Geometry & Topology

Volume 20 (2016)

Issue 6 (pages 3057–3673)

msp

GEOMETRY & TOPOLOGY

msp.org/gt

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

The subscription price for 2016 is US \$460/year for the electronic version, and \$695/year (+\$60, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscribers address should be sent to MSP. Geometry & Topology is indexed by Mathematical Reviews, Zentralblatt MATH, Current Mathematical Publications and the Science Citation Index.

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

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

PUBLISHED BY
mathematical sciences publishers
nonprofit scientific publishing
http://msp.org/

Geometry & Topology 20 (2016) 3057–3096

Asymptotic formulae for curve operators in TQFT

RENAUD DETCHERRY

The Reshetikhin–Turaev topological quantum field theories with gauge group SU_2 associate to any oriented surface Σ a sequence of vector spaces $V_r(\Sigma)$ and to any simple closed curve γ in Σ a sequence of Hermitian operators T_r^γ on the spaces $V_r(\Sigma)$. These operators are called curve operators and play a very important role in TQFT.

We show that the matrix elements of the operators T_r^γ have an asymptotic expansion in orders of $1/r$, and give a formula to compute the first two terms from trace functions, generalizing results of Marché and Paul for the punctured torus and the 4–holed sphere to general surfaces.

57R56

1 Introduction

Witten [28] proposed in 1989, by a method using Feynman path integrals, a family of new invariants of 3–manifolds derived from the Jones polynomial, together with the structure of a full topological quantum field theory. Reshetikhin and Turaev [24] formalized the ideas of Witten to construct a family $(Z_{2r}(M))_{r\in\mathbb{N}^*}$ of 3–manifolds invariants. Also they defined a TQFT-structure for these invariants in [24] and Turaev [27]. An alternative method to define these 3–manifold invariants and TQFTs using skein theory of 3–manifolds was later developed by Blanchet, Habegger, Masbaum and Vogel [11].

Let Σ be a closed oriented surface maybe with marked points p_i colored by elements $\hat{c}_i$ of $\mathcal{C}_r = \{1,\dots,r-1\}$. Neglecting the so-called framing anomaly, the construction of [11] associates a vector space $V_r(\Sigma,\hat{c})$ to $(\Sigma,\hat{c})$ and, for any cobordism (M,Σ_0,Σ_1) containing a link L, there is a morphism

$$V_r(M,L)\colon V_r(\Sigma_0)\to V_r(\Sigma_1)$$

such that for every closed orientable 3–manifold M we have $V_r(M)=Z_{2r}(M)$.

Let us recall that a multicurve on Σ is a disjoint union of simple closed curves on Σ. In particular, the construction associates to any multicurve γ on Σ a curve operator

$$T_r^\gamma = V_r\left(\Sigma\times[0,1],\gamma\times\left\{\tfrac{1}{2}\right\}\right)\in \mathrm{End}(V_r(\Sigma,\hat{c})).$$

Published: 21 December 2016 DOI: 10.2140/gt.2016.20.3057

Curve operators often play a central role in TQFT; they were used to derive the asymptotic faithfulness of quantum representations, or to relate the combinatorial and the geometric framework of TQFT; see Andersen [1; 2] or Andersen and Ueno [7; 8; 9; 10].

From the construction of [11] it follows also that each vector space $V_r(\Sigma,\hat{c})$ comes with a natural Hermitian form.

Recall that a pants decomposition of a surface Σ with marked points is a finite family of simple closed curves on Σ which cut Σ into either pair of pants containing no marked point or disks containing exactly one marked point.

We will say that a trivalent banded graph Γ inside Σ is *compatible* with a pair of pants decomposition $\mathcal{C}=(C_e)_{e\in E}$ if the following conditions are satisfied:

- Γ has a trivalent vertex v_P lying in each pair of pants P of the decomposition, and these are the only trivalent vertices of Γ.
- For every $e\in E$, Γ has exactly one edge (labeled also by e) that intersects the curve C_e. This edge is disjoint from the other curves C_f for $f\in E\setminus\{e\}$, and intersects C_e exactly once.
- The graph Γ has n univalent vertices labeled by $p_1,\dots,p_n$ corresponding to the marked points of Σ. These are the only univalent vertices of Γ.

See Figure 1 for an example of such a graph.

The construction of [11] provides the space $V_r(\Sigma,\hat{c})$ with a Hermitian basis $(\varphi_c)_{c\in U_r}$ for any choice of a pair of pants decomposition $\mathcal{C}$ of Σ and trivalent graph Γ compatible with $\mathcal{C}$. The index set U_r of this basis is the set of r–admissible colorings of the edges of Γ, defined as follows:

Let $\mathcal{C}_r=\{1,\dots,r-1\}$ be the set of colors.

An r–admissible coloring of Γ is a map $c\colon E\to\mathcal{C}_r$ such that the following conditions are met:

(1) For any $i\in\{1,\dots,n\}$, the edge adjacent to p_i is colored by $c_i=\hat{c}_i$.

(2) Let S be the set of all triples (e,f,g) such that the curves C_e, C_f and C_g bound a pair of pants (possibly two of these curves are the same). Then for any $(e,f,g)\in S$ we have

  (i) $c_e+c_f+c_g<2r$ and $c_e+c_f+c_g\equiv 1 \pmod 2$;

  (ii) $c_e<c_f+c_g$.

If we have a sequence of coloring of the marked points $\hat{c}_i=rt_i$ with $t\in\mathbb{Q}^n$, then for $c_r\in U_r$ the E–tuple c_r/r is in the set $U\subset\mathbb{R}^E$ defined by $x\in U$ if and only if

(1) $x_i = t_i$ if i is the edge adjacent to the marked point p_i; and

(2) for any $(e, f, g) \in S$, we have
  (i) $x_e + x_f + x_g < 2$,
  (ii) $x_e < x_f + x_g$.

Let γ_i be small simple closed curves encircling the marked points p_i. We introduce the SU_2–moduli space of Σ with marked points (p_i, t_i), $t_i \in [0,1]$,

$$\mathcal{M}(\Sigma, t_1, \dots, t_n) = \{\rho\colon \pi_1(\Sigma) \to \mathrm{SU}_2 \mid \mathrm{Tr}(\rho(\gamma_i)) = 2\cos(\pi t_i)\}/\mathrm{SU}_2.$$

The quotient here corresponds to the conjugation of representations by an element of SU_2.

We recall that the subset of irreducible representations in $\mathcal{M}(\Sigma)$ has a natural Atiyah–Bott–Goldman–Seshadri symplectic form, which we call ω.

Any curve γ on Σ induces a natural *trace function* f_γ on $\mathcal{M}(\Sigma)$ by the formula

$$f_\gamma\colon \rho \to -\mathrm{Tr}(\rho(\gamma)).$$

Moreover for any pants decomposition $\mathcal{C}$ of Σ, Jeffrey and Weitsman [20] introduced a momentum map $h_{\mathcal{C}}$ on $\mathcal{M}(\Sigma)$ whose image is the closure of the set U introduced above. This momentum mapping is given by the formula

$$h_{\mathcal{C}}\colon \rho \to (h_{C_e}(\rho))_{e\in E} = \left(\frac{1}{\pi}\operatorname{Acos}\left(\frac{\mathrm{Tr}(\rho(C_e))}{2}\right)\right)_{e\in E}.$$

Here U is exactly the set of regular values of the momentum map $h_{\mathcal{C}}$. Jeffrey and Weitsman showed that the h_{C_e} are independent Poisson-commuting functions, and that these Hamiltonians induce an action of a torus T on each level set. Thus the momentum map induces action-angle coordinates on the subset $h_{\mathcal{C}}^{-1}(U)$ of $\mathcal{M}(\Sigma)$: there is a map

$$R\colon U \times T \to h_{\mathcal{C}}^{-1}(U), \quad (\tau, \theta) \mapsto R(\tau_e, \theta_e).$$

The map R satisfies that $h_{\mathcal{C}}(R(\tau,\theta)) = \tau$ and $R_*(\omega) = \sum_{e\in E} d\tau_e \wedge d\theta_e$. These action-angle coordinates are unique up to a shift in angle coordinates.

Marché and Paul [21] proved from skein calculus that in the case of the once-punctured torus and the case of the four-punctured sphere, the matrix coefficients of curve operators $\langle T_r^\gamma \varphi_c, \varphi_{c+k}\rangle$ converge to the k^{th} Fourier coefficient of the trace functions

$$\theta \mapsto f_\gamma\left(R\left(\frac{c}{r}, \theta\right)\right), \quad \theta \in T.$$

They also gave an expression for the $O(1/r)$ term in the expansion of $\langle T_r^\gamma \varphi_c, \varphi_{c+k}\rangle$.

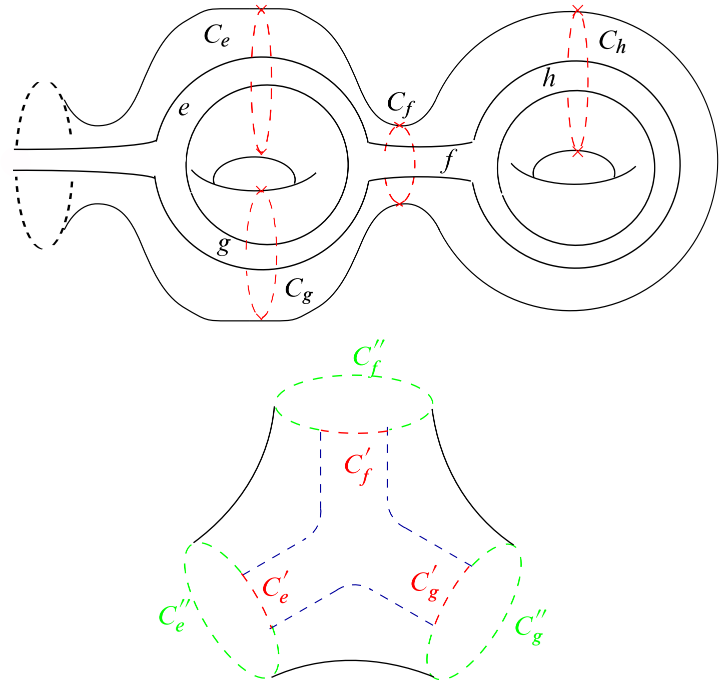


Figure 1: A banded graph compatible with a pants decomposition of Σ by curves $\{C_e\}$ and the associated cell decomposition of a pants into hexagons

Our paper aims to give a generalization of the asymptotic expansion in [21] for any marked surface Σ. We observed a new phenomenon when studying general surfaces: the asymptotic coefficients are again related to Fourier coefficients of trace functions, but they are twisted by rapidly oscillating signs.

To give an expression for these signs, we introduce some cocycles on Σ.

Equip Σ with a pants decomposition $\mathcal{C}$ and a compatible graph Γ. As we can see in the example in Figure 1, $\Sigma \setminus \Gamma$ is a trivalent banded graph diffeomorphic to Γ, so we get a continuous folding map $p\colon \Sigma \to \Gamma$ that pastes the two copies of Γ.

For any r–admissible color c we can define a multicurve L_c inside Γ: take $c_e - 1$ parallel strands at any edge e and connect at vertices in the unique way avoiding crossings.

We define a cocycle $\overline{c}$ in $H^1(\Sigma, \mathbb{Z}/2)$ by the formula

$$\overline{c}(\gamma) = L_c \cap p(\gamma).$$

Here $\cap$ is the $\cap$–product map $H_1(\Gamma, \mathbb{Z}/2) \times H_1(\Gamma, \partial\Gamma, \mathbb{Z}/2) \to \mathbb{Z}/2$, and we view $p(\gamma)$ as an element of $H_1(\Gamma, \partial\Gamma, \mathbb{Z}/2)$.

Theorem 1.1 *Let γ be a multicurve in $\Sigma \setminus \{p_1, \dots, p_n\}$.*

For $e \in E$ we write I_e^γ for the geometric intersection number of γ with C_e.

We introduce an open set $V_\gamma \subset U \times [0,1]$ by the formula

$$V_\gamma = \big\{(\tau, \hbar) \mid (\tau_e + \varepsilon_e \hbar I_e^\gamma)_{e \in E} \in U \ \text{ for all } \varepsilon \in \{\pm 1\}^E\big\}.$$

Then

1. *Whenever $k_e > I_e^\gamma$ or $k_e \neq I_e^\gamma \pmod 2$, the matrix coefficient $\langle T_r^\gamma \varphi_c, \varphi_{c+k}\rangle$ vanishes.*
2. *If $k_e \leq I_e^\gamma$ and $k_e = I_e^\gamma \pmod 2$, there exists a smooth function $(F_k^\gamma)_{k \colon E \to \mathbb{Z}}$ defined on V_γ such that, for any $c \in U_r$, the matrix coefficient $\langle T_r^\gamma \varphi_c, \varphi_{c+k}\rangle$ is $\overline{c}(\gamma) F_k^\gamma(c/r, 1/r)$.*

If we set $F_k = 0$ for any other $k \colon E \to \mathbb{Z}$, we can write

$$T_r^\gamma \varphi_c = \overline{c}(\gamma) \sum_{k \colon E \to \mathbb{Z}} F_k^\gamma\Big(\frac{c}{r}, \frac{1}{r}\Big)\varphi_{c+k}.$$

As $\overline{c}$ is an element of $H^1(\Sigma, \mathbb{Z}/2)$, $\overline{c}(\gamma)$ is just a sign. This sign factor, which did not appear in [21], will be shown to be trivial when the banded trivalent graph Γ is planar (which was the case for the punctured torus and the four-holed sphere).

The coefficients F_k^γ can be computed by hand for any multicurve γ on Σ, but to give an explicit formula for a general γ is out of reach. However, we will provide a formula for the first two terms of the Taylor expansion of F_k^γ in the second variable.

In [21], to make sense of the coefficients of T_r^γ Marché and Paul introduce a complex-valued function σ^γ, which they called the ψ–symbol of T_r^γ. We follow their approach, but the signs in our formulae lead us to define the ψ–symbol as a function with values in some algebra A_Γ, which we call the intersection algebra. We define A_Γ as follows:

Let π be the map $H^1(\Gamma, \mathbb{Z}/2) \to H^1(\Gamma, \partial\Gamma, \mathbb{Z}/2)$ and B be its image. The folding map p and the map π induce a map $p_* \colon H^1(\Sigma, \mathbb{Z}/2) \to H^1(\Gamma, \partial\Gamma, \mathbb{Z}/2)$. We define

$$A_\Gamma = \bigoplus_{[\gamma] \in B} \mathbb{C}[\gamma]$$

with the product $[\gamma][\delta] = (-1)^{\gamma \cap \widetilde{\delta}}[\gamma + \delta]$, where $\pi(\widetilde{\delta}) = [\delta]$ and $\cap$ is the intersection form $H^1(\Gamma, \partial\Gamma, \mathbb{Z}/2) \times H^1(\Gamma, \mathbb{Z}/2) \to \mathbb{Z}/2$.

Definition 1.2 Let γ be a multicurve on Σ. We define the ψ–symbol of T_r^γ as the map

$$\sigma^\gamma\colon V_\gamma \times (\mathbb{R}/2\pi\mathbb{Z}) \to A_\Gamma$$

such that

$$\sigma^\gamma(\tau,\hbar,\theta) = \sum_{k\colon E\to\mathbb{Z}} F_k(\tau,\hbar)e^{ik\cdot\theta}[p_*(\gamma)].$$

If $\chi\colon A_\Gamma \to \mathbb{C}$ is a morphism of algebras, we also introduce $\sigma_\chi^\gamma(\tau,\theta) = \chi(\sigma^\gamma)(\tau,0,\theta)$.

Let us add a few remarks on this definition:

(1) $k\cdot\theta$ stands for $\sum_{e\in E} k_e\theta_e$.

(2) The sum over $k\colon E \to \mathbb{Z}$ is actually a finite sum, as only a finite number of coefficients F_k^γ does not vanish.

(3) We will often omit the p_* and just write $[\gamma]$ for the element $[p_*(\gamma)]$, when γ is a multicurve.

(4) We will often refer to the zeroth order in $\hbar$ of the ψ–symbol, that is, $\sigma^\gamma(\tau,0,\theta)$, as the principal symbol of T_r^γ.

We use this definition to state our main result:

Theorem 1.3 *Let γ be a multicurve on Σ. The ψ–symbol $\sigma^\gamma(\tau,\hbar,\theta)$ of the curve operator T_r^γ has the following asymptotic expansion:*

$$\sigma^\gamma(\tau,\hbar,\theta) = \sigma^\gamma(\tau,0,\theta) + \frac{\hbar}{2i}\sum_{e\in E}\frac{\partial^2}{\partial\tau_e\,\partial\theta_e}\sigma^\gamma(\tau,0,\theta) + o(\hbar)$$

and, for $\chi\colon A_\Gamma \to \mathbb{C}$ a morphism of algebras, we have $\sigma_\chi^\gamma(\tau,\theta) = f_\gamma(R_\chi(\tau,\theta)) = -\mathrm{Tr}(R_\chi(\tau,\theta)(\gamma))$, where the R_χ are action-angle parametrizations on

$$\mathcal{M}(\Sigma) = \mathrm{Hom}\big(\pi_1(\Sigma\setminus\{p_1,\dots,p_n\}), \mathrm{SU}_2\big)/\mathrm{SU}_2$$

defined up to a choice of origin of the angles.

The above theorem is quite similar to results obtained by Andersen and Gammelgaard [6] in the geometric framework of the Witten–Reshetikhin–Turaev TQFT.

Recall that, for any complex structure σ on Σ representing a point in the Teichmüller space $\mathcal{T}$ of Σ, the smooth part of the moduli space of Σ has the structure of a Kähler manifold M_σ. It is then possible to identify the TQFT vector spaces $V_r(\Sigma)$ with the space of holomorphic sections $H^0(M_\sigma, L^r)$, where L is the Chern–Simons vector bundle; see Andersen and Ueno [7; 8; 9; 10].

Theorem 7 of [6] shows that curve operators T_r^γ are approximated at order 1 by Toeplitz operators of principal symbol f_γ and subprincipal symbols

$$\tfrac{1}{4}\Delta_\sigma f_\gamma + i\nabla_{X_F''} f_\gamma,$$

where X_F'' is the $(0,1)$–part of the Hamiltonian vector field for the Ricci potential.

An alternative proof of Theorem 1.3 could be to combine the results of [6] with results explaining how these Laplace operators degenerate when the complex structure on Σ converges to the pair of pants decomposition. See Andersen [5] for an outline of such techniques.

The methods in [6] rely on the geometric framework of TQFT or the Hitchin connection so they are quite different from ours, which is based on skein theory and is the continuation of the work of Marché and Paul [21].

The proof of [21] in the case where Σ is the punctured torus and the four-holed sphere relied on explicit computations for some simple set of curves that generates the Kauffman algebra of Σ, then extending the result to general curves. This approach failed in higher genus as no simple set of generators is known. Instead, we developed a more conceptual and systematic method, which relies on the study of algebraic properties of the ψ–symbol and the Kauffman algebra of Σ.

Marché and Paul [21] used the asymptotic estimation to construct a framework for curve operators on the punctured torus and the four-holed sphere as Toeplitz operators on the sphere. This allowed the application of the WKB-approximation for eigenvectors. From this they deduced asymptotic expansions of quantum invariants (such as a new proof of the asymptotic expansion of $6j$–symbols, and an expression for the punctured S–matrix). Therefore, we hope to use our asymptotic expansions for general marked surface to make a connection to the framework of curve operators as Toeplitz operators on toric varieties, or at least apply the tools of microlocal analysis. Such a Toeplitz framework for curve operators may be a useful tool to study combinatorial TQFT. Indeed, in a different approach, Andersen [1] introduced some geometrical curve operators that are Toeplitz operators to prove the asymptotic fidelity of the quantum representations of the mapping class group. We think that the idea, initiated by Andersen, of viewing the standard curve operators as Toeplitz operators is a powerful idea, as has been demonstrated in various work of his [2; 3; 4]. We believe that our result and methods, based on the BHMV approach to TQFT, could provide interesting applications in other directions.

Acknowledgements I am very thankful to my advisor Julien Marché for his guidance and advice. I also would like to thank Gregor Masbaum for pointing out an error about intersection forms.

2 A quick overview of TQFT and curve operators

In this section we will outline the BHMV approach to TQFT. Their construction relies on the notion of Kauffman bracket skein modules of 3–manifolds and Kauffman algebras of marked surfaces.

For M a compact oriented 3–manifold (which can have a boundary), we define $K(M,A)$ as the quotient of the free $\mathbb{C}[A^{\pm 1}]$–module generated by links modulo isotopy and the Kauffman relations (see Figure 2).

For $t\in\mathbb{C}^*$, we can define a Kauffman module evaluated at t: we write $K(M,t)=K(M,A)\otimes_{A=t}\mathbb{C}$.

Now, if Σ is a surface with marked points $p_1,\dots,p_n$, we denote by $K(\Sigma,A)$ the Kauffman module $K\big((\Sigma\setminus\{p_1,\dots,p_n\})\times[0,1],A\big)$.

We call a disjoint union of simple curves on Σ which is disjoint from the marked points of Σ a *multicurve* on Σ. It is easy to see that $K(\Sigma,A)$ is spanned by multicurves on Σ, and actually multicurves give a basis of this vector space, as shown in [14].

The module $K(\Sigma,A)$ has an algebra structure: the product $\gamma\cdot\delta$ of two elements of $K(\Sigma,A)$ is obtained by isotoping γ and δ so they are included in $\Sigma\times\left(\frac{1}{2};1\right]$ and $\Sigma\times\left[0;\frac{1}{2}\right)$, respectively, then gluing the two parts into $\Sigma\times[0,1]$.

For $t\in\mathbb{C}^*$, we define $K(\Sigma,t)=K(\Sigma,A)\otimes_{A=t}\mathbb{C}$, which is also an algebra, and admits the set of multicurves as a basis. Using this basis, we get a linear isomorphism between $K(\Sigma,t)$ and $K(\Sigma,-1)$ and we embed $K(\Sigma,-e^{i\pi\hbar/2})=K(\Sigma,A)\otimes_{A=-e^{i\pi\hbar/2}}\mathbb{C}[\![\hbar]\!]$ into $K(\Sigma,-1)[\![\hbar]\!]$.

The vector spaces $V_r(\Sigma,\hat{c})$ are quotients of Kauffman modules at roots of unity, as explained below:

Definition [11] Let H be a handlebody with $\partial H=\Sigma$, where Σ is a surface with marked points $p_1,\dots,p_n$.

Given a coloration $\hat{c}$ of the marked points, we choose c_i-1 points in a small neighborhood of p_i for each i, and write P for the set of all resulting points for i from 1 to n.

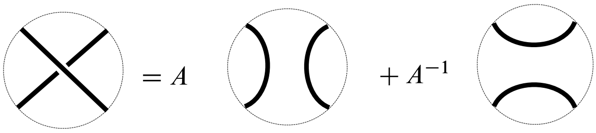


Figure 2: The first Kauffman relation. The other relation states that any trivial component is identified with $-A^2-A^{-2}$.

We define the relative Kauffman module $K(H,\hat{c},\zeta_r)$ as the $\mathbb{C}[A^{\pm 1}]$–module generated by banded tangles in H whose intersection with Σ is the set P.

For r a positive integer, we write $\zeta_r = -e^{i\pi/(2r)}$. For any embedding j of H in $\boldsymbol{S}^3$, we define the following submodule of $K(H,\hat{c},\zeta_r)$:

$$N_r^j = \left\{ x \in K(H,\hat{c},\zeta_r) \;\middle|\; \left\langle x \middle| \bigotimes_{i=1}^{r} f_{c_i-1} \middle| y \right\rangle = 0 \text{ for all } y \in K(\boldsymbol{S}^3 \setminus \mathrm{Im}(j),\hat{c},\zeta_r) \right\},$$

where we write f_k for the k^{th} Jones–Wenzl idempotent, and $\langle x|\bigotimes_{i=1}^{r} f_{c_i-1}|y\rangle$ stands for the element of $K(\boldsymbol{S}^3,\zeta_r)$ obtained from x and y by pasting H with $\boldsymbol{S}^3\setminus\mathrm{Im}(j)$, inserting the Jones–Wenzl idempotent at each marked point.

Theorem 2.1 [11] *N_r^j is in fact independent of j and of finite codimension, and we may define*

$$V_r(\Sigma,\hat{c}) = K(H,\hat{c},\zeta_r)/N_r^j.$$

With this setting, there is a simple description of the curve operator T_r^γ associated to a multicurve γ on Σ disjoint from the marked points $p_1,\dots,p_n$, or more generally to an element of $K(\Sigma,\zeta_r)$.

Indeed, we can take an element z of $K(H,\hat{c},\zeta_r)$ and stack a multicurve γ over it to obtain another element $\gamma\cdot z$ of $K(H,\hat{c},\zeta_r)$. The induced map factors through N_r^j, since for any $n\in N_r^j$ and any $z\in K(\boldsymbol{S}^3\setminus\mathrm{Im}(j),\hat{c},\zeta_r)$, we have that $\langle\gamma\cdot n|\bigotimes_{i=1}^{r} f_{c_i-1}|z\rangle = \langle n|\bigotimes_{i=1}^{r} f_{c_i-1}|\gamma\cdot z\rangle$. Thus we have defined an endomorphism T_r^γ of $V_r(\Sigma,\hat{c})$ associated to $\gamma\in K(\Sigma,\zeta_r)$.

Furthermore, the map

$$T_r^{\cdot}\colon K(\Sigma,\zeta_r)\to \mathrm{End}(V_r(\Sigma,\hat{c})), \quad \gamma\mapsto T_r^\gamma,$$

is a morphism of algebras.

In [11] it is shown that the bracket $\langle\cdot,\cdot\rangle$ that we introduced above induces a Hermitian structure on $V_r(\Sigma,\hat{c})$.

The construction of [11] provides for each admissible coloring c a vector $\varphi_c\in V_r(\Sigma,\hat{c})$. This vector is obtained by cabling the graph Γ by a specific combination of multicurves (we will detail this construction in Section 4). Moreover, the family (φ_c) when c runs over all admissible colorings is a Hermitian basis of $V_r(\Sigma,\hat{c})$.

For a multicurve γ, the operators T_r^γ are Hermitian operators for the Hermitian structure on $V_r(\Sigma,\hat{c})$ given by [11]. The spectrum and the eigenvectors of T_r^γ are known:

First, as all components of γ are disjoint, there exists a pants decomposition of Σ by a family of curves $\mathcal{C}=\{C_e\}_{e\in E}$ such that γ can be isotoped to the union of n_e parallel copies of C_e, for some integers $n_e\in\mathbb{N}$. Then the Hermitian basis (φ_c) coming from the pants decomposition $\mathcal{C}$ is an eigenbasis of T_r^γ, and we have

$$T_r^\gamma\varphi_c=\biggl(\prod_{e\in E}\Bigl(-2\cos\frac{\pi c_e}{r}\Bigr)^{n_e}\biggr)\varphi_c.$$

We should take note that the spectral radius $\|T_r^\gamma\|$ is thus always less than $2^{n(\gamma)}$, where we write $n(\gamma)$ for the number of components of the multicurve γ.

Let

$$\mathcal{M}'(\Sigma)=\mathrm{Hom}(\pi_1(\Sigma),\mathrm{SL}_2(\mathbb{C}))/\!/\mathrm{SL}_2(\mathbb{C})$$

be the space of characters of the fundamental group of $\Sigma\setminus\{p_1,\dots,p_n\}$ in $\mathrm{SL}_2(\mathbb{C})$. This space is actually an affine algebraic variety.

Also let $\mathrm{Reg}(\mathcal{M}'(\Sigma))$ be the algebra of regular functions from $\mathcal{M}'(\Sigma)$ to $\mathbb{C}$.

The following theorem, which describes the Kauffman algebra $K(\Sigma,-1)$, will have a central role in the proof of Theorem 1.3:

Theorem 2.2 *The map*

$$\sigma\colon K(\Sigma,-1)\to\mathrm{Reg}(\mathcal{M}'(\Sigma)),\qquad \gamma\mapsto f_\gamma\quad\textit{such that }f_\gamma(\rho)=-\mathrm{Tr}(\rho(\gamma)),$$

is an isomorphism of algebras.

This theorem follows from the work of various authors. Bullock [13] and Brumfiel and Hilden [12] first independently proved that the map from $K(\Sigma,-1)$ to $\mathcal{M}'(\Sigma)$ is surjective and has the nilradical of $K(\Sigma,-1)$ as kernel. It was proved later by Przytycki and Sikora [23] and independently by Charles and Marché [14] that the algebras $K(\Sigma,-1)$ are indeed reduced, which concluded the proof of Theorem 2.2.

Finally, we end this preliminary section with a formula for products of elements of the Kauffman algebra at $-e^{i\pi\hbar/2}$ to first order in $\hbar$. We recall that $\mathcal{M}'(\Sigma)$ is a Poisson manifold for the Poisson structure given in [16]. This Poisson structure depends on a choice of normalization of the symplectic structure on $\mathcal{M}(\Sigma)$. We normalize the symplectic form ω as the symplectic reduction of the form $\omega(\alpha,\beta)=(1/2\pi)\int_\Sigma\mathrm{Tr}(\alpha\wedge\beta)$ for $\alpha,\ \beta\in\Omega^1(\Sigma,\mathrm{su}_2)$. Since, by the previous theorem, it is possible to link the product of elements of $K(\Sigma,-1)$ with products of trace functions on $\mathcal{M}'(\Sigma)$, the work of Goldman [18] and Turaev [26] gives a way to think of the first order in $\hbar$ of a product of elements in $K(\Sigma,-e^{i\pi\hbar/2})$ as a Poisson bracket of trace functions.

Notice that from the fact that Kauffman algebras have the set of multicurves as a basis, as linear spaces $K(\Sigma,-e^{i\pi\hbar/2})$ is isomorphic to $K(\Sigma,-1)[\![\hbar]\!]$. This last space is isomorphic to a subspace of $\mathrm{Reg}(\mathcal{M}'(\Sigma))[\![\hbar]\!]$ via the map σ of Theorem 2.2.

Theorem 2.3 [26] *Let γ and δ be multicurves, viewed as elements of $K(\Sigma,-e^{i\pi\hbar/2})$. We have that*

$$\gamma\cdot\delta = f_\gamma f_\delta + \frac{\hbar}{i}\{f_\gamma, f_\delta\} + o(\hbar).$$

This result is due to the work of Goldman and Turaev. First Goldman [18] was able to compute the Poisson bracket of the trace functions of two simple closed curves as the sum of other trace functions. Then Turaev [26] was able to identify the terms in Goldman formula for the Poisson bracket with the order 1 terms of the product in the Kauffman algebra.

3 Algebraic properties of ψ–symbols

3.1 Some remarks on the intersection algebra

In this section, we fix a surface Σ with marked points $p_1,\dots,p_n$, with a pants decomposition $\mathcal{C}=\{C_e\}_{e\in E}$ of Σ and a compatible trivalent banded graph Γ drawn on Σ.

We see from Figure 1 that $\mathcal{C}$ and Γ give us a cell decomposition of Σ into a bunch of hexagons, their sides being the boundary components of Γ and segments of the curves C_e. For each $e\in E$, we name by C_e' (resp. C_e'') the segment $\Gamma\cap C_e$ (resp. $C_e\setminus\mathrm{Int}(C_e\cap\Gamma)$); see Figure 1.

We remark that the cocycle $\overline{c}$ of $H^1(\Sigma,\mathbb{Z}/2)$ can then be computed as

$$\overline{c}(\gamma)=\prod_{e\in E}(-1)^{(c_e-1)(C_e'^{*}(\gamma)+C_e''^{*}(\gamma))}.$$

In this formula, $C_e'^{*}$ (resp. $C_e''^{*}$) is the cellular cochain dual to ${C_e}'$ (resp. ${C_e}''$). We can directly check from the formula that $\overline{c}$ is a cocycle, as its value on the boundary of each hexagon is of the form $(-1)^{c_e+c_f+c_g-1}$ for e, f and g three adjacent edges, which equals 1 as c is an admissible color. Also it is easy to see that the formula gives exactly the intersection number $L_c\cap p_*(\gamma)$.

Now, for α and β in B, the image of $\pi\colon H_1(\Gamma,\mathbb{Z}/2)\to H_1(\Gamma,\partial\Gamma,\mathbb{Z}/2)$, we write $\langle\alpha,\beta\rangle=\widetilde{\alpha}\cap\beta$, where $\pi(\widetilde{\alpha})=\alpha$. Recall that we defined the *intersection algebra* A_Γ as

$$A_\Gamma=\bigoplus_{\alpha\in B}\mathbb{C}\cdot[\alpha],$$

with the product structure given by $[\gamma]\cdot[\delta]=(-1)^{\langle\gamma,\delta\rangle}[\gamma+\delta]$. It is not clear at this point that A_Γ is an algebra, and not even that it is well defined. This comes from the following lemma:

Lemma 3.1 *The form*

$$\langle\,,\rangle\colon B\times B\to\mathbb{Z}/2$$

given by $\langle\alpha,\beta\rangle=\tilde{\alpha}\cap\beta$ does not depend on the choice of a lift $\pi(\tilde{\alpha})=\alpha$ and is symmetric and bilinear.

Proof Indeed, two lifts of α differ by an element of $H_1(\partial\Gamma,\mathbb{Z}/2)$. Furthermore, any element γ of $H_1(\Gamma,\partial\Gamma,\mathbb{Z}/2)$ can be seen as a linear combination of closed curves and curves with extremities in $\partial\Gamma$, and $\gamma\in B$ if and only if its number of extremities in each component of $\partial\Gamma$ is even. Thus the intersection of an element of $H_1(\partial\Gamma,\mathbb{Z}/2)$ with any element of B vanishes, and the form $\langle\,,\rangle$ is independent of the choice of lift.

Actually, this shows that we can think of B as the quotient of $H_1(\Gamma,\mathbb{Z}/2)$ by the kernel of the intersection form on $H_1(\Gamma,\mathbb{Z}/2)$ and $\langle\,,\rangle$ as the corresponding quotient form.

The bilinearity of the form $\langle\,,\rangle$ is then evident.

Finally we show that the form is symmetric. Given lifts $\tilde{\alpha}$ and $\tilde{\beta}$ to $H_1(\Gamma,\mathbb{Z}/2)$ of two elements α and β in B, $\langle\alpha,\beta\rangle=\tilde{\alpha}\cap\beta$ is also the intersection number mod 2 of $\tilde{\alpha}$ and $\tilde{\beta}$, so it is symmetric. □

From the lemma we get that the product on A_Γ is well-defined, associative and commutative, so A_Γ is a commutative $\mathbb{C}$–algebra of dimension 2^d, where d is the dimension of B. This dimension can be computed using the exact sequence

$$H_1(\Gamma,\mathbb{Z}/2)\to H_1(\Gamma,\partial\Gamma\mathbb{Z}/2)\xrightarrow{\delta}H_0(\partial\Gamma,\mathbb{Z}/2)\to H_0(\Gamma,\mathbb{Z}/2)\to 0.$$

We have $B=\operatorname{Ker}\delta$ and $\dim(\operatorname{Ker}\delta)+\operatorname{rk}(\delta)=g$, where g is the genus of Γ, and $\operatorname{rk}(\delta)=b-1$, where b is the number of boundary components of Γ. Thus the dimension of B is $g-b+1$.

Note that when Γ can be embedded in the plane this dimension is 0 and $A_\Gamma=\mathbb{C}$.

As a finite-dimensional commutative $\mathbb{C}$–algebra, A_Γ is isomorphic to the algebra $\mathbb{C}^l$, where $l=\dim(A_\Gamma)=\operatorname{Card}(\widehat{A}_\Gamma)$ and we recall that $\widehat{A}_\Gamma$ is the (finite) set of algebra morphisms from A_Γ to $\mathbb{C}$. The isomorphism is given by

$$\alpha\mapsto(\chi(\alpha))_{\chi\in\widehat{A}_\Gamma}\quad\text{for }\alpha\in A_\Gamma.$$

An element χ of $\widehat{A}_\gamma$ must send each $[\alpha]$ with $\alpha \in B$ to some $(-1)^{q(\alpha)}$, with the conditions that $q(\alpha+\beta)-q(\alpha)-q(\beta) = \langle \alpha, \beta\rangle \pmod 2$. Thus $\widehat{A}_\Gamma$ is in bijective correspondence with the set of "relative spin-structures" on $(\Gamma, \partial\Gamma)$.

We end this section with the following lemma, providing a computation of products in A_Γ based on the cellular decomposition on Σ into hexagons:

Lemma 3.2 *Let γ and δ be two simple closed curves on Σ, and set*

$$i(\gamma,\delta) = \prod_{e\in E} (-1)^{I_e^\delta(C_e'^*(\gamma)+C_e''^*(\gamma))}.$$

Then $i(\gamma,\delta) = \langle p_*(\gamma), p_*(\delta)\rangle$.

Proof Let γ and δ be two curves on Σ. After an isotopy of $p(\gamma)$ and $p(\delta)$ in Γ we can arrange that $p(\delta)$ lies in the interior of Γ, and $p(\gamma)$ follows the edges of the cell decomposition of Γ. Then the intersection points lie only in the curves $p(C_e) = L_e$. The number of intersection points of $p(\gamma)$ and $p(\delta)$ in L_e is congruent modulo 2 to $\sharp(p(\delta)\cap L_e)L_e^*(p_*(\gamma))$, where L_e^* is the dual to the cell L_e.

But $L_e^*(p_*(\gamma)) = C_e'^*(\gamma)+C_e''^*(\gamma)$ and $\sharp(p(\delta)\cap L_e) = \sharp(\delta\cap C_e) \pmod 2$, hence the formula for $i(\widetilde{\gamma},\widetilde{\delta})$ computes the number of intersection points of γ and δ modulo 2, that is, $\langle p_*(\gamma), p_*(\delta)\rangle$. □

3.2 The multiplicativity property

In this section, we will temporarily assume that Theorem 1.1 holds. We can then define ψ–symbols, and we will show here that these ψ–symbol have a property of compatibility with the product in Kauffman modules. From this algebraic property alone and the theorem of Bullock, the ψ–symbols are almost constrained to have the form predicted by Theorem 1.3. Theorem 1.1 will be proved in Section 4.1 without using any of the results in this section.

For a fixed $(\tau,\hbar,\theta)$, the definition of the ψ–symbol only introduces $\gamma \mapsto \sigma^\gamma(\tau,\hbar,\theta)$ as a map from multicurves to A_Γ. We extend it by multilinearity to obtain a map

$$\sigma(\tau,\hbar,\theta)\colon K(\Sigma, -e^{i\pi\hbar/2}) \to A_\Gamma[\![\hbar]\!],$$

as $K(\Sigma,-e^{i\pi\hbar/2})$ is spanned by multicurves.

The proof of Theorem 1.3, giving an asymptotic formula for the ψ–symbol, will be the goal of Sections 5 and 6. It will rely heavily on the following property of the ψ–symbol, which explains its compatibility with the product in $K(\Sigma,-e^{i\pi\hbar/2})$:

Proposition 3.3 *Let γ and δ be two multicurves on Σ. Then we have the asymptotic expression*

$$\sigma^{\gamma\cdot\delta}(\tau,\hbar,\theta)=\left(\sigma^{\gamma}(\tau,\hbar,\theta)\sigma^{\delta}(\tau,\hbar,\theta)+\frac{\hbar}{i}\sum_{e}\partial_{\tau_e}\sigma^{\gamma}(\tau,\hbar\theta)\,\partial_{\theta_e}\sigma^{\delta}(\tau,\hbar,\theta)\right)+o(\hbar).$$

This expression is similar to the composition of symbols of Toeplitz operators. This is not a surprise, as curve operators can be approximated at order 1 by Toeplitz operators, by [6]. Theorem 8 of [6] gives the order 1 of the symbols of the composition of two such operators. It could again be possible to derive this result by degenerating the complex structure to a pair of pants decomposition.

A version of this proposition appeared already in [21] for the four-holed sphere and the pointed torus, but they worked with another definition of the ψ–symbol, which took values in $\mathbb{C}$, whereas in our definition, the ψ–symbol takes values in A_Γ.

We can however extract $\mathbb{C}$–valued functions from the ψ–symbol. As A_Γ is isomorphic to $\mathbb{C}^l$, we denote the components of the principal symbol $\sigma^\gamma(\tau,0,\theta)$ by $\sigma^\gamma_\chi(\tau,\theta)=\chi(\sigma^\gamma(\tau,0,\theta))$ for every $\chi\in\widehat{A}_\Gamma$.

Proof of Proposition 3.3 We fix $r>0$ and we take two multicurves γ and δ on Σ. The two functions appearing in the equality are smooth functions on a neighborhood of $U\times\{0\}$ in $U\times[0,1]$. We remark that any point of U can be approximated by a sequence c_r/r with $c_r\in U_r$. Hence it suffice to show that they have the same asymptotic expansion at order 1 on sequences $(c_r/r,\theta,1/r)$ where $c_r/r\to x\in U$. According to Theorem 1.1, writing $\tau=c_r/r$ and $\hbar=1/r$, the matrix coefficients of the operator T_r^γ can be written as

$$T_r^\gamma\varphi_c=\overline{c}(\gamma)\sum_{k\colon E\to\mathbb{Z}}F_k^\gamma(\tau,\hbar)\varphi_{c+k},$$

with the F_k^γ being smooth functions on V_γ such that $F_k^\gamma=0$ as soon as there is some $e\in E$ such that $|k_e|>I_e^\gamma$ or $k_e\not\equiv I_e^\gamma \pmod 2$.

As $\gamma\in K(\Sigma,-e^{i\pi/(2r)})\to T_r^\gamma\in\mathrm{End}(V_r(\Sigma))$ is an morphism of algebras, we have

$$T_r^{\gamma\cdot\delta}\varphi_c=T_r^\gamma(T_r^\delta\varphi_c)$$

and, from the above expression of the matrix coefficients, we get

$$\begin{aligned}T_r^{\gamma\cdot\delta}\varphi_c&=\sum_{m\colon E\to\mathbb{Z}}\left(\sum_{k+l=m}F_l^\gamma(\tau+k\hbar,\hbar)F_k^\delta(\tau,\hbar)\overline{c}(\delta)\overline{c+k}(\gamma)\right)\varphi_{c+m}\\&=\overline{c}(\gamma)\overline{c}(\delta)i(\gamma,\delta)\sum_{m\colon E\to\mathbb{Z}}\left(\sum_{k+l=m}F_l^\gamma(\tau+k\hbar,\hbar)F_k^\delta(\tau,\hbar)\right)\varphi_{c+m}.\end{aligned}$$

To obtain the second equality, note that $\overline{c+k}(\gamma)=\overline{c}(\gamma)\overline{k}(\gamma)$ and observe that if there exists e such that $k_e\neq I_e^\delta \pmod 2$ then, by Theorem 1.1, F_k^δ is 0.

However, if $k_e=I_e^\delta \pmod 2$ for all $e\in E$ then $\overline{k}(\gamma)=\prod_{e\in E}(-1)^{I_e^\delta(C_e'^*(\gamma)+C_e''^*(\gamma))}=i(\gamma,\delta)$ is independent of k. Hence we can factor $\overline{k}(\gamma)$ out of the sum.

Now, as $K(\Sigma,-e^{i\pi\hbar/2})$ is generated by multicurves, we can write $\gamma\cdot\delta=\sum_\lambda f_\lambda(\hbar)\lambda$, and, in this sum, $f_\lambda\neq 0$ only when $[\lambda]=[\gamma]+[\delta]\in H_1(\Sigma,\mathbb{Z}/2)$, according to the Kauffman relations. Thus we have $\overline{c}(\lambda)=\overline{c}(\gamma)\overline{c}(\delta)$. We can write another formula for the curve operator of the product:

$$T_r^{\gamma\cdot\delta}\varphi_c=\sum_m\biggl(\sum_\lambda \overline{c}(\lambda)f_\lambda(\hbar)F_m^\lambda(\tau,\hbar)\biggr)\varphi_{c+m}.$$

So, identifying coefficients in the two formulae, we get

$$\sum_\lambda f_\lambda(\hbar)F_m^\lambda(\tau,\hbar)=\biggl(\sum_{k+l=m}F_l^\gamma(\tau+k\hbar,\hbar)F_k^\delta(\tau,\hbar)\biggr)i(\gamma,\delta).$$

Now, recall that we defined the ψ–symbol of an arbitrary element of $K(\Sigma,-e^{i\pi\hbar/2})$ by extending linearly the formula for multicurves. Thus, we have

$$\sigma^{\gamma\cdot\delta}(\tau,\hbar,\theta)=\sum_m\sum_\lambda f_\lambda(\hbar)F_m^\lambda(\tau,\hbar)e^{im\theta}[\lambda],$$

recalling that $[\lambda]=[\gamma]+[\delta]$ and using the previous identity of coefficients

$$\sigma^{\gamma\cdot\delta}(\tau,\hbar,\theta)=i(\gamma,\delta)\sum_m\biggl(\sum_{k+l=m}F_l^\gamma(\tau+k\hbar,\hbar)F_k^\delta(\tau,\hbar)\biggr)e^{im\theta}[\gamma+\delta].$$

Now the Taylor expansion at order 1 in $\hbar$ of F_l^γ near $(\tau,\hbar)$ in the first variable gives

$$\begin{aligned}F_l^\gamma(\tau+k\hbar,\hbar)&=F_l^\gamma(\tau,\hbar)+\hbar\sum_{e\in E}k_e\frac{\partial}{\partial\tau_e}F_l^\gamma(\tau,\hbar)+o(\hbar)\\&=F_l^\gamma(\tau,\hbar)+\hbar\sum_{e\in E}k_e\frac{\partial}{\partial\tau_e}F_l^\gamma(\tau,0)+o(\hbar).\end{aligned}$$

Substituting into the previous equation gives us that

$$\begin{aligned}\sigma^{\gamma\cdot\delta}(\tau,\hbar,\theta)=i(\gamma,\delta)\sum_m\biggl(\sum_{k+l=m}&\biggl(F_l^\gamma(\tau,\hbar)+\hbar\sum_{e\in E}k_e\frac{\partial}{\partial\tau_e}F_l^\gamma(\tau,\hbar)\biggr)\\&\times e^{il\theta}F_k^\delta(\tau,\hbar)e^{ik\theta}\biggr)[\gamma+\delta]+o(\hbar)\end{aligned}$$

$$= i(\gamma,\delta)\langle p_*(\gamma), p_*(\delta)\rangle\Big(\sigma^\gamma(\tau,\hbar,\theta)\sigma^\delta(\tau,\hbar,\theta) \\ + \frac{\hbar}{i}\sum_{e\in E}\partial_{\tau_e}\sigma^\gamma(\tau,\hbar,\theta)\,\partial_{\theta_e}\sigma^\delta(\tau,\hbar,\theta)\Big)+o(\hbar).$$

To obtain the second equality recall that $[\gamma][\delta] = \langle p_*(\gamma), p_*(\delta)\rangle[\gamma+\delta]$ in A_Γ. From Lemma 3.2 we have that $i(\gamma,\delta) = \langle p_*(\gamma), p_*(\delta)\rangle$, which completes the proof. □

According to this proposition, the principal symbol $\sigma^{\cdot}(\tau,0,\theta)\colon K(\Sigma,-1)\to A_\Gamma$ is a morphism of algebras. Furthermore, the components $\sigma_\chi(\tau,\theta) = \chi(\sigma(\tau,0,\theta)$ are algebra morphisms from $K(\Sigma,-1)$ to $\mathbb{C}$.

Using the theorem of Bullock, we will show in Section 5.1 that these morphisms have the form $f \mapsto f(R_\chi)$, $f \in \mathrm{Reg}(\mathcal{M}'(\Sigma))$, for some representations R_χ of $\pi_1(\Sigma\setminus\{p_1,\dots,p_n\})$.

Identifying precisely the representations R_χ will come from checking the special values of the ψ–symbol on the curves C_e.

As for the computation of the first-order term, we will proceed in Section 6 in a similar fashion: first we will show, using only Proposition 3.3, that this term is related to derivations of algebras $K(\Sigma,-1)\to A$, then, by studying the values of the ψ–symbol on the curves C_e and on another family of curves D_e, we will show the first-order term is indeed given by the formula in Theorem 1.3.

4 Computations of curve operators using fusion rules

This section is devoted to the skein theory computations that will be needed in order to prove Theorem 1.1. We describe the general form of the matrix coefficients of the curve operators, and give examples of explicit computations of the coefficients F_k^γ and the ψ–symbol σ^γ for some curves γ.

4.1 Fusion rules in a pants decomposition

In this subsection, we will work with a fixed closed oriented surface Σ, along with a pants decomposition by a family of curves $\mathcal{C} = \{C_e\}_{e\in E}$. We can consider $n_e \geq 1$ parallel copies $(C_e^k)_{1\leq k\leq n_e}$ of the curves C_e such that the curves C_e^k cut the surface Σ into a collection of pants $\{P_s\}_{s\in S}$ and annuli $\{A_e^k \mid e\in E,\ 1\leq k\leq n_e-1\}$.

We recall that to this pants decomposition is associated a Hermitian basis φ_c of $V_r(\Sigma)$, of which we will recall the construction:

Let Γ be a banded trivalent graph compatible with the pants decomposition $\mathcal{C}$ of Σ as in Section 2. We recall that Γ is viewed as drawn on Σ. Given an admissible coloring $c\colon E \to C_r$, we define $\psi_c \in K(\Sigma; \hat{c}; \zeta_r)$ as follows:

- Replace each edge e of Γ by $c_e - 1$ parallel copies of e lying on Σ.
- Insert in the middle of each edge the idempotent f_{c_e-1}, where we recall that f_k is the k^{th} Jones–Wenzl idempotent.
- In the neighborhood of each trivalent vertex, join the three sets of lines in Σ in the unique possible way avoiding crossings.

This family of vectors is actually an orthogonal basis of $V_r(\Sigma, c)$ for a natural Hermitian structure defined in [11], which we do not recall here. We refer to [11, Theorem 4.11] for the proof and the formula

$$\|\psi_c\|^2 = \left(\frac{2}{r}\right)^{\chi(\Gamma)/2} \frac{\prod_P \langle c_P^1, c_P^2, c_P^3\rangle}{\prod_e \langle c_e\rangle}. \tag{1}$$

Here the first product is over all vertices P corresponding to pants of the pants decomposition, the second over the edges e of the graph Γ. We write $\langle n\rangle$ for $\sin(\pi n/r)$; $\langle n\rangle!$ for $\prod_{i=1}^n \langle i\rangle$; c_P^1, c_P^2, and c_P^3 for the colors of the 3 edges adjacent to P; and we also set

$$\langle a,b,c\rangle = \frac{\left\langle\frac{a+b+c-1}{2}\right\rangle!\left\langle\frac{a+b-c-1}{2}\right\rangle!\left\langle\frac{a-b+c-1}{2}\right\rangle!\left\langle\frac{b+c-a-1}{2}\right\rangle!}{\langle a-1\rangle!\,\langle b-1\rangle!\,\langle c-1\rangle!}.$$

As we will work with TQFT vectors locally, inside a pants of the pants decomposition for example, we will need to give a local version of this norm. Notice that if we forget the global factor $(2/r)^{\chi(\Gamma)/2}$ in the norm, we will not change the matrix coefficients of the curve operators T_r^γ.

Also, after applying fusion rules, we may get trivalent graphs with vertices other than those in the graph associated to the decomposition. We say then that a vertex is internal if it is trivalent or univalent and associated to a marked point, and that it is external otherwise. Then, we will define the square of the norm of a trivalent graph as

$$\frac{\prod_P \langle c_P^1, c_P^2, c_P^3\rangle}{\prod_{e\in E_2}\langle c_e\rangle \prod_{e\in E_1}\langle c_e\rangle^{1/2}},$$

where the products in the denominator are over E_2, the set of edges adjacent to 2 internal vertices, and E_1 the set of edges adjacent to 1 internal vertex and 1 external vertex. The other edges bear no contribution to the norm. With this definition, if we paste pieces of colored graph to get the graph Γ, we obtain the previous norm as the product of the norm of the pieces.

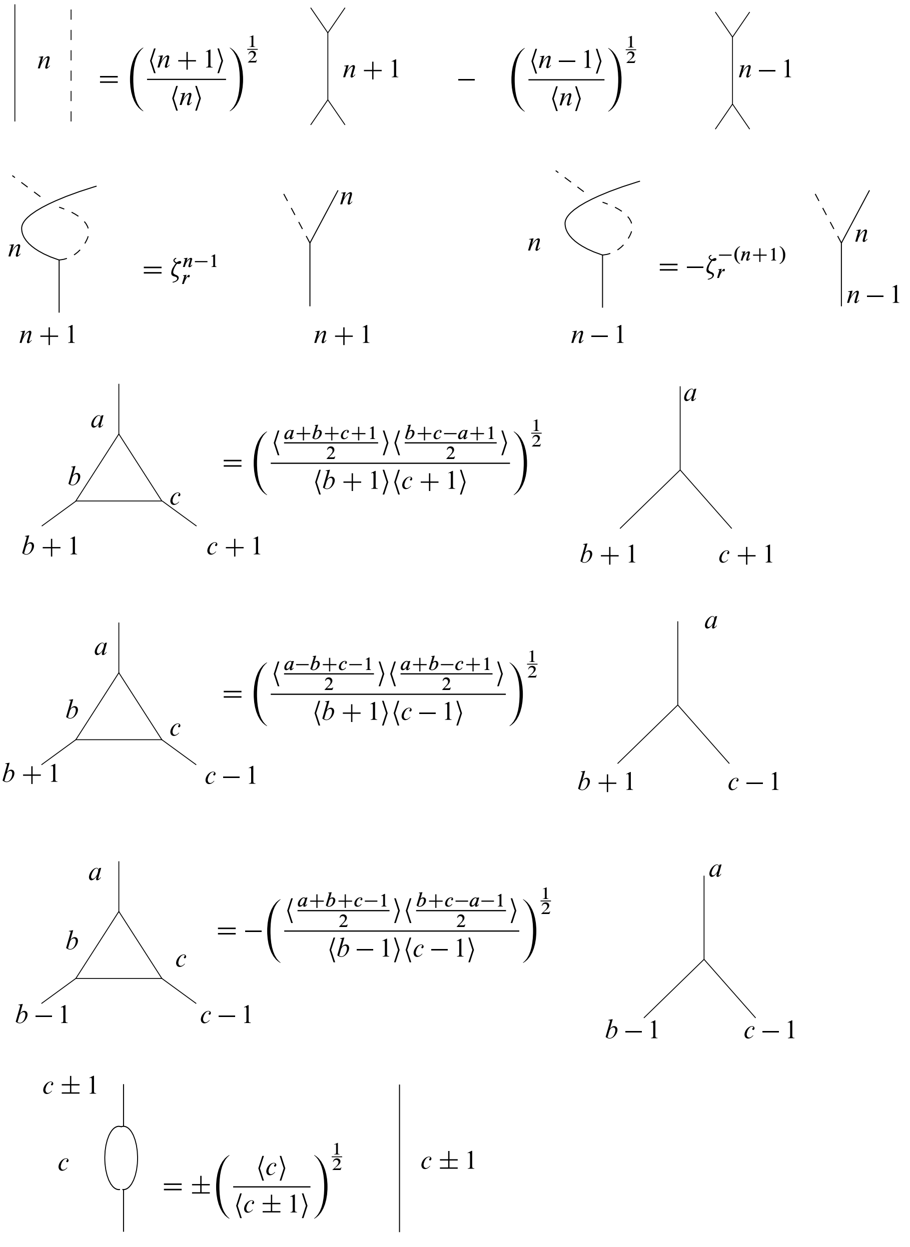

n
$= \left(\frac{\langle n+1\rangle}{\langle n\rangle}\right)^{\frac{1}{2}}$
$n+1$
$- \left(\frac{\langle n-1\rangle}{\langle n\rangle}\right)^{\frac{1}{2}}$
$n-1$
n
$n+1$
$= \zeta_r^{n-1}$
n
$n+1$
n
$n-1$
$= -\zeta_r^{-(n+1)}$
n
$n-1$
a
b
c
$b+1$
$c+1$
$= \left(\frac{\langle\frac{a+b+c+1}{2}\rangle\langle\frac{b+c-a+1}{2}\rangle}{\langle b+1\rangle\langle c+1\rangle}\right)^{\frac{1}{2}}$
a
$b+1$
$c+1$
a
b
c
$b+1$
$c-1$
$= \left(\frac{\langle\frac{a-b+c-1}{2}\rangle\langle\frac{a+b-c+1}{2}\rangle}{\langle b+1\rangle\langle c-1\rangle}\right)^{\frac{1}{2}}$
a
$b+1$
$c-1$
a
b
c
$b-1$
$c-1$
$= -\left(\frac{\langle\frac{a+b+c-1}{2}\rangle\langle\frac{b+c-a-1}{2}\rangle}{\langle b-1\rangle\langle c-1\rangle}\right)^{\frac{1}{2}}$
a
$b-1$
$c-1$
$c \pm 1$
c
$= \pm\left(\frac{\langle c\rangle}{\langle c\pm 1\rangle}\right)^{\frac{1}{2}}$
$c \pm 1$

Figure 3: Fusion rules. These "normalized" fusion rules allow us to simplify the union of a colored banded graph and a curve colored by 2. The dotted edges are colored by 2. The first rule allows to merge an edge colored by 2 with another one. The second line consists of the "half-twist formulae" of [22]. When all curves have been merged with the graph, the 3rd, 4th and 5th lines can be used to remove trigons, and the last rule to remove bigons.

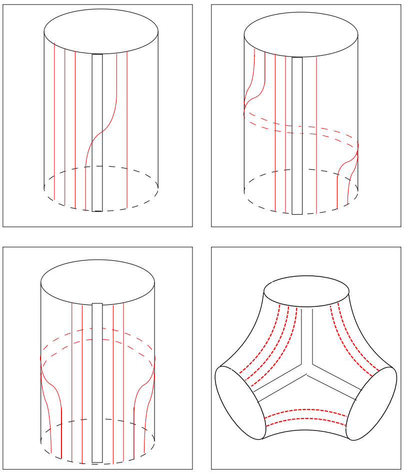


Figure 4: Dehn presentation of multicurves

With this setting, we give a normalized version of the fusion rules in TQFT. The fusion rules derived in [22], give a way to compute the image of the vector φ_c under the curves operators. We list the fusion rules that we will need in Figure 3; our version differs from the rules in [22], as we express them with the normalized vectors φ_c instead of the vectors ψ_c from [22].

We will perform the computations by using the fusion rules only locally, that is only inside of a pair of pants of the pants decomposition, or inside an annulus in the neighborhood of one of the curves C_e.

Indeed, for γ a multicurve, by a classification provided by Dehn, we can isotope γ so that the intersection of γ with each pants P_s of the decomposition looks like the 4th picture of Figure 4, and the intersection with each of the annuli A_e^k looks like one of the first three pictures of Figure 4.

Furthermore, in this isotopy class, the intersection of γ with each C_e is the smallest in the isotopy class of γ. We refer to [15, Section 4.3] for this classification.

Now, we do the computations in two steps:

First, we use fusion rules to reduce each type of piece to elements corresponding to the intersection of the graph Γ in a pants or annulus with a certain coloring, glued with "candlesticks".

A *candlestick* is an element of the TQFT vector space of an annulus that is the normalized vector associated to a banded trivalent graph in an annulus, consisting of a central edge joining the boundary components (with no twist), colored by $n \in \mathcal{C}_r$ on the bottom

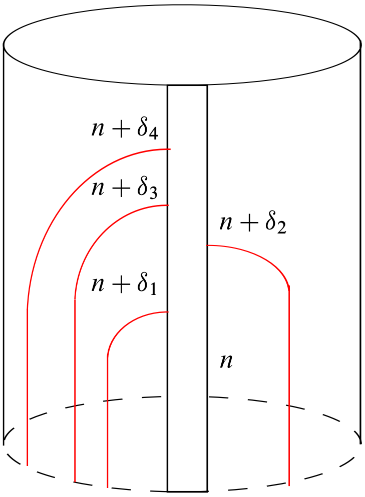
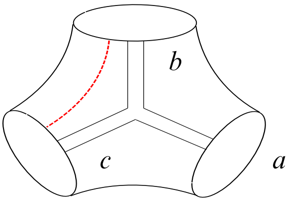


Figure 5: A candlestick $C(n,\varepsilon,\theta)$ with 4 legs. We denote by $\delta_i = \sum_{j=1}^{i} \varepsilon_j$ the partial sums of the color shifts ε_j. Notice that the legs can go alternatively to the left or to the right of the central edge.

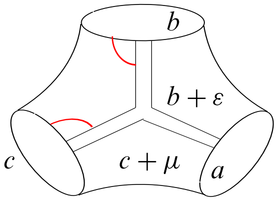
component, a collection of legs colored by 2, joining the central edge and the bottom component, as in Figure 5.

The data that defines a candlestick with k legs $C(n,\varepsilon,\Theta)$ is the color $n \in \mathcal{C}_r$ of the central edge at the bottom, the order Θ in which the legs join the central edge, and the shifts of the color of the central edge $(\varepsilon_i)_{i=1\dots k}$ when we pass each vertex corresponding to a leg.

Reduction of the different pieces Simple computations using fusion rules give us the following formulae when the pants or the annuli contain only one curve:



$$= \sum_{\varepsilon,\mu} F_{\varepsilon,\mu}(a,b,c,r)$$



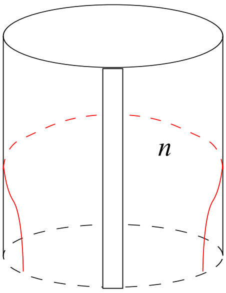
where we set

$$F_{+,+}(a,b,c,r) = \left(\frac{\left\langle \frac{a+b+c+1}{2} \right\rangle \left\langle \frac{b+c-a+1}{2} \right\rangle}{\langle b\rangle\langle c\rangle} \right)^{\frac{1}{2}},$$

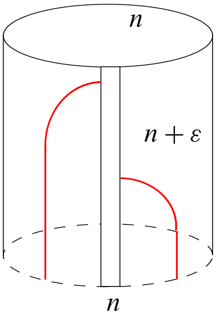
$$F_{+,-}(a,b,c,r) = F_{-,+}(a,c,b,r) = -\left(\frac{\left\langle \frac{a-b+c-1}{2} \right\rangle \left\langle \frac{a+b-c-1}{2} \right\rangle}{\langle b\rangle\langle c\rangle} \right)^{\frac{1}{2}},$$

$$F_{-,-}(a,b,c,r) = -\left(\frac{\left\langle\frac{a+b+c-1}{2}\right\rangle\left\langle\frac{b+c-a-1}{2}\right\rangle}{\langle b\rangle\langle c\rangle}\right)^{\frac{1}{2}};$$

next,



$$= \sum_{\varepsilon} G_{\varepsilon}(n,r)$$

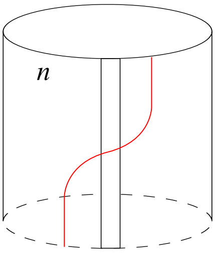


where

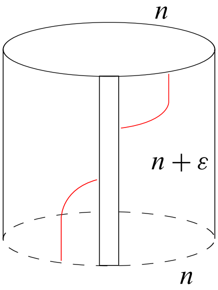
$$G_{+}(n,r) = (-1)^{n+1} e^{-i\pi(n-1)/(2r)}\left(\frac{\langle n+1\rangle}{\langle n\rangle}\right)^{\frac{1}{2}},$$

$$G_{-}(n,r) = (-1)^{n+1} e^{i\pi(n+1)/(2r)}\left(\frac{\langle n-1\rangle}{\langle n\rangle}\right)^{\frac{1}{2}};$$

third,



$$= \sum_{\varepsilon} H_{\varepsilon}(n,r)$$

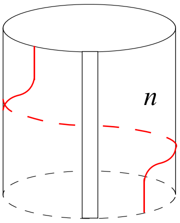


where

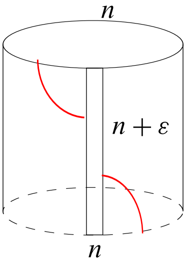
$$H_{+}(n,r) = (-1)^{n+1} e^{i\pi(n-1)/(2r)}\left(\frac{\langle n+1\rangle}{\langle n\rangle}\right)^{\frac{1}{2}},$$

$$H_{-}(n,r) = (-1)^{n+1} e^{-i\pi(n+1)/(2r)}\left(\frac{\langle n-1\rangle}{\langle n\rangle}\right)^{\frac{1}{2}};$$

and lastly



$$= \sum_{\varepsilon} L_{\varepsilon}(n,r)$$



where

$$L_+(n,r) = (-1)^{n+1} e^{i\pi(n+2)/(2r)} \left(\frac{\langle n+1\rangle}{\langle n\rangle}\right)^{\frac{1}{2}},$$

$$L_-(n,r) = (-1)^{n+1} e^{-i\pi(n-2)/(2r)} \left(\frac{\langle n-1\rangle}{\langle n\rangle}\right)^{\frac{1}{2}}.$$

All these coefficients are of the required form $\overline{c}(\gamma)F(c/r, 1/r)$ for some smooth function F defined on V_γ.

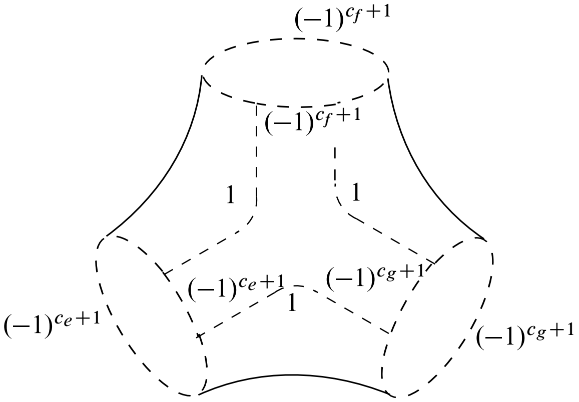
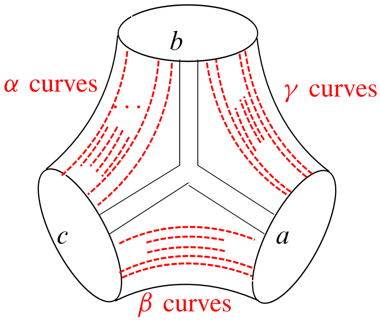
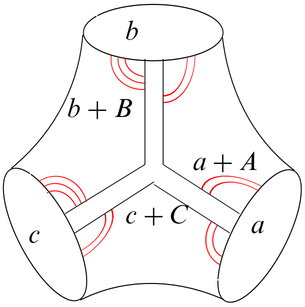


Figure 6: The cocycle $\overline{c}$ on the pants bounded by the curves C_e, C_f and C_g

If we have many curves in a pants or annulus, we only need to choose an order to make the fusions, and apply the latter formulae. For example, in the case of the pants, we obtain:



$$= \sum_{\varepsilon,\mu,\nu} P_{\varepsilon,\mu,\nu}(a,b,c,r)$$



where we use the notation $A = \sum_{i=1}^{\beta+\gamma} \varepsilon_i$, $B = \sum_{j=1}^{\alpha+\gamma} \mu_j$ and $C = \sum_{k=1}^{\alpha+\beta} \nu_k$.

Here we have first used fusion on the α curves that go from C_b to C_c, then the β curves that run from C_a to C_c, and finally the γ curves from C_a to C_c. With this order for the fusions, the coefficients $P_{\varepsilon,\mu,\nu}(a,b,c,r)$ are products of three factors corresponding to each series of fusions:

$$F_{\mu_1,\nu_1}(a,b,c,r)F_{\mu_2,\nu_2}(a,b+\mu_1,c+\nu_1,r)\cdots F_{\mu_\alpha,\nu_\alpha}\Bigl(a,b+\sum_{i=1}^{\alpha-1}\mu_i,c+\sum_{i=1}^{\alpha-1}\nu_i,r\Bigr),$$

$$F_{\nu_{\alpha+1},\varepsilon_1}\Bigl(b+\sum_{i=1}^{\alpha}\mu_i,a,c+\sum_{i=1}^{\alpha}\nu_i,r\Bigr)$$
$$\cdots F_{\nu_{\alpha+\beta},\varepsilon_\beta}\Bigl(b+\sum_{i=1}^{\alpha}\mu_i,a+\sum_{i=1}^{\beta-1}\varepsilon_i,c+\sum_{i=1}^{\alpha+\beta-1}\nu_i,r\Bigr),$$

$$F_{\mu_{\alpha+1},\varepsilon_{\beta+1}}\Bigl(c+\sum\nu,b+\sum_{i=1}^{\alpha}\mu_i,a+\sum_{i=1}^{\beta}\varepsilon_i,r\Bigr)$$
$$\cdots F_{\mu_{\alpha+\gamma},\varepsilon_{\beta+\gamma}}\Bigl(c+\sum\nu,b+\sum_{i=1}^{\alpha+\gamma-1}\mu_i,a+\sum_{i=1}^{\beta+\gamma-1}\varepsilon_i,r\Bigr).$$

Notice that, at every step of the fusion, the shifts in the color c_e are sums of ± 1 terms, one term for each arc intersecting C_e that has been merged with Γ. Thus the coefficients $P_{\varepsilon,\mu,\nu}$ are defined and smooth on the required domain $V_\gamma=\{(\tau,\hbar)\mid \tau_e\pm I_e^\gamma\hbar\in U\}$. Furthermore, in the end the shift of c_e is no greater than the number of curves that intersect C_e and of the same parity as this number.

We now only need to explain what happens when we glue together two candlesticks.

First, note that we can only paste candlesticks with the same number of legs, and the same bottom color n. Moreover, if we paste two candlesticks $C(n,\varepsilon,\Theta)$ and $C(n,\mu,\Theta')$ with $\sum_j\mu_j\neq\sum_i\varepsilon_i$, then we always obtain 0 (as the vector space $V_r(\Sigma)$ of a sphere Σ with two points marked by different colors is 0).

Proposition 4.1 *The gluing of candlesticks $C(n,\varepsilon,\Theta)$ and $C(n,\mu,\Theta')$ with k legs with $\sum_{i=1}^k\varepsilon_i=\sum_{j=1}^k\mu_j$ is proportional to a band colored by $n+\sum\varepsilon_i$ joining the two boundary components of the annulus with no twist, the proportionality constant being $G(n/r,1/r)$, where G is a smooth function on $\{(\tau,\hbar)\mid\tau\pm k\hbar\in(0,1)\}$.*

We should point out that, in this proposition, the function G depends on Θ, Θ', ε and μ.

Proof We prove this proposition by induction on the number of legs of the candlestick. If we paste two candlesticks with only one leg, this is direct from the fusion rule eliminating bigons (see Figure 3), as it only produces a factor $(\langle c\pm 1\rangle/\langle c\rangle)^{1/2}$. Now,

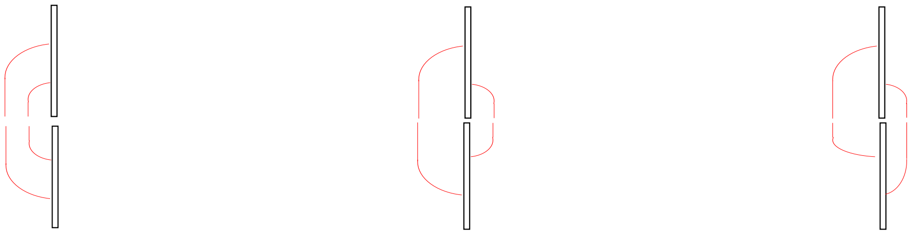
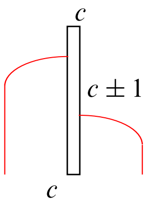


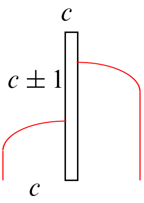
Figure 7

if $n = 2$, the only delicate case is when the legs of the two parts are positioned as in Figure 7(c).

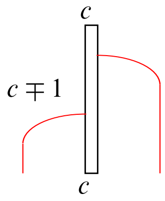
Indeed, in cases (a) and (b), we can simply eliminate two bigons. For (c), we use the following switching legs formulae:

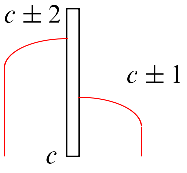


$= \mp\frac{\langle 1\rangle}{\langle c\rangle}$

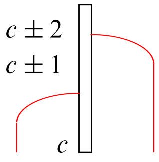


$+ \quad \frac{(\langle c+1\rangle\langle c-1\rangle)^{1/2}}{\langle c\rangle}$





$=$



To get such formulae, we have to verify that gluing the left-hand side or the right-hand side with a two-legs candlestick on the bottom, with any color shifts, we get the same result after using the fusion rules for bigons and triangles elimination. This is a straightforward computation, so we will omit it here.

This shows Proposition 4.1 for $k \leq 2$.

Now, suppose we glue two candlesticks with $k+1$ legs. We have two cases:

In Figure 8 (left), the upper leg of the upper candlestick and the bottom leg of the bottom candlestick both go to the right (or both to the left); the gluing is obtained by gluing two candlesticks with k legs, then suppressing a bigon. The factor we get is of

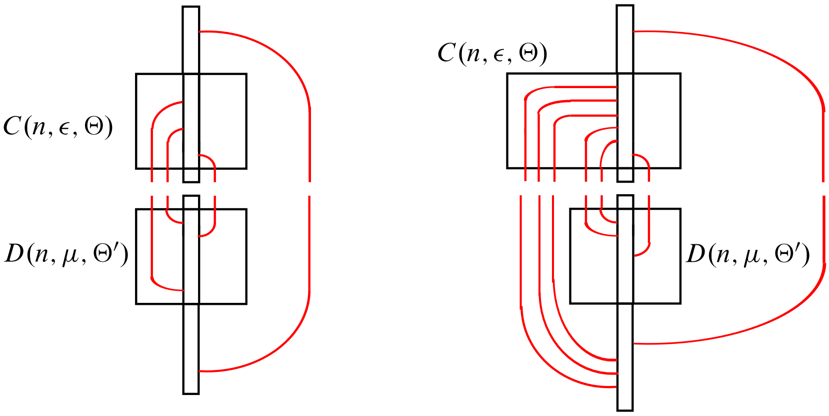


Figure 8: The two cases of pasting candlesticks with k legs

the form

$$G\Big(\frac{n}{r},\frac{1}{r}\Big)\bigg(\frac{\langle n+\sum_{i=1}^{k+1}\varepsilon_i\rangle}{\langle n+\sum_{i=1}^{k}\varepsilon_i\rangle}\bigg)^{\frac{1}{2}},$$

the factor $G(n/r,1/r)$ coming from k–leg candlestick elimination, and the other factor from the bigon elimination rule. It is indeed a function of $(n/r,1/r)$ that is smooth on the domain we claimed.

In Figure 8 (right), the upper leg of the upper part and the bottom leg of the bottom part go to different sides. We apply a sequence of switching legs formulae until the leg connected to the upper leg of the candlestick is the bottom leg of the bottom candlestick. Each of these operations yields a smooth function on V_γ as a factor; this comes from the switching legs formulae and the fact that all intermediate colors on the central edge are of the form $n+\sum_{i=1}^{j}\varepsilon_i$, with $j\le I_e^\gamma$. Then we are back to the former case. □

4.2 Examples of the ψ–symbol

We derive expressions of the ψ–symbol for two families of curves on Σ: the first family consists of the curves C_e of the pants decomposition itself, and the other of curves D_e, $e\in E$, that are in some sense dual to the curves C_e. The D_e are defined this way: if e is an internal edge that joins a vertex to itself, then D_e is a loop parallel to e. If e joins two different vertices, then D_e consists of two arcs parallel to e that we close into a loop as in Figure 9.

Note that C_e and D_f intersect each other if and only if $e=f$, and in this case they intersect once or twice. Finally, the classes in $H_1(\Gamma,\partial\Gamma,\mathbb{Z}/2)$ represented by $p_*(C_e)$ and by $p_*(D_e)$ are all zero. Note that in the case where D_e and C_e have one

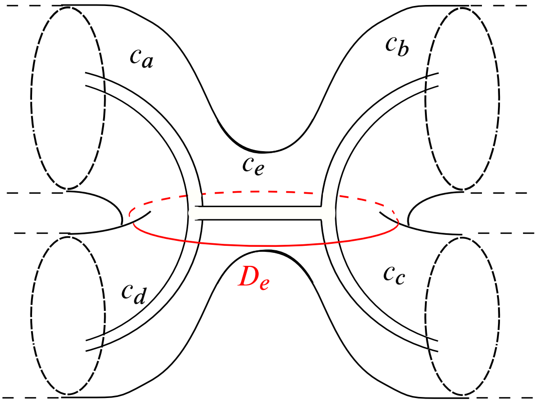


Figure 9: The curve D_e when e joins two distinct trivalent vertices of Γ

point of intersection, $p_*(D_e)$ is not zero as a class in $H_1(\Gamma, \mathbb{Z}/2)$, however it is in $H_1(\Gamma, \partial\Gamma, \mathbb{Z}/2)$ as $p(D_e)$ is homotopic to a boundary curve in the surface Γ.

Proposition 4.2 *We have, for any $e \in E$ and $c \in U_r$:*

(1) $T_r^{C_e}\varphi_c = -2\cos(\pi c_e/r)\varphi_c$ *and* $\sigma^{C_e}(\tau, \hbar, \theta) = -2\cos(\pi\tau_e)[0]$.

(2) *In the case where e is an edge joining a trivalent vertex to itself as in Figure 10 we have*

$$\sigma^{D_e}(\tau, \hbar, \theta) = \big(W(\pi\tau_e, \pi\tau_f, \hbar)e^{i\theta_e} + W(\pi\tau_e, \pi\tau_f, -\hbar)e^{-i\theta_e}\big)[0],$$

where

$$W(\tau, \alpha, \hbar) = \left(\frac{\sin(\tau + \alpha/2 + \hbar/2)\sin(\tau - \alpha/2 + \hbar/2)}{\sin\tau\,\sin(\tau + \hbar)}\right)^{\frac{1}{2}}.$$

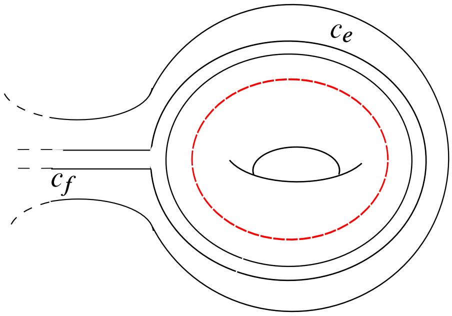


Figure 10: The curve D_e when e joins a trivalent vertex of Γ to itself

(3) *In the case where e is an edge between two distinct trivalent vertices as in Figure 9 we have*

$$\sigma^{D_e}(\tau,\hbar,\theta) = -\big(I(\pi\tau,\pi\hbar) + J(\pi\tau,\pi\hbar)e^{2i\theta_e} + J(\pi(\tau-2\hbar\delta_e),\pi\hbar)e^{-2i\theta_e}\big)[0].$$

Here, we have set δ_e for the element in $\mathbb{R}^E$ such that $\delta_{e,f}=1$ if and only if $e=f$,

$$\begin{aligned} I(\tau,\hbar) &= 2\cos(\tau_c+\tau_d-\hbar) \\ &\quad + 4\frac{\sin\frac{\tau_a+\tau_d-\tau_e-\hbar}{2}\sin\frac{\tau_a-\tau_d+\tau_e+\hbar}{2}\sin\frac{\tau_b+\tau_c-\tau_e-\hbar}{2}\sin\frac{\tau_b-\tau_c+\tau_e+\hbar}{2}}{\sin\tau_e\sin(\tau_e+\hbar)} \\ &\quad + 4\frac{\sin\frac{\tau_a+\tau_d+\tau_e-\hbar}{2}\sin\frac{-\tau_a+\tau_d+\tau_e-\hbar}{2}\sin\frac{\tau_b+\tau_c+\tau_e-\hbar}{2}\sin\frac{-\tau_b+\tau_c+\tau_e-\hbar}{2}}{\sin\tau_e\sin(\tau_e-\hbar)} \end{aligned}$$

and

$$\begin{aligned} J(\tau,\hbar) = 4\Bigg(&\frac{\sin\frac{\tau_a+\tau_d-\tau_e-\hbar}{2}\sin\frac{\tau_a-\tau_d+\tau_e+\hbar}{2}\sin\frac{\tau_b+\tau_c-\tau_e-\hbar}{2}\sin\frac{\tau_b-\tau_c+\tau_e+\hbar}{2}}{\sin\tau_e\sin(\tau_e+\hbar)} \\ &\times\frac{\sin\frac{\tau_a+\tau_d+\tau_e+\hbar}{2}\sin\frac{-\tau_a+\tau_d+\tau_e+\hbar}{2}\sin\frac{\tau_b+\tau_c+\tau_e+\hbar}{2}\sin\frac{-\tau_b+\tau_c+\tau_e+\hbar}{2}}{\sin(\tau_e+\hbar)\sin(\tau_e+2\hbar)}\Bigg)^{\frac{1}{2}}. \end{aligned}$$

The expressions of $T_r^{C_e}$ and $T_r^{D_e}$ can be derived by using the fusion rules. The computations are rather long in the last case, but straightforward.

These expressions, as well as the expressions of the ψ–symbol of the curves C_e and D_e were already given in [21]. They also checked by hand that the formulae of Theorem 1.3 were satisfied by these curves. We will only derive from the formulae that the zeroth- and first-order term for these curves are related as in Theorem 1.3, a fact that we will use later:

Proposition 4.3 *Let γ be any of the curves C_e or D_e. Then*

$$\sigma^\gamma(\tau,\hbar,\theta) = \sigma^\gamma(\tau,0,\theta) + \frac{\hbar}{2i}\sum_{e\in E}\frac{\partial^2}{\partial\tau_e\,\partial\theta_e}\sigma^\gamma(\tau,0,\theta) + o(\hbar).$$

Proof For C_e, there is not much to prove: as σ^{C_e} does not depend on $\hbar$, the first-order term vanishes, and $\partial^2\sigma^\gamma(\tau,0,\theta)/\partial\tau_e\,\partial\theta_e$ also vanishes as σ^{C_e} does not depend on θ_e.

For the curves D_e, we need to separate the case where e joins a vertex to itself, and the case where it joins two distinct vertices.

In the first case, depicted by Figure 10, we have

$$\sigma^{D_e}(\tau,\hbar,\theta) = (W(\pi\tau_e,\pi\tau_f,\pi\hbar)e^{i\theta_e} + W(\pi\tau_e,\pi\tau_f,-\pi\hbar)e^{-i\theta_e})[0].$$

Notice that we get $W(\pi\tau_e,\pi\tau_f,\pi\hbar)=W\bigl(\pi\bigl(\tau_e+\frac{1}{2}\hbar\bigr),\pi\tau_f,0\bigr)+o(\hbar)$ from the formula for W given above. Thus

$$\begin{aligned}\sigma^{D_e}&(\tau,\hbar,\theta)\\&=\sigma^{D_e}(\tau,0,\theta)+\frac{\hbar}{2}\Bigl(\frac{\partial}{\partial\tau_e}[W(\pi\tau_e,\pi\tau_f,0)e^{i\theta_e}]-\frac{\partial}{\partial\tau_e}[W(\pi\tau_e,\pi\tau_f,0)e^{-i\theta_e}]\Bigr)[0]+o(\hbar)\\&=\sigma^{D_e}(\tau,0,\theta)+\frac{\hbar}{2i}\sum_{e\in E}\frac{\partial^2}{\partial\tau_e\,\partial\theta_e}\sigma^{D_e}(\tau,0,\theta)+o(\hbar),\end{aligned}$$

as expected.

Finally, in the second case above, we have

$$\sigma^{D_e}(\tau,\hbar,\theta)=-\bigl(I(\pi\tau,\pi\hbar)+J(\pi\tau,\pi\hbar)e^{2i\theta_e}+J(\pi(\tau-2\hbar\delta_e),\pi\hbar)e^{-2i\theta_e}\bigr)[0].$$

It is easily seen that $J(\tau,\hbar)=J(\tau+\hbar\delta_e,0)$. Thus we only need to prove that $I(\tau,\hbar)=I(\tau,0)+o(\hbar)$. This is a bit more tricky:

First, notice that we can write

$$I(\tau,\hbar)=2\cos(\tau_c+\tau_d-\hbar)+\frac{1}{\sin\tau_e}(F(\tau_e+\hbar)-F(-\tau_e+\hbar))+o(\hbar),$$

where

$$\begin{aligned}F(\tau_e)&=4\frac{\sin\frac{\tau_a+\tau_d-\tau_e}{2}\sin\frac{\tau_a-\tau_d+\tau_e}{2}\sin\frac{\tau_b+\tau_c-\tau_e}{2}\sin\frac{\tau_b-\tau_c+\tau_e}{2}}{\sin\tau_e}\\&=\frac{(\cos(\tau_d-\tau_e)-\cos\tau_a)(\cos(\tau_c-\tau_e)-\cos\tau_b)}{\sin\tau_e}.\end{aligned}$$

Therefore, the first-order term for $I(\tau,\hbar)$ is

$$\hbar\biggl(2\sin(\tau_c+\tau_d)+\frac{2}{\sin\tau_e}\frac{d}{d\tau_e}\mathcal{P}(F)(\tau_e)\biggr),$$

where $\mathcal{P}(F)$ is the even part of the function F. From the formula above, we have

$$\mathcal{P}(F)(\tau_e)=\sin(\tau_c+\tau_d)\cos\tau_e-\cos\tau_a\sin\tau_c-\cos\tau_b\sin\tau_d,$$

so that $(1/\sin\tau_e)\,d\mathcal{P}(F)(\tau_e)/d\tau_e=-\sin(\tau_c+\tau_d)$, and the first order of $I(\tau,\hbar)$ vanishes. □

The computations of σ^{C_e} and σ^{D_e} were previously used in [21] to prove a version of Theorem 1.3 for the punctured torus and the 4–holed sphere. Their approach was to derive from the above formulae that the asymptotic estimate of Theorem 1.3 is valid for the curves C_e, D_e and $\tau_{C_e}(D_e)$, where τ_{C_e} denotes the Dehn twist along C_e. Then they used the compatibility of the ψ–symbol with the product in $K(\Sigma,-e^{i\pi\hbar/2})$ to prove that if Theorem 1.3 is verified for γ and δ two multicurves, then it is also true for their product $\gamma \cdot \delta$. This yielded Theorem 1.3 for all multicurves in the punctured torus and the 4–holed sphere, as the curves C_e, D_e and $\tau_{C_e}(D_e)$ were sufficient to generate the Kauffman algebra.

However, this approach fails in higher genus, as this set of curves no longer generate the Kauffman algebra. Therefore, we developed another approach to tackle the higher-genus cases, which was also more conceptual and required less computations. Our fundamental idea is to use the multiplicativity of the ψ–symbol together with the theorem of Bullock (recalled in Section 2) to view the zeroth- and first-order term of the ψ–symbol in terms of algebra morphism and derivation of algebras on $\mathrm{Reg}(\mathcal{M}'(\Sigma))$. We then only need to compare this general shape with the values of the ψ–symbol on a few curves to get the formula of Theorem 1.3. (In fact, for the zeroth-order term we will only need the values on the C_e, while the first-order term also requires the values on the D_e).

5 Principal symbol and representation spaces

This section will be centered on the study of the principal symbol $\sigma^\gamma(\tau, 0, \theta)$, that is the zeroth order of the ψ–symbol $\sigma^\gamma(\tau, \hbar, \theta)$. The goal of the first subsection is to establish the formula for the principal symbol, which is stated in our main theorem: $\sigma_\chi^\gamma(\tau, 0, \theta) = f_\gamma(R_\chi(\tau, \theta))$, where f_γ is the function on $\mathcal{M}(\Sigma)$ such that $f_\gamma(\rho) = -\mathrm{Tr}(\rho(\gamma))$ and R_χ are action-angles parametrization on $\mathcal{M}(\Sigma)$.

5.1 Principal symbol and the SL_2–character variety

This section aims to establish a link between the components of the principal symbol σ_χ and functions on the space of representations $\pi_1(\Sigma) \to \mathrm{SL}_2(\mathbb{C})$.

We will start our study of the principal symbol by the following proposition, which describes which values $\sigma_\chi^\gamma(\tau, \theta)$ can take:

Proposition 5.1 *For any multicurve γ and $\chi \in \widehat{A}_\Gamma$, we have:*

(1) $\sigma_\chi^\gamma(\tau, \theta) \in \mathbb{R}$.

(2) $|\sigma_\chi^\gamma(\tau, \theta)| \leq 2^{n(\gamma)}$, *where $n(\gamma)$ is the number of components of γ.*

Proof (1) We recall that the components of the ψ–symbol σ_χ^γ are complex-valued. The stated property comes from the fact that curve operators are Hermitian: for any multicurve γ, and every r, the operator T_r^γ is a Hermitian endomorphism of $V_r(\Sigma)$.

By definition, we have $T_r^\gamma \varphi_c = \sum_k F_k^\gamma(c/r, 1/r)\varphi_{c+k}$. As the basis $(\varphi_c)_{c\in U_r}$ is a Hermitian basis, we get

$$F_{-k}^\gamma\Big(\frac{c+k}{r}, \frac{1}{r}\Big) = \overline{F_k^\gamma\Big(\frac{c}{r}, \frac{1}{r}\Big)}$$

for all $c \in U_r$. Then for $r \to +\infty$ we have $F_{-k}^\gamma(\tau, 0) = \overline{F_k^\gamma(\tau, 0)}$.

Hence $\sigma_\chi^\gamma(\tau,\theta) = \chi(\gamma)\sum_k F_k^\gamma(\tau,0)e^{ik\cdot\theta} \in \mathbb{R}$ for all $(\tau,\theta) \in U \times (\mathbb{R}/2\pi\mathbb{Z})^E$.

(2) We want to find a bound for $|\sigma_\chi^\gamma(\tau,\theta)|$, where γ is a multicurve. By definition, we have $\sigma_\chi^\gamma(\tau,\theta) = \chi(\gamma)\sum_k F_k^\gamma(\tau,0)e^{ik\cdot\theta}$. On the one hand, we know that the coefficients F_k^γ are zero as soon as there is an e such that $|k_e| > I_e^\gamma = \sharp(\gamma \cap C_e)$. The number of nonzero coefficients is then lower than $M_\gamma = \prod_{e\in E}(2I_e^\gamma + 1)$. On the other hand, for any $r \geq 2$ and $c \in U_r$,

$$F_k^\gamma\Big(\frac{c}{r}, \frac{1}{r}\Big) = \langle T_r^\gamma \varphi_c, \varphi_{c+k}\rangle \leq \|T_r^\gamma\|.$$

We recalled in Section 2 that the spectral radius of T_r^γ is always $\leq 2^{n(\gamma)}$. Thus we have $|F_k^\gamma(c/r, 1/r)| \leq 2^{n(\gamma)}$ for every r > 0 and every $c \in U_r$. Taking the limit, we get $|F_k^\gamma(\tau, 0)| \leq 2^{n(\gamma)}$.

These two estimations only allow us to write $|\sigma_\chi^\gamma(\tau,\theta)| \leq M_\gamma 2^{n(\gamma)}$. To obtain the promised inequality, we use the multiplicativity of $\sigma_\chi^\cdot(\tau,\theta)$:

We have $|\sigma_\chi^{\gamma^p}(\tau,\theta)| = |\sigma_\chi^\gamma(\tau,\theta)|^p$ for any integer p. But γ^p is also a multicurve, obtained by taking p parallel copies of each component of γ.

So we have that $|\sigma_\chi^{\gamma^p}(\tau,\theta)| \leq M_{\gamma^p}2^{n(\gamma^p)}$.

But the number of components $n(\gamma^p)$ is just $pn(\gamma)$, and the geometric intersection numbers

$$I_e^\gamma = \sharp(\gamma \cap C_e)$$

verify $I_e^{\gamma^p} \leq pI_e^\gamma$.

From the product formula defining M_γ, we get that $M_{\gamma^p} \leq p^{|E|}M_\gamma$.

We conclude that $|\sigma_\chi^{\gamma^p}(\tau,\theta)| \leq p^{|E|}M_\gamma 2^{pn(\gamma)}$.

Then, taking the limit $p \to +\infty$, we get that $|\sigma_\chi^\gamma(\tau,\theta)| \leq 2^{n(\gamma)}$ for all (τ,θ) in $U \times (\mathbb{R}/2\pi\mathbb{Z})^E$. □

Now, recall that the components of the ψ–symbol

$$\sigma_\chi(\tau,\theta)\colon K(\Sigma,-1) \to \mathbb{C}$$

are morphisms of algebras. There is a simple description of all such morphism of algebras: indeed, by Theorem 2.2, we have an isomorphism

$$K(\Sigma,-1) \simeq \mathrm{Reg}(\mathcal{M}'(\Sigma)),$$

where $\mathcal{M}'(\Sigma)$ stands for $\mathrm{Hom}(\pi_1\Sigma, \mathrm{SL}_2(\mathbb{C}))//\mathrm{SL}_2(\mathbb{C})$, the space of characters of the fundamental group of Σ in $\mathrm{SL}_2(\mathbb{C})$. This space is an affine algebraic variety, and we are writing $\mathrm{Reg}(\mathcal{M}'(\Sigma))$ for the set of regular functions from $\mathcal{M}'(\Sigma)$ to $\mathbb{C}$. A morphism of algebras ϕ from $\mathrm{Reg}(\mathcal{M}'(\Sigma))$ to $\mathbb{C}$ is always of the form

$$\phi\colon f \mapsto f(\rho)$$

for some $\rho \in \mathcal{M}'(\Sigma)$. We deduce the existence of maps

$$R_\chi\colon U \times (\mathbb{R}/2\pi\mathbb{Z})^E \to \mathcal{M}'(\Sigma)$$

such that $\sigma_\chi^\gamma(\tau,\theta) = f_\gamma(R_\chi(\tau,\theta))$.

5.2 A system of action-angle coordinates on the SU_2–character variety

This subsection will be devoted to the study of the maps R_χ more closely, the aim being to prove that it actually gives action-angle coordinates on the character variety $\mathrm{Hom}(\pi_1(\Sigma),\mathrm{SU}_2)/\mathrm{SU}_2$, which we will denote by $\mathcal{M}(\Sigma)$.

In $\mathcal{M}(\Sigma)$ there is an open dense subset $\mathcal{M}_{\mathrm{irr}}(\Sigma)$ consisting of all conjugacy of irreducible representations. It is a well-known fact that $\mathcal{M}_{\mathrm{irr}}(\Sigma)$ consists only of smooth points of $\mathcal{M}(\Sigma)$ and it has a symplectic structure.

The maps R_χ have at first sight their image in $\mathcal{M}'(\Sigma)$. Again, we have a subset $\mathcal{M}'_{\mathrm{irr}}(\Sigma) \subset \mathcal{M}'(\Sigma)$ consisting of conjugacy classes of irreducible representations, and there is a structure of complex symplectic variety on this subspace. Moreover, $\mathcal{M}_{\mathrm{irr}}(\Sigma) \subset \mathcal{M}'_{\mathrm{irr}}(\Sigma)$.

We have two remarks:

First, we point out that $R_\chi(\tau,\theta)$ is always a noncommutative representation. Indeed, for a commutative representation, we would have, for three adjacent edges e, f and g,

$$h_{C_e}(\rho) + h_{C_f}(\rho) = h_{C_g}(\rho)$$

for one of the three orderings of e, f and g, or have $h_{C_e}(\rho) + h_{C_f}(\rho) + h_{C_g}(\rho) = 2$. This can not happen for $R_\chi(\tau,\theta)$ as $(h_{C_e})_{e\in E}$ maps it to $\tau \in U$, and we have strict inequalities $\tau_g < \tau_e + \tau_f$ and $\tau_e + \tau_f + \tau_g < 2$.

Our second point is that the map R_χ is smooth. By our first remark its image is indeed in the smooth part of $\mathcal{M}'(\Sigma)$. Note that for any $\gamma \in K(\Sigma,-1)$ the map

$(\tau,\theta)\to\sigma^{\gamma}(\tau,0,\theta)$ is smooth on $U\times(\mathbb{R}/2\pi\mathbb{Z})^E$, so $(\tau,\theta)\to\mathrm{Tr}(R_{\chi}(\tau,\theta)(\gamma))$ is smooth for every $\gamma\in\pi_1(\Sigma)$. As the space $\mathcal{M}'(\Sigma)$ can be parametrized by a finite collection of coordinates $\rho\to\mathrm{Tr}(\rho(\gamma_j))$, where $\gamma_j\in\pi_1(\Sigma)$, the map

$$R_{\chi}\colon U\times(\mathbb{R}/2\pi\mathbb{Z})^E\to\mathcal{M}'(\Sigma)$$

is smooth.

Proposition 5.2 *The maps R_{χ} take values in $\mathcal{M}_{\mathrm{irr}}(\Sigma)=\mathrm{Hom}(\pi_1\Sigma,\mathrm{SU}_2)/\mathrm{SU}_2$.*

Proof Indeed, we have seen with Proposition 5.1 that $\sigma_{\chi}^{\gamma}(\tau,\theta)$ is real-valued. We can use a well-known lemma:

Lemma *Any irreducible subgroup $G\subset\mathrm{SL}_2(\mathbb{C})$ such that the trace of all elements of G are real is conjugated to either a subgroup of $\mathrm{SL}_2(\mathbb{R})$ or a subgroup of SU_2.*

The proof of this lemma is based only on elementary algebra, manipulating trace of products of elements of G. A detailed proof can be found for example in [19, pages 3040–3041].

As we have $\sigma^{\gamma}(\tau,0,\theta)=-\mathrm{Tr}(R(\tau,\theta)(\gamma))\in\mathbb{R}$, we get that $R(\tau,\theta)$ is conjugated to either a representation in $\mathrm{SL}_2(\mathbb{R})$ or a representation in SU_2.

To prove Proposition 5.2, we still need to dismiss the case where the image of $R_{\chi}(\tau,\theta)$ would be conjugated to a subgroup of $\mathrm{SL}_2(\mathbb{R})$. To this end, we use Proposition 5.1(2), which states that $|\mathrm{Tr}(R_{\chi}(\tau,\theta)\gamma)|\le 2$ for every $\gamma\in\pi_1(\Sigma)$ representing a simple closed curve on Σ. We use the following lemma, proved in [17, Lemma 3.1.1]:

Lemma *Let $\rho\colon\pi_1(\Sigma)\to\mathrm{PSL}_2(\mathbb{C})$ be a nonelementary representation, then there exist two simple loops a and b intersecting once such that $\rho(a)$ and $\rho(b)$ are loxodromic* (meaning $|\mathrm{Tr}(\rho(a))|>2$ and $|\mathrm{Tr}(\rho(b))|>2$) *and noncommuting.*

This lemma follows from elementary considerations in hyperbolic geometry. From the lemma, we get that, since $R(\tau,\theta)(a)$ is never loxodromic, it must be an elementary representation into $\mathrm{PSL}_2(\mathbb{C})$. But if $R(\tau,\theta)$ was conjugated to a representation in $\mathrm{SL}_2(\mathbb{R})$, it would be a commutative representation, and we saw that $R(\tau,\theta)$ is not. □

Proposition 5.3 *For any $\chi\in\widehat{A}_{\Gamma}$, the map*

$$R_{\chi}\colon U\times(\mathbb{R}/2\pi\mathbb{Z})^E\to\mathcal{M}(\Sigma),\quad(\tau,\theta)\mapsto R_{\chi}(\tau,\theta),$$

gives action-angle coordinates on the symplectic variety $\mathcal{M}_{\mathrm{irr}}(\Sigma)$.

Proposition 5.2 of [20] shows that when a pants decomposition $\mathcal{C} = \{C_e\}_{e \in E}$ of Σ is given, the family of functions $h_{C_e} = \frac{1}{\pi}\operatorname{Acos}\bigl(-\frac{1}{2} f_{C_e}\bigr)$ constitutes a moment mapping h: $h^{-1}(U) \to U$ and $h^{-1}(U)$ is an open dense subset of $\mathcal{M}(\Sigma)$. The variables τ_e are the action coordinates associated to this moment mapping:

$$h_{C_e}(R_\chi(\tau,\theta)) = \frac{1}{\pi}\operatorname{Acos}\bigl(-\tfrac{1}{2} f_{C_e}(R_\chi(\tau,\theta))\bigr) = \frac{1}{\pi}\operatorname{Acos}\bigl(-\tfrac{1}{2}\sigma_\chi^{C_e}(\tau,\theta)\bigr) = \tau_e,$$

where the third equality comes from the computation of the operator $T_r^{C_e}$ given in Section 4: for any coloration c of E, we have $T_r^{C_e}\varphi_c = -2\cos(\pi c/r)\varphi_c$, so that $\sigma_\chi^{C_e}(\tau,\theta,\hbar) = F_0^{C_e}(\tau,\hbar)\chi([0]) = -2\cos(\pi\tau_e)$.

The only missing condition for (τ,θ) to be a system of action-angle coordinates on $\mathcal{M}(\Sigma)$ is that

$$R_\chi^*(\omega) = \sum_{e \in E} d\tau_e \wedge d\theta_e,$$

where ω refers to the symplectic form on the variety $\mathcal{M}(\Sigma)$.

It also amounts to the fact that the vector fields ∂_{θ_e} and $X_{h_{C_e}}$ (the symplectic gradient associated to the function h_{C_e}) on $\mathcal{M}(\Sigma)$ are equal. This equality of vector fields can be rewritten in terms of Poisson brackets:

$$\{h_{C_e}, f\} = \frac{\partial}{\partial\theta_e} f(R_\chi(\tau,\theta)) \quad \text{for all } f \in C^\infty(\mathcal{M}(\Sigma),\mathbb{C}) \text{ and all } \tau,\ \theta.$$

As the map $f \to \{h_{C_e}, f\}$ is a first-order differential operator, and any function f on $\mathcal{M}(\Sigma)$ can be approximated at order 1 near any point $\rho \in \mathcal{M}(\Sigma)$ by a linear combination of trace functions f_γ associated to multicurves, we only need to verify the equality when $f = f_\gamma$, the trace function of a multicurve γ.

To compute such Poisson brackets, we can apply Theorem 2.3:

We denote by ε the linear map

$$\varepsilon\colon K(\Sigma, -e^{i\pi\hbar/2}) \to K(\Sigma,-1) \simeq \operatorname{Reg}(\mathcal{M}'(\Sigma)),$$
$$\sum_{\gamma \text{ multicurve}} c_\gamma(\hbar)\gamma \mapsto \sum_{\gamma \text{ multicurve}} c_\gamma(0)\gamma.$$

For γ and $\delta \in K(\Sigma, -e^{i\pi\hbar/2})$ we have

$$\{f_{\varepsilon(\gamma)}, f_{\varepsilon(\delta)}\} = f_{\varepsilon((i/\hbar)[\gamma,\delta])}$$

with $[\gamma,\delta] = \gamma\cdot\delta - \delta\cdot\gamma \in K(\Sigma, -e^{i\pi\hbar/2})$.

We apply the above formula to compute $\{h_{C_e}, f_\gamma\}$ for any $\gamma \in K(\Sigma, -e^{i\pi\hbar/2})$: We recall that $h_{C_e} = \frac{1}{\pi}\operatorname{Acos}\bigl(-\frac{1}{2} f_{C_e}\bigr)$. Our strategy to compute the Poisson bracket is to approximate h_{C_e} with polynomials in f_{C_e}.

Since $\tau \in U$ we have $-2\cos(\pi\tau_e) \in (-2,2)$ and we can choose a polynomial P such that $P\big(-2\cos(\pi(\tau_e+x))\big) = x + o(x^2)$.

Now, the maps

$$\{\cdot, f_\gamma\}\colon C^\infty(\mathcal{M}(\Sigma)) \to C^\infty(\mathcal{M}(\Sigma)) \quad \text{and} \quad (i/\hbar)[\cdot,\gamma]\colon K(\Sigma,-1) \to K(\Sigma,-1)$$

being derivations of algebras, we have, by Goldman's formula,

$$\{P(f_{C_e}), f_\gamma\}(R_\chi(\tau,\theta)) = f_{\varepsilon((i/\hbar)[P(C_e),\gamma])}(R_\chi(\tau,\theta)) = \sigma_\chi^{\varepsilon((i/\hbar)[P(C_e),\gamma])}(\tau,\theta,0).$$

We compute this last quantity: we recall that we wrote $T_r^\gamma \varphi_c = \sum_k F_k^\gamma(\tau,\hbar)\varphi_{c+k}$ and we gave in Section 4.2 the expression $T_r^{C_e}\varphi_c = -2\cos(\pi\tau_e)\varphi_c$. Hence $T_r^{P(C_e)}\varphi_c = P(-2\cos(\pi\tau_e))\varphi_c$. We deduce that, for $c \in U_r$,

$$T_r^{[P(C_e),\gamma]}\varphi_c = \sum_k P\big(-2\cos(\pi(\tau_e+k_e\hbar))\big)F_k^\gamma(\tau,\hbar)\varphi_{c+k} - \sum_k P(-2\cos(\pi\tau_e))F_k^\gamma(\tau,\hbar)\varphi_{c+k}.$$

But, since $[C_e^k] = [0]$ in A_Γ,

$$\sigma_\chi^{\varepsilon((i/\hbar)[P(C_e),\gamma])}(\tau,\theta,0) = i\sum_k \frac{P\big(-2\cos(\pi(\tau_e+k_e\hbar))\big) - P(-2\cos(\pi\tau_e))}{\hbar}\bigg|_{\hbar=0} F_k^\gamma(\tau,0)e^{ik\cdot\theta}\chi(\gamma).$$

By our choice of P this reduces to

$$\sum_k ik_e F_k^\gamma(\tau,\hbar)e^{ik\cdot\theta}\chi(\gamma) = \frac{\partial}{\partial\theta_e}\sigma_\chi^\gamma(\tau,0,\theta) = \frac{\partial}{\partial\theta_e} f_\gamma(R_\chi(\tau,\theta)).$$

The last equality ends the proof: we have $\{h_{C_e}, f_\gamma\}(R_\chi(\tau,\theta)) = \partial f_\gamma(R_\chi(\tau,\theta))/\partial\theta_e$ for every multicurve γ, and R_χ gives an action-angle parametrization of $\mathcal{M}_{\mathrm{irr}}(\Sigma)$. □

5.3 Origin of angle coordinates

We want to investigate how exactly R_χ varies with $\chi \in \widehat{A_\Gamma}$. We recall that according to Section 3.2, the values of two different morphisms χ and χ' on $[\gamma]$ differ by a representation $\rho\colon H_1(\Gamma,\partial\Gamma,\mathbb{Z}/2) \to \{\pm 1\}$.

Let us also get more precise information about angle coordinates. We recall that we have a hamiltonian $h\colon \mathcal{M}_{\mathrm{irr}}(\Sigma) \to U$, given by $(h(\rho))_e = \frac{1}{\pi}\operatorname{Acos}\big(\frac{1}{2}\mathrm{Tr}(\rho(C_e))\big)$. The hamiltonian flow gives an action of $\mathbb{R}^E$ on $\mathcal{M}_{\mathrm{irr}}(\Sigma)$. This action has a kernel

$$\Lambda = \mathrm{Vect}_{\mathbb{Z}}\{(2\pi u_e)_{e\in E},\ \pi(u_e+u_f+u_g)_{(e,f,g)\in S}\},$$

where $(u_e)_{e\in E}$ is the canonical basis of $\mathbb{R}^E$, E is the set of edges of Γ and S is the set of triples of edges adjacent to the same vertex in Γ. We also define $\Lambda' = \mathrm{Vect}_{\mathbb{Z}}(\pi u_e) \supset \Lambda$. The quotient Λ'/Λ then acts on $\mathcal{M}^{\mathrm{irr}}(\Sigma)$ by $\pi u_e \cdot \rho(\gamma) = (-1)^{(C_e,\gamma)}\rho(\gamma)$, where $(\cdot,\cdot)$ is the intersection form in Σ.

Now that we know that the maps R_χ give action-angle coordinates on $\mathcal{M}_{\mathrm{irr}}(\Sigma)$, the only ambiguity is the choice of the origin of the angle part. That is, we must have, for any χ, $\chi' \in \widehat{A}_\Gamma$, that $R_{\chi'}(\tau,\theta) = R_\chi(\tau,\theta + v_{\chi,\chi'}(\tau))$, where $v_{\chi,\chi'}$ is a continuous function from U to $\mathbb{R}/\Lambda$.

We use the values of R_χ on the curves D_e to get the origin of the angle coordinates. We have $\mathrm{Tr}(R_\chi(\tau,\theta)(D_e)) = -\sigma_\chi^{D_e}(\tau,0,\theta) = -2W(\pi\tau,0)\cos\theta_e$ if e joins a vertex to itself, and $\mathrm{Tr}(R_\chi(\tau,\theta)(D_e)) = I(\pi\tau,0) + 2J(\pi\tau,0)\cos(2\theta_e)$ otherwise. We see that, in the first case, $\theta_e = 0$ is the unique minimum of $\mathrm{Tr}(R_\chi(\tau,\theta)(D_e))$, so that the origin of this coordinate is the same for all $\chi \in \widehat{A}_\Gamma$. In the second case, $\theta_e \mapsto \mathrm{Tr}(R_\chi(\tau,\theta)(D_e))$ has exactly two maxima, one for $\theta_e = 0$ and one for $\theta_e = \pi$. So θ is fixed modulo πu_e. Thus, for χ, $\chi' \in \widehat{A}_\Gamma$, we have $v_{\chi,\chi'}(\tau) \in \Lambda'/\Lambda$. Furthermore, $v_{\chi,\chi'}$ is continuous, hence it has to be constant.

Taking two elements χ and χ' in $\widehat{A}_\Gamma$, we know that they differ by a morphism

$$\rho\colon H_1(\Gamma,\partial\Gamma,\mathbb{Z}/2) \to \{\pm 1\}.$$

It is possible to recover the vector $v_{\chi,\chi'} \in \Lambda'/\Lambda$ from the representation ρ: by Poincaré duality, one can write $\rho(p_*(\gamma)) = (-1)^{\langle C,\gamma\rangle}$, where $C \in H_1(\Sigma,\mathbb{Z}/2)$, p_* is the projection $H_1(\Sigma,\mathbb{Z}/2) \to H_1(\Gamma,\partial\Gamma,\mathbb{Z}/2)$ and $\langle\cdot,\cdot\rangle$ is the intersection form in $H_1(\Sigma,\mathbb{Z}/2)$. Remember that p_* maps each C_e to zero, so that the intersection of C with each C_e must vanish. As the C_e generate a Lagrangian of $H_1(\Sigma,\mathbb{Z}/2)$, C is a linear combination of the C_e and this yields a vector $v_\rho \in \Lambda'/\Lambda$ such that $R_{\rho\chi}(\tau,\theta) = R_\chi(\tau,\theta + v_\rho)$.

We need to note that when Γ is a planar graph we can drop this complicated consideration of angle origins and we could have taken the ψ–symbol to be just $\mathbb{C}$–valued. Indeed, in this case the intersection form in $H_1(\Gamma,\partial\Gamma,\mathbb{Z}/2)$ is trivial, and the image of $H_1(\Sigma,\mathbb{Z}/2) \to H_1(\Gamma,\partial\Gamma,\mathbb{Z}/2)$ is $\{0\}$, so the ψ–symbol is $\mathbb{C}$–valued.

6 First order of the ψ–symbol

In this section, we investigate the first-order term in $\hbar$ of the asymptotic expansion of the ψ–symbol. We identify this term by linking it with the principal symbol, for which we already know a formula.

We recall that for γ a multicurve, the map $(\tau,\hbar,\theta)\mapsto\sigma^{\gamma}(\tau,\hbar,\theta)$ is defined as a finite sum of smooth functions on V_{γ}, and V_{γ} is a neighborhood of $U\times\{0\}$ in $U\times[0,1]$. We may write, for any multicurve γ,

$$\sigma^{\gamma}(\tau,\hbar,\theta)=\sigma^{\gamma}(\tau,0,\theta)+\hbar(\Delta_{\gamma}(\tau,\theta)+D_{\gamma}(\tau,\theta))+o(\hbar).$$

Here, $\Delta_{\gamma}(\tau,\theta)$ refers to the expected first order as in Theorem 1.3:

$$\Delta_{\gamma}(\tau,\theta)=\frac{1}{2i}\sum_{e}\frac{\partial^2}{\partial\tau_e\,\partial\theta_e}\sigma^{\gamma}(\tau,0,\theta).$$

Hence, what we want to prove in this section is that the remainder $D_{\gamma}(\tau,\theta)$ is zero for all γ and $(\tau,\theta)\in U\times(\mathbb{R}/2\pi\mathbb{Z})^E$.

We remark that the previous expressions define $\Delta(\tau,\theta)$ and $D(\tau,\theta)$ as maps from the set of multicurves to A_{Γ}, which we can extend by linearity to linear maps $K(\Sigma,-e^{i\pi\hbar/2})\to A_{\Gamma}[\![\hbar]\!]$.

Furthermore, Δ_{γ} and D_{γ} are some linear combinations of partial derivatives of the smooth functions F_k on V_{γ}, so they are both smooth on $U\times(\mathbb{Z}/2\pi\mathbb{Z})^E$.

Proposition 6.1 *For any multicurve γ and for all (τ,θ), the remainder term $D_{\gamma}(\tau,\theta)$ vanishes, so that the first-order term of $\sigma^{\gamma}(\tau,\hbar,\theta)$ is*

$$\Delta_{\gamma}(\tau,\theta)=\frac{1}{2i}\sum_{e}\frac{\partial^2}{\partial\tau_e\,\partial\theta_e}\sigma^{\gamma}(\tau,0,\theta).$$

The proof relies on the following two lemmas:

Lemma 6.2 *Let (τ,θ) be in $U\times(\mathbb{R}/2\pi\mathbb{Z})^E$. We will provide $\mathbb{C}$ with the structure of a $K(\Sigma,-1)$–module (or equivalently of* $\mathrm{Reg}(\mathcal{M}'(\Sigma))$*–module): for $x\in\mathbb{C}$ and $f\in\mathrm{Reg}(\mathcal{M}'(\Sigma))$, we define $f\cdot x=f(R_{\chi}(\tau,\theta))x$. Then the corresponding component of the remainder term $\gamma\mapsto\chi(D_{\gamma}(\tau,\theta))$ is a derivation of $K(\Sigma,-1)$–modules from $K(\Sigma,-1)$ to $\mathbb{C}$.*

Lemma 6.3 *With respect to the above-discussed* $\mathrm{Reg}(\mathcal{M}'(\Sigma))$*–module structure on $\mathbb{C}$ as above, we have an isomorphism* $\mathrm{Der}(\mathrm{Reg}(\mathcal{M}'(\Sigma)),\mathbb{C})\simeq T_{R_{\chi}(\tau,\theta)}\mathcal{M}(\Sigma)$ *sending a vector $X\in T_{R_{\chi}(\tau,\theta)}\mathcal{M}(\Sigma)$ to the derivation $f\to\mathcal{L}_X f(R_{\chi}(\tau,\theta))$, and the vector fields $(R_{\chi}^{*}\,\partial/\partial\tau_e, R_{\chi}^{*}\,\partial/\partial\theta_e)$ give a basis of the tangent spaces $T_{R_{\chi}(\tau,\theta)}\mathcal{M}(\Sigma)$.*

Proof of Lemma 6.2 We use Proposition 3.3 to determine how the remainder term $D(\tau,\theta)$ behaves with the product of elements in $K(\Sigma,-e^{i\pi\hbar/2})$. We work with one component σ_χ of the ψ–symbol at a time. For $\gamma \in K(\Sigma,-1)$, we will use the notation $E_\gamma = \chi(\Delta_\gamma + D_\gamma)$, so that we can write $\sigma_\chi^\gamma(\tau,\hbar,\theta) = \sigma_\chi^\gamma(\tau,0,\theta) + \hbar E_\gamma(\tau,\theta) + o(\hbar)$. Then, applying $\chi \in \widehat{A}_\Gamma$ to Proposition 3.3 we have

$$\sigma_\chi^{\gamma\cdot\delta}(\tau,\hbar,\theta) = \sigma_\chi^\gamma(\tau,\hbar,\theta)\sigma_\chi^\delta(\tau,\hbar,\theta) + \frac{\hbar}{i}\sum_e \partial_{\tau_e}\sigma_\chi^\gamma(\tau,\hbar,\theta)\,\partial_{\theta_e}\sigma_\chi^\delta(\tau,\hbar,\theta) + o(\hbar).$$

We have $\sigma_\chi^\gamma(\tau,0,\theta) = f_\gamma(R_\chi(\tau,\theta))$. Recall that, by Theorem 2.3,

$$f_{\gamma\cdot\delta} = f_\gamma f_\delta + \hbar\frac{\pi}{i}\{f_\gamma, f_\delta\} + o(\hbar).$$

So, isolating terms of order 1 in $\hbar$, we get

$$\begin{aligned}&\frac{\pi}{i}\{f_\gamma, f_\delta\}(R_\chi(\tau,\theta)) + E_{\gamma\cdot\delta}(\tau,\theta)\\ &\quad = E_\gamma(\tau,\theta) f_\delta(R_\chi(\tau,\theta)) + E_\delta(\tau,\theta) f_\gamma(R_\chi(\tau,\theta))\\ &\qquad\qquad + \frac{1}{i}\sum_e \partial_{\tau_e} f_\gamma(R_\chi(\tau,\theta))\,\partial_{\theta_e} f_\delta(R_\chi(\tau,\theta)),\end{aligned}$$

but $\{f_\gamma, f_\delta\} = (1/2\pi)\sum_e \partial_{\tau_e} f_\gamma\,\partial_{\theta_e} f_\delta - \partial_{\tau_e} f_\delta\,\partial_{\theta_e} f_\gamma$. We deduce that

$$E_{\gamma\cdot\delta} = E_\gamma\sigma_\chi^\delta + E_\delta\sigma_\chi^\gamma + \frac{1}{2i}\sum_e \partial_{\tau_e}\sigma_\chi^\gamma\,\partial_{\theta_e}\sigma_\chi^\delta + \partial_{\theta_e}\sigma_\chi^\gamma\,\partial_{\tau_e}\sigma_\chi^\delta.$$

However, as for γ, $\delta \in K(\Sigma,-1)$ we have, by Theorem 2.2, that $f_{\gamma\cdot\delta} = f_\gamma f_\delta$, and

$$\chi(\Delta_\gamma) = \frac{1}{2i}\sum_e \frac{\partial^2 f_\gamma}{\partial\tau_e\,\partial\theta_e}\circ R_\chi,$$

the Leibniz rule implies that $\chi(\Delta_\gamma)$ satisfies the same law of composition:

$$\chi(\Delta_{\gamma\cdot\delta}) = \chi(\Delta_\gamma) f_\delta + \chi(\Delta_\delta) f_\gamma + \frac{1}{2i}\sum_e \partial_{\tau_e} f_\gamma\,\partial_{\theta_e} f_\delta + \partial_{\theta_e} f_\gamma\,\partial_{\tau_e} f_\delta.$$

This concludes the proof of Lemma 6.2: $\chi\circ D$ is a derivation. □

Proof of Lemma 6.3 It is well known that $\mathcal{M}'(\Sigma)$ is an affine algebraic variety whose smooth points is the open dense subset $\mathcal{M}'_{\mathrm{irr}}(\Sigma)$ (see [25], for example). The point $R_\chi(\tau,\theta)$ is thus a smooth point of $\mathcal{M}'(\Sigma)$ for any $(\tau,\theta) \in U \times \mathbb{R}/2\pi\mathbb{Z}$.

Then the proof comes from elementary considerations of algebraic geometry: when V is an affine algebraic variety and x a point of V, we put a structure of $\mathrm{Reg}(V)$–module on $\mathbb{C}$ by defining $f\cdot\lambda = f(x)\lambda$. Then $\mathrm{Der}_x(V,\mathbb{C})$ identifies with $T_xV = m_x/(m_x)^2$, the algebraic tangent space to V at x (where $m_x = \{f \mid f(x) = 0\}$), and the algebraic tangent space at a smooth point is the same as the tangent space of V at x in the sense of differential manifolds. As the affine variety $\mathcal{M}'(\Sigma)$ is smooth on the image of R_χ, by this general property, derivations of $\mathrm{Reg}(\mathcal{M}(\Sigma))$ can be viewed as vectors of the tangent space. As $(\tau,\theta)\mapsto R_\chi(\tau,\theta)$ is a parametrization of $\mathcal{M}(\Sigma)$, the vector fields $((R_\chi)_*\,\partial/\partial\tau_e,(R_\chi)_*\,\partial/\partial\theta_e)$ give a basis of the tangent space $T_{R_\chi(\theta,\tau)}\mathcal{M}(\Sigma)$ for each (τ,θ). □

Proof of Proposition 6.1 Combining Lemmas 6.2 and 6.3 allows us to assert that $\chi(D(\tau,\theta))$, viewed as a map $\mathrm{Reg}(\mathcal{M}'(\Sigma))\to\mathbb{C}$, is of the form $f\mapsto\mathcal{L}_X f(R_\chi(\tau,\theta))$ for some $X\in T_{R_\chi(\tau,\theta)}\mathcal{M}'(\Sigma)$ and we may write $X=\sum_e a_e\,\partial/\partial_{\tau_e}+b_e\,\partial/\partial_{\theta_e}$ for some coefficients a_e, b_e: $\mathcal{M}(\Sigma)\to\mathbb{C}$. As D_γ is smooth, so are the coefficients a_e and b_e.

We want to prove that these coefficients all vanish. To this end, we recall that we proved in Section 4.2 that the remainder term vanishes for the curves C_e and D_e. Furthermore, we have the formula of Section 4:

We have $\sigma^{C_e}(\tau,\hbar,\theta)=-2\cos\pi\tau_e[0]$, so that $\chi(D_{C_e})(\tau,\theta)=2a_e\pi\sin(\pi\tau_e)$. Since the remainder term vanishes on C_e, we must have $a_e=0$.

To show the vanishing of the b_e, we use the formulae for D_e:

For the first kind of curve D_e, described in Section 4.2, we have $f_{D_e}(R_\chi(\tau,\theta))=\sigma_\chi^{D_e}(\tau,0,\theta)=2W(\pi\tau,0)\cos\theta_e$, where W does not vanish for $\tau\in U$.

We know that the remainder term D_{D_e} vanishes, so we have

$$\chi(D_{D_e}(\tau,\theta))=b_e\frac{\partial}{\partial\theta_e}f_{D_e}(R_\chi(\tau,\theta))=-2b_e\pi\sin(\theta_e)W(\pi\tau,0)=0.$$

This yields $b_e=0$.

In the second case, $f_{D_e}(R_\chi(\tau,\theta))=\sigma_\chi^{D_e}(\tau,0,\theta)=-2J(\pi\tau,0)\cos 2\theta_e-I(\pi\tau,0)$ for the functions I and J defined in Section 4.2, which are nonvanishing for $\tau\in U$.

Again since $\chi(D_{D_e}(\tau,\theta))=b_e\,\partial f_{D_e}(R_\chi(\tau,\theta))/\partial\theta_e=4\pi b_e\sin(2\theta_e)J(\pi\tau,0)$ vanishes, we must have $b_e=0$. It follows that the remainder term $\gamma\mapsto D_\gamma$ is the zero derivation on $K(\Sigma,-1)\mapsto A_\Gamma$, which is the last ingredient we needed to complete the proof of Proposition 6.1. □

References

[1] **J E Andersen**, *Asymptotic faithfulness of the quantum* SU(n) *representations of the mapping class groups*, Ann. of Math. 163 (2006) 347–368 MR2195137

[2] **J E Andersen**, *The Nielsen–Thurston classification of mapping classes is determined by TQFT*, J. Math. Kyoto Univ. 48 (2008) 323–338 MR2436739

[3] **J E Andersen**, *Asymptotics of the Hilbert–Schmidt norm of curve operators in TQFT*, Lett. Math. Phys. 91 (2010) 205–214 MR2595923

[4] **J E Andersen**, *Toeplitz operators and Hitchin's projectively flat connection*, from "The many facets of geometry" (O García-Prada, J P Bourguignon, S Salamon, editors), Oxford Univ. Press (2010) 177–209 MR2681692

[5] **J E Andersen**, *Mapping class group invariant unitarity of the Hitchin connection over Teichmüller space*, preprint (2012) `arXiv:1206.2635`

[6] **J E Andersen**, **N L Gammelgaard**, *Hitchin's projectively flat connection, Toeplitz operators and the asymptotic expansion of TQFT curve operators*, from "Grassmannians, moduli spaces and vector bundles" (D A Ellwood, E Previato, editors), Clay Math. Proc. 14, Amer. Math. Soc., Providence, RI (2011) 1–24 MR2807846

[7] **J E Andersen**, **K Ueno**, *Abelian conformal field theory and determinant bundles*, Internat. J. Math. 18 (2007) 919–993 MR2339577

[8] **J E Andersen**, **K Ueno**, *Geometric construction of modular functors from conformal field theory*, J. Knot Theory Ramifications 16 (2007) 127–202 MR2306213

[9] **J E Andersen**, **K Ueno**, *Modular functors are determined by their genus zero data*, Quantum Topol. 3 (2012) 255–291 MR2928086

[10] **J E Andersen**, **K Ueno**, *Construction of the Witten–Reshetikhin–Turaev TQFT from conformal field theory*, Invent. Math. 201 (2015) 519–559 MR3370620

[11] **C Blanchet**, **N Habegger**, **G Masbaum**, **P Vogel**, *Topological quantum field theories derived from the Kauffman bracket*, Topology 34 (1995) 883–927 MR1362791

[12] **G W Brumfiel**, **H M Hilden**, $\mathrm{SL}(2)$ *representations of finitely presented groups*, Contemporary Mathematics 187, Amer. Math. Soc. (1995) MR1339764

[13] **D Bullock**, *Rings of* $\mathrm{SL}_2(\mathbf{C})$*–characters and the Kauffman bracket skein module*, Comment. Math. Helv. 72 (1997) 521–542 MR1600138

[14] **L Charles**, **J Marché**, *Multicurves and regular functions on the representation variety of a surface in* $\mathrm{SU}(2)$, Comment. Math. Helv. 87 (2012) 409–431 MR2914854

[15] **A Fathi**, **F Laudenbach**, **V Poenaru**, *Travaux de Thurston sur les surfaces*, Astérisque 66, Société Mathématique de France, Paris (1979) MR568308

[16] **V V Fock**, **A A Rosly**, *Poisson structure on moduli of flat connections on Riemann surfaces and the* r*–matrix*, from "Moscow Seminar in Mathematical Physics" (A Y Morozov, M A Olshanetsky, editors), Amer. Math. Soc. Transl. Ser. 2 191, Amer. Math. Soc., Providence, RI (1999) 67–86 MR1730456

[17] **D Gallo**, **M Kapovich**, **A Marden**, *The monodromy groups of Schwarzian equations on closed Riemann surfaces*, Ann. of Math. 151 (2000) 625–704 MR1765706

[18] **W M Goldman**, *Invariant functions on Lie groups and Hamiltonian flows of surface group representations*, Invent. Math. 85 (1986) 263–302 MR846929

[19] **M Heusener**, **E Klassen**, *Deformations of dihedral representations*, Proc. Amer. Math. Soc. 125 (1997) 3039–3047 MR1443155

[20] **L C Jeffrey**, **J Weitsman**, *Bohr–Sommerfeld orbits in the moduli space of flat connections and the Verlinde dimension formula*, Comm. Math. Phys. 150 (1992) 593–630 MR1204322

[21] **J Marché**, **T Paul**, *Toeplitz operators in TQFT via skein theory*, Trans. Amer. Math. Soc. 367 (2015) 3669–3704 MR3314820

[22] **G Masbaum**, **P Vogel**, *3–valent graphs and the Kauffman bracket*, Pacific J. Math. 164 (1994) 361–381 MR1272656

[23] **J H Przytycki**, **A S Sikora**, *On skein algebras and* $\mathrm{Sl}_2(\mathbf{C})$*–character varieties*, Topology 39 (2000) 115–148 MR1710996

[24] **N Reshetikhin**, **V G Turaev**, *Invariants of 3–manifolds via link polynomials and quantum groups*, Invent. Math. 103 (1991) 547–597 MR1091619

[25] **A S Sikora**, *Character varieties*, Trans. Amer. Math. Soc. 364 (2012) 5173–5208 MR2931326

[26] **V G Turaev**, *Skein quantization of Poisson algebras of loops on surfaces*, Ann. Sci. École Norm. Sup. 24 (1991) 635–704 MR1142906

[27] **V G Turaev**, *Quantum invariants of knots and 3–manifolds*, de Gruyter Studies in Mathematics 18, de Gruyter, Berlin (1994) MR1292673

[28] **E Witten**, *Quantum field theory and the Jones polynomial*, Comm. Math. Phys. 121 (1989) 351–399 MR990772

Institut de Mathématiques de Jussieu, Université Paris 6
4 place Jussieu, 75005 Paris, France

detcherry@math.msu.edu

Proposed: Vaughan Jones
Seconded: Robion Kirby, Leonid Polterovich

Received: 3 September 2012
Revised: 11 September 2015

Geometry & Topology 20 (2016) 3097–3132

Trisecting 4–manifolds

DAVID T GAY
ROBION KIRBY

We show that any smooth, closed, oriented, connected 4–manifold can be trisected into three copies of $\natural^k(S^1 \times B^3)$, intersecting pairwise in 3–dimensional handlebodies, with triple intersection a closed 2–dimensional surface. Such a trisection is unique up to a natural stabilization operation. This is analogous to the existence, and uniqueness up to stabilization, of Heegaard splittings of 3–manifolds. A trisection of a 4–manifold X arises from a Morse 2–function $G\colon X \to B^2$ and the obvious trisection of B^2, in much the same way that a Heegaard splitting of a 3–manifold Y arises from a Morse function $g\colon Y \to B^1$ and the obvious bisection of B^1.

57M50, 57M99; 57R45, 57R65

1 Introduction

Consider first the 3–dimensional case of an oriented, connected, closed 3–manifold Y^3. From a Morse function $f\colon Y \to [0,3]$ with only one critical point of index 0 and one of index 3, and all critical points of index i mapping to i, we see that $f^{-1}\big(\big[0,\frac{3}{2}\big]\big)$ and $f^{-1}\big(\big[\frac{3}{2},3\big]\big)$ are solid handlebodies, $\natural^g(S^1 \times B^2)$.

For uniqueness, we use Cerf theory [3] to get a homotopy $f_t\colon Y \to [0,3]$ between f_0 and f_1 (each giving Heegaard splittings) where this homotopy introduces no new critical points of index 0 or 3. There are births and deaths of cancelling pairs of index-1 and -2 critical points, but these stabilize the Heegaard splittings by connected summing with the standard genus-1 splitting of S^3. The homotopy f_t can be chosen to keep the index-1 critical values below $\frac{3}{2}$ and the index-2 above. Then handle slides between 1–handles, or 2–handles, take one Heegaard splitting to the other. (This is a now well-known Cerf-theoretic proof of the Reidemeister–Singer theorem (see eg Saveliev [10]), which was originally proved combinatorially; see Reidemeister [9] and Singer [11].)

Recall that a Heegaard *diagram* for a Heegaard splitting is a triple (F_g,α,β), where F_g is the Heegaard surface and each of α and β is a g–tuple of simple closed curves in F_g which bounds a basis of compressing disks in each of the two handlebodies. Thus

Published: 21 December 2016 DOI: 10.2140/gt.2016.20.3097

every 3–manifold is described by a Heegaard diagram, and two Heegaard diagrams describe diffeomorphic 3–manifolds if and only if they are related by stabilization, handle slides, and diffeomorphisms of F_g.

We now set up an analogous story in dimension four: Let $Z_k = \natural^k(S^1 \times B^3)$ with $Y_k = \partial Z_k = \sharp^k(S^1 \times S^2)$. Given an integer $g \geq k$, let $Y_k = Y_{k,g}^+ \cup Y_{k,g}^-$ be the standard genus-g Heegaard splitting of Y_k obtained by stabilizing the standard genus-k Heegaard splitting $g-k$ times.

Definition 1 Given integers $0 \leq k \leq g$, a *(g,k)–trisection* (see Figure 1) of a closed, connected, oriented 4–manifold X is a decomposition of X into three submanifolds $X = X_1 \cup X_2 \cup X_3$ satisfying the following properties:

(1) For each $i = 1,2,3$, there is a diffeomorphism $\phi_i \colon X_i \to Z_k$.

(2) For each $i = 1,2,3$, taking indices mod 3,

$$\phi_i(X_i \cap X_{i+1}) = Y_{k,g}^- \quad \text{and} \quad \phi_i(X_i \cap X_{i-1}) = Y_{k,g}^+.$$

Remark 2 Note that the triple intersection $X_1 \cap X_2 \cap X_3$ is a surface of genus g and that $\chi(X) = 2 + g - 3k$. Thus k is determined by X and g, and for this reason we will often refer to a (g,k)–trisection of X simply as a *genus-g trisection* of X. Also note that, for a fixed X, different trisections thus have the same genera mod 3.

Given a (g,k)–trisection $X = X_1 \cup X_2 \cup X_3$, consider the handlebodies $H_{ij} = X_i \cap X_j$ and the central genus-g surface $F_g = X_1 \cap X_2 \cap X_3 = \partial H_{ij}$. A choice of a system of g compressing disks on F_g for each of the three handlebodies gives three collections of g curves: $\alpha = (\alpha_1, \dots, \alpha_g)$, $\beta = (\beta_1, \dots, \beta_g)$ and $\gamma = (\gamma_1, \dots, \gamma_g)$, such that compressing along α gives H_{12}, compressing along β gives H_{23} and compressing along γ gives H_{31}. Furthermore, each pair (α,β), (β,γ) and (γ,α) is a Heegaard diagram for $\sharp^k(S^1 \times S^2)$.

Definition 3 A *(g,k)–trisection diagram* is a 4–tuple $(F_g, \alpha, \beta, \gamma)$ such that each triple (F_g,α,β), (F_g,β,γ), (F_g,γ,α) is a genus-g Heegaard diagram for $\sharp^k(S^1 \times S^2)$. The 4–manifold determined in the obvious way by this trisection diagram will be denoted $X(F_g, \alpha, \beta, \gamma)$.

Theorem 4 (existence) *Every closed, connected, oriented 4–manifold X has a (g,k)–trisection for some $0 \leq k \leq g$. Moreover, g and k are such that X has a handlebody decomposition with 1 0–handle, k 1–handles, $g-k$ 2–handles, k 3–handles and 1 4–handle.*

Remark 5 There are two trivial consequences of the handle decomposition mentioned in the theorem which are worth noting:

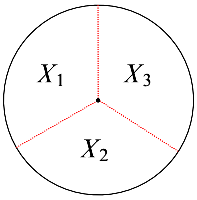


Figure 1: How the pieces of a trisection fit together

(1) If $k=0$, ie X_1, X_2 and X_3 are each 4–balls, then X has no 1– or 3–handles, and is thus simply connected.

(2) If $g=k$, then X has no 2–handles, so $X \cong \sharp^k S^1 \times S^3$.

The following is immediate:

Corollary 6 *Every closed 4–manifold is diffeomorphic to $X(F_g,\alpha,\beta,\gamma)$ for some trisection diagram $(F_g,\alpha,\beta,\gamma)$.*

Remark 7 Readers familiar with the Heegaard triples used by Ozsváth and Szabó [8] to define the Heegaard Floer 4–manifold invariants will see that a trisection diagram is a special type of Heegaard triple and may suspect that this corollary follows fairly quickly from the Heegaard triple techniques in [8]. In all fairness this is probably true; we will present two proofs of Theorem 4, one of which tells the story of how we discovered the result using Morse 2–functions, while the other is more in the spirit of [8], directly using ordinary handle decompositions. In some sense, then, our existence result can be thought of as a particularly nice packaging of the topological setup for [8].

Exactly as with Heegaard splittings in dimension 3, our uniqueness result for trisections of 4–manifolds is uniqueness up to a stabilization operation, which we now define. The idea is illustrated in Figure 2, in dimension 3.

Definition 8 (stabilization) Given a 4–manifold X with a trisection (X_1,X_2,X_3), we construct a new trisection (X_1',X_2',X_3'), as follows: For each $i,j\in\{1,2,3\}$, let H_{ij} be the handlebody $X_i\cap X_j$, with boundary $F=X_1\cap X_2\cap X_3$. Let a_{ij} be a properly embedded boundary parallel arc in each H_{ij}, such that the end points of a_{12}, a_{23} and a_{31} are disjoint in F. Let N_{ij} be a closed 4–dimensional regular neighborhood of a_{ij} in X (thus diffeomorphic to B^4), with N_{12}, N_{23} and N_{31} disjoint. Then we define

- $X_1' = (X_1 \cup N_{23}) \setminus (\mathring{N}_{31} \cup \mathring{N}_{12})$,
- $X_2' = (X_2 \cup N_{31}) \setminus (\mathring{N}_{12} \cup \mathring{N}_{23})$,
- $X_3' = (X_3 \cup N_{12}) \setminus (\mathring{N}_{23} \cup \mathring{N}_{31})$.

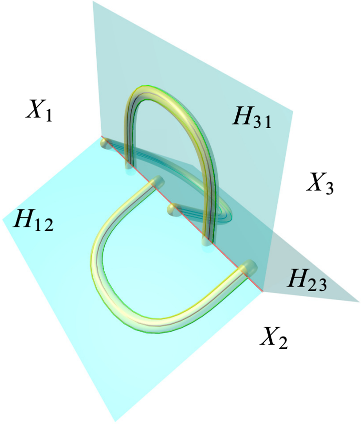


Figure 2: Stabilizing a trisection in dimension 3

The operation of replacing (X_1, X_2, X_3) with (X_1', X_2', X_3') is called *stabilization*.

Since any two boundary parallel arcs in a handlebody are isotopic, it is clear that this operation does not depend on the choice of arcs or neighborhoods.

In terms of trisection diagrams we have:

Definition 9 Given a trisection diagram $(F_g, \alpha, \beta, \gamma)$, the trisection diagram $(F'_{g'} = F_{g+3}, \alpha', \beta', \gamma')$ obtained by connected summing $(F_g, \alpha, \beta, \gamma)$ with the diagram in Figure 3 is called the *stabilization* of $(F_g, \alpha, \beta, \gamma)$.

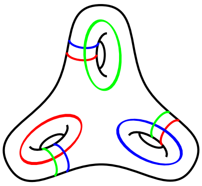


Figure 3: Stabilizing a trisection diagram means connected summing with this diagram. By itself, this describes the simplest nontrivial trisection of S^4, of genus 3. Red, blue and green indicate α, β and γ curves, respectively.

We prove the following fact at the beginning of Section 5:

Lemma 10 *If (X_1, X_2, X_3) is a genus-g trisection of X^4 with diagram $(F_g, \alpha, \beta, \gamma)$, and (X_1', X_2', X_3') is a stabilization of (X_1, X_2, X_3), then (X_1', X_2', X_3') is also a trisection of X, with genus $g' = g + 3$ and diagram $(F_{g'}, \alpha', \beta', \gamma')$, the stabilization of $(F_g, \alpha, \beta, \gamma)$.*

The reader may find Figure 2 useful in proving this lemma before reading our proof.

Theorem 11 (uniqueness) *Given two trisections (X_1, X_2, X_3) and (X_1', X_2', X_3') of X, after stabilizing each trisection some number of times there is a diffeomorphism $h\colon X \to X$ isotopic to the identity with the property that $h(X_i) = X_i'$ for each i. In particular, $h(X_i \cap X_j) = X_i' \cap X_j'$ for $i \neq j$ in $\{1, 2, 3\}$, and $h(X_1 \cap X_2 \cap X_3) = h(X_1' \cap X_2' \cap X_3')$.*

Corollary 12 *Given trisection diagrams $(F_g, \alpha, \beta, \gamma)$ and $(F_{g'}, \alpha', \beta', \gamma')$, the corresponding 4–manifolds $X(F_g, \alpha, \beta, \gamma)$ and $X(F_{g'}, \alpha', \beta', \gamma')$ are diffeomorphic if and only if $(F_g, \alpha, \beta, \gamma)$ and $(F_{g'}, \alpha', \beta', \gamma')$ are related by stabilization, handle slides, and diffeomorphism.* (Handle slides are slides of αs over αs, βs over βs and γs over γs.)

Proof Any two handle decompositions of a fixed genus-g handlebody, each with one 0–handle and g 1–handles, are related by handle slides; this is proved in Johannsson [5]. □

2 Discussion and examples

We begin with a few explicit examples of trisections and corresponding trisection diagrams.

- $S^4 \subset \mathbb{C} \times \mathbb{R}^3$ can be explicitly divided into three pieces

$$X_j = \{(re^{i\theta}, x_3, x_4, x_5) \mid 2\pi j/3 \leq \theta \leq 2\pi(j+1)/3\},$$

giving a genus-0 trisection of S^4. The diagram is S^2 with no curves.

- Stabilizing the genus-0 trisection of S^4 gives a genus-3 trisection, with trisection diagram shown in Figure 3. Since it is not known if the mapping class group of S^4 is trivial, we cannot say that the diagram determines the trisection up to isotopy, but the original description of stabilization of trisections (as opposed to stabilization of trisection diagrams) does determine this trisection up to isotopy, and thus we call this the *standard genus-*3 *trisection of* S^4.

- There is an obvious *connected sum* operation on trisected 4–manifolds, obtained by removing standardly trisected balls from each manifold and gluing along the boundary spheres so as to match the trisections. Stabilization can then also be defined as performing a connected sum with S^4 with its standard genus-3 trisection.

- The standard toric picture of $\mathbb{C}P^2$ as a right triangle gives a natural trisection into three pieces X_1, X_2, X_3 as the inverse images under the moment map of the three pieces of the right triangle shown in Figure 4. These pieces are diffeomorphic to B^4

but they intersect along solid tori all meeting along a central fiber diffeomorphic to T^2, so that this is a genus-1 trisection of $\mathbb{C}P^2$. The trisection diagram shows a $(1,0)$–, a $(0,1)$– and a $(1,1)$– curve; this is because the normals to the edges of the moment polytope tell us the direction in the torus which collapses along that edge. Alternatively, this trisection can be seen simply as the 0–handle, 2–handle and 4–handle in the standard handle decomposition of $\mathbb{C}P^2$, and the $+1$ framing on the 2–handle can be seen in the $(1,1)$–curve.

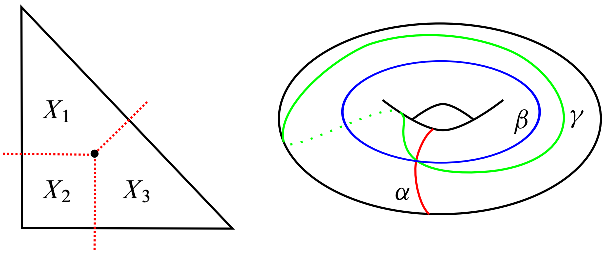


Figure 4: Trisection of $\mathbb{C}P^2$

- Reversing the orientation of the central surface in a trisection diagram reverses the orientation of the 4–manifold; ie $X(F_g,\alpha,\beta,\gamma) = -X(-F_g,\alpha,\beta,\gamma)$. Thus $\overline{\mathbb{C}P^2}$ has a genus-1 trisection, with trisection diagram given by a $(1,0)$–, $(0,1)$– and $(1,-1)$–curve.

- Looking at the standard toric picture of $S^2 \times S^2$ as a square also leads to a natural trisection of $S^2 \times S^2$ as follows: We divide the square into four regions labelled X_1, X_{2a}, X_{2b} and X_3 as indicated in Figure 5, and label the inverse images of these regions in $S^2 \times S^2$ with the same labels. Each of X_1, X_{2a}, X_{2b} and X_3 is a 4–ball,

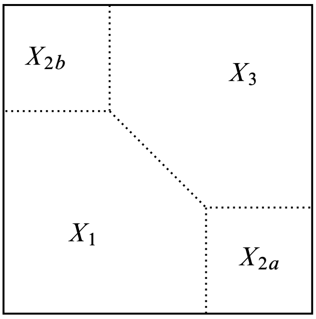


Figure 5: Trisection of $S^2 \times S^2$

and in fact they give the standard handle decomposition of $S^2 \times S^2$, with X_1 being the 0–handle, X_{2a} and X_{2b} being the 2–handles and X_3 being the 4–handle. Note that $X_1 \cap X_3$ is $T^2 \times [0,1]$, with $T^2 \times \{0\}$ being $X_1 \cap X_3 \cap X_{2a}$ and $T^2 \times \{1\}$ being $X_1 \cap X_3 \cap X_{2b}$. Let p be a point in T^2 and let a be the arc $\{p\} \times [0,1] \subset X_1 \cap X_3$. Remove a tubular neighborhood of this arc from X_1 and X_3 and add it as a tube joining X_{2a} to X_{2b}. The union of X_{2a} and X_{2b} with this tube is the X_2 of our trisection, and the new X_1 and X_3 are the results of removing the tube from the original X_1 and X_3. A little thought shows that this is a trisection with $k = 0$ (each piece is a 4–ball) and $g = 2$. (Thanks to Bob Edwards for giving us the initial picture that led to this description.)

- It may not be entirely obvious how to draw the trisection diagram for the above trisection of $S^2 \times S^2$. However, it is not hard to draw a genus-2 trisection diagram from scratch that does give $S^2 \times S^2$. In Figure 6 we show this diagram, as well as a diagram for $S^2 \tilde{\times} S^2$ and a diagram for $S^1 \times S^3$. We leave it to the reader to see how to relate these diagrams to the standard handle diagrams for these 4–manifolds. It is also an illuminating exercise, knowing that $S^2 \tilde{\times} S^2 \cong \mathbb{C}P^2 \sharp \overline{\mathbb{C}P^2}$, to verify Corollary 12in this case. The earlier discussion of connected sums and of $\pm\mathbb{C}P^2$ gives a trisection diagram for $\mathbb{C}P^2 \sharp \overline{\mathbb{C}P^2}$ and one checks that this is equivalent to that in Figure 6 for $S^2 \tilde{\times} S^2$ via handle slides and diffeomorphism of F_g. (It turns out that in this case we do not need stabilization.)

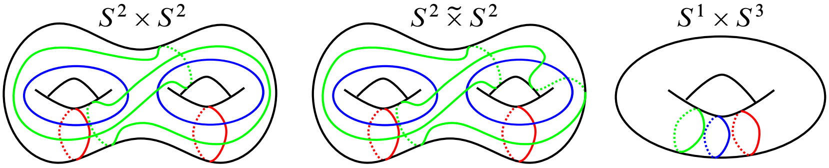


Figure 6: Various genus-2 trisection diagrams

Now we briefly discuss trisection diagrams more generally. Given a trisection diagram $(F_g, \alpha, \beta, \gamma)$, the 4–manifold $X(F_g, \alpha, \beta, \gamma)$ is constructed by attaching 4–dimensional 2–handles to $F_g \times D^2$ along $\alpha \times \{1\}$, $\beta \times \{e^{2\pi i/3}\}$ and $\gamma \times \{e^{4\pi i/3}\}$, with framings coming from $F_g \times \{p\}$, and the remainder of X is 3– and 4–handles. Recall that there is a unique way, up to diffeomorphism, to attach the 3– and 4–handles [6].

Since each of (F_g, α, β), (F_g, β, γ), (F_g, γ, α) is a Heegaard diagram for $\sharp^k(S^1 \times S^2)$, each can, after a sequence of handle slides, be made to look like the standard genus-g Heegaard diagram of $\sharp^k(S^1 \times S^2)$ [12; 5]. However, there is no reason to expect that we can simultaneously arrange for all three pairs of sets of curves to be standard.

Figure 7 illustrates a general trisection diagram (except that only one γ curve is shown) where we have made the (F_g, α, β) standard, where α is red and β is blue; the reds and blues give the standard genus-g Heegaard diagram for $\sharp^k(S^1 \times S^2)$. The important point is that most of the information about the 4–manifold X is then carried by the γ curves (one of which is drawn here in green). These green curves can be drawn anywhere with the proviso that some sequence of handle slides of the greens amongst the greens and the reds amongst reds, followed by a diffeomorphism of F_g, can make the reds and greens look like the reds and blues. The same proviso holds for the greens and blues, but a different sequence of handle slides and a different diffeomorphism may be required.

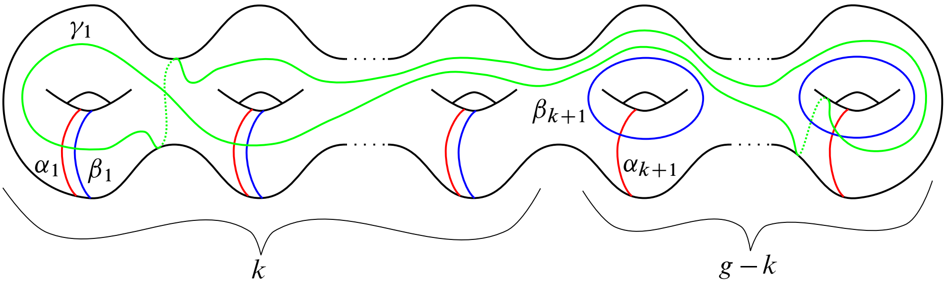


Figure 7: A general trisection diagram; only one γ curve is drawn, although there should be g of them.

In fact, if a trisection diagram is drawn so that αs and βs are standard as in Figure 7, then a framed link diagram for $X(F_g, \alpha, \beta, \gamma)$ is obtained by erasing the last $(g-k)$ αs and βs (which appear as meridian–longitude pairs) and then replacing each of the first k parallel pairs of αs and βs by a parallel dotted circle (1–handle) pushed slightly out of F_g. The γs remain as the attaching maps for 2–handles, and their framings come from the surface F_g.

An extended example: 3–manifold bundles over S^1 (Thanks to Stefano Vidussi for asking interesting questions that led to this example.) Suppose X^4 fibers over S^1, $M \hookrightarrow X \to S^1$, with fiber a closed, connected, oriented 3–manifold M^3, and monodromy $\mu\colon M \to M$.

A trisection of X is not immediately obvious, just as a bisection (Heegaard splitting) is not immediate when a 3–manifold fibers over a circle: $F_g \hookrightarrow M \to S^1$.

In the latter case, one takes two fibers over distinct points of S^1, separating M into two copies of $I \times F$. Choose a Morse function on F with one critical point of index 2 and thus one 2–handle H. Remove $I \times H$ from one $I \times F$ and add it to the other copy of $I \times F$. This turns the first copy into a handle body with $2g$ 1–handles, and adds a 1–handle to the second copy. Again let H be the 2–handle of the second copy (disjoint from the first H), and remove $I \times H$ from the second copy of $I \times F$ and add it to the first copy. Now both copies are handle bodies with $2g+1$ 1–handles and we have the desired Heegaard splitting.

In the 4–dimensional case, $X^4 = S^1 \times_\mu M$, pick a Morse function $\tau_0\colon M \to [0,3]$ with only one critical point $\hat{x}$ of index 3 and only one $\bar{x}$ of index 0 (τ_0 could give a minimal genus Heegaard splitting if desired).

Then $\tau_0\mu$ is another Morse function on M with the same kind of critical points, and μ can be isotoped so as to fix the maximum $\hat{x}$ and the minimum $\bar{x}$. Then there is a homotopy $\tau_t\colon M \to [0,3]$, $t \in [0,1]$, such that

(1) $\tau_1 = \tau_0\mu$,

(2) $\tau_t = \tau_0 = \tau_0\mu$ on $\hat{x}$ and $\bar{x}$ and there are no other definite critical points of τ_t,

(3) τ_t is a Morse function for all but a finite number of values of t at which τ_t has a birth or a death of a cancelling pair of indefinite critical points.

Since $S^1 = [0,1]/0 \sim 1$, property (1) allows us to define

$$\tau\colon X^4 = ([0,1] \times M)/(1,x) \sim (0,\mu(x)) \to S^1 \times [0,3]$$

by setting $\tau(t,x) = (t, \tau_t(x))$. To check, note that

$$\tau(1,x) = (1,\tau_1(x)) = (0,\tau_0(\mu(x))) = \tau(0,\mu(x)).$$

Thus we have a smoothly varying family of Morse functions on the fibers of X, except for the births and deaths. There are an equal number of births and deaths because τ_0 and $\tau_0\mu$ have the same number of critical points. Then we can make all the births happen earlier at $t=0$ and the deaths later at $t=1$, and furthermore by an isotopy of μ, the births and deaths can be paired off and happen at the same points of M. In that case the pairs can be merged and then τ is a family of Morse functions of the fibers of X with only one fixed maximum and minimum and g critical points of indices 1 and 2. Furthermore, it is straightforward to arrange that all critical points of index 1 (resp. 2) take values in a small neighborhood of 1 (resp. 2) for each $t \in S^1$.

Now draw a hexagonal-like grid on $[0,1] \times [0,3]$ as in Figure 8 and label the boxes with $X_i, i = 1,2,3$. Recall that the left and right ends are identified so as to have $S^1 \times [0,3]$.

The trisection of X into $X_1 \cup X_2 \cup X_3$ is to be made by tube-connect summing the preimages under τ of the X_i in Figure 8. Over each vertical line segment in Figure 8 is H_g which is defined to be a 3–dimensional handle body with g 1–handles, so over the interior vertices lie surfaces F_g. Over the diagonally sloped line segments lie 3–manifolds $I \times F_g$.

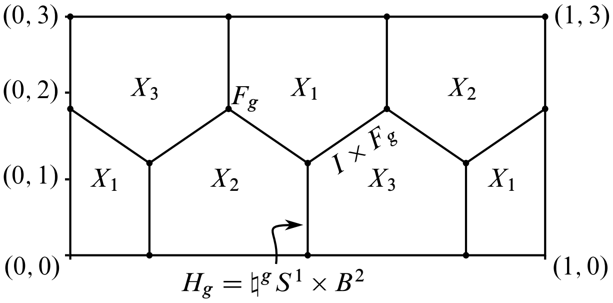


Figure 8: Fibering over S^1

Let H be the 2–handle in F_g and define a 4–dimensional 1–handle to be a thickening of $I \times H$ into the bounding X_i on either side of $I \times Fg$. Add such a 1–handle to connect each X_i to another X_i across a sloping line segment, for $i = 1, 2, 3$. Doing this twice for X_1, once along a SW-NE sloping line and once along a NW-SE sloping one as in Figure 9, we see that X_1 has become connected and is a 4–dimensional handlebody with $2g + 1$ 1–handles. Similarly with X_2 and X_3.

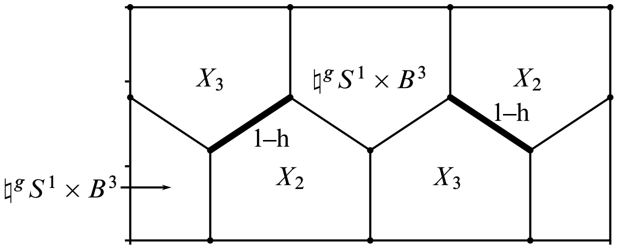


Figure 9: Connect the regions with 1–handles; here the 1–handles connecting the X_1 s are highlighted.

Next we calculate $X_1 \cap X_2$. Its various parts are shown in Figure 10. Note that the sloping edges with labels H_{2g} arise from $I \times F_g$ by having removed the $I \times H$. Thus we have $H_g \cup H_{2g} \cup H_g \cup H_{2g} \cup 4$ 1–handles, and three of the 1–handles cancel 0–handles leaving $H_{6g+1} = X_1 \cap X_2 = X_2 \cap X_3 = X_3 \cap X_1$. Then the central fiber $F_{g'}$ of the trisection has genus $g' = 6g + 1$ and gives a Heegaard splitting of $\partial X_i = \#_{2g+1} S^1 \times S^2$. Note that $k = 2g + 1$ and we can check that $\chi(X) = 0 = 2 + g' - 3k$.

(The referee for this paper pointed out an alternative, perhaps simpler, construction: By the Reidemeister–Singer theorem for 3–manifolds, there is a Heegaard splitting of M which is invariant under the monodromy μ. Then, by splitting the base S^1 into two intervals, we split X into four pieces, each a 3–dimensional handlebody crossed with an interval, or, in other words, a 4–dimensional 1–handlebody. Tubing two of

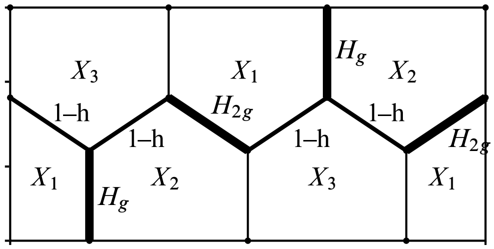


Figure 10: Understanding the pairwise intersections when fibering over S^1

these together as in Figure 5 produces a trisection, which may need a few more tubes to get the same k in each piece.)

Surface bundles over S^2 Now suppose that X^4 fibers over S^2 with fiber F a closed surface of genus g_F. We construct a trisection in a similar fashion to the preceding example.

Let $\pi\colon X \to S^2$ be the fibration. Identify S^2 with a cube and trisect S^2 as $S^2 = A_1 \cup A_2 \cup A_3$, where each A_i is the union of two opposite (closed) faces of the cube. Choose disjoint sections σ_1, σ_2 and σ_3 over A_1, A_2 and A_3, respectively, and let N_i be a closed tubular neighborhood of σ_i, for $i = 1, 2, 3$, with the N_i also disjoint. The trisection of X is $X = X_1 \cup X_2 \cup X_3$ where

$$X_i = (\pi^{-1}(A_i) \setminus \mathring{N}_i) \cup N_{i+1},$$

with indices taken mod 3.

We now verify that this is indeed a trisection, and compute g and k along the way. First, $\pi^{-1}(A_i)$ is two copies of $D^2 \times F$. Next, removing $\mathring{N}_i$ leaves us with two copies of $D^2 \times F'$, where F' has genus g_F and one boundary component. Thus $\pi^{-1}(A_i) \setminus \mathring{N}_i$ has two 0–handles and $4g_F$ 1–handles. Finally, N_{i+1} is two 1–handles connecting the two components of $\pi^{-1}(A_i) \setminus \mathring{N}_i$. Thus one of the 0–handles is cancelled by one of these two 1–handles, and we are left with one 0–handle and $k = 4g_F + 1$ 1–handles.

Now we consider the pairwise intersections. The 3–dimensional intersection $X_1 \cap X_2$ is the union of four pieces:

- $(\pi^{-1}(A_1) \setminus \mathring{N}_1) \cap (\pi^{-1}(A_2) \setminus \mathring{N}_2)$: Since A_1 and A_2 intersect along four edges of the cube, this is four copies of $[0,1] \times F''$, where F'' has genus g_F and *two* boundary components. In other words, this 3–manifold is built from four 0–handles and $4(2g_F + 1) = 8g_F + 4$ 1–handles.
- $(\pi^{-1}(A_1) \setminus \mathring{N}_1) \cap N_3$: This sits over the four edges making up $A_1 \cap A_3$, and thus contributes four 1–handles, two connecting two of the components above,

and two connecting the other two. Cancelling two of the 0–handles from the preceding step with two of these 1–handles, we are left with two 0–handles and $8g_F + 6$ 1–handles.

- $N_2 \cap (\pi^{-1}(A_2) \setminus \mathring{N}_2)$: This is just ∂N_2, which is two copies of $D^2 \times S^1$, joining up the four copies of $[0,1] \times F''$, from the first step above, in pairs, with the S^1 factor in $D^2 \times S^1$ lining up with one of the boundary components of the F'' factor of $[0,1] \times F''$. Thus we get two new 1–handles and two new 2–handles. One of the 1–handles cancels a 0–handle, and both 2–handles cancel 1–handles. This leaves us with one 0–handle and $8g_F + 6 + 1 - 2 = 8g_F + 5$ 1–handles.
- $N_2 \cap N_3$: This is empty.

Thus $X_1 \cap X_2$ is a 3–dimensional handlebody with genus $g = 8g_F + 5$, and the same holds for $X_2 \cap X_3$ and $X_3 \cap X_1$.

The triple intersection is necessarily the boundary of each pairwise intersection, so we see that we have a trisection with $k = 4g_F + 1$ and $g = 8g_F + 5$. This gives $\chi = 2 + g - 3k = 4 - 4g_F$, which is what we expect for a genus-g_F bundle over S^2.

When this technique is applied to $S^2 \times S^2$ we get the genus-5 diagram in Figure 11. With some work this can be shown to be handle slide and diffeomorphism equivalent to a single stabilization of the genus-2 diagram of $S^2 \times S^2$ in Figure 6.

Gluing maps A 4–manifold X with a trisection (X_1, X_2, X_3) is determined *up to diffeomorphism* by the data of k, g and three gluing maps between the sectors; see Figure 12. Here we discuss this gluing data carefully and show how to reduce the data

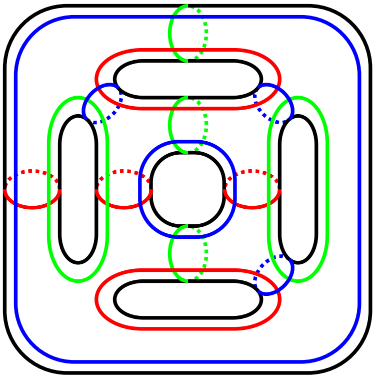


Figure 11: A genus-5 trisection diagram for $S^2 \times S^2$ obtained by seeing $S^2 \times S^2$ as an S^2 bundle over a cube. The surface shown here is naturally the boundary of a tubular neighborhood of the 1–skeleton of a cube.

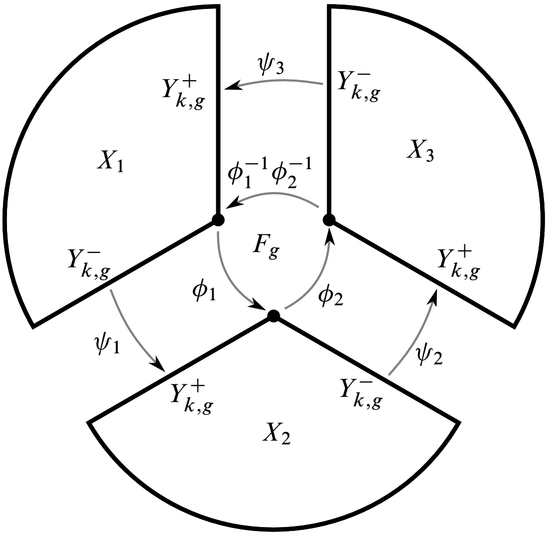


Figure 12: Gluing maps

to two elements of the mapping class group of a closed genus-g surface satisfying certain constraints.

Let X_1, X_2 and X_3 be copies of $Z_k = \natural^k(S^1 \times B^3)$. Let $Y_k = \partial Z_k = Y^+_{k,g} \cup Y^-_{k,g}$ be the standard genus-g Heegaard splitting of $Y_k = \sharp^k(S^1 \times S^2)$ with $H_{k,g} = Y^+_{k,g} \cap Y^-_{k,g}$ the Heegaard surface, with a fixed identification $H_{k,g} \cong F_g$. We can then construct a 4–manifold with three diffeomorphisms $\psi_i \colon Y^-_{k,g} \to -Y^+_{k,g}$, for $i = 1, 2, 3$, such that ψ_i glues X_i to X_{i+1} (indices taken mod. 3) by gluing the copy of $Y^-_{k,g}$ in ∂X_i to the copy of $Y^+_{k,g}$ in ∂X_{i+1}. Let $\phi_i = \psi_i|_{F_g} \colon F_g \to F_g$ and note that we need $\phi_3 \circ \phi_2 \circ \phi_1$ to be isotopic to the identity in order for the resulting manifold to close at the central fiber F_g. Furthermore, since an automorphism of a 3–dimensional handlebody is completely determined up to isotopy by its restriction to the boundary surface, this entire construction is actually determined by the two (isotopy classes of) maps $\phi_1, \phi_2 \colon F_g \to F_g$, with $\phi_3 = \phi_1^{-1} \circ \phi_2^{-1}$.

However, this characterization is slightly misleading because an arbitrary pair ϕ_1, ϕ_2 of mapping classes of F_g does not necessarily produce a trisected 4–manifold: we need that each of ϕ_1, ϕ_2 and $\phi_1^{-1} \circ \phi_2^{-1}$ extends to a diffeomorphism $\psi_i \colon Y^-_{k,g} \to -Y^+_{k,g}$, a slightly messy condition that is not entirely trivial to check.

Gluing maps from model manifolds In fact we can reduce the gluing map data to a single gluing map if we construct trisected 4–manifolds by cutting open and regluing fixed model trisected manifolds. For each $0 \leq k \leq g$ let

$$X^{k,g} = (\natural^k S^1 \times S^3) \natural (\natural^{g-k} \mathbb{C}P^2).$$

Note that $X^{k,g}$ has a standard (k,g)–trisection $X^{k,g} = (X_1^{k,g}, X_2^{k,g}, X_3^{k,g})$, because $S^1 \times S^3$ has a standard $(1,1)$–trisection and $\mathbb{C}P^2$ has a standard $(0,1)$–trisection. Also, for each such (k,g), fix an identification of $X_1^{k,g} \cap X_2^{k,g}$ with the standard genus-g handlebody $H_g = \natural^g S^1 \times B^2$. Then any other 4–manifold X with a (k,g)–trisection is obtained from $X^{k,g}$ by cutting $X_1^{k,g}$, $X_2^{k,g}$ and $X_3^{k,g}$ apart, regluing $X_1^{k,g}$ to $X_2^{k,g}$ by some automorphism ϕ of $X_1^{k,g} \cap X_2^{k,g} = H_g$, and then observing that gluing in $X_3^{k,g}$ amounts to attaching a collection of 3–handles and a 4–handle, so that no other gluing data needs to be specified. Again, not any automorphism $\phi\colon H_g \to H_g$ will work, but now one needs to verify that $\partial(X_1^{k,g} \cup_\phi X_2^{k,g})$ is diffeomorphic to $\natural^k(S^1 \times S^2)$ in order to verify that ϕ actually produces a closed trisected 4–manifold.

Lagrangians, Maslov index, signature and intersection triples Given a genus-g trisection diagram $(F_g, \alpha, \beta, \gamma)$, one can write down a triple $(Q_{\alpha\beta}, Q_{\beta\gamma}, Q_{\gamma\alpha})$ of $g \times g$ integer matrices, giving the intersection pairing between curves. Our uniqueness theorem tells us that this *intersection triple* is uniquely determined by the diffeomorphism type of $X(F_g, \alpha, \beta, \gamma)$ up to elementary row-column operations and stabilization. Here, the row-column operations are precisely those corresponding to handle slides. Thus, for example, sliding α_1 over α_2 corresponds to adding row 2 to row 1 in $Q_{\alpha\beta}$ while *simultaneously* adding column 2 to column 1 in $Q_{\gamma\alpha}$. Stabilization replaces $(Q_{\alpha\beta}, Q_{\beta\gamma}, Q_{\gamma\alpha})$ with the following triple:

$$\left(\left[\begin{array}{c|ccc} Q_{\alpha\beta} & & 0 & \\ \hline & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ & 0 & 0 & 0 \end{array} \right], \left[\begin{array}{c|ccc} Q_{\beta\gamma} & & 0 & \\ \hline & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ & 0 & 0 & 1 \end{array} \right], \left[\begin{array}{c|ccc} Q_{\gamma\alpha} & & 0 & \\ \hline & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ & 0 & 0 & 1 \end{array} \right] \right)$$

The fact that each pair of collections of curves gives a Heegaard diagram for $\natural^k S^1 \times S^2$ tells us that each of the three matrices is, independently, row-column equivalent to $\left[\begin{smallmatrix} 0_k & 0 \\ 0 & I_{g-k} \end{smallmatrix}\right]$. We thus have an invariant of 4–manifolds taking values in this set of triples, subject to this $\natural^k S^1 \times S^2$ condition, modulo an interesting equivalence relation. Of course, this invariant may contain nothing more than homological information, for example, but even if that were true it would be interesting to understand exactly how this works.

Alternatively, one can define three Lagrangian subspaces $(L_\alpha, L_\beta, L_\gamma)$ in the symplectic vector space $V = H_1(F_g; \mathbb{R})$; ie L_α is the kernel of the map $H_1(F_g; \mathbb{R}) \to H_1(H_\alpha; \mathbb{R})$ where H_α is the handlebody determined by the α curves, and so on.

One immediately recovers the three intersection matrices above as the symplectic form on V restricted to each pair of Lagrangians, relative to chosen bases on the Lagrangians. Thus our uniqueness theorem also gives us a 4–manifold invariant taking values in the set of quadruples $(V, L_\alpha, L_\beta, L_\gamma)$, where V is a symplectic vector space and the Ls are linear Lagrangian subspaces, subject again to the $\sharp^k S^1 \times S^2$ condition, modulo an equivalence relation. This equivalence relation is linear symplectomorphism and stabilization, which in this case means taking the direct sum with $(\mathbb{R}^6, \langle x_1, x_2, y_3\rangle, \langle x_1, y_2, x_3\rangle, \langle y_1, x_2, x_3\rangle)$.

This Lagrangian setup has in fact been studied in the more general context of Wall's [13] nonadditivity of the signature. A direct application of the interpretation in [2] of Wall's nonadditivity result shows that the signature of a closed 4–manifold with a trisection is precisely the Maslov index of this associated triple of Lagrangians.

However, one expects more information to be encoded in these Langrangians triples than just the signature. In particular, the Maslov index ignores the integer lattice structure of $H_1(F_g, \mathbb{Z}) \subset H_1(F_g, \mathbb{R})$. Quotienting out by this lattice gives us a triple of Lagrangian g–tori in a symplectic $2g$–torus, and one again gets a 4–manifold invariant taking values in these triples mod symplectomorphism and stabilization. It seems that a further study of this setup could be fruitful.

Curve complex perspective To record much more data than simply the homology classes of curves which bound disks in the three handlebodies, we can consider, for each handlebody H_{12}, H_{23} and H_{31}, the subsets U_{12}, U_{23} and U_{31}, respectively of the curve complex for F_g given by those essential simple closed curves which bound disks in the respective handlebody. Because each pair of handlebodies gives $\sharp^k(S^1 \times S^2)$, we know that the three intersections $U_{12} \cap U_{23}$, $U_{23} \cap U_{31}$ and $U_{31} \cap U_{12}$ are nonempty. This perspective raises many interesting questions, such as: What is the minimal area of a triangle with vertices in the three intersections? If $U_{12} \cap U_{23} \cap U_{31}$ is nonempty, what does that tell us about X? If the gluing map coming from the model manifold construction described above is, for example, pseudo-Anosov, does this tell us that the three subcomplexes are "far apart" in any sense?

3 Existence via Morse 2–functions

The proof presented in this section is an application of tools developed in [4], using Morse 2–functions. In the following section we will rewrite the proof entirely in terms of ordinary Morse functions and handle decompositions, but the trisection is so natural from the point of view of Morse 2–functions that we feel this proof is worth presenting. However, to give the basic idea for those most comfortable with the language of handle decompositions, our construction ends up putting the 0– and 1–handles of X into X_1, the 3– and 4–handles into X_3, and the 2–handles together with some "connective tissue" into X_2.

A Morse 2–function is a smooth, stable map $G\colon X^n \to \Sigma^2$; in this paper we will always map to $\mathbb{R}^2$. (Stable implies generic when mapping to dimension two.) Just like Morse functions, Morse 2–functions can be characterized by local models, and we now give these local models only in the case of $n = 4$, ie we are considering an $\mathbb{R}^2$–valued Morse 2–function G on a 4–manifold X:

(1) Each regular value $q \in \mathbb{R}^2$ has a coordinate neighborhood over which G looks like $F^2 \times B^2 \to B^2$ for some closed fiber surface F.

(2) The set of critical points of G is a smooth 1–dimensional submanifold $\mathrm{Crit}_G \subset X$ such that $G\colon \mathrm{Crit}_G \to \mathbb{R}^2$ is an immersion with isolated semicubical cusps and crossings. The noncusp points of Crit_G are called *fold points*, and arcs of such points are called *folds*.

(3) Each point $q \in G(\mathrm{Crit}_G)$ which is not a cusp or crossing has a neighborhood $U = I \times I$ with coordinates (t, y), with $G^{-1}(U)$ diffeomorphic to $I \times M^3$ for a 3–dimensional cobordism M, so that $G(t, p) = (t, g(p))$, where $g\colon M \to I$ is a Morse function on M with one critical point. The index of this critical point is then called the index of the fold, although this is only well-defined up to $i \mapsto 3 - i$. When the image of the fold is co-oriented, the index is well-defined by insisting that the y–coordinate on $I \times I$ increases in the direction of this co-orientation.

(4) Each cusp point $q \in G(\mathrm{Crit}_G)$ has a neighborhood $U = I \times I$ with coordinates (t, y), with $G^{-1}(u) = I \times M^3$, so that $G(t, p) = (t, g_t(p))$, where g_t is a 1–parameter family of Morse functions on M with no critical points for $t = 0$ and a birth of a cancelling pair of critical points at $t = \frac{1}{2}$. In our examples, these two critical points will always be of index 1 and 2.

(5) Each crossing point $q \in G(\mathrm{Crit}_G)$ has a neighborhood $U = I \times I$ with coordinates (t, y), with $G^{-1}(u) = I \times M^3$, so that $G(t, p) = (t, g_t(p))$, where g_t is a 1–parameter family of Morse functions on M with two critical points for all t, such that the critical values cross at $t = \frac{1}{2}$. In our examples, these two critical points will never be of index 0 or 3.

The basic example of a Morse 2–function is $(t, p) \mapsto (t, g_t(p))$ for an arbitrary generic homotopy g_t between two given Morse functions $g_0, g_1\colon M^3 \to [0, 1]$, and the message of the above local models is that Morse 2–functions look locally like homotopies between Morse functions, but globally we may not have a preferred "time"

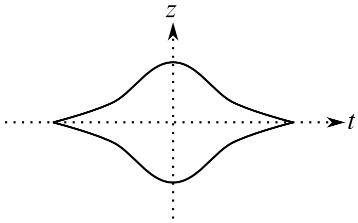


Figure 13: The "eye", a Cerf graphic in which a pair of cancelling critical points is born and then dies.

direction. When G is of the form $(t,p)\mapsto(t,g_t(p))$, we call $G(\mathrm{Crit}_G)$ a *Cerf graphic* [3]. Conversely, given a Morse 2–function $G\colon X^4\to\mathbb{R}^2$ and a rectangle $I\times I\subset\mathbb{R}^2$ in which $G(\mathrm{Crit}_G)$ has no vertical tangencies, we can find coordinates in which G is of this form $(t,p)\mapsto(t,g_t(p))$, and so again we will say that $G(\mathrm{Crit}_G)$ is a Cerf graphic in this rectangle.

There is one move on Morse 2–functions (ie local model for a generic homotopy between Morse 2–functions) that is central to this paper, which we call the "introduction of an eye". In a local chart in which a given Morse 2–function G on a 4–manifold has no critical points, we can assume G has the form $(t,x,y,z)\mapsto(t,x)$ or, equivalently, $(t,x,y,z)\mapsto(t,x^3+(t^2+1)x-y^2+z^2)$ with $t\in[-2,2]$. Introducing a parameter $s\in[-1,1]$ we get a homotopy $(t,x,y,z)\mapsto(t,x^3+(t^2-s)x-y^2+z^2)$, with $s=-1$ corresponding to the given map and $s=1$ the end result of "introducing an eye". Figure 13 shows the image of the critical locus at $s=1$, justifying the terminology. Note that this is a Cerf graphic in which, as t increases from -2 to 2, we see a Morse function on x,y,z space which starts with no critical points, develops a cancelling pair of index-1 and -2 critical points, and then the cancelling pair disappears again so that at $t=2$ there are again no critical points. Note also that the introduction of an eye takes place in a ball and is localized to a disk in the fiber cross a disk in the base; thus, as long as fibers are connected, we need only specify a disk in the base without critical points and then there is a unique, up to isotopy, way to introduce an eye in that disk.

Proof of Theorem 4 (existence) Throughout we will use coordinates (t,z) on $\mathbb{R}^2$, with t horizontal and z vertical. Here is an outline of the proof:

(1) First we will show that there is a Morse 2–function $G_1\colon X\to\mathbb{R}^2$ such that the image of the fold locus is as in Figure 14. In this and the following figures, three dots between two curves indicate that there are some number of parallel copies of the two curves in between. Fold indices are indicated with labelled transverse arrows. Boxes with folds coming in from the left and out at the right represent arbitrary Cerf graphics, with the left-right axis being time. Note that a Cerf graphic may contain left-cusps,

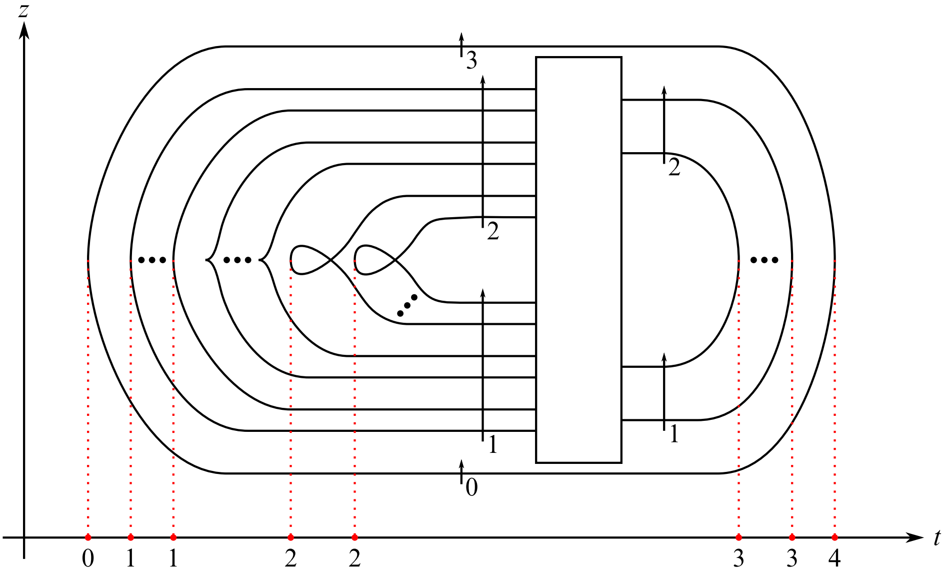


Figure 14: The image of the fold locus for G_1

right-cusps and crossings, but may not contain any vertical tangencies on the image of the fold locus.

(2) In Figure 14, the vertical tangencies of the folds are highlighted in red; these become critical points of the projection $t \circ G_1 \colon X \to \mathbb{R}$. These critical values in $\mathbb{R}$ are also indicated at the bottom of the diagram along the t–axis, with their indices.

(3) After constructing G_1, we will show how to homotope G_1 to G_2 such that the image of the fold locus for G_2 is as in Figure 15. Here the two Cerf graphics have no cusps. We have achieved two goals here: (1) Splitting the Cerf graphic into two, each involving only critical points of the same index and no cusps. (2) Replacing each kink that corresponds to an index-2 critical point of $t \circ G_1$ with a pair of cusps.

(4) Figure 16 is simply a redrawing of Figure 15 that highlights a natural trisection of $\mathbb{R}^2$ into three sectors $\mathbb{R}^2_1$, $\mathbb{R}^2_2$ and $\mathbb{R}^2_3$. Note that the critical locus over each sector consists of g components, where g is the genus of the central fiber. Also, each such component has at most one cusp. We no longer indicate the indices of the folds; the outermost fold is index-0 pointing inwards, and all other folds are index-1 pointing in.

(5) The form of the folds in Figure 16 is a special case of the form shown in Figure 17, where now we are not paying attention to which folds in a given sector, with or without cusps, connect to which folds in the next sector, with or without cusps, and we allow for arbitrary Cerf graphics (without cusps) between the sectors.

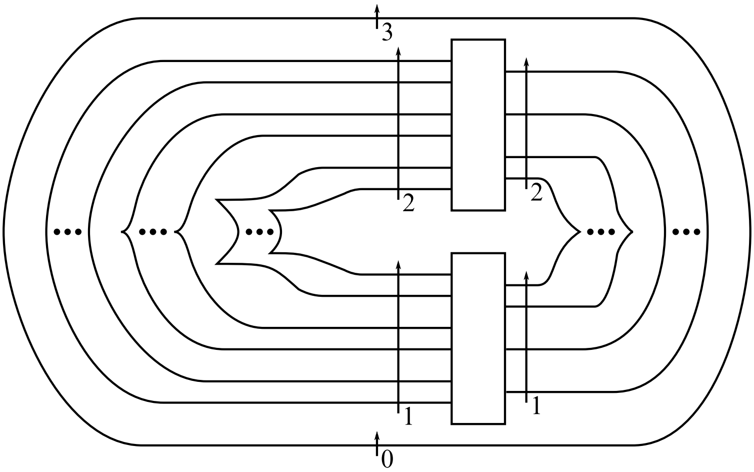


Figure 15: The image of the fold locus for G_2

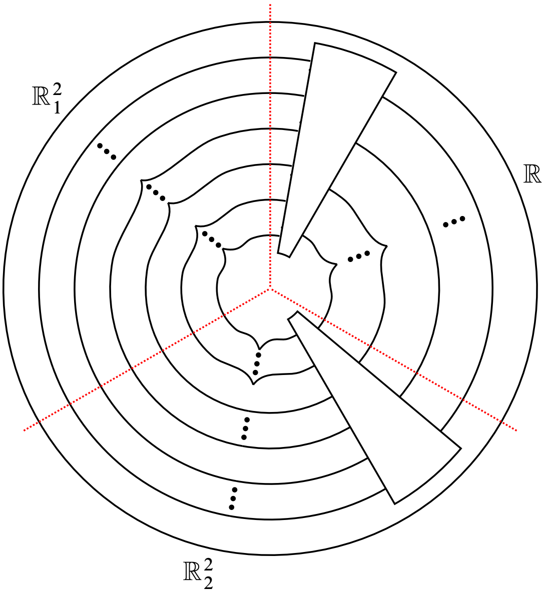


Figure 16: A more symmetric drawing of the image of the fold locus for G_2. We no longer indicate the indices of the folds; the outermost fold is index-0 going inwards, the others are index-1 going inwards.

(6) Now we have G_2 such that the image of the fold locus is as in Figure 17. At this point we could take $X_i = G_2^{-1}(\mathbb{R}^2_i)$ and we would have each X_i diffeomorphic to $\natural^{k_i} S^1 \times B^3$ for different k_i. There is one last step to arrange that the k_i are equal: In fact, k_i is equal to the number of folds in sector X_i *without cusps*. We will show how

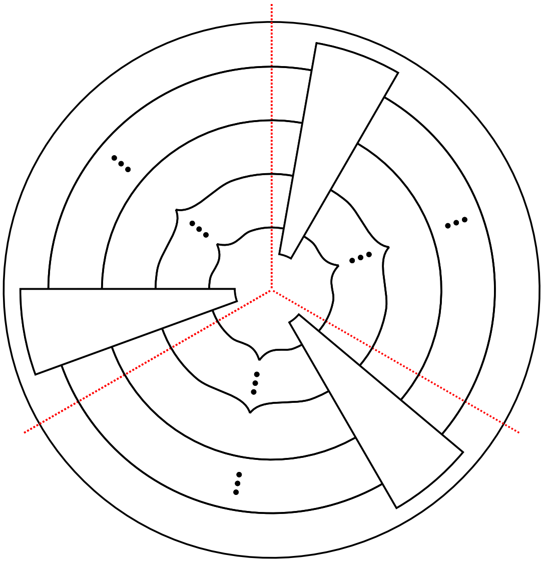


Figure 17: A slightly more general form for the image of the fold locus, which fits G_2.

to add a fold without a cusp to any one sector while adding a fold with a cusp to each of the other two sectors. This allows us to construct a homotopy from G_2 to G_3, such that G_3 has the image of its fold locus of the same form as G_2 (ie as in Figure 17), with the same number of folds without cusps in each sector, ie $k_1 = k_2 = k_3 = k$.

(7) Finally we will justify the claim that each $X_i = G_3^{-1}(\mathbb{R}^2_i)$ is diffeomorphic to $\natural^k S^1 \times B^3$ with overlap maps as advertised.

We now fill in the details.

Begin with a handle decomposition of X with one 0–handle, i_1 1–handles, i_2 2–handles, i_3 3–handles and one 4–handle. The union of the 0– and 1–handles, X_1 is diffeomorphic to $I \times (\natural^{i_1} S^1 \times B^2)$. Map this to $I \times I$ by $(t, p) \mapsto (t, g(p))$ where $g\colon \natural^{i_1} S^1 \times B^2 \to I$ is the standard Morse function with one index-0 critical point and i_1 index-1 critical points. Postcompose this map with a diffeomorphism from $I \times I$ to a half-disk and we have constructed G_1 on the union of the 0– and 1–handles so that the image of the fold locus is as in the right half of Figure 18.

Now note that $\partial X_1 = \#^{i_1}(S^1 \times S^2)$ sits over the right edge of the half disk in Figure 18 and that the vertical Morse function on ∂X_1, ie $z \circ G_1|_{\partial X_1}$ is the standard Morse function with i_1 index-1 critical points and i_1 index-2 critical points, inducing the standard genus-i_1 splitting of ∂X_1, with Heegaard surface F.

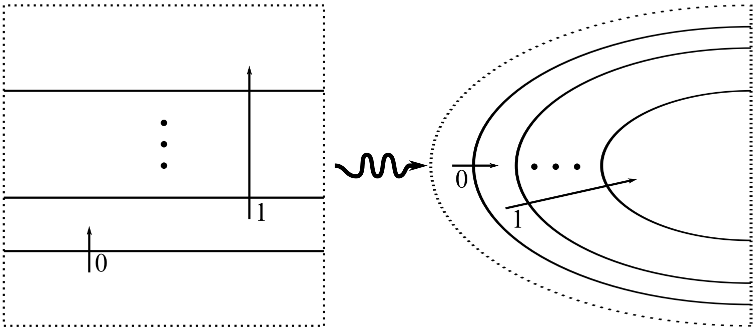


Figure 18: The first Morse 2–function, G_1, on the 0– and 1–handles of X.

Consider the framed attaching link $L \subset \partial X_1$ for the 2–handles of X. Generically L will be disjoint in ∂X_1 from the ascending 1–manifolds of the index-2 critical points of $z \circ G_1|_{\partial X_1}$ as well as the descending 1–manifolds of the index-1 critical points. Thus L can be projected onto the Heegaard surface F along gradient flow lines to give an immersed curve $\overline{L}$ in F with at worst double points. By adding kinks if necessary, we can assume that the handle framing of L agrees with the "blackboard framing" coming from $\overline{L} \subset F$. Then by stabilizing this Heegaard splitting once for each crossing of $\overline{L}$, we can resolve these crossings and get L to lie in the Heegaard surface with framing coming from the surface. This process translates into an extension of the thus-far constructed G_1 from X_1 to $X_1 \cup ([0,1] \times \partial X_1)$ with fold locus as in Figure 19, with one cusp for each stabilization. In other words, the sequence of stabilizations translates into a homotopy g_t from g_0, the standard Morse function on $\sharp^{i_1}(S^1 \times S^2)$, to g_1, the stabilized Morse function. This homotopy then becomes a Morse 2–function on the collar $[0,1] \times \partial X_1$.

Now let F refer to the stabilized Heegaard surface, in which L lies. Attaching a 4–dimensional 2–handle to X_1 along a component K of L is the same as attaching I times a 3–dimensional 2–handle to X_1 along $I \times K \subset I \times F \subset \partial X_1$. In Figure 20 we show the resulting Morse 2–function at the left, where the handle sits over a vertical rectangle. Next we bend this rectangle to make the image again a half-disk. Finally, noting that the vertical Morse function at the right edge now has an index-2 critical value below an index-1 critical value, we switch these values to get the Morse 2–function at the right side of Figure 20.

Note that everything in the preceding paragraph happened in a neighborhood of K, so that the rest of L still lies in the middle Heegaard surface for the Morse function at the right edge of the final diagram in Figure 20. Thus we can attach each 4–dimensional 2–handle this way to get the Morse 2–function at the left side of Figure 21. Each 2–handle of X corresponds to a kink in the image of the folds, ie a smoothly immersed

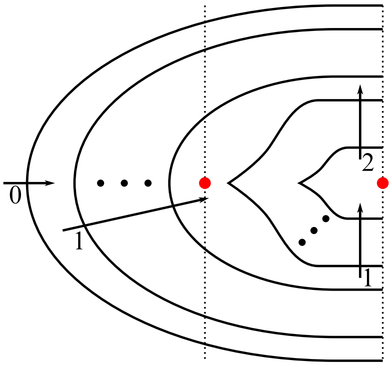


Figure 19: G_1 extended to a collar on ∂X_1. In the two vertical slices shown, both diffeomorphic to $\sharp^n(S^1 \times S^2)$, the Heegaard surface sits over the highlighted red points. The framed attaching link L for the 2–handles of X lies in the Heegaard surface for the right-most Morse function, ie over the right-most red point, with framing coming from the surface.

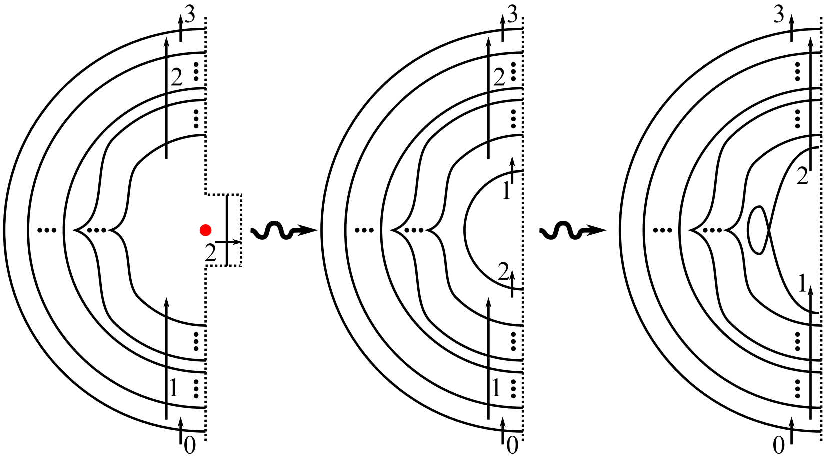


Figure 20: G_1 after attaching a 4–dimensional 2–handle

arc with a single transverse double point. Repeating our construction for X_1 with the union of the 3– and 4–handles, we construct the Morse 2–function at the right side of Figure 21. The two halves give vertical Morse functions on the boundary of the union of the 3– and 4–handles, which are related by some Cerf graphic. Putting this Cerf graphic in between the two parts of Figure 21 gives us G_1 as in Figure 14.

To get to Figure 15, first we take the Cerf graphic section of Figure 14 and pull the births (left-cusps) to the left of the Cerf graphic and the deaths (right-cusps) to the

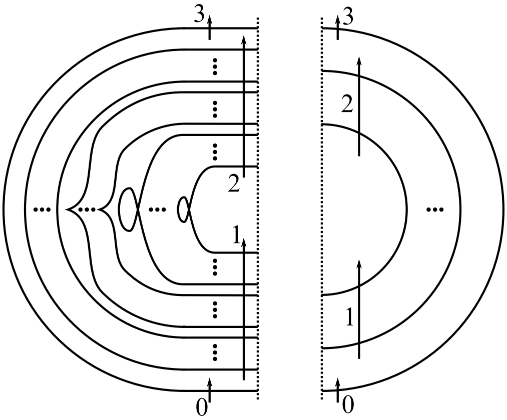


Figure 21: Two halves of G_1: the 0–, 1– and 2–handles on the left and the 3– and 4–handles on the right. Connecting them with a Cerf graphic gives Figure 14.

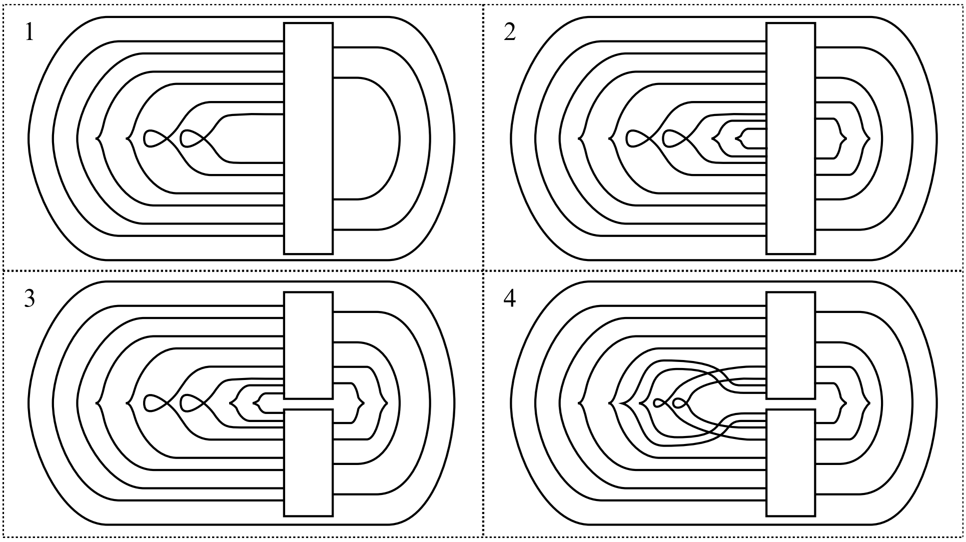


Figure 22: Pulling cusps out of the Cerf graphic. Here we suppress the "three dots" notation as well as the indices of the folds, as these are understood from earlier figures.

right, and then pull all index-1 critical points below all index-2 critical points. Then the left-cusps can be pulled further left, past the kinks which correspond to 4–dimensional 2–handles, because the 4–dimensional 2–handle attachments are independent of the 3–dimensional stabilizations corresponding to the cusps. This is shown in Figure 22. Next we homotope the kinks into pairs of cusps as in Figure 23. The first step of

Figure 23 introduces a swallowtail at the vertical tangency of the kink; this move has been discussed extensively elsewhere [7] and is a standard singularity that occurs in a homotopy between homotopies between Morse functions. The second step moves an arc of index-1 critical points in a homotopy (Cerf graphic) below an arc of index-2 critical points. This is also standard and is possible because the descending manifold for the index-1 point remains disjoint from the ascending manifold for the index-2 point throughout the homotopy. (Equivalently, in homotopies between Morse functions we never expect 1–handles to slide over 2–handles.)

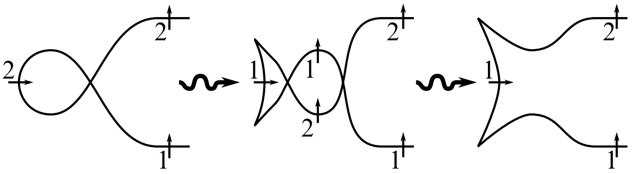


Figure 23: Turning kinks into pairs of cusps

Finally, Figure 24 shows how to add folds and cusps to a Morse 2–function as in Figure 17 so as to increase the number of folds without cusps in one of the three sectors. Here we are introducing an eye, as in Figure 13, modified by a slight isotopy. Note that the transition from the second to the third diagram in the figure is not essential, but only serves to put the resulting diagram in the form of Figure 17. Depending on how we orient the new eye with respect to the trisection of $\mathbb{R}^2$, we either add the fold without cusps to $\mathbb{R}^2_1$, $\mathbb{R}^2_2$, or $\mathbb{R}^2_3$.

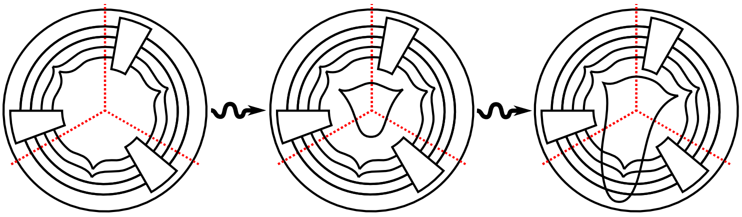


Figure 24: Adding an extra fold without cusps in one sector; again we suppress the "three dots" notation and the fold indices.

(Note that if we do this operation three times, once for each sector, we increase k by 1 and g by 3; this is precisely a stabilization of the trisection, as will be shown in Section 5.)

Now we need to show that, having put our Morse 2–function finally into the form of Figure 17, with k folds in each sector without cusps and $g-k$ folds with cusps, then for each i, $G^{-1}(\mathbb{R}^2_i) = X_i \cong \natural^k(S^1 \times B^3)$. However, we have already seen this: Each sector, ignoring the Cerf graphic block, looks just like Figure 19, which we already know is $\natural^k(S^1\times B^3)$ with a $(g-k)$–times stabilized standard Heegaard splitting on the boundary. The Cerf graphic block connecting one sector to another is a product which does not interfere with the Heegaard splitting. □

4 Trisections and handle decompositions

The techniques of the previous section lead to a relationship between trisections and handle decompositions equipped with certain extra data. We will use this relationship both to provide an alternate proof of Theorem 4 and to prove Theorem 11.

By a *system of compressing disks* for a 3–dimensional handlebody H of genus g, we mean a collection of properly embedded disks $D_1,\dots,D_g\subset H$ such that cutting H open along $D_1\cup\dots\cup D_g$ yields a 3–ball.

Lemma 13 *If $X=X_1\cup X_2\cup X_3$ is a trisection of a* 4*–manifold X, then there is a handle decomposition of X as in Theorem 4 satisfying the following properties:*

(1) *X_1 is the union of the* 0*– and* 1*–handles.*

(2) *Considering the Heegaard splitting $\partial X_1=H_{12}\cup H_{31}$ with Heegaard surface F, the attaching link L for the* 2*–handles lies in the interior of H_{12}.*

(3) *The framed attaching link $L=K_1\cup\dots\cup K_{g-k}$ is isotopic in H_{12} to a framed link $L'=K'_1\cup\dots\cup K'_{g-k}\subset F$, with framings equal to the framings induced by F.*

(4) *There is a system of compressing disks $D_1,\dots,D_g$ for H_{12} such that the curves $K'_1,\dots,K'_{g-k}$ are geometrically dual in F to the curves $\partial D_1,\dots,\partial D_{g-k}$. In other words, each K'_j intersects ∂D_j transversely once and is disjoint from all other ∂D_i.*

(5) *There is a tubular neighborhood $N=[-\epsilon,\epsilon]\times H_{12}$ of H_{12} with $[-\epsilon,0]\times H_{12}=N\cap X_1$, such that X_2 is the union of $[0,\epsilon]\times H_{12}$ with the* 2*–handles.*

Proof Each sector of the trisection of X is diffeomorphic to $\natural^k(S^1\times B^3)$ with a genus-g splitting of its boundary. Thus it has a standard Morse 2–function onto a wedge in R^2; see Figure 17. Two sectors meet at $X_i\cap X_{i+1}=\natural^k(S^1\times B^2)$, and the two Morse 2–functions on the two sectors give two Morse functions on the intersection $X_i\cap X_{i+1}$. The two Morse functions are homotopic and thus give a Cerf diagram which can be inserted into the little wedges in Figure 17. In the existence proof from the previous section we avoided cusps in the Cerf graphic boxes, but at this point we do not care; any Cerf graphic will do.

An isotopy of $\mathbb{R}^2$ makes the picture look like Figure 25. Now projection to the horizontal axis gives a Morse function in which the vertical tangencies become Morse critical points. X_1, to the left of the vertical red line, is clearly the union of the 0– and 1–handles. X_2, between the legs of the red letter h is then a handlebody H_{12}, cross I, with $g-k$ 2–handles attached. And X_3 is obviously what remains.

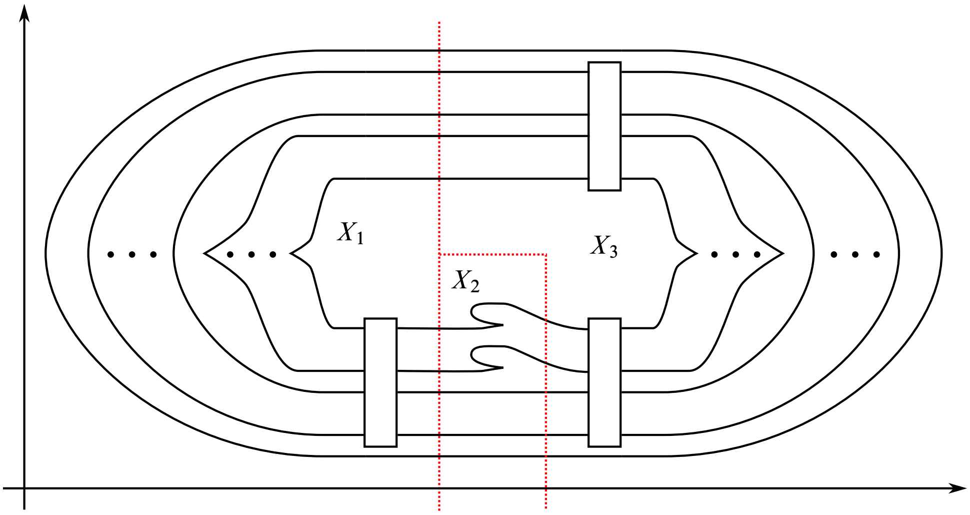


Figure 25: Extracting a handle decomposition from a trisection

We only need to show now that the attaching link for the 2–handles is as advertised. This can be seen from the fact that the attaching circle for each 2–handle, between the legs of the h, is one of a dual pair of curves on the fiber near a cusp. The other curve in the dual pair is the attaching curve for the fold that cuts across H_{12} and gives one of the compressing disks for this handlebody. This is illustrated in Figure 26, which shows a zoomed in region of Figure 25. The fiber over a specific point is drawn as a once punctured torus; this is just part of the fiber, but the rest of the fiber does not play a role in this local picture. The attaching circles for the two folds are drawn as green and blue circles on the fiber. This is just the usual picture of the fiber between the two arms of a cusp, with attaching circles being geometrically dual. Here, however, we reinterpret this picture to see the blue circle as the boundary of a compressing disk for the handlebody lying over the vertical dotted red line, and to see the green circle as the attaching circle for the 4–dimensional 2–handle coming from the vertical tangency in the fold. □

Lemma 14 *Consider a handle decomposition of a 4–manifold X^4 with one 0–handle, k 1–handles, $g-k$ 2–handles, k 3–handles and one 4–handle. Let X_1 be the union of*

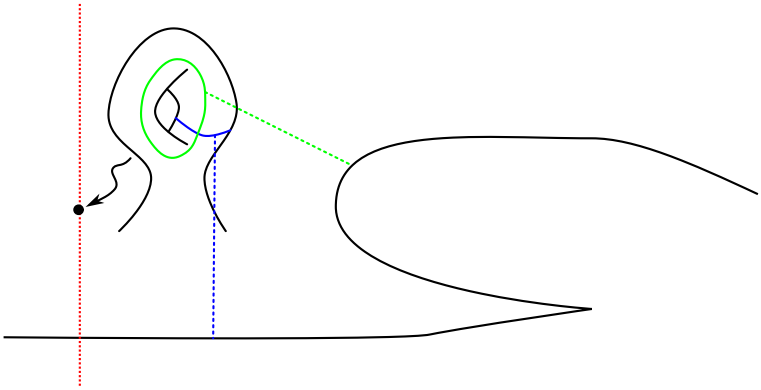


Figure 26: Zooming in on a region of Figure 25.

the 0– and 1–handles. Suppose there is a genus-g Heegaard splitting $\partial X_1 = H_{12} \cup H_{31}$ of ∂X_1 satisfying the following properties in relation to the framed attaching link L for the 2–handles:

1. *L lies in the interior of H_{12}.*
2. *L is isotopic in H_{12} to a framed link $L' \subset F$, with framing equal to the framing induced by F.*
3. *There is a system of compressing disks $D_1, \dots, D_g$ for H_{12} such that the $g-k$ components of L' are, respectively, geometrically dual in F to the curves $\partial D_1, \dots, \partial D_{g-k}$.*

Let $N = [-\epsilon, \epsilon] \times H_{12}$ be a small tubular neighborhood of H_{12} with $[-\epsilon, 0] \times H_{12} = N \cap X_1$, which the 2–handles intersect as $[0, \epsilon] \times \nu_L$, where ν_L is a tubular neighborhood of L in H_{12}. Declare X_2 to be the union of $[0, \epsilon] \times H_{12}$ with the 2–handles, and declare X_3 to be what remains (the closure of $X \setminus (X_1 \cup X_2)$). Then $X = X_1 \cup X_2 \cup X_3$ is a trisection.

Proof Almost everything we need for $X_1 \cup X_2 \cup X_3$ to be a trisection is immediate:

1. X_1 and X_3 are both diffeomorphic to $\natural^k (S^1 \times B^3)$.
2. $H_{31} = X_3 \cap X_1$ and $H_{12} = X_1 \cap X_2$ are genus-g handlebodies.
3. $F = X_1 \cap X_2 \cap X_3$ is a genus-g surface.

It remains to verify that $X_2 \cong \natural^k (S^1 \times B^3)$ and that $H_{23} = X_2 \cap X_3$ is a genus-g handlebody.

In fact X_2 is built by attaching $g-k$ 2–handles to $X_{12} \cong \natural^k (S^1 \times B^2)$ along $g-k$ copies of $S^1 \times \{0\} \subset S^1 \times B^2$ in the first $g-k$ $S^1 \times B^3$ summands. Thus the 2–handles "cancel" $g-k$ copies of $S^1 \times B^3$, giving both desired results immediately. □

Using Lemma 14, we now present a proof of the existence of trisections in the spirit of [8]:

Proof of Theorem 4 (existence) Start with a handle decomposition of X^4 with one 0–handle, k_1 1–handles, k_2 2–handles, k_3 3–handles and one 4–handle. Add cancelling 1–2 and 2–3 pairs if necessary so as to arrange that $k_1 = k_3$. Let X_1 be the union of the 0–handle and the 1–handles. Note that ∂X_1 is a connected sum of k_1 copies of $S^1 \times S^2$. Let $L \subset \partial X_1$ be the framed attaching link for the 2–handles.

Consider the genus-k_1 Heegaard splitting of ∂X_1 as $\partial X_1 = H_{12} \cup H_{31}$ with $F = H_{12} \cap H_{31}$. (We will soon be stabilizing this Heegaard splitting, but after each stabilization we will use the same names for the surface and the handlebodies.) The attaching link $L \subset \partial X_1$ can be projected onto the Heegaard surface F with transverse double points (crossings), so that the handle framing is the surface framing. (Add kinks to get the framing right.) Make sure that each component has at least one crossing using Reidemeister 2 moves if necessary. Let c be the number of crossings in this projection.

If $c \leq k_2$ then we are almost done. Stabilize the Heegaard splitting exactly k_2 times, with c of these stabilizations occuring at the crossings. Then L can be isotoped so as to resolve all the crossings by sending the over strand at each crossing over the new $S^1 \times S^1$ summand in F coming from the stabilization at that crossing. Now we have a genus-$g = k_1 + k_2$ Heegaard splitting. Letting $k = k_1$ and $g = k_1 + k_2$, and pushing L into the interior of H_{12}, we now satisfy the hypotheses of Lemma 14 and apply that lemma to produce our trisection. (We get duality to a system of meridians as follows: Each component K of L goes over at least one stabilization which no other components go over, and therefore is the unique component intersect the meridian for that stabilization. For every other meridian which K intersects, slide that meridian's compressing disk over the compressing disk corresponding to the stabilization singled out in the preceding sentence.)

If $c > k_2$ then add $c - k_2$ cancelling 1–2 pairs and $c - k_2$ cancelling 2–3 pairs to the original handle decomposition of X. Now we have $k_1' = k_1 + c - k_2$ 1–handles, and the same number of 3–handles, as well as $k_2' = 2c - k_2$ 2–handles. We consider the new $X_1' = X_1 \natural^{c-k_2} S^1 \times B^3$ with the natural genus-k_1' Heegaard splitting $\partial X_1' = H_{12}' \cup H_{31}'$ with $F' = H_{12}' \cap H_{31}'$. The original attaching link L still projects onto F' in the same way, with the same crossings, since F' is naturally $F \sharp^{c-k_2} S^1 \times S^1$.

However, we also have $2(c - k_2)$ new 2–handles. Half of these, coming from the 1–2 pairs, are attached along the meridians of the $c - k_2$ new $S^1 \times S^1$ summands in F' and thus immediately satisfy the conditions in Lemma 14. The other half, coming from the 2–3 pairs, are attached along 0–framed unknots, which project onto F' as circles bounding disks in F'.

Now stabilize the new Heegaard splitting $2c-k_2$ times: The first c of these stabilizations should happen at the crossings of L, allowing us to resolve crossings as before. The other $c-k_2$ of the stabilizations should occur next to the $c-k_2$ 0–framed unknots. Then each of these unknots is isotoped to go over the new $S^1 \times S^1$ summand coming from the adjacent stabilization. Now the entire attaching link satisfies the hypotheses of Lemma 14. The new genus of the stabilized Heegaard splitting of $\partial X_1'$ is $g' = k_1' + 2c - k_2$. To conclude the theorem by applying Lemma 14 we need that $k_2' = g' - k_1'$, and this is precisely what we have arranged. □

5 Uniqueness

We first prove that the stabilization operation of Definition 8 really does produce a new trisection. This can be done directly, but instead we will do so by showing that, from a Morse 2–function point of view, this stabilization corresponds to adding three eyes at the center of a trisected Morse 2–function. After that we can proceed with the proof of uniqueness.

Proof of Lemma 10 We are given a trisection (X_1, X_2, X_3) of X, with handlebodies $H_{ij} = X_i \cap X_j$, properly embedded arcs $A_{ij} \subset H_{ij}$, and regular neighborhoods of these arcs $N_{ij} \subset X$.

As we will see at the beginning of the proof of Theorem 11, it is easy to construct a Morse 2–function as in Figure 17 which recovers this trisection. We claim that adding three eyes arranged as in Figure 27 modifies each sector X_i exactly as in Definition 8, and since the new Morse 2–function again gives a trisection, then stabilization as defined in Definition 8 produces a trisection.

We see that the claim is true one eye at a time. Each time we add an eye, first add it away from the center straddling the intersection of two sectors, such as H_{31}, as on the left in Figure 28. We will then pull the lower fold across the central fiber to achieve the right-hand diagram in Figure 28. Up to isotopy, moving from the left to the right in this figure is the same as not moving the eye, but instead enlarging the lower sector X_2 by attaching the inverse image of the green region labelled N. This inverse image is in fact a 1–handle cobordism attached to X_2, since this fold is an index-1 fold going in towards the middle of the eye. Furthermore, the 1–handle is cancelled by a 2–handle immediately above it. The 1–handle and 2–handle are actually I cross 3–dimensional 1– and 2–handles, respectively, and thus we see that we have simply

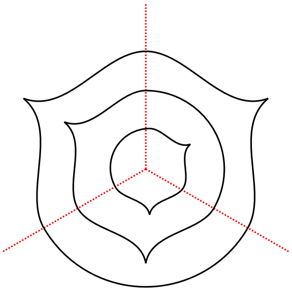


Figure 27: Stabilizing a Morse 2–function by adding three “eyes”

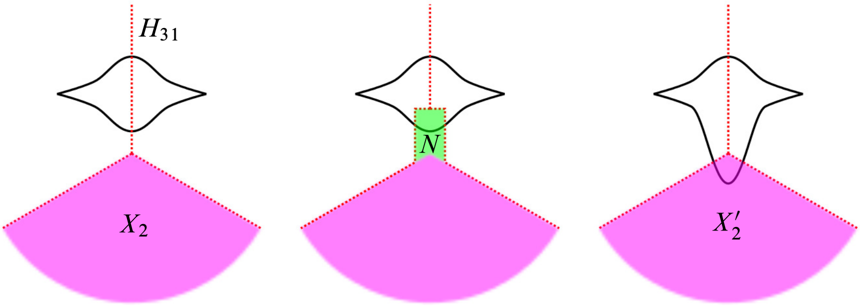


Figure 28: Adding one eye to a trisected Morse 2–function

removed a neighborhood of an arc in H_{31} from both X_1 and X_3 and added it to X_2. Repeat this for each of the three eyes. □

Proof of uniqueness, Theorem 11 Consider two trisections of the same 4–manifold: $X^4 = X_1 \cup X_2 \cup X_3 = X_1' \cup X_2' \cup X_3'$. Apply Lemma 13 to each trisection to get two handle decompositions D and D' of X, respectively, with corresponding Heegaard splittings of ∂X_1, with attaching links L and L' behaving as in Lemma 13. Cerf theory tells us that we can get from D to D' by the following operations:

(1) Add cancelling 1–2 and 2–3 pairs to both D and D'.

(2) Slide 1–handles over 1–handles, 2–handles over 2–handles and 3–handles over 3–handles.

(3) Isotope the handles and their attaching maps without sliding over any handles.

From the description of trisection stabilization in the proof of Lemma 10 above, we can see that trisection stabilization adds both a 1–2 pair and a 2–3 pair to an associated handle decomposition. Thus, after arranging that we add the same number of 1–2 pairs as 2–3 pairs, we can stabilize the two original trisections to take care of the first operation above.

Clearly sliding 1–handles over 1–handles and 3–handles over 3–handles, as well as isotoping 1–handles and 3–handles without handle slides, does not change the associated trisection.

Thus we are left to investigate the effect of 2–handle slides and 2–handle isotopies.

Suppose that we wish to perform a single 2–handle slide to the handle decomposition D. Associated to the trisection T which gives rise to D we have a Heegaard splitting $H_{12} \cup H_{31}$ for ∂X_1, with the attaching link L for the 2–handles of D lying in H_{12}. Isotope L into $\partial H_{12} = F$ so that the components of L are dual to the $g-k$ curves in a system of g meridinal curves (boundaries of compressing disks), as in Lemma 13. The handle slide involves a framed arc connecting two components K_1 and K_2 of L. This arc can be projected (following the flow of a Morse function of ∂X_1 for the given Heegaard splitting) onto F, but with crossings. We can arrange for its framing to agree with the surface framing with kinks, as usual. We want to avoid self-crossings as well as crossings between the arc and L and between the arc and the system of meridinal curves.

Stabilizing the Heegaard splitting, however, allows us to resolve the crossings. In other words, we get a new Heegaard splitting $\partial X_1 = H'_{12} \cup H'_{31}$ obtained from $H_{12} \cup H_{31}$ by Heegaard splitting stabilizations and isotopy such that L and the band lie in $\partial H'_{12} = F'$, still maintaining the property that the components of L are dual to the first $g-k$ meridinal curves in a system of meridinal curves of H'_{12}. In addition, the bands are disjoint from these $g-k$ meridinal curves. (Note that we can do this without moving L or the bands, but just by stabilizing and isotoping the Heegaard splitting.) Then sliding one component of L over another along the chosen band maintains this property; we have to change one of the meridinal curves in the system of compressing disks by a handle slide as well.

Again, from the proof of Lemma 10, we see that stabilization of the Heegaard splitting of ∂X_1 can be achieved by stabilizing the trisection, at the expense of introducing cancelling 1–2 and 2–3 pairs to the associated handle decomposition.

Thus we have shown that, if D and D' are related by handle slides supported in small neighborhoods of arcs in ∂X_1, then they are adapted to trisections related by trisection stabilization and isotopy.

Finally, suppose that D and D' are related only by an isotopy of the 2–handles and their attaching maps, without any handle slides. Then this isotopy extends to an isotopy of X with the result that we can assume that the handle decompositions are identical, and the only difference between the trisections is the Heegaard splitting of ∂X_1.

So we have two Heegaard splittings $\partial X_1 = H_{12} \cup H_{31} = H'_{12} \cup H'_{31}$, respectively, coming from T and T'. The fixed attaching link L for the 2–handles lies in both H_{12} and H'_{12}, in both cases satisfying the condition of being dual to meridinal curves.

Note that both $H_{12} \cup H_{31}$ and $H'_{12} \cup H'_{31}$ are genus-g Heegaard splittings of $\partial X_1 \cong \natural^k(S^1 \times S^2)$, so that Waldhausen's theorem [12] gives us an isotopy of ∂X_1 taking H_{12} to H'_{12}. However, this *does not* imply that the trisections T and T' are isotopic, because this isotopy will in general move the link L. If we can find an isotopy that does not move L, then we will be done, but first we will probably need to stabilize.

To see how to do this, construct two Morse functions f and f' on ∂X_1 with regular values $a < b$ such that

(1) f and f' agree on $f^{-1}(-\infty, a] = f'^{-1}(-\infty, a]$, which is a tubular neighborhood of L (thus each has $g-k$ index-0 critical points and $g-k$ index-1 critical points),

(2) $f^{-1}(-\infty, b] = H_{12}$,

(3) $f'^{-1}(-\infty, b] = H'_{12}$,

(4) f and f' have only critical values of index 1 in $[a, b]$ and critical values of index 2 and 3 in $[b, \infty)$.

