U)$, then $\rho_V^U(q)$ lies κ–close in $\mathcal{C}(V)$ to a point of $\pi_V(F)$.*

Remark 4.8 In the rest of this section, whenever we use constants M, M' and K we will assume that they are chosen as in Theorem 4.6, and so that all supporting lemmas in [8] apply. Moreover, we will impose further requirements as needed. We note that the role of K is often hidden in the statements, since it affects the set $\mathcal{U}(F)$, which in the various statements often only plays a role implicitly.

4.3 Deleting subdivision points

Now we consider how the construction of $\mathcal{Q}_{F,\boldsymbol{p}}$ is affected by the deletion of points in $\boldsymbol{p}$. If $\boldsymbol{p}_0 \subset \boldsymbol{p}$, there is a hyperplane-deletion map $h\colon \mathcal{Q}_{F,\boldsymbol{p}} \to \mathcal{Q}_{F,\boldsymbol{p}_0}$. That is, for $x \in \mathcal{Q}_{F,\boldsymbol{p}}$, the image $h(x) \in \mathcal{Q}_{F,\boldsymbol{p}_0}$ is just the orientation on the remaining walls: $h(x)(p) = x(p)$ for $p \in \boldsymbol{p}_0$. We note that the subdivisions in the following proposition need not be evenly spaced.

Proposition 4.9 *For every k and n, and M, M' and K as in Theorem 4.6, there exist K' and M'' such that if F has cardinality at most k, $\boldsymbol{p}$ is an (M, M')–subdivision, and $\boldsymbol{p}_0 \subset \boldsymbol{p}$ satisfies $|\boldsymbol{p} \setminus \boldsymbol{p}_0| \leq n$, then $\boldsymbol{p}_0$ is an (M, M'')–subdivision satisfying the conclusions of Theorem 4.6, and the diagram*

(4-3)

$$
\begin{array}{ccc}
\mathcal{Q}_{F,\boldsymbol{p}} & & \\
{\scriptstyle h}\Big\downarrow & \overset{\Phi_{F,\boldsymbol{p}}}{\searrow} & \\
\mathcal{Q}_{F,\boldsymbol{p}_0} & \xrightarrow[\Phi_{F,\boldsymbol{p}_0}]{} & \mathcal{X}
\end{array}
$$

commutes up to error K'.

Proof As above, the map $\Phi_{F,\boldsymbol{p}}(x)$ is determined by the coordinates

$$b_U = \bigcap_{p \in \boldsymbol{p}} S_{p,U}(x),$$

whereas $\Phi_{F,\boldsymbol{p}_0}(h(x))$ is determined by

$$b_{0,U} = \bigcap_{p \in \boldsymbol{p}_0} S_{p,U}(x).$$

Note that $b_U \subset b_{0,U}$, and in the other direction the diameter of $b_{0,U}$ is bounded by Theorem 4.6(1) because the new set $\boldsymbol{p}_0$ is an (M, M'')–subdivision for some $M'' \geq M'$ (and in particular $M'' \geq M_0'(M)$, so that the conclusion of Theorem 4.6 holds for $\boldsymbol{p}_0$) depending only on M, M' and the number n of deletions.

The bound on $d(\Phi_{F,\boldsymbol{p}}, \Phi_{F,\boldsymbol{p}_0} \circ h)$ then follows from Theorem 2.6 (the distance formula) since the coordinates of $\Phi_{F,\boldsymbol{p}_0}(h(x))$ and $\Phi_{F,\boldsymbol{p}}(x)$ coarsely coincide with b_U and $b_{0,U}$. □

4.4 Intersection conditions

Recall that Lemma 2.3 explains how a bijection between halfspaces that preserves intersection properties induces an isomorphism of the corresponding cube complexes. In view of this, we are interested in knowing when two of our halfspaces intersect.

We fix the setup of Section 4.2. The next lemma is the main technical support for Proposition 4.13.

Lemma 4.10 *There exists M_1, depending on $(\mathcal{X}, \mathfrak{S})$ and $|F|$, such that the following holds. Let $\boldsymbol{p}$ be an (M, M')–subdivision with $M \geq M_1$. Consider two halfspaces, $W_{p,\sigma}^V$ and $W_{q,\tau}^Z$, with associated half-trees $T_{p,\sigma}^V$ and $T_{q,\tau}^Z$, where $p \in \boldsymbol{p}_V(F)$, $q \in \boldsymbol{p}_Z(F)$ and $\tau, \sigma \in \{\pm\}$.*

Then $W_{p,\sigma}^V$ and $W_{q,\tau}^Z$ intersect if and only if one of the following holds, up to switching the roles of the half-spaces:

(1) $V \perp Z$.

(2) $V = Z$, and $T_{p,\sigma}^V \cap T_{q,\tau}^Z \neq \varnothing$.

(3) $V \pitchfork Z$, and $T_{q,\tau}^Z$ contains $\beta_F^Z(\rho_Z^V)$.

(4) $V \subsetneqq Z$, and $T_{q,\tau}^Z$ contains $\beta_F^Z(\rho_Z^V)$.

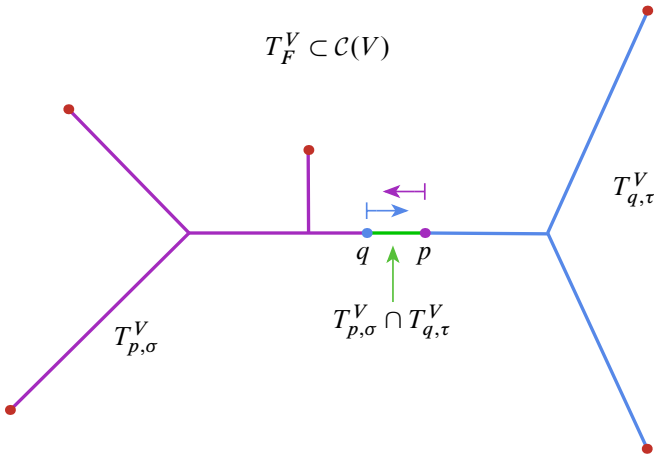*(5)* $V \subsetneqq Z$, and $T_{p,\sigma}^V$ contains $\beta_F^V(\rho_V^Z(q))$.

Figure 13: The first possibility in case (2), when $V = Z$, where p and q choose different ends of T_F^V, and $\text{diam}(T_{p,\sigma}^V \cap T_{q,\tau}^V) \geq d_{T_F^V}(p,q) \geq M$.

We note that the last three cases of the lemma boil down to the consistency inequalities for HHSs (and could even be seen as interpretations thereof).

Proof In an effort to enhance readability of the proof, we will make coarse, comparative arguments which keep track of dependencies of constants and their relative size, instead of precise quantities.

It follows from Theorem 2.7 and the fact that the various T_F^V are quasi-isometrically embedded (Theorem 3.3) that the image $\beta_F^V(H_\theta(F))$ is D–dense in the tree T_F^V (with respect to its path metric), where D only depends on the HHS structure and $|F|$. We may assume that M has been chosen greater than $10D$. Moreover, throughout the proof we will further specify conditions on M, requiring it to be suitably larger than other constants appearing in the argument. It is important to notice that, in each case, these constants depend only on $(\mathcal{X}, \mathfrak{S})$ and $|F|$. We also remark that we need to be careful when comparing distances in the trees T_F^V and the ambient spaces $\mathcal{C}(V)$.

Case (1) Suppose $V \perp Z$. The fact that in this case all pairs of halfspaces intersect is [8, Lemma 2.13].

Case (2) Suppose $V = Z$. If the halfspaces intersect, then any point x in their intersection has $\beta_F^V(x) \in T_{p,\sigma}^V \cap T_{q,\tau}^V$. Conversely, suppose $T_{p,\sigma}^V \cap T_{q,\tau}^V \neq \varnothing$. If the intersection contains both p and q, then

$$\text{diam}_{T_F^V}(T_{p,\sigma}^V \cap T_{q,\tau}^V) \geq d_{T_F^V}(p,q) \geq M.$$
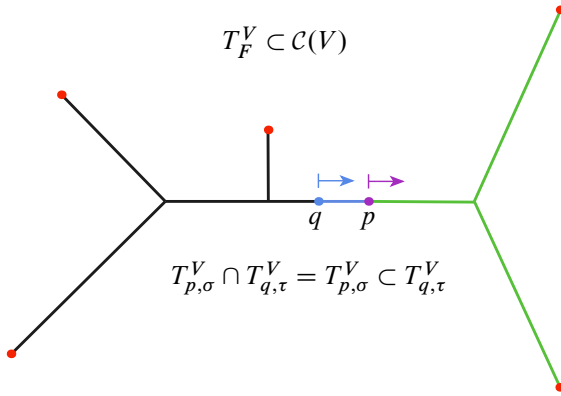
Figure 14: The other possibility in case (2), when $V = Z$, where p and q now choose the same ends of T_F^V.

If not, then the intersection contains all of $T_{p,\sigma}^V$ or all of $T_{q,\tau}^V$, and again

$$\mathrm{diam}_{T_F^V}(T_{p,\sigma}^V \cap T_{q,\tau}^V) \geq M.$$

Since $\beta_F^V(H_\theta(F))$ is D–dense in T_F^V and $M > D$, we can then find $x \in H_\theta(F)$ with $\beta_F^V(x) \in T_{p,\sigma}^V \cap T_{q,\tau}^Z$, so that $x \in W_p^V \cap W_q^Z$, as required.

Case (3)   Suppose $V \pitchfork Z$. If the halfspaces $W_{p,\sigma}^V$ and $W_{q,\tau}^Z$ intersect, then any point x in their intersection has $\beta_F^V(x) \in T_{p,\sigma}^V$ and $\beta_F^Z(x) \in T_{q,\tau}^Z$. We claim that

$$\beta_F^Z(\rho_Z^V) \notin T_{q,\tau}^Z \implies d_Z(\rho_Z^V, \pi_Z(x)) > \kappa_0,$$

where κ_0 is the constant in the transverse consistency inequality (2-1), and similarly with the roles of Z and V interchanged. Indeed, recall that Proposition 4.7(1) says that (assuming M is sufficiently large) ρ_Z^V lies within κ of a leaf of T_F^Z, where κ only depends on the HHS structure and $|F|$. Since q is at least M away from the leaves, and we can assume that M is sufficiently large compared to κ and the quasi-isometric embedding constants of T_F^Z, we see that if $\beta_F^Z(\rho_Z^V)$ is not in $T_{q,\tau}^Z$ then it must be at least $M/2$ from it (as measured in the metric of T_F^Z). Thus $\beta_F^Z(\rho_Z^V)$ is at least $M/2$ from $\beta_F^Z(x)$ in T_F^Z, and since we can assume that M is sufficiently large compared to κ, the quasi-isometric embedding constants of T_F^Z, and the distance between $\pi_Z(x)$ and (the image in $\mathcal{C}(Z)$ of) T_F^Z, we have the desired inequality. The same holds with V and Z interchanged.

However, the transverse consistency inequality (2-1) says that we cannot have both $d_Z(\rho_Z^V, \pi_Z(x)) > \kappa_0$ and $d_V(\rho_V^Z, \pi_V(x)) > \kappa_0$. Hence one of $\beta_F^Z(\rho_Z^V) \in T_{q,\tau}^Z$ or $\beta_F^V(\rho_V^Z) \in T_{p,\sigma}^V$ must hold, which is what we wanted to prove.
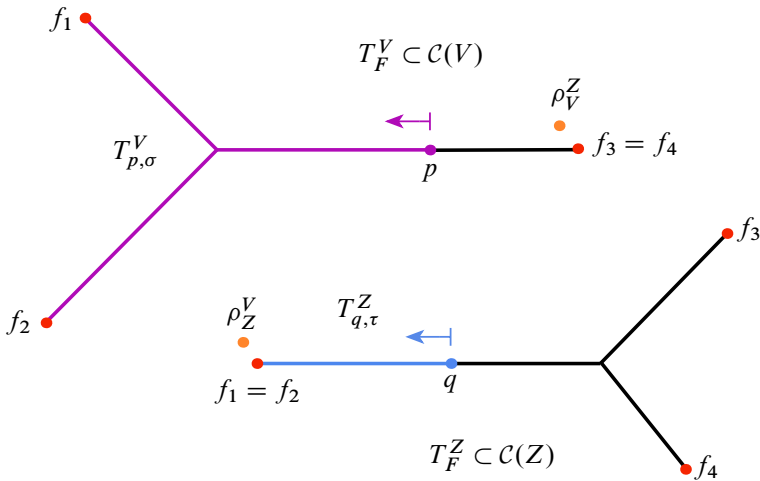
Figure 15: Case (3), when $V \pitchfork Z$.

Conversely, suppose that $T^Z_{q,\tau}$ contains $\beta^Z_F(\rho^V_Z)$. Because $\beta^V_F(H_\theta(F))$ is D–dense in T^V_F, p is far from any leaf of T^Z_F, and $\beta^Z_F(\rho^V_Z)$ is close to a leaf (Proposition 4.7(1)), we may choose $x \in H_\theta(F)$ with $\beta^V_F(x) \in T^V_{p,\sigma}$ and $d_{T^V_F}(\beta^V_F(x), \beta^V_F(\rho^Z_V)) > M/2$ (again, with distance in T^V_F).

Then $x \in W^V_{p,\sigma}$ by construction, and we claim also $x \in W^Z_{q,\tau}$. Indeed, for M sufficiently large, $d_V(x, \rho^Z_V) > \kappa_0$, so transverse consistency (2-1) implies that $d_Z(\pi_Z(x), \rho^V_Z) < \kappa_0$. Again for M sufficiently large, this gives $d_{T^Z_F}(\beta^Z_F(x), \beta^Z_F(\rho^V_Z)) \le M/2$, and since the $M/2$–neighborhood in T^Z_F of $\beta^Z_F(\rho^V_Z)$ is contained in $T^Z_{q,\tau}$, we conclude $\beta^Z_F(x) \in T^Z_{q,\tau}$.

Cases (4) and (5) Suppose $V \subsetneq Z$. Cases (4) and (5) are similar to the previous case, instead using the nested consistency inequality (2-2).

Let $x \in W^V_{p,\sigma} \cap W^Z_{q,\tau}$. Because the partition points in T^Z_F are M away from ρ^V_Z by assumption, we have again that either

- $T^Z_{q,\tau}$ contains the $M/2$–neighborhood of $\beta^Z_F(\rho^V_Z)$, or
- $d_{T^Z_F}(T^Z_{q,\tau}, \beta^Z_F(\rho^V_Z)) > M/2$.

In the first case, we are done (obtaining case (4)). In the second case, for M sufficiently large, $d_V(\pi_V(x), \rho^Z_V(\pi_Z(x))) < \kappa_0$ by the nested consistency inequality (2-2). Moreover, for M sufficiently large, there is a geodesic (in $\mathcal{C}(Z)$) from q to $\pi_Z(x)$ which is farther than κ_0 from ρ^V_Z, where κ_0 is the constant in the bounded geodesic image property (axiom (7) in Definition 2.4). In fact, the distance from $\pi_Z(x)$ to $T^Z_{q,\tau}$
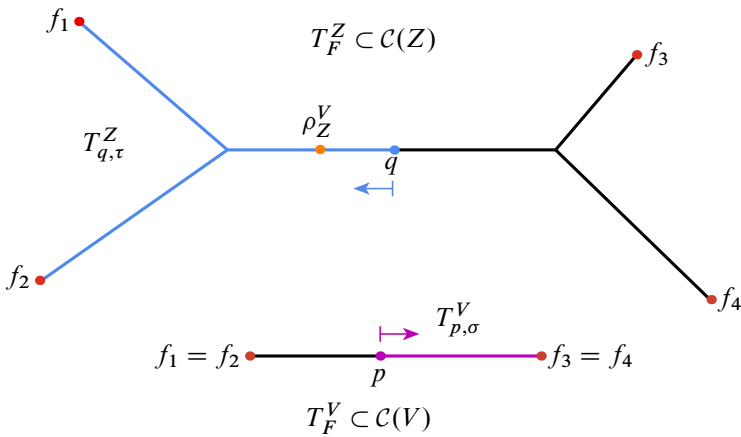
Figure 16: Case (4), when $V \subsetneqq Z$ and $T_{q,\tau}^Z$ contains $\beta_F^Z(\rho_Z^V)$.

is bounded in terms of $(\mathcal{X}, \mathfrak{S})$ and $|F|$, as is the quasiconvexity constant (of the image in $\mathcal{C}(Z)$) of $T_{q,\tau}^Z$, and thus so is the distance of any point along this geodesic from $T_{q,\tau}^Z$.

On the other hand, we can estimate $d_Z(T_{q,\tau}^Z, \rho_Z^V)$ in terms of M and the quasi-isometric embedding constants of T_F^Z (which is independent of M). Hence by choosing M large enough, we may use the bounded geodesic image property to conclude that $d_V(\rho_V^Z(\pi_Z(x)), \rho_V^Z(q)) < \kappa_0$. Since $\pi_V(x)$ is within some distance of $T_{p,\sigma}^V$ which is bounded in terms of $(\mathcal{X}, \mathfrak{S})$ and $|F|$, we see that $d_V(\beta_F^V(\rho_V^Z(q)), T_{p,\sigma}^V) < M/2$ for M sufficiently large. However, this implies that $\beta_F^V(\rho_V^Z(q))$ is in fact contained in $T_{p,\sigma}^V$, since $\rho_V^Z(q)$ is within κ of a leaf of T_F^V by [Proposition 4.7](2). Hence we are done in either case.

Now in the converse direction, suppose first that (4) holds; namely that $T_{q,\tau}^Z$ contains $\beta_F^Z(\rho_Z^V)$. Again since partition points are M away from ρ points, we know that $T_{q,\tau}^Z$ contains the $M/2$–neighborhood (in T_F^V) of $\beta_F^Z(\rho_Z^V)$. Using D–density of $\beta_F^V(H_\theta(F))$ in T_F^V and the fact that p is far from any leaf of T_F^V, we can choose $x \in H_\theta(F)$ with $\beta_F^V(x) \in T_{p,\sigma}^V$ and $\beta_F^V(x)$ at least $M/2$ from a leaf of $T_{p,\sigma}^V$.

We claim that this implies that $\beta_F^Z(x)$ has to lie within $M/2$ of $\beta_F^Z(\rho_Z^V)$, as usual for M large. If not, another application of the bounded geodesic image axiom would imply that $\rho_V^Z(\pi_Z(x))$ is within κ_0 of a leaf, and the nested consistency inequality again gives that $\rho_V^Z(\pi_Z(x))$ and $\pi_V(x)$ are within κ_0 of each other, showing that $\pi_V(x)$ lies within $2\kappa_0$ of a leaf. Since the constants involved in the preceding argument depend only on $(\mathcal{X}, \mathfrak{S})$ and $|F|$, it follows that the distance (in T_F^V) between $\beta_F^V(x)$ and a leaf is bounded independently of M, contradicting the choice of x.
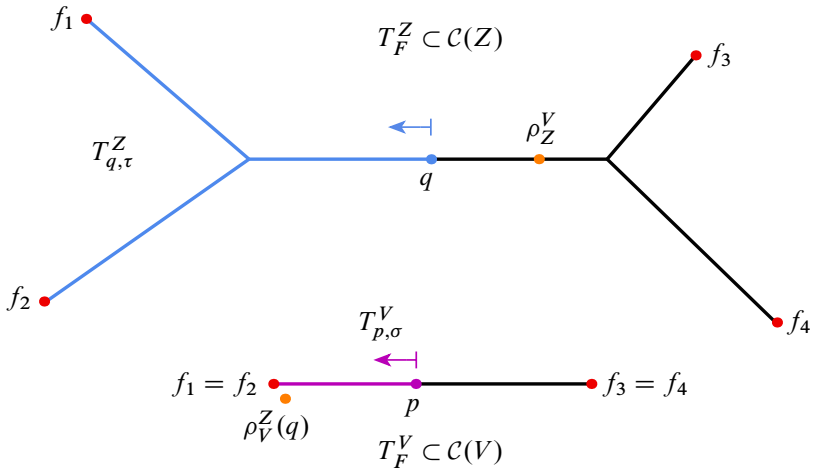
Figure 17: A cartoon for case (5), when $V \subsetneq Z$ and $T_{p,\sigma}^V$ contains $\beta_F^V(\rho_V^Z(q))$.

As a result, we find that $\beta_F^Z(x)$ is in $T_{q,\tau}^Z$ and hence $x \in W_{p,\sigma}^V \cap W_{q,\tau}^Z$, as required.

Suppose now as in (5) that $T_{p,\sigma}^V$ contains $\beta_F^V(\rho_V^Z(q))$, and assume also that (4) does not hold, so that $T_{q,\sigma}^Z$ avoids the $M/2$–neighborhood of $\beta_F^V(\rho_V^Z)$ in T_F^V. In this case, we can take $x \in H_\theta(F)$ with $\beta_F^Z(x) \in T_{q,\tau}^Z$ and $d_{T_F^Z}(\beta_F^Z(x), q) \leq D$. Then since $\beta_F^Z(\rho_V^Z)$ is at least $M/2$ from the geodesic in T_F^Z between $\beta_F^Z(x)$ and q, and since T_F^Z is quasi-isometrically embedded in $\mathcal{C}(Z)$, we may apply the bounded geodesic image axiom and the nested consistency inequality to obtain that $d_V(\beta_F^V(x), \beta_F^V(\rho_V^Z(q))) < M/2$, where as before we choose M as large as necessary. Since $\rho_V^Z(q)$ lies within κ of a leaf of T_F^V by Proposition 4.7(2) with κ depending only on $\mathfrak{S}$ and $|F|$, the $M/2$–neighborhood of $\beta_F^V(\rho_V^Z(q))$ in T_F^V must be contained in $T_{p,\sigma}^V$, and hence $\beta_F^V(x)$ is contained in $T_{p,\sigma}^V$, showing $x \in W_{p,\sigma}^V \cap W_{q,\tau}^Z$ as required.   □

4.5 Refining the subdivisions

In this section, we analyze the difference between (M, M')–evenly spaced subdivisions $\boldsymbol{p} = \wp(F)$ and $\boldsymbol{p}' = \wp(F')$ (using the fixed subdivision mechanism $\wp$ of Section 4.1). The main result here, Proposition 4.12, is that there are bounded refinements $\boldsymbol{p}_0$ and $\boldsymbol{p}_0'$ which are related by an "order-preserving" bijection $j : \boldsymbol{p}_0 \to \boldsymbol{p}_0'$ (in the sense specified below) which only moves points a bounded distance. This is where we will use the "evenly spaced" condition, which guarantees, roughly, that along corresponding edges we have the same number of partition points up to an additive, rather than multiplicative, constant.