Now Cerf theory gives us a homotopy f_t from $f_0 = f$ to $f_1 = f'$ which involves 1–2 births and deaths on $f^{-1}(b)$ and otherwise no critical values crossing b, and such that $f_t = f = f'$ on $f^{-1}(\infty, a]$. Thus, after stabilizing the Heegaard splittings away from L, there is an isotopy fixing L taking the one Heegaard splitting to the other.

Again, the Heegaard splitting stabilizations are achieved by trisection stabilizations. □

Remark 15 Morally it seems that there should be a Morse 2–function proof of uniqueness that starts with a generic homotopy between two Morse 2–functions corresponding to two given trisections. Then the proof would homotope this homotopy so as to arrange that the Cerf 2–graphic in $[0, 1] \times \mathbb{R}^2$, a surface of folds with cusps and higher codimension singularities, is in a nice position with respect to the standard trisection of $[0, 1] \times \mathbb{R}^2$. This surface of folds is, however, not trivial to work with. A good model might be the method of braid foliations used by Birman and Menasco to prove Markov's theorem in [1].

6 The relative case

When $\partial X \neq \varnothing$, we should define a trisection as the kind of subdivision of X which naturally arises from a Morse 2–function $G\colon X \to B^2$ where B^2 is trisected as in Figure 1, the locus of critical values behaves well with respect to this trisection of B^2, and the trisection of X is just G^{-1} of the three sectors of B^2. "Behaving well" should mean that the folds all have index 1 when transversely oriented towards the center of B^2, that the only tangencies to rays of B^2 are the cusps, that there is at most one cusp per fold in each sector, and that each sector has the same number of cusps. We now formulate this without mention of a Morse 2–function.

First, when M^3 has a boundary ∂M, then a Heegaard splitting is a splitting into compression bodies rather than solid handlebodies. Traditionally, a compression body is the result of attaching $n \leq k$ 3–dimensional 2–handles to $\{1\} \times F_k \subset [0,1] \times F_k$ so as to get a cobordism from F_k to F_{k-n}, where F_k is a closed surface of genus k. In fact, we can even consider the case where F is a compact surface $F_{k,b}$ of genus k with $b \geq 0$ boundary components, in which case we get a cobordism to $F_{(k-n),b}$. Note that the diffeomorphism type of such a cobordism is completely determined by k, b and n; let $C_{k,b,n}$ denote a standard model for this compression body. To summarize, both ends of $C_{k,b,n}$ are surfaces with b boundary components, the higher genus end has genus k and there are n compression disks yielding a lower genus end with genus $k-n$.

Now consider $Z_{k,b,n} = [0,1] \times C_{k,b,n}$. Part of $\partial Z_{k,b,n}$ is

$$Y_{k,b,n} = (\{0\} \times C_{k,b,n}) \cup ([0,1] \times F_{k,b}) \cup (\{1\} \times C_{k,b,n}),$$

which has a natural genus-k Heegaard splitting into two compression bodies

$$Y^+_{k,b,n} = \left(\left[\tfrac{1}{2},1\right] \times F_{k,b}\right) \cup (\{1\} \times C_{k,b,n}) \quad \text{and} \quad Y^-_{k,b,n} = (\{0\} \times C_{k,b,n}) \cup \left(\left[0,\tfrac{1}{2}\right] \times F_{k,b}\right).$$

Finally, given any $g \geq k$, let $Y_{k,b,n} = Y^+_{k,b,n,g} \cup Y^-_{k,b,n,g}$ be the genus-g Heegaard splitting obtained from the natural genus-k splitting by stabilizing $g-k$ times.

Definition 16 A trisection of a 4–manifold X with boundary is a splitting $X = X_1 \cup X_2 \cup X_3$ and integers $0 \leq k, b, n, g$ with $n \leq k \leq g$ such that each X_i is diffeomorphic to $Z_{k,b,n}$ via a diffeomorphism $\phi_i\colon X_i \to Z_{k,b,n}$ for which

$$\phi_i(X_i \cap X_{i+1}) = Y^+_{k,b,n,g} \quad \text{and} \quad \phi_i(X_i \cap X_{i-1}) = Y^-_{k,b,n,g}.$$

We leave the proof of the following to the reader:

Lemma 17 *A trisection of a* 4*–manifold* X *with nonempty boundary restricts to the boundary* $M^3 = \partial X$ *as either a fibration over* S^1 *(when* $b = 0$*) or an open book decomposition (when* $b \neq 0$*). In the first case,* $X_i \cap \partial X$ *is the inverse image under the fibration of* $[2\pi i/3, 2\pi(i+1)/3] \subset S^1$*. In the second case,* $X_i \cap \partial X$ *is the union of this inverse image and the binding.*

Remark 18 Lefschetz fibrations over B^2 can be perturbed to give examples of trisections in this relative setting. Assume that $f\colon X^4 \to B^2$ is a bundle with fiber $F_{k,b}$ except for exceptional fibers which have nodes where f is given in local coordinates (z, w) by $f(z, w) = zw$. Lekili showed in [7] that the map f could be locally perturbed so that the node is replaced by three 1–folds in the shape of a hyperbolic triangle, as in Figure 29. We need such a triangle to go around the central fiber of our trisection, so we move a cusp up to and past the central fiber. This ups the genus of the central fiber by one. Now it is easy to trisect X for the only folds are these triangles.

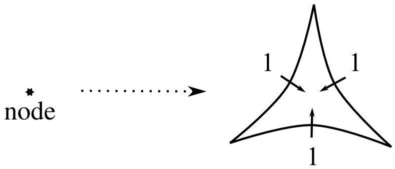


Figure 29: Perturbation of a Lefschetz node singularity

Remark 19 Given two 4–manifolds X and X', with diffeomorphic boundary, both trisected with $b = 0$, and with a gluing map $\partial X \to -\partial X'$ respecting trisections, gluing along the boundary does not immediately produce a trisection of the closed manifold $X \cup X'$. However, we naturally have six pieces which fit together like the faces of a cube. From this, the technique described in Section 2 for producing a trisection of a bundle over S^2 can be generalized to give a natural trisection of $X \cup X'$.

Theorem 20 *Given a* 4*–manifold* X *with an open book decomposition or fibration over* S^1 *on* ∂X*, there exists a trisection of* X *restricting to* ∂X *as the given fibration or open book.*

Proof Use the given boundary data to see X as a cobordism from $F \times [0,1]$ to $F \times [0,1]$, where F is either the fiber or the page. Using a handle decomposition of X compatible with this cobordism structure, repeat the second version of the proof of Theorem 4. □

Stabilization of trisections makes sense in the relative case, since it takes place inside a ball in the interior of X.

Theorem 21 *Any two trisections of a fixed* 4*–manifold X which agree on ∂X are isotopic after stabilizations.*

Proof Again, the proof of Theorem 11 works verbatim in this case, once we fix the appropriate cobordism structure on X. The key idea is that Cerf theory works perfectly well when we fix behavior on compact subsets. □

Acknowledgements This work was partially supported by a grant from the Simons Foundation (#210381 to David Gay). David Gay was partially supported by NSF grant DMS-1207721. Robion Kirby was partially supported by NSF grant DMS-0838703.

References

[1] **J S Birman**, **W W Menasco**, *On Markov's theorem*, J. Knot Theory Ramifications 11 (2002) 295–310 MR1905686

[2] **S E Cappell**, **R Lee**, **E Y Miller**, *On the Maslov index*, Comm. Pure Appl. Math. 47 (1994) 121–186 MR1263126

[3] **J Cerf**, *La stratification naturelle des espaces de fonctions différentiables réelles et le théorème de la pseudo-isotopie*, Inst. Hautes Études Sci. Publ. Math. 39 (1970) 5–173 MR0292089

[4] **D T Gay**, **R Kirby**, *Indefinite Morse* 2*–functions: broken fibrations and generalizations*, Geom. Topol. 19 (2015) 2465–2534 MR3416108

[5] **K Johannson**, *Topology and combinatorics of* 3*–manifolds*, Lecture Notes in Mathematics 1599, Springer, Berlin (1995) MR1439249

[6] **F Laudenbach**, **V Poénaru**, *A note on* 4*–dimensional handlebodies*, Bull. Soc. Math. France 100 (1972) 337–344 MR0317343

[7] **Y Lekili**, *Wrinkled fibrations on near-symplectic manifolds*, Geom. Topol. 13 (2009) 277–318 MR2469519

[8] **P Ozsváth**, **Z Szabó**, *Holomorphic triangles and invariants for smooth four-manifolds*, Adv. Math. 202 (2006) 326–400 MR2222356

[9] **K Reidemeister**, *Zur dreidimensionalen Topologie*, Abh. Math. Sem. Univ. Hamburg 9 (1933) 189–194 MR3069596

[10] **N Saveliev**, *Lectures on the topology of* 3*–manifolds*, Walter de Gruyter, Berlin (1999) MR1712769

[11] **J Singer**, *Three-dimensional manifolds and their Heegaard diagrams*, Trans. Amer. Math. Soc. 35 (1933) 88–111 MR1501673

[12] **F Waldhausen**, *Heegaard–Zerlegungen der* 3*–Sphäre*, Topology 7 (1968) 195–203 MR0227992

[13] **C T C Wall**, *Non-additivity of the signature*, Invent. Math. 7 (1969) 269–274 MR0246311

DG: *Euclid Lab, 160 Milledge Terrace*
Athens, GA 30606, United States

DG: *Department of Mathematics, University of Georgia*
Athens, GA 30602, United States

RK: *Department of Mathematics, University of California, Berkeley*
Berkeley, CA 94720-3840, United States

d.gay@euclidlab.org, kirby@math.berkeley.edu

Proposed: Yasha Eliashberg Received: 4 December 2013
Seconded: Benson Farb, Dmitri Burago Revised: 21 January 2016

Geometry & Topology 20 (2016) 3133–3217

The Picard group of topological modular forms via descent theory

AKHIL MATHEW
VESNA STOJANOSKA

This paper starts with an exposition of descent-theoretic techniques in the study of Picard groups of $\boldsymbol{E}_\infty$–ring spectra, which naturally lead to the study of Picard spectra. We then develop tools for the efficient and explicit determination of differentials in the associated descent spectral sequences for the Picard spectra thus obtained. As a major application, we calculate the Picard groups of the periodic spectrum of topological modular forms TMF and the nonperiodic and nonconnective Tmf. We find that Pic(TMF) is cyclic of order 576, generated by the suspension Σ TMF (a result originally due to Hopkins), while $\mathrm{Pic}(\mathrm{Tmf}) = \mathbb{Z} \oplus \mathbb{Z}/24$. In particular, we show that there exists an invertible Tmf–module which is not equivalent to a suspension of Tmf.

14C22, 55N34, 55P43, 55S35, 55T99; 55P47

1 Introduction

Elliptic curves and modular forms occupy a central role in modern stable homotopy theory in the guise of the variants of *topological modular forms*: the connective tmf, the periodic TMF, and Tmf, which interpolates between them. These are structured ring spectra which have demonstrated surprising connections between the arithmetic of elliptic curves and v_2–periodicity in stable homotopy. For example, tmf detects a number of 2–torsion and 3–torsion classes in the stable homotopy groups of spheres through the Hurewicz image. Even more interestingly, the more geometric-natured TMF can be used to detect and describe, using congruences between modular forms, the 2–line of the Adams–Novikov spectral sequence at primes $p \geq 5$, according to Behrens [7].

From a different perspective, the structure of topological modular forms as $\boldsymbol{E}_\infty$–ring spectra leads to symmetric monoidal ∞–categories of modules which give rise to well-behaved invariants of algebraic or algebrogeometric type. For instance, Meier [48] has studied TMF–modules which become free when certain level structures are introduced; these can be thought of as locally free sheaves with respect to a predetermined cover.

Published: 21 December 2016 DOI: 10.2140/gt.2016.20.3133

Our goal in this paper is to understand another such invariant, the Picard group. Any symmetric monoidal category has an associated group of isomorphism classes of objects invertible under the tensor product, which is called the *Picard group*. The classical examples are the Picard group $\mathrm{Pic}(R)$ of a ring R, ie of the category $\mathrm{Mod}(R)$ of R–modules, or the Picard group of a scheme X, ie of the category $\mathrm{Mod}(\mathcal{O}_X)$ of quasicoherent modules over its structure sheaf. In homotopy theory, the interest in Picard groups arose when Mike Hopkins made the observation that the homotopy categories of E_n–local and $K(n)$–local spectra have interesting Picard groups, particularly when the prime at hand is small in comparison with n. Here, E_n is the Lubin–Tate spectrum and $K(n)$ is the Morava K–theory spectrum at height n. In the few existing computations of such groups, notably those in Hopkins, Mahowald and Sadofsky [26], Hovey and Sadofsky [27], Kamiya and Shimomura [29], Goerss, Henn, Mahowald and Rezk [17] and Heard [21], one often uses that an invertible E_n–module must be a suspension of E_n itself.

The $K(2)$–localization of any of the three versions of topological modular forms gives a spectrum closely related to the Lubin–Tate spectrum E_2; namely, this localization is a finite product of homotopy fixed point spectra of finite group actions on E_2 (or slight variants of E_2 with larger residue fields). More generally, each E_n is an $\boldsymbol{E}_\infty$–ring spectrum with an action, through $\boldsymbol{E}_\infty$–ring maps, by a profinite group $\mathbb{G}_n$ called the Morava stabilizer group (see Rezk [57] for the $\boldsymbol{E}_1$–ring case). The $K(n)$–local sphere is obtained then as the Devinatz–Hopkins homotopy fixed points. However, $\mathbb{G}_n$ also has interesting finite subgroups when the prime is relatively small with respect to n. If G is such a subgroup, the homotopy fixed points E_n^{hG} are an $\boldsymbol{E}_\infty$–ring spectrum, which is in theory easier to study than the $K(n)$–local sphere, but hopefully contains a lot of information about the $K(n)$–local sphere. For instance, Hopkins has observed that in all known examples, the Picard group of E_n^{hG} (unlike that of the $K(n)$–local category) is very simple as it only contains suspensions of E_n^{hG}, and raised the following natural question.

Question (Hopkins) Let G be a finite subgroup of the Morava stabilizer group $\mathbb{G}_n$ at height n. Is it true that any invertible $K(n)$–local module over E_n^{hG} is a suspension of E_n^{hG}?

The periodic TMF is closer to its $K(2)$–localization than Tmf, and this is demonstrated by the following result, originally due to Hopkins but unpublished.

Theorem A (Hopkins) *The Picard group of* TMF *is isomorphic to* $\mathbb{Z}/576$, *generated by the suspension* $\Sigma\,\mathrm{TMF}$.

In the paper at hand, we prove Theorem A using a descent-theoretic approach. In particular, our method is different from Hopkins's. The descent-theoretic approach also enables us to prove that, nonetheless, the nonconnective, nonperiodic flavor of topological modular forms Tmf behaves differently and has a more interesting Picard group.

Theorem B *The Picard group of* Tmf *is isomorphic to* $\mathbb{Z} \oplus \mathbb{Z}/24$*, generated by the suspension* $\Sigma\,\mathrm{Tmf}$ *and a certain* 24*–torsion invertible object.*

In addition, we explicitly construct the 24–torsion module in Construction 8.4.2. We note that, after the initial submission of this paper, the preprint of Hill and Meier [23] appeared, in which the authors use techniques from C_2–equivariant stable homotopy to construct exotic torsion elements in the Picard group of $\mathrm{Tmf}_1(3)$. In contrast, our construction is given by an unusual gluing of locally trivial modules.

We hope that our method of proof of Theorems A and B, which is very general, will also be of interest to those not directly concerned with TMF. Our method is inspired by and analogous to the forthcoming work of Gepner and Lawson [15] on Galois descent of Brauer as well as Picard groups, though the key ideas are classical.

Take, for example, the periodic variant TMF. Its essential property is that it arises as the global sections of the structure sheaf $\mathcal{O}^{\mathrm{top}}$ of a regular "derived stack" $(\mathfrak{M}_{\mathrm{ell}}, \mathcal{O}^{\mathrm{top}})$ refining the moduli stack of elliptic curves M_{ell}. Thus

$$\mathrm{TMF} = \Gamma(\mathfrak{M}_{\mathrm{ell}}, \mathcal{O}^{\mathrm{top}}) = \varprojlim_{\mathrm{Spec}\, R \to M_{\mathrm{ell}}} \Gamma(\mathrm{Spec}\, R, \mathcal{O}^{\mathrm{top}}),$$

where the maps $\mathrm{Spec}\, R \to M_{\mathrm{ell}}$ range over all étale morphisms from affine schemes to M_{ell}. Moreover, the $\boldsymbol{E}_\infty$–ring spectra $\Gamma(\mathrm{Spec}\, R, \mathcal{O}^{\mathrm{top}})$ are weakly even periodic; thus we have TMF as the homotopy limit of a diagram of weakly even periodic $\boldsymbol{E}_\infty$–rings. It follows by the main result in Mathew and Meier [42] that the *module category* of TMF can also be represented as the inverse limit of the module categories $\mathrm{Mod}(\mathcal{O}^{\mathrm{top}}(\mathrm{Spec}\, R))$, that is, as quasicoherent sheaves on the derived stack. In any analogous situation, our descent techniques for calculating Picard groups apply.

Over an affine chart $\mathrm{Spec}\, R \to M_{\mathrm{ell}}$, the Picard group of $\Gamma(\mathrm{Spec}\, R, \mathcal{O}^{\mathrm{top}})$ (ie that of an *elliptic spectrum*) is purely algebraic, by a classical argument in Hopkins, Mahowald and Sadofsky [26] and Baker and Richter [4] with "residue fields". This results from the fact that the ring $\pi_*\Gamma(\mathrm{Spec}\, R, \mathcal{O}^{\mathrm{top}})$ is *homologically* simple: in particular, it has finite global dimension, which makes the study of $\Gamma(\mathrm{Spec}\, R, \mathcal{O}^{\mathrm{top}})$–modules much easier. One attempts to use this together with descent theory to compute the Picard group of TMF itself; however, doing so necessitates the consideration of higher homotopy coherences. For this, it is important to work with Picard *spectra* rather than Picard groups, as they have a better formal theory of descent.

The Picard spectrum $\mathfrak{pic}(A)$ of an $\boldsymbol{E}_\infty$–ring A is an important spectrum associated to A that deloops the space of units $\mathrm{GL}_1(A)$ of May [46]:[1] it is connective, its π_0 is the Picard group of A, and its 1–connective cover $\tau_{\geq 1}\,\mathfrak{pic}(A)$ is equivalent to $\Sigma\,\mathfrak{gl}_1(A)$ for $\mathfrak{gl}_1(A)$ the *spectrum of units* of [46]. We find that the *Picard spectrum* of TMF is the connective cover of the homotopy limit of $\mathfrak{pic}(\mathcal{O}^{\mathrm{top}}(\operatorname{Spec} R))$, taken over étale maps $\operatorname{Spec} R \to M_{\mathrm{ell}}$. This statement is a homotopy-theoretic expression of the descent theory that we need. Thus, we get a *descent spectral sequence* for the homotopy groups of $\mathfrak{pic}(\mathrm{TMF})$, which is a computational tool for understanding the aforementioned homotopy coherences concretely. We use this technique to compute $\pi_0(\mathfrak{pic}(\mathrm{TMF}))$, the group we are after.

The descent spectral sequence has many consequences in cases where it degenerates simply for dimensional reasons, or in cases where the information sought is coarse. For instance, in a specific example (Proposition 2.4.9), we show that the Picard group of the $\boldsymbol{E}_\infty$–ring $C^*(S^1;\mathbb{Q}[\epsilon]/\epsilon^2)$ is given by $\mathbb{Z}\times\mathbb{Q}$, which yields a counterexample to a general conjecture of Balmer [5, Conjecture 74] on the Picard groups of certain tensor-triangulated categories. We also prove the following general results in Sections 4 and 5.

Theorem C *Let A be a weakly even periodic Landweber exact $\boldsymbol{E}_\infty$–ring with $\pi_0 A$ regular noetherian. Let $n \geq 1$ be an integer, and let L_n denote localization with respect to the Lubin–Tate spectrum E_n. The Picard group of $L_n A$ is*

$$\operatorname{Pic}(L_n A) = \operatorname{Pic}(\pi_* A) \times \pi_{-1}(L_n A),$$

where $\operatorname{Pic}(\pi_* A)$ refers to the (algebraic) Picard group of the graded commutative ring $\pi_* A$.

Note that $\operatorname{Pic}(\pi_* A)$ sits in an extension

$$0 \to \operatorname{Pic}(\pi_0 A) \to \operatorname{Pic}(\pi_* A) \to \mathbb{Z}/2 \to 0,$$

which is split if A is *strongly* even periodic.

Theorem D *Let A be an $\boldsymbol{E}_\infty$–ring such that $\pi_0 A$ is a field of characteristic zero and such that $\pi_i A = 0$ for $i > 0$. Then $\operatorname{Pic}(A)$ is infinite cyclic, generated by ΣA.*

Theorem E *Let G be a finite group, and let $A \to B$ be a faithful G–Galois extension of $\boldsymbol{E}_\infty$–rings in the sense of Rognes [59]. Then the relative Picard group of B/A, ie the kernel of $\operatorname{Pic}(A) \to \operatorname{Pic}(B)$, is $|G|$–power torsion of finite exponent.*

[1] See Ando, Blumberg, Gepner, Hopkins and Rezk [2] for a very important application.

For TMF, the descent spectral sequence does not degenerate so nicely, and we need to work further to obtain our main results. The homotopy groups of the Picard spectrum of an $\boldsymbol{E}_\infty$–ring A, starting with π_2, are simply those of A: in fact, we have a natural equivalence of *spaces*

$$\Omega^{\infty+2}\,\mathfrak{pic}(A) \simeq \Omega^{\infty+1}A.$$

This determines the E_2–page and many of the differentials in the descent spectral sequence for Pic(TMF), but not all the ones that affect π_0. A key step in our argument is the identification of the differentials of the descent spectral sequence for the Picard spectra, in a certain *range of dimensions*, with that of the (known) descent spectral sequence for $\pi_*(\mathrm{TMF})$. We prove this in a general setting in Section 5.

At the prime 2, this technique is not sufficient to determine all the differentials in the descent spectral sequence, and we need to determine in addition the first "unstable" differential in the Picard spectral sequence (in comparison to the usual descent spectral sequence). We give a "universal" formula for this first differential in Theorem 6.1.1, which we hope will have further applications.

Conventions Throughout, we will write $\mathcal{S}$ for the ∞–category of spaces, $\mathcal{S}_*$ for the ∞–category of pointed spaces, and Sp for the ∞–category of spectra. We will frequently identify abelian groups A with their associated Eilenberg–Mac Lane spectra HA. Finally, all spectral sequences are displayed with the Adams indexing convention, ie the vertical axis represents the cohomological degree, and the horizontal axis represents the total topological degree.

Acknowledgements We would like to thank Mike Hopkins for suggesting this project, and David Gepner, Hans-Werner Henn, and Tyler Lawson for helpful discussions. We are grateful to Justin Noel for catching an error in some of the computations and to Paul Goerss for pointing us to the work of Priddy [54]. We would also like to thank the referee for numerous helpful comments. We created the spectral sequence figures using Tilman Bauer's sseq LaTeX package. Mathew is partially supported by the NSF Graduate Research Fellowship under grant DGE-110640, and Stojanoska is partially supported by NSF grant DMS-1307390.

Part I Generalities

2 Picard groups

We begin by giving an introduction to Picard groups in stable homotopy theory. General references here include [26; 47].

2.1 Generalities

Let $(\mathcal{C}, \otimes, \mathbf{1})$ be a symmetric monoidal category.

Definition 2.1.1 The *Picard group* of $\mathcal{C}$ is the group of isomorphism classes of objects $x \in \mathcal{C}$ which are *invertible*, ie such that there exists an object $y \in \mathcal{C}$ such that $x \otimes y \simeq \mathbf{1}$. We will denote this group by $\mathrm{Pic}(\mathcal{C})$.

Remark 2.1.2 If $\mathcal{C}$ is a large category, then it is not necessarily clear that the Picard group is a set. However, in all cases of interest, $\mathcal{C}$ will be *presentable* so that this will be automatic (see Remark 2.1.4).

When $\mathcal{C}$ is the category of quasicoherent sheaves on a scheme (or stack) X, then this recovers the usual Picard group of X: line bundles are precisely the invertible objects. The principal goal of this paper is to compute a Picard group in a homotopy-theoretic setting.

We will repeatedly use the following simple principle, which follows from the observation that tensoring with an invertible object induces an autoequivalence of categories.

Proposition 2.1.3 *Let $\mathcal{C}_0 \subset \mathcal{C}$ be a full subcategory that is preserved under any autoequivalence of $\mathcal{C}$. Suppose the unit object $\mathbf{1} \in \mathcal{C}$ belongs to $\mathcal{C}_0$. Then any $x \in \mathrm{Pic}(\mathcal{C})$ belongs to $\mathcal{C}_0$ as well.*

For example, if $\mathbf{1}$ is a compact object (that is, if $\mathrm{Hom}_{\mathcal{C}}(\mathbf{1}, \cdot)$ commutes with filtered colimits), then so is x.

Suppose now that, more generally, $\mathcal{C}$ is a symmetric monoidal ∞–category in the sense of [39], which is the setting that we will be most interested in. Then we can still define the Picard group $\mathrm{Pic}(\mathcal{C})$ of $\mathcal{C}$, which is the same as $\mathrm{Pic}(\mathrm{Ho}(\mathcal{C}))$. Moreover, Proposition 2.1.3 is valid, but where one is allowed to (and often should) use ∞–categorical properties.

Remark 2.1.4 The theory of *presentable* ∞–categories [34, Section 5.5] enables one to address set-theoretic concerns. If $\mathcal{C}$ is a presentable symmetric monoidal ∞–category, then the unit of $\mathcal{C}$ is κ–compact for some regular cardinal κ. Therefore, by Proposition 2.1.3 (strictly speaking, its ∞–categorical analog), every invertible object of $\mathcal{C}$ is κ–compact, and the collection of κ–compact objects of $\mathcal{C}$ is essentially small. In particular, the collection of isomorphism classes forms a set and the Picard group is well defined.

Example 2.1.5 Suppose that $\mathcal{C}$ is a symmetric monoidal stable ∞–category such that the tensor product commutes with finite colimits in each variable. Then one has a natural homomorphism

$$\mathbb{Z} \to \operatorname{Pic}(\mathcal{C}),$$

sending $n \mapsto \Sigma^n \mathbf{1}$.

Example 2.1.6 Let Sp be the ∞–category of spectra with the smash product. Then it is a classical result [26, page 90] that $\operatorname{Pic}(\mathcal{C}) \simeq \mathbb{Z}$, generated by the sphere S^1. A quick proof based on the above principle (which simplifies the argument in [26] slightly) is as follows. If $T \in \mathrm{Sp}$ is invertible, so that there exists a spectrum T' such that $T \wedge T' \simeq S^0$, then we need to show that T is a suspension of S^0.

Since the unit object $S^0 \in \mathrm{Sp}$ is compact, it follows that T is compact: that is, it is a finite spectrum. By suspending or desuspending, we may assume that T is connective,[2] and that $\pi_0 T \neq 0$. By the Künneth formula, it follows easily that $H_*(T; F)$ is concentrated in one dimension for each field F. Since $H_*(T; \mathbb{Z})$ is finitely generated, an argument with the universal coefficient theorem implies that $H_*(T; \mathbb{Z})$ is torsion-free of rank one and is concentrated in dimension zero: ie $H_0(T; \mathbb{Z}) \simeq \mathbb{Z}$. By the Hurewicz theorem, $T \simeq S^0$.

Example 2.1.7 Other variants of the stable homotopy category can have more complicated Picard groups. For instance, if $E \in \mathrm{Sp}$, one can consider the ∞–category $L_E\mathrm{Sp}$ of *E–local spectra*, with the symmetric monoidal structure given by the E–localized smash product $(X, Y) \mapsto L_E(X \wedge Y)$. The Picard group of $L_E\mathrm{Sp}$ is generally much more complicated than $\mathbb{Z}$. When E is given by the Morava E–theories E_n or the Morava K–theories $K(n)$, the resulting Picard groups have been studied in [26; 27], among other references.

Another important example of this construction arises for R an E_∞–ring, when we can consider the symmetric monoidal ∞–category $\operatorname{Mod}(R)$ of R–modules.

Definition 2.1.8 Given an E_∞–ring R, we write $\operatorname{Pic}(R)$ to denote the Picard group $\operatorname{Pic}(\operatorname{Mod}(R))$.

Using the same argument as in Example 2.1.6, it follows that any invertible R–module is necessarily compact (ie perfect): in particular, the invertible modules actually form a set rather than a proper class. Note that if R is simply an E_2–ring spectrum, then $\operatorname{Mod}(R)$ is a monoidal ∞–category, so one can still define a Picard group. This raises the following natural question.

[2] We always use "connective" to mean "(-1)–connected".

Question 2.1.9 Is there an example of an E_2–ring whose Picard group is nonabelian?

We will only work with E_∞–rings in the future, as it is for these highly commutative multiplications that we will be able to obtain good (from the point of view of descent theory) infinite loop spaces that realize $\mathrm{Pic}(R)$ on π_0.

2.2 Picard ∞–groupoids

If $(\mathcal{C}, \otimes, \mathbf{1})$ is a symmetric monoidal ∞–category, we reviewed in the previous section the *Picard group* of $\mathcal{C}$. There is, however, a more fundamental invariant of $\mathcal{C}$, where we remember all isomorphisms (and higher isomorphisms), and which behaves better with respect to descent processes.

Definition 2.2.1 Let $\mathcal{P}\mathrm{ic}(\mathcal{C})$ denote the ∞–groupoid (ie space) of *invertible objects* in $\mathcal{C}$ and equivalences between them. We will refer to this as the *Picard ∞–groupoid* of $\mathcal{C}$; it is a group-like E_∞–space, and thus [45; 60] the delooping of a connective *Picard spectrum* $\mathfrak{pic}(\mathcal{C})$.

We have in particular
$$\pi_0\,\mathcal{P}\mathrm{ic}(\mathcal{C}) \simeq \mathrm{Pic}(\mathcal{C}).$$

However, we can also describe the higher homotopy groups of $\mathcal{P}\mathrm{ic}(\mathcal{C})$. Recall that since $\mathcal{C}$ is symmetric monoidal, $\mathrm{End}(\mathbf{1})$ is canonically an E_∞–space and $\mathrm{Aut}(\mathbf{1})$ consists of the grouplike components. Since
$$\Omega\,\mathcal{P}\mathrm{ic}(\mathcal{C}) \simeq \mathrm{Aut}(\mathbf{1}),$$
we get the relations
$$\pi_1\,\mathcal{P}\mathrm{ic}(\mathcal{C}) = (\pi_0\,\mathrm{End}(\mathbf{1}))^\times \qquad \text{and} \qquad \pi_i\,\mathcal{P}\mathrm{ic}(\mathcal{C}) = \pi_{i-1}\,\mathrm{End}(\mathbf{1}) \quad \text{for } i \geq 2.$$

Example 2.2.2 Let R be an E_∞–ring. We will write
$$\mathcal{P}\mathrm{ic}(R) \overset{\mathrm{def}}{=} \mathcal{P}\mathrm{ic}(\mathrm{Mod}(R)) \quad \text{and} \quad \mathfrak{pic}(R) \overset{\mathrm{def}}{=} \mathfrak{pic}(\mathrm{Mod}(R)).$$

Then $\mathcal{P}\mathrm{ic}(R)$ is a delooping of the *space of units* $\mathrm{GL}_1(R)$ studied in [46] and more recently using ∞–categorical techniques in [2]. In particular, the homotopy groups of $\mathcal{P}\mathrm{ic}(R)$ look very much like those of R (with a shift), starting at π_2. In fact, if we take the connected components at the basepoint, we have a natural equivalence of spaces
$$\tau_{\geq 1}(\mathrm{GL}_1 R) \simeq \tau_{\geq 1}(\Omega\,\mathcal{P}\mathrm{ic}(R)) \simeq \tau_{\geq 1}(\Omega^\infty R),$$
given by subtracting 1 with respect to the group structure on the infinite loop space $\Omega^\infty R$. Nonetheless, the *spectra* $\mathfrak{pic}(R)$ and R are generally very different: that is, the infinite loop structure on $\mathcal{P}\mathrm{ic}(R)$ behaves very differently from that of $\Omega^\infty R$.

Unlike the group-valued functor Pic, both $\mathcal{P}\mathrm{ic}$ and $\mathfrak{pic}$ have the fundamental property, upon which the calculations in this paper are based, that they commute with homotopy limits.

Proposition 2.2.3 *The functor*

$$\mathfrak{pic}\colon \mathrm{Cat}^{\otimes} \to \mathrm{Sp}_{\geq 0},$$

from the ∞–category $\mathrm{Cat}^{\otimes}$ of symmetric monoidal ∞–categories to the ∞–category $\mathrm{Sp}_{\geq 0}$ of connective spectra, commutes with limits and filtered colimits, and the functor $\mathcal{P}\mathrm{ic} = \Omega^{\infty} \circ \mathfrak{pic}\colon \mathrm{Cat}^{\otimes} \to \mathcal{S}_*$ does as well.

Proof We will treat the case of limits; the case of filtered colimits is similar and easier. It suffices to show that $\mathcal{P}\mathrm{ic}$ commutes with homotopy limits, since $\Omega^{\infty}\colon \mathrm{Sp}_{\geq 0} \to \mathcal{S}_*$ creates limits. Let $\mathrm{CAlg}(\mathcal{S})$ be the ∞–category of $\boldsymbol{E}_{\infty}$–spaces. Now, $\mathcal{P}\mathrm{ic}$ is the composite $\mathrm{inv} \circ \bar{\iota}$ where:

(1) $\bar{\iota}\colon \mathrm{Cat}^{\otimes} \to \mathrm{CAlg}(\mathcal{S})$ sends a symmetric monoidal ∞–category to the symmetric monoidal ∞–groupoid (ie $\boldsymbol{E}_{\infty}$–space) obtained by excluding all noninvertible morphisms.

(2) $\mathrm{inv}\colon \mathrm{CAlg}(\mathcal{S}) \to \mathcal{S}_*$ sends an $\boldsymbol{E}_{\infty}$–space X to the union of those connected components which are invertible in the commutative monoid $\pi_0 X$, with basepoint given by the identity.

It thus suffices to show that $\bar{\iota}$ and inv both commute with limits.

(1) The functor $\iota\colon \mathrm{Cat} \to \mathcal{S}$ that sends an ∞–category $\mathcal{C}$ to its *core* $\iota\mathcal{C}$ commutes with limits: in fact, it is right adjoint to the inclusion $\mathcal{S} \to \mathrm{Cat}$ that regards a space as an ∞–groupoid. See for instance [58, Section 17.2]. Now, to see that $\bar{\iota}$ commutes with limits, we observe that limits either in $\mathrm{Cat}^{\otimes}$ or in $\mathrm{CAlg}(\mathcal{S})$ are calculated at the level of the underlying spaces (resp. ∞–categories), so the fact that ι commutes with limits implies that $\bar{\iota}$ does too.

(2) It is easy to see that inv commutes with arbitrary products. Therefore, we need to show that inv turns pullbacks in $\mathrm{CAlg}(\mathcal{S})$ into pullbacks in $\mathcal{S}_*$. We recall that if $\mathcal{A}$, $\mathcal{B}$ are complete ∞–categories, then a functor $F\colon \mathcal{C} \to \mathcal{D}$ preserves limits if and only if it preserves pullbacks and products [34, Proposition 4.4.2.7]. Suppose given a homotopy pullback

(2-1)
$$\begin{tikzcd} A \arrow[r] \arrow[d] & B \arrow[d] \\ C \arrow[r] & D \end{tikzcd}$$

in $\mathrm{CAlg}(\mathcal{S})$; we need to show that

$$\begin{tikzcd}
\mathrm{inv}(A) \arrow[r] \arrow[d] & \mathrm{inv}(B) \arrow[d] \\
\mathrm{inv}(C) \arrow[r] & \mathrm{inv}(D)
\end{tikzcd}$$

is one too, in $\mathcal{S}_*$. Given the construction of inv as a union of connected components, it suffices to show that if $x \in \pi_0 A$ has the property that x maps to invertible elements in the monoids $\pi_0 B$, $\pi_0 C$, then x itself is invertible.

To see this, consider the homotopy pullback square (2-1). Addition of x induces an endomorphism of the square. Since it acts via homotopy equivalences on B, C, D, it follows formally that it must act invertibly on A, ie that $x \in \pi_0 A$ has an inverse. □

2.3 Descent

Let $R \to R'$ be a morphism of $\boldsymbol{E}_\infty$–rings. Recall the *cobar construction*, a cosimplicial $\boldsymbol{E}_\infty$–R–algebra

$$R' \rightrightarrows R' \otimes_R R' \mathrel{\substack{\textstyle\rightarrow\\[-0.6ex]\textstyle\rightarrow\\[-0.6ex]\textstyle\rightarrow}} \cdots,$$

important in descent procedures, which receives an augmentation from R. The cobar construction is the *Čech nerve* (see [34, Section 6.1.2]) of $R \to R'$, in the opposite ∞–category.

Definition 2.3.1 [37, Definition 5.2] We say that $R \to R'$ is *faithfully flat* if the map $\pi_0 R \to \pi_0 R'$ is faithfully flat and the natural map $\pi_* R \otimes_{\pi_0 R} \pi_0 R' \to \pi_* R'$ is an isomorphism.

In this case, the theory of faithfully flat descent goes into effect. We have:

Theorem 2.3.2 [37, Theorem 6.1] *Suppose $R \to R'$ is a faithfully flat morphism of $\boldsymbol{E}_\infty$–rings. Then the symmetric monoidal ∞–category* $\mathrm{Mod}(R)$ *can be recovered as the limit of the cosimplicial diagram of symmetric monoidal ∞–categories*

$$\mathrm{Mod}(R') \rightrightarrows \mathrm{Mod}(R' \otimes_R R') \mathrel{\substack{\textstyle\rightarrow\\[-0.6ex]\textstyle\rightarrow\\[-0.6ex]\textstyle\rightarrow}} \cdots.$$

As a result, by Proposition 2.2.3, $\mathcal{P}\mathrm{ic}(R)$ can be recovered as a totalization of spaces,

$$\mathcal{P}\mathrm{ic}(R) \simeq \mathrm{Tot}(\mathcal{P}\mathrm{ic}(R'^{\otimes(\bullet+1)})). \tag{2-2}$$

Equivalently, one has an equivalence of connective spectra

$$\mathfrak{pic}(R) \simeq \tau_{\geq 0}\,\mathrm{Tot}(\mathfrak{pic}(R'^{\otimes(\bullet+1)})). \tag{2-3}$$

In this paper, we will apply a version of this, except that we will work with morphisms of ring spectra that are not faithfully flat on the level of homotopy groups. As we will see, the descent spectral sequences given by (2-2) and (2-3) are not very useful in the faithfully flat case for our purposes.

Example 2.3.3 A more classical example of this technique (eg [20, Exercise 6.9]) is as follows. Let X be a nodal cubic curve over the complex numbers $\mathbb{C}$. Then X can be obtained from its normalization $\mathbb{P}^1$ by gluing together 0 and ∞. There is a pushout diagram of schemes:

$$\begin{tikzcd}
\{0,\infty\} \arrow[r] \arrow[d] & * \arrow[d] \\
\mathbb{P}^1 \arrow[r] & X
\end{tikzcd}$$

Therefore, one would *like* to say that the category $\mathrm{QCoh}(X)$ of quasicoherent sheaves on X fits into a homotopy pullback square

$$\begin{tikzcd}
\mathrm{QCoh}(X) \arrow[r] \arrow[d] & \mathrm{QCoh}(*) \arrow[d] \\
\mathrm{QCoh}(\mathbb{P}^1) \arrow[r] & \mathrm{QCoh}(* \sqcup *)
\end{tikzcd} \tag{2-4}$$

and that therefore the Picard *groupoid* of X fits into the homotopy cartesian square:

$$\begin{tikzcd}
\mathcal{P}\mathrm{ic}(X) \arrow[r] \arrow[d] & \mathcal{P}\mathrm{ic}(*) \arrow[d] \\
\mathcal{P}\mathrm{ic}(\mathbb{P}^1) \arrow[r] & \mathcal{P}\mathrm{ic}(*) \times \mathcal{P}\mathrm{ic}(*)
\end{tikzcd} \tag{2-5}$$

Unfortunately, (2-4) is not a pullback square of categories, because restricting to a closed subscheme is not an exact functor. It is possible to remedy this (up to connectivity issues) by working with derived ∞–categories [36, Theorem 7.1], or by noting that we are working with locally free sheaves and applying a version of [49, Theorems 2.1–2.3]. In any event, one can argue that (2-5) is homotopy cartesian.

Alternatively, we obtain a homotopy pullback diagram of *connective* spectra. Using the long exact sequence on π_*, it follows that we have a short exact sequence

$$0 \to \mathbb{C}^\times \to \mathrm{Pic}(X) \to \mathrm{Pic}(\mathbb{P}^1) \simeq \mathbb{Z} \to 0.$$

The approach of this paper is essentially an elaboration of this example.

2.4 Picard groups of E_∞–rings

We now specialize to the case of interest to us in this paper. Let R be an $\boldsymbol{E}_\infty$–ring, and consider the Picard group $\mathrm{Pic}(R)$, and better yet, the Picard ∞–groupoid $\mathcal{P}\mathrm{ic}(R)$ and the Picard spectrum $\mathfrak{pic}(R)$. The first of these has been studied by Baker and Richter in the paper [4], and we start by recalling some of their results.

We start with the following useful property.

Proposition 2.4.1 *The functor $R \mapsto \mathrm{Pic}(R)$ commutes with filtered colimits in R.*

Proof This is a consequence of a form of "noetherian descent" [19, Section 8]. Given an $\boldsymbol{E}_\infty$–ring T, let $\mathrm{Mod}^\omega(T)$ denote the ∞–category of perfect T–modules. If I is a filtered ∞–category and $\{R_i\}_{i\in I}$ is a filtered system of $\boldsymbol{E}_\infty$–rings indexed by I, then the functor of symmetric monoidal ∞–categories

$$\varinjlim_{i\in I} \mathrm{Mod}^\omega(R_i) \to \mathrm{Mod}^\omega(\varinjlim_I R_i) \tag{2-6}$$

is an equivalence. We outline the proof of this below.

Assume without loss of generality that I is a filtered partially ordered set and write $R = \varinjlim_I R_i$. To see that (2-6) is an equivalence, observe that the ∞–category $\varinjlim_{i\in I} \mathrm{Mod}^\omega(R_i)$ has objects given by pairs (M, i) where $i \in I$ and $M \in \mathrm{Mod}^\omega(R_i)$. The space of maps between (M, i) and (N, j) is given by

$$\varinjlim_{k\geq i,j} \mathrm{Hom}_{\mathrm{Mod}(R_k)}(R_k \otimes_{R_i} M, R_k \otimes_{R_j} N).$$

For instance, this implies that if $i' \geq i$, the pair (M, i) is (canonically) equivalent to the pair $(R_{i'} \otimes_{R_i} M, i')$. Thus, the assertion that (2-6) is fully faithful is equivalent to the assertion that if $M, N \in \mathrm{Mod}^\omega(R_i)$ for some i, then the natural map

$$\varinjlim_{j\geq i} \mathrm{Hom}_{\mathrm{Mod}^\omega(R_j)}(R_j \otimes_{R_i} M, R_j \otimes_{R_i} N) \to \mathrm{Hom}_{\mathrm{Mod}^\omega(R)}(R \otimes_{R_i} M, R \otimes_{R_i} N) \tag{2-7}$$

is an equivalence. But (2-7) is clearly an equivalence if $M = R_i$ for *any* N. The collection of $M \in \mathrm{Mod}^\omega(R_i)$ such that (2-7) is an equivalence is closed under finite colimits, desuspensions, and retracts, and therefore it is all of $\mathrm{Mod}^\omega(R_i)$. It therefore follows that (2-6) is fully faithful.

Moreover, the image of (2-6) contains $R \in \mathrm{Mod}^\omega(R)$ and is closed under desuspensions and cofibers (thus finite colimits). Let $\mathcal{C} \subset \mathrm{Mod}^\omega(R)$ be the subcategory generated by R under finite colimits and desuspensions. We have shown the image of the fully faithful functor (2-6) contains $\mathcal{C}$. Any object $M \in \mathrm{Mod}^\omega(R)$ is a retract of an object $X \in \mathcal{C}$, associated to an idempotent map $e\colon X \to X$. We can "descend" X to some $X_i \in \mathrm{Mod}^{\omega}(R_i)$ and the map e to a self-map $e_i\colon X_i \to X_i$ such that e_i^2 is homotopic to e_i. As is classical, we use the idempotent e_i to split X_i; see [52, Proposition 1.6.8] or the older [12] and [13, Theorem 5.3]. Explicitly, form the filtered colimit Y_i of $X_i \xrightarrow{e_i} X_i \xrightarrow{e_i} \cdots$, which splits off X_i. The tensor product $R \otimes_{R_i} Y_i$ is the direct summand of X given by the idempotent e and is therefore equivalent to M.

The association $\mathcal{C} \mapsto \mathcal{P}\mathrm{ic}(\mathcal{C})$ commutes with filtered colimits of symmetric monoidal ∞–categories by Proposition 2.2.3. Taking Picard groups in the equivalence (2-6), the proposition follows. □

Purely algebraic information can be used to begin approaching $\mathrm{Pic}(R)$. Let $\mathrm{Pic}(R_*)$ be the Picard group of the symmetric monoidal category of *graded* R_*–modules. The starting point of [4] is the following.

Construction 2.4.2 There is a monomorphism

$$\Phi\colon \mathrm{Pic}(R_*) \to \mathrm{Pic}(R),$$

constructed as follows. If M_* is an invertible R_*–module, it has to be finitely generated and projective of rank one. Consequently, there is a finitely generated free R_*–module F_* of which M_* is a direct summand, ie there is a projection p_* with a section s_*:

$$\begin{tikzcd} F_* \arrow[r, shift left, "p_*"'] & M_* \arrow[l, shift left, "s_*"'] \end{tikzcd}$$

Clearly, F_* can be realized as an R–module F which is a finite wedge sum of copies of R or its suspensions. Let e_* be the idempotent given by composition $s_* \circ p_*$. Since F is free over R, e_* can be realized as an R–module map $e\colon F \to F$ which must be idempotent. Define M to be the colimit of the sequence $F \xrightarrow{e} F \xrightarrow{e} \cdots$, ie the image of the idempotent e. Observe that the homotopy groups of M are given by M_*, as desired. If M'_* is the inverse to M_* in the category of graded R_*–modules, we can construct an analogous R–module M', and clearly $M \otimes_R M' \simeq R$ by the degeneration of the Künneth spectral sequence. Thus, $M \in \mathrm{Pic}(R)$. The association $M_* \mapsto M$ defines Φ.

Note that any two R–modules that realize M_* on homotopy groups are equivalent by the degeneration of the Ext spectral sequence, and that Φ is a homomorphism by the degeneration of the Künneth spectral sequence. Observe also that Φ is clearly a monomorphism as equivalences of R–modules are detected on homotopy groups.

Definition 2.4.3 When Φ is an isomorphism, we say that $\mathrm{Pic}(R)$ is *algebraic*.

Baker and Richter [4] determine certain conditions which imply algebraicity. There are, in particular, two fundamental examples. The first one generalizes Example 2.1.6.

Theorem 2.4.4 [4] *Suppose R is a connective $\boldsymbol{E}_\infty$–ring. Then the Picard group of R is algebraic.*

Proof Since the formulation in [4, Theorem 21] assumed a coherence hypothesis on $\pi_* R$, we explain briefly how this (slightly stronger) version can be deduced from the theory of flatness of [39, Section 8.2.2]. Recall that an R–module M is *flat* if $\pi_0 M$ is a flat $\pi_0 R$–module and the natural map

$$\pi_* R \otimes_{\pi_0 R} \pi_0 M \to \pi_* M$$

is an isomorphism.

Since the Picard group commutes with filtered colimits in R, we may assume that R is finitely presented in the ∞–category of connective $\boldsymbol{E}_\infty$–rings: in particular, by [39, Proposition 8.2.5.31], $\pi_0 R$ is a finitely generated $\mathbb{Z}$–algebra and in particular noetherian; moreover, each $\pi_j R$ is a finitely generated $\pi_0 R$–module. These are the properties that will be critical for us.

Let M be an invertible R–module. We will show that $\pi_* M$ is a flat module over $\pi_* R$, which immediately implies the claim of the theorem. Localizing at a prime ideal of $\pi_0 R$, we may assume that $\pi_0 R$ is a noetherian local ring; in this case we will show the Picard group is $\mathbb{Z}$ generated by the suspension of the unit. We saw that M is perfect, so we can assume by shifting that M is connective and that $\pi_0 M \neq 0$. Now for every map[3] $R \to k$, for k a field, $\pi_*(M \otimes_R k)$ is necessarily concentrated in a single degree: in fact, $M \otimes_R k$ is an invertible object in $\mathrm{Mod}(k)$ and one can apply the Künneth formula to see that $\mathrm{Pic}(\mathrm{Mod}(k)) \simeq \mathbb{Z}$ generated by Σk. By Nakayama's lemma, since $\pi_0 M \neq 0$, the homotopy groups of $M \otimes_R k$ must be concentrated in degree zero. Thus, $M \otimes_R k \simeq k$ itself. Using Lemma 2.4.5, it follows that M is equivalent to R as an R–module, so we are done. □

Lemma 2.4.5 *Let R be a connective $\boldsymbol{E}_\infty$–ring with $\pi_0 R$ noetherian local with residue field k. Suppose moreover each $\pi_i R$ is a finitely generated $\pi_0 R$–module. Suppose M is a connective (ie (-1)–connected) perfect R–module. Then, for $n \geq 0$, the following are equivalent:*

(1) $M \simeq R^n$.

(2) $M \otimes_R k \simeq k^n$.

[3] Recall that we are using the same symbol to denote an abelian group and its Eilenberg–Mac Lane spectrum.

Proof Suppose $M \otimes_R k$ is isomorphic to k^n and concentrated in degree zero. Note that $\pi_0(M \otimes_R k) \simeq \pi_0 M \otimes_{\pi_0 R} k$. Choose a basis $\overline{x_1}, \dots, \overline{x_n}$ of this k–vector space and lift these elements to $x_1, \dots, x_n \in \pi_0 M$. These define a map $R^n \to M$ which induces an equivalence after tensoring with k, since $M \otimes_R k \simeq k^n$.

Now consider the cofiber C of $R^n \to M$. It follows that $C \otimes_R k$ is contractible. Suppose C itself is not contractible. The hypotheses on $\pi_* R$ imply that C is connective and each $\pi_j C$ is a finitely generated module over the noetherian local ring $\pi_0 R$. If j is chosen minimal such that $\pi_j C \neq 0$, then

$$0 = \pi_j(C \otimes_R k) \simeq \pi_j C \otimes_{\pi_0 R} k,$$

and Nakayama's lemma implies that $\pi_j C = 0$, a contradiction. □

Some of our analyses in the computational sections will rest upon the next result about the Picard groups of *periodic* ring spectra.

Theorem 2.4.6 (Baker and Richter [4, Theorem 37]) *Suppose R is a weakly even periodic $\boldsymbol{E}_\infty$–ring with $\pi_0 R$ regular noetherian. Then the Picard group of R is algebraic.*

The result in [4, Theorem 37] actually assumes that $\pi_0 R$ is a complete regular local ring. However, one can remove the hypotheses by replacing R with the localization $R_{\mathfrak{p}}$ for any $\mathfrak{p} \in \operatorname{Spec} \pi_0 R$ and then forming the completion at the maximal ideal.

We will need a slight strengthening of Theorem 2.4.6, though.

Corollary 2.4.7 *Suppose R is an $\boldsymbol{E}_\infty$–ring satisfying the following assumptions:*

(1) *$\pi_0 R$ is regular noetherian.*

(2) *The $\pi_0 R$–module $\pi_{2k} R$ is invertible for some $k > 0$.*

(3) *$\pi_i R = 0$ if $i \not\equiv 0 \mod 2k$.*

Then the Picard group of R is algebraic.

Proof Using the obstruction theory of [3] (as well as localization), we can construct "residue fields" in R as $\boldsymbol{E}_1$–algebras in $\operatorname{Mod}(R)$ (which will be $2k$–periodic rather than 2–periodic). After this, the same argument as in Theorem 2.4.6 goes through. □

Remark 2.4.8 If R is a ring spectrum satisfying the conditions of Corollary 2.4.7, then $\mathrm{Pic}(R) \cong \mathrm{Pic}(\pi_* R)$ sits in a short exact sequence

$$0 \to \mathrm{Pic}(\pi_0 R) \to \mathrm{Pic}(\pi_* R) \to \mathbb{Z}/(2k) \to 0.$$

The extension is such that the $(2k)^{\mathrm{th}}$ power of a set-theoretic lift of a generator of $\mathbb{Z}/(2k)$ to $\mathrm{Pic}(\pi_* R)$ is identified with the invertible $\pi_0 R$–module $\pi_{2k} R$.

An example of a nonalgebraic Picard group, based on [41, Example 7.1], is as follows.

Proposition 2.4.9 *The Picard group of the rational $\boldsymbol{E}_\infty$–ring $R = \mathbb{Q}[\epsilon_0, \epsilon_{-1}]/\epsilon_0^2$ (free on two generators ϵ_0, of degree 0, and ϵ_{-1}, of degree -1, and with the relation $\epsilon_0^2 = 0$) is given by $\mathbb{Z} \times \mathbb{Q}$.*

Proof The key observation is that R is equivalent, as an $\boldsymbol{E}_\infty$–ring, to cochains over S^1 on the (discrete) $\boldsymbol{E}_\infty$–ring $\mathbb{Q}[\epsilon_0]/\epsilon_0^2$, because $C^*(S^1; \mathbb{Q})$ is equivalent to $\mathbb{Q}[\epsilon_{-1}]$. By [40, Remark 7.9], we have a fully faithful, symmetric monoidal embedding $\mathrm{Mod}(R) \subset \mathrm{Loc}_{S^1}(\mathrm{Mod}(\mathbb{Q}[\epsilon_0]/\epsilon_0^2))$ into the ∞–category of local systems (see Definition 4.2.1 below) of $\mathbb{Q}[\epsilon_0]/\epsilon_0^2$–modules over the circle, whose image consists of those local systems of $\mathbb{Q}[\epsilon_0]/\epsilon_0^2$–modules such that the monodromy action of $\pi_1(S^1)$ is ind-unipotent.

In particular, to give an object in $\mathrm{Pic}(R)$ is equivalent to giving an element in the Picard group $\mathrm{Pic}(\mathbb{Q}[\epsilon_0]/\epsilon_0^2)$ (of which there are only the suspensions of the unit, by Theorem 2.4.4) and an ind-unipotent (monodromy) automorphism, which is necessarily given by multiplication by $1 + q\epsilon_0$ for $q \in \mathbb{Q}$. We observe that this gives the right group structure to the Picard group because $(1 + q\epsilon_0)(1 + q'\epsilon_0) = 1 + (q + q')\epsilon_0$. □

Proposition 2.4.9 provides a counterexample to [5, Conjecture 74], which states that in a tensor triangulated category generated by the unit with a local spectrum (eg with no nontrivial thick subcategories), any element $\mathcal{L}$ in the Picard group has the property that $\mathcal{L}^{\otimes n}$ is a suspension of the unit for suitable $n > 0$. In fact, one can take the (homotopy) category of perfect R–modules for R as in Proposition 2.4.9, which has no nontrivial thick subcategories by [41, Theorem 1.3].

Remark 2.4.10 Other Picard groups of interest come from the theory of *stable module ∞–categories* of a p–group G over a field k of characteristic p, which from a homotopy-theoretic perspective can be expressed as the module ∞–categories of the Tate construction k^{tG}. The Picard groups of stable module ∞–categories have been studied in the modular representation theory literature (under the name *endotrivial modules*) starting with [10], where it is proved that the Picard group is algebraic (and cyclic) in the case where G is elementary abelian. The classification for a general p–group appears in [8].

3 The descent spectral sequence

In this section, we describe a descent spectral sequence for calculating Picard groups. The spectral sequence (studied originally by Gepner and Lawson [15] in a closely related setting) is based on the observation (Proposition 2.2.3) that the association $\mathcal{C} \mapsto \mathcal{P}\mathrm{ic}(\mathcal{C})$, from symmetric monoidal ∞–categories to $\boldsymbol{E}_\infty$–spaces, commutes with homotopy limits. We will describe several examples and applications of this in the present section. Explicit computations will be considered in later parts of this paper.

For example, let $\{\mathcal{C}_U\}$ be a sheaf of symmetric monoidal ∞–categories on a site, and let $\Gamma(\mathcal{C})$ denote the global sections (ie the homotopy limit) ∞–category. Then we have an equivalence of connective spectra

$$\mathfrak{pic}(\Gamma(\mathcal{C})) \simeq \tau_{\geq 0}\Gamma(\mathfrak{pic}(\mathcal{C}_U)),$$

and one can thus use the descent spectral sequence for a sheaf of spectra to approach the computation of $\mathfrak{pic}(\Gamma(\mathcal{C}))$. We will use this approach, together with a bit of descent theory, to calculate $\mathrm{Pic(TMF)}$. The key idea is that while TMF itself has sufficiently complicated homotopy groups that results such as Theorem 2.4.6 cannot apply, the ∞–category of TMF–modules is built up as an inverse limit of module categories over $\boldsymbol{E}_\infty$–rings with better behaved homotopy groups.

3.1 Refinements

Let X be a Deligne–Mumford stack equipped with a flat map $X \to M_{\mathrm{FG}}$ to the moduli stack of formal groups. We will use the terminology of [42].

Definition 3.1.1 An *even periodic refinement* of X is a sheaf $\mathcal{O}^{\mathrm{top}}$ of $\boldsymbol{E}_\infty$–rings on the affine, étale site of X, such that for any étale map

$$\operatorname{Spec} R \to X,$$

the multiplicative homology theory associated to the $\boldsymbol{E}_\infty$–ring $\mathcal{O}^{\mathrm{top}}(\operatorname{Spec} R)$ is functorially identified with the (weakly) even-periodic Landweber-exact theory[4] associated to the formal group classified by $\operatorname{Spec} R \to X \to M_{\mathrm{FG}}$. We will denote the refinement of the ordinary stack X by $\mathfrak{X}$.

A very useful construction from the refinement $\mathfrak{X}$ is the $\boldsymbol{E}_\infty$–ring of "global sections" $\Gamma(\mathfrak{X}, \mathcal{O}^{\mathrm{top}})$, which is the homotopy limit of the $\mathcal{O}^{\mathrm{top}}(\operatorname{Spec} R)$ as $\operatorname{Spec} R \to X$ ranges over the affine étale site of X.

[4] See [35, Lecture 18] for an exposition of the theory of weakly even-periodic theories.

Example 3.1.2 When X is the moduli stack M_{ell} of elliptic curves, with the natural map $M_{\mathrm{ell}} \to M_{\mathrm{FG}}$ that assigns to an elliptic curve its formal group, fundamental work of Goerss, Hopkins, and Miller, and (later) Lurie constructs an even periodic refinement $\mathfrak{M}_{\mathrm{ell}}$. The global sections of $\mathfrak{M}_{\mathrm{ell}}$ are defined to be the $\boldsymbol{E}_\infty$–ring TMF of *topological modular forms*; for a survey, see [16]. There is a similar picture for the compactified moduli stack $\overline{M}_{\mathrm{ell}}$, whose global sections are denoted Tmf.

Definition 3.1.3 Given the refinement $\mathfrak{X}$, one has a natural symmetric monoidal stable ∞–category $\mathrm{QCoh}(\mathfrak{X})$ of *quasicoherent sheaves* on $\mathfrak{X}$, given as a homotopy limit of the (stable symmetric monoidal) ∞–categories $\mathrm{Mod}(\mathcal{O}^{\mathrm{top}}(\operatorname{Spec} R))$ for each étale map $\operatorname{Spec} R \to X$.

There is an adjunction

$$\mathrm{Mod}(\Gamma(\mathfrak{X}, \mathcal{O}^{\mathrm{top}})) \rightleftarrows \mathrm{QCoh}(\mathfrak{X}), \tag{3-1}$$

where the left adjoint "tensors up" and the right adjoint takes global sections.[5]

Our main goal in this paper is to investigate the left hand side; however, the right hand side is sometimes easier to work with, since even periodic, Landweber-exact spectra have convenient properties. Therefore, the following result will be helpful.

Theorem 3.1.4 [42, Theorem 4.1] *Suppose X is noetherian and separated, and $X \to M_{\mathrm{FG}}$ is quasiaffine. Then the adjunction* (3-1) *is an equivalence of symmetric monoidal ∞–categories.*

For example, since the map $M_{\mathrm{ell}} \to M_{\mathrm{FG}}$ is affine, it follows that Mod(TMF) is equivalent to $\mathrm{QCoh}(\mathfrak{M}_{\mathrm{ell}})$. This was originally proved by Meier, away from the prime 2, in [48]. Theorem 3.1.4 implies the analog for Tmf and the derived *compactified* moduli stack, as well [42, Theorem 7.2].

Suppose $X \to M_{\mathrm{FG}}$ is quasiaffine. In particular, it follows that there is a *sheaf* of symmetric monoidal ∞–categories on the affine, étale site of X, given by

$$(\operatorname{Spec} R \to X) \to \mathrm{Mod}(\mathcal{O}^{\mathrm{top}}(\operatorname{Spec} R)),$$

whose global sections are given by $\mathrm{Mod}(\Gamma(\mathfrak{X}, \mathcal{O}^{\mathrm{top}}))$. This diagram of ∞–categories is a sheaf in view of the descent theory of [37, Theorem 6.1], but [42, Theorem 4.1]

[5] One way to extract this from [39] is to consider the thick subcategory $\mathcal{C}$ of $\mathrm{QCoh}(\mathfrak{X}, \mathcal{O}^{\mathrm{top}})$ generated by the unit. Then, one obtains by the universal property of Ind an adjunction $\mathrm{Ind}(\mathcal{C}) \rightleftarrows \mathrm{QCoh}(\mathfrak{X}, \mathcal{O}^{\mathrm{top}})$. However, the symmetric monoidal ∞–category $\mathrm{Ind}(\mathcal{C})$ is generated under colimits by the unit, so it is by Lurie's symmetric monoidal version [39, Proposition 8.1.2.7] of Schwede–Shipley theory equivalent to modules over $\Gamma(\mathfrak{X}, \mathcal{O}^{\mathrm{top}})$, which is the ring of endomorphisms of the unit.

gives the global sections. We are now in the situation of the introduction to this section. In particular, we obtain a descent spectral sequence for $\mathfrak{pic}(\Gamma(X, \mathcal{O}^{\mathrm{top}}))$, and we turn to studying it in detail.

3.2 The Gepner–Lawson spectral sequence

Keep the notation of the previous subsection: X is a Deligne–Mumford stack equipped with a quasiaffine flat map $X \to M_{\mathrm{FG}}$, and $(\mathfrak{X}, \mathcal{O}^{\mathrm{top}})$ is an even periodic refinement.

Our goal in this subsection is to prove:

Theorem 3.2.1 *Suppose that X is a regular Deligne–Mumford stack with a quasiaffine flat map $X \to M_{\mathrm{FG}}$, and suppose $\mathfrak{X}$ is an even periodic refinement of X. There is a spectral sequence with*

$$E_2^{s,t} = \begin{cases} H^s(X, \mathbb{Z}/2) & \text{if } t = 0, \\ H^s(X, \mathcal{O}_X^{\times}) & \text{if } t = 1, \\ H^s(X, \omega^{(t-1)/2}) & \text{if } t \geq 3 \text{ is odd}, \\ 0 & \text{otherwise}, \end{cases} \tag{3-2}$$

whose abutment is $\pi_{t-s}\Gamma(\mathfrak{X}, \mathfrak{pic}(\mathcal{O}^{\mathrm{top}}))$. The differentials run $d_r\colon E_r^{s,t} \to E^{s+r,t+r-1}$.

The analogous spectral sequence for a faithful Galois extension has been studied in work of Gepner and Lawson [15], and our approach is closely based on theirs.

Proof In this situation, as we saw in the previous subsection, we get an equivalence of symmetric monoidal ∞–groupoids,

$$\mathcal{P}\mathrm{ic}(\Gamma(\mathfrak{X}, \mathcal{O}^{\mathrm{top}})) \simeq \mathrm{holim}_{\mathrm{Spec}\, R \to X}\, \mathcal{P}\mathrm{ic}(\mathcal{O}^{\mathrm{top}}(\mathrm{Spec}\, R)),$$

where $\mathrm{Spec}\, R \to X$ ranges over the affine étale maps. Equivalently, we have an equivalence of connective spectra

$$\mathfrak{pic}(\Gamma(\mathfrak{X}, \mathcal{O}^{\mathrm{top}})) \simeq \tau_{\geq 0}\big(\mathrm{holim}_{\mathrm{Spec}\, R \to X}\, \mathfrak{pic}(\mathcal{O}^{\mathrm{top}}(\mathrm{Spec}\, R))\big).$$

Let us study the descent spectral sequence associated to this. We need to understand the homotopy group *sheaves* of the sheaf of connective spectra

$$(\mathrm{Spec}\, R \to X) \mapsto \mathfrak{pic}(\mathcal{O}^{\mathrm{top}}(\mathrm{Spec}\, R)),$$

ie the sheafification of the homotopy group presheaves

$$(\mathrm{Spec}\, R \to X) \mapsto \pi_i\, \mathfrak{pic}(\mathcal{O}^{\mathrm{top}}(\mathrm{Spec}\, R)).$$

First, we know that

$$\pi_1 \mathfrak{pic}(\mathcal{O}^{\mathrm{top}}(\operatorname{Spec} R)) \simeq R^{\times},$$

and, for $i \geq 2$, we have

$$\pi_i(\mathfrak{pic}(\mathcal{O}^{\mathrm{top}}(\operatorname{Spec} R)) \simeq \pi_{i-1}\, \mathcal{O}^{\mathrm{top}}(\operatorname{Spec} R) = \begin{cases} \omega^{(i-1)/2} & \text{for } i \text{ odd,} \\ 0 & \text{for } i \text{ even.} \end{cases}$$

It remains to determine the homotopy group sheaf π_0. If X is a regular Deligne–Mumford stack, so that each ring R that enters is regular, then we can do this using Theorem 2.4.6. In fact, it follows that if R is a local ring, then $\pi_0\, \mathfrak{pic}(\mathcal{O}^{\mathrm{top}}(\operatorname{Spec} R))$ is isomorphic to $\mathbb{Z}/2$. Thus, up to suitably suspending once, invertible sheaves are locally trivial. Using the descent spectral sequence for a sheaf of spectra, we get that the above descent spectral sequence for $\Gamma(\mathfrak{X}, \mathfrak{pic}(\mathcal{O}^{\mathrm{top}}))$ is almost entirely the same as the descent spectral sequence for $\Gamma(\mathfrak{X}, \mathcal{O}^{\mathrm{top}})$ in the sense that the cohomology groups that appear for $t \geq 3$, ie $H^s(X, \omega^{(t-1)/2})$, are the same as those that appear in the descent spectral sequence for $\Gamma(\mathfrak{X}, \mathcal{O}^{\mathrm{top}})$. However, the terms for $t = 1$ are the étale cohomology of $\mathbb{G}_m$ on X. In particular, we obtain the term

$$H^1(X, \mathcal{O}_X^{\times}) \simeq \operatorname{Pic}(X),$$

which is the Picard group of the underlying ordinary stack. □

Remark 3.2.2 One may think of the spectral sequence as arising from a totalization, or rather as a filtered colimit of totalizations. Choose an étale hypercover $\mathfrak{U}$ given by $U_\bullet \to X$ by affine schemes $\{U_n\}$. For any $\boldsymbol{E}_\infty$–ring A, denote by $\mathcal{P}\mathrm{ic}^{\mathbb{Z}}(A)$ the symmetric monoidal subcategory of $\mathcal{P}\mathrm{ic}(A)$ spanned by those A–modules such that, after restricting to each connected component of $\operatorname{Spec} \pi_0 A$, become equivalent to a suspension of A. Denote by $\mathfrak{pic}^{\mathbb{Z}}(A)$ the associated connective spectrum. Then we form the totalization

$$\operatorname{Tot}\big(\mathfrak{pic}^{\mathbb{Z}}(\mathcal{O}^{\mathrm{top}}(U_\bullet))\big),$$

whose associated infinite loop space $\Omega^\infty \operatorname{Tot}\big(\mathfrak{pic}^{\mathbb{Z}}(\mathcal{O}^{\mathrm{top}}(U_\bullet))\big)$ is, by descent theory, the symmetric monoidal ∞–subgroupoid of $\mathcal{P}\mathrm{ic}(\Gamma(\mathfrak{X}, \mathcal{O}^{\mathrm{top}}))$ spanned by those invertible modules which become (up to a suspension) trivial after pullback along $U_0 \to X$. In particular, the filtered colimit of these totalizations is the spectrum we are after. The descent spectral sequence of Theorem 3.2.1 is the filtered colimit of these Tot spectral sequences.

3.3 Galois descent

We next describe the setting of the spectral sequence that was originally considered in [15]. Let $A \to B$ be a faithful G–Galois extension of $\boldsymbol{E}_\infty$–ring spectra in the sense of [59]. In particular, G acts on B in the ∞–category of $\boldsymbol{E}_\infty$–A–algebras and $A \to B^{hG}$ is an equivalence. Then $A \to B$ is an analog of a G–Galois étale cover in the sense of ordinary commutative algebra or algebraic geometry. As in ordinary algebraic geometry, there is a good theory of *Galois descent* along $A \to B$, as has been observed by several authors, for instance in [15; 48].

Theorem 3.3.1 (Galois descent) *Let $A \to B$ be a faithful G–Galois extension of $\boldsymbol{E}_\infty$–rings. Then there is a natural equivalence of symmetric monoidal ∞–categories* $\mathrm{Mod}(A) \simeq \mathrm{Mod}(B)^{hG}$.

The "strength" of the descent is in fact very good. As shown in [40, Theorem 3.36], any faithful Galois extension $A \to B$ satisfies a form of descent up to nilpotence: the thick tensor-ideal that B generates in $\mathrm{Mod}(A)$ is equal to all of $\mathrm{Mod}(A)$. This imposes strong restrictions on the descent spectral sequences that can arise.

Applying the Picard functor, we get an equivalence of spaces

$$\mathcal{P}\mathrm{ic}(A) \simeq \mathcal{P}\mathrm{ic}(B)^{hG}, \tag{3-3}$$

or an equivalence of *connective* spectra

$$\mathfrak{pic}(A) \simeq \tau_{\geq 0}\, \mathfrak{pic}(B)^{hG}. \tag{3-4}$$

Remark 3.3.2 The spectrum $\Sigma\, \mathfrak{gl}_1\, B$ is equivalent to $\tau_{\geq 1}\, \mathfrak{pic}(B)$; consider the induced map of G–homotopy fixed point spectral sequences. All the differentials involving the $t-s=0$ line will be the same for $\mathfrak{pic}\, B$ and $\Sigma\, \mathfrak{gl}_1\, B$. Hence, we obtain a short exact sequence

$$0 \to \pi_0(\Sigma\, \mathfrak{gl}_1\, B)^{hG} \to \pi_0(\mathfrak{pic}(B))^{hG} \to E_\infty^{0,0} \to 0,$$

where $E_\infty^{0,0}$ is the kernel of all the differentials supported on $H^0(G, \pi_0\, \mathfrak{pic}\, B)$. This short exact sequence exhibits $\pi_0(\Sigma\, \mathfrak{gl}_1\, B)^{hG}$ as the *relative Picard group* of $A \to B$, which consists of invertible A–modules which after smashing with B become isomorphic to B itself.

Our main interest in Galois theory, for the purpose of this paper, comes from the observation, due to Rognes, that there are numerous examples of G–Galois extensions of $\boldsymbol{E}_\infty$–rings $A \to B$ where the homotopy groups of B are significantly simpler than that of A. In particular, one hopes to understand the homotopy groups of $\mathfrak{pic}(B)$, and then use (3-3) and (3-4) together with an analysis of the associated homotopy fixed-point spectral sequence

$$H^s(G, \pi_t\, \mathfrak{pic}(B)) \Rightarrow \pi_{t-s}(\mathfrak{pic}(B))^{hG}, \tag{3-5}$$

whose abutment for $t = s$ is the Picard group $\mathrm{Pic}(A)$.

Example 3.3.3 [59, Proposition 5.3.1] The map $\mathrm{KO} \to \mathrm{KU}$ and the C_2–action on KU arising from complex conjugation exhibit KU as a C_2–Galois extension of KO.

Example 3.3.3 is fundamental and motivational to us: the study of KO–modules, which is a priori difficult because of the complicated structure of the ring $\pi_* \mathrm{KO}$, can be approached via Galois descent together with the (much easier) study of KU–modules. In particular, we obtain

$$\mathfrak{pic}(\mathrm{KO}) \simeq \tau_{\geq 0}\, \mathfrak{pic}(\mathrm{KU})^{hC_2},$$

and one can hope to use the homotopy fixed-point spectral sequence (HFPSS) to calculate $\mathfrak{pic}(\mathrm{KO})$. This approach is due to Gepner and Lawson [15],[6] and we shall give a version of it below in Section 7.1 (albeit using a different method of deducing differentials).

Other examples of Galois extensions come from the theory of topological modular forms with *level structure*.

Example 3.3.4 Let $n \in \mathbb{N}$. Let $\mathrm{TMF}(n)$ denote the periodic version of TMF for elliptic curves over $\mathbb{Z}\left[\frac{1}{n}\right]$–algebras with a *full level n structure*. Then, by [42, Theorem 7.6], $\mathrm{TMF}\left[\frac{1}{n}\right] \to \mathrm{TMF}(n)$ is a faithful $\mathrm{GL}_2(\mathbb{Z}/n)$–Galois extension. The advantage is that, if $n \geq 3$, the moduli stack of elliptic curves with level n structure is actually a regular affine scheme (by [30, Corollary 2.7.2], elliptic curves with full level $n \geq 3$ structure have no nontrivial automorphisms). In particular, $\mathrm{TMF}(n)$ is even periodic with regular π_0, and one can compute its Picard group purely algebraically by Theorem 2.4.6. One can then hope to use $\mathrm{GL}_2(\mathbb{Z}/n)$–descent to get at the Picard group of $\mathrm{TMF}\left[\frac{1}{n}\right]$. We will take this approach below.

3.4 The E_n–local sphere

In addition, descent theory can be used to give a spectral sequence for $\mathfrak{pic}(L_n S^0)$. This is related to work of Kamiya and Shimomura [29] and the upper bounds that they obtain on $\mathrm{Pic}(L_n S^0)$.

Consider the cobar construction on $L_n S^0 \to E_n$, ie the cosimplicial $\boldsymbol{E}_\infty$–ring

$$E_n \rightrightarrows E_n \wedge E_n \mathrel{\substack{\textstyle\rightarrow\\[-0.6ex]\textstyle\rightarrow\\[-0.6ex]\textstyle\rightarrow}} \cdots,$$

whose homotopy limit is $L_n S^0$. It is a consequence of the Hopkins–Ravenel smash product theorem [56, Chapter 8] that this cosimplicial diagram has "effective descent".

[6] The original calculation of the Picard group of KO, by related techniques, is unpublished work of Mike Hopkins.

Proposition 3.4.1 *The natural functor*

$$\mathrm{Mod}(L_n S^0) \to \mathrm{Tot}(\mathrm{Mod}(E_n^{\wedge(\bullet+1)})),$$

is an equivalence of symmetric monoidal ∞–categories.

Proof According to the Hopkins–Ravenel smash product theorem, the map of $\boldsymbol{E}_\infty$–rings $L_n S^0 \to E_n$ has the property that the thick tensor-ideal that E_n generates in $\mathrm{Mod}(L_n S^0)$ is all of $\mathrm{Mod}(L_n S^0)$.[7]

According to [40, Proposition 3.21], this implies the desired descent statement (the condition is there called "admitting descent"). The argument is a straightforward application of the Barr–Beck–Lurie monadicity theorem [39, Section 6.2]. □

In particular, we find that

$$\mathfrak{pic}(L_n S^0) \simeq \tau_{\geq 0} \operatorname{Tot} \mathfrak{pic}(E_n^{\wedge(\bullet+1)}).$$

Let us try to understand the associated spectral sequence.

The higher homotopy groups, π_i for $i \geq 2$, of $\mathfrak{pic}(E_n^{\wedge(\bullet+1)})$ are determined in terms of those of $E_n^{\wedge(\bullet+1)}$. Once again, it remains to determine π_0. Now E_n is an even periodic $\boldsymbol{E}_\infty$–ring whose π_0 is regular local, so $\mathrm{Pic}(E_n) \simeq \pi_0 \mathfrak{pic}(E_n) \simeq \mathbb{Z}/2$ by Theorem 2.4.6. The iterated smash products $E_n^{\wedge m}$ are also even periodic, so their Picard group contains at least a $\mathbb{Z}/2$. We do not need to know their exact Picard groups, however, to run the spectral sequence, as only the $\mathbb{Z}/2$ component is relevant for the spectral sequence (as it is all that comes from $\pi_0 \mathfrak{pic}(E_n)$).

Next, we need to determine the algebraic Picard group. After taking π_0, the simplicial scheme

$$\cdots \Rrightarrow \operatorname{Spec} \pi_0(E_n \wedge E_n) \rightrightarrows \operatorname{Spec} \pi_0 E_n$$

is a presentation of the moduli stack $M_{\mathrm{FG}}^{\leq n}$ of formal groups (over $\mathbb{Z}_{(p)}$–algebras) of height at most n.

Proposition 3.4.2 $\mathrm{Pic}(M_{\mathrm{FG}}^{\leq n}) \simeq \mathbb{Z}$, *generated by* ω.

Proof We use the presentation of M_{FG} (localized at p) via the simplicial stack

$$\cdots \Rrightarrow (\operatorname{Spec}(MU \wedge MU)_*)/\mathbb{G}_m \rightrightarrows (\operatorname{Spec} MU_*)/\mathbb{G}_m. \tag{3-6}$$

[7] The argument in [56, Chapter 8] is stated for the uncompleted Johnson–Wilson theories, but also can be carried out for the completed ones. We refer in particular to the lecture notes of Lurie [35]; Lecture 30 contains the necessary criterion for constancy of the Tot–tower.

Since the Picard group of a polynomial ring over $\mathbb{Z}_{(p)}$ is trivial,[8] and each smash power of MU has a polynomial ring for π_*, the Picard group of each of the terms in the simplicial stack *without* the $\mathbb{G}_m$–quotient is trivial, and the group of units is $\mathbb{Z}_{(p)}^\times$, constant across the simplicial object. In other words, the *Picard groupoid* of each $\operatorname{Spec}(MU^{\wedge(s+1)})_*$ is $B\mathbb{Z}_{(p)}^\times$. When we add the $\mathbb{G}_m$–quotient, we get $\mathbb{Z}\times B\mathbb{Z}_{(p)}^\times$ for the Picard groupoid of each term in the simplicial stack because of the possibility of twisting by a character of $\mathbb{G}_m$: this twisting corresponds to the powers of ω. By descent theory, this shows that $\operatorname{Pic}(M_{\mathrm{FG}}) \simeq \mathbb{Z}$, generated by ω. More precisely, the Picard groupoid of M_{FG} is the totalization of the Picard groupoids of $\operatorname{Spec}(MU^{\wedge(s+1)})_*/\mathbb{G}_m$, and each of these is $\mathbb{Z}\times B\mathbb{Z}_{(p)}^\times$: that is, the cosimplicial diagram of Picard groupoids is constant and the totalization is $\mathbb{Z}\times B\mathbb{Z}_{(p)}^\times$ again.

When we replace M_{FG} by $M_{\mathrm{FG}}^{\le n}$, we can replace the above presentation by excising from each term the closed substack cut out by $(p, v_1, \dots, v_n)$. This does not affect the Picard *groupoid* since the codimension of the substack removed is at least 2 (ie neither the Picard group nor the group of units is affected).[9] That is, when we modify each term in (3-6) to form the associated presentation of $M_{\mathrm{FG}}^{\le n}$, the Picard groupoid is unchanged. It follows by faithfully flat descent that the inclusion $M_{\mathrm{FG}}^{\le n} \to M_{\mathrm{FG}}$ induces an isomorphism on Picard groups (or groupoids) and that the Picard group is generated by ω. □

We obtain the following result.

Theorem 3.4.3 *There is a spectral sequence*

$$E_2^{s,t} = \begin{cases} \mathbb{Z}/2 & \textit{if } t = 0, \\ H^s(M_{\mathrm{FG}}^{\le n}, \mathcal{O}_{M_{\mathrm{FG}}}^\times) & \textit{if } t = 1, \\ H^s(M_{\mathrm{FG}}^{\le n}, \omega^{(t-1)/2}) & \textit{if } t \ge 3 \textit{ is odd}, \\ 0 & \textit{otherwise}, \end{cases}$$

which converges for $t-s \ge 0$ to $\pi_{t-s}\,\mathfrak{pic}(L_nS^0)$. The relevant occurrences of the second case are $H^0(M_{\mathrm{FG}}^{\le n}, \mathcal{O}_{M_{\mathrm{FG}}}^\times) \simeq \mathbb{Z}_{(p)}^\times$ and $H^1(M_{\mathrm{FG}}^{\le n}, \mathcal{O}_{M_{\mathrm{FG}}}^\times) \simeq \mathbb{Z}$.

Note in particular that the E_2–term is determined entirely in terms of the Adams–Novikov spectral sequence for the E_n–local sphere. As we will see in Section 5, many of the differentials are also determined by the ANSS.

[8] Since the Picard group commutes with filtered colimits, one reduces to the case of a polynomial ring on a finite number of variables, and here it follows from unique factorization.

[9] Once again, this is a familiar result for regular rings, and here one must pass to filtered colimits since one is working with polynomial rings on infinitely many variables.

4 First examples

In this section, we will give several examples where descent theory gives a quick calculation of the Picard group. In these examples, we will not need to analyze differentials in the descent spectral sequence (3-5). The main examples of interest, where there will be a number of differentials to determine, will be treated in the last part of this paper.

4.1 The faithfully flat case

We begin with the simplest case. Suppose $R \to R'$ is a morphism of $\boldsymbol{E}_\infty$–rings which is faithfully flat. In this case, we know from [37, Theorem 6.1] that the tensor-forgetful adjunction $\mathrm{Mod}(R) \rightleftarrows \mathrm{Mod}(R')$ is comonadic and we get a descent spectral sequence for the Picard group of R, as

$$\mathfrak{pic}(R) \simeq \tau_{\geq 0} \operatorname{Tot} \mathfrak{pic}(R'^{\otimes(\bullet+1)}).$$

This spectral sequence, however, gives essentially no new information that is not algebraic in nature. That is, the entire E_2–term $E_2^{s,t}$ for $t > 1$ vanishes, as it can be identified with the E_2–term for the cobar resolution $R'^{\otimes(\bullet+1)}$ of R, and this cobar resolution has a degenerate spectral sequence with nonzero terms only for $s = 0$ at E_2. For example, an element in $\mathrm{Pic}(R)$ is algebraic if *and only if* its image in $\mathrm{Pic}(R')$ is algebraic, by faithful flatness.

Thus, faithfully flat descent will be mostly irrelevant to us as a tool of computing the nonalgebraic parts of Picard groups. In the examples of interest, we want $\pi_* R'$ to be significantly simpler homologically than $\pi_* R$, so that we will be able to conclude (using results such as Theorem 2.4.6) that the Picard group of R' is entirely algebraic. But if $\pi_* R'$ is faithfully flat over $\pi_* R$, it cannot be much simpler homologically. (Recall for example that *regularity* descends under faithfully flat extensions of noetherian rings.)

4.2 Cochain $\boldsymbol{E}_\infty$–rings and local systems

In this subsection, we give another example of a family of $\boldsymbol{E}_\infty$–ring spectra whose Picard groups can be determined, or at least bounded.

Let X be a space and R an $\boldsymbol{E}_\infty$–ring. Let $R^X = C^*(X; R)$ be the $\boldsymbol{E}_\infty$–ring of R–valued cochains on X.

Definition 4.2.1 Let $\mathrm{Loc}_X(\mathrm{Mod}(R)) = \mathrm{Fun}(X, \mathrm{Mod}(R))$ denote the ∞–category of *local systems* of R–module spectra on X.

Then we have a fully faithful embedding of symmetric monoidal ∞–categories

$$\mathrm{Mod}^{\omega}(R^X) \subset \mathrm{Loc}_X(\mathrm{Mod}(R)),$$

which sends R^X to the constant local system at R and is determined by that. As discussed in [40, Section 7], this embedding is often useful for relating invariants of R^X to those of R. In particular, since any invertible R^X–module is perfect, we have a fully faithful functor of ∞–groupoids

$$\mathcal{P}\mathrm{ic}(R^X) \to \mathcal{P}\mathrm{ic}\big(\mathrm{Loc}_X(\mathrm{Mod}(R))\big) = \mathrm{Map}\big(X, \mathcal{P}\mathrm{ic}(\mathrm{Mod}(R))\big),$$

where the last identification follows because $\mathcal{P}\mathrm{ic}$ commutes with homotopy limits (Proposition 2.2.3). Thus, we get the following useful *upper bound* for the Picard group of R^X.

Proposition 4.2.2 *If R is an E_∞–ring and X is any space, then* $\mathrm{Pic}(R^X)$ *is a subgroup of $\pi_0(\mathfrak{pic}(R)^X)$.*

Without loss of generality, we will assume that X is connected. Note that we have a cofiber sequence

$$\Sigma\, \mathfrak{gl}_1(R) \to \mathfrak{pic}(R) \to H(\mathrm{Pic}(R)),$$

where $H(\mathrm{Pic}(R))$ is the Eilenberg–Mac Lane spectrum associated to the group $\mathrm{Pic}(R)$. If we take the long exact sequence after taking maps from X, we get an exact sequence

$$0 \to \pi_{-1}(\mathfrak{gl}_1(R)^X) \to \pi_0(\mathfrak{pic}(R)^X) \to \mathrm{Pic}(R). \tag{4-1}$$

Our object of interest, $\mathrm{Pic}(R^X)$, is a subobject of the middle term, by the above proposition.

Let us unwind the exact sequence further. First, observe that the composite map $\mathrm{Pic}(R^X) \to \pi_0(\mathfrak{pic}(R)^X) \to \mathrm{Pic}(R)$ comes from the map of E_∞–rings $R^X \to R$ given by choosing a basepoint of X. In particular, it is *split surjective* as it has a section given by $R \to R^X$ (so (4-1) is a split exact sequence). Next, using the truncation map $\mathfrak{gl}_1(R) \to HR_0^\times$, we have a map $\pi_{-1}(\mathfrak{gl}_1(R)^X) \to \pi_{-1}((HR_0^\times)^X) = \mathrm{Hom}(\pi_1(X), R_0^\times)$. We can understand this map in terms of $\mathrm{Pic}(R^X)$. Very explicitly, suppose given an invertible R^X–module M with associated local system $\mathcal{L} \in \mathrm{Loc}_X(\mathrm{Mod}(R))$. Then if the image of M in $\mathrm{Pic}(R)$ is trivial, we conclude that $\mathcal{L}_x \simeq R$ for any basepoint $x \in X$. An element in $\pi_1(X, x)$ induces a monodromy automorphism of $\mathcal{L}_x$ and thus defines an element of $R_0^\times$. This defines a map in $\mathrm{Hom}(\pi_1(X, x), R_0^\times)$. Let $\mathrm{Pic}^0(R^X)$ denote the kernel of $\mathrm{Pic}(R^X) \to \mathrm{Pic}(R)$. Then we have just described the map

$$\mathrm{Pic}^0(R^X) \xrightarrow{\phi} \mathrm{Hom}(\pi_1(X, x), R_0^\times), \tag{4-2}$$

that comes from the exact sequence (4-1).

The monodromy action cannot be arbitrary, since this local system is not arbitrary: it is in the image of $\mathrm{Mod}^{\omega}(R^X)$ and therefore belongs to the thick subcategory generated by the unit. As in [40, Section 8], it follows that the monodromy action of any element of the fundamental group must be *ind-unipotent*. In particular, fix an element M of $\mathrm{Pic}^0(R^X)$. Given any loop $\gamma \in \pi_1(X,x)$, the associated element $u = u_{\gamma,M} \in R_0^{\times}$ under the homomorphism $\phi(M)\colon \mathrm{Pic}^0(R^X) \to \mathrm{Hom}(\pi_1(X,x), R_0^{\times})$ of (4-2) must have the property that $u-1$ is nilpotent.

Hence if R_0 is a *reduced* ring, we deduce from (4-1) the following conclusion.

Corollary 4.2.3 *If R is an $\boldsymbol{E}_\infty$–ring with $\pi_0 R$ reduced, and X is any connected space, then we have a split short exact sequence*

$$0 \to A \to \mathrm{Pic}(R^X) \to \mathrm{Pic}(R) \to 0,$$

where $A \subset \pi_{-1}(\mathfrak{gl}_1(R)^X)$ is contained in $\pi_{-1}((\tau_{\geq 1}\,\mathfrak{gl}_1(R))^X) \subset \pi_{-1}((\mathfrak{gl}_1(R))^X)$. In particular, if $\pi_{-1}((\tau_{\geq 1}\,\mathfrak{gl}_1(R))^X) = 0$, then $\mathrm{Pic}(R) \to \mathrm{Pic}(R^X)$ is an isomorphism.

Again, we note that the map $\pi_{-1}((\tau_{\geq 1}\,\mathfrak{gl}_1(R))^X) \to \pi_{-1}(\mathfrak{gl}_1(R)^X)$ is injective, by the long exact sequence and the fact that $\pi_0(\mathfrak{gl}_1(R)^X) \to \pi_0((HR_0^{\times})^X) \simeq R_0^{\times}$ is surjective.

As an application, we obtain a calculation of the Picard group of a nonconnective $\boldsymbol{E}_\infty$–ring in a setting far from regularity.

Theorem 4.2.4 *Let A be any finite* abelian *group and let E_n be Morava E–theory. Then the Picard group of E_n^{BA} is $\mathbb{Z}/2$, generated by the suspension ΣE_n^{BA}. The same conclusion holds for any finite group G whose p–Sylow subgroup is abelian, where p is the prime of definition for E_n.*

Proof We induct on the p–rank of A. When A has no p–torsion, then $E_n^{BA} \simeq E_n$ and Theorem 2.4.6 implies that the Picard group is $\mathbb{Z}/2$.

If the p–rank of A is positive, write $A \simeq \mathbb{Z}/p^m \times A'$ where the p–rank of A' has smaller cardinality than that of A. The inductive hypothesis gives us that the Picard group of $E_n^{BA'}$ is $\mathbb{Z}/2$. Now $E_n^{BA} \simeq (E_n^{BA'})^{B\mathbb{Z}/p^m}$. Moreover, $E_n^{BA'}$ is well known to be even periodic (though its π_0 is not regular).[10]

[10] We refer to [25, Section 7] for a general analysis of the question of when E_n^{BG} is even-periodic for G a finite group.

We claim now that $\pi_{-1}((\tau_{\geq 1}\,\mathfrak{gl}_1(E_n^{BA}))^{B\mathbb{Z}/p^m}) = 0$. To see this, we note that the homotopy groups of $\tau_{\geq 1}\,\mathfrak{gl}_1(E_n^{BA'})$ are concentrated in even degrees and are all given by torsion-free p–complete abelian groups. Therefore, the cohomology groups $H^i(\mathbb{Z}/p^m, \pi_j\tau_{\geq 1}\,\mathfrak{gl}_1(E_n^{BA}))$ *vanish* if i is odd, since the $\mathbb{Z}/p^m$–action on them is trivial. In the homotopy fixed point spectral sequence for $(\tau_{\geq 1}\,\mathfrak{gl}_1(E_n^{BA}))^{B\mathbb{Z}/p^m}$ (ie the Atiyah–Hirzebruch spectral sequence), there is no room for contributions to π_{-1}. In fact, there is no room for differentials at all, which indicates that any $\lim^1$ terms cannot occur either. Now Corollary 4.2.3 shows that the map $E_n^{BA'} \to E_n^{BA}$ induces an equivalence on Picard groups, which completes the inductive step.

For the last claim, fix any finite group G with an abelian p–Sylow subgroup $A \subset G$. For any connected space X, denote as before $\mathrm{Pic}^0(R^X)$ the kernel of $\mathrm{Pic}(R^X) \to \mathrm{Pic}(R)$. We have a commutative square:

$$\begin{tikzcd}
\mathrm{Pic}^0(E_n^{BG}) \arrow[r] \arrow[d, hook] & \mathrm{Pic}^0(E_n^{BA}) \arrow[d, hook] \\
\pi_{-1}(\tau_{\geq 1}\,\mathfrak{gl}_1(E_n)^{BG}) \arrow[r] & \pi_{-1}(\tau_{\geq 1}\,\mathfrak{gl}_1(E_n)^{BA})
\end{tikzcd}$$

The bottom horizontal map is injective since $\tau_{\geq 1}\,\mathfrak{gl}_1(E_n)$ is p–local and BG is p–locally a wedge summand of BA in view of the transfer $\Sigma^\infty_+ BG \to \Sigma^\infty_+ BA$, which has the property that the composite $\Sigma^\infty_+ BG \to \Sigma^\infty_+ BA \to \Sigma^\infty_+ BG$ is a p–local equivalence by inspection of p–local homology. It follows that $\mathrm{Pic}^0(E_n^{BG}) \to \mathrm{Pic}^0(E_n^{BA})$ is injective, and since the latter is zero, the former must be as well. □

Recall that the spectrum E_1 is p–complete complex K–theory.

Proposition 4.2.5 *Let G be any finite group. Then the Picard group of E_1^{BG} is finite.*

Proof In fact, $\pi_{-1}(\tau_{\geq 1}\,\mathfrak{gl}_1(E_1)^{BG})$ is finite. We know that $\tau_{\geq 3}\,\mathfrak{gl}_1(E_1) \simeq \Sigma^4 ku_p^\wedge$ by a theorem of Adams and Priddy [1]. Moreover, $(ku_p^\wedge)^*(BG)$ is finite in each odd dimension, by comparing with $E_1^*(BG)$ which vanishes in odd dimensions. It follows now from Corollary 4.2.3 that the desired Picard group has to be finite. □

Question 4.2.6 Let G be any finite group. Can the Picard group of E_1^{BG} be any larger than $\mathbb{Z}/2$? What about the higher Morava E–theories?

4.3 Coconnective rational E_∞–rings

We can also determine the Picard groups of coconnective rational $\boldsymbol{E}_\infty$–ring spectra. A rational $\boldsymbol{E}_\infty$–ring R is said to be *coconnective* if:

(1) $\pi_0 R$ is a field (of characteristic zero).

(2) $\pi_i R = 0$ for $i > 0$.

Theorem D *If R is a coconnective rational $\boldsymbol{E}_\infty$–ring, then the Picard group* $\mathrm{Pic}(R)$ *is infinite cyclic, generated by ΣR.*

Proof Let $k = \pi_0 R$. We use [38, Proposition 4.3.3] to conclude that $R \simeq \mathrm{Tot}(A^\bullet)$, where $A^\bullet$ is a cosimplicial $\boldsymbol{E}_\infty$–k–algebra with each A^i of the form $k \oplus V[-1]$, where V is a discrete k–vector space; the $\boldsymbol{E}_\infty$–structure given is the "square-zero" one.

We thus begin with the case of $R = k \oplus V[-1]$: we will show that $\mathrm{Pic}(R) \simeq \mathbb{Z}$ in this case. Since Pic commutes with filtered colimits, we may assume that V is a finite-dimensional vector space. In this case,

$$R \simeq k^{S^1 \vee \cdots \vee S^1},$$

where the number of copies of S^1 in the wedge summand is equal to the dimension $n = \dim_k V$; by [38, Proposition 4.3.1], any rational $\boldsymbol{E}_\infty$–ring with these homotopy groups is equivalent to $k \oplus V[-1]$. But we can now use Corollary 4.2.3 to see that the Picard group of $k^{S^1 \vee \cdots \vee S^1}$ is $\mathbb{Z}$, generated by the suspension, because $\tau_{\geq 1} \mathfrak{gl}_1(k) = 0$.

Now suppose that R is arbitrary. As above, we have an equivalence $R \simeq \mathrm{Tot}(A^\bullet)$ where each A^i is a coconnective $\boldsymbol{E}_\infty$–ring of the form $k \oplus V[-1]$ for V a discrete k–vector space. We have seen above that $\mathrm{Pic}(A^i) \simeq \mathbb{Z}$. We know, moreover, that we have a fully faithful embedding of symmetric monoidal ∞–categories

$$\mathrm{Mod}^\omega(R) \subset \mathrm{Tot}(\mathrm{Mod}(A^\bullet)),$$

which implies that we have a fully faithful functor of ∞–groupoids

$$\mathcal{P}\mathrm{ic}(R) \to \mathrm{Tot}(\mathcal{P}\mathrm{ic}(A^\bullet)).$$

But each $\mathcal{P}\mathrm{ic}(A^i)$, as an ∞–groupoid, has homotopy groups given by

$$\pi_j \, \mathcal{P}\mathrm{ic}(A^i) \simeq \begin{cases} \mathbb{Z} & \text{if } j = 0, \\ k^\times & \text{if } j = 1, \end{cases}$$

and in particular, in the cosimplicial diagram $\mathcal{P}\mathrm{ic}(A^\bullet)$, all the maps are *equivalences*. This is a helpful consequence of coconnectivity. Therefore we find that $\mathrm{Tot}(\mathcal{P}\mathrm{ic}(A^\bullet))$ maps by equivalences to each $\mathcal{P}\mathrm{ic}(A^i)$, and we get an upper bound of $\mathbb{Z}$ for $\mathcal{P}\mathrm{ic}(R)$. This upper bound is realized by the suspension ΣR (which hits the generator of $\mathbb{Z} \simeq \pi_0 \, \mathrm{Tot}(\mathcal{P}\mathrm{ic}(A^\bullet))$). □

Remark 4.3.1 If $k = \mathbb{Q}$, then a large class of coconnective $\boldsymbol{E}_\infty$–rings with $\pi_0 \simeq \mathbb{Q}$ (eg those with reasonable finiteness hypotheses and vanishing π_{-1}) arise as cochains on a simply connected space, by Quillen and Sullivan's rational homotopy theory. The comparison with local systems can be carried out directly here to prove Theorem D for these $\boldsymbol{E}_\infty$–rings.

4.4 Quasiaffine cases

We now consider a case where the descent spectral sequence enables us to *produce* nontrivial elements in the Picard group. Let A be a weakly even-periodic $\boldsymbol{E}_\infty$–ring with $\pi_0 A$ regular noetherian, and write $\omega = \pi_2 A$. Then A leads to a sheaf of $\boldsymbol{E}_\infty$–rings on the affine étale site of $\operatorname{Spec} \pi_0 A$. That is, for every étale $\pi_0 A$–algebra A'_0, there is (functorially) associated [39, Section 8.5] an $\boldsymbol{E}_\infty$–ring A' under A with $\pi_0 A' \simeq A'_0$ and A' flat over A. We will denote this sheaf by $\mathcal{O}^{\mathrm{top}}$.

Let $a_1, \dots, a_n \in \pi_0 A$ be a regular sequence, for $n \geq 2$. We consider the complement U in $\operatorname{Spec} \pi_0 A$ of the closed subscheme $V(a_1, \dots, a_n)$ and the sections $\bar{A} = \Gamma(U, \mathcal{O}^{\mathrm{top}})$. $\bar{A}$ is an $\boldsymbol{E}_\infty$–A–algebra and is a type of localization of A, albeit not (directly) an arithmetic one.[11] Note that $\operatorname{Pic}(A)$ is algebraic by Theorem 2.4.6, but the situation for $\bar{A}$ is more complicated.

The homotopy groups $\pi_*(\bar{A})$ are given by the abutment of a descent spectral sequence

$$H^s(U, \omega^{\otimes t}) \Rightarrow \pi_{2t-s}(\bar{A}). \tag{4-3}$$

We can first determine the zero-line. We have

$$H^0(U, \omega^{\otimes t}) = H^0(\operatorname{Spec} \pi_0 A, \omega^{\otimes t}),$$

because $\operatorname{Spec} \pi_0 A$ is regular and $U \subset \operatorname{Spec} \pi_0 A$ is obtained by removing a subscheme of codimension at least two.

Proposition 4.4.1 *The only other nonzero term in the descent spectral sequence* (4-3) *occurs for $s = n-1$. The descent spectral sequence degenerates.*

Proof Cover the scheme U by the n open affine subsets $U_i = \operatorname{Spec} \pi_0(A) \setminus V(a_i)$, for $1 \leq i \leq n$. Given any quasicoherent sheaf $\mathcal{F}$ on U, it follows that the coherent cohomology $H^*(U, \mathcal{F})$ is that of the Čech complex (which starts in degree zero)

$$\bigoplus_{i=1}^n \mathcal{F}(U_i) \to \bigoplus_{i<j} \mathcal{F}(U_i \cap U_j) \to \cdots \to \mathcal{F}(U_1 \cap \cdots \cap U_n).$$

[11] Forthcoming work of Bhatt and Halpern-Leistner identifies the universal property of $\bar{A}$.

Let $R = \pi_0 A$, and suppose $\mathcal{F}$ is the restriction to $U \subset \operatorname{Spec} R$ of the quasicoherent sheaf $\widetilde{M}$ on $\operatorname{Spec} R$ for an R–module M. Then the final term is the cokernel of the map

$$\bigoplus_{i=1}^{n} M[(a_1 \cdots \widehat{a_i} \cdots a_n)^{-1}] \to M[(a_1 \cdots a_n)^{-1}],$$

where the hat denotes omission. If M is flat, the complex is exact away from degrees 0 and $n-1$ as the sequence $a_1, \dots, a_n$ is regular, using a Koszul complex argument (see [28] for a detailed treatment or [18] for a short exposition with a view towards topological applications), and the zeroth cohomology is given by M itself.

Now, in view of the map $A \to \bar{A}$, clearly everything in the zero-line of the E_2–page of the spectral sequence survives, so the spectral sequence must degenerate. □

We now study the Picard group of $\bar{A}$: as above, $\pi_* \bar{A}$ is not regular but instead has a great deal of square-zero material. Let $\mathfrak{U} = (U, \mathcal{O}^{\mathrm{top}}|_U)$ denote the derived scheme consisting of the topological space $U \subset \operatorname{Spec} \pi_0 A$, but equipped with the sheaf $\mathcal{O}^{\mathrm{top}}$ of $\boldsymbol{E}_\infty$–rings restricted to U. $\bar{A}$ arises as the global sections of the structure sheaf $\mathcal{O}^{\mathrm{top}}$ over the derived scheme $\mathfrak{U}$.

Since U is quasiaffine as an (ordinary!) scheme, it follows by [42, Corollary 3.24] that the global sections functor is the right adjoint of an inverse equivalence

$$\operatorname{Mod}(\bar{A}) \rightleftarrows \operatorname{QCoh}(\mathfrak{U}),$$

of symmetric monoidal ∞–categories. In particular, the Picard group $\operatorname{Pic}(\bar{A})$ can be computed as $\operatorname{Pic}(\operatorname{QCoh}(\mathfrak{U}))$.

As before, we have a descent spectral sequence (3-2) converging to $\pi_{t-s}\,\mathfrak{pic}(\bar{A})$. But from (3-2), we know that almost all of the terms at E_2 are identified with the descent spectral sequence for $\pi_* \bar{A}$. In addition, we know that $H^1(U, \mathcal{O}_U^\times) \simeq \operatorname{Pic}(\pi_0 A)$, as $\pi_0 A$ is regular and the complement of U has codimension ≥ 2. These classes must be permanent cycles as they are realized in $\operatorname{Pic}(\bar{A})$: in fact, they are realized in $\operatorname{Pic}(A)$ itself. Thus, the descent spectral sequence for $\mathfrak{pic}$ degenerates as well. We get three contributions to the Picard group: $\mathbb{Z}/2$ and $\operatorname{Pic}(\pi_0 A)$, which together build $\operatorname{Pic}(\pi_* A)$ (compare Remark 2.4.8), and a group that is identified with $\pi_{-1} \bar{A}$. The relevant extension problem is solved because of the map $\operatorname{Pic}(\pi_* A) \cong \operatorname{Pic}(A) \to \operatorname{Pic}(\bar{A})$ realizing the algebraic part of the Picard group. We get:

Theorem 4.4.2 *Let $\bar{A} = \Gamma(U, \mathcal{O}^{\mathrm{top}})$ as above. Then we have a natural isomorphism*

$$\operatorname{Pic}(\bar{A}) \simeq \operatorname{Pic}(\pi_* A) \times \pi_{-1}(\bar{A}).$$

Moreover, observe that

$$\text{(4-4)} \quad \pi_{-1}(\bar{A}) = \begin{cases} \operatorname{coker}\Big(\bigoplus_{i=1}^{n} \omega^{n/2-1}[(a_1 \cdots \widehat{a_i} \cdots a_n)^{-1}] \\ \qquad\qquad \to \omega^{n/2-1}[(a_1 \cdots a_n)^{-1}]\Big) & \text{for } n \geq 4 \text{ even,} \\ \qquad\qquad 0 & \text{for } n \text{ odd.} \end{cases}$$

Example 4.4.3 Let A be a Landweber-exact weakly even periodic $\boldsymbol{E}_\infty$–ring with $\pi_0 A$ regular noetherian; for instance, A could be Morava E–theory E_n. In this case, we take $a_1, \dots, a_k = p, v_1, \dots, v_{k-1}$, so that $\bar{A} \simeq L_k A$. This gives Theorem C as a special case of Theorem 4.4.2.

Part II Computational tools

5 The comparison tool in the stable range

This is a technical section in which we develop a tool that will enable us to compare many of the differentials in a Picard spectral sequence for Galois or étale descent with the analogous differentials in the corresponding descent spectral sequence before taking the Picard functor (ie for the $\boldsymbol{E}_\infty$–rings themselves). For example, in the Galois descent setting, we are given a G–Galois extension $A \to B$, and we know the descent, ie homotopy fixed point, spectral sequence for $A \simeq B^{hG}$. The tool we develop in this section will allow us to deduce many differentials in the homotopy fixed point spectral sequence for $(\mathfrak{pic}(B))^{hG}$.

For a spectrum or a pointed space X, and integers a, b, we denote by $\tau_{\geq a} X$, $\tau_{\leq b} X$ and $\tau_{[a,b]} X$ the truncations of X with homotopy groups in the designated range. Our main observation is that if R is any $\boldsymbol{E}_\infty$–ring, then for any $n \geq 2$, there is a natural equivalence of spectra

$$\tau_{[n,2n-1]} R \simeq \tau_{[n,2n-1]} \, \mathfrak{gl}_1(R).$$

This equivalence is natural at the level of ∞–categories, and enables us to identify a large number of differentials in descent spectral sequences for $\mathfrak{gl}_1$ and therefore also for $\mathfrak{pic}$. This observation, however, fails if we increase the range by 1, and an identification of the relevant discrepancy (as observed in such spectral sequences) will be the subject of the following section and the formula (6-1).

The main result of Section 5.1 is essentially a formulation of the classical concept of the "stable range" in ∞–categorical terms, as can be seen from the fact that the major ingredients of the proof are Freudenthal's suspension theorem as well as the existence of Whitehead products in the unstable setting. Nonetheless, our formulation will be extremely useful in the sequel.

5.1 Truncated spaces and spectra

Throughout, $n \geq 2$.

Definition 5.1.1 Let $\mathrm{Sp}_{[n,2n-1]} \subset \mathrm{Sp}$ denote the ∞–category of spectra with homotopy groups concentrated in degrees $[n, 2n-1]$. Let $\mathcal{S}_*$ denote the ∞–category of pointed spaces, and let $\mathcal{S}_{*,[a,b]} \subset \mathcal{S}_*$ denote the subcategory spanned by those pointed spaces whose homotopy groups are concentrated in the interval $[a, b]$.

The main goal of this subsection is to prove the following result identifying spaces and spectra whose homotopy groups are concentrated in a range of dimensions.

Theorem 5.1.2 *The functor $\Omega^\infty\colon \mathrm{Sp}_{[n,2n-1]} \to \mathcal{S}_*$ is fully faithful. The functor $\Omega^\infty\colon \mathrm{Sp}_{[n,2n-2]} \to \mathcal{S}_{*,[n,2n-2]}$ is an equivalence of ∞–categories.*

Proof Let X, $Y \in \mathrm{Sp}_{[n,2n-1]}$. We want to show that the natural map

$$\mathrm{Hom}_{\mathrm{Sp}}(X, Y) \to \mathrm{Hom}_{\mathcal{S}_*}(\Omega^\infty X, \Omega^\infty Y) \tag{5-1}$$

is a homotopy equivalence. By adjointness, we can identify this with the map

$$\mathrm{Hom}_{\mathrm{Sp}}(X, Y) \to \mathrm{Hom}_{\mathrm{Sp}}(\Sigma^\infty \Omega^\infty X, Y)$$

that arises from the counit map $\Sigma^\infty\Omega^\infty X \to X$. Observe that we have a natural equivalence $\mathrm{Hom}_{\mathrm{Sp}}(\Sigma^\infty\Omega^\infty X, Y) \simeq \mathrm{Hom}_{\mathrm{Sp}}(\tau_{\leq 2n-1}\Sigma^\infty\Omega^\infty X, Y)$ because Y is $(2n-1)$–truncated. In particular, to prove Theorem 5.1.2, it will suffice to show that the natural map of spectra

$$\tau_{\leq 2n-1}\Sigma^\infty\Omega^\infty X \to X \simeq \tau_{\leq 2n-1} X,$$

is an equivalence, for any $X \in \mathrm{Sp}_{[n,2n-1]}$. Equivalently, we need to show that for any such spectrum X, the map

$$\pi_k(\Sigma^\infty\Omega^\infty X) \to \pi_k(X) \tag{5-2}$$

is an isomorphism for $k \leq 2n-1$. But we have maps of *spaces*

$$\Omega^\infty X \to \Omega^\infty\Sigma^\infty\Omega^\infty X \to \Omega^\infty X,$$

where the composite is the identity. The first map is the unit $Y \to \Omega^\infty\Sigma^\infty Y$ applicable for any $Y \in \mathcal{S}_*$, and the second map is Ω^∞ applied to the counit. By the Freudenthal suspension theorem, the first map induces an isomorphism on homotopy groups π_k for $k \leq 2n-1$, and therefore the second map does as well. This proves the claim that (5-2) is an equivalence and the first part of the theorem.

The functor $\Omega^\infty\colon \mathrm{Sp}_{[n,2n-1]} \to \mathcal{S}_{*,[n,2n-1]}$ is not essentially surjective, because spaces with homotopy groups concentrated in degrees $[n, 2n-1]$ can still have *Whitehead products*, and spaces with nontrivial Whitehead products can never be in the image of Ω^∞. However, we claimed in the statement of the theorem that the functor $\Omega^\infty\colon \mathrm{Sp}_{[n,2n-2]} \to \mathcal{S}_{*,[n,2n-2]}$ is an equivalence of ∞–categories. To show this, it suffices to show that the functor is essentially surjective.

Given a pointed space X with homotopy groups in the desired range, we suppose inductively (on k) that $\tau_{\leq k}X$ is in the image of Ω^∞. If $k \geq 2n-2$, then we are done. Otherwise, we have a pullback square:

$$\begin{tikzcd}
\tau_{\leq k+1}X \arrow[r] \arrow[d] & * \arrow[d] \\
\tau_{\leq k}X \arrow[r] & K(\pi_{k+1}X, k+2)
\end{tikzcd}$$

Observe that the pointed spaces $\tau_{\leq k}X$, $K(\pi_{k+1}X, k+2)$ and $*$ are all in the image of Ω^∞ (the first by the inductive hypothesis), and $K(\pi_{k+1}X, k+2) \in \mathcal{S}_{*,[n,2n-1]}$. Moreover, the *maps* in the diagram are in the image of Ω^∞ by the previous part of the result. Therefore, the object $\tau_{\leq k+1}X$ is in the image of Ω^∞, as Ω^∞ preserves homotopy fiber squares. □

Given an integer k, we could precompose the functor of Theorem 5.1.2 with the equivalence $\Omega^k\colon \mathrm{Sp}_{[n+k,2n+k-1]} \to \mathrm{Sp}_{[n,2n-1]}$, and obtain the following:

Corollary 5.1.3 *For any integer k, the functor $\Omega^{\infty+k}\colon \mathrm{Sp}_{[n+k,2n+k-1]} \to \mathcal{S}_*$ is fully faithful.*

5.2 Comparisons for E_∞–rings

Our basic example for all this comes from the spectrum $\mathfrak{gl}_1(R)$ associated to an E_∞–ring R, and the comparison between the two. This comparison is the main obstacle in understanding the descent spectral sequence for the Picard group: it is generally easier to understand descent spectral sequences for the E_∞–rings themselves (eg for TMF).

We emphasize again that given an E_∞–ring R, the *spectra* R and $\mathfrak{gl}_1(R)$ are generally very different, and for an illustration we provide the following example.

Example 5.2.1 (Lawson [33]) Consider the commutative differential graded algebra $\mathbb{F}_2[x]/x^3$ where $|x| = 1$ and $dx = 0$ (so $d \equiv 0$). Let R be the associated $\boldsymbol{E}_\infty$–ring under $\mathbb{F}_2$. Then $\mathfrak{gl}_1(R)$ has homotopy groups in dimensions 1, 2 given by $\mathbb{F}_2$; however, they are connected by multiplication by η. In particular, $\mathfrak{gl}_1(R)$ is not an $\mathbb{F}_2$–module spectrum.

More generally, let R be the $\boldsymbol{E}_\infty$–ring associated to the commutative differential graded algebra $\mathbb{F}_2[x]/x^3$ where $|x| = n$ and $dx = 0$. R can also be constructed by applying the Postnikov section $\tau_{\leq 2n}$ to the free $\boldsymbol{E}_\infty$–$\mathbb{F}_2$–algebra on a class in degree n. Then $\pi_n(\mathfrak{gl}_1(R)) \simeq \pi_{2n}(\mathfrak{gl}_1(R)) \simeq \mathbb{F}_2$ and all the other homotopy groups of $\mathfrak{gl}_1(R)$ vanish. Therefore, $\mathfrak{gl}_1(R)$ is the fiber of a k–invariant map $H\mathbb{F}_2[n] \to H\mathbb{F}_2[2n+1]$. In this case, we can identify the k–invariant and thus identify $\mathfrak{gl}_1(R)$.

Proposition 5.2.2 *Given R as above, the k–invariant of $\mathfrak{gl}_1(R)$ is given by the map*

$$\mathrm{Sq}^{n+1}\colon H\mathbb{F}_2[n] \to H\mathbb{F}_2[2n+1].$$

Proof We begin by arguing, following Lawson, that $\mathfrak{gl}_1(R)$ cannot be the spectrum $H\mathbb{F}_2[n] \vee H\mathbb{F}_2[2n]$. In fact, in this case, the map of spectra $H\mathbb{F}_2[n] \to \mathfrak{gl}_1(R)$ would by adjointness [2] lead to a map of $\boldsymbol{E}_\infty$–rings

$$\Sigma^\infty_+ K(\mathbb{F}_2, n) \to R,$$

carrying the class in $\pi_n K(\mathbb{F}_2, n)$ to the nonzero class in $\pi_n R$. Smashing with $H\mathbb{F}_2$, we would get a map of $\boldsymbol{E}_\infty$–$H\mathbb{F}_2$–algebras

$$H\mathbb{F}_2 \wedge \Sigma^\infty_+ K(\mathbb{F}_2, n) \to R$$

with the same property. Now $\pi_n(H\mathbb{F}_2 \wedge \Sigma^\infty_+ K(\mathbb{F}_2, n)) \simeq \mathbb{F}_2$, with the nontrivial class coming from $\pi_n(K(\mathbb{F}_2, n))$. However, this class squares to zero by [9, Lemma 6.1, Chapter 1] while the nonzero class in $\pi_n R$ does not square to zero. This is a contradiction and proves that such a map cannot exist. Consequently, the k–invariant map for $\mathfrak{gl}_1(R)$ must be nontrivial.

On the other hand, $\Omega^\infty \mathfrak{gl}_1(R) \simeq K(\mathbb{F}_2, n) \times K(\mathbb{F}_2, 2n)$ because $\Omega^\infty \mathfrak{gl}_1(R)$ is the connected component at 1 of $\Omega^\infty R$. In particular, the k–invariant $H\mathbb{F}_2[n] \to H\mathbb{F}_2[2n+1]$ defines, upon applying Ω^∞, the trivial cohomology class in $H^{2n+1}(K(\mathbb{F}_2, n); \mathbb{F}_2)$.

So, for the k–invariant of $\mathfrak{gl}_1(R)$, we need a nonzero element ϕ of degree $n+1$ in the (mod 2) Steenrod algebra such that, if $\iota_n \in H^n(K(\mathbb{F}_2, n); n)$ is the tautological class, then $\phi\iota_n = 0$. By the calculation of the cohomology of Eilenberg–Mac Lane spaces [61] (see also [50, Chapter 9] for a textbook reference), the only possibility is Sq^{n+1}. □

Nonetheless, we will show that right below the range of the previous example, the spectra $\mathfrak{gl}_1(R)$ and R can be identified.

Corollary 5.2.3 *Let $n \geq 2$ and let R be any E_∞–ring. Then there is an equivalence of spectra, functorial in R,*

$$\tau_{[n,2n-1]}\, \mathfrak{gl}_1(R) \simeq \tau_{[n,2n-1]} R.$$

Similarly, there is an equivalence of spectra, functorial in R,

$$\tau_{[n+1,2n]}\, \mathfrak{pic}(R) \simeq \Sigma\tau_{[n,2n-1]} R.$$

Proof For any E_∞–ring R, the space $\Omega^\infty \mathfrak{gl}_1(R) = \mathrm{GL}_1(R)$ is a union of those components of $\Omega^\infty R$ that correspond to units in $\pi_0 R$. In particular, $\Omega^\infty \tau_{\geq 1}\, \mathfrak{gl}_1(R)$ is *canonically* identified with $\Omega^\infty \tau_{\geq 1} R$ in $\mathcal{S}_*$. Applying Theorem 5.1.2, we now get a canonical identification as desired in the corollary. The second half of Corollary 5.2.3 follows from the first, as $\tau_{\geq 0}\Omega\, \mathfrak{pic}(R) \simeq \mathfrak{gl}_1(R)$ as spectra. □

Take now a faithful G–Galois extension $A \to B$ of E_∞–rings, and consider the HFPSS (3-5) for the G–action on $\mathfrak{pic}(B)$. We want to understand $\pi_0(\mathfrak{pic}(B)^{hG})$, or equivalently $\pi_{-1}(\Omega\, \mathfrak{pic}(B)^{hG})$, and we can do this by understanding the HFPSS for the G–action on $\Omega\, \mathfrak{pic}(B)$. Observe first that $\pi_t \Omega\, \mathfrak{pic}(B) \simeq \pi_t B$ functorially for $t \geq 1$: in fact, $\Omega^\infty(\Omega\, \mathfrak{pic}(B)) \simeq \mathrm{GL}_1(B)$. In other words, the spectrum $\Omega\, \mathfrak{pic}(B)$ equipped with the G–action has the property that, after applying Ω^∞, it is identified with a union of connected components of $\Omega^\infty B$ (with the G–action on B).

As a result, we have a map of spaces with G–action

$$\Omega^\infty(\Omega\, \mathfrak{pic}(B)) \to \Omega^\infty B,$$

which identifies the former with a union of connected components of the latter. As a result, we can identify the respective HFPSS for the spaces $\Omega^\infty(\Omega\, \mathfrak{pic}(B))$, $\Omega^\infty B$ for $t > 0$, both at E_2 and differentials (including the "fringed" ones). This identification comes from the map $\tau_{\geq 1}\, \mathrm{GL}_1(B) \to \Omega^\infty B$ given by subtracting one.

In particular, shifting by one again, most of the differentials in the HFPSS for $\mathfrak{pic}(B)$ are determined by the HFPSS for B. More precisely, any differential out of $E_r^{s,t}$ for $t - s > 0$, $s > 0$, depends only on the G–space $\Omega\, \mathcal{P}\mathrm{ic}(B)$, so the equivalence of $\Omega\, \mathcal{P}\mathrm{ic}(B)$ with a union of connected components of $\Omega^\infty B$ implies that the differential *can be identified* with the analogous differential in the HFPSS for B.

However, to understand $\pi_0(\mathfrak{pic}(B)^{hG}) \simeq \pi_0(\mathcal{P}\mathrm{ic}(B)^{hG}) \simeq \mathrm{Pic}(A)$, we need to determine differentials out of $E_r^{s,t}$ with $t = s$. These differentials cannot be determined by $\Omega\,\mathcal{P}\mathrm{ic}(B)$, as a space with a G–action. Our strategy to determine these differentials is to use the equivalence of spectra with G–action

$$\tau_{[n+1,2n]}\,\mathfrak{pic}(B) \simeq \Sigma\tau_{[n,2n-1]}B,$$

which is a special case of Corollary 5.2.3.

Assume that $r \le t-1$. In this case, any differential $d_r\colon E_*^{s,t} \to E_*^{s+r,t+r-1}$ in the HFPSS for $\mathfrak{pic}(B)$ is determined by the G–action on $\tau_{[t,t+r-1]}\,\mathfrak{pic}(B)$. Since we have an equivalence $\tau_{[t,t+r-1]}\,\mathfrak{pic}(B) \simeq \Sigma\tau_{[t-1,t+r-2]}B$, compatible with the G–actions, we can identify the differentials.

Denote the differentials in the homotopy fixed point spectral sequence

$$H^s(G,\pi_t\,\mathfrak{pic}\,B) \Rightarrow \pi_{t-s}(\mathfrak{pic}\,B)^{hG}$$

by $d_r^{s,t}(\mathfrak{pic}\,B)$, and similarly $d_r^{s,t}(B)$ for those in the HFPSS for B. The upshot of this discussion is the following.

Comparison Tool 5.2.4 *Let $A \to B$ be a G–Galois extension of $\boldsymbol{E}_\infty$–rings. Whenever $2 \le r \le t-1$, we have an equality of differentials $d_r^{s,t}(\mathfrak{pic}\,B) = d_r^{s,t-1}(B)$.*

Of course, we also have an identification of differentials out of (s,t) if $t-s>0$, $s>0$.

Remark 5.2.5 Our original approach to the Comparison Tool 5.2.4 was somewhat more complicated than the above and has been described in [44]. Namely, our strategy was to identify the HFPSS with a Bousfield–Kan spectral sequence for a certain cosimplicial space $X^\bullet$ built from $\mathcal{P}\mathrm{ic}(B)$ with its G–action, and argue that these differentials only depended on the fiber of $\mathrm{Tot}_{t+r}(X^\bullet) \to \mathrm{Tot}_{t-1}(X^\bullet)$ (as well as the other fibers in between). In the appropriate range, these fibers depend only on $\Omega X^\bullet$ as a cosimplicial space. However, $\Omega X^\bullet$ can be (almost) identified with the analogous cosimplicial space for the G–action on $\Omega^{\infty-1}(\tau_{\ge 0}B)$ because $\Omega\,\mathcal{P}\mathrm{ic}(B)$ is a union of components of $\Omega^\infty B$. This forces the differentials to correspond to one another.

For the same reasons, we have analogous comparison results for the spectral sequence as Theorem 3.2.1. Again, any differential in the descent spectral sequence for $\mathfrak{pic}(\Gamma(X,\mathcal{O}^{\mathrm{top}}))$ that only depends on the *diagram* $\tau_{[n+1,2n]}\,\mathfrak{pic}(\mathcal{O}^{\mathrm{top}})$ can be identified with the corresponding differential in the descent spectral sequence for $\Gamma(X,\mathcal{O}^{\mathrm{top}})$, thanks to the equivalence of *diagrams* of spectra $\tau_{[n+1,2n]}\,\mathfrak{pic}(\mathcal{O}^{\mathrm{top}}) \simeq \Sigma\tau_{[n,2n-1]}\,\mathcal{O}^{\mathrm{top}}$.

Remark 5.2.6 The equivalence $\tau_{[n,2n-1]}R \simeq \tau_{[n,2n-1]}\,\mathfrak{gl}_1(R)$ resembles the following observation in commutative algebra. Let A be an ordinary commutative ring and let $I \subset A$ be a square-zero ideal. Then $1 + I \subset A^\times$ and there is an isomorphism of groups

$$I \simeq 1 + I \subset A^\times \quad \text{with } x \mapsto 1 + x.$$

This correspondence is a very degenerate version of the exponential and logarithm.

Suppose p is a prime number and $(p-1)!$ is invertible in A. Then if $J \subset A$ is an ideal with $J^p = 0$, we have $1 + J \subset A^\times$ and a natural isomorphism of groups

$$J \simeq 1 + J \quad \text{with } x \mapsto 1 + x + \frac{x^2}{2} + \cdots + \frac{x^{p-1}}{(p-1)!},$$

given by a p–truncated exponential.

Similarly, let R be an $\boldsymbol{E}_\infty$–ring with $(p-1)!$ invertible. Motivated by the above, for any $n \geq 1$, one could surmise a *functorial* equivalence of spectra

$$\tau_{[n,pn-1]} R \simeq \tau_{[n,pn-1]} \mathfrak{gl}_1(R).$$

We expect to construct such an equivalence in ongoing joint work with Clausen and Heuts.

5.3 A general result on Galois descent

As a quick application of the preceding ideas, we can prove a general result about Galois descent for Picard groups.

Theorem E *Let $A \to B$ be a faithful G–Galois extension of $\boldsymbol{E}_\infty$–rings. Then the relative Picard group of B/A is $|G|$–power torsion of finite exponent.*

Proof We know that the relative Picard group of $A \to B$ is given by $\pi_{-1}(\mathfrak{gl}_1(B)^{hG})$ (compare Remark 3.3.2). There is a HFPSS that converges to the homotopy groups, which begins with the group cohomology of G with coefficients in $\pi_*(\mathfrak{gl}_1(B))$. Every contributing term is $|G|$–power torsion: in fact, every term is a $H^i(G, \cdot)$ for $i > 0$ and is thus killed by $|G|$. However, in view of the potential infiniteness of the filtration, as well as the possibilities of nontrivial extensions, this alone does not force $\pi_{-1}(\mathfrak{gl}_1(B)^{hG})$ to be $|G|$–power torsion.

Our strategy is to compare the HFPSS for $\pi_{-1}(\mathfrak{gl}_1(B)^{hG})$ with that of $\pi_{-1}(B^{hG})$. The map $A \to B$ admits descent in the sense of [40, Definition 3.17]. In particular, by [40, Corollary 4.4], the descent spectral sequence for $A \to B$ (equivalently, the HFPSS) has a *horizontal vanishing line* at a finite stage. It follows that, above a certain filtration, everything in the HFPSS for $\pi_*(A) \simeq \pi_*(B^{hG})$ is killed by a d_k for k bounded.

In view of our Comparison Tool 5.2.4, it follows that any class in the relative Picard group has bounded filtration (though possibly the bound is weaker than the analog in $\pi_{-1}(B)$). Since every contributing term in the spectral sequence is killed by $|G|$, the theorem follows. □

6 The first unstable differential

6.1 Context

Let $R^\bullet$ be a cosimplicial $\boldsymbol{E}_\infty$–ring, and consider the Bousfield–Kan spectral sequences (BKSS) $\{E_r^{s,t}\}$ and $\{\overline{E}_r^{s,t}\}$ for the two cosimplicial objects $R^\bullet$ and $\mathfrak{gl}_1(R^\bullet)$, converging to π_{t-s} of the respective totalizations in Sp.

For $t-s\geq 0$, the spectral sequences and the differentials are mostly identified with one another, as the space $\Omega^\infty\,\mathfrak{gl}_1(R)$ is a union of connected components of $\Omega^\infty R$. But for $t-s=-1$, we get differentials

$$d_r\colon E_r^{t+1,t}\to E_r^{t+r+1,t+r-1}\quad\text{and}\quad \overline{d}_r\colon \overline{E}_r^{t+1,t}\to \overline{E}_r^{t+r+1,t+r-1}.$$

These depend on more than the spaces $\Omega^\infty R^\bullet$, $\Omega^\infty\,\mathfrak{gl}_1(R^\bullet)$: they require the one-fold deloopings. As we saw in Corollary 5.2.3, for any $n\geq 2$, in the range $[n,2n-1]$, the cosimplicial *spectra* $\tau_{[n,2n-1]}R^\bullet$ and $\tau_{[n,2n-1]}\,\mathfrak{gl}_1(R^\bullet)$ are identified. As a result, for $r\leq t$, the groups in question are (canonically) identified and $d_r=\overline{d}_r$.

But in general, $d_{t+1}\neq \overline{d}_{t+1}$. Since all the previous differentials entering or leaving this spot between the two spectral sequences were identified, the groups in question are identified. We let the correspondence $E_{t+1}^{t+1,t}\simeq \overline{E}_{t+1}^{t+1,t}$ be given as

$$x\mapsto \overline{x}.$$

Similarly, we have a correspondence $E_{t+1}^{2t+2,2t}\simeq \overline{E}_{t+1}^{2t+2,2t}$.

In this subsection, we will give a universal formula for the first differential out of the stable range. We will need this in Section 8.2 to obtain the 2–primary Picard group of TMF.

Theorem 6.1.1 *We have the formula*

$$\overline{d}_{t+1}(\overline{x})=\overline{d_{t+1}(x)+x^2}\quad \textit{for } x\in E_{t+1}^{t+1,t}. \tag{6-1}$$

Remark 6.1.2 The above formula actually makes $\overline{d}_{t+1}$ into a linear operator. This follows from the graded-commutativity of the BKSS for $R^\bullet$. Note in particular that the difference between $\overline{d}_{t+1}$ and d_{t+1} is annihilated by 2.

6.2 The universal example

The proof of (6-1) follows a standard technique in algebraic topology: we reduce to a "universal" case and show that (6-1) is essentially the only possibility. We want to consider the universal case of a cosimplicial $\boldsymbol{E}_\infty$–ring $R^\bullet$ with a class in $E_{t+1}^{t+1,t}$. This class represents an element in $\pi_{-1}\operatorname{Tot}_{2t+1}(R^\bullet)$ trivialized in $\operatorname{Tot}_t(R^\bullet)$; the differential d_{t+1} represents the obstruction to lifting to $\operatorname{Tot}_{2t+2}$. So, we need to make the analysis of differentials in the cosimplicial $\boldsymbol{E}_\infty$–ring which corepresents the functor $R^\bullet \mapsto \mathfrak{A}(R^\bullet) = \Omega^\infty\big(\Sigma^{-1}\mathrm{fib}(\operatorname{Tot}_{2t+1}(R^\bullet) \to \operatorname{Tot}_t(R^\bullet))\big)$.

The relevant cosimplicial $\boldsymbol{E}_\infty$–ring $\mathscr{X}^\bullet$ can be constructed as follows.

Definition 6.2.1 Let Lan denote the operation of left Kan extension; let $\mathrm{Lan}_{\Delta^{\le t}\to\Delta}(*)$ denote the left Kan extension of the constant functor $\Delta^{\le t} \to \mathcal{S}$ at a point to Δ. Similarly, define $\mathrm{Lan}_{\Delta^{\le 2t+1}\to\Delta}(*)$. Consider the homotopy pushout

(6-2)
$$\begin{tikzcd}
\mathrm{Lan}_{\Delta^{\le t}\to\Delta}(*)_+ \arrow[r] \arrow[d] & * \arrow[d] \\
\mathrm{Lan}_{\Delta^{\le 2t+1}\to\Delta}(*)_+ \arrow[r] & \mathscr{F}^\bullet
\end{tikzcd}$$

where $\mathscr{F}^\bullet\colon \Delta \to \mathcal{S}_*$ is a functor to the ∞–category $\mathcal{S}_*$ of *pointed* spaces.

Consider
$$\mathscr{G}^\bullet \stackrel{\text{def}}{=} \Sigma^{\infty-1}\mathscr{F}^\bullet\colon \Delta \to \mathrm{Sp}$$
and the functor
$$\mathscr{X}^\bullet = \mathrm{Free}_{\mathrm{CAlg}}(\mathscr{G}^\bullet)\colon \Delta \to \mathrm{CAlg}$$
into the ∞–category CAlg of $\boldsymbol{E}_\infty$–rings, obtained by applying the free algebra functor everywhere to $\mathscr{G}$. By construction, $\mathscr{X}^\bullet$ corepresents the functor $\mathfrak{A}\colon \mathrm{Fun}(\Delta, \mathrm{CAlg}) \to \mathcal{S}$ in which we are interested. In particular, it suffices to prove (6-1) for this particular functor. As we will see in the next paragraph, $\mathscr{G}^\bullet$ takes values in *connective* spectra and therefore so does $\mathscr{X}^\bullet$. Since we are only interested in differentials in a particular range, we may (by naturality) only consider the Postnikov section $\tau_{\le 2t}\mathscr{X}^\bullet$. We get the following basic step.

Proposition 6.2.2 *In order to prove Theorem 6.1.1, it suffices to prove it for $\tau_{\le 2t}\mathscr{X}^\bullet$ (and the tautological class).*

In fact, we have a reasonable handle on what the functor $\tau_{\le 2t}\mathscr{X}^\bullet$ looks like and can *entirely* determine the BKSS. To see this, we recall the construction of $\mathscr{F}^\bullet$; compare also the discussion in [44]. The functor

$$\operatorname*{Lan}_{\Delta^{\le t}\to\Delta}(*)\colon \Delta \to \mathcal{S}$$

sends any finite nonempty totally ordered set T to the nerve of the category $\Delta^{\le t}_{/T}$ of all order-preserving morphisms $\{S \to T\}$ where

(1) S is a finite, nonempty totally ordered set, and

(2) $|S| \le t+1$.

Proposition 6.2.3 $\mathrm{Lan}_{\Delta_{\le t}\to\Delta}(*)$ *is naturally equivalent to the functor which sends T in Δ to the nerve of the* poset $P_{\le t+1}(T)$ *of nonempty subsets of T of cardinality at most $t+1$.*

Proof In fact, for any T, there is a natural map $P_{\le t+1}(T) \to \Delta^{\le t}_{/T}$, which is a homotopy equivalence as it is right adjoint to the functor $\Delta^{\le t}_{/T} \to P_{\le t+1}(T)$ which sends $S \to T$ to $\mathrm{image}(S \to T) \subset T$. □

In view of the last proposition, one can also consider the following approach to the left Kan extension. There is a standard cosimplicial simplicial set sending $[n] \mapsto \Delta^n$. The functor of the proposition is equivalent to the barycentric subdivision of the cosimplicial simplicial set $[n] \mapsto \mathrm{sk}_t\, \Delta^n$.

As in [44], the nerve of $P_{\le t+1}(T)$, for any choice of T, is (pointwise) homotopy equivalent to a wedge of t–spheres, and contractible if $|T| \le t+1$. We get from (6-2):

Proposition 6.2.4 *The functor $\mathscr{F}^\bullet\colon \Delta \to \mathcal{S}_*$ constructed above has the following properties:*

(1) *For any T, $\mathscr{F}(T)$ is always a wedge of copies of S^{t+1} and S^{2t+1}.*

(2) *Restricted to $\Delta^{\le t}$, the functor $\mathscr{F}^\bullet$ is contractible. Restricted to $\Delta^{\le 2t}$, the functor $\mathscr{F}^\bullet$ is pointwise a wedge of copies of S^{t+1}.*

6.3 Some technical lemmas

Our first goal is to understand the BKSS for $\mathscr{G}^\bullet = \Sigma^{\infty-1}\mathscr{F}^\bullet$. Observe that pointwise, this cosimplicial spectrum is a wedge of copies of S^t and S^{2t} by Proposition 6.2.4. In order to do this, we need to understand the cosimplicial abelian group $\pi_*(\Sigma^{\infty-1}\mathscr{F}^\bullet)$. We will prove the following:

Proposition 6.3.1 *The cohomology $H^s(\pi_*(\mathscr{G}^\bullet))$ is given by*

$$H^s(\pi_*(\mathscr{G}^\bullet)) \simeq \begin{cases} \pi_* S^t & \text{if } s = t+1, \\ \pi_* S^{2t} & \text{if } s = 2(t+1). \end{cases} \tag{6-3}$$

In the spectral sequence, the differential d_{t+1} is an isomorphism.

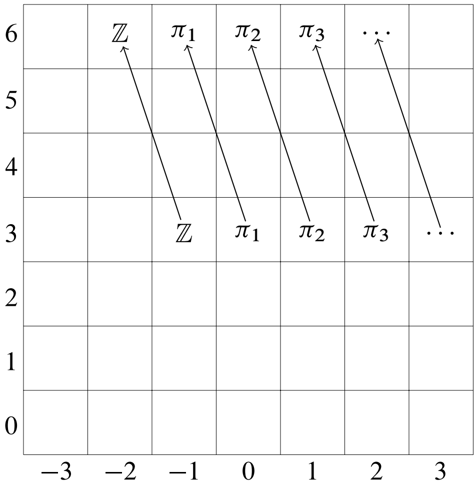


Figure 1: Bousfield–Kan spectral sequence for $\mathscr{G}^\bullet$, with $t = 2$ (π_k denotes $\pi_k S^0$)

The spectral sequence is depicted in Figure 1. The proof of Proposition 6.3.1 will take work and will be spread over two subsections. In the present subsection, our main result is that the totalization of $\mathscr{G}^\bullet$ (and related cosimplicial spectra) is contractible, and we will deduce the differentials from that. The approach to this is not computational and relies instead on ideas involving the ∞–categorical Dold–Kan correspondence of Lurie.

We recall from [34, Notation 1.2.8.4] the *cone* construction, which associates to a simplicial set K, the *cone* $K^{\lhd}$. If K is an ∞–category, $K^{\lhd}$ is as well, and is obtained by adding a new initial object to K.

Lemma 6.3.2 *Let K be a simplicial set and $\mathcal{D}$ an ∞–category with colimits. Let $F\colon K^{\lhd} \to \mathcal{D}$ be a functor with the property that F carries the cone point to an initial object of $\mathcal{D}$. Then the natural map*

$$\varinjlim_{K} F|_K \to \varinjlim_{K^{\lhd}} F$$

is an equivalence in $\mathcal{D}$.

Proof It suffices to show[12] that the natural map

$$\mathcal{D}_{K^{\lhd}/} \to \mathcal{D}_{K/} \tag{6-4}$$

is an equivalence of ∞–categories. But we have $\mathcal{D}_{K^{\lhd}/} \simeq \mathcal{D}_{(\Delta^0 \star K)/} \simeq (\mathcal{D}_{\Delta^0/})_{K/}$ in view of the definition of the overcategory [34, Section 1.2.9], where $\star$ denotes the *join* of simplicial sets [34, Section 1.2.8]. However, we also know that the projection map $\mathcal{D}_{\Delta^0/} \to \mathcal{D}$ is an equivalence since $\Delta^0 \to \mathcal{D}$ maps to an initial object. Therefore, we obtain that (6-4) is an equivalence, as desired. □

Lemma 6.3.3 *Let $\mathcal{C}$, $\mathcal{D}$ be ∞–categories and assume that $\mathcal{D}$ has colimits. Let $F\colon \mathcal{C}^{\lhd} \to \mathcal{D}$ be a functor such that F carries the cone point to an initial object of $\mathcal{D}$. Let $\mathcal{C}' \subset \mathcal{C}$ be a full subcategory. Then the following are equivalent:*

1. *$F|_{\mathcal{C}}$ is a left Kan extension of its restriction to $\mathcal{C}'$.*
2. *F is a left Kan extension of its restriction to $\mathcal{C}'^{\lhd}$.*

Proof Suppose the first condition is satisfied. Then if $c \in \mathcal{C}$ is arbitrary, the natural map

$$\varinjlim_{c' \to c \in \mathcal{C}'_{/c}} F(c') \to F(c)$$

is an equivalence. Now, we have an equivalence of ∞–categories $(\mathcal{C}'_{/c})^{\lhd} \simeq (\mathcal{C}'^{\lhd})_{/c}$, because $\lhd$ adds a new initial object. Therefore, for arbitrary $c \in \mathcal{C}$, we also get that the natural map

$$\varinjlim_{c' \to c \in (\mathcal{C}'^{\lhd})_{/c}} F(c') \simeq \varinjlim_{c' \to c \in (\mathcal{C}'_{/c})^{\lhd}} F(c') \to F(c)$$

is an equivalence, thanks to Lemma 6.3.2. At the cone point, the left Kan extension condition is automatic. Thus, it follows that F is a left Kan extension of $F|_{\mathcal{C}'^{\lhd}}$. The converse is proved in the same way. □

Proposition 6.3.4 *Let $\mathcal{C}$ be a stable ∞–category and let $F\colon \Delta^{\leq n} \to \mathcal{C}$ be any functor. Suppose F is a left Kan extension of its restriction to $\Delta^{\leq n-1}$. Then $\varprojlim_{\Delta^{\leq n}} F$ is contractible.*

Proof Observe that the cone $(\Delta^{\leq n})^{\lhd}$ is given by the category $\Delta_+^{\leq n}$ of the finite totally ordered sets $\{[i]\}_{-1 \leq i \leq n}$ since $[-1]$ is an initial object of this category. Consider the functor $\widetilde{F}\colon \Delta_+^{\leq n} \simeq (\Delta^{\leq n})^{\lhd} \to \mathcal{C}$ extending F that sends the cone point to the initial

[12] We are indebted to the referee for substantially simplifying our original argument here.

object (one can always make such an extension). To show that $\varprojlim_{\Delta^{\leq n}} F$ is contractible, it suffices to show that $\widetilde{F}$ is a right Kan extension of $F = \widetilde{F}|_{\Delta^{\leq n}}$.

Now, we recall a basic result of Lurie [39, Lemma 1.2.4.19] (which we use for the opposite category), a piece of the ∞–categorical version of the Dold–Kan correspondence: given any functor $G\colon \Delta_+^{\leq n} \to \mathcal{C}$, G is a right Kan extension of $G|_{\Delta^{\leq n}}$ if and only if G is a *left* Kan extension of $G|_{\Delta_+^{\leq n-1}}$. In our case, it follows that to show that $\widetilde{F}$ is a right Kan extension of F (as we would like to see), it suffices to show that $\widetilde{F}$ is a *left* Kan extension of $\widetilde{F}|_{\Delta_+^{\leq n-1}}$. But by Lemma 6.3.3, this follows from the fact that $\widetilde{F}|_{\Delta^{\leq n}} = F$ is a left Kan extension of $\widetilde{F}|_{\Delta^{\leq n-1}} = F|_{\Delta^{\leq n-1}}$. □

6.4 The BKSS for $\mathscr{F}$

The goal of this subsection is to complete the proof of Proposition 6.3.1. To begin with, we analyze the BKSS for the functor $\Sigma^\infty_+ \operatorname{Lan}_{\Delta^{\leq t} \to \Delta}(*)\colon \Delta \to \mathrm{Sp}$.

Proposition 6.4.1 *The BKSS for the cosimplicial spectrum $\Sigma^\infty_+ \operatorname{Lan}_{\Delta^{\leq t} \to \Delta}(*)$ satisfies*

$$E_2^{s,*} = H^s\bigl(\pi_*(\Sigma^\infty_+ \operatorname*{Lan}_{\Delta^{\leq t}\to\Delta}(*))\bigr) = \begin{cases} \pi_*(S^0) & \text{if } s = 0, \\ \pi_*(S^t) & \text{if } s = t+1. \end{cases} \tag{6-5}$$

The differential d_{t+1} is an isomorphism. (The result for $t = 2$ is displayed in Figure 2.)

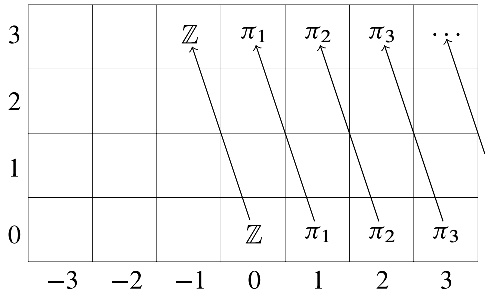


Figure 2: Bousfield–Kan spectral sequence for $\Sigma^\infty_+ \operatorname{Lan}_{\Delta^{\leq t} \to \Delta}(*)$, with $t = 2$

Proof Observe that $\operatorname{Lan}_{\Delta^{\leq t} \to \Delta}(*)$ is, pointwise, a wedge of t–spheres, so to compute the desired cohomology $H^s\bigl(\pi_*(\Sigma^\infty_+ \operatorname{Lan}_{\Delta^{\leq t} \to \Delta}(*))\bigr)$, it suffices to do this for π_t. (The disjoint basepoint contributes the $\pi_*(S^0)$ for $s = 0$ in cohomology.) In other words, we may consider the cosimplicial $H\mathbb{Z}$–module $M^\bullet = H\mathbb{Z} \wedge \Sigma^\infty_+ \operatorname{Lan}_{\Delta^{\leq t} \to \Delta}(*)$.

Now we know, for each n, that $\pi_*(M^n)$ is concentrated in degrees 0 and t, and that $\pi_0(M^\bullet)$ is the constant cosimplicial abelian group $\mathbb{Z}$. Moreover, by Proposition 6.3.4, $\mathrm{Tot}(M^\bullet)$ is contractible. A look at the spectral sequence for $\mathrm{Tot}(M^\bullet)$ shows that $H^s(\pi_t M^\bullet)$ must be concentrated in degree $s = t+1$ and must be a $\mathbb{Z}$ there. The claim about differentials also follows from contractibility of the totalization. □

Proof of Proposition 6.3.1 The definition (6-2) of $\mathcal{F}^\bullet$ and Proposition 6.4.1 together give the E_2–page of the spectral sequence, when one uses the long exact sequence in homotopy groups. The differentials are forced, again, by Proposition 6.3.4 which implies that $\mathrm{Tot}(\mathcal{G}^\bullet)$ is contractible. □

6.5 Completion of the proof

Now we need to consider the cosimplicial $\boldsymbol{E}_\infty$–ring defined earlier

$$\mathcal{Y}^\bullet \overset{\text{def}}{=} \tau_{\leq 2t}\mathcal{X}^\bullet \simeq \tau_{\leq 2t}\,\mathrm{Free}_{\mathrm{CAlg}}(\mathcal{G}^\bullet).$$

We recall that this is well defined as a cosimplicial $\boldsymbol{E}_\infty$–ring because $\mathcal{G}^\bullet$ is (pointwise) connective.

In this subsection, we will determine the relevant piece of the BKSS for $\mathcal{Y}$ and then complete the proof of Theorem 6.1.1. We have that

$$\mathcal{Y}^\bullet \simeq \tau_{\leq 2t} S^0 \vee \tau_{\leq 2t}\mathcal{G}^\bullet \vee \tau_{\leq 2t}((\mathcal{G}^\bullet)^{\wedge 2}_{h\Sigma_2}),$$

because, by a connectivity argument, no other terms contribute. In particular, the cohomology $H^s(\pi_*(\mathcal{Y}^\bullet))$ picks up a copy of $\pi_*(S^0)$ for $s=0$ (which is mostly irrelevant). In Proposition 6.3.1, we determined the BKSS for $\mathcal{G}^\bullet$; in bidegrees $(t+1,t)$ and $(2t+2,2t)$, this picks up copies of $\mathbb{Z}$ such that the first one hits the second one with a d_{t+1}. We will prove:

Proposition 6.5.1 *$E_2^{2t+2,2t} \simeq \mathbb{Z}\oplus\mathbb{Z}/2$ in the BKSS for $\mathcal{Y}^\bullet$. The $\mathbb{Z}/2$ is generated by the square of the class in bidegree $(t+1,t)$.*

Proof We will use the notation and results of Appendix C. Let $A^\bullet$ be the cosimplicial abelian group $\pi_t\mathcal{G}^\bullet$, which is levelwise free and finitely generated. As we have seen (Proposition 6.3.1), $H^{t+1}(A^\bullet)\simeq\mathbb{Z}$ and the other cohomology of $A^\bullet$ vanishes. Now, using the notation of Definition C.1,

$$\pi_{2t}(\mathcal{G}^{\bullet\wedge 2}_{h\Sigma_2}) = \begin{cases} \mathrm{Sym}_2 A^\bullet & \text{for } t \text{ even},\\ \widetilde{\mathrm{Sym}}_2 A^\bullet & \text{for } t \text{ odd}.\end{cases}$$

By Proposition C.3, we find that the $E_2^{2t+2,2t}$ term of $(\mathcal{G}^\bullet)^{\wedge 2}_{h\Sigma_2}$ is as claimed. □

We are now ready to complete the proof and determine the differential in the $\mathfrak{gl}_1$ spectral sequence. Using the notation of the beginning of this section, it follows that $E_{t+1}^{t+1,t} \simeq \mathbb{Z}$ and $E_{t+1}^{2t+1,2t} \simeq \mathbb{Z} \oplus \mathbb{Z}/2$, and similarly for $\overline{E}$. The d_{t+1} carries the $\mathbb{Z}$ into the other $\mathbb{Z}$. By naturality of the spectral sequence, it follows that there must exist a universal formula

$$\bar{d}_{t+1}(\bar{x}) = \overline{a d_{t+1}(x) + \epsilon x^2} \quad \text{for } a \in \mathbb{Z} \text{ and } \epsilon \in \{0,1\}. \tag{6-6}$$

The main claim is that $a = \epsilon = 1$. Our first goal is to compute a.

Lemma 6.5.2 *We have an equivalence of ∞–categories between the ∞–category $\mathrm{Fun}^L(\mathrm{Sp}_{\geq 0}, \mathrm{Sp}_{\geq 0})$ of cocontinuous functors $\mathrm{Sp}_{\geq 0} \to \mathrm{Sp}_{\geq 0}$ and $\mathrm{Sp}_{\geq 0}$ given by evaluating at the sphere. The inverse equivalence sends a connective spectrum Y to the functor $X \mapsto X \otimes Y$.*

Proof It suffices to show that evaluation at the sphere induces an equivalence of ∞–categories $\mathrm{Fun}^L(\mathrm{Sp}_{\geq 0}, \mathrm{Sp}) \simeq \mathrm{Sp}$ (with inverse given as above). But the ∞–category Sp is the *stabilization* [39, Section 1.4] of $\mathrm{Sp}_{\geq 0}$ (as one sees easily from the fact that Σ is *fully faithful* on $\mathrm{Sp}_{\geq 0}$ and an equivalence on Sp), so that, by [39, Corollary 1.4.4.5], we have an equivalence $\mathrm{Fun}^L(\mathrm{Sp}, \mathrm{Sp}) \simeq \mathrm{Fun}^L(\mathrm{Sp}_{\geq 0}, \mathrm{Sp})$ given by restriction. But we know that $\mathrm{Fun}^L(\mathrm{Sp}, \mathrm{Sp}) \simeq \mathrm{Sp}$ by evaluation at the sphere spectrum, with inverse given by the smash product; see [39, Section 4.8.2]. □

We need the following fact about $\mathfrak{gl}_1$.

Proposition 6.5.3 *Let X be a connective spectrum, and let $S^0 \vee X$ be the square-zero E_∞–ring. Then there is a natural equivalence of spectra,*

$$\mathfrak{gl}_1(S^0 \vee X) \simeq \mathfrak{gl}_1(S^0) \vee X.$$

On homotopy groups, this equivalence is compatible with the purely algebraic equivalence $\pi_t\, \mathfrak{gl}_1(S^0 \vee X) \simeq \pi_t(S^0 \vee X) \simeq \pi_t(S^0) \oplus \pi_t(X) \simeq \pi_t(\mathfrak{gl}_1(S^0)) \oplus \pi_t(X)$.

Proof Given the connective spectrum X, we can use the composite $S^0 \to S^0 \vee X \to S^0$, in which the second map sends X to 0, to get a natural splitting

$$\mathfrak{gl}_1(S^0 \vee X) \simeq \mathfrak{gl}_1(S^0) \vee F(X),$$

where $F\colon \mathrm{Sp}_{\geq 0} \to \mathrm{Sp}_{\geq 0}$ is a certain functor that we want to claim is naturally isomorphic to the identity. First, observe that F commutes with colimits. Namely, F

commutes with filtered colimits (as one can check on homotopy groups), F takes $*$ to $*$, and given a pushout square

$$\begin{tikzcd}
X_1 \arrow[r] \arrow[d] & X_2 \arrow[d] \\
X_3 \arrow[r] & X_4
\end{tikzcd} \tag{6-7}$$

in $\mathrm{Sp}_{\geq 0}$, the analogous diagram

$$\begin{tikzcd}
F(X_1) \arrow[r] \arrow[d] & F(X_2) \arrow[d] \\
F(X_3) \arrow[r] & F(X_4)
\end{tikzcd} \tag{6-8}$$

is a pushout square in $\mathrm{Sp}_{\geq 0}$. This in turn follows by considering long exact sequences in homotopy groups. More precisely, given the pushout square (6-7), the diagram of $\boldsymbol{E}_\infty$–rings

$$\begin{tikzcd}
S^0 \vee X_1 \arrow[r] \arrow[d] & S^0 \vee X_2 \arrow[d] \\
S^0 \vee X_3 \arrow[r] & S^0 \vee X_3
\end{tikzcd}$$

is a homotopy *pullback* in $\boldsymbol{E}_\infty$–rings, so that applying $\mathfrak{gl}_1$ (which is a *right adjoint*) leads to a pullback square

$$\begin{tikzcd}
\mathfrak{gl}_1(S^0 \vee X_1) \arrow[r] \arrow[d] & \mathfrak{gl}_1(S^0 \vee X_2) \arrow[d] \\
\mathfrak{gl}_1(S^0 \vee X_3) \arrow[r] & \mathfrak{gl}_1(S^0 \vee X_4)
\end{tikzcd}$$

and in particular, (6-8) is homotopy cartesian too in $\mathrm{Sp}_{\geq 0}$. Therefore, it is homotopy cocartesian as well if we can show that the map

$$\pi_0(\mathfrak{gl}_1(S^0 \vee X_3)) \oplus \pi_0(\mathfrak{gl}_1(S^0 \vee X_2)) \to \pi_0(\mathfrak{gl}_1(S^0 \vee X_4))$$

is surjective. This follows from the analogous fact that $\pi_0(X_3) \oplus \pi_0(X_2) \to \pi_0(X_4)$ is surjective as (6-7) is a pushout.

Therefore, as F commutes with colimits, F is necessarily of the form $X \mapsto X \otimes Y$ for some $Y \in \mathrm{Sp}_{\geq 0}$, by Lemma 6.5.2. For $X = H\mathbb{Z}$, we find $F(X) = H\mathbb{Z}$, so that

$H\mathbb{Z} \otimes Y$ is concentrated in degree zero and is isomorphic to $H\mathbb{Z}$. This forces $Y \simeq S^0$ and proves the claim. □

Proof of Theorem 6.1.1 Proposition 6.5.3 implies that in the universal formula (6-6), the constant $a = 1$. In fact, we know that if $X^\bullet$ is any cosimplicial spectrum, then the cosimplicial spectra $\mathfrak{gl}_1(S^0 \vee X^\bullet)$ and $\mathfrak{gl}_1(S^0) \vee X^\bullet$ are identified in a manner compatible with the identifications of homotopy groups. In particular, the differentials in the spectral sequence for $\mathfrak{gl}_1(S^0 \vee X^\bullet)$ and in the spectral sequence for $S^0 \vee X^\bullet$ are identified, forcing $a = 1$.

It remains to show that $\epsilon = 1$. For this, we need an example where the two differentials do *not* agree. This will be a generalization of Example 5.2.1. Consider the $\boldsymbol{E}_\infty$–ring R of Proposition 5.2.2, with $n = t$, so that, in particular, $\mathfrak{gl}_1(R)$ has homotopy groups in dimensions t and $2t$ only. Proposition 5.2.2 shows that the k–invariant is *nontrivial.*

Consider the space $X = K(\mathbb{F}_2, t+1)$, and consider the Atiyah–Hirzebruch spectral sequences for the homotopy groups of $\mathfrak{gl}_1(R)^X$ and R^X (these can be identified with BKSS's by choosing simplicial resolutions of X by points). The latter clearly degenerates because R is an Eilenberg–Mac Lane spectrum, but we claim that the former does not.

More precisely, we claim that there is no map of spectra

$$\Sigma^{-1}\Sigma^\infty K(\mathbb{F}_2, t+1) \to \mathfrak{gl}_1(R),$$

inducing an isomorphism on π_t. The degeneration of the AHSS would certainly imply the existence of such a map. To see this, it is equivalent to showing that there is no map of (pointed) spaces

$$K(\mathbb{F}_2, t+1) \to B\mathrm{GL}_1(R),$$

with the same properties. If there existed such a map, then we could combine it with the map $\tau_{\geq 2t+1} B\mathrm{GL}_1(R) \simeq K(\mathbb{F}_2, 2t+1) \to B\mathrm{GL}_1(R)$ via the infinite loop structure to obtain a map

$$K(\mathbb{F}_2, t+1) \times K(\mathbb{F}_2, 2t+1) \to B\mathrm{GL}_1(R),$$

which would be an equivalence by inspection of homotopy groups. However, this contradicts Proposition 5.2.2, which shows that the *space* $B\mathrm{GL}_1(R)$ has a nontrivial k–invariant.

This completes the proof of Theorem 6.1.1. □

Part III Computations

7 Picard groups of real K–theory and its variants

Before we embark on the lengthy computations for the Picard groups of the various versions of topological modular forms, let us work out in detail the case of real K–theory, as well as the Tate K–theory spectrum $\mathrm{KO}((q))$. In particular, these examples will illustrate our methodology without being computationally cumbersome.

7.1 Real K–theory

In this subsection, we compute the Picard group of KO using C_2–Galois descent from the C_2–Galois extension $\mathrm{KO} \to \mathrm{KU}$ and the Comparison Tool 5.2.4 (but not the universal formula of Theorem 6.1.1).

We begin with the basic case of *complex* K–theory.

Example 7.1.1 (complex K–theory) The complex K–theory spectrum has a very simple ring of homotopy groups $\mathrm{KU}_* = \mathbb{Z}[u^{\pm 1}]$ with u in degree 2. In particular, KU is even periodic with a regular noetherian π_0, so its Picard group is algebraic by Theorem 2.4.6. The inner workings of Theorem 2.4.6 would use that the only (homogeneous) maximal ideals of KU_* are generated by prime numbers p; for each p, there is a corresponding residue field spectrum, namely mod p K–theory, also known as an extension of the Morava K–theory of height one at the given prime. As the Picard group of $\mathrm{KU}_0 = \mathbb{Z}$ is trivial, and $\mathrm{Pic}(\mathrm{KU}_*) \simeq \mathbb{Z}/2$, any invertible KU–module is equivalent to either KU or $\Sigma\,\mathrm{KU}$.

To compute $\mathrm{Pic}(\mathrm{KO})$, we start with this knowledge that, thanks to Example 7.1.1, $\pi_0\,\mathfrak{pic}(\mathrm{KU}) = \mathrm{Pic}(\mathrm{KU})$ is $\mathbb{Z}/2$. We have the spectral sequence from (3-5)

$$H^*(C_2, \pi_*\,\mathfrak{pic}(\mathrm{KU})) \Rightarrow \pi_*(\mathfrak{pic}(\mathrm{KU}))^{hC_2}$$

which will allow us to compute $\pi_0(\mathfrak{pic}(\mathrm{KU}))^{hC_2} \simeq \mathrm{Pic}(\mathrm{KO})$. We note that

$$\pi_1\,\mathfrak{pic}(\mathrm{KU}) \simeq (\mathrm{KU}_0)^\times = \mathbb{Z}/2$$

and

$$H^*(C_2, \mathbb{Z}/2) = \mathbb{Z}/2[x],$$

where x is in cohomological degree 1. The higher homotopy groups of $\mathfrak{pic}(\mathrm{KU})$ coincide (as C_2–modules) with those of KU, suitably shifted by one.

Recall, moreover, that the E_2–page of the HFPSS for π_* KO is given by the bigraded ring

$$E_2^{*,*} = \mathbb{Z}[u^2, u^{-2}, h_1]/(2h_1) \quad \text{with } |u^2| = (4,0) \text{ and } |h_1| = (1,2),$$

where u^2 is the square of the Bott class in $\pi_*\,\mathrm{KU} \simeq \mathbb{Z}[u^{\pm 1}]$, and h_1 detects in homotopy the Hopf map η. The class h_1 is in bidegree $(s,t) = (1,2)$, so it is drawn using Adams indexing in the $(1,1)$ place. The differentials are determined by $d_3(u^2) = h_1^3$ and the spectral sequence collapses at E_4. For convenience, we reproduce a picture in Figure 3; the interested reader can find the detailed computation of this spectral sequence in [22, Section 5].

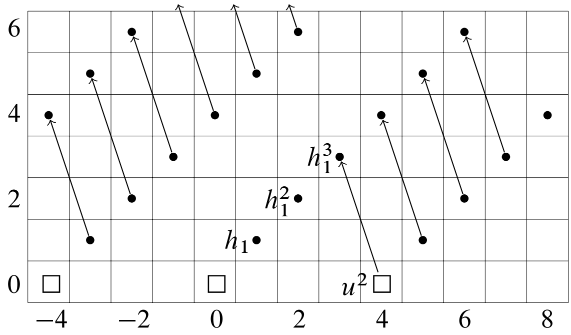


Figure 3: Homotopy fixed point spectral sequence for π_* KO $\simeq \pi_*(\mathrm{KU}^{hC_2})$ (● denotes $\mathbb{Z}/2$ and □ denotes $\mathbb{Z}$)

Therefore, the E_2–page of the spectral sequence for $(\mathfrak{pic}(\mathrm{KU}))^{hC_2}$ is as in Figure 4. To deduce differentials, we use our Comparison Tool 5.2.4: in the homotopy fixed point spectral sequence for KU, there are only (nontrivial) d_3–differentials. By the Comparison Tool 5.2.4, we conclude that we can "import" those differentials to the HFPSS for $\mathfrak{pic}(\mathrm{KU})$ when they involve terms with $t \geq 4$. In particular, we see that the differentials drawn in Figure 4 are nonzero; moreover, everything that is above the drawn range and in the $s = t$ column either supports or is the target of a nonzero differential. Note that we are not claiming that there are no other nonzero differentials, but these suffice for our purposes.

We deduce from this that $\pi_0\,\mathfrak{pic}(\mathrm{KU})^{hC_2} = \mathrm{Pic}(\mathrm{KO})$ has cardinality at most eight. On the other hand, the fact that KO is 8–periodic gives us a lower bound $\mathbb{Z}/8$ on Pic(KO). Thus we get:

Theorem 7.1.2 (Hopkins; Gepner and Lawson [15]) Pic(KO) *is precisely* $\mathbb{Z}/8$, *generated by* Σ KO.

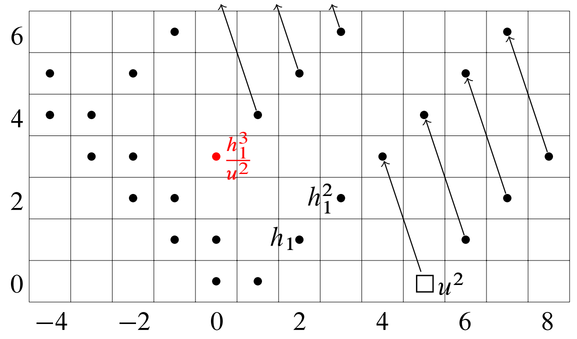


Figure 4: Homotopy fixed point spectral sequence for $\mathfrak{pic}(\mathrm{KU})^{hC_2}$

Theorem 7.1.2 was proved originally by Hopkins (unpublished) using related techniques. The approach via descent theory is due to Gepner and Lawson in [15]. Their identification of the differentials in the spectral sequence is, however, different from ours: they use an explicit knowledge of the structure of $\mathfrak{gl}_1(\mathrm{KU})$ with its C_2–action (which one does not have for TMF).

Remark 7.1.3 In view of Remark 3.3.2, we conclude that the relative Picard group of the C_2–extension $\mathrm{KO} \to \mathrm{KU}$ is $\pi_{-1}(\mathfrak{gl}_1\,\mathrm{KU})^{hC_2} \simeq \mathbb{Z}/4$.

Remark 7.1.4 In the usual descent spectral sequence for KO, the class h_1^3/u^2 (in red) supports a d_3. By Theorem 6.1.1 and the multiplicative structure of the usual SS, h_1^3/u^2 does *not* support a d_3 in the descent SS for Pic. We saw that above by counting: if h_1^3/u^2 did not survive, the Picard group of KO would be too small. For 2–local TMF, simple counting arguments will not suffice and we will actually need to use Theorem 6.1.1 as well.

Remark 7.1.5 We can also deduce from the spectral sequence that the cardinality of the relative Brauer group for KO / KU, which is isomorphic to $\pi_{-1}(\mathfrak{pic}(\mathrm{KU}))^{hC_2}$, is at most eight. However, we do not know how to construct necessarily nontrivial elements of this Brauer group in order to deduce a lower bound as in the Picard group case.

7.2 KO[q], KO[[q]] and KO((q))

We now include a variant of the above example where one adds a polynomial (resp. power series, Laurent series) generator, where we will also be able to confirm the answer using a different argument. This example can be useful for comparison with TMF using topological q–expansion maps. We begin by introducing the relevant E_∞–rings. This subsection will not be used in the sequel and may be safely skipped by the reader.

Definition 7.2.1 We write for $S^0[x]$ the suspension spectrum $\Sigma^\infty_+ \mathbb{Z}_{\geq 0}$. Since $\mathbb{Z}_{\geq 0}$ is an $\boldsymbol{E}_\infty$–monoid in spaces (in fact, a commutative topological monoid), $S^0[x]$ naturally acquires the structure of an $\boldsymbol{E}_\infty$–ring. Given an $\boldsymbol{E}_\infty$–ring R, we will write $R[x] = R \wedge S^0[x]$.

We can also derive several other variants:

(1) We will let $R[\![x]\!]$ denote the x–adic completion of $R[x]$, so its homotopy groups look like a power series ring over $\pi_* R$.

(2) We will let $R[x^{\pm 1}]$ denote the localization $R[x][1/x]$, so its homotopy groups are given by Laurent polynomials in $\pi_* R$.

(3) We will let $R(\!(x)\!) = R[\![x]\!][1/x]$, so that its homotopy groups look like formal Laurent series over $\pi_* R$.

On the one hand, $\pi_*(R[x]) \simeq (\pi_* R)[x]$ is a polynomial ring over $\pi_* R$ on a generator in degree zero. On the other hand, as an $\boldsymbol{E}_\infty$–algebra under R, the universal property of $R[x]$ is significantly more complicated than that of the "free" $\boldsymbol{E}_\infty$–R–algebra on a generator (often denoted $R\{x\}$). A map $R[x] \to R'$, for an $\boldsymbol{E}_\infty$–R–algebra R', is equivalent to an $\boldsymbol{E}_\infty$–map

$$\mathbb{Z}_{\geq 0} \to \Omega^\infty R',$$

where $\Omega^\infty R'$ is regarded as an $\boldsymbol{E}_\infty$–space under *multiplication*. In general, given a class in $\pi_0 R'$, there is no reason to expect an $\boldsymbol{E}_\infty$–map $R[x] \to R'$ carrying x to it, since $\mathbb{Z}_{\geq 0}$ as an $\boldsymbol{E}_\infty$–monoid is quite complicated. Classes for which this is possible (together with the associated maps $R[x] \to R'$) have been called "strictly commutative" by Lurie.

Example 7.2.2 There is a map $R[x] \to R$ satisfying $x \mapsto 1$. This comes from the map of $\boldsymbol{E}_\infty$–spaces $\mathbb{Z}_{\geq 0} \to * \to \Omega^\infty S^0$ where $*$ maps to the unit in $\Omega^\infty S^0$.

Example 7.2.3 There is a map $R[x] \to R$ satisfying $x \mapsto 0$.[13]

To obtain this in the universal case $R = S^0$, we consider the adjunction

$$(\Sigma^\infty, \Omega^\infty)\colon \mathcal{S}_* \rightleftarrows \mathrm{Sp}.$$

Here $\mathcal{S}_*$ and Sp are symmetric monoidal with the smash product and Σ^∞ is a symmetric monoidal functor. In particular, Σ^∞ carries commutative algebra objects in $\mathcal{S}_*$ to $\boldsymbol{E}_\infty$–ring spectra.

[13] We are grateful to the referee for suggesting this argument over our previous one.

We start with the commutative monoid M with a single element m. Then we have that $M_+ = \{*, m\} \in \mathcal{S}_*$ is a commutative algebra object of $\mathcal{S}_*$ with respect to the smash product: in fact, it is the unit S^0 as a pointed space. Similarly, $(\mathbb{Z}_{\geq 0})_+$ is a commutative algebra object of $\mathcal{S}_*$. Now we have equivalences of $\boldsymbol{E}_\infty$–ring spectra $\Sigma^\infty(M_+) \simeq S^0$ and $\Sigma^\infty(\mathbb{Z}_{\geq 0})_+ \simeq \Sigma^\infty_+ \mathbb{Z}_{\geq 0}$. There is a map of commutative monoids in $\mathcal{S}_*$

$$(\mathbb{Z}_{\geq 0})_+ \to M_+,$$

which carries $0 \in \mathbb{Z}_{\geq 0}$ to m and everything else to $*$. After applying Σ^∞, we obtain the desired map $S^0[x] \to S^0$ of $\boldsymbol{E}_\infty$–rings.

The map $R[x] \to R$ given in Example 7.2.3 has the property that it exhibits the $R[x]$–module R as the cofiber $R[x]/x$. It follows in particular that if R' is any $\boldsymbol{E}_\infty$–R–algebra and $x' \in \pi_0 R'$ is a strictly commutative element, then we can give the cofiber $R'/x' \simeq R' \otimes_{R[x]} R$ the structure of an $\boldsymbol{E}_\infty$–R'–algebra.

Remark 7.2.4 Consider the sphere spectrum S^0. No cofiber S^0/n for $n \notin \{\pm 1, 0\}$ can admit the structure of an $\boldsymbol{E}_\infty$–ring by, for example, [43, Remark 4.3].[14] It follows that the only element of $\pi_0 S^0 \simeq \mathbb{Z}$, besides 0 and 1, that can potentially be strictly commutative is -1. Now, -1 is not strictly commutative in the $K(1)$–local sphere $L_{K(1)}S^0$ at the prime 2 because of the operator θ of [24]: we have $\theta(-1) = \frac{1}{2}((-1)^2 - (-1)) = 1 \neq 0$, while power operations such as θ annihilate strictly commutative elements. Therefore, -1 cannot be strictly commutative in S^0. (One could have applied a similar argument with power operations to every other integer, too.) However, we observe that it is strictly commutative in $S^0\left[\frac{1}{2}\right]$: the obstruction is entirely 2–primary (Proposition 7.2.6 below).

Example 7.2.5 Let $a, b \in \pi_0 R$ be strictly commutative elements for R an $\boldsymbol{E}_\infty$–ring. Then ab is also strictly commutative. If a is a unit, then a^{-1} is strictly commutative. This follows because there is a natural addition on $\boldsymbol{E}_\infty$–maps $\mathbb{Z}_{\geq 0} \to \Omega^\infty R$.

Proposition 7.2.6 *Let R be an $\boldsymbol{E}_\infty$–ring with n invertible. Then any $u \in \pi_0 R$ with $u^n = 1$ (ie an n^{th} root of unity) admits the structure of a strictly commutative element.*

Proof We consider the map of $\boldsymbol{E}_\infty$–monoids $\mathbb{Z}_{\geq 0} \to \mathbb{Z}/n\mathbb{Z}$ and the induced map of $\boldsymbol{E}_\infty$–ring spectra

$$R[x] \to R \wedge \Sigma^\infty_+ \mathbb{Z}/n\mathbb{Z}. \tag{7-1}$$

[14] It is an unpublished result of Hopkins that no Moore spectrum can even admit the structure of an $\boldsymbol{E}_1$–algebra.

Since $1/n \in \pi_0 R$, we have that $R \wedge \Sigma^\infty_+ \mathbb{Z}/n\mathbb{Z}$ is étale over R and the homotopy groups are given by $\pi_* R[x]/(x^n - 1)$. We can thus produce a map of $\boldsymbol{E}_\infty$–rings $R \wedge \Sigma^\infty_+(\mathbb{Z}/n\mathbb{Z}) \to R$ sending $1 \in \mathbb{Z}/n\mathbb{Z}$ to u by étaleness.[15] Composing with (7-1) gives us the strictly commutative structure on u. □

Using these ideas, we will be able to give a direct computation of the Picard group of the $\boldsymbol{E}_\infty$–ring $\mathrm{KO}[\![q]\!]$. (We have renamed the power series variable to "q" in accordance with "q–expansions".)

Proposition 7.2.7 *The map* $\mathrm{Pic}(\mathrm{KO}) \to \mathrm{Pic}(\mathrm{KO}[\![q]\!])$ *is an isomorphism, where* q *is in degree zero.*

Proof Suppose M is an invertible $\mathrm{KO}[\![q]\!]$–module such that $M/qM \simeq M \otimes_{\mathrm{KO}[\![q]\!]} \mathrm{KO}$ is equivalent to KO. We will show that then M is equivalent to $\mathrm{KO}[\![q]\!]$ using Bocksteins. Specifically, consider the generating class in $\pi_0(M/qM) \simeq \mathbb{Z}$; we will lift this to a class in $\pi_0 M$. It will follow that the induced map $\mathrm{KO}[\![q]\!] \to M$ becomes an equivalence after tensoring with $\mathrm{KO} \simeq \mathrm{KO}[\![q]\!]/q$. Since M is q–adically complete, it will follow that $\mathrm{KO}[\![q]\!] \simeq M$.

By induction on k, suppose that:

(1) $\pi_{-1}(M/q^k M) = 0$.

(2) $\pi_0(M/q^k M) \to \pi_0(M/qM)$ is a surjection.

These conditions are clearly satisfied for $k = 1$. If these conditions are satisfied for k, then the cofiber sequence of $\mathrm{KO}[\![q]\!]$–modules

$$M/q^k M \to M/q^{k+1} M \to M/qM$$

shows that they are satisfied for $k + 1$. In the limit, we find that there is a map $\mathrm{KO}[\![q]\!] \to M$ which lifts the generator of $\pi_0(M/qM)$, which proves the claim. □

Proposition 7.2.7 can also be proved using Galois descent, but unlike for KO, we need to use Theorem 6.1.1.

Second proof of Proposition 7.2.7 The faithful C_2–Galois extension $\mathrm{KO} \to \mathrm{KU}$ induces upon base-change a faithful C_2–Galois extension $\mathrm{KO}[\![q]\!] \to \mathrm{KU}[\![q]\!]$. The Picard group of $\mathrm{KU}[\![q]\!]$, again by Theorem 2.4.6, is $\mathbb{Z}/2$ generated by the suspension.

[15] The étale obstruction theory has been developed by a number of authors; a convenient reference for the result that we need is [39, Theorem 8.5.4.2].

Consider now the descent spectral sequence for $(\mathfrak{pic}(\mathrm{KU}[\![q]\!]))^{hC_2}$, which is a modification of the descent spectral sequence for KU^{hC_2} in Figure 4. One difference is that every term with $t \geq 2$ is replaced by its tensor product over $\mathbb{Z}$ with $\mathbb{Z}[\![q]\!]$; the other is that the $t = 1$ line now contains the C_2–cohomology of the units in $\pi_0\,\mathrm{KU}[\![q]\!]$, which is a bigger module than $(\pi_0\,\mathrm{KU})^\times = \mathbb{Z}/2$. Namely, these units are $\mathbb{Z}/2 \oplus q\mathbb{Z}[\![q]\!]$, with trivial C_2–action. The resulting E_2–page is displayed in Figure 5.

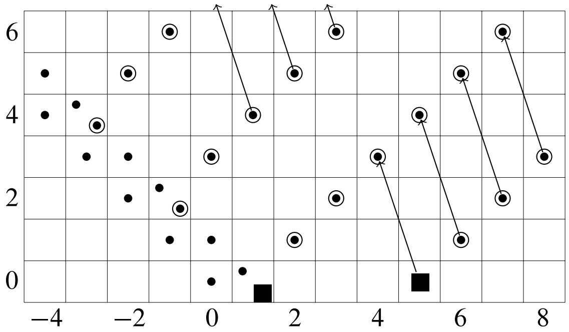


Figure 5: Homotopy fixed point spectral sequence for $\mathfrak{pic}(\mathrm{KU}[\![q]\!])^{hC_2}$ (● denotes $\mathbb{Z}/2$, ⦿ denotes $\mathbb{Z}/2[\![q]\!]$, and ■ denotes $\mathbb{Z}[\![q]\!]$)

Since the d_3 is the only differential in the ordinary HFPSS for $\pi_*\,\mathrm{KO}[\![q]\!]$, as before, it follows that the only contributions to $\mathrm{Pic}(\mathrm{KO}[\![q]\!])$ can come from the $\mathbb{Z}/2$ with $t = s = 0$ (the suspension), the $\mathbb{Z}/2$ with $(s,t) = (1,1)$ (ie the algebraic Picard group), and the $\mathbb{Z}/2[\![q]\!]$ in bidegree $(s,t) = (3,3)$.

But here, $E_2^{3,3} = \mathbb{Z}/2[\![q]\!](h_1^3/u^2)$ is infinite, so unlike previously, we do not get the automatic upper bound of eight on $|\mathrm{Pic}(\mathrm{KO}[\![q]\!])|$. On the other hand, we can use Theorem 6.1.1 to determine the d_3 supported here. Note that in the HFPSS for $(\mathrm{KU}[\![q]\!])^{hC_2}$, we have

$$d_3(f(q)(h_1^3/u^2)) = f(q)(h_1^6/u^4) \quad \text{for } f(q) \in \mathbb{Z}/2[\![q]\!].$$

Therefore, in view of (6-1), in the HFPSS for $\mathfrak{pic}(\mathrm{KU}[\![q]\!])^{hC_2}$, we have

$$d_3(f(q)(h_1^3/u^2)) = (f(q) + f(q)^2)(h_1^6/u^4).$$

(Note that a crucial point here is that in the HFPSS for KO, squaring or applying d_3 to h_1^3/u^2 yields the same result.) It follows from this that in the HFPSS, the kernel of d_3 on $E_2^{3,3}$ is $\mathbb{Z}/2$ generated by $1(h_1^3/u^2)$: the equation $f(q) + f(q)^2 = 0$ has only the solutions $f(q) \equiv 0, 1$. Therefore, we do get an upper bound of eight on the cardinality of $\mathrm{Pic}(\mathrm{KO}[\![q]\!])$ after all, as nothing else in $E_2^{3,3}$ lives to E_4. □

Corollary 7.2.8 *The maps* $\mathrm{KO} \to \mathrm{KO}[q]$ *and* $\mathrm{KO} \to \mathrm{KO}((q))$ *induce isomorphisms on Picard groups.*

Proof This result is not a corollary of Proposition 7.2.7 but rather of its second proof. In fact, the same argument shows that d_3 has a $\mathbb{Z}/2$ as kernel on the relevant term $E_2^{3,3}$, which gives an upper bound of cardinality eight on the Picard group of $\mathrm{KO}[q]$ or $\mathrm{KO}((q))$ as before. □

Remark 7.2.9 Corollary 7.2.8 cannot be proved using the Bockstein spectral sequence argument used in the first proof of Proposition 7.2.7. However, a knowledge of the Picard group of $\mathrm{KO}[\![q]\!]$ can be used to describe enough of the C_2–descent spectral sequence to make it possible to prove Corollary 7.2.8 without the explicit formula (6-1). We leave this to the reader.

8 Picard groups of topological modular forms

In the rest of the paper we proceed to use descent to compute the Picard groups of various versions of topological modular forms. We will analyze the following descent-theoretic situations:

- The Galois extension $\mathrm{TMF}\big[\frac{1}{2}\big] \to \mathrm{TMF}(2)$, with structure group $\mathrm{GL}_2(\mathbb{Z}/2)$, also known as the symmetric group on three letters.
- The Galois extension $\mathrm{TMF}\big[\frac{1}{3}\big] \to \mathrm{TMF}(3)$, with structure group $\mathrm{GL}_2(\mathbb{Z}/3)$, a group of order 48 which is a nontrivial extension of the binary tetrahedral group and C_2.
- Étale descent from the (derived) moduli stack of elliptic curves or its compactification.

In each of these cases, we will start with the knowledge of the original descent spectral sequence, computing the homotopy groups of the global sections or homotopy fixed point spectrum. This information plus some additional computation of the differing cohomology groups will provide the data for the E_2–page of the descent spectral sequence for the Picard spectrum. The additional computations are somewhat lengthy, hence we are including them separately in the appendices.

8.1 The Picard group of $\mathrm{TMF}\big[\frac{1}{2}\big]$

When 2 is inverted, the moduli stack of elliptic curves M_{ell} has a $\mathrm{GL}_2(\mathbb{Z}/2)$–Galois cover by $M_{\mathrm{ell}}(2)$, the moduli stack of elliptic curves with full level 2 structure. This remains the case for the derived versions of these stacks, and on global sections gives a faithful Galois extension $\mathrm{TMF}[\frac{1}{2}] \to \mathrm{TMF}(2)$ by [42, Theorem 7.6]. The extension is useful for the purposes of descent as the homotopy groups of TMF(2) are cohomologically very simple.

To be precise, we have that

$$\mathrm{TMF}(2)_* = \mathbb{Z}[\tfrac{1}{2}][\lambda_1^{\pm 1}, \lambda_2^{\pm 1}][(\lambda_1 - \lambda_2)^{-1}],$$

where the (topological) degree of each λ_i is four. To see this, one can use the presentation of the moduli stack $M_{\mathrm{ell}}(2)$ from [63, Section 7]. There it is computed that $\overline{M}_{\mathrm{ell}}(2)$ is equivalent to (the stacky) $\operatorname{Proj} \mathbb{Z}[\frac{1}{2}][\lambda_1, \lambda_2]$. Moreover, the substack classifying smooth curves, ie $M_{\mathrm{ell}}(2)$, is the locus of nonvanishing of $\lambda_1^2 \lambda_2^2 (\lambda_1 - \lambda_2)^2$. More precisely, $M_{\mathrm{ell}}(2)$, as a stack, is the $\mathbb{G}_m$–quotient of the ring

$$\mathbb{Z}[\tfrac{1}{2}][\lambda_1, \lambda_2, (\lambda_1^2 \lambda_2^2 (\lambda_1 - \lambda_2))^{-1}],$$

where the $\mathbb{G}_m$–action is as follows: a unit u acts as $\lambda_i \mapsto u^2 \lambda_i$ for $i = 1, 2$, so that it is an open substack of a *weighted projective stack.*

In particular, $\mathrm{TMF}(2)_*$ has a unit in degree 4, and is zero in degrees not divisible by 4. It will be helpful to write $\mathrm{TMF}(2)_*$ differently, so as to reflect this periodicity more explicitly; for example, we have that $\mathrm{TMF}(2)_* = \mathrm{TMF}(2)_0[\lambda_2^{\pm 1}]$, and

$$\mathrm{TMF}(2)_0 = \mathbb{Z}[\tfrac{1}{2}][s^{\pm 1}, (s-1)^{-1}], \tag{8-1}$$

where $s = \lambda_1 / \lambda_2$. Therefore, Corollary 2.4.7 applies to give the following conclusion.

Lemma 8.1.1 Pic(TMF(2)) *is* $\mathbb{Z}/4$, *generated by the suspension* $\Sigma\,\mathrm{TMF}(2)$.

Remark 8.1.2 The proof of Corollary 2.4.7 relies on the construction of "residue field" spectra; let us specify what they are in the case at hand. The maximal ideals in $\mathrm{TMF}(2)_0$ are $\mathfrak{m} = (p, f(s))$, where p is an odd prime and $f(s)$ a monic polynomial irreducible modulo p (and not congruent mod p to s, $s-1$). For each of these ideals, we have an associative ring spectrum (the "residue field") with homotopy groups $\mathrm{TMF}(2)_*/\mathfrak{m}$ by [3]; denote it temporarily by $\mathrm{TMF}(2)/\mathfrak{m}$. After extending scalars so that f splits, we get that $\mathrm{TMF}(2)/\mathfrak{m}$ is a product of (extensions of) mod p Morava K–theory spectra at height one or two, one for each zero of f. By [62, Chapter V, Theorem 4.1], the factor associated to the zero a of f has height two precisely when

$$\sum_{i=0}^{(p-1)/2} \binom{(p-1)/2}{i} a^i$$

is zero modulo p.

Next we use descent from TMF(2) to $\mathrm{TMF}\big[\frac{1}{2}\big]$ to obtain the following result.

Theorem 8.1.3 $\mathrm{Pic}\big(\mathrm{TMF}\big[\frac{1}{2}\big]\big)$ *is* $\mathbb{Z}/72$*, generated by the suspension* $\Sigma\,\mathrm{TMF}\big[\frac{1}{2}\big]$*. In particular, this Picard group is algebraic.*

Proof We use the homotopy fixed point spectral sequence (3-5)

$$H^s\big(\mathrm{GL}_2(\mathbb{Z}/2), \pi_t\,\mathfrak{pic}(\mathrm{TMF}(2))\big) \Rightarrow \pi_{t-s}\,\mathfrak{pic}(\mathrm{TMF}(2))^{h\mathrm{GL}_2(\mathbb{Z}/2)}. \tag{8-2}$$

To begin with, note that the homotopy groups $\pi_t\,\mathfrak{pic}(\mathrm{TMF}(2))$ for $t \geq 2$ are isomorphic to $\pi_{t-1}\,\mathrm{TMF}(2)$ as $\mathrm{GL}_2(\mathbb{Z}/2)$–modules. This tells us that the $t \geq 2$ part of the E_2–page of the HFPSS (8-2) for $\mathfrak{pic}(\mathrm{TMF}(2))$ is a shifted version of the corresponding part for TMF(2).

The latter is immediately obtained from the analogous computation for Tmf(2) depicted in [63, Figure 2], as we now describe. Recall that $\mathrm{TMF}(2) \simeq \mathrm{Tmf}(2)[\Delta^{-1}]$; the nonnegative homotopy groups $\pi_{\geq 0}\,\mathrm{Tmf}(2)$ are the graded polynomial ring $\Lambda = \mathbb{Z}\big[\frac{1}{2}\big][\lambda_1, \lambda_2]$ [63, Proposition 8.1], and the class $\Delta \in \pi_{24}\,\mathrm{Tmf}(2)$ is

$$\Delta = 16\lambda_1^2\lambda_2^2(\lambda_2 - \lambda_1)^2$$

by [63, Proposition 10.3]. Now, by [63, Proposition 10.8] we have that

$$H^*\big(\mathrm{GL}_2(\mathbb{Z}/2), \pi_*\,\mathrm{TMF}(2)\big) = H^*\big(\mathrm{GL}_2(\mathbb{Z}/2), \Lambda\big)[\Delta^{-1}].$$

In particular, the invariants $H^0\big(\mathrm{GL}_2(\mathbb{Z}/2), \Lambda\big)[\Delta^{-1}]$ are the ring of Δ–inverted modular forms

$$\mathbb{Z}\big[\tfrac{1}{2}\big][c_4, c_6, \Delta^{\pm 1}]/(12^3\Delta - c_4^3 + c_6^2).$$

The higher cohomology $H^{>0}\big(\mathrm{GL}_2(\mathbb{Z}/2), \Lambda\big)$ is computed in [63, Section 10.1], and in particular is killed by c_4 and c_6. Consequently,

$$\begin{aligned} H^{>0}\big(\mathrm{GL}_2(\mathbb{Z}/2), \pi_{\geq 0}\,\mathrm{TMF}(2)\big) &= H^{>0}\big(\mathrm{GL}_2(\mathbb{Z}/2), \Lambda\big) \\ &= H^{>0}(\mathrm{GL}_2(\mathbb{Z}/2), \pi_{\geq 0}\,\mathrm{Tmf}(2)). \end{aligned}$$

Let us recall (the names of) certain interesting classes in these cohomology groups:

1. There is the class a in $H^1(\mathrm{GL}_2(\mathbb{Z}/2), \pi_4\,\mathrm{TMF}(2)) = \mathbb{Z}/3$, hence also in $H^1\big(\mathrm{GL}_2(\mathbb{Z}/2), \pi_5\,\mathfrak{pic}(\mathrm{TMF}(2))\big)$ (so, a is in bidegree $(s,t) = (1,5)$ in the Picard HFPSS, and depicted in position $(s, t-s) = (1,4)$ using the Adams convention). In homotopy, this element detects the Greek letter element α_1 in the Hurewicz image in $\mathrm{TMF}\big[\frac{1}{2}\big]$.
2. There is b in $H^2\big(\mathrm{GL}_2(\mathbb{Z}/2), \pi_{13}\,\mathfrak{pic}(\mathrm{TMF}(2))\big) = \mathbb{Z}/3$ (b is in bidegree $(2,13)$ or position $(2,11)$); in homotopy it detects β_1.

Then, $H^{>0}\big(\mathrm{GL}_2(\mathbb{Z}/2), \mathrm{TMF}(2)_*\big)$ is precisely the ideal of $\mathbb{Z}/3[a,b][\Delta^{\pm 1}]/(a^2)$ of positive cohomological degree. For example

$$H^5\big(\mathrm{GL}_2(\mathbb{Z}/2), \pi_5\,\mathfrak{pic}(\mathrm{TMF}(2))\big) = H^5\big(\mathrm{GL}_2(\mathbb{Z}/2), \pi_4\,\mathrm{TMF}(2)\big) = \mathbb{Z}/3,$$

generated by $ab^2\Delta^{-1}$. We see this class depicted in red in Figure 6.

Next, we turn to the information which is new for the Picard HFPSS, ie the group cohomology of π_0 and π_1 of the spectrum $\mathfrak{pic}(\mathrm{TMF}(2))$. By Lemma 8.1.1, we know that the zeroth homotopy group is $\mathbb{Z}/4$, and since it is generated by the suspension $\Sigma\,\mathrm{TMF}(2)$, the action of $\mathrm{GL}_2(\mathbb{Z}/2)$ on this $\mathbb{Z}/4$ is trivial. Even though for our purposes only the invariants $H^0\big(\mathrm{GL}_2(\mathbb{Z}/2), \pi_0\,\mathfrak{pic}(\mathrm{TMF}(2))\big)$ are necessary, we can in fact compute all the cohomology groups. This is done in Lemma A.1.

The last piece of data needed for the determination of the E_2–page of the Picard HFPSS is the group cohomology with coefficients in $\pi_1\,\mathfrak{pic}(\mathrm{TMF}(2)) = (\pi_0\,\mathrm{TMF}(2))^\times$. This is done in Proposition A.2. The range $s \le 15$ and $-6 \le t-s \le 7$ of the spectral sequence is depicted in Figure 6. Note that in this range, the $t-s=0$ column has three nonzero entries: there is a $\mathbb{Z}/4$ for $s=0$, a $\mathbb{Z}/6$ for $s=1$ and a $\mathbb{Z}/3$ for $s=5$.

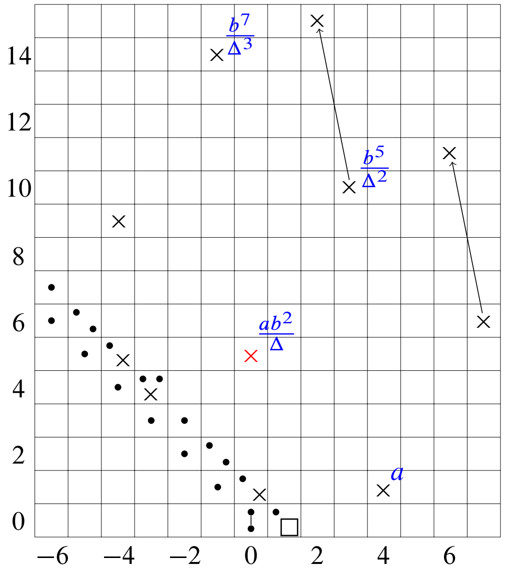


Figure 6: Homotopy fixed point spectral sequence for $\big(\mathfrak{pic}(\mathrm{TMF}(2))\big)^{h\mathrm{GL}_2(\mathbb{Z}/2)}$ (□ denotes $\mathbb{Z}$, • denotes $\mathbb{Z}/2$, and × denotes $\mathbb{Z}/3$)

Now we are ready to study the differentials in the HFPSS for $\mathfrak{pic}(\mathrm{TMF}(2))^{h\mathrm{GL}_2(\mathbb{Z}/2)}$. Comparison with the HFPSS for the $\mathrm{GL}_2(\mathbb{Z}/2)$–action on $\mathrm{TMF}(2)$ gives a number of differentials, using our Comparison Tool 5.2.4. To distinguish between the differentials in the two spectral sequences, let us denote by d_r^o those in the HFPSS of TMF(2). The superscript o stands for "original".

Recall that in the HFPSS for TMF(2), there are nonzero d_5^o and d_9^o differentials, which are obtained, for example, by a comparison with the HFPSS for Tmf(2) which is fully determined in [63]. In particular, in the HFPSS for TMF(2), the first differential is $d_5^o(\Delta) = ab^2$, and the rest of the d_5^o's are determined by multiplicativity and the fact that a and b are permanent cycles. In particular, we have

$$d_5^o\left(\frac{b^5}{\Delta^2}\right) = \frac{ab^7}{\Delta^3} \quad \text{and} \quad d_5^o\left(\frac{b^3}{\Delta}\right) = -\frac{ab^5}{\Delta^2}. \tag{8-3}$$

Next (and last) is d_9^o; we have that $d_9^o(a\Delta^2) = b^5$. Consequently, we also have

$$d_9^o\left(\frac{ab^2}{\Delta}\right) = \frac{b^7}{\Delta^3}. \tag{8-4}$$

Let us now see which of these differentials also occur in the HFPSS for $\mathfrak{pic}(\mathrm{TMF}(2))$; according to Comparison Tool 5.2.4, the d_5–differentials are imported in the $t > 5$ range, and the d_9–differentials in the $t > 9$ range. In particular, the differentials in (8-3) are the same in the Picard HFPSS; these are the two differentials drawn in Figure 6. Moreover, everything in the zero column and above the depicted region, ie such that $s = t > 16$, either supports a differential or is killed by one which originates in the $t > 9$ range. Hence, everything above the depicted region is killed in the spectral sequence and nothing survives to the E_∞–page.

Note, however, that we cannot (and should not attempt to) import the differential (8-4); this would be a d_9–differential with $t = 5$, so it does not satisfy the hypothesis of Comparison Tool 5.2.4.

Let us analyze the potentially remaining contributions to $\pi_0\,\mathfrak{pic}(\mathrm{TMF}(2))^{\mathrm{GL}_2(\mathbb{Z}/2)}$; regardless of what the rest of the differentials could possibly be, we have

- a group of order at most 4 (and dividing 4) in position $(0,0)$,
- a group of order at most 6 (and dividing 6) in position $(0,1)$, and
- a group of order at most 3 (and dividing 3) in position $(0,5)$.

Therefore $\mathrm{Pic}\big(\mathrm{TMF}\big[\tfrac{1}{2}\big]\big) = \pi_0\,\mathfrak{pic}(\mathrm{TMF}(2))^{\mathrm{GL}_2(\mathbb{Z}/2)}$ has order at most $4\times 6\times 3 = 72$, and dividing 72. This is an upper bound. But we also have a well-known lower bound: the suspension $\Sigma\,\mathrm{TMF}\big[\tfrac{1}{2}\big]$ generates a nontrivial element of $\mathrm{Pic}\big(\mathrm{TMF}\big[\tfrac{1}{2}\big]\big)$ of order 72 because $\mathrm{TMF}\big[\tfrac{1}{2}\big]$ is 72–periodic. Thus we have proven the result. □

Remark 8.1.4 Our computations give an independent proof of the result of Fulton and Olsson [14] that the Picard group of the classical moduli stack of elliptic curves M_{ell} over $\mathbb{Z}[\frac{1}{2}]$ is $\mathbb{Z}/12$. (Fulton and Olsson carry out the analysis over any base, though.) This is a toy analog of the above analysis, as we now see.

The Picard *groupoid* of the moduli stack $M_{\mathrm{ell}}[\frac{1}{2}]$ is the homotopy fixed points of the $\mathrm{GL}_2(\mathbb{Z}/2)$–action on the Picard groupoid of $M_{\mathrm{ell}}(2)$. Now the Picard *group* of $M_{\mathrm{ell}}(2)$ is $\mathbb{Z}/2$, as $M_{\mathrm{ell}}(2)$ is an open subset in a weighted projective stack over a UFD, so that quasicoherent sheaves on $M_{\mathrm{ell}}(2)$ correspond simply to graded modules over $\mathbb{Z}[\frac{1}{2}, \lambda_1, \lambda_2, (\lambda_1^2\lambda_2^2(\lambda_1 - \lambda_2))^{-1}]$ and the only nontrivial invertible object is the shift by one of the unit. Note that this is the *algebraic* setting: the generator of $\mathrm{Pic}(M_{\mathrm{ell}}(2))$ would correspond to the *two-fold* suspension of TMF(2).

Next, in the HFPSS for computing $\mathrm{Pic}(M_{\mathrm{ell}}[\frac{1}{2}])$, we see by the above computation of

$$H^1(\mathrm{GL}_2(\mathbb{Z}/2), \Gamma(M_{\mathrm{ell}}(2), \mathcal{O}^\times))$$

that one gets a contribution of order 6. Together with $\mathrm{Pic}(M_{\mathrm{ell}}(2)) = \mathbb{Z}/2$ from the previous paragraph, we get that $|\mathrm{Pic}(M_{\mathrm{ell}}[\frac{1}{2}])| \leq 12$, but we know that ω has order twelve, so we are done.

8.2 The Picard group of $\mathrm{TMF}[\frac{1}{3}]$

This section will be similar to Section 8.1, but with more complicated computations as is to be expected from 2–torsion. In this case we will use the $\mathrm{GL}_2(\mathbb{Z}/3)$–Galois extension $\mathrm{TMF}[\frac{1}{3}] \to \mathrm{TMF}(3)$, coming from the Galois cover $M_{\mathrm{ell}}(3) \to M_{\mathrm{ell}}[\frac{1}{3}]$ of the moduli stack of elliptic curves with 3 inverted by the moduli stack of elliptic curves equipped with a full level 3–structure.

From [64, Section 4.2], we can immediately compute the homotopy groups of TMF(3): the moduli stack $M_{\mathrm{ell}}(3)$ is affine, and is given as the locus of nonvanishing of

$$\Delta = 3^{-5}\zeta(1-\zeta)\gamma_1^3\gamma_2^3(\gamma_1 + \zeta\gamma_2)^3(\gamma_2 - \zeta\gamma_1)^3$$

in the compact moduli stack $\overline{M}_{\mathrm{ell}}(3) = \mathrm{Proj}\,\mathbb{Z}[\frac{1}{3}, \zeta][\gamma_1, \gamma_2]$. Here γ_i are variables in (topological) degree 2, and ζ is a primitive third root of unity, whose appearance is due to the fact that the Weil pairing on the 3–torsion points of an elliptic curve equips $\overline{M}_{\mathrm{ell}}(3)$ with a map to $\mathrm{Spec}\,\mathbb{Z}[\frac{1}{3}, \zeta]$.[16] Hence the descent spectral sequence computing $\mathrm{TMF}(3)_*$ collapses to give

$$\mathrm{TMF}(3)_* = \mathbb{Z}[\tfrac{1}{3}, \zeta][\gamma_1^{\pm 1}, \gamma_2^{\pm 1}][(\gamma_1 + \zeta\gamma_2)^{-1}, (\gamma_2 - \zeta\gamma_1)^{-1}].$$

[16] The map is given by the usual Weil pairing on the locus of smooth curves; for what it does at the cusps, see for example [11, IV.3.21].

Written differently, we have that $\mathrm{TMF}(3)_* = \mathrm{TMF}(3)_0[\gamma_2^{\pm 1}]$, and

$$\mathrm{TMF}(3)_0 = \mathbb{Z}\big[\tfrac{1}{3}, \zeta\big][t^{\pm 1}, (1-\zeta t)^{-1}, (1+\zeta^2 t)^{-1}], \tag{8-5}$$

for $t = \gamma_1/\gamma_2$. In particular $\mathrm{TMF}(3)_0$ is regular noetherian, and $\mathrm{TMF}(3)$ is even periodic. Thus, Theorem 2.4.6 (together with the fact that the ring $\mathbb{Z}[\zeta, t]$ and hence any of its localizations has unique factorization) implies the following conclusion.

Lemma 8.2.1 *The Picard group* $\mathrm{Pic}(\mathrm{TMF}(3))$ *is* $\mathbb{Z}/2$, *generated by* $\Sigma\,\mathrm{TMF}(3)$.

Naturally, we will use this lemma as an input in computing the HFPSS for the associated Picard spectra.

Theorem 8.2.2 $\mathrm{Pic}\big(\mathrm{TMF}\big[\tfrac{1}{3}\big]\big)$ *is* $\mathbb{Z}/192$, *generated by the suspension* $\Sigma\,\mathrm{TMF}\big[\tfrac{1}{3}\big]$. *In particular, this Picard group is algebraic.*

Proof As is to be expected, we use the HFPSS (3-5)

$$H^s\big(\mathrm{GL}_2(\mathbb{Z}/3), \pi_t\,\mathfrak{pic}(\mathrm{TMF}(3))\big) \Rightarrow \pi_{t-s}\,\mathfrak{pic}(\mathrm{TMF}(3))^{h\mathrm{GL}_2(\mathbb{Z}/3)}. \tag{8-6}$$

The homotopy groups $\pi_t(\mathfrak{pic}(\mathrm{TMF}(3)))$ for $t \geq 2$ are isomorphic to $\pi_{t-1}\,\mathrm{TMF}(3)$ as $\mathrm{GL}_2(\mathbb{Z}/3)$–modules; therefore the $t \geq 2$ part of the E_2–page of the HFPSS for $\mathfrak{pic}(\mathrm{TMF}(3))$ is same as the corresponding part in the HFPSS for $\mathrm{TMF}(3)$. We will use the fact that $\mathrm{TMF}(3) \simeq \mathrm{Tmf}(3)[\Delta^{-1}]$ to identify this part of the spectral sequence for $\mathrm{TMF}(3)$ and therefore for $\mathfrak{pic}(\mathrm{TMF}(3))$.

Computed in [64], and depicted in Figure 9 of loc. cit., is the E_2–page of the HFPSS computing the homotopy groups of $\mathrm{Tmf}_{\hat{2}}$ as $(\mathrm{Tmf}(3)_{\hat{2}})^{h\mathrm{GL}_2(\mathbb{Z}/3)}$. Since we are working with 3 inverted, and 2 and 3 are the only primes dividing the order of $\mathrm{GL}_2(\mathbb{Z}/3)$, we conclude that

$$H^{>0}(\mathrm{GL}_2(\mathbb{Z}/3), \pi_*\,\mathrm{Tmf}(3)) = H^{>0}(\mathrm{GL}_2(\mathbb{Z}/3), \pi_*\,\mathrm{Tmf}(3)_{\hat{2}}).$$

The invariants $H^0(\mathrm{GL}_2(\mathbb{Z}/3), \pi_{\geq 0}\,\mathrm{Tmf}(3))$ are the ring of modular forms

$$\mathbb{Z}\big[\tfrac{1}{3}\big][c_4, c_6, \Delta]/(12^3\Delta - c_4^3 + c_6^2).$$

Let Γ denote the graded ring $\mathbb{Z}\big[\tfrac{1}{3}, \zeta\big][\gamma_1, \gamma_2]$. As in the case of level 2–structures, we have that

$$H^*(\mathrm{GL}_2(\mathbb{Z}/3), \pi_*\,\mathrm{TMF}(3)) = H^*(\mathrm{GL}_2(\mathbb{Z}/3), \Gamma)[\Delta^{-1}].$$

In the group cohomology of Γ, computed and depicted in [64, Figure 7], there are a number of interesting torsion classes, including

(1) h_1 in bidegree $(s,t)=(1,2)$, depicted in position $(s,t-s)=(1,1)$, which detects (the Hurewicz image of) the Hopf map η in homotopy,

(2) h_2 in position $(1,3)$, which detects (the Hurewicz image of) the Hopf map ν,

(3) d in position $(2,14)$, which detects in homotopy the class known as κ,

(4) g in position $(4,20)$, which detects in homotopy the class $\bar{\kappa}$, and

(5) c in position $(2,8)$, which detects in homotopy the class ϵ.

The homotopy elements detected by these classes satisfy some relations; for example,

$$\eta^3 = 4\nu \quad \text{and} \quad \kappa\nu^2 = 4\bar{\kappa}.$$

Let us also name one of the less famous elements in the descent spectral sequence for $\mathrm{tmf}_{(2)}$, which also appears in the HFPSS for $\mathrm{TMF}\big[\frac{1}{3}\big]$. Namely, there is a $\mathbb{Z}/2$ in position $(1,5)$; we will denote the generating class by the generic name x (in [6] it bears the name $a_1^2h_1$).

All torsion classes with the exception of (powers of) h_1 are annihilated by c_4 and c_6. In the Picard spectral sequence, all of these classes appear shifted by one to the right; we have labeled some such classes in Figure 9. A zoomed in portion of the Picard spectral sequence is depicted in Figure 8. There, and in all of the related spectral sequences, lines of slope 1 denote h_1–multiplication, and lines of slope $\frac{1}{3}$ denote h_2–multiplication.

A zoomed out portion of the Picard HFPSS (8-6) is depicted in Figure 7; the elements that are to the right of the $t=2$ line are, of course, a shift of the corresponding elements in the spectral sequence for $\mathrm{TMF}\big[\frac{1}{3}\big]$. However, to avoid cluttering the picture, a family of classes is not shown. The family consists precisely of the h_1–power multiples of nontorsion classes. An exception is made for the elements depicted in green, namely $h_1^3c_4c_6/\Delta$ and $h_1^6c_4^2/\Delta$ (in the $(0,3)$ and $(-1,6)$ positions, respectively; these classes are also labeled in Figure 8), as well as the tower supported on 1, which do belong to this family, but are nonetheless depicted. In the zoomed in Figure 9 this family is also not shown.

More specifically, the nontorsion subring of the E_2–page of the $\mathrm{TMF}\big[\frac{1}{3}\big]$ spectral sequence is precisely the part in cohomological degree 0 and consists of the ring of modular forms $\mathrm{MF}_*\big[\frac{1}{3}\big]=\mathbb{Z}\big[\frac{1}{3}\big][c_4,c_6,\Delta^{\pm1}]/(12^3\Delta-c_4^3+c_6^2)$. On the E_2–page, these support infinite h_1–multiples, ie $\mathrm{MF}_*\big[\frac{1}{3}\big][h_1]/(2h_1)$ is a subring of the E_2–page. Note in degree zero, $\mathrm{MF}_0\big[\frac{1}{3}\big]=\mathbb{Z}\big[\frac{1}{3},j\big]$, where $j=c_4^3/\Delta$ is the classical j–invariant. What we have omitted drawing in Figure 7 and 9 are all of the elements coming from this subring, with the exception of the mentioned classes. For comparison, these elements are drawn in the smaller-range Figure 8.

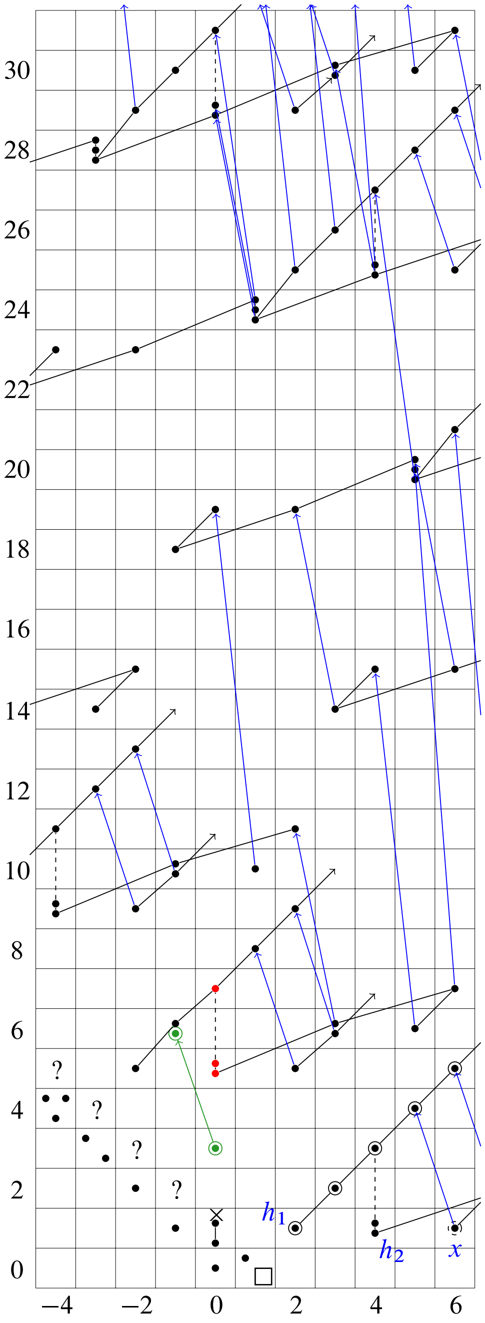


Figure 7: Homotopy fixed point spectral sequence for $\mathfrak{pic}(\mathrm{TMF}(3))^{h\mathrm{GL}_2(\mathbb{Z}/3)}$: zoomed out version with some h_1–omissions (□ denotes $\mathbb{Z}$, ● denotes $\mathbb{Z}/2$, ⦿ denotes $\mathbb{Z}/2[j]$, and × denotes $\mathbb{Z}/3$)

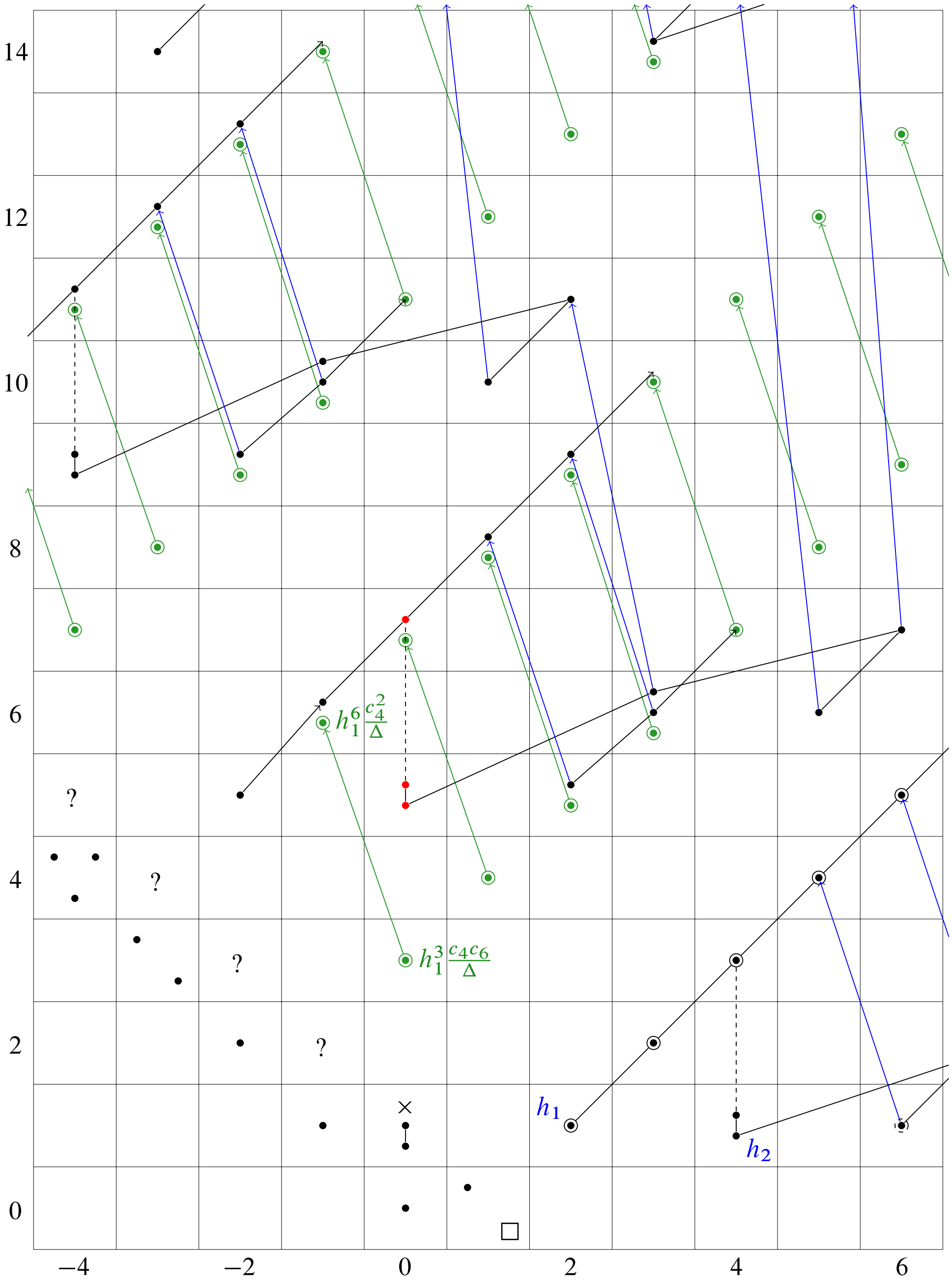


Figure 8: Homotopy fixed point spectral sequence for $\mathfrak{pic}(\mathrm{TMF}(3))^{h\mathrm{GL}_2(\mathbb{Z}/3)}$: zoomed in version without omissions (□ denotes $\mathbb{Z}$, ● denotes $\mathbb{Z}/2$, ⦿ denotes $\mathbb{Z}/2[j]$, and × denotes $\mathbb{Z}/3$)

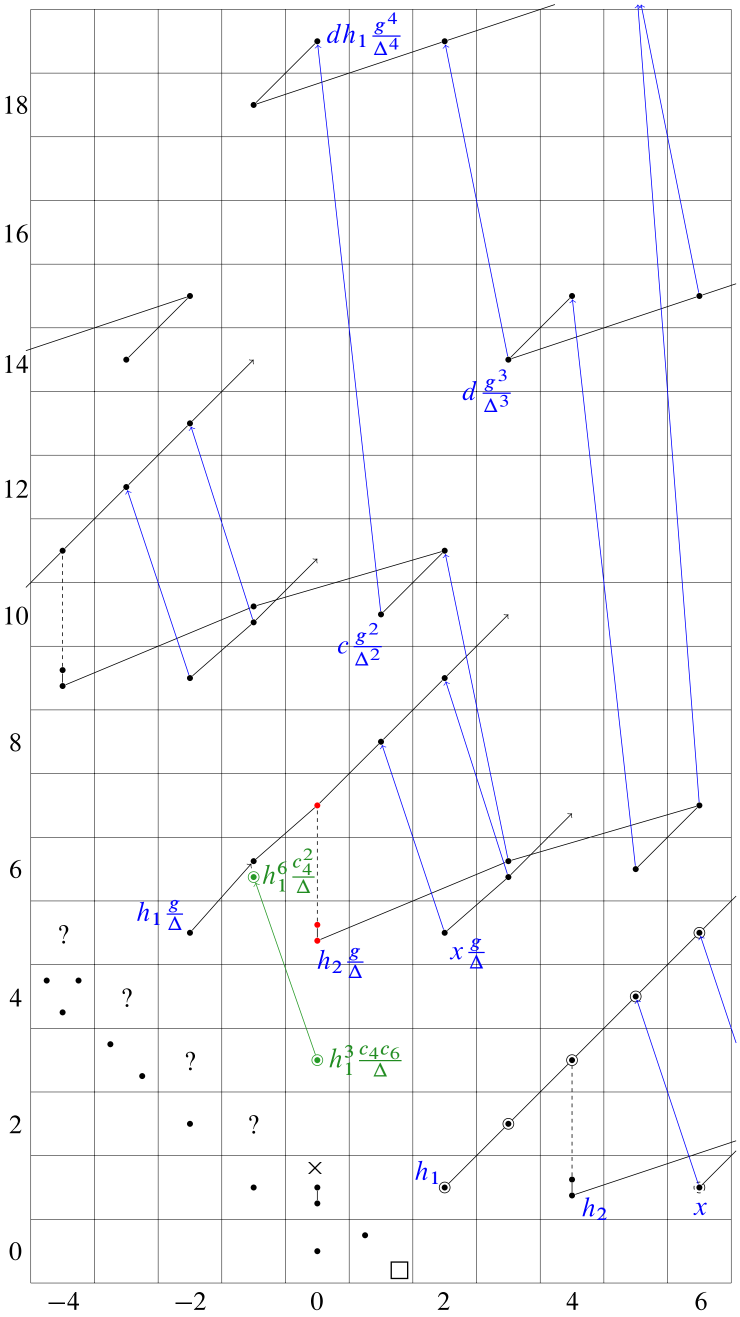


Figure 9: Homotopy fixed point spectral sequence for $\mathfrak{pic}(\mathrm{TMF}(3))^{h\mathrm{GL}_2(\mathbb{Z}/3)}$: zoomed in version with some h_1–omissions (□ denotes $\mathbb{Z}$, ● denotes $\mathbb{Z}/2$, ⦿ denotes $\mathbb{Z}/2[j]$, and × denotes $\mathbb{Z}/3$)

Remark 8.2.3 The two classes $h_1^3c_4c_6/\Delta$ and $h_1^6c_4^2/\Delta$, which we have depicted in green (in the $(0,3)$ and $(-1,6)$ positions, respectively), do not appear in the spectral sequence for $\mathrm{Tmf}\big[\frac{1}{3}\big]$, as they involve a negative power of Δ. Another difference between the Tmf and TMF situation is that in the E_2–page of the latter, there are infinite groups, isomorphic to $\mathbb{Z}/2[j]$ and generated by h_1, h_1^2, h_1^3, etc, in positions $(1,1)$, $(2,2)$, $(3,3)$, etc. Moreover, the element x in position $(1,5)$ also generates an infinite $\mathbb{Z}/2[j]$, as do all of its h_1–multiples.

Note that in the range that we are considering (namely, $t>1$), the HFPSS for the $\mathrm{GL}_2(\mathbb{Z}/3)$–action on $\mathrm{Tmf}(3)$ coincides with the descent spectral sequence for $\mathrm{Tmf}\big[\frac{1}{3}\big]$ as the sections of $\mathcal{O}^{\mathrm{top}}$ over $\overline{\mathfrak{M}}_{\mathrm{ell}}\big[\frac{1}{3}\big]$, and the differentials in the latter have been fully determined in Johan Konter's master thesis [32]. Of course, these differentials really come from the connective tmf, whose descent spectral sequence is fully computed in [6]. In these spectral sequences, d_3^o is the first nontrivial differential, followed by $d_5^o, d_7^o, d_9^o, \dots, d_{23}^o$. In particular, we have the following differentials [6, Section 8]:

$$\begin{aligned} d_3^o(c_6) &= c_4h_1^3, & d_3^o(x) &= h_1^4,\\ d_5^o(\Delta) &= gh_2, & d_7^o(4\Delta) &= gh_1^3,\\ d_9^o(\Delta^2h_1) &= g^2c, & d_{11}^o(d\Delta^2) &= g^3h_1, \end{aligned} \tag{8-7}$$

and a number of others.

Let us see now which of these differentials we can import using our Comparison Tool 5.2.4. In the $\mathrm{TMF}\big[\frac{1}{3}\big]$ spectral sequence, we have that $d_3^o(h_1^3c_4c_6/\Delta) = h_1^6c_4^2/\Delta$; in the Picard SS, the element corresponding to $h_1^3c_4c_6/\Delta$ has $t=3$, thus we *cannot* import this differential. We deal with this class later, ie in the next paragraph. However, all the other classes which are on the $s=t$ column and are h_1–power multiples of nontorsion classes, ie members of the family which we have not drawn in Figure 7, are well within the $t>3$ range, so that we can indeed conclude by Comparison Tool 5.2.4 that they either support a differential or are killed by one. For example, the h_1–multiple of the differential just discussed does happen, ie in the Picard SS we have $d_3(h_1^4c_4c_6/\Delta) = h_1^7c_4^2/\Delta$. In particular, we need not worry about these omitted classes any more.

We turn to the question of whether any differentials are supported on the $(s,t-s)=(3,0)$ position in the HFPSS for $\mathfrak{pic}(\mathrm{TMF}(3))^{h\mathrm{GL}_2(\mathbb{Z}/3)}$. For this purpose we use the universal formula (6-1) of Theorem 6.1.1, just as we did in the second proof of Proposition 7.2.7. We have that $E_2^{3,3}$ of the Picard spectrum HFPSS is $\mathbb{Z}/2[j]$ generated by $h_1^3c_4c_6/\Delta$; the corresponding element in the original HFPSS has

$$d_3^o\left(h_1^3\frac{c_4c_6}{\Delta}\right) = h_1^6\frac{c_4^2}{\Delta}.$$

Now we have that

$$\left(h_1^3\frac{c_4c_6}{\Delta}\right)^2 = h_1^6\frac{c_4^2c_6^2}{\Delta^2} = (j-12^3)h_1^6\frac{c_4^2}{\Delta} = jh_1^6\frac{c_4^2}{\Delta},$$

using the fact that $12^3\Delta = c_4^3 - c_6^2$ and that by definition, $j = c_4^3/\Delta$. Thus we conclude by (6-1) that in the Picard HFPSS, the differential $d_3\colon E_3^{3,3} \to E_3^{6,5}$ is given by

$$d_3\left(f(j)h_1^3\frac{c_4c_6}{\Delta}\right) = (f(j)+jf(j)^2)h_1^6\frac{c_4^2}{\Delta},$$

where $(f(j)h_1^3c_4c_6/\Delta)$ is an arbitrary element of $E_3^{3,3}$. However, $(f(j)+jf(j)^2)$ in $\mathbb{Z}/2[j]$ is zero only if $f(j)$ is zero, hence this d_3 is injective and has trivial kernel. (Note this is an interesting difference between the present situation and the one in Proposition 7.2.7.) Consequently, $E_4^{3,3}$ is zero.

Further use of Comparison Tool 5.2.4 determines that all the differentials we have drawn in blue in Figures 7–9 are nonzero. Note that of the classes in the $s=t$ column, ie the one which contributes to the Picard group of $\mathrm{TMF}\big[\frac{1}{3}\big]$, everything with $s \geq 8$ is killed. However, $h_2 g/\Delta$, generating a $\mathbb{Z}/4$ in $s=5$, and $h_1^3 g/\Delta$ generating a $\mathbb{Z}/2$ in $s=7$, remain. In the original spectral sequence, the first one of these supported a d_5^o and a d_{13}^o, and the second supported a d_{23}^o.

Next we need to determine the rest of the spectral sequence, ie the part which involves π_0 and π_1 of the Picard spectrum of $\mathrm{TMF}(3)$. Detailed computations for this are deferred until Appendix B. The piece in which we are most interested is $H^1\big(\mathrm{GL}_2(\mathbb{Z}/3), \pi_1\,\mathfrak{pic}(\mathrm{TMF}(3))\big)$, which is a cyclic group of order 12 according to Proposition B.1; we have also determined $H^*\big(\mathrm{GL}_2(\mathbb{Z}/3), \pi_0\,\mathfrak{pic}(\mathrm{TMF}(3))\big)$ in Proposition B.2 using a more general result of Quillen.

Now we are ready to make conclusions about the Picard group of $\mathrm{TMF}\big[\frac{1}{3}\big]$: in the $t=s$ vertical line of the HFPSS, ie the one that abuts to $\pi_0\,\mathfrak{pic}\big(\mathrm{TMF}\big[\frac{1}{3}\big]\big) = \mathrm{Pic}\big(\mathrm{TMF}\big[\frac{1}{3}\big]\big)$, nothing above the $s=7$ line survives the spectral sequence. The following might survive:

- at most a group of order 2 in position $(0,0)$,
- at most a group of order 12 in $(1,0)$,
- at most a group of order 4 in $(5,0)$, and
- at most a group of order 2 in $(7,0)$.

The upshot is that we get an upper bound of $2\times 12\times 4\times 2 = 192$ on the order of the Picard group. But $\mathrm{TMF}\big[\frac{1}{3}\big]$ is 192–periodic, so this upper bound must also be a lower bound. In conclusion, $\mathrm{Pic}\big(\mathrm{TMF}\big[\frac{1}{3}\big]\big) = \mathbb{Z}/192$, as claimed, generated by $\Sigma\,\mathrm{TMF}\big[\frac{1}{3}\big]$. □

Remark 8.2.4 As in Remark 8.1.4, we can use some of our computations to reprove Fulton and Olsson's [14] result that the moduli stack of elliptic curves $M_{\mathrm{ell}}[\frac{1}{3}]$ also has a Picard group $\mathbb{Z}/12$. Namely, we start with the knowledge that $\mathrm{Pic}(M_{\mathrm{ell}}(3))$ is trivial, as $M_{\mathrm{ell}}(3)$ is the prime spectrum of a UFD. Then, we consider the Picard HFPSS for the algebraic stack $M_{\mathrm{ell}}[\frac{1}{3}]$, which must collapse. The only contribution towards the Picard group is

$$H^1\big(\mathrm{GL}_2(\mathbb{Z}/3), \Gamma(M_{\mathrm{ell}}(3), \mathcal{O}^\times)\big),$$

which we saw by Proposition B.1 has order 12. But ω has order 12, hence $\mathrm{Pic}\big(M_{\mathrm{ell}}[\frac{1}{3}]\big)$ is cyclic of order 12.

8.3 Calculation of Pic(TMF)

In this section we will compute the Picard group of the integral periodic version of topological modular forms TMF. The result, as stated in the introduction, is:

Theorem A *The Picard group of integral* TMF *is* $\mathbb{Z}/576$*, generated by* $\Sigma\,\mathrm{TMF}$.

Proof There is no nontrivial Galois extension of the integral TMF by [40, Theorem 10.1], but we can use étale descent, as TMF is obtained as the global sections of the sheaf $\mathcal{O}^{\mathrm{top}}$ of even-periodic E_∞–rings on the moduli stack of elliptic curves. Namely, we can use Theorem 3.2.1 because the map $M_{\mathrm{ell}} \to M_{\mathrm{FG}}$ is known to be affine. The spectral sequence is

$$H^s(M_{\mathrm{ell}}, \pi_t\, \mathfrak{pic}\, \mathcal{O}^{\mathrm{top}}) \Rightarrow \pi_{t-s}\Gamma(\mathfrak{pic}\, \mathcal{O}^{\mathrm{top}}),$$

and we are interested in π_0. Using Theorem 3.2.1, the E_2–page of this spectral sequence is given by (for $t-s \geq 0$)

$$E_2^{s,t} = \begin{cases} \mathbb{Z}/2 & \text{if } t = s = 0, \\ H^s(M_{\mathrm{ell}}, \mathcal{O}^\times_{M_{\mathrm{ell}}}) & \text{if } t = 1, \\ H^s(M_{\mathrm{ell}}, \omega^{(t-1)/2}) & \text{if } t \geq 3 \text{ is odd}, \\ 0 & \text{otherwise.} \end{cases}$$

Over a field k of characteristic $\neq 2$, 3, Mumford [51] showed that

$$H^1((M_{\mathrm{ell}})_k, \mathcal{O}^\times_{M_{\mathrm{ell}}}) \simeq \mathbb{Z}/12,$$

ie the Picard group of the moduli stack is $\mathbb{Z}/12$, generated by the line bundle ω that assigns to an elliptic curve the dual of its Lie algebra. This result is also true over $\mathbb{Z}$ by the work of Fulton and Olsson [14]. However, using descent we can reprove that result. Namely, in Remarks 8.1.4 and 8.2.4 we saw that the Picard groups of both $M_{\mathrm{ell}}[\frac{1}{2}]$

and $M_{\mathrm{ell}}\big[\frac{1}{3}\big]$ are $\mathbb{Z}/12$, both generated by ω. Cover the integral stack M_{ell} by these two; their intersection is $M_{\mathrm{ell}}\big[\frac{1}{6}\big]$, which is the weighted projective stack $\operatorname{Proj}\mathbb{Z}\big[\frac{1}{6}\big][c_4, c_6]$ (with c_4 and c_6 in degrees[17] 4 and 6 respectively), and which therefore has Picard group $\mathbb{Z}/12$ also generated by ω. The descent spectral sequence for $\mathfrak{pic}$ associated to this cover gives the result.

Since $M_{\mathrm{ell}}\big[\frac{1}{6}\big]$ has no higher cohomology, the groups $H^s(M_{\mathrm{ell}}, \omega^{(t-1)/2})$, when $s > 0$, are given as the direct sum of the corresponding cohomology groups of $M_{\mathrm{ell}}\big[\frac{1}{2}\big]$ and $M_{\mathrm{ell}}\big[\frac{1}{3}\big]$. These groups, in turn, are isomorphic to

$$H^s\big(\mathrm{GL}_2(\mathbb{Z}/p), \pi_{t-1}\,\mathrm{TMF}(p)\big) = H^s\big(\mathrm{GL}_2(\mathbb{Z}/p), H^0(M_{\mathrm{ell}}(p), \omega^{(t-1)/2})\big),$$

where p is 2 or 3, as the map $M_{\mathrm{ell}}(p) \to M_{\mathrm{ell}}\big[\frac{1}{p}\big]$ is Galois, and $M_{\mathrm{ell}}(p)$ has no higher cohomology. We computed these groups in the previous examples.

The machinery of Section 5 now allows us to compare this Picard descent spectral sequence to the one which computes the homotopy groups of TMF. From Corollary 5.2.3 and an analogue of Comparison Tool 5.2.4, we conclude that the differentials involving 3–torsion classes wipe out everything above the $s = 5$ line, and those involving 2–torsion classes wipe out everything above the $s = 7$ line. These differentials are identical to what happens in the homotopy fixed point spectral sequences in the previous two examples. We conclude that the following are the only groups that can survive:

- at most a group of order 2 in $(t-s, s) = (0, 0)$,
- at most a group of order 12 in $(0, 1)$,
- at most a group of order 12 in $(0, 5)$, and
- at most a group of order 2 in $(0, 7)$.

This gives us an upper bound $2^6 3^2 = 576$ on the cardinality of π_0, which is exactly the periodicity of TMF. The spectral sequence is depicted in Figure 10. □

8.4 Calculation of Pic(Tmf)

We will now prove the following result stated in the introduction.

Theorem B *The Picard group of* Tmf *is* $\mathbb{Z} \oplus \mathbb{Z}/24$, *generated by* $\Sigma\,\mathrm{Tmf}$ *and a certain* 24*–torsion invertible module.*

[17] These are the algebraic degrees, which get doubled in topology.

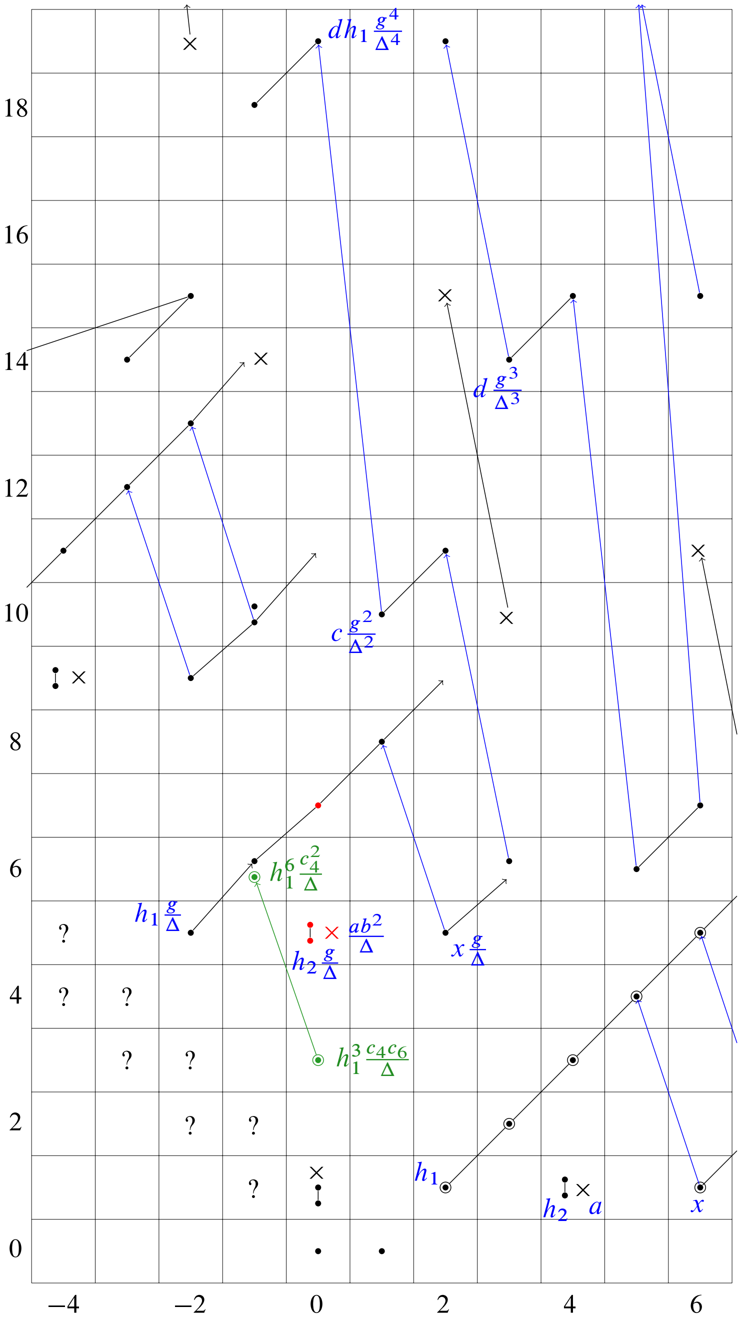


Figure 10: Descent spectral sequence for $\Gamma(\mathfrak{pic}\,\mathcal{O}^{\mathrm{top}})$ on $\mathfrak{M}_{\mathrm{ell}}$ with some h_1–omissions as in Figure 7 (□ denotes $\mathbb{Z}$, • denotes $\mathbb{Z}/2$, ⦿ denotes $\mathbb{Z}/2[j]$, and × denotes $\mathbb{Z}/3$)

While $\mathrm{Tmf}\big[\frac{1}{n}\big]$ can be described as the homotopy fixed point spectrum $\mathrm{Tmf}(n)^{h\mathrm{GL}_2(\mathbb{Z}/n)}$ for $n = 2, 3$ just as in the periodic case, the extension $\mathrm{Tmf}\big[\frac{1}{n}\big] \to \mathrm{Tmf}(n)$ is *not* Galois, and therefore we cannot use Galois descent to compute the Picard group. However, we can use Theorem 3.2.1 for the compactified moduli stack $\overline{M}_{\mathrm{ell}}$.

First, we need a lemma.

Lemma 8.4.1 *Let $\mathcal{L}$ be the line bundle on $\overline{M}_{\mathrm{ell}}$ obtained by gluing the trivial line bundles on $M_{\mathrm{ell}} = \overline{M}_{\mathrm{ell}}[\Delta^{-1}]$ and $\overline{M}_{\mathrm{ell}}[c_4^{-1}]$ via the clutching function j. Then $\mathcal{L} \simeq \omega^{-12}$.*

Proof To give a section of $\mathcal{L} \otimes \omega^{12}$ over $\overline{M}_{\mathrm{ell}}$ is equivalent to giving sections $s_1 \in \Gamma(M_{\mathrm{ell}}, \omega^{12})$ and $s_2 \in \Gamma(\overline{M}_{\mathrm{ell}}[c_4^{-1}], \omega^{12})$ such that

$$(js_1)|_{M_{\mathrm{ell}}[c_4^{-1}]} = (s_2)|_{M_{\mathrm{ell}}[c_4^{-1}]}.$$

We take $s_1 = \Delta$ and $s_2 = c_4^3$, and we get a nowhere vanishing section of $\mathcal{L} \otimes \omega^{12}$. □

Proof of Theorem B The relevant part of the Picard descent spectral sequence is similar to that of TMF, with the following exceptions: the algebraic part $H^1(\overline{M}_{\mathrm{ell}}, \mathcal{O}^{\times})$ is now $\mathbb{Z}$ generated by ω, according to Fulton and Olsson [14], and all the torsion groups are now finite, ie there are no $\mathbb{Z}/2[j]$'s appearing. In particular, $E_2^{3,3}$ is zero, and we have

- at most a group of order 2 in $(t-s, s) = (0, 0)$,
- a subquotient of $\mathbb{Z}$ in $(0, 1)$,
- at most a group of order 12 in $(0, 5)$, and
- at most a group of order 2 in $(0, 7)$

as potential contributions to the $s = t$ line of the E_∞–page. The depiction is in Figure 11.

Note that the $\mathbb{Z}/2$ in $(0, 0)$, which corresponds to a single suspension of the even-periodic spectra that Tmf is built from, is represented by $\Sigma\,\mathrm{Tmf}$ in the Picard group of Tmf. Similarly, the element $1 \in \mathbb{Z} = E_2^{0,1} = \mathrm{Pic}(\overline{M}_{\mathrm{ell}})$ corresponds to the line bundle ω, which topologically is represented by $\Sigma^2\,\mathrm{Tmf}$. Thus these groups survive to the E_∞–page and are related by an extension. The rest of the E_∞–filtration now tells us that Pic(Tmf) sits in an extension

$$0 \to A \to \mathrm{Pic}(\mathrm{Tmf}) \to \mathbb{Z} \to 0,$$

where A is a finite group of order at most 24.

We claim that $A = \mathbb{Z}/24$ and therefore $\mathrm{Pic}(\mathrm{Tmf}) = \mathbb{Z} \oplus \mathbb{Z}/24$. To see this, we will construct a line bundle $\mathcal{I}$ such that $\mathcal{I}^{\otimes 24} \simeq \mathcal{O}^{\mathrm{top}}$, but no lower power of $\mathcal{I}$ is equivalent to $\mathcal{O}^{\mathrm{top}}$.

In order to proceed with the construction, we make the preliminary observation that the modular function $j = c_4^3/\Delta$ is a homotopy class in $\pi_0\,\mathrm{TMF}[c_4^{-1}]$, ie it survives the descent spectral sequence

$$H^*(\overline{M}_{\mathrm{ell}}[\Delta^{-1}, c_4^{-1}], \omega^*) \cong H^*(M_{\mathrm{ell}}, \omega^*)[c_4^{-1}] \Rightarrow \pi_*\,\mathrm{TMF}[c_4^{-1}].$$

In fact, it is an invertible element of $\pi_0\,\mathrm{TMF}[c_4^{-1}]$. We reason as follows. The torsion in the E_2–page consists only of h_1–towers supported on the nontorsion classes, since all other torsion classes in $H^*(M_{\mathrm{ell}}, \omega^*)$ are annihilated by c_4. Therefore, when c_4 is inverted only d_3–differentials can be nonzero, and they wipe out everything above the line $s = 3$. As Δ and c_4 do not support any of those differentials, j is a permanent cycle, as is j^{-1}.

Construction 8.4.2 Consider the cover of $\overline{M}_{\mathrm{ell}}$ by $\overline{M}_{\mathrm{ell}}[\Delta^{-1}] = M_{\mathrm{ell}}$ and $\overline{M}_{\mathrm{ell}}[c_4^{-1}]$ which fit in the pushout diagram:

$$\begin{tikzcd}
\overline{M}_{\mathrm{ell}}[\Delta^{-1}, c_4^{-1}] \arrow[r] \arrow[d] & \overline{M}_{\mathrm{ell}}[\Delta^{-1}] \arrow[d] \\
\overline{M}_{\mathrm{ell}}[c_4^{-1}] \arrow[r] & \overline{M}_{\mathrm{ell}}
\end{tikzcd}$$

Let $\mathcal{J}$ be the line bundle on the *derived* moduli stack $\overline{\mathfrak{M}}_{\mathrm{ell}} = (\overline{M}_{\mathrm{ell}}, \mathcal{O}^{\mathrm{top}})$ obtained by gluing $\mathcal{O}^{\mathrm{top}}$ on $\overline{M}_{\mathrm{ell}}[\Delta^{-1}]$ and $\mathcal{O}^{\mathrm{top}}$ on $\overline{M}_{\mathrm{ell}}[c_4^{-1}]$ using the clutching function $j = c_4^3/\Delta$ on $\overline{M}_{\mathrm{ell}}[\Delta^{-1}, c_4^{-1}]$.

We claim that $\mathcal{J}$ is not a suspension of $\mathcal{O}^{\mathrm{top}}$, and that $\mathcal{I} = \Sigma^{24}\mathcal{J}$ is an element of the Picard group of order 24.

To see the first assertion, note that by Lemma 8.4.1, $\pi_0\mathcal{J}$ is ω^{-12}, so if $\mathcal{J}$ is a suspension of $\mathcal{O}^{\mathrm{top}}$, it ought to be $\Sigma^{-24}\,\mathcal{O}^{\mathrm{top}}$. However, $\Sigma^{-24}\,\mathcal{O}^{\mathrm{top}}$ restricted to $\overline{M}_{\mathrm{ell}}[\Delta^{-1}]$ is $\Sigma^{-24}\,\mathcal{O}^{\mathrm{top}}\,|_{\overline{M}_{\mathrm{ell}}[\Delta^{-1}]}$, whereas $\mathcal{J}$ restricts to $\mathcal{O}^{\mathrm{top}}\,|_{\overline{M}_{\mathrm{ell}}[\Delta^{-1}]}$.

This argument can be repeated with any power $\mathcal{J}^{\otimes m}$ such that m is not divisible by 24. In this case, $\pi_0\mathcal{J}^{\otimes m}$ is ω^{-12m}, so if $\mathcal{J}^{\otimes m}$ were a suspension of $\mathcal{O}^{\mathrm{top}}$, it would be the (–24)$m$th suspension. At the same time, $\mathcal{J}^{\otimes m}$ restricts to

$$(\mathcal{O}^{\mathrm{top}})^{\otimes m}|_{\overline{M}_{\mathrm{ell}}[\Delta^{-1}]} = \mathcal{O}^{\mathrm{top}}\,|_{\overline{M}_{\mathrm{ell}}[\Delta^{-1}]}$$

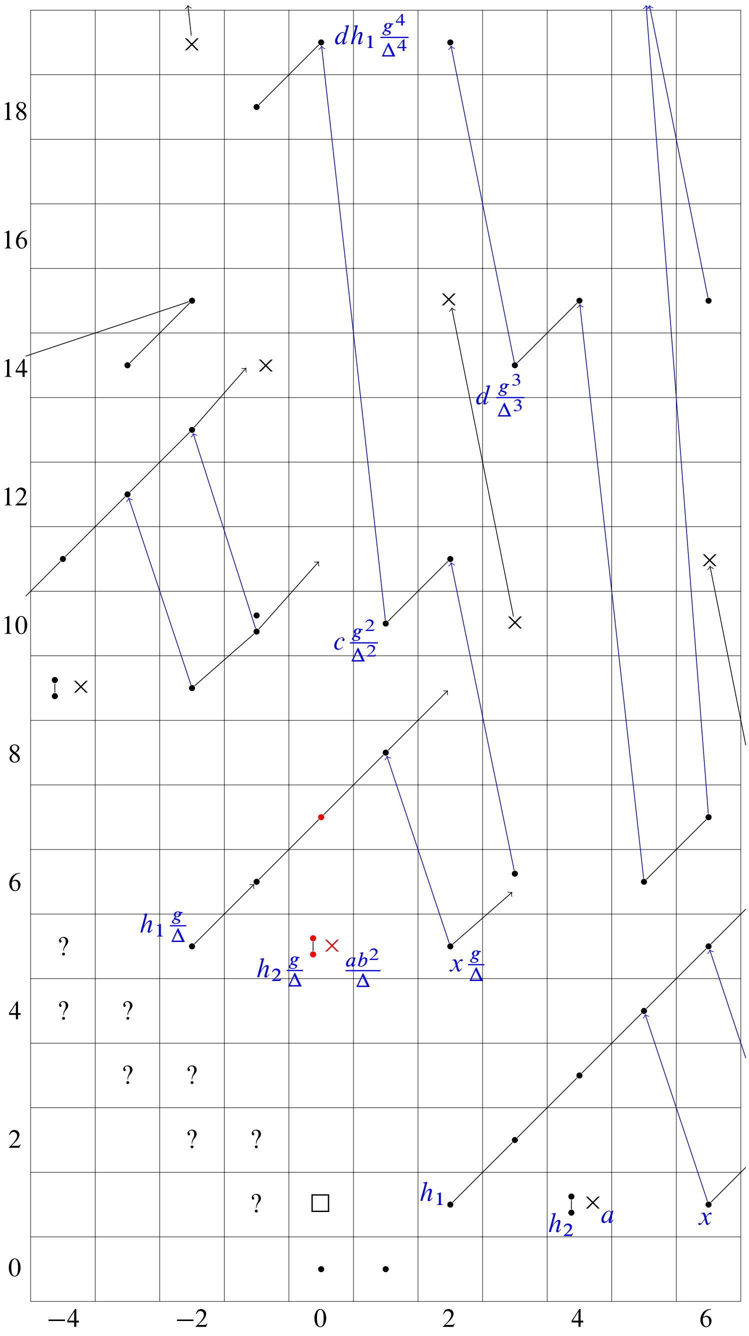


Figure 11: Descent spectral sequence for $\Gamma(\mathfrak{pic}\,\mathcal{O}^{\mathrm{top}})$ on $\overline{\mathfrak{M}}_{\mathrm{ell}}$ (□ denotes $\mathbb{Z}$, • denotes $\mathbb{Z}/2$, and × denotes $\mathbb{Z}/3$)

upon inverting Δ. If $\mathcal{J}^{\otimes m}$ were a suspension, therefore, one would have that

$$\Sigma^{-24m}\,\mathcal{O}^{\mathrm{top}}|_{\overline{M}_{\mathrm{ell}}[\Delta^{-1}]} \simeq \mathcal{O}^{\mathrm{top}}|_{\overline{M}_{\mathrm{ell}}[\Delta^{-1}]}.$$

By Theorem A, this holds if and only if m is divisible by 24.

This shows that the order of $\mathcal{J}$ in $\mathrm{Pic}(\mathcal{O}^{\mathrm{top}})/\mathbb{Z}$, where the $\mathbb{Z}$ is generated by $\Sigma\,\mathcal{O}^{\mathrm{top}}$, is at least 24. The spectral sequence argument above, however, showed that this quotient has order at most 24. □

The same analysis shows that $\mathrm{Pic}(\mathrm{Tmf}_{(2)}) = \mathbb{Z}\oplus\mathbb{Z}/8$ and $\mathrm{Pic}(\mathrm{Tmf}_{(3)}) = \mathbb{Z}\oplus\mathbb{Z}/3$, the torsion being generated by the respective localizations of $\mathcal{I}$. Moreover, when p is greater than 3, $\mathrm{Pic}(\mathrm{Tmf}_{(p)}) = \mathbb{Z}$.

8.5 Relation to the E_2–local Picard group

Notice that $\mathcal{I}$ is the only "exotic" element in all of our examples involving the various forms of topological modular forms. Let us see how it relates to the exotic piece of the Picard group of the category of E_2–local spectra, ie modules over the E_2–local sphere spectrum. The exotic phenomena only occur at $p=2$ and $p=3$, but since only the 3–primary E_2–local Picard group is known, let us concentrate on that case for the remainder of this section.

In [17], the authors compute κ_2, the exotic part of the Picard group of the category of 3–primary $K(2)$–local spectra; they show $\kappa_2 = \mathbb{Z}/3\times\mathbb{Z}/3$.

Additionally, they look at the localization map from the E_2–local category to the $K(2)$–local category and show that it induces an isomorphism $\kappa_{\mathcal{L}_2}\to\kappa_2$, where $\kappa_{\mathcal{L}_2}$ denotes the exotic E_2–local Picard group.

Consider now the commutative diagram

$$\begin{tikzcd}
\kappa_{\mathcal{L}_2} \arrow[r] \arrow[d, "t"'] & \kappa_2 \arrow[d, "t_{K(2)}"] \\
\mathrm{Pic}(\mathrm{Tmf}_{(3)}) \arrow[r] & \mathrm{Pic}(\mathrm{Tmf}_{K(2)})
\end{tikzcd}$$

in which the horizontal maps are given by $K(2)$–localization, and the vertical maps are given by smashing with Tmf and $\mathrm{Tmf}_{K(2)}$, respectively. In [17, Theorem 5.5], the authors show there is an element P of κ_2 such that $L_{K(2)}(P\wedge\mathrm{Tmf}_{K(2)})\simeq\Sigma^{48}\,\mathrm{Tmf}_{K(2)}$, ie $t_{K(2)}P = 48\in\mathbb{Z}/72\subseteq\mathrm{Pic}(\mathrm{Tmf}_{K(2)})$. Under the top horizontal isomorphism, this P lifts to an element $\widetilde{P}$ of $\kappa_{\mathcal{L}_2}$, such that $t(\widetilde{P})$ has order three in $\mathrm{Pic}(\mathrm{Tmf}_{(3)})$ and such that the $K(2)$–localization of $t(\widetilde{P})$ is $L_{K(2)}(\Sigma^{48}\,\mathrm{Tmf})$. Thus $t(\widetilde{P})$ must be twice the class of $\mathcal{I}$. In other words, the exotic element $\widetilde{P}$ of $\kappa_{\mathcal{L}_2}$ is detected as an exotic element of $\mathrm{Pic}(\mathrm{Tmf}_{(3)})$.

The other $\mathbb{Z}/3$ in κ_2, ie κ_2 modulo the subgroup generated by P, is generated by a spectrum Q such that $t_{K(2)}Q = 0$. This Q lifts to $\widetilde{Q} \in \kappa_{\mathcal{L}_2}$, still of order 3, which must map under t to an element of order 3 in $\mathrm{Pic}(\mathrm{Tmf}_{(3)})$ which is in the kernel of the bottom localization map. But there are no nontrivial elements of finite order in this kernel, hence $\widetilde{Q}$ is not detected in $\mathrm{Pic}(\mathrm{Tmf}_{(3)})$.

Perhaps at the prime 2 as well there is an element of the exotic E_2–local Picard group which is detected in the torsion of $\mathrm{Pic}(\mathrm{Tmf}_{(2)})$.

Appendices

Appendix A: Group cohomology computations for TMF(2)

In this appendix, we will compute the group cohomology for the $\mathrm{GL}_2(\mathbb{Z}/2)$–action on $\pi_0\,\mathfrak{pic}(\mathrm{TMF}(2)) = \mathbb{Z}/4$ (with trivial action), and on $\pi_1\,\mathfrak{pic}(\mathrm{TMF}(2)) = \mathrm{TMF}(2)_0^\times$ with the natural action. The group $\mathrm{GL}_2(\mathbb{Z}/2)$ is the symmetric group on three letters, so it has a (unique) normal subgroup of order 3, which we denote by C_3, with quotient C_2. We can therefore use the associated Lyndon–Hochschild–Serre spectral sequence (LHSSS)

$$H^p(C_2, H^q(C_3, M)) \Rightarrow H^{p+q}(\mathrm{GL}_2(\mathbb{Z}/2), M) \tag{A-1}$$

for $\mathrm{GL}_2(\mathbb{Z}/2)$–modules M.

Let us first deal with the easier case.

Lemma A.1 *The group cohomology for the* $\mathrm{GL}_2(\mathbb{Z}/2)$*–action on the trivial module* $\mathbb{Z}/4$ *is*

$$H^*\big(\mathrm{GL}_2(\mathbb{Z}/2), \pi_0\,\mathfrak{pic}(\mathrm{TMF}(2))\big) = \begin{cases} \mathbb{Z}/4 & \textit{if } * = 0, \\ \mathbb{Z}/2 & \textit{if } * > 0. \end{cases}$$

Proof Since 3 is invertible in $\mathbb{Z}/4$, we have that $H^*(C_3, \mathbb{Z}/4) = \mathbb{Z}/4$ concentrated in degree zero, and with trivial action by $C_2 = \mathrm{GL}_2(\mathbb{Z}/2)/C_3$. Hence the LHSSS (A-1) collapses, giving

$$H^s(\mathrm{GL}_2(\mathbb{Z}/2), \mathbb{Z}/4) = H^s(C_2, \mathbb{Z}/4),$$

which is $\mathbb{Z}/4$ for $s = 0$ and $\mathbb{Z}/2$ otherwise. □

Next we compute the group cohomology for the $\mathrm{GL}_2(\mathbb{Z}/2)$–action on $\pi_1\,\mathfrak{pic}(\mathrm{TMF}(2))$, which is the multiplicative group of units in $\pi_0\,\mathrm{TMF}(2)$. For brevity, we call this module M, and to begin with, we explicitly describe the action of $\mathrm{GL}_2(\mathbb{Z}/2)$ on M.

Let σ and τ be the generators of $\mathrm{GL}_2(\mathbb{Z}/2)$ of order 3 and 2 respectively as chosen in [63, Lemma 7.3]; of course, σ generates the normal subgroup C_3. It follows from (8-1) that M is isomorphic to $\mathbb{Z}/2\oplus\mathbb{Z}^{\oplus 3}$, where $\mathbb{Z}/2$ is multiplicatively generated by -1, and the $\mathbb{Z}$'s are multiplicatively generated by 2, s and $(s-1)$. The action is determined by [63, Lemma 7.3], where it is shown that the chosen generators σ and τ act as

$$\sigma\colon s\mapsto\frac{s-1}{s}\quad\text{and}\quad\tau\colon s\mapsto\frac{1}{s}.$$

Written additively, so that $m=(\epsilon,k,a,b)\in M$ represents $(-1)^{\epsilon}2^k s^a(s-1)^b\in\mathrm{TMF}(2)_0^{\times}$, the action is given by

$$\begin{aligned}\sigma\colon m&\mapsto(\epsilon+b,k,-a-b,a),\\ \tau\colon m&\mapsto(\epsilon+b,k,-a-b,b).\end{aligned}$$

We use this information to compute $H^*(C_3,M)$ as a C_2–module. We get that

$$H^s(C_3,M)=\begin{cases}\mathbb{Z}/2\oplus\mathbb{Z} & \text{if } s=0,\\ (\mathbb{Z}/3) & \text{if } s\equiv 0,1(4)\text{ and } s>0,\\ (\mathbb{Z}/3)_{\mathrm{sgn}} & \text{if } s\equiv 2,3(4)\text{ and } s>0.\end{cases}$$

This gives the E_2–page of the LHSSS (A-1), which must collapse and give that

$$H^s(\mathrm{GL}_2(\mathbb{Z}/2),M)=\begin{cases}\mathbb{Z}/2\oplus\mathbb{Z} & \text{if } s=0,\\ \mathbb{Z}/2\oplus\mathbb{Z}/3 & \text{if } s\equiv 1(4),\\ \mathbb{Z}/2\oplus\mathbb{Z}/2 & \text{if } s\equiv 2(4),\\ \mathbb{Z}/2 & \text{if } s\equiv 3(4),\\ \mathbb{Z}/2\oplus\mathbb{Z}/2\oplus\mathbb{Z}/3 & \text{if } s\equiv 0(4)\text{ and } s>0.\end{cases}\tag{A-2}$$

We have thus proven the following result.

Proposition A.2 *The group cohomology for the* $\mathrm{GL}_2(\mathbb{Z}/2)$*–action on* $\pi_0\,\mathfrak{pic}(\mathrm{TMF}(2))$ *is as in* (A-2). *In particular, we have that* $H^1(\mathrm{GL}_2(\mathbb{Z}/2),\mathrm{TMF}(2)_0^{\times})=\mathbb{Z}/6$.

Appendix B: Group cohomology computations for TMF(3)

This appendix is devoted to computing the group cohomology for $\mathrm{GL}_2(\mathbb{Z}/3)$ acting on $\pi_1\,\mathfrak{pic}(\mathrm{TMF}(3)_0)^{\times}$; we also determine the cohomology of $\pi_0\,\mathfrak{pic}(\mathrm{TMF}(3))=\mathbb{Z}/2$ as a simple consequence of a result of Quillen [55]. The group $\mathrm{GL}_2(\mathbb{Z}/3)$ has order 48 and has the binary tetrahedral group as a normal subgroup, in the guise of $\mathrm{SL}_2(\mathbb{Z}/3)$. We have found it difficult to compute the higher cohomology groups of $(\mathrm{TMF}(3)_0)^\times$, but since we are only using $H^1(\mathrm{GL}_2(\mathbb{Z}/3),(\mathrm{TMF}(3)_0)^\times)$ in Section 8.2, we will concentrate on computing this group only.

In this section, we denote $(\mathrm{TMF}(3)_0)^\times$ by M. From (8-1), we see that $M \subset \mathrm{TMF}(3)_0$ is isomorphic to $\mathbb{Z}/2 \oplus \mathbb{Z}/3 \oplus \mathbb{Z}^{\oplus 4}$ multiplicatively generated by $-1, \zeta$, $(1-\zeta)$, t, $(1-\zeta t)$ and $(1+\zeta^2 t)$. (To see the appearance of $(1-\zeta)$, note that $(1-\zeta)^2 = -3\zeta$.) The $\mathrm{GL}_2(\mathbb{Z}/3)$–module structure is determined in [64, Section 4.3]; to describe it, let x, y, z be the elements of $\mathrm{GL}_2(\mathbb{Z}/3)$ chosen in loc. cit. Explicitly,

$$x = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad y = \begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}, \quad z = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}.$$

Then x and y generate a quaternion group Q_8, and x, y, z generate $\mathrm{SL}_2(\mathbb{Z}/3)$. Let σ be the matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. These generate the whole group, and their action on the element $t = \gamma_1/\gamma_2$ is as determined in loc. cit.[18] to be

$$x(t) = -\frac{1}{t}, \quad y(t) = \zeta^2 \frac{1-\zeta t}{1+\zeta^2 t}, \quad z(t) = \zeta \frac{t}{1+\zeta^2 t}, \quad \sigma(t) = \frac{1}{t}.$$

The rest is determined by the fact that everything fixes $\mathbb{Z}\left[\frac{1}{3}\right] \subset \mathrm{TMF}(3)_0$, a matrix A in $\mathrm{GL}_2(\mathbb{Z}/3)$ takes ζ to $\zeta^{\det A}$, and the action respects the ring structure.

To be brutally explicit, let $m = (\epsilon, \alpha, \beta, a, b, c) \in M$ denote the element

$$(-1)^\epsilon \zeta^\alpha (1-\zeta)^\beta t^a (1-\zeta t)^b (1+\zeta^2 t)^c.$$

Then the generators x, y, z, $\sigma \in \mathrm{GL}_2(\mathbb{Z}/3)$ act as

$$\begin{aligned}
x&\colon m \mapsto (\epsilon + a + c, \alpha + b - c, \beta, -a-b-c, c, b),\\
y&\colon m \mapsto (\epsilon + b + c, \alpha - a - c, \beta, b, a, -a-b-c),\\
z&\colon m \mapsto (\epsilon, \alpha + a, \beta, a, c, -a-b-c),\\
\sigma&\colon m \mapsto (\epsilon + \beta + b, -\alpha - \beta - b + c, \beta, -a-b-c, b, c).
\end{aligned} \tag{B-1}$$

Since we know a set of generators and relations for $\mathrm{GL}_2(\mathbb{Z}/3)$, and the action is given explicitly, we can compute H^1 directly as crossed homomorphisms modulo coboundaries. We found it a little bit simpler, however, to do this for $\mathrm{SL}_2(\mathbb{Z}/3)$, and then use the Lyndon–Hochschild–Serre spectral sequence for the extension

$$1 \to \mathrm{SL}_2(\mathbb{Z}/3) \to \mathrm{GL}_2(\mathbb{Z}/3) \to C_2 \to 1,$$

[18] Actually, the formulas in loc. cit. determine a *right* action, although the *left* action that we include here is almost the same.

in which C_2 is generated by the image of $\sigma \in \mathrm{GL}_2(\mathbb{Z}/3)$. The contributions to $H^1(\mathrm{GL}_2(\mathbb{Z}/3), M)$ are from $H^1(\mathrm{SL}_2(\mathbb{Z}/3), M)^{C_2}$ and $H^1(C_2, M^{\mathrm{SL}_2(\mathbb{Z}/3)})$, and there is a potential differential

$$d_2\colon H^1(\mathrm{SL}_2(\mathbb{Z}/3), M)^{C_2} \to H^2(C_2, M^{\mathrm{SL}_2(\mathbb{Z}/3)}). \tag{B-2}$$

To compute these groups and the differential, we note that the invariants $M^{\mathrm{SL}_2(\mathbb{Z}/3)}$ are the submodule $\mathbb{Z}/2 \oplus \mathbb{Z}/3 \oplus \mathbb{Z}$ with $a = b = c = 0$. Here, $\ker(1+\sigma) = \mathrm{im}(1-\sigma)$, so that $H^1(C_2, M^{\mathrm{SL}_2(\mathbb{Z}/3)}) = 0$.

Next, suppose $f\colon \mathrm{SL}_2(\mathbb{Z}/3) \to M$ represents a class in $H^1(\mathrm{SL}_2(\mathbb{Z}/3), M)^{C_2}$, ie it is a crossed homomorphism which is σ–invariant modulo coboundaries. Since each $f(g)$ is in the kernel of the norm of g, we must have that

$$\begin{aligned}
f(x) &= (\epsilon_x, c_x, 0, a_x, -c_x, c_x),\\
f(y) &= (\epsilon_y, -a_y - c_y, 0, a_y, -a_y, c_y),\\
f(z) &= (0, \alpha_z, 0, 0, b_z, c_z).
\end{aligned}$$

The relations $x^2 = y^2$, $xyx = y$, $xz = zy^3$ and $zyx = yz$, imply that

$$a_x + c_x = a_y + c_y, \qquad b_z = -c_x, \qquad c_z = c_y, \qquad \epsilon_x = c_x + c_y, \qquad \epsilon_y = a_x.$$

One directly checks that any crossed homomorphism of this form is σ–invariant modulo coboundaries. Finally, suppose an f of this form is itself a coboundary, ie there is an $m = (\epsilon, \alpha, \beta, a, b, c) \in M$, such that $f(g) = gm - m$ for all $g \in \mathrm{SL}_2(\mathbb{Z}/3)$. Then $4b = a_x + 3c_x - 2c_y$, $a = b - a_x - c_x + 2c_y$, $c = b - c_x$ and $\alpha_z = a$. Consequently,

$$H^1(\mathrm{SL}_2(\mathbb{Z}/3), M)^{C_2} = \mathbb{Z}/12. \tag{B-3}$$

It remains to compute the differential (B-2). This is a transgression, and we have an explicit formula for it, for example in [31, Section 3.7] or [53, Section I.6]. One checks that this formula gives that d_2 is zero in our case. Thus we have proved the following.

Proposition B.1 *$H^1(\mathrm{GL}_2(\mathbb{Z}/3), \mathrm{TMF}(3)_0^\times)$ is cyclic of order 12.*

Although not directly affecting the computation of $\mathrm{Pic}\big(\mathrm{TMF}\big[\tfrac{1}{3}\big]\big)$, we record the following result of Quillen that determines a few more entries in the spectral sequence (8-6).

Proposition B.2 [55, Lemma 11] *The cohomology ring $H^*(\mathrm{GL}_2(\mathbb{Z}/3), \mathbb{Z}/2)$ is $\mathbb{Z}/2[c_1, c_2] \otimes \Lambda(e_1, e_2)$, where the cohomological degrees are $|c_i| = 2i$ and $|e_i| = 2i - 1$.*

Appendix C: Derived functors of the symmetric square

The purpose of this appendix is to prove the necessary auxiliary results on symmetric squares of cosimplicial abelian groups.

Definition C.1 Let A be an abelian group. We let $\mathrm{Sym}_2(A) = (A \otimes A)_{C_2}$ be the coinvariants for the C_2–action on $A \otimes A$ given by permuting the factors. We also let $\widetilde{\mathrm{Sym}}_2(A)$ denote the C_2–coinvariants in $(A \otimes A) \otimes \mathbb{Z}_\epsilon$ where the first factor is given the permutation action and $\mathbb{Z}_\epsilon$ is the sign representation. Note that if A is a free abelian group, then the 2–torsion in $\widetilde{\mathrm{Sym}}_2(A)$ is canonically isomorphic to $A \otimes_{\mathbb{Z}} \mathbb{F}_2$ via the "Frobenius" map

$$A/2A \to \widetilde{\mathrm{Sym}}_2(A), \quad a \mapsto a \otimes a.$$

In [54], Priddy gives a complete description of the actions of the symmetric algebra functor on cosimplicial vector spaces, or equivalently the analog of the Steenrod algebra for cosimplicial algebras. We will only need a small piece of this, which we state next. We note that the generators in question are the Steenrod squares applied to the fundamental class ι. For example, the generator in maximal degree is the cup square.

Proposition C.2 [54, Theorem 4.0.1] *Let $A^\bullet$ be a cosimplicial $\mathbb{F}_2$–vector space. Suppose that $H^{t+1}(A^\bullet) \simeq \mathbb{F}_2$ and the cohomology of $A^\bullet$ is concentrated in degree $t+1$ by a class ι. Then*

$$H^i(\mathrm{Sym}_2 A^\bullet) \simeq \begin{cases} \mathbb{F}_2 & \text{if } t+1 \le i \le 2(t+1), \\ 0 & \text{otherwise.} \end{cases}$$

Proposition C.3 *Let $t \ge 2$ and let $A^\bullet$ be a levelwise free, finitely generated cosimplicial abelian group with $H^*(A^\bullet)$ concentrated in degree $* = t+1$ and $H^{t+1}(A^\bullet) = \mathbb{Z}$ generated by ι. Then:*

(1) *If t is even, then $H^{2t+2}(\mathrm{Sym}_2 A^\bullet) \simeq \mathbb{Z}/2$, generated by ι^2.*

(2) *If t is odd, then $H^{2t+2}(\widetilde{\mathrm{Sym}}_2 A^\bullet) \simeq \mathbb{Z}/2$, generated by ι^2.*

Proof Consider first the case t even. In this case, we have maps of cosimplicial abelian groups

$$\mathrm{Sym}_2 A^\bullet \to A^\bullet \otimes A^\bullet \to \mathrm{Sym}_2 A^\bullet$$

where the first map is the norm map and the second map is projection. The composite is multiplication by two. Note that $H^{2t+2}(A^\bullet \otimes A^\bullet) \simeq \mathbb{Z}$, but since t is even, the C_2–action is the sign representation, so that the map $H^*(\mathrm{Sym}_2 A^\bullet) \to H^*(A^\bullet \otimes A^\bullet)$

must be the zero map as it lands in the C_2–invariants on cohomology. In particular, the cohomology of $\mathrm{Sym}_2(A^\bullet)$ is all annihilated by 2. By the universal coefficient theorem, it suffices to show that $H^{2t+2}(\mathrm{Sym}_2 A^\bullet \otimes_{\mathbb{Z}} \mathbb{Z}/2) \simeq \mathbb{Z}/2$ and $H^k(\mathrm{Sym}_2 A^\bullet \otimes_{\mathbb{Z}} \mathbb{Z}/2) = 0$ for $k > 2t+2$, which is the statement of Proposition C.2. In addition, we see that ι^2 is a generator, as desired, by working modulo 2.

Now suppose t is odd. Again, using the norm maps

$$\widetilde{\mathrm{Sym}}_2 A^\bullet \to A^\bullet \otimes A^\bullet \otimes \epsilon \to \widetilde{\mathrm{Sym}}_2 A^\bullet,$$

we find that the cohomology of $\widetilde{\mathrm{Sym}}_2 A^\bullet$ is annihilated by 2. We note that at the level of cosimplicial abelian groups $\widetilde{\mathrm{Sym}}_2 A^\bullet \otimes_{\mathbb{Z}} \mathbb{F}_2 \simeq \mathrm{Sym}_2 A^\bullet \otimes_{\mathbb{Z}} \mathbb{F}_2$, but working with the underived tensor product is problematic here because $\widetilde{\mathrm{Sym}}_2 A^\bullet$ has 2–torsion. If we take the derived tensor product

$$\widetilde{\mathrm{Sym}}_2(A^\bullet) \overset{\mathbb{L}}{\otimes} \mathbb{F}_2,$$

we obtain in addition a copy of $A^\bullet \otimes_{\mathbb{Z}} \mathbb{F}_2$ (ie the 2–torsion in $\widetilde{\mathrm{Sym}}_2 A^\bullet$) in π_1 that does not contribute in the relevant dimensions, so we may ignore it. Now, by Proposition C.2, we know that

$$H^k(\widetilde{\mathrm{Sym}}_2 A^\bullet \otimes_{\mathbb{Z}} \mathbb{F}_2) \simeq \begin{cases} \mathbb{F}_2 & \text{for } k = 2t+2, \\ 0 & \text{for } k > 2t+2. \end{cases}$$

So we can apply the universal coefficient theorem as in the previous case. We conclude that ι^2 is a generator similarly. □

References

[1] **J F Adams**, **S B Priddy**, *Uniqueness of BSO*, Math. Proc. Cambridge Philos. Soc. 80 (1976) 475–509 MR

[2] **M Ando**, **A J Blumberg**, **D Gepner**, **M J Hopkins**, **C Rezk**, *Units of ring spectra, orientations and Thom spectra via rigid infinite loop space theory*, J. Topol. 7 (2014) 1077–1117 MR

[3] **V Angeltveit**, *A_∞–obstruction theory and the strict associativity of E/I*, preprint (2004) Available at `http://tinyurl.com/Angeltveit-Ainfinity-pdf`

[4] **A Baker**, **B Richter**, *Invertible modules for commutative $\mathbb{S}$–algebras with residue fields*, Manuscripta Math. 118 (2005) 99–119 MR

[5] **P Balmer**, *Tensor triangular geometry*, from "Proceedings of the International Congress of Mathematicians" (R Bhatia, A Pal, G Rangarajan, V Srinivas, M Vanninathan, editors), volume 2, Hindustan Book Agency, New Delhi (2010) 85–112 MR

[6] **T Bauer**, *Computation of the homotopy of the spectrum* `tmf`, from "Groups, homotopy and configuration spaces" (N Iwase, T Kohno, R Levi, D Tamaki, J Wu, editors), Geom. Topol. Monogr. 13 (2008) 11–40 MR

[7] **M Behrens**, *Congruences between modular forms given by the divided β family in homotopy theory*, Geom. Topol. 13 (2009) 319–357 MR

[8] **J F Carlson**, **J Thévenaz**, *The classification of torsion endo-trivial modules*, Ann. of Math. 162 (2005) 823–883 MR

[9] **F R Cohen**, **T J Lada**, **J P May**, *The homology of iterated loop spaces*, Lecture Notes in Mathematics 533, Springer, Berlin (1976) MR

[10] **E Dade**, *Endo-permutation modules over p–groups, II*, Ann. of Math. 108 (1978) 317–346 MR

[11] **P Deligne**, **M Rapoport**, *Les schémas de modules de courbes elliptiques*, from "Modular functions of one variable, II" (P Deligne, W Kuyk, editors), Lecture Notes in Math. 349, Springer, Berlin (1973) 143–316 MR

[12] **P Freyd**, *Splitting homotopy idempotents*, from "Proc. Conf. Categorical Algebra" (S Eilenberg, D K Harrison, S Mac Lane, H Röhrl, editors), Springer, New York (1966) 173–176 MR

[13] **P Freyd**, *Stable homotopy*, from "Proc. Conf. Categorical Algebra" (S Eilenberg, D K Harrison, S Mac Lane, H Röhrl, editors), Springer, New York (1966) 121–172 MR

[14] **W Fulton**, **M Olsson**, *The Picard group of $\mathcal{M}_{1,1}$*, Algebra Number Theory 4 (2010) 87–104 MR

[15] **D Gepner**, **T Lawson**, *Brauer groups and Galois cohomology of commutative $\mathbb{S}$–algebras*, preprint (2016) arXiv

[16] **P G Goerss**, *Topological modular forms [after Hopkins, Miller and Lurie], Exp. No.* 1005, from "Séminaire Bourbaki 2008/2009", Astérisque 332, Société Mathématique de France, Paris (2010) 221–255 MR

[17] **P Goerss**, **H-W Henn**, **M Mahowald**, **C Rezk**, *On Hopkins' Picard groups for the prime 3 and chromatic level 2*, preprint (2012) arXiv

[18] **J P C Greenlees**, **J P May**, *Completions in algebra and topology*, from "Handbook of algebraic topology" (I M James, editor), North-Holland, Amsterdam (1995) 255–276 MR

[19] **A Grothendieck**, *Eléments de géométrie algébrique, IV: Étude locale des schémas et des morphismes de schémas, III*, Inst. Hautes Études Sci. Publ. Math. 28 (1966) 5–255

[20] **R Hartshorne**, *Algebraic geometry*, Graduate Texts in Mathematics 52, Springer, New York (1977) MR

[21] **D Heard**, *Morava modules and the $K(n)$–local Picard group*, PhD thesis, University of Melbourne (2014) Available at `http://hdl.handle.net/11343/42040`

[22] **D Heard**, **V Stojanoska**, *K–theory, reality, and duality*, J. K-Theory 14 (2014) 526–555 MR

[23] **M Hill**, **L Meier**, *All about* $\mathrm{Tmf}_1(3)$, preprint (2015) arXiv

[24] **M J Hopkins**, *$K(1)$–local E_∞–ring spectra*, from "Topological modular forms" (C L Douglas, J Francis, A G Henriques, M A Hill, editors), Math. Surveys Monogr. 201, Amer. Math. Soc., Providence, RI (2014) 287–302 MR

[25] **M J Hopkins**, **N J Kuhn**, **D C Ravenel**, *Generalized group characters and complex oriented cohomology theories*, J. Amer. Math. Soc. 13 (2000) 553–594 MR

[26] **M J Hopkins**, **M Mahowald**, **H Sadofsky**, *Constructions of elements in Picard groups*, from "Topology and representation theory" (E M Friedlander, M E Mahowald, editors), Contemp. Math. 158, Amer. Math. Soc., Providence, RI (1994) 89–126 MR

[27] **M Hovey**, **H Sadofsky**, *Invertible spectra in the $E(n)$–local stable homotopy category*, J. London Math. Soc. 60 (1999) 284–302 MR

[28] **S B Iyengar**, **G J Leuschke**, **A Leykin**, **C Miller**, **E Miller**, **A K Singh**, **U Walther**, *Twenty-four hours of local cohomology*, Graduate Studies in Mathematics 87, Amer. Math. Soc., Providence, RI (2007) MR

[29] **Y Kamiya**, **K Shimomura**, *Picard groups of some local categories*, Publ. Res. Inst. Math. Sci. 43 (2007) 303–314 MR

[30] **N M Katz**, **B Mazur**, *Arithmetic moduli of elliptic curves*, Annals of Mathematics Studies 108, Princeton Univ. Press (1985) MR

[31] **H Koch**, *Galois theory of p–extensions*, Springer, Berlin (2002) MR

[32] **J Konter**, *The homotopy groups of the spectrum Tmf*, master's thesis, Utrecht University (2012) arXiv

[33] **T Lawson**, *Strictly commutative elements of E_∞–spaces*, MathOverflow (2013) Available at `http://mathoverflow.net/q/135712`

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

[35] **J Lurie**, *Chromatic homotopy theory*, lecture notes (2010) Available at `http://math.harvard.edu/~lurie`

[36] **J Lurie**, *Derived algebraic geometry, IX: Closed immersions*, unpublished manuscript (2011) Available at `http://math.harvard.edu/~lurie`

[37] **J Lurie**, *Derived algebraic geometry, VII: Spectral schemes*, unpublished manuscript (2011) Available at `http://math.harvard.edu/~lurie`

[38] **J Lurie**, *Derived algebraic geometry, VIII: Quasi-coherent sheaves and Tannaka duality theorems*, unpublished manuscript (2011) Available at `http://math.harvard.edu/~lurie`

[39] **J Lurie**, *Higher algebra*, unpublished manuscript (2012) Available at `http://math.harvard.edu/~lurie`

[40] **A Mathew**, *The Galois group of a stable homotopy theory*, Adv. Math. 291 (2016) 403–541 MR

[41] **A Mathew**, *Residue fields for a class of rational* $\mathbf{E}_\infty$*–rings and applications*, J. Pure Appl. Algebra 221 (2017) 707–748 MR

[42] **A Mathew**, **L Meier**, *Affineness and chromatic homotopy theory*, J. Topol. 8 (2015) 476–528 MR

[43] **A Mathew**, **N Naumann**, **J Noel**, *On a nilpotence conjecture of J P May*, J. Topol. 8 (2015) 917–932 MR

[44] **A Mathew**, **V Stojanoska**, *Fibers of partial totalizations of a pointed cosimplicial space*, Proc. Amer. Math. Soc. 144 (2016) 445–458 MR

[45] **J P May**, *The geometry of iterated loop spaces*, Lectures Notes in Mathematics 271, Springer, Berlin (1972) MR

[46] **J P May**, E_∞ *ring spaces and* E_∞ *ring spectra*, Lecture Notes in Mathematics 577, Springer, Berlin (1977) MR

[47] **J P May**, *Picard groups, Grothendieck rings, and Burnside rings of categories*, Adv. Math. 163 (2001) 1–16 MR

[48] **L Meier**, *United elliptic homology*, PhD thesis, University of Bonn (2012) Available at `http://hss.ulb.uni-bonn.de/2012/2969/2969.htm`

[49] **J Milnor**, *Introduction to algebraic* K*–theory*, Annals of Mathematics Studies 72, Princeton Univ. Press (1971) MR

[50] **R E Mosher**, **M C Tangora**, *Cohomology operations and applications in homotopy theory*, Harper & Row, New York (1968) MR

[51] **D Mumford**, *Picard groups of moduli problems*, from "Arithmetical Algebraic Geometry" (O F G Schilling, editor), Harper & Row, New York (1965) 33–81 MR

[52] **A Neeman**, *Triangulated categories*, Annals of Mathematics Studies 148, Princeton Univ. Press (2001) MR

[53] **J Neukirch**, **A Schmidt**, **K Wingberg**, *Cohomology of number fields*, 2nd edition, Grundl. Math. Wissen. 323, Springer, Berlin (2008) MR

[54] **S Priddy**, Mod$-p$ *right derived functor algebras of the symmetric algebra functor*, J. Pure Appl. Algebra 3 (1973) 337–356 MR

[55] **D Quillen**, *On the cohomology and* K*–theory of the general linear groups over a finite field*, Ann. of Math. 96 (1972) 552–586 MR

[56] **D C Ravenel**, *Nilpotence and periodicity in stable homotopy theory*, Annals of Mathematics Studies 128, Princeton Univ. Press (1992) MR

[57] **C Rezk**, *Notes on the Hopkins–Miller theorem*, from "Homotopy theory via algebraic geometry and group representations" (M Mahowald, S Priddy, editors), Contemp. Math. 220, Amer. Math. Soc., Providence, RI (1998) 313–366 MR

[58] **E Riehl**, *Categorical homotopy theory*, New Mathematical Monographs 24, Cambridge Univ. Press (2014) MR

[59] **J Rognes**, *Galois extensions of structured ring spectra, Stably dualizable groups*, Mem. Amer. Math. Soc. 898, Amer. Math. Soc., Providence, RI (2008) MR

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

[61] **J-P Serre**, *Cohomologie modulo* 2 *des complexes d'Eilenberg–MacLane*, Comment. Math. Helv. 27 (1953) 198–232 MR

[62] **J H Silverman**, *The arithmetic of elliptic curves*, Graduate Texts in Mathematics 106, Springer, New York (1986) MR

[63] **V Stojanoska**, *Duality for topological modular forms*, Doc. Math. 17 (2012) 271–311 MR

[64] **V Stojanoska**, *Calculating descent for* 2*–primary topological modular forms*, from "An alpine expedition through algebraic topology" (C Ausoni, K Hess, B Johnson, W Lück, J Scherer, editors), Contemp. Math. 617, Amer. Math. Soc., Providence, RI (2014) 241–258 MR

Department of Mathematics, Harvard University
One Oxford Street, Cambridge, MA 02138, United States

Department of Mathematics, University of Illinois at Urbana-Champaign
1409 W Green Steet, Urbana, IL 61801, United States

amathew@math.harvard.edu, vesna@illinois.edu

Proposed: Mark Behrens
Seconded: Haynes Miller, Jesper Grodal

Received: 22 October 2014
Revised: 8 October 2015

msp

Geometry & Topology 20 (2016) 3219–3332

Combinatorial tangle Floer homology

INA PETKOVA
VERA VÉRTESI

We extend the idea of bordered Floer homology to knots and links in S^3: Using a specific Heegaard diagram, we construct gluable combinatorial invariants of tangles in S^3, D^3, and $I \times S^2$. The special case of S^3 gives back a stabilized version of knot Floer homology.

57M27, 57R58

1 Introduction

Knot Floer homology is a categorification of the Alexander polynomial, defined by Ozsváth and Szabó [16], and independently by Rasmussen [23], in the early 2000s. To a knot or a link one associates a filtered graded chain complex over the field of two elements $\mathbb{F}_2$ or over a polynomial ring $\mathbb{F}_2[U_1, \dots, U_n]$. The filtered chain homotopy type of this complex is a powerful invariant of the knot. For example, it detects genus (see Ozsváth and Szabó [15]), fiberedness (see Ghiggini [2] and Ni [13]), and gives a bound on the four-ball genus (see Ozsváth and Szabó [14]). The definition of knot Floer homology is based on finding a Heegaard diagram presentation for the knot and defining a chain complex by counting certain pseudoholomorphic curves in a symmetric product of the Heegaard surface. Suitable choices of Heegaard diagrams (for example, grid diagrams as in Manolescu, Ozsváth and Sarkar [11] and Manolescu, Ozsváth, Szabó, and Thurston [12], or nice diagrams as in Sarkar and Wang [25]) lead to combinatorial descriptions of knot Floer homology. However, in its nature knot Floer homology is a "global" invariant — one needs a picture of the entire knot to define it — and local modifications are only partially understood; see for example Ozsváth and Szabó [16; 20] and Manolescu [10].

Around the same time that knot Floer homology came to life, Khovanov [3] introduced another knot invariant, a categorification of the Jones polynomial now known as Khovanov homology. Khovanov's construction is somewhat simpler in nature, as one builds a chain complex generated by the different resolutions of the knot. Khovanov homology has an extension to tangles [4], thus local modifications can be understood on a categorical level.

Published: 21 December 2016 DOI: 10.2140/gt.2016.20.3219

In this paper, we extend knot Floer homology by defining a combinatorial Heegaard Floer type invariant for tangles. Note that a similar extension exists for Heegaard Floer homology, which is an invariant of closed 3–manifolds, generalizing it to manifolds with boundary; see Lipshitz, Ozsváth, and Thurston [8]. This extension is called bordered Floer homology.

1.1 Tangle Floer invariants

A tangle (see Figure 1 and Section 2.2 for precise definitions) is a properly embedded 1–manifold in D^3 or $I \times S^2$. Inspired by Lipshitz, Ozsváth, and Thurston [7], we define

- a differential graded algebra $\mathcal{A}(\mathcal{P})$ for any finite set of signed points $\mathcal{P}$ on the equator of S^2;
- a right type A module $\widehat{CFTA}(\mathcal{T})$ over $\mathcal{A}(\partial\mathcal{T})$ for any tangle $\mathcal{T}$ in D^3;
- a left type D module $\widehat{CFDT}(\mathcal{T})$ over $\mathcal{A}(-\partial\mathcal{T})$ for any tangle $\mathcal{T}$ in D^3;
- a left–right $\mathcal{A}(-\partial^0\mathcal{T})$-$\mathcal{A}(\partial^1\mathcal{T})$ type DA bimodule $\widehat{CFDTA}(\mathcal{T})$ for any tangle $\mathcal{T}$ in $I \times S^2$.

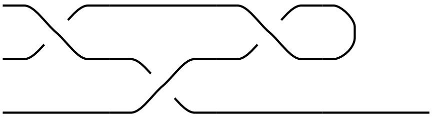


Figure 1: A projection of a tangle in $S^2 \times I$

The above (bi)modules are topological invariants of the tangle. (See Theorems 10.4, 10.2 and 10.7 for the precise statements.)

Theorem 1.1 *For a tangle $\mathcal{T}$ in D^3 the type A equivalence class of the module $\widehat{CFTA}(\mathcal{T})$ is a topological invariant of $\mathcal{T}$, and the type D equivalence class of the module $\widehat{CFDT}(\mathcal{T})$ is a topological invariant of $\mathcal{T}$. For a tangle $\mathcal{T}$ in $S^2 \times I$ the type DA equivalence class of the bimodule $\widehat{CFDTA}(\mathcal{T})$ is a topological invariant of $\mathcal{T}$.*

Furthermore, the invariants behave well under compositions of tangles. (See Theorem 12.4 and Corollary 12.5 for the precise statement.)[1]

[1] In each of the equivalences in Theorems 1.2 and 1.3, the left-hand side should also be tensored with $V^{\otimes(|\mathcal{T}_1|+|\mathcal{T}_2|-|\mathcal{T}_1\circ\mathcal{T}_2|)}$, where $V=\mathbb{F}_2\oplus\mathbb{F}_2$ has one summand in bigrading $(0,0)$ and the other summand in bigrading $(-1,-1)$. This is discussed in the full statements of the theorems, and omitted here for simplicity.

Theorem 1.2 *Suppose that $\mathcal{T}_1$ and $\mathcal{T}_2$ are tangles in $S^2 \times I$ such that $\partial^1 \mathcal{T}_1 = -\partial^0 \mathcal{T}_2$. Then, up to type DA equivalence,*

$$\widehat{CFDTA}(\mathcal{T}_1 \circ \mathcal{T}_2) \simeq \widehat{CFDTA}(\mathcal{T}_1) \mathbin{\widetilde{\otimes}} \widehat{CFDTA}(\mathcal{T}_2).$$

Thus, the above definitions give a functor from the category of oriented tangles $\mathcal{OTAN}$ to the category of bigraded type DA bimodules up to type DA equivalence. In other words, our invariant behaves like a $(0+1)$–dimensional TQFT.[2]

Note that there are analogs of Theorem 1.2 if one of the tangles is in D^3. When $\mathcal{T}_1$ and $\mathcal{T}_2$ are both in D^3, their composition $\mathcal{T}_1 \circ \mathcal{T}_2$ is a knot (or a link), and we recover knot Floer homology:

Theorem 1.3 *Suppose that $\mathcal{T}_1$ and $\mathcal{T}_2$ are tangles in D^3 with $\partial \mathcal{T}_1 = -\partial \mathcal{T}_2$, and let $K = \mathcal{T}_1 \circ \mathcal{T}_2$ be their composition. Then*

$$\widehat{CFK}(K) \otimes W \simeq \widehat{CFTA}(\mathcal{T}_1) \mathbin{\widetilde{\otimes}} \widehat{CFDT}(\mathcal{T}_2),$$

where $W = \mathbb{F}_2 \oplus \mathbb{F}_2$ with Maslov and Alexander bigradings (M, A) equal to $(0,0)$ and $(-1,0)$.

The combinatorial description of the invariants depends on the use of a certain Heegaard diagram associated to the tangle (see Figure 2). This diagram is "nice" in the sense of Sarkar and Wang [25]. The use of this diagram enables a purely combinatorial description of the generators, as partial matchings of a bipartite graph associated to the tangle. (See Figure 3 for an example.)

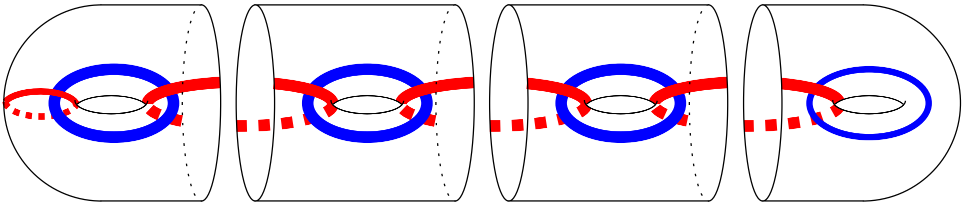


Figure 2: A Heegaard diagram associated to a tangle. The thick lines denote parallel α– and β–curves. The number of twice punctured tori in the middle depends on how complicated the tangle is. This figure shows the Heegaard diagram for a closed link. Diagrams for tangles can be obtained by deleting one or both of the once punctured tori from the sides.

[2] Note that it is not a proper TQFT as the target is not the category of vector spaces, and the functor does not respect the monoidal structure of the categories. In fact there is no obvious monoidal structure on the category of type DA structures.

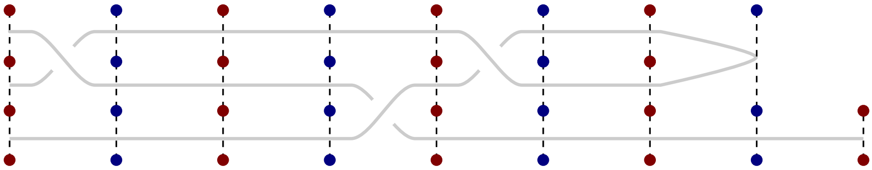


Figure 3: The bipartite graph associated to the tangle of Figure 1. The edges (not drawn) are between the consecutive vertex-sets.

Here we develop two versions of the invariants: one over $\mathbb{F}_2$, which we call a *tilde* version, and an enhanced *minus* version over $\mathbb{F}_2[U_1, \dots, U_n]$. As Theorem 1.3 depends only on a Heegaard diagram description, it holds for both versions. However, we currently only have proofs for the tilde versions of Theorems 1.1 and 1.2. This is due to the fact that our proofs rely on analytic techniques. In Section 5.3 we give evidence for the existence of completely combinatorial proofs of Theorems 1.1 and 1.2 in the minus version.

We also develop an ungraded tilde version of tangle Floer homology for tangles in arbitrary manifolds with boundary S^2 or $S^2 \amalg S^2$. Versions of the above theorems hold in this more general case too; see Theorems 10.2, 10.4, 10.7 and 12.4 and Corollary 12.5.

This TQFT-like description of knot Floer homology allows one to localize questions in Heegaard Floer homology. For instance, in a subsequent note we show that there is a skein exact sequence for tangles. The theory has the potential to help understand the change of knot Floer homology under more complicated local modifications such as mutations, or, for example, help understand the rank of the knot Floer homology of periodic knots.

We hope that our construction may provide a new bridge between Khovanov homology and knot Floer homology. Rasmussen [24] conjectures a spectral sequence connecting the two. It is possible that a relationship between the two theories can be found for simple tangles, and used to prove the conjecture.

The Jones polynomial can be defined in the Reshetikhin–Turaev way, using the vector representation of the quantum algebra $U_q(\mathfrak{sl}_2)$ and, since Khovanov's seminal work on categorifying the Jones polynomial, a program for categorification of quantum groups has begun. Similarly to the Jones polynomial construction, one can see the Alexander polynomial as a quantum invariant coming from the vector representation V of $U_q(\mathfrak{gl}(1|1))$; see Sartori [26] and Viro [27]. However, the categorification $\widehat{HFK}$ of the Alexander polynomial has not yet been understood on a representation theory level. In a future paper we show that the decategorification of tangle Floer homology is a tensor power of the vector representation of $U_q(\mathfrak{gl}(1|1))$. We believe that we can build on the structures from this paper to obtain a full categorification of the tensor powers of the vector representation of $U_q(\mathfrak{gl}(1|1))$.

1.2 Further remarks

Knot Floer homology is defined by counting holomorphic curves in a symmetric product of a Heegaard surface and, for different versions, the projection of those curves to the Heegaard surface is allowed or not allowed to cross two special sets of basepoints $\mathbb{X}$ and $\mathbb{O}$. We develop a theory for tangles that counts curves which cross only $\mathbb{O}$. While it is hard to define invariants that count curves which cross both $\mathbb{X}$ and $\mathbb{O}$, it is straightforward to modify the definitions to count curves that cross $\mathbb{X}$ or $\mathbb{O}$, but not both. Further, the invariants defined in this paper can be extended over $\mathbb{Z}$.

The structures defined in Section 3 are completely combinatorial, and an algorithm could be programmed to compute the invariants for simple tangles and obtain the knot Floer homology of some new knots. Knots with periodic behavior and knots with low bridge number relative to their grid number are especially suitable.

1.3 Organization

After a brief introduction of the relevant algebraic structures in Section 2, we turn to defining the invariants from a diagrammatic viewpoint in Section 3. In Section 4, we describe the same invariants using a class of diagrams called bordered grid diagrams, as this approach is more suited for some of the proofs and provides a bridge between Section 4 and Sections 7–12. Finally, the definitions of the tangle invariants are given in Section 5, and their relation to knot Floer homology is proved in Section 6.

Sections 7–12 are devoted to proving invariance by building up a complete holomorphic theory for tangles in 3–manifolds. The geometric structures (marked spheres) associated to the algebras are introduced in Section 7, then Section 8 describes the various Heegaard diagrams corresponding to tangles in 3–manifolds. The moduli spaces corresponding to these Heegaard diagrams are defined in Section 9. Then the definitions of the general invariants are given in Section 10. The gradings from Section 3.4 are extended to the general setting in Section 11. Section 12 contains the full statements and proofs of Theorems 1.2 and 1.3.

Acknowledgments We would like to thank Ko Honda, Mikhail Khovanov, Antony Licata, Robert Lipshitz, Peter Ozsváth, Zoltán Szabó and Yin Tian for helpful discussions. We also thank Vladimir Fock, Paolo Ghiggini, Eli Grigsby, Ciprian Manolescu and András Stipsicz for useful comments. Petkova was partially supported by the

AMS-Simons Travel Grant. Vértesi was supported by ERC Geodycon, OTKA grant number NK81203 and NSF grant number 1104690. This collaboration began in the summer of 2013 during the CRM conference "Low-dimensional topology after Floer". We thank the organizers and CRM for their hospitality. We also thank the referee for many helpful suggestions.

2 Preliminaries

2.1 Modules, bimodules, and tensor products

In this paper, we work with the same types of algebraic structures used in bordered Floer homology; see [8; 9]. Below we recall the main definitions. For more detail, see [9, Section 2].

Let A be a unital differential graded algebra with differential d and multiplication μ over a base ring $\boldsymbol{k}$. In this paper, $\boldsymbol{k}$ will always be a direct sum of copies of $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$. For the algebras we define in the later sections, the base ring for all modules and tensor products is the ring of idempotents.

A *(right)* $\mathcal{A}_\infty$*–module over* A, or a *type* A *structure over* A is a graded module M over $\boldsymbol{k}$, equipped with maps

$$m_i\colon M \otimes A[1]^{\otimes(i-1)} \to M[1]$$

for $i \geq 1$, satisfying the compatibility conditions

$$\begin{aligned} 0 = \sum_{i+j=n+1} & m_i(m_j(x, a_1, \dots, a_{j-1}), \dots, a_{n-1}) \\ & + \sum_{i=1}^{n-1} m_n(x, a_1, \dots, a_{i-1}, d(a_i), \dots, a_{n-1}) \\ & \quad + \sum_{i=1}^{n-2} m_{n-1}(x, a_1, \dots, a_{i-1}, (\mu(a_i, a_{i+1})), \dots, a_{n-1}). \end{aligned}$$

A type A structure is *strictly unital* if $m_2(x,1) = x$ and $m_i(x, a_1, \dots, a_{i-1}) = 0$ whenever $i > 2$ and some a_j is in $\boldsymbol{k}$. We assume all type A structures to be strictly unital.

We say that M is *bounded* if $m_i = 0$ for all sufficiently large i.

A (*left*) *type* D *structure* over A is a graded $\boldsymbol{k}$–module N, equipped with a homogeneous map

$$\delta\colon N \to (A \otimes N)[1]$$

satisfying the compatibility condition

$$(d \otimes \mathrm{id}_N) \circ \delta + (\mu \otimes \mathrm{id}_N) \circ (\mathrm{id}_A \otimes \delta) \circ \delta = 0.$$

We can define maps

$$\delta_k \colon N \to (A^{\otimes k} \otimes N)[k]$$

inductively by

$$\delta_k = \begin{cases} \mathrm{id}_N & \text{for } k = 0, \\ (\mathrm{id}_A \otimes \delta_{k-1}) \circ \delta & \text{for } k \geq 1. \end{cases}$$

A type D structure is *bounded* if, for any $x \in N$, $\delta_i(x) = 0$ for all sufficiently large i.

One can similarly define left type A structures and right type D structures.

If M is a right $\mathcal{A}_\infty$–module over A and N is a left type D structure, and at least one of them is bounded, we can define the *box tensor product* $M \boxtimes N$ to be the vector space $M \otimes N$ with differential

$$\partial \colon M \otimes N \to (M \otimes N)[1]$$

defined by

$$\partial = \sum_{k=1}^{\infty} (m_k \otimes \mathrm{id}_N) \circ (\mathrm{id}_M \otimes \delta_{k-1}).$$

The boundedness condition guarantees that the above sum is finite. In that case, $\partial^2 = 0$ and $M \boxtimes N$ is a graded chain complex. In general (boundedness not required), one can think of a type D structure as a left $\mathcal{A}_\infty$–module and take an $\mathcal{A}_\infty$ tensor product $\widetilde{\otimes}$; see [8, Section 2.2].

Given unital differential graded algebras A and B over $\boldsymbol{k}$ and $\boldsymbol{j}$ with differential and multiplication d_A, d_B, μ_A and μ_B, respectively, four types of bimodules can be defined in a similar way: types *DD*, *AA*, *DA* and *AD*. See [9, Section 2.2.4].

An *$\mathcal{A}_\infty$–bimodule* or *type AA bimodule over A and B* is a graded $(\boldsymbol{k}, \boldsymbol{j})$–bimodule M, together with degree 0 maps

$$m_{i,1,j} \colon A[1]^{\otimes i} \otimes M \otimes B[1]^{\otimes j} \to M[1],$$

subject to compatibility conditions analogous to those for A structures; see [9, Equation 2.2.38].

We assume all *AA* bimodules to be *strictly unital*, ie $m_{1,1,0}(1, x) = x = m_{0,1,1}(x, 1)$ and $m_{i,1,j}(a_1, \dots, a_i, x, b_1, \dots, b_j) = 0$ if $i + j > 1$ and some a_i or b_j lies in $\boldsymbol{k}$ or $\boldsymbol{j}$.

A *type DA bimodule over A and B* is a graded $(\boldsymbol{k}, \boldsymbol{j})$–bimodule M, together with degree 0, $(\boldsymbol{k}, \boldsymbol{j})$–linear maps

$$\delta^1_{1+j}\colon M \otimes B[1]^{\otimes j} \to A \otimes M[1],$$

satisfying a compatibility condition combining those for A and D structures; see [9, Definition 2.2.42].

A *type AD structure* can be defined similarly, with the roles of A and B interchanged.

A *type DD structure over A and B* is a type D structure over $A \otimes_{\mathbb{F}_2} B^{\mathrm{op}}$. In other words, it is a graded $(\boldsymbol{k}, \boldsymbol{j})$–bimodule M and a degree 0 map $\delta^1\colon M \to A \otimes M \otimes B[1]$, again with an appropriate compatibility condition.

Note that when A or B is the trivial algebra $\{1\}$, we get a left or a right A or D structure over the other algebra.

There are notions of boundedness for bimodules similar to those for one-sided modules. There are various tensor products for the various compatible pairs of bimodules. We assume that one of the factors is bounded and briefly lay out the general description. For details, see [9, Section 2.3.2].

Let M and N be two structures such that M is a module or bimodule with a right type A action by an algebra $\mathcal{A}$, and N is a left type D structure over $\mathcal{A}$, or a type *DA* or type *DD* structure over $\mathcal{A}$ on the left and some algebra on the right, with M right-bounded or N left-bounded. As a chain complex, define

$$M \boxtimes N = \mathcal{F}(M) \boxtimes \mathcal{F}(N),$$

where $\mathcal{F}(M)$ forgets the left action on M, ie turns M into a right type A structure over $\mathcal{A}$, and $\mathcal{F}(N)$ forgets the right action on N, ie turns N into a left type D structure over $\mathcal{A}$. Endow $M \boxtimes N$ with the bimodule structure maps arising from the left action on M and the right action on N. Note that this also makes sense when M is a right type A structure, or N is a left type D structure.

In general (boundedness not required), one can think of N as a structure with a left $\mathcal{A}$ action, by considering $\mathcal{A} \boxtimes N$ (where $\mathcal{A}$ is viewed as a bimodule over itself), and take an $\mathcal{A}_\infty$ tensor product $M \mathbin{\widetilde{\otimes}} N := M \mathbin{\widetilde{\otimes}} (\mathcal{A} \boxtimes N)$. Whenever they are both defined, the two tensor products yield equivalent structures; see [9, Proposition 2.3.18].

For definitions of morphisms of type A, D, *AA*, *AD*, *DA* and *DD* structures, and for definitions of the respective types of homotopy equivalences, see [9, Section 2].

2.2 Tangles

In this paper we only consider tangles in 3–manifolds with boundary S^2 or $S^2 \amalg S^2$, or in closed 3–manifolds.

Definition 2.1 An *n–marked sphere* $\mathcal{S}$ is a sphere S^2 with n oriented points $t_1, \dots, t_n$ on its equator $S^1 \subset S^2$ numbered respecting the orientation of S^1.

Definition 2.2 A *marked $2n$–tangle* $\mathcal{T}$ in an oriented 3–manifold Y with $\partial Y \cong S^2$ is a properly embedded 1–manifold T with $(\partial Y, \partial T)$ identified with a $2n$–marked sphere $\mathcal{S}$.

A *marked (m,n)–tangle* $\mathcal{T}$ in an oriented 3–manifold Y with two boundary components $\partial^0 Y \cong S^2$ and $\partial^1 Y \cong S^2$ is a properly embedded 1–manifold T with $(\partial^0 Y, \partial^0 Y \cap \partial T)$ and $(\partial^1 Y, \partial^1 Y \cap \partial T)$ each identified with an m–marked sphere and an n–marked sphere. We denote ∂T along with the ordering information by $\partial \mathcal{T} = \partial^0 \mathcal{T} \amalg \partial^1 \mathcal{T}$.

We denote the number of connected components of a tangle $\mathcal{T}$ by $|\mathcal{T}|$. Note that we allow for a tangle to also have closed components.

Given a marked sphere $\mathcal{S} = (S^2, t_1, \dots, t_n)$, we denote $(-S^2, -t_1, \dots, -t_n)$ by $-\mathcal{S}$. If $\mathcal{T}_1$ and $\mathcal{T}_2$ are two marked tangles in 3–manifolds Y_1 and Y_2, where a component of $(\partial Y_1, \partial T_1)$ is identified with a marked sphere $\mathcal{S}$ and a component of $(\partial Y_2, \partial T_2)$ is identified with $-\mathcal{S}$, we can form the union $\mathcal{T}_1 \cup_{\mathcal{S}} \mathcal{T}_2$ by identifying Y_1 and Y_2 along these two boundary components.

For a pair $(Y, \mathcal{T})$, if a component $\partial^i Y$ of the boundary of Y is identified with $\mathcal{S} = (S^2, t_1, \dots, t_n)$, so that $\partial^i \mathcal{T}$ is the ordered set of points $(t_1, \dots, t_n)$, we use $-\partial^i \mathcal{T}$ to denote $(-t_1, \dots, -t_n)$. So we can glue two tangles $\mathcal{T}_1$ and $\mathcal{T}_2$ along boundary components $\partial^i \mathcal{T}_1$ and $\partial^j \mathcal{T}_2$ exactly when $\partial^i \mathcal{T}_1 = -\partial^j \mathcal{T}_2$.

In most of this paper, we only consider tangles in product spaces, where the identification of the boundary with a marked sphere is implied, and the ordering in $\partial \mathcal{T}$ encodes all the information.

Tangles in subsets of $S^3 = \mathbb{R}^3 \cup \{*\}$, for example in D^3, $I \times S^2$ or S^3 itself, can be given by their projection to $(-\infty, c] \times \mathbb{R}$ or $[d, \infty) \times \mathbb{R}$, $[c,d] \times \mathbb{R}$ or $\mathbb{R}^2$. We can always arrange a projection to be smooth and to have no triple points, and to have only transverse intersections.

Definition 2.3 A tangle $\mathcal{T}$ is *elementary* if it contains at most one double point or vertical tangency (a tangency of the form $\{f\} \times \mathbb{R}$).

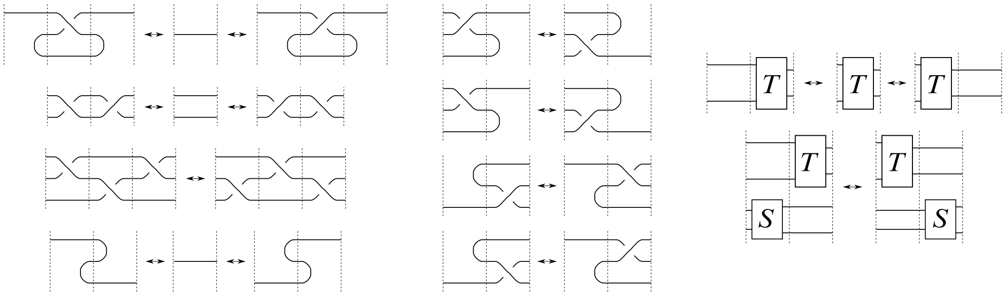


Figure 4: Relations of elementary tangles. In all diagrams there may be additional horizontal straight strands running above and/or below what is shown. Left column (top to bottom): Reidemeister I move, Reidemeister II move, Reidemeister III move, "zig-zag" move. Middle column: crossing-cap/cup slide moves. Right column (top to bottom): introducing straight strands to either side of a tangle or removing them, and sliding two vertically stacked tangles past each other.

Thus an elementary tangle can consist of straight strands (as in the first picture of Figure 6), can have one crossing (as in the second pictures of Figures 6 and 13), can be a cap (as in the third picture of Figure 6), or can be a cup (as in the last picture of Figure 6). The above examples are tangles in $[c,d]\times\mathbb{R}$. There is no elementary tangle projection in $\mathbb{R}^2$, an elementary tangle projection in $(-\infty,c]\times\mathbb{R}$ is a single cap, and an elementary tangle projection in $[d,\infty)\times\mathbb{R}$ is a single cup.

The following two propositions are well known to tangle theorists, and we do not rely on them in the paper, so we only include outlines of their proofs.

Proposition 2.4 *Any tangle projection is the concatenation of elementary tangles.*

Proof If necessary, one can isotope each tangency and/or double point slightly to the left or right, so that no two have the same horizontal coordinate. □

Further:

Proposition 2.5 *The concatenations of two sequences of elementary tangles represent isotopic tangles if and only if they are related by a finite sequence of the moves depicted in Figure 4.*

Proof The three Reidemeister moves are the standard moves that change the combinatorics of the diagram.

Using elementary Morse theory one can see that the other four types of moves are exactly the moves needed to move between two isotopic diagrams with the same combinatorics. Look at the height function obtained by projecting the tangle to the x–coordinate. The zig-zag move corresponds to canceling an index 0 critical point with an index 1 critical point or introducing a pair of such critical points. The crossing-cup slide moves are isotopies that do not change the Morse function, but slide a strand over or under a critical point. Introducing straight strands simply means taking one extra cut near one of the boundaries of a tangle. Sliding two vertically stacked tangles past each other corresponds to moving through a one-parameter family of Morse functions that changes the relative heights for the two disjoint tangles. □

In this paper, we define a (bi)module for each elementary tangle explicitly, and then define a (bi)module for any tangle by decomposing it into elementary pieces and taking the tensor product of the associated (bi)modules. We prove invariance of the decomposition using analytic techniques (the bordered Heegaard diagrams associated to isotopic tangles are related by Heegaard moves). We hope to also find a completely combinatorial proof, ie we wish to show directly that the moves from Figure 4 result in homotopy equivalent tensor products. As a first step, in Section 5.3 we show invariance under the Reidemeister II and III moves.

3 Generalized strand modules and algebras

The aim of this paper is to give a $0+1$ TQFT-like description of knot Floer homology. The description is based on a special kind of Heegaard diagram associated to a knot (or a link) disjoint union an unknot.

Given a tangle $\mathcal{T}$, by cutting it into elementary tangles like the ones in Figure 6, we can put it on a Heegaard diagram like the one depicted in Figure 2, where the genus of the diagram is the number of elementary pieces. The parts of the Heegaard diagram corresponding to the elementary pieces are depicted in Figures 18 and 24. Note that the Heegaard diagram is obtained by gluing together a once punctured torus, some twice punctured tori, and another once punctured torus. In the sequel, we will associate an algebra to each cut of the tangle, a left type A module and a right type D structure to the once punctured tori, and a type DA bimodule to each of the twice punctured tori.

In this section, we will describe the algebras, modules, and bimodules from a purely combinatorial viewpoint, with no mention of Heegaard diagrams. In Section 4, we relate these structures to bordered diagrams.

In the sequel, we define generalized strand algebras and modules whose structure depends on the extra information, encoded in a structure we will refer to as *shadow*. We define the *minus* version of the theory, and the *tilde* version can be obtained by setting all U_O to 0. In this section we describe the modules and algebras via strand diagrams, but some of the notions feel more natural in the bordered grid diagram reformulation (see Section 4). The reader who is familiar with the strand algebras of [8] should be able to understand the main idea of the definitions just by looking at the examples and the figures.

Although in this paper the main theorem is only proved for the tilde version, we have strong evidence that it holds for the minus version as well. This is why we develop both versions, but at first reading one can ignore the U–powers (ie set $U_O = 0$) and work in the tilde version.

3.1 Type *AA* structures: shadows

The objects underlying all structures are shadows:

Definition 3.1 For n, $m \in \mathbb{N}$, fix sets of integers $\boldsymbol{a} = \{1, \dots, n\}$ and $\boldsymbol{b} = \{1, \dots, m\}$, and sets of half-integers $\boldsymbol{a}_{1/2} = \left\{1\frac{1}{2}, \dots, n - \frac{1}{2}\right\}$ and $\boldsymbol{b}_{1/2} = \left\{1\frac{1}{2}, \dots, m - \frac{1}{2}\right\}$. Let $(S_{\mathbb{X}}, T_{\mathbb{X}}, \xi)$ and $(S_{\mathbb{O}}, T_{\mathbb{O}}, \omega)$ be triples such that $S_{\mathbb{X}}$, $T_{\mathbb{O}} \subset \boldsymbol{a}_{1/2}$ and $T_{\mathbb{X}}$, $S_{\mathbb{O}} \subset \boldsymbol{b}_{1/2}$, $|T_{\mathbb{X}}| = |S_{\mathbb{X}}|$ and $|T_{\mathbb{O}}| = |S_{\mathbb{O}}|$, and $\xi\colon S_{\mathbb{X}} \to T_{\mathbb{X}}$ and $\omega\colon S_{\mathbb{O}} \to T_{\mathbb{O}}$ are two bijections. The quadruple $\mathcal{P} = (m, n, \xi, \omega)$ is called a *shadow*.

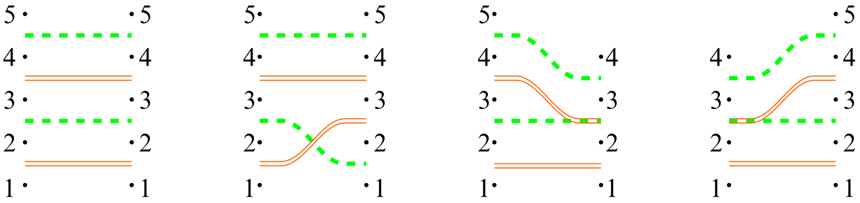


Figure 5: Examples of shadows. On each diagram $\boldsymbol{b}$ and $\boldsymbol{b}_{1/2}$ are on the left-hand side, while $\boldsymbol{a}$ and $\boldsymbol{a}_{1/2}$ are on the right-hand side. Double (orange) lines connect $\{1\} \times \{s_X\}$ with $\{0\} \times \{\xi s_X\}$ (for $s_X \in S_{\mathbb{X}}$) and dashed (green) lines connect $\{0\} \times \{s_O\}$ with $\{1\} \times \{\omega s_O\}$ (for $s_O \in S_{\mathbb{O}}$).

Note that $T_{\mathbb{X}}$, $S_{\mathbb{X}}$ and $T_{\mathbb{O}}$, $S_{\mathbb{O}}$ are suppressed from the notation. See Figure 5 for diagrams of shadows associated to elementary tangles (see Section 3.1.1). The information in the subsets $S_{\mathbb{X}}$, $T_{\mathbb{O}} \subset \boldsymbol{a}_{1/2}$ and $T_{\mathbb{X}}$, $S_{\mathbb{O}} \subset \boldsymbol{b}_{1/2}$ can be encoded as follows:

Definition 3.2 The *boundaries* of a shadow $\mathcal{P}$ are defined as

$$\boldsymbol{\epsilon}^0 = \boldsymbol{\epsilon}^0(\mathcal{P}) = (\epsilon^0_1, \dots, \epsilon^0_{m-1}) \in (2^{\{\pm 1\}})^{m-1},$$
$$\boldsymbol{\epsilon}^1 = \boldsymbol{\epsilon}^1(\mathcal{P}) = (\epsilon^1_1, \dots, \epsilon^1_{n-1}) \in (2^{\{\pm 1\}})^{n-1},$$

as follows. For a point $j+\frac{1}{2}\in \boldsymbol{b}_{1/2}$, the subset $\epsilon^0_j \subset \{\pm 1\}$ contains -1 if and only if $j+\frac{1}{2}\in S_{\mathbb{O}}$, and $+1\in\epsilon^0_j$ if and only if $j\in T_{\mathbb{X}}$. Similarly, for $j+\frac{1}{2}\in \boldsymbol{a}_{1/2}$ define the subset $\epsilon^1_j\subset\{\pm 1\}$ by $+1\in\epsilon^1_j$ if and only if $j+\frac{1}{2}\in T_{\mathbb{O}}$, and $-1\in\epsilon^0_j$ if and only if $j+\frac{1}{2}\in S_{\mathbb{X}}$.

By reversing the above process, we can recover $S_{\mathbb{X}}$, $T_{\mathbb{O}}\subset \boldsymbol{a}_{1/2}$ and $T_{\mathbb{X}}$, $S_{\mathbb{O}}\subset \boldsymbol{b}_{1/2}$ from $\boldsymbol{\epsilon}^0$ and $\boldsymbol{\epsilon}^1$ by setting $S_{\mathbb{X}}=\{j+\frac{1}{2}\in\boldsymbol{a}_{1/2}:-1\in\epsilon^1_j\}$, $T_{\mathbb{X}}=\{j+\frac{1}{2}\in\boldsymbol{b}_{1/2}:+1\in\epsilon^0_j\}$, $S_{\mathbb{O}}=\{j+\frac{1}{2}\in\boldsymbol{b}_{1/2}:-1\in\epsilon^0_j\}$ and $T_{\mathbb{O}}=\{j+\frac{1}{2}\in\boldsymbol{a}_{1/2}:+1\in\epsilon^0_j\}$. The following shadows will play an important role in our discussion.

Example 3.3 (straight lines) For $\boldsymbol{\epsilon}^0=(\epsilon^0_j)^k_{j=1}\in\{\pm 1\}^k$ let $\boldsymbol{\epsilon}^1=-\boldsymbol{\epsilon}^0$ and define $S_{\mathbb{X}}$, $T_{\mathbb{X}}$, $S_{\mathbb{O}}$ and $T_{\mathbb{O}}$ as in the previous paragraph. Consider the shadow ${}_{\boldsymbol{\epsilon}^0}\mathcal{E}_{\boldsymbol{\epsilon}^1}=(k+1,k+1,\mathrm{id}_{S_{\mathbb{X}}},\mathrm{id}_{S_{\mathbb{O}}})$. See the first picture of Figure 5 for $k=4$ and $\boldsymbol{\epsilon}^0=(+1,-1,+1,-1)$.

The next three examples correspond to elementary tangles.

Example 3.4 (crossing) For $\boldsymbol{\epsilon}^0=(\epsilon^0_j)^k_{j=1}\in\{\pm 1\}^k$ and $1<i\le k$, define $\boldsymbol{\epsilon}^1=(\epsilon^1_j)^k_{j=1}$, where

$$\epsilon^1_j=\begin{cases}-\epsilon^0_{i-1} & \text{if } j=i,\\ -\epsilon^0_i & \text{if } j=i-1,\\ -\epsilon^0_j & \text{otherwise.}\end{cases}$$

Define $S_{\mathbb{X}}$, $T_{\mathbb{X}}$, $S_{\mathbb{O}}$ and $T_{\mathbb{O}}$ as before, and for $s_O\in S_{\mathbb{O}}$ define

$$\omega s_O=\begin{cases}i+\frac{1}{2} & \text{if } s_O=i-\frac{1}{2},\\ i-\frac{1}{2} & \text{if } s_O=i+\frac{1}{2},\\ s_O & \text{otherwise.}\end{cases}$$

For $s_X\in S_{\mathbb{X}}$ define

$$\xi s_X=\begin{cases}i+\frac{1}{2} & \text{if } s_X=i-\frac{1}{2},\\ i-\frac{1}{2} & \text{if } s_X=i+\frac{1}{2},\\ s_X & \text{otherwise.}\end{cases}$$

Consider the shadow ${}_{\boldsymbol{\epsilon}^0}\mathcal{X}_{\boldsymbol{\epsilon}^1}(i)=(k+1,k+1,\xi,\omega)$. See the second picture of Figure 5 for $k=4$, $i=2$ and $\boldsymbol{\epsilon}=(+1,-1,+1,-1)$.

Example 3.5 (cap) For $\boldsymbol{\epsilon}^0=(\epsilon^0_i)^k_{i=1}\in\{\pm 1\}^k$ and $0\le i\le k$ with $\epsilon^0_{i-1}\epsilon^0_i=-1$, define $\boldsymbol{\epsilon}^1=(\epsilon^1_i)^{k-1}_{i=1}\in\{\pm 1,\{\pm 1\}\}^{k-1}$ by

$$\epsilon^1_j=\begin{cases}-\epsilon^0_j & \text{if } j<i,\\ -\epsilon^0_{j-1} & \text{if } j>i,\\ \{\pm 1\} & \text{if } j=i.\end{cases}$$

Define $S_{\mathbb{X}}$, $T_{\mathbb{X}}$, $S_{\mathbb{O}}$ and $T_{\mathbb{O}}$ as before, and for $s_O \in S_{\mathbb{O}}$ define

$$\omega s_O = \begin{cases} s_O & \text{if } s_O < i, \\ s_O - 1 & \text{if } s_O > i. \end{cases}$$

For $t_X \in T_{\mathbb{X}}$ define

$$\xi^{-1} t_X = \begin{cases} t_X & \text{if } t_X < i, \\ t_X - 1 & \text{if } t_X > i, \end{cases}$$

and consider the shadow ${}_{\boldsymbol{\epsilon}^0}\mathcal{D}_{\boldsymbol{\epsilon}^1}(i) = (k+1, k, \xi, \omega)$. See the third picture of Figure 5 for $k = 4$, $i = 3$ and $\boldsymbol{\epsilon}^0 = (+1, -1, +1, -1)$.

Example 3.6 (cup) This is the mirror of a cap. For $\boldsymbol{\epsilon}^1 = (\epsilon_i^1)_{i=1}^k \in \{\pm 1\}^k$ and $0 \le i \le k$ with $\epsilon_{i-1}^1 \epsilon_i^1 = -1$, define $\boldsymbol{\epsilon}^0 = (\epsilon_i^0)_{i=1}^{k-1} \in \{\pm 1, \{\pm 1\}\}^{k-1}$ by

$$\epsilon_j^0 = \begin{cases} -\epsilon_j^1 & \text{if } j < i, \\ -\epsilon_{j-1}^1 & \text{if } j > i, \\ \{\pm 1\} & \text{if } j = i. \end{cases}$$

Define $S_{\mathbb{X}}$, $T_{\mathbb{X}}$, $S_{\mathbb{O}}$ and $T_{\mathbb{O}}$ as before, and for $t_O \in T_{\mathbb{O}}$ define

$$\omega^{-1} t_O = \begin{cases} t_O & \text{if } t_O < i, \\ t_O - 1 & \text{if } t_O > i. \end{cases}$$

For $s_X \in S_{\mathbb{X}}$ define

$$\xi s_X = \begin{cases} s_X & \text{if } s_X < i, \\ s_X - 1 & \text{if } s_X > i, \end{cases}$$

and consider the shadow ${}_{\boldsymbol{\epsilon}^0}\mathcal{C}_{\boldsymbol{\epsilon}^1}(i) = (k, k+1, \xi, \omega)$. See the fourth picture of Figure 5 for $k = 4$, $i = 3$ and $\boldsymbol{\epsilon}^1 = (-1, +1, -1, +1)$.

Example 3.7 Given any shadow $\mathcal{P}$, one can introduce a gap at either its left- or right-hand side. We discuss the construction for the left-hand side. Given $i \in \boldsymbol{b}$, let $m' = m + 1$ and $n' = n$, and define $\mathcal{L}_i(\mathcal{P}) = (n', m', \xi', \omega')$ by $(\boldsymbol{\epsilon}^1)' := \boldsymbol{\epsilon}^1$ and $(\boldsymbol{\epsilon}^0)' = ((\epsilon_j^o)')_{j=1}^{m'}$, where

$$(\epsilon_j^0)' = \begin{cases} \epsilon_j^0 & \text{if } j < i, \\ \varnothing & \text{if } j = i, \\ \epsilon_{j-1}^0 & \text{if } j > i. \end{cases}$$

Define $S'_{\mathbb{X}}$, $T'_{\mathbb{X}}$, $S'_{\mathbb{O}}$ and $T'_{\mathbb{O}}$ as before, and for $s_O \in S_{\mathbb{O}}$ define

$$\omega' s_O = \begin{cases} \omega s_O & \text{if } s_O < i, \\ \omega s_O - 1 & \text{if } s_O > i, \end{cases}$$

For $t_X \in T_{\mathbb{X}}$ define

$$(\xi')^{-1}t_X = \begin{cases} \xi^{-1}t_X & \text{if } t_X < i, \\ \xi^{-1}t_X - 1 & \text{if } t_X > i. \end{cases}$$

Similarly, for $i \in \boldsymbol{a}$ we can introduce a gap on the right-hand side to obtain the shadow $\mathcal{R}_i(\mathcal{P})$.

3.1.1 Diagrams and tangles associated to shadows

Shadows can be best understood through their diagrams:

Definition 3.8 A *diagram* of a shadow $\mathcal{P}$ is a quadruple

$$D(\mathcal{P}) = (\{0\} \times \boldsymbol{b}_{1/2}, \{1\} \times \boldsymbol{a}_{1/2}, x, o) \subset I \times \mathbb{R},$$

where x is a set of properly embedded arcs connecting $(1, s_X)$ to $(0, \xi s_X)$ (for $s_X \in S_{\mathbb{X}}$) and o is a set of properly embedded arcs connecting $(0, s_O)$ to $(1, \omega s_O)$ (for $s_O \in S_{\mathbb{O}}$) such that there are no triple points, and the number of intersection points of all arcs is minimal within the isotopy class fixing the boundaries.

Any two diagrams of $\mathcal{P}$ are related by a sequence of Reidemeister III moves (see the first picture of Figure 8) and isotopies relative to the boundaries. We do not distinguish different diagrams of the same shadow and will refer to both the isotopy class (rel boundary) or a representative of the isotopy class as the diagram of $\mathcal{P}$.

Definition 3.9 To a shadow $\mathcal{P}$ we can associate a *tangle* $\mathcal{T}(\mathcal{P})$ as follows. Start from $D(\mathcal{P}) \subset I \times \mathbb{R}$. If $j + \frac{1}{2} \in S_{\mathbb{X}} \cap T_{\mathbb{O}}$ (that is $\epsilon_j^1 = \{\pm 1\}$) then there is one arc starting and one arc ending at $\bigl(1, j + \frac{1}{2}\bigr)$. Smooth the corner at $\bigl(1, j + \frac{1}{2}\bigr)$ by pushing the union of the two arcs slightly in the interior of $I \times \mathbb{R}$, as shown in Figure 6. Do the same at $\bigl(0, j + \frac{1}{2}\bigr)$ for $j + \frac{1}{2} \in T_{\mathbb{X}} \cap S_{\mathbb{O}}$. This process results in a smooth properly immersed set of arcs. Remove the self-intersection of the union of the above set of arcs by slightly lifting up the interior of arcs with bigger slope. After this process we obtain a tangle projection in $I \times \mathbb{R}$ or in $(0,1] \times \mathbb{R} \cong (-\infty, 1] \times \mathbb{R}$, $[0,1) \times \mathbb{R} \cong [0, \infty) \times \mathbb{R}$ or $(0,1) \times \mathbb{R} \cong \mathbb{R}^2$ if the resulting projection does not intersect $\{0\} \times \mathbb{R}$ and/or $\{1\} \times \mathbb{R}$. Then the tangle $\mathcal{T}(\mathcal{P}) = \mathcal{T}$ lives in $I \times S^2$, D^3 or in S^3 with boundaries $\partial^0\mathcal{T} = \{0\} \times \bigl\{j + \frac{1}{2} : \epsilon_j^0 = +1\bigr\} - \{0\} \times \bigl\{j + \frac{1}{2} : \epsilon_j^0 = -1\bigr\}$ and $\partial^1\mathcal{T} = \{1\} \times \bigl\{j + \frac{1}{2} : \epsilon_j^1 = +1\bigr\} - \{0\} \times \bigl\{j + \frac{1}{2} : \epsilon_j^1 = -1\bigr\}$.

The elementary tangles corresponding to Examples 3.3–3.6 are depicted in Figure 6.

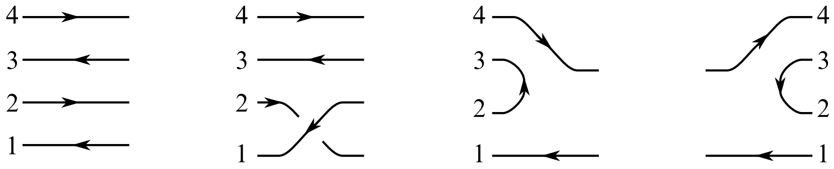


Figure 6: Elementary tangles corresponding to the shadows of Figure 5

3.1.2 Generators Now we start describing the type AA structure associated to a shadow $\mathcal{P}$. The underlying set is generated by the following elements.

Definition 3.10 For a shadow $\mathcal{P}$ let $\mathfrak{S}(\mathcal{P})$ denote the set of triples $f=(S,T,\phi)$, where $S\subset \boldsymbol{b}$, $T\subset \boldsymbol{a}$, with $|S|=|T|$ and $\phi\colon S\to T$ a bijection.

Note that we can also think of generators as partial matchings of the complete bipartite graph on the vertex sets $(\boldsymbol{a},\boldsymbol{b})$. For any generator $f=(S,T,\phi)$ we can draw a set of arcs on the diagram of $\mathcal{P}$ by connecting each $(0,s)$ to $(1,\phi s)$ with a monotone properly embed arc. See Figure 7 for diagrams of the generators. Again, in these diagrams we do not have triple points, the number of intersection points of all strands is minimal, and we do not distinguish different diagrams of the same generator. Any two diagrams with minimal intersections are related by a sequence of Reidemeister III moves (See the first picture of Figure 8). Note that the generators naturally split into subsets $\mathfrak{S}_i(\mathcal{P})=\{(S,T,\phi):|S|=|T|=i\}$. Then $\mathfrak{S}(\mathcal{P})=\bigcup_{i=1}^{\min\{n,m\}}\mathfrak{S}_i(\mathcal{P})$.

Fix a variable U_O for each pair $O=(s_O,\omega s_O)\in S_{\mathbb{O}}\times T_{\mathbb{O}}$.

Definition 3.11 Let $C^-(\mathcal{P})$ be the module generated by $\mathfrak{S}(\mathcal{P})$ over $\boldsymbol{k}=\mathbb{F}_2[U_O]_{s_O\in S_{\mathbb{O}}}$.

3.1.3 Inner differential Note that so far $C^-(\mathcal{P})$ depends only on m and n, but not on the particular structure of $(S_{\mathbb{X}},T_{\mathbb{X}},\xi)$ and $(S_{\mathbb{O}},T_{\mathbb{O}},\omega)$. The first dependence can be seen in the differential, which is described by resolutions of intersections of the

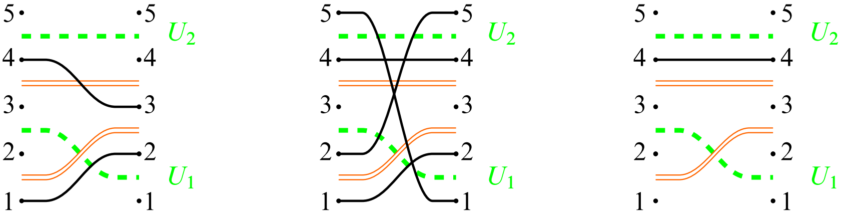


Figure 7: Diagrams of some generators $(S,T,\phi)\in\mathfrak{S}(\mathcal{P})$. Solid black lines connect s with ϕs.

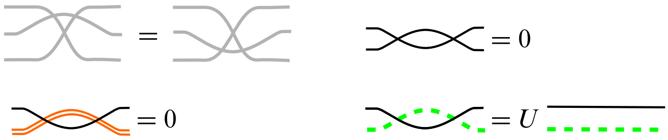


Figure 8: Relations of diagrams. In the top-left relation the strands can correspond to ϕ, ξ or ω.

diagram, subject to some relations. (See Figure 8.) The intersections of the diagram of a generator (S,T,ϕ) correspond to inversions of the partial permutation ϕ.

Let $\phi\colon S \to T$ be a bijection between subsets S and T of two ordered sets $\boldsymbol{b}$ and $\boldsymbol{a}$. Define

$$\mathrm{Inv}(\phi) = \{(s_1,s_2) \in S \times S : s_1 < s_2 \text{ and } \phi s_1 > \phi s_2\}.$$

Given ordered sets $\boldsymbol{b} \cup \boldsymbol{b}_{1/2}$ and $\boldsymbol{a} \cup \boldsymbol{a}_{1/2}$, and bijections $\phi\colon S \to T$ and $\omega\colon S_{\mathbb{O}} \to T_{\mathbb{O}}$ for $S \subset \boldsymbol{b}$, $T \subset \boldsymbol{a}$, $S_{\mathbb{O}} \subset \boldsymbol{b}_{1/2}$ and $T_{\mathbb{O}} \subset \boldsymbol{a}_{1/2}$, define

$$\mathrm{Inv}(\phi,\omega) = \{(s,s_O) \in S \times S_{\mathbb{O}} : s < s_O \text{ and } \phi s > \omega s_O, \text{ or } s > s_O \text{ and } \phi s < \omega s_O\}.$$

Define the set $\mathrm{Inv}(\phi,\xi^{-1})$ and for $s_O \in S_{\mathbb{O}}$ the set $\mathrm{Inv}(\phi,\omega|_{s_O})$ similarly. Denote the sizes of these sets by $\mathrm{inv}(\phi)$, $\mathrm{inv}(\phi,\omega)$, $\mathrm{inv}(\phi,\xi^{-1})$ and $\mathrm{inv}(\phi,\omega|_{s_O})$, respectively.

The differential of a generator (S,T,ϕ) can be given by resolving intersections. For $\tau = (s_1,s_2) \in \mathrm{Inv}(\phi)$ define the new generator (S,T,ϕ^τ), where $\phi^\tau = \phi \circ \tau$ is the *resolution of* ϕ *at* τ (for simplicity, here and throughout the paper τ denotes both the pair (s_1,s_2) and the 2–cycle permutation $(s_1 s_2)$). A resolution of $\tau = (s_1,s_2) \in \mathrm{Inv}(\phi)$ is *allowed* if $\mathrm{inv}(\phi^\tau) = \mathrm{inv}(\phi) - 1$ (Compare with the top-right picture of Figure 8.) and $\mathrm{inv}(\phi,\xi^{-1}) = \mathrm{inv}(\phi^\tau,\xi^{-1})$ (Compare with the bottom-left picture of Figure 8.). The set of inversions with allowed resolutions is denoted by $\mathrm{Inv}_0(\phi) \subset \mathrm{Inv}(\phi)$.

Given a pair $O = (s_O, \omega s_O)$ and a 2–cycle permutation τ such that $\phi \circ \tau$ is defined, define

$$n_O(\tau;\phi) = \tfrac{1}{2}(\mathrm{inv}(\phi,\omega|_{s_O}) - \mathrm{inv}(\phi^\tau,\omega|_{s_O})).$$

When ϕ is clear from the context we will omit it from the notation and will write $n_O(\tau)$ or $n_O(s_1,s_2)$ for $n_O(\tau;\phi)$. Note that $n_O(\tau)$ is always an integer. The differential is defined on generators by

$$\partial(S,T,\phi) = \sum_{\tau \in \mathrm{Inv}_0(\phi)} \Bigl(\prod_{s_O \in S_{\mathbb{O}}} U_O^{n_O(\tau)}\Bigr)(S,T,\phi^\tau).$$

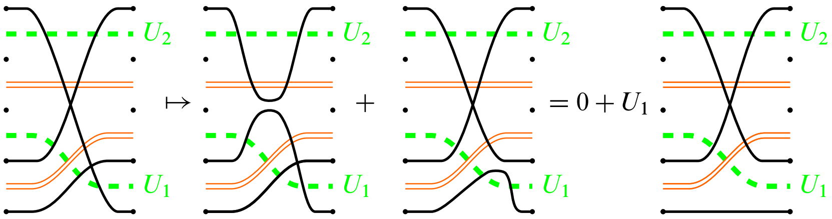


Figure 9: Example of the differential. Note that the second and the third diagrams do not have minimal intersections, thus they do not represent generators. We get the differential by removing the extra intersections using the relations of Figure 8.

Compare this equation with the bottom-right relation of Figure 8. Also see Figure 9 for an example. Extend ∂^- linearly to the whole $C^-(\mathcal{P})$.

Proposition 3.12 $(C^-(\mathcal{P}), \partial)$ *is a chain complex.*

Proof The differential first resolves intersection points and then applies the relations of Figure 8 to minimize the number of intersection points. When we apply the differential twice, then we can equivalently first resolve two intersection points and then apply the relations Figure 8 all at once. This proves that any term of

$$\partial^2(S,T,\phi) = \sum_{\tau_1 \in \mathrm{Inv}_O(\phi)} \sum_{\tau_2 \in \mathrm{Inv}_O(\phi_1^{\tau})} \prod_{s_O \in S_{\mathbb{O}}} U_O^{n_O(\tau_1;\phi)+n_O(\tau_2;\phi^{\tau_1})}(S,T,(\phi^{\tau_1})^{\tau_2})$$

appears twice with exactly the same coefficient and thus cancels. □

3.1.4 Composition of shadows: type A maps Let $\mathcal{P}_1 = (m_1, n_1, \xi_1, \omega_1)$ and $\mathcal{P}_2 = (m_2, n_2, \xi_2, \omega_2)$ be two shadows. If $n_1 = m_2$, $S_{\mathbb{X}_1} = T_{\mathbb{X}_2}$ and $T_{\mathbb{O}_1} = S_{\mathbb{O}_2}$, then we can define the *concatenation* of the shadows as $\mathcal{P}_1 * \mathcal{P}_2 = (m, n, \xi, \omega)$, where $m = m_1$, $n = n_2$, $(S_{\mathbb{X}}, T_{\mathbb{X}}, \xi) = (S_{\mathbb{X}_1}, T_{\mathbb{X}_2}, \xi_1 \circ \xi_2)$ and $(T_{\mathbb{O}}, S_{\mathbb{O}}, \omega) = (S_{\mathbb{O}_2}, T_{\mathbb{O}_1}, \omega_2 \circ \omega_1)$.

Definition 3.13 We say that $\mathcal{P}_1$ and $\mathcal{P}_2$ as above are *composable* if the numbers of intersection points add up, ie $\mathrm{inv}(\xi) = \mathrm{inv}(\xi_1) + \mathrm{inv}(\xi_2)$, $\mathrm{inv}(\omega) = \mathrm{inv}(\omega_1) + \mathrm{inv}(\omega_2)$ and $\mathrm{inv}(\omega, \xi^{-1}) = \mathrm{inv}(\omega_1, \xi_1^{-1}) + \mathrm{inv}(\omega_2, \xi_2^{-1})$. In this case $\mathcal{P}_1$ and $\mathcal{P}_2$ have a well-defined *composition* $\mathcal{P}_1 \circ \mathcal{P}_2 = \mathcal{P}_1 * \mathcal{P}_2$.

Note that on the diagram composable means that after the concatenation the resulting shadow still has minimal intersection.

Example 3.14 In Figure 5 all shadows that can be concatenated are immediately composable. However, the first two pictures of Figure 10 can be concatenated, but they are not composable.

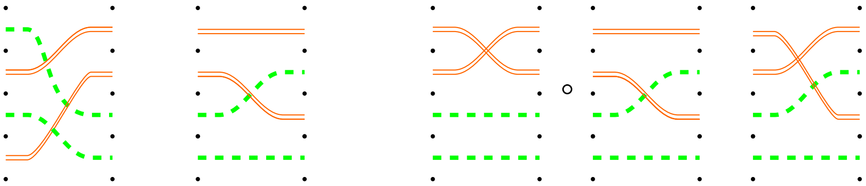


Figure 10: Two shadows that are not composable (left) and two composable shadows and their composition (right)

If $\mathcal{P}_1$ and $\mathcal{P}_2$ are composable, then there is a composition map

$$C^-(\mathcal{P}_1)\otimes C^-(\mathcal{P}_2)\to C^-(\mathcal{P}_1\circ\mathcal{P}_2),$$

denoted by $\cdot$ and defined as follows: Let $f_1=(S_1,T_1,\phi_1)$ and $f_2=(S_2,T_2,\phi_2)$ be generators of $C^-(\mathcal{P}_1)$ and $C^-(\mathcal{P}_2)$, respectively. If $T_1=S_2$, then the concatenation $(S,T,\phi)=(S_1,T_2,\phi_2\circ\phi_1)$ is well-defined. If $\operatorname{inv}(\phi)=\operatorname{inv}(\phi_1)+\operatorname{inv}(\phi_2)$ and $\operatorname{inv}(\phi,\xi^{-1})=\operatorname{inv}(\phi_1,\xi_1^{-1})+\operatorname{inv}(\phi_2,\xi_2^{-1})$, then $f_1\cdot f_2$ is defined by

$$(S_1,T_1,\phi_1)\cdot(S_2,T_2,\phi_2)=\prod_{s_O\in T_{\mathbb{O}}}U_O^{\frac{1}{2}(\operatorname{inv}(\phi_1,\omega_1|_{s_O})+\operatorname{inv}(\phi_2,\omega_2|_{\omega_1 s_O})-\operatorname{inv}(\phi,\omega|_{s_O}))}(S,T,\phi)$$

In all other cases $f_1\cdot f_2$ is defined to be 0. See Figures 11 and 22 for examples.

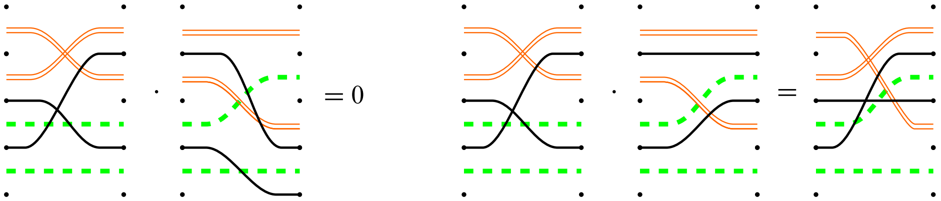


Figure 11: Composition of generators. The first composition is 0 by the third relation of Figure 8.

Note that this composition is consistent with the differential and associative:

Proposition 3.15 *Let $\mathcal{P}_1$ be composable with $\mathcal{P}_2$. Then the following square commutes:*

$$\begin{tikzcd}
C^-(\mathcal{P}_1)\otimes C^-(\mathcal{P}_2) \arrow[r, "\cdot"] \arrow[d, "\partial\otimes\mathrm{id}+\mathrm{id}\otimes\partial"] & C^-(\mathcal{P}_1\circ\mathcal{P}_2) \arrow[d, "\partial"] \\
C^-(\mathcal{P}_1)\otimes C^-(\mathcal{P}_2) \arrow[r, "\cdot"] & C^-(\mathcal{P}_1\circ\mathcal{P}_2)
\end{tikzcd}$$

If in addition $\mathcal{P}_2$ is composable with the shadow $\mathcal{P}_3$, then $\mathcal{P}_1 \circ \mathcal{P}_2$ is composable with $\mathcal{P}_3$, $\mathcal{P}_1$ is composable with $\mathcal{P}_2 \circ \mathcal{P}_3$ and the following square commutes:

$$\begin{tikzcd}
C^-(\mathcal{P}_1) \otimes C^-(\mathcal{P}_2) \otimes C^-(\mathcal{P}_3) \arrow[r, "\mathrm{id}\otimes\cdot"] \arrow[d, "\cdot\otimes\mathrm{id}"] & C^-(\mathcal{P}_1) \otimes C^-(P_2 \circ \mathcal{P}_3) \arrow[d, "\cdot"] \\
C^-(\mathcal{P}_1 \circ \mathcal{P}_2) \otimes C^-(\mathcal{P}_3) \arrow[r, "\cdot"] & C^-(\mathcal{P}_1 \circ \mathcal{P}_2 \circ \mathcal{P}_3)
\end{tikzcd}$$

Proof This statement again follows from the facts that one can first do all the operations (resolving intersections and concatenating generators) and then reduce the intersection points by the relations of Figure 8 and that both equations are obvious without the relations. □

Definition 3.16 For a shadow $\mathcal{P}$, define the shadows $\mathcal{E}_R = \mathcal{E}_R(\mathcal{P})$ and $\mathcal{E}_L = \mathcal{E}_L(\mathcal{P})$ by the quadruples $(m, m, \mathrm{id}_{T_\mathbb{X}}, \mathrm{id}_{S_\mathbb{O}})$ and $(n, n, \mathrm{id}_{S_\mathbb{X}}, \mathrm{id}_{T_\mathbb{O}})$, respectively. In general, let $\mathcal{E}$ be the shadow given by the quadruple $(n, n, \mathrm{id}_{S_\mathbb{X}}, \mathrm{id}_{S_\mathbb{O}})$, where $S_\mathbb{X} \subset \boldsymbol{b}_{1/2}$ and $S_\mathbb{O} \subset \boldsymbol{a}_{1/2}$ are any subsets. Then $\mathcal{E} \circ \mathcal{E} = \mathcal{E}$, so we call $\mathcal{E}$ an *idempotent shadow*.

Note that idempotent shadows are exactly shadows corresponding to straight lines (Example 3.3). By Proposition 3.15, the induced multiplication

$$C^-(\mathcal{E}) \times C^-(\mathcal{E}) \xrightarrow{\cdot} C^-(\mathcal{E})$$

upgrades $C^-(\mathcal{E})$ to a differential algebra:

Definition 3.17 For an idempotent shadow $\mathcal{E}$, let $\mathcal{A}(\mathcal{E})$ be the differential algebra $(C^-(\mathcal{E}), \cdot, \partial)$.

In Section 3.4 we will define a grading that turns $\mathcal{A}(\mathcal{E})$ into a differential graded algebra. Again by Proposition 3.15, $(C^-(\mathcal{P}), \partial, \cdot, \cdot)$ is a left–right $\mathcal{A}(\mathcal{E}_L)$–$\mathcal{A}(\mathcal{E}_R)$ differential module, which we can turn into a type *AA* structure:

Definition 3.18 With the above notation, let $\mathit{CATA}^-(\mathcal{P})$ be the left–right *AA* structure $(C^-(\mathcal{P}), \{m_{i,1,j}\})$ over $\mathcal{A}(\mathcal{E}_L)$ and $\mathcal{A}(\mathcal{E}_R)$, where

$$m_{i,1,j}\colon \mathcal{A}(\mathcal{E}_L)^{\otimes i} \otimes C^-(\mathcal{P}) \otimes \mathcal{A}(\mathcal{E}_R)^{\otimes j} \to C^-(\mathcal{P})$$

with $m_{i,1,j} = 0$ for $i > 1$ or $j > 1$, and nonzero maps given by

$$m_{0,1,0}(f) = \partial f, \quad m_{1,1,0}(a_L \otimes f) = a_L \cdot f, \quad m_{0,1,1}(f \otimes a_R) = f \cdot a_R.$$

The gradings of $\mathit{CATA}^-(\mathcal{P})$ will only be defined in Section 3.4. Since $\mathit{CATA}^-(\mathcal{P})$ comes from a two-sided differential module, we have:

Proposition 3.19 *For any shadow $\mathcal{P}$ the structure maps of CATA$^-(\mathcal{P})$ satisfy the type AA structure identities.*

The idempotents of $\mathcal{A}(\mathcal{E})$ are given by (S,S,id_S), where $S\subset \boldsymbol{b}$. Let $\mathcal{I}(\mathcal{A}(\mathcal{E}))$ denote the set of idempotent elements of $\mathcal{A}(\mathcal{E})$. For a generator $f=(S,T,\phi)$, define

$$\iota_L(f)=(S,S,\mathrm{id}_S)\in\mathcal{I}(\mathcal{A}(\mathcal{E}_L)),\quad \iota_R(f)=(T,T,\mathrm{id}_T)\in\mathcal{I}(\mathcal{A}(\mathcal{E}_R)).$$

These idempotents are defined so that we have $\iota_L(f)\cdot f\cdot\iota_R(f)=f$.

3.2 Type *DD* structures: mirror-shadows

To define type D structures we need to work with cochain complexes associated to "mirrors" of shadows. For a shadow $\mathcal{P}=(m,n,\xi,\omega)$, define its *mirror* $\mathcal{P}^*$ to be the same quadruple (m,n,ξ,ω). In the sequel we will always associate "dual-structures" to $\mathcal{P}^*$, which is why we make the distinction in the notation. To a mirror-shadow $\mathcal{P}^*$ we associate the cochain complex $(C^-(\mathcal{P}^*),\partial^*)=(C^-(\mathcal{P}),\partial)^*$. Thus the elements of $C^-(\mathcal{P}^*)$ are of the form $(S,T,\phi)^*$ and the codifferential ∂^* introduces intersection points

$$\partial^*(S,T,\phi)^*=\sum_{\tau\in\mathrm{Inv}_0^*(\phi)}\ \prod_{s_O\in T_{\mathbb{O}}} U_O^{-n_O(\tau;\phi)}(S,T,\phi^\tau)^*,$$

where the elements of $\mathrm{Inv}_0^*(\phi)$ are elements of $\mathrm{Inv}(\phi)^c$ such that $\mathrm{inv}(\phi^\tau)=\mathrm{inv}(\phi)+1$ and $\mathrm{inv}(\phi,\xi^{-1})=\mathrm{inv}(\phi^\tau,\xi^{-1})$.

Let $\mathcal{A}(\mathcal{E}^L)$ and $\mathcal{A}(\mathcal{E}^R)$ be the algebras corresponding to the idempotent shadows $\mathcal{E}^L=\mathcal{E}^L(\mathcal{P}^*)=(n,n,\mathrm{id}_{S^c_{\mathbb{X}}},\mathrm{id}_{T^c_{\mathbb{O}}})$ and $\mathcal{E}^R=\mathcal{E}^R(\mathcal{P}^*)=(m,m,\mathrm{id}_{T^c_{\mathbb{X}}},\mathrm{id}_{S^c_{\mathbb{O}}})$, where $\cdot^c$ denotes the complement of subsets in the appropriate set they are contained in (see Definition 3.1). Then for $f^*=(S,T,\phi)^*$ let

$$\iota^L(f^*)=(T^c,T^c,\mathrm{id}_{T^c})\in\mathcal{I}(\mathcal{A}(\mathcal{E}^L)),\quad \iota^R(f^*)=(S^c,S^c,\mathrm{id}_{S^c})\in\mathcal{I}(\mathcal{A}(\mathcal{E}^R)).$$

This definition enables us to define a bimodule structure ${}_{\mathcal{I}(\mathcal{A}(\mathcal{E}^L))}C^-(\mathcal{P}^*)_{\mathcal{I}(\mathcal{A}(\mathcal{E}^R))}$ by extending the following multiplications to $C^-(\mathcal{P}^*)$. For an idempotent generator $\iota\in\mathcal{I}(\mathcal{A}(\mathcal{E}^L))$ let

$$\iota\cdot(S,T,\phi)^*=\begin{cases}(S,T,\phi)^* & \text{if } \iota^L(S,T,\phi)^*=\iota,\\ 0 & \text{otherwise,}\end{cases}$$

and for $\iota\in\mathcal{I}(\mathcal{A}(\mathcal{E}^R))$ let

$$(S,T,\phi)^*\cdot\iota=\begin{cases}(S,T,\phi)^* & \text{if } \iota^R(S,T,\phi)^*=\iota,\\ 0 & \text{otherwise.}\end{cases}$$

3.2.1 Diagrams and tangles associated to mirror-shadows

For a mirror-shadow $\mathcal{P}^*$ we use different conventions to associate diagrams and tangles:

Definition 3.20 Let $D^*(\mathcal{P}^*)$ be the mirror of $D(\mathcal{P})$ with respect to the vertical axis $\{\frac{1}{2}\}\times\mathbb{R}$.

To indicate that we work with mirrors we put a gray background underneath $D^*(\mathcal{P}^*)$.

Definition 3.21 Let $\mathcal{T}^*(\mathcal{P}^*)$ denote the mirror (with respect to the vertical axis) of $\mathcal{T}(\mathcal{P})$ with the over-crossings changed to under-crossings.

See Figures 12 and 13 for the elementary examples.

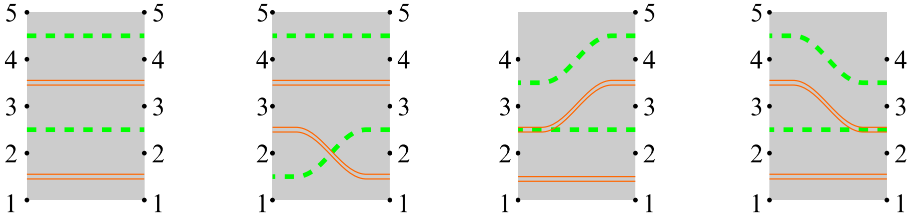


Figure 12: Examples of diagrams of mirror-shadows. On each figure, $\boldsymbol{a}$ and $\boldsymbol{a}_{1/2}$ are on the left-hand side, while $\boldsymbol{b}$ and $\boldsymbol{b}_{1/2}$ are on the right-hand side. Double (orange) lines connect $\{0\}\times\{s_X\}$ with $\{1\}\times\{\xi s_X\}$ and dashed (green) lines connect $\{1\}\times\{s_O\}$ with $\{0\}\times\{\omega s_O\}$.

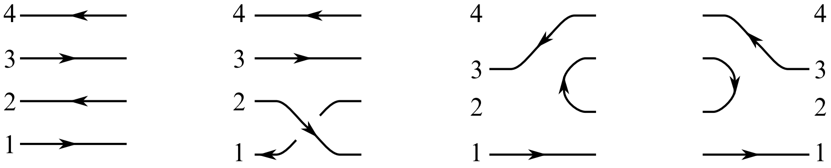


Figure 13: Elementary tangles corresponding to the mirror-shadows of Figure 12

3.2.2 Wedge product of shadows and mirror-shadows: type D maps

The mirror-shadow $\mathcal{P}_1^*$ and shadow $\mathcal{P}_2$ have a *well-defined wedge product* if $m_1 = m_2$, $T_{\mathbb{X}_1} = T_{\mathbb{X}_2}^c$ and $S_{\mathbb{O}_1} = S_{\mathbb{O}_2}^c$. This means exactly that $\mathcal{E}^R(\mathcal{P}_1^*) = \mathcal{E}_L(\mathcal{P}_2)$. Denote the ordered pair by $\mathcal{P}_1^* \wedge \mathcal{P}_2$. Diagrammatically, we indicate a wedge product by placing the corresponding diagrams next to each other. See Figure 14 for an example. Similarly, the shadow $\mathcal{P}_1$ and mirror-shadow $\mathcal{P}_2^*$ have a well-defined wedge product if $n_1 = n_2$, $S_{\mathbb{X}_1} = S_{\mathbb{X}_2}^c$ and $T_{\mathbb{O}_1} = T_{\mathbb{O}_2}^c$. The pair is denoted by $\mathcal{P}_1 \wedge \mathcal{P}_2^*$.

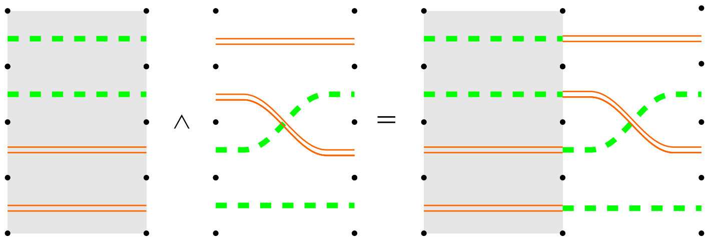


Figure 14: Wedge product of a mirror-shadow and a shadow

Let $\mathcal{I} = \mathcal{I}(\mathcal{A}(\mathcal{E}^R(\mathcal{P}_1^*))) = \mathcal{I}(\mathcal{A}(\mathcal{E}_L(\mathcal{P}_2)))$. Define

$$C^-(\mathcal{P}_1^* \wedge \mathcal{P}_2) = C^-(\mathcal{P}_1^*) \otimes_{\mathcal{I}} C^-(\mathcal{P}_2),$$

a module over $\mathbb{F}_2[U_O]_{s_O \in S_{\mathbb{O}_1} \cup S_{\mathbb{O}_2}}$. For generators $f_1^* = (S_1, T_1, \phi_1)^* \in \mathfrak{S}(\mathcal{P}_1^*)$ and $f_2 = (S_2, T_2, \phi_2) \in \mathfrak{S}(\mathcal{P}_2)$ such that $f = f_1^* \otimes f_2$ is nonzero, ie such that $S_1 = S_2^c$, define a map

$$\partial_\wedge(f_1^* \otimes f_2) = \partial^*(f_1^*) \otimes f_2 + f_1^* \otimes \partial(f_2) + \partial_{\mathrm{mix}}(f_1^* \otimes f_2),$$

where ∂^* and ∂ are the differentials on $C^-(\mathcal{P}_1^*)$ and $C^-(\mathcal{P}_2)$, respectively, and ∂_{mix} is defined below by looking at pairs of points in $S_1 \cup S_2 = \boldsymbol{b}$.

- For a pair $(p,q) \in S_1 \times S_2$ define $f^{pq} = (f_1^*)^{pq} \otimes f_2^{pq}$, where $(f_1^*)^{pq} = (S_1^{pq}, T_1^{pq}, \phi_1^{pq})^*$, $f_2^{pq} = (S_2^{pq}, T_2^{pq}, \phi_2^{pq})$. Here $S_1^{pq} = S_1 \setminus \{p\} \cup \{q\}$, $T_1^{pq} = T_1$ and, for $s_1 \in S_1^{pq}$,

$$\phi_1^{pq} s_1 = \begin{cases} \phi_1 p & \text{if } s_1 = q, \\ \phi_1 s_1 & \text{otherwise.} \end{cases}$$

Similarly, $S_2^{pq} = S_2 \setminus \{q\} \cup \{p\}$, $T_2^{pq} = T_2$ and, for $s_2 \in S_2^{pq}$,

$$\phi_2^{pq} s_2 = \begin{cases} \phi_2 q & \text{if } s_2 = p, \\ \phi_2 s_2 & \text{otherwise.} \end{cases}$$

Diagrammatically, f^{pq} is obtained from f by exchanging the p and q endpoints of the two strands ending at p and at q. The pair $(p,q) \in S_1 \times S_2$ is *exchangeable* if

- $\mathrm{Inv}(\phi_1) \supset \mathrm{Inv}({\phi_1}^{pq})$,
- $\mathrm{Inv}(\phi_2) \subset \mathrm{Inv}(\phi_2^{pq})$,
- $\mathrm{Inv}(\phi_1, \xi_1^{-1}) \supset \mathrm{Inv}({\phi_1}^{pq}, \xi_1^{-1})$, and
- $\mathrm{Inv}(\phi_2, \xi_2^{-1}) \subset \mathrm{Inv}(\phi_2^{pq}, \xi_2^{-1})$.

Diagrammatically, this means that while doing the exchange we cannot pick up crossings with black or orange strands on the $\mathcal{P}_1^*$–side and we cannot lose crossings with black or orange strands on the $\mathcal{P}_2$–side. Given such an exchangeable pair (p,q), for $O_1 = (s_{O_1}, \omega_1 s_{O_1})$ with $s_{O_1} \in S_{\mathbb{O}_1}$, let

$$n_{O_1}(pq) = \left|\operatorname{Inv}(\phi_1^{pq}, \omega_1|_{s_{O_1}}) \setminus \operatorname{Inv}(\phi_1, \omega_1|_{s_{O_1}})\right|,$$

and, for $O_2 = (s_{O_2}, \omega s_{O_2})$ with $s_{O_2} \in S_{\mathbb{O}_2}$, let

$$n_{O_2}(pq) = \left|\operatorname{Inv}(\phi_2, \omega_2|_{s_{O_2}}) \setminus \operatorname{Inv}(\phi_2^{pq}, \omega_2|_{s_{O_2}})\right|.$$

- For a pair $(p,q) \subset S_1$ with $p<q$ and $(p,q) \in \operatorname{Inv}(\phi_1)$, define $f^{pq} = (f_1^*)^{pq} \otimes f_2$, where $(f_1^*)^{pq} = (S_1, T_1, \phi_1^{(p,q)})$. The pair $(p,q) \subset S_1$ is *exchangeable* if
  - each $t \in [p,q] \cap \boldsymbol{b}$ is in S_1 and $\phi_1 t \in [\phi_1 q, \phi_1 p]$, and
  - each $t \in [p,q] \cap \boldsymbol{b}_{1/2}$ is in $T_{\mathbb{X}_1}$ and $\xi_1^{-1} t \in [\phi_1 q, \phi_1 p]$.

Diagrammatically, this means that in f each black or orange strand that ends between p and q is on the $\mathcal{P}_1^*$–side and crosses both black strands ending at p and at q. Given such an exchangeable pair (p,q), for $O_1 = (s_{O_1}, \omega_1 s_{O_1})$ with $s_{O_1} \in S_{\mathbb{O}_1}$,

$$n_{O_1}(pq) = \begin{cases} 1 & \text{if } s_{O_1} \in [p,q] \text{ and } \omega s_{O_1} \notin [\phi_1 q, \phi_1 p], \\ 0 & \text{otherwise,} \end{cases}$$

and for $O_2 = (s_{O_2}, \omega s_{O_2})$ with $s_{O_2} \in S_{\mathbb{O}_2}$ let

$$n_{O_2}(pq) = \begin{cases} 1 & \text{if } s_{O_2} \in [p,q], \\ 0 & \text{otherwise.} \end{cases}$$

- For a pair $(p,q) \subset S_2$ with $p<q$ and $(p,q) \notin \operatorname{Inv}(\phi_2)$, define $f^{pq} = f_1^* \otimes f_2^{pq}$, where $f_2^{pq} := (S_2, T_2, \phi_2^{(p,q)})$. The pair $(p,q) \subset S_2$ is *exchangeable* if
  - each $t \in [p,q] \cap \boldsymbol{b}$ is in S_2 and $\phi_2 t \in [\phi_2 p, \phi_2 p]$, and
  - each $t \in [p,q] \cap \boldsymbol{b}_{1/2}$ is in $T_{\mathbb{X}_2}$ and $\xi_2^{-1} t \in [\phi_2 p, \phi_2 q]$.

Diagrammatically, this means that in f all black and orange strands that end between p and q are on the $\mathcal{P}_2$–side, and they do not cross either of the two black strands ending at p and at q. Given such an exchangeable pair (p,q), for $O_1 = (s_{O_1}, \omega_1 s_{O_1})$ with $s_{O_1} \in S_{\mathbb{O}_1}$ let

$$n_{O_1}(pq) = \begin{cases} 1 & \text{if } s_{O_1} \in [p,q], \\ 0 & \text{otherwise,} \end{cases}$$

and, for $O_2 = (s_{O_2}, \omega s_{O_2})$ with $s_{O_2} \in S_{\mathbb{O}_2}$, let

$$n_{O_2}(pq) = \begin{cases} 1 & \text{if } s_{O_2} \in [p,q] \text{ and } \omega s_{O_2} \notin [\phi_2 p, \phi_2 q], \\ 0 & \text{otherwise.} \end{cases}$$

Denote the set of exchangeable pairs for f by $\mathrm{Exch}(f)$.

Then

$$\partial_{\mathrm{mix}}(f) = \sum_{(p,q)\in \mathrm{Exch}(f)} \prod_{s_O \in S_{\mathbb{O}_1} \cup S_{\mathbb{O}_2}} U_O^{n_O(pq)} f^{pq}.$$

See Figure 15 for an example of the mixed differential.

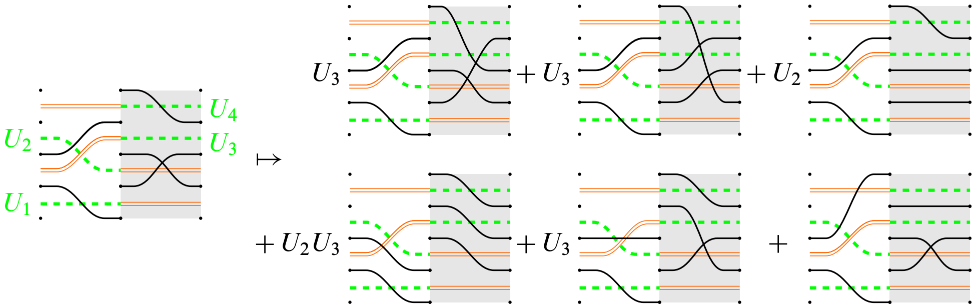


Figure 15: The differential $\partial_\wedge$. The last four terms on the right-hand side correspond to ∂_{mix}.

Extend $\partial_\wedge$ linearly to the whole module $C^-(\mathcal{P}_1^* \wedge \mathcal{P}_2)$.

Proposition 3.22 *$(C^-(\mathcal{P}_1^* \wedge \mathcal{P}_2), \partial_\wedge)$ is a chain complex.*

The proof of Proposition 3.22 is straightforward after the reformulation of the algebra to the language of bordered grid diagrams in Section 4.5 and thus it will be given there.

If $\mathcal{P}_1$ and $\mathcal{P}_2^*$ have a well-defined wedge product then $\partial_\wedge$ can be defined similarly on $C^-(\mathcal{P}_1 \wedge \mathcal{P}_2^*) = C^-(\mathcal{P}_1) \otimes_{\mathcal{I}(\mathcal{A}(\mathcal{E}_R(\mathcal{P}_1)))} C^-(\mathcal{P}_2^*)$ by

$$\partial_\wedge(f_1 \otimes f_2^*) = \partial_1(f_1) \otimes f_2^* + f_1 \otimes \partial_2^*(f_2^*) + \partial_{\mathrm{mix}}(f_1 \otimes f_2^*),$$

where the mixed differential ∂_{mix} is defined by following the same shadow and mirror-shadow rules as earlier. Specifically, we look at pairs of black strands, and exchange their endpoints in $T_1 \cup T_2$ if the following conditions are met:

- If one endpoint is in T_1 and the other in T_2, then while doing the exchange we cannot pick up crossings with black or orange strands on the $\mathcal{P}_2^*$–side and we cannot lose crossings with black or orange strands on the $\mathcal{P}_1$–side. If we pick up crossings with green strands on the $\mathcal{P}_2^*$–side or lose crossings with green strands on the $\mathcal{P}_1$–side, we record it with U_O–variables.
- If both endpoints are in T_1, then each black or orange strand that ends between the two points must be on the $\mathcal{P}_1$–side and cannot cross either of the given two black

strands. A green strand that ends between the two points but is either on the $\mathcal{P}_2^*$–side or crosses one of the two black strands is recorded with a U_O–variable.

- If both endpoints are in T_2, then each black or orange strand that ends between the two points must be on the $\mathcal{P}_2^*$–side, and crosses both of the given two black strands. A green strand that ends between the two points but either doesn't cross both black strands or is on the $\mathcal{P}_1$–side is recorded with a U_O–variable.

Then we have:

Proposition 3.23 *$(C^-(\mathcal{P}_1 \wedge \mathcal{P}_2^*), \partial_\wedge)$ is a chain complex.*

The proof of Proposition 3.23 will be given in Section 4.5 as well.

These propositions allow us to define left and right type D maps on generators $f^* = (S,T,\phi)^*$ by

$$\begin{aligned}\delta^R\colon C^-(\mathcal{P}^*) &\to C^-(\mathcal{P}^*) \otimes \mathcal{A}(\mathcal{E}^R),\\ (S,T,\phi)^* &\mapsto \partial_\wedge((S,T,\phi)^* \otimes \iota^R(S,T,\phi)^*),\end{aligned}$$

and

$$\begin{aligned}\delta^L\colon C^-(\mathcal{P}^*) &\to \mathcal{A}(\mathcal{E}^L) \otimes C^-(\mathcal{P}^*),\\ (S,T,\phi)^* &\mapsto \partial_\wedge(\iota^L(S,T,\phi)^* \otimes (S,T,\phi)^*).\end{aligned}$$

The maps δ^L and δ^R extend to the whole module $C^-(\mathcal{P}^*)$ and by merging them we can define a type *DD* structure:

Definition 3.24 With the above notation let $\mathit{CDTD}^-(\mathcal{P}^*)$ be the left–right type *DD* structure $(C^-(\mathcal{P}^*), \delta^1)$ over $\mathcal{A}(\mathcal{E}^L)$ and $\mathcal{A}(\mathcal{E}^R)$, where

$$\delta^1\colon C^-(\mathcal{P}^*) \to \mathcal{A}(\mathcal{E}^L) \otimes C^-(\mathcal{P}^*) \otimes \mathcal{A}(\mathcal{E}^R)$$

is defined via

$$\begin{aligned}\delta^1(f^*) = \iota^L(f^*) \otimes \partial^*(f^*) \otimes \iota^R(f^*) + \iota^L(f^*) \otimes \partial_{\mathrm{mix}}(f^* \otimes \iota^R(f^*))\\ + \partial_{\mathrm{mix}}(\iota^L(f^*) \otimes f^*) \otimes \iota^R(f^*).\end{aligned}$$

The type *DD* structure identities hold as a consequence of Propositions 3.22 and 3.23:

Proposition 3.25 *Let $\mathcal{P}^*$ be a mirror shadow. Then*

1. *as defined above, $(C^-(\mathcal{P}^*), \delta^L)$ is a left type D structure over $\mathcal{A}(\mathcal{E}^L)$;*
2. *as defined above, $(C^-(\mathcal{P}^*), \delta^R)$ is a right type D structure over $\mathcal{A}(\mathcal{E}^R)$;*
3. *$\mathit{CDTD}^-(\mathcal{P}^*)$ is a left–right type DD structure over $\mathcal{A}(\mathcal{E}^L)$ and $\mathcal{A}(\mathcal{E}^R)$.*

Proof As the proofs of all parts of the proposition are similar, we only prove (1). Recall that the left type D identity that we need to show is

$$(m_2 \otimes \mathrm{id}) \circ (\mathrm{id}_{\mathcal{A}} \otimes \delta^L) \circ \delta^L + (\partial_{\mathcal{A}} \otimes \mathrm{id}) \circ \delta^L = 0.$$

Let f^* be a generator of $C^-(\mathcal{P}^*)$ and let $\iota = \iota^L(f^*)$. Using $\partial\iota = 0$, we can rewrite the first term on the left-hand side as

$$(\partial_{\mathcal{A}} \otimes \mathrm{id}) \circ \delta^L(f^*) = (\partial_{\mathcal{A}} \otimes \mathrm{id}) \circ \partial_{\mathrm{mix}}(\iota \otimes f^*),$$

and using also that $(\partial^*)^2 = 0$, we can rewrite the second term on the left-hand side as

$$\begin{aligned}(m_2 \otimes \mathrm{id}) &\circ (\mathrm{id}_{\mathcal{A}} \otimes \delta^L) \circ \delta^L(f^*) \\ &= \partial_{\mathrm{mix}}(\iota \otimes \partial^* f^*) + (\mathrm{id}_{\mathcal{A}} \otimes \partial^*) \circ \partial_{\mathrm{mix}}(\iota \otimes f^*) + \partial^2_{\mathrm{mix}}(\iota \otimes f^*).\end{aligned}$$

The resulting four terms are exactly the nonzero summands of $\partial^2_\wedge(\iota \otimes f^*)$, which, since $\partial_\wedge$ is a chain map, vanishes. This finishes the proof of (1). □

This concept can be extended to multiple wedge products as follows. Let $\boldsymbol{\mathcal{P}} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ be an alternating sequence of shadows and mirror-shadows with well-defined consecutive wedge products. (Here and throughout the paper $\mathcal{P}^\circ$ indicates $\mathcal{P}$ or $\mathcal{P}^*$.) Then we can define a differential on

$$C^-(\boldsymbol{\mathcal{P}}) = C^-(\mathcal{P}_1^\circ)^\circ \otimes \dots \otimes C^-(\mathcal{P}_p^\circ)^\circ$$

by defining it on $\boldsymbol{f} = f_1^\circ \otimes \dots \otimes f_p^\circ$ as

$$\partial_\wedge \boldsymbol{f} = \sum_{j=1}^{p} f_1^\circ \otimes \dots \otimes \partial^\circ(f_j^\circ) \otimes \dots \otimes f_p^\circ + \sum_{j=1}^{p-1} f_1^\circ \otimes \dots \otimes \partial_{\mathrm{mix}}(f_j^\circ \otimes f_{j+1}^\circ) \otimes \dots \otimes f_p^\circ.$$

Observe that, depending on whether $\boldsymbol{\mathcal{P}}$ starts (ends) with a shadow or mirror-shadow, $C^-(\boldsymbol{\mathcal{P}})$ is equipped with a type *AA*, *AA*, *DA* or *DD* structure. Denote these structures by $\mathit{CATA}^-(\boldsymbol{\mathcal{P}})$, $\mathit{CATD}^-(\boldsymbol{\mathcal{P}})$, $\mathit{CDTA}^-(\boldsymbol{\mathcal{P}})$ or $\mathit{CDTD}^-(\boldsymbol{\mathcal{P}})$. Or sometimes — as the type is anyways specified by the sequence $\boldsymbol{\mathcal{P}}$ — we will refer to any of the above structures as $\mathit{CT}^-(\boldsymbol{\mathcal{P}})$.

3.2.3 Tangles associated to wedge products

Let $\boldsymbol{\mathcal{P}} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ be an alternating sequence of shadows and mirror-shadows with well-defined consecutive wedge products. Having a well-defined wedge product exactly means that the associated diagrams $\mathcal{D}(\mathcal{P}_j^\circ)$ and thus the associated tangles $\mathcal{T}(\mathcal{P}_j^\circ)$ match up. Thus let $\mathcal{D}(\boldsymbol{\mathcal{P}})$ and $\mathcal{T}(\boldsymbol{\mathcal{P}})$ be their concatenations.

3.3 One-sided modules

When a shadow or a mirror-shadow corresponds to a tangle with $\partial^0 = \varnothing$ or $\partial^1 = \varnothing$, then the left or right map can be contracted to a differential giving a one-sided right or left module. Thus, in this subsection we would like to "close up" one side of the bimodule and incorporate one of the type A (or type D) maps as a new component of the differential. (Note that this "closing up" is easier to follow in the related Section 4.6). Below we will describe in detail the closing up of the left type D map on a type DD bimodule associated to a mirror-shadow. This way we obtain a right type D structure.

Suppose that for a mirror-shadow $\mathcal{P}^*$ we have $\boldsymbol{a}_{1/2} = S_{\mathbb{X}} = T_{\mathbb{O}}$. Then we can define a new component of the differential ${}_D\partial$ that will correspond to resolving some crossings (remember that originally the type D map corresponds to introducing crossings) so that $\partial^* + {}_D\partial$ is a differential (ie has square 0) when restricted to $\mathfrak{S}_n(\mathcal{P}^*)$ (where $\mathfrak{S}_n(\mathcal{P}^*)$ consists of the generators $(S,T,\phi)^*$ with $|S| = |T| = n$).

Consider a generator $f^* = (S,T,\phi)^* \in \mathfrak{S}_n(\mathcal{P}^*)$. Suppose that for $s_1 < s_2$ the pair (s_1,s_2) is in $\operatorname{Inv}(\phi)$, ie $\phi(s_1) > \phi(s_2)$. We say that the exchange (s_1,s_2) is *allowable* if for any $t \in [\phi(s_2),\phi(s_1)]$ we have $\phi^{-1}(t) \in [s_1,s_2]$ and similarly for any $s_X \in [\phi(s_2),\phi(s_1)]$ we have $\xi(s_X) \in [s_1,s_2]$. Denote the set of such allowable pairs by ${}_D\operatorname{Exch}(\phi) \subset S \times S$. See Figure 16 for an example.

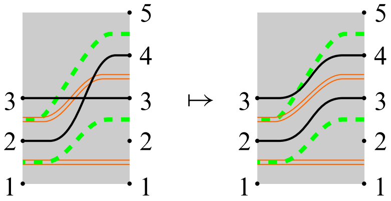


Figure 16: The differential ${}_D\partial$

For $O = (s_O, \omega s_O)$ define

$$ {}_Dn_O(s_1,s_2) = \begin{cases} 1 & \text{if } \omega s_O \in [\phi(s_2),\phi(s_1)] \text{ and } s_O \notin [s_1,s_2], \\ 0 & \text{otherwise.} \end{cases} $$

Then define

$$ {}_D\partial f^* = \sum_{(s_1,s_2)\in {}_D\operatorname{Exch}(\phi)} U_O^{{}_Dn_O(s_1,s_2)} (f^{(s_1,s_2)})^*. $$

The map $\partial^* + {}_D\partial$ can be extended to the module $C_n^-(\mathcal{P}^*)$ generated by $\mathfrak{S}_n(\mathcal{P}^*)$ over $\boldsymbol{k}$. Although $({}_D\partial)^2 \neq 0$ we have:

Lemma 3.26 *$(C_n^-(\mathcal{P}^*), \partial^* + {}_D\partial)$ is a chain complex.*

The proof of Lemma 3.26 will be given using the grid diagram reformulation of $\partial^* + {}_D\partial$ as the differential of an annular bordered grid diagram in Section 4.6.

Definition 3.27 With the above notation let $CTD^-(\mathcal{P}^*)$ be the right type D structure $(C_n^-(\mathcal{P}^*), \delta^1)$ over $\mathcal{A}(\mathcal{E}^R)$, where

$$\delta^1 \colon C_n^-(\mathcal{P}^*)^* \to C_n^-(\mathcal{P}^*)^* \otimes \mathcal{A}(\mathcal{E}^R)$$

is given by

$$f^* \mapsto \delta^R(f^*) + {}_D\partial f^* \otimes \iota^R(f^*).$$

Aside from the gradings that will be defined later, Lemma 3.26 shows that $CTD^-(\mathcal{P}^*)$ is indeed a right type D structure.

The contraction of the right type D map ∂_D can be defined similarly for mirror-shadows with $T_{\mathbb{X}} = S_{\mathbb{O}} = \boldsymbol{b}_{1/2}$ by exchanging pairs $(s_1, s_2) \in \operatorname{Inv}(\phi)$ such that any $s \in [s_1, s_2]$ has $\phi(s) \in [\phi(s_2), \phi(s_1)]$ and any $t_X \in [s_1, s_2]$ has $\xi^{-1}(t_X) \in [\phi(s_2), \phi(s_1)]$. In this way we obtain a left type D structure $CDT^-(\mathcal{P}^*)$ over $\mathcal{A}(\mathcal{E}^L)$ on $C_m^-(\mathcal{P}^*)$. In this paper we do not need to contract the type A actions, but the definitions go similarly with the only difference that ${}_A\partial$ and ∂_A introduce crossings.

Convention 3.28 Whenever the leftmost and/or rightmost shadow or mirror-shadow in a given well-defined wedge product $\boldsymbol{\mathcal{P}}$ is contractible, we will assume that the corresponding differential ∂ or ∂^* has been replaced with the appropriate map ${}_D\partial$, ∂_D, ${}_A\partial$ or ∂_A in the definition of $\partial_\wedge$, to produce a one-sided module $CTD^-(\boldsymbol{\mathcal{P}})$, $CDT^-(\boldsymbol{\mathcal{P}})$ or $CAT^-(\boldsymbol{\mathcal{P}})$, or $CTA^-(\boldsymbol{\mathcal{P}})$, or a chain complex $CT^-(\boldsymbol{\mathcal{P}})$. In these cases again we may use the notation $CT^-(\boldsymbol{\mathcal{P}})$ to refer to any of these structures, as the type is specified by the sequence $\boldsymbol{\mathcal{P}}$.

3.4 Gradings

Unlike for other bordered theories, one can define surprisingly simple absolute gradings on the structures here. For a shadow $\mathcal{P}$, we define the *Maslov* and *Alexander* gradings of a generator $f = (S, T, \phi)$ of the module as

$$\begin{aligned} M(f) &= \operatorname{inv}(\phi) - \operatorname{inv}(\phi, \omega) + \operatorname{inv}(\omega), \\ 2A(f) &= \operatorname{inv}(\phi, \xi^{-1}) - \operatorname{inv}(\phi, \omega) + \operatorname{inv}(\omega) - \operatorname{inv}(\xi^{-1}) - |T_{\mathbb{X}}|. \end{aligned}$$

For $O = (s_O, \omega s_O)$ define

$$M(U_O f) = M(f) - 2, \quad A(U_O f) = A(f) - 1.$$

This defines a grading on $C^-(\mathcal{P})$ and consequently on $CATA^-(\mathcal{P})$.

For a mirror-shadow $\mathcal{P}^*$ the gradings on $f^* = (S, T, \phi)^*$ are defined as

$$\begin{aligned} M(f^*) &= -\operatorname{inv}(\phi) + \operatorname{inv}(\phi, \omega) - \operatorname{inv}(\omega) - |S_{\mathbb{O}}|, \\ 2A(f^*) &= -\operatorname{inv}(\phi, \xi^{-1}) + \operatorname{inv}(\phi, \omega) - \operatorname{inv}(\omega) + \operatorname{inv}(\xi^{-1}) - |S_{\mathbb{O}}|, \end{aligned}$$

and again

$$M(U_O f^*) = M(f^*) - 2, \quad A(U_O f^*) = A(f^*) - 1.$$

This defines a grading on $C^-(\mathcal{P}^*)$ and consequently on $\mathit{CDTD}^-(\mathcal{P}^*)$. For an alternating sequence of shadows and mirror-shadows $\boldsymbol{\mathcal{P}} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ with well-defined consecutive wedge product, define the gradings on $\boldsymbol{f} = f_1^\circ \otimes \cdots \otimes f_p^\circ$ as the sums

$$M(\boldsymbol{f}) = \sum_{j=1}^{p} M(f_j^\circ), \quad A(\boldsymbol{f}) = \sum_{j=1}^{p} A(f_j^\circ).$$

All the differentials, multiplications and wedge products behave well with the gradings.

Theorem 3.29 *For a shadow $\mathcal{P}$, horizontal shadow $\mathcal{E}$ and composable shadows $\mathcal{P}_1$ and $\mathcal{P}_2$:*

1. *$(C^-(\mathcal{P}), \partial)$ is a graded chain complex with grading M. Moreover ∂ preserves A.*
2. *The multiplication $\cdot\colon C^-(\mathcal{P}_1) \otimes C^-(\mathcal{P}_2) \to C^-(\mathcal{P}_1 \circ \mathcal{P}_2)$ is a degree $(0,0)$ map.*
3. *$\mathcal{A}(\mathcal{E})$ is a differential graded algebra with grading M. Moreover A is preserved by both the multiplication and the differential.*
4. *$\mathit{CATA}^-(\mathcal{P})$ is a left–right differential graded bimodule over $\mathcal{A}(\mathcal{E}_L)$ and $\mathcal{A}(\mathcal{E}_R)$ (in particular a type AA structure) with grading M. Moreover A is preserved both by the multiplication and the differential.*

Theorem 3.30 *For a mirror-shadow $\mathcal{P}^*$:*

1. *$(C^-(\mathcal{P}^*), \partial^*)$ is a graded chain complex with grading M. Moreover ∂^* preserves A.*
2. *$\mathit{CDTD}^-(\mathcal{P}^*)$ is a left–right type DD structure over $\mathcal{A}(\mathcal{E}^L)$ and $\mathcal{A}(\mathcal{E}^R)$ with grading M. Moreover δ^1 preserves A.*

For tangles in $I \times S^2$ we have:

Theorem 3.31 *Suppose that $\boldsymbol{\mathcal{P}} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ is an alternating sequence of shadows and mirror-shadows with well-defined consecutive wedge product. If in addition $\mathcal{P}_1^\circ$ does not have contractible left-hand side and $\mathcal{P}_p^\circ$ does not have contractible right-hand side, then:*

(1) *If $\mathcal{P}_1$ and $\mathcal{P}_p$ are both shadows then CATA$^-(\boldsymbol{\mathcal{P}})$ is a left–right type AA structure over $\mathcal{A}(\mathcal{E}_L(\mathcal{P}_1))$ and $\mathcal{A}(\mathcal{E}_R(\mathcal{P}_p))$ with grading M. Moreover A is preserved by all multiplications $m_{0,1,0}, m_{1,1,0}$ and $m_{0,1,1}$.*

(2) *If $\mathcal{P}_1$ is a shadow and $\mathcal{P}_p^*$ is a mirror-shadow then CATD$^-(\boldsymbol{\mathcal{P}})$ is a left–right type AD structure over $\mathcal{A}(\mathcal{E}_L(\mathcal{P}_1))$ and $\mathcal{A}(\mathcal{E}^R(\mathcal{P}_p^*))$ with grading M. Moreover A is preserved by the maps δ_1^1 and δ_2^1.*

(3) *If $\mathcal{P}_1^*$ is a mirror-shadow and $\mathcal{P}_p$ is a shadow then CDTA$^-(\boldsymbol{\mathcal{P}})$ is a left–right type DA structure over $\mathcal{A}(\mathcal{E}^L(\mathcal{P}_1^*))$ and $\mathcal{A}(\mathcal{E}_R(\mathcal{P}_p))$ with grading M. Moreover A is preserved by the maps δ_1^1 and δ_2^1.*

(4) *If $\mathcal{P}_1^*$ and $\mathcal{P}_p^*$ are both mirror-shadows then CDTD$^-(\boldsymbol{\mathcal{P}})$ is a left–right type DD structure over $\mathcal{A}(\mathcal{E}^L(\mathcal{P}_1^*))$ and $\mathcal{A}(\mathcal{E}^R(\mathcal{P}_p^*))$ with grading M. Moreover A is preserved by the map δ^1.*

For tangles in D^3 and S^3:

Theorem 3.32 *Suppose that $\boldsymbol{\mathcal{P}} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ is an alternating sequence of shadows and mirror-shadows with well-defined consecutive wedge product. Then:*

(1) *If $\mathcal{P}_1^\circ$ is left-contractible, and $\mathcal{P}_p$ is a non-right-contractible shadow, then CTA$^-(\boldsymbol{\mathcal{P}})$ is a right type A structure over $\mathcal{A}(\mathcal{E}_R(\mathcal{P}_p))$ with grading M. Moreover A is preserved by all multiplications m_0 and m_1.*

(2) *If $\mathcal{P}_1^\circ$ is left-contractible, and $\mathcal{P}_p^*$ is a non-right-contractible mirror-shadow, then CTD$^-(\boldsymbol{\mathcal{P}})$ is a right type D structure over $\mathcal{A}(\mathcal{E}^R(\mathcal{P}_p^*))$ with grading M. Moreover A is preserved by the map δ^1.*

(3) *If $\mathcal{P}_p^\circ$ is right-contractible, and $\mathcal{P}_1$ is a non-left-contractible shadow, then CAT$^-(\boldsymbol{\mathcal{P}})$ is a left type A structure over $\mathcal{A}(\mathcal{E}_L(\mathcal{P}_1))$ with grading M. Moreover A is preserved by all multiplications m_0 and m_1.*

(4) *If $\mathcal{P}_p^\circ$ is right-contractible, and $\mathcal{P}_1^*$ is a non-left-contractible mirror-shadow, then CDT$^-(\boldsymbol{\mathcal{P}})$ is a left type D structure over $\mathcal{A}(\mathcal{E}^L(\mathcal{P}_1^*))$ with grading M. Moreover A is preserved by the map δ^1.*

(5) *If $\mathcal{P}_1^\circ$ is left-contractible and $\mathcal{P}_p^\circ$ is right-contractible, then CT$^-(\boldsymbol{\mathcal{P}})$ is a graded chain complex over $\boldsymbol{k}$ with grading M. Moreover ∂ preserves A.*

Proof of Theorems 3.29, 3.30, 3.31 and 3.32 Theorem 3.29 and Theorem 3.30(1) are consequences of Propositions 3.15, 3.22 and 3.23 and the definition of the grading. Theorem 3.30(2) is a consequence of Theorem 3.31, and the ungraded version of each item of Theorems 3.31 and 3.32 follows from Propositions 3.22 and 3.23. Thus, what is left to check is that $\partial_\wedge$ is a degree $(-1, 0)$ map. To keep notation simple, we will give a proof in the case of $(C^-(\mathcal{P}_1^* \wedge \mathcal{P}_2), \partial_\wedge)$. Other cases follow the same way. Given a generator $f = f_1^* \otimes f_2 = (S_1, T_1, \phi_1)^* \otimes (S_2, T_2, \phi_2)$,

$$\partial_\wedge(f_1^* \otimes f_2) = \partial_1^*(f_1^*) \otimes f_2 + f_1^* \otimes \partial_2(f_2) + \partial_{\mathrm{mix}}(f_1^* \otimes f_2).$$

For the first two terms the statement follows from Theorem 3.29 and Theorem 3.30(2). Next note that

$$\begin{aligned} M(f) &= -\operatorname{inv}(\phi_1) + \operatorname{inv}(\phi_2) + \operatorname{inv}(\phi_1, \omega_1) - \operatorname{inv}(\phi_2, \omega_2) - \operatorname{inv}(\omega_1) + \operatorname{inv}(\omega_2) - |S_{\mathbb{O}_1}|, \\ 2A(f) &= -\operatorname{inv}(\phi_1, \xi_1^{-1}) + \operatorname{inv}(\phi_2, \xi_2) + \operatorname{inv}(\phi_1, \omega_1) - \operatorname{inv}(\phi_2, \omega_2) \\ &\qquad - \operatorname{inv}(\omega_1) + \operatorname{inv}(\omega_2) + \operatorname{inv}(\xi_1^{-1}) - \operatorname{inv}(\xi_2^{-1}) - |S_{\mathbb{O}_1}| - |T_{\mathbb{X}_2}|. \end{aligned}$$

For an exchangeable pair $(p,q) \in S_1 \times S_2$ we can write the same two equations by changing ϕ_1 and ϕ_2 to ϕ_1^{pq} and ϕ_2^{pq}, respectively.

Since $S_{\mathbb{O}_1} \sqcup S_{\mathbb{O}_2} = \{1, \dots, m_1\}$ and the intersection points only change for strands that end or start between p and q, we have

$$\begin{aligned} |p-q| &= |\operatorname{Inv}(\phi_1^{pq}, \omega_1) \setminus \operatorname{Inv}(\phi_1, \omega_1)| + |\operatorname{Inv}(\phi_1, \omega_1) \setminus \operatorname{Inv}(\phi_1^{pq}, \omega_1)| \\ &\qquad + |\operatorname{Inv}(\phi_2, \omega_2) \setminus \operatorname{Inv}(\phi_2^{pq}, \omega_2)| + |\operatorname{Inv}(\phi_2^{pq}, \omega_2) \setminus \operatorname{Inv}(\phi_2, \omega_2)| \\ &= |\operatorname{Inv}(\phi_1^{pq}, \omega_1)| - |\operatorname{Inv}(\phi_1, \omega_1)| - 2|\operatorname{Inv}(\phi_1^{pq}, \omega_1) \setminus \operatorname{Inv}(\phi_1, \omega_1)| \\ &\qquad + |\operatorname{Inv}(\phi_2, \omega_2)| - |\operatorname{Inv}(\phi_2^{pq}, \omega_2)| - 2|\operatorname{Inv}(\phi_2, \omega_2) \setminus \operatorname{Inv}(\phi_2^{pq}, \omega_2)| \\ &= -2\sum_{s_O} n_O(pq) + |\operatorname{Inv}(\phi_1^{pq}, \omega_1)| - |\operatorname{Inv}(\phi_1, \omega_1)| \\ &\qquad + |\operatorname{Inv}(\phi_2, \omega_2)| - |\operatorname{Inv}(\phi_2^{pq}, \omega_2)|. \end{aligned}$$

Since the pair (p,q) is exchangeable, we have $\operatorname{Inv}(\phi_1) \subset \operatorname{Inv}(\phi_1^{pq})$, so for the inversions of ϕ_1 and ϕ_2 the analog of the above formula simplifies to

$$\operatorname{inv}(\phi_1^{pq}) - \operatorname{inv}(\phi_1) + \operatorname{inv}(\phi_2) - \operatorname{inv}(\phi_2^{pq}) = |p-q| - 1.$$

Similarly we get

$$\operatorname{inv}(\phi_1^{pq}, \xi_1^{-1}) - \operatorname{inv}(\phi_1, \xi_1^{-1}) + \operatorname{inv}(\phi_2, \xi_2^{-1}) - \operatorname{inv}(\phi_2^{pq}, \xi_2^{-1}) = |p-q|,$$

which gives

$$\begin{aligned} M(f) - M\Bigl(\prod_{s_O \in S_{\mathbb{O}_1} \cup S_{\mathbb{O}_2}} U^{n_O(pq)} f^{pq}\Bigr) &= 1, \\ A(f) - A\Bigl(\prod_{s_O \in S_{\mathbb{O}_1} \cup S_{\mathbb{O}_2}} U^{n_O(pq)} f^{pq}\Bigr) &= 0. \end{aligned}$$

Similar counting arguments work for exchangeable pairs (p,q) with $(p,q) \subset S_1$ or $(p,q) \subset S_2$. □

3.5 Pairing generalized strand modules

Taking a wedge product of a shadow and a mirror-shadow corresponds to taking the box tensor product of their algebraic structures:

Theorem 3.33 *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be shadows. Then:*

(1) *If the mirror-shadow $\mathcal{P}_1^*$ and shadow $\mathcal{P}_2$ have well-defined wedge products, then the left–right type DA structures*

$$CDTA^-(\mathcal{P}_1^* \wedge \mathcal{P}_2) \quad \text{and} \quad CDTD^-(\mathcal{P}_1^*) \boxtimes CATA^-(\mathcal{P}_2)$$

over $\mathcal{A}(\mathcal{E}^L(\mathcal{P}_1^*))$ and $\mathcal{A}(\mathcal{E}_R(\mathcal{P}_2))$ are isomorphic as type DA structures.

(2) *If the shadow $\mathcal{P}_1$ and mirror-shadow $\mathcal{P}_2^*$ have well-defined wedge products, then the left–right type AD structures*

$$CATD^-(\mathcal{P}_1 \wedge \mathcal{P}_2^*) \quad \text{and} \quad CATA^-(\mathcal{P}_1) \boxtimes CDTD^-(\mathcal{P}_2^*)$$

over $\mathcal{A}(\mathcal{E}_L(\mathcal{P}_1))$ and $\mathcal{A}(\mathcal{E}^R(\mathcal{P}_2^*))$ are isomorphic as type AD structures.

Proof This follows directly from the definition of δ^L, δ^R, and ∂_{mix}. □

Similar theorems hold for multiple wedge products of shadows and mirror-shadows.

3.6 Relations between the U–actions

Let $\mathcal{P} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ be an alternating sequence of shadows and mirror-shadows with well-defined consecutive wedge products. For $s_O \in S_{\mathbb{O}_i}$ and $s'_O \in S_{\mathbb{O}_{i'}}$ let $O = (s_O, \omega_i s_O)$ and $O' = (s'_O, \omega_{i'} s'_O)$.

Definition 3.34 The pairs O and O' are *connected by a path of length k* if there is a sequence of elements $s_O = s_0, s_1, \dots, s_k = s'_O$ such that $s_l \in S_{\mathbb{O}_{j_l}}$ and $s_{l+1} = \xi_{j'_l} \omega_{j_l} s_l$. Here, depending on whether $\mathcal{P}_{j_l}^\circ$ is a shadow or a mirror-shadow, $\omega_{j_l} s_l$ is in $S_{\mathbb{X}_{j_l}} \amalg S_{\mathbb{X}_{j_l-1}}$ or $S_{\mathbb{X}_{j_l}} \amalg S_{\mathbb{X}_{j_l+1}}$, thus j'_l equals $j_l - 1$, j_l or $j_l + 1$.

An example of a path is pictured in Figure 17.

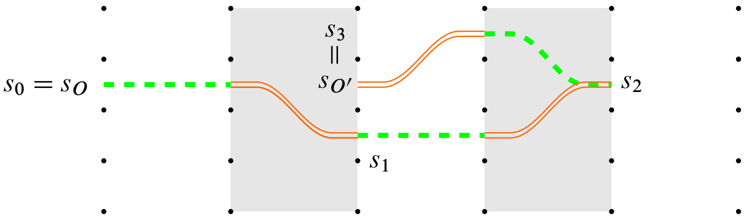


Figure 17: A path of length three

Lemma 3.35 *Suppose that O and O' are connected by a path. Then the actions of U_O and $U_{O'}$ on $CT^-(\mathcal{P})$ are equivalent.*

Here and throughout the paper "equivalent" means equivalence for the appropriate structures. Thus, it means type AA equivalence for $CATA^-(\mathcal{P})$, type DA equivalence for $CDTA^-(\mathcal{P})$, type AD equivalence for $CATD^-(\mathcal{P})$ and type DD equivalence for $CDTD^-(\mathcal{P})$.

The proof of Lemma 3.35 will be given in the next section, after introducing bordered grid diagrams.

4 Bordered grid diagrams

In what follows we introduce bordered grid diagrams and structures corresponding to bordered grid diagrams. As it will turn out, all of these notions are reformulations of notions from Section 3.

Bordered grid diagrams are a relative version of the grid diagrams used in combinatorial knot Floer homology [11; 12]. Many of the definitions below are parallel to the ones in [11; 12].

Definition 4.1 A *bordered grid diagram G* in $[c_1,c_2]\times[d_1,d_2]$ is given by a quadruple $(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbb{X},\mathbb{O})$, where $\boldsymbol{\alpha}=\{\alpha_a\}_{a\in\boldsymbol{a}}$ is a set of horizontal arcs indexed by $\boldsymbol{a}=(d_1,d_2)\cap\mathbb{Z}$ with $\alpha_a=[c_1,c_2]\times\{a\}$, and $\boldsymbol{\beta}=\{\beta_b\}_{b\in\boldsymbol{b}}$ is a set of vertical arcs indexed by $\boldsymbol{b}=(c_1,c_2)\cap\mathbb{Z}$ with $\beta_b=\{b\}\times[d_1,d_2]$. The markings $\mathbb{X}$ and $\mathbb{O}$ are subsets of $[c_1,c_2]\times[d_1,d_2]\cap\left(\mathbb{Z}+\frac{1}{2}\right)\times\left(\mathbb{Z}+\frac{1}{2}\right)$ with the property that

$$\left|[c_1,c_2]\times\left\{j+\tfrac{1}{2}\right\}\cap\mathbb{X}\right|\le 1,\quad \left|\left\{j+\tfrac{1}{2}\right\}\times[d_1,d_2]\cap\mathbb{X}\right|\le 1,$$
$$\left|[c_1,c_2]\times\left\{j+\tfrac{1}{2}\right\}\cap\mathbb{O}\right|\le 1,\quad \left|\left\{j+\tfrac{1}{2}\right\}\times[d_1,d_2]\cap\mathbb{O}\right|\le 1,$$

ie each horizontal and vertical line contains at most one X and at most one O. By identifying the edges $[c_1,c_2]\times\{d_1\}$ and $[c_1,c_2]\times\{d_2\}$ we get an *annular bordered grid diagram $G_{\boldsymbol{b}}=(\boldsymbol{\alpha},\widetilde{\boldsymbol{\beta}},\mathbb{X},\mathbb{O})$*, where $\widetilde{\boldsymbol{\beta}}$ now consists of closed curves $\widetilde{\beta}_b=\{b\}\times[d_1,d_2]/{\sim}$. Similarly, by identifying the edges $\{c_1\}\times[d_1,d_2]$ and $\{c_2\}\times[d_1,d_2]$ we get another annular bordered grid diagram $G_{\boldsymbol{a}}=(\widetilde{\boldsymbol{\alpha}},\boldsymbol{\beta},\mathbb{X},\mathbb{O})$.

A bordered grid diagram is an example of a multipointed bordered Heegaard diagram for that tangle; for the general definition of such diagrams, we refer to Section 8. In the sequel we will consider modules associated to bordered grid diagrams, annular bordered grid diagrams, and plumbings of annular bordered grid diagrams. Since all of these diagrams are "nice" in the sense of Definition 12.1, the structure maps have a combinatorial description.

4.1 Generators

For each $O \in \mathbb{O}$ fix a variable U_O, and let $C^-(G)$ be the free module generated over $\boldsymbol{k} = \mathbb{F}_2[U_O]_{O\in\mathbb{O}}$ by tuples of intersection points $\boldsymbol{x} \subset \boldsymbol{\alpha} \cap \boldsymbol{\beta}$ with the property that $|\alpha_a \cap \boldsymbol{x}| \leq 1$ and $|\beta_b \cap \boldsymbol{x}| \leq 1$. The set of generators is denoted by $\mathfrak{S}(G)$. Note that the generators naturally split into subsets $\mathfrak{S}_i(G) = \{\boldsymbol{x} : |\boldsymbol{x}| = i\}$. Then $\mathfrak{S}(G) = \bigcup_{i=1}^{\min\{c_2-c_1, d_2-d_1\}} \mathfrak{S}_i(G)$.

4.2 Inner differential

The differential can be defined by counting rectangles entirely contained in the open rectangle $(c_1, c_2) \times (d_1, d_2)$ and with boundaries on $\boldsymbol{\alpha} \cup \boldsymbol{\beta}$. For $c_1 < b_1 < b_2 < c_2$ and $d_1 < a_1 < a_2 < d_2$, the set $R = [b_1, b_2] \times [a_1, a_2]$ *is a rectangle from* $\boldsymbol{x}$ *to* $\boldsymbol{y}$ if $\boldsymbol{x} \cap R = \{(b_1, a_1), (b_2, a_2)\}$, $\boldsymbol{y} \cap R = \{(b_1, a_2), (b_2, a_1)\}$ and $\boldsymbol{x} \setminus R = \boldsymbol{y} \setminus R$. The rectangle R is empty if $\mathbb{X} \cap R = \varnothing$. The set of empty rectangles from $\boldsymbol{x}$ to $\boldsymbol{y}$ is denoted by $\mathfrak{R}_0(\boldsymbol{x}, \boldsymbol{y})$. The differential on $\boldsymbol{x} \in \mathfrak{S}(G)$ is defined by

$$\partial \boldsymbol{x} = \sum_{\boldsymbol{y}\in\mathfrak{S}(G)} \sum_{R\in\mathfrak{R}_0(\boldsymbol{x},\boldsymbol{y})} \prod_{O\in\mathbb{O}} U_O^{|R\cap O|} \boldsymbol{y}.$$

Figure 20 gives an example of the inner differential. Extend ∂ for $C^-(G)$ linearly. By the usual arguments for grid diagrams (that every domain representing a term in ∂^2 has an alternate decomposition) we have:

Proposition 4.2 *$(C^-(G), \partial)$ is a chain complex.* □

4.3 Type *AA* structures: bordered grid diagrams associated to shadows

All the structures from Section 3 have equivalent formulations via bordered grid diagrams, which will be discussed in this and the following sections. To a shadow $\mathcal{P}$ given by the quadruple (m, n, ξ, ω) we associate the following bordered grid diagram $G(\mathcal{P})$:

Definition 4.3 Define $G = G(\mathcal{P}) = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O})$ in $[-m-1, 0] \times [0, n+1] \subset \mathbb{R}^2$ as follows. For $a \in \boldsymbol{a}$ let $\alpha_a = [-m-1, 0] \times \{a\}$ and for $b \in \boldsymbol{b}$ let $\beta_b = \{-b\} \times [0, n+1]$, then let $\boldsymbol{\alpha} = \{\alpha_a\}_{a\in\boldsymbol{a}}$ and $\boldsymbol{\beta} = \{\beta_b\}_{b\in\boldsymbol{b}}$; also let $\mathbb{X} = \{(-\xi s_X, s_X)\}_{s_X\in S_{\mathbb{X}}}$ and $\mathbb{O} = \{O = (-s_O, \omega s_O)\}_{s_O\in S_{\mathbb{O}}}$.

In Figure 18 we depict the bordered grid diagrams corresponding to the shadows of Figure 5.

An equivalent way to associate a bordered grid diagram $G'(\mathcal{P})$ to the shadow $\mathcal{P}$ is to take the 180° rotation of $G(\mathcal{P})$. Thus $G'(\mathcal{P}) = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \mathbb{X}', \mathbb{O}')$ lies in the opposite

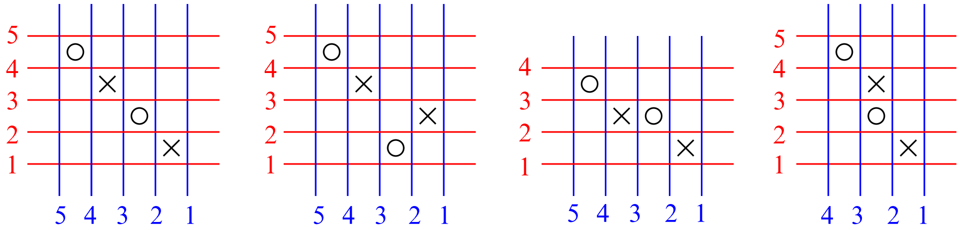


Figure 18

quadrant $[0, m+1] \times [-n-1, 0]$ with $\boldsymbol{\alpha}' = \{\alpha'_a\}_{a \in \boldsymbol{a}}$, where $\alpha'_a = [0, m+1] \times \{a'\}$, $\boldsymbol{\beta}' = \{\beta'_b\}_{b \in \boldsymbol{b}}$, where $\beta'_b = \{-b\} \times [-n-1, 0]$, $\mathbb{X} = \{(\xi s_X, -s_X)\}_{s_X \in S_{\mathbb{X}}}$ and $\mathbb{O} = \{O = (s_O, -\omega s_O)\}_{s_O \in S_{\mathbb{O}}}$. All that follows could be reformulated to $G'(\mathcal{P})$ by doing a 180° rotation to give isomorphic chain complexes and type AA structures to those for $G(\mathcal{P})$.

4.3.1 Tangles associated to $G(\mathcal{P})$ Let us complete $G(\mathcal{P})$ with some extra basepoints

$$\mathbb{X}_\partial = \{(-s, 0) : s \in S_{\mathbb{O}} \setminus T_{\mathbb{X}}\} \cup \{(0, s) : s \in T_{\mathbb{O}} \setminus S_{\mathbb{X}}\},$$
$$\mathbb{O}_\partial = \{(-s, 0) : s \in T_{\mathbb{X}} \setminus S_{\mathbb{O}}\} \cup \{(0, s) : s \in S_{\mathbb{X}} \setminus T_{\mathbb{O}}\}.$$

Then define the associated tangle $\mathcal{T}(G)$ just like one would for a closed grid diagram: connect the points $\mathbb{X} \cup \mathbb{X}_\partial$ to $\mathbb{O} \cup \mathbb{O}_\partial$ horizontally and $\mathbb{O} \cup \mathbb{O}_\partial$ to $\mathbb{X} \cup \mathbb{X}_\partial$ vertically so that vertical strands cross over horizontal strands. Then, after smoothing, $\mathcal{T}(G)$ is a tangle projection in $[-m-1, 0] \times [0, n+1]$ with boundary

$$\partial^0 = (\mathbb{X}_\partial - \mathbb{O}_\partial) \cap [-m-1, 0] \times \{0\} \quad \text{and} \quad \partial^1 = (\mathbb{X}_\partial - \mathbb{O}_\partial) \cap \{1\} \times [0, n+1].$$

See Figure 19 for some examples. Note that this tangle can be easily identified (by, for example, using polar coordinates and mapping $(r, \vartheta) \in [-m-1, 0] \times [0, n+1]$ to $(2(\pi - \vartheta)/\pi, r) \in I \times \mathbb{R}$) with a tangle in $I \times \mathbb{R}$, which we will call $\mathcal{T}(G)$ as well.

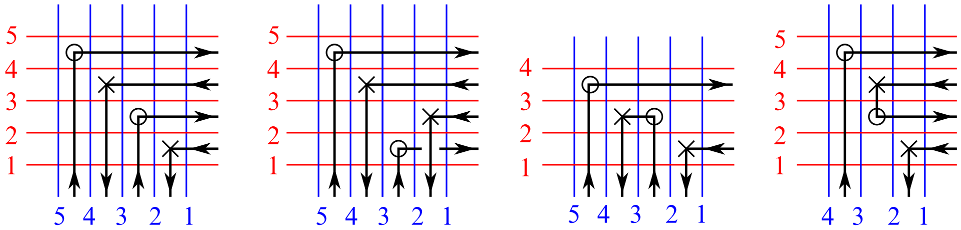


Figure 19: The tangles associated to the bordered grid diagrams of Figure 18

Proposition 4.4 *Let $\mathcal{P}$ be a shadow. Then for $G = G(\mathcal{P})$ the tangles $\mathcal{T}(\mathcal{P})$ and $\mathcal{T}(G)$ are isotopic relative to the boundary.*

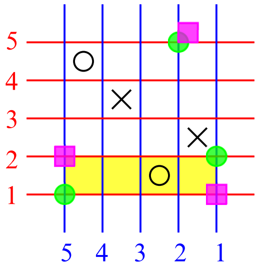


Figure 20: The inner differential for bordered grid diagrams. The generator $\boldsymbol{x}$ denoted by (green) dots corresponds to the first strand diagram of Figure 9. The only empty rectangle (in yellow) starting from $\boldsymbol{x}$ connects it to the generator $\boldsymbol{y}$ denoted by a (pink) square. The latter generator corresponds to the last strand diagram in Figure 9. The rectangle passes through the $\mathbb{O}$ marking O_1. Thus $\partial\boldsymbol{x} = U_1\boldsymbol{y}$.

Proof Let $\mathcal{T}(G) \subset I \times \mathbb{R}$ be the tangle (projection) associated to $G = G(\mathcal{P})$. If $p \in \mathcal{T}(G)$ has a vertical tangency, then depending on whether $\mathcal{T}(G)$ near p is to the right (or left) from this tangency, it is coming from an $X = (-\xi s_X, s_X)$ and an $O = (-s_O, \omega s_O)$ in the same horizontal (or vertical) line of the grid, thus $s_X = \omega s_O$ (or $\xi s_X = s_O$). If for example $s_X = \omega s_O$, then there is no further X or O in the same horizontal line of the grid, thus the point with the vertical tangency can be isotoped to $(0, s_O) \in I \times \mathbb{R}$ without altering or crossing other parts of the tangle. Do this with every point with vertical tangency and notice that the resulting tangle is $\mathcal{T}(\mathcal{P})$. □

4.3.2 Generators Recall that $C^-(G)$ is the free module generated over $\boldsymbol{k}$ by the tuples of intersection points $\boldsymbol{x} = (\alpha_{\phi s} \cap \beta_s)_{s \in S}$, where $S \subset \boldsymbol{b}$ and $\phi\colon S \to \boldsymbol{a}$ is an injection with image $T = \phi(S)$. There is a one-to-one correspondence between $\mathfrak{S}(\mathcal{P})$ and $\mathfrak{S}(G)$ given by associating $\boldsymbol{x} = (\alpha_{\phi s} \cap \beta_s)_{s \in S} \in \mathfrak{S}(G)$ to $(S, T, \phi) \in \mathfrak{S}(\mathcal{P})$.

4.3.3 Inner differential The differential of Section 4.2 translates to the following. If $s_1 < s_2$ and $t_1 < t_2$, and $\boldsymbol{x} = (\alpha_{\phi s} \cap \beta_s)_{s \in S}$ and $\boldsymbol{y} = (\alpha_{\phi^{(s_1,s_2)} s} \cap \beta_s)_{s \in S}$, where $s_1, s_2 \in S$ and $\phi\colon S \to T$ satisfies $\phi s_1 = t_2$ and $\phi s_2 = t_1$, then $R = [-s_2, -s_1] \times [t_1, t_2]$ is a rectangle from $\boldsymbol{x}$ to $\boldsymbol{y}$. Note that then automatically $(s_1, s_2) \in \mathrm{Inv}(\phi)$.

Thus, with the above definition of the inner differential:

Proposition 4.5 *The chain complexes $(C^-(G), \partial)$ and $(C^-(\mathcal{P}), \partial)$ are isomorphic.*

Moreover, if R is a rectangle from $\boldsymbol{x} = (\alpha_{\phi s} \cap \beta_s)_{s \in S}$ to $\boldsymbol{y} = (\alpha_{\phi^{(s_1,s_2)} s} \cap \beta_s)_{s \in S}$, then:

(1) $A(S, T, \phi) - A(S, T, \phi^{(s_1,s_2)}) = |R \cap \mathbb{X}| - |R \cap \mathbb{O}|$.

(2) *If $R \in \mathfrak{R}_0(\boldsymbol{x}, \boldsymbol{y})$ then* $M(S, T, \phi) - M(S, T, \phi^{(s_1,s_2)}) = 1 - 2|R \cap \mathbb{O}|$.

Proof If $(s_1,s_2)\in\operatorname{Inv}(\phi)$ then $R=[-s_2,-s_1]\times[\phi s_2,\phi s_1]$ defines a rectangle in $[-m+1,0]\times[0,n+1]$. The statement follows from the following three equations:

$$\begin{aligned}|R\cap \boldsymbol{x}| &= \bigl|\{(-s,\phi s): s\in S,\ -s_2<-s<-s_1 \text{ and } \phi s_2<\phi s<\phi s_1\}\bigr|\\ &= \bigl|\operatorname{Inv}\phi\setminus\operatorname{Inv}\phi^{(s_1,s_2)}\cup\{(s_1,s_2)\}\bigr|,\\ |R\cap\mathbb{X}| &= \bigl|\{(-\xi s_X,s_X): s_X\in S_{\mathbb{X}},\ -s_2<-\xi s_X<-s_1 \text{ and } \phi s_2<s_X<\phi s_1\}\bigr|\\ &= \bigl|\operatorname{Inv}(\phi,\xi^{-1})\setminus\operatorname{Inv}(\phi^{(s_1,s_2)},\xi^{-1})\bigr|,\\ |R\cap\mathbb{O}| &= \bigl|\{(-s_O,\omega s_O): s_O\in S_{\mathbb{O}},\ -s_2<-s_O<-s_1 \text{ and } \phi s_2<\omega s_O<\phi s_1\}\bigr|\\ &= \bigl|\operatorname{Inv}(\phi,\omega|_{s_O})\setminus\operatorname{Inv}(\phi^{(s_1,s_2)}\omega|_{s_O})\bigr|. \qquad\square\end{aligned}$$

4.3.4 Type A structures The left and right algebra actions by $\mathcal{A}(\mathcal{E}_L)$ and $\mathcal{A}(\mathcal{E}_R)$ are defined by counting sets of partial rectangles as follows. First, we will describe the right action. The left action, as will be spelled out later, is similar. For the action of $\mathcal{A}(\mathcal{E}_R)$ we consider sets of partial rectangles that intersect the left and right boundaries $\{-m-1,0\}\times(0,n+1)$. We consider the following two types of partial rectangles depending on whether the rectangle intersects the left or the right boundary edge:

- $H=[-s_1,0]\times[t_1,t_2]$ with $t_1<t_2$, or
- $H=[-m-1,-s_2]\times[t_1,t_2]$ with $t_1<t_2$,

where $s_i\in\boldsymbol{b}$ and $t_i\in\boldsymbol{a}$.

Now fix $S\subset\boldsymbol{b}$ and generators $\boldsymbol{x}=(\alpha_{\phi s}\cap\beta_s)_{s\in S}$ and $\boldsymbol{y}=(\alpha_{\phi' s}\cap\beta_s)_{s\in S}$. Let $r=(\phi(S),\phi'(S),\phi'\circ\phi^{-1})\in\mathcal{A}(\mathcal{E}_R)$. Suppose that $\boldsymbol{H}=\{H_1,\dots,H_l\}$ is a set of partial rectangles of the above two types. We say that $\boldsymbol{H}$ *connects* $\boldsymbol{x}$ *and* r *to* $\boldsymbol{y}$ if for the rectangles in $\boldsymbol{H}$, all bottom-left and top-right corners that are in the interior of G are distinct points and form the set $\boldsymbol{x}\setminus(\boldsymbol{x}\cap\boldsymbol{y})$, and all bottom-right and top-left corners that are in the interior of G are distinct points and form the set $\boldsymbol{y}\setminus(\boldsymbol{x}\cap\boldsymbol{y})$. We say that $\boldsymbol{H}$ is *allowed* if for each $H_i\in\boldsymbol{H}$ we have $H_i\cap\mathbb{X}=\varnothing$ and $H_i\cap(\boldsymbol{x}\cap\boldsymbol{y})=\varnothing$, no partial rectangle in $\boldsymbol{H}$ is completely contained in another rectangle in $\boldsymbol{H}$, and no two partial rectangles touching opposite boundary edges have overlapping interiors. See Figure 21. Note that when $\boldsymbol{H}$ consists of only one partial rectangle H, this is equivalent to the condition $\operatorname{Int}H\cap\mathbb{X}=\operatorname{Int}H\cap\boldsymbol{x}=\varnothing$.

Note that for a fixed generator $\boldsymbol{x}$ and algebra generator r, there is at most one $\boldsymbol{y}$ and at most one $\boldsymbol{H}$ as above. Thus, we can define the action of r on $\boldsymbol{x}$ as follows. If there is no set of empty partial rectangles from $\boldsymbol{x}$ and r to any $\boldsymbol{y}$, then $\boldsymbol{x}\cdot r=0$. Otherwise, let $\boldsymbol{H}$ and $\boldsymbol{y}$ be the unique objects such that $\boldsymbol{H}$ is an allowed set of partial rectangles

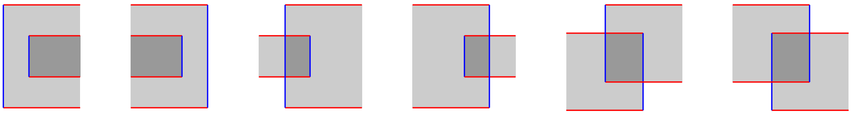


Figure 21: Forbidden pairs of partial rectangles. A set of rectangles $\boldsymbol{H}$ is allowed if no partial rectangle in it contains points in $\mathbb{X}$ or $\boldsymbol{x} \cap \boldsymbol{y}$, and no two partial rectangles in it are in relative configuration as depicted here.

connecting $\boldsymbol{x}$ and r to $\boldsymbol{y}$. Then

$$\boldsymbol{x} \cdot r = \prod_{s_O \in T_{\mathbb{O}}} U_O^{|O \cap \boldsymbol{H}|} \boldsymbol{y},$$

where $O \cap \boldsymbol{H} = \bigcup (O \cap H_i)$.

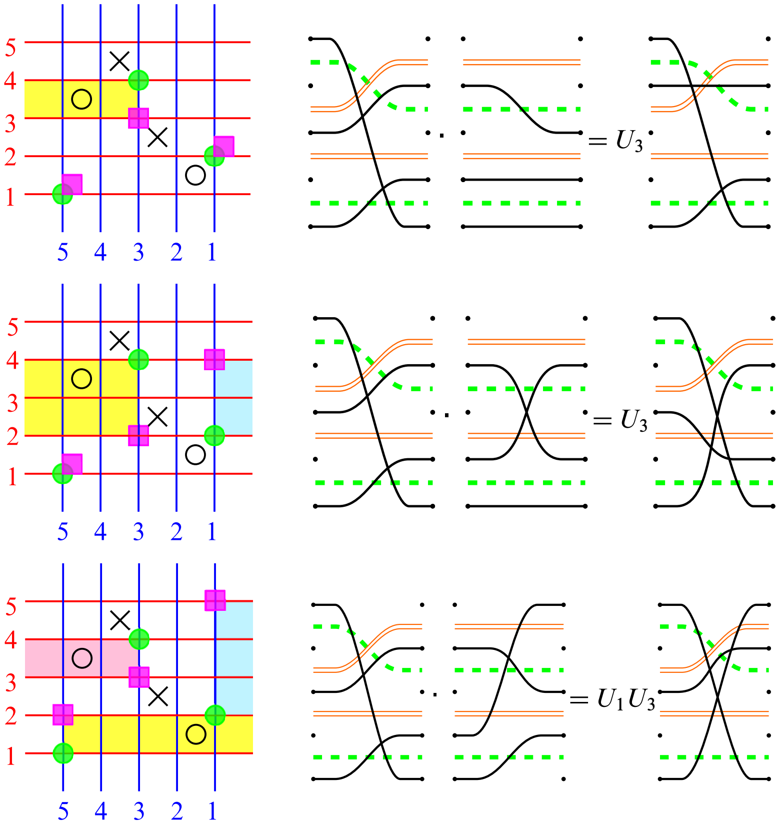


Figure 22: Examples of the right type A action. Left: examples of allowed sets of partial rectangles for the right action, starting at the generator formed by the green dots. Right: the corresponding right multiplications, viewed as concatenations of strand diagrams.

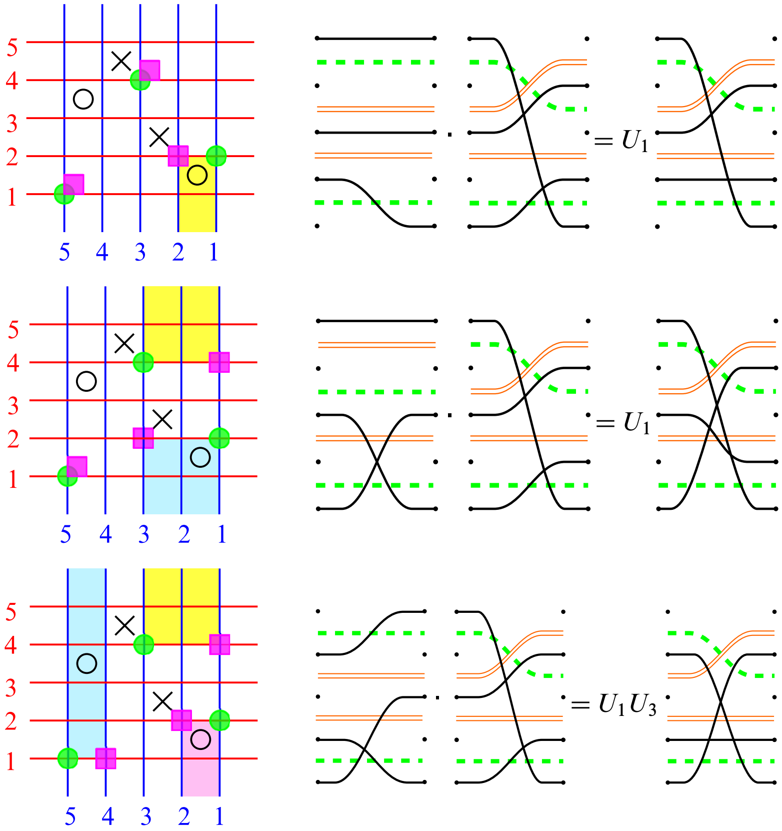


Figure 23: Examples of the left type A action. Left: examples of allowed sets of partial rectangles for the left action, starting at the generator formed by the green dots. Right: the corresponding left multiplications, viewed as concatenations of strand diagrams.

See Figure 22 for examples of the type A multiplication.

The left action can be similarly defined using partial rectangles touching the top or bottom parts of the boundary $(-m-1,0)\times\{0,n+1\}$ or by rotating the rectangles by 90°. See Figure 23.

Definition 4.6 With the above notation, let $\mathit{CATA}^-(G)$ be the left–right type AA bimodule $(C^-(\mathcal{P}),\{m_{i,1,j}\})$ over $\mathcal{A}(\mathcal{E}_L)$ and $\mathcal{A}(\mathcal{E}_R)$, where

$$m_{i,1,j}\colon \mathcal{A}(\mathcal{E}_L)^{\otimes i}\otimes C^-(\mathcal{P})\otimes \mathcal{A}(\mathcal{E}_R)^{\otimes j}\to C^-(\mathcal{P})$$

with $m_{i,1,j}=0$ when $i>1$ or $j>1$, and the nonzero maps are given by

$$m_{0,1,0}(f)=\partial f,\quad m_{1,1,0}(a_L\otimes f)=a_L\cdot f,\quad m_{0,1,1}(f\otimes a_R)=f\cdot a_R.$$

It is not immediate to see that the above definition indeed gives a type AA bimodule, but the next proposition says that it is isomorphic to $\mathit{CATA}^-(\mathcal{P})$, which, by Theorem 3.29, is a type AA structure.

Proposition 4.7 *Let $\mathcal{P}$ be a shadow and let $G = G(\mathcal{P})$. Then the one-to-one correspondence between the generators gives rise to an isomorphism of the structures $\mathit{CATA}^-(\mathcal{P})$ and $\mathit{CATA}^-(G(\mathcal{P}))$.*

Proof Observe that $\boldsymbol{H}$ connects $\boldsymbol{x}$ and r to $\boldsymbol{y}$ exactly when the strand diagrams corresponding to $\boldsymbol{x}$ and r can be concatenated. The result of the concatenation is the strand diagram corresponding to $\boldsymbol{y}$ when $\boldsymbol{H}$ is allowed, and zero otherwise. Indeed, the obstructions to $\boldsymbol{H}$ being allowed correspond to the Reidemeister II relations involving black and orange strands. Similarly, the count $O \cap \boldsymbol{H}$ corresponds to the count n_O. □

4.4 Type *DD* structures: bordered grid diagrams associated to mirror-shadows

The bordered grid diagram $G^*(\mathcal{P}^*)$ associated to the mirror-shadow $\mathcal{P}^*$ is the mirror of $G(\mathcal{P})$ with respect to a vertical axis.

Definition 4.8 $G^* = G^*(\mathcal{P}^*) = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}) \subset [0, m+1] \times [0, n+1] \subset \mathbb{R}^2$ as follows. For $a \in \boldsymbol{a}$ let $\alpha_a = [0, m+1] \times \{a\}$ and for $b \in \boldsymbol{b}$ let $\beta_b = \{b\} \times [0, n+1]$, then let $\boldsymbol{\alpha} = \{\alpha_a\}_{a \in \boldsymbol{a}}$ and $\boldsymbol{\beta} = \{\beta_b\}_{b \in \boldsymbol{b}}$. Also let $\mathbb{X} = \{(\xi s_X, s_X)\}_{s_X \in S_{\mathbb{X}}}$ and $\mathbb{O} = \{O = (s_O, \omega s_O)\}_{s_O \in S_{\mathbb{O}}}$.

Figure 24 shows the bordered grid diagrams corresponding to the mirror-shadows of Figure 12.

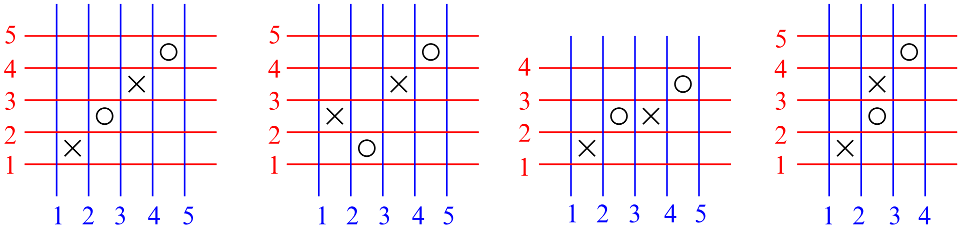


Figure 24

By mirroring $G(\mathcal{P})$ with respect to the horizontal axis instead, we get a bordered grid diagram $(G^*)'(\mathcal{P}^*)$ equivalent to $G^*(\mathcal{P}^*)$.

As in the case for $G(\mathcal{P})$, the generators $\mathfrak{S}(G^*)$ are tuples of intersection points, and similarly there is a one-to-one correspondence between $\mathfrak{S}(G^*)$ and $\mathfrak{S}(\mathcal{P}^*)$ identifying

$(S,T,\phi)^*$ with the set of intersection points $\boldsymbol{x}=(\alpha_{\phi s}\cap\beta_s)_{s\in S}$. The differential ∂ is again given by counting empty rectangles.

Proposition 4.9 *The chain complexes $(C^-(\mathcal{P}^*),\partial^*)$ and $(C^-(G^*),\partial)$ are isomorphic. Moreover if R is a rectangle from $\boldsymbol{x}=(\alpha_{\phi^{(s_1,s_2)}s}\cap\beta_s)_{s\in S}$ to $\boldsymbol{y}=(\alpha_{\phi s}\cap\beta_s)_{s\in S}$ then*

(1) $A(S,T,\phi)-A(S,T,\phi^{(s_1,s_2)})=|R\cap\mathbb{X}|-|R\cap\mathbb{O}|$;

(2) *if* $R\in\mathfrak{R}_0(\boldsymbol{x},\boldsymbol{y})$ *then* $M(S,T,\phi)-M(S,T,(\phi)^{(s_1,s_2)})=1-2|R\cap\mathbb{O}|$.

Proof This is essentially the same as the proof of Proposition 4.5. □

Associate to the bordered grid diagram G^* the tangle $\mathcal{T}^*(G^*)$ that is the mirror of $\mathcal{T}(G)$, again with respect to the vertical axis.

4.4.1 Type *D* maps Define a bimodule structure ${}_{\mathcal{I}(\mathcal{A}(\mathcal{E}^L))}C^-(G^*)_{\mathcal{I}(\mathcal{A}(\mathcal{E}^R))}$ using the one-to-one correspondence between $\mathfrak{S}(G^*)$ and $\mathfrak{S}(\mathcal{P}^*)$. In other words, if the correspondence maps $\boldsymbol{x}\in\mathfrak{S}(G^*)$ to $f^*\in\mathfrak{S}(\mathcal{P}^*)$, then define $\iota\cdot\boldsymbol{x}\cdot\iota'=\iota\cdot f^*\cdot\iota'$. For such a pair $\boldsymbol{x}$ and f^*, define $\iota^L(\boldsymbol{x})=\iota^L(f^*)$ and $\iota^R(\boldsymbol{x})=\iota^R(f^*)$. Similar to the type A maps, we define left and right type D maps

$$\delta^L\colon C^-(G^*)\to\mathcal{A}(\mathcal{E}^L)\otimes C^-(G^*),\quad \delta^R\colon C^-(G^*)\to C^-(G^*)\otimes\mathcal{A}(\mathcal{E}^R),$$

also by counting partial rectangles. In the following we describe the left type D map δ^L in detail.

Let $\boldsymbol{x}=(\alpha_{\phi s}\cap\beta_s)_{s\in S}$ be a generator. We define a map ∂^L by counting partial rectangles that intersect the left and/or right boundaries $\{0,m+1\}\times[0,n+1]$. We distinguish four types of partial rectangles as follows:

- $H=[0,s_1]\times[t_1,t_2]$, where $s_1\in S$, $t_1<t_2$ and $t_2=\phi s_1, t_1\notin\phi(S)$. Let $T_1=\phi(S)^c$, $T_2=\phi(S)^c\setminus\{t_1\}\cup\{t_2\}$, and define $\rho\colon T_1\to T_2$ by $\rho t_1=t_2$ and $\rho|_{T_1\setminus\{t_1\}}=\mathrm{id}_{T_1\setminus\{t_1\}}$. Let $r=(T_1,T_2,\rho)\in\mathcal{A}(\mathcal{E}^L)$. Let $\boldsymbol{y}$ be the set of intersection points $\boldsymbol{x}\setminus\{(s,t_2)\}\cup\{(s,t_1)\}$.
- $H=[s_2,m+1]\times[t_1,t_2]$, where $s_2\in S$, $t_1<t_2$ and $t_1=\phi s_2, t_2\notin\phi(S)$. Let $T_2=\phi(S)^c$, $T_1=\phi(S)^c\setminus\{t_2\}\cup\{t_1\}$, and define $\rho\colon T_2\to T_1$ by $\rho t_2=t_1$ and $\rho|_{T_2\setminus\{t_2\}}=\mathrm{id}_{T_2\setminus\{t_2\}}$. Let $r=(T_2,T_1,\rho)\in\mathcal{A}(\mathcal{E}^L)$ and $\boldsymbol{y}=\boldsymbol{x}\setminus\{(s,t_1)\}\cup\{(s,t_2)\}$.
- $H=[0,m+1]\times[t_1,t_2]$, where $t_1,t_2\notin\phi(S)$ and $t_1<t_2$. Let $\rho\colon\phi(S)^c\to\phi(S)^c$ be given by $(t_1t_2)\circ\mathrm{id}_{\phi(S)^c}$ and let $r=(\phi(S)^c,\phi(S)^c,\rho)\in\mathcal{A}(\mathcal{E}^L)$. Let $\boldsymbol{y}=\boldsymbol{x}$.
- $H=([0,s_1]\cup[s_2,m+1])\times[t_1,t_2]$, where $s_1<s_2$, $t_1<t_2$ and $t_1=\phi s_2, t_2=\phi s_1$. Let $r=(S^c,S^c,\mathrm{id}_{S^c})$ and $\boldsymbol{y}=(\alpha_{((t_1t_2)\circ\phi)s}\cap\beta_s)_{s\in S}$.

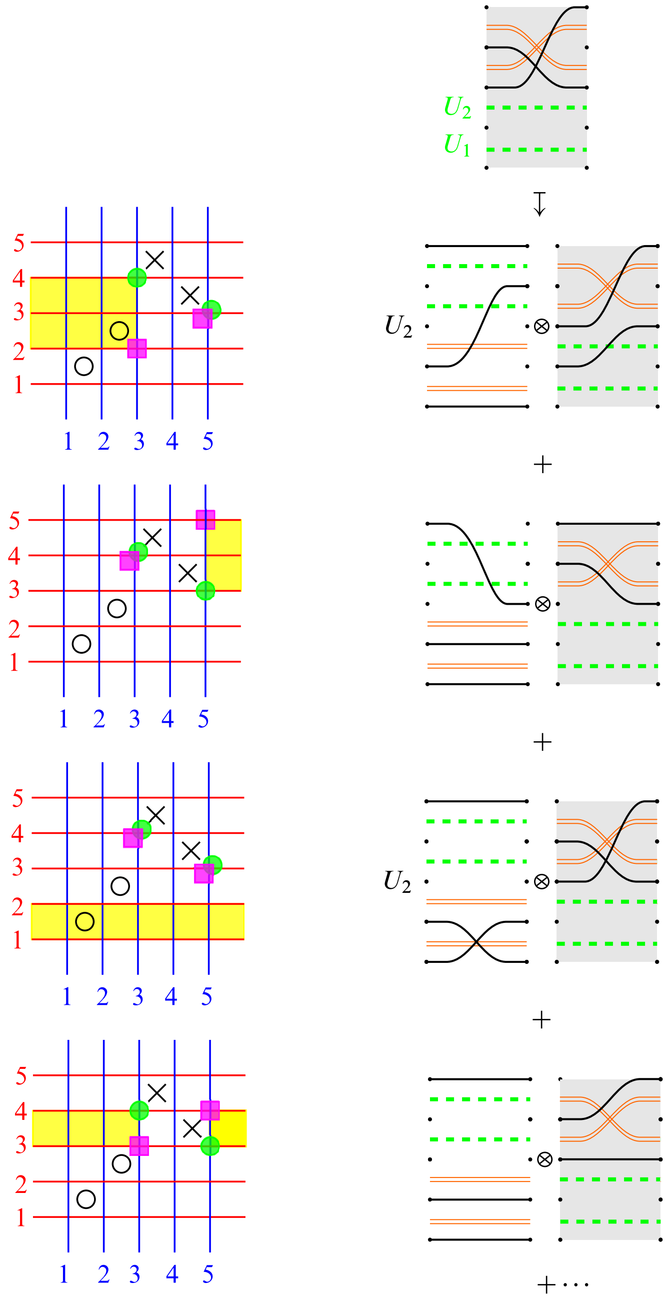


Figure 25: The four types of rectangles corresponding to the map ∂^L. Left: examples of the four types of rectangles for ∂^L applied to the generator formed by the green dots. Right: the respective terms of δ^L applied to the strand diagram corresponding to the green dots.