In the next subsection, we prove that these refined subdivisions determine isomorphic cube complexes when put through our cubulation machine.

Coarse separation We will need the following short discussion about coarse separation in quasi-isometrically embedded trees. Suppose $j : T \to X$ is a (χ, χ)–quasi-isometric embedding of a tree T into a δ–hyperbolic space X, and $p \in T$ is a point inside an edge, distance more than $\mu > 0$ from the endpoints. Let $T_{p,\pm}(\mu)$ be the two components of the complement of a μ–neighborhood of p. For any ϵ there is a $\mu(\epsilon, \delta, \chi)$ such that the images $j(T_{p,\pm}(\mu))$ are at least ϵ apart from each other. We want to compare this separation for two nearby trees:

Lemma 4.11 *Let T_1 and T_2 be trees with (χ, χ)–quasi-isometric embeddings $u_i : T_i \to X$ into a δ–hyperbolic space X whose images are within Hausdorff distance ϵ. There exists $\mu_0 = \mu_0(\delta, \chi, \epsilon) > 10\epsilon$ such that for all $\mu \geq \mu_0$ the following holds.*

Suppose that $p_i \in T_i$ are points with $d(u_1(p_1), u_2(p_2)) \leq \mu$ and each p_i is in a segment e_i contained in an edge of T_i such that $u_i|_{e_i}$ is a $(1, \chi)$–quasi-isometric embedding. Moreover, assume that p_i lies at distance more than 2μ from the endpoints of e_i. Then the labels of the components can be chosen so that

- *$u_1((T_1)_{p_1,+}(2\mu))$ is in an ϵ–neighborhood of $u_2((T_2)_{p_2,+})$,*
- *$u_1((T_1)_{p_1,+}(2\mu))$ is μ–far from $u_2((T_2)_{p_2,-})$, and*
- *the same holds if we swap $+$ and $-$.*

In what follows we apply this to trees T_F^V, and so that half-trees that here would be $(T_F^V)_{p,\sigma}$ correspond to what we wrote as $T_{p,\sigma}^V$ in Section 4.1. We will use both notations, and hope not to confuse the reader.

Common refinements from close components We are now ready to prove the refinement statement. We note that this is the main point at which we use the full power of Theorem 3.3, which provides not only that each of the pairs of trees T_F^V and $T_{F'}^V$ have Hausdorff close images in $\mathcal{C}(V)$, but that they are identical on almost every component, and that their different components can be cut into large pieces where they are coarsely identical.

Proposition 4.12 *There exists $M_2 = M_2(|F|, \mathfrak{S}) > 0$ such that if $M > M_2$, there exists $R > 0$ and:*

(1) *Subsets $p_0 \subset p$ and $p'_0 \subset p'$ with*
- *$p_0(V)$ or $p'_0(V) \neq \varnothing$ only if $V \in \mathcal{U}(F) \cap \mathcal{U}(F')$, and*
- *$|p \setminus p_0|, |p' \setminus p'_0| < R$.*

(2) *A bijection $j : p_0 \to p'_0$ with $d_V(p, j(p)) < 2M/3$ for any $p \in p_0(V)$.*

(3) *For every $p \in p_0$, a bijection j_p between the half-trees defined by p and those defined by $j(p)$ with the following property. For any V and $p, q \in p_0(V)$, if the half-tree $(T_F^V)_{p,\sigma}$ contains q, then $j_p((T_F^V)_{p,\sigma})$ contains $j(q)$.*

(4) *Moreover, let f and f' be such that either $f \in F$, $f' \in F'$, and $d_{\mathcal{X}}(f, f') \leq 1$, or $f = f' = \rho_V^U$ for some $U, V \in \mathcal{U}(F) \cap \mathcal{U}(F')$ with $U \subsetneq V$ or $U \pitchfork V$. If $\beta_F^V(f) \in (T_F^V)_{p,\sigma}$, then $\beta_{F'}^V(f') \in j_p((T_F^V)_{p,\sigma})$.*

Proof We work in each $V \in \mathfrak{S}$ separately, constructing $p_0(V)$, $p'_0(V)$ and the bijection, and combine the results.

First, if $V \in \mathcal{U}(F) \triangle \mathcal{U}(F')$, then $\operatorname{diam}_V(F)$ and $\operatorname{diam}_V(F')$ are uniformly bounded and we set $p_0(V) = p'_0(V) = \varnothing$. By Proposition 2.12, there are boundedly many such V; hence this involves deleting at most boundedly many subdivision points.

Next, if $V \in \mathcal{U}(F) \cap \mathcal{U}(F')$ is not involved in the transition from F to F' (see Definition 2.13), then all of the relevant data is constant. Hence, by our formalism for choosing subdivisions (see the definition of $\wp$ in Section 4.1), we have $p(V) = p'(V)$, and we set $p_0(V) = p'_0(V) = p(V)$.

Hence we may restrict our attention to a fixed $V \in \mathcal{U}(F) \cap \mathcal{U}(F')$ for which either $\pi_V(F) \neq \pi_V(F')$ or $\mathcal{U}^V(F) \neq \mathcal{U}^V(F')$ (or both). We note that Proposition 2.14 bounds the number of such V solely in terms of n and $\mathfrak{S}$.

Fix such a V. Recall that within $\mathcal{C}(V)$ we have $F^V = \pi_V(F)$ and

$$\mathcal{Y}^V = \{\rho_V^Y \mid Y \in \mathcal{U}^V(F)\},$$

and similarly for F'^V and $\mathcal{Y}'^V$. By Proposition 2.14, $\#(\mathcal{Y}^V \triangle \mathcal{Y}'^V) < N_1$, where $N_1 = N_1(\mathfrak{S}, k) > 0$.

Recall that Theorem 3.3 provides a constant $L = L(k, \mathfrak{S}) > 0$ and the following:

(1) Stable trees with decompositions

$$T_F^V = T_c(F^V, \mathcal{Y}^V) \cup T_e(F^V, \mathcal{Y}^V), \quad T_{F'}^V = T_c(F'^V, \mathcal{Y}'^V) \cup T_e(F'^V, \mathcal{Y}'^V);$$

we write these as $T_c^V(F), T_e^V(F)$ and $T_c^V(F'), T_e^V(F')$ for short.

(2) Two stable subsets, $T_s \subset T_e^V(F)$ and $T_s' \subset T_e^V(F')$, such that

   (a) T_s and T_s' are contained in the interiors of the edges of $T_e^V(F)$ and $T_e^V(F')$, respectively;

   (b) the complements $T_e^V(F) \setminus T_s$ and $T_e^V(F') \setminus T_s'$ each have at most L components, each of which has diameter at most L;

   (c) there is a bijective correspondence between the components of T_s and T_s';

   (d) this bijective correspondence is the identity on all but at most L components, with identical components of T_s, T_s' coming from identical components of $T_e^V(F)$ and $T_e^V(F')$;

   (e) the remaining components of T_s are within Hausdorff distance L of their corresponding components in T_s'.

We assume $M > \max(4L, 4\mu_0, 4\kappa)$, where μ_0 is given by [Lemma 4.11] when $\epsilon = L$ and the quasi-isometry constants of the trees match those for T_F^V and $T_{F'}^V$, and κ is the constant in [Proposition 4.7].

Consider the sets of subdivision points $\boldsymbol{p}_1(V) = \boldsymbol{p}(V) \cap T_s$ and $\boldsymbol{p}_1'(V) = \boldsymbol{p}'(V) \cap T_s'$ which are contained in the stable subsets. Since $\boldsymbol{p}(V) \subset T_e^V(F)$ and $\boldsymbol{p}'(V) \subset T_e^V(F')$, items (2a) and (2b), and our choice of subdivision width $M > 4L$, imply $\boldsymbol{p}(V) \setminus \boldsymbol{p}_1(V)$ and $\boldsymbol{p}'(V) \setminus \boldsymbol{p}_1'(V)$ both have cardinality bounded above by L.

By item (2d), the induced subdivisions $\boldsymbol{p}_1(V)$ and $\boldsymbol{p}_1'(V)$ agree on all but at most L components of T_s and T_s', respectively, as these components are segments in the components of $T_e^V(F)$ and $T_e^V(F')$ which are equal and hence have the same subdivisions by our setup (see [Section 4.1]).

On the remaining L components we can make bounded adjustments. Let e and e' be edge components of T_s and T_s' related by the correspondence. The closest subdivision point in e to an endpoint is at most $M'/2$ away and at least M, by definition of the subdivisions, and similarly for e'. Since $d_{\mathrm{Haus}}(e, e') < L$ by item (2e), and e and e' are quasi-isometrically embedded with multiplicative constant 1 and additive constant depending on $(\mathcal{X}, \mathfrak{S})$ and n [Theorem 3.3(c)], the difference between the number of subdivision points in e and e' is bounded in terms of $M, M', L, (\mathcal{X}, \mathfrak{S})$, and n.

We write $\boldsymbol{p}_1(e) = \boldsymbol{p}_1(V) \cap e$, and similarly for $\boldsymbol{p}_1'(e')$. Note that these sets are naturally ordered once we choose endpoints of e and e', and we order e and e' by declaring those endpoints to be minimal. We choose endpoints of e and e' within distance bounded in terms of L and the constants of the quasi-isometric embeddings of e and e' in $\mathcal{C}(V)$. We will assume that M is larger than this bound.

After deleting from $\boldsymbol{p}_1(e) \cup \boldsymbol{p}'_1(e')$ a number of points bounded by this constant (all of which occur near the endpoints of e and e') and using the fact that $\boldsymbol{p}_1(e)$ and $\boldsymbol{p}'_1(e')$ are M–evenly spaced, we can obtain refinements $\boldsymbol{p}_0(e)$ and $\boldsymbol{p}'_0(e')$, which admit an order-preserving (with respect to the aforementioned order) bijection $j_e \colon \boldsymbol{p}_0(e) \to \boldsymbol{p}'_0(e')$ which satisfies

$$d_V(p, j_e(p)) < L + \frac{M}{2} + \zeta < \frac{2M}{3},$$

where $e' \to \mathcal{C}(V)$ is a $(1, \zeta)$–quasi-isometric embedding by [Theorem 3.3(c)](#), and we have chosen M sufficiently large to guarantee the inequality. Indeed, for each p we find a nearest point in e', which is at most L away, and then move along e' at most $M/2 + \zeta$ to a point of $\boldsymbol{p}'_0(e')$. For later purposes, we can assume that points in $\boldsymbol{p}_0(e)$ do not lie within $2M$ of the endpoints of e, and similarly for $\boldsymbol{p}'_0(e')$.

If we set $\boldsymbol{p}_0(V) = \bigcup_{e \in \pi_0(T_s)} \boldsymbol{p}_0(e)$ and define $\boldsymbol{p}'_0(V)$ similarly, then the j_e maps combine to give a bijection $j_V \colon \boldsymbol{p}_0(V) \to \boldsymbol{p}'_0(V)$ which moves points by distance at most $2M/3$, as required for item (2) of the proposition.

To define the map j_p between half-trees we use [Lemma 4.11](#). Namely, we pair the half-tree $(T_F^V)_{p,\sigma}$ with the half-tree $(T_{F'}^V)_{j_V(p),\tau}$ that contains $(T_F^V)_{p,\sigma}(2M)$ in its L–neighborhood.

For item (3) of the proposition, let $p, q \in \boldsymbol{p}_0(V)$. There are two cases.

Suppose first that p and q lie in the same edge-component e of T_s. Recall that we chose the bijection j_e to be order-preserving, with respect to the order along e and e' determined by choosing endpoints $e^- \in e$ and $(e')^- \in e'$ which are a small distance (less than M) apart to be minimal in the orders. Let $(T_F^V)_{p,\sigma}$ denote the half-tree of p containing e^- and let $(T_{F'}^V)_{j_V(p),\tau}$ denote the half-tree of $j_V(p)$ containing $(e')^-$. Since we have $2M$ spacing now between p and the endpoints of e, e^- is in $(T_F^V)_{p,\sigma}(2M)$, and [Lemma 4.11](#) says that there is exactly one half-tree at $j_V(p)$ which comes within M of $(T_F^V)_{p,\sigma}(2M)$. Since e^- is within M of $(e')^-$, it follows that $(T_{F'}^V)_{j_V(p),\tau}$ is in fact the paired half-tree provided by [Lemma 4.11](#), which is our $j_p((T_F^V)_{p,\sigma})$. We conclude that $q < p$ in the order along e if and only if $q \in (T_F^V)_{p,\sigma}$ and $j_V(q) < j_V(p)$ along e' if and only if $j_V(q) \in j_p((T_F^V)_{p,\sigma})$. Since j_e is order-preserving, (3) follows in this case.

Suppose now that p and q do not lie in the same edge-component of T_s. Suppose q is contained in the half-tree $(T_F^V)_{p,\sigma}$, and let $(T_{F'}^V)_{q,\tau} = j_p((T_F^V)_{p,\sigma})$. In this case we have that q lies in $(T_F^V)_{p,\sigma}(2M)$, rather than just in $(T_F^V)_{p,\sigma}$. By [Lemma 4.11](#)

there is only one half-tree of $T_{F'}^V$ at $j_V(p)$ that comes within M of $(T_F^V)_{p,\sigma}(2M)$, and said half-tree must be $(T_{F'}^V)_{j_V(p),\tau}$. Since $j_V(q)$ lies within M of q, we have $j_V(q) \in (T_{F'}^V)_{j_V(p),\tau}$, as required.

The argument for part (4) of the proposition is essentially the same as the argument for the second case of part (3), since all we used there is that the point q of T_F^V is not close to p, but it is close to a corresponding point in $T_{F'}^V$, and the analogous properties hold in all the listed cases. $\qquad\square$

4.6 Refinements give isomorphic cube complexes

Consider the refined subdivisions $\boldsymbol{p}_0$ and $\boldsymbol{p}_0'$ for F and F', respectively, that are produced by Proposition 4.12. These are (M, M'') spaced subdivisions (though no longer evenly spaced) so each of the sets of data $(F, \boldsymbol{p}_0)$ and $(F', \boldsymbol{p}_0')$ — and their associated collections of stable trees — can be plugged into our cubulation machine to produce cube complexes $\mathcal{Q}_{F,\boldsymbol{p}_0}$ and $\mathcal{Q}_{F',\boldsymbol{p}_0'}$, respectively. We also assume that $M > \max\{M_1, M_2\}$, where M_1 and M_2 are the constants from Lemma 4.10 and Proposition 4.12, respectively, along with our other base assumptions about M.

Our next result says that these cube complexes are abstractly isomorphic and admit coarsely compatible quasi-isometric embeddings into $\mathcal{X}$. Using Proposition 4.9, we will be able to conclude that the right hand side of diagram (4-1) from Theorem 4.1 commutes.

Proposition 4.13 *There exists* $M_3 = M_3(|F|, \mathfrak{S}) > 0$, *such that if* $M > M_3$, *there exists* $B > 0$ *and a cubical isomorphism* $\hat{h} \colon \mathcal{Q}_{F,\boldsymbol{p}_0} \to \mathcal{Q}_{F',\boldsymbol{p}_0'}$ *such that the diagram*

(4-4)
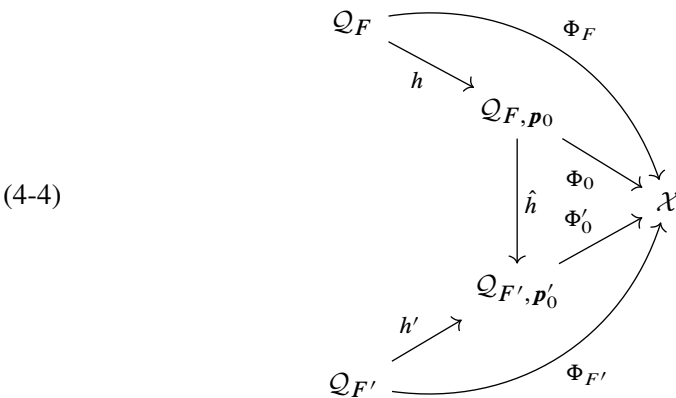
commutes up to error B.

Proof We will define the required cubical isomorphism $\hat{h}$ and then prove that the middle triangle commutes up to bounded error. This suffices for the proposition because Proposition 4.9 says that the top and bottom triangles commute up to bounded error.

In order to use Lemma 2.3 to define an isomorphism between $\mathcal{Q}_{F,\boldsymbol{p}_0}$ and $\mathcal{Q}_{F',\boldsymbol{p}_0'}$, we need a bijection between the corresponding collections of half-spaces which preserves complements and disjointness.

Let $\mathcal{W}_{F,\boldsymbol{p}_0}$ and $\mathcal{W}_{F',\boldsymbol{p}_0'}$ be the sets of half-spaces and construct a bijection

$$\iota \colon \mathcal{W}_{F,\boldsymbol{p}_0} \to \mathcal{W}_{F',\boldsymbol{p}_0'}$$

as follows. By Proposition 4.12, we need only consider $V \in \mathcal{U}(F) \cap \mathcal{U}(F')$. Any $p \in \boldsymbol{p}_0(V)$ is contained in an edge of T_F^V (in fact in T_e^V) and is at least M from its endpoints.

Let $j \colon \boldsymbol{p}_0 \to \boldsymbol{p}_0'$ be the bijection provided by Proposition 4.12, together with the corresponding maps j_p pairing the half-trees at $p \in \boldsymbol{p}_0$ with those at $j(p)$.

To define ι, let $p \in \boldsymbol{p}_0$ and $\sigma \in \{\pm\}$. If $j_p(T_{p,\sigma}^V) = T_{j_V(p),\tau}^V$ we define $\iota(W_{p,\sigma}^V) = W_{j_V(p),\tau}^V$. It is straightforward to confirm that ι respects complementation as in condition (1) of Lemma 2.3.

We now confirm condition (2) of Lemma 2.3 for ι by using the various characterizations of half-space intersections given in Lemma 4.10. We remark that the figures from the proof of Lemma 4.10 are again relevant.

Let $p \in \boldsymbol{p}_0(V)$ and $q \in \boldsymbol{p}_0(Z)$ for $Z, V \in \mathcal{U}(F) \cap \mathcal{U}(F')$, and suppose that the half-spaces corresponding to the half-trees $T_{p,\sigma}^V$ and $T_{q,\tau}^Z$ intersect nontrivially, where $\sigma, \tau \in \{\pm\}$. There are five cases, up to switching the roles of V and Z.

Case $Z \perp V$ This case follows immediately from the construction and Lemma 4.10(1), since all relevant pairs of half-spaces intersect in this case.

Case $Z = V$ In this case, Lemma 4.10(2) implies that $T_{p,\sigma}^V \cap T_{q,\tau}^V \neq \varnothing$ (recall Figures 13 and 14).

In particular, up to switching the roles of p and q, we have $q \in T_{p,\sigma}^V$. But then, in view of Proposition 4.12(3), $j(q) \in j_p(T_{p,\sigma}^V)$. In particular, $j_p(T_{p,\sigma}^V) \cap j_q(T_{q,\tau}^V) \neq \varnothing$, so that the corresponding half-spaces intersect, again by Lemma 4.10(2).

Case $Z \pitchfork V$ In this case, Lemma 4.10(3) implies, up to switching the roles of Z and V, that $(T_F^V)_{p,\sigma}$ contains $\beta_F^V(\rho_V^Z)$ (recall Figure 15). It follows immediately from part (4) of Proposition 4.12 that $\beta_{F'}^V(\rho_V^Z) \in j_p(T_{p,\sigma}^V)$, as required.

Case $V \subsetneq Z$ By Lemma 4.10, there are two subcases, up to switching the roles of V and Z: (a) when $T_{q,\tau}^Z$ contains $\beta_F^Z(\rho_Z^V)$, and (b) when $T_{p,\sigma}^V$ contains $\beta_F^V(\rho_V^Z(q))$.

In case (a), since $V \in \mathcal{U}(F) \cap \mathcal{U}(F')$ part (4) of Proposition 4.12 immediately gives that $\beta_{F'}^Z(\rho_Z^V)$ lies in $j_q(T_{q,\tau}^Z)$.

Suppose now that (b) holds. We prove that $\beta_{F'}^V(\rho_V^Z(j(q)))$ is contained in $j_p(T_{p,\sigma}^V)$ (recall Figure 17). Recall from Proposition 4.7 that $\rho_V^Z(q)$ lies close to some $\pi_V(f)$ with bound in terms of $\mathfrak{S}$ and $|F|$. Since $d_V(q, j(q)) < 2M/3$ and both q and $j(q)$ are at least $M - 2K$ from ρ_Z^V, with the $2K$ coming from the facts that $T_c^Z(F)$ is within Hausdorff distance K of the $\mathcal{Y}^Z$ and the edge of $T_e^Z(F)$ containing q is $(1, K)$–quasi-isometrically embedded in $\mathcal{C}(Z)$ (Theorem 3.3(d)), choosing M sufficiently large, we can guarantee that any geodesic in $\mathcal{C}(V)$ between p and $j_V(p)$ avoids the κ_0–neighborhood of ρ_Z^V, for κ_0 the constant of the bounded geodesic image property, which then bounds $d_V(\rho_V^Z(q), \rho_V^Z(j_Z(q))) < \kappa_0$. Hence $\rho_V^Z(j_Z(q))$ is close to both $\rho_V^Z(q)$ and a leaf of $T_{F'}^V$, and thus $\beta_{F'}^V(\rho_V^Z(j(q))) \subset j_p(T_{p,\sigma}^V)$, as required.

Since the wallspace map ι satisfies the conditions of Lemma 2.3, we obtain a cubical isomorphism $\hat{h} \colon \mathcal{Q}_{F,\boldsymbol{p}_0} \to \mathcal{Q}_{F',\boldsymbol{p}_0'}$.

Commutativity It remains to prove that

$$\Phi_{F,\boldsymbol{p}_0} \colon \mathcal{Q}_{F,\boldsymbol{p}_0} \to \mathcal{X} \quad \text{and} \quad \Phi_{F',\boldsymbol{p}_0'} \circ h \colon \mathcal{Q}_{F,\boldsymbol{p}_0} \to \mathcal{X}$$

are the same up to a bounded error depending only on k and the ambient HHS structure.

By the distance formula (Theorem 2.6), it suffices to show that for each 0–cube $x \in \mathcal{Q}_{F,\boldsymbol{p}_0}$, its respective images $\Phi_{F,\boldsymbol{p}_0}(x)$ and $\Phi_{F',\boldsymbol{p}_0'} \circ h(x)$ have coarsely the same projections to $\mathcal{C}(V)$ for each $V \in \mathcal{U}(F) \cap \mathcal{U}(F')$.