We say that the partial rectangle H *connects* $\boldsymbol{x}$ and r to $\boldsymbol{y}$, and for $O=(s_O,t_O)\in\mathbb{O}$ set $n_{t_O}(H)=|O\cap H|$. In the first three cases we say that H *connects* $\boldsymbol{x}$ *and* r *to* $\boldsymbol{y}$. H is *empty* if $H\cap\mathbb{X}=H\cap\boldsymbol{x}=\varnothing$. In the fourth case there is an extra condition on H being empty: we require that for the projection $\pi_2\colon(s,t)\mapsto t$ the images $\pi_2(\mathbb{X}\cap[s_1,s_2]\times[t_1,t_2])$ and $\pi_2(\boldsymbol{x}\cap[s_1,s_2]\times[t_1,t_2])$ are precisely $[t_1,t_2]\cap\boldsymbol{a}_{1/2}$ and $[t_1,t_2]\cap\boldsymbol{a}$. For $t_O^c\in([t_1,t_2]\cap\boldsymbol{a}_{1/2})\setminus T_{\mathbb{O}}$, let $n_{t_O^c}(H)=1$.

Given $\boldsymbol{x}$, $\boldsymbol{y}$ and r, let $\mathcal{H}_0(\boldsymbol{x},\boldsymbol{y},r)$ denote the set of empty partial rectangles connecting $\boldsymbol{x}$ and r to $\boldsymbol{y}$ (note that that set is either empty or consists of one partial rectangle). Define

$$\partial^L\boldsymbol{x}=\sum_{\substack{\boldsymbol{y}\in\mathfrak{S}(G^*)\\ r\in\mathfrak{S}(\mathcal{E}^L)}}\ \sum_{H\in\mathcal{H}_0(\boldsymbol{x},\boldsymbol{y},r)} r\otimes\prod_{t_O\in\boldsymbol{a}_{1/2}}U_O^{n_{t_O}(H)}\boldsymbol{y}.$$

See Figure 25 for an example of ∂^L.

Then the left type D map is defined on generators by

$$\delta^L\boldsymbol{x}=\iota^L(\boldsymbol{x})\otimes\partial\boldsymbol{x}+\partial^L\boldsymbol{x}.$$

In other words, δ^L is defined by counting empty rectangles in the interior of the grid, as well as empty rectangles that touch the left and/or right boundary of the grid.

The right type D map δ^R can be defined in a similar way as the sum $\delta^R=\partial\otimes\iota^R+\partial^R$ using a map ∂^R that counts partial rectangles that intersect the top and bottom boundary of $[0,n+1]\times[0,m+1]$.

The left and the right type D maps can be merged together to define a type DD map by counting all empty rectangles, interior and partial.

Definition 4.10 For $G^*=G^*(\mathcal{P}^*)$ define $\mathit{CDTD}^-(G^*)$ be the left–right type DD structure $(C^-(G^*),\delta^1\})$ over $\mathcal{A}(\mathcal{E}^L)$ and $\mathcal{A}(\mathcal{E}^R)$, where

$$\delta^1\colon C^-(G^*)\to\mathcal{A}(\mathcal{E}^L)\otimes C^-(G^*)\otimes\mathcal{A}(\mathcal{E}^R)$$

is defined via

$$\delta^1(\boldsymbol{x})=\iota^L(\boldsymbol{x})\otimes\partial^R(\boldsymbol{x})+\iota^L(\boldsymbol{x})\otimes\partial(\boldsymbol{x})\otimes\iota^R(\boldsymbol{x})+\partial^L(\boldsymbol{x})\otimes\iota^R(\boldsymbol{x}).$$

Proposition 4.11 *For $G^*=G^*(\mathcal{P}^*)$ the one-to-one correspondence between generators gives rise to an isomorphism between $\mathit{CDTD}^-(G^*)$ and $\mathit{CDTD}^-(\mathcal{P}^*)$.*

While Proposition 4.11 and the fact that $\mathit{CDTD}^-(G^*)$ satisfies the type DD identities could be proven directly, we will choose a longer way. First we understand how to glue bordered grid diagrams. Then, as is explained later, both statements are consequences of Propositions 4.12 and 4.13.

4.5 Gluing bordered grid diagrams

Suppose that $G_1 = G(\mathcal{P}_1) = (\boldsymbol{\alpha}^1, \boldsymbol{\beta}^1, \mathbb{X}_1, \mathbb{O}_1)$ and $G_2^* = G^*(\mathcal{P}_2^*) = (\boldsymbol{\alpha}^2, \boldsymbol{\beta}^2, \mathbb{X}_2, \mathbb{O}_2)$, where $\mathcal{P}_1$ and $\mathcal{P}_2^*$ have a well-defined wedge product. This means that $n_1 = n_2$, so $G = G_1 \cup G_2^*/{\sim} \subset [-m_1-1, m_2+1]\times[0, n_1+1]/{\sim}$ is a bordered grid diagram where the edges $\{-m_1-1\}\times[0, n_1+1]$ and $\{m_2+1\}\times[0, n_2+1]$ are identified. Here $\boldsymbol{\beta} = \boldsymbol{\beta}^1 \cup \boldsymbol{\beta}^2$, and the $\boldsymbol{\alpha}$–arcs are glued to form the new circles $\widetilde{\alpha}_a = [-m_1 - 1, m_2 + 1] \times \{a\}/{\sim}$. Similarly, $\mathbb{X} = \mathbb{X}_1 \cup \mathbb{X}_2$ and $\mathbb{O} = \mathbb{O}_1 \cup \mathbb{O}_2$. Note that since $\mathcal{P}_1$ and $\mathcal{P}_2^*$ have a well-defined wedge product every annulus between the alpha circles $\widetilde{\alpha}_a$ and $\widetilde{\alpha}_{a+1}$ contains exactly one element of $\mathbb{X}$ and one element of $\mathbb{O}$.

Informally, we glued G_2^* to the right of G_1 and identified the left and right edges of the resulting rectangle to obtain an annulus. Alternatively, one can shift coordinates in $\mathbb{R}^2$ and view the annulus by placing G_2^* to the left of G_1 and then identifying the left and right edges of the resulting rectangle to obtain an annulus. Abstractly, the annulus is simply the result of identifying each "$\boldsymbol{\alpha}$–boundary edge" of one grid with an $\boldsymbol{\alpha}$–boundary edge of the other grid, so that the labels on the $\boldsymbol{\alpha}$–curves match up, and the gluing respects the orientation on the two surfaces of the grids.

We define $C^-(G)$ to be the free module generated over $\mathbb{F}_2[U_O]_{O\in\mathbb{O}}$ by tuples of intersection points $\boldsymbol{x} \subset \widetilde{\boldsymbol{\alpha}} \cap \boldsymbol{\beta}$ such that there is one point on each $\widetilde{\boldsymbol{\alpha}}$–circle, and at most one point on each $\boldsymbol{\beta}$–arc. Observe that the generating set is precisely

$$\mathfrak{S}(G) = \{\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathfrak{S}(G_1)\times\mathfrak{S}(G_2^*) : |\boldsymbol{x}_1 \cap \alpha_a^1| = 1 \text{ if and only if } |\boldsymbol{x}_2 \cap \alpha_a^2| = 0\}.$$

Define a map ∂ on $\mathfrak{S}(G)$ by counting empty rectangles in the interior of G (note that rectangles may cross the newly identified edges), and extend linearly to all of $C^-(G)$. By standard grid diagram arguments, ∂ is a differential. See Figure 26 for an example of the identification where G_1 is drawn to the right.

Now there is a one-to-one correspondence between generators of $\mathcal{P}_1 \wedge \mathcal{P}_2^*$ and $\mathfrak{S}(G)$ given by mapping $(S_1, T_1, \phi_1)\otimes(S_2, T_2, \phi_2)^*$ to $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, where $\boldsymbol{x}_1 = (\alpha^1_{\phi_1 s} \cap \beta^1_s)_{s\in S_1}$ and $\boldsymbol{x}_2 = (\alpha^2_{\phi_2 s} \cap \beta^2_s)_{s\in S_2}$. We show below that under this correspondence the differential ∂ on $C^-(G)$ agrees with $\partial_\wedge$ on $C^-(\mathcal{P}_1 \wedge \mathcal{P}_2^*)$. In particular, it follows that $(C^-(\mathcal{P}_1 \wedge \mathcal{P}_2^*), \partial_\wedge)$ is a chain complex, as is stated in Proposition 3.23.

Proposition 4.12 *The structures $(C^-(\mathcal{P}_1 \wedge \mathcal{P}_2^*), \partial_\wedge)$ and $(C^-(G), \partial)$ are isomorphic.*

Proof Let $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ be the generator of $(C^-(G), \partial)$ corresponding to the element $f = f_1 \otimes f_2^* = (S_1, T_1, \phi_1) \otimes (S_2, T_2, \phi_2)^*$ in $(C^-(\mathcal{P}_1 \wedge \mathcal{P}_2^*), \partial_\wedge)$. Recall that the differential $\partial_\wedge$ of $f_1 \otimes f_2^*$ is given by the formula

$$\partial_\wedge(f_1 \otimes f_2^*) = \partial(f_1) \otimes f_2^* + f_1 \otimes \partial^*(f_2^*) + \partial_{\mathrm{mix}}(f_1 \otimes f_2^*),$$

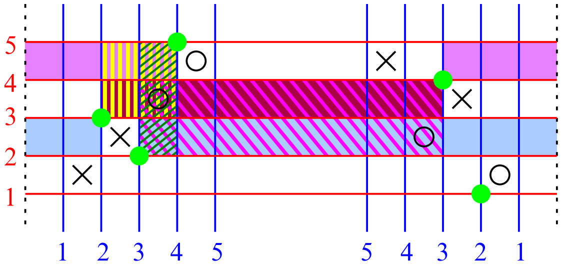


Figure 26: The differential on the annular grid diagram associated to the example of Figure 15. The dashed lines on the right- and left-hand side are glued together. The green dots corresponds to the strand diagram on the left-hand side of Figure 15 and the six rectangles to the nonzero terms in the differential of that diagram.

while the differential of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ in $(C^-(G), \partial)$ is given by counting rectangles. Suppose that the rectangle R contributes to the differential ∂. Then, depending on the position of R, the result corresponds to different components of the differential $\partial_\wedge$ as follows:

- If R is entirely contained in G_1, then R corresponds to a term of $\partial(f_1) \otimes f_2^*$.
- If R is entirely contained in G_2^*, then R corresponds to a term of $f_1 \otimes \partial^*(f_2^*)$.
- If R intersects both G_1 and G_2^*, each in a connected component, then R intersects exactly one of the vertical lines $\{0\}\times[0,n_1+1]$ or $\{-m_1-1\}\times[0,n_1+1] \sim \{m_2+1\}\times[0,n_1+1]$. In the first case $R\cap\{0\}\times[0,n_1+1]=\{0\}\times[p,q]$ for some $p<q$, and in the second case $R\cap\{m_1-1\}\times[0,n_1+1]=\{m_1-1\}\times[q,p]$ for some $q<p$. Then $(p,q)\in S_1\times S_2$ is an exchangeable pair, and R corresponds to a term of ∂_{mix}.
- If R intersects both G_1 and G_2^*, and $R\cap G_1$ has one component while $R\cap G_2^*$ has two components, then let $R\cap\{0\}\times[0,n_1+1]=\{0\}\times[p,q]$ for some $p<q$. The pair $(p,q)\subset S_2$ is exchangeable and R corresponds to a term of ∂_{mix}.
- Similarly if R intersects both G_1 and G_2^* and $R\cap G_1$ has two components while $R\cap G_2^*$ has one component, then $R\cap\{0\}\times[0,n_1+1]=\{0\}\times[p,q]$ for some $p<q$. The pair $(p,q)\subset S_1$ is exchangeable and R corresponds to a term of ∂_{mix}.

Conversely, any term of $\partial_\wedge(f_1\otimes f_2^*)$ appears in the above list, thus the statement is proved. □

Note that the writeup of the above proof uses coordinates for the case when G_1 is viewed sitting to the left of G_2^*.

Similarly, if $G_1^* = G^*(\mathcal{P}_1^*)$ and $G_2' = G_2'(\mathcal{P}_2)$, then we can glue $(G_1^*)'$ to G_2 along the x–axis, ie place G_2 above G_1^*, and identify the resulting horizontal boundaries. Alternatively, we can view the annulus by placing G_2 below G_1^* and then identifying the horizontal edges of the resulting rectangle. Abstractly, the annulus is the result of identifying $\boldsymbol{\beta}$–boundary edges. For the resulting annular grid diagram, we define a chain complex $(C^-(G), \partial)$, where again generators over $\mathbb{F}_2[U_O]_{O \in \mathbb{O}_1 \cup \mathbb{O}_2}$ are tuples of intersection points with exactly one point on each $\widetilde{\boldsymbol{\beta}}$–circle and at most one point on each $\boldsymbol{\alpha}$–arc, and the differential counts empty rectangles. Once again we have:

Proposition 4.13 *The structures* $(C^-(\mathcal{P}_1^* \wedge \mathcal{P}_2), \partial_\wedge)$ *and* $(C^-(G), \partial)$ *are isomorphic.*

Proof The proof is analogous to that of Proposition 4.12. □

As an immediate consequence we have:

Proof of Propositions 3.22 and 3.23 Both statements follow from Propositions 4.12 and 4.13 for $C^-(\mathcal{E}^L(\mathcal{P}^*) \wedge \mathcal{P}^*)$ and $C^-(\mathcal{P}^* \wedge \mathcal{E}^R(\mathcal{P}^*))$, along with the fact that ∂ is a differential for the corresponding grid diagrams. □

In general, suppose we have an alternating sequence of shadows and mirror-shadows $\boldsymbol{\mathcal{P}} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ with well-defined consecutive wedge products. We can glue the grid diagrams $G^\circ(\mathcal{P}_1^\circ), \dots, G^\circ(\mathcal{P}_p^\circ)$ by alternating the gluing along horizontal or vertical edges to obtain the nice bordered Heegaard diagram G on plumbings of annuli. We can associate a tangle to G, which is simply the concatenation of $\mathcal{T}^\circ(G^\circ(\mathcal{P}_1^\circ)), \dots, \mathcal{T}^\circ(G^\circ(\mathcal{P}_p^\circ))$. See, for example, Figure 29.

Let $C^-(G)$ be the free module over $\mathbb{F}_2[U_O]_{O \in \mathbb{O}_1 \cup \cdots \cup \mathbb{O}_p}$ generated by tuples of intersection points, one point on each $\widetilde{\boldsymbol{\alpha}}$–circle, at most one on each $\boldsymbol{\alpha}$–arc, one on each $\widetilde{\boldsymbol{\beta}}$–circle, and at most one on each $\boldsymbol{\beta}$–arc, and let ∂ be the differential on $C^-(G)$ defined by counting empty rectangles. Then:

Proposition 4.14 *The structures* $(C^-(\boldsymbol{\mathcal{P}}), \partial_\wedge)$ *and* $(C^-(G), \partial)$ *are isomorphic.*

Proof The proof is analogous to that of Proposition 4.12 (here, any empty rectangle is either fully contained in one grid or intersects two consecutive grids). □

When the gluing maps between adjacent grids are clear from the context, we will use the otherwise ambiguous notation $G^\circ(\mathcal{P}_1^\circ) \cup \cdots \cup G^\circ(\mathcal{P}_p^\circ)$ for G. We will also sometimes write $\boldsymbol{x}_1 \cup \cdots \cup \boldsymbol{x}_p$ for $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_p)$.

We are now ready to prove Proposition 4.11.

Proof of Proposition 4.11 By definition, the maps δ^L, δ^R and δ^1 on a generator f^* of $\mathit{CDTD}^-(\mathcal{P}^*)$ correspond to the map $\partial_\wedge$ on the generators $\iota^L(f^*)\otimes f^*$, $f^*\otimes\iota^R(f^*)$ and $\iota^L(f^*)\otimes f^*\otimes\iota^R(f^*)$ of $\mathcal{E}^L(\mathcal{P}^*)\wedge\mathcal{P}^*$, $\mathcal{P}^*\wedge\mathcal{E}^R(\mathcal{P}^*)$ and $\mathcal{E}^L(\mathcal{P}^*)\wedge\mathcal{P}^*\wedge\mathcal{E}^R(\mathcal{P}^*)$, respectively.

One can also see that the maps δ^L, δ^R and δ^1 on a generator $\boldsymbol{x}$ of $\mathit{CDTD}^-(G^*)$ correspond to the map ∂ on the generators $\iota^L(\boldsymbol{x})\cup\boldsymbol{x}$, $\boldsymbol{x}\cup\iota^R(\boldsymbol{x})$ and $\iota^L(\boldsymbol{x})\cup\boldsymbol{x}\cup\iota^R(\boldsymbol{x})$ of the grid diagrams $G(\mathcal{E}^L(\mathcal{P}^*))\cup G^*$, $G^*\cup G(\mathcal{E}^R(\mathcal{P}^*))$ and $G(\mathcal{E}^L(\mathcal{P}^*))\cup G^*\cup G(\mathcal{E}^R(\mathcal{P}^*))$, respectively. We outline the correspondence for δ^L here. The other cases are analogous. An empty rectangle starting at $\boldsymbol{x}$ that stays in G^* contributes to $\partial(\boldsymbol{x})$, hence to $\iota^L(\boldsymbol{x})\otimes\partial(\boldsymbol{x})$, as well as to $\partial(\iota^L(\boldsymbol{x})\cup\partial(\boldsymbol{x}))$. An empty partial rectangle starting at $\boldsymbol{x}$ in G^* of the form $[0,t_1]\times[s_1,s_2]$, $[t_2,m+1]\times[s_1,s_2]$, $[0,m+1]\times[s_1,s_2]$ or $([0,t_1]\cup[t_2,m+1])\times[s_1,s_2]$ contributes to $\partial^L(\boldsymbol{x})$ and corresponds to the empty rectangle

$$\begin{gathered}[-s_1,t_1]\times[s_1,s_2],\\ ([-n-1,s_2]\cup[t_2,m+1])\times[s_1,s_2],\\ ([-n-1,-s_2]\cup[-s_1,m+1])\times[s_1,s_2],\\ ([-n-1,t_1]\cup[t_2,m+1])\times[s_1,s_2],\end{gathered}$$

respectively, in $G(\mathcal{E}^L(\mathcal{P}^*))\cup G^*$, which contributes to $\partial(\iota^L(\boldsymbol{x})\cup\boldsymbol{x})$.

By Propositions 4.12, 4.13 and 4.14, the correspondence between generators of $\mathcal{E}^L(\mathcal{P}^*)\wedge\mathcal{P}^*$ and $G(\mathcal{E}^L(\mathcal{P}^*))\cup G^*$, $\mathcal{P}^*\wedge\mathcal{E}^R(\mathcal{P}^*)$ and $G^*\cup G(\mathcal{E}^R(\mathcal{P}^*))$, and $\mathcal{E}^L(\mathcal{P}^*)\wedge\mathcal{P}^*\wedge\mathcal{E}^R(\mathcal{P}^*)$ and $G(\mathcal{E}^L(\mathcal{P}^*))\cup G^*\cup G(\mathcal{E}^R(\mathcal{P}^*))$, respectively, carries the map $\partial_\wedge$ to the map ∂. Therefore, the structures $(C^-(G^*),\delta^L)$ and $(C^-(\mathcal{P}^*),\delta^L)$, $(C^-(G^*),\delta^R)$ and $(C^-(\mathcal{P}^*),\delta^R)$, and $(C^-(G^*),\delta^1)$ and $(C^-(\mathcal{P}^*),\delta^1)$ are pairwise isomorphic. In particular, $\mathit{CDTD}^-(G^*)$ and $\mathit{CDTD}^-(\mathcal{P}^*)$ are isomorphic. Further, by Proposition 3.25, $(C^-(G^*),\delta^L)$ is a left type D structure, $(C^-(G^*),\delta^R)$ is a right type D structure and $\mathit{CDTD}^-(G^*)$ is a left–right type DD structure. □

The above proof sums up to the following observation. For a mirror-shadow $\mathcal{P}^*$, the maps δ^L, δ^R and δ^1 on a generator f^* correspond to gluing $G^*(\mathcal{P}^*)$ to $G(\mathcal{E}^R(\mathcal{P}^*))$ along the $\boldsymbol{\beta}$–curves and/or to $G(\mathcal{E}^L(\mathcal{P}^*))$ along the $\boldsymbol{\alpha}$–curves, and then taking the inner differential of the generator of the resulting diagram corresponding to $\iota^L(f^*)\otimes f^*$, $f^*\otimes\iota^R(f^*)$ or $\iota^L(f^*)\otimes f^*\otimes\iota^R(f^*)$, respectively.

If G is the bordered Heegaard diagram corresponding to an alternating sequence of shadows and mirror-shadows $\boldsymbol{\mathcal{P}}=(\mathcal{P}_1^\circ,\dots,\mathcal{P}_p^\circ)$ with well-defined consecutive wedge products, then $C^-(G)$ has a left type A or D map depending on whether $\mathcal{P}_1^\circ$ is shadow or a mirror-shadow, defined by counting partial rectangles in $G^\circ(\mathcal{P}_1^\circ)$ as usual, and similarly it has a right type A or D map depending on whether $\mathcal{P}_p^\circ$ is a shadow or a mirror shadow. Denote the resulting structures by $\mathit{CATA}^-(G)$, $\mathit{CDTA}^-(G)$, $\mathit{CATD}^-(G)$ or $\mathit{CDTD}^-(G)$, or simply by $\mathit{CT}^-(G)$.

4.6 Self-gluing of bordered grid diagrams

In this subsection we discuss annular bordered grid diagrams corresponding to one-sided modules. Let $G^* = G^*(\mathcal{P}^*) = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O})$ correspond to a mirror-shadow $\mathcal{P}^*$ with $\boldsymbol{a}_{1/2} = S_{\mathbb{X}} = T_{\mathbb{O}}$. This means that each row of G^* contains both an X and an O, thus the annular bordered grid diagram $G_{\boldsymbol{a}}^* = (\widetilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O})$ will have an X and an O in each of its annuli. See Figure 27.

Take the subset $\mathfrak{S}_n(G_{\boldsymbol{a}}^*)$ of generators that occupy each $\widetilde{\alpha}$–circle. Then the map ∂ that also counts the rectangles which cross the line $\{0\} \times [0, m+1] \sim \{n+1\} \times [0, m+1]$ endows $C_n^-(G_{\boldsymbol{a}}^*)$ with a chain complex structure, and under the usual identification of $\mathfrak{S}_n(\mathcal{P}^*)$ with $\mathfrak{S}_n(G_{\boldsymbol{a}}^*)$ we have:

Proposition 4.15 *$(C_n^-(G_{\boldsymbol{a}}^*), \partial)$ is a chain complex isomorphic to $(C_n^-(\mathcal{P}^*), \partial^* + {}_D\partial)$.*

Proof The proof is similar to the proof of Proposition 4.5. The terms in ${}_D\partial$ correspond to those empty rectangles that cross the gluing, as follows. For the generator $f = (S, T, \phi)$ corresponding to the intersection point $\boldsymbol{x} = (\alpha_s, \beta_{\phi s})_{s \in S}$, the pair (s_1, s_2) is allowable exactly when the glued up rectangle $R = ([0, s_1] \cup [s_2, n+1]) \times [\phi(s_2), \phi(s_1)]$ is empty (ie $\boldsymbol{x} \cap R = \mathbb{X} \cap R = \varnothing$). Then R connects $\boldsymbol{x}$ to $\boldsymbol{y} = (\alpha_s, \beta_{\phi^{(s_1, s_2)} s})_{s \in S}$ and n_O measures the multiplicity of O in R. □

Lemma 3.26 now follows from Proposition 4.15.

As in Section 4.4.1, we can define a right type D map on $C_n^-(G_{\boldsymbol{a}}^*)$ by $\delta^1 \boldsymbol{x} = \partial \boldsymbol{x} \otimes \iota^R(\boldsymbol{x}) + \partial^R \boldsymbol{x}$ to obtain a right type D structure $\mathit{CTD}^-(G_{\boldsymbol{a}}^*)$, which, by arguments analogous to those for Proposition 4.12, is isomorphic to $\mathit{CTD}^-(\mathcal{P}^*)$. We can similarly define structures $\mathit{CDT}^-(G_{\boldsymbol{b}}^*)$, $\mathit{CAT}^-(G_{\boldsymbol{b}})$ and $\mathit{CTA}^-(G_{\boldsymbol{a}})$ isomorphic to $\mathit{CDT}^-(\mathcal{P})$, $\mathit{CAT}^-(\mathcal{P})$ and $\mathit{CTA}^-(\mathcal{P})$.

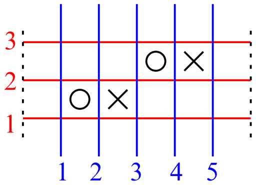


Figure 27: Self-gluing of a bordered grid diagram. The dashed lines are identified.

Convention 4.16 Similar to Convention 3.28, if $G_1^\circ \cup \cdots \cup G_p^\circ$ corresponds to an alternating sequence of shadows and mirror-shadows $\mathcal{P} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$, and G_1° and/or G_p° can be self-glued, we will always self-glue it, to produce a nice diagram G whose invariant is a one-sided module or a chain complex that agrees with $CT^-(\mathcal{P})$.

4.7 Pairing for plumbings of bordered grid diagrams

Gluing bordered grid diagrams corresponds to taking a box tensor product of their algebraic invariants:

Theorem 4.17 *Given an alternating sequence of shadows and mirror-shadows $\mathcal{P} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ with well-defined consecutive wedge products, define G_i and G_i' to be $G^\circ(\mathcal{P}_1^\circ) \cup \cdots \cup G^\circ(\mathcal{P}_i^\circ)$ and $G^\circ(\mathcal{P}_{i+1}^\circ) \cup \cdots \cup G^\circ(\mathcal{P}_p^\circ)$, respectively. The obvious identification of generators gives an isomorphism*

$$CT^-(G_i \cup G_i') \cong CT^-(G_i) \boxtimes CT^-(G_i').$$

Proof This follows from the equivalences proven earlier in this section, along with Theorem 3.33. Alternatively, one can notice that by definition of the type D and type A actions for bordered grid diagrams, pairing them via $\boxtimes$ corresponds to matching partial rectangles for the type D maps with sets of partial rectangles for the type A maps along the boundary. The possible pairings correspond to empty rectangles in the union of the two diagrams that cross the gluing. □

4.8 Relations between the U–actions

Let $\mathcal{P} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ be an alternating sequence of shadows and mirror-shadows with well-defined consecutive wedge products. Let G be the nice bordered Heegaard diagram obtained by gluing $G^\circ(\mathcal{P}_1^\circ), \dots, G^\circ(\mathcal{P}_p^\circ)$ as before.

The pairs $O = (s_O, \omega_i s_O)$ and $O' = (s_O', \omega_{i'} s_O')$ are connected by a path exactly when O and O' lie on the same component of the tangle $\mathcal{T}(G)$ associated to $\mathcal{P}$, or in other words if there is a sequence of $O = O_1$, X_1, O_2, $X_2, \dots,$ X_{k-1}, $O_k = O'$ such that O_j and X_j are in the same row, and X_{j-1} and O_j are in the same column (note that we also require that none of the X_j are in the first or last parts $G^\circ(\mathcal{P}_1^\circ)$ or $G^\circ(\mathcal{P}_p^\circ)$). Now we are ready to prove Lemma 3.35:

Proof of Lemma 3.35 First let us assume that $CT^-(\mathcal{P})$ is a type AA structure. Then we need to prove that there is a type AA map $\mathcal{H}$ such that $(U + U')\,\mathrm{id}_{CT^-(\mathcal{P})} = \partial\mathcal{H}$. It is enough to prove this statement in the case when O and O' are of distance 1 (the general case then can be obtained by adding up the homotopies for all j). This means that there is a point X which is in the row of O and in the column of O'. By definition, X is not in $G^\circ(\mathcal{P}_1^\circ)$ or $G^\circ(\mathcal{P}_p^\circ)$, thus the horizontal and vertical rows containing it are both closed up to annuli. This means that the map $\mathcal{H}_X$ that counts rectangles that cross X once consists of the single map $CT^-(\boldsymbol{\mathcal{P}}) \to CT^-(\boldsymbol{\mathcal{P}})$ with no nontrivial components of the type $\mathcal{A}(\mathcal{E}_L(\mathcal{P}_1))^{\otimes l} \otimes CT^-(\boldsymbol{\mathcal{P}}) \otimes \mathcal{A}(\mathcal{E}_R(\mathcal{P}_p))^{\otimes r} \to CT^-(\boldsymbol{\mathcal{P}})$ for l, $r > 0$. And, as in [11], the map $\mathcal{H}_X$ satisfies $(U+U')\,\mathrm{id}_{CT^-(\boldsymbol{\mathcal{P}})} = \partial\mathcal{H}_X$.

The argument goes exactly the same way for the other types of structures, with the observation that if $\boldsymbol{\mathcal{P}}$ starts or ends with a mirror-shadow, then we can complete it by adding $\mathcal{E}^R(\mathcal{P}_1^*)$ and/or $\mathcal{E}^L(\mathcal{P}_p^*)$ and denote the obtained sequence of shadows and mirror-shadows by $\boldsymbol{\mathcal{P}}'$. Then chain homotopy in $(C^-(\boldsymbol{\mathcal{P}}'), \partial)$ gives type DD (or DA, or AD) equivalence of $CT^-(\boldsymbol{\mathcal{P}})$. □

5 Modules associated to tangles

In this section we will associate a left type D structure or a right type A structure to a tangle in D^3, a type DA structure to a tangle in $I \times S^2$, and a bigraded chain complex to a knot (or link) in S^3. The main idea is to cut $\mathcal{T}$ into elementary pieces $\mathcal{T} = \mathcal{T}_1 \circ \cdots \circ \mathcal{T}_p$, associate a type A structure to $\mathcal{T}_1$ if it is in D^3, a type D structure to $\mathcal{T}_p$ if it is in D^3, and type DA structures to all the other $\mathcal{T}_j$, and then take their box-tensor product. The structures associated to elementary pieces are the structures defined earlier for wedge products of appropriate shadows and mirror-shadows. The hard part — of course — is to prove independence of the cut. Although we believe that there is a completely combinatorial proof of the independence, in this paper we will only provide a proof that uses holomorphic curve techniques; see Section 10. As a consequence of that, we can only prove independence for the tilde version of the theory.

5.1 Algebras associated to $\partial\mathcal{T}$

For a sequence of oriented points with signs $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_k)$, let $n = k+1$, and recall that the sequence $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^1$ corresponds to two complementary subsets $S_{\mathbb{X}} = \{j + \frac{1}{2} : \epsilon_j = -1\}$ and $T_{\mathbb{O}} = \{j + \frac{1}{2} : \epsilon_j = +1\}$ of the set $\{1\frac{1}{2}, \dots, n - \frac{1}{2}\}$. Set $\boldsymbol{\epsilon}^0 = -\boldsymbol{\epsilon}^1$. This determines $T_{\mathbb{X}} (= S_{\mathbb{X}})$ and $S_{\mathbb{O}} (= T_{\mathbb{O}})$ in a similar vein. Take the idempotent shadow ${}_{\boldsymbol{\epsilon}^0}\mathcal{E}_{\boldsymbol{\epsilon}^1} = (n, n, \mathrm{id}_{S_{\mathbb{X}}}, \mathrm{id}_{S_{\mathbb{O}}})$ of Example 3.3. This defines the algebra $\mathcal{A}_{\boldsymbol{\epsilon}} = \mathcal{A}({}_{\boldsymbol{\epsilon}^0}\mathcal{E}_{\boldsymbol{\epsilon}^1})$.

Given a tangle $\mathcal{T}$ with left boundary $\partial^0\mathcal{T}$ and right boundary $\partial^1\mathcal{T}$ (any of these sets can be empty if the tangle is closed from that side), let $\boldsymbol{\epsilon}^0 = \boldsymbol{\epsilon}(\partial^0\mathcal{T})$ and $\boldsymbol{\epsilon}^1 = \boldsymbol{\epsilon}(\partial^1\mathcal{T})$ be the sequences of signs of $\partial^0\mathcal{T}$ and $\partial^1\mathcal{T}$, respectively. Let $\mathcal{A}(\partial^0\mathcal{T}) = \mathcal{A}_{-\boldsymbol{\epsilon}^0}$ and

$\mathcal{A}(\partial^1\mathcal{T}) = \mathcal{A}_{\boldsymbol{\epsilon}^1}$. The minus sign in the second definition is there so that if we cut $\mathcal{T} = \mathcal{T}_1 \circ \mathcal{T}_2$, then $\boldsymbol{\epsilon}^1(\partial^1\mathcal{T}_1) = -\boldsymbol{\epsilon}^0(\partial^0\mathcal{T}_2)$, thus $\mathcal{A}(\partial^1\mathcal{T}_1) = \mathcal{A}(-\partial^0\mathcal{T}_2)$.

5.2 Invariants associated to a tangle

Given a sequence of shadows and mirror-shadows $\boldsymbol{\mathcal{P}} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ with well-defined consecutive wedge products, each $\mathcal{P}_j^\circ$ has a tangle $\mathcal{T}_j = \mathcal{T}^\circ(\mathcal{P}_j^\circ)$ associated to it. Note that if $\mathcal{P}_j$ is a shadow then at all crossings the strand with the bigger slope goes over the strand with the smaller slope, while if $\mathcal{P}_j^*$ is a mirror-shadow then at all crossings the strand with the smaller slope goes over the strand with the bigger slope. Note $\mathcal{P}_j^\circ$ and $\mathcal{P}_{j+1}^\circ$ have a well-defined wedge product — thus $\mathcal{P}_j^\circ$ is not left-contractible and $\mathcal{P}_{j+1}^\circ$ is not right-contractible — so $\partial^1\mathcal{T}_j \neq \varnothing$ and $\partial^0\mathcal{T}_{j+1} \neq \varnothing$ for $1 \leq j \leq p-1$. If $\mathcal{P}_1^\circ$ is left-contractible then $\partial^0\mathcal{T}_1 = \varnothing$ and if $\mathcal{P}_p^\circ$ is left-contractible then $\partial^1\mathcal{T}_p = \varnothing$. This means that the composition-tangle $\mathcal{T}(\boldsymbol{\mathcal{P}}) = \mathcal{T}_1 \circ \cdots \circ \mathcal{T}_p$ can be in S^3, D^3 or in $S^2 \times I$. Moreover any tangle $\mathcal{T}$ can be constructed in the above way.

Lemma 5.1 *Let $\mathcal{T}$ be a tangle in S^3, D^3 or in $S^2 \times I$. Then there is a sequence of shadows $\boldsymbol{\mathcal{P}} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ such that $\mathcal{T}$ is isotopic to $\mathcal{T}(\boldsymbol{\mathcal{P}})$ (relative to the boundary), and*

- *if $\partial^0\mathcal{T} = \varnothing$ then $\mathcal{P}_1^*$ is a mirror-shadow;*
- *if $\partial^1\mathcal{T} = \varnothing$ then $\mathcal{P}_p^*$ is a mirror-shadow;*
- *if $\partial^0\mathcal{T} \neq \varnothing$ then $\mathcal{P}_1^*$ is a mirror-shadow and $\boldsymbol{\epsilon}^0(\mathcal{P}_1^*) = \boldsymbol{\epsilon}^0(\mathcal{T})$;*
- *if $\partial^1\mathcal{T} \neq \varnothing$ then $\mathcal{P}_p$ is a shadow and $\boldsymbol{\epsilon}^1(\mathcal{P}_p) = \boldsymbol{\epsilon}^1(\mathcal{T})$.*

The first two assumptions are in the statement for cosmetic reasons (to match with the assumptions of Sections 7–12), while, as we will see later, the last two assumptions ensure that the associated invariant has the correct type and is defined over the correct algebras.

Proof The statement is clearly true for elementary tangles $\mathcal{T}$. Indeed, depending on the type of crossing in $\mathcal{T}$, or whether $\mathcal{T}$ is a cap or a cup we can always bisect $\mathcal{T}$ into two pieces $\mathcal{T}_- \circ \mathcal{T}_+$ such that one of $\mathcal{T}_-$ or $\mathcal{T}_+$ consists of straight strands (possibly with a gap) and the other one is isotopic to $\mathcal{T}$, and at the (possible) crossing of $\mathcal{T}_-$ (or $\mathcal{T}_+$) the strand with the smaller slope goes over (under) the strand with bigger slope, or $\mathcal{T}_-$ (or $\mathcal{T}_+$) is a cup (or a cap). Let $\mathcal{P}_-^*$ and $\mathcal{P}_+$ be the mirror shadow and shadow corresponding to $\mathcal{T}_-$ and $\mathcal{T}_+$ (ie $\mathcal{T}_- = \mathcal{T}^*(\mathcal{P}_-^*)$ and $\mathcal{T}_+ = \mathcal{T}(\mathcal{P}_+)$). Note that in this case the condition $\epsilon^0(\mathcal{P}_-^*) = \epsilon^0(\mathcal{T})$ is equivalent to $\mathcal{P}_-^*$ not having a gap on its left side. Similarly the condition $\epsilon^1(\mathcal{P}_+) = \epsilon^1(\mathcal{T})$ to $\mathcal{P}_+$ not having a gap on its right side.

In the general case, put $\mathcal{T}$ in a not obviously split position. This means that when cutting it up into elementary tangles $\mathcal{T} = \mathcal{T}_1 \circ \cdots \circ \mathcal{T}_p$, every cut intersects the tangle. Then, by the previous paragraph, each $\mathcal{T}_i$ is isotopic to $\mathcal{T}^*((\mathcal{P}_i)^*_-) \circ \mathcal{T}((\mathcal{P}_i)_+)$. Thus if $\partial^1\mathcal{T} \neq \varnothing$ then the decomposition $\mathcal{T} = \mathcal{T}^*((\mathcal{P}_1)^*_-) \circ \mathcal{T}((\mathcal{P}_1)_+) \circ \cdots \circ \mathcal{T}^*((\mathcal{P}_p)^*_-) \circ \mathcal{T}((\mathcal{P}_p)_+)$ works. Otherwise $\mathcal{T}_p$ is a single cap, thus it can be written as $\mathcal{T}_p = \mathcal{T}^*(\mathcal{P}^*_p)$, where $\mathcal{P}^*_p$ does not have a gap on its right. This means that the decomposition

$$\mathcal{T} = \mathcal{T}^*((\mathcal{P}_1)^*_-) \circ \mathcal{T}((\mathcal{P}_1)_+) \circ \cdots \circ \mathcal{T}^*((\mathcal{P}_{p-1})^*_-) \circ \mathcal{T}((\mathcal{P}_{p-1})_+) \circ \mathcal{T}^*(\mathcal{P}^*_p)$$

satisfies all criteria of the lemma. □

Note that by construction, if $\partial^0\mathcal{T} = \varnothing$, then $\mathcal{T}^-_1$ is left-contractible, and if $\partial^1\mathcal{T} = \varnothing$, then $\mathcal{T}^-_p$ is right-contractible.

Definition 5.2 Let $\mathcal{T}$ be a tangle given by a sequence of shadows $\boldsymbol{\mathcal{P}} = (\mathcal{P}^\circ_1, \dots, \mathcal{P}^\circ_p)$ as in Lemma 5.1.

If $\partial^0\mathcal{T} = \varnothing$ and $\partial^1\mathcal{T} = \varnothing$, then define the chain complex by

$$CT^-(\boldsymbol{\mathcal{P}}) = CTD^-(\mathcal{P}^*_1) \boxtimes \cdots \boxtimes CDT^-(\mathcal{P}^*_p).$$

If $\partial^0\mathcal{T} = \varnothing$ and $\partial^1\mathcal{T} \neq \varnothing$, then define the right type A structure over $\mathcal{A}(\partial^1\mathcal{T})$ by

$$CTA^-(\boldsymbol{\mathcal{P}}) = CTD^-(\mathcal{P}^*_1) \boxtimes \cdots \boxtimes CATA^-(\mathcal{P}_p).$$

If $\partial^0\mathcal{T} \neq \varnothing$ and $\partial^1\mathcal{T} = \varnothing$, then define the left type D structure over $\mathcal{A}(\partial^0\mathcal{T})$ by

$$CTD^-(\boldsymbol{\mathcal{P}}) = CDTD^-(\mathcal{P}^*_1) \boxtimes \cdots \boxtimes CDT^-(\mathcal{P}^*_p).$$

If $\partial^0\mathcal{T} \neq \varnothing$ and $\partial^1\mathcal{T} \neq \varnothing$, then define the left–right type DA structure over $\mathcal{A}(\partial^0\mathcal{T})$ and $\mathcal{A}(\partial^1\mathcal{T})$ by

$$CDTA^-(\boldsymbol{\mathcal{P}}) = CDTD^-(\mathcal{P}^*_1) \boxtimes \cdots \boxtimes CATA^-(\mathcal{P}_p).$$

Whenever the sequence $\boldsymbol{\mathcal{P}}$ is clear from the context, we simplify the notation of the above bimodules to $CT^-(\mathcal{T})$. In this paper we will not prove that $CT^-(\mathcal{T})$ as defined above is an invariant of $\mathcal{T}$. We will only prove it for the weaker version $\widetilde{CT}(\mathcal{T})$. From now on, we restrict ourselves to the tilde theory by setting all U_O to 0. A consequence of Theorems 12.4 and 11.15 is:

Theorem 5.3 *Suppose that $\boldsymbol{\mathcal{P}} = (\mathcal{P}^\circ_1, \dots, \mathcal{P}^\circ_p)$ and $\boldsymbol{\mathcal{Q}} = (\mathcal{Q}^\circ_1, \dots, \mathcal{Q}^\circ_q)$ give tangles* (in the sense of Lemma 5.1) *isotopic to $\mathcal{T}$. Then for some integers $k(\mathcal{P})$ and $k(\mathcal{Q})$, the (bi)modules $\widetilde{CT}(\mathcal{P}^\circ_1) \boxtimes \cdots \boxtimes \widetilde{CT}(\mathcal{P}^\circ_p) \boxtimes V^{\otimes k(\mathcal{Q})}$ and $\widetilde{CT}(\mathcal{Q}^\circ_1) \boxtimes \cdots \boxtimes \widetilde{CT}(\mathcal{Q}^\circ_q) \otimes V^{\otimes k(\mathcal{P})}$ are equivalent. Here $V = \mathbb{F}_2 \oplus \mathbb{F}_2$, where one of the $\mathbb{F}_2$ components has bigrading $(M, A) = (-1, -1)$ and the other one has bigrading $(M, A) = (0, 0)$.*

The integers $k(\mathcal{P})$ and $k(\mathcal{Q})$ in the above theorem can be computed explicitly. For a shadow $\mathcal{P}$ (or mirror-shadow $\mathcal{P}^*$), define $k(\mathcal{P}) = |S_{\mathbb{X}}|$ (or $k(\mathcal{P}^*) = |S_{\mathbb{X}}|$). For a sequence of shadows and mirror-shadows $\boldsymbol{\mathcal{P}} = (\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ)$ with a well-defined wedge product, define $k(\boldsymbol{\mathcal{P}}) = \sum_{j=1}^{p} k(\mathcal{P}_j)$.

The *DA* bimodule for the trivial tangle is equivalent to the identity bimodule, or more precisely:

Theorem 5.4 *If $\boldsymbol{\mathcal{P}} = (\mathcal{E}_1^*, \mathcal{E}_2)$ is a sequence of an idempotent mirror-shadow and shadow for a tangle $\mathcal{T}$ consisting of m straight strands, then*

$$\widetilde{CATA}(\mathcal{E}_2) \boxtimes \widetilde{CDTA}(\boldsymbol{\mathcal{P}}) \simeq \widetilde{CATA}(\mathcal{E}_2) \otimes V^{\otimes m}.$$

Proof The proof follows from the results in Sections 7–12, but we outline it here nevertheless. One can represent the sequence $(\mathcal{E}_2, \mathcal{E}_1^*, \mathcal{E}_1)$ by a plumbing of bordered grid diagrams. One can perform Heegaard moves to this plumbing to obtain the bordered grid diagram for $\mathcal{E}_2$. Every index zero/three destabilization results in an extra V factor. Observe that $\widetilde{CATA}(\mathcal{E}_2)$ is just the tilde version of the algebra $\mathcal{A}(\mathcal{E}_2)$. □

5.3 Sample invariance proofs

Although the proof of Theorem 5.3 is proved entirely in Section 10, to give evidence that the theory can be defined combinatorially we give sample proofs for statements from Theorem 5.3. Most of the arguments rely on the generalization of the commutation move for grid diagrams.

5.3.1 Generalized commutation In all the (bordered) Heegaard diagrams we have been working with, all regions (connected components of $\Sigma \setminus (\boldsymbol{\alpha} \cup \boldsymbol{\beta})$) are rectangles, and each annulus between two neighboring $\boldsymbol{\alpha}$–circles or $\boldsymbol{\beta}$–circles contains exactly one X and one O. In the following this will be our assumption on the Heegaard diagrams, and we will call these diagrams *rectangular*. Note that for rectangular diagrams the connected components of $\Sigma \setminus \boldsymbol{\alpha}$ (or $\Sigma \setminus \boldsymbol{\beta}$) are annuli or punctured spheres with at most two boundary components intersecting $\boldsymbol{\alpha}$ (or $\boldsymbol{\beta}$) and the rest of the boundary components are subsets of $\partial\Sigma$. Thus rectangular diagrams are always constructed as a plumbing of annuli.

So let $\mathcal{H} = (\Sigma, \boldsymbol{\alpha} = \boldsymbol{\alpha}^c \cup \boldsymbol{\alpha}^a, \boldsymbol{\beta} = \boldsymbol{\beta}^c \cup \boldsymbol{\beta}^a, \mathbb{X}, \mathbb{O})$ be a rectangular Heegaard diagram such that every annulus contains an X. Then in the usual way we can define a chain complex with underlying module $C^-(\mathcal{H})$ generated over $\boldsymbol{k} = \mathbb{F}[U_O]_{O \in \mathbb{O}}$ by intersection points $\boldsymbol{x} \in \mathfrak{S}(\mathcal{H})$ with one intersection point on each circle $\boldsymbol{\alpha}^c$ and each circle $\boldsymbol{\beta}^c$ and at most one intersection point on each arc $\boldsymbol{\alpha}^a$ and each arc $\boldsymbol{\beta}^a$. The

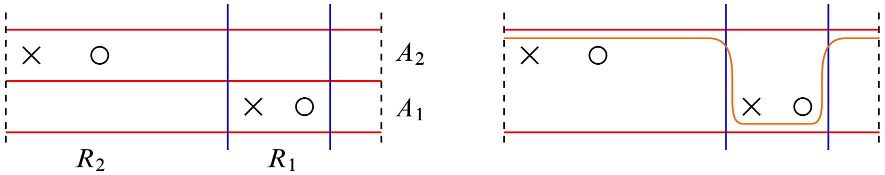


Figure 28: Generalized commutation. The left- and right-hand side of each diagram are identified.

differential is defined by counting empty rectangles: a rectangle from a generator $\boldsymbol{x}$ to a generator $\boldsymbol{y}$ is an embedded rectangle $R \subset \Sigma$ with boundary $\partial R \subset \boldsymbol{\alpha} \cup \boldsymbol{\beta}$ such that $\boldsymbol{x} \cap R$ is the two corners of R where $(T\boldsymbol{\alpha}, T\boldsymbol{\beta})$ form a positive basis of $T\Sigma$ and $\boldsymbol{y} \cap R$ is the two corners of R where $(T\boldsymbol{\alpha}, T\boldsymbol{\beta})$ form a negative basis of $T\Sigma$ (here the orientation on the tangent vectors comes from the orientation on ∂R). A rectangle R is called empty if $\mathrm{Int}(R) \cap (\boldsymbol{x} \cup \boldsymbol{y}) = \varnothing$ and $R \cap \mathbb{X} = \varnothing$. Denote the set of empty rectangles from $\boldsymbol{x}$ to $\boldsymbol{y}$ by $\mathcal{R}_0(\boldsymbol{x}, \boldsymbol{y})$. Then define

$$\partial \boldsymbol{x} = \sum_{\boldsymbol{y} \in \mathfrak{S}(\mathcal{H})} \sum_{R \in \mathcal{R}_0(\boldsymbol{x}, \boldsymbol{y})} \prod_{O \in \mathbb{O}} U^{|R \cap O|} \boldsymbol{y}.$$

This can be extended to the whole $C^-(\mathcal{H})$ and using the usual arguments we conclude:

Lemma 5.5 $(C^-(\mathcal{H}), \partial)$ *is a chain complex.* □

Take three consecutive alpha circles α_1, α_2 and α_3, so that α_1 and α_2 bound the annulus A_1 and α_2 and α_3 bound the annulus A_2. All connected components of $\boldsymbol{\beta} \cap (A_1 \cup A_2)$ are intervals. Suppose that two of these intervals corresponding to different $\boldsymbol{\beta}$–curves subdivide $A_1 \cup A_2$ into two rectangles R_1 and R_2 such that $(\mathbb{X} \cup \mathbb{O}) \cap A_1 \subset R_1$ and $(\mathbb{X} \cup \mathbb{O}) \cap A_2 \subset R_2$. Then we can define a new Heegaard diagram $\mathcal{H}'$ by changing α_2 to α_2', where α_2' is the smoothing of $(\alpha_3 \setminus \partial R_1) \cup (\partial R_1 \setminus \alpha_3)$ isotoped in the complement of $\mathbb{X} \cup \mathbb{O}$ so that it is disjoint from $\boldsymbol{\alpha} \setminus \{\alpha_2\}$, transverse to all β–curves and intersects them only once. See Figure 28. Then:

Lemma 5.6 (generalized commutation) *The complexes* $(C^-(\mathcal{H}), \partial)$ *and* $(C^-(\mathcal{H}'), \partial')$ *are chain homotopy equivalent.*

Proof The proof is literally the same as in the closed case (see [12, Section 3.1]): the chain maps count pentagons, while the homotopy counts hexagons of the triple Heegaard diagram. □

For sequences of shadows and mirror-shadows, the proof goes the same way:

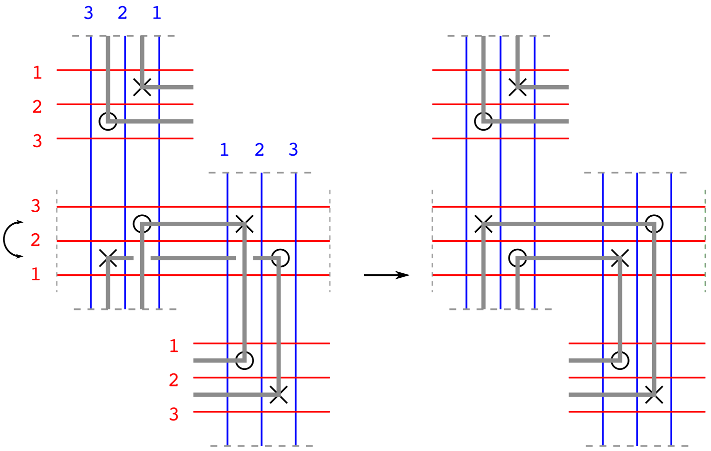


Figure 29: Diagram for simplifying a Reidemeister II move. The first picture corresponds to two canceling crossings, the arrow corresponds to a generalized commutation, and the second picture corresponds to straight strands. This image can have more straight strands that are not affected by the moves.

Lemma 5.7 *Let $\mathcal{P} = (\mathcal{P}_1^\circ, \mathcal{P}_2^\circ, \dots \mathcal{P}_p^\circ)$ and $\mathcal{Q}' = (\mathcal{Q}_1^\circ, \mathcal{Q}_2^\circ, \dots \mathcal{Q}_p^\circ)$ be sequences of shadows and mirror-shadows with well-defined wedge products. Assume that the corresponding grid diagrams $G(\mathcal{P})$ and $G(\mathcal{Q})$ are related to each other by generalized commutation. Then the associated structures $CT^-(\mathcal{P})$ and $CT^-(\mathcal{Q})$ are equivalent.*

Using Lemma 5.7, we can prove the following:

Proposition 5.8 *Let $\mathcal{P} = \{\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ\}$ and $\mathcal{Q} = \{\mathcal{Q}_1^\circ, \dots, \mathcal{Q}_p^\circ\}$ be sequences with corresponding tangles* (in the sense of Lemma 5.1) *$\mathcal{T}(\mathcal{P})$ and $\mathcal{T}(\mathcal{Q})$, respectively. Suppose that $\mathcal{T}(\mathcal{P})$ and $\mathcal{T}(\mathcal{Q})$ are related to each other by Reidemeister II and Reidemeister III moves. Then the (bi)modules $CT^-(\mathcal{T}(\mathcal{P}))$ and $CT^-(\mathcal{T}(\mathcal{Q}))$ are equivalent.*

Proof As is shown in Figure 29, a Reidemeister II move is simply a general commutation on the associated grid diagram. A Reidemeister III move can be achieved with a sequence of commutation moves; see Figure 30. □

6 Relation to knot Floer homology

This section provides the connection between CT^- and CFK^-.

Let $\mathcal{P}_1^\circ, \dots, \mathcal{P}_n^\circ$ be a sequence of shadows and mirror-shadows as in Lemma 5.1 such that the associated tangle $L = \mathcal{T}^\circ(\mathcal{P}_1^\circ) \circ \dots \circ \mathcal{T}^\circ(\mathcal{P}_n^\circ)$ is a closed link. After self-gluing

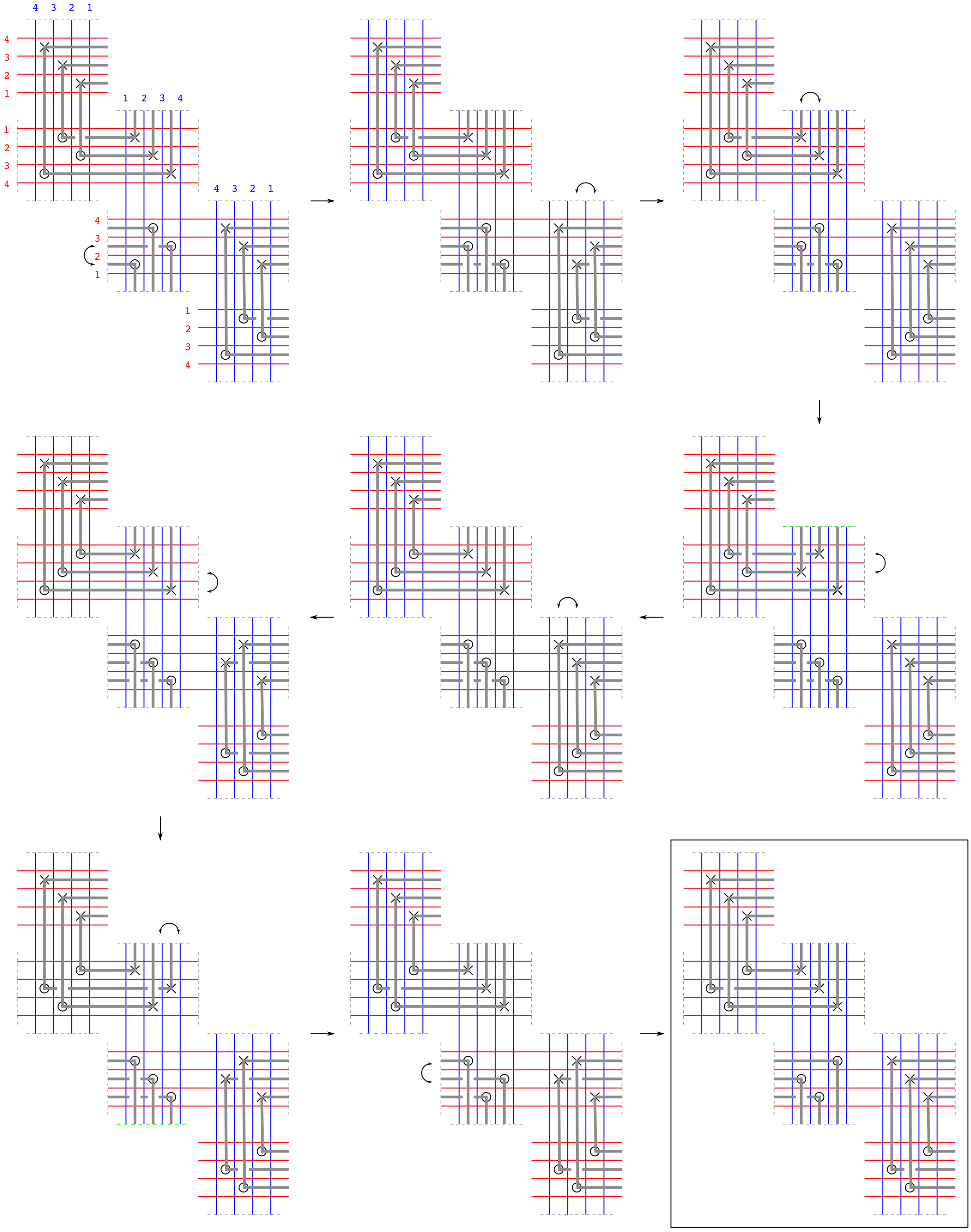


Figure 30: Commutation moves corresponding to a Reidemeister III move. Again, this image can have more straight strands that are not affected by the moves.

the first and last grid in $G^\circ(\mathcal{P}_1^\circ) \cup \cdots \cup G^\circ(\mathcal{P}_n^\circ)$, we obtain a diagram that is a plumbing of annuli and has one boundary component. Close off the boundary by gluing on a disk with one X and one O in it. The resulting closed Heegaard diagram $\mathcal{H}$ represents the link $L \cup U$, where U is an unknot unlinked from L.

Theorem 6.1 *We have a graded homotopy equivalence*

$$CT^-(\mathcal{P}_1^\circ) \boxtimes \cdots \boxtimes CT^-(\mathcal{P}_n^\circ) \simeq gCFK^-(\mathcal{H})$$

that maps a homogeneous generator in Maslov grading m and Alexander grading a to a homogeneous generator in Maslov grading $m + \frac{1}{2}|L|$ and Alexander grading $a + \frac{1}{2}|L|$.

Before we prove Theorem 6.1, we review the basic construction for knot Floer homology; see also [16; 23; 12; 21].

Let $\mathcal{H}_L = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{O}, \mathbb{X})$ be a Heegaard diagram for a knot or a link L with l components, where $\mathbb{O}$ and $\mathbb{X}$ are sets of $k \geq l$ basepoints. Let $\mathfrak{S}$ be the set of generators of $\mathcal{H}_L$. The *knot Floer complex* $CFK^-(\mathcal{H}_L)$ is generated over $\mathbb{F}_2[U_1, \dots, U_k]$ by $\mathfrak{S}$, with differential

$$\partial^-(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathfrak{S}} \sum_{\substack{B \in \tilde{\pi}_2(\boldsymbol{x},\boldsymbol{y}) \\ \operatorname{ind} B = 1}} \#\mathcal{M}^B(\boldsymbol{x}, \boldsymbol{y}) \prod_{O_i \in \mathbb{O}} (U_i^{n_{O_i}(B)}) \cdot \boldsymbol{y},$$

where $\tilde{\pi}_2(\boldsymbol{x}, \boldsymbol{y})$ is the set of homology classes from $\boldsymbol{x}$ to $\boldsymbol{y}$ which may cross both $\mathbb{O}$ and $\mathbb{X}$. The complex has a differential grading called the *Maslov* grading. As a relative grading, it is defined by

$$M'(\boldsymbol{x}) - M'(\boldsymbol{y}) = \operatorname{ind} B - 2n_{\mathbb{O}}(B),$$
$$M'(U_i \boldsymbol{x}) = M'(\boldsymbol{x}) - 2,$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathfrak{S}$ and $B \in \tilde{\pi}_2(\boldsymbol{x}, \boldsymbol{y})$. The complex also comes endowed with an *Alexander* filtration, defined by

$$A'(\boldsymbol{x}) - A'(\boldsymbol{y}) = n_{\mathbb{X}}(B) - n_{\mathbb{O}}(B),$$
$$A'(U_i \boldsymbol{x}) = A'(\boldsymbol{x}) - 1,$$

and normalized so that

$$(1) \qquad \#\{\boldsymbol{x} \in \mathfrak{S} \mid A'(\boldsymbol{x}) = a\} = \#\{\boldsymbol{x} \in \mathfrak{S} \mid A'(\boldsymbol{x}) = -a\} \mod 2.$$

The *associated graded object* $gCFK^-(\mathcal{H}_L)$ is also generated over $\mathbb{F}_2[U_1,\dots,U_k]$ by $\mathfrak{S}$, and its differential is given by

$$\partial^-(\boldsymbol{x}) = \sum_{\boldsymbol{y}\in\mathfrak{S}} \sum_{\substack{B\in\tilde{\pi}_2(\boldsymbol{x},\boldsymbol{y})\\ \operatorname{ind}B=1\\ n_{\mathbb{X}}(B)=0}} \#\mathcal{M}^B(\boldsymbol{x},\boldsymbol{y}) \prod_{O_i\in\mathbb{O}} (U_i^{n_{O_i}(B)})\cdot\boldsymbol{y}.$$

The Alexander filtration descends to a grading on $gCFK^-(\mathcal{H}_L)$. The bigraded homology

$$HFK^-(L) := H_*(gCFK^-(\mathcal{H}_L))$$

is an invariant of L.

The Maslov grading is normalized so that after setting each U_i to zero we get

$$H_*(CFK^-(\mathcal{H}_L)/(U_i{=}0)) \cong H_{*+k-1-(l-1)/2}(T^{k-1}),$$

where $*$ denotes the grading M' and we ignore the Alexander filtration on $CFK^-(\mathcal{H}_L)$.

One can also set each $U_i=0$ to obtain the filtered chain complex over $\mathbb{F}_2$,

$$\widehat{CFK}(\mathcal{H}_L) := CFK^-(\mathcal{H}_L)/(U_i{=}0).$$

The associated graded object to $\widehat{CFK}(\mathcal{H}_L)$ is $g\widehat{CFK}(\mathcal{H}_L)$, with differential

$$\hat{\partial}(\boldsymbol{x}) = \sum_{\boldsymbol{y}\in\mathfrak{S}} \sum_{\substack{B\in\tilde{\pi}_2(\boldsymbol{x},\boldsymbol{y})\\ \operatorname{ind}B=1\\ n_{\mathbb{X}}(B)=0=n_{\mathbb{O}}(B)}} \#\mathcal{M}^B(\boldsymbol{x},\boldsymbol{y})\cdot\boldsymbol{y}.$$

We denote its homology, which is an invariant of L, by $\widehat{HFK}(L) := H_*(g\widehat{CFK}(\mathcal{H}_L))$.

There is another grading, which we refer to as the $\mathbb{X}$*–normalized grading*, defined by

$$N'(\boldsymbol{x}) - N'(\boldsymbol{y}) = \operatorname{ind}B - 2n_{\mathbb{X}}(B),$$
$$N'(U_i\boldsymbol{x}) = N'(\boldsymbol{x}),$$

and normalized so that

$$H_*(gCFK^-(L)/(U_i{=}1)) \cong H_{*+k-1-(l-1)/2}(T^{k-1}),$$

where $*$ denotes the grading N'.

It turns out that

$$N' = M' - 2A' - (k-l), \tag{2}$$

so instead of using (1) to normalize the Alexander grading, we can use (2).

Next, we put the grading from Section 3.4 in the context of grid diagrams.

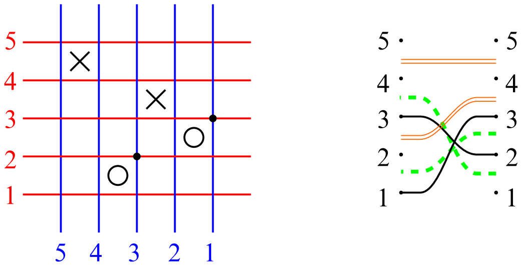


Figure 31: The generator $f_{\mathbb{O}}$ on a grid diagram G (left) and the corresponding generator on the shadow for G (right)

Let $\mathcal{P}$ be a shadow, let $G = G(\mathcal{P})$ be the corresponding grid and G^* be the grid corresponding to $\mathcal{P}^*$. We define a few special generators below.

Let $f_{\mathbb{O}}$ be the generator of G formed by picking the top-right corner of each O — see Figure 31 — and let $f'_{\mathbb{O}}$ be the generator formed by picking the bottom-left corner of each O. Similarly, let $f^*_{\mathbb{O}}$ be the generator of G^* formed by picking the bottom-left corner of each O, together with the top-right corner of the grid G^* — see Figure 32 — and let $f'^*_{\mathbb{O}}$ be the generator formed by picking the top-right corner of each O, together with the bottom-left corner of the grid G^*.

Let $f_{\mathbb{X}}$ and $f'_{\mathbb{X}}$ be the generators of G formed by picking the top-right (respectively bottom-left) corner of each X. Similarly, let $f^*_{\mathbb{X}}$ and $f'^*_{\mathbb{X}}$ be the generators of G^* formed by picking the bottom-left (respectively top-right) corner of each X, and the top-right (respectively bottom-left) corner of the grid.

Lemma 6.2 *For the generators defined above, we have*

$$\begin{aligned}
M(f_{\mathbb{O}}) &= M(f'_{\mathbb{O}}) = M(f^*_{\mathbb{O}}) = M(f'^*_{\mathbb{O}}) = -|\mathbb{O}|,\\
M(f_{\mathbb{X}}) &= M(f'_{\mathbb{X}}) = \mathrm{inv}(\xi^{-1}) - \mathrm{inv}(\xi^{-1},\omega) + \mathrm{inv}(\omega),\\
M(f^*_{\mathbb{X}}) &= M(f^*_{\mathbb{X}}) = -\mathrm{inv}(\xi^{-1}) + \mathrm{inv}(\xi^{-1},\omega) - \mathrm{inv}(\omega) - |\mathbb{O}|,\\
A(f_{\mathbb{X}}) &= \tfrac{1}{2}M(f_{\mathbb{X}}) = A(f'_{\mathbb{X}}),\\
A(f^*_{\mathbb{X}}) &= \tfrac{1}{2}M(f^*_{\mathbb{X}}) = A(f'^*_{\mathbb{X}}).
\end{aligned}$$

Proof Write out $f_{\mathbb{O}} = (S, T, \phi)$. Let $t = |\mathbb{O}|$, let $g_1, \dots, g_t$ be the dashed (green) strands in the graphical representation for the shadow $\mathcal{P}$, and let $f_1, \dots, f_t$ be the strands for f, where f_i is the strand that starts immediately below and ends immediately above g_i.

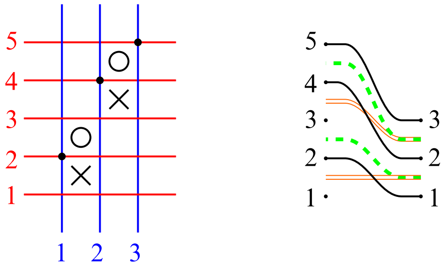


Figure 32: The generator $f^*_{\mathbb{O}}$ on a grid diagram G^* (left) and the corresponding generator on the mirror-shadow for G^* (right)

Recall that $\mathrm{inv}(\phi)$ counts intersections between pairs in $\{f_1, \dots, f_t\}$, $\mathrm{inv}(\omega)$ counts intersections between pairs in $\{g_1, \dots, g_t\}$, and $\mathrm{inv}(\phi, \omega)$ counts the total number of intersections between a strand in $\{f_1, \dots, f_t\}$ and a strand in $\{g_1, \dots, g_t\}$.

Observe that $\mathrm{inv}(\phi) = \mathrm{inv}(\omega)$, since each f_i is just a perturbation of g_i. Also, f_i intersects g_j exactly when $i \neq j$ and g_i intersects g_j, or $i = j$, so $\mathrm{inv}(\phi, \omega) = 2\,\mathrm{inv}(\omega) + |S_{\mathbb{O}}|$. Thus,

$$\begin{aligned} M(f_{\mathbb{O}}) &= \mathrm{inv}(\phi) - \mathrm{inv}(\phi, \omega) + \mathrm{inv}(\omega) \\ &= \mathrm{inv}(\omega) - 2\,\mathrm{inv}(\omega) - |S_{\mathbb{O}}| + \mathrm{inv}(\omega) \\ &= -|S_{\mathbb{O}}|. \end{aligned}$$

Similarly, write out $f^*_{\mathbb{O}} = (S, T, \phi)$. Again let $t = |\mathbb{O}|$, let $g_1, \dots, g_t$ be the dashed (green) strands in the graphical representation for the shadow $\mathcal{P}^*$, and let $f_1, \dots, f_{t+1}$ be the strands for f, where f_i is the strand that starts and ends immediately below g_i for $1 \le i \le t$, and f_{t+1} connects the highest point to the left to the highest point to the right. Clearly $\mathrm{inv}(\phi) = \mathrm{inv}(\omega)$ and $\mathrm{inv}(\phi, \omega) = 2\,\mathrm{inv}(\omega)$, since this time, for a fixed i, f_i and g_i do not intersect, so

$$M(f^*_{\mathbb{O}}) = -\mathrm{inv}(\phi) + \mathrm{inv}(\phi, \omega) - \mathrm{inv}(\omega) - |S_{\mathbb{O}}| = -|S_{\mathbb{O}}|.$$

The proof for $f'_{\mathbb{O}}$ and $f'^*_{\mathbb{O}}$ is analogous.

Now write $f_{\mathbb{X}} = (S, T, \phi)$. With notation as above, it is clear that each f_i is a perturbation of the corresponding double (orange) strand for $\mathbb{X}$. Reasoning as above, we see that

$$M(f_{\mathbb{X}}) = \mathrm{inv}(\phi) - \mathrm{inv}(\phi, \omega) + \mathrm{inv}(\omega) = \mathrm{inv}(\xi^{-1}) - \mathrm{inv}(\xi^{-1}, \omega) + \mathrm{inv}(\omega).$$

Next,

$$\begin{aligned}A(f_{\mathbb{X}}) &= \tfrac{1}{2}\bigl(\operatorname{inv}(\phi,\xi^{-1})-\operatorname{inv}(\phi,\omega)+\operatorname{inv}(\omega)-\operatorname{inv}(\xi^{-1})-|T_{\mathbb{X}}|\bigr)\\ &= \tfrac{1}{2}\bigl(2\operatorname{inv}(\xi^{-1})+|T_{\mathbb{X}}|-\operatorname{inv}(\xi^{-1},\omega)+\operatorname{inv}(\omega)-\operatorname{inv}(\xi^{-1})-|T_{\mathbb{X}}|\bigr)\\ &= \tfrac{1}{2}\bigl(\operatorname{inv}(\xi^{-1})-\operatorname{inv}(\xi^{-1},\omega)+\operatorname{inv}(\omega)\bigr)\\ &= \tfrac{1}{2}M(f_{\mathbb{X}}).\end{aligned}$$

The proof for $f_{\mathbb{X}}^*$, $f_{\mathbb{X}}'$ and $f_{\mathbb{X}}'^*$ is analogous. □

We are now ready to prove Theorem 6.1.

Proof of Theorem 6.1 Each shadow $\mathcal{P}_i^\circ$ has a corresponding grid diagram G_i°. Both for grids and for shadows, we abbreviate the notation for the (bi)modules CTA^-, $CDTD^-$, etc by CT^-. For shadows and the corresponding grids we consider the type A or type AA structures, and for mirror-shadows and the corresponding grids we consider the type D or type DD structures. By Propositions 4.7 and 4.11, the modules $CT^-(\mathcal{P}_i^\circ)$ and $CT^-(G_i^\circ)$ are isomorphic. The type A or AA structures $CT^-(G_i^\circ)$ are defined by counting empty rectangles and certain sets of half-rectangles that do not intersect $\mathbb{X}$, whereas the type D or DD structures are defined by counting empty rectangles and (individual) half-rectangles that do not intersect $\mathbb{X}$. So the differential on $CT^-(G_1^\circ)\boxtimes\cdots\boxtimes CT^-(G_n^\circ)$ counts empty rectangles in the diagram $G_1^\circ\cup\cdots\cup G_n^\circ$ that do not intersect $\mathbb{X}$, hence $CT^-(G_1^\circ)\boxtimes\cdots\boxtimes CT^-(G_n^\circ)$ is isomorphic to the complex $gCFK^-$ associated to the closure of the nice diagram $G_1^\circ\cup\cdots\cup G_n^\circ$, with an X and an O added in the new region, which represents $L\cup U$. It remains to check that this last isomorphism preserves the Maslov and Alexander gradings.

Let $\mathcal{H}$ be the Heegaard diagram obtained by closing up the plumbing of annuli $G_1^\circ\cup\cdots\cup G_n^\circ$. We argue that the absolute Maslov grading on $\mathcal{H}$ (obtained by adding the gradings on each G_i°) is correct. Let k_i be the number of Os in each grid G_i°, and let $k=\sum_{i=1}^n k_i$. Let $\boldsymbol{x}_{\mathbb{O}}=f_{\mathbb{O}_1}^*\boxtimes f_{\mathbb{O}_2}\boxtimes f_{\mathbb{O}_3}'^*\boxtimes f_{\mathbb{O}_4}'\boxtimes f_{\mathbb{O}_5}^*\boxtimes\cdots\boxtimes f_{\mathbb{O}_n}^\circ$ (the decoration $\circ$ depends on n mod 4, as specified according to the first four factors). By Lemma 6.2, $M(\boldsymbol{x}_{\mathbb{O}})=M(f_{\mathbb{O}_1}^*)+M(f_{\mathbb{O}_2})+\cdots+M(f_{\mathbb{O}_n}^\circ)=-|\mathbb{O}_1|-\cdots-|\mathbb{O}_n|=-k$.

Form a set of γ–circles $\boldsymbol{\gamma}$ by performing handleslides (which are allowed to cross $\mathbb{X}$ but not $\mathbb{O}$) of k_i of the β–circles and a perturbation of one β–circle for each G_i, as in Figure 33. We look at the holomorphic triangle map (see [19; 17]) associated to $(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\mathbb{O})$. Observe that $(\Sigma,\boldsymbol{\beta},\boldsymbol{\gamma},\mathbb{O})$ is a diagram for $(S^1\times S^2)^{\#k}$, and let Θ be the top-dimensional generator (on the diagram this is the set of intersection points at which the small bigons start). Let $\boldsymbol{y}$ be the generator of $(\Sigma,\boldsymbol{\alpha},\boldsymbol{\gamma},\mathbb{O})$ nearest to $\boldsymbol{x}_{\mathbb{O}}$. There is a holomorphic triangle that maps $\boldsymbol{x}_{\mathbb{O}}\otimes\Theta$ to $\boldsymbol{y}$; see Figure 33.

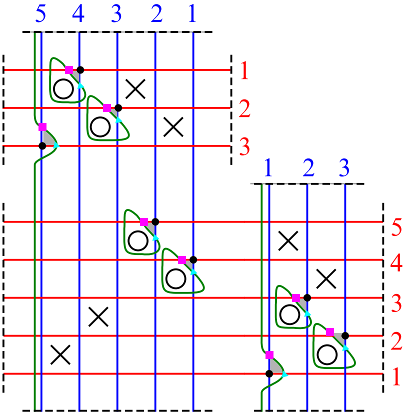


Figure 33: The union $\mathcal{H}$ of three grid diagrams: G_1^* (top), G_2 (bottom left) and G_3^* (bottom right). The black dots form the generator $\boldsymbol{x}_{\mathbb{O}}$, the purple squares form $\boldsymbol{y}$ and the cyan triangles form Θ.

Observe that $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbb{O})$ is a diagram for S^3 with 2^k generators, for which the differential vanishes (each small bigon ending at an intersection point in $\boldsymbol{y}$ is canceled by the corresponding horizontal annulus with the small region containing an O removed). By looking at the small bigons, one sees that $\boldsymbol{y}$ is the bottom-most generator of $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbb{O})$, so its Maslov grading is $-k + \frac{1}{2}l$, where $l = |L| = |L \cup U| - 1$. Since $\boldsymbol{x}_{\mathbb{O}}$, Θ, and $\boldsymbol{y}$ are connected by a Maslov index zero triangle, the Maslov grading of $\boldsymbol{x}_{\mathbb{O}}$ should be $M'(\boldsymbol{x}_{\mathbb{O}}) = -k + \frac{1}{2}l$ too.

Next, we argue that the Alexander grading on $\mathcal{H}$ is correct. For that purpose, let $\boldsymbol{x}_{\mathbb{X}} = f^*_{\mathbb{X}_1} \boxtimes f_{\mathbb{X}_2} \boxtimes f'^*_{\mathbb{X}_3} \boxtimes f'_{\mathbb{X}_4} \boxtimes \cdots \boxtimes f^\circ_{\mathbb{X}}$. A priori, $A'(\boldsymbol{x}_{\mathbb{X}}) = A(\boldsymbol{x}_{\mathbb{X}}) + s = A(f^*_{\mathbb{X}_1}) + \cdots + A(f^\circ_{\mathbb{X}_n}) + s$, where s is a constant. We show the shift s is zero. By Lemma 6.2,

$$\begin{aligned} A(\boldsymbol{x}_{\mathbb{X}}) &= A(f^*_{\mathbb{X}_1}) + A(f_{\mathbb{X}_2}) + \cdots + A(f^\circ_{\mathbb{X}_n}) \\ &= \tfrac{1}{2}M(f^*_{\mathbb{X}_1}) + \tfrac{1}{2}M(f_{\mathbb{X}_2}) + \cdots + \tfrac{1}{2}M(f^\circ_{\mathbb{X}_n}) = \tfrac{1}{2}M(\boldsymbol{x}_{\mathbb{X}}), \end{aligned}$$

and we just showed that $M \equiv M' - \frac{1}{2}l$, so $A(\boldsymbol{x}_{\mathbb{X}}) = \frac{1}{2}\big(M'(\boldsymbol{x}_{\mathbb{X}}) - \frac{1}{2}l\big)$. On the other hand, using the holomorphic triangles argument above, we see that the $\mathbb{X}$–normalized grading of $\boldsymbol{x}_{\mathbb{X}}$ is $N'(\boldsymbol{x}_{\mathbb{X}}) = -k + \frac{1}{2}l$. The closed diagram has one additional X and one O in the outside region that we closed off, for a total of $k+1$ basepoints of each type, so, by (2),

$$A'(\boldsymbol{x}_{\mathbb{X}}) = \tfrac{1}{2}\big(M'(\boldsymbol{x}_{\mathbb{X}}) - N'(\boldsymbol{x}_{\mathbb{X}}) - ((k+1) - (l+1))\big) = \tfrac{1}{2}\big(M'(\boldsymbol{x}_{\mathbb{X}}) + \tfrac{1}{2}l\big),$$

so

$$A(\boldsymbol{x}_{\mathbb{X}}) = A'(\boldsymbol{x}_{\mathbb{X}}) - \tfrac{1}{2}l. \qquad \square$$

7 Matched circles and their algebras

Just as closed 3–manifolds and knots or links in closed 3–manifolds can be represented by Heegaard diagrams, and bordered 3–manifolds can be represented by bordered Heegaard diagrams, tangles in 3–manifolds with boundary can be represented by suitable Heegaard diagrams, which we will call bordered Heegaard diagrams for tangles.

We define two types of (multipointed) bordered Heegaard diagrams for tangles in 3–manifolds with one boundary component. The reason we need two slightly different diagrams is so the result after gluing is a valid closed Heegaard diagram for a link, with the same number of α–curves as β–curves, and with the correct number of basepoints (this should become apparent once the reader goes through the relevant definitions and examples). We also define Heegaard diagrams for tangles in 3–manifolds with two boundary components. We restrict our work to the case where all boundary components are spheres.

7.1 Matched circles

An n–marked sphere $\mathcal{S} = (S^2, t_1, \dots, t_n)$ has a compatible handle decomposition as follows:

- Start with $n+2$ two-dimensional 0–handles $h_0^0, \dots, h_{n+1}^0$, where the core of h_i^0 is t_i for $1 \le i \le n$.
- Attach 1–handles $h_1^1, \dots, h_{n+1}^1$ so that h_i^1 is attached to h_{i-1}^0 and h_i^0.
- Attach a 2–handle to the resulting boundary to obtain S^2.

As a first step towards building Heegaard diagrams for tangles, we represent marked spheres by matched circles. First we define matched circles even more generally.

Definition 7.1 A *marked matched circle* $\mathcal{Z}$ is a sextuple $(Z, \boldsymbol{a}, \mu, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ of

- an oriented circle Z;
- $2n+2$ points $\boldsymbol{a} = \{a_1, \dots, a_{2n+2}\}$ on Z labeled with order induced by the orientation on Z;
- a matching μ: $\boldsymbol{a} \to [n+1]$ (where $[n+1] := \{1, \dots, n+1\}$) such that surgery on Z along the matched pairs in $\boldsymbol{a}$ yields $n+2$ circles;
- two sets of points, $\mathbb{X} = \{X_1, \dots, X_k\}$ and $\mathbb{O} = \{O_1, \dots, O_l\}$, and a pair of points $\boldsymbol{z} = \{z^-, z^+\}$ in $Z \setminus \boldsymbol{a}$ such that there is exactly one point in each circle obtained after surgery on the matched pairs in $\boldsymbol{a}$, and so that one of the points in $\boldsymbol{z}$ is in the interval (a_{2n+2}, a_1).

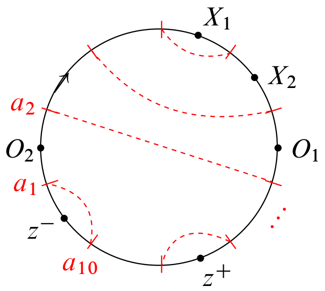


Figure 34: A marked matched circle. Here $n = 4$. The matching on $\boldsymbol{a}$ is illustrated schematically with dotted lines.

See, for example, Figure 34.

Given a marked matched circle $\mathcal{Z} = (Z, \boldsymbol{a}, \mu, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$, its *negative*, denoted $-\mathcal{Z}$, is the marked matched circle $\mathcal{Z}^*$ given by $(Z', \boldsymbol{a}', \mu', \mathbb{X}', \mathbb{O}', \boldsymbol{z}')$, where there is an orientation-reversing homeomorphism $f \colon Z \to Z'$ such that

- $f(\boldsymbol{a}) = \boldsymbol{a}'$ and $\mu = \mu' \circ f$,
- $f(z^+) = (z')^-$ and $f(z^-) = (z')^+$,
- $f(\mathbb{X}) = \mathbb{O}'$ and $f(\mathbb{O}) = \mathbb{X}'$.

In other words, $-\mathcal{Z}$ is obtained from $\mathcal{Z}$ by taking the mirror, swapping $\mathbb{X}$ and $\mathbb{O}$ and swapping z^+ and z^-. We will soon study Heegaard diagrams whose boundaries are marked matched circles, and gluing two diagrams along boundary components $\mathcal{Z}_1$ and $\mathcal{Z}_2$ will be allowed exactly when $\mathcal{Z}_1 = -\mathcal{Z}_2$.

A marked sphere $\mathcal{S} = (S^2, t_1, \dots, t_n)$ is represented by the following marked matched circle.

Definition 7.2 The *marked matched circle* $\mathcal{Z}(\mathcal{S})$ *associated to* $\mathcal{S}$ is given by the sextuple $(Z, \boldsymbol{a}, \mu, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ with $\boldsymbol{a} = \{a_1, \dots, a_{2n+2}\}$ and matching $\mu(a_i) = i = \mu(a_{2n+3-i})$ for $1 \le i \le 2n+1$. The set $\mathbb{X}$ consists of one point in each interval (a_i, a_{i+1}) on the circle Z, whenever t_i has positive orientation, and the set $\mathbb{O}$ consists of one point in each interval (a_i, a_{i+1}) on the circle Z, whenever t_i has negative orientation, for $1 \le i \le n$. The point z^- is in the interval between a_{2n+2} and a_1, and z^+ is in the interval (a_{n+1}, a_{n+2}).

See, for example, Figure 35.

We can recover the sphere $\mathcal{S}$ from $\mathcal{Z}(\mathcal{S})$ in the following way. We take a disk with boundary Z, attach 2–dimensional 1–handles along the matched pairs in $\boldsymbol{a}$, and fill the resulting $2n+2$ boundary components with 2–handles. We take $\{t_1, \dots, t_n\}$ to

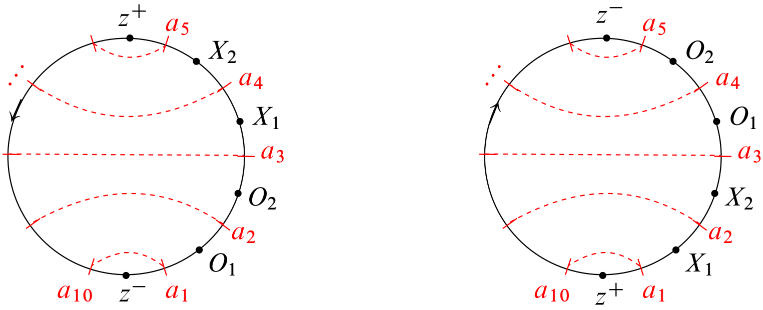


Figure 35: Examples of marked matched circles. Left: the marked matched circle $\mathcal{Z}(\mathcal{S})$ associated to $\mathcal{S} = (S^2, -, -, +, +)$. Right: the marked matched circle $\mathcal{Z}(\mathcal{S})^*$.

be the cores of the 2–handles that do not intersect (a_{2n+2}, a_1) and (a_{n+1}, a_{n+2}), and we orient t_i positively if the attaching circle for the corresponding 2–handle contains an $\mathbb{X}$ marking, and negatively if the attaching circle contains an $\mathbb{O}$ marking. This is the dual handle decomposition to the one described at the beginning of this section.

7.2 The algebra associated to a marked matched circle

Given a marked matched circle, we define an algebra similar to the algebras from [8; 28]. For marked matched circles associated to marked spheres, these algebras are precisely the ones from Section 3.1.4. The reason we give another description is that the interpretation in this section fits better with the geometric setup in the forthcoming sections. Below, we use the same notation as [8, Chapter 3] for our analogous structures, and caution the reader to remember that our matched circles are different from the ones in [8].

Definition 7.3 The *strands algebra* $\mathcal{A}(n,k,t)$ is a free $\mathbb{F}_2$–module generated by partial permutations $a = (S, T, \phi)$, where S and T are k–element subsets of the set $[2n+2] := \{1, \dots, 2n+2\}$ and $\phi\colon S \to T$ is a nondecreasing bijection such that $\phi(i) \le t$ if and only if $i \le t$. Let $\mathrm{Inv}(\phi)$ be the set of inversions of ϕ, ie the set of pairs $i, j \in S$ with $i < j$ and $\phi(j) < \phi(i)$, and $\mathrm{inv}(\phi) = \#\mathrm{Inv}(\phi)$. Multiplication on $\mathcal{A}(n,k,t)$ is given by

$$(S,T,\phi)\cdot(U,V,\psi) = \begin{cases} (S,V,\psi\circ\phi) & \text{if } T = U \text{ and } \mathrm{inv}(\phi)+\mathrm{inv}(\psi) = \mathrm{inv}(\psi\circ\phi), \\ 0 & \text{otherwise.} \end{cases}$$

For an inversion $c = (i,j)$ of ϕ, define ϕ_c by $\phi_c(i) = \phi(j)$, $\phi_c(j) = \phi(i)$, and $\phi_c(l) = \phi(l)$ for $l \ne i, j$. The differential on $\mathcal{A}(n,k,t)$ is given by

$$\partial(S,T,\phi) = \sum_{\substack{c\in\mathrm{Inv}(\phi)\\ \mathrm{inv}(\phi_c)=\mathrm{inv}(\phi)-1}} (S,T,\phi_c).$$

Compare with [8, Section 3.1.1]. We can represent a generator (S,T,ϕ) by a strands diagram of horizontal and upward-veering strands. Compare with [8, Section 3.1.2]. In this notation, the product becomes concatenation, where double crossings are set to zero. The differential corresponds to resolving crossings, subject to the same double crossing rule.

The ring of idempotents $\mathcal{I}(n,k,t) \subset \mathcal{A}(n,k,t)$ is generated by all elements of the form $I(S) := (S,S,\mathrm{id}_S)$ where S is a k–element subset of $[2n+2]$.

Fix a marked matched circle $\mathcal{Z} = (Z, \boldsymbol{a}, \mu, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ with $|\boldsymbol{a}| = 2n+2$. Recall that one of the points in $\boldsymbol{z}$ is on the interval (a_{2n+2}, a_1), and let t be the number for which the other point in $\boldsymbol{z}$ is on the interval (a_t, a_{t+1}).

If we forget the matching on the circle for a moment we can view $\mathcal{A}(n,t) = \bigoplus_i \mathcal{A}(n,i,t)$ as the algebra generated by certain sets of Reeb chords in $(Z \setminus \boldsymbol{z}, \boldsymbol{a})$: We can view a set $\boldsymbol{\rho}$ of Reeb chords, no two of which share initial or final endpoints, as a strands diagram of upward-veering strands. For such a set $\boldsymbol{\rho}$, we define the *strands algebra element associated to* $\boldsymbol{\rho}$ to be the sum of all ways of consistently adding horizontal strands to the diagram for $\boldsymbol{\rho}$, and we denote this element by $a_0(\boldsymbol{\rho}) \in \mathcal{A}(n,t)$. The basis over $\mathbb{F}_2$ from Definition 7.3 is in this terminology the nonzero elements of the form $I(S)a_0(\boldsymbol{\rho})$, where $S \subset \boldsymbol{a}$.

For a subset $\boldsymbol{s}$ of $[n+1]$, a *section* of $\boldsymbol{s}$ is a set $S \subset \mu^{-1}(\boldsymbol{s})$ such that μ maps S bijectively to $\boldsymbol{s}$. To each $\boldsymbol{s} \subset [n+1]$ we associate an idempotent in $\mathcal{A}(n,t)$ given by

$$I(\boldsymbol{s}) = \sum_{S \text{ is a section of } \boldsymbol{s}} I(S).$$

Let $\mathcal{I}(\mathcal{Z})$ be the subalgebra generated by all $I(\boldsymbol{s})$, and let $\boldsymbol{I} = \sum_{\boldsymbol{s}} I(\boldsymbol{s})$.

Definition 7.4 The *algebra* $\mathcal{A}(\mathcal{Z})$ is the subalgebra of $\mathcal{A}(n,t)$ generated (as an algebra) by $\mathcal{I}(\mathcal{Z})$ and by all $a(\boldsymbol{\rho}) := \boldsymbol{I} a_0(\boldsymbol{\rho}) \boldsymbol{I}$. We refer to $a(\boldsymbol{\rho})$ as the *algebra element associated to* $\boldsymbol{\rho}$.

Note that this definition, which is what we use for the tilde version of our invariants, does not take into account the $\mathbb{X}$ and $\mathbb{O}$ labels on $\mathcal{Z}$.

The nonzero elements $I(\boldsymbol{s})a(\boldsymbol{\rho})$ form a basis for $\mathcal{A}(\mathcal{Z})$ over $\mathbb{F}_2$. Note that for a nonzero generator $I(\boldsymbol{s})a(\boldsymbol{\rho})$, there is a unique primitive idempotent $I(\boldsymbol{t})$ such that $I(\boldsymbol{s})a(\boldsymbol{\rho}) = I(\boldsymbol{s})a(\boldsymbol{\rho})I(\boldsymbol{t})$. We can represent a generator $I(\boldsymbol{s})a(\boldsymbol{\rho})$ by a strands diagram by adding dashed horizontal strands to the strands diagram for $\boldsymbol{\rho}$, one for each horizontal strand that appears in the expansion of $I(\boldsymbol{s})a(\boldsymbol{\rho})$ as a sum of elements of $\mathcal{A}(n,t)$.

As a special case, let $\mathcal{Z}(\mathcal{S}) = (Z, \boldsymbol{a}, \mu, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ be a marked matched circle for a marked sphere $\mathcal{S}$, with $|\boldsymbol{a}| = 2n+2$. Recall the definition of a shadow (Definition 3.1), and let $\mathcal{E}$ be the idempotent shadow corresponding to the interval of $\mathcal{Z}(\mathcal{S})$ containing $a_1, \dots, a_{n+1}$, ie $(n+1, n+1, \mathrm{id}_{S_\mathbb{X}}, \mathrm{id}_{S_\mathbb{O}})$ where

$$S_\mathbb{O} = \left\{s + \tfrac{1}{2} \mid \text{there is an } X \text{ between } a_s \text{ and } a_{s+1}\right\}$$

and $S_\mathbb{X} = \left\{1\frac{1}{2}, \dots, n\frac{1}{2}\right\} \setminus S_\mathbb{O}$. Recall the definition of the algebra $\mathcal{A}(\mathcal{E})$ from Section 3.1.4. Let $\widehat{\mathcal{A}}(\mathcal{E}) := \mathcal{A}(\mathcal{E})/(U_i{=}0)$ be the algebra obtained from $\mathcal{A}(\mathcal{E})$ after setting all U_i to zero.

Proposition 7.5 *For $\mathcal{E}$ and $\mathcal{Z}(\mathcal{S})$ as above, the algebras $\widehat{\mathcal{A}}(\mathcal{E})$ and $\mathcal{A}(\mathcal{Z}(\mathcal{S}))$ are isomorphic.*

Proof As long as we do not need to keep track of the bigrading, we can think of $\widehat{\mathcal{A}}(\mathcal{E})$ simply as the algebra $\mathcal{A}(\widehat{\mathcal{E}})$ for the shadow $\widehat{\mathcal{E}} = (n+1, n+1, \mathrm{id}_{S_\mathbb{X}}, \mathrm{id}_{S_\mathbb{O}})$, where $S_\mathbb{X} = \left\{1\frac{1}{2}, \dots, n\frac{1}{2}\right\}$ and $S_\mathbb{O} = \varnothing$.

We first outline the correspondence of generators. Suppose (S, T, ϕ) is a generator for $\widehat{\mathcal{A}}(\mathcal{E})$. The corresponding element $I(\boldsymbol{s})a(\boldsymbol{\rho}) \in \mathcal{A}(\mathcal{Z}(\mathcal{S}))$ has starting idempotent $\boldsymbol{s} = S$ and the following set of Reeb chords $\boldsymbol{\rho}$: the Reeb chord from i to $\phi(i)$ if $\phi(i) > i$, and the Reeb chord from $2n+3-i$ to $2n+3-\phi(i)$ if $\phi(i) < i$.

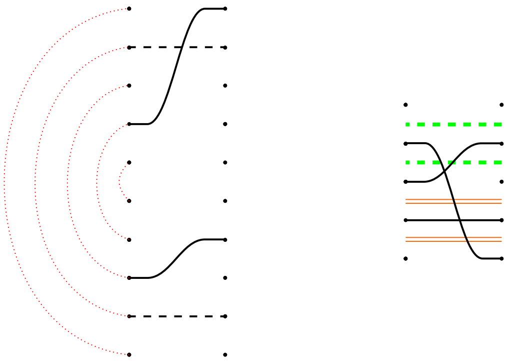


Figure 36: Example of a generator of $\mathcal{A}(\mathcal{Z})$, where $\mathcal{Z}$ is the circle in Figure 35 (left), and the corresponding generator of $\mathcal{A}(\mathcal{E})$ for the idempotent shadow $\mathcal{E}$ associated to $\mathcal{Z}$ (right)

Note that since there is a double (orange) line at every half-integer height in the diagram of $\widehat{\mathcal{E}}$, the concatenation of two strand diagrams is automatically zero whenever an upward-veering and a downward-veering strand are concatenated. Thus, the concatenation of two strand diagrams in $\widehat{\mathcal{A}}(\mathcal{E})$ is nonzero exactly when it is nonzero for the corresponding generators in $\mathcal{A}(\mathcal{Z}(\mathcal{S}))$.

The differential of $\widehat{\mathcal{A}}(\mathcal{E})$ is obtained by summing over all the ways of resolving a crossing, where resulting double crossings are set to zero. Again having a double line at every half-integer height means that resolving crossings between an upward-veering strand and a downward-veering strand is no longer allowed. The allowed resolutions are only those of crossings between two upward-veering strands, two downward-veering strands, an upward-veering and a horizontal strand, or a downward-veering and a horizontal strand. The first two kinds correspond to resolving a crossing between two Reeb chords in the lower half or upper half of a strand diagram, respectively, and the other two kinds correspond to resolving a crossing between a Reeb chord in the lower half, respectively upper half, of a strand diagram and a horizontal strand in a section of $\boldsymbol{s}$. □

8 Heegaard diagrams

We represent tangles by a type of Heegaard diagrams, which we call *multipointed bordered Heegaard diagrams for tangles*, or just tangle Heegaard diagrams. In a sense, our work in this section is a variation of the bordered Heegaard diagrams from [8; 9], and many of the statements we make and their proofs are analogous to the ones in [8; 9]. We have tried to provide detailed references, and we also encourage the reader to compare our subsections with the corresponding ones in [8, Chapter 4; 9, Chapter 5].

8.1 3–manifolds with one boundary component

Definition 8.1 A *type* 1 *multipointed bordered Heegaard diagram for a tangle*, or simply a type 1 tangle Heegaard diagram, is a sextuple $\mathcal{H}=(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta},\mathbb{X},\mathbb{O},\boldsymbol{z})$ where

- Σ is a compact surface of genus g with one boundary component;
- $\boldsymbol{\alpha}=\{\alpha_1^a,\dots,\alpha_{2n+1}^a,\alpha_1^c,\dots,\alpha_t^c\}$ is a set of pairwise disjoint, embedded curves: $2n+1$ arcs, each with boundary on $\partial\Sigma$, and t closed curves in the interior of Σ;
- $\boldsymbol{\beta}$ is a set of $t+n$ pairwise disjoint curves embedded in the interior of Σ;
- $\mathbb{X}$ and $\mathbb{O}$ are two $(t+2n-g)$–tuples of points in $\Sigma\setminus(\boldsymbol{\alpha}\cup\boldsymbol{\beta})$;
- $\boldsymbol{z}=\{z^-,z^+\}$ is a set of two oppositely oriented points on $\partial\Sigma\setminus\boldsymbol{\alpha}$;

subject to the following conditions:

- $\boldsymbol{\beta}$ spans a g–dimensional subspace of $H_1(\Sigma;\mathbb{Z})$.
- $\{\alpha_1^c,\dots,\alpha_t^c\}$ span a g–dimensional subspace of $H_1(\Sigma;\mathbb{Z})$, and along with the arcs, $\boldsymbol{\alpha}$ span a $g+1$–dimensional subspace of $H_1(\Sigma,\partial\Sigma;\mathbb{Z})$.

- $\{\alpha_1^a, \dots, \alpha_{2n+1}^a\}$ induce a concentric matching on $\partial\Sigma$. Specifically, they are labeled so that we can order the points on $\partial\boldsymbol{\alpha}$ according to the orientation of $\partial\Sigma$ as $a_1, \dots, a_{4n+2}$ so that $\partial\alpha_i^a = \{a_i, a_{4n+3-i}\}$.
- z^- lies in the interior of the segment with boundary a_{4n+2} and a_1 of $\partial\Sigma \setminus \boldsymbol{\alpha}$, and z^+ lies on the segment with boundary a_{2n+1} and a_{2n+2}.
- Each of the $t-g$ components of $\Sigma \setminus \boldsymbol{\alpha}$ that do not meet $\partial\Sigma$ contains one $X \in \mathbb{X}$ and one $O \in \mathbb{O}$, and each of the $2n$ components of $\Sigma \setminus \boldsymbol{\alpha}$ that contain two segments of $\partial\Sigma \setminus \boldsymbol{\alpha}$ contains either an X in the interior and an O on the segment of $\partial\Sigma \setminus \boldsymbol{\alpha}$ with the lower indexed endpoints, or an O in the interior and an X on the segment of $\partial\Sigma \setminus \boldsymbol{\alpha}$ with the lower indexed endpoints.
- Each of the $t+n-g$ components of $\Sigma \setminus \boldsymbol{\beta}$ that do not meet $\partial\Sigma$ contains exactly one X and one O. The unique component of $\Sigma \setminus \boldsymbol{\beta}$ that meets $\partial\Sigma$ contains n Xs and n Os on $\partial\Sigma$.

Figure 37 is an example of a type 1 Heegaard diagram for a tangle.

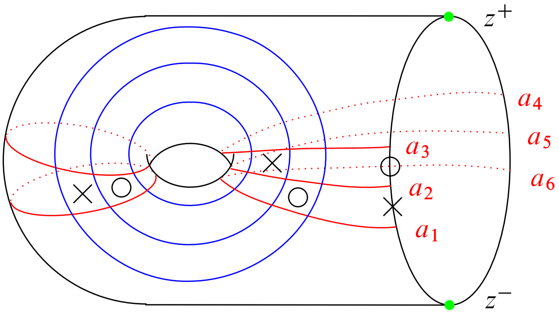


Figure 37: A type 1 tangle Heegaard diagram

A type 1 tangle Heegaard diagram gives rise to a pair $(Y, \mathcal{T})$, where Y is a 3–manifold with $\partial Y \cong S^2$ and $\mathcal{T}$ is marked $2n$–tangle in Y. We outline the topological construction below.

Let $\mathcal{S}$ be the marked sphere associated to $(Y, \mathcal{T})$. Note that $\partial\mathcal{H} \cong \mathcal{Z}(\mathcal{S})$, so we begin by building $\mathcal{S}$ from $\mathcal{Z}(\mathcal{S})$. Next, let $[-\epsilon, 0] \times Z$ be a collar neighborhood of $\partial\Sigma$, so that $\{0\} \times Z$ is identified with $\partial\Sigma$. Choose a neighborhood $Z \times [1,2]$ of Z in $\mathcal{S}$, so that $Z \times \{2\}$ is in the interior of the 0–handle from the decomposition described right after Definition 7.2. Glue $\Sigma \times [1,2]$ to $[-\epsilon, 0] \times \mathcal{S}$ so that the respective submanifolds $([-\epsilon, 0] \times Z) \times [1,2]$ and $[-\epsilon, 0] \times (Z \times [1,2])$ are identified. Call the resulting 3-manifold Y_0.

Now attach a 3–dimensional 2–handle to each $\beta_i \times \{2\} \subset \partial Y_0$ and to each $\alpha_i^c \times \{1\} \subset \partial Y_0$ to obtain a manifold Y_1. Next, join each $\alpha_i^a \times \{1\}$ to the core of the corresponding handle in $\{-\epsilon\} \times \mathcal{S}$ along their boundary to form a circle, and attach a 2–handle to each such circle. The resulting manifold, call it Y_2, has the following boundary components:

- $t+n-g$ spheres which meet $\Sigma \times \{2\}$ but do not meet $\{-\epsilon\} \times \mathcal{S}$.
- A sphere which meets both $\Sigma \times \{2\}$ and $\{-\epsilon\} \times \mathcal{S}$.
- $t-g$ spheres which meet $\Sigma \times \{1\}$ but do not meet $\{-\epsilon\} \times \mathcal{S}$.
- $2n$ spheres which meet both $\Sigma \times \{1\}$ and $\{-\epsilon\} \times \mathcal{S}$ but do not meet $\{-\epsilon\} \times \boldsymbol{z} \subset \{-\epsilon\} \times \mathcal{S}$.
- A sphere which meets both $\Sigma \times \{1\}$ and $(-\epsilon, z^-) \in \{-\epsilon\} \times \mathcal{S}$, and a sphere which meets both $\Sigma \times \{1\}$ and $(-\epsilon, z^+) \in \{-\epsilon\} \times \mathcal{S}$.
- The sphere $\{0\} \times \mathcal{S} \subset [-\epsilon, 0] \times \mathcal{S}$.

Glue 3–balls to all but the last sphere. Call the result Y.

Last, we construct a tangle $\mathcal{T} \subset Y$. Draw arcs from the Xs to the Os in $(\Sigma \setminus \boldsymbol{\beta}) \times \{\frac{3}{2}\}$, and push the interiors of the arcs into $(\Sigma \setminus \boldsymbol{\beta}) \times (\frac{3}{2}, 2]$. Draw arcs from Os to Xs in $(\Sigma \setminus \boldsymbol{\alpha}) \times \{\frac{3}{2}\}$. The union of all arcs is an oriented, marked $2n$–tangle, where the marking, ie the ordering on $\partial\mathcal{T} \subset \partial Y$ comes from the order in which those Xs and Os that are on $\partial\Sigma$ appear along $(a_1, a_{2n}) \subset Z \times \{\frac{3}{2}\} \subset \mathcal{S}$. Observe that drawing an arc from z^- to z^+ in $(\Sigma \setminus \boldsymbol{\beta}) \times \{\frac{3}{2}\}$ produces a 1–component tangle which is unlinked from $\mathcal{T}$, and, together with an arc in the 3–handle that was glued to the sphere which meets both $\Sigma \times \{2\}$ and $\{-\epsilon\} \times \mathcal{S}$, it bounds a disk away from $\mathcal{T}$ that lies entirely in that 3–handle. See, for example, Figure 38.

Definition 8.2 Given a marked sphere $\mathcal{S} = (S^2, t_1, \dots, t_n)$, we say that a Morse function f on S^2 (with an implicit choice of a Riemannian metric g) is *compatible with* $\mathcal{S}$ if

(1) $t_1, \dots, t_n$ are index 0 critical points of f;

(2) f has $n+2$ index 0 critical points in total, t_0, $t_1, \dots, t_n$, t_{n+1};

(3) f has $n+1$ index 1 critical points $p_1, \dots, p_{n+1}$, with p_i flowing down to t_{i-1} and t_i;

(4) f has a unique index 2 critical point.

Definition 8.3 Given a tangle $(Y, \mathcal{T})$, we say that a self-indexing Morse function f on Y (with an implicit choice of a Riemannian metric g) is *compatible with* $(Y, \mathcal{T})$ if

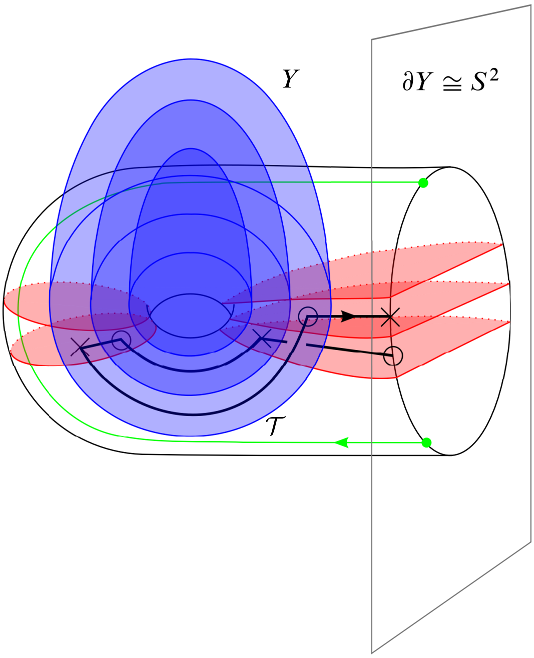


Figure 38: Building a tangle $(Y, \mathcal{T})$ from a Heegaard diagram

(1) ∂Y is totally geodesic, ∇f is parallel to ∂Y, $f|_{\partial Y}$ is a Morse function compatible with $\mathcal{S}$, and $f|_T$ is a Morse function, where $T \subset Y$ is the underlying 1–manifold for the marked tangle $\mathcal{T}$;

(2) the index 1 critical points for ∂Y are also index 1 critical points for Y;

(3) the index 0 critical points for T, along with the two additional index 0 critical points for ∂Y, are precisely the index 0 critical points for Y;

(4) the index 1 critical points for T, along with the index 2 critical point for ∂Y, are precisely the index 3 critical points for Y.

Proposition 8.4 *Every pair $(Y, \mathcal{T})$ has a type 1 Heegaard diagram.*

Proof We describe a compatible Morse function. Choose a Morse function f' and metric g' on T which takes value 0 on ∂T and is self-indexing except that it takes value 3 on the index 1 critical points. Extend to a pair (f'', g'') on $T \cup \partial Y$, so that f'' is also self-indexing on ∂Y, except that it takes value 3 on index 2 critical points of ∂Y, and is compatible with $\mathcal{S}$. Extend f'' and g'' to f and g on a neighborhood of $T \cup \partial Y$ satisfying the conditions of Definition 8.3, and extend f and g arbitrarily to a Morse function and metric on the rest of Y.

Since Y is connected, the graph formed by flows between the index 0 and index 1 critical points is connected. In fact, since the flows from the index 1 critical points on ∂Y remain on ∂Y, it follows that every index 0 critical point of $Y' := Y \setminus \nu(T \cup \partial Y)$ is connected by an edge in this graph to an index 1 critical point of Y', so we modify f in the interior of Y' to cancel every index 0 critical point of Y' with an index 1 critical point of Y'. Similarly, we eliminate all index 3 critical points of Y'.

Finally, given these f and g, we construct a type 1 tangle Heegaard diagram. Start with Heegaard surface $\Sigma = f^{-1}\left(\frac{3}{2}\right)$, oriented as the boundary of $f^{-1}\left(\left[0, \frac{3}{2}\right]\right)$. Let $\boldsymbol{\alpha}$ be the set of points on Σ that flow down to the index 1 critical points, label the arcs $\boldsymbol{\alpha}^a$ and their endpoints compatibly with $\mathcal{S}$, and let $\boldsymbol{\beta}$ be the set of points on Σ which flow up to the index 2 critical points. Mark the positive intersections of $T \cap \Sigma$ with Os, and the negative intersections with Xs. Also place an X in each region (a_i, a_{i+1}) of $\partial\Sigma \setminus \boldsymbol{\alpha} \cup (a_1, a_{2n+1})$ if the points in that region flow down to a positive endpoint t_i of the tangle T, and an O if those points flow to a negative endpoint t_i of T. Finally, place a point labeled z^- in (a_{4n+2}, a_1), and a point z^+ in (a_{2n+1}, a_{2n+2}). □

The Morse theory construction implies the following proposition.

Proposition 8.5 *Any two type* 1 *tangle Heegaard diagrams for a given tangle* $(Y, \mathcal{T})$ *are related by a sequence of* Heegaard moves*:*

- *Isotopies of the* α*–curves and* β*–curves, not crossing* $\partial\Sigma \cup \mathbb{X} \cup \mathbb{O}$.
- *Handle slides of* α*–curves over* α*–circles and* β*–circles over* β*–circles.*
- *Index one/two stabilizations (and their inverses, destabilizations) in the interior of* Σ*: forming the connected sum with a torus with one* α*–circle and one* β*–circle meeting transversely in a single point.*
- *Index zero/three stabilizations (and their inverses, destabilizations) in the interior of* Σ*: replacing a neighborhood of an* X *with one* α*–circle and one* β*–circle, isotopic to each other and intersecting in two points, and adding an* O *in the middle of the three new regions, and an* X *in each of the new side regions, or replacing a neighborhood of an* O *with such* α*– and* β*–curves, along with an* X *in the middle new region, and an* O *in each side region* (see Figure 39).

Proof The proof follows from the Morse calculus used in the proofs of [21, Proposition 3.3; 8, Proposition 4.10]. □

We also define type 2 tangle Heegaard diagrams. The definition is slightly different from that of type 1 diagrams, so that when one glues a type 1 and a type 2 diagram that agree along the boundary, the resulting closed diagram is a valid Heegaard diagram for a link.

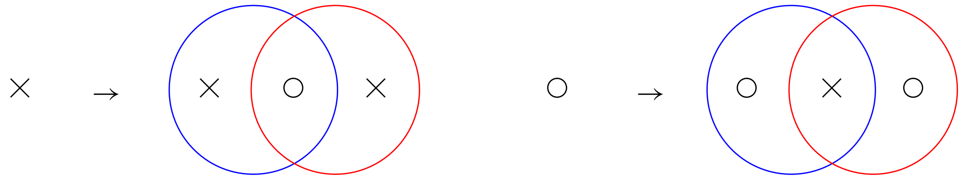


Figure 39: Index zero/three stabilization

Definition 8.6 A *type* 2 *multipointed bordered Heegaard diagram for a tangle* is a sextuple $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$, where

- Σ is a compact surface of genus g with one boundary component;
- $\boldsymbol{\alpha} = \{\alpha_1^a, \dots, \alpha_{2n+1}^a, \alpha_1^c, \dots, \alpha_t^c\}$ is a set of pairwise disjoint, embedded curves: $2n+1$ arcs, each with boundary on $\partial\Sigma$, and t closed curves in the interior of Σ;
- $\boldsymbol{\beta}$ is a set of $t+n+1$ pairwise disjoint curves embedded in the interior of Σ;
- $\mathbb{X}$ and $\mathbb{O}$ are two $(t+2n-g+1)$–tuples of points in $\Sigma \setminus (\boldsymbol{\alpha} \cup \boldsymbol{\beta})$;
- $\boldsymbol{z}$ is an oriented arc in $\Sigma \setminus (\boldsymbol{\alpha} \cup \boldsymbol{\beta})$ with boundary on $\partial\Sigma \setminus \boldsymbol{\alpha}$;

subject to the following conditions:

- $\boldsymbol{\beta}$ span a g–dimensional subspace of $H_1(\Sigma; \mathbb{Z})$.
- $\{\alpha_1^c, \dots, \alpha_t^c\}$ span a $(g-1)$–dimensional subspace of $H_1(\Sigma; \mathbb{Z})$, and along with the arcs, $\boldsymbol{\alpha}$ span a g–dimensional subspace of $H_1(\Sigma, \partial\Sigma; \mathbb{Z})$.
- $\{\alpha_1^a, \dots, \alpha_{2n+1}^a\}$ induce a concentric matching on $\partial\Sigma$, and they are labeled so that we can order the points on $\partial\boldsymbol{\alpha}$ according to the orientation of $-\partial\Sigma$ as $a_1, \dots, a_{4n+2}$ so that $\partial\alpha_i^a = \{a_i, a_{4n+3-i}\}$.
- $z^+ := \partial^+(z)$ lies in the interior of the segment with boundary a_{4n+2} and a_1 of $\partial\Sigma \setminus \boldsymbol{\alpha}$, and $z^- := \partial^-(z)$ lies on the segment with boundary a_{2n+1} and a_{2n+2}.
- Each of the $t-g+1$ components of $\Sigma \setminus \boldsymbol{\alpha}$ that do not meet $\partial\Sigma$ contains one $X \in \mathbb{X}$ and one $O \in \mathbb{O}$, and each of the $2n$ components of $\Sigma \setminus \boldsymbol{\alpha}$ that meet $\partial\Sigma$ but do not meet $\boldsymbol{z}$ contains either an X in the interior and an O on the segment of $\partial\Sigma \setminus \boldsymbol{\alpha}$ with the lower indexed endpoints, or an O in the interior and an X on the segment of $\partial\Sigma \setminus \boldsymbol{\alpha}$ with the lower indexed endpoints.
- Each of the $t+n-g+1$ components of $\Sigma \setminus \boldsymbol{\beta}$ that do not meet $\partial\Sigma$ contains exactly one X and one O. The unique component of $\Sigma \setminus \boldsymbol{\beta}$ that meets $\partial\Sigma$ contains n Xs and n Os on $\partial\Sigma$.

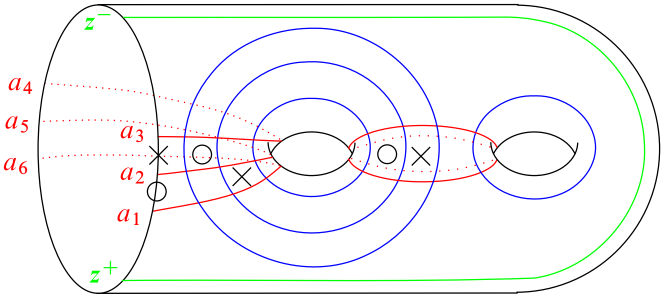


Figure 40: A type 2 tangle Heegaard diagram

Figure 40 is an example of a type 2 tangle Heegaard diagram.

A type 2 tangle Heegaard diagram gives rise to a pair $(Y, \mathcal{T})$ of a 3–manifold Y with $\partial Y \cong S^2$ and a marked tangle $\mathcal{T}$. The topological construction is similar to the one for a type 1 diagram.

We build the manifold Y_2 by following the type 1 construction, except this time $\partial\mathcal{H} \cong \mathcal{Z}(-\mathcal{S})^*$, where $\mathcal{S}$ is the marked sphere associated to $(Y, \mathcal{T})$. The difference in the types of boundary components of Y_2 is that there are now $t - g + 1$ spheres which meet $\Sigma \times \{1\}$ but do not meet $\{-\epsilon\} \times \mathcal{S}$, and there is one single sphere which meets both $\Sigma \times \{1\}$ and $\{-\epsilon\} \times \{\boldsymbol{z}^+, \boldsymbol{z}^-\} \subset \{-\epsilon\} \times \mathcal{S}$. We again glue 3–balls to all spheres except $\{0\} \times \mathcal{S}$ to obtain Y.

The tangle $\mathcal{T} \subset Y$ is again constructed by connecting the Xs and Os. This time its marking comes from the order in which the Xs and Os on $\partial\Sigma$ appear along $-\partial\Sigma$. The oriented arc $\boldsymbol{z} \times \left\{\frac{3}{2}\right\}$ is a 1–component boundary-parallel tangle which is unlinked from $\mathcal{T}$.

We cannot use Morse theory directly to prove the statements that follow. One way to explain where the problem lies is that if we start with a Morse function for ∂Y, then two index 0 critical points on ∂Y that would correspond to $\boldsymbol{z}^+$ and $\boldsymbol{z}^-$ belong to the same 0–handle in the handle decomposition for Y specified by $\mathcal{H}$.

Proposition 8.7 *Every $(Y, \mathcal{T})$ has a type 2 tangle Heegaard diagram.*

Proof Let $\mathcal{H}$ be a type 1 diagram for $(-Y, -\mathcal{T})$. We perform the following series of moves near the boundary of the diagram, as in Figure 41. Perform an index one/two stabilization near z^+ (Figure 41(b)). Denote the new α–circle by α', and the new β–circle by β'. Slide all α–arcs over α' so that now β' crosses them once each, near $a_1, \dots, a_{2n+1}$ (Figure 41(c)). Connect z^- to z^+ by an arc $\boldsymbol{z}$ that goes once over the new handle parallel to β' (Figure 41(d)). Remove α' (Figure 41(e)). Call the resulting diagram $\mathcal{H}'$. Observe that z does not intersect any α– or β–curves. The diagram $-\mathcal{H}'$ is a type 2 tangle Heegaard diagram for $(Y, \mathcal{T})$. □

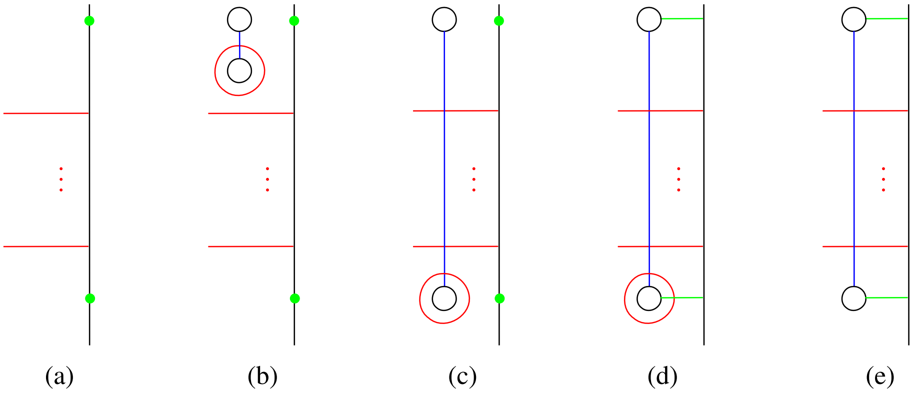


Figure 41: Transforming a type 1 diagram to a type 2 diagram

We will say that a type 2 diagram like $-\mathcal{H}'$, obtained from a type 1 diagram as above, is in *type* 1 *position*.

Proposition 8.8 *Any two type* 2 *tangle Heegaard diagrams for a given tangle* $(Y, \mathcal{T})$ *are related by a sequence of Heegaard moves:*

- *isotopies of the* α*–curves and* β*–curves, not crossing* $\partial\Sigma \cup \mathbb{X} \cup \mathbb{O} \cup \boldsymbol{z}$*;*
- *handle slides of* α*–curves over* α*–circles and* β*–circles over* β*–circles;*
- *index one/two stabilizations and destabilizations in the interior of* Σ*;*
- *index zero/three stabilizations and destabilizations in the interior of* Σ*.*

To prove this proposition, we make use of the following lemma.

Lemma 8.9 *Any type* 2 *diagram can be put in type* 1 *position.*

Proof Let $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ be a type 2 diagram for a pair $(Y, \mathcal{T})$. The idea is to find a curve on Σ which is disjoint from $\boldsymbol{\alpha}$, bounds a disk in the α–handlebody, and intersects $\boldsymbol{z}$ exactly once, and use it as a guide to modify the Heegaard diagram. We exhibit one such curve below.

Let $\alpha' \subset \Sigma$ be an embedded circle which is a push-off of the union of α^a_{2n+1} and $(a_{2n+1}, a_{2n+2}) \subset \partial\Sigma$ into Σ and does not intersect $\boldsymbol{\alpha}$, see Figure 42 and the more schematic first diagram in Figure 43. We will use α' as a guide while performing a series of Heegaard moves.

Note that α' bounds a disk in the α–handlebody. This disk is a push-off of the disk $D = D_1 \cup D_2$, where D_1 is the disk on ∂Y bounded by the interval (a_{2n+1}, a_{2n+2}) and the core of the 1–handle of ∂Y attached at a_{2n+1} and a_{2n+2}, and D_2 is the core of the 2–handle for α^a_{2n+1} from the construction of Y_2. So $\mathcal{H}' = (\Sigma, \boldsymbol{\alpha} \cup \alpha', \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ still specifies the same pair $(Y, \mathcal{T})$, or, to be more precise, $-(\Sigma, \boldsymbol{\alpha} \cup \alpha', \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \partial \boldsymbol{z})$ is a type 1 diagram for $(-Y, -\mathcal{T})$.

Also note that α' intersects $\boldsymbol{z}$ exactly once, near $\boldsymbol{z}^-$, so, since $\boldsymbol{\alpha} \cap \boldsymbol{z} = \varnothing$, no α–circle in $\boldsymbol{\alpha}$ is homologous to α' in $H_1(\Sigma; \mathbb{Z})$. This means that, after sliding α–curves over α' if necessary, we can draw $\mathcal{H}$ in the following way. Near the boundary we see $\partial\mathcal{H} \times [0, \epsilon)$, where $\partial\mathcal{H} \times \{0\}$ is the boundary of $\mathcal{H}$. There is a 1–handle for Σ with feet attached at $(\boldsymbol{z}^+, \epsilon)$ and $(\boldsymbol{z}^-, \epsilon)$, α' is a meridian of that 1–handle, $\boldsymbol{z}$ goes once over the handle. There may also be multiple β–curves going over the 1–handle. See Figure 43(b). We continue the proof with such more schematic pictures drawn in a plane.

We claim that, after an isotopy of $\boldsymbol{\beta}$ if necessary, there is some $\beta' \subset \boldsymbol{\beta}$ which intersects α' exactly once. Close $\boldsymbol{z}$ to a circle $\overline{\boldsymbol{z}}$ by connecting $\boldsymbol{z}^+$ to $\boldsymbol{z}^-$ along $\partial\Sigma$, going through $a_1, \dots, a_{2n+1}$. Since α' and $\overline{\boldsymbol{z}}$ are two circles on Σ intersecting transversely in one point, the neighborhood of $\alpha' \cup \overline{\boldsymbol{z}}$ in Σ is a punctured torus T; see Figure 43(c). Note $\boldsymbol{\beta}$ spans a g–dimensional subspace of $H_1(\Sigma)$, so $\Sigma \setminus \boldsymbol{\beta}$ only contains genus 0 pieces, so there is at least one β–circle; pick one and call it β', cutting the punctured torus into a genus 0 surface. No β can intersect $\boldsymbol{z}$ or $\partial\Sigma$, so β' cannot intersect $\overline{\boldsymbol{z}}$. Thus $\beta' \cap T$ is homologous to $\overline{\boldsymbol{z}}$ in $H_1(T, \partial T)$, so it can be isotoped to only intersect α' once. If any other β–curves intersect α', slide them over β', so that β' is the only curve intersecting α'. Now the diagram near the boundary looks like what we described

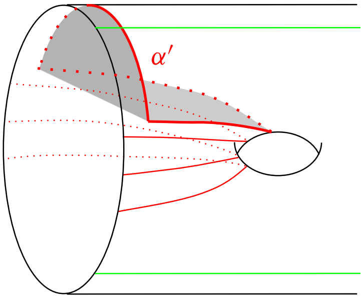


Figure 42: The circle α' and the disk it bounds

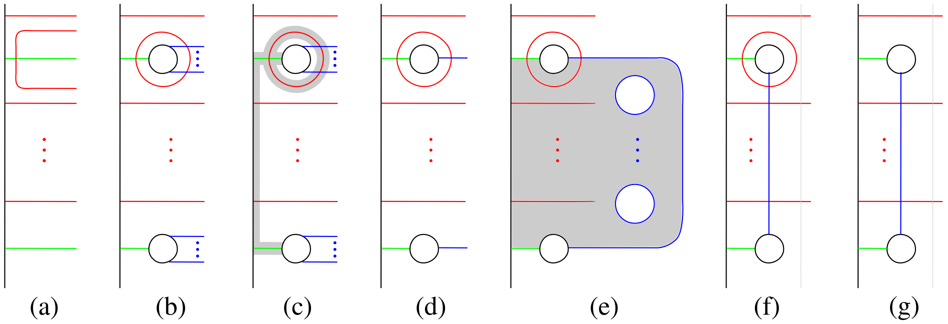


Figure 43: Putting a type 2 diagram in type 1 position. The last diagram is the mirror of the corresponding type 1 diagram.

in the previous paragraph, except there is exactly one β–curve going over the 1–handle. See Figure 43(d).

Since $\boldsymbol{\beta}$ spans a g–dimensional subspace of $H_1(\Sigma)$, all components of $\Sigma \setminus (\boldsymbol{\beta} \cup \boldsymbol{z})$ have genus zero. In particular, the region of $\Sigma \setminus (\boldsymbol{\beta} \cup \boldsymbol{z})$ that contains $\overline{z}$ is planar, with boundary components $\overline{z}$, β' and possibly some other β–circles. See Figure 43(e) (α–curves omitted from the picture away from the boundary). Slide β' over each β–circle in that region to move it close to the boundary of the diagram, ie so that it is a parallel push-off of $\overline{z}$ into the interior of Σ; see Figure 43(f). Remove α', which only served as a guide along the proof. See Figure 43(g). The resulting diagram is in type 1 position. □

Proof of Proposition 8.8 Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two type 2 diagrams for the same pair $(Y, \mathcal{T})$. By Lemma 8.9, both can be put in type 1 positions $\mathcal{H}_1'$ and $\mathcal{H}_2'$ by a sequence of the moves described in Proposition 8.8. Let $\mathcal{H}_1''$ and $\mathcal{H}_2''$ be the corresponding type 1 diagrams, so that $\mathcal{H}_i' = -\mathcal{H}_i''$ away from the boundary and the special 1–handle from Proposition 8.7.

Since $\mathcal{H}_1''$ and $\mathcal{H}_2''$ are related by a sequence of moves away from the boundaries, corresponding moves (the reflections of the original moves) can be performed between $\mathcal{H}_1'$ and $\mathcal{H}_2'$ away from the "neighborhood" of the boundary containing $\boldsymbol{z}$ and the special β–circle from the proof of Lemma 8.9, ie the β–circle shown in Figure 43(g). Thus, $\mathcal{H}_1$ and $\mathcal{H}_2$ are related by a sequence of Heegaard moves. □

8.2 3–manifolds with two boundary components

For a tangle in a manifold Y with $\partial Y \cong S^2 \amalg S^2$, we describe a Heegaard diagram with two boundary components. We will also want to keep track of a framed arc connecting the two boundary components of Y, by means of two arcs z_1 and z_2 that will connect the two boundary components of the Heegaard diagram.

Definition 8.10 A *multipointed bordered Heegaard diagram with two boundary components for a tangle* is a sextuple $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$, where

- Σ is a compact surface of genus g with two boundary components;
- $\boldsymbol{\alpha} = \{\alpha_1^0, \dots, \alpha_{m+1}^0, \alpha_1^1, \dots, \alpha_{n+1}^1, \alpha_1^c, \dots, \alpha_t^c\}$ is a set of pairwise disjoint, embedded curves: $m+n+2$ arcs (where m and n have the same parity), each with boundary on $\partial\Sigma$, and t closed curves in the interior of Σ;
- $\boldsymbol{\beta}$ is a set of $t + \frac{1}{2}(m+n) + 1$ pairwise disjoint curves embedded in the interior of Σ;
- $\mathbb{X}$ and $\mathbb{O}$ are two $(t+m+n-g+1)$–tuples of points in $\Sigma \setminus (\boldsymbol{\alpha} \cup \boldsymbol{\beta})$
- $\boldsymbol{z} = \{z_1, z_2\}$ is a set of two oriented arcs in $\Sigma \setminus (\boldsymbol{\alpha} \cup \boldsymbol{\beta})$ with boundary on $\partial\Sigma \setminus \boldsymbol{\alpha}$;

subject to the following conditions:

- $\boldsymbol{\beta}$ span a g–dimensional subspace of $H_1(\Sigma; \mathbb{Z})$.
- $\{\alpha_1^c, \dots, \alpha_t^c\}$ span a $(g-1)$–dimensional subspace of $H_1(\Sigma; \mathbb{Z})$, and along with the arcs, $\boldsymbol{\alpha}$ span a $(g+1)$–dimensional subspace of $H_1(\Sigma, \partial\Sigma; \mathbb{Z})$.
- $\{\alpha_1^0, \dots, \alpha_{m+1}^0\}$ induce a concentric matching on one component of $\partial\Sigma$, and they are labeled so that we can order their endpoints according to the orientation of $-\partial\Sigma$ as $a_1^0, \dots, a_{2m+2}^0$ so that $\partial\alpha_i^0 = \{a_i^0, a_{2m+3-i}^0\}$; $\{\alpha_1^1, \dots, \alpha_{n+1}^1\}$ induce a concentric matching on the other component of $\partial\Sigma$, and they are labeled so that we can order their endpoints according to the orientation of $\partial\Sigma$ as $a_1^1, \dots, a_{2n+2}^1$ so that $\partial\alpha_i^1 = \{a_i^1, a_{2n+3-i}^1\}$.
- $z_1^+ := \partial^+(z_1)$ lies in the interior of the segment with boundary a_{2m+2}^0 and a_1^0 of $\partial\Sigma \setminus \boldsymbol{\alpha}$, and $z_1^- := \partial^-(z_1)$ lies on the segment with boundary a_{2n+2}^1 and a_1^1; $z_2^+ := \partial^+(z_2)$ lies in the interior of the segment with boundary a_{m+1}^0 and a_{m+2}^0 of $\partial\Sigma \setminus \boldsymbol{\alpha}$, and $z_2^- := \partial^-(z_2)$ lies on the segment with boundary a_{n+1}^1 and a_{n+2}^1.
- Each of the $t-g+1$ components of $\Sigma \setminus \boldsymbol{\alpha}$ that do not meet $\partial\Sigma$ contains one $X \in \mathbb{X}$ and one $O \in \mathbb{O}$, and each of the $m+n$ components of $\Sigma \setminus \boldsymbol{\alpha}$ that meet $\partial\Sigma$ but do not meet $\boldsymbol{z}$ contains either an X in the interior and an O on the segment of $\partial\Sigma \setminus \boldsymbol{\alpha}$ with the lower indexed endpoints, or an O in the interior and an X on the segment of $\partial\Sigma \setminus \boldsymbol{\alpha}$ with the lower indexed endpoints.
- Each of the $t + \frac{1}{2}(m+n) - g + 1$ components of $\Sigma \setminus \boldsymbol{\beta}$ that do not meet $\partial\Sigma$ contains exactly one X and one O. The unique component of $\Sigma \setminus \boldsymbol{\beta}$ that meets $\partial\Sigma$ contains $\frac{1}{2}(m+n)$ Xs and $\frac{1}{2}(m+n)$ Os on $\partial\Sigma$.

We denote the component of $\partial\mathcal{H}$ containing α_i^0 by $\partial^0\mathcal{H}$, and the component of $\partial\mathcal{H}$ containing α_i^1 by $\partial^1\mathcal{H}$.

Figure 44 is an example of a tangle Heegaard diagram with two boundary components.

A tangle Heegaard diagram with two boundary components gives rise to a pair $(Y,\mathcal{T})$ of a 3–manifold Y with $\partial Y \cong S_0^2 \amalg S_1^2$ and a marked (m,n)–tangle $\mathcal{T}$, with $\partial^0\mathcal{T} \subset S_0^2$ and $\partial^1\mathcal{T} \subset S_1^2$. We describe the topological construction below.

Let $\mathcal{H}_{\mathrm{dr}}$ be the Heegaard diagram obtained from $\mathcal{H}$ by deleting a neighborhood of z_2 (this process, called *drilling*, was introduced in [9]). The boundary of this deleted neighborhood consists of the neighborhood z^0 of z_2^- on $\partial^0\mathcal{H}$, the neighborhood z^1 of z_2^+ on $\partial^1\mathcal{H}$, and two disjoint push-offs of z_2. Denote the push-off closer to a_{m+1}^0 by z_2^{front}, and the other one by z_2^{back}. The boundary of $\mathcal{H}_{\mathrm{dr}}$ is

$$\partial\mathcal{H}_{\mathrm{dr}} = (\partial^0\mathcal{H} \setminus z^0) \cup (\partial^1\mathcal{H} \setminus z^1) \cup z_2^{\mathrm{front}} \cup z_2^{\mathrm{back}}.$$

It inherits the decorations of $(\partial^0\mathcal{H} \setminus z^0)$ and $(\partial^1\mathcal{H} \setminus z^1)$. We also place a basepoint z^{front} on z_2^{front} and z^{back} on z_2^{back}.

If we ignore z^{front} and z^{back}, $\mathcal{H}_{\mathrm{dr}}$ looks like a type 2 diagram for an $(m+n)$–tangle, except that there is one extra α–arc.

We first build the pair $(Y_{\mathrm{dr}},\mathcal{T}_{\mathrm{dr}})$ for $\mathcal{H}_{\mathrm{dr}}$ as we would for any type 2 diagram. We obtain $(Y,\mathcal{T})$ from $(Y_{\mathrm{dr}},\mathcal{T}_{\mathrm{dr}})$ by attaching a 3–dimensional 2–handle to the boundary sphere along the connected sum annulus arising from the decomposition $\partial\mathcal{H}_{\mathrm{dr}} = \partial^0\mathcal{H} \# \partial^1\mathcal{H}$. More precisely, the attaching circle is the union of the two gradient flow lines from the index 2 critical point passing through z^{front} and z^{back}.

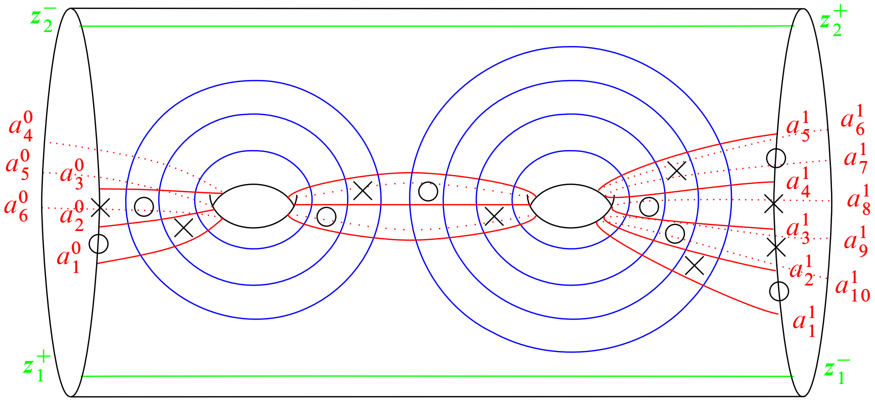


Figure 44: A tangle Heegaard diagram with two boundary components

Proposition 8.11 *Every $(Y, \mathcal{T})$ has a tangle Heegaard diagram with two boundary components.*

Proof The idea of the proof is the same as in the proof of [9, Proposition 5.8]. Choose an arc connecting $\partial^0 Y$ to $\partial^1 Y$ away from $\mathcal{T}$, and remove its neighborhood. Call the result $(Y_{\mathrm{dr}}, \mathcal{T}_{\mathrm{dr}})$, where the ordering on $\mathcal{T}_{\mathrm{dr}}$ inherits the ordering on $\partial^0\mathcal{T}$ concatenated with the reversed ordering on $\partial^1\mathcal{T}$. Let $\mathcal{H}' = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ be a type 2 diagram for $(Y_{\mathrm{dr}}, \mathcal{T}_{\mathrm{dr}})$. Add a parallel translate of α^a_{2m+1} by pushing it so that a_{2m+1} is pushed in the negative direction along $\partial\mathcal{H}'$, and call this curve α'. Call the resulting diagram $\mathcal{H}''$. Add a 1–handle to the two intervals of $\partial\Sigma \setminus \boldsymbol{\alpha}$ between α^a_{2m+1} and α'. The resulting surgery on $\partial\Sigma$ splits it into two circles. Denote the circle containing a_1 by $\partial^0\Sigma$, and the other circle by $\partial^1\Sigma$. Let z_2 be the cocore of the 1–handle, oriented from $\partial^0\Sigma$ to $\partial^1\Sigma$. Relabel $\boldsymbol{z}$ to z_1, α^a_i to α^0_i for $i \leq 2m+1$, α^a_i to $\alpha^1_{2m+2n+2-i}$ for $i > 2m+1$, and α' to α^1_{2n+1}. The resulting diagram $\mathcal{H}$ is a diagram for $(Y, \mathcal{T})$. Note that $\mathcal{H}'' = \mathcal{H}_{\mathrm{dr}}$. □

In the case of two boundary components, it is no longer true that any two diagrams for a pair $(Y, \mathcal{T})$ are related by Heegaard moves. However, if we keep better track of the parametrization of the boundary, we can still make this statement.

Definition 8.12 A *strongly marked (m, n)–tangle* $(Y, \mathcal{T}, \gamma)$ is a marked (m, n)–tangle $(Y, \mathcal{T})$ along with a framed arc γ connecting $\partial^0 Y$ to $\partial^1 Y$ in the complement of $\mathcal{T}$ such that γ and its framing λ_γ have ends on the equators of the two marked spheres, and we see $-\partial^0\mathcal{T}$, $-\partial^0\gamma$, $-\partial^0\lambda_\gamma$ and $\partial^1\mathcal{T}$, $\partial^1\gamma$, $\partial^1\lambda_\gamma$ in this order along each equator.

We say that a diagram $\mathcal{H}$ is compatible with a strongly marked tangle $(Y, \mathcal{T}, \gamma)$ if $\mathcal{H}$ describes $(Y, \mathcal{T})$, and after building $(Y, \mathcal{T})$ from $\mathcal{H}$, the arc z_1 with the framing that points into the β–handlebody yields γ.

Proposition 8.13 *If $\mathcal{H}$ and $\mathcal{H}'$ specify the same triple $(Y, \mathcal{T}, \gamma)$, then they are related by a sequence of Heegaard moves like the ones described in Proposition 8.8.*

Proof Let $\mathcal{H}_{\mathrm{dr}}$ and $\mathcal{H}'_{\mathrm{dr}}$ be the corresponding drilled diagrams. By Proposition 8.8, they are related by a sequence of moves away from the boundary of the Heegaard surface, hence away from the drilling region. Performing the inverse of the drilling operation to each diagram along the way provides a sequence of moves between $\mathcal{H}$ and $\mathcal{H}'$. □

8.3 Generators

Fix a tangle Heegaard diagram $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ of some genus g for some pair $(Y, \mathcal{T})$. Let $k := |\boldsymbol{\beta}|$.

Definition 8.14 A *generator* of $\mathcal{H}$ is a k–element subset $\boldsymbol{x} = \{x_1, \dots, x_k\}$ of points in $\boldsymbol{\alpha} \cap \boldsymbol{\beta}$ such that there is exactly one point on each β–circle, exactly one point on each α–circle and at most one point on each α–arc.

We denote the set of generators of $\mathcal{H}$ by $\mathfrak{S}(\mathcal{H})$, or simply by $\mathfrak{S}$ when $\mathcal{H}$ is fixed.

If $\mathcal{H}$ is a diagram for a $2n$–tangle, then let $o(\boldsymbol{x}) := \{i \mid \boldsymbol{x} \cap \alpha_i^a \neq \varnothing\}$ and $\overline{o}(\boldsymbol{x}) := [2n+1] \setminus o(\boldsymbol{x})$ for a generator $\boldsymbol{x} \in \mathfrak{S}$. Even though $o(\boldsymbol{x})$ and $\overline{o}(\boldsymbol{x})$ are really index sets, we often refer to them as the set of α–arcs occupied by $\boldsymbol{x}$, and the set of α–arcs not occupied by $\boldsymbol{x}$.

If $\mathcal{H}$ is a diagram for an (m, n)–tangle, then for $\boldsymbol{x} \in \mathfrak{S}$ we define

$$o^0(\boldsymbol{x}) := \{i \mid \boldsymbol{x} \cap \alpha_i^0 \neq \varnothing\}, \quad \overline{o}^0(\boldsymbol{x}) := [m+1] \setminus o^0(\boldsymbol{x}),$$
$$o^1(\boldsymbol{x}) := \{i \mid \boldsymbol{x} \cap \alpha_i^1 \neq \varnothing\}, \quad \overline{o}^1(\boldsymbol{x}) := [n+1] \setminus o^1(\boldsymbol{x}).$$

Remark If $\mathcal{H}$ is a type 1 or a type 2 diagram, then exactly n or $n+1$ of the α–arcs, respectively, are occupied by each generator. If $\mathcal{H}$ is a diagram with two boundary components, the total number of occupied α–arcs on the two sides is $\frac{1}{2}(m+n)+1$, but the number on each side may vary.

8.4 Homology classes

We will soon count pseudoholomorphic curves that connect generators. Each such curve carries a homology class, defined as follows.

Definition 8.15 Fix generators $\boldsymbol{x}$ and $\boldsymbol{y}$, and let I be the interval [0, 1]. Let $\pi_2(\boldsymbol{x}, \boldsymbol{y})$, the *homology classes from* $\boldsymbol{x}$ *to* $\boldsymbol{y}$, be the elements of

$$H_2\big(\Sigma \times I \times I, \big((\boldsymbol{\alpha} \times \{1\} \cup \boldsymbol{\beta} \times \{0\} \cup (\partial\Sigma \setminus \boldsymbol{z}) \times I) \times I\big) \cup (\boldsymbol{x} \times I \times \{0\}) \cup (\boldsymbol{y} \times I \times \{1\})\big)$$

which map to the relative fundamental class of $\boldsymbol{x} \times I \cup \boldsymbol{y} \times I$ under the composition of the boundary homomorphism and collapsing the rest of the boundary.

Definition 8.16 Given a homology class $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$, its *domain* $[B]$ is the projection of B to $H_2(\Sigma, \boldsymbol{\alpha} \cup \boldsymbol{\beta} \cup \partial\Sigma)$. We can interpret the domain of B as a linear combination of the components of $\Sigma \setminus (\boldsymbol{\alpha} \cup \boldsymbol{\beta})$, which we call *regions*.

Note that a homology class is uniquely determined by its domain.

Definition 8.17 The coefficient of each region in a domain is called its *multiplicity*. Given a point $p \in \Sigma \setminus (\boldsymbol{\alpha} \cup \boldsymbol{\beta})$, we denote by $n_p(B)$ the multiplicity of $[B]$ at the region containing p. Alternatively, $n_p(B)$ is the intersection number of B and $\{p\} \times I \times I$.

By definition, the multiplicity of $[B]$ at any region D that contains a point in $\boldsymbol{z}$ is zero.

Definition 8.18 We define the set of *empty homology classes* as

$$\widehat{\pi}_2(\boldsymbol{x}, \boldsymbol{y}) := \{B \in \pi_2(\boldsymbol{x}, \boldsymbol{y}) \mid n_X(B) = 0 \text{ and } n_O(B) = 0 \text{ for all } X \in \mathbb{X} \text{ and } O \in \mathbb{O}\}.$$

To define our Floer invariants, we will only be interested in this smaller set $\widehat{\pi}_2(\boldsymbol{x}, \boldsymbol{y})$.

Concatenation at $\boldsymbol{y} \times I$, which corresponds to addition of domains, gives products $*\colon \pi_2(\boldsymbol{x}, \boldsymbol{y}) \times \pi_2(\boldsymbol{y}, \boldsymbol{w}) \to \pi_2(\boldsymbol{x}, \boldsymbol{w})$ and $*\colon \widehat{\pi}_2(\boldsymbol{x}, \boldsymbol{y}) \times \widehat{\pi}_2(\boldsymbol{y}, \boldsymbol{w}) \to \widehat{\pi}_2(\boldsymbol{x}, \boldsymbol{w})$. This operation turns $\pi_2(\boldsymbol{x}, \boldsymbol{x})$ and $\widehat{\pi}_2(\boldsymbol{x}, \boldsymbol{x})$ into groups, called the group of *periodic domains* and the group of *empty periodic domains*, respectively.

We can split the boundary of a domain $[B]$ into three pieces, $\partial^\partial B \subset \partial\Sigma$, $\partial^\alpha B \subset \boldsymbol{\alpha}$ and $\partial^\beta B \subset \boldsymbol{\beta}$, oriented so that $\partial^\partial B + \partial^\alpha B + \partial^\beta B$ is the boundary of $[B]$. We can think of $\partial^\partial B$ as an element of $H_1(\partial\Sigma, \partial\boldsymbol{\alpha})$. For a Heegaard diagram $\mathcal{H}$ with two boundary components, we can further split $\partial^\partial B$ into two pieces, $\partial^i B \subset \partial^i \mathcal{H}$, such that $\partial^\partial B = \partial^0 B + \partial^1 B$.

Definition 8.19 A homology class B is called *provincial* if $\partial^\partial B = 0$. For a diagram with two boundary components, a homology class B is called *left-provincial* if $\partial^0 B = 0$, and *right-provincial* if $\partial^1 B = 0$. We denote the set of empty provincial homology classes from $\boldsymbol{x}$ to $\boldsymbol{y}$ by $\widehat{\pi}_2^\partial(\boldsymbol{x}, \boldsymbol{y})$.

Observe that concatenation turns $\widehat{\pi}_2^\partial(\boldsymbol{x}, \boldsymbol{x})$ into a group.

8.5 Admissibility

In order to get well-defined Heegaard–Floer invariants, we need to impose some additional conditions on the tangle Heegaard diagrams.

Definition 8.20 A tangle Heegaard diagram is called *admissible* if every nonzero empty periodic domain has both positive and negative multiplicities.

A tangle Heegaard diagram is called *provincially admissible* if every nonzero empty provincial periodic domain has both positive and negative multiplicities.

A tangle Heegaard diagram with two boundary components is called *left* (respectively *right*) *admissible* if every nonzero empty right-provincial (respectively left-provincial) periodic domain has both positive and negative multiplicities.

Proposition 8.21 *Any tangle Heegaard diagram can be made admissible by performing isotopy on $\boldsymbol{\beta}$. Further, any two admissible diagrams for a given $2n$–tangle or a strongly marked (m,n)–tangle are connected through a sequence of Heegaard moves, so that every intermediate diagram is admissible too. The same is true if we replace "admissible" by "provincially admissible".*

Proof This follows from a winding argument for the β–curves, just as in the case for closed manifolds [18, Section 5]. Alternatively, see [28, Proposition 4.11] □

Corollary 8.22 *Every tangle $(Y,\mathcal{T})$ has an admissible tangle Heegaard diagram. Similarly, Every tangle $(Y,\mathcal{T})$ has a provincially admissible tangle Heegaard diagram. The same statements hold for every strongly marked tangle.*

8.6 Gluing

Any two multipointed bordered Heegaard diagrams can be glued along a matching boundary component: if $\mathcal{H}_1$ and $\mathcal{H}_2$ are diagrams, and $\mathcal{Z}_i$ are boundary components of $\mathcal{H}_i$ with $\mathcal{Z}_1 = \mathcal{Z}_2^*$, one can glue $\mathcal{H}_1$ to $\mathcal{H}_2$ by identifying $\mathcal{Z}_1$ with $\mathcal{Z}_2^*$. In this way, one can glue a type 1 diagram to the left, ie ∂^0, boundary of a diagram with two boundary components, a type 1 diagram to a type 2 diagram, a type 2 diagram to the ∂^1 boundary of a diagram with two boundary components, or the ∂^0 boundary of a diagram with two boundary components to the ∂^1 boundary of another diagram with two boundary components.

By gluing a type 1 diagram, a sequence of diagrams with two boundary components, and a type 2 diagram together, removing the union of the z markings, and placing an X and an O in the corresponding region, one obtains a closed Heegaard diagram for the knot/link that is union of the corresponding tangles, together with an additional split unknot. See Figure 2 for a schematic example.

Below we describe in full detail how to glue Heegaard diagrams for tangles, and discuss the basic properties of the resulting diagram.

For the rest of this section, we fix two Heegaard diagrams as follows. Let $\mathcal{H}_1 = (\Sigma_1, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \mathbb{X}_1, \mathbb{O}_1, \boldsymbol{z}_1)$ be a Heegaard diagram (of type 1, or with two boundary components) for some pair $(Y_1, \mathcal{T}_1)$, and if $\mathcal{H}_1$ is of type 1, denote its boundary by $\partial^1\mathcal{H}_1$. Let $\mathcal{H}_2 = (\Sigma_2, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2, \mathbb{X}_2, \mathbb{O}_2, \boldsymbol{z}_2)$ be a Heegaard diagram (of type 2 or with two boundary components) for another pair $(Y_2, \mathcal{T}_2)$, and if $\mathcal{H}_2$ is of type 2, denote its boundary by $\partial^0\mathcal{H}_2$. Suppose $\partial^1\mathcal{H}_1 = (\partial^0\mathcal{H}_2)^*$, ie $\partial^1 Y_1$ is identified with a marked sphere $\mathcal{S}$ and $\partial^0 Y_2$ is identified with $-\mathcal{S}$.

Definition 8.23 The *union* of $\mathcal{H}_1$ and $\mathcal{H}_2$, denoted $\mathcal{H}_1 \cup \mathcal{H}_2$ is the Heegaard diagram $\mathcal{H}$ obtained in the following way: We remove all $\mathbb{X}$ and $\mathbb{O}$ markings on the boundaries of the two diagrams. We glue the two surfaces along their boundary, matching the α and z endpoints and respecting the identification $\partial^1\mathcal{H}_1 = (\partial^0\mathcal{H}_2)^*$, to obtain $\Sigma := \Sigma_1 \cup_\partial \Sigma_2$. We take $\boldsymbol{\alpha}$ to be the set of circles $\boldsymbol{\alpha}_1 \cup_\partial \boldsymbol{\alpha}_2$, and we take $\boldsymbol{\beta}$ to be $\boldsymbol{\beta}_1 \cup \boldsymbol{\beta}_2$. If $\Sigma_1 \cup \Sigma_2$ is a closed surface, we remove z_1 and z_2, place two points marked X' and O' in the same region, and let $\mathbb{X} = \mathbb{X}_1 \cup \mathbb{X}_2 \cup X'$ and $\mathbb{O} = \mathbb{O}_1 \cup \mathbb{O}_2 \cup O'$. We get a closed Heegaard diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O})$. If $\Sigma_1 \cup \Sigma_2$ has boundary, we let $\mathbb{X} = \mathbb{X}_1 \cup \mathbb{X}_2$ and $\mathbb{O} = \mathbb{O}_1 \cup \mathbb{O}_2$, and we take z to be the oriented arc(s) $z_1 \cup_\partial z_2$. We get a tangle Heegaard diagram $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, z)$.

Gluing Heegaard diagrams corresponds to gluing tangles. In the lemma below, all unions are formed by following the identifications with $\mathcal{S}$ given by the tangles.

Lemma 8.24 *When the union $\mathcal{H}_1 \cup \mathcal{H}_2$ is a diagram with one boundary component, it represents the pair $(Y_1 \cup Y_2, \mathcal{T}_1 \cup \mathcal{T}_2)$.*

When $\mathcal{H}_1 \cup \mathcal{H}_2$ is a diagram with two boundary components, it represents the triple $(Y_1 \cup Y_2, \mathcal{T}_1 \cup \mathcal{T}_2, \gamma_1 \cup \gamma_2)$.

When $\mathcal{H}_1 \cup \mathcal{H}_2$ is a closed Heegaard diagram, it represents the link $(\mathcal{T}_1 \cup \mathcal{T}_2) \cup U$ in $Y_1 \cup Y_2$, where U is an unknot unlinked from $\mathcal{T}_1 \cup \mathcal{T}_2$.

Proof This follows directly from the topological constructions of tangles from the three types of Heegaard diagrams described in Sections 8.1 and 8.2, along with the fact that adding an X and an O in one and the same region introduces an unlinked unknot. □

Generators and homology classes behave nicely under gluing. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two tangle Heegaard diagrams which agree along a boundary component. Note that given $\boldsymbol{x}_1 \in \mathfrak{S}(\mathcal{H}_1)$ and $\boldsymbol{x}_2 \in \mathfrak{S}(\mathcal{H}_2)$ such that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ occupy complementary sets of the new α–circles obtained by gluing α–arcs, the union $\boldsymbol{x}_1 \cup \boldsymbol{x}_2$ is a generator in $\mathfrak{S}(\mathcal{H}_1 \cup \mathcal{H}_2)$.

Lemma 8.25 *Given $\boldsymbol{x}_1, \boldsymbol{y}_1 \in \mathfrak{S}(\mathcal{H}_1)$ and $\boldsymbol{x}_2, \boldsymbol{y}_2 \in \mathfrak{S}(\mathcal{H}_2)$, there is a natural identification of $\pi_2(\boldsymbol{x}_1 \cup \boldsymbol{x}_2, \boldsymbol{y}_1 \cup \boldsymbol{y}_2)$ with the set of pairs (B_1, B_2) in $\pi_2(\boldsymbol{x}_1, \boldsymbol{y}_1) \times \pi_2(\boldsymbol{x}_2, \boldsymbol{y}_2)$ such that $\partial^1 B_1 = -\partial^0 B_2$. The same statement holds if we replace π_2 with $\widehat{\pi}_2$.*

Proof The proof is straightforward. □

Following notation from [8], for B_1 and B_2 which agree along the boundary as above, we denote the corresponding homology class in $\pi_2(\boldsymbol{x}_1,\boldsymbol{y}_1)\times\pi_2(\boldsymbol{x}_2,\boldsymbol{y}_2)$ by $B_1 \natural B_2$. Under this identification, the local multiplicity of B_i at a point $p\in\Sigma_i\setminus(\boldsymbol{\alpha}_i\cup\boldsymbol{\beta}_i)$ agrees with the local multiplicity of $B_1\natural B_2$ at p thought of as a point in $\Sigma_1\cup\Sigma_2$.

Lemma 8.26 *Suppose $\mathcal{H}_1$ and $\mathcal{H}_2$ are of type 1 and type 2, respectively. If one diagram is admissible, and the other one is provincially admissible, then $\mathcal{H}_1\cup\mathcal{H}_2$ is admissible.*

Proof The proof is identical to the proof of [8, Lemma 4.33], and we recall the argument here. Let $B_1\natural B_2$ be a positive periodic domain. If $\mathcal{H}_1$ is admissible, then $B_1=0$, so $\partial^\partial B_1=0$, and since $\mathcal{H}_2$ is provincially admissible, it follows that $B_2=0$. Similarly, if B_2 is admissible, it follows that $B_1=0$ and $B_2=0$. □

Lemma 8.27 (compare to [9, Lemma 5.22]) *Suppose $\mathcal{H}_1$ and $\mathcal{H}_2$ are provincially admissible multipointed bordered Heegaard diagrams with two boundary components with $\partial^1\mathcal{H}_1=(\partial^0\mathcal{H}_2)^*$, and let $\mathcal{H}=\mathcal{H}_1\cup\mathcal{H}_2$. If $\mathcal{H}_1$ is right admissible, or $\mathcal{H}_2$ is left admissible, then $\mathcal{H}$ is provincially admissible. Furthermore:*

(1) *If $\mathcal{H}_1$ and $\mathcal{H}_2$ are both left admissible (respectively right admissible), then $\mathcal{H}$ is left admissible (respectively right admissible).*

(2) *If $\mathcal{H}_1$ is admissible, then $\mathcal{H}$ is left admissible. If $\mathcal{H}_2$ is admissible, then $\mathcal{H}$ is right admissible.*

(3) *If $\mathcal{H}_1$ is admissible and $\mathcal{H}_2$ is right admissible, or if $\mathcal{H}_1$ is left admissible and $\mathcal{H}_2$ is admissible, then $\mathcal{H}$ is admissible.*

Analogous statements hold when one of the two Heegaard diagrams has one boundary component.

Proof The proof is analogous to that of [9, Lemma 5.22] □

9 Moduli spaces

In this section, we describe the holomorphic curves that will be considered in the definitions of the various invariants associated to tangle Heegaard diagrams.

Most of this discussion is a straightforward generalization of the one for bordered Floer homology [8]. We count pseudoholomorphic curves in $\Sigma\times I\times\mathbb{R}$. In the bordered Floer setting, one counts curves that avoid a basepoint $z\in\partial\Sigma$. Here, we avoid multiple basepoints, both in the interior, and on the boundary of Σ, as well as the arcs (or points) that we denote by $\boldsymbol{z}$.

9.1 Moduli spaces of holomorphic curves

Let $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ be a tangle Heegaard diagram (with one boundary component and of type 1 or type 2, or with two boundary components). We can think of the open surface $\operatorname{Int}\Sigma$ as a surface with a set of punctures $\boldsymbol{p}$ (one puncture for each boundary component of Σ). Choose a symplectic form ω_Σ such that the boundary $\partial\Sigma$ is a cylindrical end, and let j_Σ be a compatible almost complex structure. We will assume that $\boldsymbol{\alpha}^a$ is cylindrical near $\partial\Sigma$, in the following sense. There is a neighborhood $U_{\boldsymbol{p}}$ of the punctures symplectomorphic to $\partial\Sigma \times (0,\infty) \subset T^*(\partial\Sigma)$, such that j_Σ and $\boldsymbol{\alpha}^a \cap U_{\boldsymbol{p}}$ are invariant with respect to the $\mathbb{R}$–action on $\partial\Sigma \times (0,\infty)$. We write $\mathbb{D} = I \times \mathbb{R}$, and let $\omega_{\mathbb{D}}$ and $j_{\mathbb{D}}$ be the standard symplectic form and almost complex structure on $\mathbb{D} \subset \mathbb{C}$. Consider the projections

$$\begin{aligned} \pi_\Sigma &\colon \Sigma \times \mathbb{D} \to \Sigma, \\ \pi_{\mathbb{D}} &\colon \Sigma \times \mathbb{D} \to \mathbb{D}, \\ s &\colon \Sigma \times \mathbb{D} \to I, \\ t &\colon \Sigma \times \mathbb{D} \to \mathbb{R}. \end{aligned}$$

Definition 9.1 We say that an almost complex structure J on $\Sigma \times \mathbb{D}$ is *admissible* if the following conditions hold:

- $\pi_{\mathbb{D}}$ is J–holomorphic.
- $J(\partial_s) = \partial_t$ for the vector fields tangent to the fibers of π_Σ.
- The $\mathbb{R}$–action is J–holomorphic.
- J splits as $J = j_\Sigma \times j_{\mathbb{D}}$ near $\boldsymbol{p} \times \mathbb{D}$.

Definition 9.2 A *decorated source* $S^\triangleright$ consists of

- a topological type of smooth Riemann surface S with boundary, and a finite number of punctures on the boundary;
- a labeling of each puncture of S by $+$, $-$ or e;
- a labeling of each e puncture by a Reeb chord ρ in $(\partial\Sigma, \partial\boldsymbol{\alpha})$.

Given a decorated source $S^\triangleright$, we denote by $S_{\bar{e}}$ the result of filling in the e punctures of S.

We consider maps

$$u\colon (S, \partial S) \to \big(\Sigma \times \mathbb{D}, (\boldsymbol{\alpha} \times \{1\} \times \mathbb{R}) \cup (\boldsymbol{\beta} \times \{0\} \times \mathbb{R})\big)$$

such that:

(1) u is (j, J) holomorphic for some almost complex structure j on S.

(2) u: $S \to \Sigma \times \mathbb{D}$ is proper.

(3) u extends to a proper map $u_{\overline{e}}$: $S_{\overline{e}} \to \Sigma_{\overline{e}} \times \mathbb{D}$

(4) u has finite energy in the sense of Bourgeois, Eliashberg, Hofer, Wysocki and Zehnder [1].

(5) $\pi_{\mathbb{D}} \circ u$ is a g–fold branched cover.

(6) At each $+$ puncture q of S, $\lim_{z \to q} t \circ u(z) = +\infty$.

(7) At each $-$ puncture q of S, $\lim_{z \to q} t \circ u(z) = -\infty$.

(8) At each e puncture q of S, $\lim_{z \to q} \pi_\Sigma \circ u(z)$ is the Reeb chord ρ labeling q.

(9) $\pi_\Sigma \circ u$: $S \to \operatorname{Int} \Sigma$ does not cover any of the regions of $\Sigma \setminus (\boldsymbol{\alpha} \cup \boldsymbol{\beta})$ that intersect $\mathbb{X} \cup \mathbb{O} \cup \boldsymbol{z}$.

(10) For each $t \in \mathbb{R}$ and $\beta_i \in \boldsymbol{\beta}$, $u^{-1}(\beta_i \times \{0\} \times \{t\})$ consists of exactly one point; for each $t \in \mathbb{R}$ and $\alpha_i^c \in \boldsymbol{\alpha}$, $u^{-1}(\alpha_i^c \times \{1\} \times \{t\})$ consists of exactly one point; for each $t \in \mathbb{R}$ and $\alpha_i^a \in \boldsymbol{\alpha}$, $u^{-1}(\alpha_i^a \times \{1\} \times \{t\})$ consists of at most one point.

(11) u is embedded.

Under these conditions, at $-\infty$, u is asymptotic to a g–tuple of arcs $x_i \times I \times \{-\infty\}$, and at $+\infty$, u is asymptotic to a g–tuple of arcs $y_i \times I \times \{+\infty\}$, so that $\boldsymbol{x} := \{x_1, \dots, x_g\}$ and $\boldsymbol{y} := \{y_1, \dots, y_g\}$ are generators of $\mathcal{H}$. We call $\boldsymbol{x}$ the *incoming* generator, and $\boldsymbol{y}$ the *outgoing* generator for u. Such a curve u has an associated homology class $B = [u] \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$.

Definition 9.3 Given a map u from a decorated source $S^\triangleright$, the *height* of an e puncture q is the evaluation $\operatorname{ev}(q) = t \circ u_{\overline{e}}(q) \in \mathbb{R}$.

Definition 9.4 Let $E(S^\triangleright)$ be the set of e punctures of S. Let $\vec{P} = (P_1, \dots, P_m)$ be a partition of $E(S^\triangleright)$ with P_i nonempty. We say a map u is *$\vec{P}$–compatible* if for any i, all the punctures in P_i have the same height, and $\operatorname{ev}(P_i) < \operatorname{ev}(P_j)$ whenever $i < j$.

To a partition $\vec{P} = (P_1, \dots, P_m)$ we associate a sequence of sets of Reeb chords $\vec{\rho}(\vec{P}) = (\boldsymbol{\rho}_1, \dots, \boldsymbol{\rho}_m)$, where $\boldsymbol{\rho}_i := \{\rho \mid \rho \text{ labels } q,\ q \in P_i\}$. To such a sequence $\vec{\boldsymbol{\rho}}$ we can associate a homology class

$$[\vec{\boldsymbol{\rho}}] = [\boldsymbol{\rho}_1] + \cdots + [\boldsymbol{\rho}_m] \in H_1(\partial\Sigma, \partial\boldsymbol{\alpha})$$

and an algebra element

$$a(\vec{\rho}) = a(\rho_1) \cdots a(\rho_m) \in \mathcal{A}(\partial\mathcal{H}).$$

Note that $[a(\vec{\rho})] = [\vec{\rho}]$, and also if u is a $\vec{P}$–compatible map satisfying (1)–(10) with homology class $[u] = B$, then $[\vec{\rho}(\vec{P})] = \partial^{\partial} B$.

Definition 9.5 Given generators $\boldsymbol{x}$ and $\boldsymbol{y}$, a homology class $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$, and a decorated source $S^{\triangleright}$, we let

$$\widetilde{\mathcal{M}}^{B}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright})$$

denote the moduli space of curves u with source $S^{\triangleright}$ satisfying (1)–(10), asymptotic to $\boldsymbol{x}$ at $-\infty$ and to $\boldsymbol{y}$ at $+\infty$, and with homology class $[u] = B$. Given also a partition $\vec{P}$ of $E(S^{\triangleright})$, we let

$$\widetilde{\mathcal{M}}^{B}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P})$$

denote the space of $\vec{P}$–compatible maps in $\widetilde{\mathcal{M}}^{B}(\boldsymbol{x}, \boldsymbol{y}; S^{\triangleright})$, and we let

$$\widetilde{\mathcal{M}}^{B}_{\mathrm{emb}}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P})$$

denote the space of maps in $\widetilde{\mathcal{M}}^{B}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P})$ that also satisfy (11).

Many results carry over directly from the ones in [8; 28].

Proposition 9.6 (compare to [8, Proposition 5.6]) *There is a dense set of admissible J for which the spaces $\widetilde{\mathcal{M}}^{B}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P})$ are transversally cut out by the $\overline{\partial}$ equations.*

Proposition 9.7 (compare to [8, Proposition 5.8]) *The expected dimension of the space $\widetilde{\mathcal{M}}^{B}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P})$ is*

$$\operatorname{ind}(B, S^{\triangleright}, \vec{P}) = g - \chi(S) + 2e(B) + |\vec{P}|.$$

(Here $e(B)$ is the Euler measure of the domain of B and $|\vec{P}|$ is the number of parts in the partition $\vec{P}$.)

Whether a curve in $\widetilde{\mathcal{M}}^{B}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P})$ is embedded depends only on the topological data of B, $S^{\triangleright}$, and $\vec{P}$, ie there are entire components of embedded and of nonembedded curves. For embedded curves, there is another index formula that only depends on B and $\vec{P}$. Before we state this formula, we make a couple of definitions regarding Reeb chords. Even though our matched circles are different, these definitions are identical to the ones in [8, Sections 3.3.1 and 5.7.1].

Let m: $H_1(\partial\Sigma\backslash \boldsymbol{z}, \partial\boldsymbol{\alpha}; \mathbb{Z}) \times H_0(\partial\boldsymbol{\alpha}; \mathbb{Z}) \to \frac{1}{2}\mathbb{Z}$ be the map that counts local multiplicities. Specifically, for $a \in H_1(\partial\Sigma \setminus \boldsymbol{z}, \partial\boldsymbol{\alpha}; \mathbb{Z})$ and $p \in \partial\boldsymbol{\alpha}$, we define the *multiplicity* $m(a, p)$ of p in a as the average multiplicity with which a covers the regions on either side of p, and extend bilinearly.

For $a, b \in H_1(\partial\Sigma \setminus \boldsymbol{z}, \partial\boldsymbol{\alpha}; \mathbb{Z})$, define

$$L(a, b) := m(b, \partial a),$$

where ∂ is the connecting homomorphism from the homology long exact sequence. Note that $L(a, b) = -L(b, a)$ for any a and b.

For a set of Reeb chords $\boldsymbol{\rho}$ in $(\partial\Sigma \setminus \boldsymbol{z}, \partial\boldsymbol{\alpha})$, define

$$\iota(\boldsymbol{\rho}) := -\sum_{\{\rho_i, \rho_j\} \subset \boldsymbol{\rho}} |L([\rho_i], [\rho_j])| - \tfrac{1}{2}|\boldsymbol{\rho}|.$$

For a sequence of sets of Reeb chords $\vec{\boldsymbol{\rho}} = (\boldsymbol{\rho}_1, \dots, \boldsymbol{\rho}_m)$, define

$$\iota(\vec{\boldsymbol{\rho}}) := \sum_i \iota(\boldsymbol{\rho}_i) + \sum_{i<j} L([\boldsymbol{\rho}_i], [\boldsymbol{\rho}_j]).$$

Finally, we come to the index formula.

Definition 9.8 Let $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ and $\vec{\boldsymbol{\rho}}$ be a sequence of sets of Reeb chords. We define

$$\begin{aligned}\chi_{\text{emb}}(B, \vec{\boldsymbol{\rho}}) &:= g + e(B) - n_{\boldsymbol{x}}(B) - n_{\boldsymbol{y}}(B) - \iota(\vec{\boldsymbol{\rho}}),\\ \operatorname{ind}(B, \vec{\boldsymbol{\rho}}) &:= e(B) + n_{\boldsymbol{x}}(B) + n_{\boldsymbol{y}}(B) + |\vec{\boldsymbol{\rho}}| + \iota(\vec{\boldsymbol{\rho}}).\end{aligned}$$

Proposition 9.9 (compare to [8, Proposition 5.62; 28, Proposition 5.9]) *For $u \in \widetilde{\mathcal{M}}^{B}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P})$, either u is embedded, and*

$$\begin{aligned}\chi(S^{\triangleright}) &= \chi_{\text{emb}}(B, \vec{\boldsymbol{\rho}}(\vec{P})),\\ \operatorname{ind}(B, S^{\triangleright}, \vec{P}) &= \operatorname{ind}(B, \vec{\boldsymbol{\rho}}(\vec{P})),\\ \widetilde{\mathcal{M}}^{B}_{\text{emb}}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P}) &= \widetilde{\mathcal{M}}^{B}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P}),\end{aligned}$$

or u is not embedded, and

$$\begin{aligned}\chi(S^{\triangleright}) &> \chi_{\text{emb}}(B, \vec{\boldsymbol{\rho}}(\vec{P})),\\ \operatorname{ind}(B, S^{\triangleright}, \vec{P}) &< \operatorname{ind}(B, \vec{\boldsymbol{\rho}}(\vec{P})),\\ \widetilde{\mathcal{M}}^{B}_{\text{emb}}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P}) &= \varnothing.\end{aligned}$$

Each of the moduli spaces has an $\mathbb{R}$–action by translation in the t factor. For *stable* curves, ie except when the moduli space consists of a single curve u with $\pi_{\mathbb{D}} \circ u$ a trivial g–fold cover of $\mathbb{D}$ and $B = 0$, this action is free. For moduli spaces of stable curves, we quotient by this action.

Definition 9.10 Given $\boldsymbol{x}$, $\boldsymbol{y}$, $S^{\triangleright}$ and $\vec{P}$, let

$$\mathcal{M}^B(\boldsymbol{x},\boldsymbol{y},S^{\triangleright},\vec{P}) := \widetilde{\mathcal{M}}^B(\boldsymbol{x},\boldsymbol{y},S^{\triangleright},\vec{P})/\mathbb{R},$$
$$\mathcal{M}^B_{\mathrm{emb}}(\boldsymbol{x},\boldsymbol{y},S^{\triangleright},\vec{P}) := \widetilde{\mathcal{M}}^B_{\mathrm{emb}}(\boldsymbol{x},\boldsymbol{y},S^{\triangleright},\vec{P})/\mathbb{R}.$$

9.2 Degenerations

The properties of moduli spaces that are needed in order to show that the invariants are well-defined are the same as in [8]. To understand the compactifications of moduli spaces, one studies *holomorphic combs*, ie trees of homomorphic curves in $\Sigma \times \mathbb{D}$ and in $\partial\Sigma \times \mathbb{R} \times \mathbb{D}$. In the *tilde* version (when one does not allow domains that cover $\mathbb{X} \cup \mathbb{O}$), most types of degenerations are the same as in [8], and most results carry over.

The only difference is in the homological assumptions on Σ, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$. Even though the α– or the β–circles are not linearly independent, there are no new boundary degenerations, as every region of $\Sigma \setminus \boldsymbol{\alpha}$, as well as every region of $\Sigma \setminus \boldsymbol{\beta}$, contains an X or an O.

10 The modules associated to tangle Heegaard diagrams

In this section, we associate algebraic structures to tangle Heegaard diagrams. Before we proceed, recall that for any pointed matched circle $\mathcal{Z}$, the algebra $\mathcal{A}(\mathcal{Z})$ does not depend on the $\mathbb{X}$ and $\mathbb{O}$ markings on the circle.

For the remainder of this paper, we let V denote $\mathbb{F}_2 \otimes \mathbb{F}_2$.

10.1 The type D structure

We define type D structures for type 2 multipointed bordered Heegaard diagrams for tangles. The construction and results for type 1 diagrams are identical.

Suppose $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ is a provincially admissible Heegaard diagram of type 2 for some $2n$–tangle $(Y, \mathcal{T})$. Let J be an admissible almost complex structure. We define a left type D structure $\widetilde{CFTD}(\mathcal{H}, J)$ over $\mathcal{A}(-\partial\mathcal{H})$, as follows.

Let $X(\mathcal{H})$ be the $\mathbb{F}_2$ vector space spanned by $\mathfrak{S}(\mathcal{H})$. Let $I_D(\boldsymbol{x}) = I(\overline{o}(\boldsymbol{x})) \in \mathcal{I}(-\partial\mathcal{H})$. We define an action on $X(\mathcal{H})$ of $\mathcal{I}(-\partial\mathcal{H})$ by

$$I(\boldsymbol{s}) \cdot \boldsymbol{x} = \begin{cases} \boldsymbol{x} & \text{if } I(\boldsymbol{s}) = I_D(\boldsymbol{x}), \\ 0 & \text{otherwise.} \end{cases}$$

Then $\widetilde{CFTD}(\mathcal{H}, J)$ is defined as an $\mathcal{A}(-\partial\mathcal{H})$–module by

$$\widetilde{CFTD}(\mathcal{H}, J) = \mathcal{A}(-\partial\mathcal{H}) \otimes_{\mathcal{I}(-\partial\mathcal{H})} X(\mathcal{H}).$$

Given $\boldsymbol{x}$, $\boldsymbol{y} \in \mathfrak{S}(\mathcal{H})$, we define

$$a_{\boldsymbol{x},\boldsymbol{y}} := \sum_{\substack{B \in \tilde{\pi}_2(\boldsymbol{x},\boldsymbol{y}) \\ \vec{P} \text{ discrete} \\ \operatorname{ind}(B, \vec{\rho}(\vec{P}))=1}} \#\mathcal{M}^{B}_{\text{emb}}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P}) \cdot a(-P_1) \cdots a(-P_m).$$

Here all P are discrete partitions, ie partitions $P = (P_1, \dots, P_m)$ where $|P_i| = 1$.

The map $\delta\colon \widetilde{CFTD}(\mathcal{H}, J) \to \mathcal{A}(-\partial\mathcal{H}) \otimes \widetilde{CFTD}(\mathcal{H}, J)$ is defined as

$$\delta(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathfrak{S}(\mathcal{H})} a_{\boldsymbol{x},\boldsymbol{y}} \otimes \boldsymbol{y}.$$

Theorem 10.1 *Let $\mathcal{H}$ be a tangle Heegaard diagram of type 2 for a marked tangle $\mathcal{T}$ in a 3–manifold Y, equipped with an admissible almost complex structure J. If $\mathcal{H}$ is provincially admissible, then $\widetilde{CFTD}(\mathcal{H}, J)$ is a type D structure over $\mathcal{A}(-\partial\mathcal{H})$. Moreover, if $\mathcal{H}$ is admissible, then $\widetilde{CFTD}(\mathcal{H}, J)$ is bounded.*

Proof The proof follows directly from the arguments for $\widehat{CFD}$ in [8, Chapter 6]. We outline the main steps. To show that the counts of holomorphic curves are finite, we observe that in a provincially admissible diagram there are only finitely many domains that contribute to the counts, and for any diagram there are only finitely many sequences $\vec{P}$ with nonzero $a(\vec{\rho}(\vec{P})) \in \mathcal{A}(-\partial\mathcal{H})$. To show that the compatibility condition for a type D structure is satisfied, we count possible degenerations of holomorphic curves. □

Theorem 10.2 *Up to homotopy equivalence and tensoring with V, $\widetilde{CFTD}(\mathcal{H}, J)$ is independent of the choice of sufficiently generic admissible almost complex structure, and provincially admissible type 2 tangle Heegaard diagram for $(Y, \mathcal{T})$. Namely, if $\mathcal{H}_1$ and $\mathcal{H}_2$ are provincially admissible type 2 diagrams for $(Y, \mathcal{T})$ with almost complex structures J_1 and J_2, and $|\mathbb{X}_1| = |\mathbb{X}_2| + k$, then*

$$\widetilde{CFTD}(\mathcal{H}_1, J_1) \simeq \widetilde{CFTD}(\mathcal{H}_2, J_2) \otimes V^{\otimes k}.$$

Proof To show invariance, we construct chain maps corresponding to a change of almost complex structure or the various Heegaard moves. We have two Heegaard moves that do not occur in [8] — index zero/three stabilization and destabilization. Those always occur in the interior of the diagram, and result in the extra V, by the same argument as in the closed case (see [11], for example), ie if $\mathcal{H}'$ is obtained from $\mathcal{H}$ by an index zero/three stabilization, then $\widetilde{CFTD}(\mathcal{H}) \simeq \widetilde{CFTD}(\mathcal{H}') \otimes V$. □

When we write $\widetilde{CFTD}(Y, \mathcal{T})$, we mean the type D structure without the extra Vs, ie what we get from a tangle Heegaard diagram with the minimum number of basepoints, which is $|\mathbb{X} \cap \operatorname{Int} \Sigma| = |\mathcal{T}| = |\mathbb{O} \cap \operatorname{Int} \Sigma|$, or equivalently $|\mathbb{X}| = 2|\mathcal{T}| = |\mathbb{O}|$.

10.2 The type A structure

We define type A structures for type 1 multipointed bordered Heegaard diagrams for tangles. The construction and results for type 2 diagrams are identical.

Let $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ be a provincially admissible type 1 Heegaard diagram for a $2n$–tangle $(Y, \mathcal{T})$, and let J be an admissible almost complex structure. We define a type A structure $\widetilde{CFAT}(\mathcal{H}, J)$ over $\mathcal{A}(\partial\mathcal{H})$.

Define $I_A(\boldsymbol{x}) = I(o(\boldsymbol{x}))$. The module $\widetilde{CFAT}(\mathcal{H}, J)$ is generated over $\mathbb{F}_2$ by $X(\mathcal{H})$, and the right action of $\mathcal{I}(\partial\mathcal{H})$ on $\widetilde{CFAT}(\mathcal{H}, J)$ is defined on the generators by

$$\boldsymbol{x} \cdot I(\boldsymbol{s}) = \begin{cases} \boldsymbol{x} & \text{if } I(\boldsymbol{s}) = I_A(\boldsymbol{x}), \\ 0 & \text{otherwise.} \end{cases}$$

For the $\mathcal{A}_\infty$ multiplication maps, we consider partitions $P = (P_1, \dots, P_m)$ that are not necessarily discrete. When $I_A(\boldsymbol{x}) \otimes a(\boldsymbol{\rho}_1) \otimes \cdots \otimes a(\boldsymbol{\rho}_n) \neq 0$, we define $m_{n+1}\colon \widetilde{CFAT}(\mathcal{H}, J) \otimes \mathcal{A}(\partial\mathcal{H})^{\otimes n} \to \widetilde{CFAT}(\mathcal{H}, J)$ by

$$m_{n+1}(\boldsymbol{x}, a(\boldsymbol{\rho}_1), \dots, a(\boldsymbol{\rho}_n)) := \sum_{\boldsymbol{y} \in \mathfrak{S}(\mathcal{H})} \sum_{\substack{B \in \tilde{\pi}_2(\boldsymbol{x}, \boldsymbol{y}) \\ \{\vec{P} \mid \vec{\rho}(\vec{P}) = (\boldsymbol{\rho}_1, \dots, \boldsymbol{\rho}_n)\} \\ \operatorname{ind}(B, \vec{\rho}(\vec{P})) = 1}} \#\mathcal{M}^{B}_{\mathrm{emb}}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P}) \cdot \boldsymbol{y}.$$

Theorem 10.3 *Let $\mathcal{H}$ be a tangle Heegaard diagram of type 1 for a marked tangle $\mathcal{T}$ in a 3–manifold Y, equipped with an admissible almost complex structure J. If $\mathcal{H}$ is provincially admissible, then $\widetilde{CFAT}(\mathcal{H}, J)$ is an $\mathcal{A}_\infty$–module over $\mathcal{A}(\partial\mathcal{H})$. Moreover, if $\mathcal{H}$ is admissible, then $\widetilde{CFAT}(\mathcal{H}, J)$ is bounded.*

Theorem 10.4 *Up to $\mathcal{A}_\infty$ homotopy equivalence and tensoring with V, $\widetilde{CFAT}(\mathcal{H}, J)$ is independent of the choice of sufficiently generic admissible almost complex structure, and provincially admissible type 1 tangle Heegaard diagram for $(Y, \mathcal{T})$. Namely, if $\mathcal{H}_1$*

and $\mathcal{H}_2$ are provincially admissible type 1 *diagrams for $(Y, \mathcal{T})$ with almost complex structures J_1 and J_2, and $|\mathbb{X}_1| = |\mathbb{X}_2| + k$, then*

$$\widetilde{CFAT}(\mathcal{H}_1, J_1) \simeq \widetilde{CFAT}(\mathcal{H}_2, J_2) \otimes V^{\otimes k}.$$

Proof The proofs of the two theorems are analogous to those for $\widetilde{CFTD}$, except that we consider more degenerations, since we also consider sequences of sets of Reeb chords. □

When we write $\widetilde{CFAT}(Y, \mathcal{T})$, we mean the $\mathcal{A}_\infty$–module that we get from a diagram with $|\mathbb{X} \cap \operatorname{Int} \Sigma| = |\mathcal{T}|$.

10.3 The type *DA* bimodule

We define type *DA* structures for tangle Heegaard diagrams with two boundary components. One can similarly define type *AA*, *DD* and *AD* structures.

Suppose $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ is a provincially admissible diagram with two boundary components $\partial^0\mathcal{H}$ and $\partial^1\mathcal{H}$ for a strongly marked (m,n)–tangle $(Y, \mathcal{T}, \gamma)$. Let J be an admissible almost complex structure. We will define a type *DA* bimodule $\widetilde{CFDTA}(\mathcal{H}, J)$ over $\mathcal{A}(-\partial^0\mathcal{H})$ and $\mathcal{A}(\partial^1\mathcal{H})$.

As a left–right $(\mathcal{I}(-\partial^0\mathcal{H}), \mathcal{I}(\partial^1\mathcal{H}))$–bimodule, $\widetilde{CFDTA}(\mathcal{H}, J)$ is freely generated over $\mathbb{F}_2$ by $\mathfrak{S}(\mathcal{H})$, with actions of $\mathcal{I}(-\partial^0\mathcal{H})$ and $\mathcal{I}(\partial^1\mathcal{H})$ defined on the generators by

$$I(s_0) \cdot \boldsymbol{x} \cdot I(s_1) = \begin{cases} \boldsymbol{x} & \text{if } s_0 = \overline{o}^0(\boldsymbol{x}) \text{ and } s_1 = o^1(\boldsymbol{x}), \\ 0 & \text{otherwise.} \end{cases}$$

To define the type *DA* structure maps, we need to study slightly different moduli spaces than before. Given a decorated source $S^{\triangleright}$, let E_i be the set e punctures labeled by Reeb chords in $\partial^i\mathcal{H}$. We need to forget the relative heights of the punctures in E_0 to those in E_1.

Definition 10.5 Define the moduli space

$$\mathcal{M}^B_{\text{emb}}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P}_0, \vec{P}_1) = \bigcup_{\vec{P}|_{E_i} = \vec{P}_i} \mathcal{M}^B_{\text{emb}}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P}),$$

and define the index

$$\operatorname{ind}(B, \vec{\boldsymbol{\rho}}_0, \vec{\boldsymbol{\rho}}_1) = e(B) + n_{\boldsymbol{x}}(B) + n_{\boldsymbol{y}}(B) + |\vec{\boldsymbol{\rho}}_0| + |\vec{\boldsymbol{\rho}}_1| + \iota(\vec{\boldsymbol{\rho}}_0) + \iota(\vec{\boldsymbol{\rho}}_1),$$

where $\vec{\boldsymbol{\rho}}_i$ is a sequence of sets of Reeb chords in $\partial^i\mathcal{H}$.

On the $\partial^0\mathcal{H}$ side we will only allow discrete partitions. If $\vec{P}_0$ is discrete, and labeled by the sequence of Reeb chords $\vec{\rho}(\vec{P}_0) = (\rho_1, \dots, \rho_i)$, define

$$a_0(\boldsymbol{x}, \boldsymbol{y}, \vec{P}_0) := I(\bar{o}^0(\boldsymbol{x})) \cdot a(-\rho_1) \cdot \dots \cdot a(-\rho_i) \cdot I(\bar{o}^0(\boldsymbol{y})) \in \mathcal{A}(-\partial^0\mathcal{H}).$$

On the $\partial^1\mathcal{H}$ side we allow arbitrary partitions, and define

$$a_1(\boldsymbol{x}, \boldsymbol{y}, P_1) := I(o^1(\boldsymbol{x})) \cdot a(\boldsymbol{\rho}_1) \otimes \dots \otimes a(\boldsymbol{\rho}_j) \cdot I(o^1(\boldsymbol{y})) \in \mathcal{A}(\partial^1\mathcal{H})^{\otimes j},$$

where $\vec{\rho}(\vec{P}_1) = (\boldsymbol{\rho}_1, \dots, \boldsymbol{\rho}_j)$.

Finally, the structure maps are defined by

$$\begin{aligned}\delta^1_k(&\boldsymbol{x}, a_1, \dots, a_{k-1}) \\ &:= \sum_{\boldsymbol{y} \in \mathfrak{S}(\mathcal{H})} \sum_{\substack{B \in \tilde{\pi}_2(\boldsymbol{x}, \boldsymbol{y}) \\ \operatorname{ind}(B, \vec{\rho}(\vec{P}_0), \vec{\rho}(\vec{P}_1)) = 1 \\ a_1(\boldsymbol{x}, \boldsymbol{y}, P_1) = a_1 \otimes \dots \otimes a_{k-1}}} \#\mathcal{M}^B_{\mathrm{emb}}(\boldsymbol{x}, \boldsymbol{y}, S^{\triangleright}, \vec{P}_1, \vec{P}_2) \cdot a_0(\boldsymbol{x}, \boldsymbol{y}, \vec{P}_0) \otimes \boldsymbol{y}.\end{aligned}$$

Theorem 10.6 *Let $\mathcal{H}$ be a diagram with two boundary components for a strongly marked (m, n)–tangle $(Y, \mathcal{T}, \gamma)$, equipped with an admissible almost complex structure J. If $\mathcal{H}$ is provincially admissible, then $\widetilde{CFDTA}(\mathcal{H}, J)$ is a type DA bimodule over $\mathcal{A}(-\partial^0\mathcal{H})$ and $\mathcal{A}(\partial^1\mathcal{H})$. Moreover, if $\mathcal{H}$ is admissible, then $\widetilde{CFAT}(\mathcal{H}, J)$ is bounded.*

Theorem 10.7 *Up to homotopy equivalence and tensoring with $V := \mathbb{F}_2 \oplus \mathbb{F}_2$, the bimodule $\widetilde{CFDTA}(\mathcal{H}, J)$ is independent of the choice of sufficiently generic admissible almost complex structure, and provincially admissible tangle Heegaard diagram for $(Y, \mathcal{T}, \gamma)$. Namely, if $\mathcal{H}_1$ and $\mathcal{H}_2$ are provincially admissible diagrams for $(Y, \mathcal{T}, \gamma)$ with almost complex structures J_1 and J_2, and $|\mathbb{X}_1| = |\mathbb{X}_2| + k$, then*

$$\widetilde{CFDTA}(\mathcal{H}_1, J_1) \simeq \widetilde{CFDTA}(\mathcal{H}_2, J_2) \otimes V^{\otimes k}.$$

Proof The proofs are analogous to those for type D and type A structures. □

When we write $\widetilde{CFDTA}(Y, \mathcal{T}, \gamma)$, we mean the bimodule that we get from a diagram with $|\mathbb{X} \cap \operatorname{Int}\Sigma| = |\mathcal{T}|$. For a tangle $\mathcal{T}$ in $S^2 \times I$, there is a canonical framed arc γ determined by the product structure on the 3–manifold, With this framed arc, $\mathcal{T}$ becomes a strongly marked tangle, and we simply write $\widetilde{CFDTA}(\mathcal{T})$.

For here on, we suppress the almost complex structure J from the notation, and write $\widetilde{CFAT}(\mathcal{H})$, $\widetilde{CFTD}(\mathcal{H})$ and $\widetilde{CFDTA}(\mathcal{H})$.

10.4 Other diagrams and modules.

Similarly, one can associate a type A structure to a type 2 diagram, a type D structure to a type 1 diagram, or a type AA, DD or AD structure to a diagram with two boundary components.

One can also define β–bordered or α–β–bordered multipointed Heegaard diagrams for tangles, in the spirit of [6], and associate modules or bimodules, respectively. The bordered grid diagrams of Section 4 are examples of such diagrams.

11 Gradings

For now, we only discuss gradings when the tangle lies in B^3 or $S^2 \times I$. In those cases, one can define a homological grading by $\mathbb{Z}$, which we call the *Maslov* grading, and a second (internal) grading by $\frac{1}{2}\mathbb{Z}$, which we call the *Alexander* grading. In this section, all domains are assumed to avoid $\boldsymbol{z}$.

11.1 Algebra

Fix a marked matched circle $\mathcal{Z}$ for an n–marked sphere, and let $\mathcal{E} = (n+1, n+1, \mathrm{id}_{S_{\mathbb{X}}}, \mathrm{id}_{T_{\mathbb{O}}})$ be the corresponding shadow, as in Section 7.2. Recall that the algebra $\mathcal{A}(\mathcal{Z})$ does not depend on the $\mathbb{X}$ and $\mathbb{O}$ markings on $\mathcal{Z}$, and equivalently, as an ungraded algebra, $\widehat{\mathcal{A}}(\mathcal{E}) = \mathcal{A}(\mathcal{E})/(U_i = 0)$ does not depend on the sets $S_{\mathbb{X}}$ and $T_{\mathbb{O}}$. However, $\mathbb{X}$ and $\mathbb{O}$ markings play an important role in the bigrading on $\mathcal{A}(\mathcal{E})$ defined in Section 3.4.

The Maslov and Alexander gradings on $\mathcal{A}(\mathcal{E})$ defined in Section 3.4 descend to gradings on $\widehat{\mathcal{A}}(\mathcal{E})$, and thus to gradings M and A on $\mathcal{A}(\mathcal{Z})$ under the isomorphism from Proposition 7.5. The Maslov grading turns $\mathcal{A}(\mathcal{Z})$ into a differential graded algebra, and the Alexander grading is preserved by the differential and multiplication. We caution the reader that while the generators $I(\boldsymbol{s})a(\boldsymbol{\rho})$ are homogeneous with respect to the gradings, $a(\boldsymbol{\rho})$ are not.

11.2 Domains

Let $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbb{X}, \mathbb{O}, \boldsymbol{z})$ be a tangle diagram of any type. To define the Maslov and Alexander grading of a domain $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$, we will make use of the multiplicity map $m : H_1(\partial\Sigma \setminus \boldsymbol{z}, \partial\boldsymbol{\alpha}; \mathbb{Z}) \times H_0(\partial\boldsymbol{\alpha}; \mathbb{Z}) \to \frac{1}{2}\mathbb{Z}$ from Section 9.1 to record how $\partial^\partial B$ interacts with $\boldsymbol{x}$, $\boldsymbol{y}$, $\mathbb{O}$, and $\mathbb{X}$.

If $\mathcal{H}$ is a diagram with one boundary component (of type 1 or 2) for a $2n$–tangle, define the following sets. For $1 \le i \le 2n$, if there is an X on $\partial\Sigma$ between a_i and a_{i+1}, then we place an $\mathbb{O}$ marking in the interior of the segment $(a_i, a_{i+1}) \subset \partial\Sigma$ and another $\mathbb{O}$ marking in the segment (a_{4n+2-i}, a_{4n+3-i}). In other words, there is a new O on every component of $\partial\Sigma \setminus \boldsymbol{\alpha}$ that is on the boundary of a component of $\Sigma \setminus \boldsymbol{\alpha}$ with an O in the interior. Denote the set of all new $\mathbb{O}$ markings by $S^{\partial}_{\mathbb{O}}$. Define a set $S^{\partial}_{\mathbb{X}}$ analogously. Given a generator $\boldsymbol{x} \in \mathfrak{S}(\mathcal{H})$, define $S_{\boldsymbol{x}}$ to be the set of points in $\partial\boldsymbol{\alpha}$ that lie on α–arcs occupied by $\boldsymbol{x}$, and define $\overline{S}_{\boldsymbol{x}} := \partial\boldsymbol{\alpha} \setminus S_{\boldsymbol{x}}$.

If $\mathcal{H}$ is a diagram with two boundary components $\partial^i\mathcal{H}$ for $i = 0, 1$, define sets $S^i_{\mathbb{O}}$, $S^i_{\mathbb{X}}$, $S^i_{\boldsymbol{x}}$ and $\overline{S}^i_{\boldsymbol{x}}$ analogously.

When counting multiplicities below, we view a subset $S \subset \partial\boldsymbol{\alpha}$ as the element of $H_0(\partial\boldsymbol{\alpha}; \mathbb{Z})$ for which each point of S comes with coefficient $+1$, so that we can add and subtract sets.

Note that even though $S^{\partial}_{\mathbb{O}}$ is not a subset of $\partial\boldsymbol{\alpha}$, defining $m([\partial^{\partial} B], S^{\partial}_{\mathbb{O}})$ makes sense as a generalization of the multiplicity function m. Precisely, think of an interval (a_i, a_{i+1}) as a generator of $H_1(\partial\Sigma \setminus \boldsymbol{z}, \partial\boldsymbol{\alpha}; \mathbb{Z})$, and define $m([\partial^{\partial} B], S^{\partial}_{\mathbb{O}})$ as the sum of the coefficients in $[\partial^{\partial} B]$ of all intervals (a_i, a_{i+1}) that contain an $O \in S^{\partial}_{\mathbb{O}}$. Define multiplicity counts for $S^{\partial}_{\mathbb{X}}$, $S^i_{\mathbb{O}}$ and $S^i_{\mathbb{X}}$ similarly.

Let $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ be a domain for a diagram with two boundary components. Define

$$\begin{aligned} M(B) &= -e(B) - n_{\boldsymbol{x}}(B) - n_{\boldsymbol{y}}(B) + \tfrac{1}{2} m([\partial^{\partial} B], -\overline{S}^0_{\boldsymbol{x}} - \overline{S}^0_{\boldsymbol{y}} + S^1_{\boldsymbol{x}} + S^1_{\boldsymbol{y}}) \\ &\qquad + m([\partial^{\partial} B], S^0_{\mathbb{X}} - S^1_{\mathbb{O}}) + 2n_{\mathbb{O}}(B), \\ A(B) &= \tfrac{1}{2} m([\partial^{\partial} B], S^0_{\mathbb{X}} - S^0_{\mathbb{O}} + S^1_{\mathbb{X}} - S^1_{\mathbb{O}}) + n_{\mathbb{O}}(B) - n_{\mathbb{X}}(B). \end{aligned}$$

For a domain $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ on a type 1 diagram, define

$$\begin{aligned} M(B) &= -e(B) - n_{\boldsymbol{x}}(B) - n_{\boldsymbol{y}}(B) + \tfrac{1}{2} m([\partial^{\partial} B], S_{\boldsymbol{x}} + S_{\boldsymbol{y}}) - m([\partial^{\partial} B], S^{\partial}_{\mathbb{O}}) + 2n_{\mathbb{O}}(B), \\ A(B) &= \tfrac{1}{2} m([\partial^{\partial} B], S^{\partial}_{\mathbb{X}} - S^{\partial}_{\mathbb{O}}) + n_{\mathbb{O}}(B) - n_{\mathbb{X}}(B). \end{aligned}$$

For a domain $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ on a type 2 diagram, define

$$\begin{aligned} M(B) &= -e(B) - n_{\boldsymbol{x}}(B) - n_{\boldsymbol{y}}(B) - \tfrac{1}{2} m([\partial^{\partial} B], \overline{S}_{\boldsymbol{x}} + \overline{S}_{\boldsymbol{y}}) + m([\partial^{\partial} B], S^{\partial}_{\mathbb{X}}) + 2n_{\mathbb{O}}(B), \\ A(B) &= \tfrac{1}{2} m([\partial^{\partial} B], S^{\partial}_{\mathbb{X}} - S^{\partial}_{\mathbb{O}}) + n_{\mathbb{O}}(B) - n_{\mathbb{X}}(B). \end{aligned}$$

Note that the bigrading is additive under union (when we continue to view the sum of regions as a domain between the same two generators $\boldsymbol{x}$ and $\boldsymbol{y}$).

We will soon define the bigrading on the modules and bimodules. To show it is well-defined, we need to show that the bigrading on domains is additive under composition, and that it is zero on periodic domains.

Proposition 11.1 *For any periodic domain $B \in \pi_2(\boldsymbol{x},\boldsymbol{x})$, we have $M(B)=0$ and $A(B)=0$.*

Proof Since the bigrading is additive under union, we only need to show it is zero on provincial periodic domains, and on the regions of $\Sigma \setminus \boldsymbol{\alpha}$ that intersect the boundary $\partial\Sigma$. The proofs for all three types of diagrams are identical. The write-up below is for a type 1 diagram.

For a periodic domain, the bigrading simplifies to

$$\begin{aligned} M(B) &= -e(B) - 2n_{\boldsymbol{x}}(B) + m([\partial^\partial B], S_{\boldsymbol{x}}) - m([\partial^\partial B], S^\partial_{\mathbb{O}}) + 2n_{\mathbb{O}}(B),\\ A(B) &= \tfrac{1}{2} m([\partial^\partial B], S^\partial_{\mathbb{X}} - S^\partial_{\mathbb{O}}) + n_{\mathbb{O}}(B) - n_{\mathbb{X}}(B). \end{aligned}$$

Let D_i be the region of $\Sigma \setminus \boldsymbol{\alpha}$ whose α–arcs boundary consists of α_i^a and α_{i+1}^a. Geometrically, D_i is a rectangle with $t \geq 0$ disks removed from its interior, so it has Euler measure $e(D_i) = -t$. Each of the t circle boundary components is an α–circle, hence it contains a point of $\boldsymbol{x}$ on it, and contributes 1 to the count of $2n_{\boldsymbol{x}}(D_i)$. Each of the arcs α_i and α_{i+1} that is occupied by $\boldsymbol{x}$ contributes 1 to $n_{\boldsymbol{x}}(D_i)$ and 1 to $m([\partial^\partial B], S_{\boldsymbol{x}})$. There are no other contributions to $m([\partial^\partial D_i], S_{\boldsymbol{x}})$, so $-e(D_i) - 2n_{\boldsymbol{x}}(D_i) + m([\partial^\partial D_i], S_{\boldsymbol{x}}) = 0$. Last, D_i contains exactly one O or exactly one X. In either case, $m([\partial^\partial D_i], S^\partial_{\mathbb{O}}) = 2n_{\mathbb{O}}(D_i)$, and $m([\partial^\partial D_i], S^\partial_{\mathbb{X}}) = 2n_{\mathbb{X}}(D_i)$. It follows that $M(D_i) = 0$ and $A(D_i) = 0$.

For a provincial periodic domain B, the bigrading becomes

$$\begin{aligned} M(B) &= -e(B) - 2n_{\boldsymbol{x}}(B) + 2n_{\mathbb{O}}(B),\\ A(B) &= n_{\mathbb{O}}(B) - n_{\mathbb{X}}(B), \end{aligned}$$

which agrees with the bigrading for knot Floer homology, and has been shown to be zero in the case of knots and links in S^3 (note that a bordered diagram for a tangle in B^3 or $S^2 \times I$ can be completed to a diagram for a knot or a link in S^3, so we can think of B as a domain in the closed diagram). □

The proof of additivity under composition is a bit trickier, as there is linking information we need to consider.

Proposition 11.2 *If $B_1 \in \pi_2(\boldsymbol{x},\boldsymbol{y})$ and $B_2 \in \pi_2(\boldsymbol{y},\boldsymbol{w})$, then*

$$M(B_1 * B_2) = M(B_1) + M(B_2) \quad \text{and} \quad A(B_1 * B_2) = A(B_1) + A(B_2).$$

Proof Once again we write up the proof for a type 1 diagram, as the notation in this case is the lightest. The other two cases are identical.

First observe that $m([\partial^\partial B], S^\partial_{\mathbb{O}})$, $n_{\mathbb{O}}(B)$, $m([\partial^\partial B], S^\partial_{\mathbb{X}})$ and $n_{\mathbb{X}}(B)$ are all clearly additive under composition, so the statement follows for the Alexander grading.

Let $B = B_1 * B_2$. Let R_1 be a union of the regions D_i as in Proposition 11.1 with multiplicity, so that $B_1' = B_1 + R_1 \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ only covers $\partial\Sigma$ inside the interval $[a_1, a_{2n+1}]$. Similarly, let R_2 be a union of regions D_i so that $B_2' = B_2 + R_2 \in \pi_2(\boldsymbol{y}, \boldsymbol{w})$ only covers $\partial\Sigma$ inside the interval $[a_1, a_{2n+1}]$. Let $B' = B_1' * B_2'$, and note that $B' = B + R_1 + R_2$. Since the Maslov grading is additive under union, and by Proposition 11.1, we have that $M(B_i') = M(B_i)$ and $M(B') = M(B)$. So it suffices to show that $M(B') = M(B_1') + M(B_2')$.

To simplify notation, write $a = \partial^\partial B_1'$ and $b = \partial^\partial B_2'$, and note that $\partial^\partial B' = a + b$. By [8, Lemma 10.4] and since $m([\partial^\partial B], S^\partial_{\mathbb{O}})$ and $n_{\mathbb{O}}(B)$ are additive under composition,

$$\begin{aligned}
&M(B') - M(B_1') + M(B_2')\\
&\quad = L(a,b) + \tfrac{1}{2}\big(m(a+b, S_{\boldsymbol{x}}) + m(a+b, S_{\boldsymbol{w}}) - m(a, S_{\boldsymbol{x}}) - m(a, S_{\boldsymbol{y}})\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad - m(b, S_{\boldsymbol{y}}) - m(b, S_{\boldsymbol{w}})\big)\\
&\quad = L(a,b) + \tfrac{1}{2}\big(m(b, S_{\boldsymbol{x}}) + m(a, S_{\boldsymbol{w}}) - m(a, S_{\boldsymbol{y}}) - m(b, S_{\boldsymbol{y}})\big).
\end{aligned}$$

Recall that $L(a,b) = m(b, \partial a) = -m(a, \partial b)$, so $L(a,b) = \frac{1}{2}(m(b,\partial a) - m(a, \partial b))$. Thus, showing that $M(B') = M(B_1') + M(B_2')$ is equivalent to showing that

$$m(b,\partial a) - m(a,\partial b) + m(b, S_{\boldsymbol{x}}) + m(a, S_{\boldsymbol{w}}) - m(a, S_{\boldsymbol{y}}) - m(b, S_{\boldsymbol{y}}) = 0.$$

Extend the matching

$$\mu\colon \{a_1, \dots, a_{4n+2}\} \to [2n+1]$$

linearly to a function $\mu_{\mathbb{Z}}\colon H_0(\partial\boldsymbol{\alpha}; \mathbb{Z}) \to \mathbb{Z}^{2n+1}$. For a generator $\boldsymbol{x}$, think of $o(\boldsymbol{x})$ as an element of $\mathbb{Z}^{2n+1}$ where each occupied arc comes with coefficient $+1$. Since B_1 is a homology class in $\pi_2(\boldsymbol{x},\boldsymbol{y})$, we have $\partial a = o(\boldsymbol{y}) - o(\boldsymbol{x})$. Similarly, $\partial b = o(\boldsymbol{w}) - o(\boldsymbol{y})$.

Let $S_{\boldsymbol{x}}^{\text{bottom}} = S_{\boldsymbol{x}} \cap \{a_1, \dots, a_{2n+1}\}$, and let $S_{\boldsymbol{x}}^{\text{top}} = S_{\boldsymbol{x}} \cap \{a_{2n+2}, \dots, a_{4n+2}\}$. Recall that we view any subset $S \subset \partial\boldsymbol{\alpha}$ as the element of $H_0(\partial\boldsymbol{\alpha}; \mathbb{Z})$ where each point of S comes with coefficient $+1$.

Since $[a_1, a_{2n+1}] \subset \partial\Sigma$ only contains one endpoint of each α–arc, and since $\mu_{\mathbb{Z}}(\partial a) = o(\boldsymbol{y}) - o(\boldsymbol{x})$, it follows that ∂a can only be the section $S_{\boldsymbol{y}}^{\text{bottom}} - S_{\boldsymbol{x}}^{\text{bottom}}$ of $o(\boldsymbol{y}) - o(\boldsymbol{x})$. Then $m(b, \partial a) = m(b, S_{\boldsymbol{y}}^{\text{bottom}} - S_{\boldsymbol{x}}^{\text{bottom}})$. Since b only covers the "bottom" of $\partial\Sigma$, ie $[a_1, \dots, a_{2n+1}]$, the multiplicity of b at a_i is zero whenever $i \geq 2n+2$, so $m(b, \partial a) = m(b, S_{\boldsymbol{y}} - S_{\boldsymbol{x}})$. Similarly, $m(a, \partial b) = m(a, S_{\boldsymbol{w}} - S_{\boldsymbol{y}})$. This completes the proof. □

11.3 Modules and bimodules

Let $\mathcal{H}$ be a diagram of type 1 or type 2 for some pair $(B^3, \mathcal{T})$.

Proposition 11.3 *Given $\boldsymbol{x}$, $\boldsymbol{y} \in \mathfrak{S}(\mathcal{H})$, $\pi_2(\boldsymbol{x}, \boldsymbol{y})$ is nonempty.*

Proof The proof is identical to that of [8, Lemma 4.21]. Connect $\boldsymbol{x}$ to $\boldsymbol{y}$ by a union of paths $\gamma_\alpha \subset \boldsymbol{\alpha} \cup (\partial\Sigma \setminus \boldsymbol{z})$ and $\gamma_\beta \subset \boldsymbol{\beta}$. Then $\boldsymbol{x}$ and $\boldsymbol{y}$ are connected by a domain if and only if $\gamma_\alpha - \gamma_\beta$ can be made null-homologous in Σ by adding or subtracting entire α–curves and β–circles, if and only if the image of $\gamma_\alpha - \gamma_\beta$ in $H_1\big(\Sigma \times I, \boldsymbol{\alpha} \times \{1\} \cup \boldsymbol{\beta} \times \{0\} \cup (\partial\Sigma \setminus \boldsymbol{z}) \times I\big) \cong H_1(B^3, \partial B^3)$ is zero. But $H_1(B^3, \partial^3) = 0$, so this is always the case. □

Since any two generators $\boldsymbol{x}, \boldsymbol{y} \in \mathfrak{S}(\mathcal{H})$ are connected by a domain, we can define relative gradings

$$M(\boldsymbol{y}) - M(\boldsymbol{x}) = M(B),$$
$$A(\boldsymbol{y}) - A(\boldsymbol{x}) = A(B),$$

where $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$. We can assume B does not cross $\boldsymbol{z}$: if any domain B' intersects $\boldsymbol{z}$, we can add copies of the periodic domain(s) that are the region(s) of $\Sigma \setminus \boldsymbol{\alpha}$ containing the points/arc $\boldsymbol{z}$, to obtain a domain $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ that avoids $\boldsymbol{z}$.

When $\mathcal{H}$ is a diagram for $(S^2 \times I, \mathcal{T})$, it is no longer true that any two generators are connected by a domain. However, the *DA* bimodule splits as

$$\widetilde{CFDTA}(\mathcal{H}) \cong \bigoplus_{i=0}^{2m+1} \widetilde{CFDTA}_i(\mathcal{H}),$$

where $\widetilde{CFDTA}_i(\mathcal{H})$ is generated by $\mathfrak{S}_i := \{\boldsymbol{x} \in \mathfrak{S} : |o^0(\boldsymbol{x})| = i\}$.

Lemma 11.4 *For a fixed i, and for any $\boldsymbol{x}$, $\boldsymbol{y} \in \widetilde{CFDTA}_i(\mathcal{H})$, we have $\pi_2(\boldsymbol{x}, \boldsymbol{y}) \neq \varnothing$.*

Proof Let $\boldsymbol{x}_{\mathrm{dr}}$ and $\boldsymbol{y}_{\mathrm{dr}}$ be the generators corresponding to $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{H}_{\mathrm{dr}}$. There is some domain $B_{\mathrm{dr}} \in \pi_2(\boldsymbol{x}_{\mathrm{dr}}, \boldsymbol{y}_{\mathrm{dr}})$, since $\mathcal{H}_{\mathrm{dr}}$ is a diagram for B^3. Add copies of the two periodic regions of $\Sigma_{\mathrm{dr}} \setminus \boldsymbol{\alpha}$ containing $\boldsymbol{z}_1$ and $\{z^{\mathrm{front}}, z^{\mathrm{back}}\}$, to obtain a domain $B'_{\mathrm{dr}} \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ with zero multiplicity at $\boldsymbol{z}_1$ and z^{back}, resulting in some multiplicity p at z^{front}. Write $\boldsymbol{\alpha}^0$ for the set $\{\alpha^0_1, \dots, \alpha^0_{m+1}\}$, and $\boldsymbol{\alpha}^1$ for the set $\{\alpha^1_1, \dots, \alpha^1_{n+1}\}$. Let $S \in \mathbb{Z}\langle\partial(\boldsymbol{\alpha}^0 \cup \boldsymbol{\alpha}^1)\rangle$ be the set of points (with sign and multiplicity) in the boundary of $\partial^\partial B'_{\mathrm{dr}}$. The matching μ for the pointed matched circle $\partial\mathcal{H}_{\mathrm{dr}}$ extends bilinearly to a map $\mu_*\colon \mathbb{Z}\langle\partial\boldsymbol{\alpha}\rangle \to \mathbb{Z}\langle[m+n+2]\rangle$. Since $\mu_*(S) = o(\boldsymbol{y}_{\mathrm{dr}}) - o(\boldsymbol{x}_{\mathrm{dr}})$, and $o^0(\boldsymbol{x})$ and $o^0(\boldsymbol{y})$ both have cardinality i, it follows that $p = 0$, so after attaching a 1–handle at $\{z^{\mathrm{front}}, z^{\mathrm{back}}\}$, B'_{dr} becomes a domain B' on $\mathcal{H}$. □

Define relative gradings on $\widetilde{CFDTA}_i(\mathcal{H})$ by

$$M(\boldsymbol{y}) - M(\boldsymbol{x}) = M(B),$$
$$A(\boldsymbol{y}) - A(\boldsymbol{x}) = A(B),$$

where $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ (again arrange for B to have zero multiplicities at z_1 and z_2 by adding copies of the corresponding regions of $\Sigma \setminus \boldsymbol{\alpha}$, if necessary).

By Proposition 11.1, the relative bigrading on the modules and bimodules is well-defined.

Proposition 11.5 *The various structures defined in this section are graded (A, D or DA) (bi)modules with respect to the grading M. Further, the internal grading A is preserved by all structure maps.*

The proof is based on understanding the relation between the bigrading on a domain with a compatible sequence of sets of Reeb chords and the bigrading on the algebra elements associated to the Reeb chords. We start by relating the Maslov grading of algebra generators to ι.

Lemma 11.6 *Let $a = I(\boldsymbol{s})a(\boldsymbol{\rho})I(\boldsymbol{t})$ be a generator for $\mathcal{A}(\mathcal{Z})$. Then*

$$M(a) - \iota(\boldsymbol{\rho}) = \tfrac{1}{2} m([\boldsymbol{\rho}], S + T) - m([\boldsymbol{\rho}], S^{\partial}_{\mathbb{O}}),$$

where $S = \mu^{-1}(\boldsymbol{s})$ and $T = \mu^{-1}(\boldsymbol{t})$.

Proof Let $a' = (\boldsymbol{s}, \boldsymbol{t}, \phi)$ be the element in $\widehat{\mathcal{A}}(\bar{\mathcal{E}})$ corresponding to a under the isomorphism $\mathcal{A}(\mathcal{Z}) \cong \widehat{\mathcal{A}}(\bar{\mathcal{E}})$ discussed earlier. Recall that $M(a') = \mathrm{inv}(\phi) - \mathrm{inv}(\phi, \omega) + \mathrm{inv}(\omega)$. Decompose $\boldsymbol{s}$ as $\boldsymbol{s}^+ \sqcup \boldsymbol{s}^- \sqcup \boldsymbol{s}^0$, so that $\phi^+ := \phi|_{\boldsymbol{s}^+}$ is increasing, $\phi^- := \phi|_{\boldsymbol{s}^-}$ is decreasing, and $\phi^0 := \phi|_{\boldsymbol{s}^0}$ is the identity. Then

$$\begin{aligned} M(a') &= \mathrm{inv}(\phi) - \mathrm{inv}(\phi, \omega) \\ &= \mathrm{inv}(\phi^+) + \mathrm{inv}(\phi^-) + \mathrm{inv}(\phi^+, \phi^0) + \mathrm{inv}(\phi^-, \phi^0) + \mathrm{inv}(\phi^+, \phi^-) - \mathrm{inv}(\phi, \omega) \\ &= \mathrm{inv}(\boldsymbol{\rho}) + \mathrm{inv}(\phi^+, \phi^0) + \mathrm{inv}(\phi^-, \phi^0) + \mathrm{inv}(\phi^+, \phi^-) - \mathrm{inv}(\phi, \omega). \end{aligned}$$

By [8, Lemma 5.57], $\iota(\boldsymbol{\rho})$ can be written as

$$\iota(\boldsymbol{\rho}) = \mathrm{inv}(\boldsymbol{\rho}) - m([\boldsymbol{\rho}], S(\boldsymbol{\rho})),$$

where $S(\boldsymbol{\rho})$ is the set of initial endpoints of $\boldsymbol{\rho}$.

The upward-veering strands in ϕ, ie the strands for ϕ^+, correspond to the set of Reeb chords $\boldsymbol{\rho}^+ \subset \boldsymbol{\rho}$ contained in $[a_{2n+2}, a_{4n+2}]$, and the downward-veering strands in ϕ

correspond to the Reeb chords $\boldsymbol{\rho}^- \subset \boldsymbol{\rho}$ contained in $[a_1, a_{2n+1}]$. The horizontal strands of ϕ correspond to the projection under the matching μ of the dashed horizontal strands in the strands diagram for a. Let $S(\boldsymbol{\rho}^+)$ and $S(\boldsymbol{\rho}^-)$ be the sets of initial endpoints of $\boldsymbol{\rho}^+$ and $\boldsymbol{\rho}^-$, respectively. Note that $S(\boldsymbol{\rho}^+)$ is the section of $\boldsymbol{s}^+$ contained in $[a_1, a_{2n+1}]$, and $S(\boldsymbol{\rho}^-)$ is the section of $\boldsymbol{s}^-$ contained in $[a_{2n+2}, a_{4n+2}]$. Equivalently, $S(\boldsymbol{\rho}^+) = \mu^{-1}(\boldsymbol{s}^+) \cap [a_1, a_{2n+1}]$ and $S(\boldsymbol{\rho}^-) = \mu^{-1}(\boldsymbol{s}^-) \cap [a_{2n+2}, a_{4n+2}]$.

Decompose S as $S = S^+ \sqcup S^- \sqcup S^0$, where $S^+ = \mu^{-1}(\boldsymbol{s}^+)$, $S^+ = \mu^{-1}(\boldsymbol{s}^-)$, and S^0 is the set of initial points for the dashed horizontal strands. Decompose T similarly by the type of final endpoints as $T = T^+ \sqcup T^- \sqcup T^0$. Note that the multiplicity of a Reeb chord in $[a_1, a_{2n+1}]$ is zero at any point in $[a_{2n+2}, a_{4n+2}]$, and similarly the multiplicity of a Reeb chord in $[a_{2n+2}, a_{4n+2}]$ is zero at any point in $[a_1, a_{2n+1}]$, so $m([\boldsymbol{\rho}^+], S(\boldsymbol{\rho})) = m([\boldsymbol{\rho}^+], S(\boldsymbol{\rho}^+)) = m([\boldsymbol{\rho}^+], S^+)$, and similarly $m([\boldsymbol{\rho}^-], S(\boldsymbol{\rho})) = m([\boldsymbol{\rho}^-], S^-)$. Since $[\boldsymbol{\rho}] = [\boldsymbol{\rho}^+] + [\boldsymbol{\rho}^-]$,

$$\iota(\boldsymbol{\rho}) = \operatorname{inv}(\boldsymbol{\rho}) - m([\boldsymbol{\rho}^+], S^+) - m([\boldsymbol{\rho}^-], S^-).$$

Next,we express $M(a')$ in terms of $\boldsymbol{\rho}$. Observe that

$$\operatorname{inv}(\phi^+, \phi^0) + \operatorname{inv}(\phi^-, \phi^0) = m([\boldsymbol{\rho}^+], S^0) + m([\boldsymbol{\rho}^-], S^0)$$

and

$$\operatorname{inv}(\phi, \omega) = m([\boldsymbol{\rho}], S^{\partial}_{\mathbb{O}}).$$

It remains to understand $\operatorname{inv}(\phi^+, \phi^-)$. Let s^- and s^+ be a downward-veering and an upward-veering strand, and let ρ^- and ρ^+ be the corresponding Reeb chords on $\mathcal{Z}$. The strands s^- and s^+ cross exactly when one of the following happens:

- The initial endpoint of s^+ is between the initial and final endpoints of s^-. This happens exactly when
$$m\big([\rho^-], \mu^{-1}\big(\mu(S(\rho^+))\big)\big) = 1 \quad \text{and} \quad m\big([\rho^+], \mu^{-1}\big(\mu(S(\rho^-))\big)\big) = 0.$$
- The initial endpoint of s^- is between the initial and final endpoints of s^+. Equivalently, $m\big([\rho^+], \mu^{-1}\big(\mu(S(\rho^-))\big)\big) = 1$ and $m\big([\rho^-], \mu^{-1}\big(\mu(S(\rho^+))\big)\big) = 0$.
- The initial endpoint of s^+ is the final endpoint of s^-, ie
$$m\big([\rho^+], \mu^{-1}\big(\mu(S(\rho^-))\big)\big) = \tfrac{1}{2} \quad \text{and} \quad m\big([\rho^-], \mu^{-1}\big(\mu(S(\rho^+))\big)\big) = \tfrac{1}{2}.$$

The strands do not cross if and only if

$$m\big([\rho^+], \mu^{-1}\big(\mu(S(\rho^-))\big)\big) = 0 = m\big([\rho^-], \mu^{-1}\big(\mu(S(\rho^+))\big)\big).$$

By linearity then,

$$\begin{aligned}\operatorname{inv}(\phi^+,\phi^-) &= \sum_{\rho^-\in\boldsymbol{\rho}^-,\,\rho^+\in\boldsymbol{\rho}^+} m\big([\rho^-],\mu^{-1}\big(\mu(S(\rho^+))\big)+m\big([\rho^+],\mu^{-1}\big(\mu(S(\rho^-))\big)\big)\big)\\ &= m([\boldsymbol{\rho}^-],S^+)+m([\boldsymbol{\rho}^+],S^-).\end{aligned}$$

So,

$$\begin{aligned}M(a') &= \operatorname{inv}(\boldsymbol{\rho})+\operatorname{inv}(\phi^+,\phi^0)+\operatorname{inv}(\phi^-,\phi^0)+\operatorname{inv}(\phi^+,\phi^-)-\operatorname{inv}(\phi,\omega)\\ &= \operatorname{inv}(\boldsymbol{\rho})+m([\boldsymbol{\rho}^+],S^0)+m([\boldsymbol{\rho}^-],S^0)+m([\boldsymbol{\rho}^-],S^+)+m([\boldsymbol{\rho}^+],S^-)\\ &\qquad -m([\boldsymbol{\rho}],S^\partial_{\mathbb{O}})\\ &= \operatorname{inv}(\boldsymbol{\rho})+m([\boldsymbol{\rho}^+],S)+m([\boldsymbol{\rho}^-],S)-m([\boldsymbol{\rho}^-],S^-)-m([\boldsymbol{\rho}^+],S^+)-m([\boldsymbol{\rho}],S^\partial_{\mathbb{O}})\\ &= \iota(\boldsymbol{\rho})+m([\boldsymbol{\rho}^+],S)+m([\boldsymbol{\rho}^-],S)-m([\boldsymbol{\rho}],S^\partial_{\mathbb{O}})\\ &= \iota(\boldsymbol{\rho})+m([\boldsymbol{\rho}],S)-m([\boldsymbol{\rho}],S^\partial_{\mathbb{O}}).\end{aligned}$$

It is not hard to see that $m([\boldsymbol{\rho}],S)=m([\boldsymbol{\rho}],T)$, so

$$M(a')-\iota(\boldsymbol{\rho})=\tfrac{1}{2}m([\boldsymbol{\rho}],S+T)-m([\boldsymbol{\rho}],S^\partial_{\mathbb{O}}). \qquad \square$$

Let $B\in\pi_2(\boldsymbol{x},\boldsymbol{y})$ be a domain for a diagram $\mathcal{H}$ with two boundary components, let $\vec{\boldsymbol{\rho}}_0=(\rho_1,\dots,\rho_i)$ be a sequence of Reeb chords on $\partial^0\mathcal{H}$, and let $\vec{\boldsymbol{\rho}}_1=(\boldsymbol{\rho}_1,\dots,\boldsymbol{\rho}_j)$ be a sequence of sets of Reeb chords on $\partial^1\mathcal{H}$, both compatible with B. Recall that we write

$$a_0:=a_0(\boldsymbol{x},\boldsymbol{y},\vec{P}_0)=I(\bar{o}^0(\boldsymbol{x}))\cdot a(-\rho_1)\cdots a(-\rho_i)\cdot I(\bar{o}^0(\boldsymbol{y}))\in\mathcal{A}(-\partial^0\mathcal{H})$$

and

$$a_1(\boldsymbol{x},\boldsymbol{y},P_1)=I(o^1(\boldsymbol{x}))\cdot a(\boldsymbol{\rho}_1)\otimes\cdots\otimes a(\boldsymbol{\rho}_j)\cdot I(o^1(\boldsymbol{y}))\in\mathcal{A}(\partial^1\mathcal{H})^{\otimes j},$$

and observe that we can equivalently write $a_1(\boldsymbol{x},\boldsymbol{y},P_1)$ as

$$a_1(\boldsymbol{x},\boldsymbol{y},P_1)=I(o^1(\boldsymbol{x}))a(\boldsymbol{\rho}_1)I_1\otimes I_1a(\boldsymbol{\rho}_2)I_2\cdots\otimes I_{j-1}a(\boldsymbol{\rho}_j)I(o^1(\boldsymbol{y})).$$

Denote $I(o^1(\boldsymbol{x}))a(\boldsymbol{\rho}_1)I_1,\dots,I_{j-1}a(\boldsymbol{\rho}_j)I(o^1(\boldsymbol{y}))$ by $a_1,\dots,a_j$.

Proposition 11.7 *For the triple $(B,\vec{\boldsymbol{\rho}}_0,\vec{\boldsymbol{\rho}}_1)$ we have*

$$M(B,\vec{\boldsymbol{\rho}}_0,\vec{\boldsymbol{\rho}}_1)=|\vec{\boldsymbol{\rho}}_1|-\operatorname{ind}(B,\vec{\boldsymbol{\rho}}_0,\vec{\boldsymbol{\rho}}_1)+\sum_{t=1}^{j}M(a_t)-M(a_0)+2n_{\mathbb{O}}(B),$$

$$A(B,\vec{\boldsymbol{\rho}}_0,\vec{\boldsymbol{\rho}}_1)=\sum_{t=1}^{j}A(a_t)-A(a_0)+n_{\mathbb{O}}(B)-n_{\mathbb{X}}(B).$$

Proof The equality for the Alexander grading follows immediately from the definition.

For the Maslov grading, denote the right-hand side of the equation by R. Note that while $a_1 \otimes \cdots \otimes a_j \neq 0$, it may be that $a_1 \cdots a_j = 0$. Resolve crossings in each a_t if necessary to get a nonzero product $a' = a'_1 \cdots a'_j$. Note that $\iota(a'_t) = \iota(a_t) - c_t$ and $M(a'_t) = M(a_t) - c_t$, where c_t is the number of resolved crossings to get from a_t to a'_t, and $L([a_s],[a_t]) = L([a'_s],[a'_t])$, since resolving crossings does not change the homology class. Then

$$\begin{aligned}\sum_{t=1}^{j} M(a_j) - \iota(\vec{\rho}_1) &= \sum_{t=1}^{j} M(a_j) - \sum_{t=1}^{j} \iota(\boldsymbol{\rho}_t) - \sum_{s<t} L([\boldsymbol{\rho}_s],[\boldsymbol{\rho}_t]) \\ &= \sum_{t=1}^{j} M(a_j) - \sum_{t=1}^{j} \iota(a_t) - \sum_{s<t} L([a_s],[a_t]) \\ &= \sum_{t=1}^{j} M(a'_j) - \sum_{t=1}^{j} \iota(a'_t) - \sum_{s<t} L([a'_s],[a'_t]) \\ &= M(a') - \iota(a').\end{aligned}$$

By [22, Lemma 18], $\iota(a_0) = -|\vec{\rho}_0| - \iota(\vec{\rho}_0)$.

Substituting the definition of ind in R, we get

$$\begin{aligned}R &= -e(B) - n_{\boldsymbol{x}}(B) - n_{\boldsymbol{y}}(B) - |\vec{\rho}_0| - \iota(\vec{\rho}_0) - \iota(\vec{\rho}_1) + \sum_{t=1}^{j} M(a_t) - M(a_0) + 2n_{\mathbb{O}}(B) \\ &= -e(B) - n_{\boldsymbol{x}}(B) - n_{\boldsymbol{y}}(B) - |\vec{\rho}_0| - \iota(\vec{\rho}_0) + M(a') - \iota(a') - M(a_0) + 2n_{\mathbb{O}}(B) \\ &= -e(B) - n_{\boldsymbol{x}}(B) - n_{\boldsymbol{y}}(B) + \iota(a_0) + M(a') - \iota(a') - M(a_0) + 2n_{\mathbb{O}}(B).\end{aligned}$$

Applying Lemma 11.6 to a' and a_0, and since $[a'] = [\vec{\rho}_1]$, we get $R = M(B, \vec{\rho}_0, \vec{\rho}_1)$. □

The equalities below for a type 1 or type 2 diagram are a special case of Proposition 11.7, and follow immediately.

Proposition 11.8 *For a domain $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ on a type 2 diagram, and a sequence of Reeb chords $\vec{\rho}$,*

$$\begin{aligned}M(B, \vec{\rho}) &= -\operatorname{ind}(B, \vec{\rho}) - M(-\vec{\rho}) + 2n_{\mathbb{O}}(B), \\ A(B, \vec{\rho}) &= -A(-\vec{\rho}) + n_{\mathbb{O}}(B) - n_{\mathbb{X}}(B).\end{aligned}$$

Proposition 11.9 *For a domain $B \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$ on a type 1 diagram, and a sequence of sets of Reeb chords $\vec{\boldsymbol{\rho}} = (\boldsymbol{\rho}_1, \dots, \boldsymbol{\rho}_l)$,*

$$M(B, \vec{\boldsymbol{\rho}}) = |\vec{\boldsymbol{\rho}}| - \operatorname{ind}(B, \vec{\boldsymbol{\rho}}) + \sum_{i=1}^{l} M(\boldsymbol{\rho}_i) + 2n_{\mathbb{O}}(B),$$

$$A(B, \vec{\boldsymbol{\rho}}) = \sum_{i=1}^{l} A(\boldsymbol{\rho}_i) + n_{\mathbb{O}}(B) - n_{\mathbb{X}}(B).$$

Proposition 11.5 follows:

Proof of Proposition 11.5 All algebraic structures here are defined by counting curves of index 1. The claim follows directly by substituting 1 for the index in the grading formulas from Propositions 11.7, 11.8, and 11.9. □

11.4 Tensor products

It is easy to see that the bigrading on domains is additive under gluing.

Proposition 11.10 *If $\mathcal{H}_1$ and $\mathcal{H}_2$ are diagrams with $\partial^1\mathcal{H}_1 = -\partial^0\mathcal{H}_2$, and B is a domain on $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$ that decomposes as $B = B_1 \times B_2$, with B_i a domain on $\mathcal{H}_i$, then $M(B) = M(B_1) + M(B_2)$ and $A(B) = A(B_1) + A(B_2)$.*

Proof This follows directly from the definitions of M and A. □

Thus, for a generator $\boldsymbol{x} = \boldsymbol{x}_1 \cup \boldsymbol{x}_2 \in \mathfrak{S}(\mathcal{H})$, where $\boldsymbol{x}_i \in \mathfrak{S}(\mathcal{H}_i)$, the bigrading on $\boldsymbol{x}$ agrees with the bigrading on $\boldsymbol{x}_1 \boxtimes \boldsymbol{x}_2$.

11.5 Absolute gradings

We finish this section by turning the relative grading into an absolute one.

First, for any type of diagram, it is straightforward to verify that the homotopy equivalences from Theorems 10.2, 10.4, and 10.7 preserve the relative bigrading.

Next, recall that under the correspondence between bordered grid diagrams and shadows, bordered grid diagrams inherit the bigrading defined in Section 3.4. A plumbing G of a sequence of grid diagrams can be completed to a multipointed bordered Heegaard diagram $\mathcal{H}_G$ in a natural way, by embedding it on a smooth surface, as in Figure 2, and adding the appropriate $\boldsymbol{z}$ decoration in the region(s) outside the image of the embedding. Under the natural correspondence of generators and maps, the resulting diagram $\mathcal{H}_G$ has an associated type A, D or DA structure, which we simply call $\widetilde{CFT}(\mathcal{H}_G)$, identical to $\widetilde{CT}(G)$. The bigrading that $\widetilde{CFT}(\mathcal{H}_G)$ inherits from $\widetilde{CT}(G)$ agrees with the relative bigrading on $\widetilde{CFT}(\mathcal{H}_G)$ defined in this section. We turn the bigrading from this section into an absolute one by requiring that it agrees with the one on $\widetilde{CT}$ for a chosen plumbing of grid diagrams.

Definition 11.11 Given a tangle Heegaard diagram $\mathcal{H}$ of any type for a tangle $\mathcal{T}$, let $\mathcal{H}_G$ be a Heegaard diagram of the same type arising from a plumbing G of grid diagrams representing $\mathcal{T}$, and let $h\colon \widetilde{CFT}(\mathcal{H}_G) \to \widetilde{CFT}(\mathcal{H})$ be the homotopy equivalence corresponding to a chosen sequence of Heegaard moves between $\mathcal{H}_G$ and $\mathcal{H}$. Define the absolute bigrading on $\mathcal{H}$ by requiring that h preserves gradings.

We need to show that the absolute grading in Definition 11.11 is independent of the choice of grid decomposition G, and also independent of the choice of sequence of Heegaard moves, ie of h.

Lemma 11.12 *Fix $\mathcal{H}$ and $\mathcal{H}_G$ as in Definition 11.11, let s and s' be two sequences of Heegaard moves from $\mathcal{H}_G$ to $\mathcal{H}$, and let h', $h''\colon \widetilde{CFT}(\mathcal{H}_G) \to \widetilde{CFT}(\mathcal{H})$ be the homotopy equivalences corresponding to s' and s''. The two bigradings gr' and gr'' induced by h' and h'' according to Definition 11.11 agree.*

Proof We simplify notation and denote the bigrading (M, A) from Section 3.4 by gr.

We will complete $\mathcal{H}_G$ to a closed Heegaard diagram $\overline{\mathcal{H}}_G$ for a link, by gluing to it one (if $\mathcal{H}$ is of type 1 or 2) or two (if $\mathcal{H}$ had two boundary components) plumbings of bordered grid diagrams. The proof in each case is analogous, so from here on we assume that $\mathcal{H}$ is a type 1 diagram. Let H be some plumbing of grids so that $G \cup H$ represents a closed knot or link. Let $\mathcal{H}_H$ be the type 2 Heegaard diagram corresponding to H, and let $\overline{\mathcal{H}}_G = \mathcal{H}_G \cup \mathcal{H}_H$.

Complete each diagram obtained along the sequences s' and s'' to a closed one by gluing to it $\mathcal{H}_H$. The sequences of moves s' and s'' extend to sequences of moves $\overline{s}'$ and $\overline{s}''$ connecting $\overline{\mathcal{H}}_G$ to $\overline{\mathcal{H}} := \mathcal{H} \cup \mathcal{H}_H$, by fixing the $\mathcal{H}_H$ area of each closed diagram and performing the moves specified by s' and s'' outside the $\mathcal{H}_H$ area. Observe that the resulting homotopy equivalences $\overline{h}', \overline{h}''\colon \widetilde{CFT}(\overline{\mathcal{H}}_G) \to \widetilde{CFT}(\overline{\mathcal{H}})$ are exactly the maps $h' \boxtimes \mathrm{id}_{\mathcal{H}_H}, h'' \boxtimes \mathrm{id}_{\mathcal{H}_H}\colon \widetilde{CFT}(\mathcal{H}_G) \boxtimes \widetilde{CFT}(\mathcal{H}_H) \to \widetilde{CFT}(\mathcal{H}) \boxtimes \widetilde{CFT}(\mathcal{H}_H)$. So the gradings induced by $\overline{h}'$ and $\overline{h}''$ are exactly the gradings $\mathrm{gr}' \boxtimes \mathrm{gr}$ and $\mathrm{gr}'' \boxtimes \mathrm{gr}$.

By Theorem 6.1, the grading on $\widetilde{CFT}(\overline{\mathcal{H}}_G) \cong \widetilde{CT}(G \cup H)$ from Section 3.4, which is given by $\mathrm{gr}(\boldsymbol{x}_G \cup \boldsymbol{x}_H) = \mathrm{gr}(\boldsymbol{x}_G) + \mathrm{gr}(\boldsymbol{x}_H)$, agrees with the grading on $\widetilde{CFK}(\overline{\mathcal{H}}_G)$. Since $\overline{h}'$ and $\overline{h}''$ are homotopy equivalences arising from sequences of Heegaard moves, it follows that the gradings they induce on $\widetilde{CFT}(\overline{\mathcal{H}})$ agree with the grading on $\widetilde{CFK}(\overline{\mathcal{H}})$ too. In particular, $\mathrm{gr}' \boxtimes \mathrm{gr} = \mathrm{gr}'' \boxtimes \mathrm{gr}$, so $\mathrm{gr}' = \mathrm{gr}''$. □

Lemma 11.13 *Let $\mathcal{H}$ be a Heegaard diagram for a tangle $\mathcal{T}$. Let $\boldsymbol{\mathcal{P}} = \{\mathcal{P}_1^\circ, \dots, \mathcal{P}_p^\circ\}$ and $\boldsymbol{\mathcal{Q}} = \{\mathcal{Q}_1^\circ, \dots, \mathcal{Q}_q^\circ\}$ be two sequences of shadows for $\mathcal{T}$, let G' and G'' be the corresponding plumbings of bordered grid diagrams, and let $h'\colon \widetilde{CFT}(\mathcal{H}_{G'}) \to \widetilde{CFT}(\mathcal{H})$ and $h''\colon \widetilde{CFT}(\mathcal{H}_{G''}) \to \widetilde{CFT}(\mathcal{H})$ be the homotopy equivalences corresponding to some two sequences of Heegaard moves s' and s'' from $\mathcal{H}_{G'}$ and $\mathcal{H}_{G''}$, respectively, to $\mathcal{H}$. The two bigradings gr' and gr'' induced by h' and h'' according to Definition 11.11 agree.*

Proof Assume $\mathcal{H}$ is a type 1 diagram. The other cases are analogous.

Again denote the bigrading (M, A) from Section 3.4 by gr.

Fix a plumbing H of bordered grid diagrams, as in the proof of Lemma 11.12, so that $G' \cup H$ and $G'' \cup H$ represent a closed knot or link. Let $\overline{\mathcal{H}}_{G'} = \mathcal{H}_{G'} \cup \mathcal{H}_H$, $\overline{\mathcal{H}}_{G''} = \mathcal{H}_{G''} \cup \mathcal{H}_H$, $\overline{\mathcal{H}} = \mathcal{H} \cup \mathcal{H}_H$.

We now apply the same reasoning as in the proof of Lemma 11.12. We get homotopy equivalences $\bar{h}'\colon \widetilde{CFT}(\overline{\mathcal{H}}_{G'}) \to \widetilde{CFT}(\overline{\mathcal{H}})$ and $\bar{h}''\colon \widetilde{CFT}(\overline{\mathcal{H}}_{G''}) \to \widetilde{CFT}(\overline{\mathcal{H}})$. By Theorem 6.1, the grading on $\widetilde{CFT}(\overline{\mathcal{H}}_{G'}) \cong \widetilde{CT}(G' \cup H)$ from Section 3.4 agrees with the grading on $\widetilde{CFK}(\overline{\mathcal{H}}_{G'})$, so the grading $\mathrm{gr}' \boxtimes \mathrm{gr}$ induced by $\bar{h}'$ on $\widetilde{CFT}(\overline{\mathcal{H}})$ agrees with the grading on $\widetilde{CFK}(\overline{\mathcal{H}})$ too. Similarly, the grading $\mathrm{gr}'' \boxtimes \mathrm{gr}$ induced by $\bar{h}''$ on $\widetilde{CFT}(\overline{\mathcal{H}})$ agrees with the grading on $\widetilde{CFK}(\overline{\mathcal{H}})$. Thus, $\mathrm{gr}' \boxtimes \mathrm{gr} = \mathrm{gr}'' \boxtimes \mathrm{gr}$, so $\mathrm{gr}' = \mathrm{gr}''$. □

Proposition 11.14 *The bigrading from Definition 11.11 is well-defined.*

Proof Lemmas 11.12 and 11.13 show that Definition 11.11 is independent of the choices made. This completes the proof. □

We can now conclude that for tangles in B^3 or $S^2 \times I$, the homotopy equivalences from Theorems 10.2, 10.4 and 10.7 are graded. In other words, $\widetilde{CFAT}$, $\widetilde{CFTD}$ and $\widetilde{CFDTA}$ are graded tangle invariants. Below, $V = \mathbb{F}_2 \oplus \mathbb{F}_2$, with one summand in grading $(0, 0)$ and the other summand in grading $(-1, -1)$.

Theorem 11.15 *Up to graded homotopy equivalence and tensoring with V, the modules defined in Section 10 are independent of the choices made in their definitions. Namely:*

If $\mathcal{H}_1$ and $\mathcal{H}_2$ are provincially admissible type 2 diagrams for a $2n$–tangle $\mathcal{T}$ in B^3 with almost complex structures J_1 and J_2, and $|\mathbb{X}_1| = |\mathbb{X}_2| + k$, then there is a graded type D homotopy equivalence

$$\widetilde{CFTD}(\mathcal{H}_1, J_1) \simeq \widetilde{CFTD}(\mathcal{H}_2, J_2) \otimes V^{\otimes k}.$$

If $\mathcal{H}_1$ and $\mathcal{H}_2$ are provincially admissible type 1 diagrams for a $2n$–tangle $\mathcal{T}$ in B^3 with almost complex structures J_1 and J_2, and $|\mathbb{X}_1| = |\mathbb{X}_2| + k$, then there is a graded type A homotopy equivalence

$$\widetilde{CFAT}(\mathcal{H}_1, J_1) \simeq \widetilde{CFAT}(\mathcal{H}_2, J_2) \otimes V^{\otimes k}.$$

If $\mathcal{H}_1$ and $\mathcal{H}_2$ are provincially admissible diagrams for an (m,n)–tangle $\mathcal{T}$ in $S^2 \times I$ with almost complex structures J_1 and J_2, and $|\mathbb{X}_1| = |\mathbb{X}_2| + k$, then there is a graded type DA homotopy equivalence

$$\widetilde{CFDTA}(\mathcal{H}_1, J_1) \simeq \widetilde{CFDTA}(\mathcal{H}_2, J_2) \otimes V^{\otimes k}.$$

Thus, given a marked $2n$–tangle $\mathcal{T}$ in B^3, if $\mathcal{H}$ is a type 1 or a type 2 diagram for $\mathcal{T}$ with $|\mathbb{X} \cap \operatorname{Int} \Sigma| = |\mathcal{T}|$, we get an invariant of the tangle

$$\widehat{CFAT}(\mathcal{T}) := \widetilde{CFAT}(\mathcal{H})$$

up to type A homotopy equivalence, or

$$\widehat{CFTD}(\mathcal{T}) := \widetilde{CFTD}(\mathcal{H})$$

up to type D homotopy equivalence, respectively.

Similarly, given an (m,n)–tangle $\mathcal{T}$ in $S^2 \times I$, if $\mathcal{H}$ is a diagram with two boundary components for $\mathcal{T}$, we get an invariant of the tangle

$$\widehat{CFDTA}(\mathcal{T}) := \widetilde{CFDTA}(\mathcal{H})$$

up to type DA homotopy equivalence.

Similar results hold for the various other modules from Section 10.4.

12 Pairing (nice diagrams)

Sarkar and Wang [25] introduced a class of Heegaard diagrams for 3–manifolds called *nice*. These were used in [7] to prove a pairing theorem in bordered Floer homology. In a similar vein, here we define nice Heegaard diagrams for tangles, and use them to prove a pairing theorem.

Definition 12.1 A tangle Heegaard diagram is called *nice* if every region that does not contain an interior X or O and does not intersect $\boldsymbol{z}$ is a disk with at most 4 corners.

Proposition 12.2 *Any tangle Heegaard diagram can be turned into a nice diagram via a sequence of isotopies and handleslides of the β–curves in the interior of the Heegaard surface.*

Proof The proof uses "finger moves" and is analogous to the proof of [8, Proposition 8.2]. □

Lemma 12.3 *If $\mathcal{H}$ is nice, then $\mathcal{H}$ is admissible.*

Proof The proof is a straightforward generalization of the one for the closed case [5, Corollary 3.2]. Suppose D is a nontrivial domain in $\Sigma \setminus (\mathbb{X} \cup \mathbb{O} \cup \boldsymbol{z})$ with only nonnegative multiplicities, and its boundary is a linear combination of entire α– and β–curves. Consider a curve that appears in ∂D with nonzero multiplicity, and orient it so that all regions directly to its left have positive multiplicity. If that curve is an α–circle or a β–circle, then [5, Lemma 3.1] applies, ie one of these regions contains a basepoint, which gives a contradiction. So suppose that curve is an α–arc, call it α_i. We verify that the argument in [5, Lemma 3.1] can be used again to show that one of these regions contains a basepoint.

Suppose one of the regions directly to the left of α_i is a bigon. Then the other edge of that region is part of a β–circle, call it β_j. On the other side of β_j there is a square (a bigon would imply α_i is a circle, not an arc) and the edge of that square across from β_j is either part of a β–circle again, or part of $\partial\Sigma$. In the first case, there is yet another square on the other side, and we look at that square. Eventually we reach a square with an edge on $\partial\Sigma$. The union of all these regions forms a component of $\Sigma \setminus \boldsymbol{\alpha}$ (with two corners), so we reach a contradiction, since every component of $\Sigma \setminus \boldsymbol{\alpha}$ contains a point in $\mathbb{X} \cup \mathbb{O} \cup \boldsymbol{z}$.

Now suppose there are no bigon regions directly to the left of α_i. Then all those regions are squares, and they must form a chain that starts and ends at $\partial\Sigma$. The edges across from α_i on those squares form a complete α–arc, and the union of the squares is a component of $\Sigma \setminus \boldsymbol{\alpha}$ (with four corners). This again is a contradiction. □

Since nice diagrams are admissible, there are only a few types of holomorphic curves, as one only counts domains that are squares or bigons. Specifically, for $\widehat{CFAT}$, all multiplication maps m_n for $n > 2$ are zero, and for $\widehat{CFDTA}$ all structure maps δ^1_{1+j} for $j > 1$ are zero.

We are now ready to state and prove a pairing theorem. By invariance (Theorem 11.15), assume that all diagrams below are nice.

Theorem 12.4 *The following equivalences hold:*

(1) *If $\mathcal{H}_1 \cup \mathcal{H}_2$ is the union of a type 1 Heegaard diagram $\mathcal{H}_1$ and a Heegaard diagram with two boundary components $\mathcal{H}_2$ along $\partial\mathcal{H}_1$ and $-\partial^0\mathcal{H}_2$, then*
$$\widetilde{CFAT}(\mathcal{H}_1) \boxtimes \widetilde{CFDTA}(\mathcal{H}_2) \simeq \widetilde{CFAT}(\mathcal{H}_1 \cup \mathcal{H}_2).$$

(2) *If $\mathcal{H}_1 \cup \mathcal{H}_2$ is the union of Heegaard diagrams $\mathcal{H}_1$ and $\mathcal{H}_2$ with two boundary components along $\partial^1\mathcal{H}_1$ and $-\partial^0\mathcal{H}_2$, then*
$$\widetilde{CFDTA}(\mathcal{H}_1) \boxtimes \widetilde{CFDTA}(\mathcal{H}_2) \simeq \widetilde{CFDTA}(\mathcal{H}_1 \cup \mathcal{H}_2).$$

(3) *If $\mathcal{H}_1 \cup \mathcal{H}_2$ is the union of a Heegaard diagram $\mathcal{H}_1$ with two boundary components and a Heegaard diagram $\mathcal{H}_2$ of type 2 along $\partial^1\mathcal{H}_1$ and $-\partial\mathcal{H}_2$, then*
$$\widetilde{CFDTA}(\mathcal{H}_1) \boxtimes \widetilde{CFTD}(\mathcal{H}_2) \simeq \widetilde{CFTD}(\mathcal{H}_1 \cup \mathcal{H}_2).$$

(4) *If $\mathcal{H}_1 \cup \mathcal{H}_2$ is the union of a Heegaard diagram $\mathcal{H}_1$ of type 1 and a Heegaard diagram $\mathcal{H}_2$ of type 2 along $\partial\mathcal{H}_1$ and $-\partial\mathcal{H}_2$, then*
$$\widetilde{CFAT}(\mathcal{H}_1) \boxtimes \widetilde{CFTD}(\mathcal{H}_2) \simeq \widetilde{CFK}(\mathcal{H}_1 \cup \mathcal{H}_2).$$

Moreover, when the underlying manifolds are B^3, $S^2 \times I$ or S^3, the homotopy equivalences respect the bigrading.

Proof The proof is analogous to that for bordered Heegaard Floer homology [9, Theorem 11]. We outline it for the first case. First note that $\mathcal{H}_1 \cup \mathcal{H}_2$ is automatically a type 1 Heegaard diagram. Since $\mathcal{H}_1$ and $\mathcal{H}_2$ are nice diagrams, then both diagrams are admissible, so the corresponding type A and type DA structures are bounded, and their box tensor product is well-defined. There is a correspondence between generators of $\widetilde{CFAT}(\mathcal{H}_1) \boxtimes \widetilde{CFDTA}(\mathcal{H}_2)$ and $\widetilde{CFAT}(\mathcal{H}_1 \cup \mathcal{H}_2)$.

The differential on $\widetilde{CFAT}(\mathcal{H}_1) \boxtimes \widetilde{CFDTA}(\mathcal{H}_2)$ counts bigons and rectangles that are provincial in $\mathcal{H}_1$ (corresponding to the differential m_1 on $\widetilde{CFAT}(\mathcal{H}_1)$), provincial in $\mathcal{H}_2$ (corresponding to the "differential" on $\widetilde{CFDTA}(\mathcal{H}_2)$, ie the part of δ^1_1 that outputs an idempotent algebra element), or provincial in $\mathcal{H}_1 \cup \mathcal{H}_2$ but crossing the common boundary of $\mathcal{H}_1$ and $\mathcal{H}_2$ (for $(m_2 \otimes \mathrm{id}) \circ (\mathrm{id} \otimes \delta^1_1)$ when δ^1_1 outputs a nonidempotent algebra element). The third kind can only be a rectangle. These are exactly all the provincial domains for $\widetilde{CFAT}(\mathcal{H}_1 \cup \mathcal{H}_2)$. So the differentials on $\widetilde{CFAT}(\mathcal{H}_1) \boxtimes \widetilde{CFDTA}(\mathcal{H}_2)$ and $\widetilde{CFAT}(\mathcal{H}_1 \cup \mathcal{H}_2)$ agree.

Half-rectangles on $\mathcal{H}_1 \cup \mathcal{H}_2$ that cross $\partial^1\mathcal{H}_2$ are entirely contained (left provincial) in $\mathcal{H}_2$, and the same sets of these half-rectangles are counted for the right multiplications m_2 on $\widetilde{CFAT}(\mathcal{H}_1) \boxtimes \widetilde{CFDTA}(\mathcal{H}_2)$ and on $\widetilde{CFAT}(\mathcal{H}_1 \cup \mathcal{H}_2)$.

Thus, the type A structures $\widetilde{CFAT}(\mathcal{H}_1) \boxtimes \widetilde{CFDTA}(\mathcal{H}_2)$ and $\widetilde{CFAT}(\mathcal{H}_1 \cup \mathcal{H}_2)$ are isomorphic.

The other cases are analogous. □

In particular, tangle Floer homology recovers knot Floer homology. For tangles in B^3 and $S^2 \times I$, this result is simply a restatement of Theorem 6.1.

If $\mathcal{H}_1$ or $\mathcal{H}_2$ is not a nice diagram, the corresponding structure may not be bounded. In that case, the box tensor product is not defined, and we need to look at the $\mathcal{A}_\infty$ tensor product $\widetilde{CFT}(\mathcal{H}_1) \widetilde{\otimes} \big(\mathcal{A}(-\partial^0\mathcal{H}_2) \boxtimes \widetilde{CFT}(\mathcal{H}_2)\big)$. So by [9, Proposition 2.3.18], invariance, and the above theorem, $\widetilde{CFT}(\mathcal{H}_1) \widetilde{\otimes} \big(\mathcal{A}(-\partial^0\mathcal{H}_2) \boxtimes \widetilde{CFT}(\mathcal{H}_2)\big) \simeq \widetilde{CFT}(\mathcal{H}_1 \cup \mathcal{H}_2)$, or using the shorter notation, $\widetilde{CFT}(\mathcal{H}_1) \widetilde{\otimes} \widetilde{CFT}(\mathcal{H}_2) \simeq \widetilde{CFT}(\mathcal{H}_1 \cup \mathcal{H}_2)$. Here $\widetilde{CFT}$ stands for any of the structures in Theorem 12.4.

Corollary 12.5 *The following equivalences hold:*

(1) *If $(Y_1, \mathcal{T}_1)$ is a $2m$–tangle where ∂Y_1 is identified with a marked sphere $\mathcal{S}$, $(Y_2, \mathcal{T}_2, \gamma)$ is a strongly marked $(2m, 2n)$–tangle with $\partial^0 Y_2$ identified with $-\mathcal{S}$, and $(Y, \mathcal{T})$ is their union along $\mathcal{S}$, then*

$$\widehat{CFAT}(Y_1, \mathcal{T}_1) \boxtimes \widehat{CFDTA}(Y_2, \mathcal{T}_2, \gamma) \simeq \widehat{CFAT}(Y, \mathcal{T}) \otimes V^{\otimes(|\mathcal{T}_1|+|\mathcal{T}_2|-|\mathcal{T}|)}.$$

(2) *If $(Y_1, \mathcal{T}_1, \gamma_1)$ is a strongly marked (m, n)–tangle with $\partial^1 Y_1$ identified with a marked sphere $\mathcal{S}$, $(Y_2, \mathcal{T}_2, \gamma_2)$ is a strongly marked (n, l)–tangle with $\partial^0 Y_2$ identified with $-\mathcal{S}$, and $(Y, \mathcal{T}, \gamma)$ is their union along $\mathcal{S}$, then*

$$\widehat{CFDTA}(Y_1, \mathcal{T}_1, \gamma_1) \boxtimes \widehat{CFDTA}(Y_2, \mathcal{T}_2, \gamma_2) \simeq \widehat{CFDTA}(Y, \mathcal{T}, \gamma) \otimes V^{\otimes(|\mathcal{T}_1|+|\mathcal{T}_2|-|\mathcal{T}|)}.$$

(3) *If $(Y_1, \mathcal{T}_1, \gamma)$ is a strongly marked $(2m, 2n)$–tangle with $\partial^1 Y_1$ identified with a marked sphere $\mathcal{S}$, $(Y_2, \mathcal{T}_2)$ is a $2n$–tangle with ∂Y_2 identified with $-\mathcal{S}$, and $(Y, \mathcal{T})$ is their union along $\mathcal{S}$, then*

$$\widehat{CFDTA}(Y_1, \mathcal{T}_1, \gamma) \boxtimes \widehat{CFTD}(Y_2, \mathcal{T}_2) \simeq \widehat{CFTD}(Y, \mathcal{T}) \otimes V^{\otimes(|\mathcal{T}_1|+|\mathcal{T}_2|-|\mathcal{T}|)}.$$

(4) *If $(Y_1, \mathcal{T}_1)$ is a $2n$–tangle with ∂Y_1 identified with a marked sphere $\mathcal{S}$, $(Y_2, \mathcal{T}_2)$ is a $2n$–tangle with ∂Y_2 identified with $-\mathcal{S}$, and $(Y, \mathcal{T})$ is their union along $\mathcal{S}$, then*

$$\widehat{CFAT}(Y_1, \mathcal{T}_1) \boxtimes \widehat{CFTD}(Y_2, \mathcal{T}_2) \simeq \widehat{CFK}(Y, \mathcal{T}) \otimes V^{\otimes(|\mathcal{T}_1|+|\mathcal{T}_2|-|\mathcal{T}|)} \otimes W,$$

where $W = \mathbb{F}_2 \oplus \mathbb{F}_2$.

Moreover, when the underlying manifolds are B^3, $S^2 \times I$ or S^3, the homotopy equivalences respect the bigrading, where the two summands of V are in (M, A) bigradings $(0,0)$ and $(-1,-1)$, and the two summands of W are in bigradings $(0,0)$ and $(-1,0)$.

Proof In each case, for a choice of nice Heegaard diagrams, we have an equivalence of tilde modules as in the proof of Theorem 12.4. To have precisely the "hat" modules for $\mathcal{T}_1$ and $\mathcal{T}_2$, pick nice Heegaard diagrams $\mathcal{H}_i$ with $|\mathbb{X}_i \cap \operatorname{Int} \Sigma_i| = |\mathcal{T}_i|$. Note that on $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$ we have $|\mathbb{X} \cap \operatorname{Int}(\Sigma_1 \cup \Sigma_2)| = |\mathcal{T}_1| + |\mathcal{T}_2|$, and we need a diagram such that $|\mathbb{X} \cap \operatorname{Int} \Sigma| = |\mathcal{T}|$ to obtain the "hat" module for $\mathcal{T}$, so $\mathcal{H}$ produces a module equivalent to the "hat" module tensored with $|\mathcal{T}_1| + |\mathcal{T}_2| - |\mathcal{T}|$ copies of V.

Note that in the fourth case $\mathcal{H}_1 \cup \mathcal{H}_2$ is a Heegaard diagram for the link $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$ union a split unknot U in Y (see Lemma 8.24), so

$$\begin{aligned}\widetilde{CFK}(\mathcal{H}_1 \cup \mathcal{H}_2) &\simeq \widetilde{CFK}(Y, \mathcal{T} \cup U) \otimes V^{\otimes(|\mathcal{T}_1|+|\mathcal{T}_2|-|\mathcal{T}|)} \\ &\simeq \widehat{CFK}(Y, \mathcal{T} \cup U) \otimes V^{\otimes(|\mathcal{T}_1|+|\mathcal{T}_2|-|\mathcal{T}|)} \otimes W.\end{aligned}$$

The second equivalence is a known fact in Heegaard Floer theory. □

Similar results hold for the various other modules from Section 10.4.

References

[1] **F Bourgeois**, **Y Eliashberg**, **H Hofer**, **K Wysocki**, **E Zehnder**, *Compactness results in symplectic field theory*, Geom. Topol. 7 (2003) 799–888 MR2026549

[2] **P Ghiggini**, *Knot Floer homology detects genus-one fibred knots*, Amer. J. Math. 130 (2008) 1151–1169 MR2450204

[3] **M Khovanov**, *A categorification of the Jones polynomial*, Duke Math. J. 101 (2000) 359–426 MR1740682

[4] **M Khovanov**, *A functor-valued invariant of tangles*, Algebr. Geom. Topol. 2 (2002) 665–741 MR1928174

[5] **R Lipshitz**, **C Manolescu**, **J Wang**, *Combinatorial cobordism maps in hat Heegaard Floer theory*, Duke Math. J. 145 (2008) 207–247 MR2449946

[6] **R Lipshitz**, **P Ozsváth**, **D Thurston**, *Heegaard Floer homology as morphism spaces, II*, in preparation

[7] **R Lipshitz**, **P Ozsváth**, **D Thurston**, *Slicing planar grid diagrams: a gentle introduction to bordered Heegaard Floer homology*, from "Proceedings of Gökova geometry/topology conference 2008" (S Akbulut, T Önder, R J Stern, editors), GGT, Gökova, Turkey (2009) 91–119 MR2500575

[8] **R Lipshitz**, **P Ozsváth**, **D Thurston**, *Bordered Heegaard Floer homology: invariance and pairing*, preprint (2011) `arXiv:0810.0687v4`

[9] **R Lipshitz**, **P S Ozsváth**, **D P Thurston**, *Bimodules in bordered Heegaard Floer homology*, Geom. Topol. 19 (2015) 525–724 MR3336273

[10] **C Manolescu**, *An unoriented skein exact triangle for knot Floer homology*, Math. Res. Lett. 14 (2007) 839–852 MR2350128

[11] **C Manolescu**, **P Ozsváth**, **S Sarkar**, *A combinatorial description of knot Floer homology*, Ann. of Math. 169 (2009) 633–660 MR2480614

[12] **C Manolescu**, **P Ozsváth**, **Z Szabó**, **D Thurston**, *On combinatorial link Floer homology*, Geom. Topol. 11 (2007) 2339–2412 MR2372850

[13] **Y Ni**, *Knot Floer homology detects fibred knots*, Invent. Math. 170 (2007) 577–608 MR2357503

[14] **P Ozsváth**, **Z Szabó**, *Knot Floer homology and the four-ball genus*, Geom. Topol. 7 (2003) 615–639 MR2026543

[15] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and genus bounds*, Geom. Topol. 8 (2004) 311–334 MR2023281

[16] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and knot invariants*, Adv. Math. 186 (2004) 58–116 MR2065507

[17] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and three-manifold invariants: properties and applications*, Ann. of Math. 159 (2004) 1159–1245 MR2113020

[18] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. 159 (2004) 1027–1158 MR2113019

[19] **P Ozsváth**, **Z Szabó**, *Holomorphic triangles and invariants for smooth four-manifolds*, Adv. Math. 202 (2006) 326–400 MR2222356

[20] **P Ozsváth**, **Z Szabó**, *On the skein exact squence for knot Floer homology*, preprint (2007) `arXiv:0707.1165`

[21] **P Ozsváth**, **Z Szabó**, *Holomorphic disks, link invariants and the multi-variable Alexander polynomial*, Algebr. Geom. Topol. 8 (2008) 615–692 MR2443092

[22] **I Petkova**, *An absolute $\mathbb{Z}/2$ grading on bordered Heegaard Floer homology*, preprint (2014) `arXiv:1401.2670`

[23] **J A Rasmussen**, *Floer homology and knot complements*, PhD thesis, Harvard University (2003) MR2704683 Available at `http://search.proquest.com/docview/305332635`

[24] **J Rasmussen**, *Knot polynomials and knot homologies*, from "Geometry and topology of manifolds" (H U Boden, I Hambleton, A J Nicas, B D Park, editors), Fields Inst. Commun. 47, Amer. Math. Soc., Providence, RI (2005) 261–280 MR2189938

[25] **S Sarkar**, **J Wang**, *An algorithm for computing some Heegaard Floer homologies*, Ann. of Math. 171 (2010) 1213–1236 MR2630063

[26] **A Sartori**, *The Alexander polynomial as quantum invariant of links*, Ark. Mat. 53 (2015) 177–202 MR3319619

[27] **O Y Viro**, *Quantum relatives of the Alexander polynomial*, Algebra i Analiz 18 (2006) 63–157 MR2255851 In Russian; translated in St. Petersburg Math. J. 18 (2007) 391–457

[28] **R Zarev**, *Bordered Floer homology for sutured manifolds*, preprint (2009) arXiv:0908.1106

Department of Mathematics, Dartmouth College
Hanover, NH 03755, United States

Institut de Recherche Mathématique Avancée, Université de Strasbourg
7 rue René Decartes, 67087 Strasbourg, France

ina.petkova@dartmouth.edu, vertesi@math.unistra.fr

http://www.math.dartmouth.edu/~ina/,
http://www-irma.u-strasbg.fr/~vertesi/

Proposed: Ciprian Manolescu
Seconded: Ronald Stern, Ian Agol

Received: 20 November 2014
Revised: 19 October 2015

msp

Geometry & Topology 20 (2016) 3333–3430

Persistent homology and Floer–Novikov theory

MICHAEL USHER
JUN ZHANG

We construct “barcodes” for the chain complexes over Novikov rings that arise in Novikov’s Morse theory for closed one-forms and in Floer theory on not-necessarily-monotone symplectic manifolds. In the case of classical Morse theory these coincide with the barcodes familiar from persistent homology. Our barcodes completely characterize the filtered chain homotopy type of the chain complex; in particular they subsume in a natural way previous filtered Floer-theoretic invariants such as boundary depth and torsion exponents, and also reflect information about spectral invariants. Moreover, we prove a continuity result which is a natural analogue both of the classical bottleneck stability theorem in persistent homology and of standard continuity results for spectral invariants, and we use this to prove a C^0–robustness result for the fixed points of Hamiltonian diffeomorphisms. Our approach, which is rather different from the standard methods of persistent homology, is based on a nonarchimedean singular value decomposition for the boundary operator of the chain complex.

53D40; 55U15

1 Introduction

Persistent homology is a well-established tool in the rapidly developing field of topological data analysis. On an algebraic level, the subject studies “persistence modules”, ie structures $\mathbb{V}$ consisting of a module V_t associated to each $t \in \mathbb{R}$ with homomorphisms $\sigma_{st}\colon V_s \to V_t$ whenever $s \le t$ satisfying the functoriality properties that $\sigma_{ss} = I_{V_s}$, the identity map on module V_s, and $\sigma_{su} = \sigma_{tu} \circ \sigma_{st}$ (more generally $\mathbb{R}$ could be replaced by an arbitrary partially ordered set, but this generalization will not be relevant to this paper). Persistence modules arise naturally in topology when one considers a continuous function $f\colon X \to \mathbb{R}$ on a topological space X; for a field $\mathcal{K}$ one can then let $V_t = H_*(\{f \le t\}; \mathcal{K})$ be the homology of the t–sublevel set, with the σ_{st} being the inclusion-induced maps. For example if $X = \mathbb{R}^n$ and the function $f\colon \mathbb{R}^n \to \mathbb{R}$ is given by the minimal distance to a finite collection of points sampled from some subset $S \subset \mathbb{R}^n$, then V_t is the homology of the union of balls of radius t around the points of the sample; the structure of the associated persistence module has been used

Published: 21 December 2016 DOI: 10.2140/gt.2016.20.3333

effectively to make inferences about the topological structure of the set S in some real-world situations; see eg Carlsson [7].

Under finiteness hypotheses on the modules V_t (for instance finite-type as in Zomorodian and Carlsson [46] or more generally pointwise finite-dimensionality as in Crawley-Boevey [12]), provided that the coefficient ring for the modules V_t is a field $\mathcal{K}$, it can be shown that the persistence module $\mathbb{V}$ is isomorphic in the obvious sense to a direct sum of "interval modules" $\mathcal{K}_I$, where $I \subset \mathbb{R}$ is an interval and by definition $(\mathcal{K}_I)_t = \mathcal{K}$ for $t \in I$ and $\{0\}$ otherwise, and the morphisms σ_{st} are the identity on $\mathcal{K}$ when $s, t \in I$ and 0 otherwise. The *barcode* of $\mathbb{V}$ is then defined to be the multiset of intervals appearing in this direct sum decomposition. When $\mathbb{V}$ is obtained as the filtered homology of a finite-dimensional chain complex, [46] gives a worst-case-cubic-time algorithm that computes the barcode given the boundary operator on the chain complex.

If $f\colon X \to \mathbb{R}$ is a Morse function on a compact smooth manifold, a standard construction (see eg Schwarz [39]) yields a "Morse chain complex" $(\mathrm{CM}_*(f), \partial)$. The degree-$k$ part $\mathrm{CM}_k(f)$ of the complex is formally spanned (say over the field $\mathcal{K}$) by the critical points of f having index k. The boundary operator $\partial\colon \mathrm{CM}_{k+1}(f) \to \mathrm{CM}_k(f)$ counts (with appropriate signs) negative gradient flowlines of f which are asymptotic as $t \to -\infty$ to an index-$(k+1)$ critical point and as $t \to \infty$ to an index-k critical point. For any $t \in \mathbb{R}$, if we consider the subspace $\mathrm{CM}_*^t(f) \le \mathrm{CM}_*(f)$ spanned only by those critical points p of f with $f(p) \le t$, then the fact that f decreases along its negative gradient flowlines readily implies that $\mathrm{CM}_*^t(f)$ is a subcomplex of $\mathrm{CM}_*(f)$. So taking homology gives filtered Morse homology groups $\mathrm{HM}_*^t(f)$, with inclusion-induced maps $\mathrm{HM}_*^s(f) \to \mathrm{HM}_*^t(f)$ when $s \le t$ that satisfy the usual functoriality properties. Thus the filtered Morse homology groups associated to a Morse function yield a persistence module; given a formula for the Morse boundary operator one could then apply the algorithm from [46] to compute its barcode. In fact, standard results of Morse theory show that this persistence module is (up to isomorphism) simply the persistence module comprising the sublevel homologies $H_*(\{f \le t\}; \mathcal{K})$ with the inclusion-induced maps.

There are a variety of situations in which one can do some form of Morse theory for a suitable function $\mathcal{A}\colon \mathcal{C} \to \mathbb{R}$ on an appropriate infinite-dimensional manifold $\mathcal{C}$. Indeed, Morse himself [32] applied his theory to the energy functional on the loop space of a Riemannian manifold in order to study its geodesics. Floer [15; 16; 17] discovered some rather different manifestations of infinite-dimensional Morse theory involving functions $\mathcal{A}$ which, unlike the energy functional, are unbounded above and below and have critical points of infinite index. In these cases, one still obtains a Floer chain complex analogous to the Morse complex of the previous paragraph and can still speak of the filtered homologies HF^t with their inclusion-induced maps

$\mathrm{HF}^s \to \mathrm{HF}^t$; however it is no longer true that these filtered homology groups relate directly to classical topological invariants; rather, they are new objects. Thus Floer's construction gives (taking filtrations into account as above) a persistence module. If the persistence module satisfies appropriate finiteness conditions one then obtains a barcode by the procedure indicated earlier; however, as we will explain below the finiteness conditions only hold in rather restricted circumstances. While the filtered Floer groups have been studied since the early 1990s and have been a significant tool in symplectic topology since that time (see eg Floer and Hofer [18], Schwarz [40], Entov and Polterovich [13], Oh [34], Usher [45] and Humilière, Leclercq and Seyfaddini [26]), it is only very recently that they have been considered from a persistent-homological point of view. Namely, Polterovich and Shelukhin [36] apply ideas from persistent homology to prove interesting results about autonomous Hamiltonian diffeomorphisms of symplectic manifolds, subject to a topological restriction that is necessary to guarantee the finiteness property that leads to a barcode. This paper will generalize the notion of a barcode to more general Floer-theoretic situations. In particular, this opens up the possibility of extending the results from [36] to manifolds other than those considered therein; this is the subject of work in progress by the second author.

The difficulty with applying the theory of barcodes to general Floer complexes lies in the fact that, typically, Floer theory is more properly viewed as an infinite dimensional version of Novikov's Morse theory for closed one-forms (see Novikov [33] and Farber [14]) rather than of classical Morse theory. Here one considers a closed 1–form α on some manifold M which vanishes transversely with finitely many zeros, and takes a regular covering $\pi\colon \widetilde{M} \to M$ on which we have $\pi^*\alpha = d\tilde{f}$ for some function $\tilde{f}\colon \widetilde{M} \to \mathbb{R}$. Then $\tilde{f}$ will be a Morse function whose critical locus consists of the preimage of the (finite) zero locus of α under π; in particular, if the de Rham cohomology class of α is nontrivial then $\pi\colon \widetilde{M} \to M$ will necessarily have infinite fibers and so $\tilde{f}$ will have infinitely many critical points.

One then attempts to construct a Morse-type complex $\mathrm{CN}_*(\tilde{f})$ by setting $\mathrm{CN}_k(\tilde{f})$ equal to the span over $\mathcal{K}$ of the index-k critical points[1] of $\tilde{f}$, with boundary operator $\partial\colon \mathrm{CN}_{k+1}(\tilde{f}) \to \mathrm{CN}_k(\tilde{f})$ given by setting, for an index-$(k+1)$ critical point p of $\tilde{f}$,

$$\partial p = \sum_{\mathrm{ind}_{\tilde{f}}(q)=k} n(p,q)q,$$

where $n(p,q)$ is a count of negative gradient flowlines for $\tilde{f}$ (with respect to a suitably generic Riemannian metric pulled back to $\widetilde{M}$ from M) asymptotic to p in negative time and to q in positive time. However the above attempt does not quite work because the

[1] "Index" means Morse index in the finite-dimensional case (see eg Schwarz [39]), and typically some version of the Maslov index in the Floer-theoretic case (see eg Robbin and Salamon [37]).

sum on the right-hand side may have infinitely many nonzero terms; thus it is necessary to enlarge $\mathrm{CN}_k(\tilde{f})$ to accommodate certain formal infinite sums. The correct definition is, denoting by $\mathrm{Crit}_k(\tilde{f})$ the set of critical points of $\tilde{f}$ with index k,

$$\mathrm{CN}_k(\tilde{f}) = \left\{ \sum_{p\in\mathrm{Crit}_k(\tilde{f})} a_p p \;\middle|\; a_p \in \mathcal{K} \text{ and } \#\{p \mid a_p \neq 0,\ \tilde{f}(p) > C\} < \infty \text{ for all } C \in \mathbb{R} \right\}. \tag{1}$$

Then under suitable hypotheses it can be shown that the definition of ∂ above gives a well-defined map $\partial\colon \mathrm{CN}_{k+1}(\tilde{f}) \to \mathrm{CN}_k(\tilde{f})$ such that $\partial^2 = 0$. This construction can be carried out in many contexts, including the classical Novikov complex where M is compact and various Floer theories where M is infinite-dimensional. In the latter case, the zeros of α are typically some objects of interest, such as closed orbits of a Hamiltonian flow, on some other finite-dimensional manifold. In these cases, just as in Morse theory, ∂ preserves the $\mathbb{R}$–filtration given by, for $t \in \mathbb{R}$, letting $\mathrm{CN}^t_k(\tilde{f})$ consist of only those formal sums $\sum_p a_p p$ where each $\tilde{f}(p)$ is at most t. In this way we obtain filtered Novikov homology groups $\mathrm{HN}^t_*(\tilde{f})$ with inclusion-induced maps $\mathrm{HN}^s(\tilde{f}) \to \mathrm{HN}^t(\tilde{f})$ satisfying the axioms of a persistence module over $\mathcal{K}$.

However, when the cover $\tilde{M} \to M$ is nontrivial, this persistence module over $\mathcal{K}$ does not satisfy the hypotheses of many of the major theorems of persistent homology; the maps $\mathrm{HN}^s(\tilde{f}) \to \mathrm{HN}^t(\tilde{f})$ generally have infinite rank over $\mathcal{K}$ (due to a certain "lifting" scenario which is described later in this paragraph) and so the persistence module is not "q-tame" in the sense of Chazal, de Silva, Glisse and Oudot [9]. As is well-known, to get a finite-dimensional object out of the Novikov complex one should work not over $\mathcal{K}$ but over a suitable Novikov ring. From now on we will assume that the cover $\pi\colon \tilde{M} \to M$ is minimal subject to the property that $\pi^*\alpha$ is exact; in other words, the covering group coincides with the kernel of the homomorphism $I_\alpha\colon \pi_1(M) \to \mathbb{R}$ induced by integrating α over loops. This will lead to our Novikov ring being a field. Given this assumption, let $\hat{\Gamma} \le \mathbb{R}$ be the image of I_α. Then by, for any $g \in \hat{\Gamma}$, lifting loops in M with integral equal to $-g$ to paths in $\tilde{M}$, we obtain an action of $\hat{\Gamma}$ on the critical locus of $\tilde{f}$ such that $\tilde{f}(p) - \tilde{f}(gp) = g$. In some Floer-theoretic situations this action can shift the index by $s(g)$ for some homomorphism $s\colon \hat{\Gamma} \to \mathbb{Z}$. For instance, in Hamiltonian Floer theory s is given by evaluating twice the first Chern class of the symplectic manifold on spheres, whereas in the classical case of the Novikov chain complex of a closed one-form on a finite-dimensional manifold, s is zero. Now let $\Gamma = \ker s$, so that Γ acts on the index-k critical points of $\tilde{f}$, and this action then gives rise to an action of the following *Novikov field* on $\mathrm{CN}_k(\tilde{f})$:

$$\Lambda^{\mathcal{K},\Gamma} = \left\{ \sum_{g\in\Gamma} a_g T^g \;\middle|\; a_g \in \mathcal{K} \text{ and } \#\{g \mid a_g \neq 0,\ g < C\} < \infty \text{ for all } C \in \mathbb{R} \right\}.$$

It follows from the description that $\mathrm{CN}_k(\tilde{f})$ is a vector space over $\Lambda^{\mathcal{K},\Gamma}$ of (finite!) dimension equal to the number of zeros of our original $\alpha \in \Omega^1(M)$ that admit lifts to index-k critical points for $\tilde{f}$ in $\tilde{M}$. Indeed, if the set $\{\tilde{p}_1, \dots, \tilde{p}_{m_i}\} \subset \tilde{M}$ consists of exactly one such lift of each of these zeros of α then $\{\tilde{p}_1, \dots, \tilde{p}_{m_i}\}$ is a $\Lambda^{\mathcal{K},\Gamma}$–basis for $\mathrm{CN}_k(\tilde{f})$.

Now since the action by an element g of Γ shifts the value of $\tilde{f}$ by $-g$, the filtered groups $\mathrm{CN}_k^t(\tilde{f})$ are not preserved by multiplication by scalars in $\Lambda^{\mathcal{K},\Gamma}$, and so the aforementioned persistence module $\{\mathrm{HF}^t(\tilde{f})\}$ over $\mathcal{K}$ can *not* be viewed as a persistence module over $\Lambda^{\mathcal{K},\Gamma}$, unless of course $\Gamma = \{0\}$, in which case $\Lambda^{\mathcal{K},\Gamma} = \mathcal{K}$. Our strategy in this paper is to understand filtered Novikov and Floer complexes not through their induced persistence modules on homology (see Remark 1.1 below) but rather through the *nonarchimedean geometry* that the filtration induces on the chain complexes. This will lead to an alternative theory of barcodes which recovers the standard theory in the case that $\Gamma = \{0\}$ (see Zomorodian and Carlsson [46], Chazal, de Silva, Glisse and Oudot [9] and, for a different perspective, Barannikov [2]) but which also makes sense for arbitrary Γ, while continuing to enjoy various desirable properties.

We should mention that, in the case of Morse–Novikov theory for a function $f\colon X \to S^1$, a different approach to persistent homology is taken in Burghelea and Dey [5] and Burghelea and Haller [6]. These works are based around the notion of the (zigzag) persistent homology of level sets of the function; this is a rather different viewpoint from ours, as in order to obtain insight into Floer theory we only use the algebraic features of the Floer chain complex, and in a typical Floer theory there is nothing that plays the role of the homology of a level set. Rather, we construct what could be called an algebraic simulation of the more classical sublevel set persistence, even though (as noted in [5]) from a geometric point of view it does not make sense to speak of the sublevel sets of an S^1–valued function. Also our theory, unlike that of [5] and [6], applies to the Novikov complexes of closed one-forms that have dense period groups. Notwithstanding these differences, there are some indications (see in particular the remark after [6, Theorem 1.4]) that the constructions may be related on their common domains of applicability; it would be interesting to understand this further.

1.1 Outline of the paper and summary of main results

With the exception of an application to Hamiltonian Floer theory in Section 12, the entirety of this paper is written in a general algebraic context involving chain complexes of certain kinds of nonarchimedean normed vector spaces over Novikov fields $\Lambda = \Lambda^{\mathcal{K},\Gamma}$. (In particular, no knowledge of Floer theory is required to read the large majority of the paper, though it may be helpful as motivation.) The definitions necessary for our theory are somewhat involved and so will not be included in detail in this introduction, but they make use of the standard notion of orthogonality in nonarchimedean normed vector spaces, a subject which is reviewed in Section 2. Our first key result is Theorem 3.4, which shows that any linear map $A\colon C \to D$ between two finite-dimensional nonarchimedean normed vector spaces C and D over Λ having orthogonal bases admits a *singular value decomposition*: there are orthogonal bases B_C for C and B_D for D such that A maps each member of B_C either to zero or to one of the elements of B_D. In the case that C and D admit ortho*normal* bases and not just orthogonal ones this was known (see Kedlaya [27, Section 4.3]); however, Floer complexes typically admit orthogonal but not orthonormal bases (unless one extends coefficients, which leads to a loss of information), and in this case Theorem 3.4 appears to be new.

In Definition 4.1 we introduce the notion of a "Floer-type complex" (C_*, ∂, ℓ) over a Novikov field Λ; this is a chain complex of Λ–vector spaces (C_*, ∂) with a nonarchimedean norm e^{ℓ} on each graded piece C_k that induces a filtration which is respected by ∂. We later construct our versions of the barcode by consideration of singular value decompositions of the various graded pieces of the boundary operator. Singular value decompositions are rather nonunique, but we prove a variety of results reflecting that data about filtrations of the elements involved in a singular value decomposition is often independent of choices and so gives rise to invariants of the Floer-type complex (C_*, ∂, ℓ). The first instance of this appears in Theorem 4.11, which relates the boundary depth of Usher [44; 45], as well as generalizations thereof, to singular value decompositions. Theorem 4.13 shows that these generalized boundary depths are equal to (an algebraic abstraction of) the torsion exponents from Fukaya, Oh, Ohta and Ono [20]. Since the definition of the torsion exponents in [20] requires first extending coefficients to the universal Novikov field (with $\Gamma = \mathbb{R}$), whereas our definition in terms of singular value decompositions does not require such an extension, this implies new restrictions on the values that the torsion exponents can take: in particular, they all must be equal to differences between filtration levels of chains in the original Floer complex.

1.1.1 Barcodes Our fundamental invariants of a Floer-type complex, the "verbose barcode" and the "concise barcode", are defined in Definition 6.3. The verbose barcode in any given degree is a *finite* multiset of elements $([a], L)$ of the Cartesian product $(\mathbb{R}/\Gamma) \times [0, \infty]$, where $\Gamma \le \mathbb{R}$ is the subgroup described above and involved in the definition of the Novikov field $\Lambda = \Lambda^{\mathcal{K},\Gamma}$. The concise barcode is simply the submultiset of the verbose barcode consisting of elements $([a], L)$ with $L > 0$. Both barcodes are constructed in an explicit way from singular value decompositions of the graded pieces of the boundary operator on a Floer-type complex.

To be a bit more specific, as is made explicit in Proposition 7.4, a singular value decomposition can be thought of as expressing the Floer-type complex as an *orthogonal* direct sum of very simple complexes[2] having the form

$$\text{(2)}\quad \cdots \to 0 \to \mathrm{span}_\Lambda\{y\} \to \mathrm{span}_\Lambda\{\partial y\} \to 0 \to \cdots \qquad \text{or} \quad \cdots \to 0 \to \mathrm{span}_\Lambda\{x\} \to 0 \to \cdots ,$$

and the verbose barcode consists of the elements $([\ell(\partial y)], \ell(y) - \ell(\partial y))$ for summands of the first type and $([\ell(x)], \infty)$ for summands of the second type. The concise barcode discards those elements coming from summands with $\ell(\partial y) = \ell(y)$ (as these do not affect any of the filtered homology groups).

To put these barcodes into context, suppose that $\Gamma = \{0\}$ and that our Floer-type complex (C_*, ∂, ℓ) is given by the Morse complex $\mathrm{CM}_*(f)$ of a Morse function f on a compact manifold X (with ℓ recording the highest critical value attained by a given chain in the Morse chain complex). Then standard persistent homology methods associate to f a barcode, which is a collection of intervals $[a,b)$ with $a < b \leq \infty$, given the interpretation that each interval $[a,b)$ in the collection corresponds to a topological feature of X which is "born" at the level $\{f = a\}$ and "dies" at the level $\{f = b\}$ (or never dies if $b = \infty$). Theorem 6.2 proves that, when $\Gamma = \{0\}$ (so that $\mathbb{R}/\Gamma = \mathbb{R}$), our concise barcode is equivalent to the classical persistent homology barcode under the correspondence that sends a pair (a, L) in the concise barcode to an interval $[a, a+L)$. (Thus the second coordinates L in our elements of the concise barcode correspond to the lengths of bars in the persistent homology barcode.) To relate this back to the persistence module $\{\mathrm{HM}_*^t(f)\}_{t\in\mathbb{R}} \cong \{H_*(\{f \leq t\}; \mathcal{K})\}_{t\in\mathbb{R}}$ discussed earlier in the introduction, each $\mathrm{HM}_k^t(f)$ has dimension equal to the number of elements (a, L) in the degree-k concise barcode such that $a \leq t < a + L$, and the rank of the inclusion-induced map $\mathrm{HM}_k^s(f) \to \mathrm{HM}_k^t(f)$ is equal to the number of such elements with $a \leq s \leq t < a + L$.

When Γ is a nontrivial subgroup of $\mathbb{R}$, a Floer-type complex over Λ is more akin to the Morse–Novikov complex of a multivalued function f, where the ambiguity of the values of f is given by the group Γ (for instance, identifying $S^1 = \mathbb{R}/\mathbb{Z}$, for an S^1–valued function we would have $\Gamma = \mathbb{Z}$). While this situation lies outside the scope of classical persistent homology barcodes for reasons indicated earlier in the introduction, on a naive level it should be clear that if a topological feature of X is born where $f = a$ and dies where $f = b$ (corresponding to a bar $[a,b)$ in a hypothetical

[2] The "Morse–Barannikov complex" described in Barannikov [2] and Le Peutrec, Nier and Viterbo [28, Section 2] can be seen as a special case of this direct sum decomposition when $\Gamma = \{0\}$ and the Floer-type complex is the Morse complex of a Morse function whose critical values are all distinct; see Remark 5.6 for details.

barcode), then it should equally be true that, for any $g \in \Gamma$, a topological feature of X is born where $f = a + g$ and dies where $f = b + g$. So bars would come in Γ–parametrized families with Γ acting on both endpoints of the interval; such families in turn can be specified by the coset $[a]$ of the left endpoint a in $\mathbb{R}/\Gamma$ together with the length $L = b - a \in [0, \infty]$. This motivates our definition of the verbose and concise barcodes as multisets of elements of $(\mathbb{R}/\Gamma) \times [0, \infty]$. In terms of the summands in (2), the need to quotient by Γ simply comes from the fact that the elements y and x are only specified up to the scalar multiplication action of $\Lambda \setminus \{0\}$, which can affect their filtration levels by an arbitrary element of Γ. The following classification results are two of the main theorems of this paper.

Theorem A *Two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ are filtered chain isomorphic to each other if and only if they have identical verbose barcodes in all degrees.*

Theorem B *Two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ are filtered chain homotopy equivalent to each other if and only if they have identical concise barcodes in all degrees.*

Theorem A includes the statement that the verbose (and hence also the concise) barcode is independent of the singular value decomposition used to define it; indeed this statement is probably the hardest part of Theorems A and B to prove. We prove these theorems in Section 7.

As should already be clear from the above discussion, the only distinction between the verbose and concise barcodes of a Floer-type complex (C_*, ∂, ℓ) arises from elements $y \in C_*$ with $\ell(\partial y) = \ell(y)$. While our definition of a Floer-type complex only imposes the inequality $\ell(\partial y) \leq \ell(y)$, in many of the most important examples, including the Morse complex of a Morse function or the Hamiltonian Floer complex of a nondegenerate Hamiltonian, one in fact always has a strict inequality $\ell(\partial y) < \ell(y)$ for all $y \in C_* \setminus \{0\}$. For complexes satisfying this latter property the verbose and concise barcodes are equal, and so Theorems A and B show that the filtered chain isomorphism classification of such complexes is exactly the same as their filtered chain homotopy equivalence classification. (This fact can also be proven in a more direct way; see for instance the argument at the end of Usher [44, Proof of Lemma 3.8].)

In Remark 4.3 we mention some examples of naturally occurring Floer-type complexes in which an equality $\ell(\partial y) = \ell(y)$ can sometimes hold. In these complexes the verbose and concise barcodes are generally different, and thus the filtered chain homotopy equivalence classification is coarser than the filtered chain isomorphism classification. For many purposes the filtered chain isomorphism classification is likely too fine, in that it may depend on auxiliary choices made in the construction of the complex (for instance, in the Morse–Bott complex as constructed in Frauenfelder [19], it would depend on the choices of Morse functions on the critical submanifolds of the Morse–Bott function under consideration). The filtered chain homotopy type (and thus, by Theorem B, the concise barcode) is generally insensitive to such choices, and moreover is robust in a sense made precise in Theorem 1.4.

When $\Gamma = \{0\}$, Theorem B may be seen as an analogue of standard results from persistent homology theory (like [46, Corollary 3.1]) which imply that the degree-k barcode of a Floer-type complex completely classifies the persistence module obtained from its filtered homologies $H_k^t(C_*)$. Of course, the filtered chain homotopy type of a filtered chain complex is sufficient to determine its filtered homologies. Conversely, still assuming that $\Gamma = \{0\}$, by using the description of finite-type persistence modules as $\mathcal{K}[t]$–modules in Zomorodian and Carlsson [46], and taking advantage of the fact that (because $\mathcal{K}[t]$ is a PID) chain complexes of free $\mathcal{K}[t]$–modules are classified up to chain homotopy equivalence by their homology, one can show that the filtered chain homotopy type of a Floer-type complex is determined by its filtered homology persistence module. Thus, although the persistent homology literature generally focuses on homological invariants rather than classification of the underlying chain complexes up to filtered isomorphism or filtered homotopy equivalence, when $\Gamma = \{0\}$ Theorem B can be deduced from [46] together with a little homological algebra and Theorem 6.2.

For any choice of the group Γ, the concise barcode contains information about various numerical invariants of Floer-type complexes that have previously been used in filtered Floer theory. In particular, by Theorems 4.11 and 4.13 and the definition of the concise barcode, the torsion exponents from Fukaya, Oh, Ohta and Ono [20] are precisely the second coordinates L of elements $([a], L)$ of the concise barcode having $L < \infty$, written in decreasing order; the boundary depth of [44] is just the largest of these. In Section 6.1 we show that the concise barcode also carries information about the spectral invariants as in Schwarz [40] and Oh [34]. In particular, a number a arises as the spectral invariant of some class in the homology of the complex if and only if there is an element of form $([a], \infty)$ in the concise barcode. By contrast, the numbers a appearing in elements $([a], L)$ of the concise barcode with $L < \infty$ do not seem to have standard analogues in Floer theory, and so could be considered as new invariants. Whereas the spectral invariants and boundary depth have the notable feature of varying in Lipschitz fashion with respect to the Hofer norm on the space of Hamiltonians, these numbers a have somewhat more limited robustness properties, which can be understood in terms of our stability results such as Corollary 1.5 below.

In Section 6.2 we show how the verbose (and hence also the concise) barcodes of a Floer-type complex in various degrees are related to those of its dual complex, and to those of the complex obtained by extending the coefficient field by enlarging the group Γ. The relationships are rather simple; in the case of the dual complex they can be seen as extending results from Usher [43] on the Floer theory side and from de Silva, Morozov and Vejdemo-Johansson [41] on the persistent homology side.

Remark 1.1 Our approach differs from the conventional approach in the persistent homology literature in that we work almost entirely at the chain level; for the most part our theorems do not directly discuss the homology persistence modules $\{H_k^t(C_*)\}_{t\in\mathbb{R}}$. The primary reason for this is that, when $\Gamma \neq \{0\}$, such homology persistence modules are unlikely to fit into any reasonable classification scheme. The basic premise of the original introduction of barcodes in [46] is that a finite-type persistence module over a field $\mathcal{K}$ can be understood in terms of the classification of finitely generated $\mathcal{K}[x]$–modules; however, when $\Gamma \neq \{0\}$ our persistence modules are infinitely generated over $\mathcal{K}$, leading to infinitely generated $\mathcal{K}[x]$–modules and suggesting that one should work with a larger coefficient ring than $\mathcal{K}$. Since the action of the Novikov field does not preserve the filtration on the chain complex, the $H_k^t(C_*)$ are not modules over the full Novikov field Λ. They are however modules over the subring $\Lambda_{\geq 0}$ consisting of elements $\sum_g a_g T^g$ with all $g \geq 0$, and if Γ is nontrivial and discrete (in which case $\Lambda_{\geq 0}$ is isomorphic to a formal power series ring $\mathcal{K}[\![t]\!]$) then each $H_k^t(C_*)$ is a finitely generated $\Lambda_{\geq 0}$–module. But then the approach from [46] leads to the consideration of finitely generated $\mathcal{K}[\![t]\!][x]$–modules, which again do not admit a simple description in terms of barcode-type data since $\mathcal{K}[\![t]\!][x]$ is not a PID.

Our chain-level approach exploits the fact that the chain groups C_k in a Floer-type complex, unlike the filtered homologies, are finitely generated vector spaces over a field (namely Λ), which makes it more feasible to obtain a straightforward classification. It does follow from our results that the filtered homology persistence module of a Floer-type complex can be expressed as a finite direct sum of filtered homology persistence modules of the building blocks $\mathcal{E}(a, L, k)$ depicted in (2). However, since the filtered homology persistence modules of the $\mathcal{E}(a, L, k)$ are themselves somewhat complicated (as the interested reader may verify by direct computation) it is not clear whether this is a useful observation. For instance, we do not know whether the image on homology of a filtered chain map between two Floer-type complexes can always likewise be written as a direct sum of these basic persistence modules; if this is true then it might be possible to adapt arguments from Bauer and Lesnick [3] or Chazal, de Silva, Glisse and Oudot [9, Section 3.4] to remove the factor of 2 in Theorem 1.4.

1.1.2 Stability Among the most important theorems in persistent homology theory is the bottleneck stability theorem, which in its original form (see Cohen-Steiner, Edelsbrunner and Harer [10]) shows that the barcodes of the sublevel persistence modules $\{H_*(\{f \leq t\}; \mathcal{K}\}_{t\in\mathbb{R}}$ associated to suitably tame functions $f\colon X \to \mathbb{R}$ on a fixed topological space X depend in 1–Lipschitz fashion on f, where we use the C^0–norm to measure the distance between functions and the bottleneck distance (recalled below) to measure distances between barcodes. Since in applications there is inevitably some imprecision in the function f, some sort of result along these lines is evidently important in order to ensure that the barcode detects robust information. More recently, a number of extensions and new proofs of the bottleneck stability theorem have appeared, for instance in Chazal, Cohen-Steiner, Glisse, Guibas and Oudot [8], Chazal, de Silva, Glisse and Oudot [9] and Bauer and Lesnick [3]; these have recast the theorem as an essentially algebraic result about persistence modules satisfying a finiteness condition such as q–tameness or pointwise finite-dimensionality (see [3, pages 163, 167] for precise definitions). When recast in this fashion the stability theorem can be improved to an isometry theorem, stating that two natural metrics on an appropriate class of persistence modules are equal.

Hamiltonian Floer theory (see Floer [17], Hofer and Salamon [24], Liu and Tian [30], Fukaya and Ono [22] and Pardon [35]) associates a Floer-type complex to any suitably nondegenerate Hamiltonian $H\colon S^1 \times M \to \mathbb{R}$ on a compact symplectic manifold (M, ω). A well-established and useful principle in Hamiltonian Floer theory is that many aspects of the filtered Floer complex are robust under C^0–small perturbations of the Hamiltonian; for instance, various $\mathbb{R}$–valued quantities that can be extracted from the Floer complex such as spectral invariants and boundary depth are Lipschitz with respect to the C^0–norm on Hamiltonian functions (see Schwarz [40], Oh [34] and Usher [44]). Naively this is rather surprising since C^0–perturbing a Hamiltonian can dramatically alter its Hamiltonian flow. Our notion of the concise barcode — which by Theorem B gives a complete invariant of the filtered chain homotopy type of a Floer-type complex — allows us to obtain a more complete understanding of this C^0–rigidity property, as an instance of a general algebraic result which extends the bottleneck stability/isometry theorem to Floer-type complexes for general subgroups $\Gamma \leq \mathbb{R}$.

In order to formulate our version of the stability theorem we must explain the notions of distance that we use between Floer-type complexes on the one hand and concise barcodes on the other. Beginning with the latter, consider two multisets $\mathcal{S}$ and $\mathcal{T}$ of elements of $(\mathbb{R}/\Gamma) \times [0, \infty]$. For $\delta \geq 0$, a δ*-matching* between $\mathcal{S}$ and $\mathcal{T}$ consists of the following data:

(i) Submultisets $\mathcal{S}_{\mathrm{short}}$ and $\mathcal{T}_{\mathrm{short}}$ such that the second coordinate L of every element $([a], L) \in \mathcal{S}_{\mathrm{short}} \cup \mathcal{T}_{\mathrm{short}}$ obeys $L \leq 2\delta$.

(ii) A bijection $\sigma\colon \mathcal{S} \setminus \mathcal{S}_{\mathrm{short}} \to \mathcal{T} \setminus \mathcal{T}_{\mathrm{short}}$ such that, for each $([a], L) \in \mathcal{S} \setminus \mathcal{S}_{\mathrm{short}}$ (where $a \in \mathbb{R}$, $L \in [0, \infty]$) we have $\sigma([a], L) = ([a'], L')$, where for all $\epsilon > 0$ the representative a' of the coset $[a'] \in \mathbb{R}/\Gamma$ can be chosen such that both $|a' - a| \leq \delta + \epsilon$ and either $L = L' = \infty$ or $|(a' + L') - (a + L)| \leq \delta + \epsilon$.

Thus, viewing elements $([a], L)$ as corresponding to intervals $[a, a+L)$ (modulo Γ–translation), a δ–matching is a matching which shifts both endpoints of each interval by at most δ, with the proviso that we allow an interval I to be matched with a fictitious zero-length interval at the center of I.

Definition 1.2 If $\mathcal{S}$ and $\mathcal{T}$ are two multisets of elements of $(\mathbb{R}/\Gamma) \times [0, \infty]$ then the *bottleneck distance* between $\mathcal{S}$ and $\mathcal{T}$ is

$$d_B(\mathcal{S}, \mathcal{T}) = \inf\{\delta \geq 0 \mid \text{there exists a } \delta\text{–matching between } \mathcal{S} \text{ and } \mathcal{T}\}.$$

If $\mathcal{S} = \{\mathcal{S}_k\}_{k \in \mathbb{Z}}$ and $\mathcal{T} = \{\mathcal{T}_k\}_{k \in \mathbb{Z}}$ are two $\mathbb{Z}$–parametrized families of multisets of elements of $(\mathbb{R}/\Gamma) \times [0, \infty]$ then we write

$$d_B(\mathcal{S}, \mathcal{T}) = \sup_{k \in \mathbb{Z}} d_B(\mathcal{S}_k, \mathcal{T}_k).$$

It is easy to see that in the special case where $\Gamma = \{0\}$ the above definition agrees with the notion of bottleneck distance in Cohen-Steiner, Edelsbrunner and Harer [10]. Note that the value d_B can easily be infinity. For instance this occurs if $\mathcal{S} = \{([a], \infty)\}$ and $\mathcal{T} = \{([a], L)\}$, where $L < \infty$.

On the Floer complex side, we make the following definition, which is a slight modification of Usher [45, Definition 3.7]. As is explained in the appendix, this is very closely related to the notion of *interleaving* of persistence modules from [8].

Definition 1.3 Let $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ be two Floer-type complexes, and $\delta \geq 0$. A *δ–quasiequivalence* between C_* and D_* is a quadruple (Φ, Ψ, K_1, K_2), where:

- $\Phi\colon C_* \to D_*$ and $\Psi\colon D_* \to C_*$ are chain maps, with
$$\ell_D(\Phi c) \leq \ell_C(c) + \delta \quad \text{and} \quad \ell_C(\Psi d) \leq \ell_D(d) + \delta$$
for all $c \in C_*$ and $d \in D_*$.
- $K_C\colon C_* \to C_{*+1}$ and $K_D\colon D_* \to D_{*+1}$ obey the homotopy equations
$$\Psi \circ \Phi - I_{C_*} = \partial_C K_C + K_C \partial_C \quad \text{and} \quad \Phi \circ \Psi - I_{D_*} = \partial_D K_D + K_D \partial_D,$$
and for all $c \in C_*$ and $d \in D_*$ we have
$$\ell_C(K_C c) \leq \ell_C(c) + 2\delta \quad \text{and} \quad \ell_D(K_D d) \leq \ell_D(d) + 2\delta.$$

The *quasiequivalence distance* between $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ is then defined to be

$$d_Q\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big) = \inf\big\{\delta \geq 0 \mid \exists\ \delta\text{–quasiequivalence between} \\ (C_*, \partial_C, \ell_C) \text{ and } (D_*, \partial_D, \ell_D)\big\}.$$

We will prove the following as Theorems 8.17 and 8.18 in Sections 9 and 10:

Theorem 1.4 *Given a Floer-type complex $(C_*, \partial_C, \ell_C)$, denote its concise barcode by $\mathcal{B}(C_*, \partial_C, \ell_C)$ and the degree-k part of its concise barcode by $\mathcal{B}_{C,k}$. Then the bottleneck and quasiequivalence distances obey, for any Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$, the following conditions:*

(i) $$d_Q\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big) \leq d_B\big(\mathcal{B}(C_*, \partial_C, \ell_C), \mathcal{B}(D_*, \partial_D, \ell_D)\big) \\ \leq 2 d_Q\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big).$$

(ii) *For $k \in \mathbb{Z}$ let $\Delta_{D,k} > 0$ denote the smallest second coordinate L of all of the elements of $\mathcal{B}_{D,k}$. If $d_Q\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big) < \frac{1}{4}\Delta_{D,k}$, then*

$$d_B(\mathcal{B}_{C,k}, \mathcal{B}_{D,k}) \leq d_Q\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big).$$

Thus the map from filtered chain homotopy equivalence classes of Floer-type complexes to concise barcodes is at least bi-Lipschitz, with Lipschitz constant 2. We expect that it is always an isometry; in fact when $\Gamma = \{0\}$ this can be inferred from [9, Theorem 4.11] and Theorem 6.2, and as mentioned in Remark 9.15 it is also true in the opposite extreme case when Γ is dense.

Our proof that the bottleneck distance d_B obeys the upper bounds of Theorem 1.4 is roughly divided into two parts. First, in Proposition 9.3, we prove the sharp inequality $d_B \leq d_Q$ in the special case that the Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ have the same underlying chain complex, and differ only in their filtration functions ℓ_C and ℓ_D. In the rest of Section 9 we approximately reduce the general case to this special case, using a mapping cylinder construction to obtain two different filtration functions on a single chain complex, one of which has concise barcode equal to that of $(D_*, \partial_D, \ell_D)$ (see Proposition 9.12), and the other of which has concise barcode consisting of the concise barcode of $(C_*, \ell_C, \partial_C)$ together with some "extra" elements $([a], L) \in (\mathbb{R}/\Gamma) \times [0, \infty]$ all having $L \leq 2 d_Q\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big)$ (see Proposition 9.13). These constructions are quickly seen in Section 9.5 to yield the upper bounds on d_B in the two parts of Theorem 1.4; the factor of 2 in part (i) arises from the "extra" bars in the concise barcode of the Floer-type complex from Proposition 9.13.

In contrast, the proof of the other inequality $d_Q \le d_B$ in Theorem 1.4(i) is considerably simpler, and is carried out by a direct construction in Section 10.

As mentioned earlier, it is likely that the factor of 2 in Theorem 1.4(i) is unnecessary, ie that the map from Floer-type complexes to concise barcodes is an isometry with respect to the quasiequivalence distance d_Q on Floer-type complexes and the bottleneck distance d_B on concise barcodes. Although we do not prove this, by taking advantage of Theorem 1.4(ii) we show in Section 11 that, if d_Q is replaced by a somewhat more complicated distance d_P that we call the interpolating distance, then the map is indeed an isometry (see Theorem 11.2). The expected isometry between d_Q and d_B is then equivalent to the statement that $d_P = d_Q$. Consistently with this, our experience in concrete situations has been that methods which lead to bounds on one of d_P or d_Q often also produce identical bounds on the other.

The final section of the body of the paper applies our general algebraic results to Hamiltonian Floer theory, the relevant features of which are reviewed at the beginning of that section.[3] Combining Theorem 11.2 with standard results from Hamiltonian Floer theory proves the following:

Corollary 1.5 *If H_0 and H_1 are two nondegenerate Hamiltonians on any compact symplectic manifold (M,ω), then the bottleneck distance between the concise barcodes of $(\mathrm{CF}_*(H_0), \partial_{H_0}, \ell_{H_0})$ and $(\mathrm{CF}_*(H_1), \partial_{H_1}, \ell_{H_1})$ is less than or equal to $\int_0^1 \|H_1(t,\cdot) - H_0(t,\cdot)\|_{L^\infty}\, dt$.*

To summarize, we have shown how to associate to the Hamiltonian Floer complex combinatorial data, in the form of the concise barcode, which completely classifies the complex up to filtered chain homotopy equivalence, and which is continuous with respect to variations in the Hamiltonian in a way made precise in Corollary 1.5. Given the way in which torsion exponents, the boundary depth, and spectral invariants are encoded in the concise barcode, this continuity can be seen as a simultaneous extension of continuity results for those quantities; see Fukaya, Oh, Ohta and Ono [20, Theorem 6.1.25], Usher [44, Theorem 1.1(ii)] and Schwarz [40, (12)].

We then apply Corollary 1.5 to prove our main application, Theorem 12.2, concerning the robustness of the fixed points of a nondegenerate Hamiltonian diffeomorphism under C^0–perturbations of the Hamiltonian: roughly speaking, as long as the perturbation is small enough (as determined by the concise barcode of the original Hamiltonian), the perturbed Hamiltonian, if it is still nondegenerate, will have at least as many fixed

[3] While we focus on Hamiltonian Floer theory in Section 12, very similar results would apply to the Hamiltonian-perturbed Lagrangian Floer chain complexes or to the chain complexes underlying Novikov homology.

points as the original one, with actions that are close to the original actions. Moreover, depending in a precise way on the concise barcode, fixed points with certain actions may be identified as enjoying stronger robustness properties (in the sense that a larger perturbation is required to eliminate them) than general fixed points of the same map. While C^0–robustness of fixed points is a familiar idea in Hamiltonian Floer theory (see eg Cornea and Ranicki [11, Theorem 2.1]), Theorem 12.2 goes farther than previous results both in its control over the actions of the perturbed fixed points and in the way that it gives stronger bounds for the robustness of unperturbed fixed points with certain actions (see Remark 12.3).

Finally, the appendix identifies the quasiequivalence distance d_Q that features in Theorem 1.4 with a chain level version of the interleaving distance that is commonly used (eg in [8]) in the persistent homology literature.

Acknowledgements Usher thanks the TDA Study Group in the UGA Statistics Department for introducing him to persistent homology, and K Ono for pointing out the likely relationship between torsion thresholds and boundary depth. Both authors are grateful to L Polterovich for encouraging us to study Floer theory from a persistent-homological point of view, for comments on an initial version of the paper, and for various useful conversations. Some of these conversations occurred when Zhang visited Tel Aviv University in Fall 2014; he is indebted to its hospitality and also to a guided reading course overseen by L Polterovich. In particular, discussions with D Rosen during this visit raised the question considered in the appendix. The authors also thank the anonymous referee for a careful reading and many suggestions. This work was partially supported by NSF grants DMS-1105700 and DMS-1509213.

2 Nonarchimedean orthogonality

2.1 Nonarchimedean normed vector spaces

Fixing a ground field $\mathcal{K}$ and an additive subgroup $\Gamma \leq \mathbb{R}$ as in the introduction, we will consider vector spaces over the *Novikov field* defined as

$$\Lambda = \Lambda^{\mathcal{K},\Gamma} = \left\{ \sum_{g \in \Gamma} a_g T^g \;\middle|\; a_g \in \mathcal{K} \text{ and } \#\{g \mid a_g \neq 0, g < C\} < \infty \text{ for all } C \in \mathbb{R} \right\},$$

where T is a formal symbol and we use the obvious "power series" addition and multiplication. This Novikov field adapts the ring used by Novikov in his version of Morse theory for multivalued functions; see [24] both for some of its algebraic properties and for its use in Hamiltonian Floer homology. Note that when Γ is the trivial group, Λ reduces to the ground field $\mathcal{K}$.

First, we need the following classical definition.

Definition 2.1 A *valuation* ν on a field $\mathcal{F}$ is a function $\nu\colon \mathcal{F} \to \mathbb{R} \cup \{\infty\}$ such that

(V1) $\nu(x) = \infty$ if and only if $x = 0$;

(V2) $\nu(xy) = \nu(x) + \nu(y)$ for any $x, y \in \mathcal{F}$;

(V3) $\nu(x + y) \geq \min\{\nu(x), \nu(y)\}$ for any $x, y \in \mathcal{F}$.

Moreover, we call a valuation ν *trivial* if $\nu(x) = 0$ for $x \neq 0$ and $\nu(x) = \infty$ precisely when $x = 0$.

For $\mathcal{F} = \Lambda$ defined as above, we can associate a valuation simply by

$$\nu\left(\sum_{g\in\Gamma} a_g T^g\right) = \min\{g \mid a_g \neq 0\},$$

where we use the standard convention that the minimum of the empty set is ∞. It is easy to see that this ν satisfies conditions (V1), (V2) and (V3). Note that the finiteness condition in the definition of Novikov field ensures that the minimum exists. If $\Gamma = \{0\}$, then the valuation ν is trivial.

Definition 2.2 A *nonarchimedean normed vector space* over Λ is a pair (C, ℓ), where C is a vector space over Λ endowed with a filtration function $\ell\colon C \to \mathbb{R} \cup \{-\infty\}$ satisfying the following axioms:

(F1) $\ell(x) = -\infty$ if and only if $x = 0$;

(F2) $\ell(\lambda x) = \ell(x) - \nu(\lambda)$ for any $\lambda \in \Lambda$ and $x \in C$;

(F3) $\ell(x + y) \leq \max\{\ell(x), \ell(y)\}$ for any $x, y \in C$.

In terms of Definition 2.2, the standard convention would be that the norm on a nonarchimedean normed vector space (C, ℓ) is e^{ℓ}, not ℓ. The phrasing of the above definition reflects the fact that we will focus on the function ℓ, not on the norm e^{ℓ}.

We record the following standard fact:

Proposition 2.3 *If (C, ℓ) is a nonarchimedean normed vector space over Λ and the elements $x, y \in C$ satisfy $\ell(x) \neq \ell(y)$, then*

$$\ell(x + y) = \max\{\ell(x), \ell(y)\}. \tag{3}$$

Proof Of course the inequality "$\leq$" in (3) is just (F3). For "$\geq$" we assume without loss of generality that $\ell(x) > \ell(y)$, so we are to show that $\ell(x+y) \geq \ell(x)$. Now (F2) implies that $\ell(-y) = \ell(y)$, so $\ell(x) = \ell((x+y)+(-y)) \leq \max\{\ell(x+y), \ell(y)\}$. Thus since we have assumed that $\ell(x) > \ell(y)$ we indeed must have $\ell(x) \leq \ell(x+y)$. □

Example 2.4 (Rips complexes) Let X be a collection of points in euclidean space. We will define a one-parameter family of "Rips complexes" associated to X as follows. Let $\mathrm{CR}_*(X)$ be the simplicial chain complex over $\mathcal{K}$ of the complete simplicial complex on the set X, so that $\mathrm{CR}_k(X)$ is the free $\mathcal{K}$–vector space generated by the k–simplices all of whose vertices lie in X. Define $\ell\colon \mathrm{CR}_*(X) \to \mathbb{R} \cup \{-\infty\}$ by setting $\ell(\sum_i a_i\sigma_i)$ equal to the largest diameter of any of the simplices σ_i with $a_i \neq 0$ (and to $-\infty$ when $\sum_i a_i\sigma_i = 0$). Then $(\mathrm{CR}_*(X), \ell)$ is a nonarchimedean vector space over $\Lambda^{\mathcal{K},\{0\}} = \mathcal{K}$. For any $\epsilon > 0$ we define the Rips complex with parameter ϵ, $\mathrm{CR}_*(X;\epsilon)$, to be the subcomplex of C_* with degree-k part given by

$$\mathrm{CR}_k(X;\epsilon) = \{c \in \mathrm{CR}_k(X) \mid \ell(x) \leq \epsilon\}.$$

Thus $\mathrm{CR}_*(X;\epsilon)$ is spanned by those simplices with diameter at most ϵ. The standard simplicial boundary operator maps $\mathrm{CR}_k(X;\epsilon)$ to $\mathrm{CR}_{k-1}(X;\epsilon)$, yielding Rips homology groups $HR_k(X;\epsilon)$, and the dependence of these homology groups on ϵ is a standard object of study in applied persistent homology, as in [46].

Example 2.5 (Morse complex) Suppose we have a closed manifold X and f is a Morse function on X. We may then consider its Morse chain complex $\mathrm{CM}_*(X;f)$ over the field $\mathcal{K} = \Lambda^{\mathcal{K},\{0\}}$ as in [39]. Let $C = \bigoplus_k \mathrm{CM}_k(X;f)$. For any element $x \in C$, by the definition of the Morse chain complex, $x = \sum_i a_i p_i$, where each p_i is a critical point and $a_i \in \mathcal{K}$. Then define $\ell\colon C \to \mathbb{R} \cup \{-\infty\}$ by

$$\ell\Big(\sum\nolimits_i a_i p_i\Big) = \max\{f(p_i) \mid a_i \neq 0\},$$

with the usual convention that the maximum of the empty set is $-\infty$. It is easy to see that ℓ satisfies (F1), (F2) and (F3) above. Therefore, $\big(\bigoplus_k \mathrm{CM}_k(X;f), \ell\big)$ is a nonarchimedean normed vector space over $\mathcal{K} = \Lambda^{\mathcal{K},\{0\}}$.

Example 2.6 Given a closed one-form α on a closed manifold M, let $\pi\colon \widetilde{M} \to M$ denote the regular covering space of M that is minimal subject to the property that $\pi^*\alpha$ is exact, and choose $\tilde{f}\colon \widetilde{M} \to \mathbb{R}$ such that $d\tilde{f} = \pi^*\alpha$. The graded parts $\mathrm{CN}_k(\tilde{f})$ of the Novikov complex (see (1)) can likewise be seen as nonarchimedean vector spaces over $\Lambda = \Lambda^{\mathcal{K},\Gamma}$, where the group $\Gamma \leq \mathbb{R}$ consists of all possible integrals of α around loops in M. Namely, just as in the previous two examples we put

$$\ell\Big(\sum a_p p\Big) = \max\{\tilde{f}(p) \mid a_p \neq 0\}.$$

We leave verification of axioms (F1), (F2), and (F3) to the reader.

2.2 Orthogonality

We use the standard notions of orthogonality in nonarchimedean normed vector spaces (see [31]).

Definition 2.7 Let (C,ℓ) be a nonarchimedean normed vector space over a Novikov field Λ.

- Two subspaces V and W of C are said to be *orthogonal* if for all $v\in V$ and $w\in W$, we have
$$\ell(v+w)=\max\{\ell(v),\ell(w)\}.$$
- A finite ordered collection $(w_1,\dots,w_r)$ of elements of C is said to be *orthogonal* if, for all $\lambda_1,\dots,\lambda_r\in\Lambda$, we have
$$\ell\Bigl(\sum_{i=1}^{r}\lambda_i w_i\Bigr)=\max_{1\le i\le r}\ell(\lambda_i w_i). \tag{4}$$

In particular a pair (v,w) of elements of C is orthogonal if and only if the spans $\langle v\rangle_\Lambda$ and $\langle w\rangle_\Lambda$ are orthogonal as subspaces of C. Of course, by (F2), the criterion (4) can equivalently be written as

$$\ell\Bigl(\sum_{i=1}^{r}\lambda_i w_i\Bigr)=\max_{1\le i\le r}(\ell(w_i)-\nu(\lambda_i)). \tag{5}$$

Example 2.8 Here is a simple example illustrating the notion of orthogonality. Let $\Gamma=\{0\}$ so that $\Lambda=\mathcal{K}$ has the trivial valuation defined in Definition 2.1. Let C be a two-dimensional $\mathcal{K}$–vector space, spanned by elements x,y. We may define a filtration function ℓ on C by declaring (x,y) to be an orthogonal basis with, say, $\ell(x)=1$ and $\ell(y)=0$; then in accordance with (5) and the definition of the trivial valuation ν we will have

$$\ell(\lambda x+\eta y)=\begin{cases}1 & \text{if } \lambda\neq 0,\\ 0 & \text{if } \lambda=0,\ \eta\neq 0,\\ -\infty & \text{if } \lambda=\eta=0.\end{cases}$$

The ordered basis $(x+y,y)$ will likewise be orthogonal: indeed for $\lambda,\eta\in\mathcal{K}$ we have

$$\ell(\lambda(x+y)+\eta y)=\ell(\lambda x+(\lambda+\eta)y)=\begin{cases}1 & \text{if } \lambda\neq 0,\\ 0 & \text{if } \lambda=0,\ \lambda+\eta\neq 0,\\ -\infty & \text{if } \lambda=\eta=0,\end{cases}$$

which is indeed equal to the maximum of $\ell(\lambda(x+y))$ and $\ell(\eta y)$ (the former being 1 if $\lambda \neq 0$ and $-\infty$ otherwise, and the latter being 0 if $\eta \neq 0$ and $-\infty$ otherwise).

On the other hand the pair $(x, x+y)$ is *not* orthogonal: letting $\lambda = -1$ and $\eta = 1$ we see that $\ell(\lambda x + \eta(x+y)) = \ell(y) = 0$ whereas $\max\{\ell(\lambda x), \ell(\eta(x+y))\} = 1$.

Here are some simple but useful observations that follow directly from Definition 2.7.

Lemma 2.9 *If (C, ℓ) is an nonarchimedean normed vector space over Λ, then:*

(i) *If two subspaces U and V are orthogonal, then U intersects V trivially.*

(ii) *For subspaces U, V, W, if U and V are orthogonal, and $U \oplus V$ and W are orthogonal, then U and $V \oplus W$ are orthogonal.*

(iii) *If U and V are orthogonal subspaces of C, and if $(u_1, \dots, u_r)$ is an orthogonal ordered collection of elements of U while $(v_1, \dots, v_s)$ is an orthogonal ordered collection of elements of V, then $(u_1, \dots, u_r, v_1, \dots, v_s)$ is orthogonal in $U \oplus V$.*

Proof For (i), if $w \in U \cap V$, then noting that (F2) implies that $\ell(-w) = \ell(w)$, we see that, since $w \in U$ and $-w \in V$, where U and V are orthogonal,

$$-\infty = \ell(0) = \ell(w + (-w)) = \max\{\ell(w), \ell(w)\} = \ell(w),$$

and so $w = 0$ by (F1). So indeed U intersects V trivially.

For (ii), first note that if $U \oplus V$ and W are orthogonal, then in particular, V and W are orthogonal. For any elements $u \in U, v \in V$ and $w \in W$, we have

$$\begin{aligned} \ell(u + (v + w)) &= \ell((u + v) + w) \\ &= \max\{\ell(u + v), \ell(w)\} \\ &= \max\{\ell(u), \ell(v), \ell(w)\} \\ &= \max\{\ell(u), \ell(v + w)\}. \end{aligned}$$

The second equality comes from orthogonality between $U \oplus V$ and W; the third equality comes from orthogonality between U and V; and the last equality comes from orthogonality between V and W.

Part (iii) is an immediate consequence of the definitions. □

Definition 2.10 *An orthogonalizable Λ–space* (C, ℓ) is a finite-dimensional nonarchimedean normed vector space over Λ such that there exists an orthogonal basis for C.

Example 2.11 $(\Lambda, -\nu)$ is an orthogonalizable Λ–space.

Example 2.12 $(\Lambda^n, -\vec{\nu})$ is an orthogonalizable Λ–space, where $\vec{\nu}$ is defined as $\vec{\nu}(\lambda_1, \dots, \lambda_n) = \min_{1 \le i \le n} \nu(\lambda_i)$. Moreover, fixing some vector $\vec{t} = (t_1, \dots, t_n) \in \mathbb{R}^n$, the shifted version $(\Lambda^n, -\vec{\nu}_{\vec{t}})$ is also an orthogonalizable Λ–space, where $\vec{\nu}_{\vec{t}}$ is defined as

$$\vec{\nu}_{\vec{t}}(\lambda_1, \dots, \lambda_n) = \min_{1 \le i \le n} (\nu(\lambda_i) - t_i).$$

Specifically, an orthogonal ordered basis is given by the standard basis $(e_1, \dots, e_n)$ for Λ^n: indeed, we have $-\vec{\nu}_{\vec{t}}(e_i) = t_i$, and

$$-\vec{\nu}_{\vec{t}}\left(\sum_{i=1}^{n} \lambda_i e_i\right) = \max_{1 \le i \le n} (t_i - \nu(\lambda_i)) = \max_{1 \le i \le n} (-\vec{\nu}_{\vec{t}}(e_i) - \nu(\lambda_i)).$$

In Example 2.6 above, if we let $\{\tilde{p}_i\}_{i=1}^n \subset \widetilde{M}$ consist of one point in every fiber of the covering space $\widetilde{M} \to M$ that contains an index-k critical point, then it is easy to see that we have a vector space isomorphism $\mathrm{CN}_k(\tilde{f}) \cong \Lambda^n$, with the filtration function ℓ on $\mathrm{CN}_k(\tilde{f})$ mapping to the shifted filtration function $-\vec{\nu}_{\vec{t}}$, where $t_i = \tilde{f}(\tilde{p}_i)$.

Remark 2.13 In fact, using (F2) and the definition of orthogonality, it is easy to see that *any* orthogonalizable Λ–space (C, ℓ) is isomorphic in the obvious sense to some $(\Lambda^n, -\vec{\nu}_{\vec{t}})$: if $(v_1, \dots, v_n)$ is an ordered orthogonal basis for (C, ℓ) then mapping v_i to the i^{th} standard basis vector for Λ^n gives an isomorphism of vector spaces which sends ℓ to $-\vec{\nu}_{\vec{t}}$, where $t_i = \ell(v_i)$.

2.3 Nonarchimedean Gram–Schmidt process

In classical linear algebra, the Gram–Schmidt process is applied to modify a set of linearly independent elements into an orthogonal set. A similar procedure can be developed in the nonarchimedean context. The key part of this process comes from the following theorem, which we state using our notations in this paper (see Remark 2.13).

Theorem 2.14 [42, Theorem 2.5] *Suppose (C, ℓ) is an orthogonalizable Λ–space and $W \le C$ is a Λ–subspace. Then for any $x \in C \setminus W$ there exists some $w_0 \in W$ such that*

$$\ell(x - w_0) = \inf\{\ell(x - w) \mid w \in W\}. \tag{6}$$

Thus w_0 achieves the minimal distance to x among all elements of W. Note that (in contrast to the situation with more familiar notions of distance such as the euclidean distance on $\mathbb{R}^n$) the element w_0 is generally not unique. However, similarly to the case of the euclidean distance, solutions to this distance-minimization problem are closely related to orthogonality, as the following lemma shows.

Lemma 2.15 *Let (C,ℓ) be a nonarchimedean normed vector space over Λ, and let $W \le C$ be a Λ–subspace and $x \in C \backslash W$. Then W and $\langle x\rangle_\Lambda$ are orthogonal if and only if $\ell(x) = \inf\{\ell(x-w) \mid w \in W\}$.*

Proof Suppose W and $\langle x\rangle_\Lambda$ are orthogonal. Then for any $w \in W$, by orthogonality,

$$\ell(x-w) = \max\{\ell(x),\ell(w)\} \ge \ell(x).$$

Therefore, taking an infimum, we get $\inf\{\ell(x-w) \mid w \in W\} \ge \ell(x)$. Moreover, by taking $w = 0$, we have $\inf\{\ell(x-w) \mid w \in W\} \le \ell(x-0) = \ell(x)$. Therefore, $\ell(x) = \inf\{\ell(x-w) \mid w \in W\}$.

Conversely, suppose that $\ell(x) = \inf\{\ell(x-w) \mid w \in W\}$ and let $y = w + \mu x$ be a general element of $W \oplus \langle x\rangle_\Lambda$. We must show that $\ell(y) = \max\{\ell(w),\ell(\mu x)\}$; in fact, the inequality "$\le$" automatically follows from (F3), so we just need to show that $\ell(y) \ge \max\{\ell(w),\ell(\mu x)\}$. If $\mu = 0$ this is obvious since then $y = w$, so assume from now on that $\mu \ne 0$. Then

$$\ell(y) = \ell\big(\mu(\mu^{-1}w + x)\big) = \ell(\mu^{-1}w + x) - \nu(\mu) \ge \ell(x) - \nu(\mu) = \ell(\mu x),$$

where the inequality uses the assumed optimality property of x. If $\ell(\mu x) \ge \ell(w)$ this proves that $\ell(y) \ge \max\{\ell(w),\ell(\mu x)\}$. On the other hand if $\ell(\mu x) < \ell(w)$ then the fact that $\ell(y) \ge \max\{\ell(w),\ell(\mu x)\}$ simply follows by Proposition 2.3. □

Theorem 2.16 *(nonarchimedean Gram–Schmidt process). Let (C,ℓ) be an orthogonalizable Λ–space and let $\{x_1,\dots,x_r\}$ be a basis for a subspace $V \le C$. Then there exists an* orthogonal *ordered basis $(x_1',\dots,x_r')$ for V whose members have the form*

$$\begin{aligned} x_1' &= x_1, \\ x_2' &= x_2 - \lambda_{2,1}x_1, \\ &\vdots \\ x_r' &= x_r - \lambda_{r,r-1}x_{r-1} - \lambda_{r,r-2}x_{r-2} - \cdots - \lambda_{r,1}x_1, \end{aligned}$$

where the $\lambda_{\alpha,\beta}$ are constants in Λ. Moreover if the first i elements of the initial basis are such that $(x_1,\dots,x_i)$ are orthogonal, then we can take $x_j' = x_j$ for $j = 1,\dots,i$.

Proof We proceed by induction on the dimension r of V. If V is one-dimensional then we simply take $x_1' = x_1$. Assuming the result to be proven for all k–dimensional subspaces, let $(x_1,\dots,x_{k+1})$ be an ordered basis for V, with $(x_1,\dots,x_i)$ orthogonal for some $i\in\{1,\dots,k+1\}$. If $i=k+1$ then we can set $x'_j=x_j$ for all j and we are done. Otherwise apply the inductive hypothesis to the span W of $\{x_1,\dots,x_k\}$ to obtain an orthogonal ordered basis $(x'_1,\dots,x'_k)$ for W, with $x'_j=x_j$ for all $j\in\{1,\dots,i\}$. Now apply Theorem 2.14 to W and the element x_{k+1} to obtain some $w_0\in W$ such that $\ell(x_{k+1}-w_0)=\inf\{\ell(x_{k+1}-w)\mid w\in W\}$. Let $x'_{k+1}=x_{k+1}-w_0$. It then follows from Lemma 2.15 that W and $\langle x'_{k+1}\rangle_\Lambda$ are orthogonal, and so by Lemma 2.9(iii) $(x'_1,\dots,x'_k,x'_{k+1})$ is an orthogonal ordered basis for V. Moreover since $x'_{k+1}=x_{k+1}-w_0$, where w_0 lies in the span of $x_1,\dots,x_k$, it is clear that x_{k+1} has the form required in the theorem. This completes the inductive step and hence the proof. □

Corollary 2.17 *If (C,ℓ) is an orthogonalizable Λ–space, then for every subspace $W\le C$, $(W,\ell|_W)$ is also an orthogonalizable Λ–space.*

Proof Apply Theorem 2.16 to an arbitrary basis for W to obtain an orthogonal ordered basis for W. □

Corollary 2.18 *If (C,ℓ) is an orthogonalizable Λ–space and $V\le W\le C$, any orthogonal ordered basis of V may be extended to an orthogonal basis of W.*

Proof By Corollary 2.17, we have an orthogonal ordered basis $(v_1,\dots,v_i)$ for V. Extend it arbitrarily to a basis $\{v_1,\dots,v_i,v_{i+1},\dots,v_r\}$ for W, and then apply Theorem 2.16 to obtain an orthogonal ordered basis for W whose first i elements are $v_1,\dots,v_i$. □

Corollary 2.19 *Suppose that (C,ℓ) is an orthogonalizable Λ–space and $U\le C$. Then there exists a subspace V such that $U\oplus V=C$ and U and V are orthogonal.* (We call any such V an *orthogonal complement* of U).

Proof By Corollary 2.17, we have an orthogonal ordered basis $(u_1,\dots,u_k)$ for subspace U. By Corollary 2.18, extend it to an orthogonal ordered basis for C, say $(u_1,\dots,u_k,v_1,\dots,v_l)$ (so $\dim(C)=k+l$). Then $V=\mathrm{span}_\Lambda\{v_1,\dots,v_l\}$ satisfies the desired properties. □

Orthogonal complements are generally not unique, as is illustrated by Example 2.8, in which $\langle x+ay\rangle_{\mathcal{K}}$ is an orthogonal complement to $\langle y\rangle_{\mathcal{K}}$ for any $a\in\mathcal{K}$.

2.4 Duality

Given a nonarchimedean normed vector space (C,ℓ), the dual space C^* (over Λ) becomes a nonarchimedean normed vector space if we associate a filtration function $\ell^*\colon C^*\to\mathbb{R}\cup\{\infty\}$ defined by

$$\ell^*(\phi)=\sup_{0\neq x\in C}(-\ell(x)-\nu(\phi(x))).$$

Indeed, for ϕ and ψ in C^* and $x\in C$, we have

$$\begin{aligned}-\ell(x)-\nu(\phi(x)+\psi(x))&\leq-\ell(x)-\min\{\nu(\phi(x)),\nu(\psi(x))\}\\&=\max\{-\ell(x)-\nu(\phi(x)),-\ell(x)-\nu(\psi(x))\}\\&\leq\max\{\ell^*(\phi),\ell^*(\psi)\}\end{aligned}$$

and so taking the supremum over x shows that $\ell^*(\phi+\psi)\leq\max\{\ell^*(\phi),\ell^*(\psi)\}$, and it is easy to check the other axioms (F1) and (F3) required of ℓ^*. The following proposition demonstrates a relation between bases of the original space and its dual space.

Proposition 2.20 *If (C,ℓ) is an orthogonalizable Λ–space with orthogonal ordered basis $(v_1,\dots,v_n)$, then (C^*,ℓ^*) is an orthogonalizable Λ–space with an orthogonal ordered basis given by the dual basis $(v_1^*,\dots,v_n^*)$. Moreover, for each i, we have*

$$\ell^*(v_i^*)=-\ell(v_i). \tag{7}$$

Proof For any $x\in C$, written as $\sum_{j=1}^n\lambda_iv_i$, we have $v_i^*x=\lambda_i$ for each i, so if $\lambda_i=0$ then $-\ell(x)-\nu(v_i^*x)=-\infty$, while otherwise

$$\begin{aligned}-\ell(x)-\nu(v_i^*x)&=-\max_{1\leq j\leq n}(\ell(v_j)-\nu(\lambda_j))-\nu(\lambda_i)\\&\leq-(\ell(v_i)-\nu(\lambda_i))-\nu(\lambda_i)=-\ell(v_i).\end{aligned}$$

Equality holds in the above when $x=v_i$, so $\ell^*(v_i^*)=-\ell(v_i)$.

To prove orthogonality, given any $\lambda_1,\dots,\lambda_n\in\Lambda$, choose i_0 to maximize the quantity $-\ell(v_i)-\nu(\lambda_i)$ over $i\in\{1,\dots,n\}$. Then

$$\begin{aligned}\ell^*\left(\sum_{i=1}^n\lambda_iv_i^*\right)&\geq-\ell(v_{i_0})-\nu\left(\left(\sum_{i=1}^n\lambda_iv_i^*\right)v_{i_0}\right)\\&=-\ell(v_{i_0})-\nu(\lambda_{i_0})=\max_{1\leq i\leq n}(\ell^*(v_i^*)-\nu(\lambda_i)).\end{aligned}$$

The reverse direction immediately follows from the nonarchimedean triangle inequality (F3) in Definition 2.2. Therefore, we have proven the orthogonality of the dual basis. □

2.5 Coefficient extension

This is a somewhat technical subsection which is not used for most of the main results — mainly we are including it in order to relate our barcodes to the torsion exponents from [20] — so it could reasonably be omitted on first reading.

Throughout most of this paper we consider a fixed subgroup $\Gamma \le \mathbb{R}$, with associated Novikov field $\Lambda = \Lambda^{\mathcal{K},\Gamma}$, and we consider orthogonalizable Λ–spaces over this fixed Novikov field Λ. Suppose now that we consider a larger subgroup $\Gamma' \ge \Gamma$ (still with $\Gamma' \le \mathbb{R}$). The inclusion $\Gamma \hookrightarrow \Gamma'$ induces in obvious fashion a field extension $\Lambda \hookrightarrow \Lambda^{\mathcal{K},\Gamma'}$, and so for any Λ vector space C we obtain a $\Lambda^{\mathcal{K},\Gamma'}$–vector space

$$C' = C \otimes_\Lambda \Lambda^{\mathcal{K},\Gamma'}.$$

If (C,ℓ) is an orthogonalizable Λ–space with orthogonal ordered basis $(w_1,\dots,w_n)$ then $\{w_1 \otimes 1,\dots,w_n \otimes 1\}$ is a basis for C' and so we can make C' into an orthogonalizable $\Lambda^{\mathcal{K},\Gamma'}$–space (C',ℓ') by putting

$$\ell'\left(\sum_{i=1}^{n} \lambda'_i w_i \otimes 1\right) = \max_i(\ell(w_i) - \nu(\lambda'_i))$$

for all $\lambda'_1,\dots,\lambda'_n \in \Lambda^{\mathcal{K},\Gamma'}$; in other words we are defining ℓ' by declaring $(w_1 \otimes 1,\dots,w_n \otimes 1)$ to be an orthogonal ordered basis for (C',ℓ'). The following proposition might be read as saying that this definition is independent of the choice of orthogonal basis $(w_1,\dots,w_n)$ for (C,ℓ).

Proposition 2.21 *With the above definition, if $(x_1,\dots,x_n)$ is any orthogonal ordered basis for (C,ℓ) then $(x_1 \otimes 1,\dots,x_n \otimes 1)$ is an orthogonal ordered basis for (C',ℓ').*

Proof Let $(w_1,\dots,w_n)$ denote the orthogonal basis that was used to define ℓ'. Let $N \in \mathrm{GL}_n(\Lambda)$ be the basis change matrix from $(w_1,\dots,w_n)$ to $(x_1,\dots,x_n)$, ie the matrix characterized by the fact that for $j \in \{1,\dots,n\}$ we have $x_j = \sum_i N_{ij} w_i$. Then for $\vec{\lambda}' = (\lambda'_1,\dots,\lambda'_n) \in (\Lambda^{\mathcal{K},\Gamma'})^n$ we have

$$\ell'\left(\sum_{j=1}^{n} \lambda'_j x_j \otimes 1\right) = \ell\left(\sum_{i=1}^{n} (N\vec{\lambda}')_i w_i\right) = \max_i(\ell(w_i) - \nu((N\vec{\lambda}')_i)). \tag{8}$$

Now the vector $\vec{\lambda}' \in (\Lambda^{\mathcal{K},\Gamma'})^n$ is a formal sum $\vec{\lambda}' = \sum_{g\in\Gamma'} \vec{v}_g T^g$ where $\vec{v}_g \in \mathcal{K}^n$ and where the set of g with $\vec{v}_g \neq 0$ is discrete and bounded below. Let $S_{\vec{\lambda}'} \subset \Gamma'$ consist of those $g \in \Gamma'$ such that g is the minimal element in its coset $g + \Gamma \subset \Gamma'$ having

$\vec{v}_g \neq 0$. We can then reorganize the above sum as

$$\vec{\lambda}' = \sum_{g \in S_{\vec{\lambda}'}} \vec{\lambda}_g T^g,$$

where now $\vec{\lambda}_g \in \Lambda^n$, and where the set $S_{\vec{\lambda}'}$ is discrete and bounded below and has the property that distinct elements of $S_{\vec{\lambda}'}$ belong to distinct cosets of Γ in Γ'.

Now since N has its coefficients in Λ, we will have

$$N\vec{\lambda}' = \sum_{g \in S_{\vec{\lambda}'}} N\vec{\lambda}_g T^g,$$

where each $N\vec{\lambda}_g \in \Lambda^n$. For each i the various $\nu((N\vec{\lambda}_g)_i T^g)$ are equal to $g + \nu((N\vec{\lambda}_g)_i)$ and so belong to distinct cosets of Γ in Γ' (in particular, they are distinct from each other) and so we have for each i

$$\nu((N\vec{\lambda}')_i) = \min_{g \in S_{\vec{\lambda}'}} (g + \nu((N\vec{\lambda}_g)_i)),$$

and similarly $\nu(\lambda'_j) = \min_g(g + \nu((\vec{\lambda}_g)_j))$ for each j. Combining this with (8) and using the orthogonality of $(w_1, \dots, w_n)$ and $(x_1, \dots, x_n)$ with respect to ℓ and the fact that the $\vec{\lambda}_g$ belong to Λ^n gives

$$\begin{aligned}
\ell'\biggl(\sum_{j=1}^n \lambda'_j x_j \otimes 1\biggr) &= \max_{i,g}(\ell(w_i) - g - \nu((N\vec{\lambda}_g)_i)) \\
&= \max_g\Bigl(-g + \max_i(\ell(w_i) - \nu((N\vec{\lambda}_g)_i))\Bigr) \\
&= \max_g\biggl(-g + \ell\biggl(\sum_i (N\vec{\lambda}_g)_i w_i\biggr)\biggr) \\
&= \max_g\biggl(-g + \ell\biggl(\sum_j (\vec{\lambda}_g)_j x_j\biggr)\biggr) \\
&= \max_g\Bigl(-g + \max_j(\ell(x_j) - \nu((\vec{\lambda}_g)_j))\Bigr) \\
&= \max_j\Bigl(\ell(x_j) - \min_g(g + \nu((\vec{\lambda}_g)_j))\Bigr) \\
&= \max_j(\ell(x_j) - \nu(\lambda'_j)),
\end{aligned}$$

proving the orthogonality of $(x_1 \otimes 1, \dots, x_n \otimes 1)$ since it follows directly from the original definition of ℓ' in terms of $(w_1, \dots, w_n)$ that $\ell'(x \otimes 1) = \ell(x)$ whenever $x \in C$. □

3 (Nonarchimedean) singular value decompositions

Recall that in linear algebra over $\mathbb{C}$ with its standard inner product, a singular value decomposition for a linear transformation $A\colon \mathbb{C}^n \to \mathbb{C}^m$ is typically defined to be a factorization $A = X\Sigma Y^*$ where $X \in U(m)$, $Y \in U(n)$, and $\Sigma_{ij} = 0$ when $i \neq j$ while each $\Sigma_{ii} \geq 0$. The "singular values" of A are by definition the diagonal entries $\sigma_i = \Sigma_{ii}$, and then we have an orthonormal basis $(y_1, \dots, y_n)$ for $\mathbb{C}^n$ (given by the columns of Y) and an orthonormal basis $(x_1, \dots, x_m)$ for $\mathbb{C}^m$ (given by the columns of X) with $Ay_i = \sigma_i x_i$ for all i with $\sigma_i \neq 0$, and $Ay_i = 0$ otherwise.

An analogous construction for linear transformations between orthogonalizable Λ–spaces will play a central role in this paper. In the generality in which we are working, we should not ask for the bases $(y_1, \dots, y_n)$ to be ortho*normal*, since an orthogonalizable Λ–space may not even admit an orthonormal basis (for the examples $(\Lambda^n, -\vec{\nu}_{\vec{t}})$ of Example 2.12, an orthonormal basis exists if and only if each t_i belongs to the value group Γ). However in the classical case asking for a singular value decomposition is equivalent to asking for orthogonal bases $(y_1, \dots, y_n)$ for the domain and $(x_1, \dots, x_m)$ for the codomain such that for all i either $Ay_i = x_i$ or $Ay_i = 0$; the singular values could then be recovered as the numbers $\|Ay_i\|/\|y_i\|$. This is precisely what we will require in the nonarchimedean context. For the case in which the spaces in question do admit orthonormal bases (and so are equivalent to $(\Lambda^n, -\vec{\nu}))$ such a construction can be found in [27, Section 4.3].

3.1 Existence of (nonarchimedean) singular value decomposition

Definition 3.1 Let (C, ℓ_C), (D, ℓ_D) be orthogonalizable Λ–spaces and let $A\colon C \to D$ be a linear map with rank r. A *singular value decomposition of* A is a choice of *orthogonal* ordered bases $(y_1, \dots, y_n)$ for C and $(x_1, \dots, x_m)$ for D such that

(i) $(y_{r+1}, \dots, y_n)$ is an orthogonal ordered basis for $\ker A$;

(ii) $(x_1, \dots, x_r)$ is an orthogonal ordered basis for $\operatorname{Im} A$;

(iii) $Ay_i = x_i$ for $i \in \{1, \dots, r\}$;

(iv) $\ell_C(y_1) - \ell_D(x_1) \geq \dots \geq \ell_C(y_r) - \ell_D(x_r)$.

Remark 3.2 Consistently with the remarks at the start of the section, the singular values of A would then be the quantities $e^{\ell_D(x_i) - \ell_C(y_i)}$ for $1 \leq i \leq r$, as well as 0 if $r < n$. So the quantities $\ell_C(y_i) - \ell_D(x_i)$ from (iv) are the negative logarithms of the singular values.

Remark 3.3 Occasionally it will be useful to consider data $((y_1,\dots,y_n),(x_1,\dots,x_m))$ which satisfy all of the conditions of Definition 3.1 except condition (iv); such a pair will be called an *unsorted singular value decomposition.* Of course passing from an unsorted singular value decomposition to a genuine singular value decomposition is just a matter of sorting by the quantity $\ell_C(y_i)-\ell_C(x_i)$.

The rest of this subsection will be devoted to proving the following existence theorem:

Theorem 3.4 *If (C,ℓ_C) and (D,ℓ_D) are orthogonalizable Λ–spaces, then any Λ–linear map $A\colon C\to D$ has a singular value decomposition.*

We will prove Theorem 3.4 by providing an algorithm (with proof) for producing a singular value decomposition of linear map A between orthogonalizable Λ–spaces. The algorithm is essentially Gaussian elimination, but with a carefully designed rule for pivot selection which allows us to achieve the desired orthogonality properties. In this respect it is similar to the algorithm from [46] (that computes barcodes in classical persistent homology); however [46] uses a pivot-selection rule which does not adapt well to our context, where the value group Γ may be nontrivial, leading us to use a different such rule. Like the algorithm from [46], our algorithm requires a number of field operations that is at most cubic in the dimensions of the relevant vector spaces, and can be expected to do better than this in common situations where the matrix representing the linear map is sparse. Of course, when working over a Novikov field there is an additional concern regarding how one can implement arithmetic operations in this field on a computer; we do not attempt to address this here.

Theorem 3.5 (algorithmic version of Theorem 3.4) *Let (C,ℓ_C) and (D,ℓ_D) be orthogonalizable Λ–spaces, let $A\colon C\to D$ be a Λ–linear map, and let $(v_1,\dots,v_n)$ be an orthogonal ordered basis for C. Then one may algorithmically construct an orthogonal ordered basis $(v'_1,\dots,v'_n)$ of C such that*

- (i) *$\ell_C(v'_i)=\ell_C(v_i)$ and $\ell_D(Av'_i)\le\ell_D(Av_i)$ for each i;*
- (ii) *Let $\mathcal{U}=\{i\in\{1,\dots,n\}\mid Av'_i\neq 0\}$. Then the ordered subset $(Av'_i\mid i\in\mathcal{U})$ is orthogonal in D.*

Remark 3.6 In particular, $(v'_i\mid i\notin\mathcal{U})$ then gives an orthogonal ordered basis for $\ker A$.

Proof Fix throughout the algorithm an orthogonal ordered basis $(w_1,\dots,w_m)$ for D. Represent A by a matrix (A_{ij}) with respect to these bases, so that $Av_j=\sum_i A_{ij}w_i$. Note that v_j changes as the algorithm proceeds (though the w_i do not), so the elements

$A_{ij} \in \Lambda$ will likewise change in a corresponding way. Initialize the set of "unused column indices" to be $\mathcal{J} = \{1, \dots, n\}$, and the set of "pivot pairs" to be $\mathcal{P} = \varnothing$; at each step an element will be removed from $\mathcal{J}$ and an element will be added to $\mathcal{P}$. Here is the algorithm:

```
while $(\exists j \in \mathcal{J})(Av_j \neq 0)$ do
    Choose $i_0 \in \{1, \dots, m\}$ and $j_0 \in \mathcal{J}$ which maximize the quantity
        $\ell_D(w_i) - \nu(A_{ij}) - \ell_C(v_j)$ over all $(i,j) \in \{1 \dots, m\} \times \mathcal{J}$.
    Add $(i_0, j_0)$ to the set $\mathcal{P}$.
    Remove $j_0$ from the set $\mathcal{J}$.
    For each $j \in \mathcal{J}$, replace $v_j$ by $v_j' := v_j - \frac{A_{i_0 j}}{A_{i_0 j_0}} v_{j_0}$.
    For each $j \in \mathcal{J}$ and $i \in \{1, \dots, m\}$, replace $A_{ij}$ by $A_{ij}' := A_{ij} - \frac{A_{i_0 j} A_{i j_0}}{A_{i_0 j_0}}$ (thus
        restoring the property that $Av_j = \sum_{i=1}^m A_{ij} w_i$).
end
```

Note that the while loop predicate implies that in each iteration there is some $(i,j) \in \{1, \dots, m\} \times \mathcal{J}$ such that $A_{ij} \neq 0$, so in particular $A_{i_0 j_0} \neq 0$ (otherwise $A = 0$) and so the divisions by $A_{i_0 j_0}$ in the last two steps of the iteration are not problematic. The ordered basis $(v_1', \dots, v_n')$ promised in the statement of this theorem is then simply the tuple to which $(v_1, \dots, v_n)$ has evolved upon the termination of the while loop. To prove that this satisfies the required properties it suffices to prove that, in each iteration of the while loop, the following assertions hold:

Claim 3.7 *If the initial basis $(v_1, \dots, v_n)$ is orthogonal, then so is the basis obtained by replacing v_j by*

$$v_j' = v_j - \frac{A_{i_0 j}}{A_{i_0 j_0}} v_{j_0}$$

for each $j \in \mathcal{J} \setminus \{j_0\}$. Moreover $\ell_C(v_j') = \ell_C(v_j)$ while $\ell_D(Av_j') \le \ell_D(Av_j)$.

Claim 3.8 *After each iteration, the ordered set $(Av_j \mid j \notin \mathcal{J}) \subset D$ is orthogonal.*

Proof of Claim 3.7 For any $j \in \mathcal{J} \setminus \{j_0\}$, by the orthogonality of $(v_1, \dots, v_n)$ and the definition of v_j', we have

$$\ell_C(v_j') = \max\left\{\ell_C(v_j), \ell_C\left(\frac{A_{i_0 j}}{A_{i_0 j_0}} v_{j_0}\right)\right\}.$$

Because (i_0, j_0) is chosen to satisfy

$$\ell_D(w_{i_0}) - \nu(A_{i_0 j_0}) - \ell_C(v_{j_0}) \ge \ell_D(w_i) - \nu(A_{ij}) - \ell_C(v_j)$$

for all i and j, it in particular holds that

$$\ell_D(w_{i_0}) - \nu(A_{i_0 j_0}) - \ell_C(v_{j_0}) \geq \ell_D(w_{i_0}) - \nu(A_{i_0 j}) - \ell_C(v_j),$$

which can be rearranged to give

$$\ell_C\left(\frac{A_{i_0 j}}{A_{i_0 j_0}} v_{j_0}\right) \leq \ell_C(v_j). \tag{9}$$

So we get

$$\ell_C(v'_j) = \ell_C(v_j). \tag{10}$$

As for the statement about $\ell_D(Av'_j)$, note that

$$\ell_D(Av_{j_0}) = \ell_D\left(\sum_{i=1}^{m} A_{ij_0} w_i\right) = \max_i(\ell_D(w_i) - \nu(A_{ij_0})) = \ell_D(w_{i_0}) - \nu(A_{i_0 j_0}),$$

where the last equation follows from the optimality criterion satisfied by (i_0, j_0). Therefore,

$$\begin{aligned}\ell_D\left(\frac{A_{i_0 j}}{A_{i_0 j_0}} Av_{j_0}\right) &= \ell_D(w_{i_0}) - \nu(A_{i_0 j}) \leq \max_{1 \leq i \leq n} \ell_D(A_{ij} w_i) \\ &= \ell_D\left(\sum_{i=1}^{n} A_{ij} w_i\right) = \ell_D(Av_j)\end{aligned}$$

and hence

$$\ell_D(Av'_j) \leq \max\left\{\ell_D(Av_j), \ell_D\left(\frac{A_{i_0 j}}{A_{i_0 j_0}} Av_{j_0}\right)\right\} = \ell_D(Av_j).$$

It remains to prove orthogonality of the basis obtained by replacing the v_j by v'_j for $j \in \mathcal{J}$. Here and for the rest of the proof we use the variable values as they are after the third step of the given iteration of the while loop — thus the v_j have not been changed but j_0 has been removed from $\mathcal{J}$. The new basis will be $\{v'_1, \dots, v'_n\}$, where $v'_j = v_j$ if $j \notin \mathcal{J}$ and $v'_j = v_j - (A_{i_0 j}/A_{i_0 j_0}) v_{j_0}$ otherwise. Let $\lambda_1, \dots, \lambda_n \in \Lambda$ and observe that, by the orthogonality of $\{v_1, \dots, v_n\}$,

$$\begin{aligned}\ell_C\left(\sum_{j=1}^{n} \lambda_j v'_j\right) &= \ell_C\left(\sum_{j=1}^{n} \lambda_j v_j - \sum_{j \in \mathcal{J}} \lambda_j \frac{A_{i_0 j}}{A_{i_0 j_0}} v_{j_0}\right) \\ &= \max\left\{\ell_C\left(\left(\lambda_{j_0} - \sum_{k \in \mathcal{J}} \lambda_k \frac{A_{i_0 k}}{A_{i_0 j_0}}\right) v_{j_0}\right), \max_{j \neq j_0} \ell_C(\lambda_j v_j)\right\}.\end{aligned} \tag{11}$$

If $\ell(\lambda_{j_0}v'_{j_0}) > \ell(\lambda_j v'_j)$ for all $j \neq j_0$, then of course

$$\ell_C\left(\sum_{j=1}^n \lambda_j v'_j\right) = \ell_C(\lambda_{j_0}v'_{j_0}) = \max_j\{\ell_C(\lambda_j v'_j)\}.$$

Otherwise, there is $j_1 \neq j_0$ such that

$$\max_j \ell_C(\lambda_j v'_j) = \ell_C(\lambda_{j_1}v'_{j_1}). \tag{12}$$

Now by (10) and the optimality condition (12), we have

$$\ell_C(\lambda_{j_1}v_{j_1}) = \ell_C(\lambda_{j_1}v'_{j_1}) \geq \ell_C(\lambda_{j_0}v'_{j_0}) = \ell_C(\lambda_{j_0}v_{j_0}). \tag{13}$$

Also, by (9) and (12), for all $k \in \mathcal{J}$,

$$\ell_C(\lambda_{j_1}v_{j_1}) \geq \ell_C\left(\lambda_k \frac{A_{i_0k}}{A_{i_0j_0}} v_{j_0}\right).$$

Thus

$$\ell_C(\lambda_{j_1}v_{j_1}) \geq \ell_C\left(\left(\lambda_{j_0} - \sum_{k\in\mathcal{J}} \lambda_k \frac{A_{i_0k}}{A_{i_0j_0}}\right) v_{j_0}\right). \tag{14}$$

So combining (11), (12), and (14), we have

$$\ell_C\left(\sum_{j=1}^n \lambda_j v'_j\right) = \max_j \ell_C(\lambda_j v'_j),$$

proving the orthogonality of $(v'_1, \dots, v'_n)$. This completes the proof of Claim 3.7. □

Proof of Claim 3.8 For $k \geq 1$ let (i_k, j_k) denote the pivot pair that is added to the set $\mathcal{P}$ during the k^{th} iteration of the while loop. In particular j_k is removed from $\mathcal{J}$ during the k^{th} iteration, and after this removal we have $\mathcal{J} = \{1, \dots, n\} \setminus \{j_1, \dots, j_k\}$. So the column operation in the last step of the k^{th} iteration replaces the matrix entries A_{i_kj} for $j \notin \{j_1, \dots, j_k\}$ by

$$A_{i_kj} - \frac{A_{i_kj}A_{i_kj_k}}{A_{i_kj_k}} = 0.$$

Moreover for $j \notin \{j_1, \dots, j_k\}$ and any $i \in \{1, \dots, m\}$ such that after the prior iteration we had $A_{ij_k} = A_{ij} = 0$ (for instance this applies, inductively, to any $i \in \{i_1, \dots, i_{k-1}\}$), the fact that $A_{ij} = 0$ will be preserved after the k^{th} iteration. Thus,

(15) after the k^{th} iteration, $A_{i_lj} = 0$ for $l \in \{1, \dots, k\}$ and $j \notin \{j_1, \dots, j_l\}$.

We now show that, after the k^{th} iteration, the ordered set $(Av_{j_1}, \dots, Av_{j_k})$ is orthogonal; this is evidently equivalent to the statement of the claim. Note that, for $1 \leq l \leq k$, neither the element v_{j_l} nor the j_l^{th} column of the matrix (A_{ij}) changes during or after the l^{th} iteration of the while loop, due to the removal of j_l from $\mathcal{J}$ during that iteration. For $l \in \{1,\dots,k\}$, the optimality condition satisfied by the pair (i_l, j_l) guarantees that $\ell_D(w_i) - \nu(A_{ij_l}) \le \ell_D(w_{i_l}) - \nu(A_{i_l j_l})$ for all i and hence

$$\ell_D(Av_{j_l}) = \max_i(\ell_D(A_{ij_l} w_i)) = \ell_D(A_{i_l j_l} w_{i_l}). \tag{16}$$

Given $\lambda_1,\dots,\lambda_k \in \Lambda$ we shall show that $\ell_D\bigl(\sum_{l=1}^k \lambda_l Av_{j_l}\bigr) = \max_l \ell_D(\lambda_l Av_{j_l})$. Let l_0 be the *smallest* element of $\{1,\dots,k\}$ with the property that

$$\ell_D(\lambda_{l_0} A_{i_{l_0} j_{l_0}} w_{i_{l_0}}) = \max_{1\le l\le k} \ell_D(\lambda_l A_{i_l j_l} w_{i_l}).$$

For all $i \in \{1,\dots,m\}$ and $l \in \{1,\dots,k\}$ we have, by the choice of (i_l, j_l),

$$\ell_D(\lambda_l A_{ij_l} w_i) \le \ell_D(\lambda_l A_{i_l j_l} w_{i_l}) \le \ell_D(\lambda_{l_0} A_{i_{l_0} j_{l_0}} w_{i_{l_0}}).$$

Using (15), $A_{i_{l_0} j_l} \neq 0$ only for $l \le l_0$, and so

$$\sum_l \lambda_l A_{i_{l_0} j_l} w_{i_{l_0}} = \lambda_{l_0} A_{i_{l_0} j_{l_0}} w_{i_{l_0}} + \sum_{l<l_0} \lambda_l A_{i_{l_0} j_l} w_{i_{l_0}}.$$

Each term $\lambda_l A_{i_{l_0} j_l} w_{i_{l_0}}$ has filtration level bounded above by $\ell_D(\lambda_l A_{i_l j_l} w_{i_l})$ by the second equality in (16), and this latter filtration level is, for $l < l_0$, *strictly lower than* $\ell_D(\lambda_{l_0} A_{i_{l_0} j_{l_0}} w_{i_{l_0}})$ because we chose l_0 as the smallest maximizer of $\ell_D(\lambda_l A_{i_l j_l} w_{i_l})$. So we in fact have

$$\ell_D\biggl(\sum_l \lambda_l A_{i_{l_0} j_l} w_{i_{l_0}}\biggr) = \ell_D(\lambda_{l_0} A_{i_{l_0} j_{l_0}} w_{i_{l_0}}).$$

By the orthogonality of the ordered basis $(w_1,\dots,w_m)$ we therefore have

$$\begin{aligned}\ell_D\biggl(\sum_{l=1}^k \lambda_l Av_{j_l}\biggr) &= \ell_D\biggl(\sum_{l=1}^k\sum_{i=1}^m \lambda_l A_{ij_l} w_i\biggr)\\ &= \max_{1\le i\le m} \ell_D\biggl(\sum_{l=1}^k \lambda_l A_{ij_l} w_i\biggr) \ge \ell_D(\lambda_{l_0} A_{i_{l_0} j_{l_0}} w_{i_{l_0}})\\ &= \max_l \ell_D(\lambda_l A_{i_l j_l} w_{i_l}) = \max_l \ell_D(\lambda_l Av_{j_l}),\end{aligned}$$

where in the first equality in the third line we use the defining property of l_0 and in the last equality we use (16). Since the reverse inequality

$$\ell_D\biggl(\sum_l \lambda_l Av_{j_l}\biggr) \le \max_l \ell_D(\lambda_l Av_{j_l})$$

is trivial this completes the proof of the orthogonality of $(Av_{j_1},\dots,Av_{j_k})$. □

As noted earlier, Claims 3.7 and 3.8 directly imply that the basis for C obtained at the termination of the while loop satisfies the required properties, thus completing the proof of Theorem 3.5. □

Proof of Theorem 3.4 First reorder the elements v_i' produced by the Theorem 3.5 so that $Av_i' \neq 0$ if and only if $i \in \{1, \dots, r\}$, where r is the rank of A, and such that

$$\ell_C(v_1') - \ell_D(Av_1') \geq \cdots \geq \ell_C(v_r') - \ell_D(Av_r').$$

If A is surjective, then $((v_1', \dots, v_n'), (Av_1', \dots, Av_r'))$ will immediately be a singular value decomposition for A. More generally, we may use Corollary 2.19 to find an orthogonal complement of $\mathrm{Im}(A)$ in D, and by Corollary 2.17 this orthogonal complement has some orthogonal ordered basis $(x_{r+1}, \dots, x_m)$. We thus conclude that $((v_1', \dots, v_n'), (Av_1', \dots, Av_r', x_{r+1}, \dots, x_m))$ is a singular value decomposition for A. □

3.2 Duality and coefficient extension for singular value decompositions

Proposition 2.20 allows us to easily convert a singular value decomposition for a map $A\colon C \to D$ to one for the adjoint map $A^*\colon D^* \to C^*$. Explicitly:

Proposition 3.9 *Let (C, ℓ_C) and (D, ℓ_D) be two orthogonalizable Λ–spaces and $A\colon C \to D$ be a Λ–linear map with rank r. Suppose $((y_1, \dots, y_n), (x_1, \dots, x_m))$ is a singular value decomposition for A. Then $((x_1^*, \dots, x_m^*), (y_1^*, \dots, y_n^*))$ is a singular value decomposition for its adjoint map $A^*\colon D^* \to C^*$.*

Proof By the first assertion of Proposition 2.20, $(x_1^*, \dots, x_m^*)$ is an orthogonal ordered basis for D^* and $(y_1^*, \dots, y_n^*)$ is an orthogonal ordered basis for C^*. By the definition of a singular value decomposition, $Ay_i = x_i$ for $i \in \{1, \dots, r\}$ and $Ay_i = 0$ for $i \in \{r+1, \dots, n\}$, so $A^*x_i^* = y_i^*$ for $i \in \{1, \dots, r\}$ and $A^*x_i^* = 0$ for $i \in \{r+1, \dots, m\}$. Therefore $(x_{r+1}^*, \dots, x_m^*)$ is an orthogonal ordered basis for $\ker A^*$ and $\{y_1, \dots, y_r\} = \{A^*x_1^*, \dots, A^*x_r^*\}$ is an orthogonal ordered basis for $\mathrm{Im}\, A^*$. Finally, for $i \in \{1, \dots, r\}$, by the second assertion of Proposition 2.20, we have

$$\ell_{D^*}^*(x_i^*) - \ell_{C^*}^*(y_i^*) = -\ell_D(x_i) + \ell_C(y_i) = \ell_C(y_i) - \ell_D(x_i).$$

So the ordering of $\ell_C(y_i) - \ell_D(x_i)$ implies the desired ordering for $\ell_{D^*}^*(x_i^*) - \ell_{C^*}^*(y_i^*)$. □

Similarly, Proposition 2.21 implies that singular value decompositions are well-behaved under coefficient extension.

Proposition 3.10 *Consider two subgroups $\Gamma \le \Gamma' \le \mathbb{R}$, and write $\Lambda = \Lambda^{\mathcal{K},\Gamma}$ and $\Lambda' = \Lambda^{\mathcal{K},\Gamma'}$. Let (C, ℓ_C) and (D, ℓ_D) be orthogonalizable Λ–spaces and let $A\colon C \to D$ be a Λ–linear map, with singular value decomposition $((y_1, \dots, y_n), (x_1, \dots, x_m))$. Then if $C \otimes_\Lambda \Lambda'$ and $D \otimes_\Lambda \Lambda'$ are endowed with the filtration functions ℓ'_C and ℓ'_D as in Section 2.5, the map $A \otimes 1\colon C \otimes_\Lambda \Lambda' \to D \otimes_\Lambda \Lambda'$ has singular value decomposition given by $((y_1 \otimes 1, \dots, y_n \otimes 1), (x_1 \otimes 1, \dots, x_m \otimes 1))$.*

Proof The ordered sets $(y_1 \otimes 1, \dots, y_n \otimes 1)$ and $(x_1 \otimes 1, \dots, x_m \otimes 1)$ are orthogonal by Proposition 2.21. Moreover by definition of the relevant filtration functions we have $\ell'_C(y_i \otimes 1) = \ell_C(y_i)$ and $\ell'_D(x_i \otimes 1) = \ell_D(x_i)$ for all i such that these are defined. Once these facts are known it is a trivial matter to check each of the conditions (i)-(iv) in the definition of a singular value decomposition. □

4 Boundary depth and torsion exponents via singular value decompositions

The *boundary depth* as defined in [44] or [45] is a numerical invariant of a filtered chain complex that, in the case of the Hamiltonian and Lagrangian Floer complexes, has been effectively used to obtain applications in symplectic topology. A closely related notion is that of the *torsion threshold* and more generally the *torsion exponents* that were introduced in [20, Section 6.1] for the Lagrangian Floer complex over the universal Novikov ring and were used in [21] to obtain lower bounds for the displacement energies of polydisks. We will see in this section that, for complexes like those that arise in Floer theory, both of these notions are naturally encoded in the (nonarchimedean) singular value decomposition of the boundary operator of the chain complex. In particular our discussion will show that the boundary depth coincides with the torsion threshold when both are defined, and that certain natural generalizations of the boundary depth likewise coincide with the rest of the torsion exponents. This implies new restrictions on the values that the torsion exponents can take. Our generalized boundary depths will be part of the data that comprise the concise barcode of a Floer-type complex, our main invariant to be introduced in Section 6.

For the rest of the paper, we will always work with what we call a *Floer-type complex* over a Novikov field Λ, defined as follows:

Definition 4.1 A *Floer-type complex* $(C_*, \partial_C, \ell_C)$ over a Novikov field $\Lambda = \Lambda^{\mathcal{K},\Gamma}$ is a chain complex $\left(C_* = \bigoplus_{k \in \mathbb{Z}} C_k, \partial_C\right)$ over Λ together with a function $\ell_C\colon C_* \to \mathbb{R} \cup \{-\infty\}$ such that each $(C_k, \ell|_{C_k})$ is an orthogonalizable Λ–space, and for each $x \in C_k$ we have $\partial_C x \in C_{k-1}$ with $\ell_C(\partial_C x) \le \ell_C(x)$.

Example 4.2 According to Example 2.12, the Morse, Novikov, and Hamiltonian Floer chain complexes are all Floer-type complexes. In each case the boundary operator is defined by counting connecting trajectories between two critical points for some function, which satisfy a certain differential equation (see eg [38, Section 1.5] for the Hamiltonian Floer case).

Remark 4.3 In fact in many Floer-type complexes including the Morse, Novikov, and Hamiltonian Floer complexes one has the strict inequality $\ell_C(\partial_C x) < \ell_C(x)$. However it is also often useful in Morse and Floer theory to consider complexes where the inequality is not necessarily strict; for instance the Biran–Cornea pearl complex [4] with appropriate coefficients can be described in this way, as can the Morse–Bott complex built from moduli spaces of "cascades" in [19, Appendix A]. Also our definition allows other, non-Floer-theoretic, constructions such as the Rips complex (see Example 2.4), and the mapping cylinders which play a crucial role in the proofs of Theorem B and Theorem 1.4, to be described as Floer-type complexes, whereas requiring $\ell_C(\partial_C x) < \ell_C(x)$ would rule these out. In the case that one does have a strict inequality for the effect of the boundary operator on the filtration, the verbose and concise barcodes that we define later are easily seen to be equal to each other.

Definition 4.4 Given two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$, a *filtered chain isomorphism* between these two complexes is a chain isomorphism $\Phi\colon C_* \to D_*$ such that $\ell_D(\Phi(x)) = \ell_C(x)$ for all $x \in C_*$.

Definition 4.5 Given two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$, two chain maps Φ, $\Psi\colon C_* \to D_*$ are called *filtered chain homotopic* if there exists $K\colon C_* \to D_{*+1}$ such that $\Phi - \Psi = \partial_D K + K\partial_D$ and K *preserves* filtration, ie $\ell_D(K(x)) \le \ell_C(x)$ for all x, and both Φ and Ψ preserve filtration as well.

We say that $(C_*, \partial_C, \ell_C)$ is *filtered homotopy equivalent* to $(D_*, \partial_D, \ell_D)$ if there exist chain maps $\Phi\colon C_* \to D_*$ and $\Psi\colon D_* \to C_*$ which both preserve filtration such that $\Psi \circ \Phi$ is filtered chain homotopic to identity I_C while $\Phi \circ \Psi$ is filtered chain homotopic to the I_D.

In order to cut down on the number of indices that appear in our formulas, we will sometimes work in the following setting:

Definition 4.6 A *two-term Floer-type complex* $(C_1 \xrightarrow{\partial} C_0)$ is a Floer-type complex of the form

$$\cdots \to 0 \to C_1 \xrightarrow{\partial} C_0 \to 0 \to \cdots .$$

Given any Floer-type complex $(C_*, \partial_C, \ell_C)$, fixing a degree k, we can consider the two-term Floer-type complex

$$(\widetilde{C}_1^{(k)} \xrightarrow{\partial|_{C_k}} \widetilde{C}_0^{(k)}),$$

where $\widetilde{C}_1^{(k)} = C_k$ and $\widetilde{C}_0^{(k)} = \ker(\partial|_{C_{k-1}})(\leq C_{k-1})$.

For the rest of this section, we will focus mainly on two-term Floer-type complexes; consistently with the above discussion this roughly corresponds to focusing on a given degree in one of the multiterm chain complexes that we are ultimately interested in. For a two-term Floer-type complex $(C_1 \xrightarrow{\partial} C_0)$, by Theorem 3.4 we may fix a singular value decomposition $((y_1, \dots, y_n), (x_1, \dots, x_m))$ for the boundary map $\partial\colon C_1 \to C_0$. Denote the rank of ∂ by r. We will see soon that the numbers $\{\ell(y_i) - \ell(x_i)\}$ for $i \in \{1, \dots, r\}$ (which have earlier been described as the negative logarithms of the singular values of ∂) can be characterized in terms of the following notion of *robustness* of the boundary operator.

Definition 4.7 Let $\delta \in \mathbb{R}$. An element $x \in C_0$ is said to be *δ–robust* if for all $y \in C_1$ such that $\partial y = x$ it holds that $\ell(y) > \ell(x) + \delta$. A subspace $V \leq C_0$ is said to be δ–robust if every $x \in V \setminus \{0\}$ is δ–robust.

Example 4.8 When $(C_1 \xrightarrow{\partial} C_0)$ is the two-term Floer-type complex $\widetilde{\mathrm{CM}}_*^{(k)}(f)$ induced by the degree-k and degree-$(k-1)$ parts of the Morse complex $\mathrm{CM}_*(f)$ of a Morse function on a compact manifold, the reader may verify that each nonzero element of C_0 is δ–robust for all $\delta < \delta_k$, where δ_k is the minimal positive difference between a critical value of an index-k critical point and a critical value of an index-$(k-1)$ critical point. Because a strict inequality is required in the definition of robustness, there may be elements of C_0 which are not δ_k–robust.

In the presence of our singular value decomposition $((y_1, \dots, y_n), (x_1, \dots, x_m))$, the following simple observation is useful for checking δ–robustness:

Lemma 4.9 *Let $x = \sum_{i=1}^r \lambda_i x_i$ be any element of* $\operatorname{Im}\partial$, *and suppose $y \in C_1$ obeys $\partial y = x$. Then*

$$\ell(y) \geq \ell\left(\sum_{i=1}^r \lambda_i y_i\right) = \max\{\ell(y_i) - \nu(\lambda_i) \mid 1 \leq i \leq r\}.$$

Proof Since $\partial y_i = x_i$ for $1 \leq i \leq r$ and $\partial y_i = 0$ for $i > r$, and since the x_i are linearly independent, the elements $y \in C_1$ such that $\partial y = x$ are precisely those of the form $\sum_{i=1}^{r}\lambda_i y_i+\sum_{i=r+1}^{n}\mu_i y_i$ for arbitrary $\mu_{r+1},\dots,\mu_n\in\Lambda$. The proposition then follows directly from the fact that $(y_1,\dots,y_n)$ is an orthogonal ordered basis for C_1. □

Definition 4.10 Given a two-term chain complex $(C_1\xrightarrow{\partial}C_0)$ and a positive integer k, let

$$\beta_k(\partial)=\sup\bigl(\{0\}\cup\{\delta\geq 0\mid\exists\ \delta\text{–robust subspace } V\leq\operatorname{Im}\partial \text{ with } \dim(V)=k\}\bigr).$$

Note that $\beta_k(\partial)=0$ if ∂ is the zero map or if $k>\dim(\operatorname{Im}\partial)$. It is easy to see that, when $k\leq\dim(\operatorname{Im}\partial)$, $\beta_k(\partial)$ can be rephrased as

$$\beta_k(\partial)=\sup_{\substack{V\leq\operatorname{Im}\partial\\ \dim(V)=k}}\ \inf_{x\in V\setminus\{0\}}\{\ell(y)-\ell(x)\mid\partial y=x\}.$$

When $k=1$, this is exactly the definition of boundary depth in [45] (see [45, (24)]), and so we can view the $\beta_k(\partial)$ as generalizations of the boundary depth. Clearly one has

$$\beta_1(\partial)\geq\beta_2(\partial)\geq\cdots\geq\beta_k(\partial)\geq 0$$

for all k. We will prove the following theorem which relates the $\beta_k(\partial)$ to singular value decompositions.

Theorem 4.11 *Given a singular value decomposition $((y_1,\dots,y_n),(x_1,\dots,x_m))$ for a two-term chain complex $(C_1\xrightarrow{\partial}C_0)$, the numbers $\beta_k(\partial)$ are given by*

$$\beta_k(\partial)=\begin{cases}\ell(y_k)-\ell(x_k) & \text{if } 1\leq k\leq r,\\ 0 & \text{if } k>r,\end{cases}$$

where r is the rank of ∂.

Proof For each $k\in\{1,\dots,r\}$, we will show that there exists a k–dimensional δ–robust subspace of $\operatorname{Im}\partial$ for any $\delta<\ell(y_k)-\ell(x_k)$, but that no k–dimensional subspace is $(\ell(y_k)-\ell(x_k))$–robust. This clearly implies the result by the definition of $\beta_k(\partial)$.

Considering the subspace $V_k=\operatorname{span}_\Lambda\{x_1,\dots,x_k\}$, let $x=\sum_{i=1}^{k}\lambda_i x_i$ be any nonzero element in V_k. Let $i_0\in\{1,\dots,k\}$ maximize the quantity $\ell(x_i)-\nu(\lambda_i)$ over all $i\in\{1,\dots,k\}$, so that by the orthogonality of the x_i we have $\ell(x)=\ell(x_{i_0})-\nu(\lambda_{i_0})$. Then, using the orthogonality of the y_i,

$$\begin{aligned}\ell\Bigl(\sum_{i=1}^{k}\lambda_i y_i\Bigr)-\ell(x)&=\max_i(\ell(y_i)-\nu(\lambda_i))-(\ell(x_{i_0})-\nu(\lambda_{i_0}))\\&\geq(\ell(y_{i_0})-\nu(\lambda_{i_0}))-(\ell(x_{i_0})-\nu(\lambda_{i_0}))=\ell(y_{i_0})-\ell(x_{i_0})\\&\geq\ell(y_k)-\ell(x_k),\end{aligned}$$

where the last inequality follows from our ordering convention for the x_i. But then by Lemma 4.9, it follows that whenever $\partial y = x$ we have $\ell(y) - \ell(x) \geq \ell(y_k) - \ell(x_k)$. Since this holds for an arbitrary element $x \in \mathrm{span}_\Lambda\{x_1, \dots, x_k\} \setminus \{0\}$ we obtain that $\mathrm{span}\{x_1, \dots, x_k\}$ is δ–robust for all $\delta < \ell(y_k) - \ell(x_k)$.

Next, for any k–dimensional subspace $V \leq \mathrm{Im}\,\partial$, let $W = \mathrm{span}_\Lambda\{x_k, x_{k+1}, \dots, x_r\}$. Since W has codimension $k-1$ in $\mathrm{Im}\,\partial$, the intersection $V \cap W$ contains some nonzero element x. Since $x \in W$ we can write $x = \sum_{i=k}^{r} \lambda_i x_i$ where not all λ_i are zero. Choose $i_0 \in \{k, \dots, r\}$ to maximize the quantity $\ell(y_i) - \nu(\lambda_i)$ over $i \in \{k, \dots, r\}$. Let $y = \sum_{i=k}^{r} \lambda_i y_i$. Then we have $\partial y = x$, and

$$\begin{aligned}\ell(y) - \ell(x) &= (\ell(y_{i_0}) - \nu(\lambda_{i_0})) - \max_i(\ell(x_i) - \nu(\lambda_i))\\ &\leq (\ell(y_{i_0}) - \nu(\lambda_{i_0})) - (\ell(x_{i_0}) - \nu(\lambda_{i_0}))\\ &= \ell(y_{i_0}) - \ell(x_{i_0}) \leq \ell(y_k) - \ell(x_k)\end{aligned}$$

by our ordering convention for the x_i. So since $x \in V \setminus \{0\}$ (and since the inequality required in the definition of δ–robustness is strict) this proves that V is not $(\ell(y_k) - \ell(x_k))$–robust.

Finally, when $k > r$, there is no $V \leq \mathrm{Im}\,\partial$ such that $\dim(V) = k$ (since $\dim(\mathrm{Im}\,\partial) = r$). Then by definition of $\beta_k(\partial)$, it is zero. □

Note that Definition 4.10 makes clear that $\beta_k(\partial)$ is independent of the choice of singular value decomposition; thus we deduce the nonobvious fact that the difference $\ell(y_k) - \ell(x_k)$ is likewise independent of the choice of singular value decomposition for each $k \in \{1, \dots, r\}$. Note also that any filtration-preserving Λ–linear map A between two orthogonalizable Λ–spaces C and D can just as well be viewed as a two-term chain complex $(C \xrightarrow{A} D)$, and so we obtain generalized boundary depths $\beta_k(A)$. Theorem 3.4 or Theorem 3.5 provides a systematic way to compute $\beta_k(A)$. It is also clear from the definition that if $A\colon C \to D$ has image contained in some subspace $D' \leq D$ then $\beta_k(A)$ is the same regardless of whether we regard A as a map $C \to D$ or as a map $C \to D'$. For instance if $(C_*, \partial_C, \ell_C)$ is a Floer-type complex, for any i we could consider either of the two-term complexes

$$(C_i \xrightarrow{\partial|_{C_i}} C_{i-1}) \quad \text{or} \quad (C_i \xrightarrow{\partial|_{C_i}} \ker(\partial_C|_{C_{i-1}}))$$

and obtain the same values of β_k.

We conclude this section by phrasing the torsion exponents of [20; 21] in our terms and proving that these torsion exponents coincide with our generalized boundary depths β_k. We will explain this just for two-term Floer-type complexes $(C_1 \xrightarrow{\partial} C_0)$; this represents no loss of generality, as for a general Floer-type complex $(C_*, \partial_C, \ell_C)$ one may apply the discussion below to the various two-term Floer-type complexes

$$(C_{i+1} \xrightarrow{\partial|_{C_{i+1}}} \ker(\partial_C|_{C_i}))$$

in order to relate the torsion exponents and generalized boundary depths in any degree $i \in \mathbb{Z}$.

So let $(C_1 \xrightarrow{\partial} C_0)$ be a two-term Floer-type complex over $\Lambda = \Lambda^{\mathcal{K},\Gamma}$. We first define the torsion exponents (in degree zero) in our language, leaving it to readers familiar with [20] to verify that our definition is consistent with theirs. Write $\Lambda^{\text{univ}} = \Lambda^{\mathcal{K},\mathbb{R}}$ for the "universal" Novikov field, so named because regardless of the choice of Γ we have a field extension $\Lambda^{\mathcal{K},\Gamma} \hookrightarrow \Lambda^{\text{univ}}$. Also define

$$\Lambda_0^{\text{univ}} = \{\lambda \in \Lambda^{\text{univ}} \mid \nu(\lambda) \geq 0\};$$

thus Λ_0^{univ} is the subring of Λ^{univ} consisting of formal sums $\sum_g a_g T^g$ with each $g \geq 0$.

As in Section 2.5, for $j = 0, 1$ let $C_j' = C_j \otimes_\Lambda \Lambda^{\text{univ}}$, and endow C_j' with the filtration function obtained by choosing an orthogonal ordered basis $(w_1, \dots, w_a)$ for C_j and putting $\ell'\big(\sum_i \lambda_i' w_i \otimes 1\big) = \max_i(\ell(w_i) - \nu(\lambda_i'))$ for any $\lambda_1', \dots, \lambda_a' \in \Lambda^{\text{univ}}$. By Proposition 2.21 this definition is independent of the choice of orthogonal basis $(w_1, \dots, w_a)$.

Now, for $j = 0, 1$, define

$$\bar{C}_j' = \{c \in C_j' \mid \ell'(c) \leq 0\}$$

and observe that $\bar{C}_j$ is a module over the subring Λ_0^{univ} of Λ^{univ}. Moreover, again taking Proposition 2.21 into account, it is easy to see that if $(w_1, \dots, w_a)$ is *any* orthogonal ordered basis for C_j, then the elements $\bar{w}_i = w_i \otimes T^{\ell(w_i)}$ form a basis for $\bar{C}_j'$ as a Λ_0^{univ}–module.

The fact that $\ell(\partial c) \leq \ell(c)$ implies that the coefficient extension $\partial \otimes 1\colon C_1' \to C_0'$ restricts to $\bar{C}_1'$ as a map to $\bar{C}_0'$. So we have a (two-term) chain complex of Λ_0^{univ}–modules

$$(\bar{C}_1' \xrightarrow{\partial \otimes 1} \bar{C}_0').$$

Fukaya, Oh, Ohta, and Ono show [20, Theorem 6.1.20] that the zeroth homology of this complex (ie the quotient $\bar{C}_0'/(\partial \otimes 1)\bar{C}_1'$) is isomorphic to

$$(\Lambda_0^{\text{univ}})^q \oplus \bigoplus_{k=1}^{s} (\Lambda_0^{\text{univ}}/T^{\lambda_k}\Lambda_0^{\text{univ}}) \tag{17}$$

for some natural numbers q, s and positive real numbers $\lambda_i, \dots, \lambda_s$.

Definition 4.12 [20] Order the summands in the decomposition (17) of $\bar{C}_0'/(\partial\otimes 1)\bar{C}_1'$ so that $\lambda_1 \geq \cdots \geq \lambda_s$. For a positive integer k, the k^{th} *torsion exponent* of the two-term Floer-type complex $(C_1 \xrightarrow{\partial} C_0)$ is λ_k if $k \leq s$ and 0 otherwise. The first torsion exponent is also called the *torsion threshold.*

Theorem 4.13 *For each positive integer k the k^{th} torsion exponent of $(C_1 \xrightarrow{\partial} C_0)$ is equal to the generalized boundary depth $\beta_k(\partial)$.*

Proof Let $((y_1, \dots, y_n), (x_1, \dots, x_m))$ be a singular value decomposition for the map $\partial\colon C_1 \to C_0$. By Proposition 3.10, $((y_1 \otimes 1, \dots, y_n \otimes 1), (x_1 \otimes 1, \dots, x_m \otimes 1))$ is a singular value decomposition for $\partial \otimes 1\colon C_1' \to C_0'$. Let r denote the rank of ∂ (equivalently, that of $\partial \otimes 1$).

Let us determine the image $(\partial \otimes 1)(\bar{C}_1') \subset C_0'$. A general element x of C_0' can be written as $x = \sum_{i=1}^m \lambda_i x_i \otimes 1$, where $\lambda_i \in \Lambda^{\text{univ}}$. By the definition of a singular value decomposition, in order for x to be in the image of $\partial \otimes 1$ we evidently must have $\lambda_i = 0$ for $i > r$. Given that this holds, we will have $(\partial \otimes 1)\big(\sum_{i=1}^r \lambda_i y_i \otimes 1\big) = x$, and moreover by Lemma 4.9, $\sum_{i=1}^r \lambda_i y_i \otimes 1$ has the lowest filtration level among all preimages of x under $\partial \otimes 1$. Now

$$\ell'\left(\sum_{i=1}^r \lambda_i y_i \otimes 1\right) = \max_i(\ell(y_i) - \nu(\lambda_i)),$$

so we conclude that $x = \sum_{i=1}^m \lambda_i x_i \otimes 1$ belongs to $(\partial \otimes 1)(\bar{C}_1')$ if and only if both $\lambda_i = 0$ for $i > r$ and $\nu(\lambda_i) \geq \ell(y_i)$ for $i = 1, \dots, r$.

Recall that the elements $\bar{x_i} = x_i \otimes T^{\ell(x_i)}$ form a Λ_0^{univ}–basis for $\bar{C}_0'$. Letting $\mu_i = T^{-\ell(x_i)}\lambda_i$, the conclusion of the above paragraph can be rephrased as saying that $(\partial \otimes 1)(\bar{C}_1')$ consists precisely of elements $\sum_{i=1}^m \mu_i \bar{x}_i$ such that $\mu_i = 0$ for $i > r$ and $\nu(\mu_i) \geq \ell(y_i) - \ell(x_i)$ for $i = 1, \dots, r$. Now for any $\mu \in \Lambda^{\text{univ}}$ and $c \in \mathbb{R}$, one has $\nu(\mu) \geq c$ if and only if $\mu \in T^c \Lambda_0^{\text{univ}}$. So we conclude that

$$(\partial \otimes 1)(\bar{C}_1') = \operatorname{span}_{\Lambda_0^{\text{univ}}}\{T^{\ell(y_1)-\ell(x_1)}\bar{x}_1, \dots, T^{\ell(y_r)-\ell(x_r)}\bar{x}_r\},$$

while as mentioned earlier

$$\bar{C}_0' = \operatorname{span}_{\Lambda_0^{\text{univ}}}\{\bar{x}_1, \dots, \bar{x}_m\}.$$

These facts immediately imply that

$$\frac{\bar{C}_0'}{(\partial \otimes 1)(\bar{C}_1')} = (\Lambda_0^{\text{univ}})^{m-r} \oplus \bigoplus_{k=1}^r (\Lambda_0^{\text{univ}}/T^{\ell(y_k)-\ell(x_k)}\Lambda_0^{\text{univ}}).$$

Comparing with (17) we see that the numbers that we have denoted by s and r are equal to each other, and that the k^{th} torsion exponent is equal to $\ell(y_k)-\ell(x_k)$ for $1\le k\le r$ and to zero otherwise. By Theorem 4.11 this is the same as $\beta_k(\partial)$. □

5 Filtration spectrum

The filtration spectrum of an orthogonalizable Λ–space is an algebraic abstraction of the set of critical values of a Morse function or the action spectrum of a Hamiltonian diffeomorphism (see [40]).

In the definition below and elsewhere, our convention is that $\mathbb{N}$ is the set of nonnegative integers (so includes zero).

Definition 5.1 A *multiset* M is a pair (S,μ), where S is a set and $\mu\colon S\to\mathbb{N}\cup\{\infty\}$ is a function, called the *multiplicity function* of M. If T is some other set, a *multiset of elements of* T is a multiset (S,μ) such that $S\subset T$.

For $s\in S$, the value $\mu(s)$ should be interpreted as "the number of times that s appears" in the multiset M. By abuse of notation we will sometimes denote multisets in set-theoretic notation with elements repeated: for instance $\{1,3,1,2,3\}$ denotes a multiset with $\mu(1)=\mu(3)=2$ and $\mu(2)=1$. The cardinality of the multiset (S,μ) is by definition $\sum_{s\in S}\mu(S)$. (For notational simplicity we are not distinguishing between different infinite cardinalities in our definition; in fact, for nearly all of the multisets that appear in this paper the multiplicity function will only take finite values.)

Also, if $S\subset T$ and $\mu\colon T\to\mathbb{N}\cup\{\infty\}$ is a function with $\mu|_{T\setminus S}\equiv 0$ then we will not distinguish between the multisets (T,μ) and $(S,\mu|_S)$.

Definition 5.2 Let (C,ℓ) be an orthogonalizable Λ–space with a fixed orthogonal ordered basis $(v_1,\dots,v_n)$. The *filtration spectrum* of (C,ℓ) is the multiset $(\mathbb{R}/\Gamma,\mu)$, where

$$\mu(s)=\#\{v_i\in\{v_1,\dots,v_n\}\mid \ell(v_i)\equiv s \mod \Gamma\}.$$

Remark 5.3 When Γ is trivial, the filtration spectrum is just the set $\{\ell(v_1),\dots,\ell(v_n)\}$ and multiplicity function is just defined by setting $\mu(s)$ equal to the number of i such that $\ell(v_i)=s$.

Example 5.4 Let $\Gamma=\mathbb{Z}$ and $C=\operatorname{span}_\Lambda\{v_1,v_2\}$, where v_1,v_2 are orthogonal with $\ell(v_1)=2.5$ and $\ell(v_2)=0.5$. Then for $[0.5]\in\mathbb{R}/\Gamma$ we have $\mu([0.5])=2$, while for $[0.7]\in\mathbb{R}/\Gamma$ we have $\mu([0.7])=0$. The filtration spectrum is then the multiset $\{[0.5],[0.5]\}$.

While Definition 5.2 relies on a choice of an orthogonal basis for (C,ℓ), the following proposition shows that the filtration spectrum can be reformulated in a way that is manifestly independent of the choice of orthogonal basis, and so is in fact an invariant of the orthogonalizable Λ–space (C,ℓ).

Proposition 5.5 *Let (C,ℓ) be an orthogonalizable $\Lambda^{\mathcal{K},\Gamma}$–space and let $(\mathbb{R}/\Gamma,\mu)$ be the filtration spectrum of (C,ℓ) (as determined by an arbitrary orthogonal basis). Then for any $s\in\mathbb{R}/\Gamma$,*

$$\mu(s)=\max\bigl\{k\in\mathbb{N}\mid\exists\, V\le C \textit{ with } \dim(V)=k \textit{ and } \ell(v)\equiv s \bmod \Gamma \textit{ for all } v\in V\setminus\{0\}\bigr\}.$$

Proof Let $(v_1,\dots,v_n)$ be an orthogonal ordered basis of C and let μ be the multiplicity of some element $s\in\mathbb{R}/\Gamma$ in the filtration spectrum of C. So by definition there are precisely μ elements $i_1,\dots,i_\mu\in\{1,\dots,n\}$ such that each $\ell(v_{i_j})\equiv s \bmod \Gamma$ for $j=1,\dots,\mu$. Any nonzero element u in the μ–dimensional subspace spanned by the v_{i_j} can be written as $u=\sum_j \lambda_j v_{i_j}$, where $\lambda_j\in\Lambda$ are not all zero, and then $\ell(u)=\max_j\{\ell(v_{i_j})-\nu(\lambda_j)\}\equiv s \bmod \Gamma$ since $\nu(\lambda_j)$ all belong to Γ. This proves that μ is less than or equal to right hand side in the statement of the proposition.

For the reverse inequality, suppose that $V\le C$ has dimension greater than μ. For $i_1,\dots,i_\mu$ as in the previous paragraph, let $W=\operatorname{span}_\Lambda\{v_i\mid i\notin\{i_1,\dots,i_\mu\}\}$. Since W has codimension μ and $\dim V>\mu$, V and W intersect nontrivially. So there is some nonzero element $v=\sum_{i\notin\{i_1,\dots,i_\mu\}}\lambda_i v_i\in V\cap W$. Since the v_i are orthogonal, $\ell(v)$ has the same reduction modulo Γ as one of the v_i with $i\notin\{i_1,\dots,i_\mu\}$, and so this reduction is not equal to s. Thus no subspace of dimension greater than μ can have the property indicated in the statement of the proposition. □

Remark 5.6 Let us now relate our singular value decompositions to the Morse–Barannikov complex $\mathcal{C}(f)$ of an excellent Morse function $f\colon M\to\mathbb{R}$ on a Riemannian manifold as described in [28, Section 2], where the term "excellent" means in particular that the restriction of f to its set of critical points is injective.

This latter assumption means, in our language, that the filtration spectrum of the orthogonalizable $\mathcal{K}$–space $(\mathrm{CM}_*(f),\ell)$ consists of the index-k critical values of f, each occurring with multiplicity one, since (essentially by definition) $(\mathrm{CM}_*(f),\ell)$ has an orthogonal basis given by the critical points of f, with filtrations given by their corresponding critical values. So in view of Proposition 5.5, the filtration function ℓ will restrict to any other orthogonal basis of $(\mathrm{CM}_*(f),\ell)$ as a bijection to the set of critical values of f.

Denoting by ∂ the boundary operator on $\mathrm{CM}_*(f)$, Theorem 3.4 allows us to construct an orthogonal ordered basis $(x_1,\dots,x_r,y_1,\dots,y_r,z_1,\dots,z_h)$ for $\mathrm{CM}_*(f)$ such that $\mathrm{span}\{x_1,\dots,x_r\} = \mathrm{Im}(\partial)$, $\mathrm{span}\{x_1,\dots,x_r,z_1,\dots,z_h\} = \ker(\partial)$, and $\partial y_i = x_i$. By the previous paragraph, then, each critical value c of f can then be written in exactly one way as $c = \ell(x_i)$ or $c = \ell(y_i)$ or $c = \ell(z_i)$.

For $\lambda \in \mathbb{R}$, let C_*^λ denote the subcomplex of $\mathrm{CM}_*(f)$ spanned by the critical points with critical value at most λ. Observe that C_*^λ is equal to the subcomplex of $\mathrm{CM}_*(f)$ spanned by the x_i, y_i, z_i having $\ell \le \lambda$ (indeed the latter is clearly a subspace of C_*^λ, but Proposition 5.5 implies that their dimensions are the same). Now the treatment of the Barannikov complex in [28] involves separating the critical values c of f into three types, where ϵ represents a small positive number:

- The *lower* critical values, for which the natural map
$$H_*(C_*^{c+\epsilon}/C_*^{c-\epsilon}) \to H_*(\mathrm{CM}_*(f)/C_*^{c-\epsilon})$$
vanishes;
- The *upper* critical values, for which the natural map
$$H_*(C_*^{c+\epsilon}) \to H_*(C_*^{c+\epsilon}, C_*^{c-\epsilon})$$
vanishes (equivalently, $H_*(C_*^{c-\epsilon}) \to H_*(C_*^{c+\epsilon})$ is surjective);
- All other critical values, called *homological* critical values.

If w is any of x_i, y_i, or z_i and if $\ell(w) = c$, one has $C_*^{c+\epsilon} = C_*^{c-\epsilon} \oplus \langle w \rangle$. Consequently it is easy to see that c is a lower critical value if and only if $c = \ell(x_i)$ for some i, that c is an upper critical value if and only if $c = \ell(y_i)$ for some i, and that c is a homological critical value if and only if $c = \ell(z_i)$ for some i. Moreover, in the case that c is an upper critical value so that $c = \ell(y_i)$ for some i, the natural map $H_*(C_*^{c+\epsilon}/C_*^\lambda) \to H_*(C_*^{c+\epsilon}/C_*^{c-\epsilon})$ vanishes precisely for $\lambda \le \ell(x_i)$.

In [28, Definition 2.9], the Morse–Barannikov complex $(\mathcal{C}(f), \partial_B)$ is described as the chain complex generated by the critical values of f, with boundary operator given by $\partial_B c = 0$ if c is a lower critical value or a homological critical value, and
$$\partial_B c = \sup\{\lambda \mid H_*(C_*^{c+\epsilon}/C_*^\lambda) \to H_*(C_*^{c+\epsilon}/C_*^{c-\epsilon}) \text{ is the zero map}\}$$
if c is an upper critical value. The foregoing discussion shows that the unique linear map $(\mathrm{CM}_*(f), \partial) \to (\mathcal{C}(f), \partial_B)$ that sends the basis elements x_i, y_i, z_i to their respective filtration levels $\ell(x_i), \ell(y_i), \ell(z_i)$ defines an isomorphism of chain complexes. In particular, the Morse–Barannikov complex can be recovered quite directly from a singular value decomposition.

6 Barcodes

Recall from the introduction that a persistence module $\mathbb{V}=\{V_t\}_{t\in\mathbb{R}}$ over the field $\mathcal{K}$ is a system of $\mathcal{K}$–vector spaces V_t with suitably compatible maps $V_s\to V_t$ whenever $s\le t$.

A special case of a persistence module is obtained by choosing an interval $I\subset\mathbb{R}$ and defining

$$(\mathcal{K}_I)_t=\begin{cases}\mathcal{K} & \text{if } t\in I,\\ 0 & \text{if } t\notin I,\end{cases}$$

with the maps $(\mathcal{K}_I)_s\to(\mathcal{K}_I)_t$ defined to be the identity when $s,t\in I$ and to be zero otherwise.

A persistence module $\mathbb{V}$ is called *pointwise finite-dimensional* if each V_t is finite-dimensional. Such persistence modules obey the following structure theorem.

Theorem 6.1 [46; 12] *Every pointwise finite-dimensional persistence module $\mathbb{V}$ can be uniquely decomposed into the following normal form:*

$$\mathbb{V}\cong\bigoplus_{\alpha}\mathcal{K}_{I_\alpha} \tag{18}$$

for certain intervals $I_\alpha\subset\mathbb{R}$

The (persistent homology) *barcode* of $\mathbb{V}$ is then by definition the multiset (S,μ), where S is the set of intervals I for which $\mathcal{K}_I$ appears in (18) and $\mu(I)$ is the number of times that $\mathcal{K}_I$ appears. As follows from the discussion at the end of the introduction in [12], the barcode is a complete invariant of a pointwise finite-dimensional persistence module.

In classical persistent homology, where the persistence module is constructed from the filtered homologies of the Čech or Rips complexes associated to a point cloud, [46] provides an algorithm computing the resulting barcode (cf Theorem 3.5). In this case the intervals in the barcode are all half-open intervals $[a,b)$ (with possibly $b=\infty$). See eg [23, Figure 4] and [7, page 278] for some nice illustrations of barcodes.

Returning to the context of the Floer-type complexes (C_*,∂,ℓ) considered in this paper, for any $t\in\mathbb{R}$, if we let $C_k^t=\{c\in C_k\mid \ell(c)\le t\}$ the assumption on the effect of ∂ on ℓ shows that we have a subcomplex C_*^t; just as discussed in the introduction for any k the degree-k homologies $H_k^t(C_*)$ of these complexes yield a persistence module over the base field $\mathcal{K}$. Typically $H_k^t(C_*)$ can be infinite-dimensional (and also may not satisfy the weaker descending chain condition which appears in [12]), so Theorem 6.1 usually does not apply to these persistence modules. The exception to this is when the subgroup $\Gamma \le \mathbb{R}$ used in the Novikov field $\Lambda = \Lambda^{\mathcal{K},\Gamma}$ is the trivial group, in which case we just have $\Lambda = \mathcal{K}$ and the chain groups C_k (and so also the homologies) are finite-dimensional over $\mathcal{K}$. So when $\Gamma = \{0\}$, Theorem 6.1 does apply to show that the persistence module $\{H_k^t(C_*)\}_{t\in\mathbb{R}}$ decomposes as a direct sum of interval modules $\mathcal{K}_I$; by definition the degree-k part of the barcode of C_* is then the multiset of intervals appearing in this direct sum decomposition. We have:

Theorem 6.2 *Assume that $\Gamma = \{0\}$ and let (C_*, ∂, ℓ) be a Floer-type complex over $\Lambda^{\mathcal{K},\{0\}} = \mathcal{K}$. For each $k \in \mathbb{Z}$ write $\partial_{k+1}\colon C_{k+1} \to C_k$ for the degree-$(k+1)$ part of the boundary operator ∂, and write $Z_k = \ker \partial_k$, so that ∂_{k+1} has image contained in Z_k. Let $((y_1, \dots, y_n), (x_1, \dots, x_m))$ be a singular value decomposition for $\partial_{k+1}\colon C_{k+1} \to Z_k$. Then if $r = \operatorname{rank}(\partial_{k+1})$, the degree-$k$ part of the barcode of C_* consists precisely of:*

- *an interval $[\ell(x_i), \ell(y_i))$ for each $i \in \{1, \dots, r\}$ such that $\ell(y_i) > \ell(x_i)$; and*
- *an interval $[\ell(x_i), \infty)$ for each $i \in \{r+1, \dots, m\}$.*

Proof As explained earlier, $\{H_k^t(C_*)\}_{t\in\mathbb{R}}$ is a pointwise finite-dimensional persistence module. Therefore by Theorem 6.1, we have a normal form $\bigoplus_\alpha \mathcal{K}_{I_\alpha}$. Given a singular value decomposition $((y_1, \dots, y_n), (x_1, \dots, x_m))$ as in the hypothesis, we first claim that, for all $t \in \mathbb{R}$,

$$H_k^t(C_*) = \operatorname{span}_{\mathcal{K}}\left\{ [x_i] \;\middle|\; \begin{array}{ll} \ell(x_i) \le t < \ell(y_i) & \text{if } i \in \{1, \dots, r\} \\ \ell(x_i) \le t & \text{if } i \in \{r+1, \dots, m\} \end{array} \right\}. \tag{19}$$

In fact, $(x_1, \dots, x_m)$ is an orthogonal ordered basis for $\ker \partial_k$, so $\{x_i \mid \ell(x_i) \le t\}$ is an orthogonal basis for $\ker(\partial_k|_{C_k^t})$. By Lemma 4.9, when $\Gamma = \{0\}$ (so that ν vanishes on all nonzero elements of Λ), an element $x = \sum_{i=1}^m \lambda_i x_i$ lies in $\partial_{k+1}(C_{k+1}^t)$ if and only if it holds both that $\lambda_i = 0$ for all $i > r$ and that $\ell\big(\sum_{i=1}^r \lambda_i y_i\big) \le t$, ie if and only if $x \in \operatorname{span}_{\mathcal{K}}\{x_i \mid 1 \le i \le r,\ \ell(y_i) \le t\}$. So we have bases $\{x_i \mid \ell(x_i) \le t\}$ for $Z_k \cap C_k^t$ and $\{x_i \mid 1 \le i \le r,\ \ell(y_i) \le t\}$ for $\partial_{k+1}(C_{k+1}^t)$, from which the expression (19) for $H_k^t(C_*)$ immediately follows.

Write V_t for the right hand side of (19). For $s \le t$, the inclusion-induced map $\sigma_{st}\colon H_k^s(C_*) \to H_k^t(C_*)$ is identified with the map $\sigma_{st}\colon V_s \to V_t$ defined as follows, for any generator $[x_i]$ of V_s:

$$\sigma_{st}([x_i]) = \begin{cases} [x_i] & \text{if } \ell(y_i) > t \text{ or } i \in \{r+1, \dots, s\}, \\ 0 & \text{if } \ell(y_i) \le t. \end{cases} \tag{20}$$

Clearly, this is a $\mathcal{K}$–linear homomorphism. It is easy to check that $\sigma_{ss} = I_{V_s}$ and for $s \le t \le u$, $\sigma_{su} = \sigma_{tu} \circ \sigma_{st}$. Therefore, $\mathbb{V} = \{V_t\}_{t\in\mathbb{R}}$ is a persistence module, which is (tautologically) *isomorphic*, in the sense of persistence modules, to $\{H_k^t(C_*)\}_{t\in\mathbb{R}}$.

On the other hand, the normal form of $\mathbb{V}$ can be explicitly written out as follows:

$$\mathbb{V} \cong \bigoplus_{1\le i\le r} \mathcal{K}_{[\ell(x_i),\ell(y_i))} \oplus \bigoplus_{r+1\le j\le m} \mathcal{K}_{[\ell(x_j),\infty)}. \tag{21}$$

Indeed the indicated isomorphism of persistence modules can be obtained by simply mapping $1 \in (\mathcal{K}_{[\ell(x_i),\ell(y_i))]})_t = \mathcal{K}$ to the class $[x_i]$ for $t \in [\ell(x_i), \ell(y_i))$ and $i = 1, \dots, r$, and similarly for the $\mathcal{K}_{[\ell(x_i),\infty)}$ for $i > r$. □

Thus in the "classical" $\Gamma = \{0\}$ case the barcode can be read off directly from the filtration levels of the elements involved in a singular value decomposition; in particular, these filtration levels are independent of the choice of singular value decomposition, consistently with Theorem 7.1 below. For nontrivial Γ there is clearly some amount of arbitrariness of the filtration levels of the elements of a singular value decomposition: if $((y_1, \dots, y_n), (x_1, \dots, x_m))$ is a singular value decomposition, then $((T^{g_1}y_1, \dots, T^{g_r}y_r, y_{r+1}, \dots, y_n), (T^{g_1}x_1, \dots, T^{g_m}x_m))$ is also a singular value decomposition for any $g_1, \dots, g_m \in \Gamma$; based on Theorem 6.2 one would expect this to result in a change of the positions of each of the intervals in the barcode. Note that this change moves the endpoints of the intervals but does not alter their lengths. This suggests the following definition, related to the ideas of boundary depth and filtration spectrum:

Definition 6.3 Let (C_*, ∂, ℓ) be a Floer-type complex over $\Lambda = \Lambda^{\mathcal{K},\Gamma}$ and for each $k \in \mathbb{Z}$ write $\partial_k = \partial|_{C_k}$ and $Z_k = \ker \partial_k$. Given any $k \in \mathbb{Z}$ choose a singular value decomposition $((y_1, \dots, y_n), (x_1, \dots, x_m))$ for the Λ–linear map $\partial_{k+1}\colon C_{k+1} \to Z_k$ and let r denote the rank of ∂_{k+1}. Then the degree-k *verbose barcode* of (C_*, ∂, ℓ) is the multiset of elements of $(\mathbb{R}/\Gamma) \times [0, \infty]$ consisting of

(i) a pair $(\ell(x_i) \bmod \Gamma, \ell(y_i) - \ell(x_i))$ for $i = 1, \dots, r$;

(ii) a pair $(\ell(x_i) \bmod \Gamma, \infty)$ for $i = r + 1, \dots, m$.

The *concise barcode* is the submultiset of the verbose barcode consisting of those elements whose second element is positive.

Thus in the case that $\Gamma = \{0\}$ elements $[a, b)$ of the persistent homology barcode correspond according to Theorem 6.2 to elements $(a, b - a)$ of the concise barcode. In general we think of an element $([a], L)$ of the (verbose or concise) barcode as corresponding to an interval with left endpoint a and length L, with the understanding that the left endpoint is only specified up to the additive action of Γ.

Definition 6.3 appears to depend on a choice of singular value decomposition, but we will see in Theorem 7.1 that different choices of singular value decompositions yield the same verbose (and hence also concise) barcodes. Of course in the case that $\Gamma = \{0\}$ this already follows from Theorem 6.2; in the opposite extreme case that $\Gamma = \mathbb{R}$ (in which case the first coordinates of the pairs in the verbose and concise barcodes carry no information) it can easily be inferred from Theorem 4.13.

Remark 6.4 Our reduction modulo Γ in Definition 6.3(i) and (ii) is easily seen to be necessary if there is to be any hope of the verbose and concise barcodes being independent of the choice of singular value decomposition, for the reason indicated in the paragraph before Definition 6.3. Namely, acting on the elements involved in the singular value decomposition by appropriate elements of Λ could change the various quantities $\ell(x_i)$ involved in the barcode by arbitrary elements of Γ.

Remark 6.5 In the spirit of Theorem 3.5, we outline the procedure for computing the degree-k verbose barcode for a Floer-type complex (C_*, ∂, ℓ):

- First, by applying the algorithm in Theorem 3.5 to $\partial_k \colon C_k \to C_{k-1}$ or otherwise, obtain an orthogonal ordered basis $(w_1, \dots, w_m)$ for $\ker \partial_k$.
- Express $\partial_{k+1} \colon C_{k+1} \to \ker \partial_k$ in terms of an orthogonal basis for C_{k+1} and the basis $(w_1, \dots, w_m)$ for $\ker \partial_k$, and apply Theorem 3.5 to obtain data $(v_1', \dots, v_n')$ and $\mathcal{U}$ as in the statement of that theorem.
- The degree-k verbose barcode consists of one element $([\ell(Av_i')], \ell(v_i') - \ell(Av_i'))$ for each $i \in \mathcal{U}$, and one element $([a], \infty)$ for each $[a]$ lying in the multiset complement $\{[\ell(w_1)], \dots, [\ell(w_m)]\} \setminus \{[\ell(Av_i')] \mid i \in \mathcal{U}\}$.

6.1 Relation to spectral invariants

Following a construction that is found in [40; 34] in the context of Hamiltonian Floer theory (and which is closely related to classical minimax-type arguments in Morse theory), we may describe the *spectral invariants* associated to a Floer-type complex (C_*, ∂, ℓ): letting $H_k(C_*)$ denote the degree-k homology of C_*, these invariants take the form of a map $\rho \colon H_k(C_*) \to \mathbb{R} \cup \{-\infty\}$ defined by, for $\alpha \in H_k(C_*)$,

$$\rho(\alpha) = \inf\{\ell(c) \mid c \in C_k,\ [c] = \alpha\}$$

(where $[c]$ denotes the homology class of c). In a more general context the main result of [42] shows that the infimum in the definition of $\rho(\alpha)$ is always attained.

The spectral invariants are reflected in the concise barcode in the following way.

Proposition 6.6 *Let $\mathcal{B}_{C,k}$ be the degree-k part of the concise barcode of a Floer-type complex (C_*, ∂, ℓ), obtained from a singular value decomposition of $\partial_{k+1} \colon C_{k+1} \to \ker \partial_k$. Then:*

(i) *There is a basis* $\{\alpha_1,\dots,\alpha_h\}$ *for* $H_k(C_*)$ *over* Λ *such that the submultiset of* $\mathcal{B}_{C,k}$ *consisting of elements with second coordinate equal to* ∞ *is equal to* $\{([\rho(\alpha_1)],\infty),\dots,([\rho(\alpha_h)],\infty)\}$, *where for each* i, $[\rho(\alpha_i)]$ *denotes the reduction of* $\rho(\alpha_i)$ *modulo* Γ.

(ii) *For any class* $\alpha\in H_k(C_*)$, *if we write* $\alpha=\sum_{i=1}^h \lambda_i\alpha_i$, *where* $\lambda_i\in\Lambda$ *and* $\{\alpha_1,\dots,\alpha_h\}$ *is the basis from* (i), *then* $\rho(\alpha)=\max_i(\rho(\alpha_i)-\nu(\lambda_i))$. *In particular, if* $\alpha\neq 0$, *then the concise barcode* $\mathcal{B}_{C,k}$ *contains an element of the form* $([\rho(\alpha)],\infty)$.

Proof Let $((y_1,\dots,y_m),(x_1,\dots,x_n))$ be a singular value decomposition of the map $\partial_{k+1}\colon C_{k+1}\to\ker\partial_k$. In particular, if $r=\operatorname{rank}\partial_{k+1}$, then $\operatorname{span}_\Lambda\{x_{r+1},\dots,x_m\}$ is an orthogonal complement to $\operatorname{Im}\partial_{k+1}$. Hence the classes $\alpha_i=[x_{r+i}]$ (for $1\le i\le m-r$) form a basis for $H_k(C_*)$, and the dimension of the $H_k(C_*)$ over Λ is $h=m-r$. By definition, the submultiset of $\mathcal{B}_{C,k}$ consisting of elements with second coordinate equal to ∞ is $\{([\ell(x_{r+1})],\infty),\dots,([\ell(x_m)],\infty)\}$, so both part (i) and the first sentence of part (ii) of the proposition will follow if we show that, for any $\lambda_1,\dots,\lambda_{m-r}\in\Lambda$ we have

$$\rho\left(\sum_{i=1}^{m-r}\lambda_i\alpha_i\right)=\max_i(\ell(x_{r+i})-\nu(\lambda_i)) \tag{22}$$

(indeed the special case of (22) in which $\lambda_i=\delta_{ij}$ implies that $\rho(\alpha_j)=\ell(x_{r+j})$).

To prove (22), simply note that any class $\alpha=\sum_i\lambda_i\alpha_i\in H_k(C_*)$ is represented by the chain $\sum_i\lambda_i x_{r+i}$, and that the general representative of α is given by $x=y+\sum_i\lambda_i x_{r+i}$ for $y\in\operatorname{Im}\partial_{k+1}$. So since $\{x_{r+1},\dots,x_m\}$ is an orthogonal basis for an orthogonal complement to $\operatorname{Im}\partial_{k+1}$ it follows that

$$\ell(x)=\max\left\{\ell(y),\ell\left(\sum_i\lambda_i x_{r+i}\right)\right\}\ge\ell\left(\sum_i\lambda_i x_{r+i}\right)=\max_i(\ell(x_{r+i})-\nu(\lambda_i)),$$

with equality if $y=0$. Thus the minimal value of ℓ on any representative x of $\sum_{i=1}^{m-r}\lambda_i\alpha_i$ is equal to $\max_i(\ell(x_{r+i})-\nu(\lambda_i))$, proving (22).

As noted earlier, (22) directly implies (i) and the first sentence of (ii). But then the second sentence of (ii) also follows immediately, since each $\lambda\in\Lambda\setminus\{0\}$ has $\nu(\lambda)\in\Lambda$, and so if $\alpha=\sum_i\lambda_i\alpha_i\neq 0$ it follows from (22) that $\rho(\alpha)$ is congruent mod Γ to one of the $\rho(\alpha_i)$. □

6.2 Duality and coefficient extension for barcodes

Given a Floer-type complex (C_*,∂,ℓ) over $\Lambda=\Lambda^{\mathcal{K},\Gamma}$ one obtains a dual complex $(C_*^\vee,\delta,\ell^*)$ by taking $C_k^\vee$ to be the dual over Λ of C_{-k}, $\delta\colon C_k^\vee\to C_{k-1}^\vee$ to be the adjoint of $\partial\colon C_{-k+1} \to C_{-k}$ and defining ℓ^* as in Section 2.4. The following can be seen as a generalization both of [43, Corollary 1.6] and of [41, Proposition 2.4]

Proposition 6.7 *For all k, denote by $\widetilde{\mathcal{B}}_{C,k}$ the degree-k verbose barcode of (C_*, ∂, ℓ). Then the degree-k verbose barcode of $(C_*^\vee, \delta, \ell^*)$ is given by*

$$\begin{aligned}(23)\quad \widetilde{\mathcal{B}}_{C^\vee,k} = \big\{([-a], \infty) \mid ([a], \infty) \in \widetilde{\mathcal{B}}_{C,-k}\big\} \\ \cup \big\{([-a-L], L) \mid L < \infty \text{ and } ([a], L) \in \widetilde{\mathcal{B}}_{C,-k-1}\big\}.\end{aligned}$$

Proof Suppose that

$$\begin{aligned} r &= \operatorname{rank}\big(\partial_{-k}\colon C_{-k} \to C_{-k-1}\big),\\ s &= \operatorname{rank}\big(\partial_{-k+1}\colon C_{-k+1} \to C_{-k}\big),\\ t &= \dim\ker\big(\partial_{-k-1}\colon C_{-k-1} \to C_{-k-2}\big),\end{aligned}$$

and note that $t \geq r$. Using the Gram–Schmidt process in Theorem 2.16 if necessary, we can modify a singular value decomposition $((y_1, \dots, y_n), (x_1, \dots, x_m))$ of $\partial_{-k}\colon C_{-k} \to C_{-k-1}$ so that it has the following additional properties:

(i) $(x_1, \dots, x_t)$ is an orthogonal ordered basis for $\ker \partial_{-k-1}$, so that in particular $((y_1, \dots, y_n), (x_1, \dots, x_t))$ is a singular value decomposition for $\partial_{-k}\colon C_{-k} \to \ker \partial_{-k-1}$.

(ii) $(y_{n-s+1}, \dots, y_n)$ is an orthogonal ordered basis for $\operatorname{Im}\partial_{-k+1}$, so that the elements $([a], L)$ of $\widetilde{\mathcal{B}}_{C,-k}$ having $L = \infty$ are precisely the $([\ell(y_i)], \infty)$ for $i \in \{r+1, \dots, n-s\}$.

By Proposition 2.20, a singular value decomposition for $\delta_{k+1}\colon C_{k+1}^\vee \to C_k^\vee$ is given by $((x_1^*, \dots, x_m^*), (y_1^*, \dots, y_n^*))$, where the x_i^* and y_j^* form dual bases for the bases $(x_1, \dots, x_m)$ and $(y_1, \dots, y_n)$, respectively. Moreover by (ii) above, the kernel of $\delta_k\colon C_k^\vee \to C_{k-1}^\vee$ (ie the annihilator of the image of ∂_{-k+1}) is precisely the span of $y_1^*, \dots, y_{n-s}^*$, and so $((x_1^*, \dots, x_m^*), (y_1^*, \dots, y_{n-s}^*))$ is a singular value decomposition for $\delta_{k+1}\colon C_{k+1}^\vee \to \ker \delta_k$. Since by (7) we have $\ell^*(x_i^*) = -\ell(x_i)$ and $\ell^*(y_i^*) = -\ell(y_i)$ it follows that

$$\widetilde{\mathcal{B}}_{C^\vee,k} = \big\{([-\ell(y_i)], \ell(y_i) - \ell(x_i)) \mid i = 1, \dots, r\big\} \cup \big\{([-\ell(y_i)], \infty) \mid i = r+1, \dots, n-s\big\},$$

which precisely equals the right hand side of (23). □

The effect on the verbose barcode of extending the coefficient field of a Floer-type complex by enlarging the value group Γ is even easier to work out, given our earlier results.

Proposition 6.8 *Let (C_*, ∂, ℓ) be a Floer-type complex over $\Lambda = \Lambda^{\mathcal{K},\Gamma}$, let $\Gamma' \leq \mathbb{R}$ be a subgroup containing Γ, and consider the Floer-type complex $(C'_*, \partial \otimes 1, \ell')$ over $\Lambda^{\mathcal{K},\Gamma'}$ given by letting $C'_k = C_k \otimes_\Lambda \Lambda^{\mathcal{K},\Gamma'}$ and defining ℓ' as in Section 2.5. Let $\widetilde{\mathcal{B}}_{C,k}$ be the verbose barcode of (C_*, ∂, ℓ) in degree k and let $\pi\colon \mathbb{R}/\Gamma \to \mathbb{R}/\Gamma'$ be the projection. Then the verbose barcode of $(C'_*, \partial \otimes 1, \ell')$ in degree k is*

$$\{(\pi([a]), L) \mid ([a], L) \in \widetilde{\mathcal{B}}_{C,k}\}.$$

Proof This follows directly from Proposition 3.10 and the definitions. □

7 Classification theorems

In the spirit of the structure theorem (Theorem 6.1) for pointwise finite-dimensional persistence modules, we will use the verbose and concise barcodes to classify Floer-type complexes up to filtered chain isomorphism and filtered homotopy equivalence. Specifically, we will prove the following two key theorems, stated earlier in the introduction.

Theorem A *Two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ are filtered chain isomorphic to each other if and only if they have identical verbose barcodes in all degrees.*

Theorem B *Two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ are filtered homotopy equivalent to each other if and only if they have identical concise barcodes in all degrees.*

7.1 Classification up to filtered isomorphism

We will assume the following important theorem first, and then Theorem A will follow quickly.

Theorem 7.1 *For any $k \in \mathbb{Z}$, the degree-k verbose barcode of any Floer-type complex is independent of the choice of singular value decomposition for $\partial_{k+1}\colon C_{k+1} \to Z_k$.*

Proof of Theorem A On the one hand, a filtered chain isomorphism $C_* \to D_*$ maps a singular value decomposition for $(\partial_C)_{k+1}\colon C_{k+1} \to \ker(\partial_C)_k$ to a singular value decomposition for $(\partial_D)_{k+1}\colon D_{k+1} \to \ker(\partial_D)_k$, while keeping all filtration levels the same. Therefore, the "only if" part of Theorem A is a direct consequence of Theorem 7.1.

To prove the "if" part of Theorem A we begin by introducing some notation that will also be useful to us later. Given a collection of Floer-type complexes $\mathcal{C}_\alpha = (C_{\alpha *}, \partial_\alpha, \ell_\alpha)$

we define $\bigoplus_\alpha \mathcal{C}_\alpha$ to be the triple $\left(\bigoplus_\alpha C_{\alpha *}, \bigoplus_\alpha \partial_\alpha, \tilde{\ell}\right)$, where $\tilde{\ell}((c_\alpha)) = \max_\alpha \ell_\alpha(c_\alpha)$. Provided that, for each $k \in \mathbb{Z}$, only finitely many of the $C_{\alpha k}$ are nontrivial, $\bigoplus_\alpha \mathcal{C}_\alpha$ is also a Floer-type complex.

Definition 7.2 Fix $\Gamma \leq \mathbb{R}$ and the associated Novikov field $\Lambda = \Lambda^{\mathcal{K},\Gamma}$. For $a \in \mathbb{R}$, $L \in [0, \infty]$, and $k \in \mathbb{Z}$ define the *elementary Floer-type complex* $\mathcal{E}(a, L, k)$ to be the Floer-type complex $(E_*, \partial_E, \ell_E)$ given as follows:

- If $L = \infty$ then
$$E_m = \begin{cases} \Lambda & \text{if } m = k, \\ 0 & \text{otherwise,} \end{cases}$$
$\partial_E = 0$, and $\ell(\lambda) = a - \nu(\lambda)$ for $\lambda \in E_m = \Lambda$.
- If $L \in [0, \infty)$, then E_k is the one-dimensional Λ–vector space generated by a symbol x, E_{k+1} is the one-dimensional Λ–vector space generated by a symbol y, and $E_m = \{0\}$ for $m \notin \{k, k+1\}$. Also, $\partial_E \colon E_* \to E_*$ is defined by $\partial_E(\lambda x + \mu y) = \mu x$, and $\ell_E(\lambda x + \mu y) = \max\{a - \nu(\lambda), (a + L) - \nu(\mu)\}$.

Remark 7.3 If $b - a \in \Gamma$, then there is a filtered chain isomorphism $\mathcal{E}(a, L, k) \to \mathcal{E}(b, L, k)$ given by scalar multiplication by the element $T^{b-a} \in \Lambda$.

Proposition 7.4 *Let (C_*, ∂, ℓ) be a Floer-type complex and denote by $\tilde{\mathcal{B}}_{C,k}$ the degree-k verbose barcode of (C_*, ∂, ℓ). Then there is a filtered chain isomorphism*

$$(C_*, \partial, \ell) \cong \bigoplus_{k \in \mathbb{Z}} \bigoplus_{([a], L) \in \tilde{\mathcal{B}}_{C,k}} \mathcal{E}(a, L, k)$$

(where for each $([a], L) \in \tilde{\mathcal{B}}_{C,k}$ we choose an arbitrary representative $a \in \mathbb{R}$ of the coset $[a] \in \mathbb{R}/\Gamma$).

Proof of Proposition 7.4 For each k let

$$((y_1^k, \dots, y_{r_k}^k, \dots, y_{r_k + m_{k+1}}^k), (x_1^k, \dots, x_{m_k}^k))$$

be an arbitrary singular value decomposition for $(\partial_C)_{k+1} \colon C_{k+1} \to \ker(\partial_C)_k$, where r_k is the rank of $(\partial_C)_{k+1}$ and $m_k = \dim(\ker(\partial_C)_k)$ for each degree $k \in \mathbb{Z}$. We will first modify these singular value decompositions for various k to be related to each other in a convenient way. Specifically, since $(x_1^{k+1}, \dots, x_{m_{k+1}}^{k+1})$ is an orthogonal ordered basis for $\ker(\partial_C)_{k+1}$, the tuple

$$((y_1^k, \dots, y_{r_k}^k, x_1^{k+1}, \dots, x_{m_{k+1}}^{k+1}), (x_1^k, \dots, x_{m_k}^k))$$

is also a singular value decomposition for $(\partial_C)_{k+1}\colon C_{k+1} \to \ker(\partial_C)_k$. So letting

$$(a_i^k, L_i^k) = \begin{cases} (\ell(x_i^k), \ell(y_i^k) - \ell(x_i^k)) & \text{if } 1 \le i \le r_k, \\ (\ell(x_i^k), \infty) & \text{if } r_k + 1 \le i \le m_k, \end{cases}$$

we have $\mathcal{B}_{C,k} = \{([a_i^k], L_i^k) | 1 \le i \le m_k\}$ and the proposition states that (C_*, ∂, ℓ) is filtered chain isomorphic to $\bigoplus_k \bigoplus_{i=1}^{m_k} \mathcal{E}(a_i^k, L_i^k, k)$. Now for each i and k there is an obvious embedding $\phi_{i,k}\colon \mathcal{E}(a_i^k, L_i^k, k) \to C_*$ defined by

- $\phi_{i,k}(\lambda) = \lambda x_i^k$ when $L_i^k = \infty$;
- $\phi_{i,k}(\lambda x + \mu y) = \lambda x_i^k + \mu y_i^k$ when $L_i^k < \infty$.

From the definition of the filtration and boundary operator on $\mathcal{E}(a_i^k, L_i^k, k)$ this embedding is a chain map which exactly preserves filtration levels. Then

$$\bigoplus_{i,k} \phi_{i,k}\colon \bigoplus_{i,k} \mathcal{E}(a_i^k, L_i^k, k) \to C_*$$

is also a chain map. Finally, for each k, the fact that $(y_1^k, \dots, y_{r_k}^k, x_1^{k+1}, \dots, x_{m_{k+1}}^{k+1})$ is an orthogonal ordered basis for C_{k+1} readily implies that $\bigoplus_{i,k} \phi_{i,k}$ is in fact a filtered chain isomorphism. □

Since, by Remark 7.3, the filtered isomorphism type of $\mathcal{E}(a, L, k)$ only depends on $[a], L, k$, and since quite generally filtered chain isomorphisms $\Phi_\alpha\colon \mathcal{C}_\alpha \to \mathcal{D}_\alpha$ between Floer-type complexes induce a filtered chain isomorphism $\bigoplus_\alpha\colon \bigoplus_\alpha \mathcal{C}_\alpha \to \bigoplus_\alpha \mathcal{D}_\alpha$, Proposition 7.4 shows that the filtered chain isomorphism type of a Floer-type complex is determined by its verbose barcode, proving the "if" part of Theorem A. □

The remainder of this subsection is directed toward the proof of Theorem 7.1. We will repeatedly apply the following criterion for testing whether a subspace is an orthogonal complement of a given subspace.

Lemma 7.5 *Let (C, ℓ) be an orthogonalizable Λ–space, and let $U, U', V \le C$ be subspaces such that U is an orthogonal complement to V and $\dim U' = \dim U$. Consider the projection $\pi_U\colon C \to U$ associated to the direct sum decomposition $C = U \oplus V$. Then U' is an orthogonal complement of V if and only if $\ell(\pi_U x) = \ell(x)$ for all $x \in U'$.*

Proof Assume that U' is an orthogonal complement to V. Then for $x \in U'$, we of course have

$$x = \pi_U x + (x - \pi_U x),$$

where $\pi_U x \in U$ and $x - \pi_U x \in V$. Because U and V are orthogonal, it follows that $\ell(x) = \max\{\ell(\pi_U x), \ell(x - \pi_U x)\}$. In particular,

$$\ell(x) \geq \ell(\pi_U x). \tag{24}$$

On the other hand, since

$$\pi_U x = x - (x - \pi_U x),$$

where $x \in U'$, $x - \pi_U x \in V$, and U' and V are orthogonal, we have $\ell(\pi_U x) = \max\{\ell(x), \ell(x - \pi_U x)\}$. In particular, $\ell(\pi_U x) \geq \ell(x)$. Combined with (24), this shows $\ell(x) = \ell(\pi_U x)$.

Conversely, suppose that $\ell(\pi_U x) = \ell(x)$ for all $x \in U'$. To show that U' is an orthogonal complement to V we just need to show that U' and V are orthogonal, that is, for any $x \in U'$ and $v \in V$ we have $\ell(x + v) = \max\{\ell(x), \ell(v)\}$ (indeed if we show this, then by Lemma 2.9(i) U' and V will have trivial intersection and so dimensional considerations will imply that $C = U' \oplus V$). Now write $x \in U'$ as

$$x = \pi_U x + (x - \pi_U x),$$

where $\pi_U x \in U$ and $x - \pi_U x \in V$. Because U and V are orthogonal, our assumption shows that $\ell(x) = \ell(\pi_U x) \geq \ell(x - \pi_U x)$. Now

$$x + v = \pi_U x + (v + (x - \pi_U x)),$$

where $\pi_U x$ in U and $v + (x - \pi_U x) \in V$. Again, U and V are orthogonal, so we have

$$\begin{aligned}\ell(x + v) &= \max\{\ell(\pi_U x), \ell(v + (x - \pi_U x))\}\\ &= \max\{\ell(x), \ell(v + (x - \pi_U x))\}.\end{aligned}$$

Now if $\ell(v) > \ell(x)$ then $\ell(x + v) = \ell(v) = \max\{\ell(x), \ell(v)\}$, as desired. On the other hand if $\ell(v) \leq \ell(x)$ then $\ell(v + (x - \pi_U x)) \leq \max\{\ell(v), \ell(x - \pi_U x)\} \leq \ell(x)$, and so $\ell(x + v) = \ell(x) = \max\{\ell(x), \ell(v)\}$. So in any case we indeed have $\ell(x + v) = \max\{\ell(x), \ell(v)\}$ for any $x \in U', v \in V$, and so U' and V are orthogonal. □

Notation 7.6 Let $((y_1, \dots, y_n), (x_1, \dots, x_m))$ be a singular value decomposition for a two-term Floer-type complex $(C_1 \xrightarrow{\partial} C_0)$, and let r be the rank of ∂. Denote $k_1, \dots, k_p \in \{1, \dots, r\}$ to be the increasing finite sequence of integers defined by the property that $k_1 = 1$ and, for $i \in \{1, \dots, p\}$, either $\beta_{k_i}(\partial) = \beta_{k_i+1}(\partial) = \cdots = \beta_r(\partial)$ (in which case $p = i$) or else $\beta_{k_i}(\partial) = \cdots = \beta_{k_{i+1}-1}(\partial) > \beta_{k_{i+1}}(\partial)$. Also let $k_{p+1} = r + 1$. We emphasize that the numbers k_i are independent of the choice of singular value decomposition (since the $\beta_k(\partial)$ are likewise independent thereof; see Definition 4.10).

The proof of Theorem 7.1 inductively uses the following lemma, which is an application of Lemma 7.5.

Lemma 7.7 *Let $((y_1,\dots,y_n),(x_1,\dots,x_m))$ be a singular value decomposition for $(C_1 \xrightarrow{\partial} C_0)$ and $r=\operatorname{rank}(\partial)$, and let $k_1,\dots,k_{p+1}$ be the integers in Notation 7.6. Let $i\in\{1,\dots,p\}$, and suppose that $V,W\le \operatorname{Im}\partial\le C_0$ satisfy the following conditions:*

(i) *$\dim V=k_i-1$, V is δ–robust for all $\delta<\beta_{k_{i-1}}(\partial)$, and V is orthogonal to $\operatorname{span}_\Lambda\{x_{k_i},\dots,x_m\}$. (If $i=1$ these conditions mean $V=\{0\}$.)*

(ii) *$\dim W=k_{i+1}-k_i$, W is orthogonal to V, and $V\oplus W$ is δ–robust for all $\delta<\beta_{k_i}(\partial)$.*

Now let $X=\operatorname{span}_\Lambda\{x_{k_i},\dots,x_{k_{i+1}-1}\}$ and $X'=\operatorname{span}_\Lambda\{x_{k_{i+1}},\dots,x_m\}$. Then $V\oplus W$ is orthogonal to X', and there is an isomorphism of filtered vector spaces $W\cong X$.

Proof Since V is orthogonal to $X\oplus X'$ and X is orthogonal to X', by Lemma 2.9, we have an orthogonal direct sum decomposition $C_0=X\oplus(X'\oplus V)$. We will first show that the projection $\pi_X\colon C_0\to X$ associated to this direct sum decomposition has the property that $\pi_X|_W$ exactly preserves filtration levels.

Let $w\in W$, and write $w=v+x+x'$, where $v\in V$, $x\in X$, and $x'\in X'$, so our goal is to show that $\ell(w)=\ell(x)$. Of course this is trivial if $w=0$, so assume $w\neq 0$. Now

$$\ell(w)=\max\{\ell(x+x'),\ell(v)\}$$

since V is orthogonal to $X\oplus X'$. Since $x+x'=w-v$ and V and W are orthogonal we have $\ell(x+x')=\max\{\ell(v),\ell(w)\}\ge\ell(v)$. So $\ell(w)=\ell(x+x')=\max\{\ell(x),\ell(x')\}$. (In particular x and x' are not both zero.) Now expand $w-v=x+x'$ in terms of the basis $\{x_j\}$ as

$$w-v=\sum_{j=k_i}^{r}\lambda_j x_j.$$

The fact that we can take the sum to start at k_i follows from the definitions of X and X', and the sum terminates at r because $w-v\in V\oplus W\le\operatorname{Im}\partial$. Then $\ell(w-v)=\max\{\ell(\lambda_j x_j)\mid j\in\{k_i,\dots,r\}\}$. By Lemma 4.9, the infimal filtration level of any $\tilde y\in C_1$ such that $\partial\tilde y=x+x'$ is attained by $\tilde y=y+y'$, where $y=\sum_{j=k_i}^{k_{i+1}-1}\lambda_j y_j$ and $y'=\sum_{j=k_{i+1}}^{r}\lambda_j y_j$; by the assumption that $V\oplus W$ is δ–robust for all $\delta<\beta_{k_i}(\partial)$, we will have

$$\ell(y+y')\ge\ell(w-v)+\beta_{k_i}(\partial)=\ell(x+x')+\beta_{k_i}(\partial).$$

Thus by the orthogonality of the bases $\{x_j\}$ and $\{y_j\}$,

$$\beta_{k_i}(\partial) \leq \ell(y+y') - \ell(x+x') = \max\{\ell(y), \ell(y')\} - \max\{\ell(x), \ell(x')\}. \tag{25}$$

Now if we choose j_0 to maximize the quantity $\ell(\lambda_j y_j)$ over all $j \in \{k_{i+1}, \dots, r\}$ we will have

$$\ell(y') = \ell(\lambda_{j_0} y_{j_0}) = \ell(\lambda_{j_0} x_{j_0}) + \beta_{j_0}(\partial) \leq \ell(x') + \beta_{j_0}(\partial).$$

So

$$\ell(y') - \max\{\ell(x), \ell(x')\} \leq \ell(y') - \ell(x') \leq \beta_{j_0}(\partial) < \beta_{k_i}(\partial)$$

since $j_0 \geq k_{i+1}$. Thus in view of (25) we must have $\ell(y) > \ell(y')$ and so by Proposition 2.3 $\ell(y + y') = \ell(y)$. Similarly, choose $i_0 \in \{k_i, \dots, k_{i+1} - 1\}$ to maximize the quantity $\ell(\lambda_j x_j)$, so that $\ell(x) = \ell(\lambda_{i_0} x_{i_0})$. Then

$$\ell(y) - \ell(x) \geq \ell(\lambda_{i_0} y_{i_0}) - \ell(\lambda_{i_0} x_{i_0}) = \beta_{i_0}(\partial).$$

Symmetrically, choose $i_1 \in \{k_i, \dots, k_{i+1} - 1\}$ to maximize the quantity $\ell(y) = \sum_{k_i}^{k_{i+1}-1} \lambda_i y_i$, that is $\ell(y) = \ell(\lambda_{i_1} y_{i_1})$. Then

$$\ell(y) - \ell(x) \leq \ell(\lambda_{i_1} y_{i_1}) - \ell(\lambda_{i_1} x_{i_1}) = \beta_{i_1}(\partial).$$

Because $\beta_{k_i}(\partial) = \cdots = \beta_{k_{i+1}-1}(\partial)$ and $i_0, i_1 \in \{k_i, \dots, k_{i+1} - 1\}$, the above inequalities imply that $\beta_{i_0}(\partial) = \beta_{i_i}(\partial) = \beta_{k_i}(\partial)$. Thus we necessarily have $\ell(y) - \ell(x) = \beta_{k_i}(\partial)$. So we cannot have $\ell(x') > \ell(x)$, since if this were the case then $\ell(y+y') - \ell(x+x') = \ell(y) - \max\{\ell(x), \ell(x')\}$ would be strictly smaller than $\beta_{k_i}(\partial)$, a contradiction to condition (ii). Thus $\ell(x) \geq \ell(x')$. So since we have seen that $\ell(w) = \max\{\ell(x), \ell(x')\}$ this proves that $\ell(w) = \ell(x)$.

Thus the projection $\pi_X \colon C_0 \to X$ associated to the direct sum decomposition $X \oplus (V \oplus X')$ has $\ell(\pi_X w) = \ell(w)$ for all $w \in W$, and in particular it is injective because 0 is the only element with filtration level $-\infty$. So dimensional considerations prove the last statement of the lemma. By Lemma 7.5, this also implies that W is an orthogonal complement to $V \oplus X'$. Since X' is orthogonal to V and $V \oplus X'$ is orthogonal to W it follows from Lemma 2.9(ii) that $V \oplus W$ is orthogonal to X', which is precisely the remaining conclusion of the lemma. □

Corollary 7.8 *Let $((z_1, \dots, z_n), (w_1, \dots, w_m))$ and $((y_1, \dots, y_n), (x_1, \dots, x_m)))$ be two singular value decompositions for $(C_1 \xrightarrow{\partial} C_0)$. Then for each $i \in \{1, \dots, p\}$*

there is a commutative diagram

$$\begin{tikzcd}
\mathrm{span}_\Lambda\{z_{k_i},\dots,z_{k_{i+1}-1}\} \arrow[r] \arrow[d, "\partial"] & \mathrm{span}_\Lambda\{y_{k_i},\dots,y_{k_{i+1}-1}\} \arrow[d, "\partial"] \\
\mathrm{span}_\Lambda\{w_{k_i},\dots,w_{k_{i+1}-1}\} \arrow[r] & \mathrm{span}_\Lambda\{x_{k_i},\dots,x_{k_{i+1}-1}\}
\end{tikzcd}$$

where the horizontal arrows are isomorphisms of filtered vector spaces.

Proof Consider the ascending sequence

$$\{0\} = V_0 \le V_1 \le V_2 \le \cdots \le V_p = \mathrm{Im}\,\partial$$

of subspaces of $\mathrm{Im}\,\partial$, where $V_i = \mathrm{span}\{w_1,\dots,w_{k_{i+1}-1}\}$. Each V_i is δ–robust for all $\delta < \beta_{k_i}(\partial)$ by Lemma 4.9. Also let $W_i = \mathrm{span}_\Lambda\{w_{k_i},\dots,w_{k_{i+1}-1}\}$, so we have an orthogonal direct sum decomposition $V_i = V_{i-1} \oplus W_i$.