Recall from the end of Section 4.2 that the maps $\Phi_{F,\boldsymbol{p}_0}$ and $\Phi_{F',\boldsymbol{p}_0'}$ are defined domainwise by intersecting certain collections of half-trees of T_F^V and $T_{F'}^V$, for each $V \in \mathfrak{S}$, and hence the same is true for $\Phi_{F',\boldsymbol{p}_0'} \circ h$. By chasing the relevant definitions, of h and j especially, we see that the collections of half-trees involved in the definition of $\Phi_{F,\boldsymbol{p}_0}(x)$ and $\Phi_{F',\boldsymbol{p}_0'} \circ h(x)$ are in bijection with each other, with corresponding half-trees lying within bounded Hausdorff distance depending only on k and the ambient HHS structure. Hence their intersections in T_F^V and $T_{F'}^V$ coarsely coincide, as required. This completes the proof of the proposition. $\qquad\square$
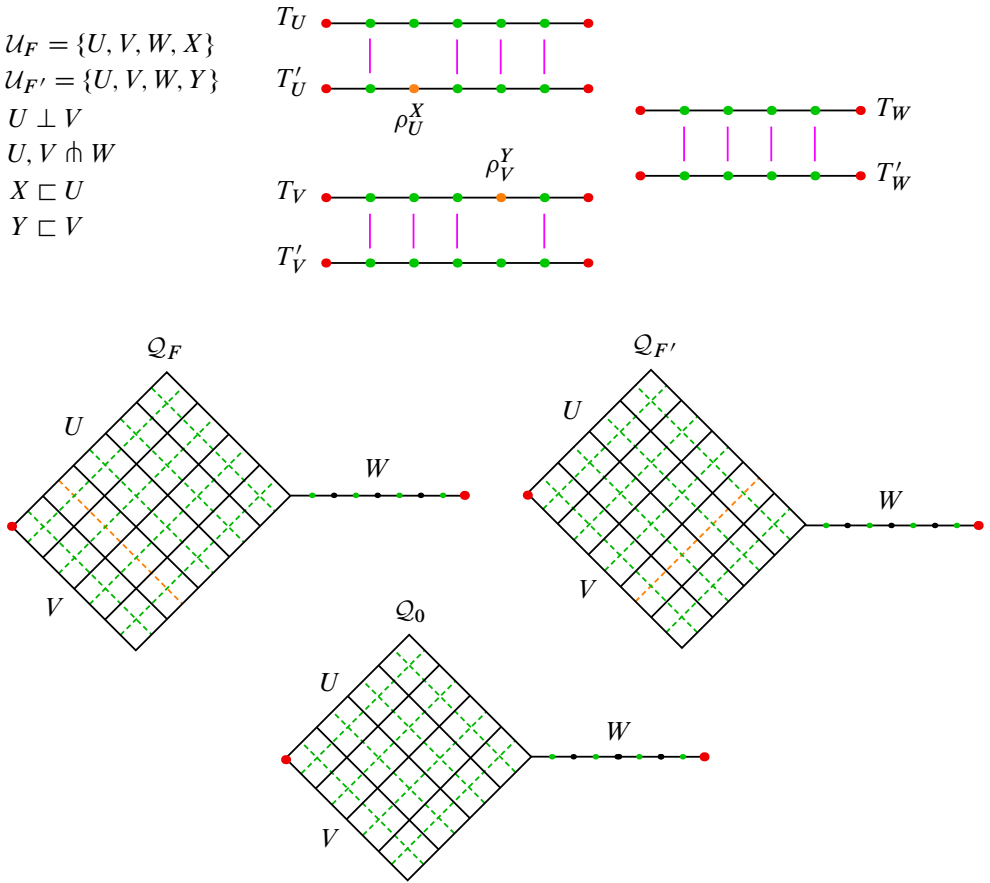
Figure 18: A simple example of how subdivision bijections determine hyperplane deletions. The relevance of X for F' but not F requires deleting a subdivision point to obtain the bijection $j_U \colon \boldsymbol{p}_U \to \boldsymbol{p}'_U$ (indicated in pink), and similarly with Y for $j_V \colon \boldsymbol{p}_V \to \boldsymbol{p}'_V$. Deleting subdivision points results in hyperplane deletions when passing from $\mathcal{Q}_F$ and $\mathcal{Q}_{F'}$ to $\mathcal{Q}_0$. Note that since $X, Y \notin \mathcal{U}_F \cap \mathcal{U}_{F'}$, neither domain determines any subdivision points.

4.7 Proof of Theorem 4.1

Let F and F' be as in the statement.

The CAT(0) cube complexes $\mathcal{Q}_F$, $\mathcal{Q}_{F'}$ are constructed in Section 4.2, using subdivisions $\boldsymbol{p} = \wp(F)$, $\boldsymbol{p}' = \wp(F')$ within each of the relevant stable trees produced in Section 3. Items (1) and (2) are proven in [8], as recalled in Theorem 4.6 above.

For (3), we may define a map $\psi_F \colon F \to \mathcal{Q}_F$ as in [8, Proof of Theorem 2.1]. For each $f \in F$, let $\psi_F(f)$ be the orientation of the walls on hull $H_\theta(F)$ obtained by choosing, for each wall $(W^V_{p,+}, W^V_{p,-})$, the halfspace containing f. We define $\psi_{F'}$ similarly. That ψ_F and $\psi_{F'}$ satisfies (3) now follows from Theorem 4.6.

We now prove the stability statements. We will only consider the case that g is the identity for the same reason as in Remark 3.4, namely that what we have to prove is that the choices that we made along the way only affect the output in the way predicted by the statement, and such choices can be "translated" by automorphisms. The choices we are referring to are those of the trees T^V_F, of the evenly spaced subdivisions, and of points projecting coarsely in specified places in the various $\mathcal{C}(Y)$, as in Theorem 4.6.(2).

Let $\boldsymbol{p}_0 \subset \boldsymbol{p}$ and $\boldsymbol{p}'_0 \subset \boldsymbol{p}'$ be the refinements provided by Proposition 4.12, where at most N subdivision points are deleted.

The cube complex $\mathcal{Q}_0$ as in the statement of Theorem 4.1 can be taken to be either of the cube complexes $\mathcal{Q}_{F,\boldsymbol{p}_0}$ or $\mathcal{Q}_{F',\boldsymbol{p}'_0}$ that are produced by Proposition 4.13. So take $\mathcal{Q}_0 = \mathcal{Q}_{F',\boldsymbol{p}'_0}$ and let $\Phi_0 \colon \mathcal{Q}_0 \to \mathcal{X}$ be the map $\Phi_{F',\boldsymbol{p}'_0} \colon \mathcal{Q}_{F',\boldsymbol{p}'_0} \to \mathcal{X}$ given by Proposition 4.13. Finally, let $\eta' \equiv h' \colon \mathcal{Q}_{F'} \to \mathcal{Q}_0$ be as in Proposition 4.13 and define $\eta \colon \mathcal{Q}_F \to \mathcal{Q}_0$ by $\eta = \hat{h} \circ h$.

By Proposition 4.13, these maps each satisfy the required properties and the right part of diagram (4-1) commutes up to bounded error, as required.

To see exact commutativity of the left part of the diagram, it remains to prove that $\eta \circ \psi_F \circ \iota_F = \eta' \circ \psi_{F'} \circ \iota_{F'}$. Recall that $\psi_F \colon F \to \mathcal{Q}_F$ is defined by sending f to the coherent orientation on the wallspace defined by $\boldsymbol{p}$ which, for each $p \in \boldsymbol{p}(V)$, chooses the half-tree of $T^V_F \setminus p$ containing $\beta^V_F(f)$, for each $V \in \mathcal{U}(F)$. Since h is a hyperplane deletion map, $h \circ \psi_F$ makes, for $p \in \boldsymbol{p}_0$, the same choice as ψ_F.

Fix $f \in F$ (the argument for $f \in F'$ is similar). Then the two sides of the equation are coherent orientations on the wallspace defining $\mathcal{Q}_{F',\boldsymbol{p}'_0}$, and we have to check that they coincide on every halfspace. Pick $p \in \boldsymbol{p}_0$ and let $p' = j_V(p)$, the map defined in Proposition 4.12. As above the orientation of $h \circ \psi_F \circ \iota_F(f) = h \circ \psi_F(f)$ on the wall associated to p is the one that chooses the half-tree of $T^V_F \setminus p$, call it $(T^V_F)_{p,+}$, that contains $\beta^V_F(f)$. The map $\hat{h}$, by the construction in Proposition 4.13, takes this to the orientation that chooses the half-tree of p' in $T^V_{F'}$ given by $j_p((T^V_F)_{p,+})$ (where j_p is the bijection of half-trees provided by Proposition 4.12). Letting $f' = \iota_{F'}(f)$ we have $d_{\mathcal{X}}(f', f) \leq 1$, so part (4) of Proposition 4.12 tells us that $\beta^V_{F'}(f') \in j_p((T^V_F)_{p,+})$,

which means that $j_p((T_F^V)_{p,+})$ is the half-tree selected by $\eta' \circ \psi_{F'}(f')$. Thus we conclude that $\eta(\psi_F(\iota_F(f))) = \eta'(\psi_{F'}(\iota_{F'}(f)))$.

This completes the proof of Theorem 4.1.   □

5 Generalizing normal paths to find barycenters

In this section, we describe a variation on the "normal paths" construction due to Niblo and Reeves [46]. We remark that they used these normal paths to build a biautomatic structure for any cubical group, with the paths playing the central role of the bicombing in that structure.

For the case of two points, our construction gives a "symmetrized" version of the Niblo–Reeves construction. The main difference with their construction, however, is that ours is adapted to allow for multiple points, which we need for our barycenter application (Theorem E).

The reader may want to refer to Section 2.1 for the various notions and notations relating to cube complexes that we will use throughout this section

Let X be a CAT(0) cube complex, $\mathcal{H}$ its set of hyperplanes, and $f \colon P \to X^{(0)}$ a (not necessarily injective) map from a finite set P into the vertices of X. Roughly, we will find a barycenter for the set $f(P)$ in X by an iterative sequence of contractions which behaves stably under hyperplane deletions. The main result of this section, and the only statement from this section that we need to prove our main theorems, is the following:

Theorem 5.1 *Let X be a CAT(0) cube complex, and let $\mathcal{H}$ be its set of hyperplanes. For each $f \colon P \to X^{(0)}$, where P is a finite set, there is a finite sequence*

$$\{f_i \colon P \to X^{(0)} \mid i = 0, \ldots, n = n_f\}$$

with the following properties:

(1) $f_0 = f$ *and* $\mathrm{diam}_\infty(f_n(P)) \le 1$.

(2) *For each* $p \in P$ *and* $0 \le i \le n-1$ *we have* $d_\infty(f_i(p), f_{i+1}(p)) \le 1$.

(3) *For each* p, *there is an* ℓ^1–*geodesic going through the vertices*

$$f_0(p), f_1(p), \ldots, f_n(p)$$

in this order.

(4) *No hyperplane separates every point of* $f(P)$ *from a point in* $f_n(P)$.

(5) If $g\colon Q \to P$ is surjective, then $f_i \circ g = (f \circ g)_i$ for all i.

(6) If G is any hyperplane of X, the hyperplane deletion map $\mathrm{Res}_{\mathcal{H}\backslash G}\colon X \to X(\mathcal{H}\backslash G)$ satisfies

$$|n_f - n_{\mathrm{Res}_{\mathcal{H}\backslash G}\circ f}| \leq 1 \quad \text{and} \quad d_\infty\big(\mathrm{Res}_{\mathcal{H}\backslash G}(f_i(p)), (\mathrm{Res}_{\mathcal{H}\backslash G} \circ f)_i(p)\big) \leq 1.$$

Recall that d_∞ is the metric generated by the sup metric on each cube in the ambient complex.

The proof of Theorem 5.1 occurs in parts over the remainder of this section. We tie them together in Section 5.4 below.

We will mostly ignore the ambient cube complexes and focus on the hyperplane set $\mathcal{H}$ and regard maps f as above as maps associating to $p \in P$ an orientation on $\mathcal{H}$. For each pair $(f, \mathcal{H})$ we consider a number of operations.

First, let $\mathcal{H}_f$ be the set of hyperplanes of X that *separate* $f(P)$. That is, $\mathcal{H}_f$ is the set of hyperplanes $H \in \mathcal{H}$ for which there exist $p, p' \in P$ such that $f(p)$ and $f(p')$ are on different sides of H.

Let

$$\mathrm{Trim}(f, \mathcal{H}) = (\mathrm{Res}_{\mathcal{H}_f}(f), \mathcal{H}_f)$$

be the restriction to $\mathcal{H}_f$. Note that the quotient complex $X(\mathcal{H}_f)$ actually embeds in $X(\mathcal{H})$ (with image the subcomplex spanned by all vertices that lie in the intersection of all the halfspaces of $X(\mathcal{H})$ that contain $f(P)$), and that it is finite even if $X(\mathcal{H})$ is not.

If G is a collection of mutually crossing hyperplanes, we let

$$\mathrm{del}_G(f, \mathcal{H}) = (\mathrm{Res}_{\mathcal{H}\backslash G}(f), \mathcal{H} \backslash G)$$

This "deletion map" corresponds to composing f with the quotient by G; that is,

$$\mathrm{Res}_{\mathcal{H}\backslash G} \circ f\colon P \to X(\mathcal{H} \backslash G).$$

We will also write $\mathrm{Res}_{\mathcal{H}\backslash G}(f)$ as $f|_{\mathcal{H}\backslash G}$, in a slight abuse of notation.

5.1 Extremal and transitional hyperplanes

As stated above, our generalized normal paths give a series of contractions of our set $f(P)$ to a bounded diameter set in X. This is accomplished by iteratively jumping the points of $f(P)$ (and their subsequent contracted images) over a sequence of hyperplanes in $\mathcal{H}_f$. Thus we are led to understand which hyperplanes are next in line to be jumped.
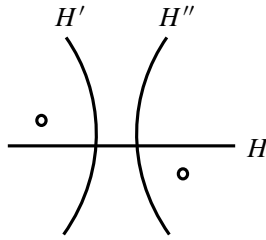
Figure 19: H is extremal but not transitional, while H' and H'' are both transitional.

Definition 5.2 Let X be a cube complex and $\mathcal{H}$ its set of hyperplanes.

- A point $p \in X$ is *adjacent* to a hyperplane $H \in \mathcal{H}$ if there are no hyperplanes separating p from H.

- A hyperplane $H \in \mathcal{H}_f$ is *extremal* if on one side of H every point of $f(P)$ is adjacent to H in X. We let $\boldsymbol{E}(f, \mathcal{H}) \subset \mathcal{H}_f$ denote the set of extremal hyperplanes.

- A hyperplane $H \in \boldsymbol{E}(X, f)$ is *transitional* if it is extremal and on one side of H *not* every point of $f(P)$ is adjacent to H; we let $\boldsymbol{T}(f, \mathcal{H}) \subset \boldsymbol{E}(f, \mathcal{H})$ be the set of transitional hyperplanes.

If H is transitional, we write $P = P_0(H) \sqcup P_1(H) = P_0 \sqcup P_1$ where $f(P_0)$ is adjacent to H on one side, $f(P_1)$ is on the other side, and at least one point of $f(P_1)$ is not adjacent to H.

We note that $\boldsymbol{E}(f, \mathcal{H})$ is always nonempty when P is nonempty. In fact, for any $p \in P$, it is readily shown that H is extremal whenever H is a hyperplane in $\mathcal{H}_f$ such that the number of hyperplanes separating $f(p)$ from H is maximal.

Moreover, for any hyperplane $H \in \boldsymbol{E}(f, \mathcal{H}) \setminus \boldsymbol{T}(f, \mathcal{H})$, every point of $f(P)$ must be adjacent to H.

Lemma 5.3 *The set $f(P)$ is contained in a single cube of X if and only if $\boldsymbol{T}(f, \mathcal{H}) = \varnothing$.*

Proof First of all, we observe that $f(P)$ being contained in a single cube is equivalent to $\mathcal{H}_f$ being mutually crossing. Moreover, it is clear from the definitions that if $\mathcal{H}_f$ is mutually crossing, then $\boldsymbol{T}(f, \mathcal{H}) = \varnothing$.

In the other direction, we suppose that $f(P)$ is not contained in a single cube and then produce a transitional hyperplane. The fact that $f(P)$ is not contained in a single cube implies that there exists $p \in P$ and a hyperplane $H \in \mathcal{H}_f$ with $f(p)$ not adjacent to H.
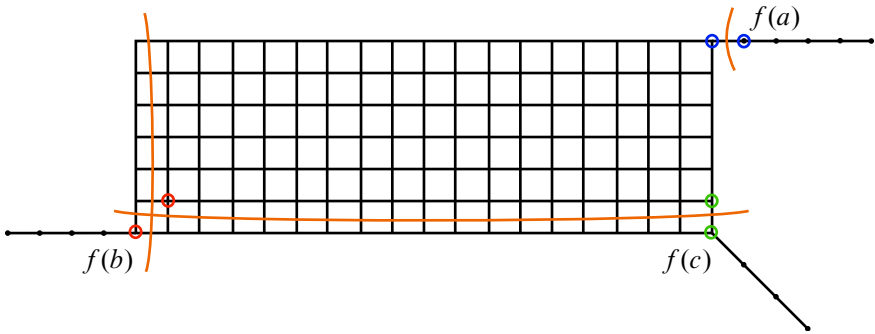
Figure 20: An example of a single move. $f(P)$ consists of three points, and the points of $f_1(P)$, in matching colors, are on the other side of the transitional hyperplanes, which are indicated in orange.

Choose now $p \in P$ and $H \in \mathcal{H}_f$ so that the number of hyperplanes separating $f(p)$ from H is maximal; the argument above implies that the number of these hyperplanes is positive, so $f(p)$ is not adjacent to H. As observed before the lemma, it follows that H is extremal, and since $f(p)$ is not adjacent to H, we have that H is in fact transitional, as required. $\qquad\square$

5.2 The move sequence

We now build our sequence of contractions, which we call moves.

Roughly speaking, a *move* is an operation on $(f, \mathcal{H})$ in which, for every $H \in \boldsymbol{T}(f, \mathcal{H})$, the points of $P_0(H)$ cross H to the opposite side. The resulting pair

$$(f_1, \mathcal{H}) = \mathrm{Move}(f, \mathcal{H})$$

is a map for which $\mathcal{H}_{f_1} = \mathcal{H}_f \setminus \boldsymbol{T}(f, \mathcal{H})$, so the image of the new map $f_1(P)$ is strictly contained within the subcomplex spanned by $f(P)$.

To define f_1, we need the following observation:

Lemma 5.4 *For each $p \in P$, the set*

$$J(p) = \{H \in \boldsymbol{T}(f, \mathcal{H}) \mid p \in P_0(H)\}$$

is mutually crossing.

Proof Suppose, by way of contradiction, that $H_1, H_2 \in J(p)$ and that H_1 does not cross H_2. Let s_1 be the side of H_1 on which $f(P_0(H_1))$ lies (and hence is adjacent)
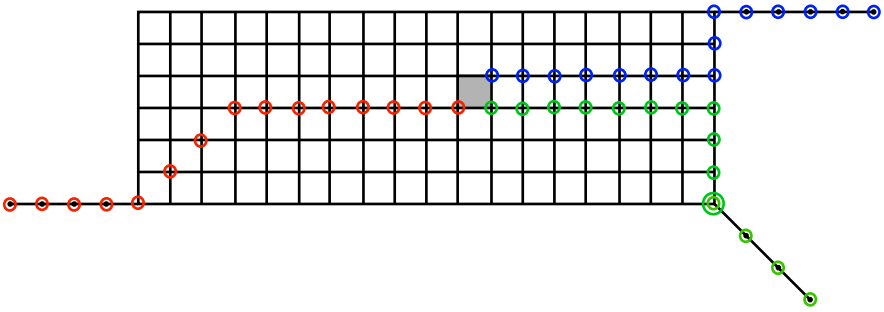
Figure 21: An example of a move sequence terminating in all points lying in a single cube. Concentric circles around a vertex indicate, in this case, a point p with $f_i(p) = f_{i+1}(p)$.

and let s_2 be the corresponding side for H_2. Since $p \in P_0(H_1) \cap P_0(H_2)$, we have that s_1 and s_2 intersect. Since H_1 does not cross H_2, up to swapping indices we have that H_2 lies in s_1 and separates H_1 from $f(P_0(H_2))$, which contains p and thus contradicts the choice of s_1.   □

We can now define $f_1(p)$. By Lemma 5.4, $f(p)$ is the corner of a unique maximal cube, each of whose midcubes is contained in some hyperplane of $J(p)$, and we can choose $f_1(p)$ to be the diagonally opposite corner. Equivalently, $f_1(p)$ can be defined as the point obtained by flipping the orientation that $f(p)$ gives to all the hyperplanes in $J(p)$. Either definition gives the new map f_1.

Note that there may be p for which $J(p) = \varnothing$, and for these $f_1(p) = f(p)$. Moreover, since the operation moves all points of $P_0(H)$ across H for each $H \in \boldsymbol{T}(f, \mathcal{H})$, we have the following consequence:

Lemma 5.5 $\mathcal{H}_{f_1} = \mathcal{H}_f \setminus \boldsymbol{T}(f, \mathcal{H}).$

Now we are ready to consider the *move sequence* of $(f, \mathcal{H})$. For $i \geq 0$ define

$$(f_i, \mathcal{H}) = \text{Move}^i(f, \mathcal{H}).$$

Lemma 5.6   *The move sequence* $\{f_i\}$ *is eventually constant, with constant image all contained in one cube.*

Proof   By Lemma 5.5, the set $\mathcal{H}_{f_i}$ becomes strictly smaller for each i as long as $\boldsymbol{T}(f_i, \mathcal{H}) \neq \varnothing$, so after a finite number of steps we must have $\boldsymbol{T}(f_n, \mathcal{H}) = \varnothing$. Thereafter, the f_i are all the same and their images are all contained in a cube by Lemma 5.3.   □
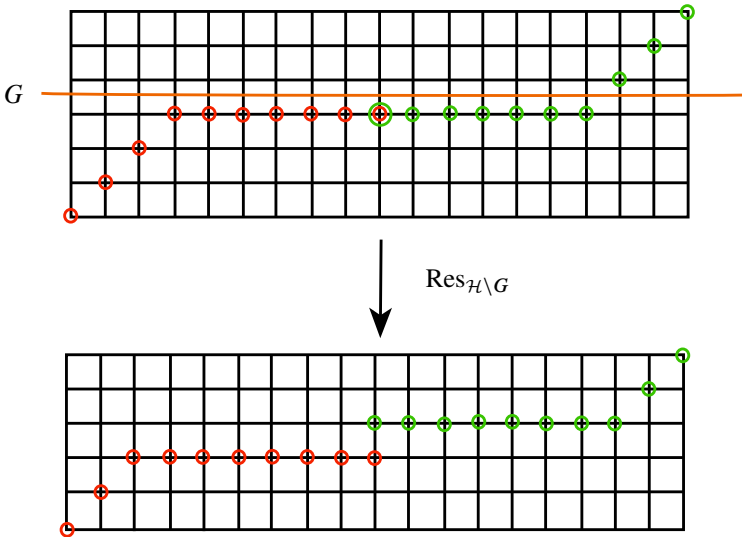
Figure 22: An illustration of Proposition 5.7, where P has two points.

Our main result about this sequence is its stability under the operation del_G. This is part of Theorem 5.1, which we rephrase here in our new language:

Proposition 5.7 *Let G be a mutually crossing set in $\mathcal{H}_f$. Let*

$$(f_i, \mathcal{H}) = \mathrm{Move}^i(f, \mathcal{H})$$

and

$$(f_i', \mathcal{H} \setminus G) = \mathrm{Move}^i(\mathrm{del}_G(f, \mathcal{H}))$$

be the move sequences for $(f, \mathcal{H})$ and $\mathrm{del}_G(f, \mathcal{H})$. Then for each $p \in P$,

$$d_\infty(\mathrm{Res}_{\mathcal{H} \setminus G} \circ f_i(p), f_i'(p)) \leq 1.$$

The proposition is in fact a generalization of what we need, since it deals with a mutually crossing set.

5.3 Proof of stability of move sequences

We begin by studying the structure of the extremal and transitional hyperplanes for a pair $(f, \mathcal{H})$, and the way in which they are affected by hyperplane deletions.

Lemma 5.8 *Every $J \in E(f, \mathcal{H}) \setminus T(f, \mathcal{H})$ crosses every hyperplane in $\mathcal{H}_f \setminus \{J\}$.*

Proof Since J is extremal but not transitional, $f(P)$ is adjacent to it on both sides; see Figure 19. This means that no other hyperplane can separate any $f(p)$ from J, and this implies that every $H \in \mathcal{H}_f \setminus \{J\}$ crosses J. □

The next lemma explains how the extremal and transitional hyperplanes change after a deletion step:

Lemma 5.9 *Let G be a mutually crossing hyperplane set in $\mathcal{H}_f$. Then*

$$(5\text{-}1) \qquad\qquad \boldsymbol{E}(f, \mathcal{H}) \setminus G \subset \boldsymbol{E}(\mathrm{del}_G(f, \mathcal{H}))$$

and

$$(5\text{-}2) \qquad \boldsymbol{E}(\mathrm{del}_G(f, \mathcal{H})) \setminus \boldsymbol{E}(f, \mathcal{H}) = \boldsymbol{T}(\mathrm{del}_G(f, \mathcal{H})) \setminus \boldsymbol{T}(f, \mathcal{H}).$$

Moreover, if $G \cap \boldsymbol{E}(f, \mathcal{H}) = \varnothing$, then

$$(5\text{-}3) \qquad\qquad \boldsymbol{E}(f, \mathcal{H}) = \boldsymbol{E}(\mathrm{del}_G(f, \mathcal{H}))$$

and

$$(5\text{-}4) \qquad\qquad \boldsymbol{T}(f, \mathcal{H}) = \boldsymbol{T}(\mathrm{del}_G(f, \mathcal{H})).$$

Proof The inclusion (5-1) is clear from the definitions.

For (5-2), if $J \in \boldsymbol{E}(\mathrm{del}_G(f, \mathcal{H})) \setminus \boldsymbol{E}(f, \mathcal{H})$, then $f(P)$ is not adjacent to J on either side, but on at least one side the only hyperplanes separating J from $f(P)$ are in G. In fact this happens on exactly one side since G is a mutually crossing set and its members cannot be separated by J. In particular this means $J \in \boldsymbol{T}(\mathrm{del}_G(f, \mathcal{H}))$, and therefore $J \in \boldsymbol{T}(\mathrm{del}_G(f, \mathcal{H})) \setminus \boldsymbol{T}(f, \mathcal{H})$ since $\boldsymbol{T}(f, \mathcal{H}) \subset \boldsymbol{E}(f, \mathcal{H})$. This situation is indicated in Figure 23, left.

Conversely if $J \in \boldsymbol{T}(\mathrm{del}_G(f, \mathcal{H})) \setminus \boldsymbol{T}(f, \mathcal{H})$, then either $f(P)$ is not adjacent to J on either side, in which case we are in the same situation as above, or $f(P)$ is adjacent to J on both sides. But in the latter case this adjacency remains true after deletion of G, which contradicts $J \in \boldsymbol{T}(\mathrm{del}_G(f, \mathcal{H}))$.

Thus, $J \in \boldsymbol{E}(\mathrm{del}_G(f, \mathcal{H})) \setminus \boldsymbol{E}(f, \mathcal{H})$ if and only if $J \in \boldsymbol{T}(\mathrm{del}_G(f, \mathcal{H})) \setminus \boldsymbol{T}(f, \mathcal{H})$, which gives (5-2).

In the description of $J \in \boldsymbol{E}(\mathrm{del}_G(f, \mathcal{H})) \setminus \boldsymbol{E}(f, \mathcal{H})$, we note that the hyperplanes of G separating J from $f(P)$ cannot themselves be separated from $f(P)$ (on the side not containing J) by any other hyperplanes, because this would contradict $J \in \boldsymbol{E}(\mathrm{del}_G(f, \mathcal{H}))$.
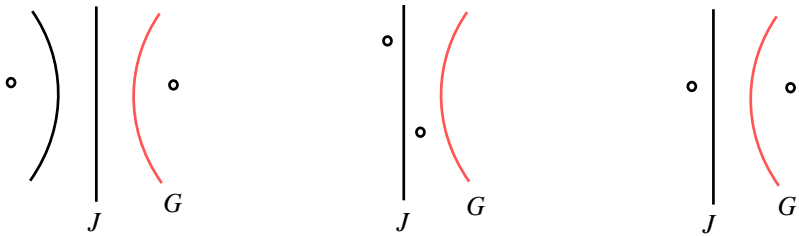
Figure 23: Changes caused by deleting G. Left, J is not in $E(f, \mathcal{H})$ but is in $T(\mathrm{del}_G(f, \mathcal{H}))$. Middle, J is in $E \setminus T$ both before and after. Right, J is in $T(f, \mathcal{H})$ but not in $T(\mathrm{del}_G(f, \mathcal{H}))$.

Thus those hyperplanes are themselves extremal. We conclude that, if $G \cap E(f, \mathcal{H}) = \varnothing$, then $E(\mathrm{del}_G(f, \mathcal{H})) \setminus E(f, \mathcal{H}) = \varnothing$, giving (5-3).

Finally to show (5-4) when $G \cap E(f, \mathcal{H}) = \varnothing$, note first that (5-2) and (5-3) imply that $T(\mathrm{del}_G(f, \mathcal{H})) \subset T(f, \mathcal{H})$. Now if $J \in T(f, \mathcal{H}) \setminus T(\mathrm{del}_G(f, \mathcal{H}))$, then on the side of J where $f(P)$ is not adjacent there must only be hyperplanes of G separating J from $f(P)$, whose deletion makes $f(P)$ adjacent on that side; see Figure 23, right. But this contradicts the assumption that $G \cap E(f, \mathcal{H}) = \varnothing$. □

We can now obtain the following lemma, which in the simplest case shows that moves and deletions commute.

Lemma 5.10 *In the notation of Proposition 5.7, if $G \cap E(f, \mathcal{H}) = \varnothing$, then the following diagram commutes:*

$$
\begin{array}{ccc}
(f, \mathcal{H}) & \xrightarrow{\ \mathrm{del}_G\ } & (f', \mathcal{H} \setminus G) \\
\Big\downarrow{\scriptstyle \text{Move}} & & \Big\downarrow{\scriptstyle \text{Move}} \\
(f_1, \mathcal{H}) & \xrightarrow{\ \mathrm{del}_G\ } & (f_1', \mathcal{H} \setminus G)
\end{array}
$$

Proof By Lemma 5.9, $T(f, \mathcal{H}) = T(\mathrm{del}_G(f, \mathcal{H}))$. This means that the Move operation on both $(f, \mathcal{H})$ and $(f', \mathcal{H} \setminus G)$ affects exactly the same set of hyperplanes, and in exactly the same way. That is, for $J \in T(f, \mathcal{H})$, the subset $P_0 \subset P$ whose f–image is on the adjacent side of J is also the subset whose f'–image is on the adjacent side, since the deletion of G does not affect this. The lemma follows. □

Now we consider the general situation, where some hyperplanes of G may be in $E(f, \mathcal{H})$.
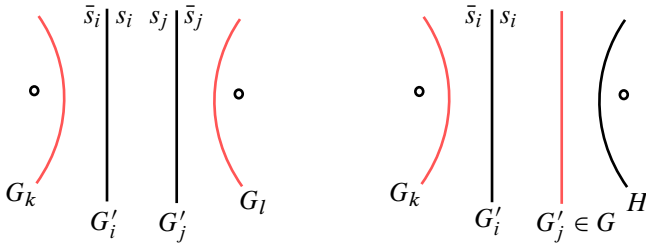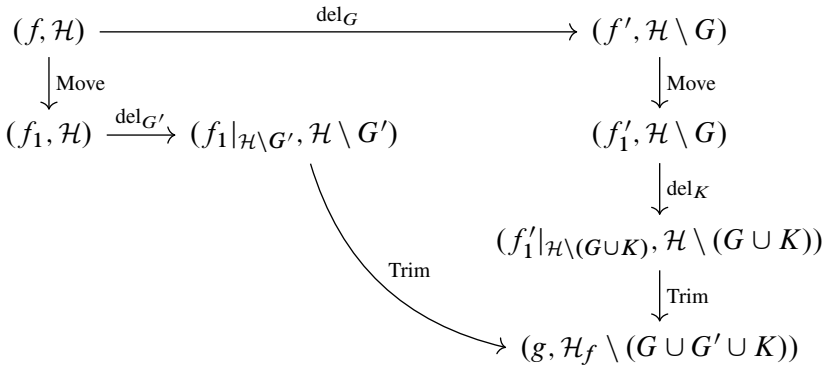
Figure 24: The contradictions arising from hyperplanes in G' not being mutually crossing.

Lemma 5.11 *Let G be a mutually crossing set in $\mathcal{H}$ and define*

$$G' = (G \cup \mathbf{T}(\mathrm{del}_G(f, \mathcal{H}))) \setminus \mathbf{T}(f, \mathcal{H}) \quad \text{and} \quad K = \mathbf{T}(f, \mathcal{H}) \setminus (G \cup \mathbf{T}(\mathrm{del}_G(f, \mathcal{H}))).$$

Then G' is a mutually crossing set, every hyperplane $H \in K$ crosses all hyperplanes of $\mathcal{H}_f \setminus (G \cup \{H\})$, and there exists a map $g \colon P \to X(\mathcal{H}_f \setminus (G \cup G' \cup K))$ such that the following diagram commutes:

$$
\begin{array}{ccc}
(f, \mathcal{H}) & \xrightarrow{\quad \mathrm{del}_G \quad} & (f', \mathcal{H} \setminus G) \\
\downarrow{\scriptstyle \text{Move}} & & \downarrow{\scriptstyle \text{Move}} \\
(f_1, \mathcal{H}) \xrightarrow{\mathrm{del}_{G'}} (f_1|_{\mathcal{H} \setminus G'}, \mathcal{H} \setminus G') & & (f_1', \mathcal{H} \setminus G) \\
& & \downarrow{\scriptstyle \mathrm{del}_K} \\
& & (f_1'|_{\mathcal{H} \setminus (G \cup K)}, \mathcal{H} \setminus (G \cup K)) \\
& {\scriptstyle \text{Trim}} & \downarrow{\scriptstyle \text{Trim}} \\
& & (g, \mathcal{H}_f \setminus (G \cup G' \cup K))
\end{array}
$$

Proof First we prove that $G' = \{G_1', \ldots, G_k'\}$ is a mutually crossing set. Suppose that G_i' and G_j' do not cross. Then they cannot both be in G by hypothesis.

Assume first that G_i' and G_j' are in $\mathbf{T}(\mathrm{del}_G(f, \mathcal{H})) \setminus \mathbf{T}(f, \mathcal{H})$, which is the same as $\mathbf{E}(\mathrm{del}_G(f, \mathcal{H})) \setminus \mathbf{E}(f, \mathcal{H})$ by Lemma 5.9. Let s_i be the side of G_i' that contains G_j', and define s_j similarly (Figure 24, left). Then $f'(P)$ cannot be adjacent to G_i' on the side s_i, since part of $f'(P)$ is separated from G_i' by G_j'. Therefore, since $G_i' \in \mathbf{E}(\mathrm{del}_G(f, \mathcal{H}))$, $f'(P)$ must be adjacent to G_i' on the opposite side, $\bar{s}_i$. Similarly $f'(P)$ must be adjacent to G_j' on the opposite side, $\bar{s}_j$. Note that $\bar{s}_i$ and $\bar{s}_j$ are disjoint. On the other hand, since G_i' and G_j' are not in $\mathbf{E}(f, \mathcal{H})$, there must be some $G_k \in G$ in $\bar{s}_i$ separating

G'_i from some point of $f(P)$, and similarly $G_l \in G$ in $\bar{s}_j$ separating G'_j from some point of $f(P)$. But this is not possible since G_k crosses G_l.

Now assume $G'_i \in E(\mathrm{del}_G(f, \mathcal{H})) \setminus E(f, \mathcal{H})$ and $G'_j \in G \setminus T(f, \mathcal{H})$ (Figure 24, right). If $G'_j \in E(f, \mathcal{H}) \setminus T(f, \mathcal{H})$ then it crosses G'_i by Lemma 5.8, so we may assume $G'_j \notin E(f, \mathcal{H})$. Define s_i as before. Now since $G'_j \notin E(f, \mathcal{H})$, on the side of G'_j contained in s_i, there must be another hyperplane H separating G'_j from a point of $f(P)$. This H cannot be in G since G'_j crosses H, so it is not deleted and hence $f'(P)$ is not adjacent to G'_i on the s_i side. As above there must therefore be a $G_k \in G$ on the $\bar{s}_i$ side. This G_k cannot cross G'_j; again a contradiction.

To see that each hyperplane of K crosses all other hyperplanes of $\mathcal{H}_f \setminus G$, note that K is contained in $E(\mathrm{del}_G(f, \mathcal{H})) \setminus T(\mathrm{del}_G(f, \mathcal{H}))$ by (5-1) of Lemma 5.9, and then use Lemma 5.8.

To finish the argument, we claim that all we have to check is that the set of hyperplanes that are either transitional for a Move operation or deleted along each side of the diagram is the same. This is because of the relations

$$\mathrm{Trim} \circ \mathrm{Move}(f, \mathcal{H}) = \mathrm{del}_{T(f, \mathcal{H})} \circ \mathrm{Trim}(f, \mathcal{H})$$

and

$$\mathrm{Trim} \circ \mathrm{del}_G = \mathrm{del}_G \circ \mathrm{Trim}.$$

which follow directly from the definitions. With these relations, we can simplify each side of the diagram to a single $\mathrm{Trim} \circ \mathrm{del}_V$ where V is the union of hyperplanes from all the deletion and Move steps on that side.

Thus, comparing the left side of the diagram with the top arrow and right side, it remains to check that

(5-5) $$T(f, \mathcal{H}) \cup G' = G \cup T(\mathrm{del}_G(f, \mathcal{H})) \cup K.$$

Using the definitions of G' and K, we see that both sides are equal to

$$G \cup T(\mathrm{del}_G(f, \mathcal{H})) \cup T(f, \mathcal{H}). \qquad \square$$

Every hyperplane $H \in \mathcal{H}_f$ becomes extremal at some point along the move sequence. We want to understand how deletions affect when this occurs. Note that when a hyperplane $H \in \mathcal{H}_f$ becomes extremal, it need not become transitional, and what happens can change with the deletion of a nearby hyperplane.

For a hyperplane $H \in \mathcal{H}_f$, define

$$e_H(f, \mathcal{H})$$

to be the first index i such that $H \in \boldsymbol{E}(f_i, \mathcal{H})$.

Lemma 5.12 *For any mutually crossing set $G \subset \mathcal{H}_f$, and any $H \in \mathcal{H}_f \setminus G$,*

(5-6)
$$e_H(f, \mathcal{H}) = e_H(\mathrm{del}_G(f, \mathcal{H})) + \delta$$

for some $\delta \in \{0, 1\}$.

Proof If $e_H(f, \mathcal{H}) = 0$ then H is already in $\boldsymbol{E}(f, \mathcal{H})$, which implies $H \in \boldsymbol{E}(\mathrm{del}_G(f, \mathcal{H}))$ by (5-1) of Lemma 5.9. The equality (5-6) follows with $\delta = 0$.

Thus we may assume $e_H(f, \mathcal{H}) > 0$. Suppose that $e_H(\mathrm{del}_G(f, \mathcal{H})) = 0$. This means that $H \in \boldsymbol{E}(\mathrm{del}_G(f, \mathcal{H})) \setminus \boldsymbol{E}(f, \mathcal{H})$, which implies (as in the proof of Lemma 5.11) that there are some elements G_i of G which separate H from $f(P)$ on one side s_i, so that $f(P)$ is adjacent to G_i on the s_i side. But $f(P)$ is not adjacent to G_i on the other side because H is there, which means $G_i \in \boldsymbol{T}(f, \mathcal{H})$. But this implies that, in Move$(f, \mathcal{H})$, all G_i as above are no longer in the set of separating hyperplanes, and hence $H \in \boldsymbol{E}(\mathrm{Move}(f, \mathcal{H}))$, so $e_H(f, \mathcal{H}) = 1$. This gives (5-6) with $\delta = 1$.

From now on we can assume $e_H(f, \mathcal{H}) > 0$ and $e_H(\mathrm{del}_G(f, \mathcal{H})) > 0$, and prove the statement by induction on the cardinality of $\mathcal{H}_f$ (the case $|\mathcal{H}_f| = 2$ is easy, and already covered by the previous paragraphs).

Let $(f', \mathcal{H}') = \mathrm{del}_G(f, \mathcal{H})$, $(f_1, \mathcal{H}) = \mathrm{Move}(f, \mathcal{H})$, and $(f_1', \mathcal{H}') = \mathrm{Move}(f', \mathcal{H}')$. By definition (since $e_H(f, \mathcal{H}) > 0$ and $e_H(f', \mathcal{H}) > 0$),

$$e_H(f, \mathcal{H}) = e_H(f_1, \mathcal{H}) + 1 \quad \text{and} \quad e_H(f', \mathcal{H}') = e_H(f_1', \mathcal{H}') + 1.$$

Thus it will suffice to prove

(5-7)
$$e_H(f_1, \mathcal{H}) = e_H(f_1', \mathcal{H}') + \delta.$$

Consider the warmup case when $G \cap \boldsymbol{E}(f, \mathcal{H}) = \varnothing$. By Lemma 5.10,

$$(f_1', \mathcal{H}') = \mathrm{del}_G(f_1, \mathcal{H}).$$

Since $|\mathcal{H}_{f_1}| < |\mathcal{H}_f|$, the inductive hypothesis gives us

$$e_H(f_1, \mathcal{H}) = e_H(f_1', \mathcal{H}') + \delta$$

for $\delta = 0$ or 1, proving (5-7) and hence (5-6).

Now in the general case, we use the diagram of Lemma 5.11. Note first that the value of e_H is not affected by a Trim operation. This is because Trim does not affect the set $\mathcal{H}_f$, or the membership in E or T.

The value of e_H is also unaffected by the del_K arrow on the right side. This is because each hyperplane $H \in K$ crosses every hyperplane in $\mathcal{H}_f \setminus (G \cup \{H\})$, which implies that any hyperplane in K cannot affect the membership in E or T of any other hyperplane in $\mathcal{H}_f \setminus G$. Therefore, we see that del_K commutes with the Move sequence on $(f_1', \mathcal{H}_f \setminus G)$.

The remaining arrow is labeled by $\mathrm{del}_{G'}$, and G' is a mutually crossing set. Thus by induction we know

$$e_H(f_1, \mathcal{H}) = e_H(\mathrm{del}_{G'}(f_1, \mathcal{H})) + \delta$$

for $\delta = 0$ or 1. Again the equality (5-7) follows. $\qquad \square$

We are now ready to prove the stability result for move sequences.

Proof of Proposition 5.7 We need to prove the following statement: for each i and $p \in P$, the set of $H \in \mathcal{H}_f \setminus G$ such that H separates $f_i(p)$ from $f_i'(p)$ is mutually crossing.

Note that when a sequence $(f_i(p))$ crosses a hyperplane H, it can only happen in the transition from f_j to f_{j+1} where $j = e_H(f, \mathcal{H})$. Moreover it must be that $H \in T(f_j, \mathcal{H})$, and that $f_j(p)$ is on the side of H where $f_j(P)$ is adjacent. If the sequence $(f_i(p))$ does not cross H, then H will either remain in $\mathcal{H}_{f_i}$ for all $i \geq j$, or it will be crossed only by points on the other side and not by $f_i(p)$. The same holds for the sequence (f_i').

Now suppose that H_1 and H_2 separate $f_i(p)$ from $f_i'(p)$, for some i, and H_1 does not cross H_2. Since $f_0(p)$ and $f_0'(p)$ are on the same side of both hyperplanes, we may assume that H_1 separates them from H_2. Hence there exists a (different) time i at which one of them is still on the original side of H_1, whereas the other has crossed both hyperplanes; fix such an i.

Suppose first that $f_i'(p)$ is the one which lies on the other side of H_2. Let $j = e_{H_1}(f', \mathcal{H})$. Then since H_1 does not cross H_2, $f_{j+1}'(p)$ has not yet crossed H_2. Thus we must have

$$j = e_{H_1}(f', \mathcal{H}) < e_{H_2}(f', \mathcal{H}) < i$$

and $H_1 \in T(f_j', \mathcal{H})$, with $f_j'(p)$ on the side where $f_j'(P)$ is adjacent to H_1.