We claim by induction on i that V_i is orthogonal to $\mathrm{span}_\Lambda\{x_{k_{i+1}},\dots,x_m\}$. Indeed for $i = 0$ this is trivial, and assuming that it holds for the value $i-1$ then applying Lemma 7.7 with $V = V_{i-1}$ and $W = W_i$ proves the claim for the value i. Given this fact, for any i we may again apply Lemma 7.7 to obtain a filtered isomorphism $W_i \to \mathrm{span}_\Lambda\{x_{k_i},\dots,x_{k_{i+1}-1}\}$, which serves as the bottom arrow in the diagram in the statement of the Corollary.

Since the side arrows and the bottom arrow are all linear isomorphisms, there is a unique top arrow that makes the diagram commute. Moreover the bottom arrow exactly preserves filtration, and the side arrows both decrease the filtration levels of all nonzero elements by *exactly* $\beta_{k_i}(\partial)$, so it follows that the top arrow is an isomorphism of filtered vector spaces as well. □

Proof of Theorem 7.1 Let $((z_1,\dots,z_n),(w_1,\dots,w_m))$, $((y_1,\dots,y_n),(x_1,\dots,x_m))$ be two singular value decompositions. Both

$$\mathrm{span}_\Lambda\{w_{r+1},\dots,w_m\} \quad \text{and} \quad \mathrm{span}_\Lambda\{x_{r+1},\dots,x_m\}$$

are orthogonal complements to $\mathrm{Im}\,\partial$, where $r = \mathrm{rank}(\partial)$, so they are filtered isomorphic by Lemma 7.5 and so they have the same filtration spectra by Proposition 5.5. The subspaces $\mathrm{span}_\Lambda\{w_{k_i},\dots,w_{k_{i+1}-1}\}$ and $\mathrm{span}_\Lambda\{x_{k_i},\dots,x_{k_{i+1}-1}\}$ are filtered isomorphic for each $i \in \{1,\dots,p\}$ by Corollary 7.8, so they likewise have the same filtration spectra. The conclusion now follows immediately from the description of verbose barcode, using Theorem 4.11. □

7.2 Classification up to filtered homotopy equivalence

Now we move on to the classification of the filtered chain homotopy equivalence class of a Floer-type complex. First, we will prove the "if" part, which is the easier direction.

Proposition 7.9 *For any Floer-type complex $(C_*, \partial_C, \ell_C)$, let $\mathcal{B}_{C,k}$ be the degree-k concise barcode of $(C_*, \partial_C, \ell_C)$. For each $([a], L) \in \mathcal{B}_{C,k}$, choose a representative a of the coset $[a] \in \mathbb{R}/\Gamma$. Then $(C_*, \partial_C, \ell_C)$ is filtered homotopy equivalent to*

$$\bigoplus_{k\in\mathbb{Z}} \bigoplus_{([a],L)\in\mathcal{B}_{C,k}} \mathcal{E}(a, L, k).$$

Proof For each k let $\widetilde{\mathcal{B}}_{C,k}$ denote the degree-k verbose barcode of $(C_*, \partial_C, \ell_C)$ and $\mathcal{B}_{C,k}$ the degree-k concise barcode, so $\mathcal{B}_{C,k} = \{([a], L) \in \widetilde{\mathcal{B}}_{C,k} \mid L > 0\}$.

By Proposition 7.4, if for each $([a], L) \in \widetilde{\mathcal{B}}_{C,k}$ we choose a representative a of the coset $[a] \in \mathbb{R}/\Gamma$, $(C_*, \partial_C, \ell_C)$ is filtered chain isomorphic to

$$\left(\bigoplus_{k} \bigoplus_{([a],L)\in\mathcal{B}_{C,k}} \mathcal{E}(a, L, k)\right) \oplus \left(\bigoplus_{k} \bigoplus_{([a],0)\in\widetilde{\mathcal{B}}_{C,k}\setminus\mathcal{B}_{C,k}} \mathcal{E}(a, 0, k)\right). \tag{26}$$

Recall the definition of $\mathcal{E}(a, 0, k)$ as the triple $(E_*, \partial_E, \ell_E)$, where E_* is spanned over Λ by elements $y \in E_{k+1}$ and $x \in E_k$ with $\partial_E y = x$ and $\ell_E(y) = \ell_E(x) = a$. If we define $K\colon E_* \to E_{*+1}$ to be the Λ–linear map defined by $Kx = -y$ and $K|_{E_m} = 0$ for $m \neq k$, we see that $\ell_E(Ke) \leq \ell_E(e)$ for all $e \in E_*$, that $(\partial_E K + K\partial_E)x = -\partial_E y = -x$, and that $(\partial_E K + K\partial_E y) = Kx = -y$. So K defines a filtered chain homotopy between 0 and the identity, in view of which $\mathcal{E}(a, 0, k)$ is filtered homotopy equivalent to the zero chain complex. Since a direct sum of filtered homotopy equivalences is a filtered homotopy equivalence, the Floer-type complex in (26) (and hence also $(C_*, \partial_C, \ell_C)$) is filtered homotopy equivalent to $\bigoplus_{k\in\mathbb{Z}} \bigoplus_{([a],L)\in\mathcal{B}_{C,k}} \mathcal{E}(a, L, k)$. □

Recall from Remark 7.3 that the filtered isomorphism type of $\mathcal{E}(a, L, k)$ only depends on $([a], L, k)$, so that up to filtered chain isomorphism $\bigoplus_{k\in\mathbb{Z}} \bigoplus_{([a],L)\in\mathcal{B}_{C,k}} \mathcal{E}(a, L, k)$ is independent of the choices a of representatives of the cosets $[a]$. In light of this, the "if" part of Theorem B follows directly from Proposition 7.9.

7.2.1 Mapping cylinders We review here the standard homological algebra construction of the mapping cylinder of a chain map between two chain complexes; the special case where the chain map is a homotopy equivalence will be used both in the proof of the "only if" part of Theorem B and in the proof of the stability theorem.

For a chain complex (C_*, ∂_C) we use $(C[1]_*, \partial_C)$ to denote the chain complex obtained by shifting the degree of C_* by 1: $C[1]_k = C_{k-1}$, with boundary operator given tautologically by the boundary operator of C_*.

Definition 7.10 Let (C_*, ∂_C) and (D_*, ∂_D) be two chain complexes over an arbitrary ring, and let $\Phi\colon C_* \to D_*$ be a chain map. The *mapping cylinder* of Φ is the chain complex $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}})$ defined by $\mathrm{Cyl}(\Phi)_* = C_* \oplus D_* \oplus C[1]_*$ and, for $(c, d, e) \in \mathrm{Cyl}(\Phi)_*$, $\partial_{\mathrm{cyl}}(c, d, e) = (\partial_C c - e, \partial_D d + \Phi e, -\partial_C e)$. Thus, in block form,

$$\partial_{\mathrm{cyl}} = \begin{pmatrix} \partial_C & 0 & -I_{C_*} \\ 0 & \partial_D & \Phi \\ 0 & 0 & -\partial_C \end{pmatrix}.$$

It is a routine matter to check that $\partial_{\mathrm{cyl}}^2 = 0$, so $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}})$ as defined above is indeed a chain complex.

For the moment we will work at the level of chain complexes, not of filtered chain complexes, the reason being that we will later use Lemma 7.12 below under a variety of different kinds of assumptions about filtration levels.

Definition 7.11 Given two chain complexes (C_*, ∂_C) and (D_*, ∂_D), a homotopy equivalence between (C_*, ∂_C) and (D_*, ∂_D) is a quadruple (Φ, Ψ, K_C, K_D) such that $K_C\colon C_* \to C_{*+1}$, $K_D\colon D_* \to D_{*+1}$ are linear maps shifting degree by $+1$ and $\Phi\colon C_* \to D_*$, $\Psi\colon D_* \to C_*$ are chain maps, obeying $\Psi\Phi - I_{C_*} = \partial_C K_C + K_C \partial_C$ and $\Phi\Psi - I_{D_*} = \partial_D K_D + K_D \partial_D$.

(In particular our convention is to consider the homotopies part of the data of a homotopy equivalence.)

Lemma 7.12 *Let (Φ, Ψ, K_C, K_D) be a homotopy equivalence between (C_*, ∂_C) and (D_*, ∂_D). Then:*

(i) *Suppose that $i_D\colon D_* \to \mathrm{Cyl}(\Phi)_*$ is the inclusion, $\alpha\colon \mathrm{Cyl}(\Phi)_* \to D_*$ is defined by $\alpha(c, d, e) = \Phi c + d$, and $K\colon \mathrm{Cyl}(\Phi)_* \to \mathrm{Cyl}(\Phi)_{*+1}$ is defined by $K(c, d, e) = (0, 0, c)$. Then the quadruple $(i_D, \alpha, 0, K)$ is a homotopy equivalence between (D_*, ∂_D) and $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}})$.*

(ii) *Suppose that $i_C\colon C_* \to \mathrm{Cyl}(\Phi)_*$ is the inclusion, $\beta\colon \mathrm{Cyl}(\Phi)_* \to C_*$ is defined by $\beta(c, d, e) = c + \Psi d + K_C e$, and $L\colon \mathrm{Cyl}(\Phi)_* \to \mathrm{Cyl}(\Phi)_{*+1}$ is defined by*

$$L(c, d, e) = (-K_C c, K_D(\Phi c + d), c - \Psi(\Phi c + d)).$$

Then the quadruple $(i_C, \beta, 0, L)$ is a homotopy equivalence between (C_*, ∂_C) and $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}})$.

Proof The proof requires only a series of routine computations to show that i_D, α, i_C, β are all chain maps and that the various chain homotopy equations hold. We will do only the most nontrivial of these, namely the proof of the identity $i_C\beta - I_{\mathrm{Cyl}(\Phi)_*} = \partial_{\mathrm{Cyl}}L + L\partial_{\mathrm{Cyl}}$, leaving the rest to the reader. We see that, for $(c,d,e) \in \mathrm{Cyl}(\Phi)_*$,

$$(i_C\beta - I_{\mathrm{Cyl}(\Phi)_*})(c,d,e) = (\Psi d + K_C e, -d, -e)$$

while

$$\begin{aligned}\partial_{\mathrm{cyl}}L(c,d,e) &= \partial_{\mathrm{cyl}}\big(-K_C c, K_D(\Phi c + d), c - \Psi(\Phi c + d)\big)\\ &= \big(-\partial_C K_C c - c + \Psi\Phi c + \Psi d, \partial_D K_D(\Phi c + d) + \Phi c - \Phi\Psi(\Phi c + d),\\ &\qquad -\partial_C c + \partial_C \Psi(\Phi c + d)\big)\\ &= \big(K_C\partial_C c + \Psi d, -K_D\partial_D\Phi c + (\partial_D K_D - \Phi\Psi)d,\\ &\qquad -\partial_C c + \partial_C\Psi(\Phi c + d)\big),\end{aligned}$$

where we have used the facts that $\Psi\Phi - I_{C_*} = \partial_C K_C + K_C\partial_C$ and $\Phi\Psi - I_{D_*} = \partial_D K_D + K_D\partial_D$. Furthermore,

$$\begin{aligned}L\partial_{\mathrm{cyl}}(c,d,e) &= L(\partial_C c - e, \partial_D d + \Phi e, -\partial_C e)\\ &= \big(-K_C\partial_C c + K_C e, K_D(\Phi\partial_C c + \partial_D d), \partial_C c - e - \Psi(\Phi\partial_C c + \partial_D d)\big).\end{aligned}$$

So

$$\begin{aligned}(\partial_{\mathrm{cyl}}L + L\partial_{\mathrm{cyl}})(c,d,e) &= \big(\Psi d + K_C e, (\partial_D K_D - \Phi\Psi + K_D\partial_D)d, -e\big)\\ &= (\Psi d + K_C e, -d, -e) = (i_C\beta - I_{\mathrm{Cyl}(\Phi)_*})(c,d,e),\end{aligned}$$

where in the first equation we have used the fact that Φ and Ψ are chain maps and in the second equation we have again used that $\Phi\Psi - I_{D_*} = \partial_D K_D + K_D\partial_D$. So indeed $i_C\beta - I_{\mathrm{Cyl}(\Phi)_*} = \partial_{\mathrm{Cyl}}L + L\partial_{\mathrm{Cyl}}$; as mentioned earlier the remaining identities are easier to prove and so are left to the reader. □

We can now fill in the last part of our proofs of the main classification results.

Proof of Theorem B One implication has already been proven in Proposition 7.9. For the other direction, let $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ be two filtered homotopy equivalent Floer-type complexes. Thus there is a homotopy equivalence (Φ, Ψ, K_C, K_D) satisfying the additional properties that, for all $c \in C_*$ and $d \in D_*$, we have

$$\ell_D(\Phi c) \le \ell_C(c), \quad \ell_C(\Psi d) \le \ell_D(d), \quad \ell_C(K_C c) \le \ell_C(c), \quad \ell_D(K_D d) \le \ell_D(d). \tag{27}$$

Now form the mapping cylinder $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}})$ as described earlier, and define $\ell_{\mathrm{cyl}}\colon \mathrm{Cyl}(\Phi)_* \to \mathbb{R} \cup \{-\infty\}$ by

$$\ell_{\mathrm{cyl}}(c,d,e) = \max\{\ell_C(c), \ell_D(d), \ell_C(e)\}.$$

It is easy to see that $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_{\mathrm{cyl}})$ is then a Floer-type complex.[4] Moreover, $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_{\mathrm{cyl}})$ has a concise barcode in each degree; we will show that this concise barcode is both the same as that of $(C_*, \partial_C, \ell_C)$ and the same as that of $(D_*, \partial_D, \ell_D)$, which will suffice to prove the result.

Using the notation of Lemma 7.12, since $\alpha\colon \mathrm{Cyl}(\Phi)_* \to D_*$ is a chain map with $\alpha i_D = I_{D_*}$, we have a direct sum decomposition *of chain complexes* $\mathrm{Cyl}(\Phi)_* = D_* \oplus \ker\alpha$. We claim that D_* and $\ker\alpha$ are orthogonal (with respect to the filtration function ℓ_{cyl}). Now

$$\ker\alpha = \bigl\{(c,d,e) \in \mathrm{Cyl}(\Phi)_* \mid d = -\Phi c\bigr\} = \bigl\{(c,-\Phi c,e) \mid (c,e) \in C_* \oplus C[1]_*\bigr\}.$$

Since D_* is an orthogonal complement to $C_* \oplus C[1]_*$ in $\mathrm{Cyl}(\Phi)_*$, and since in each grading k the dimensions of the degree-k part of $\ker\alpha$ and of $C_k \oplus C[1]_k$ are the same, by Lemma 7.5 in order to show that $\ker\alpha$ is orthogonal to D_* it suffices to show that, writing $\pi\colon \mathrm{Cyl}(\Phi)_* \to C_* \oplus C[1]_*$ for the orthogonal projection $(c,d,e) \mapsto (c,e)$, one has $\ell_{\mathrm{cyl}}(\pi x) = \ell_{\mathrm{cyl}}(x)$ for all $x \in \ker\alpha$. But any $x \in \ker\alpha$ has $x = (c,-\Phi c,e)$ for some $(c,e) \in C_* \oplus C[1]_*$, and $\ell_D(-\Phi c) \le \ell_C(c)$, so we indeed have $\ell_{\mathrm{cyl}}(\pi x) = \max\{\ell_C(c), \ell_C(e)\} = \ell_{\mathrm{cyl}}(x)$. So indeed D_* and $\ker\alpha$ are orthogonal.

In view of the orthogonal direct sum decomposition of chain complexes $\mathrm{Cyl}(\Phi)_* = D_* \oplus \ker\alpha$, for every degree k we can obtain a singular value decomposition for $(\partial_{\mathrm{cyl}})_{k+1}\colon \mathrm{Cyl}(\Phi)_{k+1} \to \ker(\partial_{\mathrm{cyl}})_k$ by simply combining singular value decompositions for the restrictions of $(\partial_{\mathrm{cyl}})_{k+1}$ to D_{k+1} and to $(\ker\alpha)_{k+1}$. Then by Theorem 7.1, the verbose barcode of $\mathrm{Cyl}(\Phi)_*$ is the union of the verbose barcodes of D_* and of $\ker\alpha$.

To describe the latter of these, we will show presently that every element in $\ker(\partial_{\mathrm{cyl}}|_{\ker\alpha})$ is the boundary of an element having the same filtration level. In fact, for any $x \in \ker(\partial_{\mathrm{cyl}}|_{\ker\alpha})$, the equation $i_D\alpha - I_{\mathrm{Cyl}(\Phi)_*} = \partial_{\mathrm{cyl}}K + K\partial_{\mathrm{cyl}}$ shows that $x = \partial_{\mathrm{cyl}}(-Kx)$. Moreover,

$$\ell_{\mathrm{cyl}}(x) = \ell_{\mathrm{cyl}}(\partial_{\mathrm{cyl}}(-Kx)) \le \ell_{\mathrm{cyl}}(-Kx) \le \ell_{\mathrm{cyl}}(x),$$

where the last inequality comes from the formula for K in Lemma 7.12. Therefore $\ell_{\mathrm{cyl}}(x) = \ell_{\mathrm{cyl}}(-Kx)$.

[4] For comparison with what we do later it is worth noting that the fact that $\ell_{\mathrm{cyl}}(\partial_{\mathrm{cyl}}x) \le \ell_{\mathrm{cyl}}(x)$ for all x is crucially dependent on the first inequality of (27).

Consequently, every element $([a], s)$ of the verbose barcode of $\ker\alpha$ has $s = 0$ (or, said differently, the concise barcode of $\ker\alpha$ is empty in every degree). Thus the verbose barcode of $\mathrm{Cyl}(\Phi)_*$ may be obtained from the verbose barcode of D_* by adding elements with second coordinate equal to zero; consequently the concise barcodes of $\mathrm{Cyl}(\Phi)_*$ and of D_* are equal.

The proof that the concise barcodes of $\mathrm{Cyl}(\Phi)_*$ and C_* are likewise equal is very similar. We have a direct sum decomposition of chain complexes $\mathrm{Cyl}(\Phi)_* = C_* \oplus \ker\beta$, where $\ker\beta = \{(-\Psi d - K_C e, d, e) \mid (d, e) \in D_* \oplus C[1]_*\}$. Let $\pi'\colon \mathrm{Cyl}(\Phi)_* \to D_* \oplus C[1]_*$ be the projection associated to the orthogonal direct sum decomposition $\mathrm{Cyl}(\Phi)_* = C_* \oplus (D_* \oplus C[1]_*)$. The inequalities (27) imply that $\ell_{\mathrm{cyl}}(\pi' x) = \ell_{\mathrm{cyl}}(x)$ for all $x \in \ker\beta$. Hence by applying Lemma 7.5 degree-by-degree we see that $\mathrm{Cyl}(\Phi)_* = C_* \oplus \ker\beta$ is an *orthogonal* direct sum decomposition of chain complexes, and hence that in any degree k the verbose barcode of $\mathrm{Cyl}(\Phi)_*$ is the union of the degree-k verbose barcodes of C_* and of $\ker\beta$. Any cycle x in $\ker\beta$ obeys $x = -\partial_{\mathrm{cyl}} Lx$, where the formula for L (together with (27)) shows that $\ell_{\mathrm{cyl}}(-Lx) \le \ell_{\mathrm{cyl}}(x)$. While Lx might not be an element of $\ker\beta$, the orthogonality of C_* and $\ker\beta$ together with Lemma 4.9 allow one to find $y \in \ker\beta$ with $\partial y = x$ and $\ell_{\mathrm{cyl}}(y) \le \ell_{\mathrm{cyl}}(-Lx) \le \ell_{\mathrm{cyl}}(x)$. Just as above, this proves that all elements $([a], s)$ of the verbose barcode of $\ker\beta$ have second coordinate s equal to zero, and so once again the concise barcode of $\mathrm{Cyl}(\Phi)_*$ coincides with that of C_*. □

8 The stability theorem

The stability theorem (or a closely related statement sometimes called the isometry theorem) is the one of the most important theorems in the theory of persistent homology. It successfully transfers the problem of relating the filtered homology groups constructed by different methods (eg different Morse functions on a given manifold) to a combinatorial problem based on the associated barcodes. The result was originally established for the persistence modules associated to "tame" functions on topological spaces in [10]; since then a variety of different proofs and generalizations have appeared (see eg [8; 3]), and it now generally understood as an algebraic statement in the abstract context of persistence modules. In this section, we will introduce some basic notations and definitions in order to state our version of the stability theorem, which unlike previous versions applies to Floer-type complexes over general Novikov fields $\Lambda^{\mathcal{K},\Gamma}$. In the special case that $\Gamma = \{0\}$ the result follows from recent more algebraic formulations of the stability theorem like that in [3], though we would say that our proof is conceptually rather different.

The following is an abstraction of the filtration-theoretic properties satisfied by the "continuation maps" in Hamiltonian Floer theory that relate the Floer-type complexes associated to different Hamiltonian functions; namely such maps are homotopy equivalences which shift the filtration by a certain amount which is related to an appropriate distance (the Hofer distance) between the Hamiltonians (see [45, Propositions 5.1, 5.3 and 6.1]).

Definition 8.1 Let $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ be two Floer-type complexes over Λ, and $\delta \geq 0$. A *δ-quasiequivalence* between C_* and D_* is a quadruple (Φ, Ψ, K_C, K_D), where:

(i) (Φ, Ψ, K_C, K_D) is a homotopy equivalence (see Definition 7.11).

(ii) For all $c \in C_*$ and $d \in D_*$ we have

$$\begin{aligned} \ell_D(\Phi c) &\leq \ell_C(c) + \delta, & \ell_C(\Psi d) &\leq \ell_D(d) + \delta, \\ \ell_C(K_C c) &\leq \ell_C(c) + 2\delta, & \ell_D(K_D d) &\leq \ell_D(d) + 2\delta. \end{aligned} \tag{28}$$

The *quasiequivalence distance* between $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ is then defined to be

$$d_Q\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big) = \inf\big\{\delta \geq 0 \mid \exists\ \delta\text{–quasiequivalence between } (C_*, \partial_C, \ell_C) \text{ and } (D_*, \partial_D, \ell_D)\big\}.$$

Of course, $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ are said to be δ–quasiequivalent provided that there exists a δ–quasiequivalence between them. Note that a 0–quasiequivalence is the same thing as a filtered homotopy equivalence.

Remark 8.2 It is easy to see that if $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ are δ_0–quasiequivalent and $(D_*, \partial_D, \ell_D)$ and $(E_*, \partial_E, \ell_E)$ are δ_1–quasiequivalent then $(C_*, \partial_C, \ell_C)$ and $(E_*, \partial_E, \ell_E)$ are $(\delta_0 + \delta_1)$–quasiequivalent. Thus d_Q satisfies the triangle inequality. In particular, if $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ are δ–quasiequivalent then $(C_*, \partial_C, \ell_C)$ is also δ–quasiequivalent to any Floer-type complex that is filtered homotopy equivalent to $(D_*, \partial_D, \ell_D)$.

Example 8.3 Take (F_1, g_1) and (F_2, g_2) to be two Morse functions together with suitably generic Riemannian metrics on a closed manifold X. Let $\delta = \|F_1 - F_2\|_{L^\infty}$. Then it is well-known (and can be deduced from constructions in [39], for instance) that the associated Morse chain complexes, over the ground field $\mathcal{K} = \Lambda^{\mathcal{K},\{0\}}$, $\mathrm{CM}_*(X; F_1, g_1)$ and $\mathrm{CM}_*(X; F_2, g_2)$ are δ–quasiequivalent.

Example 8.4 Take (H_1, J_1) and (H_2, J_2) to be two generic Hamiltonian functions together with compatible almost complex structures on a closed symplectic manifold (M, ω). Then, as is recalled in greater detail at the start of Section 12, one has Hamiltonian Floer complexes $(\mathrm{CF}_*(M; H_1, J_1))$ and $(\mathrm{CF}_*(M; H_2, J_2))$ over the Novikov field $\Lambda^{\mathcal{K},\Gamma}$ where $\Gamma \leq \mathbb{R}$ is defined in (40). Define

$$E_+(H) = \int_0^1 \max_M H(t,\cdot)\,dt \quad \text{and} \quad E_-(H) = -\int_0^1 \min_M H(t,\cdot)\,dt$$

and let $\delta = \max\{E_+(H_2 - H_1), E_-(H_2 - H_1)\}$. Then the Hamiltonian Floer complexes $(\mathrm{CF}_*(M; H_1, J_1))$ and $(\mathrm{CF}_*(M; H_2, J_2))$ are δ–quasiequivalent. The maps in the corresponding quadruple (Φ, Ψ, K_1, K_2) are constructed by counting solutions of certain partial differential equations (see [1, Chapter 11]).

Remark 8.5 One could more generally define, for $\delta_1, \delta_2 \in \mathbb{R}$, a (δ_1, δ_2)–quasi-equivalence by replacing (28) by the conditions

$$\begin{aligned} \ell_D(\Phi c) &\leq \ell_C(c) + \delta_1, & \ell_C(\Psi d) &\leq \ell_D(d) + \delta_2, \\ \ell_C(K_C c) &\leq \ell_C(c) + \delta_1 + \delta_2, & \ell_D(K_D d) &\leq \ell_D(d) + \delta_1 + \delta_2. \end{aligned}$$

(So in this language a δ–quasiequivalence is the same as a (δ, δ)–quasiequivalence.) Then in Example 8.4 one has the somewhat sharper statement that $(\mathrm{CF}_*(M; H_1, J_1))$ and $(\mathrm{CF}_*(M; H_2, J_2))$ are $(E_+(H_2 - H_1), E_-(H_2 - H_1))$–quasiequivalent. However since adding a suitable constant to H_1 has the effect of reducing to the case that $E_+(H_2 - H_1)$ and $E_-(H_2 - H_1)$ are equal to each other while changing the filtration on the Floer complex (and hence changing the barcode) by a simple uniform shift, for ease of exposition we will restrict attention to the more symmetric case of a δ–quasiequivalence.

Remark 8.6 We will explain in the appendix that quasiequivalence is closely related with the notion of *interleaving* of persistent homology from [3]. In particular, the quasiequivalence distance d_Q is equal to a natural chain-level version of the interleaving distance from [3].

Our first step toward the stability theorem will be a continuity result for the quantities β_k from Definition 4.10. Recall that for $i \in \mathbb{Z}$ the degree-i part of the (verbose or concise) barcode of $(C_*, \partial_C, \ell_C)$ is obtained from a singular value decomposition of the map $(\partial_C)_{i+1}\colon C_{i+1} \to \ker(\partial_C)_i$.

Lemma 8.7 *Let (Φ, Ψ, K_C, K_D) be a δ–quasiequivalence and let $\eta \geq 2\delta$. If $V \leq \ker(\partial_C)_i$ is η–robust then $\Phi|_V$ is injective and $\Phi(V)$ is $(\eta - 2\delta)$–robust.*

Proof If $v \in V$ and $\Phi v = 0$ then

$$v = v - \Psi\Phi v = \partial_C(-K_C v),$$

where $\ell_C(-K_C v) \le \ell_C(v) + 2\delta$; by the definition of η–robustness (see Definition 4.7) this implies that $v = 0$ since $\eta \ge 2\delta$. So indeed $\Phi|_V$ is injective.

Now suppose that $0 \ne w = \Phi v \in \Phi(V)$ with $\partial_D y = w$. Then

$$\partial_C \Psi y = \Psi \partial_D y = \Psi\Phi v = v + \partial_C K_C v$$

(where we've used the fact that $V \le \ker \partial_C$). So $v = \partial_C(\Psi y - K_C v)$. By the definition of η–robustness we have $\ell_C(\Psi y - K_C v) > \ell_C(v) + \eta$. Since $\ell_C(K_C v) \le \ell_C(v) + 2\delta \le \ell_C(v) + \eta$ this implies that

$$\ell_C(\Psi y) > \ell_C(v) + \eta.$$

But $\ell_D(y) \ge \ell_C(\Psi y) - \delta$, and $\ell_D(w) = \ell_D(\Phi v) \le \ell_C(v) + \delta$, which combined with the displayed inequality above shows that $\ell_D(y) > \ell_D(w) + (\eta - 2\delta)$. Since w was an arbitrary nonzero element of $\Phi(V)$ this proves that $\Phi(V)$ is $(\eta - 2\delta)$–robust. □

Corollary 8.8 *Suppose that $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ are δ–quasiequivalent. Then for all $i \in \mathbb{Z}$ and $k \in \mathbb{N}$, we have $|\beta_k((\partial_C)_{i+1}) - \beta_k((\partial_D)_{i+1})| \le 2\delta$.*

Proof By definition $\beta_k((\partial_C)_{i+1})$ is the supremal $\eta \ge 0$ such that there exists a k–dimensional η–robust subspace of $\mathrm{Im}((\partial_D)_{i+1})$, or is zero if no such subspace exists for any η. If $\beta_k((\partial_C)_{i+1}) > 2\delta$, then given $\epsilon > 0$ there is a k–dimensional subspace $V \le \mathrm{Im}(\partial_C)_{i+1}$ which is $(\beta_k((\partial_C)_{i+1}) - \epsilon)$–robust, and then (for small enough ϵ) Lemma 8.7 shows that $\Phi(V) \le \mathrm{Im}((\partial_D)_{i+1})$ is k–dimensional and $(\beta_k((\partial_C)_{i+1}) - \epsilon - 2\delta)$–robust. Since this construction applies for all sufficiently small $\epsilon > 0$ it follows that

$$\beta_k((\partial_D)_{i+1}) \ge \beta_k((\partial_C)_{i+1}) - 2\delta \tag{29}$$

provided that $\beta_k((\partial_C)_{i+1}) > 2\delta$. But of course if $\beta_k((\partial_C)_{i+1}) \le 2\delta$ then (29) still holds for the trivial reason that $\beta_k((\partial_D)_{i+1})$ is by definition nonnegative. So (29) holds in any case. But this argument may equally well be applied with the roles of the complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ reversed (as the relation of δ–quasiequivalence is symmetric), yielding $\beta_k((\partial_C)_{i+1}) \ge \beta_k((\partial_D)_{i+1}) - 2\delta$, which together with (29) directly implies the corollary. □

In order to state our stability theorem we must explain the bottleneck distance, which is a measurement of the distance between two barcodes in common use at least since [10]. First we will define some notions related to matchings between multisets, similar to what can be found in eg [9]. We initially express this in rather general terms in order to make clear that our notion of a partial matching can be identified with corresponding notions found elsewhere in the literature. Recall below that a pseudometric space is a generalization of a metric space in which two distinct points are allowed to be a distance zero away from each other, and an extended pseudometric space is a generalization of a pseudometric space in which the distance between two points is allowed to take the value ∞.

Definition 8.9 Let (X,d) be an extended pseudometric space equipped with a "length function" $\lambda\colon X\to[0,\infty]$, and let $\mathcal{S}$ and $\mathcal{T}$ be two multisets of elements of X.

- A *partial matching* between $\mathcal{S}$ and $\mathcal{T}$ is a triple $\mathfrak{m}=(\mathcal{S}_{\text{short}},\mathcal{T}_{\text{short}},\sigma)$ where $\mathcal{S}_{\text{short}}$ and $\mathcal{T}_{\text{short}}$ are submultisets of $\mathcal{S}$ and $\mathcal{T}$, respectively, and $\sigma\colon \mathcal{S}\setminus\mathcal{S}_{\text{short}}\to\mathcal{T}\setminus\mathcal{T}_{\text{short}}$ is a bijection. (The elements of $\mathcal{S}_{\text{short}}$ and $\mathcal{T}_{\text{short}}$ will sometimes be called "unmatched".)

- For $\delta\in[0,\infty]$, a *δ–matching* between $\mathcal{S}$ and $\mathcal{T}$ is a partial matching $(\mathcal{S}_{\text{short}},\mathcal{T}_{\text{short}},\sigma)$ such that for all $x\in\mathcal{S}_{\text{short}}\cup\mathcal{T}_{\text{short}}$ we have $\lambda(x)\le\delta$ and for all x in $\mathcal{S}\setminus\mathcal{S}_{\text{short}}$ we have $d(\sigma(x),x)\le\delta$.

- If $\mathfrak{m}$ is a partial matching between $\mathcal{S}$ and $\mathcal{T}$, the *defect* of $\mathfrak{m}$ is
$$\delta(\mathfrak{m})=\inf\{\delta\ge 0\mid \mathfrak{m}\text{ is a }\delta\text{–matching}\}.$$

Example 8.10 Let $\mathcal{H}=\{(x,y)\in(-\infty,\infty]^2\mid x<y\}$ with extended metric
$$d_{\mathcal{H}}((a,b),(c,d))=\max\{|c-a|,|d-b|\}$$
and $\lambda_{\mathcal{H}}((a,b))=\frac{1}{2}(b-a)$. Then our notion of a δ–matching between multisets of elements of $\mathcal{H}$ is readily verified to be the same as that used in [9, Section 4] or [3, Section 3.2].

Example 8.11 Consider $\mathbb{R}\times(0,\infty]$ with the extended metric
$$d((a,L),(a',L'))=\max\{|a-a'|,|(a+L)-(a'+L')|\}$$
and the length function $\lambda(a,L)=L/2$. Then the bijection $f\colon\mathbb{R}\times(0,\infty]\to\mathcal{H}$ defined by $f(a,L)=(a,a+L)$ pulls back $d_{\mathcal{H}}$ and $\lambda_{\mathcal{H}}$ from the previous example to d and λ, respectively, so giving a δ–matching $\mathfrak{m}$ between multisets of elements of $\mathbb{R}\times(0,\infty]$ is equivalent to giving a δ–matching $f_*\mathfrak{m}$ between the corresponding multisets of elements of $\mathcal{H}$.

Example 8.12 Our main concern will be δ–matchings between concise barcodes of Floer-type complexes, which are by definition multisets of elements of $(\mathbb{R}/\Gamma)\times(0,\infty]$

for a subgroup $\Gamma \leq \mathbb{R}$. For this purpose we use the length function $\lambda\colon (\mathbb{R}/\Gamma)\times(0,\infty]\to\mathbb{R}$ defined by $\lambda([a],L) = L/2$ and the extended pseudometric

$$d(([a],L),([a'],L')) = \inf_{g\in\Gamma} \max\{|a+g-a'|, |(a+g+L)-(a'+L')|\}.$$

In the case that $\Gamma = \{0\}$ this evidently reduces to Example 8.11.

For convenience, we rephrase the definition of a δ–matching between concise barcodes:

Definition 8.13 Consider two concise barcodes $\mathcal{S}$ and $\mathcal{T}$ (viewed as multisets of elements of $(\mathbb{R}/\Gamma)\times(0,\infty]$). A *$\delta$–matching* between $\mathcal{S}$ and $\mathcal{T}$ consists of the following data:

(i) submultisets $\mathcal{S}_{\text{short}}$ and $\mathcal{T}_{\text{short}}$ such that the second coordinate L of every element $([a],L) \in \mathcal{S}_{\text{short}} \cup \mathcal{T}_{\text{short}}$ obeys $L \leq 2\delta$.

(ii) A bijection $\sigma\colon \mathcal{S}\setminus\mathcal{S}_{\text{short}} \to \mathcal{T}\setminus\mathcal{T}_{\text{short}}$ such that, for each $([a],L) \in \mathcal{S}\setminus\mathcal{S}_{\text{short}}$ (where $a\in\mathbb{R}$, $L\in[0,\infty]$) we have $\sigma([a],L) = ([a'],L')$, where for all $\epsilon > 0$ the representative a' of the coset $[a'] \in \mathbb{R}/\Gamma$ can be chosen such that both $|a'-a| \leq \delta+\epsilon$ and either $L = L' = \infty$ or $|(a'+L')-(a+L)| \leq \delta+\epsilon$.

It follows from the discussion in Example 8.11 that our definition agrees in the case that $\Gamma = \{0\}$ (via the map $(a,L)\mapsto(a,a+L)$) to the definitions in, for example, [9] or [3].

Definition 8.14 If $\mathcal{S}$ and $\mathcal{T}$ are two multisets of elements of $(\mathbb{R}/\Gamma)\times(0,\infty]$ then the *bottleneck distance* between $\mathcal{S}$ and $\mathcal{T}$ is

$$d_B(\mathcal{S},\mathcal{T}) = \inf\{\delta \geq 0 \mid \text{There exists a } \delta\text{–matching between } \mathcal{S} \text{ and } \mathcal{T}\}.$$

Our constructions associate to a Floer-type complex a concise barcode *for every* $k\in\mathbb{Z}$, so the appropriate notion of distance for this entire collection of data is:

Definition 8.15 Let $\mathcal{S} = \{\mathcal{S}_k\}_{k\in\mathbb{Z}}$ and $\mathcal{T} = \{\mathcal{T}_k\}_{k\in\mathbb{Z}}$ be two families of multisets of elements of $(\mathbb{R}/\Gamma)\times(0,\infty]$. The bottleneck distance between $\mathcal{S}$ and $\mathcal{T}$ is then

$$d_B(\mathcal{S},\mathcal{T}) = \sup_{k\in\mathbb{Z}} d_B(\mathcal{S}_k,\mathcal{T}_k).$$

Remark 8.16 It is routine to check that d_B is indeed an extended pseudometric. In particular, it satisfies the triangle inequality.

We can now formulate another of this paper's main results, the stability theorem.

Theorem 8.17 (stability theorem) *Given a Floer-type complex $(C_*, \partial_C, \ell_C)$ and $k \in \mathbb{Z}$, denote its degree-k concise barcode by $\mathcal{B}_{C,k}$; moreover let $\mathcal{B}_C = \{\mathcal{B}_{C,k}\}_{k\in\mathbb{Z}}$ denote the indexed family of concise barcodes for all gradings k. Then the bottleneck and quasiequivalence distances obey, for any two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$, the inequality*

$$d_B(\mathcal{B}_C, \mathcal{B}_D) \le 2d_Q\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big). \tag{30}$$

Moreover, for any $k \in \mathbb{Z}$, if we let $\Delta_{D,k} > 0$ denote the smallest second coordinate L of all of the elements of $\mathcal{B}_{D,k}$, and if $d_Q((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)) < \frac{1}{4}\Delta_{D,k}$, then

$$d_B(\mathcal{B}_{C,k}, \mathcal{B}_{D,k}) \le d_Q\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big). \tag{31}$$

We will also prove an inequality in the other direction, analogous to [9, (4.11″)].

Theorem 8.18 (converse stability theorem) *With the same notation as in Theorem 8.17, we have an inequality*

$$d_Q\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big) \le d_B(\mathcal{B}_C, \mathcal{B}_D).$$

Thus, with respect to the quasiequivalence and bottleneck distances, the map from Floer-type complexes to concise barcodes is globally at least bi-Lipschitz, and moreover is a local isometry (at least among complexes having a uniform positive lower bound on the parameters $\Delta_{D,k}$ as k varies through $\mathbb{Z}$; for instance this is true for the Hamiltonian Floer complexes). We expect that the factor of two in (30) is unnecessary so that the map is always a global isometry (as is the case when Γ in trivial by [9, Theorem 4.11]). In Section 11, we will see this becomes true if the quasiequivalence distance d_Q is replaced by more complicated distance called the *interpolating distance*.

We prove the stability theorem in the following section, and the (easier) converse stability theorem in Section 10.

9 Proof of the stability theorem

9.1 Varying the filtration

The proof of the stability theorem will involve first estimating the bottleneck distance between two Floer-type complexes having the same underlying chain complex but different filtration functions, and then using a mapping cylinder construction to reduce the general case to this special case. We begin with a simple combinatorial lemma:

Lemma 9.1 *Suppose that A and B are finite sets and that $\sigma, \tau\colon A \to B$ are bijections and $f\colon A \to \mathbb{R}$ and $g\colon B \to \mathbb{R}$ are functions such that, for some $\delta \geq 0$, we have $f(a) - g(\sigma(a)) \leq \delta$ and $g(\tau(a)) - f(a) \leq \delta$ for all $a \in A$. Then there is a bijection $\eta\colon A \to B$ such that $|f(a) - g(\eta(a))| \leq \delta$ for all $a \in A$.*

Proof Denote the elements of A as $a_1, \dots, a_n$, ordered in such a way that $f(a_1) \leq \cdots \leq f(a_n)$; likewise denote the elements of B as $b_1, \dots, b_n$, ordered such that $g(b_1) \leq \cdots \leq g(b_n)$. Our bijection $\eta\colon A \to B$ will then be given by $\eta(a_i) = b_i$ for $i = 1, \dots, n$.

Given $i \in \{1, \dots, n\}$, write $\tau(a_i) = b_m$ and suppose first that $m \geq i$. Then $g(b_m) \geq g(b_i)$, so $g(b_i) - f(a_i) \leq g(b_m) - f(a_i) \leq \delta$ by the hypothesis on τ. On the other hand if $m < i$ then there must be some $j \in \{1, \dots, i-1\}$ such that $\tau(a_j) = b_k$ with $k \geq i$ (for otherwise τ would give a bijection between $\{a_1, \dots, a_i\}$ and a subset of $\{b_1, \dots, b_{i-1}\}$). In this case since $j < i \leq k$ we have

$$g(b_i) - f(a_i) \leq g(b_k) - f(a_j) = g(\tau(a_j)) - f(a_j) \leq \delta.$$

So in any event $g(b_i) - f(a_i) \leq \delta$ for all i. A symmetric argument (using σ^{-1} in place of τ) shows that likewise $f(a_i) - g(b_i) \leq \delta$ for all i. So indeed our permutation η defined by $\eta(a_i) = b_i$ obeys $|f(a) - g(\eta(a))| \leq \delta$ for all $a \in A$. □

Lemma 9.2 *Let (C, ℓ_C), (D, ℓ_D) be orthogonalizable Λ–spaces and $A\colon C \to D$ a Λ–linear map with unsorted singular value decomposition $((y_1, \dots, y_n), (x_1, \dots, x_m))$. Let $\ell'_D\colon D \to \mathbb{R} \cup \{-\infty\}$ be another filtration function such that (D, ℓ'_D) is an orthogonalizable Λ–space, and let $\delta > 0$ be such that $|\ell_D(d) - \ell'_D(d)| \leq \delta$ for all $d \in D$. Then there is an unsorted singular value decomposition $((y'_1, \dots, y'_n), (x'_1, \dots, x'_m))$ for the map A with respect to ℓ_C and the new filtration function ℓ'_D, such that:*

(i) $\ell_C(y'_i) = \ell_C(y_i)$ *for each i.*

(ii) $|\ell'_D(x'_i) - \ell_D(x_i)| \leq \delta$ *for each $i \leq \mathrm{rank}(A)$.*

Proof To simplify matters later, we shall assume the following:

(32) for all i, j, if $\ell_C(y_i) \equiv \ell_C(y_j) \bmod \Gamma$, then $\ell_C(y_i) = \ell_C(y_j)$.

There is no loss of generality in this assumption, as it may be arranged to hold by multiplying the various y_i, x_i by appropriate field elements T^{g_i} (and then correspondingly multiplying the elements y'_i, x'_i constructed in the proof of the lemma by T^{-g_i}).

Let us first apply the algorithm described in Theorem 3.5 to A, viewed as a map between the nonarchimedean normed vector spaces (C, ℓ_C) and (D, ℓ'_D). That algorithm takes as input orthonormal bases for both the domain and the codomain of A; for the domain (C, ℓ_C) we use the ordered basis $(y_1, \dots, y_n)$ from the given singular value decomposition (for A as a map from (C, ℓ_C) to (D, ℓ_D)), while we use an arbitrary orthogonal basis for the codomain.

Denote the rank of A by r. Since $Ay_i = 0$ for $i = r+1, \dots, n$, inspection of the algorithm in the proof of Theorem 3.5 shows that, for $i = r+1, \dots, m$, the element y_i is unchanged throughout the running of the algorithm. Thus the ordered basis $(y'_1, \dots, y'_n)$ for C that is output by the algorithm has $y'_i = y_i$ for $i = r+1, \dots, m$. So since r is the rank of A and $Ay'_i = Ay_i = 0$ for $i > r$, it follows that $Ay'_i \neq 0$ for $i \in \{1, \dots, r\}$. In fact, setting $x'_i = Ay'_i$ for $i \in \{1, \dots, r\}$, the tuple $(x'_1, \dots, x'_r)$ gives an orthogonal ordered basis for $\operatorname{Im}(A)$. Moreover, according to Theorem 3.5, we have $\ell_C(y'_i) = \ell_C(y_i)$ for all i, while

$$\ell'_D(x'_i) \leq \ell'_D(x_i) \quad \text{for } i \in \{1, \dots, r\}. \tag{33}$$

Taking $(x'_{r+1}, \dots, x'_m)$ to be an arbitrary ℓ'_D–orthogonal basis for an orthogonal complement to $\operatorname{Im}(A)$, it follows that $((y'_1, \dots, y'_n), (x'_1, \dots, x'_m))$ is an unsorted singular value decomposition for A considered as a map from (C, ℓ_C) to (D, ℓ'_D), which moreover satisfies property (i) in the statement of the lemma.

We will show that, possibly after replacing y'_i, x'_i by $y'_{\eta(i)}, x'_{\eta(i)}$ for some permutation η of $\{1, \dots, r\}$ having $\ell_C(y_i) = \ell_C(y_{\eta(i)})$ for each i, this singular value decomposition also satisfies property (ii). In this direction, symmetrically to the previous paragraph, apply the algorithm from Theorem 3.5 to A as a map from (C, ℓ_C) to (D, ℓ_D), using as input the basis $(y'_1, \dots, y'_n)$ for C that we obtained above. This yields a new unsorted singular value decomposition $((y''_1, \dots, y''_n), (x''_1, \dots, x''_m))$ for A as a map from (C, ℓ_C) to (D, ℓ_D), having

$$\ell_C(y''_i) = \ell_C(y'_i) = \ell_C(y_i) \quad \text{for all } i$$

and

$$\ell_D(x''_i) \leq \ell_D(x'_i) \quad \text{for } i \in \{1, \dots, r\}. \tag{34}$$

Now by Theorem 7.1 and our assumption (32), there is an equality of multisets of elements of $\mathbb{R}^2$:

$$\{(\ell_C(y_i), \ell_D(x_i)) \mid i = 1, \dots, r\} = \{(\ell_C(y''_i), \ell_D(x''_i)) \mid i = 1, \dots, r\}. \tag{35}$$

Indeed, each of these multisets corresponds to the finite-length bars in the verbose barcode of the two-term Floer-type complex $(C \xrightarrow{A} D)$, and the condition (32) and the fact that $\ell_C(y''_i) = \ell_C(y_i)$ ensure that an equality of some $\ell_C(y_i)$ and $\ell_C(y''_j)$

modulo Γ implies an equality in $\mathbb{R}$. For any $z \in \{\ell_C(y_1), \dots, \ell_C(y_r)\}$, let

$$I_z = \{i \in \{1, \dots, r\} \mid \ell_C(y_i) = z\}$$

and define functions $f, g\colon I_z \to \mathbb{R}$ by $f(i) = \ell'_D(x'_i)$ and $g(i) = \ell_D(x_i)$. Using (33), for each $i \in I_z$ we then have

$$f(i) \le \ell'_D(x_i) \le \ell_D(x_i) + \delta = g(i) + \delta.$$

On the other hand, by (35) there is a permutation τ of I_z such that $\ell_D(x_{\tau(i)}) = \ell_D(x''_i)$ for all $i \in I_z$, and so by (34) we have

$$g(\tau(i)) = \ell_D(x_{\tau(i)}) = \ell_D(x''_i) \le \ell_D(x'_i) \le \ell'_D(x'_i) + \delta = f(i) + \delta.$$

So we can apply Lemma 9.1 to obtain a permutation η_z of I_z such that

$$|\ell'_D(x_i) - \ell_D(x_{\eta_z(i)})| = |f(i) - g(\eta_z(i))| \le \delta$$

for all i. Repeating this process for each $z \in \{\ell_C(y_1), \dots, \ell_C(y_r)\}$, and reordering the tuples $(y'_1, \dots, y'_r)$ and $(x'_i, \dots, x'_r)$ using the permutation η of $\{1, \dots, r\}$ that restricts to each I_z as η_z, we obtain a singular value decomposition for A as a map $(C, \ell_C) \to (D, \ell'_D)$ satisfying the desired properties. □

We now prove a version of the stability theorem in the case that the Floer-type complexes in question arise from the same underlying chain complex, with different filtration functions.

Proposition 9.3 *Let (C_*, ∂) be a chain complex of Λ–vector spaces and let*

$$\ell_0, \ell_1\colon C_* \to \mathbb{R} \cup \{-\infty\}$$

be two filtration functions such that both (C_*, ∂, ℓ_0) and (C_*, ∂, ℓ_1) are Floer-type complexes. Assume that $\delta \ge 0$ is such that $|\ell_1(c) - \ell_0(c)| \le \delta$ for all $c \in C_*$. Then denoting by $\mathcal{B}^0_C$ and $\mathcal{B}^1_C$ the concise barcodes of (C_*, ∂, ℓ_0) and (C_*, ∂, ℓ_1), respectively, we have $d_B(\mathcal{B}^0_C, \mathcal{B}^1_C) \le \delta$.

Proof Fix a grading k, let $r = \operatorname{rank} \partial|_{C_{k+1}}$, and let $((y_1, \dots, y_n), (x_1, \dots, x_m))$ be a singular value decomposition for $\partial|_{C_{k+1}}$, considered as a map $(C_{k+1}, \ell_0) \to (C_k, \ell_0)$. In particular, the finite-length bars of the degree-k part of $\mathcal{B}^0_C$ are given by $([\ell_0(x_i)], \ell_0(y_i) - \ell_0(x_i))$ for $1 \le i \le r$, and the infinite-length bars of the degree-$(k+1)$ part of $\mathcal{B}^0_C$ are given by $([\ell_0(y_i)], \infty)$ for $r + 1 \le i \le n$.

We may then apply Lemma 9.2 to obtain an unsorted singular value decomposition $((y'_1, \dots, y'_n), (x'_1, \dots, x'_m))$ for $\partial|_{C_{k+1}}$, considered as a map $(C_{k+1}, \ell_0) \to (C_k, \ell_1)$, such that $\ell_0(y'_i) = \ell_0(y_i)$ for all i and $|\ell_1(x'_i) - \ell_0(x_i)| \le \delta$.

Now consider the adjoint $\partial^*\colon (C_k)^* \to (C_{k+1})^*$ and the dual filtration functions ℓ_0^*, ℓ_1^* as defined in Section 2.4. It follows immediately from the definitions of ℓ_0^*, ℓ_1^* and the assumption that $|\ell_1(c)-\ell_0(c)| \le \delta$ for all $c \in C_*$ that, likewise, $|\ell_1^*-\ell_0^*|$ is uniformly bounded above by δ. Moreover by Proposition 3.9, the collection of dual basis elements $((x_1'^*,\dots,x_m'^*),(y_1'^*,\dots,y_n'^*))$ gives an unsorted singular value decomposition for ∂^* considered as a map from $((C_k)^*,\ell_1^*)$ to $((C_{k+1})^*,\ell_0^*)$. Thus Lemma 9.2 applies to give an unsorted singular value decomposition $((\xi_1,\dots,\xi_m),(\eta_1,\dots,\eta_n))$ for ∂^* considered as a map $((C_k)^*,\ell_1^*) \to ((C_{k+1})^*,\ell_1^*)$, with $\ell_1^*(\xi_i) = \ell_1^*(x_i'^*)$ for all i and $|\ell_1^*(\eta_i) - \ell_0^*(y_i'^*)| \le \delta$ for all $i \in \{1,\dots,r\}$. Again using Proposition 3.9 (and using the canonical identification of $(C_i)^{**}$ with C_i for $i = k, k+1$), it follows that $((\eta_1^*,\dots,\eta_n^*),(\xi_1^*,\dots,\xi_m^*))$ is a singular value decomposition for ∂ considered as a map $(C_{k+1},\ell_1^{**}) \to (C_k,\ell_1^{**})$. It is easy to see (for instance by using (7) twice) that $\ell_1^{**} = \ell_1$. Thus the finite-length bars in the degree-k part of $\mathcal{B}_C^1$ are given by $([\ell_1(\xi_i^*)], \ell_1(\eta_i^*) - \ell_1(\xi_i^*))$.

Now using (7) we have

$$|\ell_1(\xi_i^*)-\ell_0(x_i)| \le |-\ell_1^*(\xi_i)-\ell_1(x_i')|+|\ell_1(x_i')-\ell_0(x_i)| \le |-\ell_1^*(\xi_i)+\ell_1^*(x_i'^*)|+\delta = \delta$$

and similarly

$$|\ell_1(\eta_i^*) - \ell_0(y_i)| = |-\ell_1^*(\eta_i) - \ell_0(y_i')| = |-\ell_1^*(\eta_i) + \ell_0^*(y_i'^*)| \le \delta.$$

Thus we obtain a δ–matching between the finite-length bars in the degree-k parts of $\mathcal{B}_C^0$ and $\mathcal{B}_C^1$ by pairing each $([\ell_0(x_i)], \ell_0(y_i) - \ell_0(x_i))$ with $([\ell_1(\xi_i^*)], \ell_1(\eta_i^*) - \ell_1(\xi_i^*))$ for $i = 1,\dots,r$.

It now remains to similarly match the infinite-length bars in the degree-k parts of the $\mathcal{B}_C^i$. Let us write

$$\ker(\partial|_{C_k}) = \mathrm{Im}(\partial|_{C_{k+1}}) \oplus V_0 = \mathrm{Im}(\partial|_{C_{k+1}}) \oplus V_1,$$

where $\mathrm{Im}(\partial|_{C_{k+1}})$ is orthogonal to V_0 with respect to ℓ_0 and $\mathrm{Im}(\partial|_{C_{k+1}})$ is orthogonal to V_1 with respect to ℓ_1. For $i = 0, 1$, the infinite-length bars in the degree-k parts of $\mathcal{B}_C^i$ are then given by (c,∞) as c varies through the filtration spectrum of V_i.

For $i = 0, 1$, let $\pi_i\colon \ker(\partial|_{C_k}) \to V_i$ denote the projections associated to the above direct sum decompositions. Note that $\pi_1|_{V_0}\colon V_0 \to V_1$ is a linear isomorphism, with inverse given by $\pi_0|_{V_1}$. So for $v_0 \in V_0$ we obtain

$$\ell_1(\pi_{V_1} v) \le \ell_1(v) \le \ell_0(v) + \delta$$

while

$$\ell_0(v) = \ell_0(\pi_{V_0}\pi_{V_1}(v)) \le \ell_0(\pi_{V_1} v) \le \ell_1(\pi_{V_1} v) + \delta.$$

So the linear isomorphism $\pi_{V_1}|_{V_0}\colon V_0 \to V_1$ obeys $|\ell_1(\pi_{V_1}v)-\ell_0(v)| \le \delta$ for all $v \in V$. A singular value decomposition for the map $\pi_{V_1}|_{V_0}\colon (V_0, \ell_0|_{V_0}) \to (V_1, \ell_1|_{V_1})$ precisely gives orthogonal ordered bases $(w_1, \dots, w_{m-r})$ and $(\pi_{V_1}w_1, \dots \pi_{V_1}w_{m-r})$ for $(V_0, \ell_0|_{V_0})$ and $(V_1, \ell_1|_{V_1})$, respectively, and the matching which sends $([\ell_0(w_i)], \infty)$ to $([\ell_1(\pi_{V_1}w_i)], \infty)$ then has defect at most δ. Combining this matching of the infinite-length bars in the degree-k parts of $\mathcal{B}^0_C$ and $\mathcal{B}^1_C$ with the matching of the finite-length bars that we constructed earlier, and letting k vary through $\mathbb{Z}$, we conclude that indeed $d_B(\mathcal{B}^0_C, \mathcal{B}^1_C) \le \delta$. □

9.2 Splittings

Our proof of Theorem 8.17 will involve, given a δ–quasiequivalence (Φ, Ψ, K_C, K_D), applying Proposition 9.3 to a certain pair of filtrations on the mapping cylinder $\mathrm{Cyl}(\Phi)_*$. It turns out that our arguments can be made sharper if we assume that the quasiequivalence (Φ, Ψ, K_C, K_D) satisfies a certain condition; in this subsection we introduce this condition and prove that there is no loss of generality in asking for it to be satisfied.

Definition 9.4 Let (C_*, ∂, ℓ) be a Floer-type complex. A *splitting* of C_* is a graded vector space $F^C_* = \oplus_{k\in\mathbb{Z}} F^C_k$ such that each F^C_k is an orthogonal complement in C_k to $\ker \partial_k (= \ker \partial|_{C_k})$.

Clearly splittings always exist, as already follows from Corollary 2.19. One can read off a splitting from singular value decompositions of the boundary operator in various degrees: if $((y^{k-1}_1, \dots, y^{k-1}_n), (x^{k-1}_1, \dots, x^{k-1}_m))$ is a singular value decomposition for $\partial_k\colon C_k \to \ker \partial_{k-1}$ and if r_k is the rank of ∂_k then we may take $F^C_k = \mathrm{span}_\Lambda\{y^{k-1}_1, \dots, y^{k-1}_{r_k}\}$.

Definition 9.5 If $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ are Floer-type complexes with splittings F^C_* and F^D_*, respectively, a chain map $\Phi\colon C_* \to D_*$ is said to be *split* provided that $\Phi(F^C_*) \subset F^D_*$.

Lemma 9.6 *Let $\Phi\colon C_* \to D_*$ be a chain map between two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ having splittings F^C_* and F^D_*, and let $\pi_C\colon C_* \to F^C_*$ and $\pi_D\colon D_* \to F^D_*$ be the projections associated to the direct sum decompositions $C_* = F^C_* \oplus \ker(\partial_C)_*$ and $D_* = F^D_* \oplus \ker(\partial_D)_*$. Define*

$$\Phi^\pi = \pi_D \Phi \pi_C + \Phi(I_C - \pi_C).$$

Then this map satisfies following properties:

(i) Φ^π *is a chain map;*

(ii) Φ^π *is split, and* $\Phi^\pi|_{\ker\partial_C} = \Phi|_{\ker\partial_C}$*;*

(iii) *If* $\delta \geq 0$ *and* $\ell_D(\Phi(x)) \leq \ell_C(x) + \delta$ *for all* $x \in C_*$*, then likewise* $\ell_D(\Phi^\pi(x)) \leq \ell_C(x) + \delta$ *for all* $x \in C_*$*.*

Proof For (i), since $\partial_C(I_C - \pi_C) = 0$, we see that $\partial_C\pi_C = \partial_C$ and similarly, $\partial_D\pi_D = \partial_D$. Then using that Φ is a chain map, we get

$$\partial_D\Phi^\pi = \partial_D\pi_D\Phi\pi_C + \partial_D\Phi(I_C - \pi_C) = \Phi\partial_C\pi_C + \Phi\partial_C(I_C - \pi_C) = \Phi\partial_C.$$

Moreover, $\operatorname{Im}\partial_C \leq \ker\partial_C$, so $\pi_C\partial_C = 0$, and

$$\Phi^\pi\partial_C = \pi_D\Phi\pi_C\partial_C + \Phi(I_C - \pi_C)\partial_C = \Phi\partial_C.$$

So Φ^π is a chain map.

For (ii), for $x \in F_k^C$, $\pi_C x = x$ and so $(I_C - \pi_C)x = 0$. So $\Phi^\pi x = \pi_D\Phi\pi_C x = \pi_D\Phi x \in F_k^D$, proving that Φ^π is split. Furthermore, for $x \in \ker(\partial_C)_k$, we have $\pi_C x = 0$ and so $\Phi^\pi x = \pi_D\Phi\pi_C x + \Phi(I_C - \pi_C)x = \Phi x$.

For (iii), note first that since π_D (being a projection) obeys $\pi_D^2 = \pi_D$, we have

$$\pi_D\Phi^\pi = \pi_D\Phi\pi_C + \pi_D\Phi(I_C - \pi_C) = \pi_D\Phi$$

while

$$(I_D - \pi_D)\Phi^\pi = (I_D - \pi_D)\Phi(I - \pi_C).$$

So since F_k^D and $\ker(\partial_D)_k$ are orthogonal, for all $x \in C_k$ we have

$$\begin{aligned}\ell_D(\Phi^\pi x) &= \max\{\ell_D(\pi_D\Phi^\pi x), \ell_D((I_D - \pi_D)\Phi^\pi x\} \\ &= \max\{\ell_D(\pi_D\Phi x), \ell_D((I_D - \pi_D)\Phi(I_C - \pi_C)x\} \\ &\leq \max\{\ell_D(\Phi x), \ell_D(\Phi(I_C - \pi_C)x)\}.\end{aligned}$$

But, assuming that $\ell_D(\Phi x) \leq \ell_C(x) + \delta$ for any $x \in C_k$, the orthogonality of F_k^C and $\ker(\partial_C)_k$ implies that

$$\ell_D(\Phi(I_C - \pi_C)x) \leq \ell_C(x - \pi_C x) + \delta \leq \ell_C(x) + \delta.$$

Thus $\ell_D(\Phi^\pi x) \leq \ell_C(x) + \delta$ for all $x \in C_k$. □

Proposition 9.7 *Let* (C_*, ∂, ℓ) *be a Floer-type complex with a splitting* F_*^C *and let* $\pi\colon C_* \to F_*^C$ *be the projection associated to the direct sum decomposition* $C_* = F_*^C \oplus \ker\partial_*$*. Suppose that* $A, A'\colon C_* \to C_*$ *are two chain maps such that:*

(i) *$A = \partial K + K\partial$ for some $K\colon C_* \to C_{*+1}$ such that there is $\epsilon \geq 0$ with the property that $\ell(Kx) \leq \ell(x) + \epsilon$ for all $x \in C_*$.*

(ii) *A' is split.*

(iii) *$A|_{\ker\partial} = A'|_{\ker\partial}$.*

Then for $K' = \pi K(I_C - \pi)$, we have $A' = \partial K' + K'\partial$ and $\ell(x) \leq \ell(K'x) + \epsilon$ for all $x \in C_*$.

Proof The statement that $\ell(x) \leq \ell(K'x) + \epsilon$ follows directly from the corresponding assumption on K and the fact that π and $I_C - \pi$ are orthogonal projections. So we just need to check that $A' = \partial K' + K'\partial$; we will check this separately on elements of $\ker\partial_*$ and elements of F^C_*.

For the first of these, note that just as in the proof of the preceding lemma we have $\partial\pi = \partial$, and if $x \in \ker\partial_*$ then $(I_C - \pi)x = x$. Hence, by assumption (iii),

$$A'x = Ax = \partial Kx + K\partial x = \partial Kx = \partial\pi Kx = \partial K'x = \partial K'x + K'\partial x,$$

as desired.

On the other hand if $x \in F^C_*$ we first observe that

$$\partial A'x = A'\partial x = A\partial x = \partial Ax = \partial K\partial x,$$

where the second equality again follows from (iii). Now since $\partial\pi = \partial$ and since $I_C - \pi$ is the identity on $\operatorname{Im}\partial$ we have

$$\partial K\partial x = \partial\pi K(I - \pi)\partial x = \partial K'\partial x.$$

Thus $\partial A'x = \partial K'\partial x$. But both A' and K' have image in F^C_*, on which ∂ is injective, so $A'x = K'\partial x$. Since we are assuming in this paragraph that $x \in F^C_*$, we have $(I_C - \pi)x = 0$ and so $K'x = 0$. So indeed $A'x = (\partial K' + K'\partial)x$.

Since A' and $\partial K' + K'\partial$ coincide on both summands $\ker\partial_C$ and F^C_* of C_* we have shown that they are equal. □

Corollary 9.8 *Given two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ with splittings F^C_* and F^D_*, the quasiequivalence distance $d_Q((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D))$ is equal to*

$$\inf\big\{\delta \geq 0 \mid \exists\ \delta\text{–quasiequivalence } (\Phi, \Psi, K_C, K_D) \\ \text{between } (C_*, \partial_C, \ell_C) \text{ and } (D_*, \partial_D, \ell_D) \text{ such that } \Phi \text{ and } \Psi \text{ are split}\big\}.$$

Proof It suffices to show that if (Φ, Ψ, K_C, K_D) is a δ–quasiequivalence then there is another δ–quasiequivalence $(\Phi', \Psi', K'_C, K'_D)$ such that Φ' and Ψ' are split. For this purpose we can take $\Phi' = \Phi^\pi$ and $\Psi' = \Psi^\pi$ to be the maps provided by Lemma 9.6. We can then apply Proposition 9.7 with $A = \Psi\Phi - I_C$ and $A' = \Psi'\Phi' - I_C$ to obtain $K'_C \colon C_* \to C_{*+1}$ with $\Psi'\Phi' - I_C = \partial_C K'_C + K'_C \partial_C$ and $\ell_C(K'_C x) \le \ell_C(x) + 2\delta$. Similarly applying Proposition 9.7 with $A = \Phi\Psi - I_D$ and $A' = \Phi'\Psi' - I_D$ yields a map $K'_D \colon D_* \to D_{*+1}$, and the conclusions of Lemma 9.6 and Proposition 9.7 readily imply that $(\Phi', \Psi', K'_C, K'_D)$ is, like (Φ, Ψ, K_C, K_D), a δ–quasiequivalence. □

Let us briefly describe the strategy of the rest of the proof of Theorem 8.17. In the following two subsections we will introduce a filtration function ℓ_{co} on the mapping cone $\mathrm{Cone}(\Phi)_*$ of a δ–quasiequivalence $\Phi \colon C_* \to D_*$, and two filtration functions ℓ_0, ℓ_1 on the mapping cylinder $\mathrm{Cyl}(\Phi)_*$, with ℓ_0 and ℓ_1 obeying a uniform bound $|\ell_1 - \ell_0| \le \delta$. Moreover $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_0)$ will be filtered homotopy equivalent to D_*, while $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_1)$ will be filtered homotopy equivalent to $C_* \oplus \mathrm{Cone}(\Phi)_*$. Combined with Proposition 9.10 below which places bounds on the barcode of $\mathrm{Cone}(\Phi)_*$ when Φ is split, these constructions will quickly yield Theorem 8.17 in Section 9.5.

9.3 Filtered mapping cones

Fix throughout this section a nonnegative real number δ. We will make use of the following algebraic structure, related to the mapping cylinder introduced earlier.

Definition 9.9 Given two chain complexes (C_*, ∂_C) and (D_*, ∂_D) and a chain map $\Phi \colon C_* \to D_*$ define the *mapping cone* of Φ, $(\mathrm{Cone}(\Phi)_*, \partial_{\mathrm{co}})$ by

$$\mathrm{Cone}(\Phi)_* = D_* \oplus C[1]_*$$

with boundary operator $\partial_{\mathrm{co}}(d, e) = (\partial_D d - \Phi e, -\partial_C e)$, ie in block form

$$\partial_{\mathrm{co}} = \begin{pmatrix} \partial_D & -\Phi \\ 0 & -\partial_C \end{pmatrix}.$$

Assuming additionally that $\ell_D(\Phi x) \le \ell_C(x) + \delta$ for all $x \in C_*$, define the *filtered mapping cone* $(\mathrm{Cone}(\Phi)_*, \partial_{\mathrm{co}}, \ell_{\mathrm{co}})$, where the filtration function ℓ_{co} is given by $\ell_{\mathrm{co}}(d, e) = \max\{\ell_D(d) + \delta, \ell_C(e) + 2\delta\}$.[5]

[5] One could equally well define $\ell_{\mathrm{co}}(d, e) = \max\{\ell_D(d) + t, \ell_C(e) + t + \delta\}$ for any $t \in \mathbb{R}$ (the δ is included to ensure that ℓ_{co} does not increase under ∂_{co}). Although $t = 0$ might seem to be the most natural choice, we use $t = \delta$ here in order to make the proofs of Propositions 9.10 and 9.13 more reader-friendly.

It is routine to check that $\partial_{\mathrm{co}}^2 = 0$ and that $\ell_{\mathrm{co}}(\partial_{\mathrm{co}}(d,e)) \le \ell_{\mathrm{co}}(d,e)$ for all $(d,e) \in \mathrm{Cone}_*(\Phi)$. In the case that Φ is part of a δ–quasiequivalence (Φ, Ψ, K_C, K_D), we will require some information about the concise barcode of $\mathrm{Cone}(\Phi)_*$; we will be able to make an especially strong statement when Φ is split in the sense of the previous subsection. Specifically:

Proposition 9.10 *Let $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ be Floer-type complexes with splittings F_*^C and F_*^D, and let (Φ, Ψ, K_C, K_D) be a δ–quasiequivalence such that Φ and Ψ are split. Then all elements $([a], L)$ of the concise barcode of $(\mathrm{Cone}(\Phi)_*, \partial_{\mathrm{co}}, \ell_{\mathrm{co}})$ have second coordinate obeying $L \le 2\delta$.*

Proof The desired conclusion is an easy consequence of the following statement:

(36) $\forall x \in \ker(\partial_{\mathrm{co}}),\ \exists y \in \mathrm{Cone}(\Phi)_*$ such that $\partial_{\mathrm{co}} y = x$ and $\ell_{\mathrm{co}}(y) \le \ell_{\mathrm{co}}(x) + 2\delta$.

Indeed, by definition, the elements $([a], L)$ of the concise barcode with $L < \infty$ each correspond to pairs $y_i, x_i = \partial y_i$ from a singular value decomposition for ∂_{co}, with $a = \ell_{\mathrm{co}}(x)$ and $L = \ell_{\mathrm{co}}(y_i) - \ell_{\mathrm{co}}(x_i)$, and by Lemma 4.9 any element y with $\partial y = x_i$ has $\ell(y) \ge \ell(y_i)$. Thus (36) implies that $L \le 2\delta$ provided that $L < \infty$. There can be no bars with $L = \infty$ since such bars arise from elements of an orthogonal complement to $\mathrm{Im}(\partial_{\mathrm{co}})$ in $\ker(\partial_{\mathrm{co}})$ but (36) implies that $\mathrm{Im}(\partial_{\mathrm{co}}) = \ker(\partial_{\mathrm{co}})$.

We now prove (36). Let $x = (d,e) \in \ker(\partial_{\mathrm{co}})_*$; thus $\partial_{\mathrm{co}}(d,e) = (\partial_D d - \Phi e, -\partial_C e) = 0$. Therefore,

$$\partial_D d = \Phi e \quad \text{and} \quad \partial_C e = 0.$$

Split d according to the direct sum decomposition $D_* = F_*^D \oplus \ker(\partial_D)_*$ as $d = d_F + d_K$ and let $\lambda = \ell_{\mathrm{co}}(x)$. Then $\ell_D(d) \le \lambda - \delta$ and $\ell_C(e) \le \lambda - 2\delta$. So since F_*^D and $\ker(\partial_D)_*$ are orthogonal, $\ell_D(d_K) \le \lambda - \delta$ and $\ell_D(d_F) \le \lambda - \delta$. Moreover, since $\partial_C e = 0$, the equation $\Psi\Phi - I_C = \partial_C K_C + K_C \partial_C$ implies that $\partial(K_C e) = \Psi\Phi e - e$, where $\ell_C(K_C e) \le \ell_C(e) + 2\delta \le \lambda$.

Write $K_C e = a + a'$ with $a \in F_C^*$ and $a' \in \ker(\partial_C)_*$. Then by the orthogonality of F_C^* and $\ker(\partial_C)_*$ we have $\ell_C(a) \le \ell_C(K_C e) \le \lambda$, and $\partial_C a = \partial_C K_C e = (\Psi\Phi - I_D)e$.

We then find that

$$\begin{aligned} \partial_D(\Phi\Psi d_F - d_F - \Phi a) &= \Phi\Psi\partial_D d_F - \partial_D d_F - \Phi\partial_C a \\ &= (\Phi\Psi - I_D)\Phi e - \Phi\partial_C a = 0. \end{aligned} \tag{37}$$

On the other hand, because Φ and Ψ are split we have $\Phi\Psi d_F - d_F - \Phi a \in F_*^D$, so since $\partial_D|_{F_*^D}$ is injective (37) implies that

$$\Phi a = \Phi\Psi d_F - d_F.$$

Since $\partial_D d_K = 0$, the element $b = K_D d_K \in D_{*+1}$ obeys

$$\partial_D b = (\Phi\Psi - I_D) d_K$$

and $\ell_D(b) \le \ell_D(d_K) + 2\delta \le \lambda - \delta + 2\delta = \lambda + \delta$. Let $y = (-b, a - \Psi d)$. We claim that this y obeys the desired conditions stated at the start of the proof. In fact,

$$\begin{aligned}
\partial_{\mathrm{co}}(y) &= (\partial_D(-b) - \Phi(a - \Psi d), -\partial_C(a - \Psi d)) \\
&= (-\partial_D b - \Phi a + \Phi\Psi d, -\partial_C a + \partial_C \Psi d) \\
&= (d_K - \Phi\Psi d_K - \Phi a + \Phi\Psi d, e - \Psi\Phi e + \Psi\partial_D d) \\
&= (d_K - \Phi\Psi d_K - \Phi\Psi d_F + d_F + \Phi\Psi d, e) \\
&= (d, e) = x.
\end{aligned}$$

Moreover, the filtration level of y obeys

$$\begin{aligned}
\ell_{\mathrm{co}}(y) &= \ell_{\mathrm{co}}((-b, a - \Psi d)) \\
&= \max\{\ell_D(-b) + \delta, \ell_C(a - \Psi d) + 2\delta\} \\
&\le \max\{\lambda + 2\delta, \max\{\ell_C(a), \ell_C(d) + \delta\} + 2\delta\} \\
&= \lambda + 2\delta = \ell_{\mathrm{co}}(x) + 2\delta.
\end{aligned}$$

So $\partial_{\mathrm{co}} y = x$ and $\ell_{\mathrm{co}}(y) \le \ell_{\mathrm{co}}(x) + 2\delta$, as desired. Since x was an arbitrary element of $\ker(\partial_{\mathrm{co}})_*$ this implies the result. □

Remark 9.11 If one drops the hypothesis that Φ and Ψ are split, then it is possible to construct examples showing that the largest second coordinate in an element of the concise barcode of $\mathrm{Cone}(\Phi)_*$ can be as large as 4δ.

9.4 Filtered mapping cylinders

Recall the definition of the mapping cylinder $\mathrm{Cyl}(\Phi)_*$ of a chain map $\Phi\colon C_* \to D_*$ from Section 7.2.1, and the homotopy equivalences $(i_D, \alpha, 0, K)$ between D_* and $\mathrm{Cyl}(\Phi)_*$ and $(i_C, \beta, 0, L)$ between C_* and $\mathrm{Cyl}(\Phi)_*$ from Lemma 7.12 (the first of these exists for any chain map Φ, while the second requires Φ to be part of a homotopy equivalence, as is indeed the case in our present context). The "only if" direction of Theorem B was proven by, in the case that (Φ, Ψ, K_C, K_D) is a filtered homotopy equivalence, exploiting the behavior of a suitable filtration function on $\mathrm{Cyl}(\Phi)_*$ with respect to $(i_D, \alpha, 0, K)$ and $(i_C, \beta, 0, L)$. In the case that (Φ, Ψ, K_C, K_D) is instead a δ–quasiequivalence, we will follow a similar strategy, but using different filtration functions on $\mathrm{Cyl}(\Phi)_*$ for the two homotopy equivalences.

Proposition 9.12 *Given two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ and a δ–quasiequivalence (Φ, Ψ, K_C, K_D) between them, define a filtration function $\ell_0\colon \mathrm{Cyl}(\Phi)_* \to \mathbb{R} \cup \{-\infty\}$ by*

$$\ell_0(c, d, e) = \max\{\ell_C(c) + \delta, \ell_D(d), \ell_C(e) + \delta\}.$$

Then:

(i) *$\ell_0(\partial_{\mathrm{cyl}} x) \le \ell_0(x)$ for all $x \in \mathrm{Cyl}(\Phi)_*$. Thus $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_0)$ is a Floer-type complex.*

(ii) *Let $(i_D, \alpha, 0, K)$ be as defined in Lemma 7.12. Then $(i_D, \alpha, 0, K)$ is a filtered homotopy equivalence between $(D_*, \partial_D, \ell_D)$ and $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_0)$.*

Proof For (i), if $(c, d, e) \in \mathrm{Cyl}(\Phi)_*$ we have

$$\ell_0(\partial_{\mathrm{cyl}}(c, d, e)) = \max\{\ell_C(\partial_C c - e) + \delta, \ell_D(\partial_D d + \Phi e), \ell_C(\partial_C e) + \delta\}$$

while $\ell_0(c, d, e) = \max\{\ell_C(c) + \delta, \ell_D(d), \ell_C(e) + \delta\}$. So (i) follows from the facts that:

- $\ell_C(\partial_C c - e) + \delta \le \max\{\ell_C(c) + \delta, \ell_C(e) + \delta\}$;
- $\ell_D(\partial_D d + \Phi e) \le \max\{\ell_D(d), \ell_D(\Phi e)\} \le \max\{\ell_D(d), \ell_C(e) + \delta\}$;
- $\ell_C(\partial_C e) + \delta \le \ell_C(e) + \delta$.

By Lemma 7.12, $(i_D, \alpha, 0, K)$ is a homotopy equivalence, so to prove (ii) we just need to check that each of the maps preserves filtration. We see that:

- Clearly $\ell_0(i_D d) = \ell_D(d)$ for all $d \in D_*$, by definition of ℓ_0.
- For $(c, d, e) \in \mathrm{Cyl}(\Phi)_*$,
$$\ell_D(\alpha(c, d, e)) = \ell_D(\Phi c + d) \le \max\{\ell_C(c) + \delta, \ell_D(d)\} \le \ell_0(c, d, e).$$
- For $(c, d, e) \in \mathrm{Cyl}(\Phi)_*$, $\ell_0(K(c, d, e)) = \ell_0(0, 0, c) = \ell_C(c) + \delta \le \ell_0(c, d, e)$.

Thus $(i_D, \alpha, 0, K)$ is indeed a filtered homotopy equivalence. □

Proposition 9.13 *Given two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ having splittings F_*^C and F_*^D and a δ–quasiequivalence (Φ, Ψ, K_C, K_D) where Φ and Ψ are split, define a new filtration function ℓ_1 on $\mathrm{Cyl}(\Phi)_*$ by*

$$\ell_1(c, d, e) = \max\{\ell_C(c), \ell_D(d) + \delta, \ell_C(e) + 2\delta\}.$$

Then, with β as defined in Lemma 7.12:

(i) $\ell_1(\partial_{\mathrm{cyl}}(c,d,e)) \le \ell_1(c,d,e)$ *for all* $(c,d,e) \in \mathrm{Cyl}(\Phi)_*$, *so* $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_1)$ *is a Floer-type complex.*

(ii) $i_C(C_*)$ *and* $\ker\beta$ *are orthogonal complements with respect to* ℓ_1*.*

(iii) *The second coordinates of all elements of the concise barcode of* $(\ker\beta, \partial_{\mathrm{cyl}}, \ell_1)$ *are at most* 2δ*.*

Proof Part (i) follows just as in the proof of Proposition 9.12(i) (which only depended on the fact that the shift $\ell_0(0,0,e) - \ell_C(e)$ in the filtration level of $\ell_C(e)$ in the definition of ℓ_0 was greater than or equal to both $\ell_0(c,0,0) - \ell_C(c)$ and $\delta + \ell_0(0,d,0) - \ell_D(d)$; this condition also holds with ℓ_1 in place of ℓ_0).

For part (ii), first note that $\ker\beta$ consists precisely of elements of the form $(-\Psi d - K_C e, d, e)$ for $(d,e) \in D_* \oplus C[1]_*$. We will apply Lemma 7.5 with $V = i_C(C_*)$, $U = \{0\} \oplus D_* \oplus C[1]_*$, and $U' = \ker\beta$. Clearly U and V are orthogonal with respect to ℓ_1, and the projection $\pi_U\colon \mathrm{Cyl}(\Phi)_* \to U$ is given by $(c,d,e) \mapsto (0,d,e)$, so

$$\ell_1(-\Psi d - K_C e, d, e) = \max\{\ell_D(d) + \delta, \ell_C(e) + 2\delta\} = \ell_1(0,d,e)$$

which shows that $\ell_1(\pi_U x) = \ell_1(x)$ for all $x \in \ker\beta$. Thus $\ker\beta$ is indeed an orthogonal complement to $V = i_C(C_*)$.

For part (iii), define a map $f\colon \ker\beta \to \mathrm{Cone}_*(-\Phi)$ by

$$f(-\Psi d - K_C e, d, e) = (d,e).$$

We claim that f is a filtered chain isomorphism. By definition, we have

$$(f \circ \partial_{\mathrm{cyl}})(-\Psi d - K_C e, d, e) = (\partial_D d + \Phi e, -\partial_C e).$$

Furthermore,

$$(\partial_{\mathrm{co}} \circ f)(-\Psi d - K_C e, d, e) = (\partial_D d + \Phi e, -\partial_C e).$$

Therefore, f is a chain map. As for the filtrations,

$$\begin{aligned}\ell_{\mathrm{co}}(f(-\Psi d - K_C e, d, e)) &= \ell_{\mathrm{co}}(d,e)\\ &= \max\{\ell_D(d) + \delta, \ell_C(c) + 2\delta\}\\ &= \ell_1(-\Psi d - K_C e, d, e).\end{aligned}$$

Thus f defines an isomorphism between $(\ker\beta, \partial_{\mathrm{cyl}}, \ell_1)$ and $(\mathrm{Cone}_*(-\Phi), \partial_{\mathrm{co}}, \ell_{\mathrm{co}})$ as Floer-type complexes. Moreover, replacing (Φ, Ψ, K_C, K_D) by $(-\Phi, -\Psi, K_C, K_D)$ does not change the homotopy equations and also it has no effect on the filtration relations. Therefore, the conclusion follows from Theorem A and Proposition 9.10. □

9.5 End of the proof of Theorem 8.17

Assume that $\delta \geq 0$ and that (Φ, Ψ, K_C, K_D) is a δ–quasiequivalence which is split with respect to splittings F_*^C and F_*^D for the Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$. The preceding subsection gives filtration functions $\ell_0, \ell_1\colon \mathrm{Cyl}(\Phi)_* \to \mathbb{R} \cup \{-\infty\}$ which evidently satisfy the bound $|\ell_1(x) - \ell_0(x)| \leq \delta$ for all $x \in \mathrm{Cyl}(\Phi)_*$. Hence by Proposition 9.3, we have a bound

$$d_B(\mathcal{B}_{\mathrm{Cyl},\ell_0}, \mathcal{B}_{\mathrm{Cyl},\ell_1}) \leq \delta \tag{38}$$

for the bottleneck distance between the concise barcodes of the Floer-type complexes $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_0)$ and $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_1)$.

Corollary 9.14 *If two Floer-type complexes $(C_*, \partial_C, \ell_C)$, $(D_*, \partial_D, \ell_D)$, are δ–quasi-equivalent, then we have $d_B(\mathcal{B}_C, \mathcal{B}_D) \leq 2\delta$. Therefore, in particular,*

$$d_B(\mathcal{B}_C, \mathcal{B}_D) \leq 2 d_Q((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)).$$

Proof By Corollary 9.8, the assumption implies that there is a δ–quasiequivalence (Φ, Ψ, K_C, K_D) which moreover is split with respect to some splittings for $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$.

By Proposition 9.13(ii), $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_1)$ decomposes as an orthogonal direct sum of subcomplexes $(i_C(C_*), \partial_{\mathrm{cyl}}, \ell_1)$ and $(\ker\beta, \partial_{\mathrm{cyl}}, \ell_1)$, so in any degree a singular value decomposition for $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_1)$ may be obtained by combining singular value decompositions for $(i_C(C_*), \partial_{\mathrm{cyl}}, \ell_1)$ and $(\ker\beta, \partial_{\mathrm{cyl}}, \ell_1)$. Thus the concise barcode for $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_1)$ is the union of the concise barcodes for these two subcomplexes.

Now i_C embeds $(C_*, \partial_C, \ell_C)$ filtered isomorphically as $(i_C(C_*), \partial_{\mathrm{cyl}}, \ell_1)$, so the concise barcode of $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_1)$ consists of the concise barcode of $(C_*, \partial_C, \ell_C)$ together with the concise barcode of $(\ker\beta, \partial_{\mathrm{cyl}}, \ell_1)$. By Proposition 9.13(iii), all elements $([a], L)$ in the second of these barcodes have $L \leq 2\delta$. Thus by matching the elements of the concise barcode of $(C_*, \partial_C, \ell_C)$ with themselves and leaving the elements of the concise barcode $(\ker\beta, \partial_{\mathrm{cyl}}, \ell_1)$ unmatched, we obtain, in each degree, a partial matching between the concise barcodes of $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_1)$ and of $(C_*, \partial_C, \ell_C)$ with defect at most δ. Thus, in obvious notation,

$$d_B(\mathcal{B}_C, \mathcal{B}_{\mathrm{Cyl},\ell_1}) \leq \delta.$$

Finally, by Proposition 9.12(ii) and Theorem B, we know

$$\mathcal{B}_{\mathrm{Cyl},\ell_0} = \mathcal{B}_D.$$

Therefore, by the triangle inequality and (38), we get

$$d_B(\mathcal{B}_C, \mathcal{B}_D) \leq d_B(\mathcal{B}_C, \mathcal{B}_{\mathrm{Cyl},\ell_1}) + d_B(\mathcal{B}_{\mathrm{Cyl},\ell_1}, \mathcal{B}_{\mathrm{Cyl},\ell_0}) + d_B(\mathcal{B}_{\mathrm{Cyl},\ell_0}, \mathcal{B}_D) \leq 2\delta. \qquad \square$$

We have thus proven the inequality (30).

For the last assertion in Theorem 8.17, let $\lambda = d_Q((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D))$, so there are arbitrarily small $\epsilon > 0$ such that there exists a (split) $(\lambda+\epsilon)$–quasiequivalence (Φ, Ψ, K_C, K_D) between $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$. So by (38) with $\delta = \lambda+\epsilon$, there is a δ–matching $\mathfrak{m}$ between the concise barcodes of $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_0)$ and $(\mathrm{Cyl}(\Phi)_*, \partial_{\mathrm{cyl}}, \ell_1)$. Just as in the proof of Corollary 9.14, the first of these concise barcodes is, in any given degree k, the same as that of $(D_*, \partial_D, \ell_D)$, while the second of these is the union of the concise barcode of $(C_*, \partial_C, \ell_C)$ with a multiset $\mathcal{S}$ of elements all having second coordinate at most $2(\lambda+\epsilon)$. For a grading k in which $\lambda < \frac{1}{4}\Delta_{D,k}$, let us take ϵ so small that still $\delta = \lambda+\epsilon < \frac{1}{4}\Delta_{D,k}$. Now by definition, the image of any element $([a], L)$ which is not unmatched under a δ–matching must have second coordinate at most $L + 2\delta$. Since $\delta < \frac{1}{4}\Delta_{D,k}$, the concise barcode $\mathcal{B}_{D,k}$ has *no* elements with second coordinate at most 4δ, all of the elements of our multiset $\mathcal{S}$ (each of which have second coordinate less than or equal to 2δ) must be unmatched under $\mathfrak{m}$. But since all elements of $\mathcal{S}$ are unmatched, we can discard them from the domain of $\mathfrak{m}$ and so restrict $\mathfrak{m}$ to a matching between the barcodes $\mathcal{B}_{C,k}$ and $\mathcal{B}_{D,k}$, still having defect at most $\delta = \lambda+\epsilon$. So $d_B(\mathcal{B}_{C,k}, \mathcal{B}_{D,k}) \leq \lambda+\epsilon$, and since $\epsilon > 0$ can be taken arbitrarily small this implies that

$$d_B(\mathcal{B}_{C,k}, \mathcal{B}_{D,k}) \leq \lambda = d_Q((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)).$$

Remark 9.15 In the case that Γ is dense, a simpler argument based on Corollary 8.8 suffices to prove the stability theorem, in fact with the stronger inequality $d_B \leq d_Q$. Indeed, if Γ is dense then the extended pseudometric d from Example 8.12 is easily seen to simplify to $d(([a], L), ([a'], L')) = \frac{1}{2}|L - L'|$. If two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_C, \ell_C)$ are δ–quasiequivalent, then we can obtain a partial matching of defect at most δ between the concise barcodes $\mathcal{B}_C$ and $\mathcal{B}_D$ by first sorting the respective barcodes in descending order by the size of the second coordinate L and then matching elements in corresponding positions on the two sorted lists. It follows easily from Theorem 4.11 and Corollary 8.8 that, when Γ is dense, this partial matching has defect at most δ.

10 Proof of converse stability

Recall the elementary Floer-type complexes $\mathcal{E}(a, L, k)$ from Definition 7.2.

Lemma 10.1 *If $\delta \in [0,\infty)$, $|a-a'| \le \delta$, and either*

$$L = L' = \infty \quad \text{or} \quad |(a+L)-(a'+L')| \le \delta,$$

then $\mathcal{E}(a,L,k)$ is δ–quasiequivalent to $\mathcal{E}(a',L',k)$. Moreover, if $L \le 2\delta$, then $\mathcal{E}(a,L,k)$ is δ–quasiequivalent to the zero chain complex.

Proof In the case that $L = L' = \infty$, the chain complexes underlying $\mathcal{E}(a,L,k)$ and $\mathcal{E}(a',L',k)$ are just one-dimensional, consisting of a copy of Λ in degree k, with filtrations given by $\ell(\lambda) = a - \nu(\lambda)$ and $\ell'(\lambda) = a' - \nu(\lambda)$. Let $\mathbb{I}$ denote the identity on Λ. The fact that $|a-a'| \le \delta$ then readily implies that $(\mathbb{I},\mathbb{I},0,0)$ is a δ–quasiequivalence.

Similarly if L and hence (under the hypotheses of the lemma) L' are both finite, the underlying chain complexes of $\mathcal{E}(a,L,k)$ and $\mathcal{E}(a',L',k)$ are both Λ–vector spaces generated by an element x in degree k and an element y in degree $k+1$, with filtration functions ℓ and ℓ' given by saying that (x,y) is an orthogonal ordered set with $\ell(x) = a$, $\ell(y) = a+L$, $\ell'(x) = a'$, and $\ell'(y) = a'+L'$. The hypotheses imply that $|\ell(x)-\ell'(x)| \le \delta$ and $|\ell(y)-\ell'(y)| \le \delta$, and if $\mathbb{I}$ now denotes the identity on the two-dimensional vector space spanned by x and y, $(\mathbb{I},\mathbb{I},0,0)$ is again a δ–quasiequivalence.

Finally, if similarly to the proof of Proposition 7.9 we define a linear transformation K on $\mathrm{span}_\Lambda\{x,y\}$ by $Kx = -y$ and $Ky = 0$, then $(0,0,K,0)$ is readily seen to be a δ–quasiequivalence between $\mathcal{E}(a,L,k)$ and the zero chain complex for all $\delta \ge L/2$, proving the last sentence of the lemma. □

Proof of Theorem 8.18 Let $\delta = d_B(\mathcal{B}_C,\mathcal{B}_D)$; it suffices to prove the result under the assumption that $\delta < \infty$.

For any $k \in \mathbb{Z}$, $d_B(\mathcal{B}_{C,k},\mathcal{B}_{D,k}) \le \delta$. By the definition of the bottleneck distance (and using the fact that there are only finitely many partial matchings between the finite multisets $\mathcal{B}_{C,k}$ and $\mathcal{B}_{D,k}$, so the infimum in the definition is attained), there exists a partial matching $\mathfrak{m}_k = (\mathcal{B}_{C,k,\text{short}},\mathcal{B}_{D,k,\text{short}},\sigma_k)$ between $\mathcal{B}_{C,k}$ and $\mathcal{B}_{D,k}$ having defect $\delta(\mathfrak{m}_k) \le \delta$.

We claim that, for all $\epsilon > 0$,

$$\bigoplus_k \bigoplus_{([a],L)\in\mathcal{B}_{C,k}} \mathcal{E}(a,L,k) \quad \text{and} \quad \bigoplus_k \bigoplus_{([a'],L')\in\mathcal{B}_{D,k}} \mathcal{E}(a',L',k)$$

are $(\delta+\epsilon)$–quasiequivalent, for some representatives a and a' of the various cosets $[a]$ and $[a']$ in $\mathbb{R}/\Gamma$. By Proposition 7.9 and Remark 8.2 this will imply that (C_*,∂_C,ℓ_C)

and $(D_*, \partial_D, \ell_D)$ are $(\delta+\epsilon)$–quasiequivalent, which suffices to prove the theorem since by the definition of the quasiequivalence distance, it will show that

$$d_Q\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big) \le \delta + \epsilon = d_B(\mathcal{B}_C, \mathcal{B}_D) + \epsilon \quad \text{for all } \epsilon > 0.$$

To prove our claim, note that by Lemma 10.1 and the fact that $\delta(\mathfrak{m}_k) \le \delta$, each $\mathcal{E}(a, L, k)$ for $([a], L) \in \mathcal{B}_{C,k,\mathrm{short}} \cup \mathcal{B}_{D,k,\mathrm{short}}$ is $(\delta+\epsilon)$–quasiequivalent to the zero chain complex (as these $\mathcal{E}(a, L, k)$ all have $L \le 2\delta$). Also, for $([a], L) \in \mathcal{B}_{C,k} \setminus \mathcal{B}_{C,k,\mathrm{short}}$, if we write $([a'], L') = \sigma_k([a], L)$, where σ_k is the bijection from the partial matching $\mathfrak{m}_k$, then there are representatives a and a' of the cosets $[a]$ and $[a']$ such that $|a - a'| \le \delta + \epsilon$ and $|(a + L) - (a' + L')| \le \delta + \epsilon$. So by Lemma 10.1, the associated summands $\mathcal{E}(a, L, k)$ and $\mathcal{E}(a', L', k)$ are $(\delta+\epsilon)$–quasiequivalent.

Moreover, it is straightforward from the definitions that a direct sum of $(\delta+\epsilon)$–quasiequivalences is a $(\delta+\epsilon)$–quasiequivalence. So we obtain a $(\delta+\epsilon)$–quasiequivalence between $\bigoplus_k \bigoplus_{([a],L)\in\mathcal{B}_{C,k}} \mathcal{E}(a, L, k)$ and $\bigoplus_k \bigoplus_{([a'],L')\in\mathcal{B}_{D,k}} \mathcal{E}(a', L', k)$ by taking a direct sum of:

- a $(\delta+\epsilon)$–quasiequivalence between $\mathcal{E}(a, L, k)$ and $\mathcal{E}(a', L', k)$ for each $([a], L) \in \mathcal{B}_{C,k} \setminus \mathcal{B}_{C,k,\mathrm{short}}$, where $([a'], L') = \sigma_k([a], L)$;
- a $(\delta+\epsilon)$–quasiequivalence between $\bigoplus_k \bigoplus_{([a],L)\in\mathcal{B}_{C,k,\mathrm{short}}} \mathcal{E}(a, L, k)$ and the zero chain complex;
- a $(\delta+\epsilon)$–quasiequivalence between the zero chain complex and
$$\bigoplus_k \bigoplus_{([a'],L')\in\mathcal{B}_{D,k,\mathrm{short}}} \mathcal{E}(a', L', k). \qquad \square$$

11 The interpolating distance

In this section we introduce a somewhat more complicated distance function on Floer-type complexes, the interpolating distance d_P, and prove the isometry result Theorem 11.2 between this distance and the bottleneck distance between barcodes. We think that it is likely that d_P is always equal to the quasiequivalence distance d_Q, and indeed in the case that Γ is dense this equality can be inferred from our results (specifically, Theorem 11.2, Remark 9.15, and Theorem 8.18), while in the case that Γ is trivial it can be inferred from Theorem 11.2 and [9, Theorem 4.11].

The definition of the distance d_P will be based on a strengthening of the notion of quasiequivalence, asking not only for a quasiequivalence between the two complexes C_* and D_* but also for a one parameter family of complexes that interpolates between C_* and D_* in a suitably "efficient" way. Our interest in d_P is based on the facts that, on the one hand, we can prove Theorem 11.2 about it, and on the other hand standard arguments in Hamiltonian Floer theory (and other Floer theories) that give bounds for the quasiequivalence distance can be refined to give bounds on d_P, as we use in Section 12.

Definition 11.1 A *δ–interpolation* between two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ is a family of Floer-type complexes $(C_*^s, \partial^s, \ell^s)$ indexed by a parameter s that varies through $[0,1]\backslash S$ for some finite subset $S \subset (0,1)$, such that:

- $(C_*^0, \partial^0, \ell^0) = (C_*, \partial_C, \ell_C)$ and $(C_*^1, \partial^1, \ell^1) = (D_*, \partial_D, \ell_D)$; and
- for all $s,t \in [0,1]\backslash S$, $(C_*^s, \partial^s, \ell^s)$ and $(C_*^t, \partial^t, \ell^t)$ are $\delta|s-t|$–quasiequivalent.

The *interpolating distance d_P* between Floer-type complexes is then defined by

$$d_P\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big) = \inf\big\{\delta \ge 0 \mid \exists\ \delta\text{–interpolation between } (C_*, \partial_C, \ell_C) \text{ and } (D_*, \partial_D, \ell_D)\big\}.$$

The following theorem gives a global isometry result between the bottleneck and interpolating distances.

Theorem 11.2 *For any two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ we have*

$$d_B(\mathcal{B}_C, \mathcal{B}_D) = d_P\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big).$$

Proof First, we will prove that for any degree $k \in \mathbb{Z}$,

$$d_B(\mathcal{B}_{C,k}, \mathcal{B}_{D,k}) \le d_P\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big),$$

which will imply that $d_B(\mathcal{B}_C, \mathcal{B}_D) \le d_P\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big)$ by taking the supremum over k. Let $\lambda = d_P\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big)$, so by definition, given any $\epsilon > 0$, there exists a δ–interpolation between $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$ with $\delta \le \lambda + \epsilon$, denoted as $(C_*^1, \partial^s, \ell^s)$ with a finite singular set S.

For any $p \in [0,1] \setminus S$ and any degree $k \in \mathbb{Z}$, choose $\epsilon_{p,k} > 0$ such that $\Delta_{C_k^p} > 4\delta\epsilon_{p,k}$, where the meaning of $\Delta_{C_k^p}$ is as in the last statement of Theorem 8.17. By the definition of a δ–interpolation, for any $s \in (p-\epsilon_{p,k}, p]$, $(C_*^s, \partial^s, \ell^s)$ and $(C_*^p, \partial^p, \ell^p)$ are $(\delta(p-s))$–quasiequivalent, which implies that

$$d_Q\big((C_*^s, \partial^s, \ell^s), (C_*^p, \partial^p, \ell^p)\big) < \tfrac{1}{4}\Delta_{C_k^p}.$$

Then, again assuming that $s \in (p-\epsilon_{p,k}, p]$, the last assertion from Theorem 8.17 implies that

$$d_B(\mathcal{B}_{C^s,k}, \mathcal{B}_{C^p,k}) = d_Q\big((C_*^s, \partial^s, \ell^s), (C_*^p, \partial^p, \ell^p)\big) \le \delta(p-s).$$

Symmetrically, for any $s' \in [p, p+\epsilon_{p,k})$,

$$d_B(\mathcal{B}_{C^p,k}, \mathcal{B}_{C^{s'},k}) = d_Q\big((C_*^p, \partial^p, \ell^p), (C_*^{s'}, \partial^{s'}, \ell^{s'})\big) \le \delta(s'-p).$$

Therefore, by the triangle inequality, for s, s' such that $p-\epsilon_{p,k} < s \le p \le s' < p+\epsilon_{p,k}$, we have $d_B(\mathcal{B}_{C^s,k}, \mathcal{B}_{C^{s'},k}) \le \delta(s'-s)$.

Now we claim that for any closed interval $[s,t] \subset [0,1]$ with $s, t \notin S$, the following estimate holds:

$$d_B(\mathcal{B}_{C^s,k}, \mathcal{B}_{C^t,k}) \le (t-s)\delta. \tag{39}$$

We will prove this by induction on the cardinality of $S \cap [s,t]$. First, when $S \cap [s,t]$ is empty, by considering a covering $\{(p-\epsilon_{p,k}, p+\epsilon_{p,k})\}_{p\in[s,t]}$ of $[s,t]$ where the $\epsilon_{p,k}$ are as above, we may take a finite subcover to obtain $s = s_0 < s_1 < \cdots < s_N = t$ such that $d_B(\mathcal{B}_{C^{s_{i-1}},k}, \mathcal{B}_{C^{s_i},k}) \le \delta(s_i - s_{i-1})$. Therefore, by the triangle inequality again,

$$d_B(\mathcal{B}_{C^s,k}, \mathcal{B}_{C^t,k}) \le \sum_{i=1}^{N} d_B(\mathcal{B}_{C_k^{s_{i-1}}}, \mathcal{B}_{C_k^{s_i}}) \le (t-s)\delta.$$

Now inductively, we will assume that (39) holds when $|S \cap [s,t]| \le m$. For the case that $|S \cap [s,t]| = m+1$, denote the smallest element of $S \cap [s,t]$ by p^* and consider the intervals $[s, p^*-\epsilon']$ and $[p^*+\epsilon', t]$ for any sufficiently small $\epsilon' > 0$. Applying the inductive hypothesis on both intervals,

$$d_B(\mathcal{B}_{C^s,k}, \mathcal{B}_{C^{p^*-\epsilon'},k}) \le (p^*-\epsilon'-s)\delta \quad \text{and} \quad d_B(\mathcal{B}_{C^{p^*+\epsilon'},k}, \mathcal{B}_{C^t,k}) \le (t-p^*-\epsilon')\delta.$$

By the first conclusion of Theorem 8.17,

$$d_B(\mathcal{B}_{C^{p^*-\epsilon'},k}, \mathcal{B}_{C^{p^*+\epsilon'},k}) \le 2d_Q(\mathcal{B}_{C^{p^*-\epsilon'},k}, \mathcal{B}_{C^{p^*+\epsilon'},k}) \le 4\epsilon'\delta.$$

Together, we get

$$d_B(\mathcal{B}_{C^s,k}, \mathcal{B}_{C^t,k}) \le (p^*-\epsilon'-s)\delta + (t-p^*-\epsilon')\delta + 4\epsilon'\delta = (t-s)\delta + 2\epsilon'\delta.$$

Since ϵ' is arbitrarily small, it follows that $d_B(\mathcal{B}_{C^s,k}, \mathcal{B}_{C^t,k}) \le (t-s)\delta$ whenever $s \le t$ and $s, t \in [0,1] \setminus S$. So we have proven (39).

In particular, letting $s=0$ and $t=1$, we get $d_B(\mathcal{B}_{C,k}, \mathcal{B}_{D,k}) \le \delta \le \lambda + \epsilon$. Since ϵ is arbitrarily small, this shows that $d_B(\mathcal{B}_{C,k}, \mathcal{B}_{D,k}) \le \lambda = d_P\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big)$.

Now we will prove the converse direction:

$$d_P\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big) \le d_B(\mathcal{B}_C, \mathcal{B}_D).$$

Let $\delta = d_B(\mathcal{B}_C, \mathcal{B}_D)$. It is sufficient to prove the result under the assumption that $\delta < \infty$. For any $k \in \mathbb{Z}$, $d_B(\mathcal{B}_{C,k}, \mathcal{B}_{D,k}) \le \delta$. By definition, there exists a partial matching

$\mathfrak{m}_k = (\mathcal{B}_{C,k,\text{short}}, \mathcal{B}_{D,k,\text{short}}, \sigma_k)$ between $\mathcal{B}_{C,k}$ and $\mathcal{B}_{D,k}$ such that $\delta(\mathfrak{m}_k) \le \delta$. We will prove that, for all $\epsilon > 0$, there exists a $(\delta+\epsilon)$–interpolation between $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$.

For each $([a], L) \in \mathcal{B}_{C,k,\text{short}}$, choose a representative a of $[a]$; also if $([a], L) \in \mathcal{B}_{C,k} \setminus \mathcal{B}_{C,k,\text{short}}$ write $\sigma([a], L) = ([a'], L')$, where the representative a' is chosen so that both $|a' - a| \le \delta + \epsilon$ and $|(a + L) - (a' + L')| \le \delta + \epsilon$. Now for $t \in (0, 1)$ consider the Floer-type complex $(C^t_*, \partial^t, \ell^t)$ given by

$$\bigoplus_{k\in\mathbb{Z}}\Bigg(\Bigg(\bigoplus_{([a'],L')\in\mathcal{B}_{D,k,\text{short}}} \mathcal{E}(a' + (1-t)L'/2, tL', k)\Bigg) \oplus \Bigg(\bigoplus_{([a],L)\in\mathcal{B}_{C,k,\text{short}}} \mathcal{E}(a + tL/2, (1-t)L, k)\Bigg) \oplus \Bigg(\bigoplus_{([a],L)\in\mathcal{B}_{C,k}\setminus\mathcal{B}_{C,k,\text{short}}} \mathcal{E}((1-t)a + ta', (1-t)L + tL', k)\Bigg)\Bigg).$$

It is easy to see by Lemma 10.1 that, for $t_0, t_1 \in (0, 1)$, the t_0–version of each of these summands is $(\delta+\epsilon)|t_0 - t_1|$–quasiequivalent to its corresponding t_1–version. So since the direct sum of $(\delta+\epsilon)|t_0-t_1|$–quasiequivalences is a $(\delta+\epsilon)|t_0-t_1|$–quasiequivalence this shows that $(C^{t_0}_*, \partial^{t_0}, \ell^{t_0})$ and $(C^{t_1}_*, \partial^{t_1}, \ell^{t_1})$ are $(\delta+\epsilon)|t_0-t_1|$–quasiequivalent for $t_0, t_1 \in (0, 1)$. Moreover $\mathcal{E}(a' + (1-t)L'/2, tL', k)$ is $t\delta$–quasiequivalent to the zero chain complex for each $([a'], L') \in \mathcal{B}_{D,k,\text{short}}$, and likewise $\mathcal{E}(a + tL/2, (1-t)L, k)$ is $(1-t)\delta$–quasiequivalent to the zero chain complex for each $([a], L) \in \mathcal{B}_{C,k,\text{short}}$. In view of Proposition 7.9 it follows that $(C_*, \ell_C, \partial_C)$ is $t(\delta+\epsilon)$–quasiequivalent to $(C^t_*, \partial^t, \ell^t)$, and that $(D_*, \ell_D, \partial_D)$ is $(1-t)(\delta+\epsilon)$–quasiequivalent to $(C^t_*, \partial^t, \ell^t)$. So extending the family $(C^t_*, \partial^t, \ell^t)$ to all $t \in [0, 1]$ by setting $(C^0_*, \partial^0, \ell^0) = (C_*, \partial_C, \ell_C)$ and $(C^1_*, \partial^1, \ell^1) = (D_*, \partial_D, \ell_D)$, $\{(C^t_*, \partial^t, \ell^t)\}_{t\in[0,1]}$ gives the desired $(\delta+\epsilon)$–interpolation between $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$. □

12 Applications in Hamiltonian Floer theory

We now bring our general algebraic theory into contact with Hamiltonian Floer theory on compact symplectic manifolds, leading to a rigidity result for fixed points of Hamiltonian diffeomorphisms. First we quickly review the geometric content of the Hamiltonian Floer complex; see eg [17; 24; 1] for more background, details, and proofs.

Let (M, ω) be a compact symplectic manifold. Identifying $S^1 = \mathbb{R}/\mathbb{Z}$, a smooth function $H \colon S^1 \times M \to \mathbb{R}$ induces a family of diffeomorphisms $\{\phi^t_H\}_{t\in\mathbb{R}}$ obtained as the flow of the time-dependent vector field $X_{H(t,\cdot)}$ that is characterized by the property that, for all t, $\omega(\cdot, X_{H(t,\cdot)}) = d(H(t,\cdot))$. Let

$$\mathcal{P}(H) = \{\gamma\colon S^1 \to M \mid \gamma(t) = \phi_H^t(\gamma(0)),\ \gamma \text{ is contractible}\},$$

so that in particular $\mathcal{P}(H)$ is in bijection with a subset of the fixed point set of ϕ_H^1 via the map $\gamma \mapsto \gamma(0) \in M$. The Hamiltonian H is called *nondegenerate* if for each $\gamma \in \mathcal{P}(H)$ the linearized map $(d\phi_H^1)_{\gamma(0)}\colon T_{\gamma(0)}M \to T_{\gamma(0)}M$ has all eigenvalues distinct from 1. Generic Hamiltonians H satisfy this property. We will assume in what follows that H is nondegenerate, which guarantees in particular that $\mathcal{P}(H)$ is a finite set.

Viewing S^1 as the boundary of the disk D^2 in the usual way, given $\gamma \in \mathcal{P}(H)$ and a map $u\colon D^2 \to M$ with $u|_{S^1} = \gamma$, one has a well-defined "action"

$$\int_0^1 H(t,\gamma(t))\,dt - \int_{D^2} u^*\omega$$

and Conley–Zehnder index. Define $\widetilde{\mathcal{P}}(H)$ to be the set of equivalence classes $[\gamma, u]$ of pairs (γ, u) where $\gamma \in \mathcal{P}(H)$, $u\colon D^2 \to M$ has $u|_{S^1} = \gamma$, and (γ, u) is equivalent to (γ', v) if and only if $\gamma = \gamma'$ and the map $u \# \bar{v}\colon S^2 \to M$ obtained by gluing u and v along γ has both vanishing ω–area and vanishing first Chern number. Then there are well-defined maps $\mathcal{A}_H\colon \widetilde{\mathcal{P}}(H) \to \mathbb{R}$ and $\mu\colon \widetilde{\mathcal{P}}(H) \to \mathbb{Z}$ defined by setting $\mathcal{A}_H([\gamma,u]) = \int_0^1 H(t,\gamma(t))\,dt - \int_{D^2} u^*\omega$ and $\mu([\gamma,u])$ equal to the Conley–Zehnder index of the path of symplectic matrices given by expressing $\{(d\phi_H^t)_{\gamma(0)}\}_{t\in[0,1]}$ in terms of a symplectic trivialization of u^*TM.

The degree-k part of the Floer chain complex $\mathrm{CF}_k(H)$ is then by definition (using the ground field $\mathcal{K}$)

$$\left\{ \sum_{\substack{[\gamma,u]\in\widetilde{\mathcal{P}}(H)\\ \mu([\gamma,u])=k}} a_{[\gamma,u]}[\gamma,u] \;\middle|\; a_{[\gamma,u]} \in \mathcal{K} \text{ and } \#\Theta_C < \infty \text{ for all } C \in \mathbb{R} \right\},$$

where

$$\Theta_C = \{[\gamma,u] \mid a_{[\gamma,u]} \neq 0,\ \mathcal{A}_H([\gamma,u]) > C\}.$$

Let

$$\Gamma = \left\{ \int_{S^2} w^*\omega \;\middle|\; w\colon S^2 \to M,\ \langle c_1(TM), w_*[S^2]\rangle = 0 \right\}. \tag{40}$$

Then $\mathrm{CF}_k(H)$ is a vector space over $\Lambda = \Lambda^{\mathcal{K},\Gamma}$, with the scalar multiplication obtained from the action of Γ on $\mathcal{P}(H)$ given by, for $g \in \Gamma$ and $[\gamma,u] \in \widetilde{\mathcal{P}}(H)$, gluing a sphere of Chern number zero and area g to u.

We make $\mathrm{CF}_k(H)$ into a nonarchimedean normed vector space over Λ by setting

$$\ell_H\Big(\sum a_{[\gamma,u]}[\gamma,u]\Big)=\max\{\mathcal{A}_H([\gamma,u])\mid a_{[\gamma,u]}\neq 0\}.$$

Define

$$\mathcal{P}_k(H)=\big\{\gamma\in\mathcal{P}(H)\,|\,\text{there exists } u\colon D^2\to M \text{ with } u|_{S^1}=\gamma,\ \mu([\gamma,u])=k\big\}. \tag{41}$$

Then it is easy to see that an orthogonal ordered basis for $\mathrm{CF}_k(H)$ is given by $([\gamma_1,u_1],\dots,[\gamma_{n_k},u_{n_k}])$, where $\gamma_1,\dots,\gamma_{n_k}$ are the elements of $\mathcal{P}_k(H)$ and, for each i, u_i is an arbitrarily chosen map $D^2\to M$ with $u_i|_{\partial D^2}=\gamma_i$ and $\mu([\gamma_i,u_i])=k$. In particular $(\mathrm{CF}_k(H),\ell_H)$ is an orthogonalizable Λ–space.

The function $\mathcal{A}_H$ introduced above could just as well have been defined on the cover of the entire space of contractible loops of M obtained by dropping the condition that $\gamma\in\mathcal{P}(H)$; then $\widetilde{\mathcal{P}}(H)$ is the set of critical points of this extended functional. The degree-k part of the Floer boundary operator $(\partial_H)_k\colon \mathrm{CF}_k(H)\to\mathrm{CF}_{k-1}(H)$ is constructed by counting isolated formal negative gradient flowlines of this extended version of $\mathcal{A}_H$ in the usual way indicated in the introduction. It is a deep but (at least when (M,ω) is semipositive, but see [35] for the more general case) by now standard fact that ∂_H can indeed be defined in this way, so that the resulting triple $(\mathrm{CF}_*(H),\partial_H,\ell_H)$ obeys the axioms of a Floer-type complex; thus in every degree k we obtain a concise barcode $\mathcal{B}_{\mathrm{CF}_*(H),k}$. The construction of ∂_H depends on some auxiliary choices, but the filtered chain isomorphism type of $(\mathrm{CF}_*(H),\partial_H,\ell_H)$ is independent of these choices (see eg [44, Lemma 1.2]), so $\mathcal{B}_{\mathrm{CF}_*(H),k}$ is an invariant of H.

Proposition 12.1 *For nondegenerate Hamiltonians $H_0,H_1\colon S^1\times M\to\mathbb{R}$ on a compact symplectic manifold, the associated Floer chain complexes $(\mathrm{CF}_*(H_0),\partial_{H_0},\ell_{H_0})$ and $(\mathrm{CF}_*(H_1),\partial_{H_1},\ell_{H_1})$ obey*

$$d_P\big((\mathrm{CF}_*(H_0),\partial_{H_0},\ell_{H_0}),(\mathrm{CF}_*(H_1),\partial_{H_1},\ell_{H_1})\big)\le\int_0^1\|H_1(t,\cdot)-H_0(t,\cdot)\|_{L^\infty}\,dt.$$

Proof Write $\delta=\int_0^1\|H_1(t,\cdot)-H_0(t,\cdot)\|_{L^\infty}\,dt$ and let $\epsilon>0$; we will show that there exists a $(\delta+\epsilon)$–interpolation between $(\mathrm{CF}_*(H_0),\partial_{H_0},\ell_{H_0})$ and $(\mathrm{CF}_*(H_1),\partial_{H_1},\ell_{H_1})$.

Define $\widehat{H}^0\colon[0,1]\times S^1\times M\to\mathbb{R}$ by $\widehat{H}^0(s,t,m)=sH_1(t,m)+(1-s)H_0(t,m)$. A standard argument with the Sard–Smale theorem (see eg [29, Propositions 6.1.2 and 6.1.3]) shows that, arbitrarily close to $\widehat{H}^0$ in the C^1–norm, there is a smooth map $\widehat{H}\colon[0,1]\times S^1\times M\to\mathbb{R}$ such that

- $\widehat{H}(0,t,m) = H_0(t,m)$ and $\widehat{H}(1,t,m) = H_1(t,m)$ for all $(t,m) \in S^1 \times M$, and
- there are only finitely many $s \in [0,1]$ with the property that $H(s,\cdot,\cdot)\colon S^1 \times M \to \mathbb{R}$ fails to be nondegenerate.

In particular we can take $\widehat{H}$ to be so C^1–close to $\widehat{H}^0$ that $\|\partial\widehat{H}/\partial s - \partial\widehat{H}^0/\partial s\|_{L^\infty} < \epsilon$.

For $s \in [0,1]$ write $\widehat{H}_s(t,m) = \widehat{H}(s,t,m)$. Then for $0 \le s_0 \le s_1 \le 1$ and $(t,m) \in S^1 \times M$ we have

$$\begin{aligned}|\widehat{H}_{s_1}(t,m) - \widehat{H}_{s_0}(t,m)| &= \left|\int_{s_0}^{s_1} \frac{\partial\widehat{H}}{\partial s}(s,t,m)\,ds\right| \\ &\le \epsilon(s_1 - s_0) + \int_{s_0}^{s_1} \left|\frac{\partial\widehat{H}^0}{\partial s}(s,t,m)ds\right| ds \\ &= (\epsilon + |H_1(t,m) - H_0(t,m)|)(s_1 - s_0).\end{aligned}$$

Thus, for any $s_0, s_1 \in [0,1]$,

$$\begin{aligned}(42)\quad \int_0^1 \|\widehat{H}_{s_1}(t,\cdot) - \widehat{H}_{s_0}(t,\cdot)\|_{L^\infty}\,dt& \\ &\le \left(\epsilon + \int_0^1 \|H_1(t,\cdot) - H_0(t,\cdot)\|_{L^\infty}\,dt\right)|s_1 - s_0| \\ &= (\delta + \epsilon)|s_1 - s_0|.\end{aligned}$$

Let $S = \{s \in [0,1] \mid \widehat{H}_s \text{ is not nondegenerate}\}$, so by construction S is a finite set, and for $s \in [0,1] \setminus S$ we have a Floer-type complex $(\mathrm{CF}_*(\widehat{H}_s), \partial_{\widehat{H}_s}, \ell_{\widehat{H}_s})$. Standard facts from filtered Hamiltonian Floer theory (summarized for instance in [45, Proposition 5.1], though note that the definition of quasiequivalence there is slightly different from ours) show that, for $s_0, s_1 \in [0,1] \setminus S$, the Floer-type complexes $(\mathrm{CF}_*(\widehat{H}_{s_0}), \partial_{\widehat{H}_{s_0}}, \ell_{\widehat{H}_{s_0}})$ and $(\mathrm{CF}_*(\widehat{H}_{s_1}), \partial_{\widehat{H}_{s_1}}, \ell_{\widehat{H}_{s_1}})$ are $\left(\int_0^1 \|\widehat{H}_{s_1}(t,\cdot) - \widehat{H}_{s_0}(t,\cdot)\|_{L^\infty}\,dt\right)$–quasiequivalent, and hence $(\delta+\epsilon)|s_1 - s_0|$–quasiequivalent by (42).

Thus we see that the family $(\mathrm{CF}_*(\widehat{H}_s), \partial_{\widehat{H}_s}, \ell_{\widehat{H}_s})$ defines a $(\delta + \epsilon)$–interpolation between $(\mathrm{CF}_*(H_0), \partial_{H_0}, \ell_{H_0})$ and $(\mathrm{CF}_*(H_1), \partial_{H_1}, \ell_{H_1})$. Since this construction can be carried out for all $\epsilon > 0$ the result immediately follows. □

Combining this proposition with Theorem 11.2, we immediately get the following result:

Corollary 1.5 *If H_0 and H_1 are two nondegenerate Hamiltonians on any compact symplectic manifold (M,ω), then the bottleneck distance between the concise barcodes of $(\mathrm{CF}_*(H_0), \partial_{H_0}, \ell_{H_0})$ and $(\mathrm{CF}_*(H_1), \partial_{H_1}, \ell_{H_1})$ is less than or equal to $\int_0^1 \|H_1(t,\cdot) - H_0(t,\cdot)\|_{L^\infty}\,dt$.*

Similar results apply to the way in which the barcodes of Lagrangian Floer complexes $\mathrm{CF}(L_0, \phi_H^1(L_1))$ depend on the Hamiltonian H, or for that matter to the dependence of Novikov complexes $\mathrm{CN}_*(\widetilde{f})$ on the function $\widetilde{f}\colon \widetilde{M} \to \mathbb{R}$. When Γ is nontrivial these facts do not follow from previously known results. (When Γ is trivial they can be inferred from [8] and standard Floer-theoretic results like [45, Proposition 5.1].)

We now give an application of Corollary 1.5 to fixed points of Hamiltonian diffeomorphisms. Apart from its intrinsic interest, we also intend this as an illustration of how to use the methods developed in this paper.

It will be relevant that the Floer-type complex $(\mathrm{CF}_*(H), \partial_H, \ell_H)$ of a nondegenerate Hamiltonian on a compact symplectic manifold obeys the additional property that $\ell_H(\partial_H c) < \ell_H(c)$ for all $c \in \mathrm{CF}_*(H)$, rather than the weaker inequality "$\leq$" which is generally required in the definition of a Floer-type complex (this standard fact follows because the boundary operator ∂_H counts nonconstant formal negative gradient flowlines of $\mathcal{A}_H$, and the function $\mathcal{A}_H$ strictly decreases along such flowlines). Consequently there can be no elements of the form $([a], 0)$ in the verbose barcode of $(\mathrm{CF}_*(H), \partial_H, \ell_H)$ in any degree k, as such an element would correspond to elements $x \in \mathrm{CF}_k(H)$ and $y \in \mathrm{CF}_{k+1}(H)$ with $\partial_H y = x$ and $\ell_H(y) = \ell_H(x)$. In other words, for each degree k, the verbose barcode $\widetilde{\mathcal{B}}_{\mathrm{CF}_*(H),k}$ of $(\mathrm{CF}_*(H), \partial_H, \ell_H)$ is equal to its concise barcode $\mathcal{B}_{\mathrm{CF}_*(H),k}$.

To state the promised result, recall the notation $\mathcal{P}_k(H)$ from (41), and for any subset $E \subset \mathbb{R}$, define

$$\mathcal{P}_k^E(H_0) = \{\gamma \in \mathcal{P}_k(H) \mid \exists u\colon D^2 \to M \text{ with } u|_{S^1} = \gamma,\ \mathcal{A}_{H_0}([\gamma, u]) \in E,\ \mu([\gamma, u]) = k\}.$$

Theorem 12.2 *Let $H_0\colon S^1 \times M \to \mathbb{R}$ be a nondegenerate Hamiltonian on a compact symplectic manifold (M, ω), let $k \in \mathbb{Z}$, let $E \subset \mathbb{R}$ be any subset, and let $\Delta^E > 0$ be the minimum of the following two quantities:*

- *The smallest second coordinate L of any element $([a], L)$ of the degree-k part $\mathcal{B}_{\mathrm{CF}_*(H_0),k}$ of the concise barcode such that some representative a of the coset $[a]$ belongs to E.*
- *The smallest second coordinate of any $([a], L) \in \mathcal{B}_{\mathrm{CF}_*(H_0),k-1}$ such that some $a \in [a]$ has $a + L \in E$.*

Let $H\colon S^1 \times M \to \mathbb{R}$ be any nondegenerate Hamiltonian with

$$\int_0^1 \|H(t,\cdot) - H_0(t,\cdot)\|_{L^\infty}\, dt < \tfrac{1}{2}\Delta^E.$$

Then there is an injection $f\colon \mathcal{P}_k^E(H_0)\to\mathcal{P}_k(H)$ and, for each $\gamma\in\mathcal{P}_k(H_0)$, maps $u,\tilde{u}\colon D^2\to M$ with $u|_{S^1}=\gamma$ and $\tilde{u}|_{S^1}=f(\gamma)$ such that

$$|\mathcal{A}_H([f(\gamma),\tilde{u}])-\mathcal{A}_{H_0}([\gamma,u])|\le\int_0^1\|H(t,\cdot)-H_0(t,\cdot)\|_{L^\infty}\,dt.$$

Proof As in the proof of Proposition 7.4, we can find singular value decompositions for $(\partial_{H_0})_{k+1}\colon \mathrm{CF}_{k+1}(H_0)\to\ker(\partial_{H_0})_k$ and $(\partial_{H_0})_k\colon \mathrm{CF}_k(H_0)\to\ker(\partial_{H_0})_{k-1}$ having the forms

$$((y_1^k,\dots,y_{r_k}^k,x_1^{k+1},\dots,x_{m_{k+1}}^{k+1}),(x_1^k,\dots,x_{m_k}^k))$$

and

$$((y_1^{k-1},\dots,y_{r_{k-1}}^{k-1},x_1^k,\dots,x_{m_k}^k),(x_1^{k-1},\dots,x_{m_{k-1}}^{k-1})),$$

respectively. In particular $(y_1^{k-1},\dots,y_{r_{k-1}}^{k-1},x_1^k,\dots,x_{m_k}^k)$ is an orthogonal ordered basis for $\mathrm{CF}_k(H_0)$. Write the elements of $\mathcal{P}_k(H_0)$ as $\gamma_1,\dots,\gamma_n$, ordered in such a way that $\mathcal{P}_k^E(H_0)=\{\gamma_1,\dots,\gamma_s\}$ for some $s\le n$. As discussed before the statement of the theorem, if for each $i\in\{1,\dots,n\}$ we choose an arbitrary $u_i\colon D^2\to M$ with $u_i|_{S^1}=\gamma_i$ and $\mu([\gamma_i,u_i])=k$, and moreover $\mathcal{A}_{H_0}([\gamma_i,u_i])\in E$ for $i=1,\dots,s$, then $([\gamma_1,u_1],\dots,[\gamma_n,u_n])$ will be an orthogonal ordered basis for $\mathrm{CF}_k(H_0)$. So by Proposition 5.5 and the definition of ℓ_{H_0}, there is a bijection $\alpha\colon\mathcal{P}_k(H_0)\to\{y_1^{k-1},\dots,y_{r_{k-1}}^{k-1},x_1^k,\dots,x_{m_k}^k\}$ such that $\ell_{H_0}(\alpha(\gamma_i))\equiv\mathcal{A}_{H_0}([\gamma_i,u_i])\pmod{\Gamma}$.

If $\alpha(\gamma_i)=y_{j_i}^{k-1}$ for some $j_i\in\{1,\dots,r_{k-1}\}$, then the element

$$([a_i],L_i):=([\ell_{H_0}(x_{j_i}^{k-1})],\ell_{H_0}(y_{j_i}^{k-1})-\ell_{H_0}(x_{j_i}^{k-1}))$$

of the degree-$(k-1)$ verbose barcode of $(\mathrm{CF}_*(H_0),\partial_{H_0},\ell_{H_0})$ corresponds to a capped orbit $[\gamma_i,u_i]$ having filtration $\mathcal{A}_H([\gamma_i,u_i])\equiv a_i+L_i\pmod{\Gamma}$. Otherwise, $\alpha(\gamma_i)=x_{j_i}^k$ for some $j_i\in\{1,\dots,m_k\}$, and then we have an element $([a_i],L_i)$ of the degree-k verbose barcode of $(\mathrm{CF}_*(H_0),\partial_{H_0},\ell_{H_0})$, where

$$a_i=\ell_{H_0}(x_{j_i}^k)\quad\text{and}\quad L_i=\begin{cases}\ell_{H_0}(y_{j_i}^k)-\ell_{H_0}(x_{j_i}^k) & \text{if } 1\le i\le m_k,\\ \infty & \text{otherwise;}\end{cases}$$

in this case $\mathcal{A}_H([\gamma_i,u_i])\equiv a_i\pmod{\Gamma}$. As noted before the theorem, the verbose barcode of $(\mathrm{CF}_*(H_0),\partial_{H_0},\ell_{H_0})$ is the same in every degree as its concise barcode, so in particular these elements (a_i,L_i) of the verbose barcodes belong to the concise barcodes $\mathcal{B}_{\mathrm{CF}_*(H_0),k}$ or $\mathcal{B}_{\mathrm{CF}_*(H_0),k-1}$.

Considering now our new Hamiltonian H, write

$$\delta=\int_0^1\|H(t,\cdot)-H_0(t,\cdot)\|_{L^\infty}.$$

Our hypothesis, along with the fact that $\mathcal{A}_{H_0}([\gamma_i,u_i])\in E$ for $i=1,\dots,s$, then guarantees that, for $i=1,\dots,s$, the elements $([a_i],L_i)$ of the concise barcodes $\mathcal{B}_{\mathrm{CF}_*(H_0),k}$ or $\mathcal{B}_{\mathrm{CF}_*(H_0),k-1}$ described in the previous paragraph all have $L_i\ge\Delta^E>2\delta$. On the other hand Corollary 1.5 implies that there is a partial matching $\mathfrak{m}_k$ between $\mathcal{B}_{\mathrm{CF}_*(H_0),k}$ and $\mathcal{B}_{\mathrm{CF}_*(H),k}$, and likewise a partial matching $\mathfrak{m}_{k-1}$ between $\mathcal{B}_{\mathrm{CF}_*(H_0),k-1}$ and $\mathcal{B}_{\mathrm{CF}_*(H),k-1}$, with both $\mathfrak{m}_k$ and $\mathfrak{m}_{k-1}$ having defects at most δ. So since each $L_i>2\delta$, none of the elements $([a_i],L_i)$ for $i=1,\dots,s$ can be unmatched under these partial matchings. So each of them is matched to an element, say $([\tilde{a}_i],\tilde{L}_i)$, of the degree-$k$ or $k-1$ concise barcode of $(\mathrm{CF}_*(H),\partial_H,\ell_H)$. We will denote the multiset of all such "targets" by

$$\mathcal{T}_{k,k-1}=\{([\tilde{a}_i],\tilde{L}_i)\mid i=1,\dots,s\}. \tag{43}$$

Since the defect of our partial matching is at most δ, we can each choose $\tilde{a}_i$ within its Γ–coset so that $|\tilde{a}_i-a_i|\le\delta$ and either $\tilde{L}_i=L_i=\infty$ or $|(\tilde{a}_i+\tilde{L}_i)-(a_i+L_i)|\le\delta$.

We now apply the reasoning that was used at the start of the proof to $\mathrm{CF}_*(H)$ in place of $\mathrm{CF}_*(H_0)$. We may consider singular value decompositions for the maps $(\partial_H)_{k+1}$ and $(\partial_H)_k$ on $\mathrm{CF}_*(H)$ having the forms

$$((z_1^k,\dots,z_{r_k'}^k,w_1^{k+1},\dots,w_{m_{k+1}'}^{k+1}),(w_1^k,\dots,w_{m_k'}^k))$$

and

$$((z_1^{k-1},\dots,z_{r_{k-1}'}^{k-1},w_1^k,\dots,w_{m_k'}^k),(w_1^{k-1},\dots,w_{m_{k-1}'}^{k-1})),$$

respectively. Then if the elements of $\mathcal{P}_k(H)$ are written as $\{\eta_1,\dots,\eta_p\}$, we may choose $v_j\colon D^2\to M$ with $v_j|_{S^1}=\eta_j$ for each $j\in\{1,\dots,p\}$ in such a way that the multiset of real numbers $\mathcal{A}_H([\eta_j,v_j])$ is equal to the multiset

$$\{\ell_H(z_j^{k-1})\mid 1\le j\le r_{k-1}'\}\cup\{\ell_H(w_j^k)\mid 1\le j\le m_k'\}.$$

This equality of multisets gives an injection ι from the submultiset $\mathcal{T}_{k,k-1}\subset\mathcal{B}_{\mathrm{CF}_*(H),k}\cup\mathcal{B}_{\mathrm{CF}_*(H),k-1}$ described in (43) to $\mathcal{P}_k(H)$. Specifically:

- For $i\in\{1,\dots,s\}$ such that $\alpha(\gamma_i)=y_{j_i}^{k-1}$, the element $([\tilde{a}_i],\tilde{L}_i)$ belongs to $\mathcal{B}_{\mathrm{CF}_*(H),k-1}$, and $\iota([\tilde{a}_i],\tilde{L}_i])$ will be some $\eta_{q_i}\in\mathcal{P}_k(H)$ with $\mathcal{A}_H([\eta_{q_i},v_{q_i}])=\tilde{a}_i+\tilde{L}_i$;
- For $i\in\{1,\dots,s\}$ such that $\alpha(\gamma_i)=x_{j_i}^k$, the element $([\tilde{a}_i],\tilde{L}_i)$ belongs to $\mathcal{B}_{\mathrm{CF}_*(H),k}$, and $\iota([\tilde{a}_i],\tilde{L}_i])$ will be some η_{q_i} with $\mathcal{A}_H([\eta_{q_i},v_{q_i}])=\tilde{a}_i$.

The map $f\colon\mathcal{P}_k^E(H_0)\to\mathcal{P}_k(H)$ promised in the theorem is then the one which sends each γ_i to η_{q_i}; the fact that this obeys the required properties follows directly from the inequalities $|\tilde{a}_i-a_i|\le\delta$ and $|(\tilde{a}_i+\tilde{L}_i)-(a_i+L_i)|\le\delta$ and the fact that the value of $\mathcal{A}_H([\gamma_{q_i}, v_{q_i}])$ can be varied within its Γ–coset, without changing the grading k, by using a different choice of capping disk v_{q_i}. □

Remark 12.3 Theorem 12.2 may be applied with $E = \mathbb{R}$, in which case it shows that if $\int_0^1 \|H(t,\cdot) - H_0(t,\cdot)\|_{L^\infty}\, dt$ is less than half of the minimal second coordinate of the concise barcode of $\mathrm{CF}_*(H_0)$ in any degree, then the time-one flow of the perturbed Hamiltonian H will have at least as many fixed points[6] as that of the original Hamiltonian H_0. This may appear somewhat surprising, as a C^0–small perturbation of the Hamiltonian function H can still rather dramatically alter the Hamiltonian vector field X_H, which depends on the derivative of H. However this basic phenomenon is by now rather well-known in symplectic topology; see in particular [11, Theorem 2.1] and [44, Corollary 2.3], though these other results do not give control over the values of $\mathcal{A}_H$ on $\widetilde{\mathcal{P}}_k(H)$ as in Theorem 12.2.

For a more general choice of E our result does not appear to have analogues in the literature, particularly when $\Gamma \neq \{0\}$; this generalization is of interest when Δ^E, thought of as the minimal length of a barcode interval with endpoint lying in E, is larger than the minimal length $\Delta^{\mathbb{R}}$ of all barcode intervals, in which case the Theorem shows that fixed points of $\phi^1_{H_0}$ with action lying in E enjoy a robustness that the other fixed points of $\phi^1_{H_0}$ may not. For instance in the case that $E = \{a_0\}$ is a singleton and there is just one element $[\gamma_0, u_0]$ of $\widetilde{\mathcal{P}}_k$ having $\mathcal{A}_H([\gamma_0, u_0]) = a_0$, then Δ^E is bounded below by the lowest energy of a Floer trajectory converging to γ_0 in positive or negative time, whereas $\Delta^{\mathbb{R}}$ is bounded below by the lowest energy of *all* Floer trajectories, which might be much smaller.

In the special case that both $\Gamma = \{0\}$ and $E = \{a_0\}$ a version of Theorem 12.2 can be obtained using a standard argument in terms of the "action window" Floer homologies $\mathrm{HF}_*^{[a,b]}(H)$ of the quotient complexes

$$\frac{\{c \in \mathrm{CF}_*(H) \mid \ell_H(c) \le b\}}{\{c \in \mathrm{CF}_*(H) \mid \ell_H(c) < a\}}.$$

Indeed, for any $\delta \in \mathbb{R}$ such that

$$\int_0^1 \|H(t,\cdot) - H_0(t,\cdot)\|_{L^\infty}\, dt < \delta < \tfrac{1}{2}\Delta^E,$$

we will have a commutative diagram of continuation maps (induced by appropriate monotone homotopies; see [25, Section 6.6])

[6] The fixed points have contractible orbit under ϕ^t_H, though one can drop this restriction by using a straightforward variant of the Floer complex built from noncontractible orbits.

$$\begin{tikzcd}
\mathrm{HF}_k^{[a_0-\delta,a_0+\delta]}(H_0+\delta) \arrow[rr, "\Phi"] \arrow[dr] & & \mathrm{HF}_k^{[a_0-\delta,a_0+\delta]}(H_0-\delta) \\
& \mathrm{HF}_k^{[a_0-\delta,a_0+\delta]}(H) \arrow[ur] &
\end{tikzcd}$$

and the hypothesis on the barcode can be seen to imply that the above map Φ has rank at least equal to $\#\mathcal{P}_k^E(H_0)$, whence $\mathrm{HF}_k^{[a_0-\delta,a_0+\delta]}(H)$ has dimension at least equal to $\#\mathcal{P}_k^E(H_0)$. When $\Gamma=\{0\}$ this last statement implies that the number of fixed points of the time-one flow of H with action in the interval $[a_0-\delta,a_0+\delta]$ is at least $\#\mathcal{P}_k^E(H_0)$. However for $\Gamma\neq\{0\}$ the implication in the previous sentence may not be valid, since the above argument only estimates the dimension of $\mathrm{HF}_k^{[a_0-\delta,a_0+\delta]}(H)$ over $\mathcal{K}$, and the contribution of a single fixed point to $\dim_{\mathcal{K}}\mathrm{HF}_k^{[a_0-\delta,a_0+\delta]}(H)$ might be greater than one due to recapping.

Thus Theorem 12.2 provides a way of avoiding difficulties with recapping that arise in arguments with action window Floer homology when $\Gamma\neq\{0\}$. Even when $\Gamma=\{0\}$, if E consists of, say, of two or more real numbers that are a distance less than $\Delta^E/2$ away from each other, then Theorem 12.2 can be seen to give sharper results than are obtained by action window arguments such as those described in the previous paragraph.

Appendix: Interleaving distance

In this brief appendix, we will discuss the relation of our quasiequivalence distance d_Q to the notion of *interleaving*, which is often used (eg in [8]) as a measure of proximity between persistence modules. Because the main objects of the paper are Floer-type complexes, rather than the persistence modules given by their filtered homologies, we will use the following definition; on passing to homology this gives (at least in principle) a slightly different notion than that used in [8], as the maps on filtered homology in [8] are not assumed to be induced by maps on the original chain complexes.

Definition A.1 For $\delta\geq 0$, a *chain level δ–interleaving* of two Floer-type complexes (C_*,∂_C,ℓ_C) and (D_*,∂_D,ℓ_D) is a pair (Φ,Ψ) of chain maps $\Phi\colon C_*\to D_*$ and $\Psi\colon D_*\to C_*$ such that:

- $\ell_D(\Phi c)\leq \ell_C(c)+\delta$ for all $c\in C_*$.
- $\ell_D(\Psi d)\leq \ell_D(d)+\delta$ for all $d\in D_*$.
- For all $\lambda\in\mathbb{R}$ the compositions $\Psi\Phi\colon C_*^\lambda\to C_*^{\lambda+2\delta}$ and $\Phi\Psi\colon D_*^\lambda\to D_*^{\lambda+2\delta}$ induce the same maps on homology as the respective inclusions.

It is easy to see that a chain level δ–interleaving induces maps $\Phi_*\colon H^\lambda(C_*) \to H^{\lambda+\delta}(D_*)$ and $\Psi_*\colon H^\lambda(D_*) \to H^{\lambda+\delta}(C_*)$ (as λ varies through $\mathbb{R}$) which give a strong δ–interleaving between the persistence modules $\{H^\lambda(C_*)\}$ and $\{H^\lambda(D_*)\}$ in the sense of [8]. It is also easy to see that if (Φ, Ψ, K_C, K_D) is a δ–quasiequivalence between $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$, then (Φ, Ψ) is a chain level δ–interleaving. We will see that the converse of this latter statement is true provided that Φ and Ψ are split in the sense of Section 9.2.

Lemma A.2 *Let F_*^C be a splitting of a Floer-type complex $(C_*, \partial_C, \ell_C)$, and suppose that $A\colon C_* \to C_*$ is a chain map which is split with respect to this splitting, such that there exists $\epsilon > 0$ such that $\ell_C(Ac) \le \ell_C(c) + \epsilon$ for all $c \in C_*$ and, for all $\lambda \in \mathbb{R}$, the induced map $A_*\colon H_*(C_*^\lambda) \to H_*(C_*^{\lambda+\epsilon})$ is zero. Then there exists a map $K\colon C_* \to C_{*+1}$ such that $\ell_C(Kc) \le \ell_C(c) + \epsilon$ for all $c \in C_*$ and $A = \partial_C K + K\partial_C$.*

Proof Let $B_* = \mathrm{Im}(\partial_C)_{*+1}$. Then the boundary operator ∂_C restricts as an isomorphism $(\partial_C)_{*+1}\colon F_{*+1}^C \to B_*$. Let $L_* = \oplus_k L_k$, where each L_k is a complement to B_k in $\ker(\partial_C)_k$, so that $\ker(\partial_C)_* = B_* \oplus L_*$,

Let $s\colon C_* \to C_{*+1}$ be the linear map such that $s|_{L_* \oplus F_*} = 0$ and $s|_{B_*} = (\partial_C|_{F_{*+1}})^{-1}$. Therefore, $\partial_C s|_{B_*}$ is the identity map on B_*, and for any $b \in B_*$, $s(b)$ is the unique element of F_*^C such that $\partial_C s(b) = b$. Moreover, because F_{*+1}^C is orthogonal to $\ker(\partial_C)_{*+1}$ we have

$$\ell_C(s(b)) = \inf\{\ell_C(c) \mid c \in C_{*+1}, \partial_C c = b\}. \tag{44}$$

Now let $K = sA$; we will check that $A = \partial_C K + K\partial_C$. Indeed:

(i) For $x \in \ker(\partial_C)_*$, we have $(\partial_C K + K\partial_C)x = \partial_C Kx = \partial_C sAx = Ax$, since $Ax \in B_*$ by the hypothesis on $A_*\colon H_*(C_*^\lambda) \to H_*(C_*^{\lambda+\epsilon})$.

(ii) For $y \in F_*^C$, since A is split and so $Ay \in F_*^C$, $Ky = sAy = 0$. Therefore, $(\partial_C K + K\partial_C)y = sA\partial_C y = s\partial_C Ay = Ay$, where the last equality comes from the fact that $\partial_C s\partial_C Ay = \partial_C Ay$ and that both $s\partial_C Ay$ and Ay belong to F_*^C, together with the injectivity of $\partial_C|_{F_*^C}$.

Finally, by the hypothesis that each $A_*\colon H_*(C_*^\lambda) \to H_*(C_*^{\lambda+\epsilon})$ is zero, for any $x \in \ker(\partial_C)_*$, there exists some $z \in C_{*+1}$ such that $\partial_C z = Ax$ and $\ell_C(z) \le \ell_C(x) + \epsilon$. Since $Kx = sAx$ also obeys $\partial_C Kx = Ax$, (44) implies that

$$\ell_C(Kx) \le \ell_C(z) \le \ell_C(x) + \epsilon.$$

More generally any $c \in C_*$ can be written $c = x + f$ with $x \in \ker(\partial_C)_*$ and $f \in F_*^C$, and by definition $Kf = 0$, so

$$\ell_C(Kc) = \ell_C(Kx) \le \ell(x) + \epsilon \le \ell_C(c) + \epsilon,$$

where the final inequality follows from the orthogonality of $\ker(\partial_C)_*$ and F_*^C. □

Corollary A.3 *If there is a chain-level δ–interleaving between the Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$, then there exists a δ–quasiequivalence between $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$.*

Proof By Lemma 9.6, we can replace both Φ and Ψ by Φ^π and Ψ^π which are split with respect to splittings F_*^C and F_*^D of our two complexes; then we will have

$$(\Psi^\pi \circ \Phi^\pi - I_C)(F_*^C) \subset F_*^C \quad \text{and} \quad (\Phi^\pi \circ \Psi^\pi - I_D)(F_*^D) \subset F_*^D.$$

Note that, due to condition (ii) in Lemma 9.6, Φ^π and Ψ^π induce the same maps on homology as do Φ and Ψ, so the fact that (Φ, Ψ) is a chain level δ–interleaving implies that the maps

$$\Psi_*^\pi \Phi_*^\pi - I_{C*}\colon H^\lambda(C_*) \to H^{\lambda+2\delta}(C_*) \quad \text{and} \quad \Phi_*^\pi \Psi_*^\pi - I_{D*}\colon H^\lambda(D_*) \to H^{\lambda+2\delta}(D_*)$$

are all zero. Hence applying Lemma A.2 to $\Psi^\pi \Phi^\pi - I_C$ and to $\Phi^\pi \Psi^\pi - I_D$ gives maps K_C and K_D such that $(\Phi^\pi, \Psi^\pi, K_C, K_D)$ is a δ–quasiequivalence. □

In other words, if we define the (chain-level) interleaving distance d_I by, for any two Floer-type complexes $(C_*, \partial_C, \ell_C)$ and $(D_*, \partial_D, \ell_D)$,

$$d_I\big((C_*, \partial_C, \ell_C), (D_*, \partial_D, \ell_D)\big) = \inf\big\{\delta \ge 0 \mid \exists \text{ chain-level } \delta\text{–interleaving between } (C_*, \partial_C, \ell_C) \text{ and } (D_*, \partial_D, \ell_D)\big\},$$

then we have an equality of distance functions $d_I = d_Q$, where d_Q is the quasiequivalence distance.

References

[1] **M Audin**, **M Damian**, *Morse theory and Floer homology*, Springer, London (2014) MR3155456

[2] **S A Barannikov**, *The framed Morse complex and its invariants*, from "Singularities and bifurcations" (V I Arnol'd, editor), Adv. Soviet Math. 21, Amer. Math. Soc., Providence, RI (1994) 93–115 MR1310596

[3] **U Bauer**, **M Lesnick**, *Induced matchings and the algebraic stability of persistence barcodes*, J. Comput. Geom. 6 (2015) 162–191 MR3333456

[4] **P Biran**, **O Cornea**, *Rigidity and uniruling for Lagrangian submanifolds*, Geom. Topol. 13 (2009) 2881–2989 MR2546618

[5] **D Burghelea**, **T K Dey**, *Topological persistence for circle-valued maps*, Discrete Comput. Geom. 50 (2013) 69–98 MR3070542

[6] **D Burghelea**, **S Haller**, *Topology of angle valued maps, bar codes and Jordan blocks*, preprint (2013) `arXiv:1303.4328`

[7] **G Carlsson**, *Topology and data*, Bull. Amer. Math. Soc. 46 (2009) 255–308 MR2476414

[8] **F Chazal**, **D Cohen-Steiner**, **M Glisse**, **L Guibas**, **S Oudot**, *Proximity of persistence modules and their diagrams*, from "Computational geometry", ACM (2009) 237–246

[9] **F Chazal**, **V de Silva**, **M Glisse**, **S Oudot**, *The structure and stability of persistence modules*, preprint (2012) `arXiv:1207.3674`

[10] **D Cohen-Steiner**, **H Edelsbrunner**, **J Harer**, *Stability of persistence diagrams*, Discrete Comput. Geom. 37 (2007) 103–120 MR2279866

[11] **O Cornea**, **A Ranicki**, *Rigidity and gluing for Morse and Novikov complexes*, J. Eur. Math. Soc. 5 (2003) 343–394 MR2017851

[12] **W Crawley-Boevey**, *Decomposition of pointwise finite-dimensional persistence modules*, J. Algebra Appl. 14 (2015) art. id. 1550066, 8 pp. MR3323327

[13] **M Entov**, **L Polterovich**, *Calabi quasimorphism and quantum homology*, Int. Math. Res. Not. 2003 (2003) 1635–1676 MR1979584

[14] **M Farber**, *Topology of closed one-forms*, Mathematical Surveys and Monographs 108, Amer. Math. Soc., Providence, RI (2004) MR2034601

[15] **A Floer**, *An instanton-invariant for* 3*–manifolds*, Comm. Math. Phys. 118 (1988) 215–240 MR956166

[16] **A Floer**, *Morse theory for Lagrangian intersections*, J. Differential Geom. 28 (1988) 513–547 MR965228

[17] **A Floer**, *Symplectic fixed points and holomorphic spheres*, Comm. Math. Phys. 120 (1989) 575–611 MR987770

[18] **A Floer**, **H Hofer**, *Symplectic homology, I: Open sets in* $\mathbb{C}^n$, Math. Z. 215 (1994) 37–88 MR1254813

[19] **U Frauenfelder**, *The Arnold–Givental conjecture and moment Floer homology*, Int. Math. Res. Not. 2004 (2004) 2179–2269 MR2076142

[20] **K Fukaya**, **Y-G Oh**, **H Ohta**, **K Ono**, *Lagrangian intersection Floer theory: anomaly and obstruction, Volume I*, AMS/IP Studies in Advanced Mathematics 46, Amer. Math. Soc., Providence, RI (2009) MR2553465

[21] **K Fukaya**, **Y-G Oh**, **H Ohta**, **K Ono**, *Displacement of polydisks and Lagrangian Floer theory*, J. Symplectic Geom. 11 (2013) 231–268 MR3046491

[22] **K Fukaya**, **K Ono**, *Arnold conjecture and Gromov–Witten invariant*, Topology 38 (1999) 933–1048 MR1688434

[23] **R Ghrist**, *Barcodes: the persistent topology of data*, Bull. Amer. Math. Soc. 45 (2008) 61–75 MR2358377

[24] **H Hofer**, **D A Salamon**, *Floer homology and Novikov rings*, from "The Floer memorial volume" (H Hofer, C H Taubes, A Weinstein, E Zehnder, editors), Progr. Math. 133, Birkhäuser, Basel (1995) 483–524 MR1362838

[25] **H Hofer**, **E Zehnder**, *Symplectic invariants and Hamiltonian dynamics*, Birkhäuser, Basel (1994) MR1306732

[26] **V Humilière**, **R Leclercq**, **S Seyfaddini**, *Coisotropic rigidity and C^0–symplectic geometry*, Duke Math. J. 164 (2015) 767–799 MR3322310

[27] **K S Kedlaya**, *p–adic differential equations*, Cambridge Studies in Advanced Mathematics 125, Cambridge Univ. Press (2010) MR2663480

[28] **D Le Peutrec**, **F Nier**, **C Viterbo**, *Precise Arrhenius law for p–forms: the Witten Laplacian and Morse–Barannikov complex*, Ann. Henri Poincaré 14 (2013) 567–610 MR3035640

[29] **Y-J Lee**, *Reidemeister torsion in Floer–Novikov theory and counting pseudo-holomorphic tori, I*, J. Symplectic Geom. 3 (2005) 221–311 MR2199540

[30] **G Liu**, **G Tian**, *Floer homology and Arnold conjecture*, J. Differential Geom. 49 (1998) 1–74 MR1642105

[31] **A F Monna**, **T A Springer**, *Sur la structure des espaces de Banach non-archimédiens*, Nederl. Akad. Wetensch. Proc. Ser. A 27 (1965) 602–614 MR0187061

[32] **M Morse**, *The calculus of variations in the large*, Amer. Math. Soc. Colloq. Publ. 18, Amer. Math. Soc., New York (1934) MR1451874

[33] **S P Novikov**, *Multivalued functions and functionals: an analogue of the Morse theory*, Dokl. Akad. Nauk SSSR 260 (1981) 31–35 MR630459 In Russian

[34] **Y-G Oh**, *Construction of spectral invariants of Hamiltonian paths on closed symplectic manifolds*, from "The breadth of symplectic and Poisson geometry" (J E Marsden, T S Ratiu, editors), Progr. Math. 232, Birkhäuser, Boston (2005) 525–570 MR2103018

[35] **J Pardon**, *An algebraic approach to virtual fundamental cycles on moduli spaces of pseudo-holomorphic curves*, Geom. Topol. 20 (2016) 779–1034 MR3493097

[36] **L Polterovich**, **E Shelukhin**, *Autonomous Hamiltonian flows, Hofer's geometry and persistence modules*, Selecta Math. 22 (2016) 227–296 MR3437837

[37] **J Robbin**, **D Salamon**, *The Maslov index for paths*, Topology 32 (1993) 827–844 MR1241874

[38] **D Salamon**, *Lectures on Floer homology*, lecture notes (1997) Available at `https://people.math.ethz.ch/~salamon/PREPRINTS/floer.pdf`

[39] **M Schwarz**, *Morse homology*, Progress in Mathematics 111, Birkhäuser, Basel (1993) MR1239174

[40] **M Schwarz**, *On the action spectrum for closed symplectically aspherical manifolds*, Pacific J. Math. 193 (2000) 419–461 MR1755825

[41] **V de Silva**, **D Morozov**, **M Vejdemo-Johansson**, *Dualities in persistent (co)homology*, Inverse Problems 27 (2011) art. id. 124003, 1–17 MR2854319

[42] **M Usher**, *Spectral numbers in Floer theories*, Compos. Math. 144 (2008) 1581–1592 MR2474322

[43] **M Usher**, *Duality in filtered Floer–Novikov complexes*, J. Topol. Anal. 2 (2010) 233–258 MR2652908

[44] **M Usher**, *Boundary depth in Floer theory and its applications to Hamiltonian dynamics and coisotropic submanifolds*, Israel J. Math. 184 (2011) 1–57 MR2823968

[45] **M Usher**, *Hofer's metrics and boundary depth*, Ann. Sci. Éc. Norm. Supér. 46 (2013) 57–128 MR3087390

[46] **A Zomorodian**, **G Carlsson**, *Computing persistent homology*, Discrete Comput. Geom. 33 (2005) 249–274 MR2121296

Department of Mathematics, University of Georgia
Athens, GA 30602, United States

School of Mathematical Sciences, Tel Aviv University
Ramat Aviv, Tel Aviv 69978, Israel

usher@math.uga.edu, jzhang4518@gmail.com

http://alpha.math.uga.edu/~usher/, http://junzhangsite.wordpress.com/

Proposed: Leonid Polterovich
Seconded: Gang Tian, Yasha Eliashberg

Received: 9 April 2015
Revised: 9 December 2015

msp

Geometry & Topology 20 (2016) 3431–3517

Deformations of colored $\mathfrak{sl}_N$ link homologies via foams

DAVID E V ROSE
PAUL WEDRICH

We prove a conjectured decomposition of deformed $\mathfrak{sl}_N$ link homology, as well as an extension to the case of colored links, generalizing results of Lee, Gornik, and Wu. To this end, we use foam technology to give a completely combinatorial construction of Wu's deformed colored $\mathfrak{sl}_N$ link homologies. By studying the underlying deformed higher representation-theoretic structures and generalizing the Karoubi envelope approach of Bar-Natan and Morrison, we explicitly compute the deformed invariants in terms of undeformed type A link homologies of lower rank and color.

17B37, 57M25, 81R50

1 Introduction

1.1 Statement of results

Khovanov [16] introduced a homology theory categorifying the Jones polynomial. This homology theory for links in S^3 has proven to be a powerful topological invariant, leading eg to Rasmussen's combinatorial proof of the Milnor conjecture on the slice genus of torus knots [37]. Rasmussen's work built on earlier results of Lee [26], who studied a deformed version of Khovanov's link invariant. Khovanov's theory is controlled by the Frobenius algebra $\mathbb{C}[X]/\langle X^2\rangle$, which appears as the invariant of the unknot, and Lee showed that deforming this algebra to $\mathbb{C}[X]/\langle X^2-1\rangle$ leads to a link homology theory which at first glance seems trivial, assigning the direct sum of two copies of the vector space $\mathbb{C}$ to any knot. However, this link invariant surprisingly contains highly nontrivial topological information: Rasmussen shows how to define a concordance invariant from a filtration on the deformed link homology, which in particular gives a lower bound on the smooth slice genus of the knot.

Khovanov and Rozansky [22] used the theory of matrix factorizations to generalize Khovanov homology to a link homology theory (now called Khovanov–Rozansky homology) which categorifies the $\mathfrak{sl}_N$ link polynomial. This was later extended by Wu [43] and Yonezawa [44] to a categorified invariant of links whose components are colored by fundamental representations $\bigwedge^k\mathbb{C}^N$ of $\mathfrak{sl}_N$ for $0\leq k\leq N$. In these

Published: 21 December 2016 DOI: 10.2140/gt.2016.20.3431

theories, the underlying Frobenius algebra is isomorphic to $\mathbb{C}[X]/\langle X^N\rangle$. Following work of Gornik [12], Rasmussen [37] and Krasner [23], Wu defined deformed versions of $\mathfrak{sl}_N$ link homology [42], in which this algebra is deformed to $\mathbb{C}[X]/\langle P(X)\rangle$, where $P(X)$ is an arbitrary degree-N polynomial. Gornik [12] and Wu [41] showed that if $P(X)$ has simple roots, this invariant assigns the direct sum of N copies of the vector space $\mathbb{C}$ to any 1–colored knot. This result as well as Lee's, and their generalizations to the case of links, can be interpreted as saying that when $P(X)$ has simple roots, the 1–colored deformed homology of a link decomposes into the direct sum of $\mathfrak{sl}_1$ homologies of various sublinks, which are always 1–dimensional. Other deformations have been studied for $N=2$ by Khovanov [17] and $N=3$ by Mackaay and Vaz [34].

In this paper, we prove a vast generalization of these results, showing that the deformation of colored $\mathfrak{sl}_N$ link homology corresponding to a general degree-N monic polynomial $P(X)$ with root multiset Σ decomposes into type A link homologies of lower rank and color. To this end, we use foam technology to define deformed colored $\mathfrak{sl}_N$ link homologies $\mathrm{KhR}^{\Sigma}(-)$ and compare them to the undeformed colored $\mathfrak{sl}_M$ link homologies $\mathrm{KhR}^{\mathfrak{sl}_M}(-)$ constructed by Queffelec and Rose [36]. Precisely, we show:

Theorem 1.1 *Let $\mathcal{L}(a_1,\dots,a_k)$ be a k–component oriented, framed link with the i^{th} component colored by the fundamental $\mathfrak{sl}_N$ representations $\bigwedge^{a_i}\mathbb{C}^N$. Let Σ be an N–element multiset of complex numbers consisting of l distinct numbers occurring with multiplicities $N_1,\dots,N_l$. There is an isomorphism of vector spaces*

$$\mathrm{KhR}^{\Sigma}(\mathcal{L}(a_1,\dots,a_k)) \cong \bigoplus_{\substack{\sum_{j=1}^{l} b_{i,j}=a_i\\ 0\le b_{i,j}\le N_j}} \bigotimes_{j=1}^{l} \mathrm{KhR}^{\mathfrak{sl}_{N_j}}(\mathcal{L}(b_{1,j},\dots,b_{k,j})) \tag{1-1}$$

which preserves the homological grading.

Remark 1.2 An intended feature of the decomposition formulas and (1-2) is that there are no homological grading shifts on the right-hand side. Lee's, Gornik's and Wu's deformation results, on the other hand, require such grading shifts due to a different normalization arising since they work in the unframed setting.

Example Let K be a 1–colored knot. Then the Σ–deformed $\mathfrak{sl}_N$ homology of K splits into the direct sum of undeformed $\mathfrak{sl}_M$ homologies of K, and there is one $\mathfrak{sl}_M$ summand for every root of multiplicity M in Σ:

$$\mathrm{KhR}^{\Sigma}(K) \cong \bigoplus_{j=1}^{l} \mathrm{KhR}^{\mathfrak{sl}_{N_j}}(K). \tag{1-2}$$

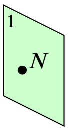
This has been a widely believed conjecture in the link homology community — see eg Gukov and Walcher [15] — however, to our knowledge, no proof has appeared until now.

Example Let K be a knot and write K^0, K^1 and K^2 for its 0–, 1– and 2–colored variants, respectively. Let $\Sigma = \{\lambda_1, \lambda_1, \lambda_2, \lambda_2, \lambda_2\}$ for complex numbers $\lambda_1 \neq \lambda_2$. Then the Σ–deformed $\mathfrak{sl}_5$ homology of K^2 is

$$\begin{aligned}\mathrm{KhR}^{\Sigma}(K^2) &\cong \big(\mathrm{KhR}^{\mathfrak{sl}_2}(K^2) \otimes \mathrm{KhR}^{\mathfrak{sl}_3}(K^0)\big) \oplus \big(\mathrm{KhR}^{\mathfrak{sl}_2}(K^1) \otimes \mathrm{KhR}^{\mathfrak{sl}_3}(K^1)\big) \\ &\qquad\qquad \oplus \big(\mathrm{KhR}^{\mathfrak{sl}_2}(K^0) \otimes \mathrm{KhR}^{\mathfrak{sl}_3}(K^2)\big) \\ &\cong \mathbb{C} \oplus \big(\mathrm{KhR}^{\mathfrak{sl}_2}(K^1) \otimes \mathrm{KhR}^{\mathfrak{sl}_3}(K^1)\big) \oplus \mathrm{KhR}^{\mathfrak{sl}_3}(K^2)\end{aligned}$$

up to shifts in homological degree on the first direct summand.

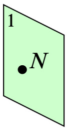
Our main tool is the $\mathfrak{sl}_N$ foam 2–category $N\mathbf{Foam}$ constructed in Queffelec and Rose [36], as well as its relation to the Khovanov–Lauda diagrammatic categorification of quantum $\mathfrak{sl}_m$ [19]; see also work of Rouquier [39]. The former can be viewed as the universal framework for the definition of categorified $\mathfrak{sl}_N$ Reshetikhin–Turaev invariants of tangles τ colored by the fundamental representations of $\mathfrak{sl}_N$. More specifically, given any colored tangle τ, there exists an invariant $[\![\tau]\!]$ taking values in the homotopy category of chain complexes over Hom–categories in $N\mathbf{Foam}$, which consist of trivalent graphs called webs and decorated, singular cobordisms between them called foams. Passing to a quotient 2–category $N\mathbf{Foam}^{\bullet}$ obtained by introducing an additional foam relation on decorated 1–labeled foam facets

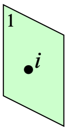
(1-3) 

$= 0$

it is shown in [36] that the resulting bigraded link invariant $\mathrm{KhR}^{\mathfrak{sl}_N}(-)$ essentially agrees with Wu's and Yonezawa's colored generalization of $\mathfrak{sl}_N$ Khovanov–Rozansky link homology. Equation (1-3) corresponds to the fact that X^N is the derivative of the polynomial used to give the potentials for the matrix factorizations in their construction.

In this paper, we analogously define the deformed colored $\mathfrak{sl}_N$ link invariants $\mathrm{KhR}^{\Sigma}(-)$ for an N–element multiset Σ of complex deformation parameters by working in a deformed foam 2–category $N\mathbf{Foam}^{\Sigma}$. It is defined as the quotient 2–category of $N\mathbf{Foam}$ by the additional relation

(1-4) 

$= \displaystyle\sum_{i=0}^{N-1} (-1)^{N-i-1} e_{N-i}(\Sigma)$  where $e_i(\Sigma)$ denotes the i^{th} elementary symmetric polynomial in N variables, evaluated at the multiset Σ. This is motivated by the relation between $N\mathbf{Foam}$ and categorified quantum groups and by Wu's construction of deformed $\mathfrak{sl}_N$ link homology, which utilizes matrix factorizations whose potential is built from a polynomial with derivative $P(X) = \sum_{i=0}^{N}(-1)^{N-i}e_{N-i}(\Sigma)X^i$ with root multiset Σ.

We prove Theorem 1.1 for the invariants constructed via the Σ–deformed $\mathfrak{sl}_N$ foam 2–categories $N\mathbf{Foam}^{\Sigma}$ and undeformed $\mathfrak{sl}_{N_j}$ foam 2–categories $N_j\mathbf{Foam}^{\bullet}$. To this end, we adapt Bar-Natan and Morrison's Karoubi envelope technology [2], originally used to give a "local" proof of Lee's deformation result, to the setting of foams; see Section 2.5. The relation to Wu's deformed Khovanov–Rozansky link homology is then provided by the following generalization of [36, Theorem 4.11]:

Theorem 1.3 *The invariant* $\mathrm{KhR}^{\Sigma}(\mathcal{L})$ *constructed from* $N\mathbf{Foam}^{\Sigma}$ *is (up to grading shifts) isomorphic to Wu's colored, deformed Khovanov–Rozansky homology of the mirror link* $\mathcal{L}'$ *with respect to deformation parameters* Σ.

In [36], the identification of the link invariants defined via foams and matrix factorizations is proven using results of Mackaay and Yonezawa [35], which imply the existence of a 2–representation of $N\mathbf{Foam}$ on $\mathfrak{sl}_N$ matrix factorizations. Rather than adapt their results to the deformed case, we instead give a new, streamlined proof utilizing the theory of stabilization of matrix factorizations to give a 2–representation of $N\mathbf{Foam}^{\Sigma}$ on a 2–category of deformed matrix factorizations. We believe this result might be of independent interest; see Section 4.4.

1.2 Outlook

There are several possible applications of the results in this paper.

The first concerns the definition and study of concordance invariants in the spirit of Rasmussen's s–invariant [37]. Lobb [28; 29] has used Gornik's generic deformation of $\mathfrak{sl}_N$ link homology to define concordance invariants that are analogous to Rasmussen's invariant. Lewark [27] has recently proved independence results for these concordance invariants. It would be interesting to see whether deformations of colored $\mathfrak{sl}_N$ link homologies also give rise to concordance invariants and whether foam technology can be used to prove (in)dependence properties between them.

The next application concerns relations between type A link homology theories of different rank and color. Both experimental computations and physical reasons suggest that the type A link homology package carries a very rigid structure, which is only partially visible on the decategorified level of Reshetikhin–Turaev $\mathfrak{sl}_N$ invariants; see

Dunfield, Gukov and Rasmussen [10], Gukov and Walcher [15], Gukov and Stošić [14] and Gorsky, Gukov and Stošić [13]. One feature of this structure is the stabilization of $\mathfrak{sl}_N$ link homologies as $N \to \infty$ to a triply graded link homology theory that categorifies the HOMFLY-PT polynomial. The flip-side of this feature provides specialization spectral sequences from the triply graded homology to $\mathfrak{sl}_N$ homology for every N. Both of these features have been proven for the 1–colored case by Rasmussen [38].

Many other aspects of the conjectured structure have not been rigorously proven yet. One, however, that seems to be in reach is the existence of spectral sequences, or "differentials", between $\mathfrak{sl}_N$ and $\mathfrak{sl}_M$ link homologies for $N > M$. In analogy to the Lee–Rasmussen spectral sequence that links Khovanov homology to Lee's deformation, Wu [42] has defined spectral sequences connecting the ordinary $\mathfrak{sl}_N$ link homology to its deformations. Together with Theorems 1.1 and 1.3, which identify Wu's deformed invariants in terms of undeformed invariants, it should be possible to construct the desired spectral sequences.

Wu [42] has further proved that the deformed $\mathfrak{sl}_N$ link homologies inherit a quantum filtration from the bigraded undeformed invariant. We have ignored this filtration in this paper, but tracking it through the computation of the deformed invariants should significantly improve our understanding of the Rasmussen-type concordance invariants and the deformation spectral sequences.

We also note that there are bigraded equivariant versions of $\mathfrak{sl}_N$ link homology, in which the deformation parameters are not specialized to complex numbers but kept as graded variables. The $\mathfrak{sl}_2$ and $\mathfrak{sl}_3$ equivariant theories have been studied by Khovanov [16] and Mackaay and Vaz [34]. Krasner [23] has introduced a version for 1–colored $\mathfrak{sl}_N$ link homology for general N, which has been subsequently generalized by Wu [42] to arbitrary colorings by fundamental representations. It is an interesting question whether these equivariant theories also admit a definition via foam technology. This in turn would help to understand the quantum filtration on the deformed invariants.

Daniel Tubbenhauer has informed us that the deformations studied in this paper could be useful for writing down explicit isomorphisms between the centers of $\mathfrak{sl}_N$ web algebras and cohomology rings of certain generalizations of Springer fibers, whose existence is guaranteed by Mackaay [32, Corollary 7.10]. This would generalize work of Mackaay, Pan and Tubbenhauer [33] on the case of $\mathfrak{sl}_3$.

Structure of this paper We begin by introducing the necessary technology and graphical calculi in Section 2. In particular, we discuss foams, categorified quantum groups, and the Karoubi envelope technology of Bar-Natan and Morrison. In Section 3 we study deformations of the higher representation-theoretic structures that control deformed link invariants and prove a version of Theorem 1.1 for the unknot. Armed with this tool, we prove splitting relations in the deformed foam 2–category $N\mathbf{Foam}^{\Sigma}$ and introduce a suitable idempotent completion $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ in Section 4. This section also establishes a 2–representation of the deformed foam 2–category on matrix factorizations, which is necessary for the proof of Theorem 1.3. Finally, Section 5 contains the definition of the deformed link invariants $\mathrm{KhR}^{\Sigma}(\mathcal{L})$ and the proofs of Theorems 1.1 and 1.3.

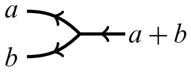
Acknowledgements The authors would like to thank Hanno Becker, Sabin Cautis, Eugene Gorsky, Aaron Lauda, Daniel Murfet, Jake Rasmussen and Ben Webster for helpful conversations and email exchanges during the course of this work. Special thanks go to Marko Stošić and Daniel Tubbenhauer for numerous useful comments on a draft of this paper. Rose would like to especially thank Sabin Cautis and Aaron Lauda, as preliminary work on this topic began in collaboration with them. Wedrich would like to especially thank Jake Rasmussen for his support and guidance.[1]

2 Technology review

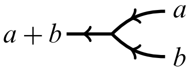
In this section, we recall the relevant machinery needed to prove Theorems 1.1 and 1.3. Explicitly, we discuss $\mathfrak{sl}_N$ foams, categorified quantum groups and their deformations, as well as the Karoubi envelope technology used in [2].

2.1 Foams

Recall from [36] that a natural setting for a combinatorial formulation for Khovanov–Rozansky's $\mathfrak{sl}_N$ link homology is the 2–category $N\mathbf{Foam}$. In this 2–category, objects are given by sequences $\boldsymbol{a} = (a_1, \dots, a_m)$ for $m > 0$ with $a_i \in \{1, \dots, N\}$, 1–morphisms are formal direct sums of enhanced $\mathfrak{sl}_N$ webs — leftward-oriented, labeled[2] trivalent graphs generated by



and

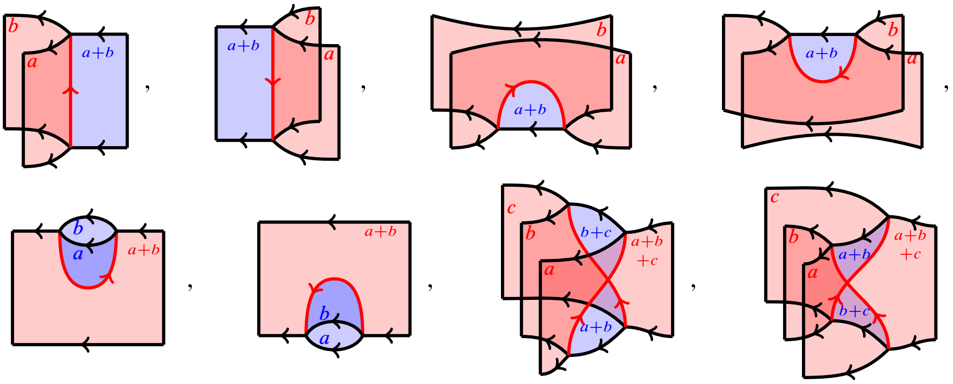


which we view as mapping from the sequence determined by the labeled points on the right boundary to the one determined by the left. The 2–morphisms are matrices of

[1] Wedrich's PhD studies at the Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, have been supported by the ERC grant ERC-2007-StG-205349 held by Ivan Smith and an EPSRC department doctoral training grant. The proof of Theorem 1.1 in this paper is part of Wedrich's thesis.

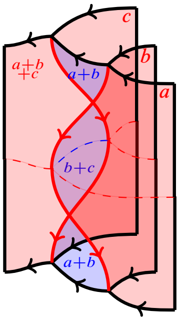
[2] These labels correspond to the "colorings" of tangle components by fundamental representations of $\mathfrak{sl}_N$. We reserve the word "color" for certain idempotent decorations on foams and webs; see below.

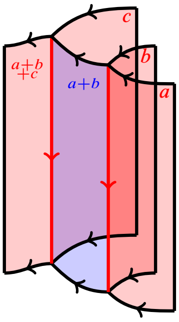
enhanced $\mathfrak{sl}_N$ foams, singular cobordisms between such webs generated by

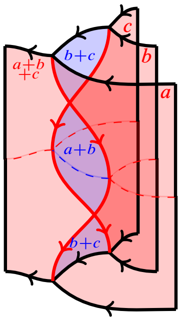


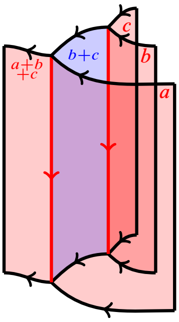
modulo isotopy and local relations.[3] By convention, we view foams as mapping from the web determined by the bottom boundary to that on the top. The facets of these foams again carry labelings by elements in $\{1,\ldots,N\}$, and a k–labeled facet may also be decorated by elements from the ring of symmetric functions in k variables. Note that in [36] the authors utilize the fact that this 2–category admits a grading; however, as we will eventually pass to quotients $N\mathbf{Foam}^{\Sigma}$ where this grading is broken, we won't concern ourselves with these issues.

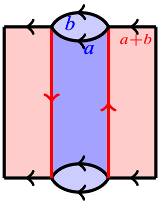
Rather than recall the complete list of local relations, we refer the reader to [36] for full details, and list only a few that will play a substantial role in this paper:

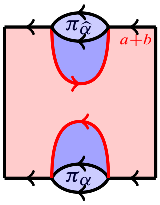
(2-1)


$=$


,


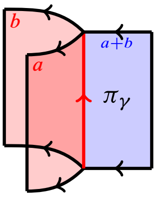
$=$


(2-2)


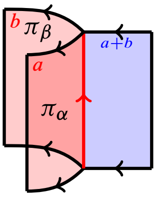
$= \sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}$


[3] The colors red and blue in the foam graphics here and in [36] have no special significance. Later we will use specific colorings of foam facets to indicate decorations by idempotents; see Convention 4.3.

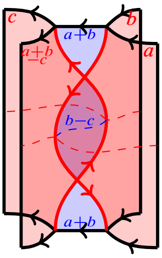
(2-3)



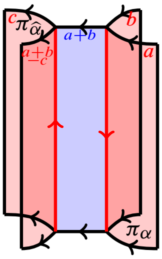
$$= \sum_{\alpha,\beta} c^{\gamma}_{\alpha,\beta}$$



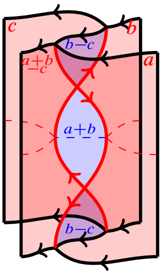
(2-4)



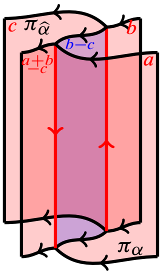
$$= \sum_{\alpha\in P(a,c)} (-1)^{|\hat{\alpha}|}$$



(2-5)



$$= \sum_{\alpha\in P(a,c)} (-1)^{|\hat{\alpha}|}$$



Here $P(a,b)$ denotes the set of partitions of length $\leq a$ with each part $\leq b$, π_α denotes the Schur function corresponding to the partition α and the $c^{\gamma}_{\alpha,\beta}$ are the corresponding Littlewood–Richardson coefficients.

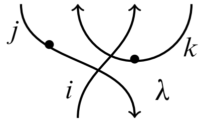
A tangle diagram whose components are labeled by elements in $\{1,\dots,N\}$ determines a complex in N**Foam**, which is, up to homotopy equivalence, an invariant of the underlying framed tangle. In the case that the tangle is a link, passing to the quotient N**Foam**$^\bullet$ and applying a representable functor yields a complex of vector spaces whose homology is isomorphic (up to shifts and grading conventions) to the $\mathfrak{sl}_N$ link homology defined by Khovanov and Rozansky and generalized to the colored case by Wu and Yonezawa.

2.2 Higher representation theory

The construction of N**Foam** was motivated by a desired relation to higher representation theory. The categorified quantum group $\mathcal{U}_Q(\mathfrak{sl}_m)$ is the 2–category whose objects are given by $\mathfrak{sl}_m$ weights λ, and whose 1–morphisms are formal direct sums of (shifts $\{k\}$ of) compositions of

$$\mathbf{1}_\lambda, \quad \mathbf{1}_{\lambda+\alpha_i}\mathcal{E}_i = \mathbf{1}_{\lambda+\alpha_i}\mathcal{E}_i\mathbf{1}_\lambda = \mathcal{E}_i\mathbf{1}_\lambda \quad \text{and} \quad \mathbf{1}_{\lambda-\alpha_i}\mathcal{F}_i = \mathbf{1}_{\lambda-\alpha_i}\mathcal{F}_i\mathbf{1}_\lambda = \mathcal{F}_i\mathbf{1}_\lambda$$

for $i \in \{1, \dots, m-1\}$ and where the α_i are the simple $\mathfrak{sl}_m$ roots. The 2–morphisms are given by matrices of linear combinations of (degree zero) string diagrams — dotted, immersed oriented curves colored by elements $i \in \{1, \dots, m-1\}$ with top and bottom boundary, eg



modulo local relations. The domain 1–morphism of such a diagram is given (up to grading shifts) by considering the orientations and labelings of the strands incident upon the bottom boundary, reading an upward strand as $\mathcal{E}$ and a downward strand as $\mathcal{F}$, and similarly for the codomain by considering the top boundary. For example, the domain and codomain of the above string diagram are (up to shifts) $\mathcal{E}_i \mathcal{F}_j \mathbf{1}_\lambda$ and $\mathcal{F}_j \mathcal{E}_k \mathcal{E}_i \mathcal{F}_k \mathbf{1}_\lambda$.

We refer the reader to the work of Lauda [24] and Khovanov and Lauda [18; 19; 20] (see also independent work of Rouquier [39]) for a detailed discussion on categorified quantum $\mathfrak{sl}_m$. The main result of [19] is that the 2–category $\dot{\mathcal{U}}_Q(\mathfrak{sl}_m)$, obtained by passing to the Karoubi envelope in each Hom–category of $\mathcal{U}_Q(\mathfrak{sl}_m)$, categorifies quantum $\mathfrak{sl}_m$. Explicitly, they show that the Lusztig idempotent form $\dot{\mathbf{U}}_q(\mathfrak{sl}_m)$ of the quantum group is isomorphic to the category obtained by taking the Grothendieck group K_0 in each Hom–category of $\dot{\mathcal{U}}_Q(\mathfrak{sl}_m)$. We will assume some familiarity with categorified quantum groups for the duration, and utilize the conventions and notation from [36].

The 2–category N**Foam** is constructed to give a 2–representation of $\mathcal{U}_Q(\mathfrak{sl}_m)$ via categorical skew Howe duality. Recall that work of Cautis, Kamnitzer and Licata [7] and Cautis, Kamnitzer and Morrison [8] shows that the commuting (skew Howe dual) actions of quantum $\mathfrak{gl}_m$ and $\mathfrak{sl}_N$ on the vector spaces $\bigwedge_q^k(\mathbb{C}_q^m \otimes \mathbb{C}_q^N)$ induce a functor

$$\varphi_m\colon U_q(\mathfrak{gl}_m) \to \mathrm{Rep}(U_q(\mathfrak{sl}_N))$$

which sends a $\mathfrak{gl}_m$ weight $\boldsymbol{a} = (a_1, \dots, a_m)$ to the tensor product of fundamental quantum $\mathfrak{sl}_N$ representations $\bigwedge_q^{a_1}\mathbb{C}_q^N \otimes \dots \otimes \bigwedge_q^{a_m}\mathbb{C}_q^N$. In fact, Cautis, Kamnitzer and Morrison use this to give a completely combinatorial description for the full subcategory of quantum $\mathfrak{sl}_N$ representations generated by the fundamental representations. In their description, objects are given as in N**Foam** and morphisms are given by linear combinations of $\mathfrak{sl}_N$ webs, modulo planar isotopy and relations.

By design, the 2–category N**Foam** gives a categorification of this result, ie it admits a 2–functor $\Phi_m\colon \mathcal{U}_Q(\mathfrak{gl}_m) \to N$**Foam** so that the diagram

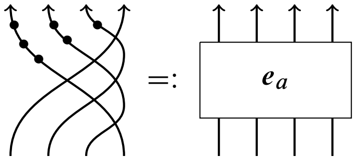
$$\begin{tikzcd}
\mathcal{U}_Q(\mathfrak{gl}_m) \arrow[r, "\Phi_m"] \arrow[d, "\mathrm{K}_0"] & N\mathbf{Foam} \arrow[d, "\mathrm{K}_0"] \\
\dot{\mathbf{U}}_q(\mathfrak{gl}_m) \arrow[r, "\varphi_m"] & \mathrm{Rep}(U_q(\mathfrak{sl}_N))
\end{tikzcd}$$

commutes, where $\mathcal{U}_Q(\mathfrak{gl}_m)$ is the direct sum of an infinite number of copies of $\mathcal{U}_Q(\mathfrak{sl}_m)$ and admits a similar description in which $\mathfrak{sl}_m$ weights are replaced by $\mathfrak{gl}_m$ weights.

2.3 Thick calculus

The 2–functor $\mathcal{U}_Q(\mathfrak{gl}_m) \to N\mathbf{Foam}$ actually extends to a certain full 2–subcategory $\check{\mathcal{U}}_Q(\mathfrak{gl}_m) \subset \dot{\mathcal{U}}_Q(\mathfrak{gl}_m)$. In the case $m = 2$, $\check{\mathcal{U}}_Q(\mathfrak{gl}_2) = \dot{\mathcal{U}}_Q(\mathfrak{gl}_2)$, and this category is described[4] graphically by Khovanov, Lauda, Mackaay and Stošić in [21]. Recall that the objects in the Karoubi envelope of a category $\boldsymbol{C}$ are given by pairs (c,e) where $c \in \mathrm{Ob}(\boldsymbol{C})$ and $c \xrightarrow{e} c$ is an idempotent morphism. In the case of $\check{\mathcal{U}}_Q(\mathfrak{gl}_2)$, consider the idempotent morphism $\mathcal{E}^a\mathbf{1}_\lambda \xrightarrow{\boldsymbol{e}_a} \mathcal{E}^a\mathbf{1}_\lambda$ where $\boldsymbol{e}_a$ is given by decorating any string diagram giving a reduced expression for the longest word in the symmetric group on a elements with a specific pattern of dots, starting with $a-1$ dots on the top left-most strand, and placing one fewer dot on each strand as we head to the right.[5] The following depicts the case $a = 4$:

(2-6)

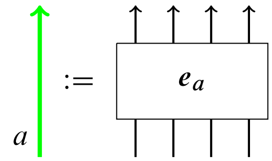


where we use the box notation from [21] for the 2–morphism (here, we do not depict the strand labels, as there is only one possible in the $\mathfrak{gl}_2$ case). Khovanov, Lauda, Mackaay and Stošić show that the 1–morphisms $\mathcal{E}^{(a)}\mathbf{1}_\lambda := \big(\mathcal{E}^a\mathbf{1}_\lambda\{\frac{1}{2}a(a-1)\}, \boldsymbol{e}_a\big)$ and their biadjoints $\mathbf{1}_\lambda\mathcal{F}^{(a)}$ generate $\dot{\mathcal{U}}_Q(\mathfrak{gl}_2)$, and also introduce a "thick calculus" to describe this 2–category. In the Karoubi envelope of a category $\boldsymbol{C}$, a morphism between two objects (c,e) and (c',e') is given by a 1–morphism $f\colon c \to c'$ in $\boldsymbol{C}$ such that $e'f = f = fe$. The map $\boldsymbol{e}_a$, which gives the identity 2–morphism on $\mathcal{E}^{(a)}\mathbf{1}_\lambda$ in $\dot{\mathcal{U}}_Q(\mathfrak{gl}_2)$, is depicted by a thick, colored upward strand

[4] Technically, they describe $\dot{\mathcal{U}}_Q(\mathfrak{sl}_2)$, but the only difference in passing to $\dot{\mathcal{U}}_Q(\mathfrak{gl}_2)$ is that we use $\mathfrak{gl}_2$ weights.

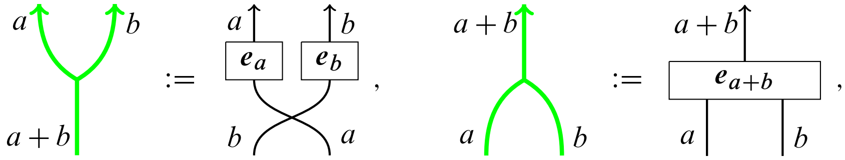
[5] We use the boldface notation $\boldsymbol{e}_a$ for the nil-Hecke idempotents to distinguish them clearly from elementary symmetric polynomials e_i.

(2-7)

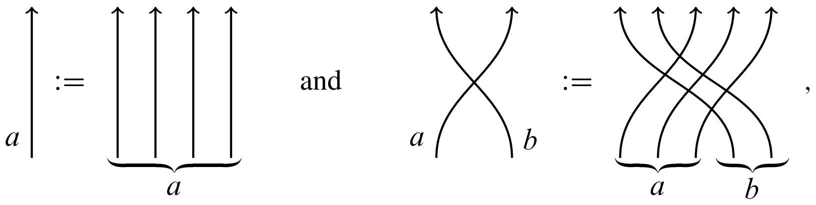


and the remainder of the 2–morphisms in $\dot{\mathcal{U}}_Q(\mathfrak{gl}_2)$ are generated by splitter and merger maps

(2-8)

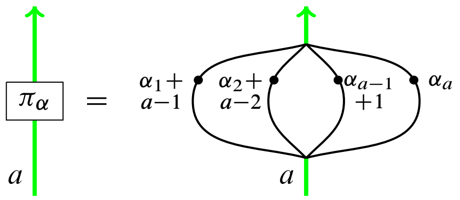


where



which are maps $\mathcal{E}^{(a+b)}\mathbf{1}_\lambda \to \mathcal{E}^{(a)}\mathcal{E}^{(b)}\mathbf{1}_\lambda\{-ab\}$ and $\mathcal{E}^{(a)}\mathcal{E}^{(b)}\mathbf{1}_\lambda \to \mathcal{E}^{(a+b)}\mathbf{1}_\lambda\{-ab\}$. Thick strands also may carry decorations by elements of the ring of symmetric functions in a variables (depicted by placing a box containing the function on such a strand). Schur functions π_α satisfy the relation

(2-9)

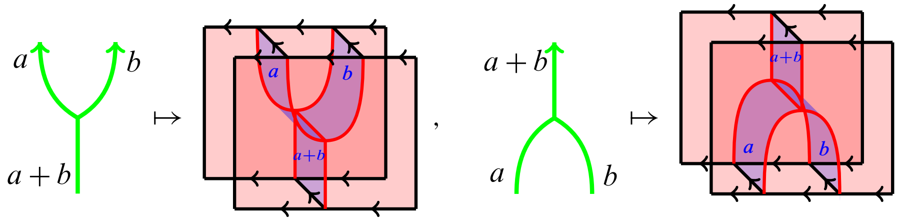


in which the morphisms which split and merge thickness-a strands into thin (thickness-1) strands are given by any of the possible compositions of the above mergers and splitters; the relations for $\dot{\mathcal{U}}_Q(\mathfrak{gl}_2)$ given in [21] guarantee that they are the same.

There is not currently a completely diagrammatic description for $\dot{\mathcal{U}}_Q(\mathfrak{gl}_m)$ for $m \geq 3$, hence we instead work with $\check{\mathcal{U}}_Q(\mathfrak{gl}_m)$, the full 2–subcategory generated by $\mathcal{E}_i^{(a)}\mathbf{1}_\lambda := (\mathcal{E}_i^a\mathbf{1}_\lambda\{\frac{1}{2}a(a-1)\}, e_a)$ and their biadjoints, where here e_a is as above, but with all strands i–labeled. We refer the reader to [36, Section 3.2] for details about the 2–functor $\Phi_m \colon \check{\mathcal{U}}_Q(\mathfrak{gl}_m) \to N\mathbf{Foam}$, but note here that it acts on splitter/merger morphisms in

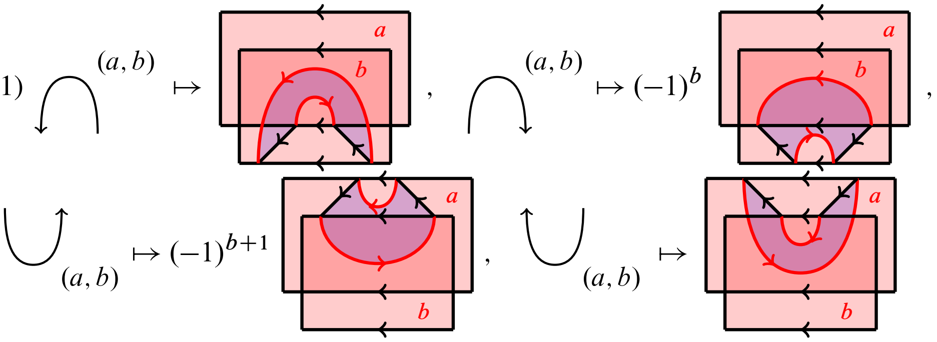
$\check{\mathcal{U}}_Q(\mathfrak{gl}_2)$ via

(2-10)



and on (thin) cap/cup morphisms by

(2-11)



since these will be explicitly used later in our description of the link invariant.

The 2–representation $\Phi_m\colon \check{\mathcal{U}}_Q(\mathfrak{gl}_m) \to N\mathbf{Foam}$ necessarily maps $\mathfrak{gl}_m$ weights whose entries don't lie in $\{0,\dots,N\}$ to zero, hence factors through the quotient $\check{\mathcal{U}}_Q^{0\le N}(\mathfrak{gl}_m)$, where we kill (the identity 2–morphism on the identity 1–morphism of) these weights. As we will see in Section 3.1, it is exactly the procedure of taking this quotient which gives rise to deformation parameters controlling the deformed link invariants.

2.4 Quantum Weyl group action and Rickard complexes

A crucial observation of Cautis, Kamnitzer and Licata [7] is that the braiding on the category of quantum $\mathfrak{sl}_N$ representations (which gives rise to $\mathfrak{sl}_N$ link polynomials) can be recovered from the functor $\phi_m\colon \dot{\mathbf{U}}_q(\mathfrak{gl}_m) \to \mathrm{Rep}(U_q(\mathfrak{sl}_N))$. Indeed, Lusztig's "quantum Weyl group" elements

$$T_i 1_{\boldsymbol{a}} = \begin{cases} \displaystyle\sum_{\substack{j_1,j_2\ge 0\\ j_1-j_2=a_i-a_{i+1}}} (-q)^{j_2} F_i^{(j_1)} E_i^{(j_2)} 1_{\boldsymbol{a}} & \text{if } a_i \ge a_{i+1},\\ \displaystyle\sum_{\substack{j_1,j_2\ge 0\\ j_1-j_2=a_i-a_{i+1}}} (-q)^{j_1} E_i^{(j_2)} F_i^{(j_1)} 1_{\boldsymbol{a}} & \text{if } a_i \le a_{i+1}, \end{cases}$$

generate a braid group action on any finite-dimensional representation of quantum $\mathfrak{sl}_m$; see [30, Section 5.1.1; 8]. Under φ_m, these elements map to the braiding between fundamental $\mathfrak{sl}_N$ representations; explicitly, the element $T1_{(a,b)}$ gives the braiding $\bigwedge_q^a\mathbb{C}_q^N \otimes \bigwedge_q^b\mathbb{C}_q^N \to \bigwedge_q^b\mathbb{C}_q^N \otimes \bigwedge_q^a\mathbb{C}_q^N$.

The Rickard complexes, introduced in the $q=1$ case by Chuang and Rouquier [9], categorify these elements, and generate a categorical braid group action on any (integrable) 2–representation of $\check{\mathcal{U}}_Q(\mathfrak{gl}_m)$. These complexes $\mathcal{T}_i\mathbf{1}_{\boldsymbol{a}}$ take the form

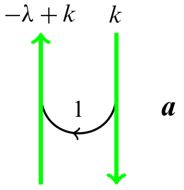
$$\underline{\mathcal{F}_i^{(a_i-a_{i+1})}\mathbf{1}_{\boldsymbol{a}}} \xrightarrow{d_1} \mathcal{F}_i^{(a_i-a_{i+1}+1)}\mathcal{E}_i\mathbf{1}_{\boldsymbol{a}}\{1\} \xrightarrow{d_2} \cdots \xrightarrow{d_s} \mathcal{F}_i^{(a_i-a_{i+1}+s)}\mathcal{E}_i^{(s)}\mathbf{1}_{\boldsymbol{a}}\{s\}\cdots \tag{2-12}$$

when $a_i \geq a_{i+1}$ and

$$\underline{\mathcal{E}_i^{(a_{i+1}-a_i)}\mathbf{1}_{\boldsymbol{a}}} \xrightarrow{d_1} \mathcal{E}_i^{(a_{i+1}-a_i+1)}\mathcal{F}_i\mathbf{1}_{\boldsymbol{a}}\{1\} \xrightarrow{d_2} \cdots \xrightarrow{d_s} \mathcal{E}_i^{(a_{i+1}-a_i+s)}\mathcal{F}_i^{(s)}\mathbf{1}_{\boldsymbol{a}}\{s\}\cdots \tag{2-13}$$

when $a_i \leq a_{i+1}$. Here and throughout, we've underlined in blue the term in homological degree zero. The differential d_k that appears in the second complex is conveniently expressed in thick calculus as

$$d_k =$$



where all strands are colored by the index $i \in I$ and $\lambda = a_i - a_{i+1}$. The differential in the first complex is defined similarly, and in both cases the equality $d^2=0$ follows directly from thick calculus relations.

Recall that the images of the Rickard complexes under any integrable 2–representation are invertible, up to homotopy, with inverses $\mathbf{1}_{\boldsymbol{a}}\mathcal{T}_i^{-1}$ given by the images of the complexes

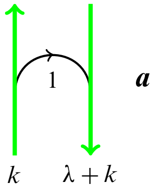
$$\cdots\mathbf{1}_{\boldsymbol{a}}\mathcal{F}_i^{(s)}\mathcal{E}_i^{(a_i-a_{i+1}+s)}\{-s\} \xrightarrow{d_s^*} \cdots \xrightarrow{d_2^*} \mathbf{1}_{\boldsymbol{a}}\mathcal{F}_i\mathcal{E}_i^{(a_i-a_{i+1}+1)}\{-1\} \xrightarrow{d_1^*} \underline{\mathbf{1}_{\boldsymbol{a}}\mathcal{E}_i^{(a_i-a_{i+1})}} \tag{2-14}$$

when $a_i \geq a_{i+1}$ and

$$\cdots\mathbf{1}_{\boldsymbol{a}}\mathcal{E}_i^{(s)}\mathcal{F}_i^{(a_{i+1}-a_i+s)}\{-s\} \xrightarrow{d_s^*} \cdots \xrightarrow{d_2^*} \mathbf{1}_{\boldsymbol{a}}\mathcal{E}_i\mathcal{F}_i^{(a_{i+1}-a_i+1)}\{-1\} \xrightarrow{d_1^*} \underline{\mathbf{1}_{\boldsymbol{a}}\mathcal{F}_i^{(a_{i+1}-a_i)}} \tag{2-15}$$

when $a_i \leq a_{i+1}$. In both cases the differential is given by a composition of splitters with a thickness-1 cap 2–morphisms, eg for (2-15):

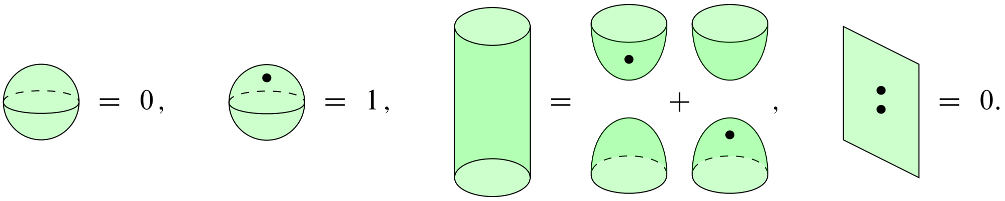
$$d_k^* =$$



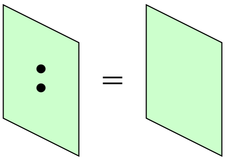
In Section 5, we will use these complexes to define our tangle invariant.

2.5 Karoubi envelope technology

In his famous paper [1], Bar-Natan shows that Khovanov homology can be constructed locally, by working in the homotopy category of chain complexes over a certain $(1+1)$–dimensional cobordism category. Objects of this category are formal direct sums of 1–manifolds embedded in the plane (possibly with boundary) and equipped with a formal $\mathbb{Z}$–grading. Morphisms are matrices of linear combinations of cobordisms between 1–manifolds, decorated with dots, modulo the following local relations:



In [2] Bar-Natan and Morrison explain that Lee's deformed $\mathfrak{sl}_2$ link homology [26] arises from the same kind of construction, after modifying the final "sheet" relation above to



so that the operator given by adding a dot to a cobordism is no longer nilpotent.

To analyze the effects of this deformation, consider the algebra of endomorphisms of a strand, denoting the identity by $\mathbb{1}$, a sheet decorated by a dot by X, and extend linearly so that polynomials in X denote linear combinations of decorated sheets. The undeformed sheet relation can then be expressed as $X^2 = 0$ and the deformed relation is $X^2 - \mathbb{1} = 0$. From this it is clear that in the deformed case the operator of placing a dot on a sheet has eigenvalues 1 and -1 with corresponding eigenvectors $\mathbb{1}_{+1} := \frac{1}{2}(\mathbb{1} + X)$ and $\mathbb{1}_{-1} := \frac{1}{2}(\mathbb{1} - X)$.

The decomposition into eigenspaces for the action of adding a dot splits the deformed cobordism category: every connected component of a cobordism can be written as a sum of the two decorations $\mathbb{1} = \mathbb{1}_{+1} + \mathbb{1}_{-1}$, which are orthogonal (ie $\mathbb{1}_{+1}\mathbb{1}_{-1} = 0$), idempotent (ie $\mathbb{1}_{\pm 1}\mathbb{1}_{\pm 1} = \mathbb{1}_{\pm 1}$), and obviously commute.

Next, Bar-Natan and Morrison enlarge the cobordism category by proceeding to its Karoubi envelope.[6] Practically, this means allowing objects, ie planar 1–manifolds, to

[6] See the explanation in Section 2.3.

be "colored" by $\mathbb{1}_{+1}$ and $\mathbb{1}_{-1}$ as well. Any uncolored 1–manifold is isomorphic to the direct sum of the $\mathbb{1}_{+1}$ and $\mathbb{1}_{-1}$ versions, and colored cobordisms between colored 1–manifolds are only nonzero if the corresponding idempotent decorations agree.

Using this splitting of the deformed cobordism category, Bar-Natan and Morrison compute a decomposition for the chain complexes arising in the definition of the deformed link invariant, whose objects are planar 1–manifolds that arise as resolutions of the link diagram. The second result is that each such coloring contributes only one generator to the link homology. This reproduces Lee's result that the deformed $\mathfrak{sl}_2$ homology of a l–component link is 2^l–dimensional. Alternatively, we could say that it is a direct sum of tensor products of $\mathfrak{sl}_1$ homologies, where $\mathfrak{sl}_1$ homology assigns the 1–dimensional vector space $\mathbb{C}$ to any link. More precisely, we have one summand for each coloring of components of the link by $\mathbb{1}_{+1}$ or $\mathbb{1}_{-1}$, and the tensorands are the $\mathfrak{sl}_1$ homologies of the $\mathbb{1}_{+1}$- and $\mathbb{1}_{-1}$–colored sublinks, respectively.

Gornik's generalization [12] of the generic deformation result to $\mathfrak{sl}_N$ can be understood along very similar lines. Again, there is a Frobenius algebra $\mathbb{C}[X]/\langle X^N\rangle$ of local decorations which is being deformed to $\mathbb{C}[X]/\langle X^N-\beta^N\rangle \cong \mathbb{C}\oplus\cdots\oplus\mathbb{C}$, with one summand for each of the N roots of the polynomial $X^N-\beta^N$. The idempotents that project onto the N summands then split the category underlying the chain complexes in the construction of the link homology. The resulting invariant for a knot is a direct sum of N copies of its $\mathfrak{sl}_1$ homology. Similarly, for l–component links one gets a N^l–dimensional vector space which can be understood as a direct sum over possible root-colorings of components of 1–dimensional tensor products of $\mathfrak{sl}_1$ homologies of sublinks, one for each different root.

In order to prove our decomposition result Theorem 1.1, we start by computing the algebra of decorations on foam facets in the deformed foam 2–category $N\mathbf{Foam}^{\Sigma}$. In fact, the algebra of decorations on a k–labeled facet is isomorphic to the deformed link homology of the $\bigwedge^k\mathbb{C}^N$–colored unknot. We compute it, and hence prove Theorem 1.1 in the special case of the unknot, in Section 3. In particular, the algebra of decorations on a k–labeled foam facet decomposes into a direct sum of local pieces indexed by k–element multisubsets of the set of roots Σ. This gives idempotent foam decorations along which the link invariant splits into a direct sum, which is proved in Section 5.1. Similarly as for the generic deformation in the 1–colored case, the only nonzero contributions to the deformed link invariant come from idempotent colorings that are consistent along link components. This is shown in Lemma 5.10. However, in the case of general deformations of colored invariants there are two new features that have not been rigorously addressed in the literature. One appears because we allow higher colors, the other because we allow nongeneric deformations.

Higher color Foam facets are not colored by roots, ie elements of Σ, anymore, but by multisubsets of Σ of size corresponding to the label of the facet. Such a multisubset can contain several different roots and in this case we need a new way to split this facet into parts colored by single roots. This is where we use the full power of the foam technology; see Section 4 for preparatory work and Section 5.2 for the actual tensor product decomposition of the link invariant.

Nongeneric deformation A root λ can occur in Σ with a multiplicity $N_\lambda \geq 1$. The λ–colored part of a direct summand of the deformed link invariants is essentially the $\mathfrak{sl}_{N_\lambda}$ homology of the λ–colored sublink, ie in particular it is usually not trivially 1–dimensional. To see this we need to check that after all splitting procedures, the λ–colored foams behave like $\mathfrak{sl}_{N_\lambda}$–foams. This is done in Section 5.3.

3 Deforming nil-Hecke algebra quotients

The nil-Hecke algebra $\mathcal{NH}_a$ plays a fundamental role in higher representation theory. Indeed, this algebra is given by the algebra of (not necessarily degree zero) 2–endomorphisms of the a–fold composition of $\mathcal{E}_i$ with itself in the positive half of $\mathcal{U}_Q(\mathfrak{sl}_m)$. In this section, we will review the nil-Hecke algebra, and then proceed to study certain deformations of its cyclotomic quotients, which control the deformed Khovanov–Rozansky homologies of colored unknots.

Definition 3.1 The nil-Hecke algebra on a strands, $\mathcal{NH}_a$ admits an algebraic presentation as the graded $\mathbb{C}$–algebra of endomorphisms of the abelian group $\mathbb{C}[X_1,\dots,X_a]$ generated by operators

- ξ_i of degree 2 for $1 \leq i \leq a$ acting by multiplication by X_i,
- ∂_i of degree -2 for $1 \leq i \leq a-1$ acting as divided difference,
  ie for $p(X_1,\dots,X_a) \in \mathbb{C}[X_1,\dots,X_a]$
$$\partial_i(p(X_1,\dots,X_a)) = \frac{p(\dots,X_i,X_{i+1},\dots) - p(\dots,X_{i+1},X_i,\dots)}{X_i - X_{i+1}},$$

which satisfy the complete set of relations

- $\xi_i\xi_j = \xi_j\xi_i$,
- $\xi_i\partial_j = \partial_j\xi_i$ if $i \notin \{j, j+1\}$,
- $\partial_i\partial_i = 0$,
- $\partial_i\partial_{i+1}\partial_i = \partial_{i+1}\partial_i\partial_{i+1}$,
- $\xi_i\partial_i - \partial_i\xi_{i+1} = 1 = \partial_i\xi_i - \xi_{i+1}\partial_i$.

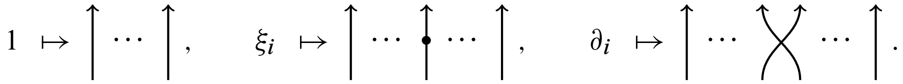
The following result is due to Lauda:

Proposition 3.2 [24, Proposition 3.5] (1) *The center of $\mathcal{NH}_a$ is* $Z(\mathcal{NH}_a) \cong \mathbb{C}[\xi_1, \dots, \xi_a]^{S_a} =: \mathrm{Sym}(\xi_1, \dots, \xi_a)$.

(2) $\mathcal{NH}_a$ *is graded isomorphic to the algebra of $a! \times a!$ matrices over its center:*

$$\mathcal{NH}_a \cong \mathrm{Mat}(a!, Z(\mathcal{NH}_a)).$$

The homomorphism $\mathcal{NH}_a \to \mathrm{END}(\mathcal{E}_i^a \mathbf{1}_\lambda)$ is given by identifying the generator ξ_i with the string diagram consisting of a upward strands with a dot on the i^{th} strand and the generator ∂_i with a crossing between the i^{th} and $(i+1)^{\text{st}}$ strands:



Multiplication is given by composition of 2–morphisms in $\mathcal{U}_Q(\mathfrak{sl}_m)$, ie by stacking diagrams vertically. An arbitrary element of $\mathcal{NH}_a$ can be written as a $\mathbb{C}$–linear combination of such stacked string diagrams.

We will also utilize the "thick calculus" for the nil-Hecke algebra, detailed in [21], which corresponds to the algebra of upward strands in $\check{\mathcal{U}}_Q(\mathfrak{sl}_m)$ having varying thickness. Set

$$D_a := (\partial_1 \partial_2 \cdots \partial_{a-1})(\partial_1 \cdots \partial_{a-2}) \cdots (\partial_1)$$

and let $\Delta_X = \prod_{1 \le i < j \le a}(X_i - X_j)$ be the Vandermonde determinant. The action of D_a on polynomials $p \in \mathbb{C}[X_1, \dots, X_a]$ is given by

$$D_a(p(X_1, \dots, X_a)) = \frac{1}{\Delta_X} \sum_{w \in S_a} \epsilon(w) p(X_{w(1)}, \dots, X_{w(a)}),$$

where $\epsilon(w) \in \{\pm 1\}$ is the sign of the permutation w. In other words, D_a antisymmetrizes a polynomial and then divides by the Vandermonde determinant, resulting in a symmetric polynomial. Divided differences not only act on elements of $\mathbb{C}[X_1, \dots, X_a]$, but also on the subring $\mathbb{C}[\xi_1, \dots, \xi_a]$ of $\mathcal{NH}_a$. In particular, if $f \in \mathbb{C}[\xi_1, \dots, \xi_a]$, we denote by $D_a(f)$ the action of the product of divided differences on f. The following compatibility relation holds:

$$D_a f(\xi_1, \dots, \xi_a) D_a = D_a(f)(\xi_1, \dots, \xi_a) D_a.$$

However, we point out that this is only true in the presence of the D_a on the right.

Define $\delta_a := \xi_1^{a-1} \xi_2^{a-2} \cdots \xi_{a-1}$. It is easy to compute that

$$\Delta_\xi = \prod_{1 \le i < j \le a} (\xi_i - \xi_j) = \sum_{w \in S_a} \epsilon(w) \xi_{w(1)}^{a-1} \xi_{w(2)}^{a-2} \cdots \xi_{w(a-1)}$$

and hence $D_a(\delta_a) = \Delta_\xi / \Delta_\xi = 1$ and $\boldsymbol{e}_a = \delta_a D_a$ is idempotent in $\mathcal{NH}_a$:

$$\boldsymbol{e}_a^2 = \delta_a D_a \delta_a D_a = \delta_a D_a(\delta_a) D_a = \delta_a D_a = \boldsymbol{e}_a.$$

In fact, this is exactly the idempotent $\boldsymbol{e}_a$ defined in the introduction, and depicted graphically (in the case $a = 4$) in (2-6).

One can use this idempotent to explicitly describe the isomorphism between the center $Z(\mathcal{NH}_a) \cong \mathbb{C}[\xi_1, \dots, \xi_a]^{S_a}$ and the direct summand $\boldsymbol{e}_a \mathcal{NH}_a \boldsymbol{e}_a \subset \mathcal{NH}_a$ via

$$Z(\mathcal{NH}_a) \cong Z(\mathcal{NH}_a)\boldsymbol{e}_a = \boldsymbol{e}_a \mathcal{NH}_a \boldsymbol{e}_a, \quad y \mapsto y\boldsymbol{e}_a.$$

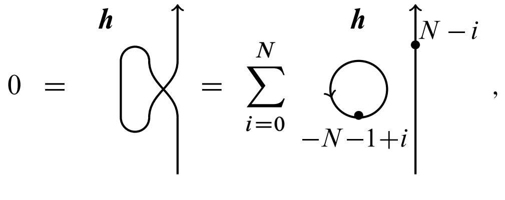
If $\alpha = (\alpha_1, \dots, \alpha_a)$ is a partition of length $\leq a$ and $\pi_\alpha(\xi_1, \dots, \xi_a)$ is the Schur polynomial associated to α, then $D_a(\xi_1^{a-1+\alpha_1} \xi_2^{a-2+\alpha_2} \cdots \xi_a^{\alpha_a}) = \pi_\alpha(\xi_1, \dots, \xi_a)$. Hence, under the above isomorphism we have

$$\begin{aligned}(3\text{-}1)\qquad \pi_\alpha(\xi_1, \dots, \xi_a) \mapsto \pi_\alpha(\xi_1, \dots, \xi_a)\boldsymbol{e}_a &= \delta_a D_a(\xi_1^{a-1+\alpha_1} \xi_2^{a-2+\alpha_2} \cdots \xi_a^{\alpha_a}) D_a \\ &= \delta_a D_a \xi_1^{\alpha_1} \xi_2^{\alpha_2} \cdots \xi_a^{\alpha_a} \delta_a D_a \\ &= \boldsymbol{e}_a \xi_1^{\alpha_1} \xi_2^{\alpha_2} \cdots \xi_a^{\alpha_a} \boldsymbol{e}_a.\end{aligned}$$

Compare this with the thick calculus relation (2-9).

3.1 Quotients of the nil-Hecke algebra

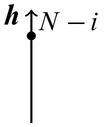
A certain quotient of the nil-Hecke algebra will be relevant to our study of deformed link homology. Recall that the 2–functor $\check{\mathcal{U}}_Q(\mathfrak{gl}_m) \to N\mathbf{Foam}$ factors through the quotient $\check{\mathcal{U}}_Q^{0\leq N}(\mathfrak{gl}_m)$, where we kill the $\mathfrak{gl}_m$ weights whose entries lie outside the set $\{0, \dots, N\}$. Consider $\boldsymbol{h} = (N, \dots, N, 0, \dots, 0)$, which is a highest weight in this quotient, and note that



where the string diagrams are colored by the number of N's in $\boldsymbol{h}$. The first equality holds since the region inside the "left-curl" is zero in $\check{\mathcal{U}}_Q^{0\leq N}(\mathfrak{gl}_m)$. Note that the individual summands on the right-hand side are not (necessarily) zero, since these bubbles are fake, in the sense of [24]. The infinite Grassmannian relation [24] implies that these bubbles generate the endomorphism algebra of the highest weight object $\boldsymbol{h}$, hence we can view the positive degree fake bubbles as (graded) parameters.

Under the 2–functor $\check{\mathcal{U}}_Q^{0\le N}(\mathfrak{gl}_m) \to N\mathbf{Foam}$, this highest weight endomorphism algebra maps to the endomorphism algebra of an N–labeled facet, with the fake bubble of degree i mapping to the decoration by a signed elementary symmetric polynomial $(-1)^i e_i$ by [36, Equation 3.33]. This endomorphism algebra in turn determines the ground ring over which the link homology theory is defined, ie the invariant of a link will be a module over this algebra. In [36], it is shown that by setting the (images of) the bubble deformation parameters to zero yields a link homology theory isomorphic to Khovanov–Rozansky homology. Setting these parameters to other values should thus correspond to a deformed version of Khovanov–Rozansky homology.

We thus arrive at the relation

(3-2)
$$\sum_{i=0}^{N} c_i$$


$$= \; 0,$$

where the $c_i \in \mathbb{C}$ are the specializations of the fake bubbles. This corresponds to a relation on 1–labeled foams facets which meet N–labeled foam facets (see [36, Section 4.1] for a general discussion about N–labeled facets). Since we would like foam relations to be local, this motivates studying this relation for all weights, not just for $\boldsymbol{a} = \boldsymbol{h}$.

To this end, let Σ be a multiset of N complex numbers and

$$P(X) = \prod_{s\in\Sigma}^{N}(X-s) = X^N + \sum_{i=0}^{N-1} c_{N-i}X^i \tag{3-3}$$

be the monic degree N polynomial with root multiset Σ and coefficients $c_i = (-1)^i e_i(\Sigma)$.

Definition 3.3 The Σ–deformed quotient of the nil-Hecke algebra $\mathcal{NH}_a^\Sigma$ is the quotient algebra of $\mathcal{NH}_a$ modulo the ideal generated by $P(\xi_1)$.

In the case where $\Sigma = \{0^N\}$ (ie $P(X) = X^N$), this algebra is known as the level N cyclotomic quotient of $\mathcal{NH}_a$, which we denote by $\mathcal{NH}_a^N$. We aim to now generalize the following result of Lauda:

Proposition 3.4 [25, Proposition 5.3] *There are isomorphisms of graded algebras*

(1) $Z(\mathcal{NH}_a^N) \cong \mathrm{H}^*(\mathrm{Gr}(a,N))$,

(2) $\mathcal{NH}_a^N \cong \mathrm{Mat}(a!, Z(\mathcal{NH}_a^N))$.

(Here $\mathrm{H}^*(\mathrm{Gr}(a,N))$ denotes the cohomology ring with coefficients in $\mathbb{C}$ of the Grassmannian of complex a–planes in $\mathbb{C}^N$.)

To generalize this to arbitrary Σ, we will adapt Lauda's method of proof to our setting.

Definition 3.5 Let $\mathbb{X}=\{\xi_1,\dots,\xi_a\}$ and $\mathbb{Y}=\{y_1,\dots y_b\}$ be two alphabets of variables. We denote the ring of symmetric polynomials in $\mathbb{X}$ by $\mathrm{Sym}(\mathbb{X})$ and the ring of polynomials separately symmetric in $\mathbb{X}$ and $\mathbb{Y}$ by $\mathrm{Sym}(\mathbb{X}\,|\,\mathbb{Y})$. For layout we sometimes abbreviate Sym by S. The *complete symmetric polynomials* $h_i(\mathbb{X})$ in $\mathbb{X}$ can be defined via their generating function:

$$\sum_{i=0}^{\infty} h_i(\mathbb{X})t^i=\prod_{\xi\in\mathbb{X}}(1-t\xi)^{-1}.$$

The *elementary symmetric polynomials* $e_i(\mathbb{X})$ in $\mathbb{X}$ are defined by

$$\sum_{i=0}^{\infty} e_i(\mathbb{X})t^i=\prod_{\xi\in\mathbb{X}}(1+t\xi),$$

and finally we define the *complete symmetric functions in* $\mathbb{X}-\mathbb{Y}$, denoted $h_i(\mathbb{X}-\mathbb{Y})$, by

$$\sum_{i=0}^{\infty} h_i(\mathbb{X}-\mathbb{Y})t^i=\frac{\prod_{y\in\mathbb{Y}}(1-ty)}{\prod_{\xi\in\mathbb{X}}(1-t\xi)}.$$

Note that this gives the explicit formula

$$h_k(\mathbb{X}-\mathbb{Y})=\sum_{i=0}^{k}(-1)^i e_i(\mathbb{Y})h_{k-i}(\mathbb{X}).$$

Definition 3.6 Let $\mathbb{X}=\{\xi_1,\dots,\xi_a\}$ be an alphabet of a variables (of degree 2) and $\mathbb{B}=\{b_1,\dots,b_N\}$ an alphabet of N variables (of degree 2). The following is an explicit description of the $\mathrm{GL}(N)$–equivariant cohomology (with $\mathbb{C}$ coefficients) of the Grassmannian $\mathrm{Gr}(a,N)$ of complex a–planes in $\mathbb{C}^N$:

$$\mathrm{H}^*_{\mathrm{GL}(N)}(\mathrm{Gr}(a,N))\cong\frac{\mathrm{Sym}(\mathbb{X}\,|\,\mathbb{B})}{\langle h_{N-a+1}(\mathbb{X}-\mathbb{B}),\dots,h_N(\mathbb{X}-\mathbb{B})\rangle}.$$

This is a rank $\binom{N}{a}$ graded free module over $\mathrm{Sym}(\mathbb{B})\cong\mathrm{H}^*_{\mathrm{GL}(N)}(*)$; see [42, Section 2.3] and references therein. If we quotient $\mathrm{H}^*_{\mathrm{GL}(N)}(\mathrm{Gr}(a,N))$ by the relations $b_i=0$ we recover the well-known description of the ordinary cohomology ring of the Grassmannian

$$\mathrm{H}^*(\mathrm{Gr}(a,N))\cong\frac{\mathrm{Sym}(\mathbb{X})}{\langle h_{N-a+1}(\mathbb{X}),\dots,h_N(\mathbb{X})\rangle}.$$

We can also quotient $\mathrm{H}^*_{\mathrm{GL}(N)}(\mathrm{Gr}(a,N))$ by sending $\mathbb{B}$ to Σ, an arbitrary multisubset of N complex numbers. The result is the $\mathbb{C}$–algebra

$$H_a^\Sigma := \frac{\mathrm{Sym}(\mathbb{X})}{\langle h_{N-a+1}(\mathbb{X}-\Sigma),\dots,h_N(\mathbb{X}-\Sigma)\rangle},$$

which we call the Σ*–deformed cohomology ring of* $\mathrm{Gr}(a,N)$. It is a flat deformation of $\mathrm{H}^*(\mathrm{Gr}(a,N))$, in particular it has complex dimension $\binom{N}{a}$. We use the following notation for its defining ideal:

$$I_a^\Sigma := \langle h_{N-a+1}(\mathbb{X}-\Sigma),\dots,h_N(\mathbb{X}-\Sigma)\rangle \subset \mathrm{Sym}(\mathbb{X}).$$

Proposition 3.7 *There are isomorphisms of algebras*

(1) $Z(\mathcal{NH}_a^\Sigma) \cong H_a^\Sigma$,

(2) $\mathcal{NH}_a^\Sigma \cong \mathrm{Mat}(a!, H_a^\Sigma)$.