Now $k = e_{H_1}(f, \mathcal{H}) \leq j + 1$ by Lemma 5.12. Since $j + 1 < i$ and $f_i(p)$ has not crossed H_1, it must be that the side of H_1 where $f_k(P)$ is adjacent is the one opposite from $f_k(p)$, the one containing H_2.

But this means that H_2 can no longer be in $\mathcal{H}_{f_k}$, which can only be if

$$e_{H_2}(f, \mathcal{H}) \leq k - 1 \leq j.$$

Thus, again by Lemma 5.12, $e_{H_2}(f', \mathcal{H}) \leq j$, which is a contradiction.

We conclude that H_1 does not cross H_2, which is what we wanted.

The case where $f_i(p)$ crosses the hyperplanes and $f_i'(p)$ does not is handled similarly. The main difference is that in this case, instead of using $e_{H_1}(f, \mathcal{H}) \leq e_{H_1}(f', \mathcal{H}) + 1$, we use $e_{H_1}(f', \mathcal{H}) \leq e_{H_1}(f, \mathcal{H})$, which creates a "+1" there, which is then lost in the other application of Lemma 5.12. □

5.4 Completing the proof of Theorem 5.1

Property (1) follows from Lemma 5.3, while property (2) follows from the construction, where for both properties we use the fact that cubes have diameter 1 in the d_∞ metric. Property (5) is also easily seen to hold by construction, and more specifically it follows from the fact that the sets $E(f, \mathcal{H})$ and $T(f, \mathcal{H})$ only depend on the image of f (and hence that this will remain true throughout the move sequence).

For property (3), observe that for a fixed $p \in P$, no hyperplane H separates $f_i(p)$ from $f_{i+1}(p)$ for two different values of i. It follows that any combinatorial path obtained by concatenating a choice of geodesics from $f_i(p)$ to $f_{i+1}(p)$ for $0 \leq i < n$ is an ℓ^1–geodesic in X. This proves (3).

For property (4), by definition of the contraction sequence, the hyperplanes that separate any given $f(p)$ from $f_n(q)$ for $p, q \in P$ are contained in $\mathcal{H}_f$. On the other hand, a hyperplane $H \in \mathcal{H}_f$ cannot separate every point in $f(P)$ from any fixed vertex of X, because there are elements of $f(P)$ on both sides of H, by definition of $\mathcal{H}_f$. Hence, there can be no hyperplane separating $f(P)$ from a point in $f_n(P)$. This gives property (4).

Property (6) is a direct consequence of Proposition 5.7 and Lemma 5.6. □

6 Proofs of the main theorems

We are now almost ready to prove our main theorems, Theorem E and Theorem B. The main bit of work here is Proposition 6.1, which compiles our preceding stability results into a useful form for our current purposes.

Proposition 6.1 *Let $(\mathcal{X}, \mathfrak{S})$ be a G–colorable HHS for $G < \mathrm{Aut}(\mathfrak{S})$. For any $k \in \mathbb{N}$, there exists $K_3 = K_3(k, \mathfrak{S}) > 0$ such that the following holds.*

Suppose that $F, F' \subset \mathcal{X}$ are finite subsets satisfying $|F|, |F'| \leq k$, let $g \in G$, and suppose that $d_{\mathrm{Haus}}(gF, F') \leq 1$. Choose any map $\iota_F \colon F \sqcup F' \to F$ such that $\iota_F(f) = f$ if $f \in F$ and $d_{\mathcal{X}}(g\iota_F(f), f) \leq 1$ if $f \in F'$. Also, choose a map $\iota_{F'} \colon F \sqcup F' \to F'$ such that $\iota_{F'}(f) = f$ if $f \in F'$ and $d_{\mathcal{X}}(gf, \iota_{F'}(f)) \leq 1$ if $f \in F$. Consider:

- *The cube complexes $\mathcal{Q}_F, \mathcal{Q}_{F'}$ produced by Theorem 4.1 with associated maps $\Phi_F, \Phi_{F'}$ to $\mathcal{X}$, and $\psi_F, \psi_{F'}$ from F, F' to $\mathcal{Q}_F, \mathcal{Q}_{F'}$.*

- *The sequences of contractions $\{(\psi_F)_i = \psi_i\}_{i \leq n_{\psi_F}}$ and $\{(\psi_{F'})_i = \psi'_i\}_{i \leq n_{\psi_{F'}}}$ produced by Theorem 5.1. Set $n_{\psi_F} = n_F$ and $n_{\psi_{F'}} = n_{F'}$.*

Then

(1) *$|n_F - n_{F'}| < K_3$,*

(2) *for each $i \in \{1, \ldots, \max\{n_F, n_{F'}\}\}$ and any $f \in F \sqcup F'$,*

$$d_{\mathcal{X}}\big(g \circ \Phi_F \circ \psi_i(\iota_F(f)), \Phi_{F'} \circ \psi'_i(\iota_{F'}(f))\big) \leq K_3,$$

(3) *$\mathrm{diam}_{\mathcal{X}}(\Phi_F(\psi_{n_F}(F))) < K_3$.*

More visually, item (2) says that the diagram

(6-1)

coarsely commutes.

Proof We will use the output and notation of Theorem 4.1, and in particular the CAT(0) cube complex $\mathcal{Q}_0$ obtained from both $\mathcal{Q}_F$ and $\mathcal{Q}_{F'}$ by collapsing at most $N = N(k, \mathfrak{S}) > 0$ hyperplanes, with hyperplanes collapse maps h and h'.

We will also use the notation of Theorem 5.1, in particular the notation $\{f_i \mid i \leq n_f\}$ for the sequence of maps starting with f and ending with a map with bounded image.

We have $\psi := h \circ \psi_F \circ \iota_F = h' \circ \psi_{F'} \circ \iota_{F'}$, as stated in Theorem 4.1. By Theorem 5.1, composing, say, ψ_F with a hyperplane deletion map affects the length of the corresponding sequence of maps by at most 1. In particular, $|n_F - n_\psi| \leq N$ and, similarly, $|n_{F'} - n_\psi| \leq N$ (notice that $n_{\psi_F} = n_{\psi_F \circ \iota_F}$ by Theorem 5.1(5), and a similar statement holds for F'). Hence, conclusion (1) holds for any K_3 larger than $2N$.

We now prove conclusion (2). By Theorem 4.1, diagram (4-4) commutes with error at most $K = K(k, \mathfrak{S})$. For convenience, we reproduce the diagram here:

(6-2)

For any $f \in F \sqcup F'$,

$$d_{\mathcal{X}}(g \circ \Phi_F \circ (\psi_F)_i (\iota_F(f)), \Phi_0 \circ \eta \circ (\psi_F)_i (\iota_F(f))) \leq K.$$

By Theorem 5.1, $d_\infty(\eta \circ (\psi_F)_i (\iota_F(f)), (\eta \circ \psi_F)_i (\iota_F(f))) \leq N$, and hence

$$d_{\mathcal{X}}(\Phi_0 \circ \eta \circ (\psi_F)_i (f), \Phi_0 \circ (\eta \circ \psi_F)_i (f)) \leq K' = K'(k, \mathfrak{S})$$

since Φ_0 is a quasi-isometric embedding with controlled constants (and the dimension of $\mathcal{Q}_0$ is bounded in terms of $\mathfrak{S}$ by Theorem 4.1, so that the ℓ^∞ and ℓ^1 metrics on it are uniformly quasi-isometric). The triangle inequality then gives

$$d_{\mathcal{X}}(g \circ \Phi_F \circ (\psi_F)_i (\iota_F(f)), \Phi_0 \circ (\eta \circ \psi_F)_i (\iota_F(f))) \leq K' + K.$$

Similarly, we get

$$d_{\mathcal{X}}(\Phi_{F'} \circ (\psi_{F'})_i (\iota_{F'}(f)), \Phi_0 \circ (\eta' \circ \psi_{F'})_i (\iota_{F'}(f))) \leq K' + K.$$

By Theorem 5.1(5), $(\eta' \circ \psi_{F'})_i (\iota_{F'}(f)) = (\eta' \circ \psi_{F'} \circ \iota_{F'})_i (f) = (\eta \circ \psi_F \circ \iota_F)_i (f)$, and hence conclusion (2) holds for any K_3 larger than $2(K' + K)$.

Finally, to prove (3), we now bound the diameter of $\Phi_F(\psi_{n_F}(F))$. Similarly to above, Φ_F is a quasi-isometric embedding with constants controlled in terms of k and $\mathfrak{S}$ even when we endow $\mathcal{Q}_F$ with the ℓ^∞ metric. Since $\mathrm{diam}_{\mathcal{Q}_F}(\psi_{n_F}(F)) \leq 1$ (in the ℓ^∞ metric) by Theorem 5.1, this gives the required bound on $\mathrm{diam}_{\mathcal{X}}(\Phi_F(\psi_{n_F}(F)))$. $\square$

6.1 Barycenters: proof of Theorem E

Our next goal is to prove Theorem E. To do so, we'll need the precise definition of stable barycenter:

Definition 6.2 For a metric space X a *stable barycenter map* for k points is a map $\tau \colon X^k \to X$ which is

- *permutation invariant*, meaning $\tau \circ \pi = \tau$ for any $\pi \colon X^k \to X^k$ that is a permutation of the factors;
- *coarsely Lipschitz*, meaning there exists $\kappa_1 > 0$ such that for $x, x' \in X^k$,

$$d_X(\tau(x), \tau(x')) \leq \kappa_1 d_{X^k}(x, x') + \kappa_1.$$

We further say that τ is *coarsely equivariant* with respect to a group Γ acting on X by isometries if there exists $\kappa_1 > 0$ such that for all $g \in \Gamma$

$$d_X(g\tau(x), \tau(gx)) \leq \kappa_1,$$

where Γ acts on X^k diagonally.

We now prove that colorable HHSs admit stable coarsely equivariant barycenters, with the following version slightly more general than Theorem E:

Theorem 6.3 *Let $(\mathcal{X}, \mathfrak{S})$ be a G–colorable HHS for $G < \mathrm{Aut}(\mathfrak{S})$. Then $\mathcal{X}$ admits coarsely G–equivariant stable barycenters for k points, for any $k \geq 1$. Moreover, the coarse barycenter of a set F is contained in the hierarchical hull of F.*

Proof We use the notation from the statement of Proposition 6.1.

To define a barycenter $\tau(f_1, \ldots, f_k)$, we consider $F = \{f_i\}$, set $x_F = \Phi_F(\psi_{n_F}(F))$, and let $\tau(f_1, \ldots, f_k)$ be an arbitrary point in x_F; we make the choice depending on the set F only, so that permutation invariance is achieved.

This choice does not matter for our purposes since $\mathrm{diam}_{\mathcal{X}}(\Phi_F(\psi_{n_F}(F))) < K_3$ by Proposition 6.1.

Now suppose that $(f_1', \ldots, f_k')$ is such that $d_{\mathrm{Haus}}(\{f_i'\}, \{f_i\}) \leq 1$, and set $F' = \{f_i'\}$. Without loss of generality, assume that $n_F \geq n_{F'}$, where we note that $n_F - n_{F'} < K_3$ by part (1) of Proposition 6.1. Part (2) of Proposition 6.1 now implies that for any $f \in F \sqcup F'$,

$$d_{\mathcal{X}}\big(\Phi_F(\psi_{n_F}(\iota_F(f))), \Phi_{F'}(\psi'_{n_F}(\iota_{F'}(f)))\big) < K_3.$$

But $\psi'_{n_{F'}} = \psi'_{n_F}$ since $n_F \geq n_{F'}$, so we can conclude that

$$d_{\mathcal{X}}\big(\Phi_F(\psi_{n_F}(\iota_F(f))), \Phi_{F'}(\psi'_{n_{F'}}(\iota_{F'}(f)))\big) < K_3.$$

Finally, the fact that $\mathrm{diam}_{\mathcal{X}}(\Phi_F(\psi_{n_F}(F))) < K_3$ and $\mathrm{diam}_{\mathcal{X}}(\Phi_{F'}(\psi'_{n_{F'}}(F'))) < K_3$ gives that

$$\mathrm{diam}_{\mathcal{X}}(x_F \cup x_{F'}) < 3K_3.$$

Setting $\kappa_1 = 3K_3$, we get that τ is κ_1–coarsely Lipschitz.

Finally, coarse equivariance follows similarly, applying Proposition 6.1 with $F' = gF$, as follows. First, as above we can assume $n_F \geq n_{gF}$, for otherwise we can swap the roles of F and gF, by considering the automorphism g^{-1}. We still have $n_F - n_{gF} < K_3$. Part (2) of Proposition 6.1 implies that for any $f \in F \sqcup gF$,

$$d_{\mathcal{X}}\big(g \circ \Phi_F(\psi_{n_F}(\iota_F(f))), \Phi_{gF}(\psi'_{n_{gF}}(\iota_{gF}(f)))\big) < K_3.$$

As above, we conclude that

$$\mathrm{diam}_{\mathcal{X}}(g(x_F) \cup x_{gF}) < 3K_3,$$

which completes the proof. $\qquad\square$

6.2 Bicombability: Proof of Theorem B

We begin with the formal definition of bicombing which is appropriate for our context; see [1]. In the following definition, we adopt the convention that if $\phi\colon [0, a] \to X$ is a map, then we trivially extend ϕ by $\phi(t) = \phi(a)$ for all $t > a$.

Definition 6.4 A *discrete*, *bounded*, *quasigeodesic bicombing* of a metric space X consists of a family of discrete paths $\{\Omega_{x,y}\}_{x,y \in X}$ which are, for some constant $\kappa_2 > 0$,

(1) *quasigeodesic*, meaning that for any $x, y \in X$ with $d = d_X(x, y)$, there exists $n_{x,y} \leq \kappa_2 d + \kappa_2$ such that the path $\Omega_{x,y}\colon \{0, \dots, n_{x,y}\} \to \mathcal{X}$ is a (κ_2, κ_2)–quasi-isometric embedding with $\Omega_{x,y}(0) = x$ and $\Omega_{x,y}(n_{x,y}) = y$; and

(2) *fellow-traveling*, meaning that if $x', y' \in X$ with $d' = d_X(x', y')$ and

$$d_X(x, x'), d_X(y, y') \leq 1,$$

then for all $t \in \{0, \dots, \max\{n_{x,y}, n_{x',y'}\}\}$,

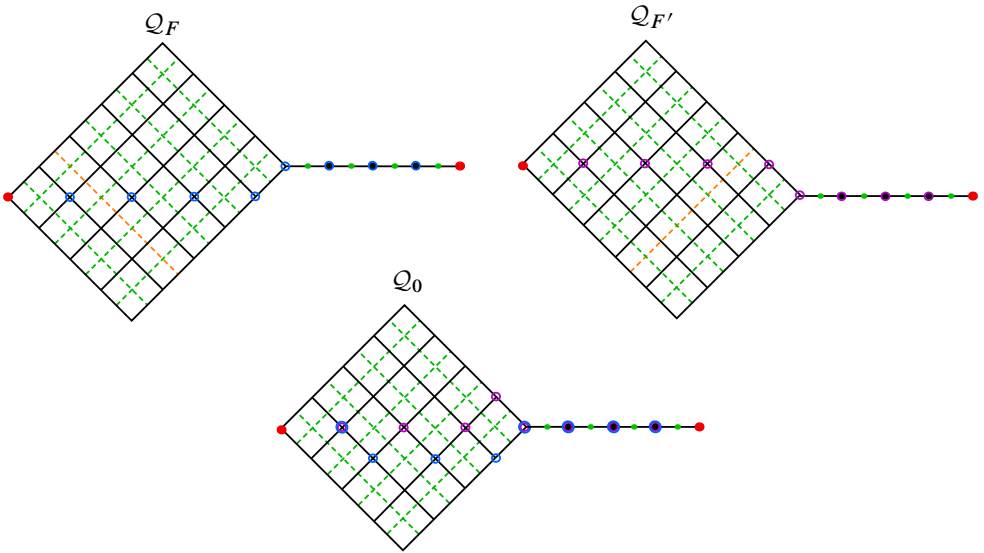$$d_X(\Omega_{x,y}(t), \Omega_{x',y'}(t)) \leq \kappa_2.$$

Figure 25: A simple example of bicombing paths, building on the hierarchical setup from Figure 18. Deleting the orange hyperplanes from $\mathcal{Q}_F$ and $\mathcal{Q}_{F'}$ results in perturbing the contraction paths in $\mathcal{Q}_0$.

In addition, we say that $\{\Omega_{x,y}\}_{x,y \in X}$ is Γ–*coarsely equivariant* with respect to a group $\Gamma < \mathrm{Isom}(X)$ if for any $g \in \Gamma$ and $x, y \in X$ and $t \in \{0, \ldots, \max\{n_{x,y}, n_{x',y'}\}\}$,

$$d_X(g \cdot \Omega_{x,y}(t), \Omega_{g \cdot x, g \cdot y}(t)) < \kappa_2.$$

Finally, we recall the following definition from [7], which was inspired by the paths constructed in [43]:

Definition 6.5 For $D \geq 1$, a path γ in $\mathcal{X}$ is a D–*hierarchy path* if

(1) γ is a (D, D)–quasigeodesic,

(2) for each $W \in \mathfrak{S}$, $\pi_W \circ \gamma$ is an unparametrized (D, D)–quasigeodesic.

We can now prove that colorable HHSs admit discrete, bounded, quasigeodesic, coarsely equivariant bicombings by hierarchy paths.

Theorem 6.6 Let $(\mathcal{X}, \mathfrak{S})$ be a G–colorable HHS with $G < \mathrm{Aut}(\mathfrak{S})$. Then there exists $D > 0$ such that $(\mathcal{X}, \mathfrak{S})$ admits a coarsely G–equivariant, discrete, bounded, quasigeodesic bicombing by D–hierarchy paths.

Proof Let $(\mathcal{X}, \mathfrak{S})$ be a colorable HHS. We again use the notation from the statement of Proposition 6.1, where now $F = \{x, y\}$ and $F' = \{x', y'\}$ with $d_{\mathcal{X}}(x, x') \leq 1$ and $d_{\mathcal{X}}(y, y') \leq 1$. We make a blanket observation that $k = 2$ and so the constant K_3 in Proposition 6.1 depends only on $(\mathcal{X}, \mathfrak{S})$.

Coarse equivariance can be obtained using the argument below, setting $F' = gF$. We omit the details for readability.

Construction of the bicombing paths Let $\psi = \psi_0$ and define a map

$$\omega_{x,y} \colon \{0, \ldots, 2n_{x,y}\} \to \mathcal{Q}_F$$

by

$$\omega_{x,y}(i) = \begin{cases} \psi_i(x) & \text{if } i \in \{0, \ldots, n_{x,y}\}, \\ \psi_{2n_{x,y}-i}(y) & \text{if } i \in \{n_{x,y}+1, \ldots, 2n_{x,y}\}. \end{cases}$$

We claim that $\omega_{x,y}$ is a (C, C)–quasigeodesic in the ℓ^1 metric on $\mathcal{Q}_F$, for some uniform C. First, the points $\omega_{x,y}(0), \ldots, \omega_{x,y}(n_{x,y})$ appear on an ℓ^1–geodesic γ_1 from $\omega_{x,y}(0)$ to $\omega_{x,y}(n_{x,y})$ in the given order by Theorem 5.1(3), and the same holds for $\omega_{x,y}(n_{x,y}+1), \ldots, \omega_{x,y}(2n_{x,y})$ for some ℓ^1–geodesic γ_2 from $\omega_{x,y}(n_{x,y}+1)$ to $\omega_{x,y}(2n_{x,y})$. Moreover, consecutive $\omega_{x,y}(i)$ are uniformly close to each other, since they are at distance at most 1 in the ℓ^∞ metric, which is uniformly quasi-isometric to the ℓ^1 metric with constant only depending on the dimension of $\mathcal{Q}_F$, which in turn only depends on $\mathfrak{S}$.

Let γ be the concatenation of γ_1, an ℓ^1–geodesic from $\omega_{x,y}(n_{x,y})$ to $\omega_{x,y}(n_{x,y}+1)$, and γ_2. Since no hyperplane can separate $\{x, y\}$ from $\psi_{n_{x,y}}(x)$ or $\psi_{n_{x,y}}(y)$, again by Theorem 5.1, we see that γ crosses each hyperplane at most once, and is therefore an ℓ^1–geodesic. Since $\omega_{x,y}(n_{x,y})$ and $\omega_{x,y}(n_{x,y}+1)$ are just opposite corners of a cube, we see that the $\omega_{x,y}(i)$ appear along an ℓ^1–geodesic in the given order, and with uniformly spaced gaps. This shows that $\omega_{x,y}$ is a (C, C)–quasigeodesic in the ℓ^1 metric, for C depending only on the dimension of $\mathcal{Q}_F$ and hence only on $\mathfrak{S}$.

It follows then that the composition

$$\Omega_{x,y} = \Phi_F \circ \omega_{x,y} \colon \{0, \ldots, 2n_{x,y}\} \to \mathcal{X}$$

is a (K_4, K_4)–quasigeodesic in $\mathcal{X}$ with $K_4 = K_4(\mathfrak{S})$. We can perturb it a uniformly bounded amount at the endpoints to make sure that the endpoints are x and y; with a slight abuse of notation we still denote the perturbation by $\Omega_{x,y}$ and the quasi-isometry constants by K_4.

This proves that Definition 6.4(1) holds for the family $\{\Omega_{x,y}\}_{x,y\in\mathcal{X}}$.

Fellow-traveling We now prove the fellow-traveling condition in Definition 6.4(2) holds. Once again adopting our previous notation, we want to prove that there exists $\kappa_2 = \kappa_2(\mathcal{X}, \mathfrak{S}) > 0$ such that for any $t \in \{0, \dots, \max\{2n_{x,y}, 2n_{x',y'}\}\}$,

$$(6\text{-}3) \qquad\qquad d_{\mathcal{X}}(\Omega_{x,y}(t), \Omega_{x',y'}(t)) < \kappa_2.$$

Without loss of generality, suppose that $n_{x,y} \geq n_{x',y'}$ and recall that Proposition 6.1(1) gives that $\delta = n_{x,y} - n_{x',y'} < K_3$, where K_3 depends only on $(\mathcal{X}, \mathfrak{S})$. There are four cases to consider:

  (i) When $0 \leq i \leq n_{x',y'}$, where $\Omega_{x,y}$ and $\Omega_{x',y'}$ are defined using x and x', respectively.

  (ii) When $n_{x',y'} < j \leq n_{x,y}$, where $\Omega_{x,y}$ is defined using x whereas $\Omega_{x',y'}$ is defined using y'.

  (iii) When $n_{x,y} < q \leq 2n_{x',y'}$, where both $\Omega_{x,y}$ and $\Omega_{x',y'}$ are nonconstant and defined using y and y', respectively.

  (iv) When $2n_{x',y'} < r \leq 2n_{x,y}$, where $\Omega_{x,y}$ is nonconstant but

$$\Omega_{x',y'}(j) = \Omega_{x',y'}(2n_{x',y'})$$

    for all such j.

In what follows, we will repeatedly use the fact that $\Omega_{x,y}$ and $\Omega_{x',y'}$ are (K_4, K_4)–quasigeodesics. Also, set $K_5 = K_4 \cdot (2\delta) + K_4$.

In case (i), equation (6-3) follows immediately from Proposition 6.1(2) with $\kappa_2 = K_3$.

In case (ii),

$$d_{\mathcal{X}}(\Omega_{x,y}(n_{x',y'}), \Omega_{x,y}(j)) < K_5 \quad \text{and} \quad d_{\mathcal{X}}(\Omega_{x',y'}(n_{x',y'}), \Omega_{x',y'}(j)) < K_5,$$

while Proposition 6.1(2) gives

$$d_{\mathcal{X}}(\Omega_{x,y}(j + 2\delta), \Omega_{x',y'}(j)) < K_3,$$

so the triangle inequality implies that (6-3) holds in this case with $\kappa_2 = 2K_5 + K_3$.

In case (iii),

$$d_{\mathcal{X}}(\Omega_{x,y}(q), \Omega_{x,y}(q + 2\delta)) < K_5$$

and Proposition 6.1(2) provides

$$d_{\mathcal{X}}(\Omega_{x,y}(q + 2\delta), \Omega_{x',y'}(q)) < K_3,$$

so the triangle inequality implies that (6-3) holds with $\kappa_2 = K_3 + K_5$.

Finally, in case (iv),

$$d_{\mathcal{X}}(\Omega_{x,y}(2n_{x',y'}), \Omega_{x,y}(2n_F)) < K_5$$

and Proposition 6.1(2) provides

$$d_{\mathcal{X}}(\Omega_{x,y}(2n_{x',y'}), \Omega_{x',y'}(2n_{x',y'})) < K_3.$$

Since $\Omega_{x',y'}(2n_F) = \Omega_{x',y'}(2n_{x',y'})$ by convention, the triangle inequality implies that (6-3) holds with $\kappa_2 = K_3 + K_5$.

Hence we may set $\kappa_2 = 2K_5 + K_3$ to complete the proof of the fellow-traveling condition in Definition 6.4(2). This completes the proof that these paths gives a bicombing.

Hierarchy paths To finish the proof, we now show that $\Omega_{x,y}$ is a D–hierarchy path for some $D = D(\mathfrak{S}) > 0$ (Definition 6.5). We will use that Φ_F is a K–median map (Theorem 4.1(2)); we now recall what this means.

In a CAT(0) cube complex $\mathcal{Q}$ one can define a map $m_{\mathcal{Q}} \colon \mathcal{Q}^3 \to \mathcal{Q}$ (called median), and the only property of this map that we need here is that if x, y, and z appear in this order along an ℓ^1–geodesic, then $m(x, y, z) = y$. Also, in an HHS $\mathcal{X}$, there is a map $m_{\mathcal{X}} \colon \mathcal{X}^3 \to \mathcal{X}$ called coarse median, whose definition we do not need, and Φ_F being K–median means that for all $x, y, z \in \mathcal{Q}_F$,

$$d_{\mathcal{X}}\big(\Phi_F(m_{\mathcal{Q}_F}(x, y, z)), m_{\mathcal{X}}(\Phi_F(x), \Phi_F(y), \Phi_F(z))\big) \leq K.$$

This inequality implies that, for all $i < j < k$, we have that $\Omega_{x,y}(j)$ lies uniformly close to $m_{\mathcal{X}}(\Omega_{x,y}(i), \Omega_{x,y}(j), \Omega_{x,y}(k))$. This is enough to guarantee that the quasi-isometric embedding $\Omega_{x,y}$ is a hierarchy path by [8, Lemma 1.37]. $\qquad\square$

References

[1] **J M Alonso**, **M R Bridson**, *Semihyperbolic groups*, Proc. London Math. Soc. 70 (1995) 56–114 MR Zbl

[2] **G Baumslag**, **S M Gersten**, **M Shapiro**, **H Short**, *Automatic groups and amalgams — a survey*, from "Algorithms and classification in combinatorial group theory" (G Baumslag, C F Miller, III, editors), Math. Sci. Res. Inst. Publ. 23, Springer (1992) 179–194 MR Zbl

[3] **J A Behrstock**, *Asymptotic geometry of the mapping class group and Teichmüller space*, Geom. Topol. 10 (2006) 1523–1578 MR Zbl

[4] **J Behrstock**, **M Hagen**, **A Martin**, **A Sisto**, *A combinatorial take on hierarchical hyperbolicity and applications to quotients of mapping class groups*, preprint (2020) arXiv 2005.00567

[5] **J Behrstock**, **M F Hagen**, **A Sisto**, *Asymptotic dimension and small-cancellation for hierarchically hyperbolic spaces and groups*, Proc. Lond. Math. Soc. 114 (2017) 890–926 MR Zbl

[6] **J Behrstock**, **M F Hagen**, **A Sisto**, *Hierarchically hyperbolic spaces, I: Curve complexes for cubical groups*, Geom. Topol. 21 (2017) 1731–1804 MR Zbl

[7] **J Behrstock**, **M Hagen**, **A Sisto**, *Hierarchically hyperbolic spaces, II: Combination theorems and the distance formula*, Pacific J. Math. 299 (2019) 257–338 MR Zbl

[8] **J Behrstock**, **M F Hagen**, **A Sisto**, *Quasiflats in hierarchically hyperbolic spaces*, Duke Math. J. 170 (2021) 909–996 MR Zbl

[9] **J Behrstock**, **B Kleiner**, **Y Minsky**, **L Mosher**, *Geometry and rigidity of mapping class groups*, Geom. Topol. 16 (2012) 781–888 MR Zbl

[10] **J A Behrstock**, **Y N Minsky**, *Centroids and the rapid decay property in mapping class groups*, J. Lond. Math. Soc. 84 (2011) 765–784 MR Zbl

[11] **F Berlai**, **B Robbio**, *A refined combination theorem for hierarchically hyperbolic groups*, Groups Geom. Dyn. 14 (2020) 1127–1203 MR Zbl

[12] **M Bestvina**, **K Bromberg**, **K Fujiwara**, *Constructing group actions on quasi-trees and applications to mapping class groups*, Publ. Math. Inst. Hautes Études Sci. 122 (2015) 1–64 MR Zbl

[13] **M Bestvina**, **K Bromberg**, **K Fujiwara**, *Proper actions on finite products of quasi-trees*, Ann. H. Lebesgue 4 (2021) 685–709 MR Zbl

[14] **M Bestvina**, **K Bromberg**, **K Fujiwara**, **A Sisto**, *Acylindrical actions on projection complexes*, Enseign. Math. 65 (2019) 1–32 MR Zbl

[15] **B H Bowditch**, *Coarse median spaces and groups*, Pacific J. Math. 261 (2013) 53–93 MR Zbl

[16] **B Bowditch**, *Convex hulls in coarse median spaces*, preprint (2018) Available at http://homepages.warwick.ac.uk/~masgak/papers/hulls-cms.pdf

[17] **M R Bridson**, *Semisimple actions of mapping class groups on CAT(0) spaces*, from "Geometry of Riemann surfaces" (F P Gardiner, G González-Diez, C Kourouniotis, editors), London Math. Soc. Lecture Note Ser. 368, Cambridge Univ. Press (2010) 1–14 MR Zbl

[18] **M R Bridson**, **A Haefliger**, *Metric spaces of non-positive curvature*, Grundl. Math. Wissen. 319, Springer (1999) MR Zbl

[19] **P-E Caprace**, **M Sageev**, *Rank rigidity for CAT(0) cube complexes*, Geom. Funct. Anal. 21 (2011) 851–891 MR Zbl

[20] **J Chalopin**, **V Chepoi**, **A Genevois**, **H Hirai**, **D Osajda**, *Helly groups*, preprint (2020) arXiv 2002.06895

[21] **I Chatterji**, **G Niblo**, *From wall spaces to* CAT(0) *cube complexes*, Internat. J. Algebra Comput. 15 (2005) 875–885 MR Zbl

[22] **V Chepoi**, *Graphs of some* CAT(0) *complexes*, Adv. in Appl. Math. 24 (2000) 125–179 MR Zbl

[23] **M Chesser**, *Stable subgroups of the genus 2 handlebody group*, Algebr. Geom. Topol. 22 (2022) 919–971 MR Zbl

[24] **M T Clay**, **C J Leininger**, **J Mangahas**, *The geometry of right-angled Artin subgroups of mapping class groups*, Groups Geom. Dyn. 6 (2012) 249–278 MR Zbl

[25] **S Dowdall**, **M G Durham**, **C J Leininger**, **A Sisto**, *Extensions of Veech groups*, *I*: *A hyperbolic action*, preprint (2020) arXiv 2006.16425

[26] **S Dowdall**, **M G Durham**, **C J Leininger**, **A Sisto**, *Extensions of Veech groups*, *II*: *Hierarchical hyperbolicity and quasi-isometric rigidity*, preprint (2021) arXiv 2111.00685

[27] **M G Durham**, *The augmented marking complex of a surface*, J. Lond. Math. Soc. 94 (2016) 933–969 MR Zbl

[28] **M G Durham**, *Elliptic actions on Teichmüller space*, Groups Geom. Dyn. 13 (2019) 415–465 MR Zbl

[29] **M G Durham**, **M F Hagen**, **A Sisto**, *Boundaries and automorphisms of hierarchically hyperbolic spaces*, Geom. Topol. 21 (2017) 3659–3758 MR Zbl

[30] **M G Durham**, **M F Hagen**, **A Sisto**, *Correction to the article "Boundaries and automorphisms of hierarchically hyperbolic spaces"*, Geom. Topol. 24 (2020) 1051–1073 MR Zbl

[31] **D B A Epstein**, **J W Cannon**, **D F Holt**, **S V F Levy**, **M S Paterson**, **W P Thurston**, *Word processing in groups*, Jones and Bartlett, Boston, MA (1992) MR Zbl

[32] **A Eskin**, **H Masur**, **K Rafi**, *Large-scale rank of Teichmüller space*, Duke Math. J. 166 (2017) 1517–1572 MR Zbl

[33] **S M Gersten**, **H B Short**, *Rational subgroups of biautomatic groups*, Ann. of Math. 134 (1991) 125–158 MR Zbl

[34] **M Gromov**, *Hyperbolic groups*, from "Essays in group theory" (S M Gersten, editor), Math. Sci. Res. Inst. Publ. 8, Springer (1987) 75–263 MR Zbl

[35] **T Haettel**, **N Hoda**, **H Petyt**, *Coarse injectivity*, *hierarchical hyperbolicity*, *and semi-hyperbolicity*, preprint (2020) arXiv 2009.14053

[36] **M Hagen**, *Non-colorable hierarchically hyperbolic groups*, Internat. J. Algebra Comput. 33 (2023) 337–350 MR Zbl

[37] **M Hagen**, **A Martin**, **A Sisto**, *Extra-large type Artin groups are hierarchically hyperbolic*, Math. Ann. (2022)

[38] **M F Hagen**, **T Susse**, *On hierarchical hyperbolicity of cubical groups*, Israel J. Math. 236 (2020) 45–89 MR Zbl

[39] **F Haglund**, **F Paulin**, *Simplicité de groupes d'automorphismes d'espaces à courbure négative*, from "The Epstein birthday schrift" (I Rivin, C Rourke, C Series, editors), Geom. Topol. Monogr. 1, Geom. Topol., Coventry (1998) 181–248 MR Zbl

[40] **U Hamenstädt**, *Geometry of the mapping class group*, *II*: *A biautomatic structure*, preprint (2009) arXiv 0912.0137

[41] **G C Hruska**, **D T Wise**, *Finiteness properties of cubulated groups*, Compos. Math. 150 (2014) 453–506 MR Zbl

[42] **H A Masur**, **Y N Minsky**, *Geometry of the complex of curves, I: Hyperbolicity*, Invent. Math. 138 (1999) 103–149 MR Zbl

[43] **H A Masur**, **Y N Minsky**, *Geometry of the complex of curves, II: Hierarchical structure*, Geom. Funct. Anal. 10 (2000) 902–974 MR Zbl

[44] **B Miesch**, *Injective metrics on cube complexes*, preprint (2014) arXiv 1411.7234

[45] **L Mosher**, *Mapping class groups are automatic*, Ann. of Math. 142 (1995) 303–384 MR Zbl

[46] **G A Niblo**, **L D Reeves**, *The geometry of cube complexes and the complexity of their fundamental groups*, Topology 37 (1998) 621–633 MR Zbl

[47] **B Nica**, *Cubulating spaces with walls*, Algebr. Geom. Topol. 4 (2004) 297–309 MR Zbl

[48] **K Rafi**, *A combinatorial model for the Teichmüller metric*, Geom. Funct. Anal. 17 (2007) 936–959 MR Zbl

[49] **K Rafi**, *Hyperbolicity in Teichmüller space*, Geom. Topol. 18 (2014) 3025–3053 MR Zbl

[50] **K Rafi**, **Y Verberne**, *Geodesics in the mapping class group*, Algebr. Geom. Topol. 21 (2021) 2995–3017 MR Zbl

[51] **B Robbio**, **D Spriano**, *Hierarchical hyperbolicity of hyperbolic-2–decomposable groups*, preprint (2020) arXiv 2007.13383

[52] **J Russell**, *Extensions of multicurve stabilizers are hierarchically hyperbolic*, preprint (2021) arXiv 2107.14116

[53] **J Russell**, **D Spriano**, **H C Tran**, *Convexity in hierarchically hyperbolic spaces*, Algebr. Geom. Topol. 23 (2023) 1167–1248 MR

[54] **M Sageev**, *Ends of group pairs and non-positively curved cube complexes*, Proc. London Math. Soc. 71 (1995) 585–617 MR Zbl

[55] **M Sageev**, CAT(0) *cube complexes and groups*, from "Geometric group theory" (M Bestvina, M Sageev, K Vogtmann, editors), IAS/Park City Math. Ser. 21, Amer. Math. Soc., Providence, RI (2014) 7–54 MR Zbl

[56] **D Spriano**, *Hyperbolic HHS*, *II*: *Graphs of hierarchically hyperbolic groups*, preprint (2018) arXiv 1801.01850

[57] **P A Storm**, *The Novikov conjecture for mapping class groups as a corollary of Hamenstädt's theorem*, preprint (2005) arXiv math/0504248

[58] **J Tao**, *Linearly bounded conjugator property for mapping class groups*, Geom. Funct. Anal. 23 (2013) 415–466 MR Zbl

[59] **A J Tromba**, *On a natural algebraic affine connection on the space of almost complex structures and the curvature of Teichmüller space with respect to its Weil–Petersson metric*, Manuscripta Math. 56 (1986) 475–497 MR Zbl

[60] **S A Wolpert**, *Chern forms and the Riemann tensor for the moduli space of curves*, Invent. Math. 85 (1986) 119–145 MR Zbl

Department of Mathematics, University of California Riverside
Riverside, CA, United States

Department of Mathematics, Yale University
New Haven, CT, United States

Department of Mathematics, Heriot-Watt University
Edinburgh, United Kingdom

mdurham@ucr.edu, yair.minsky@yale.edu, a.sisto@hw.ac.uk

Smallest noncyclic quotients of
braid and mapping class groups

SUDIPTA KOLAY

We show that the smallest noncyclic quotients of braid groups are symmetric groups, proving a conjecture of Margalit. Moreover, we recover results of Artin and Lin about the classification of homomorphisms from braid groups on n strands to symmetric groups on k letters, where k is at most n. Unlike the original proofs, our method does not use the Bertrand–Chebyshev theorem, answering a question of Artin. Similarly, for mapping class group of closed orientable surfaces, the smallest noncyclic quotient is given by the mod two reduction of the symplectic representation. We provide an elementary proof of this result, originally due to Kielak and Pierro, which proves a conjecture of Zimmermann.

1 Introduction

The goal of this paper is to show that, with some obvious exceptions, the smallest noncyclic quotients of the braid and mapping class groups, are given by natural projections $\pi\colon B_n \to S_n$ (forgetful map) and $\Phi\colon \mathrm{Mod}(\Sigma_g) \to \mathrm{Sp}(2g, \mathbb{Z}_2)$ (mod two reduction of the symplectic representation). We begin by stating our main result for the Artin braid groups B_n.

Theorem 1 *Suppose $n = 3$ or $n \geq 5$. If G is a noncyclic quotient of B_n, then either $|G| > |S_n| = n!$ or G is isomorphic to S_n. Moreover, in the latter case the quotient map $B_n \to G$ is obtained by postcomposing the natural map π with an automorphism of S_n.*

There are no noncyclic quotients of B_n for $n \leq 2$, and for $n = 4$ the smallest noncyclic quotient is S_3, which is proved in Claim 7 in Section 2. Hence the hypothesis $n = 3$ or $n \geq 5$ is necessary in the theorem above.

The first statement of this theorem proves a conjecture of Margalit—see Chudnovsky, Kordek, Li and Partin [7] and Scherich and Verberne [18]—stating that the smallest noncyclic quotient of B_n is S_n for $n \geq 5$. For the nontrivial cases $n \in \{5, 6\}$, this was first proved by Caplinger and Kordek [4], and several recent papers [7; 4; 18] prove lower bounds for the order of noncyclic quotients of braid groups, using totally symmetric sets — see Kordek and Margalit [13, Section 2] — towards proving Margalit's conjecture. Our work builds further upon the idea of totally symmetric sets; see the discussion after Lemma 8.

Since the automorphisms of symmetric groups are well understood, the second statement in the theorem above immediately implies, for $n \neq 4$, the characterization of noncyclic[1] homomorphisms from $B_n \to S_k$, with $k \leq n$, originally due to Artin [1] for $k = n$ (and transitive homomorphisms) and improved by Lin [16, Theorem 3.9] for the remaining cases.

Corollary 2 *For $n \geq 3$ and $n \neq 4, 6$, all noncyclic homomorphisms $f : B_n \to S_n$ are conjugate to the standard projection π. Also, the only exceptional (up to conjugation) homomorphism $f : B_6 \to S_6$ comes from composing π with the only nontrivial (up to conjugation) outer automorphism of S_6, defined by $(12) \mapsto (1, 2)(3, 4)(5, 6)$ and $(1, 2, 3, 4, 5, 6) \mapsto (1, 2, 3)(4, 5)$.*

Artin noted that his proof in [1] "uses the existence of a prime between $\frac{1}{2}n$ and $n - 2$ for $n > 7$ but it would be preferable if a proof could be found that does not make use of this fact". This fact, known as the Bertrand–Chebyshev theorem [5], is also crucial for Lin's proof of the above result [16, Theorem 3.9]. Our proof here does not use this fact (and, to the best of our knowledge, this is the first such proof).

Remark 3 (exceptional case $n = 4$; Artin [1])   For completeness, we will record here the exceptional noncyclic homomorphisms (up to conjugations) from B_4 to S_k with $k \leq 4$. Let σ_1, σ_2 and σ_3 denote the Artin generators of B_4 and let $\alpha = \sigma_3 \sigma_2 \sigma_1$. We see that B_4 is generated by σ_1 and α. Then we have

(1) $f_1 : B_4 \to S_4$ defined by $\sigma_1 \mapsto (1, 2, 3, 4)$ and $\alpha \mapsto (1, 2)$;

(2) $f_2 : B_4 \to S_4$ defined by $\sigma_1 \mapsto (1, 3, 2, 4)$ and $\alpha \mapsto (1, 2, 3, 4)$;

(3) $f_3 : B_4 \to A_4 \subset S_4$ defined by $\sigma_1 \mapsto (1, 2, 3)$ and $\alpha \mapsto (1, 2)(3, 4)$ (here A_4 denotes alternating group on four letters, which uniquely embeds in S_4);

(4) $f_4 : B_4 \to S_3 \ (\subset S_4)$ defined by $\sigma_1 \mapsto (1, 2)$ and $\alpha \mapsto (1, 3)$.

---

[1]By this we mean the image of the homomorphism is not cyclic.

Our main result for mapping class groups $\text{Mod}(\Sigma_g)$ of closed orientable surfaces parallels Theorem 1, and is essentially the same as the result of Kielak and Pierro [12], using other methods.

Theorem 4 *Let $g \geq 1$. For any noncyclic quotient H of $\text{Mod}(\Sigma_g)$, either $|H| > |\text{Sp}(2g, \mathbb{Z}_2)|$ or H is isomorphic to $\text{Sp}(2g, \mathbb{Z}_2)$. Moreover, in the latter case the quotient map $\text{Mod}(\Sigma_g) \to H$ is obtained by postcomposing Φ with an automorphism of $\text{Sp}(2g, \mathbb{Z}_2)$.*

Zimmermann [19] proved that, for $g \in \{3, 4\}$, the smallest nontrivial[2] quotient of $\text{Mod}(\Sigma_g)$ is $\text{Sp}(2g, \mathbb{Z}_g)$, and conjectured the same statement holds for arbitrary $g \geq 3$. This conjecture was first proved by Kielak and Pierro [12] using the classification of finite simple groups and representation theory of mapping class groups. Moreover, Kielak and Pierro proved the same result holds for quotients of $\text{Mod}(\Sigma_g^b)$, where b is the number of boundary components, and we further extend their result here by allowing punctures as well.

Theorem 5 *Let $g \geq 3$. The smallest nontrivial quotient of $\text{Mod}(\Sigma_{g,n}^b)$ is $\text{Sp}(2g, \mathbb{Z}_2)$ for $n \in \{0, 1\}$, and $\mathbb{Z}_2$ for $n \geq 2$. If we furthermore assume $n \geq 5$, any noncyclic quotient of $\text{Mod}(\Sigma_{g,n}^b)$ of smallest order is isomorphic to either S_n or $\text{Sp}(2g, \mathbb{Z}_2)$ (depending on which group is smaller). Moreover, in any of the above cases, any epimorphism to a quotient of smallest order is the standard projection, postcomposed with an automorphism of the image.*

As indicated, some of the results above were previously known, but our proofs are considerably easier. For example, we do not use the classification of finite simple groups or the Bertrand–Chebyshev theorem. We use an *inductive orbit stabilizer method*, described in Section 3, which should also be applicable in other settings. Our approach is similar to that of Chudnovsky, Kordek, Li and Partin [7], Caplinger and Kordek [4] and particularly Scherich and Verberne [18], in that we all consider some group actions of the quotient (of braid groups), and use the orbit stabilizer theorem to find a bound on the size of the quotient. The advantage of our approach is that we prove an optimal lower bound on orbit size (by looking at the corresponding orbit size in the candidate smallest quotient), and moreover use induction to find the stabilizer size. For the two families of groups we consider here, this not only gives us the optimal

---

[2]For $g \geq 3$, $\text{Mod}(\Sigma_g)$ is perfect and therefore its smallest nontrivial and noncyclic quotients are the same.

lower bounds for size of the smallest quotient at the numerical level, but we also obtain the smallest quotient group up to isomorphism, and moreover a characterization of all possible minimal quotient maps.