Proof To explain the context we first go through a proof of Proposition 3.2, following the exposition in [25, Section 5].

Let $\mathbb{X} := \{\xi_1,\dots,\xi_a\}$ be an alphabet of a variables and denote by $\mathcal{H}_a$ the abelian subgroup of $\mathbb{C}[\mathbb{X}] := \mathbb{C}[\xi_1,\dots,\xi_a]$ generated by all monomials $\xi_1^{\alpha_1}\cdots\xi_a^{\alpha_a}$ with $0\le\alpha_i\le a-i$. Then $\mathcal{H}_a$ has rank $a!$ and $\mathbb{C}[\mathbb{X}]\cong\mathcal{H}_a\otimes\mathrm{Sym}(\mathbb{X})$ as graded $\mathrm{Sym}(\mathbb{X})$–modules. In particular, the generators of $\mathcal{H}_a$ give a basis for $\mathbb{C}[\mathbb{X}]$ as a free graded $\mathrm{Sym}(\mathbb{X})$–module and $\mathrm{End}_{\mathrm{Sym}(\mathbb{X})}(\mathbb{C}[\mathbb{X}])\cong\mathrm{Mat}(a!,\mathrm{Sym}(\mathbb{X}))$. It is easy to check that the nil-Hecke generators ξ_i and ∂_i act as $\mathrm{Sym}(\mathbb{X})$–module endomorphisms of $\mathbb{C}[\mathbb{X}]$ and hence there is a homomorphism

$$\theta\colon \mathcal{NH}_a \to \mathrm{Mat}(a!,\mathrm{Sym}(\mathbb{X})).$$

Lauda [24] has shown that this is an isomorphism of graded algebras, which proves Proposition 3.2.

Let $\alpha=(\alpha_2,\dots,\alpha_a)$ be a sequence with $0\le\alpha_i\le a-i$ and write $\overline{\xi}^\alpha := \xi_2^{\alpha_2}\cdots\xi_a^{\alpha_a}$, then we can partition the above basis for $\mathcal{H}_a$ into $(a-1)!$ ordered subsets $B_\alpha := \{\xi_1^{a-1}\overline{\xi}^\alpha,\dots,\xi_1\overline{\xi}^\alpha,\overline{\xi}^\alpha\}$ indexed by sequences α as above. The orders on B_α extend to a total order on the basis of $\mathcal{H}_\alpha$ and with respect to this ordered basis the action of ξ_1 under the isomorphism θ is given by a block diagonal matrix of $(a-1)!$ identical blocks (the restriction to the span of the B_α) of the form

$$\theta(\xi_1) = \begin{pmatrix} e_1 & 1 & 0 & \cdots & 0 \\ -e_2 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & 0 & 1 \\ (-1)^{a-1}e_a & 0 & \cdots & \cdots & 0 \end{pmatrix},$$

where we write $e_i := e_i(\mathbb{X})$ for the i^{th} elementary symmetric polynomials in $\mathbb{X}$.

The image of the ideal $\langle P(\xi_1)\rangle$ under the isomorphism θ is determined by the matrix equation $0 = \theta(P(\xi_1)) = P(\theta(\xi_1))$. To explicitly compute the right-hand side of this matrix equation we first describe powers of $\theta(\xi_1)$.

Lemma 3.8 *If we write*

$$\theta(\xi_1)^k = \begin{pmatrix} b^k_{1,1} & \cdots & b^k_{1,a} \\ \vdots & & \vdots \\ b^k_{a,1} & \cdots & b^k_{a,a} \end{pmatrix},$$

then the $b^k_{i,j}$ satisfy (and are completely determined by) the relations

$$b^k_{i,j} = \begin{cases} h_{k+i-j} - \sum_{l=1}^{i-1} h_{i-l} b^k_{l,j} & \text{for } j \le k, \\ \delta_{i+k,j} & \text{for } j > k. \end{cases}$$

(Here we use the shorthand $h_i := h_i(\mathbb{X})$ for the i^{th} complete symmetric polynomial in $\mathbb{X}$. In particular, the first row of $\theta(\xi_1)^k$ has entries $h_k, \dots, h_{k+1-a}$.)

Proof Provided that the $b^k_{i,j}$ satisfy the above set of relations, it is clear that they are completely determined by it. We prove the former by induction on k. For $k = 1$ the second relation is immediate, so we only have to check that the entries in the first column of $\theta(\xi_1)$ satisfy the first relation (with $j = 1$). This holds since

$$h_{1+i-1} - \sum_{l=1}^{i-1} h_{i-l}(-1)^{l-1} e_l = \underbrace{\sum_{l=0}^{i} h_{i-l}(-1)^l e_l}_{=0} - (-1)^i e_i = (-1)^{i-1} e_i.$$

For the induction step, we assume the relations hold for $b^k_{i,j}$ and will deduce the relations for $b^{k+1}_{i,j}$. First note that since $\theta(\xi_1)$ has the identity matrix as $(a-1)\times(a-1)$–minor, we get $b^{k+1}_{i,j+1} = b^k_{i,j}$ for all $0 \le i \le a$ and $0 \le j < a$. It is then immediate that the $b^{k+1}_{i,j}$ for all $0 \le i \le a$ and $j \ge 2$ satisfy the required relations, and we only have to check the relations between the entries $b^{k+1}_{i,1}$ in the first column by induction on i. The case $i = 1$ is given by

$$b^{k+1}_{1,1} = \sum_{l=1}^{a} (-1)^{l-1} e_l b^k_{1,l} = \sum_{l=1}^{a} (-1)^{l-1} e_l h_{k+1-l} = 0 - (-1)^{-1} e_0 h_{k+1} = h_{k+1}$$

and, assuming it holds for all smaller indices, the case for $i+1$ is given by

$$\begin{aligned}
b^{k+1}_{i+1,1} &= \sum_{l=1}^{a}(-1)^{l-1}e_l b^k_{i+1,l} = \sum_{l=1}^{a}(-1)^{l-1}e_l\Bigl(h_{k+i+1-l} - \sum_{r=1}^{i} h_{i+1-r}b^k_{r,l}\Bigr)\\
&= \sum_{l=1}^{a}(-1)^{l-1}e_l(h_{k+i+1-l}) - \sum_{r=1}^{i} h_{i+1-r}\Bigl(\sum_{l=1}^{a}(-1)^{l-1}e_l b^k_{r,l}\Bigr)\\
&= h_{(k+1)+(i+1)-1} - \sum_{r=1}^{i} h_{i+1-r}b^{k+1}_{r,1}. \qquad \square
\end{aligned}$$

Lemma 3.9 *Denoting*

$$\theta(P(\xi_1)) = \begin{pmatrix} c_{1,1} & \cdots & c_{1,a} \\ \vdots & & \vdots \\ c_{a,1} & \cdots & c_{a,a}\end{pmatrix},$$

we have $c_{1,i} = h_{N+1-i}(\mathbb{X}-\Sigma)$, and all other $c_{i,j}$ lie in the ideal I^{Σ}_a generated by the entries of the first row.

Proof Since θ is an algebra isomorphism, we have

$$\theta(P(\xi_1)) = P(\theta(\xi_1)) = \sum_{l=0}^{N}(-1)^l e_l(\Sigma)\theta(\xi_1)^{N-l},$$

and hence the entries are given by

$$c_{i,j} = \sum_{l=0}^{N}(-1)^l e_l(\Sigma)b^{N-l}_{i,j}.$$

We then compute that the entries in the first row are

$$c_{1,j} = \sum_{l=0}^{N}(-1)^l e_l(\Sigma)b^{N-l}_{1,j} = \sum_{l=0}^{N}(-1)^l e_l(\Sigma)h_{N-l+1-j} = h_{N-j+1}(\mathbb{X}-\Sigma).$$

All other entries $c_{i,j}$ are determined by the entries in the first row by a similar recursion to the case of $\theta(\xi_1)^k$; assume $i>1$, then we have

$$\begin{aligned}
c_{i,j} &= \sum_{l=0}^{N}(-1)^l e_l(\Sigma)b^{N-l}_{i,j} = \sum_{l=0}^{N}(-1)^l e_l(\Sigma)\Bigl(h_{N-l+i-j} - \sum_{r=1}^{i-1} h_{i-r}b^{N-l}_{r,j}\Bigr)\\
&= h_{N+i-j}(\mathbb{X}-\Sigma) - \sum_{r=1}^{i-1} h_{i-r}\sum_{l=0}^{N}(-1)^l e_l(\Sigma)b^{N-l}_{r,j}
\end{aligned}$$

$$= h_{N+i-j}(\mathbb{X}-\Sigma) - \sum_{r=1}^{i-1} h_{i-r} c_{r,j}.$$

The following (by induction on s) shows that $h_{N+s}(\mathbb{X}-\Sigma) \in I_a^\Sigma$ for every $s > 0$:

$$\begin{aligned} h_{N+s}(\mathbb{X}-\Sigma) &= \sum_{l=0}^{N+s} (-1)^l e_l(\Sigma) h_{N+s-l} = \sum_{l=0}^{N} (-1)^l e_l(\Sigma) h_{N+s-l} \\ &= \sum_{l=0}^{N} (-1)^l e_l(\Sigma) \Big(- \sum_{r=1}^{N} (-1)^r e_r h_{N+s-l-r} \Big) \\ &= - \sum_{r=1}^{N} (-1)^r e_r h_{N+s-r}(\mathbb{X}-\Sigma). \end{aligned}$$

It then follows (again by induction on i) that $c_{i,j} \in I_a^\Sigma$ for all $i > 1$. □

Since the (two-sided) ideal generated by a matrix A is equal to the ideal of matrices with entries taking values in the ideal generated by the entries of A, Lemma 3.9 shows that taking the quotient of $\mathrm{Mat}(a!, \mathrm{Sym}(\mathbb{X}))$ by the ideal $\theta(\langle P(\xi_1)\rangle)$ is equal to the quotient of $\mathrm{Mat}(a!, \mathrm{Sym}(\mathbb{X}))$ by matrices with entries in the ideal I_a^Σ. This shows that $\mathcal{NH}_a^\Sigma \cong \mathrm{Mat}(a!, H_a^\Sigma)$. Moreover $Z(\mathcal{NH}_a^\Sigma)$ is isomorphic via θ to $Z(\mathrm{Mat}(a!, H_a^\Sigma)) = H_a^\Sigma \,\mathrm{id}_{a!} \cong H_a^\Sigma$. □

Remark 3.10 Note that as far as the center $Z(\mathcal{NH}_a^\Sigma) = \boldsymbol{e}_a \mathcal{NH}_a^\Sigma \boldsymbol{e}_a$ is concerned, there is nothing special about ξ_1: in $\boldsymbol{e}_a \mathcal{NH}_a^\Sigma \boldsymbol{e}_a$ the relation $\boldsymbol{e}_a P(\xi_j) \boldsymbol{e}_a = 0$ holds for every $1 \le j \le a$.

3.2 Decomposing the Σ–deformed Grassmannian cohomology ring

The following is equivalent to Theorem 1.1 in the special case of the $\bigwedge^a \mathbb{C}^N$–colored unknot.

Theorem 3.11 *Let $\lambda_1, \dots, \lambda_l$ be pairwise distinct complex numbers and $N_1, \dots, N_l$ natural numbers such that $\sum_{i=1}^l N_i = N$ and let $\Sigma = \{\lambda_1^{N_1}, \dots, \lambda_l^{N_l}\}$ be the multiset containing λ_i exactly N_i times. There is an isomorphism of $\mathbb{C}$–algebras*

$$H_a^\Sigma \cong \bigoplus_{\substack{\sum a_j = a \\ 0 \le a_j \le N_j}} \bigotimes_{j=1}^{l} H_{a_j}^{N_j}.$$

Definition 3.12 Let $\mathbb{X} = \{\xi_1, \dots, \xi_a\}$ be an alphabet of a variables and $H_1^\Sigma = \mathbb{C}[\xi]/\langle P(\xi)\rangle$. Then we define $T_a^\Sigma := \langle P(\xi_1), \dots P(\xi_a)\rangle$ and identify

$$\bigotimes_{i=1}^{a} H_1^\Sigma \cong \frac{\mathbb{C}[\mathbb{X}]}{T_a^\Sigma} =: R_a^\Sigma.$$

The symmetric group S_a acts on this by permuting tensor factors or, in other words, by permuting the ξ_i. Denote by $\bigwedge^a H_1^\Sigma$ the vector space of antisymmetric tensors in $\bigotimes_{i=1}^a H_1^\Sigma$ and by $\bigwedge^a \mathbb{C}[\xi]$ the vector space of antisymmetric tensors in $\bigotimes_{i=1}^a \mathbb{C}[\xi]$. The latter we identify with antisymmetric polynomials in $\mathbb{C}[\mathbb{X}]$. In both cases, we denote the antisymmetrization map by

$$\mathrm{Antisym}(-) = \frac{1}{a!} \sum_{w \in S_a} \epsilon(w) w(-).$$

Recall that $\Delta_\xi = \prod_{1 \le i < j \le a} (\xi_j - \xi_i)$ denotes the Vandermonde determinant. Multiplying by Δ_ξ is a vector space isomorphism from $\mathrm{Sym}(\mathbb{X})$ to $\bigwedge^a \mathbb{C}[\xi]$ and equips the latter with the pushforward algebra structure: if $\Delta_\xi f, \Delta_\xi g \in \bigwedge^a \mathbb{C}[\xi]$ for $f, g \in \mathrm{Sym}(\mathbb{X})$, then

$$(\Delta_\xi f) * (\Delta_\xi g) := \Delta_\xi (fg).$$

Lemma 3.13 *The pushforward algebra structure on $\bigwedge^a \mathbb{C}[\xi]$ descends to the quotient $\bigwedge^a H_1^\Sigma$, and multiplication by Δ_ξ descends to an algebra isomorphism*

$$H_a^\Sigma \xrightarrow{\cong} \textstyle\bigwedge^a H_1^\Sigma.$$

Proof It suffices to check that $\Delta_\xi \cdot I_a^\Sigma \subset T_a^\Sigma$. We then have the composition of linear maps

$$\begin{aligned} H_a^\Sigma = \frac{\mathrm{Sym}(\mathbb{X})}{I_a^\Sigma} \overset{\Delta_\xi}{\cong} \frac{\bigwedge^a \mathbb{C}[\xi]}{\Delta_\xi \cdot I_a^\Sigma} \twoheadrightarrow \frac{\bigwedge^a \mathbb{C}[\xi]}{\bigwedge^a \mathbb{C}[\xi] \cap T_a^\Sigma} &\cong \frac{\bigwedge^a \mathbb{C}[\xi] + T_a^\Sigma}{T_a^\Sigma} \\ &= \mathrm{Antisym}\bigg(\bigotimes_{i=1}^a H_1^\Sigma \bigg) = \textstyle\bigwedge^a H_1^\Sigma, \end{aligned}$$

which is surjective, and hence must be an isomorphism for dimensional reasons.

We now check that the generators of I_a^Σ are mapped into T_a^Σ under multiplication by Δ_ξ. Let $1 \le j \le a$; then we have

$$\Delta_\xi h_{N-a+j}(\mathbb{X} - \Sigma) = \sum_{i=0}^{N} (-1)^i e_i(\Sigma) \Delta_\xi h_{N-a+j-i}(\mathbb{X})$$

$$= \sum_{i=0}^{N} (-1)^i e_i(\Sigma) \sum_{w \in S_a} \epsilon(w) \xi_{w(a)}^{a-1+N-a+j-i} \xi_{w(a-1)}^{a-2} \cdots \xi_{w(2)}^{1}$$
$$= \sum_{w \in S_a} \epsilon(w) (\xi_{w(a)}^{j-1} P(\xi_{w(a)})) \xi_{w(a-1)}^{a-2} \cdots \xi_{w(2)}^{1} \in T_a^{\Sigma}.$$

Here we have used the identity $\Delta_\xi h_k(\mathbb{X}) = \sum_{w \in S_a} \epsilon(w) \xi_{w(a)}^{a-1+k} \xi_{w(a-1)}^{a-2} \cdots \xi_{w(2)}^{1}$, which is clear from the defining formula for the Schur polynomials

$$\pi_\alpha(\mathbb{X}) := \frac{\det_{1 \le i,j, \le m}(\xi_i^{\alpha_j + a - j})}{\det_{1 \le i,j, \le m}(\xi_i^{a-j})} = \frac{\det_{1 \le i,j, \le m}(\xi_i^{\alpha_j + a - j})}{\Delta_\xi}$$

and the identity $h_k(\mathbb{X}) = \pi_{(k)}(\mathbb{X})$. □

Proof of Theorem 3.11 By the Chinese remainder theorem we know that

$$\frac{\mathbb{C}[\xi]}{\langle P(\xi) \rangle} \cong \bigoplus_{i=1}^{l} \frac{\mathbb{C}[\xi]}{\langle (\xi - \lambda_i)^{N_i} \rangle},$$

so let $\mathbb{1}_\mu(\xi) \in \mathbb{C}[\xi]$ be a representative for the idempotent that picks out the summand corresponding to the root $\mu \in \Sigma$. Thus we get the algebra isomorphism

$$H_1^{\Sigma} = \frac{\mathbb{C}[\xi]}{\langle P(\xi) \rangle} \cong \bigoplus_{i=1}^{l} \mathbb{1}_{\lambda_i}(\xi) H_1^{\Sigma}.$$

In the following we make liberal use of the canonical isomorphism

$$\bigotimes_{j=1}^{a} H_1^{\Sigma} \cong \frac{\mathbb{C}[\mathbb{X}]}{T_a^{\Sigma}} = R_a^{\Sigma}.$$

A set of minimal idempotents in R_a^{Σ} is given by $\{\mathbb{1}_{\bar\mu} := \prod_{j=1}^{a} \mathbb{1}_{\mu_j}(\xi_j)\}$, where $\bar\mu = (\mu_1, \dots, \mu_a)$ ranges over all a–tuples of roots appearing in Σ. The symmetric group S_a acts on $\mathbb{C}[\mathbb{X}]$ and R_a^{Σ} by permuting the indices of the variables ξ_i and on tuples $\bar\mu$ by permuting roots. For $w \in S_a$ we have

$$w(\mathbb{1}_{\bar\mu}) = \prod_{j=1}^{a} \mathbb{1}_{\mu_j}(\xi_{w(j)}) = \prod_{j=1}^{a} \mathbb{1}_{\mu_{w^{-1}(j)}}(\xi_j) = \mathbb{1}_{w^{-1}(\bar\mu)}.$$

Given an a–element multiset $A = \{\lambda_1^{a_1}, \dots, \lambda_l^{a_l}\}$ we write $\mu_A := (\lambda_1, \dots, \lambda_2, \dots, \lambda_l)$ for the corresponding tuple ordered by index. Every $\bar\mu$ can be written as $\bar\mu = \tau^{-1}(\mu_A)$

for a $\tau \in S_a$ and a multiset A, and this presentation is unique if we restrict the choice of τ to a set of coset representatives[7] T of $\prod_{i=1}^{l} S_{a_i}$ in S_a.

With these conventions in place, we can decompose R_a^Σ into S_a–invariant direct summands:

$$R_a^\Sigma \cong \bigoplus_{\substack{a\text{–element multisets A}\\ \text{of roots}}} \underbrace{\bigoplus_{\tau\in T} \tau(\mathbb{1}_{\mu_A}) R_a^\Sigma}_{S_a\text{–invariant}}. \tag{3-4}$$

Taking antisymmetric components respects the decomposition on the right-hand side into S_a–invariant direct summands. Thus, our goal is to compute the antisymmetric component of an (outer) summand on the right-hand side. Consider the projection

$$\bigoplus_{\tau\in T} \tau(\mathbb{1}_{\mu_A}) R_a^\Sigma \to \mathbb{1}_{\mu_A} R_a^\Sigma$$

which is given by multiplying by the idempotent $\mathbb{1}_{\mu_A}$. An elementary computation shows that this restricts to a vector space isomorphism

$$\pi\colon X_1 := \mathrm{Antisym}_{S_a}\biggl(\bigoplus_{\tau\in T} \tau(\mathbb{1}_{\mu_A}) R_a^\Sigma\biggr) \to \mathrm{Antisym}_{\prod_{i=1}^{l} S_{a_l}}(\mathbb{1}_{\mu_A} R_a^\Sigma) =: X_2, \tag{3-5}$$

where the right-hand side denotes the vector space of tensors y in $\mathbb{1}_{\mu_A} R_a^\Sigma$ which are antisymmetric for the action of $\prod_{i=1}^{l} S_{a_l} \subset S_a$; that is, $w(y) = \epsilon(w)y$ for all $w \in \prod_{i=1}^{l} S_{a_l}$. The inverse for π is given by $\psi(y) := \sum_{\tau\in T} \epsilon(\tau)\tau(\mathbb{1}_{\mu_A} y)$.

Fix $A = \{\lambda_1^{a_1}, \dots, \lambda_l^{a_l}\}$ as above; then for $1 \le i \le l$ we denote

$$\mathbb{X}_i := \Bigl\{\xi_{1+\sum_{k=1}^{i-1} a_k}, \dots, \xi_{\sum_{k=1}^{i} a_k}\Bigr\}, \quad T_a^{\lambda_i\in\Sigma} := \langle P(\xi) \mid \xi \in \mathbb{X}_i\rangle,$$
$$R_i := \frac{\mathbb{C}[\mathbb{X}_i]}{T_a^{\lambda_i\in\Sigma}} \quad \text{and} \quad \mathbb{1}_{\lambda_i} := \prod_{\xi\in\mathbb{X}_i} \mathbb{1}_{\lambda_i}(\xi).$$

Under the canonical isomorphism $R_a^\Sigma \cong \bigotimes_{i=1}^{l} R_i$ we have $\mathbb{1}_{\mu_A} R_a^\Sigma \cong \bigotimes_{i=1}^{l} \mathbb{1}_{\lambda_i} R_i$ and

$$X_2 = \mathrm{Antisym}_{\prod_{i=1}^{l} S_{a_l}}(\mathbb{1}_{\mu_A} R_a^\Sigma) \cong \bigotimes_{i=1}^{l} \mathrm{Antisym}_{S_{a_i}} \mathbb{1}_{\lambda_i} R_i =: X_3. \tag{3-6}$$

[7] For convenience we choose T to be simultaneously a set of right and left coset representatives.

Since $\mathbb{1}_A := \sum_{\tau \in T} \tau(\mathbb{1}_{\mu_A})$ is S_a–invariant and $\mathbb{1}_{\mu_A}$ is $\prod_{i=1}^{l} S_{a_i}$–invariant, we may note for later use that we also have

$$X_1 = \operatorname{Antisym}_{S_a}\biggl(\bigoplus_{\tau \in T} \tau(\mathbb{1}_{\mu_A}) R_a^{\Sigma}\biggr) = \mathbb{1}_A \operatorname{Antisym}_{S_a}(R_a^{\Sigma}),$$
$$X_2 = \operatorname{Antisym}_{\prod_{i=1}^{l} S_{a_l}}(\mathbb{1}_{\mu_A} R_a^{\Sigma}) = \mathbb{1}_{\mu_A} \operatorname{Antisym}_{\prod_{i=1}^{l} S_{a_i}}(R_a^{\Sigma}),$$

with respect to the multiplication in R_a^{Σ}.

From the Chinese remainder theorem we know that $\xi \mapsto w + \lambda_i$ gives an algebra isomorphism

$$\phi\colon \mathbb{1}_{\lambda_i}(\xi)\frac{\mathbb{C}[\xi]}{\langle P(\xi)\rangle} \to \frac{\mathbb{C}[\xi]}{\langle (\xi-\lambda_i)^{N_i}\rangle} \to \frac{\mathbb{C}[w]}{\langle w^{N_i}\rangle}$$

and this extends to an S_{a_i}–equivariant algebra isomorphism

$$\phi\colon \mathbb{1}_{\lambda_i} R_i = \mathbb{1}_{\lambda_i}\frac{\mathbb{C}[\mathbb{X}_i]}{T_a^{\lambda_i \in \Sigma}} \to \frac{\mathbb{C}[\mathbb{X}_i]}{\langle (\xi-\lambda_i)^{N_i} \mid \xi \in \mathbb{X}_i\rangle} \to \frac{\mathbb{C}[\mathbb{W}_i]}{\langle w^{N_i} \mid w \in \mathbb{W}_i\rangle},$$

where $\mathbb{W}_i = \{w_1, \dots, w_{a_i}\}$ is an auxiliary alphabet. It follows from Lemma 3.13 that, when restricted to the antisymmetric component, ϕ gives the vector space isomorphism

$$\phi\colon X_3 = \bigotimes_{i=1}^{l} \operatorname{Antisym}_{S_{a_i}} \mathbb{1}_{\lambda_i} R_i \to \bigotimes_{i=1}^{l} H_{a_i}^{N_i}. \tag{3-7}$$

The composition of the vector space isomorphisms in equations (3-5), (3-6) and (3-7) thus gives a decomposition of the S_a–invariant direct summands of (3-4), as required by the statement of the theorem. However, we further must check that the composition is an algebra isomorphism. In fact it is not, but it is close and the discrepancy is not hard to fix.

To see this, we compute the pushforward of the multiplication $*$ on X_1 under π. Let $x, y \in X_1$ be represented by antisymmetric polynomials in $\mathbb{C}[\mathbb{X}]$ and denote by xy their product in R_a^{Σ} and by $x * y$ their product in $\bigwedge^a H_1^{\Sigma}$. We compute

$$\pi(x) * \pi(y) := \pi(x * y) = \pi\biggl(\frac{xy}{\Delta_\xi}\biggr) = \mathbb{1}_{\mu_A}\frac{xy}{\Delta_\xi} = c\frac{(\mathbb{1}_{\mu_A}x)(\mathbb{1}_{\mu_A}y)}{\prod_{i=1}^{l}\Delta_i},$$

where

$$\Delta_i := \prod_{1+\sum_{k=1}^{i-1} a_k \le r < s \le \sum_{k=1}^{i} a_k} (\xi_r - \xi_s)$$

is the Vandermonde determinant in the subalphabet $\mathbb{X}_i \subset \mathbb{X}$ and $c = \mathbb{1}_{\mu_A}\bigl(\prod_{i=1}^{l}\Delta_l\bigr)/\Delta_\xi$. We will see in Lemma 3.14 that c represents a unit in $\mathbb{1}_{\mu_A} R_a^{\Sigma}$ and clearly it is $\prod_{i=1}^{l} S_{a_i}$–invariant. It follows that π/c is still a vector space isomorphism $X_1 \to X_2$, and the pushforward of the multiplication $*$ on X_1 under it is given by

$$(\pi/c)(x) * (\pi/c)(y) := (\pi/c)(x * y) = \frac{(\mathbb{1}_{\mu_A} x)(\mathbb{1}_{\mu_A} y)}{\prod_{i=1}^{l} \Delta_i}. \tag{3-8}$$

We now equip each tensorand $\operatorname{Antisym}_{S_{a_i}} \mathbb{1}_{\lambda_i} R_i$ of X_3 — see (3-6) — with the multiplication $*$ given by multiplying representing antisymmetric polynomials and then dividing by the appropriate Vandermonde determinant Δ_i. Then (3-8) says that $\pi/c\colon X_1 \to X_2$ composed with the canonical isomorphism $X_2 \to X_3$ is an algebra isomorphism with respect to the tensor product algebra structure on X_3. Since ϕ sends Δ_i to $\prod_{0 \le r < s \le a_i}(w_r + \lambda_i - w_s - \lambda_i) = \Delta_w$, an easy check shows that ϕ in (3-7) is also an algebra isomorphism.

To summarize the proof, we assemble the algebra isomorphisms:

$$H_a^\Sigma \cong {\textstyle\bigwedge}^a H_1^\Sigma \cong \bigoplus_{\substack{a\text{–element multisets } A\\ \text{of roots}}} \mathbb{1}_A \operatorname{Antisym}_{S_a}(\mathbb{1}_A R_a^\Sigma)$$

$$\cong \bigoplus_{\substack{\sum a_j = a\\ A=\{\lambda_1^{a_1},\dots,\lambda_l^{a_l}\}}} \bigotimes_{i=1}^{l} \operatorname{Antisym}_{S_{a_i}} \mathbb{1}_{\lambda_i} R_i \cong \bigoplus_{\substack{\sum a_j = a\\ 0 \le a_j \le N_j}} \bigotimes_{i=1}^{l} H_{a_i}^{N_i}.$$

The first isomorphism was introduced in Lemma 3.13, and the second one comes from the direct sum decomposition of $(H_1^\Sigma)^{\otimes a}$ into S_a–invariant summands. The third isomorphism is assembled from the isomorphisms π/c from (3-8) on summands composed with the canonical isomorphism in (3-6), and the last one comes from the Chinese remainder theorem and the inverse of the isomorphism from Lemma 3.13; see (3-7). The last isomorphism also shows that a summand indexed by a multiset A of roots is nonzero if and only if A is actually a multisubset of Σ. □

In the proof we have claimed that $c = \mathbb{1}_{\mu_A}\big(\prod_{i=1}^{l} \Delta_l\big)/\Delta_\xi$ represents a unit in $\mathbb{1}_{\mu_A} R_a^\Sigma$. This is clear from the following useful lemma:

Lemma 3.14 *Let R be a finite-dimensional quotient of a polynomial ring $R = \mathbb{C}[x_1, \dots, x_a]/I$ and let $V(I) \subset \mathbb{C}^a$ be the vanishing set of I. Then we have the decomposition*

$$R \cong \bigoplus_{v \in V(I)} \mathbb{1}_v R,$$

where $\mathbb{1}_v$ are minimal idempotents and $\mathbb{1}_v R$ is isomorphic to R_{p_v}, the localization of R at the complement of the maximal ideal $(x_1 - v_1, \dots, x_a - v_a)/I$. For elements

$\bar{f} \in \mathbb{1}_v R$ we have

$$\bar{f} \text{ is not a unit} \iff \bar{f} \text{ is a zero divisor} \iff f(v) = 0, \tag{3-9}$$

where f is any lift of $\bar{f}$ to $\mathbb{C}[x_1, \dots, x_a]$.

Proof Since R is a commutative Artinian ring, it decomposes uniquely into local commutative Artinian rings, one for each maximal ideal of R. Maximal ideals of R are in bijection with maximal ideals of $\mathbb{C}[x_1, \dots, x_a]$ that contain I. The maximal ideals of $\mathbb{C}[x_1, \dots, x_a]$ are exactly $I_v := (x_1 - v_1, \dots, x_a - v_a)$ for $v \in \mathbb{C}^a$ and $I \subset I_v$ if and only if $f(v) = 0$ for all $f \in I$, which holds if and only if $v \in V(I)$. It follows that

$$R \cong \bigoplus_{v \in V(I)} R_{p_v} \cong \bigoplus_{v \in V(I)} \mathbb{1}_v R,$$

where $p_v := I_v / I$, R_{p_v} denotes the localization of R at $R \setminus p_v$ and $\mathbb{1}_v \in R$ is the idempotent corresponding to the summand R_{p_v}. The statement about nonunits is then clear from the explicit description of the local ring R_{p_v}. □

Now, in the case of $c \in \mathbb{1}_{\mu_A} R_a^\Sigma$ for $\mu_A = (\mu_1, \dots, \mu_a) = (\lambda_1, \dots, \lambda_1, \lambda_2, \dots, \lambda_l)$ we have

$$c^{-1}\big|_{\xi_i \mapsto \mu_i} = \mathbb{1}_{\mu_A} \frac{\Delta_\xi}{\prod_{i=1}^{l} \Delta_l}\bigg|_{\xi_i \mapsto \mu_i} = \mathbb{1}_{\mu_A} \prod_{\mu_i \neq \mu_j, i<j} (\mu_i - \mu_j) \neq 0$$

and (3-9) shows that c^{-1}, hence also c, is a unit.

Remark 3.15 We have the isomorphism

$$H_a^\Sigma = \frac{\operatorname{Sym}(\mathbb{X})}{\langle h_{N-a+1}(\mathbb{X} - \Sigma), \dots, h_N(\mathbb{X} - \Sigma)\rangle} \cong \frac{\mathbb{C}[e_1(\mathbb{X}), \dots, e_a(\mathbb{X})]}{\langle h_{N-a+1}(\mathbb{X} - \Sigma), \dots, h_N(\mathbb{X} - \Sigma)\rangle}$$

and it follows by considering the generating function of $h_j(\mathbb{X} - \Sigma)$ that the vanishing set of this ideal is given by $\{(e_1(A), \dots, e_a(A)) \mid A \subset \Sigma,\ |A| = a\} \subset \mathbb{C}^a$. Applying Lemma 3.14 reproves the fact that the minimal idempotents of H_a^Σ are indexed by a–element multisubsets $A \subset \Sigma$. However, we should check that the idempotent corresponding to A identified in this remark — call it $\mathbb{1}'_A$ — equals $\mathbb{1}_A$ as defined in the proof of Theorem 3.11. For this it suffices to check that $\mathbb{1}_A(\mathbb{X})|_{\mathbb{X} \mapsto A} \neq 0 \in \mathbb{C}$. Recall that, by definition, $\mathbb{1}_{\lambda_i}(\xi) = \mathbb{1}'_{\lambda_i}(\xi)$ in $\mathbb{C}[\xi]/\langle P(\xi)\rangle$, and hence $\mathbb{1}_{\lambda_i}(\lambda_j) = \delta_i^j$. Further, $\mu_A = (\mu_1, \dots, \mu_a)$ was defined as the a–tuple consisting of elements λ_i

of A, ordered by index i, so we compute

$$\begin{aligned}\mathbb{1}_A(\mathbb{X})|_{\mathbb{X}\mapsto A} = \mathbb{1}_A(\mathbb{X})|_{(\xi_1,\dots,\xi_a)\mapsto\mu_A} &= \sum_{\tau\in T}\tau(\mathbb{1}_{\mu_A})(\mu_A)\\ &= \sum_{\tau\in T}\prod_{j=1}^{a}\mathbb{1}_{\mu_{\tau(j)}}(\mu_j) = \prod_{j=1}^{a}\mathbb{1}_{\mu_j}(\mu_j) = 1.\end{aligned}$$

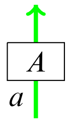
Corollary 3.16 *Let A be an a–element multisubset of Σ and $f \in \mathrm{Sym}(\mathbb{X})$; then f represents a unit in $\mathbb{1}_A H_a^\Sigma$ if and only if $f(A) \neq 0$.*

Proof This is immediate from (3-9) and Remark 3.15. □

3.3 Thick calculus for nil-Hecke quotients

We now deduce relations for the nil-Hecke quotients $\mathcal{NH}_a^\Sigma$ using the thick graphical calculus introduced in [21] and detailed above in Section 2.3. Note that in the quotients $\mathcal{NH}_a^\Sigma$ the element $\boldsymbol{e}_a$ is still an idempotent and it projects onto a direct summand isomorphic to $Z(\mathcal{NH}_a^\Sigma) \cong H_a^\Sigma$, but in general it is not a minimal idempotent due to the decomposition of H_a^Σ given in Theorem 3.11.

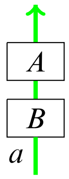
Corollary 3.17 *The collection of symmetric polynomials*



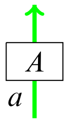
$$:= \mathbb{1}_A = \sum_{\tau\in T}\tau(\mathbb{1}_{\mu_A}) \in \mathrm{Sym}(\mathbb{X})$$

for $A \subset \Sigma$ and $|A| = a$, which were introduced in the proof of Theorem 3.11, give a complete collection of minimal orthogonal idempotents of $Z(\mathcal{NH}_a^\Sigma) \cong H_a^\Sigma$. In other words, in H_a^Σ we have that for a–element multisubsets A and B of Σ,

(3-10)



$=$

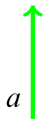


if $A = B$,

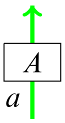
0 *if $A \neq B$,*

and the thick edge decomposes in H_a^Σ into a sum of $\mathbb{1}_A$–decorated thick edges:

(3-11)



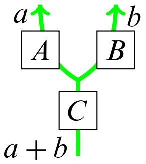
$$= \sum_{\substack{a\text{–element multisets}\\ A\subset\Sigma}}$$



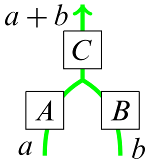
Proof This is immediate. □

Proposition 3.18 (nonadmissible colorings by multisubsets) *Let A, B and C be a–, b– and $(a+b)$–element multisubsets of Σ; then in $\mathcal{NH}^{\Sigma}_{a+b}$ we have*

(3-12)



=



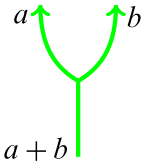
$= 0$ *if* $A \uplus B \neq C$

and we call such a coloring ***nonadmissible.*** (Here $\uplus$ denotes the multiset sum, or disjoint union, of multisets.)

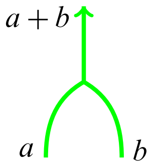
Labelings by idempotents corresponding to multisubsets of Σ that "add up" at mergers and splitters (ie $A \uplus B = C$) are called *admissible.*

Proof Denote by $\mathbb{X}_1$, $\mathbb{X}_2$ and $\mathbb{X}$ the alphabets of operators ξ_j on the strands of thickness a, b and $a+b$, respectively, and by $H^{\Sigma}_a(\mathbb{X}_1)$, $H^{\Sigma}_b(\mathbb{X}_2)$ and $H^{\Sigma}_{a+b}(\mathbb{X})$ the algebras of decorations on these strands.

Equation (2.61) in [21] then implies that the algebras of decorations on the diagrams



and



are both given by

$$\frac{H^{\Sigma}_{a+b}(\mathbb{X}) \otimes H^{\Sigma}_a(\mathbb{X}_1) \otimes H^{\Sigma}_b(\mathbb{X}_2)}{\langle e_i(\mathbb{X}) - e_i(\mathbb{X}_1 \sqcup \mathbb{X}_2) \mid i > 0\rangle}.$$

In the following we write $\langle \mathbb{X} = \mathbb{X}_1 \sqcup \mathbb{X}_2\rangle$ for the ideal $\langle e_i(\mathbb{X}) - e_i(\mathbb{X}_1 \sqcup \mathbb{X}_2) \mid i > 0\rangle$. Let A, B and C be a–, b– and $(a+b)$–element multisubsets of Σ, respectively. Then the algebra of additional decorations on the idempotent-decorated diagrams in (3-12) is

$$\frac{\mathbb{1}_C(\mathbb{X})H^{\Sigma}_{a+b}(\mathbb{X}) \otimes \mathbb{1}_A(\mathbb{X}_1)H^{\Sigma}_a(\mathbb{X}_1) \otimes \mathbb{1}_B(\mathbb{X}_2)H^{\Sigma}_b(\mathbb{X}_2)}{\langle \mathbb{X} = \mathbb{X}_1 \sqcup \mathbb{X}_2\rangle \cap \big(\mathbb{1}_C(\mathbb{X})H^{\Sigma}_{a+b}(\mathbb{X}) \otimes \mathbb{1}_A(\mathbb{X}_1)H^{\Sigma}_a(\mathbb{X}_1) \otimes \mathbb{1}_B(\mathbb{X}_2)H^{\Sigma}_b(\mathbb{X}_2)\big)}. \tag{3-13}$$

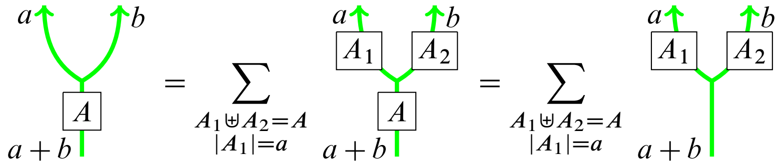
The numerator here is a direct summand of $H^{\Sigma}_{a+b}(\mathbb{X}) \otimes H^{\Sigma}_a(\mathbb{X}_1) \otimes H^{\Sigma}_b(\mathbb{X}_2)$ that can be picked out by localizing at the complement of the maximal ideal

$$\langle e_i(\mathbb{X}) - e_i(C), e_i(\mathbb{X}_1) - e_i(A), e_i(\mathbb{X}_2) - e_i(B) \mid i > 0\rangle.$$

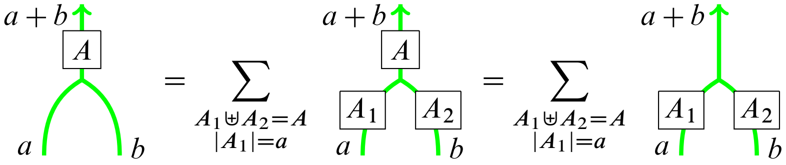
If $C \neq A \uplus B$ then there is a $j \in \mathbb{N}$ such that $e_j(C) - e_j(A \uplus B) \neq 0 \in \mathbb{C}$; thus, by Corollary 3.16, $e_j(\mathbb{X}) - e_j(\mathbb{X}_1 \sqcup \mathbb{X}_2)$ is a unit in the numerator. Taking the quotient in (3-13) then collapses the direct summand, and (3-12) then follows. □

Corollary 3.19 (idempotent decoration migration) *Let A be an $(a{+}b)$–element multisubset of Σ; then in $\mathcal{NH}^{\Sigma}_{a+b}$ we have:*

(3-14)

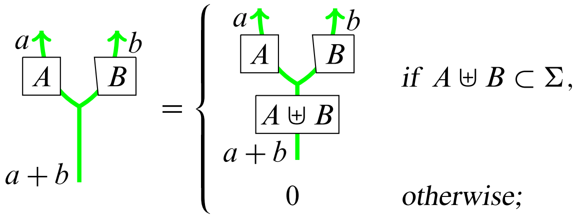


(3-15)

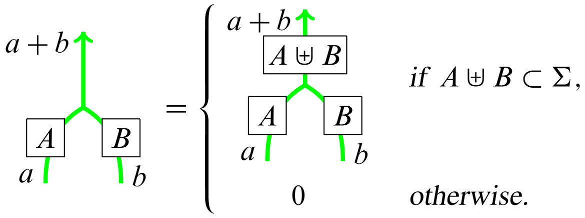


In particular, for multisubsets A, $B \subset \Sigma$ with $|A| = a$ and $|B| = b$, we have:

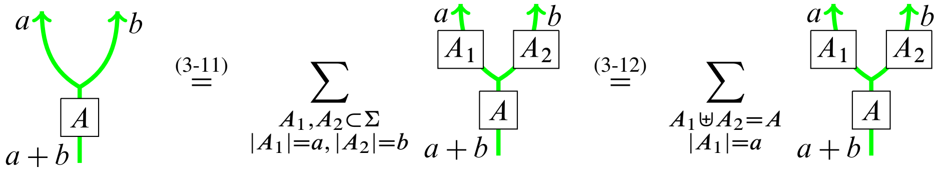
(3-16)



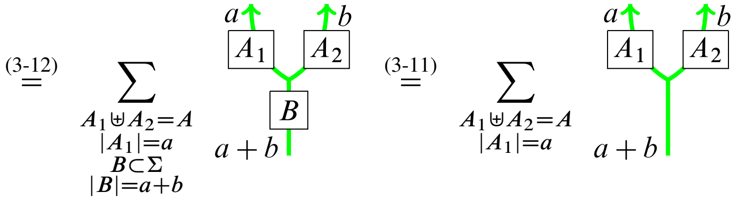
(3-17)



Proof For (3-14) we compute



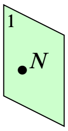
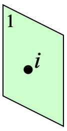
a
b
A
a + b
$\overset{(3\text{-}11)}{=}$
$\sum_{A_1, A_2 \subset \Sigma,\ |A_1| = a, |A_2| = b}$
A_1
A_2
$\overset{(3\text{-}12)}{=}$
$\sum_{A_1 \uplus A_2 = A,\ |A_1| = a}$


(3-12) =
$\sum_{A_1 \uplus A_2 = A, |A_1|=a, B\subset\Sigma, |B|=a+b}$
a
b
A_1
A_2
B
$a+b$
(3-11) =
$\sum_{A_1 \uplus A_2 = A, |A_1|=a}$
a
b
A_1
A_2
$a+b$

and the proof of (3-15) is analogous. Equations (3-16) and (3-17) follow similarly. □

4 The Σ-deformed foam category $N\mathbf{Foam}^{\Sigma}$

We define the 2–category $N\mathbf{Foam}^{\Sigma}$ of Σ–deformed $\mathfrak{sl}_N$–foams as the quotient of the foam 2–category $N\mathbf{Foam}$, described in Section 2.1, by the following additional relation on 1–labeled foam facets:

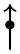
(4-1)  $= \sum_{i=0}^{N-1} (-1)^{N-i-1} e_{N-i}(\Sigma)$ 

Since this equation is not degree-homogeneous, we hence ignore the grading on foams (ie to be precise we first pass to the ungraded version of $N\mathbf{Foam}$, then impose this relation to pass to $N\mathbf{Foam}^{\Sigma}$).

This quotient is motivated by the deformed nil-Hecke algebra quotient introduced in the last section. Indeed, the 2–representation $\check{\mathcal{U}}_Q(\mathfrak{gl}_m) \to N\mathbf{Foam}$ gives an action of the nil-Hecke algebra on the latter, and in order to obtain an action of the Σ–deformed nil-Hecke quotient, we impose this local foam analog of (3-2).

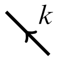
Definition 4.1 We let $\Phi_\Sigma\colon \check{\mathcal{U}}_Q(\mathfrak{gl}_m) \to N\mathbf{Foam}^{\Sigma}$ be the composition of the foamation 2–functor $\Phi_m\colon \check{\mathcal{U}}_Q(\mathfrak{gl}_m) \to N\mathbf{Foam}$ and the quotient 2–functor $N\mathbf{Foam} \to N\mathbf{Foam}^{\Sigma}$.

It follows that the 2–functor $\Phi_\Sigma\colon \check{\mathcal{U}}_Q(\mathfrak{gl}_m) \to N\mathbf{Foam}^{\Sigma}$ factors through the quotient of $\check{\mathcal{U}}_Q(\mathfrak{gl}_m)$ in which we've imposed the relation that dots satisfy the equation

$P\Big($  $\Big) = 0;$

hence, the thick calculus equations in Corollaries 3.17 and 3.19 and Proposition 3.18 correspond to analogous foam relations in $N\mathbf{Foam}^{\Sigma}$. In fact, the thick calculus relations can be seen as intersections of foam relations with planes. More precisely, we get:

Lemma 4.2 *The algebra of decorations of a k–labeled foam facet, or alternatively, the endomorphism algebra of the k–labeled web edge, carries an action of H_k^Σ. In fact, from the 2–representation on deformed matrix factorizations in Section 4.4 it follows that there is an isomorphism*

End() $\cong H_k^\Sigma.$

Compare with [36, Remark 4.1]. Moreover, we have the following important consequences:

- Every k–labeled foam facet in $N\mathbf{Foam}^\Sigma$ splits into a sum over foam facets colored by minimal idempotent decorations corresponding to k–element multisubsets of Σ.
- Equation (3-12) then implies that a foam is zero whenever it contains a seam whose adjacent facets are nonadmissibly colored by idempotents. Here, similar to the case of thick calculus diagrams, we say that a foam is *admissibly colored* precisely when around any seam the sum of the multisets of the idempotents coloring two of the facets equals the multiset coloring the third. Consequently, foam relations analogous to those of Corollary 3.19 hold in a neighborhood of any seam.

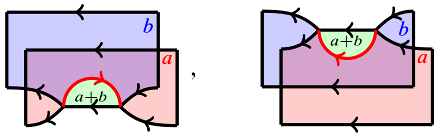
4.1 Foam splitting relations

Convention 4.3 Let $A, B \subset \Sigma$ be disjoint multisubsets of roots:

$$\lambda \in A \implies \lambda \notin B \qquad \text{and} \qquad \mu \in B \implies \mu \notin A.$$

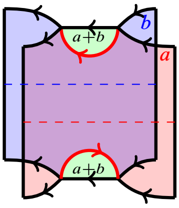
For the duration, unless otherwise stated, we use red and blue colored foam facets to denote facets decorated by the orthogonal idempotents $\mathbb{1}_A$ and $\mathbb{1}_B$, respectively. We use green as a generic color for both undecorated foam facets and for decorations by $\mathbb{1}_{A \uplus B}$.

Lemma 4.4 *The following foams are invertible as 2–morphisms in $N\mathbf{Foam}^\Sigma$:*

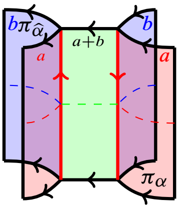


Proof Decorating the $b=c$ case of the foam relations in (2-4) by red and blue idempotents, we get:

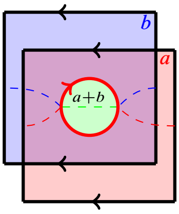
(4-2)



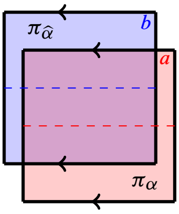
$$= \sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}$$



(4-3)



$$= \sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}$$

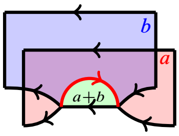
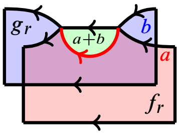


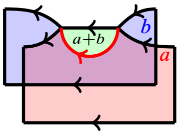
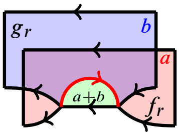
Let $\mathbb{X}$ and $\mathbb{Y}$ be the alphabets assigned to the red and blue foam facets where π_α and $\pi_{\hat{\alpha}}$ are placed. We check that $\sum_{\alpha\in P(a,b)}(-1)^{|\hat{\alpha}|}\pi_\alpha(\mathbb{X})\pi_{\hat{\alpha}}(\mathbb{Y})$ represents a unit in $\mathbb{1}_A H_a^\Sigma(\mathbb{X})\otimes\mathbb{1}_B H_b^\Sigma(\mathbb{Y})$ by using the criterion in Corollary 3.16:

$$\begin{aligned}\sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}\pi_\alpha(\mathbb{X})\pi_{\hat{\alpha}}(\mathbb{Y})\Big|_{\substack{\mathbb{X}\mapsto A\\ \mathbb{Y}\mapsto B}} &= \sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}\pi_\alpha(A)\pi_{\hat{\alpha}}(B)\\ &= \prod_{\lambda\in A}\prod_{\mu\in B}(\mu-\lambda)\neq 0\in\mathbb{C}.\end{aligned}$$

A proof for the second equality can eg be found in [31, Example 5, page 65], and the product is nonzero because A and B consist of distinct roots.

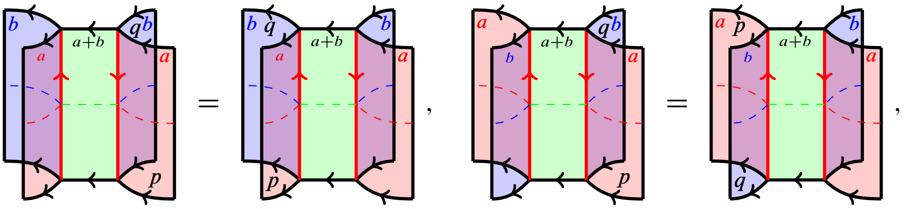
Let $\sum_r f_r(\mathbb{X})g_r(\mathbb{Y})$ be a representative of $\big(\sum_{\alpha\in P(a,b)}(-1)^{|\hat{\alpha}|}\pi_\alpha(\mathbb{X})\pi_{\hat{\alpha}}(\mathbb{Y})\big)^{-1}$ in the ring $H_a^\Sigma(\mathbb{X})\otimes H_b^\Sigma(\mathbb{Y})$; then the following are explicit inverses for the decorated unzip and zip foams:

$\Bigg($  $\Bigg)^{-1} = \sum_r$  ,

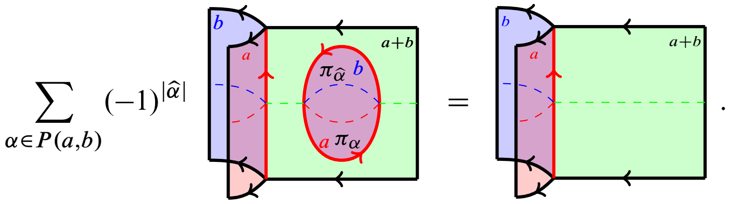
$\Bigg($  $\Bigg)^{-1} = \sum_r$  . □

Lemma 4.5 *Let p and q be symmetric polynomials in a and b variables, respectively. Then the following relations hold:*

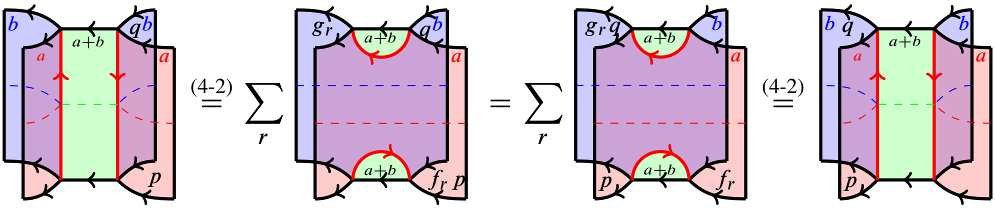
(4-4)



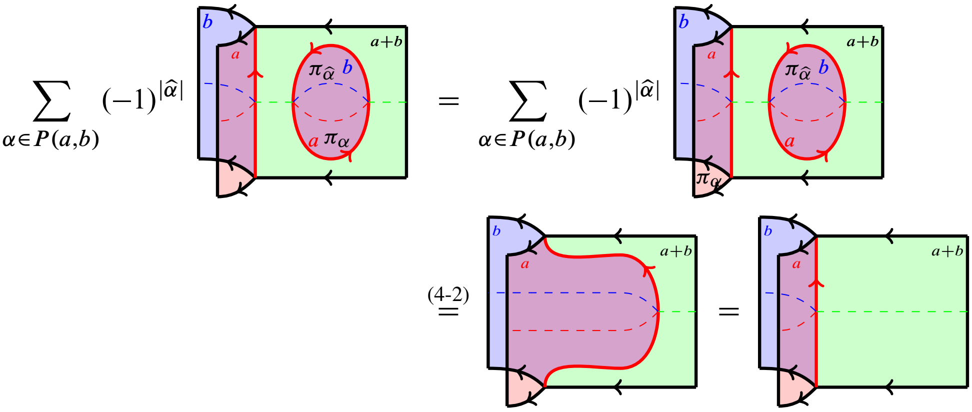
(4-5)



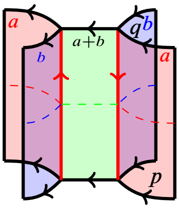
Proof We again use $\sum_r f_r(\mathbb{X})g_r(\mathbb{Y})$, which is a representative of the inverse of $\sum_{\alpha\in P(a,b)}(-1)^{|\hat\alpha|}\pi_\alpha(\mathbb{X})\pi_{\hat\alpha}(\mathbb{Y})$ in $H_a^\Sigma(\mathbb{X})\otimes H_b^\Sigma(\mathbb{Y})$. For the first relation in (4-4) we compute:



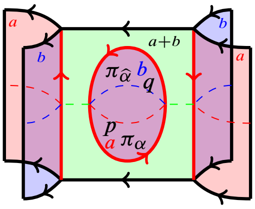
Equation (4-5) then follows via:


$\sum_{\alpha\in P(a,b)} (-1)^{|\hat\alpha|}$
b
a
a+b
$\pi_{\hat\alpha}$ b
a π_α
=
$\sum_{\alpha\in P(a,b)} (-1)^{|\hat\alpha|}$
π_α
$\overset{(4\text{-}2)}{=}$
=

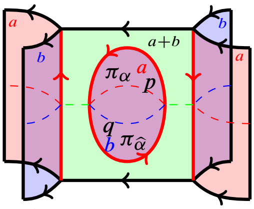
For the second relation in (4-4) we now have:



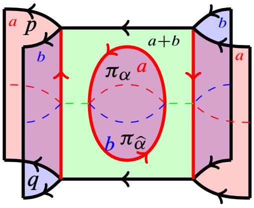
$$= \sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}$$



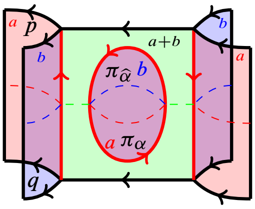
$$= \sum_{\alpha\in P(a,b)} (-1)^{ab+|\hat{\alpha}|}$$



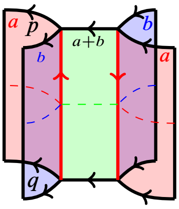
$$= \sum_{\alpha\in P(a,b)} (-1)^{ab+|\hat{\alpha}|}$$



$$= \sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}$$

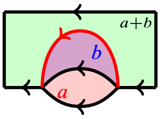


$=$

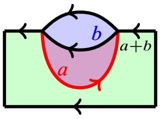


□

Lemma 4.6 *The following foams are invertible as 2–morphisms in* $N\mathbf{Foam}^{\Sigma}$*:*

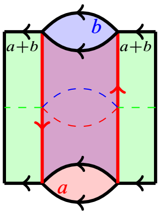


,

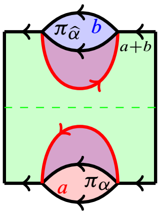


Proof Given the relations[8]

(4-6)



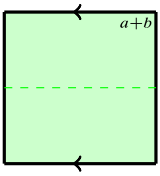
$$= \sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}$$



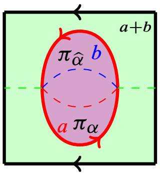
,

[8] In relation (4-7) the green shading is meant to indicate a decoration by the mixed idempotent $\mathbb{1}_{A\uplus B}$.

(4-7)



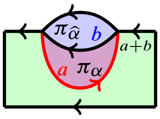
$$= \sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}$$



,

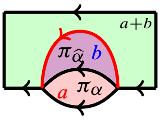
and the decoration migration relations (4-4), it follows immediately that

$$\sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}$$



and

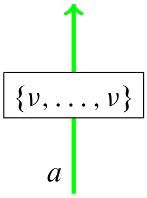
$$\sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}$$



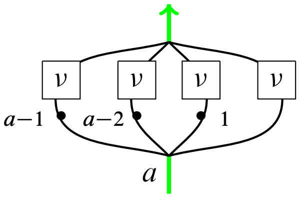
are inverse to the digon removal and creation foams, respectively.

Equation (4-6) is just an idempotent decorated version of relation (2-2). Equation (4-7) is a stronger, more local version of (4-5), but we cannot use the same trick to deduce it. To de-clutter the pictures, we compute this relation in $\mathcal{NH}^{\Sigma}_{a+b}$; the result can then be transferred using the foamation functor Φ_Σ. Alternatively, one can interpret the following nil-Hecke pictures as $2d$–slices through the corresponding foams.

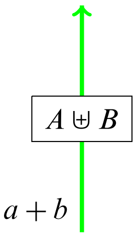
We begin by using (2-9) to explode a thick edge into thin edges, and then combine this relation with Corollary 3.19 to slide the decoration by the multiset onto the thin edges. In the simplest case, where the multiset contains only one root ν, we have:



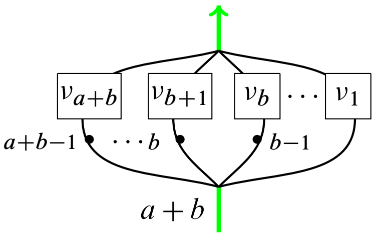
$=$



Now suppose that $A=\{\lambda,\dots,\lambda\}$ and $B=\{\mu,\dots,\mu\}$ with $\lambda\neq\mu$, then similarly we have:

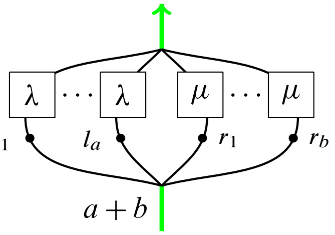


$$= \sum_{\substack{(\nu_{a+b},\dots,\nu_1)\\ \text{is a reordering of}\\ (\lambda,\dots,\lambda,\mu,\dots,\mu)}}$$



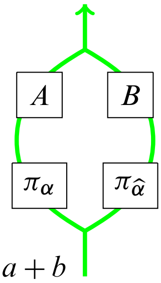
Next we reorder the decorations on the strands (at the expense of signs) so that all λ idempotents lie on the left, all μ idempotents on the right, and in both groups of strands the number of additional dots decreases from left to right, so the right-hand side above becomes:

$$\sum_{l_1>\cdots>l_a} \pm$$



Here the sum is taken over all strictly decreasing sequences $a+b-1 \geq l_1 > \cdots > l_a \geq 0$ and $r_1, \dots, r_b$ are the remaining b numbers between 0 and $a+b-1$ in decreasing order. Clearly the set of such sequences $(l_1, \dots, l_a)$ is in bijection with partitions $(l_1-(a-1), l_2-(a-2), \dots, l_a)$ whose Young diagrams fit into a $a \times b$ box. If $(l_1, \dots, l_a)$ corresponds to a partition $\alpha \in P(a,b)$, then it is easy to check that $(r_1, \dots, r_b)$ corresponds to $\hat{\alpha} \in P(b,a)$ and the sign introduced by reordering decorations on strands is $(-1)^{|\hat{\alpha}|}$. Finally, we use (2-9) to express the a strands on the left and the b strands on the right in terms of strands of thickness a and b, respectively. This expresses the decorations $(l_1, \dots, l_a)$ and $(r_1, \dots, r_b)$ on the thin strands as Schur polynomials π_α and $\pi_{\hat{\alpha}}$, and using Corollary 3.19 we can slide the idempotents onto the thick strands to obtain:

$$\sum_{\alpha \in P(a,b)} (-1)^{|\hat{\alpha}|}$$



This gives the thick calculus version of (4-7) for this choice of A and B.

The case of general A and B is very similar. The main difference is that there are more possible reorderings of the decoration by roots on thin strands. However, if we interpret a nil-Hecke picture decorated by idempotents λ and μ as a sum over all possible ways of replacing the instances of λ by elements of A and of the μ by elements of B, then the proof of the special case immediately carries over to the general setting. □

4.2 Karoubi envelope technology

Let W be a web, ie a 1–morphism in $N\mathbf{Foam}^\Sigma$; then the foam versions of Proposition 3.18 and Corollary 3.19 show that the identity 2–morphism id_W decomposes into a sum of idempotent foams — one for each coloring of the edges of W by multisubsets of roots that is compatible at vertices. We now proceed to a 2–category $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ in which these idempotents split.

Definition 4.7 Let $\mathrm{Kar}(N\mathbf{Foam}^{\Sigma})$ denote the 2–category obtained by passing to the Karoubi envelope in each Hom–category of $N\mathbf{Foam}^{\Sigma}$. We define $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ to be a certain full 2–subcategory of $\mathrm{Kar}(N\mathbf{Foam}^{\Sigma})$ that contains as 1–morphisms all the pairs (W, F_W) where W is a web in $N\mathbf{Foam}^{\Sigma}$ and F_W is a decorated identity foam on W in $N\mathbf{Foam}^{\Sigma}$ such that each a–labeled facet is decorated by an idempotent $\mathbb{1}_A$ corresponding to an a–element multisubset $A \subset \Sigma$. More precisely, $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ has the same objects as $N\mathbf{Foam}^{\Sigma}$ and has Hom–categories given by the full subcategories of the corresponding Hom–categories of $\mathrm{Kar}(N\mathbf{Foam}^{\Sigma})$ that contain all formal direct sums of pairs (W, F_W).

Note that, in particular, $N\mathbf{Foam}^{\Sigma}$ embeds as a full 2–subcategory of $(N\mathbf{Foam}^{\Sigma})^{\wedge}$, since the identity foam over any web can be expressed as the sum over all possible colorings of its facets. Practically speaking, $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ can be viewed as the 2–category in which

- objects are sequences $\boldsymbol{a} = (a_1, \dots, a_m)$ for $m \geq 0$ as in $N\mathbf{Foam}^{\Sigma}$,
- 1–morphisms are formal direct sums of webs where, in addition to a labeling, each a–labeled edge is colored by an idempotent $\mathbb{1}_A$ corresponding to an a–element multisubset $A \subset \Sigma$, and
- 2–morphisms are matrices of linear combinations of foams as in $N\mathbf{Foam}^{\Sigma}$, but with each facet incident upon a web edge decorated by the idempotent coloring the edge.

As in the case of thick calculus diagrams and foams, we call a web *admissibly colored* if at each trivalent vertex the union of the multisets coloring two of the edges equals the third. Since nonadmissibly colored foams are zero, it follows that a nonadmissibly colored web is isomorphic to the "zero web" (ie the zero object in the relevant Hom–category).

We now point out that in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ there are three[9] ways of composing morphisms, and establish our notation for them:

- Sequences, webs, and foams can be placed side by side (ie on objects this is concatenation of sequences). We denote this operation by $\sqcup$.
- Webs and foams can be composed in the 1–morphism direction, ie glued horizontally along their left and right boundaries, and we denote this by $\otimes$.
- Foams can be composed in the 2–morphism direction by gluing vertically, and we write $\circ$ for this operation.

[9] This is due to the fact that $N\mathbf{Foam}$ is secretly a 3–category.

We will also utilize two notions of "equivalence" for colored webs in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$. A 1–morphism $W\colon \boldsymbol{o}_1 \to \boldsymbol{o}_2$ in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ is *isomorphic* to a 1–morphism $V\colon \boldsymbol{o}_1 \to \boldsymbol{o}_2$ if there exist 2–morphisms $F_1\colon W \to V$ and $F_2\colon V \to W$ in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ such that

$$F_2 \circ F_1 = \mathrm{id}_W \quad \text{and} \quad F_1 \circ F_2 = \mathrm{id}_V.$$

In this case we write $V \cong W$. Next, a 1–morphism $W\colon \boldsymbol{o}_1 \to \boldsymbol{o}_2$ in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ is *weakly equivalent* to a 1–morphism $V\colon \boldsymbol{u}_1 \to \boldsymbol{u}_2$ if there exist 1–morphisms $L\colon \boldsymbol{o}_2 \to \boldsymbol{u}_2$, $L^{-1}\colon \boldsymbol{u}_2 \to \boldsymbol{o}_2$, $R\colon \boldsymbol{u}_1 \to \boldsymbol{o}_1$ and $R^{-1}\colon \boldsymbol{o}_1 \to \boldsymbol{u}_1$ such that

$$L \otimes W \otimes R \cong V, \quad L^{-1} \otimes L \cong 1_{\boldsymbol{o}_2}, \quad L \otimes L^{-1} \cong 1_{\boldsymbol{u}_2}, \quad R \otimes R^{-1} \cong 1_{\boldsymbol{o}_1}, \quad R^{-1} \otimes R \cong 1_{\boldsymbol{u}_1}.$$

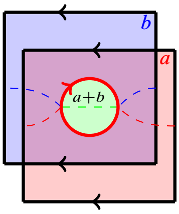
We now aim to use these notions of equivalence to "split" the foam 2–category $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ into pieces in which webs and foams are colored by multisubsets of Σ containing only one root $\lambda \in \Sigma$. Although we do not prove a full decomposition theorem (see Remark 4.28), we will see in Section 5.2 that the splitting results obtained here suffice to decompose the link invariant as in Theorem 1.1.

Let F be a foam with an admissible coloring of facets by multisubsets of Σ and let $\lambda \in \Sigma$ be a root. We want to define the foam F_λ that results from forgetting everything in F that is not colored by λ. More precisely, consider the underlying CW-complex of F; in it we erase all 2–cells that are colored with multisubsets not containing λ and smoothen out all seams that have become obsolete. We define a foam structure on the resulting CW-complex by setting the label of each remaining 2–cell to be the (positive) multiplicity of λ in the corresponding color on F. This is again a foam by admissibility of the original coloring. Finally we decorate each facet with the idempotent of the multisubset containing only instances of λ.

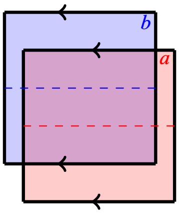
Definition 4.8 The *λ–component* of an admissibly colored foam F, denoted by F_λ, is the foam in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ constructed via the procedure just described.

Example 4.9 If $\lambda_1 \neq \lambda_2$ are two roots in Σ and colors red, blue and green indicate decorations with idempotents corresponding to multisets $\{\lambda_1^a\}$, $\{\lambda_2^b\}$ and $\{\lambda_1^a, \lambda_2^b\}$, respectively, then we have, for example:

$F =$



, $F_{\lambda_2} \sqcup F_{\lambda_1} =$

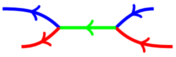
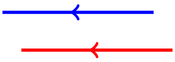


In the following, we will use the shorthand $\bigsqcup_\lambda F_\lambda := F_{\lambda_l} \sqcup \cdots \sqcup F_{\lambda_1}$.

Definition 4.10 Let W be a colored web in $(N\mathbf{Foam}^\Sigma)^\wedge$.

- The *λ–component W_λ of W* is the (co)domain of $(\mathrm{id}_W)_\lambda$, the λ–component of the identity foam on W. As for foams, we define the shorthand $\bigsqcup_\lambda W_\lambda := W_{\lambda_l} \sqcup \cdots \sqcup W_{\lambda_1}$.
- W is called *split* if $W = \bigsqcup_\lambda W_\lambda$. More generally, for any colored web W', the split web $\bigsqcup_\lambda W'_\lambda$ is called the *split web associated to W'*.

Example 4.11 With coloring conventions as in Example 4.9 we have, for example:

$W =$  , $W_{\lambda_2} \sqcup W_{\lambda_1} =$ 

Next, let $\boldsymbol{o} = (a_1, \ldots, a_m)$ be an object in $(N\mathbf{Foam}^\Sigma)^\wedge$ and suppose that for every entry a_i of $\boldsymbol{o}$ we are given an a_i–element multisubset $A_i = \{\lambda_1^{a_{i,1}}, \ldots, \lambda_l^{a_{i,l}}\} \subset \Sigma$; we call such a collection $\boldsymbol{A} = (A_1, \ldots, A_m)$ an *incidence condition* for $\boldsymbol{o}$. We then consider the identity web on $\boldsymbol{o}$ with strands colored by multisubsets A_i, and use the following notation for the (co)domain of the associated split web:

$$\bigsqcup_\lambda \boldsymbol{o}_\lambda := (a_{1,l}, \ldots, a_{m,l}, \ldots, a_{1,1}, \ldots, a_{m,1}). \tag{4-8}$$

Definition 4.12 Let $L\colon \boldsymbol{o} \to \bigsqcup_\lambda \boldsymbol{o}_\lambda$ be the combinatorially simplest web from $\boldsymbol{o}$ to $\bigsqcup_\lambda \boldsymbol{o}_\lambda$ that is colored with the multiset A_i on the strand starting at the entry a_i of $\boldsymbol{o}$ and colored with the multiset $\{\lambda_j^{a_{i,j}}\}$ on the strand terminating at the entry $a_{i,j}$ of $\bigsqcup_\lambda \boldsymbol{o}_\lambda$. Analogously we define $R\colon \bigsqcup_\lambda \boldsymbol{o}_\lambda \to \boldsymbol{o}$ to be the combinatorially simplest web from $\bigsqcup_\lambda \boldsymbol{o}_\lambda$ to $\boldsymbol{o}$ that is colored with $\{\lambda_j^{a_{i,j}}\}$ on the strand starting at the entry $a_{i,j}$ of $\bigsqcup_\lambda \boldsymbol{o}_\lambda$ and colored with A_i on the strand terminating at the entry a_i of $\boldsymbol{o}$.

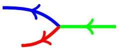
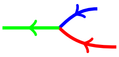
We now explicitly describe L (and R) before giving illustrations in Examples 4.13 and 4.14. L is given as a composition $L := L_{l-1} \otimes \cdots \otimes L_1$ with one component L_j for each root λ_j, except the last one. Each L_j itself can be decomposed as $L_j = L_{1,j} \otimes \cdots \otimes L_{m,j}$, where $L_{m,1}$ splits off the λ_1–component from the strand coming out of a_m and continues it below the remainder of the a_m strand. $L_{m-1,1}$ splits off the λ_1–component from the strand coming out of a_{m-1}, merges it with the remainder of the a_m strand, which contains no λ_1 any more, and splits it off on the other side. In general $L_{i,j}$ splits the λ_j–component off the remainder of the a_i–strand, and passes it through the remainders of all a_k–strands with $k > i$, which contain no λ_j any more. The composite L_j thus is the combinatorially simplest web that splits the λ_j–components off all a_k strands and continues them as a bundle of parallel strands below the a_i remainder strands and above the bundles of $\lambda_{j'}$–colored parallel strands for $j' < j$ that have been split off by $L_{j'}$. It is not hard to see that the composition L is, up to planar isotopy, the combinatorially simplest web from $\boldsymbol{o}$ to $\bigsqcup_\lambda \boldsymbol{o}_\lambda$ with the prescribed boundary colorings. The colored web R can be obtained similarly, or simply by reflecting L horizontally.

Given a colored web $W\colon \boldsymbol{o}_1 \to \boldsymbol{o}_2$, we will mostly be interested in the webs

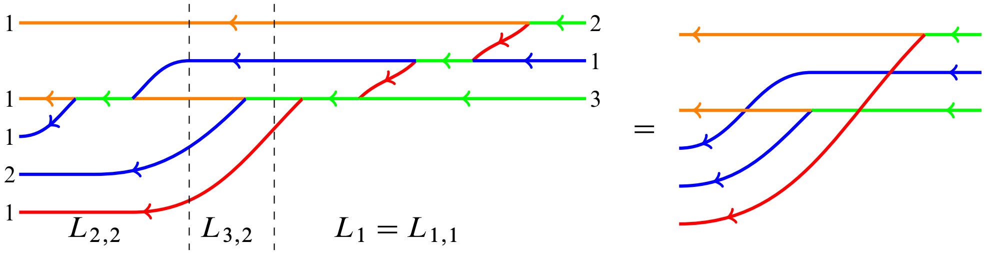
$$L\colon \boldsymbol{o}_2 \to \bigsqcup_\lambda \boldsymbol{o}_{2,\lambda} \quad \text{and} \quad R\colon \bigsqcup_\lambda \boldsymbol{o}_{1,\lambda} \to \boldsymbol{o}_1$$

constructed from the incidence conditions for $\boldsymbol{o}_2$ and $\boldsymbol{o}_1$ determined by the coloring of left and right boundary edges of the colored web W. In particular, we can then consider the colored web $L \otimes W \otimes R$.

Example 4.13 In the case of the identity web $1_{\boldsymbol{o}}$ on $\boldsymbol{o} = (a+b)$, which is colored by the multisubset $\{\lambda_1^a, \lambda_2^b\} \subset \Sigma$, and using the coloring conventions from Example 4.9, we have the following prototypical example:

$L =$  , $R =$ 

Example 4.14 For a slightly more generic example, let $\boldsymbol{o} = (2,1,3)$ with the incidence condition $A = (\{\lambda_1,\lambda_3\},\{\lambda_2\},\{\lambda_2^2,\lambda_3\})$, then $\bigsqcup_\lambda \boldsymbol{o}_\lambda = (1,1,1,2,1)$ and L takes the following form:



where we use Convention 4.19 below in the second diagram to write the first colored web more succinctly. Here web strands colored by multisets containing multiple roots are green, and those containing only one of λ_1, λ_2 or λ_3 are red, blue, and orange (respectively).

Lemma 4.15 *Suppose $\boldsymbol{o}$ is an object in $(N\mathbf{Foam}^\Sigma)^\wedge$ and fix an incidence condition for $\boldsymbol{o}$. Let L and R be the corresponding webs constructed in Definition 4.12. Then we have*

$$R \otimes L \cong 1_{\boldsymbol{o}} \quad \textit{and} \quad L \otimes R \cong 1_{\bigsqcup_\lambda \boldsymbol{o}_\lambda}.$$

Proof L and R are both compositions of mergers (and splitters) whose two incoming (outgoing) strands are colored with disjoint multisubsets. Moreover, splitters and mergers in R are paired up with mergers and splitters in L — in reverse order. Repeated application of Lemmas 4.4 and 4.6 allows the construction of foams giving the isomorphism (see also (4-9) below). □

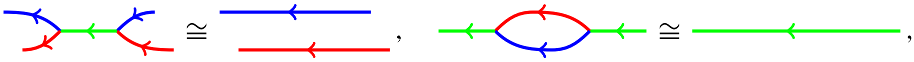
Definition 4.16 Let W be a colored web in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$. W is called *boundary-split* if it is of the form $L \otimes W' \otimes R$ for some colored web $W'\colon \boldsymbol{o}_1 \to \boldsymbol{o}_2$ in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ and for $L\colon \boldsymbol{o}_2 \to \bigsqcup_{\lambda} \boldsymbol{o}_{2,\lambda}$ and $R\colon \bigsqcup_{\lambda} \boldsymbol{o}_{1,\lambda} \to \boldsymbol{o}_1$ as in Definition 4.12. $L \otimes W' \otimes R$ is then called the *boundary-split web associated to* W'.

Remark 4.17 Lemma 4.15 shows that every web W' in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ is weakly equivalent to its associated boundary-split web $L \otimes W' \otimes R$.

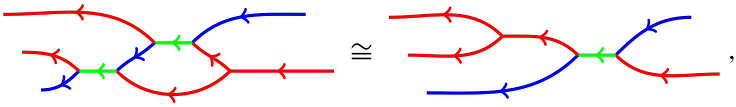
Our goal is now to show that a boundary-split web W is isomorphic to its associated split web $\bigsqcup_{\lambda} W_{\lambda}$. Unless stated otherwise, we use red and blue colors to denote colorings of web edges with disjoint multisubsets of Σ. Green denotes mixed or arbitrary colorings.

Lemma 4.18 *The following isomorphisms hold in* $(N\mathbf{Foam}^{\Sigma})^{\wedge}$*:*

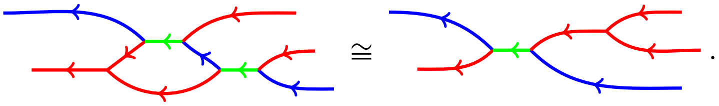
(4-9)



(4-10)



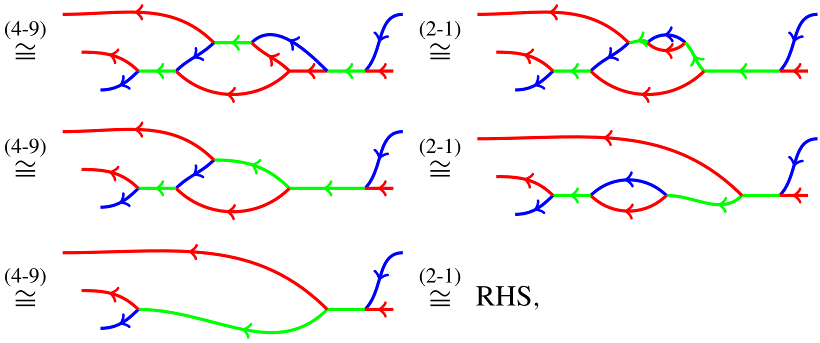
(4-11)



The reflections of these relations across the horizontal axis in the plane also hold.

Proof Equation (4-9) follows from Lemmas 4.4 and 4.6. For (4-10) we have

LHS

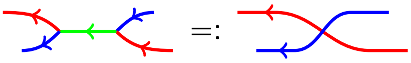


RHS,

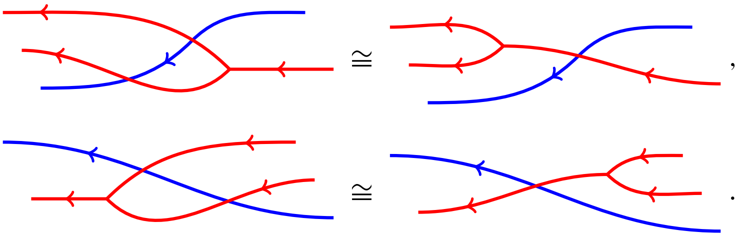
where we have used (the splitter version of) relation (2-1) three times. Equation (4-11) follows similarly. □

Convention 4.19 We define the following shorthand for “crossings” of web edges colored by disjoint multisubsets:

(4-12)



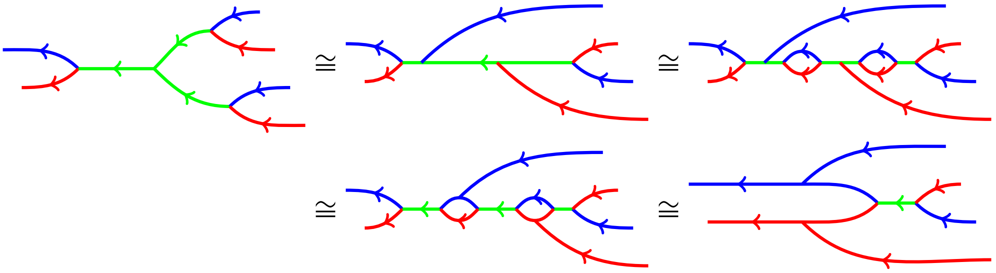
Using this, (4-10) and (4-11) take the form



Definition 4.20 We call a web *semisplit* if it is boundary-split and each edge is either colored by a multisubset containing a single root or a multisubset containing exactly two distinct roots, in which case the edge (green) is required to have a neighborhood as on the left-hand side of (4-12).

Lemma 4.21 *Every boundary-split web $L \otimes W \otimes R$ in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ is isomorphic to a semisplit web W'.*

Proof We inductively split off roots, starting with λ_1. For this we draw in red edges colored with multisubsets of the single root λ_1, in blue edges colored with multisubsets not containing λ_1, and in green edges colored by mixed multisubsets. By (4-9) we can open a red–blue digon in every green edge of $L \otimes W \otimes R$ and get an isomorphic web. Next we replace all vertices that are adjacent to at least two green edges with isomorphic webs that only contain green edges of type (4-12), eg for an all-green merger web:



color	roots occurring in multisubset (with some multiplicity)
red	only λ_i
orange	exactly one λ_k with $k < i$
magenta	exactly two distinct λ_k with $k \leq i$
blue	some roots λ_k with $k > i$
green	λ_i and at least one λ_k with $k > i$
cyan	exactly one λ_k with $k < i$ and at least one with $k > i$
black	λ_i, exactly one λ_k with $k < i$ and at least one with $k > i$

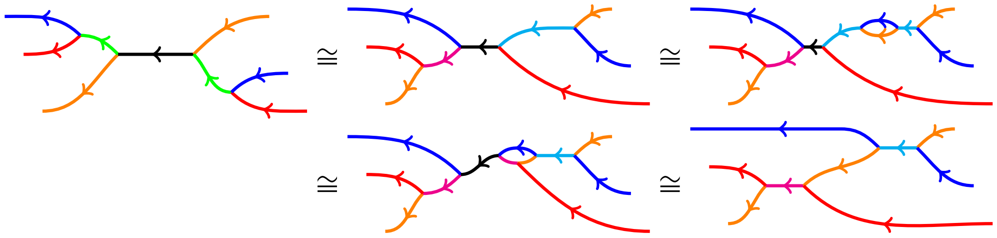
Table 1

The case of an all-green splitter web is completely analogous, and vertices with only two adjacent green edges are even easier to split. The local replacements of green vertices as above patch together to give an isomorphism to a web in which green edges are flanked by red–blue mergers and splitters in the crossing configuration from (4-12).

For the induction step $i-1 \mapsto i$ we use the coloring on edges given in Table 1

We can assume that only these colorings are present. Moreover, orange strands can interact with {other orange, red, blue, green} strands only in crossing configurations around {magenta, magenta, cyan, black} edges, respectively. Furthermore, such crossing configurations are the only occurrences of magenta, cyan and black edges.

The goal for the induction step is to split red edges off green and black edges. As before we introduce red–blue digons in every green edge and locally replace green vertices. Every remaining green edge is in red–blue crossing configuration or bounds red–blue on one side and orange–black on the other side (and every black bounds orange-green on both sides). We get rid of all black edges by splitting off their red component:

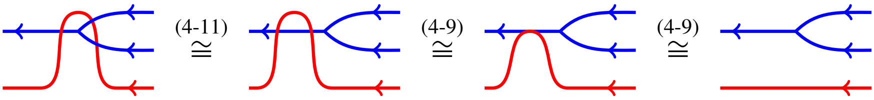


Note that now red and orange strands interact with each other and with strands that contain higher-index roots (blue) only in crossing configurations (around magenta, green and cyan edges), as required in the induction step. For the next step old {blue, green, cyan} edges become {green, black, black} or {blue, cyan, cyan} depending on whether they contain λ_{i+1} or not. Orange stays orange, red becomes orange, and magenta stays magenta. This colored web satisfies the induction hypothesis for the next step. After repeating this process for each root, it terminates in a semisplit web W'. □

Proposition 4.22 *Every boundary-split web $L \otimes W \otimes R$ is isomorphic to its associated split web $\bigsqcup_\lambda W_\lambda$.*

Proof The proof proceeds in two steps; first we use Lemma 4.21 to find an isomorphism from $L \otimes W \otimes R$ to a semisplit web W'. Clearly W, $L \otimes W \otimes R$ and W' have equal associated split webs. It remains to completely separate the λ_i–components in W'. Again we proceed by induction and start by peeling off the λ_1–component W'_{λ_1}. For this, consider a web-isotopy $t \mapsto W'_{\lambda_1}(t)$ for $t \in [0,1]$, ie an ambient isotopy of W'_{λ_1} in the plane which preserves the left-directedness of web edges and which moves W'_{λ_1} off the rest, $W' \setminus W'_{\lambda_1}$. If we superimpose $W'_{\lambda_1}(t)$ and $W' \setminus W'_{\lambda_1}$ we get a homotopy $t \mapsto W'(t)$ of graphs of valence ≤ 6. If the original web-isotopy is generic, the graphs $W'(t)$ actually are of valence ≤ 5 and there are only finitely many t for which the valence is 5 — these correspond to the moves in (4-10) and (4-11). In general, 4–valent vertices in $W(t)$ should be understood as composition of a merge- and a split-3–valent vertex, either in crossing configuration as in (4-12), or splittable as in (4-9). Thus, $t \mapsto W'(t)$ is a web-isotopy except in finitely many points t where the number and valence of vertices changes locally. It is not hard to see that the possible local changes are exactly the ones from Lemma 4.18 and hence can be realized by isomorphism foams.

For example, the following illustrates how to move a single red web edge across a blue vertex:



A composition of the appropriate local isomorphism foams, thus, splits off W'_{λ_1} from W'. One then proceeds to split off, in exactly the same way, W'_{λ_2} and so forth up to $W'_{\lambda_{l-1}}$. The result then follows since $W'_{\lambda_i} = W_{\lambda_i}$ for $1 \leq i \leq l$. □

Remark 4.23 Proposition 4.22 and Lemma 4.15 together show that every web W in $(N\mathbf{Foam}^\Sigma)^\wedge$ is weakly equivalent to its associated split web $\bigsqcup_\lambda W_\lambda$.

4.3 A web splitting functor

We now extend Proposition 4.22 to the 2–categorical level. Ideally, we would like a 2–endofunctor of $(N\mathbf{Foam}^\Sigma)^\wedge$ which fully splits foams into pieces carrying colorings of only one root, but the naïve splitting procedure does not give a well-defined 2–functor (a counterexample can be constructed which sends the left- and right-hand sides of (4-3) to unequal multiples of each other). Instead, we take a more direct approach and define a family of functors between Hom–categories in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ using compositions with explicit webs and foams, which will suffice to split the complex assigned to a tangle.

We begin by fixing, for each colored web, an isomorphism between its associated boundary split and split webs. Precisely, let $W\colon \boldsymbol{o}_1 \to \boldsymbol{o}_2$ be a colored web in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ and suppose that $L\colon \boldsymbol{o}_2 \to \bigsqcup_\lambda \boldsymbol{o}_{2,\lambda}$ and $R\colon \bigsqcup_\lambda \boldsymbol{o}_{1,\lambda} \to \boldsymbol{o}_1$ are the webs given in Definition 4.12. Proposition 4.22 guarantees that there is an isomorphism $T_W\colon L \otimes W \otimes R \to \bigsqcup_\lambda W_\lambda$, so fix one and denote its inverse by B_W. We have some freedom in choosing T_W, and in Section 5.2 we will specify a convenient choice for webs that arise as resolutions of tangle diagrams.

For the next definition, suppose $F\colon W_1 \to W_2$ is a foam between colored webs $W_1, W_2\colon \boldsymbol{o}_1 \to \boldsymbol{o}_2$ in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ with identical incident conditions on the boundary sequences $\boldsymbol{o}_1$ and $\boldsymbol{o}_2$, respectively. Further, consider the webs L and R and the isomorphism foams T_{W_2} and B_{W_1} described above.

Definition 4.24 Let $\phi := \phi_2 \circ \phi_1$ be the composition of

$$\begin{aligned}\phi_1\colon \operatorname{Hom}(W_1, W_2) &\to \operatorname{Hom}(L \otimes W_1 \otimes R, L \otimes W_2 \otimes R),\\ F &\mapsto \mathrm{id}_L \otimes F \otimes \mathrm{id}_R,\end{aligned}$$

and

$$\begin{aligned}\phi_2\colon \operatorname{Hom}(L \otimes W_1 \otimes R, L \otimes W_2 \otimes R) &\to \operatorname{Hom}\Big(\bigsqcup_\lambda W_{1,\lambda}, \bigsqcup_\lambda W_{2,\lambda}\Big),\\ F &\mapsto T_{W_2} \circ F \circ B_{W_1}.\end{aligned}$$

Proposition 4.25 *Fix objects $\boldsymbol{o}_1, \boldsymbol{o}_2 \in (N\mathbf{Foam}^{\Sigma})^{\wedge}$. Then the maps*

$$\phi\colon \operatorname{Hom}(W_1, W_2) \to \operatorname{Hom}\Big(\bigsqcup_\lambda W_{1,\lambda}, \bigsqcup_\lambda W_{2,\lambda}\Big)$$

for colored webs $W_1, W_2\colon \boldsymbol{o}_1 \to \boldsymbol{o}_2$ with identical incident conditions on $\boldsymbol{o}_1$ and $\boldsymbol{o}_2$, respectively, are vector space isomorphisms that respect the composition $\circ$ of foams.

Proof It is clear that the maps ϕ_1 respect composition of foams and for ϕ_2 it follows from the definition of B_W as the inverse of T_W.

Next, note that ϕ_2 is clearly a vector space isomorphism, since it is pre- and post-composition with isomorphism foams. To see that ϕ_1 is as well, let L^{-1} and R^{-1} be the webs obtained by reflecting L and R horizontally. We have isomorphism foams

$\phi_L\colon L^{-1}\otimes L\to 1_{o_2}$ and $\phi_R\colon R\otimes R^{-1}\to 1_{o_1}$, and an inverse for ϕ_1 is then given by $\psi\colon G\mapsto(\phi_L\otimes \mathrm{id}_{W_2}\otimes\phi_R)\circ(\mathrm{id}_{L^{-1}}\otimes G\otimes \mathrm{id}_{R^{-1}})\circ(\phi_L^{-1}\otimes\mathrm{id}_{W_1}\otimes\phi_R^{-1})$. □

Finally, suppose that $\boldsymbol{A}$ and $\boldsymbol{B}$ are incidence conditions for objects $\boldsymbol{o}_1$ and $\boldsymbol{o}_2$ in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$, respectively. By expressing each facet incident upon a left or right boundary as a sum over colorings, we see that the Hom–categories in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ split into direct sums

$$\mathrm{Hom}(\boldsymbol{o}_1,\boldsymbol{o}_2)\cong\bigoplus_{\boldsymbol{A},\boldsymbol{B}}\mathrm{Hom}^{\boldsymbol{A}\to\boldsymbol{B}}(\boldsymbol{o}_1,\boldsymbol{o}_2),\tag{4-13}$$

where the sum is over all incidence conditions $\boldsymbol{A}$ and $\boldsymbol{B}$ and $\mathrm{Hom}^{\boldsymbol{A}\to\boldsymbol{B}}(\boldsymbol{o}_1,\boldsymbol{o}_2)$ denotes the full subcategory of $\mathrm{Hom}(\boldsymbol{o}_1,\boldsymbol{o}_2)$ generated by webs that are colored with the multisets prescribed by $\boldsymbol{A}$ and $\boldsymbol{B}$ on the right and left boundary edges, respectively.

Definition 4.26 Let W be a web in $\mathrm{Hom}^{\boldsymbol{A}\to\boldsymbol{B}}(\boldsymbol{o}_1,\boldsymbol{o}_2)$ and suppose that $\bigsqcup_\lambda \boldsymbol{o}_{1,\lambda}$ and $\bigsqcup_\lambda \boldsymbol{o}_{2,\lambda}$ are the objects given in (4-8); then we also denote by ϕ the functor

$$\mathrm{Hom}^{\boldsymbol{A}\to\boldsymbol{B}}(\boldsymbol{o}_1,\boldsymbol{o}_2)\to\mathrm{Hom}\Big(\bigsqcup_\lambda \boldsymbol{o}_{1,\lambda},\bigsqcup_\lambda \boldsymbol{o}_{2,\lambda}\Big)$$

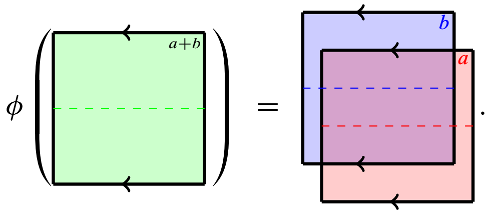
defined on webs W and foams F in $\mathrm{Hom}^{\boldsymbol{A}\to\boldsymbol{B}}(\boldsymbol{o}_1,\boldsymbol{o}_2)$ by

$$\phi(W):=\bigsqcup_\lambda W_\lambda\quad\text{and}\quad\phi(F):=T_{W_2}\circ(\mathrm{id}_L\otimes F\otimes\mathrm{id}_R)\circ B_{W_1}.$$

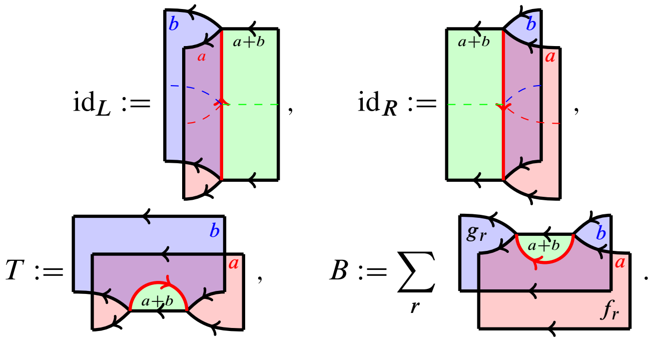
We call the functors ϕ *web splitting functors*. Note that their definition depends on our choice of isomorphism foam T_W for every colored web W in $\mathrm{Hom}^{\boldsymbol{A}\to\boldsymbol{B}}(\boldsymbol{o}_1,\boldsymbol{o}_2)$. In Section 5.2 we show that with a suitable choice of T_W the functors ϕ not only split webs, but also certain foams between them. We give a prototypical example of this:

Example 4.27 With coloring conventions as in Example 4.9 we have

(4-14)



Indeed, this follows from Lemma 4.4 using



Remark 4.28 For colored webs W_1 and W_2 in $(N\mathbf{Foam}^\Sigma)^\wedge$ we conjecture that the map

$$\bigotimes_\lambda \mathrm{Hom}(W_{1,\lambda}, W_{2,\lambda}) \to \mathrm{Hom}\Big(\bigsqcup_\lambda W_{1,\lambda}, \bigsqcup_\lambda W_{2,\lambda}\Big),$$

given by placing foams colored by individual roots side by side, is an isomorphism of vector spaces. In particular, this would mean that every foam between split webs can be split into noninteracting colored components, possibly with additional decorations. One could prove such a result by extending Proposition 4.22 to the 2–categorical level, finding local foam moves which move the λ_i–colored component away from everything colored by λ_j for $j > i$.

Having done this, we could compose the web splitting functor ϕ with the inverse of the above isomorphism to produce an honest foam splitting functor. However, in order to extend this functor to a 2–endofunctor of $(N\mathbf{Foam}^\Sigma)^\wedge$, we must verify compatibility with horizontal composition, which will depend on our choice of the T_W. Additionally, decorations will arise while pulling the foams apart which are difficult to control. In Section 5.2 we carry out this analysis in the limited case of foams arising as differentials in the complex assigned to a tangle, and use this to prove Theorem 1.1.

4.4 A 2–representation of $N\mathbf{Foam}^\Sigma$

In this section, we prove that the deformed $\mathfrak{sl}_N$ foam 2–category $N\mathbf{Foam}^\Sigma$ is sufficiently nondegenerate, by constructing a 2–representation onto a version of Wu's deformed matrix factorizations [42]. Indeed, let **HMF** denote the 2–category given as follows:

- objects are pairs (R, w) where R is a $\mathbb{C}$–algebra and $w \in R$,
- 1–morphisms $(R, w) \to (S, v)$ are matrix factorizations X over $R \otimes_{\mathbb{C}} S$ with potential $v - w$, and
- 2–morphisms $X \to Y$ are morphisms in the homotopy category of matrix factorizations.

We will assume the basics concerning matrix factorizations, which can eg be found in Khovanov and Rozansky [22]; see Carqueville and Murfet [5] for details about the 2–category of matrix factorizations.

Our result is the following:

Theorem 4.29 *There is a 2–representation from the deformed foam 2–category $N\mathbf{Foam}^{\Sigma}$ to the 2–category of matrix factorizations. Moreover, this 2–representation assigns to a web in $N\mathbf{Foam}^{\Sigma}$ the same matrix factorization as in Wu's construction of deformed link homology.*

Of course, it suffices to assign pairs (R, w) to sequences, the same matrix factorizations as in [42] to generating webs, and morphisms of matrix factorizations to generating foams, and then check that the images of the foam relations hold in $\mathbf{HMF}$. However, we can simplify this check using an argument similar to that in [36]. Indeed, there it is shown that the undeformed foam category $N\mathbf{Foam}$ is equivalent to a certain 2–subcategory of the quotient of $\dot{\mathcal{U}}_Q(\mathfrak{gl}_\infty)$ by the N–bounded weights. Since the 2–category of matrix factorizations is idempotent complete, it suffices to construct a 2–representation of $\mathcal{U}_Q(\mathfrak{gl}_\infty)$ sending non-N–bounded weights to zero and satisfying (the preimage of) the additional foam relation in $N\mathbf{Foam}^{\Sigma}$, which then induces a 2–functor from $N\mathbf{Foam}^{\Sigma}$.

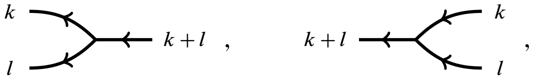
Practically speaking, this shows that we need only check the foam relations coming from relations in $\mathcal{U}_Q(\mathfrak{gl}_\infty)$ and not those coming from the thick calculus in $\dot{\mathcal{U}}_Q(\mathfrak{gl}_\infty)$, which are used to split certain idempotent foams in $N\mathbf{Foam}$. This simplifies the number of relations needed to be checked (more details below).

We hence begin by following Wu, assigning a pair (R, w) to an object $(a_1, \dots, a_k)$ in $N\mathbf{Foam}^{\Sigma}$. We set $R = \mathrm{Sym}(\mathbb{X}_1 \,|\cdots|\, \mathbb{X}_k)$, the $\mathbb{C}$–algebra of partially symmetric functions in the alphabets $\mathbb{X}_1, \dots, \mathbb{X}_k$, where $\mathbb{X}_i$ consists of a_i variables. We let $w = Q(\mathbb{X}_1 \cup \cdots \cup \mathbb{X}_k)$, where $Q'(X) = (N+1)P(X)$ with $P(X)$ as in (3-3), $Q(0) = 0$, and for a polynomial $T(X) = \sum_{i=0}^{k} c_i X^i \in \mathbb{C}[X]$ we set $T(\mathbb{X}) = \sum_{i=0}^{k} c_i p_i(\mathbb{X})$, where $p_i(\mathbb{X})$ denotes the i^{th} power sum symmetric polynomial in the alphabet $\mathbb{X}$.

Given sequences $\boldsymbol{a}$ and $\boldsymbol{b}$ of elements of a $\mathbb{C}$–algebra, we will follow Khovanov and Rozansky [22] and denote by $\{\boldsymbol{a}, \boldsymbol{b}\}$ the Koszul matrix factorization they determine. We then assign the Koszul matrix factorizations

$$\{(U_i)_{i=1}^{k+l}, (e_i(\mathbb{W} \cup \mathbb{X}) - e_i(\mathbb{Y}))_{i=1}^{k+l}\}, \quad \{(-U_i)_{i=1}^{k+l}, (e_i(\mathbb{Y}) - e_i(\mathbb{W} \cup \mathbb{X}))_{i=1}^{k+l}\} \tag{4-15}$$

over $\mathrm{Sym}(\mathbb{W}|\mathbb{X}|\mathbb{Y})$ with potentials $Q(\mathbb{W}\cup\mathbb{X})-Q(\mathbb{Y})$ and $Q(\mathbb{Y})-Q(\mathbb{W}\cup\mathbb{X})$ (respectively) to the generating webs



where $|\mathbb{W}|=k$, $|\mathbb{X}|=l$ and $|\mathbb{Y}|=k+l$. Here the polynomials U_i are chosen so that

$$Q(\mathbb{W}\cup\mathbb{X})-Q(\mathbb{Y})=\sum_{i=1}^{k+l}\bigl(e_i(\mathbb{W}\cup\mathbb{X})-e_i(\mathbb{Y})\bigr)U_i.$$

Note that these are the same matrix factorizations that Wu assigns to trivalent vertices.

We now assign a morphism of matrix factorizations to each generating foam. To do so, we utilize the concept of stabilization of linear factorizations. Recall from Carqueville and Murfet [4] that a linear factorization L over a ring R with potential $w\in R$ is a $\mathbb{Z}/2\mathbb{Z}$–graded R–module, equipped with an odd degree differential d satisfying $d^2=w\,\mathrm{id}$. Informally, a linear factorization is a matrix factorization where we loosen the requirement that the R–module be free. In particular, matrix factorizations give examples of linear factorizations.

Following [4], define the *stabilization* of a linear factorization L over (R,w) to be a finite-rank matrix factorization M_L over (R,w) together with a morphism of linear factorizations $\pi\colon M_L\to L$ inducing a quasi-isomorphism of $\mathbb{Z}/2\mathbb{Z}$–graded complexes

$$\mathrm{Hom}_R(K,M_L)\xrightarrow{\pi\circ}\mathrm{Hom}_R(K,L) \tag{4-16}$$

for any finite-rank matrix factorization K over (R,w).

We use stabilizations as follows: suppose that we are given linear factorizations L_1 and L_2 with corresponding stabilizations M_{L_i}; then, the diagram

$$\begin{tikzcd} M_{L_1} \arrow[r, "\pi_1"] & L_1 \arrow[d] \\ M_{L_2} \arrow[r, "\pi_2"] & L_2 \end{tikzcd} \tag{4-17}$$

induces a map on homology

$$\mathrm{H}_*(\mathrm{Hom}_R(L_1,L_2))\to\mathrm{H}_*(\mathrm{Hom}_R(M_{L_1},L_2))\to\mathrm{H}_*(\mathrm{Hom}_R(M_{L_1},M_{L_2})).$$

Since H_0 gives the morphisms in the homotopy category of matrix (or linear) factorizations, we can construct a morphism $\mathrm{stab}(\varphi)\in\mathrm{Hom}_{\mathbf{HMF}}(M_{L_1},M_{L_2})$ from a morphism

$\varphi\colon L_1 \to L_2$ which is the unique (up to homotopy) morphism such that the diagram

$$\begin{tikzcd} M_{L_1} \arrow[r, "\pi_1"] \arrow[d, "\mathrm{stab}(\varphi)"'] & L_1 \arrow[d, "\varphi"] \\ M_{L_2} \arrow[r, "\pi_2"] & L_2 \end{tikzcd}$$

commutes. We will use this to define the morphisms of matrix factorizations assigned to generating foams, and to check that the foam relations are satisfied.

In doing so, we utilize facts about the stabilization of Koszul matrix factorizations. Let $\{\boldsymbol{a}, \boldsymbol{b}\}$ be a Koszul factorization over a $\mathbb{C}$–algebra R, then there exists a morphism of linear factorizations $\{\boldsymbol{a}, \boldsymbol{b}\} \to R/(\boldsymbol{b})$, where the latter is viewed as a linear factorization concentrated in degree zero.

Proposition 4.30 [4, Corollary D.3] *If $\boldsymbol{b}$ is a regular sequence in R, then the map $\{\boldsymbol{a}, \boldsymbol{b}\} \to R/(\boldsymbol{b})$ is a stabilization.*

Convention 4.31 In the following we use a large number of quotient rings of the form

$$\frac{\mathrm{Sym}(\mathbb{X}_1|\cdots|\mathbb{X}_a|\mathbb{X}_{a+1}|\cdots|\mathbb{X}_{a+b})}{\langle e_i(\mathbb{X}_1\cup\cdots\cup\mathbb{X}_a) - e_i(\mathbb{X}_{a+1}\cup\cdots\cup\mathbb{X}_{a+b}) \mid i>0\rangle}$$

where $\mathrm{Sym}(\mathbb{X}_1|\cdots|\mathbb{X}_a|\mathbb{X}_{a+1}|\cdots|\mathbb{X}_{a+b})$ denotes the subring of $\mathbb{C}[\mathbb{X}_1\cup\cdots\cup\mathbb{X}_{a+b}]$ of polynomials symmetric in each of the alphabets $\mathbb{X}_1,\dots,\mathbb{X}_{a+b}$ separately. Since the quotient has the effect of identifying symmetric polynomials in the alphabets $\mathbb{X}_1\cup\cdots\cup\mathbb{X}_a$ and $\mathbb{X}_{a+1}\cup\cdots\cup\mathbb{X}_{a+b}$, we use the shorthand

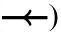
$$\frac{\mathrm{Sym}(\mathbb{X}_1|\cdots|\mathbb{X}_a|\mathbb{X}_{a+1}|\cdots|\mathbb{X}_{a+b})}{\langle \mathbb{X}_1\cup\cdots\cup\mathbb{X}_a = \mathbb{X}_{a+1}\cup\cdots\cup\mathbb{X}_{a+b}\rangle}$$

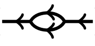
for such a quotient ring. We further abbreviate by writing m for a 1–element alphabet $\mathbb{X}=\{m\}$ in this notation.

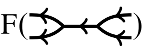
Proposition 4.30 implies that the matrix factorizations appearing in (4-15) are stabilizations of the linear factorizations

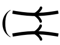
$$\mathrm{Sym}(\mathbb{W}|\mathbb{X}|\mathbb{Y})/\langle \mathbb{W}\cup\mathbb{X}=\mathbb{Y}\rangle \quad\text{and}\quad \mathrm{Sym}(\mathbb{W}|\mathbb{X}|\mathbb{Y})/\langle \mathbb{Y}=\mathbb{W}\cup\mathbb{X}\rangle.$$

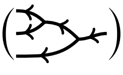
Moreover, denoting the matrix factorization associated to a web W by $\mathrm{MF}(W)$, we have that, for the maps

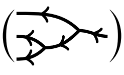
MF() $\xrightarrow{\pi} \dfrac{\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{Y})}{\langle \mathbb{V} = \mathbb{Y}\rangle},$

(4-18) MF() $\xrightarrow{\pi} \left(\dfrac{\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{L}\,|\,\mathbb{M})}{\langle \mathbb{V} = \mathbb{L}\cup\mathbb{M}\rangle}\right) \otimes_{\mathrm{Sym}(\mathbb{L}\,|\,\mathbb{M})} \left(\dfrac{\mathrm{Sym}(\mathbb{L}\,|\,\mathbb{M}\,|\,\mathbb{Y})}{\langle \mathbb{L}\cup\mathbb{M} = \mathbb{Y}\rangle}\right),$

MF() $\xrightarrow{\pi} \left(\dfrac{\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{W}\,|\,\mathbb{L})}{\langle \mathbb{V}\cup\mathbb{W} = \mathbb{L}\rangle}\right) \otimes_{\mathrm{Sym}(\mathbb{L})} \left(\dfrac{\mathrm{Sym}(\mathbb{L}\,|\,\mathbb{X}\,|\,\mathbb{Y})}{\langle \mathbb{L} = \mathbb{X}\cup\mathbb{Y}\rangle}\right),$

(4-19) MF() $\xrightarrow{\pi} \dfrac{\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{X})}{\langle \mathbb{V} = \mathbb{X}\rangle} \otimes_{\mathbb{C}} \dfrac{\mathrm{Sym}(\mathbb{W}\,|\,\mathbb{Y})}{\langle \mathbb{W} = \mathbb{Y}\rangle},$

MF() $\xrightarrow{\pi} \left(\dfrac{\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{W}\,|\,\mathbb{L})}{\langle \mathbb{V}\cup\mathbb{W} = \mathbb{L}\rangle}\right) \otimes_{\mathrm{Sym}(\mathbb{L})} \left(\dfrac{\mathrm{Sym}(\mathbb{L}\,|\,\mathbb{X}\,|\,\mathbb{Y})}{\langle \mathbb{L}\cup\mathbb{X} = \mathbb{Y}\rangle}\right),$

MF() $\xrightarrow{\pi} \left(\dfrac{\mathrm{Sym}(\mathbb{W}\,|\,\mathbb{X}\,|\,\mathbb{M})}{\langle \mathbb{W}\cup\mathbb{X} = \mathbb{M}\rangle}\right) \otimes_{\mathrm{Sym}(\mathbb{M})} \left(\dfrac{\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{M}\,|\,\mathbb{Y})}{\langle \mathbb{V}\cup\mathbb{M} = \mathbb{Y}\rangle}\right),$

the matrix factorizations are (homotopy equivalent to) stabilizations of the indicated linear factorizations. The fact that these maps are stabilizations follows in each case, except for the digon web in the second line, since the matrix factorizations are homotopy equivalent to Koszul factorizations, and the indicated linear factorization is isomorphic to the corresponding linear factorization which the Koszul factorization stabilizes.

The matrix factorization assigned to the digon web is a tensor product of Koszul factorizations, and we must slightly generalize Proposition 4.30 to show that it stabilizes the tensor product of the corresponding linear factorizations. Recall from [4, Proposition D.1] that Proposition 4.30 can be proven as follows. One first considers the Koszul complex $\{\boldsymbol{b}\}$ over R given by the regular sequence $\boldsymbol{b}$. There exists a homotopy equivalence (over $\mathbb{C}$) between $\{\boldsymbol{b}\}$ and $R/(\boldsymbol{b})$ which specifies a deformation retract datum. Tensoring with the finite-rank matrix factorization $K^\vee$ (the dual of the matrix factorization K) and applying perturbation gives a deformation retract datum over $\mathbb{C}$ between $K^\vee \otimes R/(\boldsymbol{b})$ and $K^\vee \otimes \{\boldsymbol{a},\boldsymbol{b}\}$ which gives the quasi-isomorphism in (4-16). Here we utilize the isomorphism of matrix factorizations $K^\vee \otimes_R M \cong \mathrm{Hom}_R(K, M)$.

This same method (which is adapted from the results in Dyckerhoff and Murfet [11]) shows that the stabilization result for the digon web follows provided the tensor product of Koszul complexes associated to the web only has homology in degree zero, and which equals the corresponding tensor product of linear factorizations. We hence consider the Koszul complexes $C_1 = \{e_i(\mathbb{V}) - e_i(\mathbb{L}\cup\mathbb{M})\}$ and $C_2 = \{e_i(\mathbb{L}\cup\mathbb{M}) - e_i(\mathbb{Y})\}$ over the rings $\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{L}\,|\,\mathbb{M})$ and $\mathrm{Sym}(\mathbb{L}\,|\,\mathbb{M}\,|\,\mathbb{Y})$, respectively. Let $S = \mathrm{Sym}(\mathbb{L}\,|\,\mathbb{M})$, then the homology of $C_1 \otimes_S C_2$ is computed using the Künneth spectral sequence to be

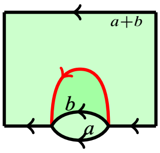
$$\mathrm{H}_i(C_1 \otimes_S C_2) \cong \left(\mathrm{H}_i(C_1) \otimes_S \frac{\mathrm{Sym}(\mathbb{L}\,|\,\mathbb{M}\,|\,\mathbb{Y})}{\langle \mathbb{L}\cup\mathbb{M} = \mathbb{Y}\rangle}\right) \oplus \mathrm{Tor}_1^S\left(\mathrm{H}_{i-1}(C_1), \frac{\mathrm{Sym}(\mathbb{L}\,|\,\mathbb{M}\,|\,\mathbb{Y})}{\langle \mathbb{L}\cup\mathbb{M} = \mathbb{Y}\rangle}\right),$$

which is only nonzero when $i = 0$ (since $\mathrm{Sym}(\mathbb{V}|\mathbb{L}|\mathbb{M})/\langle\mathbb{V} = \mathbb{L}\cup\mathbb{M}\rangle$ is a free S–module) in which case it equals

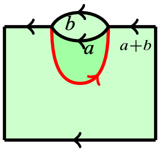
$$\left(\frac{\mathrm{Sym}(\mathbb{V}|\mathbb{L}|\mathbb{M})}{\langle\mathbb{V} = \mathbb{L}\cup\mathbb{M}\rangle}\right)\otimes_S\left(\frac{\mathrm{Sym}(\mathbb{L}|\mathbb{M}|\mathbb{Y})}{\langle\mathbb{L}\cup\mathbb{M} = \mathbb{Y}\rangle}\right),$$

as desired.

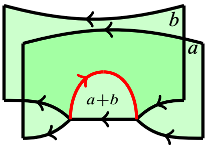
We now use (4-17) to assign a morphism of matrix factorizations to each generating foam, noting that the domain web of each generator is mapped to a matrix factorization which is homotopy equivalent to a finite-rank matrix factorization. We send



$$\mapsto\left(\overline{\pi^{\mathbb{L}}_{\lambda}}\otimes\overline{1}\mapsto\begin{cases}\overline{1} & \text{if } \lambda = b^a,\\ 0 & \text{all other } \lambda\in P(a,b),\end{cases}\right),$$

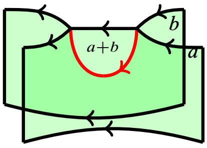


$$\mapsto(\overline{1}\mapsto\overline{1}\otimes\overline{1}),$$

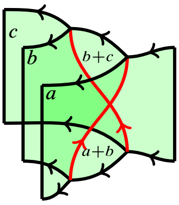


$$\mapsto(\overline{1}\otimes\overline{1}\mapsto\overline{1}\otimes\overline{1}),$$

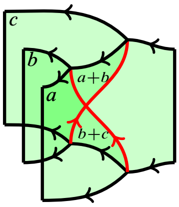
(4-20)



$$\mapsto\left(\overline{1}\otimes\overline{1}\mapsto\sum_{\alpha\in P(a,b)}(-1)^{|\widehat{\alpha}|}\overline{\pi^{\mathbb{V}}_{\widehat{\alpha}}}\otimes\overline{\pi^{\mathbb{Y}}_{\alpha}}\right),$$



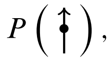
$$\mapsto(\overline{1}\otimes\overline{1}\mapsto\overline{1}\otimes\overline{1}),$$



$$\mapsto(\overline{1}\otimes\overline{1}\mapsto\overline{1}\otimes\overline{1}),$$

where in each case the map on the right-hand side describes a morphism between the linear factorizations from (4-19) corresponding to the top and bottom webs and $\overline{f}$ denotes the equivalence class of f in the quotient. In these formulae, $\pi^{\mathbb{W}}_{\lambda}$ denotes the Schur polynomial in the alphabet $\mathbb{W}$ corresponding to the partition λ, and $b^a = (b,\dots,b)$, the partition of ab given by a sequence of b's of length a. We can now proceed with the proof of Theorem 4.29.

Proof It suffices to show that the foam relations hold in **HMF**. As we mentioned above, rather than check them all by hand, we will instead adopt a method of proof from [36]. By an argument similar to that in Section 4 of that paper, it suffices to construct a family of 2–functors $\Phi_m\colon \mathcal{U}_Q(\mathfrak{gl}_m) \to \mathbf{HMF}$ which kill non-N–bounded weights and the 2–morphism



and so that the triangles

$$\begin{tikzcd}
\mathcal{U}_Q(\mathfrak{gl}_m) \arrow[r] \arrow[dr, "\Gamma_m"'] & \mathcal{U}_Q(\mathfrak{gl}_{m+1}) \arrow[d, "\Gamma_{m+1}"] \\
 & \mathbf{HMF}
\end{tikzcd}$$

commute. From the definition of the foamation 2–functor in [36] and our above assignments to webs and foams, it is clear how such 2–functors should be defined. To see that they are well-defined, we must check that all relations in $\mathcal{U}_Q(\mathfrak{gl}_m)$ are satisfied. This in turn implies that we need only check the foam relations which are the analogs of the relations in $\mathcal{U}_Q(\mathfrak{gl}_m)$. Since (4-20) implies that the image in **HMF** of the "Matveev–Piergallini (M–P) foam relations" from (2-1) are satisfied, things simplify even more, and we finally deduce that we need only check a subset of the general foam relations, which we verify below.

To do so, we will again employ stabilization. The matrix factorizations through which the (images of the) foam relations factor are all given as tensor products of Koszul factorizations assigned to trivalent webs, and we can consider the corresponding tensor product of the linear factorizations they stabilize. This gives a diagram

$$\begin{tikzcd}
\bigotimes_i M_{L_{i,1}} \arrow[r, "\pi_1"] \arrow[d, "\mathrm{stab}(\varphi_1)"'] & \bigotimes_i L_{i,1} \arrow[d, "\varphi_1"] \\
\bigotimes_j M_{L_{j,2}} \arrow[r, "\pi_2"] \arrow[d, "\mathrm{stab}(\varphi_2)"'] & \bigotimes_j L_{j,2} \arrow[d, "\varphi_2"] \\
\vdots \arrow[d, "\mathrm{stab}(\varphi_{l-1})"'] & \vdots \arrow[d, "\varphi_{l-1}"] \\
\bigotimes_k M_{L_{k,l}} \arrow[r, "\pi_k"] & \bigotimes_k L_{k,l}
\end{tikzcd}$$

which commutes up to homotopy. Each side of a foam relation gives rise to such a diagram, and the morphism of matrix factorizations is uniquely determined by the morphism of linear factorizations, provided the matrix factorizations assigned to the bottom webs are homotopic to ones which are finite-rank, and provided that the matrix factorizations assigned to the top webs (ie the bottom left in the above diagram) are homotopic to ones which stabilize the corresponding tensor product of linear factorizations.

The finite-rank condition for the bottom webs follows similarly to results of Wu [43] in the undeformed case. To see that the matrix factorizations corresponding to the top webs (are homotopic to ones which) stabilize the corresponding linear factorizations, we note that we've already shown this for [36, Equations (3.9)–(3.12)]. For the remainder of the relations we argue as for the digon web above. It again suffices to show that the tensor product of Koszul complexes associated to the top web has homology only in degree zero, and equal to the corresponding tensor product of linear factorizations. In each case, this follows from (possibly repeated) use of the Künneth spectral sequence, and the fact that $\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{W}\,|\,\mathbb{X})/\langle \mathbb{V}\cup\mathbb{W}=\mathbb{X}\rangle$ is a free module, over both $\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{W})$ and $\mathrm{Sym}(\mathbb{X})$. Note that this is essentially a version of results of Becker [3, Theorem 2] and Webster [40, Theorem 2.5] for deformed potentials.

We now check the requisite foam relations (with numbering and notation from [36] for the remainder of this section) by confirming that the corresponding maps of linear factorizations agree.

[36, Equation (3.9)] By [36, Remark 3.2], this only needs to be checked when $\pi_\gamma = e_s$, and then follows since multiplication by $e_i(\mathbb{X})$ on $\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{W}\,|\,\mathbb{X})/\langle \mathbb{V}\cup\mathbb{W}=\mathbb{X}\rangle$ is equal to multiplication by $e_i(\mathbb{V}\cup\mathbb{W})$.

[36, Equation (3.10), first relation] Again by [36, Remark 3.2], it suffices to check the case when $\pi_\alpha = 1$. Let $\mathbb{V}$ and $\mathbb{Y}$ be alphabets with $k+1$ variables. It suffices to show that the morphism corresponding to the right-hand side is the identity, hence we compute

$$\begin{tikzcd}
\frac{\mathrm{S}(\mathbb{V}|\mathbb{Y})}{\langle\mathbb{V}=\mathbb{Y}\rangle} \arrow[r] & \frac{\mathrm{S}(\mathbb{V}|m|\mathbb{M})}{\langle\mathbb{V}=m\cup\mathbb{M}\rangle}\otimes\frac{\mathrm{S}(m|\mathbb{M}|\mathbb{Y})}{\langle m\cup\mathbb{M}=\mathbb{Y}\rangle} \arrow[r] & \frac{\mathrm{S}(\mathbb{V}|m|\mathbb{M})}{\langle\mathbb{V}=m\cup\mathbb{M}\rangle}\otimes\frac{\mathrm{S}(m|\mathbb{M}|\mathbb{Y})}{\langle m\cup\mathbb{M}=\mathbb{Y}\rangle} \arrow[r] & \frac{\mathrm{S}(\mathbb{V}|\mathbb{Y})}{\langle\mathbb{V}=\mathbb{Y}\rangle}, \\
\overline{1} \arrow[r, mapsto] & \overline{1}\otimes\overline{1} \arrow[r, mapsto] & \overline{m^k}\otimes\overline{1} \arrow[r, mapsto] & \overline{1},
\end{tikzcd}$$

which verifies the relation. (Recall that S stands for Sym here.)

[36, Equation (3.10), second relation] It suffices to check the case when with $a = 1 = b$, by the M–P relation, isotopy, and [36, (3.9)]. We compute the left-hand side:

$$\begin{tikzcd}
\frac{\mathrm{S}(\mathbb{W}|\mathbb{X})}{\langle\mathbb{W}=\mathbb{X}\rangle} \arrow[r] & \frac{\mathrm{S}(\mathbb{W}|m|n)}{\langle\mathbb{W}=\{m,n\}\rangle}\otimes\frac{\mathrm{S}(m|n|\mathbb{X})}{\langle\{m,n\}=\mathbb{X}\rangle} \arrow[r] & \frac{\mathrm{S}(\mathbb{W}|m|n)}{\langle\mathbb{W}=\{m,n\}\rangle}\otimes\frac{\mathrm{S}(m|n|\mathbb{X})}{\langle\{m,n\}=\mathbb{X}\rangle} \arrow[r] & \frac{\mathrm{S}(\mathbb{W}|\mathbb{X})}{\langle\mathbb{W}=\mathbb{X}\rangle}, \\
\overline{1} \arrow[r, mapsto] & \overline{1}\otimes\overline{1} \arrow[r, mapsto] & \overline{f(m)g(n)}\otimes\overline{1}, & \\
 & & \overline{m}\otimes\overline{1} \arrow[r, mapsto] & \overline{1}, \\
 & & \overline{1}\otimes\overline{1} \arrow[r, mapsto] & 0,
\end{tikzcd}$$

while the right-hand side is the negative of the map which is the same as the above, but with the second map given instead by $\overline{1}\otimes\overline{1}\mapsto\overline{g(m)f(n)}\otimes\overline{1}$. Equivalently, this is the negative of the map which is the same as the above, but instead with the third map given by $\overline{n}\otimes\overline{1}\mapsto\overline{1}$ and $\overline{1}\otimes\overline{1}\mapsto 0$. Since $\overline{n}\otimes\overline{1}=\overline{e_1(\mathbb{W})}\otimes\overline{1}-\overline{m}\otimes\overline{1}$ and $\overline{e_1(\mathbb{W})}\otimes\overline{1}\mapsto 0$ under the final map, this confirms the relation.

[36, Equation (3.11)] The $a=1=b$ case of this relation is used to deduce that the image of the 3rd nil-Hecke relation is satisfied. A careful analysis of the proof of [36, Lemma 3.7] shows that the only remaining version of this relation required are those when $a=1$, $b=2$ and $a=2$, $b=1$, which are used to prove the 2nd Reidemeister III-like nil-Hecke relation.

In the $a=1=b$ case, the right-hand side corresponds to the sum of the map

$$\begin{tikzcd}
\frac{\mathrm{Sym}(\mathbb{W}|m|n)}{\langle\mathbb{W}=\{m,n\}\rangle}\otimes\frac{\mathrm{Sym}(m|n|\mathbb{X})}{\langle\{m,n\}=\mathbb{X}\rangle} \arrow[r] & \frac{\mathrm{Sym}(\mathbb{W}|\mathbb{X})}{\langle\mathbb{W}=\mathbb{X}\rangle} \arrow[r] & \frac{\mathrm{Sym}(\mathbb{W}|m|n)}{\langle\mathbb{W}=\{m,n\}\rangle}\otimes\frac{\mathrm{Sym}(m|n|\mathbb{X})}{\langle\{m,n\}=\mathbb{X}\rangle}, \\
\overline{1}\otimes\overline{1} \arrow[r, mapsto] & \overline{1} \arrow[r, mapsto] & \overline{1}\otimes\overline{1}, \\
\overline{m}\otimes\overline{1} \arrow[r, mapsto] & \overline{e_1(\mathbb{W})} \arrow[r, mapsto] & \overline{m+n}\otimes\overline{1},
\end{tikzcd}$$

and the negative of the map

$$\begin{tikzcd}
\frac{\mathrm{Sym}(\mathbb{W}|m|n)}{\langle\mathbb{W}=\{m,n\}\rangle}\otimes\frac{\mathrm{Sym}(m|n|\mathbb{X})}{\langle\{m,n\}=\mathbb{X}\rangle} \arrow[r] & \frac{\mathrm{Sym}(\mathbb{W}|\mathbb{X})}{\langle\mathbb{W}=\mathbb{X}\rangle} \arrow[r] & \frac{\mathrm{Sym}(\mathbb{W}|m|n)}{\langle\mathbb{W}=\{m,n\}\rangle}\otimes\frac{\mathrm{Sym}(m|n|\mathbb{X})}{\langle\{m,n\}=\mathbb{X}\rangle}, \\
\overline{1}\otimes\overline{1} \arrow[r, mapsto] & 0 \arrow[r, mapsto] & 0, \\
\overline{m}\otimes\overline{1} \arrow[r, mapsto] & \overline{1} \arrow[r, mapsto] & \overline{n}\otimes\overline{1},
\end{tikzcd}$$

which confirms that this map equals the identity, as desired.

For the $a=1$, $b=2$ case, let $|\mathbb{V}|=3=|\mathbb{Y}|$ and $|\mathbb{M}|=2$. The right-hand side corresponds to the sum of the map

$$\begin{tikzcd}
\frac{\mathrm{Sym}(\mathbb{V}|m|\mathbb{M})}{\langle\mathbb{V}=m\cup\mathbb{M}\rangle}\otimes\frac{\mathrm{Sym}(m|\mathbb{M}|\mathbb{Y})}{\langle m\cup\mathbb{M}=\mathbb{Y}\rangle} \arrow[r] & \frac{\mathrm{Sym}(\mathbb{V}|\mathbb{Y})}{\langle\mathbb{V}=\mathbb{Y}\rangle} \arrow[r] & \frac{\mathrm{Sym}(\mathbb{V}|m|\mathbb{M})}{\langle\mathbb{V}=m\cup\mathbb{M}\rangle}\otimes\frac{\mathrm{Sym}(m|\mathbb{M}|\mathbb{Y})}{\langle m\cup\mathbb{M}=\mathbb{Y}\rangle}, \\
\overline{1}\otimes\overline{1} \arrow[r, mapsto] & \overline{1} \arrow[r, mapsto] & \overline{1}, \\
\overline{m}\otimes\overline{1} \arrow[r, mapsto] & \overline{e_1(\mathbb{V})} \arrow[r, mapsto] & \overline{m+e_1(\mathbb{M})}\otimes\overline{1}, \\
\overline{m^2}\otimes\overline{1} \arrow[r, mapsto] & \overline{e_1(\mathbb{V})^2-e_2(\mathbb{V})} \arrow[r, mapsto] & \overline{m^2+me_1(\mathbb{M})+e_1(\mathbb{M})^2-e_2(\mathbb{M})}\otimes\overline{1},
\end{tikzcd}$$

the negative of the map

$$\begin{tikzcd}
\frac{\mathrm{Sym}(\mathbb{V}|m|\mathbb{M})}{\langle \mathbb{V}=m\cup\mathbb{M}\rangle}\otimes\frac{\mathrm{Sym}(m|\mathbb{M}|\mathbb{Y})}{\langle m\cup\mathbb{M}=Z\rangle} \arrow[r] & \frac{\mathrm{Sym}(\mathbb{V}|\mathbb{Y})}{\langle \mathbb{V}=\mathbb{Y}\rangle} \arrow[r] & \frac{\mathrm{Sym}(\mathbb{V}|m|\mathbb{M})}{\langle \mathbb{V}=m\cup\mathbb{M}\rangle}\otimes\frac{\mathrm{Sym}(m|\mathbb{M}|\mathbb{Y})}{\langle m\cup\mathbb{M}=Z\rangle}, \\
\bar{1}\otimes\bar{1} \arrow[r, mapsto] & 0 \arrow[r, mapsto] & 0, \\
\bar{m}\otimes\bar{1} \arrow[r, mapsto] & \bar{1} \arrow[r, mapsto] & \overline{e_1(\mathbb{M})}\otimes\bar{1}, \\
\overline{m^2}\otimes\bar{1} \arrow[r, mapsto] & \overline{e_1(\mathbb{V})} \arrow[r, mapsto] & \overline{me_1(\mathbb{M})+e_1(\mathbb{M})^2}\otimes\bar{1},
\end{tikzcd}$$

and the map

$$\begin{tikzcd}
\frac{\mathrm{Sym}(\mathbb{V}|m|\mathbb{M})}{\langle \mathbb{V}=m\cup\mathbb{M}\rangle}\otimes\frac{\mathrm{Sym}(m|\mathbb{M}|\mathbb{Y})}{\langle m\cup\mathbb{M}=Z\rangle} \arrow[r] & \frac{\mathrm{Sym}(\mathbb{V}|\mathbb{Y})}{\langle \mathbb{V}=\mathbb{Y}\rangle} \arrow[r] & \frac{\mathrm{Sym}(\mathbb{V}|m|\mathbb{M})}{\langle \mathbb{V}=m\cup\mathbb{M}\rangle}\otimes\frac{\mathrm{Sym}(m|\mathbb{M}|\mathbb{Y})}{\langle m\cup\mathbb{M}=Z\rangle}, \\
\bar{1}\otimes\bar{1} \arrow[r, mapsto] & 0 \arrow[r, mapsto] & 0, \\
\bar{m}\otimes\bar{1} \arrow[r, mapsto] & 0 \arrow[r, mapsto] & 0, \\
\overline{m^2}\otimes\bar{1} \arrow[r, mapsto] & \bar{1} \arrow[r, mapsto] & \overline{e_2(\mathbb{M})}\otimes\bar{1},
\end{tikzcd}$$

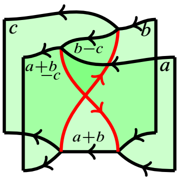
which confirms that this map is the identity. The case $a=2$, $b=1$ follows similarly.

[36, Equation (3.12)] Both sides of this relation are given by

$$\begin{tikzcd}
\frac{\mathrm{Sym}(\mathbb{A}\,|\,\mathbb{L}\,|\,\mathbb{V})}{\langle \mathbb{A}=\mathbb{L}\cup\mathbb{V}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{L}\,|\,\mathbb{W}\,|\,\mathbb{M})}{\langle \mathbb{L}=\mathbb{W}\cup\mathbb{M}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{M}\,|\,\mathbb{X}\,|\,\mathbb{Y})}{\langle \mathbb{M}=\mathbb{X}\cup\mathbb{Y}\rangle} \arrow[d] & \bar{1}\otimes\bar{1}\otimes\bar{1} \arrow[d, mapsto] \\
\frac{\mathrm{Sym}(\mathbb{A}\,|\,\mathbb{S}\,|\,\mathbb{Y})}{\langle \mathbb{A}=\mathbb{S}\cup\mathbb{Y}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{S}\,|\,\mathbb{T}\,|\,\mathbb{X})}{\langle \mathbb{S}=\mathbb{T}\cup\mathbb{X}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{T}\,|\,\mathbb{V}\,|\,\mathbb{W})}{\langle \mathbb{T}=\mathbb{V}\cup\mathbb{W}\rangle}, & \bar{1}\otimes\bar{1}\otimes\bar{1}.
\end{tikzcd}$$

Hence, they are equal. In the above, the tensor products are each taken over symmetric polynomials in the common alphabets between the tensor factors.

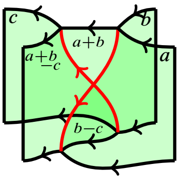
[36, Equations (3.13) and (3.14)] It suffices to prove these relations in the case when $a=1=c$; however, it isn't much more difficult to verify the general relation. To check this, we first note that both of the possible ways to construct the crossing



correspond to the morphism of linear factorizations

$$\frac{\mathrm{Sym}(\mathbb{V}|\mathbb{W}|\mathbb{L})}{\langle \mathbb{V}\cup\mathbb{W}=\mathbb{L}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{L}|\mathbb{X}|\mathbb{Y})}{\langle \mathbb{L}=\mathbb{X}\cup\mathbb{Y}\rangle}\to\frac{\mathrm{Sym}(\mathbb{W}|\mathbb{M}|\mathbb{Y})}{\langle \mathbb{W}=\mathbb{M}\cup\mathbb{Y}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{V}|\mathbb{M}|\mathbb{X})}{\langle \mathbb{V}\cup\mathbb{M}=\mathbb{X}\rangle},$$
$$\overline{1}\otimes\overline{1}\longmapsto\overline{1}\otimes\overline{1},$$

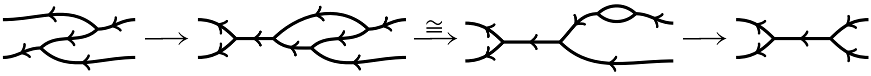
and similarly both ways of constructing the crossing



give the map

$$\frac{\mathrm{Sym}(\mathbb{W}|\mathbb{M}|\mathbb{Y})}{\langle \mathbb{W}=\mathbb{M}\cup\mathbb{Y}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{V}|\mathbb{M}|\mathbb{X})}{\langle \mathbb{V}\cup\mathbb{M}=\mathbb{X}\rangle}\longrightarrow\frac{\mathrm{Sym}(\mathbb{V}|\mathbb{W}|\mathbb{L})}{\langle \mathbb{V}\cup\mathbb{W}=\mathbb{L}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{L}|\mathbb{X}|\mathbb{Y})}{\langle \mathbb{L}=\mathbb{X}\cup\mathbb{Y}\rangle},$$
$$\overline{1}\otimes\overline{1}\longmapsto\sum_{\alpha\in P(a,c)}(-1)^{|\hat{\alpha}|}\,\overline{\pi^{\mathbb{V}}_{\hat{\alpha}}}\otimes\overline{\pi^{\mathbb{Y}}_{\alpha}}.$$

The first is clear, and the second follows, for example, since one way of constructing the sideways crossing is given by the composition:



The corresponding morphism of linear factorizations is the composition

$$\begin{tikzcd}
\frac{\mathrm{S}(\mathbb{W}|\mathbb{M}|\mathbb{Y})}{\langle \mathbb{W}=\mathbb{M}\cup\mathbb{Y}\rangle}\otimes\frac{\mathrm{S}(\mathbb{V}|\mathbb{M}|\mathbb{X})}{\langle \mathbb{V}\cup\mathbb{M}=\mathbb{X}\rangle} \arrow[d] & \overline{1}\otimes\overline{1} \arrow[d, mapsto] \\
\frac{\mathrm{S}(\mathbb{V}|\mathbb{W}|\mathbb{L})}{\langle \mathbb{V}\cup\mathbb{W}=\mathbb{L}\rangle}\otimes\frac{\mathrm{S}(\mathbb{L}|\mathbb{S}|\mathbb{T})}{\langle \mathbb{L}=\mathbb{S}\cup\mathbb{T}\rangle}\otimes\frac{\mathrm{S}(\mathbb{T}|\mathbb{M}|\mathbb{Y})}{\langle \mathbb{T}=\mathbb{M}\cup\mathbb{Y}\rangle}\otimes\frac{\mathrm{S}(\mathbb{S}|\mathbb{M}|\mathbb{X})}{\langle \mathbb{S}\cup\mathbb{M}=\mathbb{X}\rangle} \arrow[d] & \sum_{\alpha}(-1)^{|\hat{\alpha}|}\overline{\pi^{\mathbb{V}}_{\hat{\alpha}}}\otimes\overline{1}\otimes\overline{1}\otimes\overline{\pi^{\mathbb{M}}_{\alpha}} \arrow[d, mapsto] \\
\frac{\mathrm{S}(\mathbb{V}|\mathbb{W}|\mathbb{L})}{\langle \mathbb{V}\cup\mathbb{W}=\mathbb{L}\rangle}\otimes\frac{\mathrm{S}(\mathbb{L}|\mathbb{P}|\mathbb{Y})}{\langle \mathbb{L}=\mathbb{P}\cup\mathbb{Y}\rangle}\otimes\frac{\mathrm{S}(\mathbb{P}|\mathbb{S}|\mathbb{M})}{\langle \mathbb{P}=\mathbb{S}\cup\mathbb{M}\rangle}\otimes\frac{\mathrm{S}(\mathbb{S}|\mathbb{M}|\mathbb{X})}{\langle \mathbb{S}\cup\mathbb{M}=\mathbb{X}\rangle} \arrow[d] & \sum_{\alpha,\beta,\gamma}(-1)^{|\hat{\alpha}|}c^{\alpha}_{\beta,\gamma}\overline{\pi^{\mathbb{V}}_{\hat{\alpha}}}\otimes\overline{\pi^{\mathbb{Y}}_{\gamma}}\otimes\overline{\pi^{\mathbb{M}}_{\beta}}\otimes\overline{1} \arrow[d, mapsto] \\
\frac{\mathrm{S}(\mathbb{V}|\mathbb{W}|\mathbb{L})}{\langle \mathbb{V}\cup\mathbb{W}=\mathbb{L}\rangle}\otimes\frac{\mathrm{S}(\mathbb{L}|\mathbb{X}|\mathbb{Y})}{\langle \mathbb{L}=\mathbb{X}\cup\mathbb{Y}\rangle}, & \sum_{\gamma\in P(a,c)}(-1)^{|\hat{\gamma}|}\overline{\pi^{\mathbb{V}}_{\hat{\gamma}}}\otimes\overline{\pi^{\mathbb{Y}}_{\gamma}},
\end{tikzcd}$$

where in the summations $\alpha\in P(a+b-c,c)$, $\beta\in P(b-c)$ and $\gamma\in P(a)$, and we use the fact that $\overline{\pi^{\mathbb{M}}_{\beta}}\otimes\overline{1}\mapsto 0$ under the last map if $|\beta|\le c(b-c)$. Given this, the only time the Littlewood–Richardson coefficient $c^{\alpha}_{\beta,\gamma}$ is nonzero is when $\beta=c^{b-c}$

(so $\gamma \in P(a,c)$ and $\hat{\alpha} = \hat{\gamma}$), in which case it equals one. Both of the relations then follow from the descriptions of these maps.

[36, Equations (3.15) and (3.16)] The linear factorization stabilized by the matrix factorization corresponding to the top and bottom webs in [36, (3.15)] is

$$\left(\frac{\mathrm{Sym}(\mathbb{P}\,|\,l\,|\,w)}{\langle \mathbb{P} = \{l,w\}\rangle} \otimes \frac{\mathrm{Sym}(w\,|\,\mathbb{W}\,|\,\mathbb{L})}{\langle w \cup \mathbb{W} = \mathbb{L}\rangle}\right) \otimes \left(\frac{\mathrm{Sym}(\mathbb{L}\,|\,\mathbb{M}\,|\,z)}{\langle \mathbb{L} = \mathbb{M} \cup z\rangle} \otimes \frac{\mathrm{Sym}(l\,|\,\mathbb{M}\,|\,\mathbb{X})}{\langle l \cup \mathbb{M} = \mathbb{X}\rangle}\right),$$

where all of the tensor products are over polynomials partially symmetric in the common variables. The map between linear factorizations corresponding to the first term on the left-hand side of [36, (3.15)] is determined by the fact that

$$\overline{1} \otimes \overline{1} \otimes \overline{1} \otimes \overline{1} \mapsto \overline{1} \otimes \overline{1} \otimes \overline{1} \otimes \overline{1} \quad \text{and} \quad \overline{w} \otimes \overline{1} \otimes \overline{1} \otimes \overline{1} \mapsto \overline{1} \otimes \overline{1} \otimes \overline{z} \otimes \overline{1}$$

and the second term is determined by

$$\overline{1} \otimes \overline{1} \otimes \overline{1} \otimes \overline{1} \mapsto 0 \quad \text{and} \quad \overline{w} \otimes \overline{1} \otimes \overline{1} \otimes \overline{1} \mapsto \overline{1} \otimes \overline{1} \otimes \overline{z} \otimes \overline{1} - \overline{w} \otimes \overline{1} \otimes \overline{1} \otimes \overline{1}.$$

The difference between these two maps is thus the identity, confirming the relation. The check of [36, (3.16)] is completely analogous.

[36, Equations (3.17)–(3.20)] The left-hand side of [36, (3.17)] corresponds to the morphism of linear factorizations

$$\begin{aligned}\frac{\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{L}\,|\,\mathbb{M})}{\langle \mathbb{V} = \mathbb{L} \cup \mathbb{M}\rangle} \otimes \frac{\mathrm{Sym}(\mathbb{P}\,|\,\mathbb{L}\,|\,\mathbb{W})}{\langle \mathbb{P} \cup \mathbb{L} = \mathbb{W}\rangle} \otimes \frac{\mathrm{Sym}(\mathbb{M}\,|\,\mathbb{X}\,|\,\mathbb{Y})}{\langle \mathbb{M} = \mathbb{X} \cup \mathbb{Y}\rangle} \\ \to \frac{\mathrm{Sym}(\mathbb{P}\,|\,\mathbb{V}\,|\,\mathbb{S})}{\langle \mathbb{P} \cup \mathbb{V} = \mathbb{S}\rangle} \otimes \frac{\mathrm{Sym}(\mathbb{S}\,|\,\mathbb{T}\,|\,\mathbb{Y})}{\langle \mathbb{S} = \mathbb{T} \cup \mathbb{Y}\rangle} \otimes \frac{\mathrm{Sym}(\mathbb{T}\,|\,\mathbb{W}\,|\,\mathbb{X})}{\langle \mathbb{T} = \mathbb{W} \cup \mathbb{X}\rangle}\end{aligned}$$

given by

$$\overline{1} \otimes \overline{1} \otimes \overline{1} \mapsto \sum_{\alpha \in P(b,d),\, \beta \in P(a,d)} (-1)^{|\hat{\alpha}|+|\hat{\beta}|} \overline{\pi^{\mathbb{P}}_{\hat{\alpha}} \pi^{\mathbb{P}}_{\hat{\beta}}} \otimes \overline{\pi^{\mathbb{Y}}_{\beta}} \otimes \overline{\pi^{\mathbb{X}}_{\alpha}},$$

while the right-hand side is given by

$$\overline{1} \otimes \overline{1} \otimes \overline{1} \mapsto \sum_{\gamma \in P(a+b,d)} c^{\gamma}_{\alpha,\beta} (-1)^{|\hat{\gamma}|} \overline{\pi^{\mathbb{P}}_{\hat{\gamma}}} \otimes \overline{\pi^{\mathbb{Y}}_{\beta}} \otimes \overline{\pi^{\mathbb{X}}_{\alpha}}.$$

The relation then holds since

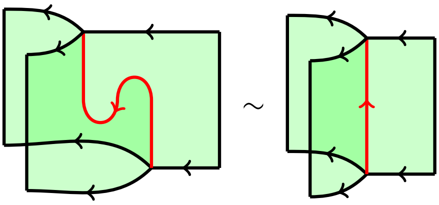
$$\pi^{\mathbb{P}}_{\hat{\alpha}} \pi^{\mathbb{P}}_{\hat{\beta}} = \sum_{\hat{\gamma}} c^{\hat{\gamma}}_{\hat{\alpha}\hat{\beta}} \pi^{\mathbb{P}}_{\hat{\gamma}} \quad \text{and} \quad c^{\hat{\gamma}}_{\hat{\alpha}\hat{\beta}} = c^{\gamma}_{\alpha,\beta}.$$

Relation [36, (3.18)] holds since both sides are given by the map

$$\frac{\mathrm{Sym}(\mathbb{P}\,|\,\mathbb{V}\,|\,\mathbb{S})}{\langle \mathbb{P}\cup\mathbb{V}=\mathbb{S}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{S}\,|\,\mathbb{T}\,|\,\mathbb{Y})}{\langle \mathbb{S}=\mathbb{T}\cup\mathbb{Y}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{T}\,|\,\mathbb{W}\,|\,\mathbb{X})}{\langle \mathbb{T}=\mathbb{W}\cup\mathbb{X}\rangle}$$
$$\to\frac{\mathrm{Sym}(\mathbb{V}\,|\,\mathbb{L}\,|\,\mathbb{M})}{\langle \mathbb{V}=\mathbb{L}\cup\mathbb{M}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{P}\,|\,\mathbb{L}\,|\,\mathbb{W})}{\langle \mathbb{P}\cup\mathbb{L}=\mathbb{W}\rangle}\otimes\frac{\mathrm{Sym}(\mathbb{M}\,|\,\mathbb{X}\,|\,\mathbb{Y})}{\langle \mathbb{M}=\mathbb{X}\cup\mathbb{Y}\rangle}$$

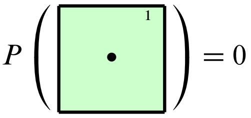
sending $\overline{1}\otimes\overline{1}\otimes\overline{1}\mapsto\overline{1}\otimes\overline{1}\otimes\overline{1}$. The final two relations follow via similar computations.

Isotopy relations All isotopy relations follow from the fact that both way to construct the "sideways crossings" give the same map in **HMF**, and the fact that the foam relation



and its analogs are satisfied in **HMF**. Both are direct computations.

Dot relation Finally, the foam relation



holds via a direct computation that multiplication by $P(X)$ is null-homotopic in the endomorphism algebra of the Koszul factorization $\{Q(X)-Q(Y),X-Y\}$ over $\mathbb{C}[X,Y]$. □

5 The link invariant

In this section, we assign a complex $[\![\tau]\!]^{\Sigma}$ of webs and foams to certain labeled[10] tangle diagrams τ, which, up to homotopy equivalence, is an invariant of the corresponding labeled tangle. We then show how to obtain a link homology isomorphic to that defined by Wu [42] from this invariant, proving Theorem 1.3. Finally, we use the foam technology to prove Theorem 1.1.

[10] In the study of quantum invariants, links and tangles are usually referred to as "colored" by representations of a Lie algebra (or, more precisely, a quantum group). Since we reserve the word colored for webs and foams colored by idempotents, recall that we instead use the nonstandard terminology "labeled", which agrees with our use of this word for webs.

The most precise setting for this invariant is in a certain limiting version of $N\mathbf{Foam}^{\Sigma}$. Note that $N\mathbf{Foam}^{\Sigma}$ is the direct sum of foam categories $N\mathbf{Foam}^{\Sigma}(K)$, where $K = \sum_{i=1}^{m} a_i$ is the sum of the entries in an object $(a_1, \dots, a_m)$. We have a 2–functor $N\mathbf{Foam}^{\Sigma}(K) \to N\mathbf{Foam}^{\Sigma}(K+N)$ given by taking disjoint union with an N–labeled edge/facet. The natural setting for the tangle invariant[11] is the direct limit

$$N\mathbf{Foam}(k + N\infty)^{\Sigma} := \varinjlim_{s} N\mathbf{Foam}^{\Sigma}(k + Ns);$$

however, the invariant can be viewed in $N\mathbf{Foam}^{\Sigma}(k + Ns)$ for s sufficiently large.

We begin by defining $[\![\tau]\!]^{\Sigma}$ on generating tangles, and then explain how to define the invariant for general tangles. Given a labeled, oriented tangle diagram τ, let $c_1, \dots, c_r$ be the labels of the right endpoints and $d_1, \dots, d_l$ be the labels of the left endpoints. Set

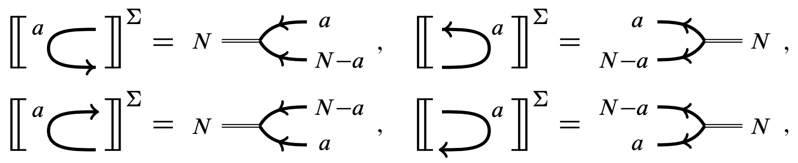
$$\mathcal{O}_R(c_i) = \begin{cases} c_i & \text{if } \tau \text{ is directed out from the } i^{\text{th}} \text{ endpoint,} \\ N - c_i & \text{if } \tau \text{ is directed into the } i^{\text{th}} \text{ endpoint;} \end{cases}$$
$$\mathcal{O}_L(d_i) = \begin{cases} d_i & \text{if } \tau \text{ is directed into the } i^{\text{th}} \text{ endpoint,} \\ N - d_i & \text{if } \tau \text{ is directed out from the } i^{\text{th}} \text{ endpoint;} \end{cases}$$

then $[\![\tau]\!]$ is defined to be a complex in the Hom–category

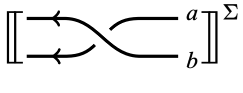
$$\mathrm{Hom}\big((N, \dots, N, \mathcal{O}_R(c_1), \dots, \mathcal{O}_R(c_r)), (N, \dots, N, \mathcal{O}_L(d_1), \dots, \mathcal{O}_L(d_l))\big)$$

of $N\mathbf{Foam}\big(\sum_{i=1}^{r} \mathcal{O}_R(c_i) + Ns\big)^{\Sigma}$.

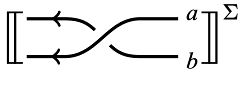
For labeled cap and cup tangles we set



and for labeled, left-directed crossings we use homological shifts of the Rickard complexes $\mathcal{T}\mathbf{1}_{(a,b)}$ from (2-12) and (2-13) to set



$$= \Phi_{\Sigma}(\mathcal{T}\mathbf{1}_{(a,b)})[\min(a,b)],$$

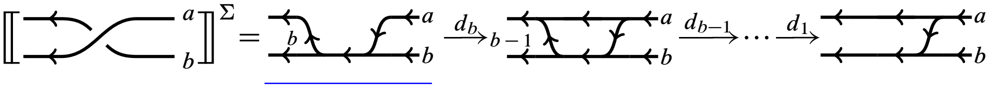


$$= \Phi_{\Sigma}(\mathcal{T}^{-1}\mathbf{1}_{(a,b)})[-\min(a,b)],$$

[11] Here k depends on the boundary and labeling of the tangle.

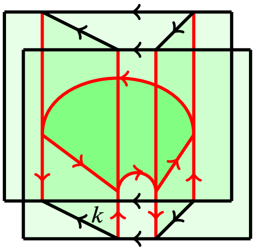
where here [−] here denotes a shift in homological degree. (In [36], the crossing also involved a shift in the quantum grading; however, we omit it from this definition since this grading is broken in $N\mathbf{Foam}^{\Sigma}$.)

Example 5.1 The complex assigned to a negative crossing with $a \geq b$ — compare with (2-15) — is

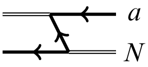


where the underlined term is in homological degree zero and the differential is given by:

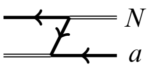
d_k :=



Every labeled tangle admits a diagram given as the horizontal composition ⊗ of tangles which are the disjoint union ⊔ of labeled, directed identity tangles with one of the tangles on which we've already defined the invariant. We define $[\![\tau]\!]^{\Sigma}$ on the disjoint union of identity tangles and a crossing by first taking the disjoint union of $\Phi_{\Sigma}(\widetilde{\mathcal{T}}\mathbf{1}_{(a,b)})$ or $\Phi_{\Sigma}(\widetilde{\mathcal{T}}^{-1}\mathbf{1}_{(a,b)})$ with the identity webs (resp. foams) corresponding to the identity tangle, then taking the disjoint union with N–labeled strands (resp. facets). Finally, we define the invariant on the disjoint union of identity tangles with a cap or cup by taking the disjoint union of the relevant webs with the corresponding identity webs then repeatedly horizontally composing ⊗ with webs

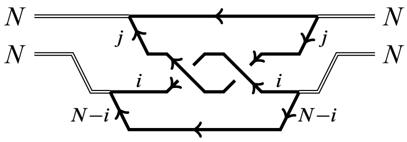


or



to obtain a web mapping between objects where the top-most label is N in both the domain and codomain, and then taking the disjoint union with N–labeled strands. Note that the action of ⊔ as well as ⊗ on complexes is modeled on the tensor product of chain complexes, exactly as in Bar-Natan's canopolis formalism [1].

Example 5.2 For the Hopf link we use the ladder-type link diagram:



Proposition 5.3 *Given a oriented, framed, labeled tangle τ, the complex $[\![\tau]\!]^{\Sigma}$ is independent, up to homotopy, of the diagram used.*

Proof Exactly the same as in [36, Theorem 4.8]. □

In the case that the tangle is actually a labeled link $\mathcal{L}$, all of the boundary points in the complex $[\![\mathcal{L}]\!]^{\Sigma}$ are N–labeled and all webs in it are endomorphisms of a highest weight object of the form $\boldsymbol{o}^{\mathrm{top}} := (N, \dots, N)$. Hence we can apply the representable functor

$$\mathrm{taut}(-) := \mathrm{Hom}(1_{\boldsymbol{o}^{\mathrm{top}}}, -)$$

to $[\![\mathcal{L}]\!]^{\Sigma}$ to obtain a complex of vector spaces. Moreover, we claim that each term in this complex is finite-dimensional. Indeed, every web in $\mathrm{End}(\boldsymbol{o}^{\mathrm{top}})$ is isomorphic to a (finite) direct sum of identity webs $1_{\boldsymbol{o}^{\mathrm{top}}}$. Foam facets with label N are additively indecomposable, since the only admissible coloring by idempotents is given by the full multiset Σ. It follows that endomorphisms of $1_{\boldsymbol{o}^{\mathrm{top}}}$ are all given by the images of closed diagrams in $\check{\mathcal{U}}_Q(\mathfrak{gl}_m)$, which act by scalars in $N\mathbf{Foam}^{\Sigma}$, confirming our claim. Denote by $\mathrm{KhR}^{\Sigma}(\mathcal{L})$ the homology of this complex.

Theorem 5.4 *Up to shifts in homological degree, $\mathrm{KhR}^{\Sigma}(\mathcal{L})$ is isomorphic to Wu's colored, deformed Khovanov–Rozansky homology of the mirror link $\mathcal{L}'$.*

Proof This result follows in the spirit of the proof of [36, Theorem 4.12]. We cannot directly apply the methods there, however, since the 2–functor $\check{\mathcal{U}}_Q(\mathfrak{gl}_m) \to N\mathbf{Foam}^{\Sigma}$ is not a 2–representation in the strict sense, as it doesn't preserve the grading.

Nevertheless, we can consider the 2–category $\check{\mathcal{U}}_Q^{0\le N}(\mathfrak{gl}_m)^{\Sigma}$ where we've imposed relation (3-2) for each weight. This implies the specifications of fake bubble parameters in highest weight to elementary symmetric functions evaluated at Σ. Given a labeled link $\mathcal{L}$, we can pull the complex $[\![\mathcal{L}]\!]^{\Sigma}$ back to $\check{\mathcal{U}}_Q^{0\le N}(\mathfrak{gl}_m)^{\Sigma}$ and simplify until each term the complex only consists of direct sums of the identity 1–morphism on the highest weight $(N, \dots, N, 0, \dots, 0)$ in $\check{\mathcal{U}}_Q^{0\le N}(\mathfrak{gl}_m)^{\Sigma}$ (which maps to the object $\boldsymbol{o}^{\mathrm{top}}$ under Φ_{Σ}).

The homology of the link can be computed entirely in the context of $\check{\mathcal{U}}_Q^{0\le N}(\mathfrak{gl}_m)^{\Sigma}$. Moreover, similarly to the case discussed in [36], any link homologies defined using the images of the Rickard complexes in a "skew Howe" 2–representation factoring through $\check{\mathcal{U}}_Q^{0\le N}(\mathfrak{gl}_m)^{\Sigma}$ must agree (see the work of Cautis [6] for the first appearance of this idea).

Deformed foams give such a 2–representation, as do deformed matrix factorizations, via the 2–functors Γ_m from the proof of Theorem 4.29. Note that the link homology theory defined using Γ_m is not defined in exactly the same way as in Wu's work. Indeed, there are the following differences: Wu's assignment of complexes to link crossings is opposite to ours, he has shifts in homological degree for some crossings in order to obtain invariance under the first Reidemeister move, and he does not use matrix factorizations associated to N–labeled web edges.

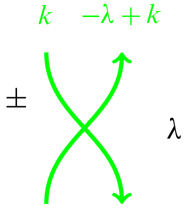
Nevertheless it is easy to see the relation between our invariant and Wu's. Given a labeled braid, the 2–functor Γ_m assigns a to it a complex of matrix factorizations which, up to shifts in homological degree, agrees with the complex of matrix factorizations Wu assigns to the mirror image of the braid. It hence suffices to show that the vector spaces and differentials in this complex after closing the braid agrees with the those obtained by closing using N–labeled edges and taking Hom from $1_{\boldsymbol{o}^{\mathrm{top}}}$. This follows exactly as in [36, Theorem 4.12]. □

Remark 5.5 As a variation of (2-14), where $a_i \geq a_{i+1}$, let $\mathbf{1}_{\boldsymbol{a}}\mathcal{T}_i'^{-1}$ denote the complex

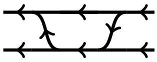
$$\cdots \mathbf{1}_{\boldsymbol{a}}\mathcal{E}_i^{(a_i-a_{i+1}+s)}\mathcal{F}_i^{(s)}\{-s\} \xrightarrow{d_s'} \cdots \xrightarrow{d_2'} \mathbf{1}_{\boldsymbol{a}}\mathcal{E}_i^{(a_i-a_{i+1}+1)}\mathcal{F}_i\{-1\} \xrightarrow{d_1'} \underline{\mathbf{1}_{\boldsymbol{a}}\mathcal{E}_i^{(a_i-a_{i+1})}}$$

with the underlined term (as usual) in homological degree zero and differentials given by compositions of splitters and thickness-1 cap 2–morphisms.

It is easy to check that $\mathbf{1}_{\boldsymbol{a}}\mathcal{T}_i'^{-1}$ is isomorphic to $\mathbf{1}_{\boldsymbol{a}}\mathcal{T}_i^{-1}$ via the chain map given on objects by



for a suitable choice of signs. Analogously, the Rickard complexes (2-12) are isomorphic to complexes with objects $\mathcal{E}_i^{(s)}\mathcal{F}_i^{(a_i-a_{i+1}+s)}$ and in general we may assume that the complexes $[\![-]\!]^{\Sigma}$ associated to crossings consist of webs of shape:



We now proceed with the decomposition of the invariant. Consider an oriented, labeled tangle diagram τ or, more specifically, an oriented, labeled link diagram $\mathcal{L}$. Our goal is to understand the dependence of $[\![\tau]\!]^{\Sigma}$ and $\mathrm{KhR}^{\Sigma}(\mathcal{L}) = \mathrm{H}_*(\mathrm{taut}([\![\mathcal{L}]\!]^{\Sigma}))$ on Σ. This is done in four steps:

(1) In Section 5.1 we show that $[\![\tau]\!]^\Sigma$, regarded as a complex over $(N\mathbf{Foam}^\Sigma)^\wedge$, decomposes into a direct sum of complexes $[\![\tau_f]\!]^\Sigma$ indexed by colorings f of the tangle components by multisubsets of Σ.

(2) In Section 5.2 we show that the summands $[\![\tau_f]\!]^\Sigma$ from the first step correspond under the splitting functor ϕ from Section 4.3 to a tensor product with one tensorand $[\![\tau_{\lambda\in f}]\!]^\Sigma$ for every different root $\lambda \in \Sigma$.

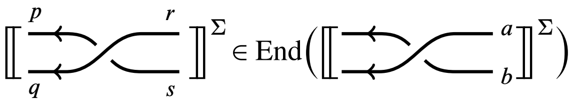
(3) In Section 5.3 we show that foams colored with only one root λ behave like $\mathfrak{sl}_{N_\lambda}$ foams.

(4) In Section 5.4 we assemble the previous results for $\tau = \mathcal{L}$ and track them through relatives of the functor taut to prove Theorem 1.1.

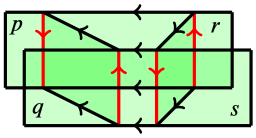
5.1 The direct sum decomposition of the invariant

We already know that if we work in $(N\mathbf{Foam}^\Sigma)^\wedge$, all webs in the complex $[\![\tau]\!]^\Sigma$ split into direct sums under coloring web edges with multisubsets of Σ. The goal of this section is to show in Lemma 5.10 that the colorings that contribute to $[\![\tau]\!]^\Sigma$ are the ones that are consistent along tangle components. This follows from the orthogonality of idempotents coming from inconsistent colorings, see Corollary 5.8, after observing in Proposition 5.7 that decorations "slide through crossings".

Definition 5.6 Let p, q, r and s be symmetric polynomials of the appropriate number of variables. Then we define endomorphisms of the chain complexes for negative crossings

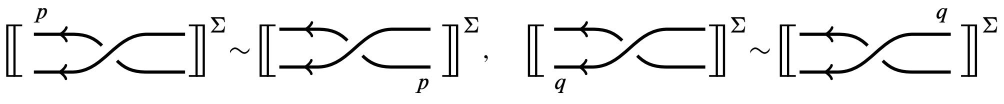


given by



on the webs appearing in the complex. Here we have assumed $a \geq b$. For the cases of $a \leq b$ and for the positive crossings we make analogous definitions.

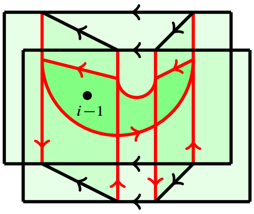
Proposition 5.7 *Let p and q be symmetric polynomials in the appropriate number of variables. Then the following chain maps are homotopic:*



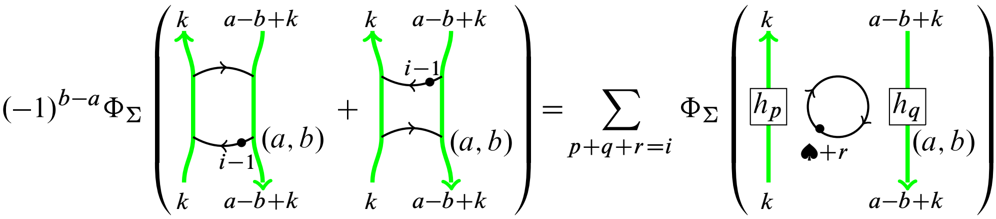
Analogous statements also hold for the positive crossing.

Proof From the foam description of these chain maps, it is easy to see that composing such chain maps is equivalent to multiplying decorations on the foam facets. Since null-homotopic chain maps form an ideal in the ring of endomorphisms of the crossing complex, it suffices to find homotopies in the cases where p (or q) is a complete symmetric polynomial h_i. We will only prove the first homotopy, as the other case is analogous. Denote h_i acting on the left as h_i^l and on the right as h_i^r. Recall that the differential for negative crossing complexes is given using 1–labeled cap foams. We now prove by induction on $i \geq 1$ that the foams

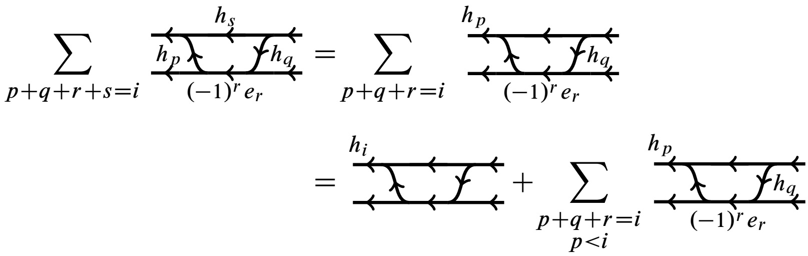
$\eta_i \;:=$

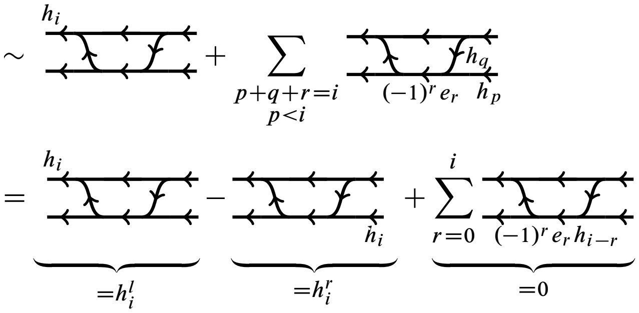


constructed using $(i-1)$–dotted 1–labeled cup foams, assemble to a chain homotopy from h_i^l to h_i^r. We start the computation with an equation from [21, Lemma 4.6.4], which under the foamation functor Φ_Σ gives

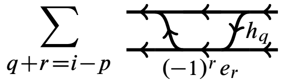


which is a foam identity where the left-hand side is $\pm(\eta_i d + d\eta_i)$. We continue the computation but, since from the next step onwards all foams are identity foams with decorations, we only draw the underlying webs and write decorations next to the corresponding web edges. Using [36, Equation (3.32)] to resolve the "bubble" in the previous step, we get that this equals:



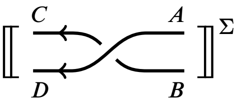


In the case where $i = 1$, the homotopy $\sim$ at the beginning of the second line is an equality. This constitutes the start of the induction. For the induction step we use the homotopy of h_p^l and h_p^r for $p < i$ to proceed to the second line. This is possible because of the fact, which can easily be checked via the decoration migration relations on foams (see (2-3)), that



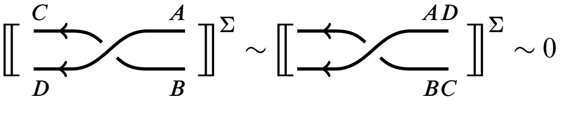
is a chain map. □

Corollary 5.8 *Let A, B, C and D be multisubsets of Σ of the appropriate size and, by abuse of notation, we denote the associated idempotents with the same letter. Then*



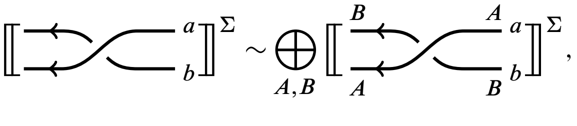
is an idempotent chain map. Furthermore, if $A \neq D$ or $B \neq C$, then it is null-homotopic. Analogous statements hold for positive crossings.

Proof We have already noted that composition of such chain maps corresponds to multiplication of decorations. Thus, the chain map is clearly idempotent. Now suppose that $A \neq D$ or $B \neq C$. Then, using Proposition 5.7, we have



since A and D or B and C are orthogonal idempotents. □

Lemma 5.9 *There is a homotopy equivalence of complexes over $(N\mathbf{Foam}^{\Sigma})^{\wedge}$*



where the summands on the right-hand side denote the subcomplexes of the complex on the left-hand side obtained by coloring webs and foams by idempotents A and B at the indicated positions. In the direct sum, A and B range over all multisubsets of Σ of the correct size. The analogous statements hold for positive crossings.

Proof The objects of the complex on the left-hand side, which are webs, split into direct sums according to the definition of $(N\mathbf{Foam}^{\Sigma})^{\wedge}$ when one colors all boundary edges by idempotents. The differential clearly respects this decomposition since it locally looks like an identity foam around the decoration by idempotents. Finally, Corollary 5.8 shows that summands are null-homotopic if they do not come from a coloring that is consistent along the strands in the crossing, and working in the homotopy category of such complexes we immediately cancel null-homotopic summands. □

The following global version of this lemma follows directly:

Lemma 5.10 *Let τ be a labeled, oriented tangle diagram. The complex $[\![\tau]\!]^{\Sigma}$, regarded over $(N\mathbf{Foam}^{\Sigma})^{\wedge}$, splits into a direct sum of complexes $[\![\tau_f]\!]^{\Sigma}$, and there is one such piece for every coloring f of tangle components by idempotents corresponding to multisubsets of Σ of the appropriate size.*

5.2 The tensor product decomposition of the summands

In this section we show that the functor ϕ from Section 4.3 can be used to split idempotent colored summands $[\![\tau_f]\!]^{\Sigma}$ of the chain complex associated to a tangle diagram — as described at the end of the previous subsection — into the tensor product complex $\bigotimes_{\lambda}[\![\tau_{\lambda\in f}]\!]^{\Sigma}$ of their λ–components. We now define these concepts:

Definition 5.11 Let $[\![\tau_{\lambda\in f}]\!]^{\Sigma}$, *the λ–component* of $[\![\tau_f]\!]^{\Sigma}$, be the sequence of colored webs and foams between them obtained by taking the λ–component of every web and foam appearing in $[\![\tau_f]\!]^{\Sigma}$. It is easy to check that $[\![\tau_{\lambda\in f}]\!]^{\Sigma}$ is itself a chain complex over $(N\mathbf{Foam}^{\Sigma})^{\wedge}$.

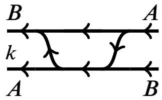
Let $\bigotimes_{\lambda}[\![\tau_{\lambda\in f}]\!]^{\Sigma}$ be the tensor product complex of $[\![\tau_{\lambda\in f}]\!]^{\Sigma}$ given on webs by taking disjoint union $\sqcup$. In particular, the webs in this chain complex are exactly the associated split webs $\bigsqcup_{\lambda} W_{\lambda}$ of webs W in $[\![\tau_f]\!]^{\Sigma}$. The foams giving the components of the differential in $\bigotimes_{\lambda}[\![\tau_{\lambda\in f}]\!]^{\Sigma}$ are (up to a sign) the disjoint union of the λ–components of the differential foams in $[\![\tau_f]\!]^{\Sigma}$. The sign is the usual sign that is necessary to make the differential in the tensor product complex square to zero.

Applying the web splitting functor ϕ from Section 4.3 to the chain complex $[\![\tau_f]\!]^\Sigma$ results in a chain complex consisting of exactly the same split webs as $\bigotimes_\lambda [\![\tau_{\lambda\in f}]\!]^\Sigma$. Furthermore, there exists a natural choice of homological grading on $[\![\tau_{\lambda\in f}]\!]^\Sigma$ that makes the bijection between webs in $\phi([\![\tau_f]\!]^\Sigma)$ and $\bigotimes_\lambda [\![\tau_{\lambda\in f}]\!]^\Sigma$ grading-preserving. This is explained for the local case of a single crossing in Remark 5.12 and immediately generalizes to $[\![\tau_f]\!]^\Sigma$.

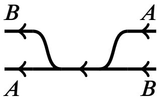
The main task in this section is to prove Theorem 5.14, which states that the isomorphism foams T_W in the definition of ϕ can be chosen so that the differential of $\phi([\![\tau_f]\!]^\Sigma)$ equals the differential of $\bigotimes_\lambda [\![\tau_{\lambda\in f}]\!]^\Sigma$, and we have

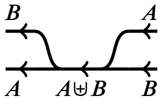
$$\phi([\![\tau_f]\!]^\Sigma) = \bigotimes_\lambda [\![\tau_{\lambda\in f}]\!]^\Sigma.$$

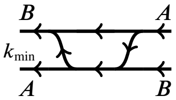
Remark 5.12 Consider the webs in the chain complex associated to a crossing, eg with cap differentials for the sake of concreteness, where we have already placed idempotents on all boundary edges of the webs:

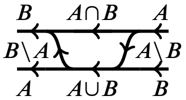
$W_k :=$ 

Without loss of generality, we assume that $|A| \geq |B|$. If such a web is not isomorphic to the zero web, it decomposes into a direct sum by coloring the interior edges of the web with various idempotents. The crossing complex starts with $W_{k_{\max}} = W_{|B|}$ in homological degree zero – which is isomorphic to the zero web if and only if $A \uplus B \not\subset \Sigma$, but which is indecomposable otherwise. Further, there exists a minimal $k_{\min} = |B \setminus A|$ such that $W_{k_{\min}}$ is nonzero and indecomposable:

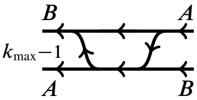
$W_{k_{\max}} =$ 

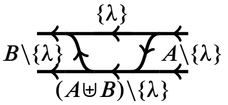
$=$ 

, $W_{k_{\min}} =$ 

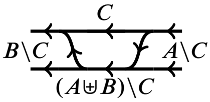
$=$ 

Now consider the target $W_{k_{\max}-1}$ of the differential on $W_{k_{\max}}$:

$W_{k_{\max}-1} =$ 

$= \bigoplus_\lambda$ 

More generally, any nonzero web W_k decomposes into a direct sum of webs which differ in labels and colorings from $W_{k_{\max}}$ by a rerouting of a multisubset C of $A \cap B$ around the square:

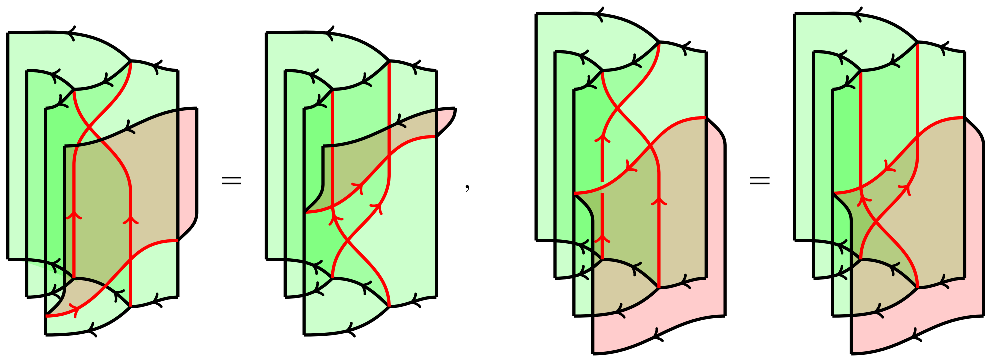
$W_C :=$ 

Such an indecomposable web is nonzero if and only if $(A \uplus B) \setminus C \subset \Sigma$. Clearly $C = A \cap B$ satisfies this because we assume that $A, B \subset \Sigma$. Since this condition can be checked for every root individually, there is a minimal $C_{\min}$ such that $W_{C_{\min}}$ and every W_C for $C_{\min} \subset C \subset A \cap B$ is nonzero. We can think of the set of admissible C as lying on the lattice $\mathbb{Z}^l$ with the k^{th} coordinate indicating the multiplicity of the k^{th} root in C. Then it is clear that the homological grading of a web is the sum of the coordinates and the support of the nonzero W_C is an l–dimensional box. Components of the differential are caps colored by a single root λ_k and hence map between summands in which the rerouting sets C differ by λ_k, ie map between lattice points which differ by 1 in the k^{th} coordinate only. The differentials in this complex already appear as the ones coming from a tensor product of complexes — one for each root λ_k — with homological grading the k^{th} coordinate in the lattice and with differential corresponding to the λ_k colored cap differential.

In the next lemma we collect commutation relations needed in the proof of Theorem 5.14. Red facets are colored with a multisubset containing a single root λ, blue facets are colored with a multisubset not containing λ, and the coloring of the green facets is uniquely determined or arbitrary — generically, they contain both λ and other roots.

Lemma 5.13 *Splitting off or merging a red facet commutes with arbitrary M–P foams, red–blue digon creation, digon removal, zip and unzip foams, up to certain units.* (The graphics in the following proof illustrate and make precise these statements.)

Proof Throughout the proof of this lemma, the displayed graphics are to be interpreted as local foam pieces. First we consider the case of M–P foams, by which we mean the elementary foams between the two possible two-splitter (two-merger) webs. They are shown in green in the following graphics:



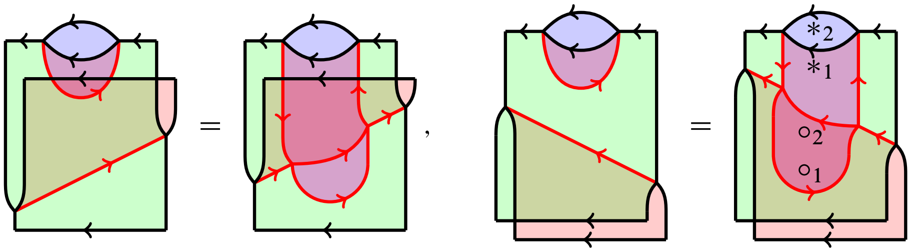
The first commutation relation follows from a version of the foam relation [36, (3.12)] and repeated use of relation (2-1). The second commutation relation holds because it is an isotoped version of the pitchfork relation [36, (3.19)]. There are analogous versions of these relations where the seam attaching the red facet to the rest of the foam is reoriented and inclined the other way, and another four relations hold for a red facet split off on the back-side of the green foam. Clearly, red mergers and splitters then also commute with the inverse M–P foams. These 16 commutation relations describe all possible interactions of red splitters and mergers with a M–P foam between splitter webs. The cases of M–P foams between merger webs is handled similarly.

While the commutation relations with M–P foams are independent of the coloring of foam facets with idempotents, this is in general no longer the case for digon creation, digon removal, zip and unzip foams. Instead, we get commutation up to unit decorations, using the relations in Section 4.1. If we denote the foams that split off or merge a red facet by d and the foam across which we want to commute it by X, then the relations we get take the form

$$X \circ d = u_1 \circ d \circ u_2 \circ X',$$

where u_1 and u_2 are identity foams with decorations that are invertible under the composition $\circ$ in the 2–morphism direction, and X' is a foam that is equal to X as a CW-complex, but might have different labels on facets.[12] Practically this means that we can commute the red facet past the foams mentioned in the statement of the lemma at the expense of invertible decorations. Furthermore, we will see that we can keep the red facet clear of all such decorations.

First we look at the case of a digon creation. The following graphics represent the local piece around the digon creation:



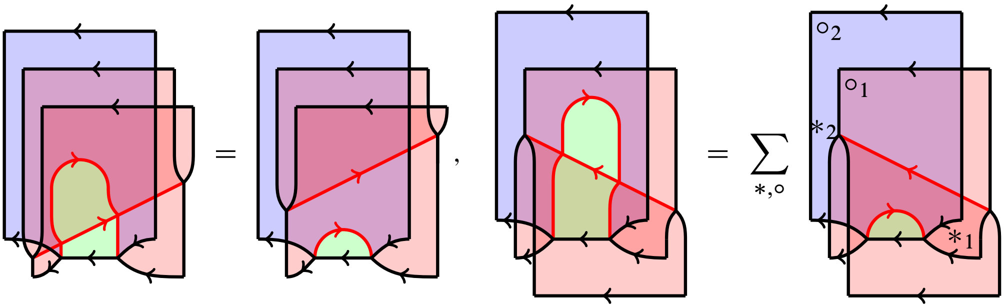
We suppress the precise description of the unit decorations, since they are not immediately relevant for the following discussion and can easily be reconstructed from the description here and the relations in Section 4.1. We do, however, keep track of where

[12] Facets which would have label 0 have to be erased in X'.

the decorations are placed and of their type: we place a pair $\circ_1$, $\circ_2$ on faces with alphabets $\mathbb{X}$ and $\mathbb{Y}$ respectively for a decoration of the form $\sum_{\alpha \in P(-,-)} (-1)^{|\widehat{\alpha}|} \pi_\alpha(\mathbb{X}) \pi_{\widehat{\alpha}}(\mathbb{Y})$ and $*_1$, $*_2$ for the corresponding inverse decoration.

The first commutation relation holds because it is a M–P foam with its inverse; see relation (2-1). For the second one we first introduce a red–blue blister via relation (4-7) below the seam of the red facet, next we slide this seam across the seam of the blister using relation (2-1), and finally we use (4-6) (with inverse units $*$) to join the blister and the digon creation in the upper region of the foam.[13] Analogous identities hold for sliding a facet past a digon creation on the other side.

Next we consider the case of an unzip:



The first commutation relation again holds because it is a M–P foam with its inverse. For the second one, we first break the green strip in the lower half of the diagram on the left-hand side of the relation using relation (4-2). The seam bounding the upper green region can then be moved upward across the seam of the red facet using relation (2-1), and finally the whole upper green region can be removed via relation (4-3) at the expense of a unit $\circ$ acting on top. Similar commutation relations hold for digon removal and zip foams. □

Theorem 5.14 *Let $[\![\tau_f]\!]^\Sigma$ be an idempotent colored summand of the chain complex associated to a tangle diagram, then there exists a choice of isomorphism foams T_W used to define the functor ϕ* (see Definition 4.26) *such that $\phi([\![\tau_f]\!]^\Sigma) = \bigotimes_\lambda [\![\tau_{\lambda \in f}]\!]^\Sigma$.*

Proof Recall from Definition 4.24 and the proof of Proposition 4.25 that ϕ is the composition of functors ϕ_2 and ϕ_1. The latter acts on complexes by replacing colored webs W by $L \otimes W \otimes R$ and foams d by $\mathrm{id}_L \otimes d \otimes \mathrm{id}_R$. We prove this theorem by constructing splitter isomorphism foams

$$T_W \colon \phi_1(W) = L \otimes W \otimes R \to \bigsqcup_\lambda W_\lambda = \phi(W)$$

[13] Here, we avoid the use of relation [36, (3.13)], which would put decorations on the split off red facet.

for each colored web W in $[\![\tau_f]\!]^\Sigma$ that give an isomorphism of chain complexes $\phi_1([\![\tau_f]\!]^\Sigma) \to \bigotimes_\lambda [\![\tau_{\lambda \in f}]\!]^\Sigma$. That is, we have to check that the T_W assemble to a chain map with respect to the differential $d_1 := \mathrm{id}_L \otimes d \otimes \mathrm{id}_R$ on $\phi_1([\![\tau_f]\!]^\Sigma)$ and the differential d_2 on $\bigotimes_\lambda [\![\tau_{\lambda \in f}]\!]^\Sigma$. If $d_1\colon \phi_1(W_1) \to \phi_1(W_2)$, then we need

$$T_{W_2} \circ d_1 = d_2 \circ T_{W_1}. \tag{5-1}$$

Actually, it suffices to construct isomorphism foams $T''_W\colon \phi_1(W) \to \phi(W)$ such that, for every web W in $[\![\tau_f]\!]^\Sigma$, there is an identity foam with unit decoration $u_W\colon \phi(W) \to \phi(W)$ such that

$$T''_{W_2} \circ d_1 = u_{W_2} \circ (\pm d_2) \circ u^{-1}_{W_1} \circ T''_{W_1}. \tag{5-2}$$

Then setting $T'_W := u^{-1}_W \circ T''_W$ gives isomorphism foams that satisfy

$$T'_{W_2} \circ d_1 = (\pm d_2) \circ T'_{W_1}$$

and with a suitable choice of signs $T_W := \pm T'_W$ will satisfy (5-1). That such a sign assignment always exists is well-known and can be proved along similar lines as the fact that Khovanov homology is independent of the numbering of the crossings in a link diagram.

It remains to construct web splitting isomorphism foams T''_W that satisfy (5-2). They are systematically built in three steps:

(1) The resolutions of a crossing in the tangle diagram are ladder webs (see [8] or [36] for this terminology) with two rungs. The first step splits the rungs in every crossing ladder web and sorts them into groups according to their root coloring.

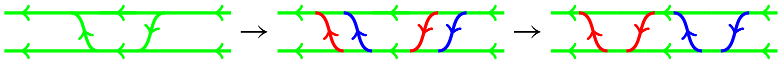
(2) The second step splits the uprights in every crossing ladder. The result is a semisplit web.

(3) The third step is of a global nature; it completely separates the colored components, as in the proof of Proposition 4.22.

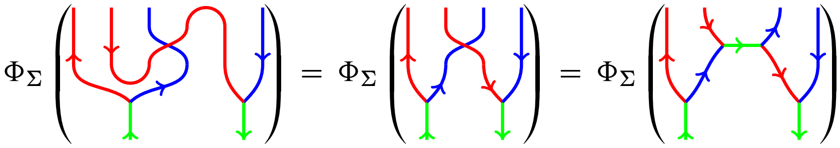
Every step corresponds to a foam that splits the web further and T'' is then defined as their composition. In the following we show that the foams in every step satisfy an equation of type (5-2). That is, the cap (or cup) differential can be moved through the splitting foam at the expense of signs and unit decorations which only depend on the identity foam on which they are placed. In this case we say that the differential *commutes* with the splitting foam *up to canonical units*. If every step satisfies this then so does the composite T'', since unit decorations slide through such isomorphism foams via relations (4-4).

In each of the three steps we only treat the case of colorings by two orthogonal idempotents, which are indicated by red and blue colorings. The same argument implies that we can split off one root at a time from the rest, and induction on the number of distinct roots in Σ then proves the theorem. Furthermore, we only consider the case of cap differentials, as the cup differential case is completely analogous.

Step 1 For every crossing, we consider the corresponding ladder web. First we split the rungs of the ladder into components:

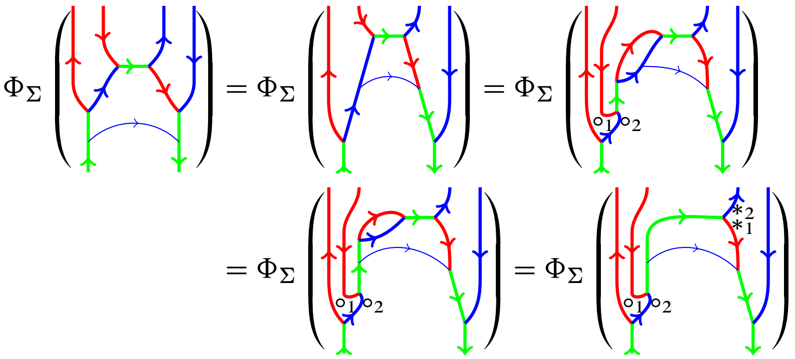


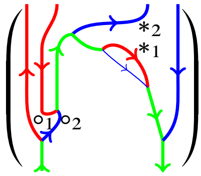
We choose this foam to be the image under the foamation functor Φ_Σ of certain categorified quantum group 2–morphisms in the quotient $\check{\mathcal{U}}_Q^{0\leq N}(\mathfrak{gl}_m)^\Sigma$. For this, we use thick calculus, but we omit the weights and thicknesses of strands. Here, we use colors blue and red to indicate decorations by idempotents corresponding to disjoint multisubsets of Σ, whereas green is the generic color which is used for mixed colorings. Let the reader be warned again that the following graphics show categorified quantum group 2–morphisms and *not* webs. The foam above is given by:

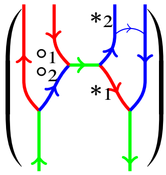


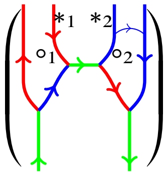
Using the $\mathcal{U}_Q(\mathfrak{gl}_2)$ relations, it is not hard to see that this 2–morphism is invertible via the vertically flipped 2–morphism with some unit decorations. The only nontrivial observation is that the oppositely oriented Reidemeister II-type move can be undone at the expense of a sign, because all error terms are killed by orthogonal idempotents.

In the following we investigate how this 2–morphism commutes with red and blue thickness-1 cap 2–morphisms respectively. This computation immediately transfers to the corresponding foams via Φ_Σ:

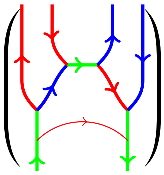
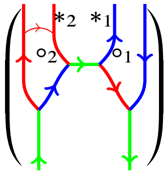


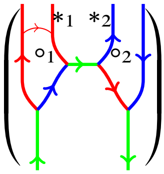
$= \Phi_\Sigma$ 

$= (-1)^r \Phi_\Sigma$ 

$= (-1)^r \Phi_\Sigma$ 

Here and in the following, r (resp. b) is the thickness of the right red (resp. left blue) component in the bottom green strands. An analogous computation shows:

Φ_Σ 
$= (-1)^b \Phi_\Sigma$ 

$= \Phi_\Sigma$ 

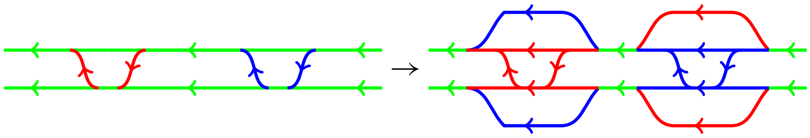
The last equation holds because we can swap the positions $\circ_1$ and $\circ_2$ on strands of thickness r and b at the expense of multiplying by $(-1)^{rb}$. This is immediate from

$$\sum_{\alpha\in P(r,b)} (-1)^{|\hat{\alpha}|}\pi_\alpha(\mathbb{X})\pi_{\hat{\alpha}}(\mathbb{Y}) = \sum_{\hat{\alpha}\in P(b,r)} (-1)^{|\hat{\alpha}|}\pi_{\hat{\alpha}}(\mathbb{Y})\pi_{\hat{\hat{\alpha}}}(\mathbb{X})$$
$$= (-1)^{rb} \sum_{\beta\in P(b,r)} (-1)^{|\hat{\beta}|}\pi_\beta(\mathbb{Y})\pi_{\hat{\beta}}(\mathbb{X})$$

and the analogous statement holds for the inverse decorations on positions $*_1$ and $*_2$ similarly.

We conclude that the foams from the first step commute with the differential up to canonical units.

Step 2 We further split the crossing resolutions into semisplit webs:

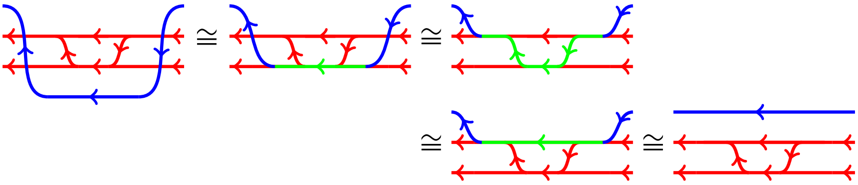


A foam that peels off the outer strands can be constructed from the building blocks studied in Lemma 5.13. According to it, a differential with this target web commutes with the peeling foam up to canonical unit decorations. All these local foams glue together and can be further composed with unzips (if necessary) to have as target semisplit webs. These unzips are placed far away from crossing sites and thus don't change the commutation behavior.

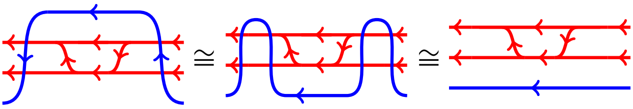
Step 3 Finally, we construct a foam from the semisplit webs of Step 2 to the completely split webs $\phi(W)$. This can be done as in the proof of Proposition 4.22, but we further assume that the "squares" in which the differentials are supported only interact with edges far away from other colored crossing sites during the homotopy. In other words, we assume that the vertices and edges from other colored squares never cross each other during the homotopy. We thus check that the local move of isotoping a red square through a blue edge commutes with the red cap differential up to canonical units.

Each of the isomorphisms

(5-3)



except the third is a composite of red–blue zip, unzip, digon creation, digon removal, and M–P foams as in Lemma 5.13, and hence they commute with the differential up to canonical units. The third isomorphism can be realized as a blue cap in thick calculus, which commutes with a red cap on the same square, up to a sign. The red cap differential then also commutes, up to sign and canonical units, with the inverse of the above isomorphism, and with pulling a blue facet across the square in the opposite direction:



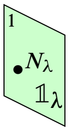
□

5.3 Identifying the tensorands

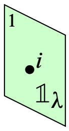
Definition 5.15 Let $(N\mathbf{Foam}^{\lambda\in\Sigma})^\wedge$ be the 2–subcategory of $(N\mathbf{Foam}^{\Sigma})^\wedge$ consisting of only those 1–morphisms and 2–morphisms colored by idempotents $\mathbb{1}_\lambda$ corresponding to multisubsets of Σ which only contain the root λ.

Lemma 5.16 $(N\mathbf{Foam}^{\lambda\in\Sigma})^\wedge$ *is generated as a 2–category by the same elementary foams as* $N\mathbf{Foam}$, *but with idempotent decorations* $\mathbb{1}_\lambda$ *on each web edge and foam facet. It satisfies the same relations as* $N\mathbf{Foam}$ *and additionally:*

(5-4)



$$= -\sum_{i=0}^{N_\lambda-1}\binom{N_\lambda}{i}(-\lambda)^{N_\lambda-i}$$



Proof All relations except (5-4) are directly inherited from $N\mathbf{Foam}$ via its quotient $N\mathbf{Foam}^{\Sigma}$. The decorations $\mathbb{1}_\lambda$ are idempotent and can be moved around freely. For relation (5-4), we write the action of a dot as ξ and will show the equivalent formulation $\mathbb{1}_\lambda(\xi-\lambda)^{N_\lambda}=0$. To see this, we consider the algebra of decorations of a 1–labeled facet in $N\mathbf{Foam}^{\Sigma}$, which is given by $\mathbb{C}[\xi]/\langle P(\xi)\rangle$. Under the algebra isomorphism

$$\mathbb{C}[\xi]/\langle P(\xi)\rangle \to \bigoplus_{k=1}^{l} \mathbb{C}[\xi]/\langle(\xi-\lambda_k)^{N_{\lambda_k}}\rangle,$$
$$p(\xi)+\langle P(\xi)\rangle \mapsto \big(p(\xi)+\langle(\xi-\lambda_1)^{N_{\lambda_1}}\rangle,\dots,p(\xi)+\langle(\xi-\lambda_l)^{N_{\lambda_l}}\rangle\big),$$

$\mathbb{1}_\lambda$ is sent to the vector having a single entry $1+\langle(\xi-\lambda)^{N_\lambda}\rangle$ and zero everywhere else, hence $\mathbb{1}_\lambda(\xi-\lambda)^{N_\lambda}$ is sent to zero. □

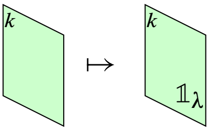
Proposition 5.17 *Let N_λ be the multiplicity of λ in Σ, then there is an isomorphism of 2–categories*

$$N_\lambda\mathbf{Foam}^\bullet \cong (N\mathbf{Foam}^{\lambda\in\Sigma})^\wedge. \tag{5-5}$$

Proof Let $\iota_\lambda\colon N_\lambda\mathbf{Foam}^\bullet \to (N\mathbf{Foam}^{\lambda\in\Sigma})^\wedge$ be the 2–functor which is defined on

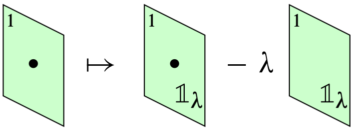
- objects by sending a sequence $\boldsymbol{a}$ to itself,
- 1–morphisms by sending webs to the same webs, but with additional coloring by multisets containing only λ on the edges, and
- 2–morphisms by sending a foam to the foam which is topologically identical but has a decoration by a λ–idempotent $\mathbb{1}_\lambda$ added on every facet:

(5-6)



A decoration on a foam facet, interpreted as a symmetric polynomial in an alphabet $x_1,\dots,x_k$, is sent to the same symmetric polynomial, but in the alphabet $x_1-\lambda,\dots,x_k-\lambda$. Formally, it suffices to define:

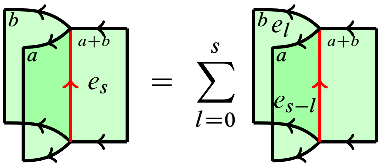
(5-7)



We now check that ι_λ exactly maps the defining relations of $N_\lambda\mathbf{Foam}^\bullet$ — see [36] — to the set of relations that determine $(N\mathbf{Foam}^{\lambda\in\Sigma})^\wedge$, which was identified in (5-4).

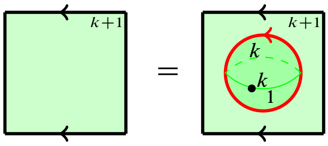
All relations that do not involve decorations are preserved by ι_λ; these are [36, (3.8), (3.12), (3.15)–(3.20)]. We examine the remaining relations:

- A minimal version of [36, (3.9)] is

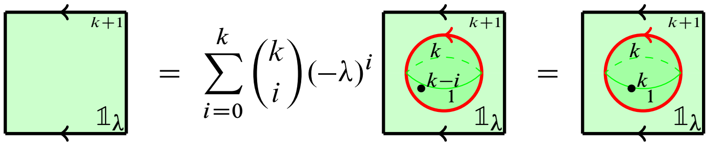


and the collection of all instances of this relation (for $1 \leq s \leq a+b$) has the effect of identifying symmetric polynomials in the alphabet $\{x_1, \dots x_{a+b}\}$ on the $a+b$ facet with those in the alphabet $\{y_1, \dots, y_{a+b}\}$, which is the union of the alphabets on the other two facets. The 2–functor ι_λ maps these relations to relations that identify symmetric polynomials in $\{x_1 - \lambda, \dots, x_{a+b} - \lambda\}$ with symmetric polynomials in $\{y_1 - \lambda, \dots, y_{a+b} - \lambda\}$. They generate the same ideal, and hence are equivalent sets of relations.

- A minimal version of [36, (3.10)] is

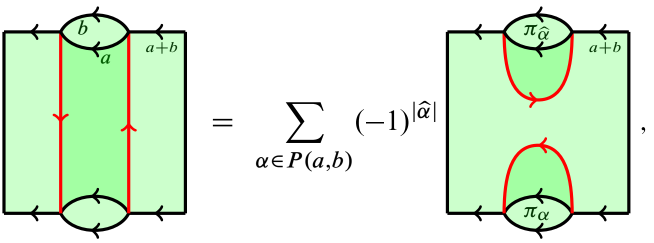


and under ι_λ it is sent to



where the final equality holds since all terms in the middle except the one with k dots is zero.

- Relations [36, (3.11)],



are sent by ι_λ to relations of the same form, where π_α and $\pi_{\hat{\alpha}}$ are now interpreted as symmetric polynomials in the new alphabet which is shifted by λ. Denote the alphabets on the facets in the original relation on which the decorations are placed by $\mathbb{X}$ and $\mathbb{Y}$, then we can write the decoration on the right-hand side of the original relation as

$$\sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}\pi_\alpha(\mathbb{X})\pi_{\hat{\alpha}}(\mathbb{Y}) = \prod_{x\in\mathbb{X}}\prod_{y\in\mathbb{Y}}(y-x).$$

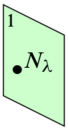
Under ι_λ this is sent to

$$\prod_{x\in\mathbb{X}}\prod_{y\in\mathbb{Y}}((y-\lambda)-(x-\lambda)) = \prod_{x\in\mathbb{X}}\prod_{y\in\mathbb{Y}}(y-x) = \sum_{\alpha\in P(a,b)} (-1)^{|\hat{\alpha}|}\pi_\alpha(\mathbb{X})\pi_{\hat{\alpha}}(\mathbb{Y}),$$

so ι_λ preserves the relation.

- Relations [36, (3.13) and (3.14)] are also preserved by ι_λ; the proofs are completely analogous to the case of [36, (3.11)].

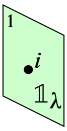
- The relation



$$= \quad 0$$

is mapped by ι_λ to

$$\sum_{i=0}^{N_\lambda}\binom{N_\lambda}{i}(-\lambda)^{N_\lambda-i}$$



$$= \quad 0,$$

which is precisely relation (5-4).

Furthermore, the functor ι_λ is clearly invertible (forget idempotents, shift decorations back) and similar arguments as above show that all relations in $(N\mathbf{Foam}^{\lambda\in\Sigma})^\wedge$ are sent by the inverse to relations of $N_\lambda\mathbf{Foam}^\bullet$. □

5.4 Proof of the decomposition theorem

For this section let $\mathcal{L}$ be an oriented, labeled link diagram. Recall that the complex $[\![\mathcal{L}]\!]^\Sigma$ over $N\mathbf{Foam}^\Sigma$ is, up to homotopy equivalence, an invariant of the corresponding oriented, framed, labeled link and each web appearing in $[\![\mathcal{L}]\!]^\Sigma$ has endomorphisms of $\boldsymbol{o}^{\mathrm{top}} := (N,\dots,N)$ as objects. We can view $[\![\mathcal{L}]\!]^\Sigma$ as a complex in $(N\mathbf{Foam}^\Sigma)^\wedge$ (since $N\mathbf{Foam}^\Sigma$ embeds as a full 2–subcategory) where it splits into a direct sum of complexes $[\![\mathcal{L}_f]\!]^\Sigma$, one for each coloring f of link components with a multisubset of Σ of the correct size. The objects of a summand $[\![\mathcal{L}_f]\!]^\Sigma$ are again endomorphisms of

$\boldsymbol{o}^{\mathrm{top}}$ in $(N\mathbf{Foam}^{\Sigma})^{\wedge}$, and there are natural isomorphisms of chain complexes of vector spaces:

$$\text{(5-8)}\quad \mathrm{taut}([\![\mathcal{L}]\!]^{\Sigma}) = \underbrace{\mathrm{Hom}(1_{\boldsymbol{o}^{\mathrm{top}}}, [\![\mathcal{L}]\!]^{\Sigma})}_{\text{over } N\mathbf{Foam}^{\Sigma}} \cong \underbrace{\bigoplus_f \mathrm{Hom}(1_{\boldsymbol{o}^{\mathrm{top}}}, [\![\mathcal{L}_f]\!]^{\Sigma})}_{\text{over } (N\mathbf{Foam}^{\Sigma})^{\wedge}} = \bigoplus_f \mathrm{taut}([\![\mathcal{L}_f]\!]^{\Sigma}).$$

It follows that taut$(-)$ respects the direct sum decomposition, and it remains to describe the summands $\mathrm{taut}([\![\mathcal{L}_f]\!]^{\Sigma})$.

The only nonzero coloring of the identity web $1_{\boldsymbol{o}^{\mathrm{top}}}$ is the one where every N–labeled strand is colored by the full multiset Σ. With respect to this coloring, we will use the object $\bigsqcup_{\lambda} \boldsymbol{o}_{\lambda}^{\mathrm{top}}$, the (co)domain of the associated split web $\bigsqcup_{\lambda}(1_{\boldsymbol{o}^{\mathrm{top}}})_{\lambda}$, and the object $\boldsymbol{o}_{\lambda}^{\mathrm{top}}$, the (co)domain of the λ–component $(1_{\boldsymbol{o}^{\mathrm{top}}})_{\lambda}$ of $1_{\boldsymbol{o}^{\mathrm{top}}}$. For endomorphism webs on these objects and foams between them, we define the representable functors $\mathrm{taut}_{\mathrm{split}}(-) := \mathrm{Hom}(\bigsqcup_{\lambda}(1_{\boldsymbol{o}^{\mathrm{top}}})_{\lambda}, -)$ and $\mathrm{taut}_{\lambda}(-) := \mathrm{Hom}((1_{\boldsymbol{o}^{\mathrm{top}}})_{\lambda}, -)$, respectively. Further, we now need the webs L and R for the object $\boldsymbol{o}^{\mathrm{top}}$ with the only possible incidence condition, as given in Definition 4.12.

Recall that $[\![\mathcal{L}_f]\!]^{\Sigma}$ is a complex of colored webs W and foams d between them. Then $\phi_1([\![\mathcal{L}_f]\!]^{\Sigma})$ is the complex consisting of webs $L \otimes W \otimes R$ and foams $\mathrm{id}_L \otimes d \otimes \mathrm{id}_R$ between them and $\phi([\![\mathcal{L}_f]\!]^{\Sigma}) = \phi_2\phi_1([\![\mathcal{L}_f]\!]^{\Sigma})$ is the complex consisting of webs $\bigsqcup_{\lambda} W_{\lambda}$ and foams $T_* \circ (\mathrm{id}_L \otimes d \otimes \mathrm{id}_R) \circ B_*$. We first show:

Lemma 5.18 *There are isomorphisms of chain complexes of vector spaces*

$$\text{(5-9)}\qquad \mathrm{taut}([\![\mathcal{L}_f]\!]^{\Sigma}) \cong \mathrm{taut}_{\mathrm{split}}\big(\phi_1([\![\mathcal{L}_f]\!]^{\Sigma})\big) \cong \mathrm{taut}_{\mathrm{split}}\big(\phi([\![\mathcal{L}_f]\!]^{\Sigma})\big).$$

Proof Proposition 4.25 provides the following isomorphisms between the objects of these chain complexes:

$$\begin{aligned}\mathrm{taut}(W) &= \mathrm{Hom}(1_{\boldsymbol{o}^{\mathrm{top}}}, W)\\ &\overset{\phi_1}{\cong} \mathrm{Hom}(L \otimes 1_{\boldsymbol{o}^{\mathrm{top}}} \otimes R, L \otimes W \otimes R)\\ &\cong \mathrm{Hom}\Big(\bigsqcup_{\lambda}(1_{\boldsymbol{o}^{\mathrm{top}}})_{\lambda}, L \otimes W \otimes R\Big) = \mathrm{taut}_{\mathrm{split}}(\phi_1(W))\\ &\cong \mathrm{Hom}\Big(\bigsqcup_{\lambda}(1_{\boldsymbol{o}^{\mathrm{top}}})_{\lambda}, \bigsqcup_{\lambda} W_{\lambda}\Big) = \mathrm{taut}_{\mathrm{split}}(\phi(W)),\end{aligned}$$

where the last two isomorphisms are given by composition with $B_{1_{\boldsymbol{o}^{\mathrm{top}}}}$ and T_W. Under these isomorphisms, the differentials transform as required for (5-9):

$$
\begin{aligned}
&\mathrm{taut}(d)\\
&\quad= \big(\mathrm{Hom}(1_{\boldsymbol{o}^{\mathrm{top}}}, W) \xrightarrow{d\circ} \mathrm{Hom}(1_{\boldsymbol{o}^{\mathrm{top}}}, W')\big)\\
&\quad\mapsto \big(\mathrm{Hom}(L \otimes 1_{\boldsymbol{o}^{\mathrm{top}}} \otimes R, L \otimes W \otimes R) \xrightarrow{(\mathrm{id}_L \otimes d \otimes \mathrm{id}_R)\circ} \mathrm{Hom}(L \otimes 1_{\boldsymbol{o}^{\mathrm{top}}} \otimes R, L \otimes W' \otimes R)\big)\\
&\quad\mapsto \left(\mathrm{Hom}\left(\bigsqcup_{\lambda} (1_{\boldsymbol{o}^{\mathrm{top}}})_{\lambda}, L \otimes W \otimes R\right) \xrightarrow{(\mathrm{id}_L \otimes d \otimes \mathrm{id}_R)\circ} \mathrm{Hom}\left(\bigsqcup_{\lambda} (1_{\boldsymbol{o}^{\mathrm{top}}})_{\lambda}, L \otimes W' \otimes R\right)\right)\\
&\quad\mapsto \left(\mathrm{Hom}\left(\bigsqcup_{\lambda} (1_{\boldsymbol{o}^{\mathrm{top}}})_{\lambda}, \bigsqcup_{\lambda} W_{\lambda}\right) \xrightarrow{T_* \circ (\mathrm{id}_L \otimes d \otimes \mathrm{id}_R) \circ B_* \circ} \mathrm{Hom}\left(\bigsqcup_{\lambda} (1_{\boldsymbol{o}^{\mathrm{top}}})_{\lambda}, \bigsqcup_{\lambda} W'_{\lambda}\right)\right)
\end{aligned}
$$

since the last two lines give $\mathrm{taut}_{\mathrm{split}}(\phi_1(d))$ and $\mathrm{taut}_{\mathrm{split}}(\phi(d))$, respectively. □

Theorem 5.14 shows that there is a consistent choice of isomorphism foams T_* and B_* in the definition of ϕ such that $\phi(\llbracket \mathcal{L}_f \rrbracket^{\Sigma}) = \bigotimes_{\lambda} \llbracket \mathcal{L}_{\lambda \in f} \rrbracket^{\Sigma}$. Now we have

$$
\begin{aligned}
\text{(5-10)}\qquad \mathrm{taut}(\llbracket \mathcal{L}_f \rrbracket^{\Sigma}) \cong \mathrm{taut}_{\mathrm{split}}\big(\phi(\llbracket \mathcal{L}_f \rrbracket^{\Sigma})\big) &= \mathrm{taut}_{\mathrm{split}}\left(\bigotimes_{\lambda} \llbracket \mathcal{L}_{\lambda \in f} \rrbracket^{\Sigma}\right)\\
&\cong \bigotimes_{\lambda} \mathrm{taut}_{\lambda}(\llbracket \mathcal{L}_{\lambda \in f} \rrbracket^{\Sigma}).
\end{aligned}
$$

The last isomorphism is clear from the definition of the two versions of taut and the tensor product structure given by disjoint union of webs and foams.

It remains to identify the tensorands $\mathrm{taut}_{\lambda}(\llbracket \mathcal{L}_{\lambda \in f} \rrbracket^{\Sigma})$. To this end, recall the notation $\mathcal{L}(a_1, \dots, a_k)$ introduced in the statement of Theorem 1.1, which makes explicit that we consider $\mathcal{L}$ with the i^{th} component labeled by the fundamental $\mathfrak{sl}_N$ representation $\bigwedge^{a_i} \mathbb{C}^N$. Let $b_{i,j}$ be the multiplicity of the root λ_j in the multisubset of Σ that the coloring f assigns to the i^{th} component of $\mathcal{L}$. Further, recall that N_j denotes the multiplicity of λ_j in Σ.

The complex $\llbracket \mathcal{L}_{\lambda_j \in f} \rrbracket^{\Sigma}$ is a complex over the 2–subcategory $(N\mathbf{Foam}^{\lambda_j \in \Sigma})^{\wedge}$. Under the isomorphism to the 2–category $N_j\mathbf{Foam}^{\bullet}$, which Proposition 5.17 established, this complex corresponds to the undeformed $\mathfrak{sl}_{N_j}$ complex $\llbracket \mathcal{L}(b_{1,j}, \dots, b_{k,j}) \rrbracket^{\{0,\dots,0\}}$ of the relabeled sublink $\mathcal{L}(b_{1,j}, \dots, b_{k,j})$. As we have seen in Remark 5.12, this correspondence preserves the homological grading. Clearly, $\mathrm{taut}_{\lambda_j}(\llbracket \mathcal{L}_{\lambda_j \in f} \rrbracket^{\Sigma})$ is isomorphic to the image of $\llbracket \mathcal{L}(b_{1,j}, \dots, b_{k,j}) \rrbracket^{\{0,\dots,0\}}$ under the appropriate representable functor, and the homology of this complex is $\mathrm{KhR}^{\mathfrak{sl}_{N_j}}(\mathcal{L}(b_{1,j}, \dots, b_{k,j}))$. Finally, by (5-10), $\mathrm{taut}(\llbracket \mathcal{L}_f \rrbracket^{\Sigma})$ is isomorphic to the tensor product of these complexes, and since we are working over $\mathbb{C}$, the Künneth theorem gives that the homology of this tensor product complex is isomorphic to the tensor product of the respective homologies. This completes the proof of Theorem 1.1.

References

[1] **D Bar-Natan**, *Khovanov's homology for tangles and cobordisms*, Geom. Topol. 9 (2005) 1443–1499 MR

[2] **D Bar-Natan**, **S Morrison**, *The Karoubi envelope and Lee's degeneration of Khovanov homology*, preprint (2009) arXiv

[3] **H Becker**, *Khovanov–Rozansky homology via Cohen–Macaulay approximations and Soergel bimodules*, preprint (2011) arXiv

[4] **N Carqueville**, **D Murfet**, *Computing Khovanov–Rozansky homology and defect fusion*, Algebr. Geom. Topol. 14 (2014) 489–537 MR

[5] **N Carqueville**, **D Murfet**, *Adjunctions and defects in Landau–Ginzburg models*, Adv. Math. 289 (2016) 480–566 MR

[6] **S Cautis**, *Clasp technology to knot homology via the affine Grassmannian*, Math. Ann. 363 (2015) 1053–1115 MR

[7] **S Cautis**, **J Kamnitzer**, **A Licata**, *Categorical geometric skew Howe duality*, Invent. Math. 180 (2010) 111–159 MR

[8] **S Cautis**, **J Kamnitzer**, **S Morrison**, *Webs and quantum skew Howe duality*, Math. Ann. 360 (2014) 351–390 MR

[9] **J Chuang**, **R Rouquier**, *Derived equivalences for symmetric groups and $\mathfrak{sl}_2$–categorification*, Ann. of Math. 167 (2008) 245–298 MR

[10] **N M Dunfield**, **S Gukov**, **J Rasmussen**, *The superpolynomial for knot homologies*, Experiment. Math. 15 (2006) 129–159 MR

[11] **T Dyckerhoff**, **D Murfet**, *Pushing forward matrix factorizations*, Duke Math. J. 162 (2013) 1249–1311 MR

[12] **B Gornik**, *Note on Khovanov link cohomology*, preprint (2004) arXiv

[13] **E Gorsky**, **S Gukov**, **M Stošić**, *Quadruply-graded colored homology of knots*, preprint (2013) arXiv

[14] **S Gukov**, **M Stošić**, *Homological algebra of knots and BPS states*, from "Proceedings of the Freedman Fest" (R Kirby, V Krushkal, Z Wang, editors), Geom. Topol. Monogr. 18 (2012) 309–367 MR

[15] **S Gukov**, **J Walcher**, *Matrix factorizations and Kauffman homology*, preprint (2005) arXiv

[16] **M Khovanov**, *A categorification of the Jones polynomial*, Duke Math. J. 101 (2000) 359–426 MR

[17] **M Khovanov**, *Link homology and Frobenius extensions*, Fund. Math. 190 (2006) 179–190 MR

[18] **M Khovanov**, **A D Lauda**, *A diagrammatic approach to categorification of quantum groups, I*, Represent. Theory 13 (2009) 309–347 MR

[19] **M Khovanov**, **A D Lauda**, *A categorification of quantum* $\mathrm{sl}(n)$, Quantum Topol. 1 (2010) 1–92 MR

[20] **M Khovanov**, **A D Lauda**, *A diagrammatic approach to categorification of quantum groups, II*, Trans. Amer. Math. Soc. 363 (2011) 2685–2700 MR

[21] **M Khovanov**, **A D Lauda**, **M Mackaay**, **M Stošić**, *Extended graphical calculus for categorified quantum* $\mathrm{sl}(2)$, Mem. Amer. Math. Soc. 1029, Amer. Math. Soc., Providence, RI (2012) MR

[22] **M Khovanov**, **L Rozansky**, *Matrix factorizations and link homology*, Fund. Math. 199 (2008) 1–91 MR

[23] **D Krasner**, *Equivariant* $\mathrm{sl}(n)$*–link homology*, Algebr. Geom. Topol. 10 (2010) 1–32 MR

[24] **A D Lauda**, *A categorification of quantum* $\mathrm{sl}(2)$, Adv. Math. 225 (2010) 3327–3424 MR

[25] **A D Lauda**, *An introduction to diagrammatic algebra and categorified quantum* $\mathfrak{sl}_2$, Bull. Inst. Math. Acad. Sin. 7 (2012) 165–270 MR

[26] **E S Lee**, *An endomorphism of the Khovanov invariant*, Adv. Math. 197 (2005) 554–586 MR

[27] **L Lewark**, *Rasmussen's spectral sequences and the* $\mathfrak{sl}_N$*–concordance invariants*, Adv. Math. 260 (2014) 59–83 MR

[28] **A Lobb**, *A slice genus lower bound from* $\mathrm{sl}(n)$ *Khovanov–Rozansky homology*, Adv. Math. 222 (2009) 1220–1276 MR

[29] **A Lobb**, *A note on Gornik's perturbation of Khovanov–Rozansky homology*, Algebr. Geom. Topol. 12 (2012) 293–305 MR

[30] **G Lusztig**, *Introduction to quantum groups*, Progress in Mathematics 110, Birkhäuser, Boston (1993) MR

[31] **I G Macdonald**, *Symmetric functions and Hall polynomials*, 2nd edition, Clarendon, New York (1995) MR

[32] **M Mackaay**, *The* $\mathfrak{sl}_N$*–web algebras and dual canonical bases*, J. Algebra 409 (2014) 54–100 MR

[33] **M Mackaay**, **W Pan**, **D Tubbenhauer**, *The* $\mathfrak{sl}_3$*–web algebra*, Math. Z. 277 (2014) 401–479 MR

[34] **M Mackaay**, **P Vaz**, *The universal* sl_3*–link homology*, Algebr. Geom. Topol. 7 (2007) 1135–1169 MR

[35] **M Mackaay**, **Y Yonezawa**, $\mathfrak{sl}(N)$*–web categories*, preprint (2013) arXiv

[36] **H Queffelec**, **D E V Rose**, *The* $\mathfrak{sl}_n$ *foam* 2*–category: a combinatorial formulation of Khovanov–Rozansky homology via categorical skew Howe duality*, preprint (2014) arXiv

[37] **J Rasmussen**, *Khovanov homology and the slice genus*, Invent. Math. 182 (2010) 419–447 MR

[38] **J Rasmussen**, *Some differentials on Khovanov–Rozansky homology*, Geom. Topol. 19 (2015) 3031–3104 MR

[39] **R Rouquier**, 2*–Kac–Moody algebras*, preprint (2008) arXiv

[40] **B Webster**, *Khovanov–Rozansky homology via a canopolis formalism*, Algebr. Geom. Topol. 7 (2007) 673–699 MR

[41] **H Wu**, *Generic deformations of the colored* $\mathfrak{sl}(N)$*–homology for links*, Algebr. Geom. Topol. 11 (2011) 2037–2106 MR

[42] **H Wu**, *Equivariant colored* $\mathfrak{sl}(N)$*–homology for links*, J. Knot Theory Ramifications 21 (2012) art. id. 1250012, 104 pp MR

[43] **H Wu**, *A colored* $\mathfrak{sl}(N)$ *homology for links in* S^3, Dissertationes Math. (Rozprawy Mat.) 499 (2014) 217 MR

[44] **Y Yonezawa**, *Quantum* $(\mathfrak{sl}_n, \wedge V_n)$ *link invariant and matrix factorizations*, Nagoya Math. J. 204 (2011) 69–123 MR

University of North Carolina at Chapel Hill, Mathematics Department
120 E Cameron Avenue, CB #3250, 329 Phillips Hall, Chapel Hill, NC 27599, United States

Department of Mathematics, Imperial College London
London, SW7 2AZ, United Kingdom

davidrose@unc.edu, p.wedrich@gmail.com

http://www.unc.edu/~davidev/, http://paul.wedrich.at

Proposed: Ciprian Manolescu Received: 10 May 2015
Seconded: Tomasz Mrowka, Ronald Stern Revised: 6 October 2015

msp

Geometry & Topology 20 (2016) 3519–3569

Cylindrical contact homology and topological entropy

MARCELO R R ALVES

We establish a relation between the growth of the cylindrical contact homology of a contact manifold and the topological entropy of Reeb flows on this manifold. We show that if a contact manifold (M,ξ) admits a hypertight contact form λ_0 for which the cylindrical contact homology has exponential homotopical growth rate, then the Reeb flow of every contact form on (M,ξ) has positive topological entropy. Using this result, we provide numerous new examples of contact 3–manifolds on which every Reeb flow has positive topological entropy.