Let us note that, if $G \to H$ and $H \to I$ are surjective group homomorphisms and I is smallest noncyclic (respectively nontrivial) quotient of G, then I is also the smallest noncyclic (respectively nontrivial) quotient of H. Thus, an immediate consequence of Theorems 4 and 5 is the following result:

Corollary 6 *For $g \geq 1$ (respectively $g \geq 3$), $\mathrm{Sp}(2g, \mathbb{Z}_2)$ is the smallest noncyclic (respectively nontrivial) quotient of $\mathrm{Sp}(2g, \mathbb{Z})$.*

2   Background

In this section we will collect several necessary definitions and results. We will also prove a claim, which will serve as base cases for our inductive proofs later.

Braid groups

The most well-known quotient of the braid group B_n [2] on n strands is the symmetric group S_n on n letters, obtained by forgetting all crossing information. This quotient map $\pi\colon B_n \to S_n$ can alternatively be described as adding the relations $\sigma_i^2 = 1$ (here the σ_i are half twists) to the Artin presentation [2] of the braid group B_n:

$$B_n = \{\sigma_1, \ldots, \sigma_{n-1} : \sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1} \text{ for all } 1 \leq i < n-1,$$
$$\sigma_i\sigma_j = \sigma_j\sigma_i \text{ if } |i - j| > 1\}.$$

Consider B_n as the mapping class group of the closed unit disc with n marked points $p_1, \ldots, p_n$ with increasing first coordinates and identical second coordinate.

Consider, for all $1 \leq i \neq j \leq n$, the arcs $\gamma_{i,j} = \gamma_{j,i}$ joining the p_i and p_j going over all p_k between p_i and p_j, and let $\rho_{i,j}$ denote the right-handed half twists about $\gamma_{i,j}$. For $1 \leq i < j \leq n$, the various $\rho_{i,j}$ are the *Birman–Ko–Lee generators* [3] generators of the braid group B_n, and we note that $\sigma_i = \rho_{i,i+1}$.

Mapping class groups

Let $\Sigma_{g,n}^b$ denote the orientable surface of genus g with n punctures and b boundary components (where we will drop n and b from the notation if they are zero), and denote its mapping class group by $\mathrm{Mod}(\Sigma_{g,n}^b)$. Our convention is that mapping classes preserve orientation, fix boundary components, and can permute the punctures. The subgroup $\mathrm{PMod}(\Sigma_{g,n}^b)$ will denote the pure mapping class group, consisting of mapping classes that fix the punctures.

We get an epimorphism Φ from $\mathrm{Mod}(\Sigma_g^b)$ by composing the capping homomorphism [8, Section 3.6.2] with the symplectic representation [8, Section 6.3] and the mod two reduction

$$\mathrm{Mod}(\Sigma_g^b) \xrightarrow{\text{capping}} \mathrm{Mod}(\Sigma_g) \xrightarrow{\text{symplectic}} \mathrm{Sp}(2g, \mathbb{Z}) \xrightarrow{\text{reduce}} \mathrm{Sp}(2g, \mathbb{Z}_2).$$

More generally, for $\Sigma_{g,n}$, let us consider the action of the mapping class group on homology. If we take a free basis of $H_1(\Sigma_{g,n}, \mathbb{Z})$ by taking a standard symplectic basis curves for each genus and a the class of a loop surrounding each puncture, the action of any mapping class can be represented by an invertible integral matrix in $\mathrm{GL}(2g+n, \mathbb{Z})$. Moreover, for any such matrix, the top left block is a symplectic matrix, the top right block is zero, and the bottom right block will be a permutation matrix. Thus, by projecting to diagonal blocks, we obtain epimorphisms from $\mathrm{Mod}(\Sigma_{g,n})$ (and thus from $\mathrm{Mod}(\Sigma_{g,n}^b)$ as well by capping) to $\mathrm{Sp}(2g, \mathbb{Z})$ (and hence to $\mathrm{Sp}(2g, \mathbb{Z}_2)$) and S_n. We will call these homomorphisms standard projections from $\mathrm{Mod}(\Sigma_{g,n}^b)$ to $\mathrm{Sp}(2g, \mathbb{Z}_2)$ and S_n. It can be seen that this standard projection from $\mathrm{Mod}(\Sigma_{g,n}^b)$ to S_n is the same as the induced action of the mapping classes on the punctures.

Some facts about symmetric and symplectic groups It is well known that, for $n \geq 5$, the only nontrivial quotient of S_n is $\mathbb{Z}_2$ (obtained by modding out by the simple group A_n). Also, it is known — see [9, Chapter 3] — that the symplectic group $\mathrm{Sp}(2g, \mathbb{Z}_2)$ is simple for $g \geq 3$, and for the exceptional cases we have the isomorphisms $\mathrm{Sp}(2, \mathbb{Z}_2) \cong S_3$ and $\mathrm{Sp}(4, \mathbb{Z}_2) \cong S_6$.

The following claim gives the base cases for our inductive proofs later:

Claim 7 *The smallest noncyclic quotient of B_3, B_4 and $\mathrm{Mod}(\Sigma_1) = \mathrm{SL}(2, \mathbb{Z})$ is S_3. Moreover, all epimorphisms from these three groups to S_3 are related by a conjugation of S_3.*

Proof The natural homomorphism π and f_4 from Remark 3 show S_3 is a quotient of B_3 and B_4, respectively. Moreover, it is easy to see that $\pi: B_3 \to S_3$ factors through $B_3 / Z(B_3) \cong \mathrm{PSL}(2, \mathbb{Z})$, and thus S_3 is a quotient of $\mathrm{PSL}(2, \mathbb{Z})$ and hence $\mathrm{SL}(2, \mathbb{Z})$. We note that all the groups except S_3 of order at most $|S_3| = 6$ are abelian (the only noncyclic group among them is the Klein four group), and thus cannot be a noncyclic quotient of a group with cyclic abelianization (such as braid groups or $\mathrm{SL}(2, \mathbb{Z})$). The last statement of the claim follows[3] by noting that the only pair of noncommuting elements in S_3 satisfying the braid relation are the transpositions. $\qquad \square$

3 The inductive orbit stabilizer method

The orbit stabilizer theorem is widely used in computing orders of finite groups which naturally act on a space, and, as this paper illustrates, it is also useful for determining orders of smallest noncyclic[4] quotients of groups. In our context we work with an infinite family of groups, and we can use the orbit stabilizer theorem inductively. We formulate the steps of the method below. While this method may not be new, proofs of similar results in the literature seem to rely on more complicated methods, as mentioned in the introduction.

Suppose we have a nested family of groups $(G_n)_{n \geq 1}$ with cyclic abelianizations. If we want to show the smallest noncyclic quotient is the family of groups $(H_n)_{n \geq 1}$, with a family of quotient maps $\pi_n: G_n \to H_n$, it suffices to carry out the following steps (after checking base cases):

(1) **Lower bound on orbit size** Find the size k of an orbit of the conjugation action of H_n. Find a suitable collection of elements $x_1, \ldots, x_k$ in G_n whose images generate the orbit, and show that the normal closure of each $x_i x_j^{-1}$ contains the commutator subgroup G_n' of G_n (equivalently, under any noncyclic quotient of G_n, the quotient classes $\bar{x}_i$ are all distinct).

(2) **Inductively find size of stabilizer** For some noncyclic quotient $q: G_n \to I_n$, inductively bound the size of the stabilizer of the quotient class of $q(x_1)$ in I_n, so

---

[3]For $n = 4$, a similar (but more tedious) check verifies Remark 3.

[4]It may be possible to adapt this method to find smallest nontrivial/nonabelian/nonsolvable quotients.

that the orbit stabilizer theorem implies $|I_n| \geq |H_n|$. For instance, if the centralizer of x_1 contains $\langle x_1 \rangle \times G_{n-i}$, it may be possible to get the desired result by applying the inductive hypothesis on the induced quotient $G_{n-i} \to q(G_{n-i})/Z(q(G_{n-i}))$. Finally, if $|I_n| = |H_n|$, show that I_n is isomorphic to H_n. This follows if the kernel of q contains the kernel of π_n, which moreover shows any epimorphism from G_n to H_n is π_n composed with an automorphism of H_n.

Some modifications, such as considering a different group action, may be needed to make this method work in a particular situation, and we will see one such modification for the mapping class groups case later.

4 Smallest noncyclic quotients of braid groups

We will carry out the steps of the inductive orbit stabilizer method here for Artin braid groups, and show that smallest noncyclic quotients are symmetric groups.

Lower bounds for size of orbit

Let us begin by observing that the conjugacy class of all transpositions in S_n consists of $\binom{n}{2}$ elements. We will take the x_i to be the Birman–Ko–Lee generators of the braid group, as mentioned in Section 2. The following lemma will complete the first step:

Lemma 8 *For $n \geq 5$ and a noncyclic quotient of B_n, the $\binom{n}{2}$ quotient classes $\overline{\rho}_{i,j}$ are distinct.*

We should note that the lemma does not hold for $n = 4$, as there is an exceptional homomorphism from B_4 to B_3 (which can be further quotiented to obtain $f_4 \colon B_4 \to S_3$, as mentioned in Remark 3) defined by $\sigma_1 \mapsto \sigma_1$, $\sigma_2 \mapsto \sigma_2$ and $\sigma_3 \mapsto \sigma_1$.

Totally symmetric sets are subsets of a group with the property that any homomorphism restricts to an injective map on that set or to a trivial map on that set (that is not the definition, but a consequence; see [13, Lemma 2.1]). Lemma 8 can be similarly phrased as saying that the set $\{\rho_{i,j}\}$ satisfies this same property. We will give two proofs of this lemma; the first is essentially in [6, Lemma 4.2], and the second is more hands-on.

Proof Suppose we have $\gamma_{i,j}$ and $\gamma_{k,l}$ with $\{i, j\} \neq \{k, l\}$ having the same quotient class. Since $n \geq 5$, we can find an arc δ between two marked points disjoint from $\gamma_{k,l}$ and sharing an endpoint with $\gamma_{i,j}$. It follows that δ and its image under $\rho_{i,j}\rho_{k,l}^{-1}$ share one endpoint and have disjoint interiors. Thus, by a change of coordinates principle

[8, Section 1.3.2], the commutator of ρ_δ (the right-handed half twist about δ) and $\rho_{i,j}\rho_{k,l}^{-1}$ is conjugate to $\sigma_1\sigma_2^{-1}$. Now, as $\rho_{i,j}\rho_{k,l}^{-1}$ is in the kernel of the quotient map, so is its commutator with ρ_δ, and thus so is $\sigma_1\sigma_2^{-1}$. The result now follows since $\sigma_1\sigma_2^{-1}$ normally generates B_n' (which is a direct consequence of the braid and far commutation relations), using the fact that B_n/B_n' is cyclic.   □

Alternative proof We will repeatedly use the following two observations:

(1) If two elements x and y in any group satisfy both the braid and far commutation relations, then $xyx = yxy \Longrightarrow xyx = xyy \Longrightarrow x = y$, ie x and y must coincide.

(2) For any distinct i, j and k, if $\overline{\rho}_{i,j}$ is same as $\overline{\rho}_{j,k}$, then, by the partial commutation relation,[5] they are also equal to $\overline{\rho}_{i,k}$.

Now, let us suppose the lemma is not true; let us first consider the case $\overline{\rho}_{i,j} = \overline{\rho}_{j,k}$ with distinct i, j and k, and, by the second observation above, we may assume $i < j < k$. For any l distinct from i, j and k, we see that, if l is (respectively is not) between i and j, then by the first observation we have $\overline{\rho}_{k,l} = \overline{\rho}_{j,k}$ (respectively $\overline{\rho}_{k,l} = \overline{\rho}_{i,j}$). By repeatedly applying the second observation, we see all the $\overline{\rho}_{i,j}$ must coincide, and thus the quotient is cyclic (as B_n is generated by the half twists σ_i), a contradiction.

Let us now consider the case $\overline{\rho}_{i,j} = \overline{\rho}_{k,l}$ for distinct i, j, k and l. Since $n + 1 \geq 5$, we can find m distinct from all of i, j, k and l. Let $o \in \{i, j, k, l\}$ be such that $|o - m|$ is smallest. By symmetry, without loss of generality, we may assume that $o \in \{i, j\}$. We see that $\overline{\rho}_{i,j}$ and $\overline{\rho}_{o,m}$ satisfies both the braid relation (as o is common) and the far commutation relation (as $\overline{\rho}_{i,j} = \overline{\rho}_{k,l}$, and $\gamma_{k,l}$ and $\gamma_{o,m}$ are disjoint). By the first observation, we must have $\overline{\rho}_{i,j} = \overline{\rho}_{o,m}$, and, by our discussion in the previous paragraph, all the $\overline{\rho}_{i,j}$ must be the same, again leading to a contradiction.   □

Inductive step

Now we will use induction to prove Theorem 1. (we repeat the statement below for convenience):

Inductive hypothesis Suppose $n = 3$ or $n \geq 5$. If G is a noncyclic quotient of B_n, then either $|G| > |S_n| = n!$ or G is isomorphic to S_n. Moreover, in the latter case the quotient map $B_n \to G$ is obtained by postcomposing the natural map π with an automorphism of S_n.

---

[5]For an appropriate $\epsilon \in \{-1, 1\}$, depending on the relative position among i, j and k, we have $\rho_{i,j}^\epsilon(\gamma_{j,k}) = \gamma_{i,k}$ and hence we get the partial commutation relation $\rho_{i,k} = \rho_{j,k}^\epsilon \rho_{i,j} \rho_{j,k}^{-\epsilon}$.

We will use induction on n in steps of two, and we will use the base case $n = 3$ from Claim 7, and the base case $n = 6$ from the computer-assisted proof of Caplinger and Kordek [4]. But we can also do the $n = 6$ case by hand with a separate argument similar to the inductive proof, as explained after this proof.

Proof idea Note that the centralizer of a transposition $(1, 2)$ in S_n is $\{1, (1, 2)\} \times S_{n-2}$, where S_{n-2} is the symmetric group on the letters $3, \ldots, n$. Similarly, we see that the centralizer of $x_1 = \sigma_1$ in B_n contains $\langle \sigma_1 \rangle \times B_{n-2}$, which projects to $\{1, (1, 2)\} \times S_{n-2}$ under π. If, under some noncyclic quotient of B_n, the centralizer of $\bar{x}_1$ is $\langle \bar{x}_1 \rangle \times \bar{B}_{n-2}$, then use the inductive hypothesis on the size of $\bar{B}_{n-2}$. But $\langle \bar{x}_1 \rangle$ and $\bar{B}_{n-2}$ may not intersect trivially; however, we see that their intersection is central in $\bar{B}_{n-2}$. Therefore, we can use the inductive hypothesis on $\bar{B}_{n-2}/Z(\bar{B}_{n-2})$.

Proof of Theorem 1 As mentioned above, we will use the base cases $n \in \{3, 6\}$, and use induction on n in steps of two, and this will imply the result for all odd $n \geq 5$ and even $n \geq 8$.

We will assume the inductive hypothesis is true for $k = n - 1$ and prove the statement for $k = n + 1$ (with $n + 1 \geq 5$). Suppose $q \colon B_{n+1} \to G$ is a noncyclic quotient of smallest order. By Lemma 8, it follows that all the $\frac{1}{2}(n + 1)n$ quotient classes $\bar{\rho}_{i,j}$ must be distinct for noncyclic G. It is known that all the $\rho_{i,j}$ are conjugate in B_{n+1}, so the $\bar{\rho}_{i,j}$ are conjugate in G. Therefore, if we consider the group action of G on itself by conjugation, the orbit stabilizer theorem tells us

$$(1) \qquad\qquad |G| = |O||C| \geq \tfrac{1}{2}(n + 1)n|C|,$$

where C denotes the centralizer (ie stabilizer of the conjugation action) of the element $\bar{\rho}_{1,2}$ and O denotes its conjugacy class (ie the image of the half twists). Since $\sigma_1 = \rho_{1,2}$ commutes with the subgroup $V_{1,2}$ of B_{n+1} generated by $\sigma_3, \ldots, \sigma_n$ (thus $V_{1,2}$ is isomorphic to B_{n-1}), we see C contains $H_{1,2} := q(V_{1,2})$ as a subgroup, and clearly it also contains $\bar{\rho}_{1,2}$. It follows from Lemma 8 that $H_{1,2}$ is not cyclic, and so we can apply the inductive hypothesis to any noncyclic quotient of $H_{1,2}$.

Let M denote the cyclic subgroup generated by $\bar{\rho}_{1,2}$ in G. We see that $Y = H_{1,2} \cap M$ is in the center Z of $H_{1,2}$ as $\bar{\rho}_{1,2}$ commutes with all elements $H_{1,2}$. If $H_{1,2}/Z$ is cyclic, we know that $H_{1,2}$ is abelian, but, as $H_{1,2}$ is a quotient of $V_{1,2} \cong B_{n-1}$, it has to factor through the abelianization and is therefore cyclic, contradicting Lemma 8. Hence, $H_{1,2}/Z$ is a noncyclic quotient of B_{n-1}, and so, by the inductive hypothesis for

$k = n - 1$, we have $|H_{1,2}/Z| \geq (n-1)!$. Thus, we have $|H_{1,2}| \geq |Z|(n-1)! \geq |Y|(n-1)!$. Also, if D denotes the subgroup of C generated by $H_{1,2}$ and M, we see that M is in the center of D and thus $|D| = |M/Y||H_{1,2}| \geq |M|(n-1)!$.

By combining with (1), we see that

$$(2) \quad |G| \geq \tfrac{1}{2}(n+1)n|C| \geq \tfrac{1}{2}(n+1)n|D| \geq \tfrac{1}{2}(n+1)n|M|(n-1)! = (n+1)! \cdot \tfrac{1}{2}|M|.$$

Thus, the only way $|G| \leq (n+1)!$ is if $|M| = 1$ (in this case $\overline{\rho}_{1,2} = 1$, so G is the trivial group, a contradiction) or $|M| = 2$. If the latter case happens then $q(\sigma_i^2) = 1$ for all i, and thus q factors through the standard quotient map $\pi \colon B_{n+1} \to S_{n+1}$. Since the only proper quotient of S_{n+1} (for $n+1 \geq 5$) is $\mathbb{Z}_2$, it must be the case that G is isomorphic to S_{n+1}, as required. Moreover, this shows that q is a composition of the standard map π with an automorphism of S_{n+1}. □

Proof of Theorem 1 for $n = 6$ We will show the desired result for this case using a similar argument as above, and we use the same notation. Let m denote the order of $\overline{\rho}_{1,2}$ in G (a noncyclic quotient of B_6 of smallest order). If $m = 2$, we know $q \colon B_6 \to G$ factors through S_6, and therefore the desired result holds, so we will assume $m > 2$ hereafter. By (1), we have $|G| \geq 15|C| \geq 15|H_{1,2}|$. The following claim gives a lower bound on $|H_{1,2}|$ which implies $|G| \geq 6!$, and thus $|G| = 6!$:

Claim 9 *For $m > 2$, we have $|H_{1,2}| \geq 48$, and equality holds only if $m = 4$ and $\overline{\sigma}_3^2 = \overline{\sigma}_5^2$.*

Proof We see that the $\binom{4}{2} = 6$ elements $\overline{\rho}_{i,j}$ are distinct for $3 \leq i < j \leq 6$ (we are applying Lemma 8 for $n = 6$, and not 4). Thus, by the orbit stabilizer theorem, we have $|V_{1,2}| = |\widehat{O}||\widehat{C}|$, where $\widehat{O}$ and $\widehat{C}$ denote the orbit and centralizer of the element $\overline{\rho}_{3,4} = \overline{\sigma}_3$ in $V_{1,2}$. We see that $\widehat{C}$ contains the cyclic subgroups generated by the commuting elements $\overline{\sigma}_3$ and $\overline{\sigma}_5$.

If these subgroups coincide, we will have $\overline{\sigma}_5 = \overline{\sigma}_3^p$ for some p, and, by an appropriate conjugation in G (by the image of a periodic braid), we get $\overline{\sigma}_3 = \overline{\sigma}_1^p$ and $\overline{\sigma}_4 = \overline{\sigma}_2^p$. It would therefore follow that G is generated by $\overline{\sigma}_1$ and $\overline{\sigma}_2$, but then the stabilizer of $\overline{\sigma}_4$ is all of G, contradicting that we have a nontrivial orbit of $\overline{\sigma}_4$. Thus, $\widehat{C}$ properly contains the cyclic subgroup generated by $\overline{\sigma}_5$, and so $|\widehat{C}| \geq 2m$. For $m = 3$, we see the subgroups generated by $\overline{\sigma}_3$ and $\overline{\sigma}_5$ cannot intersect (or otherwise they coincide) and therefore $|\widehat{C}| = 9$, and thus $|V_{1,2}| \geq 6 \cdot 9 = 54$. Lastly, for $m \geq 4$, we have $|\widehat{C}| \geq 2m$ and so $|V_{1,2}| = |\widehat{O}||\widehat{C}| \geq 6 \cdot 2m = 12m \geq 48$. Moreover, it is easily checked that $|V_{1,2}| = 48$ if and only if $m = 4$ and $\overline{\sigma}_3^2 = \overline{\sigma}_5^2$. □

It remains to consider the case $|G| = 6!$, $m = 4$ and $\overline{\sigma}_3^2 = \overline{\sigma}_5^2$. By conjugation by the image of a periodic braid, it follows that $\overline{\sigma}_1^2 = \overline{\sigma}_3^2$. The nontrivial (since $m \neq 2$) element $\overline{\sigma}_1^2$ (commuting with $\overline{\sigma}_1$, $\overline{\sigma}_3$, $\overline{\sigma}_4$ and $\overline{\sigma}_5$) is in the center of G, as $\overline{\sigma}_2$ commutes with $\overline{\sigma}_5^2$ $(= \overline{\sigma}_1^2)$. Thus, G has nontrivial center $Z(G)$, and so $G/Z(G)$ must be a strictly smaller noncyclic quotient of B_6, a contradiction. □

We will now see how Theorem 1 implies Artin and Lin's results.

Proof of Corollary 2 If $f : B_n \to S_k$ is a noncyclic homomorphism, by Theorem 1, we must have $k = n$ and we have $f = g \circ \pi$, where $g : S_n \to S_n$ is an automorphism. Now we use the fact, due to Hölder [11], that for $n \neq 2, 6$ all automorphisms of S_n are inner, and there is exactly one outer automorphism of S_6 up to conjugation, which is mentioned in the statement of the corollary. □

5   Smallest noncyclic quotients of mapping class groups

We will use a slightly modified form of the inductive orbit stabilizer method here. In the inductive step, it will be more convenient to look at the conjugation action on a pair of elements (instead of a single element) of the quotient.

Lower bounds for size of orbit

We note that the orbit of all transvections in $\mathrm{Sp}(2g, \mathbb{Z}_2)$ is $2^{2g} - 1$, since these are in bijection with primitive vectors in $(\mathbb{Z}/2\mathbb{Z})^{2g}$. In this case we will take the x_i to be suitable right-handed Dehn twists about simple closed curves, so that their mod two homology classes give us all primitive vectors in $(\mathbb{Z}/2\mathbb{Z})^{2g}$. Corresponding to each primitive vector v with zeroes and ones in the first homology $H_1(\Sigma_g, \mathbb{Z})$, we will construct a simple closed curve α_v realizing this homology class, and denote by T_v the
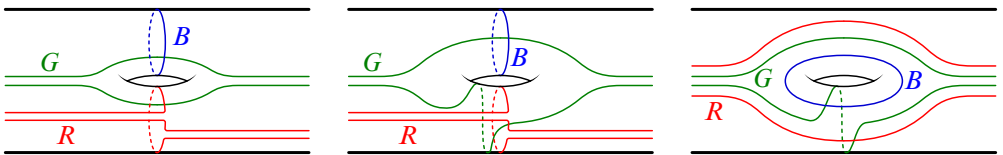

Figure 1: Illustrative examples of finding a new simple closed curve B (in blue) disjoint from the red curve R and having geometric intersection number one with the green curve G.

right-handed Dehn twist about the curve α_v. Starting at the first entry of v, for each nonzero pair (p, q) of entries we can draw the (p, q) curve on the corresponding genus, and we can join these curves by standard bands running straight across. For instance, the red and green curves in Figure 1, left, show the $(1, 0)$ and $(0, 1)$ curves on a genus, which is then band-summed with the other (p, q) curves. It is easy to see that, if we have two binary vectors v and w which differ on the same pair of entries, then, by localizing to the corresponding genus, we can find a third simple closed curve β which intersects exactly one of α_v or α_w once and is disjoint from the other, as illustrated in Figure 1.

Remark 10 As Dan Margalit pointed out to us, the above construction can also be done using double branched covers, which can be more useful in certain situations. By quotienting out by the hyperelliptic involution, we can consider Σ_g as a double branched cover over the sphere Σ_0, with $2g + 2$ branch points $y_1, \ldots, y_{2g+2}$. We can think of the branched cover of the pair of branch points y_{2g+1} and y_{2g+2} as forming a tube connecting two disjoint Σ_0^{g+1}'s, and the rest of the pairs correspond to adding genus. For each subset of the first $2g$ branch points, we can consider a simple closed curve in $\Sigma_{0,2g+2}$ enclosing these points (we think of the region containing y_{2g+2} as outside), and if necessary y_{2g+1} so that the total number of points is even. The lift of this curve realizes the mod two homology class of the binary vector corresponding to which branch points were chosen (in fact there is a bijection between $H_1(\Sigma_g; \mathbb{Z}_2)$ and the even subgroup of $H_1(\Sigma_{0,2g+2}; \mathbb{Z}_2)$). Lastly, let us observe that, given any two mod two nonhomologous simple closed curves in $\Sigma_{0,2g+2}$, it is possible to choose an arc joining two branch points which intersects one and is disjoint from the other, and its lift is a simple closed curve in Σ_g having the same property.

Lemma 11 *For $g \geq 1$ and any noncyclic quotient of* $\mathrm{Mod}(\Sigma_g)$*, the* $2^{2g} - 1$ *quotient classes* $\overline{T}_v$ *must be distinct.*

Proof Suppose we have two different binary vectors v and w such that $\overline{T}_v = \overline{T}_w$. By our above discussion, we can find a simple closed curve β such that T_β commutes with one of $\overline{T}_v$ or $\overline{T}_w$ and satisfies the braid relation with the other. Hence, by the first observation in the alternative proof of Lemma 8, we see that, for two simple closed curves c and d with geometric intersection number one, we have $\overline{T}_c = \overline{T}_d$. By [15, Lemma 2.1], the quotient must be abelian (and hence cyclic since all abelianizations of $\mathrm{Mod}(\Sigma_g)$ are cyclic [8, Chapter 5]), which gives a contradiction.   □

Inductive step

In this step, we will consider the conjugation action on a pair of group elements, but the size of the orbit readily follows from the conjugation action considered in the first step.

Proof of Theorem 4 Let us first recall the statement we are going to prove:

Inductive hypothesis Let $g \geq 1$. For any noncyclic quotient H of $\mathrm{Mod}(\Sigma_g)$, either $|H| > |\mathrm{Sp}(2g, \mathbb{Z}_2)|$ or H is isomorphic to $\mathrm{Sp}(2g, \mathbb{Z}_2)$. Moreover, in the latter case the quotient map $\mathrm{Mod}(\Sigma_g) \to H$ is obtained by postcomposing Φ with an automorphism of $\mathrm{Sp}(2g, \mathbb{Z}_2)$.

We will use induction on g, and we note that the base case $g = 1$ follows from Claim 7. We will inductively assume the statement is true for $k = g - 1$ (with $g \geq 2$), and prove it for $k = g$. Let $q \colon \mathrm{Mod}(\Sigma_g) \to H$ be a quotient of smallest order. Let R and S denote the right-handed Dehn twists about the simple closed curves α_{e_1} and α_{e_2} (we use e_i to denote the ith standard basis vector in $\mathbb{Z}^{2g}$, and the same notation as in the previous section). By Lemma 11, we know that the conjugacy class of the quotient class $\overline{R}$ in H has size at least $2^{2g} - 1$. We will consider the conjugation action of G on the set of all ordered pairs of elements in G. Using our original collection of curves α_v, we have $2^{2g-1}(2^{2g} - 1)$ ordered pairs with geometric intersection number one, so, by the change of coordinates principle [8, Section 1.3.3], we see that the orbit of the ordered pair $(\overline{R}, \overline{S})$ under the conjugation action is at least $2^{2g-1}(2^{2g} - 1)$. We see the stabilizer of $(\overline{R}, \overline{S})$ contains the image I under q of $\mathrm{Mod}(\Sigma_{g-1}^1)$ (where Σ_{g-1}^1 is obtained by cutting Σ_g along the separating curve which is the boundary of a regular neighborhood of α_{e_1} and α_{e_2}, ie we are deleting the leftmost genus containing α_{e_1} and α_{e_2}), since $\mathrm{Mod}(\Sigma_g^1)$ fixes α_{e_1} and α_{e_2}. If $Z(I)$ denotes the center of this image I, we see that $I/Z(I)$ is a noncyclic quotient (otherwise I must be abelian, and thus the various conjugate T_v must map to the same element, contradicting Lemma 11) of $\mathrm{Mod}(\Sigma_{g-1}^1)$. Since the boundary parallel Dehn twist in $\mathrm{Mod}(\Sigma_{g-1}^1)$ is central, it follows that $I/Z(I)$ is also a noncyclic quotient of $\mathrm{Mod}(\Sigma_{g-1})$. By the inductive hypothesis for $k = g - 1$, we have that $|I| \geq |I/Z(I)| \geq |\mathrm{Sp}(2g - 2, \mathbb{Z}_2)|$. Thus, by the orbit stabilizer theorem, we have

$$(3) \quad |H| \geq 2^{2g-1}(2^{2g} - 1)|I| \geq 2^{2g-1}(2^{2g} - 1)|\mathrm{Sp}(2g - 2, \mathbb{Z}_2)| = |\mathrm{Sp}(2g, \mathbb{Z}_2)|.$$

Thus, we get the desired result at the numerical level and, moreover, in the case of equality above, we see that $Z(I)$ is trivial. Moreover, from the inductive hypothesis

we have that I is isomorphic to $\mathrm{Sp}(2g-2,\mathbb{Z}_2)$. It follows that separating twists and, for $g \geq 3$, genus one bounding pairs are in the kernel of q. By results of Birman, Powell and Johnson [10], for $g \geq 3$ (respectively $g = 2$), genus one bounding pairs (respectively separating twists) normally generate the Torelli group, so we see that q factors through $q_1 \colon \mathrm{Sp}(2g,\mathbb{Z}) \to H$. Moreover, by the inductive hypothesis, some $\overline{T}_v$ has order 2, and so the kernel of q_1 contains squares of all transvections, and thus, by [17, Proposition A3], the kernel of q_1 contains the level two congruence subgroup. Consequently, q in fact factors through $\mathrm{Sp}(2g,\mathbb{Z}_2)$, and the result follows. $\qquad\square$

6 Allowing punctures and boundary components

In this final section, we will see some results about smallest noncyclic/nontrivial quotients of $\mathrm{Mod}(\Sigma_{g,n}^b)$. These results are consequences of our main results and facts about the abelianizations of mapping class groups, discussed below:

Abelianization of mapping class groups It is known [14, Theorem 5.1] that the abelianization of the pure[6] mapping class group $\mathrm{PMod}(\Sigma_{g,n})$ is

(1) $\mathbb{Z}_{12}$ if $g = 1$ and $b = 0$;

(2) $\mathbb{Z}^b$ if $g = 1$ and $b \geq 1$;

(3) $\mathbb{Z}_{10}$ if $g = 2$, and

(4) trivial if $g \geq 3$.

This implies the following result (likely known, but we could not find it in the literature):

Lemma 12 *The abelianization of* $\mathrm{Mod}(\Sigma_{g,n}^b)$ *equals* $\mathbb{Z}_2$ *for* $g \geq 3$ *and* $n \geq 2$.

Proof By the above result and the change of coordinates principle, we see under the abelianization map of $\mathrm{Mod}(\Sigma_{g,n}^b)$, all essential Dehn twists map to the identity, and all right-handed half twists map to the same element. If we consider the subsurface Σ_g^c of $\Sigma_{g,n}^b$ such that almost all the additional boundary components added consist of standard loops enclosing exactly two punctures (and one containing a single puncture if n is odd), we see that squares of half twists must also map to the identity in the abelianization of $\mathrm{Mod}(\Sigma_{g,n}^b)$. The result follows by noting that the abelianization cannot be trivial since we have an epimorphism from $\mathrm{Mod}(\Sigma_{g,n}^b)$ to S_n, and hence to $\mathbb{Z}_2$. $\quad\square$

---

[6]We caution the reader that the reference we are citing follows the convention that mapping classes fix punctures and thus their mapping class group coincides with our pure mapping class group.

We now find the smallest nontrivial quotient of $\mathrm{Mod}(\Sigma_{g,n}^b)$ for $g \geq 1$ and arbitrary n and b.

Theorem 13 *The smallest nontrivial quotient of* $\mathrm{Mod}(\Sigma_{g,n}^b)$ *of smallest order is*

(1) $\mathbb{Z}_2$ *for* $n \geq 2$ *or* $g \in \{1, 2\}$, *and arbitrary* b;

(2) $\mathrm{Sp}(2g, \mathbb{Z}_2)$ *for* $g \geq 3$ *and* $n \in \{0, 1\}$, *and arbitrary* b.

Proof For $n \geq 2$, we get an epimorphism $\mathrm{Mod}(\Sigma_{g,n}^b) \to S_n$ by considering the action on the punctures, and we can further quotient to the unique smallest nontrivial group $\mathbb{Z}_2$. Thus, it only remains to consider $n \in \{0, 1\}$, and so all mapping classes are pure. From the aforementioned result about abelianization, we see that, for $g \in \{1, 2\}$, the smallest nontrivial quotient is $\mathbb{Z}_2$. Also, the same result tells us that, for $g \geq 3$, there can be no nontrivial abelian quotients. Hence, all boundary parallel and puncture surrounding Dehn twists (which are central) must map to the identity under any nontrivial quotient of smallest order (otherwise we get an even smaller nontrivial quotient by quotienting by the center), and thus we reduce to the case in Theorem 4. □

We also find the smallest noncyclic quotient of $\mathrm{Mod}(\Sigma_{g,n}^b)$ for a wide range of cases.

Theorem 14 *Any noncyclic quotient of* $\mathrm{Mod}(\Sigma_{g,n}^b)$ *of smallest order is*

(1) *the smaller of the groups among* S_n *and* $\mathrm{Sp}(2g, \mathbb{Z}_2)$ *for* $g \geq 3$, $n \geq 5$ *and arbitrary* b;

(2) S_3 *for* $g \geq 3$, $n \in \{3, 4\}$ *and arbitrary* b;

(3) $\mathrm{Sp}(2g, \mathbb{Z}_2)$ *for* $g \geq 2$, $n \in \{0, 1\}$ *and arbitrary* b (*also for* $g = 1$ *and* $n, b \in \{0, 1\}$);

(4) $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ *for* $g = 1$, $n \in \{0, 1\}$ *and* $b \geq 2$.

Proof Let us consider the center of a noncyclic quotient of $\mathrm{Mod}(\Sigma_{g,n}^b)$ of smallest order. The only way this center is nontrivial is if the quotient is noncyclic abelian (otherwise we get a strictly smaller noncyclic quotient). This situation does happen for $g = 1$, $n \in \{0, 1\}$ and $b \geq 2$, where the abelianization of $\mathrm{Mod}(\Sigma_{g,n}^b)$ is $\mathbb{Z}^b$, which has the Klein four group (the unique noncyclic group of smallest order) as a quotient.

Also, the above is the only case (among the ones mentioned in the statement) where this can happen, since the abelianization of $\mathrm{Mod}(\Sigma_{g,n}^b)$ is $\mathbb{Z}_2$ for $g \geq 3$ and $n \geq 2$, and $\mathbb{Z}/10\mathbb{Z}$ for $g = 2$ and $n \in \{0, 1\}$. Thus, for these cases, the smallest noncyclic quotient

must necessarily be nonabelian. Moreover, all boundary parallel Dehn twists must map to the trivial element in the quotient, and so we reduce to the case $b = 0$ (and, if $n = 1$, the Dehn twist about the curve surrounding the puncture is also central, so we can also reduce to the case $n = 0$). Hence, for $g \geq 2$, $n \in \{0, 1\}$ and arbitrary b (and also for $g = 1$ and $n, b \in \{0, 1\}$), we reduce to the case $n = b = 0$, and we get the desired result by Theorem 4.

If $n \in \{3, 4\}$ and $g \geq 3$, we see that S_3 is a quotient of $\mathrm{Mod}(\Sigma_{g,n}^b)$ (using the induced action on the punctures and the exceptional homomorphism $S_4 \to S_3$). As we saw earlier, S_3 must be the smallest quotient in this case as it is the unique smallest nonabelian group.

Finally, we now consider the case $g \geq 3$, $n \geq 5$ and $b = 0$. Suppose we have a quotient of $\mathrm{Mod}(\Sigma_{g,n})$ such that the restriction to both $\mathrm{Mod}(\Sigma_g^1)$ and $B_n \cong \mathrm{Mod}(\Sigma_{0,n}^1)$ are both cyclic. Then, by Theorem 13, it must be the case that the image of $\mathrm{Mod}(\Sigma_g^1)$ is trivial. Moreover, by the braid relation, all half twists in $\mathrm{Mod}(\Sigma_{g,n})$ must map to a single element. Given any Dehn twist in $\mathrm{Mod}(\Sigma_{g,n})$, by a change of coordinate we can find a half twist commuting with it. So we see that the image of each half twist is a central element in the quotient, as $\mathrm{Mod}(\Sigma_{g,n})$ is generated by Dehn twists and half twists. This contradicts our observation earlier, so one of the restrictions to $\mathrm{Mod}(\Sigma_g^1)$ or B_n is noncyclic, giving us the desired result by using Theorems 1 and 13. □

To complete the proof of Theorem 5, it remains to verify the statement about maps. However, let us note that the corresponding statement is not true for all the cases in Theorem 14. For instance, for $b \geq 3$, there are multiple epimorphisms from $\mathrm{Mod}(\Sigma_1^b)$ to $\mathbb{Z}_2 \oplus \mathbb{Z}_2$, even up to postcomposing by automorphisms of the image.

Proof of Theorem 5 For $g \geq 3$ and $n \geq 5$, let us first consider the case that the quotient of $\mathrm{Mod}(\Sigma_{g,n}^b)$ of smallest order is S_n. We know from the proof of Theorem 14 that we can reduce to the case $b = 0$, and the restriction of this quotient on $\mathrm{Mod}(\Sigma_{0,n}^1)$ is S_n as well. Since $\mathrm{Mod}(\Sigma_{0,n}^1)$ commutes with $\mathrm{Mod}(\Sigma_g^1)$, and S_n is centerless, it follows that $\mathrm{Mod}(\Sigma_g^1)$ is in the kernel of this quotient map. As all Dehn twists in $\mathrm{Mod}(\Sigma_{g,n})$ are conjugate, it follows that the kernel contains the pure mapping class group $\mathrm{PMod}(\Sigma_{g,n})$. Consequently, the quotient map factors through $\mathrm{Mod}(\Sigma_{g,n})/\mathrm{PMod}(\Sigma_{g,n}) \cong S_n$, and the desired result follows.

For $g \geq 3$ and $n \geq 5$, let us now consider the case that the quotient of $\mathrm{Mod}(\Sigma_{g,n}^b)$ of smallest order is $\mathrm{Sp}(2g, \mathbb{Z}_2)$. Similar to our above discussion, we see that the quotient

map restricted to $\mathrm{Mod}(\Sigma_g^1)$ is surjective, and all half twists are in the kernel of the quotient map. We know the epimorphism from $\mathrm{Mod}(\Sigma_g^1)$ to $\mathrm{Sp}(2g, \mathbb{Z}_2)$ has to send the boundary parallel Dehn twists to the identity, and so it factors through $\mathrm{Mod}(\Sigma_g)$. By Theorem 4, we know this map is the standard projection Φ, up to an automorphism h of $\mathrm{Sp}(2g, \mathbb{Z}_2)$. By looking at the action on $\mathbb{Z}_2^{2g}$, we see that all the Dehn twists about curves not contained in Σ_g^1 (next to the punctures) in [8, Figure 4.10] must map to the same element. Since we also know that all half twists are in the kernel of the quotient map $\mathrm{Mod}(\Sigma_{g,n}^b) \to \mathrm{Sp}(2g, \mathbb{Z}_2)$, it follows that this map coincides with the standard projection, postcomposed with the same automorphism h of $\mathrm{Sp}(2g, \mathbb{Z}_2)$. ,

For $g \geq 3$ and $n \in \{0, 1\}$, the result follows by the same argument as in the last paragraph. Lastly, for $g \geq 3$ and $n \geq 2$, any homomorphism from $\mathrm{Mod}(\Sigma_{g,n}^b)$ to an abelian group must factor through the abelianization of $\mathrm{Mod}(\Sigma_1^b)$, which, by Lemma 12, is $\mathbb{Z}_2$. Hence, the result follows, and moreover this map is unique since $\mathbb{Z}_2$ does not have a nontrivial automorphism. □

References

[1] **E Artin**, *Braids and permutations*, Ann. of Math. 48 (1947) 643–649 MR Zbl

[2] **E Artin**, *Theory of braids*, Ann. of Math. 48 (1947) 101–126 MR Zbl

[3] **J Birman**, **K H Ko**, **S J Lee**, *A new approach to the word and conjugacy problems in the braid groups*, Adv. Math. 139 (1998) 322–353 MR Zbl

[4] **N Caplinger**, **K Kordek**, *Small quotients of braid groups*, preprint (2020) arXiv 2009.10139

[5] **P Chebyshev**, *Mémoire sur les nombres premiers*, J. Math. Pures Appl. 17 (1852) 366–390

[6] **L Chen**, **K Kordek**, **D Margalit**, *Homomorphisms between braid groups*, preprint (2019) arXiv 1910.00712

[7] **A Chudnovsky**, **K Kordek**, **Q Li**, **C Partin**, *Finite quotients of braid groups*, Geom. Dedicata 207 (2020) 409–416 MR Zbl

[8] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR Zbl

[9] **L C Grove**, *Classical groups and geometric algebra*, Graduate Studies in Math. 39, Amer. Math. Soc., Providence, RI (2002) MR Zbl

[10] **A Hatcher**, **D Margalit**, *Generating the Torelli group*, Enseign. Math. 58 (2012) 165–188 MR Zbl

[11] **O Hölder**, *Bildung zusammengesetzter Gruppen*, Math. Ann. 46 (1895) 321–422 MR Zbl

[12] **D Kielak**, **E Pierro**, *On the smallest non-trivial quotients of mapping class groups*, Groups Geom. Dyn. 14 (2020) 489–512 MR Zbl

[13] **K Kordek**, **D Margalit**, *Homomorphisms of commutator subgroups of braid groups*, Bull. Lond. Math. Soc. 54 (2022) 95–111 MR Zbl

[14] **M Korkmaz**, *Low-dimensional homology groups of mapping class groups: a survey*, Turkish J. Math. 26 (2002) 101–114 MR Zbl

[15] **J Lanier**, **D Margalit**, *Normal generators for mapping class groups are abundant*, Comment. Math. Helv. 97 (2022) 1–59 MR Zbl

[16] **V Lin**, *Braids and permutations*, preprint (2004) arXiv math/0404528

[17] **D Mumford**, *Tata lectures on theta, I*, Progr. Math. 28, Birkhäuser, Boston, MA (1983) MR Zbl

[18] **N Scherich**, **Y Verberne**, *Finite image homomorphisms of the braid group and its generalizations*, preprint (2020) arXiv 2012.01378

[19] **B P Zimmermann**, *On minimal finite quotients of mapping class groups*, Rocky Mountain J. Math. 42 (2012) 1411–1420 MR Zbl

School of Mathematics, Georgia Institute of Technology
Atlanta, GA, United States

sudiptakolay.gt@gmail.com

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