37B40, 53D35, 53D42, 37J05

1 Introduction

The aim of this paper is to establish a relation between the behaviour of cylindrical contact homology and the topological entropy of Reeb flows. The topological entropy is a nonnegative number associated to a dynamical system which measures the complexity of the orbit structure of the system. Positivity of the topological entropy means that the system possesses some type of exponential instability. We show that if the cylindrical contact homology of a contact 3–manifold is "complicated enough" from a homotopical viewpoint, then every Reeb flow on this contact manifold has positive topological entropy.

1.1 Basic definitions and history of the problem

We first recall some basic definitions from contact geometry. A 1–form λ on a $(2n+1)$–dimensional manifold Y is called a *contact form* if $\lambda \wedge (d\lambda)^n$ is a volume form on Y. The hyperplane $\xi = \ker\lambda$ is called the *contact structure*. For us a *contact manifold* will be a pair (Y,ξ) such that ξ is the kernel of some contact form λ on Y (these are usually called co-oriented contact manifolds in the literature). When λ satisfies $\xi = \ker\lambda$, we will say that λ is a contact form on (Y,ξ). On any contact manifold there always exist infinitely many different contact forms. Given a contact form λ, its *Reeb vector field* is the unique vector field X_λ satisfying $\lambda(X_\lambda)=1$ and $i_{X_\lambda} d\lambda = 0$. The *Reeb flow* ϕ_{X_λ} of λ is the flow generated by the vector field X_λ. We will refer

Published: 21 December 2016 DOI: 10.2140/gt.2016.20.3519

to the periodic orbits of ϕ_{X_λ} as *Reeb orbits* of λ. The action $A(\gamma)$ of a Reeb orbit is defined by $A(\gamma) := \int_\gamma \lambda$.

We study the topological entropy of Reeb flows from the point of view of contact topology. More precisely, we search for conditions on the topology of a contact manifold (M,ξ) that force *all* Reeb flows on (M,ξ) to have positive topological entropy. The condition we impose is on the behaviour of a contact topological invariant called cylindrical contact homology. We show that if a contact manifold (M,ξ) admits a contact form λ_0 for which the cylindrical contact homology has *exponential homotopical growth*, then all Reeb flows on (M,ξ) have positive topological entropy.

The notion of exponential homotopical growth of cylindrical contact homology, which is introduced in this paper, differs from the notion of growth of contact homology studied by Colin and Honda [12] and by Vaugon [40]. For reasons explained in Section 2, the growth of contact homology is not well adapted to study the topological entropy of Reeb flows, while the notion of homotopical growth rate is (as we show) well suited for this purpose. We begin by explaining the results which were previously known relating the behaviour of contact topological invariants to the topological entropy of Reeb flow.

The study of contact manifolds all of whose Reeb flows have positive topological entropy was initiated by Macarini and Schlenk [36]. They showed that if Q is an energy hyperbolic manifold and ξ_{geo} is the contact structure on the unit tangent bundle T_1Q associated to the geodesic flows, then every Reeb flow on (T_1Q,ξ_{geo}) has positive topological entropy. Their work was based on previous ideas of Frauenfelder and Schlenk [20; 21] which related the growth rate of Lagrangian Floer homology to entropy invariants of symplectomorphisms. The strategy to estimate the topological entropy used in [36] can be briefly sketched as follows:

$$\begin{gathered}\text{Exponential growth of Lagrangian Floer homology of the tangent fibre } (TQ)|_p\\ \Downarrow\\ \text{Exponential volume growth of the unit tangent fibre } (T_1Q)|_p\\ \text{for all Reeb flows in } (T_1Q,\xi_{\text{geo}})\\ \Downarrow\\ \text{Positivity of the topological entropy for all Reeb flows in } (T_1Q,\xi_{\text{geo}}).\end{gathered}$$

To obtain the first implication, Macarini and Schlenk use the fact that (T_1Q,ξ_{geo}) has the structure of a Legendrian fibration, and apply the geometric idea of [20; 21] to show that the number of trajectories connecting a Legendrian fibre to another Legendrian fibre can be used to obtain a volume growth estimate. The second implication in this scheme follows from Yomdin's theorem, which states that exponential volume growth of a submanifold implies positivity of topological entropy.[1]

In the author's Ph D thesis [2; 3], this approach was extended to deal with 3–dimensional contact manifolds which are not unit tangent bundles. This was done by designing a localized version of the geometric idea of [20; 21]. Globally most contact 3–manifolds are not Legendrian fibrations, but a sufficiently small neighbourhood of a given Legendrian knot in a contact 3–manifold can always be given the structure of a Legendrian fibration. It turns out that this is enough to conclude that if the linearized Legendrian contact homology of a pair of Legendrian knots in a contact 3–manifold (M^3, ξ) grows exponentially, then the length of these Legendrian knots grows exponentially for any Reeb flow on (M^3, ξ). We then apply Yomdin's theorem to obtain that all Reeb flows on (M^3, ξ) have positive topological entropy.

One drawback of these approaches is that they only give lower entropy bounds for C^∞–smooth Reeb flows. The reason is that Yomdin's theorem holds only for C^∞–smooth flows. The approach presented in the present paper *does not* use Yomdin's theorem and gives lower bounds for the topological entropy of C^1–smooth Reeb flows.

Another advantage is that the cylindrical contact homology is usually easier to compute than the linearized Legendrian contact homology. In fact, to apply the strategy of [2; 3] to a contact 3–manifold (M^3, ξ), one must first find a pair of Legendrian curves which "should" have exponential growth of linearized Legendrian contact homology. This is highly nontrivial since on any contact 3–manifolds there exist many Legendrian links for which the linearized Legendrian contact homology does not even exist. On the other hand, the definition of cylindrical contact homology involves only the contact manifold (M^3, ξ), and no Legendrian submanifolds.

1.2 Main results

Our results are inspired by the philosophy that a "complicated" topological structure should force chaotic behaviour for dynamical systems associated to this structure. Two important examples of this phenomenon are: the fact that on manifolds with complicated loop space the geodesic flow always has positive topological entropy (see Paternain [38]), and the fact that every diffeomorphism of a surface which is isotopic to a pseudo-Anosov diffeomorphism has positive topological entropy (see Fel'shtyn [16]).

To state our results we introduce some notation. Let M be a manifold and X be a C^k $(k \geq 1)$ vector field. Our first result relates the topological entropy of the flow ϕ_X to

[1] The same scheme was used by Frauenfelder and Schlenk [22] and by Frauenfelder, Labrousse and Schlenk [19] to obtain positive lower bounds for the intermediate and slow entropies of Reeb flows on unit tangent bundles.

the growth (relative to T) of the number of distinct homotopy classes which contain periodic orbits of ϕ_X with period at most T. More precisely, let Λ_X^T be the set of free homotopy classes of M which contain a periodic orbit of ϕ_X with period at most T. We denote by $N_X(T)$ the cardinality of Λ_X^T.

Theorem 1 *If for real numbers $a > 0$ and b there is a sequence $T_n \to +\infty$ such that*

$$N_X(T_n) \geq e^{aT_n+b}$$

for all T_n, then $h_{\mathrm{top}}(\phi_X) \geq a$.

Theorem 1 might be a folklore result in the theory of dynamical systems. However, as we have not found it in the literature, we provide a complete proof in Section 2. It contains as a special case Ivanov's inequality for surface diffeomorphisms; see Jiang [31]. Our motivation for proving this result is to apply it to Reeb flows. Contact homology allows one to carry over information about the dynamical behaviour of one special Reeb flow on a contact manifold to all other Reeb flows on the same contact manifold. In Section 4, we introduce the notion of exponential homotopical growth of cylindrical contact homology. As we already mentioned, this growth rate differs from the ones previously considered in the literature and is specially designed to allow one to use Theorem 1 to obtain results about the topological entropy of Reeb flows. Recall that a contact form is called hypertight if its Reeb flow has no contractible closed orbits. We prove the following result:

Theorem 8 *Let λ_0 be a hypertight contact form on a contact manifold (M, ξ), and assume that the cylindrical contact homology of λ_0 has exponential homotopical growth with exponential weight $a > 0$. Then for every C^k $(k \geq 2)$ contact form λ on (M, ξ), the Reeb flow of X_λ has positive topological entropy. More precisely, if f_λ is the function such that $\lambda = f_\lambda \lambda_0$, then*

$$h_{\mathrm{top}}(\phi_{X_\lambda}) \geq \frac{a}{\max f_\lambda}. \tag{1-1}$$

Notice that Theorem 8 allows us to conclude the positivity of the topological entropy for *all* Reeb flows on a given contact manifold (M, ξ), once we show that (M, ξ) admits one special hypertight contact form for which the cylindrical contact homology has exponential homotopical growth. It is worth remarking that our proof of Theorem 8 is carried out in full rigour, and does *not* make use of the polyfold technology which is being developed by Hofer, Wysocki and Zehnder. The reason is that we do not use the linearized contact homology considered by Bourgeois, Ekholm and Eliashberg [7] and Vaugon [40], but resort to a topological idea used by Hryniewicz, Momin and Salomão [30] to prove existence of Reeb orbits in prescribed homotopy classes.

Theorem 8 allows one to obtain estimates for the topological entropy for C^1–smooth Reeb flows. As previously observed, the strategy used in [36; 2; 3] produces estimates for the topological entropy only for C^∞–smooth contact forms as they depend on Yomdin's theorem, which fails for finite regularity.

Our other results are concerned with the existence of examples of contact manifolds which have a contact form with exponential homotopical growth rate of cylindrical contact homology. We show that in dimension 3 they exist in abundance, and it follows from Theorem 8 that every Reeb flow on these contact manifolds has positive topological entropy. In Section 5, we use a construction of Colin and Honda [12] to obtain many such examples of contact 3–manifolds. In these examples, the underlying differentiable 3–manifold has nontrivial JSJ decomposition and a hyperbolic component that fibres over the circle.

Theorem 9 *Let M be a closed connected oriented 3–manifold which can be cut along a nonempty family of incompressible tori into a family $\{M_i,\ 0 \le i \le q\}$ of irreducible manifolds with boundary such that*

- *M_0 is the mapping torus of a diffeomorphism $h\colon S \to S$ with pseudo-Anosov monodromy on a surface S with nonempty boundary.*

Then M can be given infinitely many nondiffeomorphic contact structures ξ_k such that for each ξ_k, there exists a hypertight contact form λ_k on (M, ξ_k) which has exponential homotopical growth of cylindrical contact homology. It follows that on each (M, ξ_k), all Reeb flows have positive topological entropy.

The contact structures studied in Theorem 9 are among the tight contact structures constructed by Colin and Honda [12] in closed connected irreducible toroidal 3–manifolds.

In Section 6, we study the cylindrical contact homology of contact 3–manifolds $(M, \xi_{(q,\mathfrak{r})})$ obtained via a special integral Dehn surgery on the unit tangent bundle $(T_1S, \xi_{\mathrm{geo}})$ of a hyperbolic surface (S, g). This Dehn surgery is performed on a neighbourhood of a Legendrian curve $L_{\mathfrak{r}}$ which is the Legendrian lift of a simple closed separating geodesic $\mathfrak{r}$. The surgery we consider is the contact version of Handel–Thurston surgery, which was introduced by Foulon and Hasselblatt in [18] to produce nonalgebraic Anosov Reeb flows on 3–manifolds. We call this contact surgery the Foulon–Hasselblatt surgery. This surgery produces not only a contact 3–manifold $(M, \xi_{(q,\mathfrak{r})})$, but also a special contact form, which we denote by λ_{FH}, on $(M, \xi_{(q,\mathfrak{r})})$. In [18], the authors restrict their attention to integer surgeries with positive surgery coefficient q and prove that, in this case, the Reeb flow of λ_{FH} is Anosov. Our methods also work for negative coefficients as the Anosov condition on λ_{FH} does not play a role in our results. We obtain:

Theorem 16 *Let $(M, \xi_{(q,\mathfrak{r})})$ be the contact manifold obtained from performing the Foulon–Hasselblat q–surgery on the Legendrian curve $L_{\mathfrak{r}} \subset (T_1 S, \xi_{\text{geo}})$, and denote by λ_{FH} the contact form on $(M, \xi_{(q,\mathfrak{r})})$ obtained from this surgery. Then λ_{FH} is hypertight, and its cylindrical contact homology has exponential homotopical growth. It follows that every Reeb flow on $(M, \xi_{(q,\mathfrak{r})})$ has positive topological entropy.*

Organization of the paper In Section 2, we recall one of the definitions of the topological entropy and present the proof of Theorem 1. In Section 3, we recall the definition of cylindrical contact homology and its basic properties. In Section 4, we introduce the notion of exponential homotopical growth of cylindrical contact homology and prove Theorem 8. Section 5 is devoted to the proof of Theorem 9. In Section 6, we present the definition of the integral Foulon–Hasselblatt surgery and prove Theorem 16. In Section 7, we discuss the results obtained in this paper and propose some questions for future research.

Remark We again would like to point out that all the results above *do not* depend on the polyfolds technology which is being developed Hofer, Wysocki and Zehnder. This is the case because the versions of contact homology used for proving the results above involve only somewhere injective pseudoholomorphic curves. In this situation, transversality can be achieved by "classical" perturbation methods as in Dragnev [13].

Acknowledgements I especially thank my advisors Frédéric Bourgeois and Chris Wendl for their guidance, support and our many discussions which were crucial for the development of this paper, which is a part of my Ph D thesis developed under their supervision. I would like to thank professor Pedro Salomão for many helpful discussions and for explaining to me the techniques used in [30] which made it possible to avoid dealing with transversality problems arising from multiply covered pseudoholomorphic curves. My thanks to professor Felix Schlenk for his interest in this work, his suggestions for improving the exposition and for suggesting many directions for future work. My personal thanks to Ana Nechita, André Alves, Hilda Ribeiro and Lucio Alves for their unconditional personal support. This work was partially supported by a FRIA fellowship from FNRS-Belgium.

2 Homotopic growth of periodic orbits and topological entropy

Throughout this section, M will denote a compact manifold. We endow M with an auxiliary Riemannian metric g, which induces a distance function d_g on M, whose injectivity radius we denote by ϵ_g. Let $\widetilde{M}$ be the universal cover of M, $\tilde{g}$ be the Riemannian metric that makes the covering map $\pi\colon \widetilde{M} \to M$ an isometry, and $d_{\tilde{g}}$ be the distance induced by the metric $\tilde{g}$.

Let X be a vector field on M with no singularities and ϕ_X^t the flow generated by X. We call $P^X(T)$ the number of periodic orbits of ϕ^t with period in $[0,T]$. For us, a periodic orbit of X is a pair $([\gamma]_c, T)$, where $[\gamma]_c$ is the set of parametrizations of a given *immersed* curve $c\colon S^1 \to M$, and T is a positive real number (called the period of the orbit), such that

- $\gamma \in [\gamma]_c$ if and only if $\gamma\colon \mathbb{R} \to M$ parametrizes c and $\dot{\gamma}(t) = X(\gamma(t))$,
- for all $\gamma \in [\gamma]_c$, we have $\gamma(T+t) = \gamma(t)$ and $\gamma([0,T]) = c$.

We say that a periodic orbit $([\gamma]_c, T)$ is in a free homotopy class l of M if $c \in l$.

By a parametrized periodic orbit (γ, T) we mean a periodic orbit $([\gamma]_c, T)$ with a fixed choice of parametrization $\gamma \in [\gamma]_c$. A parametrized periodic orbit (γ, T) is said to be in a free homotopy class l when the underlying periodic orbit $([\gamma]_c, T)$ is in l.

We now recall a definition of topological entropy due to Bowen [10] which will be very useful for us. Let T and δ be positive real numbers. A set S is said to be T,δ–separated if for all $q_1 \neq q_2 \in S$, we have

$$\max_{t\in[0,T]} d_g\big(\phi_X^t(q_1), \phi_X^t(q_2)\big) > \delta. \tag{2-1}$$

We denote by $n^{T,\delta}$ the maximal cardinality of a T,δ–separated set for the flow ϕ_X. Then we define the δ–entropy $h_\delta(\phi_X)$ as

$$h_\delta(\phi_X) = \limsup_{T\to+\infty} \frac{\log(n^{T,\delta})}{T}. \tag{2-2}$$

The topological entropy h_{top} is then defined by

$$h_{\mathrm{top}}(\phi_X) = \lim_{\delta\to 0} h_\delta(\phi_X).$$

One can prove that the topological entropy does not depend on the metric d_g but only on the topology determined by the metric. For these and other structural results about topological entropy, we refer the reader to any standard textbook in dynamics such as [34] and [39].

From the work of Kaloshin and others it is well known that the exponential growth rate of periodic orbits,

$$\limsup_{T\to+\infty} \frac{\log(P^X(T))}{T}, \tag{2-3}$$

can be much bigger than the topological entropy. This implies that the growth rate (2-3) does not give a lower bound for the topological entropy of an arbitrary flow. There is, however, a different growth rate that measures how quickly periodic orbits appear in different free homotopy classes, which can be used to give such a lower bound of the topological entropy of a flow.

Let Λ denote the set of free homotopy classes of loops in M, and $\Lambda_0 \subset \Lambda$ the subset of primitive free homotopy classes. We define the set $\Lambda_X^T \subset \Lambda$ in the following way: $\varrho \in \Lambda_X^T$ if and only if there exists a periodic orbit of ϕ_X^t with period at most T that belongs to ϱ. We denote by $N_X(T)$ the cardinality of Λ_X^T.

Let $\{(\gamma_i, T_i) : 1 \leq i \leq n\}$ be a finite set of parametrized periodic orbits of X. For a number T satisfying $T \geq T_i$ for all $i \in \{1, \dots, n\}$ and a constant $\delta > 0$, we denote by $\Lambda_X^{T,\delta}((\gamma_1, T_1), \dots, (\gamma_n, T_n))$ the subset of Λ such that

- $l \in \Lambda_X^{T,\delta}((\gamma_1, T_1), \dots, (\gamma_n, T_n))$ if and only if there exist a parametrized periodic orbit $(\widehat{\gamma}, \widehat{T})$ with period $\widehat{T} \leq T$ in the free homotopy class l and a number $i_l \in \{1, \dots, n\}$ for which $\max_{t\in[0,T]}\bigl(d_g(\gamma_{i_l}(t), \widehat{\gamma}(t))\bigr) \leq \delta$.

Notice that

$$\Lambda_X^{T,\delta}((\gamma_1, T_1), \dots, (\gamma_n, T_n)) = \bigcup_{i\in\{1,\dots,n\}} \Lambda_X^{T,\delta}((\gamma_i, T_i)). \tag{2-4}$$

We are ready to prove the main result in this section. Theorem 1 below is well known to be true in the particular cases when ϕ_X is a geodesic flow, where it follows from Manning's inequality (see [33] and [38]), and when ϕ_X is the suspension of a surface diffeomorphism with pseudo-Anosov monodromy, where it follows from Ivanov's theorem (see [31]). It can be seen as a generalization of these results in the sense that it includes them as particular cases and that it applies to many other situations. Our argument is inspired by the remarkable proof of Ivanov's inequality given by Jiang in [31].

Theorem 1 *If for real numbers $a > 0$ and b there is a sequence $T_n \to +\infty$ such that*

$$N_X(T_n) \geq e^{aT_n+b}$$

for all T_n, then $h_{\mathrm{top}}(\phi_X) \geq a$.

Proof The theorem will follow if we prove that for all $0 < \delta < \epsilon_g/32$, we have $h_\delta(\phi_X) \geq a$. From now on, fix $0 < \delta < \epsilon_g/32$.

Step 1 For any point $p \in M$, let $V_{4\delta}(p)$ be the 4δ–neighbourhood of $\pi^{-1}(p)$. Because $\delta < \epsilon_g/32$, it is clear that $V_{4\delta}(p)$ is the disjoint union

$$V_{4\delta}(p) = \bigcup_{\widetilde{p}\in\pi^{-1}(p)} B_{4\delta}(\widetilde{p}), \tag{2-5}$$

where the ball $B_{4\delta}(\widetilde{p})$ is taken with respect to the metric $\widetilde{g}$.

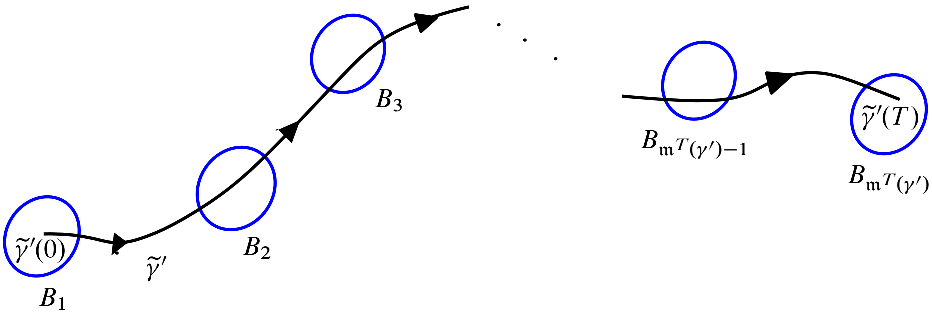


Figure 1: The set $\{B_j : 1 \leq j \leq \mathfrak{m}^T(\gamma', T')\}$

Because of our choice of $\delta < \epsilon_g/32$, it is clear that there exists a constant $k_1 > 0$, which does not depend on p, such that if B and B' are two distinct connected components of $V_{4\delta}(p)$, we have $d_{\widetilde{g}}(B, B') > k_1$.

Because of compactness of M, we know that the vector field $\widetilde{X} := \pi^* X$ is bounded in the norm given by the metric $\widetilde{g}$. Combining this with the inequality in the last paragraph, one obtains the existence of a constant $k_2 > 0$, which again does not depend on p, such that if $\widetilde{\upsilon}\colon [0, R] \to \widetilde{M}$ is a parametrized trajectory of $\phi_{\widetilde{X}}$ with $\widetilde{\upsilon}(0) \in B$ and $\widetilde{\upsilon}(R) \in B'$, then $R > k_2$.

From the last assertion, we deduce the existence of a constant $\widetilde{K}$, depending only g and X, such that for every $p \in M$ and every parametrized trajectory $\widetilde{\upsilon}\colon [0, T] \to \widetilde{M}$ of $\phi_{\widetilde{X}}$, the number $L^T(p, \widetilde{\upsilon})$ of distinct connected components of $V_{4\delta}(p)$ intersected by the curve $\widetilde{\upsilon}([0, T])$ satisfies

$$L^T(p, \widetilde{\upsilon}) < \widetilde{K}T + 1. \tag{2-6}$$

Step 2 We claim that for every parametrized periodic orbit (γ', T') of X, we have

$$\#\big(\Lambda_X^{T,\delta}((\gamma', T'))\big) < \widetilde{K}T + 1 \tag{2-7}$$

for all $T > T'$.

To see this, take a lift $\widetilde{\gamma}'$ of γ' to $\widetilde{M}$, and let $p' = \gamma'(0)$ and $\widetilde{p}' = \widetilde{\gamma}'(0)$. We consider (see Figure 1) the set $\{B_j : 1 \leq j \leq \mathfrak{m}^T(\gamma', T')\}$ of connected components of $V_{4\delta}(p')$ satisfying:

- $B_j \neq B_k$ if $j \neq k$,
- if B is a connected component of $V_{4\delta}(p')$ which intersects $\widetilde{\gamma}'([0, T])$, then $B = B_j$ for some $j \in \{1, \dots, \mathfrak{m}^T(\gamma', T')\}$,
- if $j < i$, then B_j is visited by the trajectory $\widetilde{\gamma}'\colon [0, T] \to \widetilde{M}$ before B_i.

From step 1, we know that $\mathfrak{m}^T(\gamma', T') < \widetilde{K}T + 1$.

For each $l \in \Lambda_X^{T,\delta}((\gamma', T'))$, pick a parametrized periodic orbit (χ_l, T_l) in l which satisfies $d_g(\chi_l(t), \gamma'(t)) < \delta$ for all $t \in [0, T]$. There exists a lift $\widetilde{\chi}_l$ of χ_l satisfying $d_{\widetilde{g}}(\widetilde{\chi}_l(t), \widetilde{\gamma}'(t)) < \delta$ for all $t \in [0, T]$.

From the triangle inequality, it is clear that the point $q_l = \widetilde{\chi}_l(0)$ is in the connected component B_1 which contains $\widetilde{p}'$. We will show that $\widetilde{\chi}_l(T_l)$ is contained in B_j for some $j \in \{1, \dots, \mathfrak{m}^T(\gamma')\}$. Because $\pi(\widetilde{\chi}_l(0)) = \pi(\widetilde{\chi}_l(T_l))$, we have

$$d_{\widetilde{g}}\big(\widetilde{\chi}_l(T_l), \pi^{-1}(p')\big) = d_{\widetilde{g}}\big(\widetilde{\chi}_l(0), \pi^{-1}(p')\big) < \delta, \tag{2-8}$$

which already implies that $\widetilde{\chi}_l(T_l) \in V_{4\delta}(p')$. We denote by $\widetilde{p}_l{}'$ the unique element in $\pi^{-1}(p')$ for which we have $d_{\widetilde{g}}(\widetilde{\chi}_l(T_l), \widetilde{p}_l{}') < \delta$. Using the triangle inequality we now obtain

$$d_{\widetilde{g}}\big(\widetilde{\gamma}'(T_l), \widetilde{p}_l{}'\big) \leq d_{\widetilde{g}}\big(\widetilde{\gamma}'(T_l), \widetilde{\chi}_l(T_l)\big) + d_{\widetilde{g}}\big(\widetilde{\chi}_l(T_l), \widetilde{p}_l{}'\big) < \delta + \delta.$$

From the inequalities above we conclude that $\widetilde{\gamma}'(T_l)$ and $\widetilde{\chi}_l(T_l)$ are in the connected component of $V_{4\delta}(p')$ that contains $\widetilde{p}_l{}'$. Because this connected component contains $\widetilde{\gamma}'(T_l)$, it is therefore one of the B_j for $j \in \{1, \dots, \mathfrak{m}^T(\gamma', T')\}$ as we wanted to show. We can thus define a map

$$\Upsilon^{T,\delta}_{(\gamma',T')} \colon \Lambda_X^{T,\delta}((\gamma', T')) \to \{1, \dots, \mathfrak{m}^T((\gamma', T'))\} \tag{2-9}$$

which associates to each $l \in \Lambda_X^{T,\delta}(\gamma')$ the unique $j \in \{1, \dots, \mathfrak{m}^T(\gamma', T')\}$ for which $\widetilde{\chi}_l(T_l) \in B_j$.

We now claim that if $l \neq l'$, then $\widetilde{\chi}_l(T_l)$ and $\widetilde{\chi}_{l'}(T_{l'})$ are in different connected components of $V_{4\delta}(p')$. To see this, notice that both $\widetilde{\chi}_l(0)$ and $\widetilde{\chi}_{l'}(0)$ are in the component B_1. Therefore, it is clear, because $\delta < \epsilon_g/32$, that if $\widetilde{\chi}_l(T_l)$ and $\widetilde{\chi}_{l'}(T_{l'})$ are in the same component of $V_{4\delta}(p')$, then the closed curves $\chi_l([0, T_l])$ and $\chi_{l'}([0, T_{l'}])$ are freely homotopic. This contradicts our choice of (χ_l, T_l) and $(\chi_{l'}, T_{l'})$ and the fact that $l \neq l'$.

We thus conclude that the map (2-9) is injective, which implies that $\#(\Lambda_X^{T,\delta}((\gamma', T'))) \leq \mathfrak{m}^T(\gamma', T') < \widetilde{K}T + 1$.

Step 3 (inductive step) As an immediate consequence of step 2, we have that if $\{(\gamma_i, T_i) : 1 \leq i \leq m\}$ is a set of parametrized periodic orbits of X, we have

$$\#(\Lambda_X^{T,\delta}((\gamma_1, T_1), \dots, (\gamma_m, T_m))) \leq m(\widetilde{K}T + 1).$$

Inductive claim *Fix $T > 0$, and suppose that $S_m^T = \{(\gamma_i, T_i) : 1 \le i \le m\}$ is a set of parametrized periodic orbits such that $T \ge T_i$ for every $i \in \{1, \dots, m\}$, and that satisfies:*

(a) *The free homotopy classes l_i of (γ_i, T_i) and l_j of (γ_j, T_j) are distinct if $i \neq j$.*

(b) *For every $i \neq j$ we have $\max_{t \in [0,T]} d_g(\gamma_i(t), \gamma_j(t)) > \delta$.*

Then, if

$$m < \frac{N_X(T)}{\widetilde{K}T + 1},$$

there exists a parametrized periodic orbit $(\gamma_{m+1}, T_{m+1} \le T)$ such that its homotopy class l_{m+1} does not belong to the set $\{l_i : 1 \le i \le m\}$ and such that

$$\max_{t \in [0,T]} d_g(\gamma_{m+1}(t), \gamma_i(t)) > \delta \tag{2-10}$$

for all $i \in 1, \dots, m$.

Proof of claim First, recall that $\#(\Lambda_X^{T,\delta}((\gamma_1, T_1), \dots, (\gamma_m, T_m))) \le m(\widetilde{K}T + 1)$. Therefore, because $m < N_X(T)/(\widetilde{K}T + 1)$, there exists a free homotopy class $l_{m+1} \in \Lambda_X^T \backslash \Lambda_X^{T,\delta}((\gamma_1, T_1), \dots, (\gamma_m, T_m))$. Choose a parametrized periodic orbit (γ_{m+1}, T_{m+1}) with $T_{m+1} \le T$ in the homotopy class l_{m+1}.

As $l_{m+1} \notin \Lambda_X^{T,\delta}((\gamma_1, T_1), \dots, (\gamma_m, T_m))$, we must have (2-10) for all $i \in 1, \dots, m$, thus completing the proof of the claim. □

Step 4 Obtaining a T, δ–separated set.

As usual, we denote by $\lfloor N_X(T)/(\widetilde{K}T + 1) \rfloor$ the largest integer which is at most $N_X(T)/(\widetilde{K}T + 1)$. The strategy is now to use the inductive step to obtain a set $S_X^T = \{(\gamma_i, T_i) : 1 \le i \le \lfloor N_X(T)/(\widetilde{K}T + 1) \rfloor\}$, satisfying conditions (a) and (b) above, with the maximum possible cardinality. We start with a set $S_1^T = \{(\gamma_1, T_1)\}$, which clearly satisfies conditions (a) and (b), and if $1 < \lfloor N_X(T)/(\widetilde{K}T + 1) \rfloor$ we apply the inductive step to obtain a parametrized periodic orbit $(\gamma_2, T_2 \le T)$ such that $S_2^T = \{(\gamma_1, T_1), (\gamma_2, T_2 \le T)\}$ satisfies (a) and (b). We can go on applying the inductive step to produce sets $S_m^T = \{(\gamma_i, T_i) : 1 \le i \le m\}$ satisfying the desired conditions (a) and (b) as long as $m - 1$ is smaller than $\lfloor N_X(T)/(\widetilde{K}T + 1) \rfloor$. By this process, we can construct a set $S_X^T = \{(\gamma_i, T_i) : 1 \le i \le \lfloor N_X(T)/(\widetilde{K}T + 1) \rfloor\}$ such that for all $i, j \in \{1, \dots, \lfloor N_X(T)/(\widetilde{K}T + 1) \rfloor\}$, (a) and (b) above hold true.

For each $i \in \{1, \dots, \lfloor N_X(T)/(\widetilde{K}T + 1) \rfloor\}$, let $q_i = \gamma_i(0)$. We define the set $P_X^T := \{q_i : 1 \le i \le \lfloor N_X(T)/(\widetilde{K}T) \rfloor + 1\}$. The condition (b) satisfied by S_X^T implies that P_X^T is a T, δ–separated set. It then follows from the definition of the δ–entropy h_δ that

$$h_\delta(\phi_X) \ge \limsup_{T \to +\infty} \frac{\log\bigl(\lfloor N_X(T)/(\widetilde{K}T + 1) \rfloor\bigr)}{T}. \tag{2-11}$$

Step 5 From the hypothesis of the theorem, we know that for the real numbers $a > 0$ and b, there exists a sequence $T_n \to +\infty$ such that $N_X(T_n) \geq e^{aT_n+b}$ for all T_n.

For every $\epsilon > 0$, we know that for T_n big enough we have $e^{\epsilon T_n} > \widetilde{K}T_n + 1$. This implies that

$$\limsup_{T_n\to+\infty} \frac{\log\bigl(\lfloor N_X(T_n)/(\widetilde{K}T_n+1)\rfloor\bigr)}{T_n} \geq \limsup_{T_n\to+\infty} \frac{\log(\lfloor e^{aT_n+b}/e^{\epsilon T_n}\rfloor)}{T_n} = \limsup_{T_n\to+\infty} \frac{\log(\lfloor e^{(a-\epsilon)T_n+b}\rfloor)}{T_n}. \tag{2-12}$$

It is clear that $\limsup_{T_n\to+\infty} \log(\lfloor e^{(a-\epsilon)T_n+b}\rfloor)/T_n = a-\epsilon$. Combining this with (2-11), we conclude that under the hypothesis of the theorem, $h_\delta(\phi_X) \geq a-\epsilon$. Because ϵ can be taken arbitrarily small, we obtain

$$h_\delta(\phi_X) \geq a. \tag{2-13}$$

Step 6 So far, we have shown that for all $\delta < \epsilon_g/32$, we have $h_\delta(\phi_X) \geq a$. It then follows that

$$h_{\text{top}}(\phi_X) = \lim_{\delta\to 0} h_\delta(\phi_X) \geq a, \tag{2-14}$$

finishing the proof of the theorem. □

Remark One could naively believe that there exists a constant $\delta_g > 0$ depending only on the metric g such that if two parametrized closed curves $\sigma_1\colon \mathbb{R} \to M$ of period T_1 and $\sigma_2\colon \mathbb{R} \to M$ of period T_2 satisfy $\sup_{t\in[0,\max\{T_1,T_2\}]}\{d_g(\sigma_1(t),\sigma_2(t))\} < \delta_g$, then (γ_1, T_1) and (γ_2, T_2) are freely homotopic to each other. This would make the proof of Theorem 1 much shorter. However such a constant does not exist. One can easily find for any $\delta > 0$ two parametrized curves in the 3–torus which are in different primitive free homotopy classes and satisfy $\sup_{t\in[0,\max\{T_1,T_2\}]}\{d_g(\sigma_1(t),\sigma_2(t))\} < \delta$. We sketch the construction below.

Consider coordinates $(x,y,z) \in (\mathbb{R}/\mathbb{Z})^3$ on the three-dimensional torus $\mathbb{T}^3$. Figure 2 above represents the universal cover of the two-dimensional torus $\mathbb{T}^2 \subset \mathbb{T}^3$ obtained by fixing the coordinate $z = 0$ in $\mathbb{T}^3$. The dotted points p_0, $\widehat{p}$, p_1 and p_2 in Figure 2 represent lifts of a point $p \in \mathbb{T}^2$. It is then clear that the curve c represented in Figure 2 projects to a smooth immersed curve in $\mathbb{T}^2 \subset \mathbb{T}^3$.

We consider a parametrization by arc length $\varsigma_1\colon [0,T_1] \to \mathbb{R}^2$ of the piece of c connecting p_0 and p_1. We can extend ς_1 periodically to $\mathbb{R}$ by demanding that $\varsigma_1(t+T_1) = \varsigma_1(t) + (1,2)$ for all $t \in \mathbb{R}$. This extension is a lift to $\mathbb{R}^2$ of the closed immersed curve

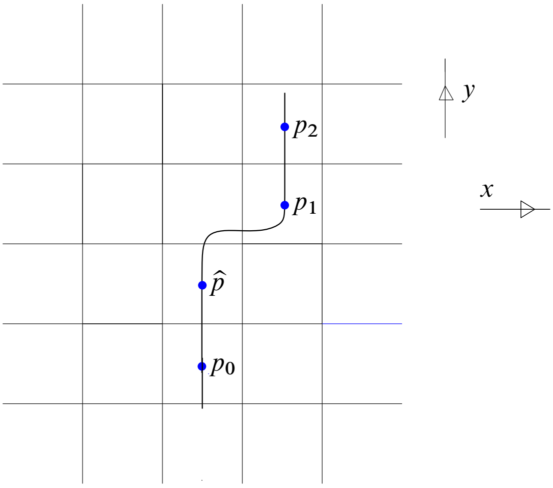


Figure 2: The universal cover of $\mathbb{T}^2 \subset \mathbb{T}^3$

obtained by projecting $\varsigma_1([0, T_1])$ to $\mathbb{T}^2$. By a very small perturbation of the projection of $\varsigma_1([0, T_1])$, we can produce a closed smooth embedded curve $\sigma_1\colon [0, T_1] \to \mathbb{T}^3$ which closes at the point $(p, 0) = \sigma_1(0) = \sigma_1(T_1)$. We consider the natural extension of σ_1 to $\mathbb{R}$ obtained by demanding that $\sigma_1(t) = \sigma_1(t - T_1)$ for all $t \in \mathbb{R}$.

Analogously, we consider a parametrization by arc length $\varsigma_2\colon [0, T_1 + 1] \to \mathbb{R}^2$ of the piece of c connecting p_0 and p_2. We can also extend ς_2 periodically to $\mathbb{R}$, this time demanding that $\varsigma_2(t + T_1 + 1) = \varsigma_2(t) + (1, 3)$. By making a very small perturbation of ς_2, we can produce a closed smooth embedded curve $\sigma_2\colon [0, T_1 + 1] \to \mathbb{T}^3$ which closes at the point $(p, \delta/K) = \sigma_2(0) = \sigma_2(T_1 + 1)$ and which is disjoint from the image of σ_1. We consider the natural extension of σ_2 to $\mathbb{R}$ obtained by demanding that $\sigma_2(t) = \sigma_2(t - (T_1 + 1))$ for all $t \in \mathbb{R}$.

We point out that the extensions $\varsigma_1\colon \mathbb{R} \to \mathbb{R}^2$ and $\varsigma_2\colon \mathbb{R} \to \mathbb{R}^2$ coincide on the interval $[0, T_1 + 1]$. To see this just notice that the piece of c connecting p_0 and $\widehat{p}$ and the piece of c connecting p_1 and p_2 project to the same circle in $\mathbb{T}^2$.

Now let $\sigma_0\colon [0, T_1 + 1] \to \mathbb{T}^2$ be the parametrized curve obtained by projecting $\varsigma_1\colon [0, T_1 + 1] \to \mathbb{R}^2$, which equals $\varsigma_2\colon [0, T_1 + 1] \to \mathbb{R}^2$, to the torus $\mathbb{T}^2$. The curves $\sigma_1|_{[0,T_1+1]}$ and $\sigma_2|_{[0,T_1+1]}$ are both perturbations of the parametrized curve σ_0. By making the perturbations sufficiently small we can guarantee that $\sigma_1|_{[0,T_1+1]}$ and $\sigma_2|_{[0,T_1+1]}$ are arbitrarily close. It is immediate to see that $\sigma_1|_{[0,T_1+1]}$ and $\sigma_2|_{[0,T_1+1]}$ are in distinct homotopy classes.

3 Contact homology

3.1 Pseudoholomorphic curves in symplectic cobordisms

To define the contact homologies used in this paper, we use pseudoholomorphic curves in symplectizations of contact manifolds and symplectic cobordisms. Pseudoholomorphic curves in symplectic manifolds were introduced by Gromov in [24] and adapted to symplectizations and symplectic cobordisms by Hofer [26]; see also [8] as a general reference for pseudoholomorphic curves in symplectic cobordisms.

3.1.1 Cylindrical almost complex structures Let (Y,ξ) be a contact manifold and λ a contact form on (Y,ξ). The symplectization of (Y,ξ) is the product $\mathbb{R}\times Y$ with the symplectic form $d(e^s\lambda)$ (where s denotes the $\mathbb{R}$ coordinate in $\mathbb{R}\times Y$). The 2–form $d\lambda$ restricts to a symplectic form on the vector bundle ξ, and it is well known that the set $\mathfrak{j}(\lambda)$ of $d\lambda$–compatible almost complex structures on the symplectic vector bundle ξ is nonempty and contractible. Notice that if Y is 3–dimensional, the set $\mathfrak{j}(\lambda)$ does not depend on the contact form λ on (Y,ξ).

For $j\in\mathfrak{j}(\lambda)$, we can define an $\mathbb{R}$–invariant almost complex structure J on $\mathbb{R}\times Y$ by demanding that

$$J\partial_s = X_\lambda \quad\text{and}\quad J|_\xi = j. \tag{3-1}$$

We will denote by $\mathcal{J}(\lambda)$ the set of almost complex structures in $\mathbb{R}\times Y$ that are $\mathbb{R}$–invariant, $d(e^s\lambda)$–compatible and satisfy (3-1) for some $j\in\mathfrak{j}(\lambda)$.

3.1.2 Exact symplectic cobordisms with cylindrical ends An exact symplectic cobordism is, roughly, an exact symplectic manifold (W,ϖ) that, outside a compact subset, is like the union of cylindrical ends of symplectizations. We restrict our attention to exact symplectic cobordisms having only one positive end and one negative end.

Let $(W,\varpi=d\kappa)$ be an exact symplectic manifold without boundary, and let (Y^+,ξ^+) and (Y^-,ξ^-) be contact manifolds with contact forms λ^+ and λ^-. We say that $(W,\varpi=d\kappa)$ is an exact symplectic cobordism from λ^+ to λ^- if there exist subsets W^-, W^+ and $\widehat{W}$ of W and diffeomorphisms $\Psi^+\colon W^+\to[0,+\infty)\times Y^+$ and $\Psi^-\colon W^-\to(-\infty,0]\times Y^-$, such that

$$\begin{gathered}\widehat{W}\text{ is compact},\quad W=W^+\cup\widehat{W}\cup W^-,\quad W^+\cap W^-=\varnothing,\\ (\Psi^+)^*(e^s\lambda^+)=\kappa\quad\text{and}\quad(\Psi^-)^*(e^s\lambda^-)=\kappa.\end{gathered}\tag{3-2}$$

In such a cobordism, we say that an almost complex structure $\bar{J}$ is cylindrical if

$$\bar{J} \text{ coincides with } J^+ \in \mathcal{J}(C^+\lambda^+) \text{ in the region } W^+, \tag{3-3}$$

$$\bar{J} \text{ coincides with } J^- \in \mathcal{J}(C^-\lambda^-) \text{ in the region } W^-, \tag{3-4}$$

$$\bar{J} \text{ is compatible with } \varpi \text{ in } \widehat{W}, \tag{3-5}$$

where $C^+ > 0$ and $C^- > 0$ are constants.

For fixed $J^+ \in \mathcal{J}(C^+\lambda^+)$ and $J^- \in \mathcal{J}(C^-\lambda^-)$, we denote by $\mathcal{J}(J^-, J^+)$ the set of cylindrical almost complex structures in $(\mathbb{R} \times Y, \varpi)$ coinciding with J^+ on W^+ and J^- on W^-. It is well known that $\mathcal{J}(J^-, J^+)$ is nonempty and contractible. We will write $\lambda^+ \succ_{\mathrm{ex}} \lambda^-$ if there exists an exact symplectic cobordism from λ^+ to λ^- as above. We remind the reader that $\lambda^+ \succ_{\mathrm{ex}} \lambda$ and $\lambda \succ_{\mathrm{ex}} \lambda^-$ implies $\lambda^+ \succ_{\mathrm{ex}} \lambda^-$, or in other words that the exact symplectic cobordism relation is transitive; see [8] for a detailed discussion on symplectic cobordisms with cylindrical ends. Notice that a symplectization is a particular case of an exact symplectic cobordism.

Remark We point out to the reader that in many references in the literature, a slightly different definition of cylindrical almost complex structures is used: instead of demanding that $\bar{J}$ satisfies conditions (3-3) and (3-4), the stronger condition that $\bar{J}$ coincides with $J^\pm \in \mathcal{J}(\lambda^\pm)$ in the region $W^\pm$ is demanded. We need to consider this more relaxed definition of cylindrical almost complex structures when we study the cobordism maps of cylindrical contact homologies in Section 3.2.3.

3.1.3 Splitting symplectic cobordisms Let λ^+, λ and λ^- be contact forms on (Y, ξ) such that $\lambda^+ \succ_{\mathrm{ex}} \lambda$ and $\lambda \succ_{\mathrm{ex}} \lambda^-$. For $\epsilon > 0$ sufficiently small, it is easy to see that one also has $\lambda^+ \succ_{\mathrm{ex}} (1+\epsilon)\lambda$ and $(1-\epsilon)\lambda \succ_{\mathrm{ex}} \lambda^-$. Then for each $R > 0$, we can construct an exact symplectic form $\varpi_R = d\kappa_R$ on $W = \mathbb{R} \times Y$, where

$$\kappa_R = e^{s-R-2}\lambda^+ \quad \text{in } [R+2, +\infty) \times Y, \tag{3-6}$$

$$\kappa_R = f(s)\lambda \quad \text{in } [-R, R] \times Y, \tag{3-7}$$

$$\kappa_R = e^{s+R+2}\lambda^- \quad \text{in } (-\infty, -R-2] \times Y, \tag{3-8}$$

and $f\colon [-R, R] \to [1-\epsilon, 1+\epsilon]$ satisfies $f(-R) = 1-\epsilon$, $f(R) = 1+\epsilon$ and $f' > 0$. In $(\mathbb{R} \times Y, \varpi_R)$, we consider a compatible cylindrical almost complex structure $\widetilde{J}_R$, but we demand an extra condition on $\widetilde{J}_R$:

$$\widetilde{J}_R \text{ coincides with } J \in \mathcal{J}(\lambda) \text{ in } [-R, R] \times Y. \tag{3-9}$$

Again we divide W into regions: $W^+ = [R+2, +\infty) \times Y$, $W(\lambda^+, \lambda) = [R, R+2] \times Y$, $W(\lambda) = [-R, R] \times Y$, $W(\lambda, \lambda^-) = [-R-2, -R] \times Y$ and $W^- = (-\infty, -R-2] \times Y$.

The family of exact symplectic cobordisms with cylindrical almost complex structures $(\mathbb{R}\times Y, \varpi_R, \widetilde{J}_R)$ is called a splitting family from λ^+ to λ^- along λ.

3.1.4 Pseudoholomorphic curves Let (S,i) be a closed Riemann surface without boundary and $\Gamma\subset S$ a finite set. Let λ be a contact form on (Y,ξ), and let $J\in\mathcal{J}(\lambda)$. A finite energy pseudoholomorphic curve in the symplectization $(\mathbb{R}\times Y, J)$ is a map $\widetilde{w}=(s,w)\colon S\setminus\Gamma\to\mathbb{R}\times Y$ that satisfies

$$\bar\partial_J(\widetilde{w}) = d\widetilde{w}\circ i - J\circ d\widetilde{w} = 0 \tag{3-10}$$

and

$$0 < E(\widetilde{w}) = \sup_{q\in\mathcal{E}}\int_{S\setminus\Gamma}\widetilde{w}^* d(q\lambda), \tag{3-11}$$

where $\mathcal{E}=\{q\colon\mathbb{R}\to[0,1] : q'\geq 0\}$. The quantity $E(\widetilde{w})$ is called the Hofer energy and was introduced in [26]. The operator $\bar\partial_J$ above is called the Cauchy–Riemann operator for the almost complex structure J.

For an exact symplectic cobordism (W,ϖ) from λ^+ to λ^- as considered above, and for $\bar J\in\mathcal{J}(J^-,J^+)$, a finite energy pseudoholomorphic curve is again a map $\widetilde{w}\colon S\setminus\Gamma\to W$ satisfying

$$d\widetilde{w}\circ i = \bar J\circ d\widetilde{w} \tag{3-12}$$

and

$$0 < E_{\lambda^-}(\widetilde{w}) + E_c(\widetilde{w}) + E_{\lambda^+}(\widetilde{w}) < +\infty, \tag{3-13}$$

where

$$E_{\lambda^-}(\widetilde{w}) = \sup_{q\in\mathcal{E}}\int_{\widetilde{w}^{-1}(W^-))}\widetilde{w}^* d(q\lambda^-),$$
$$E_{\lambda^+}(\widetilde{w}) = \sup_{q\in\mathcal{E}}\int_{\widetilde{w}^{-1}(W^+)}\widetilde{w}^* d(q\lambda^+),$$
$$E_c(\widetilde{w}) = \int_{\widetilde{w}^{-1}(W(\lambda^+,\lambda^-))}\widetilde{w}^*\varpi.$$

These energies were also introduced in [26].

In splitting symplectic cobordisms, we use a slightly modified version of energy. Instead of demanding $0 < E_-(\widetilde{w}) + E_c(\widetilde{w}) + E_+(\widetilde{w}) < +\infty$, we demand that

$$0 < E_{\lambda^-}(\widetilde{w}) + E_{\lambda^-,\lambda}(\widetilde{w}) + E_{\lambda}(\widetilde{w}) + E_{\lambda,\lambda^+}(\widetilde{w}) + E_{\lambda^+}(\widetilde{w}) < +\infty, \tag{3-14}$$

where

$$E_{\lambda}(\widetilde{w}) = \sup_{q\in\mathcal{E}} \int_{\widetilde{w}^{-1}W(\lambda)} \widetilde{w}^* d(q\lambda),$$

$$E_{\lambda^-,\lambda}(\widetilde{w}) = \int_{\widetilde{w}^{-1}(W(\lambda,\lambda^-))} \widetilde{w}^*\varpi,$$

$$E_{\lambda,\lambda^+}(\widetilde{w}) = \int_{\widetilde{w}^{-1}(W(\lambda^+,\lambda))} \widetilde{w}^*\varpi,$$

and where $E_{\lambda^-}(\widetilde{w})$ and $E_{\lambda^+}(\widetilde{w})$ are as above.

The elements of the set $\Gamma \subset S$ are called punctures of the pseudoholomorphic curve $\widetilde{w}$. According to [26; 27], punctures fall into two classes, positive and negative, according to the behaviour of $\widetilde{w}$ in the neighbourhood of the puncture. Before presenting this classification, we introduce some notation. Let $B_\delta(z)$ be the ball of radius δ centred at the puncture z, and denote by $\partial(B_\delta(z))$ its boundary. We can describe the types of punctures as follows:

- $z \in \Gamma$ is called a positive interior puncture if $\lim_{z'\to z} s(z') = +\infty$ and there exist a sequence $\delta_n \to 0$ and a Reeb orbit γ^+ of X_{λ^+} such that $w(\partial(B_{\delta_n}(z)))$ converges in C^∞ to γ^+ as $n \to +\infty$,
- $z \in \Gamma$ is called a negative interior puncture if $\lim_{z'\to z} s(z') = -\infty$, and there exist a sequence $\delta_n \to 0$ and a Reeb orbit γ^- of X_{λ^-} such that $w(\partial(B_{\delta_n}(z)))$ converges in C^∞ to γ^- as $n \to +\infty$.

The results in [26] and [27] imply that these are indeed the only possibilities we need to consider for the behaviour of $\widetilde{w}$ near punctures. Intuitively, we have that at the punctures, the pseudoholomorphic curve $\widetilde{w}$ detects Reeb orbits. When for a puncture z, there is a subsequence δ_n such that $w(\partial(B_{\delta_n}(z)))$ converges to a Reeb orbit γ, we will say that $\widetilde{w}$ is asymptotic to this Reeb orbit γ at the puncture z.

If a pseudoholomorphic curve $\widetilde{w}$ is asymptotic to a nondegenerate Reeb orbit at a puncture z, more can be said about its asymptotic behaviour in neighbourhoods of this puncture. Take a neighbourhood $U \subset S$ of z that admits a holomorphic chart $\psi_U\colon (U,z) \to (\mathbb{D},0)$. Using polar coordinates $(r,t) \in (0,+\infty)\times S^1$, we can write $x \in (\mathbb{D}\setminus 0)$ as $x = e^{-r}t$. With this notation, it is shown in [26; 27], that if z is a positive interior puncture at which $\widetilde{w}$ is asymptotic to a nondegenerate Reeb orbit γ^+ of X_{λ^+}, then $\widetilde{w}\circ\psi_U^{-1}(r,t) = (s(r,t), w(r,t))$ satisfies

- $w^r(t) = w(r,t)$ converges in C^∞ to a Reeb orbit γ^+ of X_{λ^+}, exponentially in r and uniformly in t.

Similarly, if z is a negative interior puncture at which $\widetilde{w}$ is asymptotic to a nondegenerate Reeb orbit γ^- of X_{λ^-}, then $\widetilde{w}\circ\psi_U^{-1}(r,t) = (s(r,t), w(r,t))$ satisfies

- $w^r(t) = w(r,t)$ converges in C^∞ to a Reeb orbit γ^- of $-X_{\lambda^-}$ as $r \to +\infty$, exponentially in r and uniformly in t.

Remark The fact that the convergence of pseudoholomorphic curves near punctures to Reeb orbits is of exponential nature is a consequence of the asymptotic formula obtained in [27]. Such formulas are necessary for the Fredholm theory developed in [28] that gives the dimension of the space of pseudoholomorphic curves with fixed asymptotic data.

The discussion above can be summarized by saying that, near punctures, the finite energy pseudoholomorphic curves detect Reeb orbits. It is exactly this behaviour that makes these objects useful for the study of dynamics of Reeb vector fields.

For us, it will be important to consider the moduli spaces $\mathcal{M}(\gamma, \gamma_1', \dots, \gamma_m'; J)$ of genus-0 pseudoholomorphic curves, modulo biholomorphic reparametrization, with one positive puncture asymptotic to a nondegenerate Reeb orbit γ, and negative punctures asymptotic to nondegenerate Reeb orbits $\gamma_1', \dots, \gamma_m'$. It is well known that the linearization $D\bar{\partial}_J$ of $\bar{\partial}_J$ at any element of $\mathcal{M}(\gamma, \gamma_1', \dots, \gamma_m'; J)$ is a Fredholm map (we remark that this property is valid for more general moduli spaces of curves with prescribed asymptotic behaviour). One would like to conclude that the dimension of a connected component of $\mathcal{M}(\gamma, \gamma_1', \dots, \gamma_m'; J)$ is given by the Fredholm index of an element of $\mathcal{M}(\gamma, \gamma_1', \dots, \gamma_m'; J)$. However, this is not always the case if the moduli space contains multiply covered pseudoholomorphic curves.

Fact As a consequence of the exactness of the symplectic cobordisms considered above, we obtain that the energy $E(\tilde{w})$ of $\tilde{w}$ satisfies $E(\tilde{w}) \leq 5A(\tilde{w})$, where $A(\tilde{w})$ is the sum of the action of the Reeb orbits detected by the punctures of $\tilde{w}$ counted with multiplicity; see [8; 29].

3.2 Contact homologies

Contact homologies were introduced in [14] as homology theories which are topological invariants of contact manifolds. In Sections 3.2.1 and 3.2.2, we give an introduction to the more basic and well-known versions of contact homologies. This serves mainly as a motivation to Section 3.2.3, where we present the version of contact homology that will be used in this paper.[2]

[2] We stress that while the versions of contact homology presented in Sections 3.2.1 and 3.2.2 do depend on the polyfold technology currently being developed by Hofer, Wysocki and Zehnder, the version of contact homology which we use in this paper and present in Section 3.2.3 *does not* depend on polyfold and can be constructed in complete rigour with technology that is available in the literature. See the detailed discussion in Section 3.2.3 below.

3.2.1 Full contact homology

Full contact homology was introduced in [14] as an important invariant of contact structures. We refer the reader to [14] and [6] for detailed presentations of the material contained in this subsection.

Let $(Y^{2\mathrm{n}+1}, \xi)$ be a contact manifold with λ a nondegenerate contact form. We denote by $\mathcal{P}(\lambda)$ the set of good periodic orbits of the Reeb vector field X_λ. To each orbit $\gamma \in \mathcal{P}(\lambda)$, we define a $\mathbb{Z}_2$–degree $|\gamma| = (\mu_{CZ}(\gamma) + (\mathrm{n}-2)) \bmod 2$. An orbit γ is called good if it is either simple, or if $\gamma = (\gamma')^i$ for a simple orbit γ' with $|\gamma| = |\gamma'|$.

$\mathfrak{A}(Y,\lambda)$ is defined to be the supercommutative, $\mathbb{Z}_2$–graded $\mathbb{Q}$–algebra with unit generated by $\mathcal{P}(\lambda)$ (an algebra with these properties is called a commutative superalgebra or a super-ring). The $\mathbb{Z}_2$–grading on the elements of the algebra is obtained by considering (on the generators) the grading mentioned above and extending it to $\mathfrak{A}(Y,\lambda)$.

$\mathfrak{A}(Y,\lambda)$ can be equipped with a differential d_J. Denote by $\mathcal{M}^k(\gamma,\gamma'_1,\dots,\gamma'_m;J)$ the moduli space of finite energy pseudoholomorphic curves of genus 0 and Fredholm index k, modulo reparametrization, with one positive puncture asymptotic to γ and negative punctures asymptotic to $\gamma'_1,\dots,\gamma'_m$ in the symplectization $(\mathbb{R}\times Y, J)$. As the almost complex structure J is $\mathbb{R}$–invariant in $\mathbb{R}\times Y$, we have an $\mathbb{R}$–action on $\mathcal{M}^k(\gamma,\gamma'_1,\dots,\gamma'_m;J)$, and we write

$$\widehat{\mathcal{M}}^k(\gamma,\gamma'_1,\dots,\gamma'_m;J) = \mathcal{M}^k(\gamma,\gamma'_1,\dots,\gamma'_m;J)/\mathbb{R}.$$

Lastly, we denote by $\overline{\mathcal{M}}^k(\gamma,\gamma'_1,\dots,\gamma'_m;J)$, as presented in [8], the compactification of $\widehat{\mathcal{M}}^k(\gamma,\gamma'_1,\dots,\gamma'_m;J)$. The compactified moduli space $\overline{\mathcal{M}}^k(\gamma,\gamma'_1,\dots,\gamma'_m;J)$ also involves pseudoholomorphic buildings that appear as limits of a sequence of curves in $\widehat{\mathcal{M}}^k(\gamma,\gamma'_1,\dots,\gamma'_m;J)$ that "breaks"; we refer the reader to [8] for a more detailed description of these moduli spaces. To define our differential, we need the following hypothesis:

Hypothesis H *There exists an abstract perturbation of the Cauchy–Riemann operator ∂_J such that the compactified moduli spaces $\overline{\mathcal{M}}(\gamma,\gamma'_1,\dots,\gamma'_m;J)$ of solutions of the perturbed equation are unions of branched manifolds with corners and rational weights whose dimension is given by the Conley–Zehnder index of the asymptotic orbits and the relative homology class of the solution.*

The proof that Hypothesis H is true is still not written. Establishing its validity is one of the main reasons for the development of the polyfold technology by Hofer, Wysocki and Zehnder. We define

$$d_J\gamma = m(\gamma)\sum_{\gamma'_1,\dots,\gamma'_m}\frac{C(\gamma,\gamma'_1,\dots,\gamma'_m)}{m!}\gamma'_1\gamma'_2\dots\gamma'_m, \tag{3-15}$$

where $C(\gamma,\gamma'_1,\dots,\gamma'_m)$ is the algebraic count of points in the 0–dimensional manifold

$$\widehat{\mathcal{M}}^1(\gamma,\gamma'_1,\dots,\gamma'_m;J), \tag{3-16}$$

and $m(\gamma)$ is the multiplicity of γ. The map d_J is extended to the whole algebra by the Leibnitz rule. Under Hypothesis H, it was proved in [14] that $(d_J)^2=0$. We therefore have that $(\mathfrak{A}(Y,\lambda),d_J)$ is a differential $\mathbb{Z}_2$–graded supercommutative algebra.

Definition 2 The *full contact homology* $\mathbb{CH}(\lambda,J)$ of λ is the homology of the complex $(\mathfrak{A},d_J)$.

Under Hypothesis H, it was also proved in [14] that the full contact homology does not depend on the contact form λ on (Y,ξ), nor on the choice of the cylindrical almost complex structure $J\in\mathcal{J}(\lambda)$.

3.2.2 Cylindrical contact homology

Suppose now that (Y,ξ) is a contact manifold, and λ is a nondegenerate hypertight contact form on (Y,ξ). Fix a cylindrical almost complex structure $J\in\mathcal{J}(\lambda)$. For hypertight contact manifolds, we can define a simpler version of contact homology called cylindrical contact homology. We denote by $\mathrm{CH}_{\mathrm{cyl}}(\lambda)$ the $\mathbb{Z}_2$–graded $\mathbb{Q}$–vector space generated by the elements of $\mathcal{P}(\lambda)$. The differential $d_J^{\mathrm{cyl}}\colon \mathrm{CH}_{\mathrm{cyl}}(\lambda)\to\mathrm{CH}_{\mathrm{cyl}}(\lambda)$ will count elements in the moduli space $\widehat{\mathcal{M}}^1(\gamma,\gamma';J)$. For the generators $\gamma\in\mathcal{P}(\lambda)$, we define

$$d_J^{\mathrm{cyl}}(\gamma)=\mathrm{cov}(\gamma)\sum_{\gamma'\in\mathcal{P}(\lambda)}C(\gamma,\gamma';J)\gamma', \tag{3-17}$$

where $C(\gamma,\gamma';J)$ is the algebraic count of elements in $\widehat{\mathcal{M}}^1(\gamma,\gamma';J)$ and $\mathrm{cov}(\gamma)$ is the covering number of γ. For λ hypertight, and assuming Hypothesis H is true, Eliashberg, Givental and Hofer proved in [14] that $(d_J^{\mathrm{cyl}})^2=0$.

Definition 3 The *cylindrical contact homology* $\mathbb{CH}_{\mathrm{cyl}}(\lambda)$ of λ is the homology of the complex $(\mathrm{CH}_{\mathrm{cyl}}(\lambda),d_J^{\mathrm{cyl}})$.

Under Hypothesis H, the cylindrical contact homology does not depend on the hypertight contact form λ on (Y,ξ), nor on the cylindrical almost complex structure $J\in\mathcal{J}(\lambda)$.

Denote by Λ the set of free homotopy classes of Y. It is easy to see that for each $\rho\in\Lambda$, the subspace $\mathrm{CH}^{\rho}_{\mathrm{cyl}}(\lambda)\subset\mathrm{CH}_{\mathrm{cyl}}(\lambda)$ generated by the set $\mathcal{P}_\rho(\lambda)$ of good periodic orbits in ρ is a subcomplex of $(\mathrm{CH}_{\mathrm{cyl}}(\lambda),d_J^{\mathrm{cyl}})$. This follows from the fact that the number $C(\gamma,\gamma';J)$ can only be nonzero for Reeb orbits γ' that are freely homotopic to γ, which implies that the restriction $d_J^{\mathrm{cyl}}|_{\mathrm{CH}^{\rho}_{\mathrm{cyl}}}$ has image in $\mathrm{CH}^{\rho}_{\mathrm{cyl}}(\lambda)$. From now on, we will denote the restriction $d_J^{\mathrm{cyl}}|_{\mathrm{CH}^{\rho}_{\mathrm{cyl}}}\colon \mathrm{CH}^{\rho}_{\mathrm{cyl}}(\lambda) \to \mathrm{CH}^{\rho}_{\mathrm{cyl}}(\lambda)$ by d_J^{ρ}. Denoting by $\mathrm{C}\mathbb{H}^{\rho}_{\mathrm{cyl}}$ the homology of $(\mathrm{CH}^{\rho}_{\mathrm{cyl}}(\lambda), d_J^{\rho})$, we thus have

$$\mathrm{C}\mathbb{H}_{\mathrm{cyl}}(\lambda) = \bigoplus_{\rho \in \Lambda} \mathrm{C}\mathbb{H}^{\rho}_{\mathrm{cyl}}(\lambda). \tag{3-18}$$

The fact that we can define partial versions of cylindrical contact homology restricted to certain free homotopy classes will be of crucial importance for us. It will allow us to obtain our results without resorting to Hypothesis H. This is explained in the next subsection.

3.2.3 Cylindrical contact homology in special homotopy classes

Maintaining the notation of the previous sections, we denote by (Y, ξ) a contact manifold endowed with a hypertight contact form λ.

Let Λ_0 denote the set of primitive free homotopy classes of Y. Let $\rho \in \Lambda$ be either an element of Λ_0, or a free homotopy class which contains only simple Reeb orbits of λ. Assume that all Reeb orbits in $\mathcal{P}_\rho(\lambda)$ are nondegenerate. By the work of Dragnev [13], we know that there exists a generic subset $\mathcal{J}^{\rho}_{\mathrm{reg}}(\lambda)$ of $\mathcal{J}(\lambda)$ such that for all $J \in \mathcal{J}^{\rho}_{\mathrm{reg}}(\lambda)$ we have:

- For all Reeb orbits $\gamma_1, \gamma_2 \in \rho$, the moduli space of pseudoholomorphic cylinders $\mathcal{M}(\gamma_1, \gamma_2; J)$ is transverse, ie the linearized Cauchy–Riemann operator $D\bar{\partial}_J(\widetilde{w})$ is surjective for all $\widetilde{w} \in \mathcal{M}(\gamma_1, \gamma_2; J)$.
- For all Reeb orbits $\gamma_1, \gamma_2 \in \rho$, each connected component $\mathcal{L}$ of the moduli space $\mathcal{M}(\gamma_1, \gamma_2; J)$ is a manifold whose dimension is given by the Fredholm index of any element $\widetilde{w} \in \mathcal{L}$.

In this case, for $J \in \mathcal{J}^{\rho}_{\mathrm{reg}}(\lambda)$, we define

$$d_J^{\rho}(\gamma) = \mathrm{cov}(\gamma) \sum_{\gamma' \in \mathcal{P}_\rho(\lambda)} C^{\rho}(\gamma, \gamma'; J)\gamma' = \sum_{\gamma' \in \mathcal{P}_\rho(\lambda)} C^{\rho}(\gamma, \gamma'; J)\gamma', \tag{3-19}$$

where $C^{\rho}(\gamma, \gamma'; J)$ is the number of points of the moduli space $\widehat{\mathcal{M}}^1(\gamma, \gamma'; J)$. The second equality follows from the fact that all Reeb orbits in ρ are simple, which implies $\mathrm{cov}(\gamma) = 1$.

For λ and ρ as above and $J \in \mathcal{J}^{\rho}_{\mathrm{reg}}(\lambda)$, the differential $d_J^{\rho}\colon \mathrm{CH}^{\rho}_{\mathrm{cyl}}(\lambda) \to \mathrm{CH}^{\rho}_{\mathrm{cyl}}(\lambda)$ is well-defined and satisfies $(d_J^{\rho})^2 = 0$. Thus, in this situation, we can define the cylindrical contact homology $\mathrm{C}\mathbb{H}^{\rho;J}_{\mathrm{cyl}}(\lambda)$ without imposing Hypothesis H. Once the transversality for J has been achieved, and using coherent orientations constructed in [9], the proof that d_J^{ρ} is well-defined and that $(d_J^{\rho})^2 = 0$ is a combination of compactness and gluing, similar to the proof of the analogous result for Floer homology. For the convenience of the reader, we sketch these arguments below:

Claim *For ρ as above, $d^{\rho}_{J}\colon \mathrm{CH}^{\rho}_{\mathrm{cyl}}(\lambda) \to \mathrm{CH}^{\rho}_{\mathrm{cyl}}(\lambda)$ is well-defined, and for every $\gamma \in \mathcal{P}_{\rho}(\lambda)$, the differential $d^{\rho}_{J}(\gamma)$ is a finite sum.*

Proof The moduli space $\widehat{\mathcal{M}}^1(\gamma, \gamma'; J)$ can be nonempty only if $A(\gamma') \leq A(\gamma)$. It then follows from the nondegeneracy of λ that, for a fixed γ, the numbers $C^{\mathrm{cyl}}(\gamma, \gamma'; J)$ can be nonzero for only finitely many γ'. To see that $C^{\mathrm{cyl}}(\gamma, \gamma'; J)$ is finite for every $\gamma' \in \rho$, suppose by contradiction that there is a sequence $\widetilde{w}_i$ of distinct elements of $\widehat{\mathcal{M}}^1(\gamma, \gamma'; J)$. By the SFT compactness theorem [8], such a sequence has a convergent subsequence that converges to a pseudoholomorphic building $\widetilde{w}$ which has Fredholm index 1. Because of the hypertightness of λ, no bubbling can occur and all the levels $\widetilde{w}^1, \dots, \widetilde{w}^k$ of the building $\widetilde{w}$ are pseudoholomorphic cylinders. As all Reeb orbits of λ in ρ are simple, it follows that all these cylinders are somewhere injective pseudoholomorphic curves, and the regularity of J implies that they must all have Fredholm index at least 1. As a result, we have $1 = I_F(\widetilde{w}) = \sum (I_F(\widetilde{w}^l)) \geq k$, which implies $k = 1$. Thus $\widetilde{w} \in \widehat{\mathcal{M}}^1(\gamma, \gamma'; J)$, and it is the limit of a sequence of distinct elements of $\widehat{\mathcal{M}}^1(\gamma, \gamma'; J)$. This is absurd, because $\widehat{\mathcal{M}}^1(\gamma, \gamma'; J)$ is a 0–dimensional manifold. We thus conclude that the numbers $C^{\mathrm{cyl}}(\gamma, \gamma'; J)$ are all finite. □

Claim *For ρ as above, $(d^{\rho}_{J})^2 = 0$.*

Proof If we write

$$d^{\rho}_{J} \circ d^{\rho}_{J}(\gamma) = \sum_{\gamma'' \in \mathcal{P}_{\rho}(\lambda)} m_{\gamma,\gamma''} \gamma'', \tag{3-20}$$

we know that $m_{\gamma,\gamma''}$ is the number of two-level pseudoholomorphic buildings $\widetilde{w} = (\widetilde{w}^1, \widetilde{w}^2)$ such that $\widetilde{w}^1 \in \widehat{\mathcal{M}}^1(\gamma, \gamma'; J)$ and $\widetilde{w}^2 \in \widehat{\mathcal{M}}^1(\gamma', \gamma''; J)$ for some $\gamma' \in \mathcal{P}_{\rho}(\lambda)$. Because of transversality of $\widetilde{w}^1$ and $\widetilde{w}^2$, we can perform gluing. This implies that $\widetilde{w}$ is in the boundary of the moduli space $\overline{\mathcal{M}}^2(\gamma, \gamma''; J)$. Taking a sequence $\widetilde{w}_i$ of elements in $\widehat{\mathcal{M}}^2(\gamma, \gamma''; J)$ converging to the boundary of $\overline{\mathcal{M}}^2(\gamma, \gamma''; J)$ and arguing similarly as above, we have that this sequence converges to a pseudoholomorphic building $\widetilde{w}_\infty$ whose levels are somewhere injective pseudoholomorphic cylinders. Using that $I_F(\widetilde{w}_\infty) = 2$, we obtain that $\widetilde{w}_\infty$ can have at most 2 levels. As $\widetilde{w}_\infty$ is in the boundary of $\overline{\mathcal{M}}^2(\gamma, \gamma''; J)$, it cannot have only one level, and is therefore a two-level pseudoholomorphic building whose levels have Fredholm index 1. Summing up, $\widetilde{w}_\infty = (\widetilde{w}^1_\infty, \widetilde{w}^2_\infty)$, where $\widetilde{w}^1_\infty \in \widehat{\mathcal{M}}^1(\gamma, \gamma'; J)$ and $\widetilde{w}^2_\infty \in \widehat{\mathcal{M}}^1(\gamma', \gamma''; J)$, for some $\gamma' \in \mathcal{P}_{\rho}(\lambda)$.

The discussion above implies that $m_{\gamma,\gamma''}$ is the count with signs of boundary components of the compactified moduli space $\overline{\mathcal{M}}^2(\gamma,\gamma''; J)$ which is homeomorphic to a one-dimensional manifold with boundary. Because the signs of this count are determined by coherent orientations of $\overline{\mathcal{M}}^2(\gamma,\gamma''; J)$, it follows that $m_{\gamma,\gamma''} = 0$. □

These claims give us the following:

Proposition 4 *Let (Y,ξ) be a contact manifold with a hypertight contact form λ. Let $\rho \in \Lambda$ be either an element of Λ_0 or a free homotopy class which contains only simple Reeb orbits of λ. Assume that all Reeb orbits in $\mathcal{P}_\rho(\lambda)$ are nondegenerate and pick $J \in \mathcal{J}^\rho_{\mathrm{reg}}(\lambda)$. Then d^ρ_J is well defined and $(d^\rho_J)^2 = 0$. Under these conditions we define $\mathrm{C}\mathbb{H}^\rho_{\mathrm{cyl}}(\lambda)$ as the homology of the pair $(\mathrm{CH}^\rho_{\mathrm{cyl}}(\lambda), d^\rho_J)$.*

Exact symplectic cobordisms induce homology maps for the SFT-invariants. We describe how this is done for the version of cylindrical contact homology considered in this section. Let (Y^+,ξ^+) and (Y^-,ξ^-) be contact manifolds, with hypertight contact forms λ^+ and λ^-. Let (W,ω) be an exact symplectic cobordism from λ^+ to λ^-. Assume that ρ is either a primitive free homotopy class or that all the closed Reeb orbits of both λ^+ and λ^- which belong to ρ are simple. Assume moreover that all Reeb orbits of both $\mathcal{P}_\rho(\lambda^+)$ and $\mathcal{P}_\rho(\lambda^-)$ are nondegenerate. Choose almost complex structures $J^+ \in \mathcal{J}^\rho_{\mathrm{reg}}(\lambda^+)$ and $J^- \in \mathcal{J}^\rho_{\mathrm{reg}}(\lambda^-)$. From the work of Dragnev [13] (see also Section 2.3 in [37]) we know that there is a generic subset $\mathcal{J}^\rho_{\mathrm{reg}}(J^-,J^+) \in \mathcal{J}(J^-,J^+)$ such that for $\widehat{J} \in \mathcal{J}^\rho_{\mathrm{reg}}(J^-,J^+)$, $\gamma^+ \in \mathcal{P}_\rho(\lambda^+)$ and $\gamma^- \in \mathcal{P}_\rho(\lambda^-)$, we have that

- all the curves $\widetilde{w}$ in the moduli spaces $\mathcal{M}(\gamma^+,\gamma^-; \widehat{J})$ are Fredholm regular,
- the connected components $\mathcal{V}$ of $\mathcal{M}(\gamma^+,\gamma^-; \widehat{J})$ have dimension equal to the Fredholm index of any pseudoholomorphic curve in $\mathcal{V}$.

In this case, we can define a map $\Phi^{\widehat{J}}\colon \mathrm{CH}^\rho_{\mathrm{cyl}}(\lambda^+) \to \mathrm{CH}^\rho_{\mathrm{cyl}}(\lambda^-)$, given on elements of $\mathcal{P}_\rho(\lambda^+)$, by

$$\Phi^{\widehat{J}}(\gamma^+) = \sum_{\gamma^- \in \mathcal{P}_\rho(\lambda^-)} n_{\gamma^+,\gamma^-}\gamma^-, \tag{3-21}$$

where n_{γ^+,γ^-} is the number of pseudoholomorphic cylinders with Fredholm index 0, positively asymptotic to γ^+ and negatively asymptotic to γ^-. Using a combination of compactness and gluing (see [6]) one proves that $\Phi^{\widehat{J}} \circ d^\rho_{J^+} = d^\rho_{J^-} \circ \Phi^{\widehat{J}}$. As a result we obtain a map $\Phi^{\widehat{J}}\colon \mathrm{C}\mathbb{H}^{\rho,J^+}_{\mathrm{cyl}}(\lambda^+) \to \mathrm{C}\mathbb{H}^{\rho,J^-}_{\mathrm{cyl}}(\lambda^-)$ on the homology level.

We study the cobordism map in the following situation: take $(V = \mathbb{R}\times Y, \varpi)$ to be an exact symplectic cobordism from $C\lambda$ to $c\lambda$, where $C > c > 0$ and λ is a hypertight contact form. Suppose that one can make an isotopy of exact symplectic cobordisms $(\mathbb{R}\times Y,\varpi_t)$ from $C\lambda$ to $c\lambda$, with ϖ_t satisfying $\varpi_0=\varpi$ and $\varpi_1=d(e^s\lambda_0)$. We consider the space $\widetilde{\mathcal{J}}(J,J)$ of smooth homotopies

$$J_t\in\mathcal{J}(J,J),\quad t\in[0,1], \tag{3-22}$$

such that $J_0=J_V$, $J_1\in\mathcal{J}_{\mathrm{reg}}(\lambda)$, and J_t is compatible with ϖ_t for every $t\in[0,1]$. Here J_t is a deformation of J_0 to J_1 through asymptotically cylindrical almost complex structures in the cobordisms $(\mathbb{R}\times Y,\varpi_t)$. For Reeb orbits $\gamma,\gamma'\in\mathcal{P}_\rho(\lambda)$, we consider the moduli space

$$\widetilde{\mathcal{M}}^1(\gamma,\gamma';J_t)=\bigl\{(t,\tilde{w})\mid t\in[0,1]\text{ and }\tilde{w}\in\widehat{\mathcal{M}}^1(\gamma,\gamma';J_t)\bigr\}. \tag{3-23}$$

By using the techniques of [13], we know that there is a generic subset $\widetilde{\mathcal{J}}_{\mathrm{reg}}(J,J)\subset\widetilde{\mathcal{J}}(J,J)$ such that $\widetilde{\mathcal{M}}^1(\gamma,\gamma';J_t)$ is a 1–dimensional smooth manifold with boundary. The crucial condition that makes this valid is again the fact that all the pseudoholomorphic curves that make part of this moduli space are somewhere injective.

The following proposition follows from combining the work of Eliashberg, Givental and Hofer [14] and Dragnev [13].

Proposition 5 *Let (Y,ξ) be a contact manifold with a hypertight contact form λ. Let $\lambda^+=C\lambda$ and $\lambda^-=c\lambda$, where $C>c>0$ are constants, and let ρ be either a primitive free homotopy class or a free homotopy class in which all Reeb orbits of λ are simple. Assume that all Reeb orbits in $\mathcal{P}_\rho(\lambda)$ are nondegenerate. Choose an almost complex structure $J\in\mathcal{J}^\rho_{\mathrm{reg}}(\lambda)$, and set $J^+=J^-=J$. Let $(W=\mathbb{R}\times Y,\varpi)$ be an exact symplectic cobordism from $C\lambda$ to $c\lambda$, and choose a regular almost complex structure $\widehat{J}\in\mathcal{J}^\rho_{\mathrm{reg}}(J^-,J^+)$. Then, if there is a homotopy $(\mathbb{R}\times Y,\varpi_t)$ of exact symplectic cobordisms from $C\lambda$ to $c\lambda$ with $\varpi_0=\varpi$ and $\varpi_1=d(e^s\lambda)$, it follows that the map*

$$\Phi^{\widehat{J}}\colon \mathrm{CH}^{\rho,J}_{\mathrm{cyl}}(\lambda)\to\mathrm{CH}^{\rho,J}_{\mathrm{cyl}}(\lambda)$$

is chain homotopic to the identity.

The proof is again a combination of compactness and gluing, and we sketch it below. We refer the reader to [6] and [14] for the details.

Sketch of the proof We initially define the map

$$K\colon \mathrm{CH}^{\rho}_{\mathrm{cyl}}(\lambda)\to\mathrm{CH}^{\rho}_{\mathrm{cyl}}(\lambda) \tag{3-24}$$

that counts finite energy Fredholm index-(-1) pseudoholomorphic cylinders in the cobordisms $(\mathbb{R}\times Y,\varpi_t)$ for $t\in[0,1]$. Because of the regularity of our homotopy, the moduli space of index -1 cylinders whose positive puncture detects a fixed Reeb orbit γ is finite, and therefore the map K is well defined.

Notice that for $t=1$, the cobordism map $\Phi^{\hat{J}_1}$ is the identity, and the pseudoholomorphic curves that define it are just trivial cylinders over Reeb orbits. For $t=0$, the map $\Phi^{\hat{J}_0}=\Phi^{\hat{J}}$ counts index-0 cylinders in the cobordism $(\mathbb{R}\times Y,\varpi)$. From the regularity of J_0, J_1 and the homotopy J_t, we have that the pseudoholomorphic cylinders involved in these two maps belong to the 1–dimensional moduli spaces $\widetilde{\mathcal{M}}^1(\gamma,\gamma';J_t)$.

By using a combination of compactness and gluing we can show that the boundary of the moduli space $\widetilde{\mathcal{M}}^1(\gamma,\gamma';J_t)$ is exactly the set of pseudoholomorphic buildings $\widetilde{w}$ with two levels, $\widetilde{w}_{\text{cob}}$ and $\widetilde{w}_{\text{symp}}$, such that $\widetilde{w}_{\text{cob}}$ is an index-(-1) cylinder in a cobordism $(\mathbb{R}\times Y,\varpi_t)$, and $\widetilde{w}_{\text{symp}}$ is an index-1 pseudoholomorphic cylinder in the symplectization of λ above or below $\widetilde{w}_{\text{cob}}$. Such two-level buildings are exactly the ones counted in the map $K\circ d_J^{\text{cyl}}+d_J^{\text{cyl}}\circ K$. As a consequence, one has that the difference between the maps $\Phi^{\hat{J}_1}=\text{Id}$ and $\Phi^{\hat{J}}$ is equal to $K\circ d_J^{\text{cyl}}+d_J^{\text{cyl}}\circ K$. This implies that $\Phi^{\hat{J}}$ is chain homotopic to the identity. □

The result above can be used to show that $\mathrm{CH}^{\rho}_{\text{cyl}}(\lambda)$ does not depend on the regular almost complex structure J used to define the differential d_J.

4 Exponential homotopical growth rate of cylindrical contact homology and estimates for h_{top}

In this section, we define the exponential homotopical growth of contact homology and relate it to the topological entropy of Reeb vector fields. The basic idea is to use the fact that the cylindrical contact homology of (M,ξ) in a free homotopy class is nonvanishing to obtain existence of Reeb orbits in such a homotopy class for any contact form on (M,ξ); this idea is present in [30; 37]. It is straightforward to see that the period and action of a Reeb orbit are equal, and in the sequel, we will use the same notation to refer to period and action of Reeb orbits.

Definition 6 Let (M,ξ) be a contact manifold and λ_0 a hypertight contact form on (M,ξ). We denote by $\Lambda(M)$ the set of free homotopy classes of loops in M. For $T>0$, we define the subset $\widetilde{\Lambda}_T(\lambda_0)\subset\Lambda(M)$ by the following condition:

- $\rho\in\widetilde{\Lambda}_T(\lambda_0)$ if and only if all Reeb orbits of X_{λ_0} in ρ are simply covered, nondegenerate, have action/period at most T, and $\mathrm{CH}^{\rho}_{\text{cyl}}(\lambda_0)\neq 0$.

We define $N_T^{\text{cyl}}(\lambda_0):=\#\widetilde{\Lambda}_T(\lambda_0)$.

Definition 7 We say that the cylindrical contact homology of λ_0 has exponential homotopical growth with exponential weight $a > 0$ if there exist a number b and a sequence $T_n \to +\infty$ such that $N^{\mathrm{cyl}}_{T_n}(\lambda_0) \geq e^{aT_n+b}$ for all T_n.

Remark Notice that in Definition 7, we do not demand that λ_0 is nondegenerate. We only demand the weaker condition that the Reeb orbits of λ_0 belonging to some free homotopy classes are nondegenerate.

The main result of this section is the following:

Theorem 8 *Let λ_0 be a hypertight contact form on a contact manifold (M,ξ), and assume that the cylindrical contact homology of λ_0 has exponential homotopical growth with exponential weight $a > 0$. Then for every C^k $(k \geq 2)$ contact form λ on (M,ξ), the Reeb flow of X_λ has positive topological entropy. More precisely, if f_λ is the function such that $\lambda = f_\lambda \lambda_0$, then*

$$h_{\mathrm{top}}(\phi_{X_\lambda}) \geq \frac{a}{\max f_\lambda}. \tag{4-1}$$

Proof We write $E = \max f_\lambda$.

Step 1 We assume first that λ is nondegenerate and C^∞. For every $\epsilon > 0$ we can construct an exact symplectic cobordism from $(E+\epsilon)\lambda_0$ to λ. Analogously, for $\epsilon' > 0$ small enough, it is possible to construct an exact symplectic cobordism from λ to $\epsilon'\lambda_0$.

Using these cobordisms, we can construct a splitting family $(\mathbb{R}\times M, \varpi_R, J_R)$ from $(E+\epsilon)\lambda_0$ to $\epsilon'\lambda_0$, along λ, such that for every $R > 0$, we have that $(\mathbb{R}\times M, \varpi_R, J_R)$ is homotopic to the symplectization of λ_0. For a fixed $\rho \in \widetilde{\Lambda}_T(\lambda_0)$, we pick a regular almost complex structure $J_0 \in \mathcal{J}^{\rho}_{\mathrm{reg}}(\lambda_0)$ and $J \in \mathcal{J}(\lambda)$, and demand that J_R coincides with J_0 in the positive and negative ends of the cobordism, and with J on $[-R,R]\times M$.

We claim that for every R, there exists a finite energy pseudoholomorphic cylinder $\widetilde{w}$ in $(\mathbb{R}\times M, J_R)$ that is positively asymptotic to a Reeb orbit in $\mathcal{P}_\rho(\lambda_0)$ and negatively asymptotic to an orbit in $\mathcal{P}_\rho(\lambda_0)$.

If this was not true for a certain $R > 0$, then because of the absence of pseudoholomorphic cylinders asymptotic to Reeb orbits in $\mathcal{P}_\rho(\lambda_0)$, we would have that $J_R \in \mathcal{J}^{\rho}_{\mathrm{reg}}(J_0, J_0)$. Therefore, the map $\Phi^{J_R}\colon \mathbb{CH}^{\rho}_{\mathrm{cyl}}(\lambda_0) \to \mathbb{CH}^{\rho}_{\mathrm{cyl}}(\lambda_0)$ induced by $(\mathbb{R}\times M, \varpi_R, J_R)$ is well-defined. But because there are no pseudoholomorphic cylinders asymptotic to Reeb orbits in $\mathcal{P}_\rho(\lambda_0)$, we have that Φ^{J_R} vanishes. On the other hand, from Proposition 5 in Section 3.2.3, we know that Φ^{J_R} is the identity. As Φ^{J_R} vanishes and is the identity, we conclude that $\mathbb{CH}^{\rho}_{\mathrm{cyl}}(\lambda_0) = 0$, contradicting that $\rho \in \widetilde{\Lambda}_T(\lambda_0)$.

Step 2 Let $\rho \in \widetilde{\Lambda}_T(\lambda_0)$, let $R_n \to +\infty$ be a strictly increasing sequence, and let $\widetilde{w}_n\colon (S^1 \times \mathbb{R}, i) \to (\mathbb{R} \times M, J_{R_n})$ be a sequence of pseudoholomorphic cylinders with one positive puncture asymptotic to an orbit in $\mathcal{P}_\rho(\lambda_0)$ and one negative puncture asymptotic to an orbit in $\mathcal{P}_\rho(\lambda_0)$. Notice that because of the properties of ρ, the energy of $\widetilde{w}_n$ is uniformly bounded.

Therefore, we can apply the SFT compactness theorem to obtain a subsequence of $\widetilde{w}_n$ which converges to a pseudoholomorphic building $\widetilde{w}$. Notice that in order to apply the SFT compactness theorem, we need to use the nondegeneracy of λ. Moreover, we can give a very precise description of the building.

Let $\widetilde{w}^k$ for $k \in \{1, \dots, m\}$ be the levels of the pseudoholomorphic building $\widetilde{w}$. Because the topology of our curve $\widetilde{w}$ does not change after breaking, we have the following picture:

- The upper level $\widetilde{w}^1$ is composed of one connected pseudoholomorphic curve, which has one positive puncture asymptotic to an orbit $\gamma_0 \in \mathcal{P}_\rho(\lambda_0)$, and several negative punctures. All of the negative punctures detect contractible orbits, except one that detects a Reeb orbit γ_1 which is also in ρ.
- On every other level $\widetilde{w}^k$, there is a special pseudoholomorphic curve which has one positive puncture asymptotic to a Reeb orbit γ_{k-1} in ρ, and at least one, but possibly several, negative punctures. Of the negative punctures, there is one that is asymptotic to an orbit γ_k in ρ, while all the others detect contractible Reeb orbits.

Because of the splitting behaviour of the cobordisms $(\mathbb{R} \times M, J_{R_n})$, it is clear that there exists a k_0 such that the level $\widetilde{w}^{k_0}$ is in an exact symplectic cobordism from $(E+\epsilon)\lambda_0$ to λ. This implies that the special orbit γ_{k_0} is a Reeb orbit of X_λ in the homotopy class ρ.

Notice that $A(\gamma_0) \le (E+\epsilon)T$. This implies that all the other orbits appearing as punctures of the building $\widetilde{w}$ have action smaller than $(E+\epsilon)T$ and, in particular, that γ_{k_0} has action smaller than $(E+\epsilon)T$.

As we can do the construction above for any $\epsilon > 0$, we can obtain a sequence of Reeb orbits γ_j^ρ which are all in ρ such that $A(\gamma_j^\rho) \le (E + 1/j)T$. Using the Arzela–Ascoli theorem, one can extract a convergent subsequence of γ_j^ρ. Its limit γ_ρ is clearly a Reeb orbit of λ in the free homotopy class ρ, with action at most ET.

Step 3 (estimating $N_{X_\lambda}(T)$)[3] From step 2, we know that if $\rho \in \widetilde{\Lambda}_T(\lambda_0)$, then there is a Reeb orbit γ_ρ of the Reeb flow of X_λ with $A(\gamma_\rho) \le ET$. Recalling that the

[3] Recall from Section 2 that $N_{X_\lambda}(T)$ is the number of free homotopy classes of M that contain periodic orbits of X_λ with period at most T.

period and the action of a Reeb orbit coincide, we obtain that $N_{X_\lambda}(T) \geq \#\widetilde{\Lambda}_{T/E}(\lambda_0)$. Under the hypothesis of the theorem, there exists a sequence $T_n \to +\infty$ such that $\#\widetilde{\Lambda}_{T/E}(\lambda_0) \geq e^{aT_n/E+b}$ for all T_n. We then conclude that

$$N_{X_\lambda}(T_n) \geq e^{aT_n/E+b} \tag{4-2}$$

for all elements of the sequence T_n. Applying Theorem 1, we obtain $h_{\mathrm{top}}(\phi_{X_\lambda}) \geq a/E$. This proves the theorem in the case that λ is C^∞ and nondegenerate.

Step 4 Here we pass to the case of a general $C^{k\geq 2}$ contact form λ (the case where λ is degenerate is included here).

Let λ_i be a sequence of nondegenerate smooth contact forms converging in the C^k–topology to a contact form λ which is C^k $(k \geq 2)$ and possibly degenerate. For every $\epsilon > 0$, there is i_0 such that for $i > i_0$, there exists an exact symplectic cobordism from $(E+\epsilon)\lambda_0$ to λ_i.

Fixing a homotopy class $\rho \in \widetilde{\Lambda}_T(\lambda_0)$, we know, by the previous steps, that there exists a Reeb orbit $\gamma_\rho(i)$ of λ_i in the homotopy class ρ with action smaller than $(E+\epsilon)T$. By applying the Arzela–Ascoli theorem to $\gamma_\rho(i)$, we obtain a subsequence which converges to a Reeb orbit $\gamma_{\epsilon,\rho}$ of X_λ with $A(\gamma_{\epsilon,\rho}) \leq (E+\epsilon)T$. Notice that here we use that λ is at least C^2 (so that X_λ is at least C^1) in order to be able to use the Arzela–Ascoli theorem.

Because $\epsilon > 0$ above can be taken arbitrarily close to 0, we can actually obtain a sequence $\gamma_{j,\rho}$ of Reeb orbits of X_λ, whose homotopy class is ρ, such that the actions $A(\gamma_{j,\rho})$ converges to a number at most ET. Again applying the Arzela–Ascoli theorem, we obtain that the sequence $\gamma_{j,\rho}$ has a convergent subsequence which converges to an orbit γ_ρ satisfying $A(\gamma_\rho) \leq ET$.

Reasoning as in step 3 above, we conclude that $N_{X_\lambda}(T_n) \geq e^{aT_n/E+b}$ for all elements of the sequence $T_n \to +\infty$. Applying Theorem 1, we obtain the desired estimate for the topological entropy. This finishes the proof of the theorem. □

5 Contact 3–manifolds with a hyperbolic component

In this section, we prove the following theorem:

Theorem 9 *Let M be a closed connected oriented 3–manifold which can be cut along a nonempty family of incompressible tori into a family $\{M_i, 0 \leq i \leq q\}$ of irreducible manifolds with boundary, such that*

- *M_0 is the mapping torus of a diffeomorphism $h\colon S \to S$ with pseudo-Anosov monodromy on a surface S with nonempty boundary.*

Then M can be given infinitely many nondiffeomorphic contact structures ξ_k such that for each ξ_k, there exists a hypertight contact form λ_k on (M, ξ_k) which has exponential homotopical growth of cylindrical contact homology. It follows that on each (M, ξ_k), all Reeb flows have positive topological entropy.

We denote by S a compact surface with nonempty boundary and by ω a symplectic form on S. Let h be a symplectomorphism of (S, ω) to itself, with pseudo-Anosov monodromy and which is the identity on a neighbourhood of ∂S. We follow a well-known recipe to construct a suitable contact form on the mapping torus $\Sigma(S, h)$.

We choose a primitive β for ω such that, for coordinates $(r, \theta) \in [-\epsilon, 0] \times S^1$ in a neighbourhood V of ∂S, we have $\beta = f(r)d\theta$, where $f > 0$ and $f' > 0$. We pick a smooth nondecreasing function $F_0\colon \mathbb{R} \to [0, 1]$ which satisfies $F_0(t) = 0$ for $t \in \left(-\infty, \frac{1}{100}\right)$ and $F_0(t) = 1$ for $t \in \left(\frac{1}{100}, +\infty\right)$. For $i \in \mathbb{Z}$, define $F_i(t) = F_0(t-i)$. Fixing $\epsilon > 0$, we define a 1–form $\widetilde{\alpha}$ on $\mathbb{R} \times S$ by

$$\widetilde{\alpha} = dt + \epsilon(1 - F_i(t))(h^i)^*\beta + \epsilon F_i(t)(h^{i+1})^*\beta \quad \text{for } t \in [i, i+1). \tag{5-1}$$

This defines a smooth 1–form on $\mathbb{R} \times S$, and a simple computation shows that if ϵ is small enough, the 1–form $\widetilde{\alpha}$ is a contact form. For $t \in [0, 1]$, the Reeb vector field $X_{\widetilde{\alpha}}$ is equal to $\partial_t + v(p, t)$, where $v(p, t)$ is the unique vector tangent to S that satisfies $\omega(v(p, t), \cdot) = F_0'(t)(\beta - h^*\beta)$.

Consider the diffeomorphism $H\colon \mathbb{R} \times S \to \mathbb{R} \times S$ defined by $H(t, p) = (t-1, h(p))$. The mapping torus $\Sigma(S, h)$ is defined by

$$\Sigma(S, h) := (\mathbb{R} \times S)/(t, p) \sim H(t, p), \tag{5-2}$$

and we denote by $\pi\colon \mathbb{R} \times S \to \Sigma(S, h)$ the associated covering map.

Because $H^*\widetilde{\alpha} = \widetilde{\alpha}$, there exists a unique contact form α on $\Sigma(S, h)$ such that $\pi^*\alpha = \widetilde{\alpha}$. Notice that in the neighbourhood $S^1 \times V$ of $\partial\Sigma(S, h)$, we have $\alpha = dt + \epsilon f(r)d\theta$, which implies that X_α is tangent to $\partial\Sigma(S, h)$.

The Reeb vector field X_α on $\Sigma(S, h)$ is transverse to the surfaces $\{t\} \times S$ for $t \in \mathbb{R}/\mathbb{Z}$. This implies that $\{0\} \times S$ is a global surface of section for the Reeb flow of α, and by our expression of $X_{\widetilde{\alpha}}$, the first return map of the Reeb flow of α is isotopic to h.

It follows from [1, Theorem 13] that we can make a perturbation of α supported in the interior $\breve{\Sigma}(S, h)$ of $\Sigma(S, h)$ to obtain a contact form $\widehat{\alpha}$ on $\Sigma(S, h)$ (whose kernel coincides with that of α) satisfying that all Reeb orbits of $\widehat{\alpha}$ which are contained in $\breve{\Sigma}(S, h)$ are nondegenerate. By doing the perturbation small enough, we can also guarantee that $\{0\} \times S$ is still a global surface of section for the flow of $X_{\widehat{\alpha}}$. Since the perturbation is supported in the interior of $\Sigma(S, h)$, the Reeb flow of $\widehat{\alpha}$ is also tangent to the boundary of $\Sigma(S,h)$. It is clear that the first return map $\widehat{h}\colon \{0\}\times S \to \{0\}\times S$ of $\phi_{X_{\widehat{\alpha}}}$ is a diffeomorphism isotopic to h.

5.1 Contact 3–manifolds containing $(\Sigma(S,h),\widehat{\alpha})$ as a component

Let M be a closed connected oriented 3–manifold which can be cut along a nonempty family of incompressible tori into a family $\{M_i, 0 \le i \le q\}$ of irreducible manifolds with boundary, such that the component M_0 is diffeomorphic to $\Sigma(S,h)$. Then it is possible to construct hypertight contact forms on M which match with $\widehat{\alpha}$ in the component M_0. More precisely, we have the following result due to Colin and Honda, and Vaugon:

Proposition 10 [12; 40] *Let M be a closed connected oriented 3–manifold which can be cut along a nonempty family of incompressible tori into a family $\{M_i, 0 \le i \le q\}$ of irreducible manifolds with boundary, such that the component M_0 is diffeomorphic to $\Sigma(S,h)$. Then, there exists an infinite family $\{\xi_k, k \in \mathbb{Z}\}$ of nondiffeomorphic contact structures on M such that*

- *for each $k \in \mathbb{Z}$, there exists a hypertight contact form λ_k on (M,ξ_k) which coincides with $\widehat{\alpha}$ on the component M_0.*

We briefly recall the construction of the contact forms λ_k and refer the reader to [12; 40] for the details. For $i \ge 1$, we apply [12, Theorem 1.3] to obtain a hypertight contact form α_i on M_i which is compatible with the orientation of M_i, and whose Reeb vector field X_{α_i} is tangent to the boundary of M_i. On the special piece M_0, we consider the contact form α_0 equal to $\widehat{\alpha}$ constructed above.

Let $\{\mathfrak{T}_j \mid 1 \le j \le m\}$ be the family of incompressible tori along which we cut M to obtain the pieces M_i. Then the contact forms α_i give a hypertight contact form on each component of $M \setminus \bigcup_{j\ge 1}^m \mathbb{V}(\mathfrak{T}_j)$, where $\mathbb{V}(\mathfrak{T}_j)$ is a small open neighbourhood of $\mathfrak{T}_j$. This gives a contact form $\widehat{\lambda}$ on $M \setminus \bigcup_{j\ge 1}^m \mathbb{V}(\mathfrak{T}_j)$. Using an interpolation process (see [40, Section 7]), one can construct contact forms on the neighbourhoods $\overline{\mathbb{V}(\mathfrak{T}_j)}$ which coincide with $\widehat{\lambda}$ on $\partial\overline{\mathbb{V}(\mathfrak{T}_j)}$. The interpolation process is not unique and can be done in ways so as to produce an infinite family of distinct contact forms $\{\lambda_k \mid k \in \mathbb{Z}\}$ on M that extend $\widehat{\lambda}$, and which are associated to contact structures $\xi_k := \ker \lambda_k$ that are all nondiffeomorphic. The contact topological invariant used to distinguish the contact structures ξ_k is the *Giroux torsion*; see [40, Section 7].

5.2 Proof of Theorem 9

It is clear that Theorem 9 will follow if we establish that the cylindrical contact homology of λ_k has exponential homotopical growth. This is the content of the following:

Proposition 11 *The cylindrical contact homology of λ_k has exponential homotopical growth.*

Before proving the proposition, we introduce some necessary ideas and notation. The first return map of $X_{\widehat{\alpha}}$ is a diffeomorphism $\widehat{h}\colon S \to S$ which is homotopic to h and, therefore, to a pseudo-Anosov map $\psi\colon S \to S$. The Reeb orbits of $X_{\widehat{\alpha}}$ are in one-to-one correspondence with periodic orbits of $\widehat{h}$. Moreover, we have that two Reeb orbits γ_1 and γ_2 of $X_{\widehat{\alpha}}$ are freely homotopic if and only if their associated periodic orbits are in the same Nielsen class. Thus there is an injective map Ξ from the set $\mathcal{N}$ of Nielsen classes to the set $\Lambda(\Sigma(S,h))$ of free homotopy classes of Reeb orbits in $\Sigma(S,h)$.

We now recall some facts about Nielsen theory for pseudo-Anosov maps in surfaces with boundary, which the reader can find in [11; 15; 16]. Let P_n be the set of periodic orbits of ψ with period n which are contained in the interior of S. Because pseudo-Anosov maps have Markov partitions [15; 16], we know that there exist numbers $a > 0$ and b such that

$$\#P_n > e^{an+b}$$

for every $n \in \mathbb{N}$. It follows from [11, Lemma 1.1] that all periodic orbits in P_n belong to distinct Nielsen classes, and that these Nielsen classes are unrelated to the boundary of S. By this, we mean that for every periodic orbit in P_n, its suspension is a curve in $\Sigma(S,h)$ which cannot be homotoped to a curve completely contained in the boundary of $\Sigma(S,h)$.

We denote by $\mathcal{N}_n$ the set of Nielsen classes associated to the periodic orbits P_n of ψ. Notice that $\mathcal{N}$ equals the disjoint union $\bigcup_{n\in\mathbb{N}} \mathcal{N}_n$. It follows from the discussion above that

$$\#\mathcal{N}_n > e^{an+b}$$

for all $n \in \mathbb{N}$. It is immediate to see that the fixed points of $\widehat{h}$ belong to a finite number of Nielsen classes, and we denote by c the number of elements in $\mathcal{N}_1$. We write $\mathcal{N}_1 = \{\nu_1, \dots, \nu_c\}$. For each $\nu_i \in \mathcal{N}_1$ we will denote by ν_i^n the Nielsen class in $\mathcal{N}_n$ which n–covers ν_i in the following sense: if x_i is a fixed point in ν_i, then ν_i^n is the Nielsen class that contains the periodic orbit of period n that "covers" x_i.

As observed previously, there exists an injective map $\Xi\colon \mathcal{N} \to \Lambda(\Sigma(S,h))$. Let p be a prime number, and let $\rho \in \Xi(\mathcal{N}_p)$. Then there are two possibilities:

(a) ρ contains only simple Reeb orbits,

(b) ρ contains a Reeb orbit γ which is a p–cover of a simple orbit γ_0 that intersects $\{0\} \times S$ once.

The reason why these are the only two possibilities is that every Reeb orbit $\gamma \in \rho$ intersects $\{0\} \times S$ exactly p times. If γ is a multiple cover of a simple orbit γ_0, then the number of intersections of γ_0 with $\{0\} \times S$ must be a divisor of p. As p is prime, this number is either p, which implies that γ is simple, or 1. It is clear that if $\rho \in \Xi(\mathcal{N}_p)$ satisfies (b), then $\rho = \Xi(\nu_i^p)$ for some $\nu_i \in \mathcal{N}_1$. We denote by $\mathcal{N}_p^{\mathrm{simp}}$ the set $\mathcal{N}_p \setminus \{\nu_1^p, \dots, \nu_i^p, \dots, \nu_q^p\}$. As a consequence we conclude that if $\Lambda_{\mathrm{simp}}^p := \Xi(\mathcal{N}_p^{\mathrm{simp}})$ is the set of elements in $\Xi(\mathcal{N}_p)$ satisfying (a), then

$$\#\Lambda_{\mathrm{simp}}^p = \#\mathcal{N}_p - c$$

for every prime number p. Since $\#\mathcal{N}_p > e^{ap+b}$ for every prime p, we conclude that there exists a prime number p_0 such that for every prime $p \geq p_0$,

$$\#\Lambda_{\mathrm{simp}}^p \geq e^{ap+q}.$$

Let $\mathfrak{x}$ be a periodic orbit of $\widehat{h}$ of period n. Viewing $\widehat{h}$ as the first return map for a global surface of section of the Reeb flow $\phi_{X_{\widehat{\alpha}}}$ we know that there is a Reeb orbit $\gamma_{\mathfrak{x}}$ of $\widehat{\alpha}$ (and also of λ_k) which is the suspension of $\mathfrak{x}$. Because of the compactness of S, we know that there exists a number $\eta > 0$, depending only on $\widehat{h}$ and $\widehat{\alpha}$, such that $A(\gamma_{\mathfrak{x}}) \leq \eta n$.

We are now ready for the proof of Proposition 11. The main ideas of the argument are due to Vaugon, who estimated in [40] a different growth rate of the cylindrical contact homology λ_k.

Proof of Proposition 11 **Step 1** Let $i\colon \Sigma(S,h) \to M$ be the injection we obtain from viewing $\Sigma(S,h)$ as a component of M. Because of the incompressibility of $\partial\Sigma(S,h)$ in M, the associated map $i_*\colon \Lambda(\Sigma(S,h)) \to \Lambda(M)$ is injective (here $\Lambda(M)$ denotes the free loop space of M).

For each prime number p, we define $T_p := \eta p$. Recall that if $\rho \in \Lambda_{\mathrm{simp}}^p$, then ρ does not contain curves completely contained in the boundary of $\Sigma(S,h)$. It follows from this and from the incompressibility of $\partial\Sigma(S,h)$ in M, that if $\varrho \in i_*(\Lambda_{\mathrm{simp}}^p)$, then every loop in ϱ must intersect the interior $\Sigma(S,h)$.

Using that the Reeb flow of λ_k is tangent to $\partial\Sigma(S,h)$, it follows that if $\varrho \in i_*(\Lambda_{\mathrm{simp}}^p)$, then all Reeb orbits of $\phi_{X_{\lambda_k}}$ that belong to ϱ are contained in the interior of $\Sigma(S,h)$. This implies that ϱ contains only nondegenerate[4] Reeb orbits of $\phi_{X_{\lambda_k}}$. Combining this with the injectivity of i_* and Ξ, we conclude that every Reeb orbit λ_k in ϱ is the suspension of a periodic orbit of $\widehat{h}$ in the Nielsen class $\nu := (i_* \circ \Xi)^{-1}\varrho \in \mathcal{N}_p^{\mathrm{simp}}$. This implies that

[4] Recall that because of our choice of $\widehat{\alpha}$, Reeb orbits contained in $\mathrm{int}(\Sigma(S,h))$ are nondegenerate.

(c) all Reeb orbits of λ_k in the free homotopy class ϱ are nondegenerate and simple,

(d) all Reeb orbits of λ_k in the free homotopy class ϱ have action $\leq T_p$.

Hypertightness of λ_k and (c) imply that if $\varrho \in i_*(\Lambda^p_{\text{simp}})$, then $\mathrm{C}\mathbb{H}^{\varrho}_{\text{cyl}}(\lambda_k)$ is well defined.

Step 2 For every $\varrho \in i_*(\Lambda^p_{\text{simp}})$, we have $\mathrm{C}\mathbb{H}^{\varrho}_{\text{cyl}}(\lambda_k) \neq 0$. Indeed, Vaugon showed (see the proofs of Lemma 7.11 and Theorems 1.3 and 1.2 in [40]) that the number of Reeb orbits in ϱ of even and odd degree differ. For Euler characteristic reasons, this implies that $\mathrm{C}\mathbb{H}^{\varrho}_{\text{cyl}}(\lambda_k) \neq 0$. Combining this with (d) from step 1, we conclude that every $\varrho \in i_*(\Lambda^p_{\text{simp}})$ belongs to the set $\widetilde{\Lambda}_{T_p}(\lambda_k)$ as defined in Definition 6.

Step 3 Recall that in Definition 6 of Section 4, we defined $N^{\text{cyl}}_T(\lambda_k)$ as the cardinality of $\#\widetilde{\Lambda}_T(\lambda_k)$. That is, $N^{\text{cyl}}_T(\lambda_k)$ is the number of free homotopy classes ϱ in $\Lambda(M)$ which contain only nondegenerate simple Reeb orbits with action smaller than T and that satisfy $\mathrm{C}\mathbb{H}^{\varrho}_{\text{cyl}}(\lambda_k) \neq 0$.

Because of the injectivity of i_*, we know that $\# i_*(\Lambda^p_{\text{simp}}) = \#\Lambda^p_{\text{simp}}$. Combining this with steps 1 and 2, it follows that for every element of the sequence $T_p \to +\infty$,

$$N^{\text{cyl}}_{T_p}(\lambda_k) \geq \# i_*(\Lambda^p_{\text{simp}}) = \#\Lambda^p_{\text{simp}} \geq e^{aT_p/\eta + b}, \tag{5-3}$$

which establishes the proposition. □

Proof of Theorem 9 As mentioned previously, Theorem 9 follows directly from combining Proposition 11, Proposition 10 and Theorem 8. □

It would be interesting to obtain an upper bound on the constant η above. This could provide a more precise estimate for the homotopical growth rate of the cylindrical contact homology of λ_k.

6 Graph manifolds and Handel–Thurston surgery

In [25], Handel and Thurston used Dehn surgery to obtain nonalgebraic Anosov flows in 3–manifolds. Their surgery was adapted to the contact setting by Foulon and Hasselblatt in [18], who interpreted it as a Legendrian surgery and used it to produce nonalgebraic Anosov Reeb flows on 3–manifolds. In this section, we apply the Foulon–Hasselblatt Legendrian surgery to obtain more examples of contact 3–manifolds which are distinct from unit tangent bundles, and on which every Reeb flow has positive topological entropy.

Some clarifications regarding the surgeries we consider are in order. On one hand, we restrict our attention to the Foulon–Hasselblatt surgery on Legendrian lifts of embedded separating geodesics on hyperbolic surfaces. This is an important restriction, since Foulon and Hasselblatt perform their surgery on the Legendrian lift of any immersed closed geodesic on a hyperbolic surface. On the other hand, for this restricted class of Legendrian knots, the surgery we consider is a bit more general than the one in [18]. They restrict their attention to Dehn surgeries with positive integer coefficients, while we consider the case of any integer coefficient, as is explained in Section 6.1.

6.1 The surgery

We start by fixing some notation. Let (S,g) be an oriented hyperbolic surface and $\mathfrak{r}\colon S^1 \to S$ an embedded oriented separating geodesic of g. We let $\pi\colon (\mathbb{D},g)\to(S,g)$ denote a locally isometric covering of (S,g) by the hyperbolic disc $(\mathbb{D},g)$ with the property that $(-1,1)\times\{0\}\subset\pi^{-1}(\mathfrak{r}(S^1))$. Such a covering always exists since the segment $(-1,1)\times\{0\}$ of the real axis is a geodesic in $(\mathbb{D},g)$. We denote by $v(\theta)$ the unique unitary vector field along $\mathfrak{r}(\theta)$ satisfying $\angle(\mathfrak{r}'(\theta),v(\theta))=-\pi/2$. Our orientation convention is chosen so that for coordinates $z=x+iy\in\mathbb{D}$, the lift of $v(\theta)$ to $(-1,1)\times\{0\}$ is a positive multiple of the vector field $-\partial_y$ along $(-1,1)\times\{0\}$. Also, let $\Pi\colon T_1S\to S$ denote the base point projection.

Because $\mathfrak{r}$ is a separating geodesic, we can cut S along $\mathfrak{r}$ to obtain two oriented hyperbolic surfaces with boundary which we denote by S_1 and S_2. Our labelling is chosen so that the vector field $v(\theta)$ points into S_2 and out of S_1. This decomposition of S induces a decomposition of T_1S into T_1S_1 and T_1S_2. Both T_1S_1 and T_1S_2 are 3–manifolds whose boundary is the torus formed by the unit fibres over $\mathfrak{r}$.

Denote by $V_{\mathfrak{r},\delta}$ the closed $\delta-$neighbourhood of the geodesic $\mathfrak{r}$ for the hyperbolic metric g. For $\delta>0$ sufficiently small, we have that $V_{\mathfrak{r},\delta}$ is an annulus such that the only closed geodesics contained in $V_{\mathfrak{r},\delta}$ are the covers of $\mathfrak{r}$, and such that $V_{\mathfrak{r},\delta}$ satisfies the following convexity property: if $\check{V}$ is the connected component of $\pi^{-1}(V_{\mathfrak{r},\delta})$ containing $(-1,1)\times\{0\}$, then every segment of a hyperbolic geodesic starting and ending in $\check{V}$ is completely contained in $\check{V}$. It also follows from the conventions adopted above that, if we denote by U^+ the upper hemisphere of $\mathbb{D}$ composed of points with positive imaginary part and by U^- the lower hemisphere of the $\mathbb{D}$ composed of points with negative imaginary part, we have

$$\check{V}\cap U^+\subset\pi^{-1}(S_1)\quad\text{and}\quad\check{V}\cap U^-\subset\pi^{-1}(S_2). \tag{6-1}$$

This fact has the following important consequence: if $\nu([0,K])$ is a hyperbolic geodesic segment starting and ending at $V_{\mathfrak{r},\delta}$ and contained in one of the S_i, then $[\nu]$ is a nontrivial homotopy class in the relative fundamental group $\pi_1(S_i,V_{\mathfrak{r},\delta})$.

On the unit tangent bundle T_1S, we consider the contact form λ_g whose Reeb vector field is the geodesic vector field for the hyperbolic metric g. It is well known that the lifted curve $L_{\mathfrak{r}}(\theta) = (\mathfrak{r}(\theta), v(\theta))$ in T_1S is Legendrian on the contact manifold $(T_1S, \ker\lambda_g)$. The geodesic vector field X_{λ_g} along $L_{\mathfrak{r}}$ coincides with the horizontal lift of v (see [38, Section 1.3]), points into T_1S_2 and out of T_1S_1, and is normal to $\partial T_1S_2 = \partial T_1S_1$ for the Sasaki metric on T_1S.

Moreover, if $\delta > 0$ is small enough, we know that for every $\vartheta \in L_{\mathfrak{r}}$, there exist numbers $t_1 < 0$ and $t_2 > 0$ such that

$$\phi^{t_1}_{\lambda_g}(\vartheta) \in T_1S_1 \setminus \Pi^{-1}(V_{\mathfrak{r},\delta}), \tag{6-2}$$

$$\phi^{t_2}_{\lambda_g}(\vartheta) \in T_1S_2 \setminus \Pi^{-1}(V_{\mathfrak{r},\delta}). \tag{6-3}$$

Following [18], we know that there exists a neighbourhood $B^{3\eta}_{2\epsilon}$ of $L_{\mathfrak{r}}$ on which we can find coordinates $(t, s, w) \in (-3\eta, 3\eta) \times S^1 \times (-2\epsilon, 2\epsilon)$ such that

$$\lambda_g = dt + w\,ds, \tag{6-4}$$

$$L_{\mathfrak{r}} = \{0\} \times S^1 \times \{0\}, \tag{6-5}$$

where $\{0\} \times \{\vartheta\} \times (-2\epsilon, 2\epsilon)$ is a local parametrization of the unitary fibre over $\vartheta \in L_{\mathfrak{r}}$, and $\epsilon < \eta/(4|q|\pi)$, with q being a fixed integer. Let $\mathcal{W}^- = \{-3\eta\} \times S^1 \times (-2\epsilon, 2\epsilon)$ and $\mathcal{W}^+ = \{+3\eta\} \times S^1 \times (-2\epsilon, 2\epsilon)$. It is clear that $\Pi(\mathcal{W}^-) \subset S_1$ and $\Pi(\mathcal{W}^+) \subset S_2$. Because on $\overline{B}^{3\eta}_{2\epsilon}$, the Reeb vector field X_{λ_g} is given by ∂_t, it is clear that for every point $p \in B^{3\eta}_{2\epsilon}$, there are $p^- \in \mathcal{W}^-$, $p^+ \in \mathcal{W}^+$, $t^- \in (-6\eta, 0)$ and $t^+ \in (0, 6\eta)$ for which

$$\phi^{t^-}_{X_{\lambda_g}}(p) = p^- \quad \text{and} \quad \phi^{t^+}_{X_{\lambda_g}}(p) = p^+. \tag{6-6}$$

This means that trajectories of the flow of X_{λ_g} that enter the box $B^{3\eta}_{2\epsilon}$ enter through $\mathcal{W}^-$ and exit through $\mathcal{W}^+$. They cannot stay inside $B^{3\eta}_{2\epsilon}$ for a very long positive or negative interval of time. We can say even more about these trajectories.

For $\sigma = (\mathfrak{p}, \dot{\mathfrak{p}}) \in S \times T_pS$ in $\mathcal{W}^+ \cup \mathcal{W}^-$ let $\tilde{\sigma} = (\tilde{\mathfrak{p}}, \dot{\tilde{\mathfrak{p}}})$ be a lift of σ to the unit tangent bundle $T_1\mathbb{D}$ such that $\tilde{\mathfrak{p}} \in \breve{V}$. The geodesic vector field X_{λ_g} at $\tilde{\sigma}$ coincides with the horizontal lift of $\dot{\mathfrak{p}}$ [38, Section 1.3]. For δ, $\eta > 0$ and $\epsilon < \eta/(4|q|\pi)$ sufficiently small, we can guarantee that

- $\Pi(B^{3\eta}_{2\epsilon})$ is contained in $V_{\mathfrak{r},\delta}$,
- for the lifts $\tilde{\sigma} = (\tilde{\mathfrak{p}}, \dot{\tilde{\mathfrak{p}}})$ of points in $\mathcal{W}^+ \cup \mathcal{W}^-$ as above, the vector $\dot{\tilde{\mathfrak{p}}}$ (which is the projection of the geodesic vector field $X_{\lambda_g}(\tilde{\sigma})$) satisfies $\angle(\dot{\tilde{\mathfrak{p}}}, -\partial_y) < \delta$.

With such a choice of $\delta > 0$, $\eta > 0$ and $0 < \epsilon < \eta/(4|q|\pi)$, we obtain that for every $\sigma^+ \in \mathcal{W}^+$ there exists $t_{\sigma^+} > 0$, and for every $\sigma^- \in \mathcal{W}^-$ there exists $t_{\sigma^-} < 0$, such that

$$\phi^{t_{\sigma^+}}_{X_{\lambda_g}}(\sigma^+) \in (T_1 S_2) \setminus \Pi^{-1}(V_{\mathfrak{r},\delta}) \quad \text{and} \quad \forall t \in [0, t_{\sigma^+}],\ \phi^t_{X_{\lambda_g}}(\sigma^+) \notin B^{3\eta}_{2\epsilon}, \tag{6-7}$$

$$\phi^{t_{\sigma^-}}_{X_{\lambda_g}}(\sigma^-) \in (T_1 S_1) \setminus \Pi^{-1}(V_{\mathfrak{r},\delta}) \quad \text{and} \quad \forall t \in [t_{\sigma^-}, 0],\ \phi^t_{X_{\lambda_g}}(\sigma^+) \notin B^{3\eta}_{2\epsilon}. \tag{6-8}$$

To prove the last condition above one uses the fact that $\angle(\dot{\tilde{\mathfrak{p}}}, -\partial_y) < \delta$ is small and studies the behaviour of geodesics in $(\mathbb{D}, g)$ starting at points close to the real axis and with initial velocity close to $-\partial_y$. It is easy to see that such geodesics have to intersect the region $V_{\mathfrak{r},\delta}$ and visit the interior of both $S_1 \setminus V_{\mathfrak{r},\delta}$ and $S_2 \setminus V_{\mathfrak{r},\delta}$. From now on we will assume that $\delta > 0$, $\eta > 0$ and $0 < \epsilon < \eta/(8|q|\pi)$ are such that all the properties described above hold simultaneously.

Consider the map $F\colon B^{2\eta}_{2\epsilon} \setminus \bar{B}^{\eta}_{\epsilon} \to B^{2\eta}_{2\epsilon} \setminus \bar{B}^{\eta}_{\epsilon}$ defined by

$$F(t,s,w) = (t, s + f(w), w) \quad \text{for} \quad (t,s,w) \in (\eta, 2\eta) \times S^1 \times (-2\epsilon, 2\epsilon), \tag{6-9}$$

where $f(w) = -q\mathcal{R}(w/\epsilon)$ (for our previously chosen integer q) and $\mathcal{R}\colon [-1,1] \to [0, 2\pi]$ satisfies $\mathcal{R} = 0$ on a neighbourhood of -1, $\mathcal{R} = 2\pi$ on a neighbourhood of 1, $0 \le \mathcal{R}' \le 4$ and $\mathcal{R}'$ is an even function.

Our new 3–manifold M is obtained by gluing $T_1 S \setminus \bar{B}^{\eta}_{\epsilon}$ and $B^{2\eta}_{2\epsilon}$ using the map F:

$$M = (T_1 S \setminus \bar{B}^{\eta}_{\epsilon}) \cup B^{2\eta}_{2\epsilon} \big/ (x \in B^{2\eta}_{2\epsilon} \setminus \bar{B}^{\eta}_{\epsilon}) \sim (F(x) \in T_1 S \setminus \bar{B}^{\eta}_{\epsilon}). \tag{6-10}$$

Notice that

$$T_1 S = (T_1 S \setminus \bar{B}^{\eta}_{\epsilon}) \cup B^{2\eta}_{2\epsilon} \big/ (x \in B^{2\eta}_{2\epsilon} \setminus \bar{B}^{\eta}_{\epsilon}) \sim (x \in T_1 S \setminus \bar{B}^{\eta}_{\epsilon}).$$

This clarifies our construction of M and shows that M is obtained from $T_1 S$ via a Dehn surgery on $L_{\mathfrak{r}}$. We follow [18] to endow M with a contact form which coincides with λ_g outside $B^{2\eta}_{2\epsilon}$. As a preparation, we define the function $\beta\colon (-3\eta, 3\eta) \to \mathbb{R}$:

- β is equal to 1 on an open neighbourhood of $[-2\eta, 2\eta]$,
- $|\beta'| \le \pi/\eta$ and $\operatorname{supp}\beta$ is contained in $[-3\eta, 3\eta]$.

Using β we define

$$r(t,w) = \frac{\beta(t)}{2} \int_{-2\epsilon}^{w} x f'(x)\, dx. \tag{6-11}$$

We point out that $\operatorname{supp}(r)$ is contained in $B^{3\eta}_{\epsilon}$, and therefore, so is $\operatorname{supp}(dr)$. Notice also that in $B^{2\eta}_{2\epsilon} \setminus \bar{B}^{\eta}_{\epsilon}$, one has $dr = \frac{1}{2} w f'(w) dw$.

Again following [18], we define in $T_1S \setminus \bar{B}^{\eta}_{\epsilon}$ the 1–form A_r:

$$A_r = dt + w\,ds + dr \quad \text{for } t \in (-3\eta, -\eta), \tag{6-12}$$

$$A_r = dt + w\,ds - dr \quad \text{for } t \in (\eta, 3\eta), \tag{6-13}$$

$$A_r = \lambda_g \quad \text{otherwise.} \tag{6-14}$$

Notice that because $\operatorname{supp}(dr)$ is contained in $B^{3\eta}_{\epsilon}$, the 1–form A_r is well defined.

On the box $B^{2\eta}_{2\epsilon}$, we define

$$\widetilde{A} = dt + w\,ds + dr. \tag{6-15}$$

A direct computation shows that $F^*(A_r) = \widetilde{A}$, which means that the gluing map F allows us to glue the 1–forms A_r and $\widetilde{A}$. We denote by λ_{FH} the 1–form on M obtained by gluing $\widetilde{A}$ and A_r. We will denote by $\widetilde{B}$ the following region:

$$\widetilde{B} = ((B^{3\eta}_{2\epsilon} \setminus \bar{B}^{\eta}_{\epsilon}) \subset M) \cup B^{2\eta}_{2\epsilon} \big/ (x \in B^{2\eta}_{2\epsilon} \setminus \bar{B}^{\eta}_{\epsilon}) \sim (F(x) \in (B^{3\eta}_{2\epsilon} \setminus \bar{B}^{\eta}_{\epsilon}))\cdot \tag{6-16}$$

The importance of this region lies in the fact that in $M \setminus \widetilde{B} = T_1S \setminus B^{3\eta}_{2\epsilon}$, the contact form λ_{FH} coincides with λ_g.

Following [18], one shows by a direct computation that $(dt + w\,ds \pm dr) \wedge (dw \wedge ds) = (1 \pm \partial r/\partial t) dt \wedge dw \wedge ds$. Using the fact that $\epsilon < \eta/(8\pi|q|)$, one gets that $|\partial r/\partial t| < 1$, thus obtaining that $(dt + w\,ds \pm dr)$ is a contact form. It follows from this that A_r and $\widetilde{A}$ are contact forms in their respective domains, and therefore, λ_{FH} is a contact form on M. More strongly, Foulon and Hasselblatt proceed to show that if q is nonnegative, the Reeb flow of λ_{FH} is Anosov.

6.2 Hypertightness and exponential homotopical growth of contact homology of λ_{FH}

For $q \in \mathbb{N}$, the hypertightness of λ_{FH} follows from the fact that its Reeb flow is Anosov [17]. In this subsection, we give an independent and completely geometrical proof of the hypertightness of λ_{FH}, which is valid for every $q \in \mathbb{Z}$.

To understand the topology of Reeb orbits of λ_{FH}, we will study trajectories that enter the surgery region $\widetilde{B}$. We start by studying trajectories in $B^{2\eta}_{2\epsilon}$. In this region, we have

$$X_{\lambda_{\mathrm{FH}}} = \frac{\partial_t}{1 + \partial_t r}. \tag{6-17}$$

This implies, similarly to what happens for λ_g, that for points $p \in B^{2\eta}_{2\epsilon}$, the trajectory $\phi^t_{X_{\lambda_{\mathrm{FH}}}}(p)$ leaves the box $B^{2\eta}_{2\epsilon}$ in forward and backward times. More precisely, there exists a constant $\widetilde{a} > 0$, depending only on λ_{FH}, such that for $p \in B^{2\eta}_{2\epsilon}$, there are $\breve{p}^- \in \breve{\mathcal{W}}^- = \{-2\eta\} \times S^1 \times [-2\epsilon, 2\epsilon]$, $\breve{p}^+ \in \breve{\mathcal{W}}^+ = \{+2\eta\} \times S^1 \times [-2\epsilon, 2\epsilon]$, $\breve{t}^- \in (-\widetilde{a}, 0]$ and $\breve{t}^+ \in [0, \widetilde{a})$ such that

$$\begin{gathered} \phi^t_{X_{\lambda_{\mathrm{FH}}}}(\breve{p}) \text{ is in the interior of } B^{2\eta}_{2\epsilon} \text{ for every } t \in (t^-, t^+), \\ \phi^{t^-}_{X_{\lambda_{\mathrm{FH}}}}(\breve{p}) = \breve{p}^- \quad \text{and} \quad \phi^{t^+}_{X_{\lambda_{\mathrm{FH}}}}(\breve{p}) = \breve{p}^+. \end{gathered} \tag{6-18}$$

We now analyse the trajectories of points $\breve{p}^- \in \breve{\mathcal{W}}^-$ and $\breve{p}^+ \in \breve{\mathcal{W}}^+$. For this, we first notice that on $\widetilde{B} \setminus B^{\eta}_{\epsilon}$, the contact form λ_{FH} is given by $dt + w\,ds \pm dr$, and therefore, we have in this region

$$X_{\lambda_{\mathrm{FH}}} = \frac{\partial_t}{1 \pm \partial_t r}, \tag{6-19}$$

which is still a positive multiple of ∂_t.

This implies that for every $\breve{p}^- \in \breve{\mathcal{W}}^-$ and $\breve{p}^+ \in \breve{\mathcal{W}}^+$, there exist $t^{\breve{p}^-} < 0$ and $t^{\breve{p}^+} > 0$ such that

$$\phi^{t^{\breve{p}^-}}_{X_{\lambda_{\mathrm{FH}}}}(\breve{p}^-) \in \mathcal{W}^- \quad \text{and} \quad \phi^{t^{\breve{p}^+}}_{X_{\lambda_{\mathrm{FH}}}}(\breve{p}^+) \in \mathcal{W}^+. \tag{6-20}$$

Again using that $X_{\lambda_{\mathrm{FH}}}$ is a positive multiple of ∂_t on $\widetilde{B} \setminus B^{2\eta}_{2\epsilon}$, we have that for every point p in $\widetilde{B} \setminus B^{2\eta}_{2\epsilon}$ whose t coordinate is in $[2\eta, 3\eta]$, the trajectory of the flow $\phi^t_{X_{\lambda_{\mathrm{FH}}}}$ going through p is a straight line, with fixed coordinates s and w, that goes from $\breve{\mathcal{W}}^+$ to $\mathcal{W}^+$. Analogously, for every point p in $\widetilde{B} \setminus B^{2\eta}_{2\epsilon}$ whose t coordinate is in $[-3\eta, -2\eta]$, the trajectory of the backward flow of $\phi^t_{X_{\lambda_{\mathrm{FH}}}}$ going through p is a straight line from $\breve{\mathcal{W}}^-$ to $\mathcal{W}^-$.

Summing up, with all the cases considered above, we have showed that for every point $p \in \widetilde{B}$, the trajectory of the flow $\phi^t_{X_{\lambda_{\mathrm{FH}}}}$ going through p for $t = 0$ intersects $\mathcal{W}^-$ for nonpositive time and $\mathcal{W}^+$ for nonnegative time. In other words, all trajectories that intersect $\widetilde{B}$ enter through $\mathcal{W}^-$ and leave through $\mathcal{W}^+$, which means that for all $\breve{p} \in \widetilde{B}$, there exist times $\breve{t}^- \leq 0$ and $\breve{t}^+ \geq 0$ such that

$$\phi^{\breve{t}^+}_{X_{\lambda_{\mathrm{FH}}}}(\breve{p}) \in \mathcal{W}^+, \tag{6-21}$$

$$\phi^{\breve{t}^-}_{X_{\lambda_{\mathrm{FH}}}}(\breve{p}) \in \mathcal{W}^-, \tag{6-22}$$

$$\phi^{t}_{X_{\lambda_{\mathrm{FH}}}}(\breve{p}) \in \widetilde{B} \quad \text{for all } t \in [\breve{t}^-, \breve{t}^+]. \tag{6-23}$$

Now, because on $M \setminus \widetilde{B} = T_1 S \setminus B^{3\eta}_{2\epsilon}$, the contact form λ_{FH} coincides with λ_g, we have that trajectories of $X_{\lambda_{\mathrm{FH}}}$ starting at $\mathcal{W}^-$ at time $t = 0$ have to leave $M \setminus N$ (with N defined as in (6-26) below) as time diminishes before reentering on $\widetilde{B}$. Similarly, the trajectories starting at $\mathcal{W}^+$ have to leave $M \setminus N$ for positive time before reentering to $\widetilde{B}$. More precisely, one can use (6-7) and (6-8) to show that for $p^- \in \mathcal{W}^-$ and $p^+ \in \mathcal{W}^+$, there exist $t_{p^-} < 0$ and $t_{p^+} > 0$ such that

$$\phi^{t_{p^+}}_{X_{\lambda_{\mathrm{FH}}}}(p^+) \in M_2 \setminus N \quad \text{and} \quad \forall t \in [0, t_{p^+}],\ \phi^{t}_{X_{\lambda_{\mathrm{FH}}}}(p^+) \notin \widetilde{B}, \tag{6-24}$$

$$\phi^{t_{p^-}}_{X_{\lambda_{\mathrm{FH}}}}(p^-) \in M_1 \setminus N \quad \text{and} \quad \forall t \in [t_{p^-}, 0],\ \phi^{t}_{X_{\lambda_{\mathrm{FH}}}}(p^-) \notin \widetilde{B}, \tag{6-25}$$

where, denoting

$$B^{2\eta}_{2\epsilon}(-) = [-2\eta, 0] \times S^1 \times (-2\epsilon, 2\epsilon) \quad \text{and} \quad B^{2\eta}_{2\epsilon}(+) = [0, 2\eta] \times S^1 \times (-2\epsilon, 2\epsilon),$$

the submanifolds M_1, M_2 and N of M are defined as follows:

$$M_1 = {(T_1 S_1 \setminus B^{\eta}_{\epsilon}) \cup B^{2\eta}_{2\epsilon}(-)}\big/\big(x \in B^{2\eta}_{2\epsilon}(-) \setminus \overline{B}^{\eta}_{\epsilon}\big) \sim \big(F(x) \in ((B^{2\eta}_{2\epsilon} \cap T_1 S_1) \setminus \overline{B}^{\eta}_{\epsilon})\big),$$

$$M_2 = {(T_1 S_2 \setminus B^{\eta}_{\epsilon}) \cup B^{2\eta}_{2\epsilon}(+)}\big/\big(x \in B^{2\eta}_{2\epsilon}(+) \setminus \overline{B}^{\eta}_{\epsilon}\big) \sim \big(F(x) \in ((B^{2\eta}_{2\epsilon} \cap T_1 S_2) \setminus \overline{B}^{\eta}_{\epsilon})\big),$$

and

$$N = {(\Pi^{-1}(V_{\mathfrak{c},\delta}) \setminus B^{\eta}_{\epsilon}) \cup B^{2\eta}_{2\epsilon}(-)}\big/{x \sim F(x)}, \tag{6-26}$$

with $x \in B^{2\eta}_{2\epsilon}(-) \setminus \overline{B}^{\eta}_{\epsilon}$ and $F(x) \in ((B^{2\eta}_{2\epsilon} \cap T_1 S_1) \setminus \overline{B}^{\eta}_{\epsilon})$.

Remark It is not hard to see that

$$M = {M_1 \cup M_2}\big/(x \in \partial M_1) \sim (\widetilde{F}(x) \in \partial M_2).$$

Here $\widetilde{F}$ is a Dehn twist which coincides with $(s + f(w), w)$ for $w \in [-2\epsilon, 2\epsilon]$ and is the identity elsewhere. This picture of M is closer to the one in the paper [25] and shows that M is a graph manifold (a graph manifold is one whose JSJ decomposition consists of Seifert S^1 bundles). By using this description of M and applying van Kampen's theorem to analyse the fundamental group of M, Handel and Thurston show that, for q not belonging to a finite subset of $\mathbb{Z}$, no finite cover of M is a Seifert manifold, thus obtaining that M is an "exotic" graph manifold.

From their definition, one sees that as manifolds, $M_1 \cong T_1 S_1$ and $M_2 \cong T_1 S_2$. This implies that ∂M_1 and ∂M_2 are incompressible tori in M_1 and M_2, respectively. If we look at M_1 and M_2 as submanifolds of M, their boundary $\mathbb{T}$ coincides and is also incompressible in M. We remark that $M_i \setminus N$ is diffeomorphic to $T_1 S_i \setminus \Pi^{-1}(V_{c,\delta})$, which is diffeomorphic to $T_1 S_i$ for $i = 1, 2$.

In a similar way, we can describe the topology of N. Let $N_i = M_i \cap N$. Reasoning identically as one does to show that M_i is diffeomorphic to $T_1 S_i$, one shows that N_i

is diffeomorphic to a thickened two torus $\mathcal{T}^2 \times [-1,1]$. As N is obtained from N_1 and N_2 by gluing them along $\mathbb{T}$ (which is a boundary component of both of them), we have that N is also diffeomorphic to the product $\mathcal{T}^2 \times [-1,1]$.

The discussion above proves the following:

Lemma 12 *For all $\breve{p} \in \widetilde{B}$, the trajectory $\{\phi^t_{X_{\lambda_{\mathrm{FH}}}}(\breve{p}) \mid t \in \mathbb{R}\}$ intersects $M_1 \setminus N$ and $M_2 \setminus N$.*

Proof We have already established that for $\breve{p} \in \widetilde{B}$, its trajectory intersects $\mathcal{W}^+$ for some nonnegative time and $\mathcal{W}^-$ for some nonpositive time, as shown in (6-21) and (6-22). One now applies (6-24) and (6-25) to finish the proof of the lemma. □

Notice that trajectories can only enter in $\widetilde{B}$ through the wall $\mathcal{W}^-$, which is contained in M_1, and can only exit $\widetilde{B}$ through the wall $\mathcal{W}^+$, which is contained in M_2. We also point out that all trajectories of the flow $\phi^t_{X_{\lambda_{\mathrm{FH}}}}$ are transversal to $\mathbb{T}$, with the exception of the two Reeb orbits which correspond to parametrizations of the hyperbolic geodesic $\mathfrak{r}$ (they continue to exist as periodic orbits after the surgery because they are out of the surgery region).

We will deduce, from the previous discussion, the following important lemma.

Lemma 13 *Let $\gamma([0,T'])$ be a trajectory of $X_{\lambda_{\mathrm{FH}}}$ such that $\gamma(0), \gamma(T') \in \mathbb{T}$ and for all $t \in (0,T')$ we have $\gamma(t) \notin \mathbb{T}$ (notice that in such a situation, $\gamma([0,T']) \subset M_i$ for $i=1$ or $i=2$). Then $\gamma([0,T']) \cap (M_i \setminus N)$ is nonempty.*

Proof We divide the proof in 3 possible scenarios.

Case 1 Suppose that $\gamma([0,T']) \cap \widetilde{B}$ is empty. In this case, $\gamma([0,T'])$ is a hyperbolic geodesic with endpoints on the closed geodesic $\mathfrak{r}$. It follows from the convexity of the hyperbolic metric that $[\gamma([0,T'])] \in \pi_1(T_1 S_i, \mathbb{T})$ is nontrivial. This implies that $[\gamma([0,T'])] \in \pi_1(M_i, \mathbb{T})$ is nontrivial, which can be true only if $\gamma([0,T']) \cap (M_i \setminus N)$ is nonempty since N is a tubular neighbourhood of $\mathbb{T}$.

Case 2 Suppose that $\gamma([0,T']) \cap \widetilde{B}$ is nonempty and $\gamma([0,T']) \subset M_2$. Take $\hat{t} \in [0,T']$ such that $\gamma(\hat{t}) \in \widetilde{B}$. We know from our previous discussion that there are $\hat{t}_1 \leq \hat{t} \leq \hat{t}_2$ such that $\gamma([\hat{t}_1,\hat{t}_2]) \subset \widetilde{B}$, $\gamma(\hat{t}_1) \in (\mathbb{T} \cap \widetilde{B})$ and $\gamma(\hat{t}_2) \in \mathcal{W}^+$; notice that in the coordinates (t,s,w) for $\widetilde{B}$ considered previously, $\mathbb{T} \cap \widetilde{B}$ is the annulus $\{0\} \times S^1 \times (-2\epsilon, 2\epsilon)$. From this picture, it is clear that for t smaller that $\hat{t}_1$, the trajectory enters M_1. Therefore, we must have $\hat{t}_1 = 0$ and $\gamma([0,\hat{t}_2]) \subset \widetilde{B}$. Notice also that for all t slightly bigger than $\hat{t}_2$, the trajectory is outside $\widetilde{B}$. Because trajectories of $X_{\lambda_{\mathrm{FH}}}$ can only enter $\widetilde{B}$

in M_1, we obtain that $\gamma([\hat{t}_2, T'])$ does not intersect the interior of $\widetilde{B}$ and, therefore, is a hyperbolic geodesic in T_1S_2. Now, using (6-7) and (6-8), we obtain that, because $\gamma(\hat{t}_2) \in \mathcal{W}^+$, the trajectory $\gamma\colon [\hat{t}_2, T'] \to M_2$ has to intersect $M_2 \setminus N$ before hitting $\mathbb{T}$ at $t = T'$. Thus there is some $t \in (\hat{t}_2, T')$ for which $\gamma(t) \in M_2 \setminus N$.

Case 3 The proof in the case where $\gamma([0, T']) \cap \widetilde{B}$ is nonempty and $\gamma([0, T']) \subset M_1$ is analogous to the one of case 2.

These three cases exhaust all possibilities and, therefore, prove the lemma. □

Our reason for introducing the above decomposition of M into M_1 and M_2, and for proving the lemmas above, is to introduce the following representation of Reeb orbits of λ_{FH}. Let (γ, T) be a Reeb orbit of λ_{FH} which intersects both $M_1 \setminus N$ and $M_2 \setminus N$. We can assume that the chosen parametrization of γ is such that $\gamma(0) \in \partial N$, and that there are $t_+ > 0$ and $t_- < 0$ such that

$$\gamma(t_+) \in M_1 \setminus N \quad \text{and} \quad \gamma([0, t_+]) \in M_1 \cup N, \tag{6-27}$$

$$\gamma(t_-) \in M_2 \setminus N \quad \text{and} \quad \gamma([t_-, 0]) \in M_2 \cup N. \tag{6-28}$$

This means that in an interval of the origin, γ is coming from $M_2 \setminus N$ and going to $M_1 \setminus N$. It follows from Lemma 13 that there exists a unique sequence $0 = t_0 < t_{1/2} < t_1 < t_{3/2} < \cdots < t_n = T$ such that for all $k \in \{0, \dots, n-1\}$,

- $\gamma([t_k, t_{k+(1/2)}]) \subset M_i$ for $i = 1$ or $i = 2$,
- $\gamma([t_{k+(1/2)}, t_{k+1}]) \in N$ and there is a unique $\widetilde{t}_k \in [t_{k+(1/2)}, t_{k+1}]$ such that $\gamma(\widetilde{t}_k) \in \mathbb{T}$,
- if $\gamma([t_k, t_{k+(1/2)}]) \subset M_i$, then $\gamma([t_{k+1}, t_{k+(3/2)}]) \subset M_j$ for $j \neq i$.

Notice that $\gamma([t_0, t_{1/2}]) \subset M_1$ and $\gamma([t_{n-1}, t_{n-(1/2)}]) \subset M_2$. This implies that n is even, so we can write $n = 2n'$, and that $\gamma([t_k, t_{k+(1/2)}]) \subset M_1$ for k even and $\gamma([t_k, t_{k+(1/2)}]) \subset M_2$ for k odd. For each $k \in \{0, \dots, 2n'-1\}$, the existence of the unique $\widetilde{t}_k$ in the interval $[t_{k+(1/2)}, t_{k+1}]$ for which $\gamma(\widetilde{t}_k) \in \mathbb{T}$ is guaranteed by Lemma 13 and the fact that $\mathbb{T}$ is the hypersurface that separates M_1 and M_2.

In order to obtain information on the free homotopy class of (γ, T), we observe that $\gamma([t_k, t_{k+(1/2)}])$ coincides with a hyperbolic geodesic segment in T_1S_i starting and ending in $V_{\tau,\delta}$. Therefore, as we have previously seen, the homotopy class $[\gamma([t_k, t_{k+(1/2)}])]$ in $\pi_1(T_1S_i, V_{\tau,\delta})$ is nontrivial, which implies that $\gamma([t_k, t_{k+(1/2)}])$ is a nontrivial relative homotopy class in $\pi_1(M_i, N)$. We consider now the curve $\gamma([\widetilde{t}_k, \widetilde{t}_{k+1}])$: it is the concatenation of 3 curves, the first and the third ones being completely contained in N and the middle one being $\gamma([t_k, t_{k+(1/2)}])$. From this description and the fact that $\gamma([t_k, t_{k+(1/2)}])$ is a nontrivial relative homotopy class in $\pi_1(M_i, N)$ it is clear that $\gamma([\widetilde{t}_k, \widetilde{t}_{k+1}])$ is also nontrivial in $\pi_1(M_i, N)$ (and also nontrivial in $\pi_1(M_i, \mathbb{T})$).

We now denote by $\widetilde{M}$ the universal cover of M and $\widehat{\pi}\colon \widetilde{M} \to M$ the covering map. From the incompressibility of $\mathbb{T}$, it follows that every lift of $\mathbb{T}$ is an embedded plane in $\widetilde{M}$. We denote by $\widetilde{N}^0$ a lift of N. Because N is a thickened neighbourhood of an incompressible torus, it follows that $\widetilde{N}^0$ is diffeomorphic to $\mathbb{R}^2 \times [-1, 1]$, ie it is a thickened neighbourhood of an embedded plane in $\widetilde{M}$. Because N separates M into two components, it follows that $\widetilde{N}^0$ separates $\widetilde{M}$ into two connected components. Now, $\partial \widetilde{N}^0$ is the union of two embedded planes, P^0_- and P^0_+, which are characterized by the fact that there are neighbourhoods V_- and V_+ of P^0_- and P^0_+, respectively, such that $\widehat{\pi}(V_-) \subset M_1$ and $\widehat{\pi}(V_+) \subset M_2$. We will denote by C^0_- the connected component of $\widetilde{M} \setminus \widetilde{N}^0$ which intersects V_-, and by C^0_+ the connected component of $\widetilde{M} \setminus \widetilde{N}^0$ which intersects V_+.

As seen earlier, $[\gamma([t_k, t_{k+(1/2)}])]$ is a nontrivial relative homotopy class in $\pi_1(M_i, N)$. We show that this class remains nontrivial when seen in $\pi_1(M, N)$. Let $\mathbb{T}_i = \partial N \cap M_i$. Because N is obtained by attaching over each point of $\mathbb{T}_i$ a small compact interval (ie it is a bundle over $\mathbb{T}_i$ whose fibres are intervals), it follows that $[\gamma([t_k, t_{k+(1/2)}])]$ is trivial in $\pi_1(M_i, \mathbb{T}_i)$ if and only if it is trivial in $\pi_1(M_i, N)$, which is not the case. As $\mathbb{T}_i$ is isotopic to $\mathbb{T}$, it is also an incompressible torus that divides M into two components. Now, $[\gamma([t_k, t_{k+(1/2)}])]$ is trivial in $\pi_1((M_i \setminus \operatorname{int}(N)), \mathbb{T}_i)$ if and only if there exists a curve $\mathfrak{c}$ in $\mathbb{T}_i$, with endpoints $\gamma(t_k)$ and $\gamma(t_{k+(1/2)})$, such that the concatenation $\gamma * \mathfrak{c}$ is contractible in $M_i \setminus \operatorname{int}(N)$. Because of the incompressibility of $\mathbb{T}_i$, such a curve $\gamma * \mathfrak{c}$ is contractible in $M_i \setminus \operatorname{int}(N)$ if and only if it is contractible in M. This implies that $[\gamma([t_k, t_{k+(1/2)}])]$ is trivial in $\pi_1(M, \mathbb{T}_i)$ if and only if it is trivial in $\pi_1((M_i \setminus \operatorname{int}(N)), \mathbb{T}_i)$, which we know not to be the case. Lastly, again because N is an interval bundle over $\mathbb{T}_i$, it is clear that as $[\gamma([t_k, t_{k+(1/2)}])]$ is not trivial in $\pi_1(M, \mathbb{T}_i)$, it cannot be trivial in $\pi_1(M, N)$, as we wished to show.

Let $\widetilde{\gamma}$ be a lift of γ such that $\widetilde{\gamma}(0) \in \widetilde{N}^0$. We know that $\widetilde{\gamma}([t_{2n'-(1/2)} - T, t_{1/2}]) \subset \widetilde{N}^0$. It will be useful to define the sequence

$$\widetilde{t}_i = q_i T + t_{r_i}, \tag{6-29}$$

where q_i and $r_i < 2n'$ are the unique integers such that $i = q_i(2n') + r_i$. To $\widetilde{t}_i$ we associate the lift $\widetilde{N}^i$ of N which is determined by the property that $\widetilde{\gamma}(\widetilde{t}_i) \in \widetilde{N}^i$. It is clear that the sequence $\widetilde{N}^i$ contains all lifts of N which are intersected by the curve $\widetilde{\gamma}(\mathbb{R})$. For the lifts $\widetilde{N}^i$, we define the connected components C^i_- and C^i_+ of $\widetilde{M} \setminus \widetilde{N}^i$, and the planes P^i_- and P^i_+ in the same way as for $\widetilde{N}^0$. A priori it could be that, for $i \neq j$, we have $\widetilde{N}^i = \widetilde{N}^j$. We will show that this cannot happen.

Firstly, $\widetilde{N}^0 \neq \widetilde{N}^1$ because $\gamma([\widetilde{t}_0, \widetilde{t}_1])$ is nontrivial in $\pi_1(M, N)$. Also, we have that $\widetilde{N}^1 \subset C_-^0$ because $\gamma([t_0, t_{1/2}]) \subset M_1$. The same reasoning shows that $\widetilde{N}^2 \neq \widetilde{N}^1$ and

$$\widetilde{N}^2 \subset C_+^1. \tag{6-30}$$

On the other hand, we have that $\widetilde{N}^0 \subset C_-^1$, because $\widetilde{\gamma}([\widetilde{t}_0, t_{1/2}])$ is a path totally contained in $\widetilde{M} \setminus \widetilde{N}^1$ connecting $\widetilde{N}^0$ and P_-^1. As $\widetilde{N}^2 \subset C_+^1$ and $\widetilde{N}^0 \subset C_-^1$, we must have $\widetilde{N}^2 \neq \widetilde{N}^0$. In the same way, one shows that $\widetilde{N}^3 \neq \widetilde{N}^1$ and, more generally, that $\widetilde{N}^{i+2} \neq \widetilde{N}^i$ and $\widetilde{N}^{i+1} \neq \widetilde{N}^i$. Now for $\widetilde{N}^3$, we have that $\widetilde{N}^3 \subset C_-^2$. As $\widetilde{\gamma}([\widetilde{t}_0, t_{3/2}])$ is a path completely contained in $\widetilde{M} \setminus \widetilde{N}^2$ connecting $\widetilde{N}^0$ and P_+^2, we obtain that $\widetilde{N}^0 \subset C_+^2$ and, therefore, $\widetilde{N}^3 \neq \widetilde{N}^0$.

Proceeding inductively along this line, one obtains that $\widetilde{N}^i \neq \widetilde{N}^0$ for all $i \neq 0$ and, more generally, $\widetilde{N}^i \neq \widetilde{N}^j$ for all $i \neq j$. As a consequence, we obtain that the curve $\widetilde{\gamma}(\mathbb{R})$ cannot be homeomorphic to a circle, and therefore, $\gamma(\mathbb{R})$ cannot be contractible. We are ready for the main result of this subsection.

Proposition 14 λ_{FH} *is hypertight.*

Proof There are two possibilities for Reeb orbits.

Possibility 1 The Reeb orbit γ visits both $M_1 \setminus N$ and $M_2 \setminus N$. In this case, we have just showed that γ is not contractible.

Possibility 2 The Reeb orbit γ is completely contained in M_i for $i = 1$ or $i = 2$. In this case, γ does not visit the surgery region $\widetilde{B}$. Therefore, it also existed before the surgery as a closed hyperbolic geodesic in $M_i \setminus \widetilde{B} = T_1 S_i \setminus B_{2\epsilon}^{3\eta}$. Such a closed geodesic is noncontractible in $T_1 S_i$, which is diffeomorphic to M_i. Thus $\gamma \subset M_i$ is noncontractible in M_i.

Now looking at M_i as a submanifold with boundary of M, we recall that ∂M_i is an incompressible torus in M. This implies that every noncontractible closed curve in M_i remains noncontractible in M. Therefore, γ is also a noncontractible Reeb orbit in this case. □

6.2.1 Exponential homotopical growth of cylindrical contact homology for λ_{FH}

We now obtain more information on the properties of periodic orbits of $X_{\lambda_{\mathrm{FH}}}$.

Lemma 15 *If a Reeb orbit (γ, T) of λ_f visits both $M_1 \setminus N$ and $M_2 \setminus N$, then any curve freely homotopic to (γ, T) must intersect $\mathbb{T}$.*

Proof As we saw earlier, the lift $\widetilde{\gamma}$ intersects all the elements of the sequence $\widetilde{N}_i$ (of lifts of N), which satisfy $\widetilde{N}_i \neq \widetilde{N}_j$ for all $i \neq j$.

Introducing an auxiliary distance d on the compact manifold M (coming from a Riemannian metric), we obtain an auxiliary distance $\widetilde{d}$ on $\widetilde{M}$ by pulling d back by the covering map. It is clear that for i sufficiently large, the $\widetilde{d}$–distance between $\widetilde{N}_{\pm i}$ and $\widetilde{N}_0$ becomes arbitrarily large. As a consequence, one obtains that for each $K > 0$, there exists $t_K > 0$ such that $\widetilde{d}(\widetilde{\gamma}(\pm t_K), \widetilde{N}_0) > K$.

Now let $\zeta\colon [0,T] \to M$ be a closed curve freely homotopic to $\gamma([0,T])$. A homotopy $H\colon [0,T] \times [0,1] \to M$ generates a homotopy $\widetilde{H}\colon \mathbb{R} \times [0,1] \to \widetilde{M}$ from a lift $\widetilde{\gamma}$ to a lift $\widetilde{\zeta}$. Using the fact that H is uniformly continuous, one sees that there exists a constant $\mathfrak{C} > 0$ such that $\widetilde{d}(\widetilde{H}(\{t\} \times [0,1]), \widetilde{\gamma}(t)) < \mathfrak{C}$ for all $t \in \mathbb{R}$.

Now take $K > 2\mathfrak{C}$. Using the inequalities

$$\widetilde{d}((\widetilde{H}(\{t\} \times [0,1])), \widetilde{\gamma}(t)) < \mathfrak{C}, \quad \widetilde{d}(\widetilde{\gamma}(\pm t_K), \widetilde{N}_0) > K,$$

and the triangle inequality, we obtain that $H(\{t_K\} \times [0,1])$ is always in the connected component of $\widetilde{\gamma}(t_K)$. This implies that $\widetilde{\zeta}(\mathbb{R})$ visits both connected components of $\widetilde{M} \setminus \widetilde{N}_0$ and must thus intersect $\widetilde{N}_0$. Even more, because $\widetilde{\zeta}(\mathbb{R})$ intersects both components of $\partial \widetilde{N}_0$, we have that ζ visits both components of $M \setminus N$ and, therefore, has to intersect $\mathbb{T}$. This completes the proof of the lemma. □

We are now ready for the main result of this section:

Theorem 16 *Let $(M, \xi_{(q,\mathfrak{r})})$ be the contact manifold obtained from performing the Foulon–Hasselblat q–surgery on the Legendrian curve $L_{\mathfrak{r}} \subset (T_1 S, \xi_{\text{geo}})$, and denote by λ_{FH} the contact form on $(M, \xi_{(q,\mathfrak{r})})$ obtained from this surgery. Then λ_{FH} is hypertight, and its cylindrical contact homology has exponential homotopical growth. It follows that every Reeb flow on $(M, \xi_{(q,\mathfrak{r})})$ has positive topological entropy.*

Proof It suffices to show that the cylindrical contact homology of λ_{FH} has exponential homotopical growth, since this combined with Theorem 1 establishes the last assertion of the theorem.

Step 1 (a special class of Reeb orbits) We will obtain our estimate by looking at Reeb orbits which are completely contained in the component M_1. As we saw previously, such orbits never cross the surgery region $\widetilde{B}$. Thus they are in a region where λ_{FH} coincides with λ_g, and such Reeb orbits are closed geodesics in (S_1, g). Conversely, every closed geodesic in (S_1, g) does not cross the region $B^{3\eta}_{2\epsilon}$ and thus is a Reeb orbit of λ_{FH}. This gives a bijective correspondence between closed geodesics of (S_1, g) which are not homotopic to a multiple of ∂S_1 and Reeb orbits of λ_{FH} which are contained in M_1.

Let $\Lambda(S_1)$ denote the set of free homotopy classes of curves in S_1 which are not covers of $[\partial S_1]$. We know that each $\rho \in \Lambda(S_1)$ contains exactly one closed geodesic $\mathfrak{c}_\rho$. The canonical lift γ_ρ of $\mathfrak{c}_\rho$ to T_1S_1 is a Reeb orbit of λ_g. As we saw above, each γ_ρ can also be seen as a Reeb orbit of λ_{FH}. Because of the negative curvature of g we know that the geodesic $\mathfrak{c}_\rho$ is hyperbolic. This implies that γ_ρ is a nondegenerate Reeb orbit of λ_g, and as λ_{FH} coincides with λ_g on a neighbourhood of γ_ρ, we conclude that γ_ρ is also nondegenerate when viewed as a Reeb orbit of λ_{FH}.

We will denote by $\Lambda(S_1)^{\leq T}$ the set of primitive of free homotopy classes in $\Lambda(S_1)$ whose unique closed geodesic has period smaller or equal to T. Because g is hyperbolic, it is a well known fact that there exist constants $a > 0$ and b such that $\#(\Lambda(S_1)^{\leq T}) \geq e^{aT+b}$. The map $\Theta\colon \Lambda(S_1) \to \Lambda(T_1S_1)$ (where $\Lambda(T_1S_1)$ is the free loop space of T_1S_1) associating with $\mathfrak{c}_\rho$ the Reeb orbit γ_ρ in T_1S_1 is easily seen to be injective. Because T_1S_1 is diffeomorphic to M_1, we can also view $\Theta(\Lambda(S_1))$ as a subset of the free loop space $\Lambda(M_1)$ of M_1.

Step 2 Let $i\colon M_1 \to M$ be the injection. As seen before, the boundary $\partial(i(M_1)) = \mathbb{T}$ is an incompressible torus in M. We consider the induced map of free loop spaces $i_*\colon \Lambda(M_1) \to \Lambda(M)$. As a consequence of the incompressibility of $\partial(i(M_1))$, the restriction of i_* to $\Theta(\Lambda(S_1))$ is injective.

To see this, it suffices to show the following claim: if ζ and ζ' are curves in M_1 which cannot be isotoped to a curve in ∂M_1 and which are in the same free homotopy class in M, then ζ and ζ' are freely homotopic in M_1. For ζ and ζ' satisfying the hypothesis of our claim, there is a cylinder cyl in M, whose boundary components are ζ and ζ', which intersects ∂M_1 transversely. Then cyl intersects ∂M_1 in a finite collection of curves $\{w_n\}$ which are all contractible in M; the contractibility of these curves is due to the fact that both ζ and ζ' cannot be isotoped to a curve contained in ∂M_1. The incompressibility of ∂M_1 implies that these $\{w_n\}$ are all contractible in ∂M_1. Now we cut the discs in cyl whose boundary are the curves w_n and substitute them by discs contained in ∂M_1. This produces a cylinder cyl$'$ completely contained in M_1 whose boundaries are ζ and ζ'. This implies that ζ and ζ' are already in the same free homotopy class in M_1, as we wished to show.

From step 1, we know that for each $\rho \in i_*(\Theta(\Lambda(S_1)))$, there is a Reeb orbit γ_ρ in ρ.

Step 3 We will show that for each $\rho \in i_*(\Theta(\Lambda(S_1)))$, the Reeb orbit γ_ρ considered in step 1 is the unique Reeb orbit of λ_{FH} in ρ.

Let γ be a Reeb orbit in ρ. If it is contained in M_1, we know that γ is a closed geodesic in (S_1, g). Using an argument as in step 2, it is easy to show that γ and γ_ρ are freely homotopic in M_1 and, therefore, also in T_1S_1. Projecting to S_1, we obtain that γ and γ_ρ are lifts of geodesics of (S_1, g) in the same free homotopy class of S_1. But for each free homotopy class of S_1, there is a unique closed geodesic of (S_1, g); this implies that $\gamma = \gamma_\rho$.

Step 3 will now follow if we prove the following:

Claim *Every Reeb orbit of λ_{FH} in ρ is completely contained in M_1.*

Proof of the claim If γ was contained in M_2, then it would be possible to isotope γ_ρ to a curve contained in ∂M_1. This is impossible by the definition of $\Lambda(S_1)$.

The only remaining possibility is that γ visits both M_1 and M_2. In this case, it has to visit both $M_1 \setminus N$ and $M_2 \setminus N$ (indeed, if γ is completely contained in $M_i \cup N$, convexity of the hyperbolic metric implies that γ is in M_i). We then know from Lemma 15 that every curve which is freely homotopic to γ has to intersect the torus $\mathbb{T}$. But γ_ρ, which is freely homotopic to γ, does not intersect $\mathbb{T}$. This contradiction rules out the possibility that γ visits both M_1 and M_2, and establishes the claim. □

Step 4 From the previous steps, we know that for each $\rho \in i_*(\Theta(\Lambda(S_1)))$, there exists a unique nondegenerate[5] Reeb orbit $\gamma_\rho \in \rho$. Hence for such ρ, the cylindrical contact homology $\mathbb{CH}^{\rho}_{\mathrm{cyl}}(\lambda_{\mathrm{FH}})$ is well-defined, and for Euler characteristic reasons, $\mathbb{CH}^{\rho}_{\mathrm{cyl}}(\lambda_{\mathrm{FH}}) \neq 0$.

Let $\rho \in i_*(\Theta(\Lambda(S_1)^{\leq T}))$. Then as we showed in the previous steps, the unique Reeb orbit of λ_{FH} in ρ has action at most T, and $\mathbb{CH}^{\rho}_{\mathrm{cyl}}(\lambda_{\mathrm{FH}}) \neq 0$. This implies that

$$N_T^{\mathrm{cyl}}(\lambda_{\mathrm{FH}}) \geq \#(i_*(\Theta(\Lambda(S_1)^{\leq T}))). \tag{6-31}$$

As i_* restricted to $\Theta(\Lambda(S_1)^{\leq T})$ is injective, and Θ is injective, we conclude that

$$\#(i_*(\Theta(\Lambda(S_1)^{\leq T}))) = \#(\Lambda(S_1)^{\leq T}) \geq e^{aT+b}. \tag{6-32}$$

Combining formulas (6-31) and (6-32), we obtain

$$N_T^{\mathrm{cyl}}(\lambda_{\mathrm{FH}}) \geq e^{aT+b}. \tag{6-33}$$

□

7 Conclusion

The works of Katok [32; 33] and of Lima and Sarig [35] imply that if ϕ is a smooth flow on a 3–manifold, generated by a nonvanishing vector field, then ϕ has positive topological entropy if and only if there exists a Smale "horseshoe" as a subsystem of

[5] Recall that we established in step 1 that γ_ρ is nondegenerate.

the flow. For a flow, a "horseshoe" is a compact invariant set where the dynamics is conjugate to that of the suspension of a shift map. In particular, the number of hyperbolic periodic orbits on a "horseshoe" of a 3–dimensional flow ϕ grows exponentially with respect to the period. We remark that the result obtained in the recent work of Lima and Sarig [35] is stronger: they show that there exists a compact invariant set $\mathcal{K}$ of ϕ where the dynamics is nonuniformly hyperbolic and such that $h_{\mathrm{top}}(\phi_{\mathcal{K}}) = h_{\mathrm{top}}(\phi)$.[6]

As a consequence, for the contact 3–manifolds (M, ξ) considered in Theorems 9 and 16, we have that for *every* Reeb flow on (M, ξ), the number of hyperbolic Reeb orbits grows exponentially with the action. This can be summarized by saying that all Reeb flows on these contact manifolds posses a "complicated" orbit structure which is forced to exist by the "complicated" contact topology of these contact manifolds.

An interesting property of the entropy estimate used in this paper, and also in [3] and [36], is that it gives estimates on the growth of the number of hyperbolic Reeb orbits for degenerate contact forms as well. This kind of information is not obtainable just by studying the growth rate of contact homology.

It is known that the consequences of positivity of topological entropy in higher dimensions are not as strong as in the low-dimensional case. In particular, positive topological entropy for a flow in dimension greater than 3 does not imply the existence of a "horseshoe" in the flow. It is, however, natural to ask the following question.

Question 1 In dimension greater than or equal to 5, does exponential homotopical growth of periodic orbits for a Reeb flow imply the existence of a compact invariant set where the dynamics is conjugate to a shift?

In another direction, one would like to know if it is possible to obtain more dynamical information about the Reeb flows on the contact manifolds covered by Theorems 9 and 16.

Question 2 Let (M, ξ) be a manifold satisfying the hypothesis of Theorem 9 or 16, and let λ be a contact form on (M, ξ). Is it true that for the Reeb flow ϕ_{X_λ}, there exists an invariant region of positive measure (with respect to the measure $\lambda \wedge d\lambda$) on which the dynamics of the Reeb flow is ergodic?

One important property of many of the contact 3–manifolds covered in Theorem 9 is that they have positive Giroux torsion. By a theorem of Gay [23] (see also [41]),

[6] We remark that in [32], Katok proves analogous results for diffeomorphisms on surfaces and only states the results for flows on 3–manifolds in [33]. To the best of our knowledge, the complete proofs of all the results mentioned above for 3–dimensional flows with positive topological entropy only appeared in [35], which builds on the ideas of [32; 33].

manifolds with positive Giroux torsion are not strongly fillable. This implies that many of the contact manifolds satisfying the claims of Theorem 9 are not strongly fillable and therefore different from the unit tangent bundles studied in [36], which are fillable. It would be interesting to know if such examples also exist in higher dimensions.

Question 3 Are there examples of nonsymplectically fillable contact manifolds, with dimension at least 5, on which every Reeb flow has positive topological entropy? Are there examples, in dimension at least 5, of manifolds which admit infinitely many different contact structures such that, on all of them, every Reeb flow has positive topological entropy?

We remark also that in Theorem 9, we showed the existence of 3–manifolds with hyperbolic components which can be given infinitely many different contact structures whose Reeb flows always have positive topological entropy. From the perspective of 3–dimensional topology, it would be interesting to have examples of contact structures on hyperbolic 3–manifolds on which every Reeb flow has positive topological entropy.

Question 4 Are there examples of contact structures on closed hyperbolic 3–manifolds on which every Reeb flow has positive topological entropy?[7] Are there hyperbolic 3–manifolds which admit multiple nondiffeomorphic contact structures, on which every Reeb flow has positive topological entropy?

Lastly we mention that the techniques used in this paper, and in [3], can also be used in combination with the ideas of Momin [37] to establish chaotic behaviour of Reeb flows on $(S^3, \xi_{\text{tight}})$ when these Reeb flows have a special link as a Reeb orbit. This and similar results will appear in [5].

References

[1] **P Albers**, **B Bramham**, **C Wendl**, *On nonseparating contact hypersurfaces in symplectic 4–manifolds*, Algebr. Geom. Topol. 10 (2010) 697–737 MR

[2] **M R R Alves**, *Growth rate of Legendrian contact homology and dynamics of Reeb flows*, PhD thesis, Université Libre de Bruxelles (2014) Available at `http://tinyurl.com/ulb-MRRAlves-thesis-2014`

[3] **M R R Alves**, *Legendrian contact homology and topological entropy*, preprint (2014) arXiv

[7] Examples of contact structures on closed hyperbolic 3–manifolds on which every Reeb flow has positive topological entropy have recently been constructed by the author in [4].

[4] **M R R Alves**, *Positive topological entropy for Reeb flows on* 3*–dimensional Anosov contact manifolds*, preprint (2015) arXiv To appear in J. Mod. Dyn.

[5] **M R R Alves**, **P A S Salomão**, *Legendrian contact homology on the complement of Reeb orbits and topological entropy*, in preparation

[6] **F Bourgeois**, *A survey of contact homology*, 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) 45–71 MR

[7] **F Bourgeois**, **T Ekholm**, **Y Eliashberg**, *Effect of Legendrian surgery*, Geom. Topol. 16 (2012) 301–389 MR

[8] **F Bourgeois**, **Y Eliashberg**, **H Hofer**, **K Wysocki**, **E Zehnder**, *Compactness results in symplectic field theory*, Geom. Topol. 7 (2003) 799–888 MR

[9] **F Bourgeois**, **K Mohnke**, *Coherent orientations in symplectic field theory*, Math. Z. 248 (2004) 123–146 MR

[10] **R Bowen**, *Topological entropy and axiom* A, from "Global Analysis" (S-S Chern, S Smale, editors), Amer. Math. Soc., Providence, RI (1970) 23–41 MR

[11] **P Boyland**, *Isotopy stability of dynamics on surfaces*, from "Geometry and topology in dynamics" (M Barge, K Kuperberg, editors), Contemp. Math. 246, Amer. Math. Soc., Providence, RI (1999) 17–45 MR

[12] **V Colin**, **K Honda**, *Constructions contrôlées de champs de Reeb et applications*, Geom. Topol. 9 (2005) 2193–2226 MR

[13] **D L Dragnev**, *Fredholm theory and transversality for noncompact pseudoholomorphic maps in symplectizations*, Comm. Pure Appl. Math. 57 (2004) 726–763 MR

[14] **Y Eliashberg**, **A Givental**, **H Hofer**, *Introduction to symplectic field theory*, Geom. Funct. Anal. (2000) 560–673 MR

[15] **A Fathi**, **F Laudenbach**, **V Poenaru** (editors), *Travaux de Thurston sur les surfaces*, Astérisque 66, Société Mathématique de France, Paris (1979) MR

[16] **A Fel'shtyn**, *Dynamical zeta functions, Nielsen theory and Reidemeister torsion*, Mem. Amer. Math. Soc. 699, Amer. Math. Soc., Providence, RI (2000) MR

[17] **S R Fenley**, *Homotopic indivisibility of closed orbits of* 3*–dimensional Anosov flows*, Math. Z. 225 (1997) 289–294 MR

[18] **P Foulon**, **B Hasselblatt**, *Contact Anosov flows on hyperbolic* 3*–manifolds*, Geom. Topol. 17 (2013) 1225–1252 MR

[19] **U Frauenfelder**, **C Labrousse**, **F Schlenk**, *Slow volume growth for Reeb flows on spherizations and contact Bott–Samelson theorems*, J. Topol. Anal. 7 (2015) 407–451 MR

[20] **U Frauenfelder**, **F Schlenk**, *Volume growth in the component of the Dehn–Seidel twist*, Geom. Funct. Anal. 15 (2005) 809–838 MR

[21] **U Frauenfelder**, **F Schlenk**, *Fiberwise volume growth via Lagrangian intersections*, J. Symplectic Geom. 4 (2006) 117–148 MR

[22] **U Frauenfelder**, **F Schlenk**, *Filtered Hopf algebras and counting geodesic chords*, Math. Ann. 360 (2014) 995–1020 MR

[23] **D T Gay**, *Four-dimensional symplectic cobordisms containing three-handles*, Geom. Topol. 10 (2006) 1749–1759 MR

[24] **M Gromov**, *Pseudoholomorphic curves in symplectic manifolds*, Invent. Math. 82 (1985) 307–347 MR

[25] **M Handel**, **W P Thurston**, *Anosov flows on new three manifolds*, Invent. Math. 59 (1980) 95–103 MR

[26] **H Hofer**, *Pseudoholomorphic curves in symplectizations with applications to the Weinstein conjecture in dimension three*, Invent. Math. 114 (1993) 515–563 MR

[27] **H Hofer**, **K Wysocki**, **E Zehnder**, *Properties of pseudoholomorphic curves in symplectisations, I: Asymptotics*, Ann. Inst. H. Poincaré Anal. Non Linéaire 13 (1996) 337–379 MR

[28] **H Hofer**, **K Wysocki**, **E Zehnder**, *Properties of pseudoholomorphic curves in symplectizations, III: Fredholm theory*, from "Topics in nonlinear analysis" (J Escher, G Simonett, editors), Progr. Nonlinear Differential Equations Appl. 35, Birkhäuser, Basel (1999) 381–475 MR

[29] **H Hofer**, **K Wysocki**, **E Zehnder**, *Finite energy foliations of tight three-spheres and Hamiltonian dynamics*, Ann. of Math. 157 (2003) 125–255 MR

[30] **U Hryniewicz**, **A Momin**, **P A S Salomão**, *A Poincaré–Birkhoff theorem for tight Reeb flows on S^3*, Invent. Math. 199 (2015) 333–422 MR

[31] **B Jiang**, *Estimation of the number of periodic orbits*, Pacific J. Math. 172 (1996) 151–185 MR

[32] **A Katok**, *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms*, Inst. Hautes Études Sci. Publ. Math. 51 (1980) 137–173 MR

[33] **A Katok**, *Entropy and closed geodesics*, Ergodic Theory Dynam. Systems 2 (1982) 339–365 MR

[34] **A Katok**, **B Hasselblatt**, *Introduction to the modern theory of dynamical systems*, Encyclopedia of Math. and its Applications 54, Cambridge Univ. Press (1995) MR

[35] **Y Lima**, **O Sarig**, *Symbolic dynamics for three-dimensional flows with positive topological entropy*, preprint (2014) arXiv

[36] **L Macarini**, **F Schlenk**, *Positive topological entropy of Reeb flows on spherizations*, Math. Proc. Cambridge Philos. Soc. 151 (2011) 103–128 MR

[37] **A Momin**, *Contact homology of orbit complements and implied existence*, J. Mod. Dyn. 5 (2011) 409–472

[38] **G P Paternain**, *Geodesic flows*, Progress in Mathematics 180, Birkhäuser, Boston (1999) MR

[39] **C Robinson**, *Dynamical systems: Stability, symbolic dynamics, and chaos*, CRC Press, Boca Raton, FL (1995) MR

[40] **A Vaugon**, *On growth rate and contact homology*, Algebr. Geom. Topol. 15 (2015) 623–666 MR

[41] **C Wendl**, *Strongly fillable contact manifolds and J–holomorphic foliations*, Duke Math. J. 151 (2010) 337–384 MR

Institut de Mathématiques, Université de Neuchâtel,
Rue Emile-Argand 11, CH-2000 Neuchâtel, Switzerland

marcelorralves@gmail.com

Proposed: Yasha Eliashberg — Received: 18 August 2015
Seconded: Ciprian Manolescu, Leonid Polterovich — Revised: 14 November 2015

Geometry & Topology 20 (2016) 3571–3621

A 1–parameter family of spherical CR uniformizations of the figure eight knot complement

MARTIN DERAUX

We describe a simple fundamental domain for the holonomy group of the boundary unipotent spherical CR uniformization of the figure eight knot complement, and deduce that small deformations of that holonomy group (such that the boundary holonomy remains parabolic) also give a uniformization of the figure eight knot complement. Finally, we construct an explicit 1–parameter family of deformations of the boundary unipotent holonomy group such that the boundary holonomy is twist-parabolic. For small values of the twist of these parabolic elements, this produces a 1–parameter family of pairwise nonconjugate spherical CR uniformizations of the figure eight knot complement.

22E40, 32V05, 57M50

1 Introduction

The existence of a complete hyperbolic structure on a 3–manifold has important topological consequences. For instance, this gives a definition of the volume of a knot (when a knot admits a complete hyperbolic structure, that structure is unique by Mostow rigidity, so the volume of that metric is a well-defined invariant).

In this paper, we focus on another kind of geometric structures on 3–manifolds, namely structures modeled on the boundary of a symmetric space X of negative curvature (transition maps are required to be locally given by isometries of X). The visual boundary $\partial_\infty X$ is then a 3–dimensional sphere if $X = H^4_{\mathbb{R}}$ or $H^2_{\mathbb{C}}$.

The first case gives rise to the theory of flat conformal structures, and the second one to the theory spherical CR structures. In the first case, one considers the unit ball model of $H^4_{\mathbb{R}}$, so the visual boundary is $S^3 \subset \mathbb{R}^4$, and the group of isometries of $H^4_{\mathbb{R}}$ acts as Möbius transformations (ie transformations that map spheres into spheres, of possibly infinite radius). Alternatively, one can use stereographic projection and think of S^3 as $\mathbb{R}^3 \cup \{\infty\}$; this would also correspond to using the upper half plane model for $H^3_{\mathbb{R}}$.

In the second case, using the ball model $\mathbb{B}^2 \subset \mathbb{C}^2$, one can identify $\partial_\infty H^2_{\mathbb{C}}$ with the unit sphere $S^3 \subset \mathbb{C}^2$. The action on the boundary is best understood in stereographic

Published: 21 December 2016 DOI: 10.2140/gt.2016.20.3571

projection, and identifying $S^3 \setminus \{p_\infty\} \simeq \mathbb{R}^3 \simeq \mathbb{C} \times \mathbb{R}$ with the Heisenberg group. Isometries of $H^2_{\mathbb{C}}$ fixing p_∞ then act as automorphisms of the Heisenberg group. Of course, the Heisenberg group acting on itself by left translations gives many automorphisms (which correspond to the action of unipotent matrices in $U(2,1)$), and one gets the full automorphism group by adjoining a rotation in $\mathbb{C} \times \mathbb{R}$ around the $\mathbb{R}$ factor, and a scaling of the form $(z,t) \mapsto (\lambda z, \lambda^2 t)$ (which corresponds to a loxodromic isometry); see Section 3B.

Even though a lot of partial results have been obtained (see Kamishima and Tsuboi [18], and Goldman [13], for instance), the classification of 3–manifolds that admit a spherical CR structure is far from understood. When a manifold admits a spherical CR structure, the moduli space of such structures is also quite mysterious.

In this paper, we will be interested in a special kind of spherical CR structures, namely spherical CR *uniformizations* (in the literature, these are sometimes called complete spherical CR structures). These are characterized by the fact that the developing map of the structure is a diffeomorphism onto its image, which is an open set in S^3. In that case, the holonomy group is a discrete subgroup $\Gamma \subset \mathrm{PU}(2,1)$, and the image of the developing map is the domain of discontinuity Ω_Γ of Γ (ie the largest open set where the action is proper). The quotient $\Gamma \setminus \Omega_\Gamma$ is called the *manifold at infinity* of Γ.

The classification of 3–manifolds that admit a spherical CR uniformization is also an open problem. Recall that $H^2_{\mathbb{C}}$ is homogeneous under the action of $\mathrm{PU}(2,1)$, and the isotropy group of a point is isomorphic to $U(2)$. In particular, finite subgroups of $U(2)$ such that nontrivial elements fix only the origin (in other words the groups should not contain any complex reflection) yield spherical CR uniformizable 3–manifolds with finite fundamental group.

In a similar vein, quotients of the Heisenberg group yield Nil manifolds that trivially admit a spherical CR uniformization such that the holonomy group has a global fixed point, which is now in $\partial_\infty H^2_{\mathbb{C}}$ instead of $H^2_{\mathbb{C}}$.

It is also natural to consider stabilizers of totally geodesic subspaces in $H^2_{\mathbb{C}}$, namely copies of $H^2_{\mathbb{R}}$ or $H^1_{\mathbb{C}}$. In that setting, Fuchsian groups (ie discrete subgroups of $\mathrm{SO}(2,1)$ or $\mathrm{SU}(1,1)$, seen as subgroups of $\mathrm{SU}(2,1)$) produce as their manifold at infinity a circle bundle over a surface (or more generally over a 2–orbifold). This class is more interesting than the previous one, because it is known that the corresponding groups often admit deformations (but not always: see Toledo [29]). We will summarize the results in this well developed line of research by saying simply that many Seifert 3–manifolds admit spherical CR uniformizations; see Goldman and Kapovich [15], Anan'in, Grossi and Gusevskii [1], Parker and Platis [20], Will [30] and others.

The class of *hyperbolic* manifolds that admit a spherical CR uniformization is also far from being understood. In a number of beautiful results that appeared in the last decade, Schwartz [25; 27; 28] discovered that many hyperbolic manifolds admit spherical CR uniformizations. His starting point was to consider representations of triangle groups into $\mathrm{PU}(2,1)$ (see Schwartz [26]), and to determine the manifold at infinity of well-chosen such representations.

More recently, the figure eight knot complement was shown to admit a spherical CR uniformization by the author and Falbel [7] through a somewhat different strategy, namely, it was found as a byproduct of Falbel's program for finding representations of fundamental groups of triangulated 3–manifolds into $\mathrm{PU}(2,1)$ (see Falbel [9]), or in $\mathrm{PGL}(3,\mathbb{C})$ (see Bergeron, Falbel and Guilloux [3]).

Falbel's construction turned out to produce lots of representations, and in fact, so many that the geometric properties of the resulting representations are, in general, difficult to analyze. In order to make the list more tractable (and also for other reasons related to the study of Bloch groups), the search is often restricted to representations such that peripheral subgroups are mapped to unipotent matrices (matrices with 1 as their only eigenvalue). The boundary unipotent representations for noncompact 3–manifolds with low complexity (ie those that can be built by gluing up to three ideal tetrahedra) are listed in Falbel, Koseleff and Rouillier [11], and the geometry of some of these representations are analyzed in [7] and by the author in [6]. It turns out very few representations in that list are discrete.

It is quite clear, however, that the unipotent restriction is somewhat artificial. Part of the point of the present paper is to show that, at least in some cases, there are many boundary parabolic representations that are not unipotent, and that these representations carry just as much interesting geometric information about the 3–manifold.

Let M denote the figure eight knot complement. The main goal of this paper is to show that M admits a 1–parameter family of pairwise nonconjugate spherical CR uniformizations.

We will build on the fact that M admits a unique spherical CR uniformization with unipotent boundary holonomy, as was shown in [7]. For future reference, we will refer to that structure simply as *the* boundary unipotent uniformization of M (see the precise uniqueness statement in [7]), and we denote the corresponding holonomy representation by ρ. In view of Schwartz's spherical CR Dehn surgery theorem [28], one expects that small deformations of the boundary unipotent holonomy representation should still be discrete, and they should have a manifold at infinity given by *some* Dehn filling of the figure eight knot complement.

In order to turn this into a proof, one could try and prove that the boundary unipotent representation satisfies the hypotheses of Schwartz's theorem, ie that its image is a horotube group (without exceptional parabolic elements), and that its limit set is porous. If that works, then it is enough to show that the group admits deformations, and to study the type of the deformed unipotent element; Schwartz's surgery formula shows, in particular, that (under some technical assumptions) if there are deformations where the unipotent peripheral holonomy stays parabolic, then the manifold at infinity should not change at all in small deformations.

Although a few examples of noncompact hyperbolic manifolds are known to admit spherical CR uniformizations (see [25; 27; 7]), the deformation theory of the holonomy representations of these examples is still quite mysterious. In particular, there are only two examples where nontrivial deformations are known to exist such that peripheral elements map to parabolic elements. These two examples are the figure eight knot complement and the Whitehead link complement. The results announced by Parker and Will [21] say that there are at least two different spherical CR uniformizations of the Whitehead link complement, and that there is a 1–parameter family of representations interpolating between their holonomy representations.

Our first result gives an explicit construction of twist-parabolic deformations.

Theorem 1.1 *There is a continuous* 1*–parameter family of irreducible representations* $\rho_t\colon \pi_1(M)\to \mathrm{PU}(2,1)$*, such that* ρ_t*, for each* t*, maps peripheral subgroups of* M *onto a cyclic group generated by a single parabolic element with eigenvalues* e^{it}, e^{it}, e^{-2it}.

Given the eigenvalue condition, it should be clear that the representations ρ_t are pairwise nonconjugate. We will choose ρ_t so that ρ_0 is the holonomy of the boundary unipotent spherical CR uniformization.

Note that the existence of such parabolic deformations was independently discovered by Pierre–Vincent Koseleff, using a variant of the method devised by Falbel to parametrize boundary unipotent representations of 3–manifolds; see [9; 3; 11], for instance. An alternative parametrization of this family can also be obtained from the description of the full character variety; see Falbel, Guilloux, Koseleff, Rouillier and Thistlethwaite [10], and also Heusener, Muñoz and Porti [17].

We will use a more naïve construction, which is closer in spirit to the parametrization of the character variety of the figure eight knot group (or more generally 2–bridge knot groups) into $\mathrm{PSL}_2(\mathbb{C})$ by Riley [23].

Our main result is the following.

Theorem 1.2 *There exists a $\delta > 0$ such that for $|t| < \delta$, ρ_t is the holonomy of a spherical CR uniformization of the figure eight knot complement.*

In order to show this, we will study the Ford domain for the image of ρ_0, and we will show that it is generic enough for its combinatorics to be preserved under small deformations of ρ_0. Note that this argument turns out to fail for the Ford domain of the holonomy of the spherical CR uniformization of the Whitehead link complement announced by Parker and Will in [21]. Indeed, their Ford domain has the same local combinatorial structure as the Dirichlet domain described in [7], and in particular, it has lots of tangent spinal spheres.

It will be clear to the reader familiar with the notion of horotubes [28] that the Ford domain exhibits an explicit horotube structure for the group, but since our construction of horotubes is actually very close to proving Theorem 1.2, we will give a detailed argument that does not quote Schwartz's result. Of course, in many places, our proof parallels some of the intermediate results in [28].

We will not attempt to give an explicit allowable range of parameters t in Theorem 1.2, although it would certainly be interesting to do so (and also to try and make this range optimal).

The bulk of the work will be to describe the Ford domain for the holonomy group of the unipotent uniformization of M, and to study in detail the generic character of the intersection of its sides, along facets of all dimensions. The genericity that we will prove is genericity at infinity, namely, we will show that each ideal vertex in the Ford domain lies on precisely three sides that intersect transversely at that point. For finite vertices, no genericity is to be expected, since the group is known to contain elliptic elements of orders 3 and 4; see [7]. In fact, all the deformations we consider will preserve the conjugacy classes of these elliptic elements, and we will show that they do not affect the nongeneric character of the fundamental domains at these points:

Proposition 1.3 *The image of ρ_t is a triangle group. More specifically, for all t, we have*

$$\rho_t(g_2)^4 = \rho_t(g_1 g_2)^3 = \rho_t(g_2 g_1 g_2)^3 = \mathrm{id}.$$

Acknowledgements This work was partly supported by the ANR, through the grant SGT ("Structures Géométriques Triangulées"). I also benefited from generous support from the GEAR network (NSF grants DMS 1107452, 1107263, 1107367), via funding of a long term visit at ICERM. I am pleased to thank ICERM for its hospitality, and the participants of the semester program entitled "Low-dimensional Topology, Geometry, and Dynamics" for inspiring interactions. Finally, I would like to thank the referee for several suggestions that helped improve the readability of the paper.

2 The real hyperbolic Ford domain

Throughout this section, we denote by M the figure eight knot complement. We review the description of a cusp neighborhood for M. This is probably familiar to most readers, but the details will be used in the identification of the manifold at infinity of our complex hyperbolic groups. Moreover, quite remarkably, the local combinatorics of the real hyperbolic Ford domain turn out to be exactly the same as the local combinatorics of our fundamental domain for the action of the group on the domain of discontinuity.

Recall that the fundamental group $\pi_1(M)$ has a presentation of the form

$$\langle g_1, g_2, g_3 \mid g_2 = [g_3, g_1^{-1}],\ g_1 g_2 = g_2 g_3 \rangle,$$

with peripheral subgroup generated by g_3^{-1} and $g_1(g_1g_2)^{-1}g_3g_2g_3^{-1}$.

From this, one can find all type-preserving representations of $\pi_1(M)$ up to conjugation, as in [22]. Indeed, the generators g_1 and g_3 should be parabolic elements in $\mathrm{SL}_2(\mathbb{C})$, which we denote by G_1 and G_3. We may assume G_1 (resp. G_3) fixes 0 (resp. ∞), and since all parabolic elements are conjugate, we may also assume

$$G_1 = \begin{pmatrix} 1 & 0 \\ -\omega & 1 \end{pmatrix} \quad \text{and} \quad G_3 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

for some $\omega \in \mathbb{C}$. The relation $G_1[G_3, G_1^{-1}] = [G_3, G_1^{-1}]G_3$ in $\mathrm{PSL}_2(\mathbb{C})$ is easily seen to imply $\omega^2 + \omega + 1$, so we may take

$$\omega = \frac{-1 + i\sqrt{3}}{2}.$$

The stabilizer of ∞ in $\mathrm{PSL}_2(\mathbb{Z}[\omega])$ is clearly given by translations by Eisenstein integers, but the stabilizer in the group generated by G_1 and G_3 is slightly smaller, it can be checked to be generated by translations by 1 and $2i\sqrt{3}$; see [22] for more details.

Recall that the Ford isometric sphere of an element

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is bounded by the circle $|cz + d| = 1$. The Ford domain turns out to be the intersection of the exteriors of all spheres of radius 1 centered at Eisenstein integers. A schematic picture is shown in Figure 1, where the sides corresponding to $G_1^{\pm 1}$ are shaded in the same color, so the corresponding 2–faces get identified by the corresponding isometries, and similarly for $G_2^{\pm 1} = [G_3, G_1^{-1}]^{\pm 1}$. The complete description of identifications on bottom face of the prism is given in Figure 2, and there are also identifications on the vertical sides of the prism, which are simply given by translations whenever these

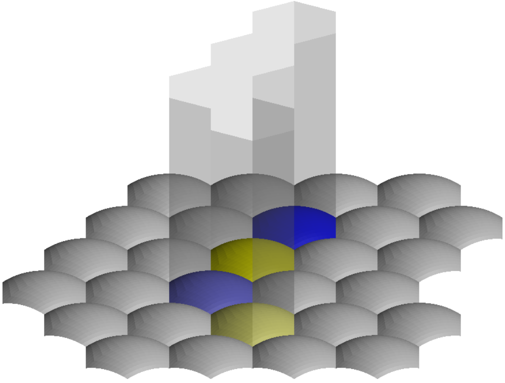


Figure 1: A fundamental domain for the action of Γ is an infinite chimney over the union of four hexagons, each hexagon living in a unit hemisphere around the appropriate Eisenstein integer.

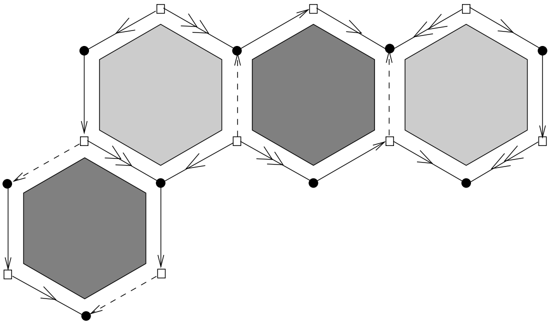


Figure 2: Bottom of the prism (spine of the figure eight knot complement)

sides are parallel. Note that these identifications are described in [22]; using current computer technology, they can also be found using the pictures produced by SnapPy.

3 Basic complex hyperbolic geometry

In this section, we review some basic material about the complex hyperbolic plane. The reader can find more details in [14].

Recall that $\mathbb{C}^{2,1}$ denotes $\mathbb{C}^3$ equipped with a Hermitian form of signature $(2,1)$. The standard such form is given by $\langle V, W\rangle = V_1\overline{W}_3 + V_2\overline{W}_2 + V_3\overline{W}_1 = W^*JV$, where

$$J = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

We denote by $U(2,1)$ the subgroup of $\mathrm{GL}(3,\mathbb{C})$ that preserves that Hermitian form, and by $\mathrm{PU}(2,1)$ the same group modulo scalar matrices. It is sometimes convenient to work with $\mathrm{SU}(2,1)$, which is a 3–fold cover of $\mathrm{PU}(2,1)$.

The complex hyperbolic plane $H^2_{\mathbb{C}}$ is the set of negative complex lines in $\mathbb{C}^{2,1}$, equipped with a Kähler metric that is invariant under the action of $\mathrm{PU}(2,1)$. Such a metric is unique up to scaling, and it turns out to have constant holomorphic sectional curvature (which one can choose to be -1).

It is well known that the maximal totally geodesic submanifolds of $H^2_{\mathbb{C}}$ are copies of $H^1_{\mathbb{C}}$ (with curvature -1) and copies of $H^2_{\mathbb{R}}$ (with curvature $-1/4$).

3A Bisectors

The corresponding distance function is given by

$$\cosh^2 \tfrac{1}{2}d(z,w) = \frac{|\langle Z,W\rangle|^2}{\langle Z,Z\rangle\langle W,W\rangle},$$

where Z (resp. W) denotes a representative of z (resp. w). Given two distinct points $p,q \in H^2_{\mathbb{C}}$, the locus $\mathcal{B}(p,q)$ of points that are equidistant of p and q is called a bisector. Beware that isometries switching p and q do not fix the corresponding bisector pointwise, and in fact bisectors are not totally geodesic. The copies of $H^1_{\mathbb{C}}$ (resp. $H^2_{\mathbb{R}}$) in $\mathcal{B}(p,q)$ are called its complex (resp. real) slices. All real slices intersect along the same real geodesic, called the *real spine* of the bisector; see [14].

Every bisector in $H^2_{\mathbb{C}}$ is diffeomorphic to the unit ball in $\mathbb{R}^3$ in such a way that the vertical axis is the real spine, complex slices are horizontal disks, and real slices are disks in vertical planes containing the vertical axis. One way to do this explicitly for the bisector $\mathcal{B}(p,q)$ is to scale q by a complex number of modulus one so that $\langle p,q\rangle$ is real and negative. Then an orthogonal basis for $\mathbb{C}^{2,1}$ is given by $v_0 = p+q$, $v_1 = p-q$, $v_2 = v_0 \boxtimes v_1$ ($\boxtimes$ denotes the Hermitian cross product; see page 43 of [14]). Of course, this basis can be made Lorentz orthonormal by scaling its vectors so that $\langle v_0,v_0\rangle = -1$, $\langle v_1,v_1\rangle = 1$ and $\langle v_2,v_2\rangle = 1$. The bisector then can be parametrized by $(z,t) \in \mathbb{C}\times\mathbb{R}$ by taking vectors of the form

$$v_0 + itv_1 + zv_2.$$

Given a set $S \subset H^2_{\mathbb{C}}$, we write $\mathcal{B}(S)$ for the locus equidistant of all point in S, which can be thought of as an intersection of bisectors.

The intersection of two bisectors is usually not totally geodesic, but it can be in some rare instances. When p, q and r are not in a common complex line (ie when lifts of these vectors are linearly independent), the locus $\mathcal{B}(p,q,r)$ of points equidistant of p, q and r is a smooth disk that is not totally geodesic, and is often called a Giraud disk; see [12]. The following property is crucial when studying fundamental domains; see [12; 14].

Theorem 3.1 *If p, q and r are not in a common complex line, then $\mathcal{B}(p,q,r)$ is contained in precisely three bisectors, namely $\mathcal{B}(p,q)$, $\mathcal{B}(q,r)$ and $\mathcal{B}(q,r)$.*

Note that checking whether an isometry maps a Giraud disk to another is equivalent to checking that the corresponding triple of points are mapped to each other.

In order to study Giraud disks, we will use spinal coordinates. The complex slices of $\mathcal{B}(p,q)$ are given explicitly by choosing a lift $\widetilde{p}$ (resp. $\widetilde{q}$) of p (resp. q).

When $p,q \in H^2_{\mathbb{C}}$, we simply choose lifts such that $\langle \widetilde{p},\widetilde{p}\rangle = \langle \widetilde{q},\widetilde{q}\rangle$. In this paper, we will mainly use these parametrizations when $p,q \in \partial_\infty H^2_{\mathbb{C}}$. In that case, the condition $\langle \widetilde{p},\widetilde{p}\rangle = \langle \widetilde{q},\widetilde{q}\rangle$ is vacuous, since all lifts are null vectors; we then choose some fixed lift $\widetilde{p}$ for the center of the Ford domain, and we take $\widetilde{q} = G\widetilde{p}$ for some $G \in U(2,1)$. If a different matrix $G' = SG$, with S a scalar matrix, note that the diagonal element of S is a unit complex number, so $\widetilde{q}$ is well defined up to a unit complex number.

The complex slices of $\mathcal{B}(p,q)$ are obtained as (the set of negative lines in) $(\bar{z}\widetilde{p}-\widetilde{q})^\perp$ for some arc of values of $z \in S^1$, which is determined by requiring that $\langle \bar{z}\widetilde{p}-\widetilde{q}, \bar{z}\widetilde{p}-\widetilde{q}\rangle > 0$.

Since a point of the bisector is on precisely one complex slice, we can parametrize $\mathcal{B}(p,q,r)$ by $(z_1,z_2) \in S^1 \times S^1$ via

$$V(z_1,z_2) = (\bar{z}_1 p - q) \boxtimes (\bar{z}_2 p - r) = q \boxtimes r + z_1 r \boxtimes p + z_2 p \boxtimes q. \tag{1}$$

The Giraud disk corresponds to the $(z_1,z_2) \in S^1 \times S^1$ with $\langle V(z_1,z_2), V(z_1,z_2)\rangle < 0$ (it follows from the fact that the bisectors are covertical that this region is a topological disk, but this is not obvious; see Chapters 8 and 9 in [14]).

The boundary at infinity $\partial_\infty \mathcal{B}(p,q,r)$ is a circle, given in spinal coordinates by the equation

$$\langle V(z_1,z_2), V(z_1,z_2)\rangle = 0. \tag{2}$$

Note that the choice of two lifts of q and r affects the spinal coordinates by rotation on each of the S^1 factors.

A defining equation for the trace of another bisector $\mathcal{B}(a,b)$ on the Giraud disk $\mathcal{B}(p,q,r)$ can be written in the form

$$|\langle V(z_1,z_2), a\rangle| = |\langle V(z_1,z_2), b\rangle|, \tag{3}$$

provided that a and b are suitably chosen lifts. The expressions $\langle V(z_1,z_2),a\rangle$ and $\langle V(z_1,z_2),b\rangle$ are affine in z_1, z_2.

These triple bisector intersections can be parametrized fairly explicitly, because one can solve the equation $|\langle V(z_1,z_2),a\rangle|^2 = |\langle V(z_1,z_2),b\rangle|^2$ for one of the variables z_1 or z_2 simply by solving a quadratic equation. A detailed explanation of how this works can be found in Section 2.3 of [7]; we will also review this in Section 5C3.

Note that our parameters also give a parametrization of the intersection in $P^2_{\mathbb{C}}$ of the extors extending the bisectors; see Chapter 8 of [14]. The Giraud disk is a disk in the intersection of the extors, which is a torus.

3B The Siegel domain and the Heisenberg group

The complex analogue of the upper half space model for $H^n_{\mathbb{R}}$ is the Siegel domain, which is obtained by sending the line spanned by $(1,0,0)$ to infinity. We denote the corresponding point of $\partial_\infty H^2_{\mathbb{C}}$ by p_∞.

More precisely, we take affine coordinates $z_1 = Z_1/Z_3$ and $z_2 = Z_2/Z_3$, and a negative complex line has a unique representative of the form $z = (z_1, z_2, 1)$ with

$$z^* J z = 2\,\mathfrak{Re}(z_1) + |z_2|^2 < 0.$$

Since we are interested in geometric structures modeled on $\partial_\infty H^2_{\mathbb{C}}$, we will use mainly the boundary of the Siegel domain, which is given by points $z = (z_1, z_2, 1)$ with $2\,\mathfrak{Re}(z_1) + |z_2|^2 = 0$. It is best understood in terms of Heisenberg geometry, as we now briefly recall.

A large part of the stabilizer of the point at infinity is given by unipotent upper triangular matrices. One easily checks that such a matrix preserves the Hermitian form J if and only if it can be written as

$$\begin{pmatrix} 1 & -\bar{a}\sqrt{2} & -|a|^2+is \\ 0 & 1 & a\sqrt{2} \\ 0 & 0 & 1 \end{pmatrix}$$

for some $(a,s) \in \mathbb{C}\times\mathbb{R}$. Since these upper triangular matrices form a group, we get a group law on $\mathbb{C}\times\mathbb{R}$, given by

$$(a,s) * (a',s') = (a+a', s+s'+2\,\mathfrak{Im}(a\bar{a}')). \tag{4}$$

This is the so-called Heisenberg group law.

The action of the unipotent stabilizer of p_∞ is simply transitive on $\partial_\infty H^2_{\mathbb{C}} - \{p_\infty\}$, so we will often identify the latter with $\mathbb{C}\times\mathbb{R}$.

The boundary at infinity of totally geodesic subspaces can be seen in somewhat simple terms in $\mathbb{C}\times\mathbb{R}$. The boundary of a copy of $H^1_{\mathbb{C}}$ (which is the intersection of an affine line in $\mathbb{C}^2$ with the Siegel half space) is called a $\mathbb{C}$–circle. These are ellipses that project to circles in $\mathbb{C}$ (or possibly vertical lines, if they go through p_∞).

The boundary of copies of $H^2_{\mathbb{R}}$ (which are images under arbitrary isometries of the set of real points in the Siegel half space) intersect the boundary at infinity in a so-called $\mathbb{R}$–circle. In the Heisenberg group, these are curves that project to lemniscates in $\mathbb{C}$ (or possibly straight lines when they go through p_∞). For more on this, see Chapter 4 of [14], for instance.

The full stabilizer of p_∞ is generated by the above unipotent group, together with the isometries of the forms

$$\begin{pmatrix}1&0&0\\0&e^{i\theta}&0\\0&0&1\end{pmatrix}\quad\text{and}\quad\begin{pmatrix}\lambda&0&0\\0&1&0\\0&0&1/\lambda\end{pmatrix},$$

where $\theta,\lambda\in\mathbb{R}$ and $\lambda\neq 0$. The first acts on Heisenberg as a rotation with vertical axis:

$$(a,s)\mapsto(e^{i\theta}a,s),$$

whereas the second one acts as

$$(a,s)\mapsto(\lambda a,\lambda^2 s).$$

There is a natural invariant metric on the Heisenberg group, called the Cygan metric, given by $d(g,g')=\|g^{-1}g'\|$, and the norm of an element of the Heisenberg group is given by

$$\|(z,t)\|=\left||z|^2+it\right|^{1/2}. \tag{5}$$

The Cygan sphere with center (z_0,t_0) and radius r has equation

$$\left||z-z_0|^2+i(t-t_0+2\,\mathfrak{Im}(z\overline{z}_0))\right|=r^2. \tag{6}$$

3C Ford domains and the Poincaré polyhedron theorem

Let Γ be a subgroup of $\mathrm{PU}(2,1)$, let $q\in\partial_\infty H^2_{\mathbb{C}}$ and let Q denote a lift of q in $\mathbb{C}^{2,1}$.

Definition 3.2 *The Ford domain for Γ centered at q is the set $F_{\Gamma,q}$ of points $z\in H^2_{\mathbb{C}}$ such that*

$$|\langle Z,Q\rangle|\leq|\langle Z,G(Q)\rangle|,$$

where G is a matrix representative of some element $g\in\Gamma$.

The inequality is actually independent of the lift $G \in U(2,1)$ chosen for $g \in \mathrm{PU}(2,1)$. For a given $g \in \Gamma$ and lift $G \in U(2,1)$, we denote by $\mathcal{B}_g$ the bisector given in homogeneous coordinates by

$$|\langle Z, Q\rangle| = |\langle Z, G(Q)\rangle|. \tag{7}$$

For concreteness, we mention that the boundary at infinity of $\mathcal{B}_g$ can be described as a Cygan sphere in the Heisenberg group; see Section 3B. The Cygan sphere corresponding to an element G has radius $\sqrt{2/|g_{31}|}$ (note that G fixes p_∞ if and only if $g_{31}=0$) and center $(\bar{g}_{32}/\bar{g}_{31}, 2\,\mathfrak{Im}(\bar{g}_{33}/\bar{g}_{31}))$; see (6).

We let $b_g = \mathcal{B}_g \cap F$; ie b_g is the side of F that lies on the bisector $\mathcal{B}_g$, and we refer to it as *the side corresponding to the group element* g. For a general $g \in \Gamma$, it may be that b_g has dimension smaller than 3 (in fact, it is often empty). A bisector of the form $\mathcal{B}_g$ such that b_g has dimension three will be called a *bounding bisector*.

The basic fact is that if q has trivial stabilizer in Γ, then $F = F_{\Gamma,q}$ is a fundamental domain for its action. However, it is customary to take q to have a nontrivial stabilizer $H \subset \Gamma$, in which case F is only a fundamental domain modulo the action of H. In other words, in that case, F is a fundamental domain for the decomposition of Γ into cosets of H.

It is usually very hard to determine F explicitly; in order to prove that a given polyhedron is equal to F, the main tool is the Poincaré polyhedron theorem. The basic idea is that the sides of F should be paired by isometries, and the images of F under these so-called side-pairing maps should give a local tiling of $H^2_{\mathbb{C}}$. If they do (and if the quotient of F by the identifications given by the side-pairing maps is complete), then the Poincaré polyhedron theorem implies that the images of F actually give a global tiling.

Once a fundamental domain is obtained, one gets an explicit presentation of Γ in terms of the generators given by the side-pairing maps together with a generating set for the stabilizer H, the relations corresponding to so-called ridge cycles (which correspond to the local tiling near each codimension-two face).

For more details on this theorem, see [7; 8; 19].

4 A boundary parabolic family of representations

In this section, we parametrize a neighborhood of the unipotent solution in the character variety $\chi(\pi_1(M), \mathrm{PU}(2,1))$. We will use the presentation

$$\langle g_1, g_2, g_3 \mid g_1g_2 = g_2g_3,\ g_2 = [g_3, g_1^{-1}]\rangle.$$

In order to describe representations, we seek to parametrize triples G_1, G_2, G_3 of matrices in $\mathrm{SU}(2,1)$ that satisfy the same relations as g_1, g_2, g_3 (possibly up to multiplication by a scalar matrix, since we are really after representations in $\mathrm{PU}(2,1)$).

If the fixed points of G_1 and G_3 are distinct, we may assume

$$G_1 = \begin{pmatrix} \lambda & a & b \\ 0 & \bar{\lambda}^2 & c \\ 0 & 0 & \lambda \end{pmatrix} \quad \text{and} \quad G_3 = \begin{pmatrix} \lambda & 0 & 0 \\ f & \bar{\lambda}^2 & 0 \\ e & d & \lambda \end{pmatrix}, \tag{8}$$

where $|\lambda| = 1$.

Note that the representation considered in [7] is obtained by taking

$$\lambda = 1, \quad a = d = 1, \quad c = f = -1, \quad b = \bar{e} = -\frac{1+i\sqrt{7}}{2}$$

in (8).

The fact that G_1 and G_3 are isometries of the form J implies

$$\begin{cases} c = -\bar{a}\bar{\lambda}, & |d|^2 + \bar{e}\lambda + e\bar{\lambda} = 0, \\ f = -\bar{d}\bar{\lambda}, & |a|^2 + \bar{b}\lambda + b\bar{\lambda} = 0. \end{cases} \tag{9}$$

We then compute the commutator $G_2 = [G_3, G_1^{-1}]$ and consider the system of equations given by $R = 0$, where

$$R = G_1 G_2 - G_2 G_3. \tag{10}$$

Note that this already restricts the character variety, since we only consider representations into $U(2,1)$ rather than $\mathrm{PU}(2,1)$, but this is fine if we are after a neighborhood of the boundary unipotent solution, where the relation (10) holds in $U(2,1)$.

The $(1,1)$–entry of R is given by

$$(|a|^2 e - |d|^2 b)(1 + \bar{a}d - \lambda^3 - \bar{\lambda}^3). \tag{11}$$

The first factor does not vanish for the boundary unipotent solution, so in its component we must have

$$1 + \bar{a}d = \lambda^3 + \bar{\lambda}^3. \tag{12}$$

Note that by conjugation by a diagonal matrix with diagonal entries k_1, k_2, k_3, we can assume that $a \in \mathbb{R}$ (and we can also impose that $|b|$ is given by any positive real number). Then (12) implies that d is real as well, so from this point on we assume

$$a, d \in \mathbb{R}.$$

The $(2,2)$–entry of R can then be written as

$$-(|a|^2e-|d|^2b)(a^2e\overline{\lambda}^4+a^2d^2\overline{\lambda}^3-ad+be\overline{\lambda}^5-1+bd^2\overline{\lambda}^4),$$

so we get the expression

$$a^2e\overline{\lambda}^4+a^2d^2\overline{\lambda}^3-ad+be\overline{\lambda}^5-1+bd^2\overline{\lambda}^4. \tag{13}$$

Using the relations (9) and (12), we have that (13) can be rewritten as

$$be\lambda=\lambda^3+\overline{\lambda}^3. \tag{14}$$

As mentioned above, by conjugation by a diagonal matrix, we can adjust $|b|$, for instance, so that

$$|b|^2=\lambda^3+\overline{\lambda}^3,$$

and in that case, (14) implies

$$|e|^2=|b|^2.$$

We will now show that, given λ, the following system has precisely two solutions:

$$\begin{cases} a^2+\overline{b}\lambda+b\overline{\lambda}=0,\\ d^2+\overline{e}\lambda+e\overline{\lambda}=0,\\ \qquad 1+ad=\lambda^3+\overline{\lambda}^3,\\ \qquad\; eb\lambda=\lambda^3+\overline{\lambda}^3,\\ \qquad\; |b|^2=\lambda^3+\overline{\lambda}^3. \end{cases} \tag{15}$$

In order to do that, note that the first four imply

$$b\overline{e}+\overline{b}e=1-2(\lambda^3+\overline{\lambda}^3),$$

and the last two imply

$$e=\overline{b\lambda}.$$

Putting these two together, we get

$$\mathfrak{Re}(b^2\lambda)=\tfrac{1}{2}-2\kappa, \tag{16}$$

where we have written

$$\kappa=(\lambda^3+\overline{\lambda}^3)/2. \tag{17}$$

The equation $\mathfrak{Re}(z)=\frac{1}{2}-2\kappa$ has a solution with $|z|=2\kappa$ if and only if

$$2\kappa\geq\tfrac{1}{2}-2\kappa,$$

and in that case one gets a simple formula for the solutions (intersect a vertical line with the circle of radius $|2\kappa|$ centered at the origin).

We get that (16) has solutions if and only if $\kappa \geq \frac{1}{8}$, and the solutions are given by

$$b^2\lambda = \tfrac{1}{2} - 2\kappa \pm i\sqrt{\tfrac{1}{2}\bigl(4\kappa - \tfrac{1}{2}\bigr)}. \tag{18}$$

This determines b up to its sign, opposite values clearly giving conjugate groups (they differ by conjugation by a diagonal matrix). The two values also yield isomorphic groups, obtained from each other by complex conjugation.

We will choose the solution to match the notation for the unipotent solution given in [7], which corresponds to $\lambda = 1$, $a = d = 1$, $b = -\frac{1}{2}(1 + i\sqrt{7})$ and $e = -\frac{1}{2}(1 - i\sqrt{7})$.

As a consequence, we take

$$b = -\frac{1 + i\sqrt{8\kappa - 1}}{2\sqrt{\lambda}},$$

where we take the square root to vary continuously near $\lambda = 1$.

The system (15) then gives values for the other parameters, namely

$$e = 2\kappa/b\lambda = -\frac{1 - i\sqrt{8\kappa - 1}}{2\sqrt{\lambda}},$$

and one easily writes explicit formulas for a and d (once again, these are determined only up to sign, but changing a to $-a$ can be effected by conjugation by a diagonal matrix). The formulas are

$$a = \sqrt{(4\mu^2-3)\mu + \sqrt{8\kappa-1}(4\mu^2-1)\nu}, \quad d = \sqrt{(4\mu^2-3)\mu - \sqrt{8\kappa-1}(4\mu^2-1)\nu},$$

where we have written $\sqrt{\lambda} = \mu + i\nu$ with μ, ν real. In terms of this new parameter, the condition $\kappa > \frac{1}{8}$ translates into

$$\mu > \cos\bigl(\tfrac{1}{3}\arctan\tfrac{\sqrt{7}}{3}\bigr) = 0.9711209254\ldots.$$

In fact, in order to get a and d to be real, we also need

$$(4\mu^2 - 3)\mu - \sqrt{8\kappa - 1}(4\mu^2 - 1)\nu \geq 0,$$

which translates into $\mu \geq \cos(\pi/18)$. The value $\mu = \cos(\pi/18)$ corresponds to a situation where $d = 0$.

4A Triangle group relations

The following matrices can be computed explicitly:

$$G_2=\begin{pmatrix}1+\lambda^3 & a\bar{\lambda}-\bar{b}d & (e+b)\bar{\lambda}\\ ab-d\bar{\lambda}^2 & -\lambda^3 & 0\\ (e+b)\bar{\lambda} & 0 & 0\end{pmatrix},$$

$$G_1G_2=\begin{pmatrix}\lambda & a(1-\lambda^3)-ed\lambda^2 & (e+b)\\ -\bar{\lambda}^2(ae+d\bar{\lambda}^2) & -\lambda & 0\\ (e+b) & 0 & 0\end{pmatrix},$$

$$G_1^2G_2=\begin{pmatrix}\bar{\lambda} & -\lambda^3(a\lambda+ed) & (e+b)\lambda\\ \lambda^2(ab+d\lambda) & -\bar{\lambda} & 0\\ (e+b)\lambda & 0 & 0\end{pmatrix}.$$

In particular,

$$\operatorname{tr}(G_2)=1,\quad \operatorname{tr}(G_1G_2)=0,\quad \operatorname{tr}(G_2G_1G_2)=0,$$

or in other words,

$$G_2^4=\mathrm{id},\quad (G_1G_2)^3=\mathrm{id},\quad (G_1^2G_2)^3=\mathrm{id}\,.$$

The last two relations imply that

$$(G_2G_1G_2)^3=\mathrm{id}\,.$$

Proposition 4.1 *Throughout the twist parabolic deformation, we have $G_1G_2=G_2G_3$, $G_2=[G_3,G_1^{-1}]$, $G_2^4=\mathrm{id}$, $(G_1G_2)^3=\mathrm{id}$, $(G_2G_1G_2)^3=\mathrm{id}$.*

4B Fixed points of elliptic elements

Note also that for each of the three matrices G_2, G_1G_2 and $G_1^2G_2$, the negative eigenvector is the one with eigenvalue 1 (indeed, this is true for the unipotent solution, so it holds throughout the corresponding component of the character variety).

For future reference, we give explicit formulas for these fixed points:

$$p_2=\big(1+\lambda^3,ab-d\bar{\lambda}^2,(\bar{\lambda}+\lambda^2)(e+b)\big),$$
$$p_{12}=\big(1+\lambda,-\bar{\lambda}^2(ae+d\bar{\lambda}^2),(1+\bar{\lambda})(e+b)\big),$$
$$p_{112}=\big(1+\bar{\lambda},\lambda^2(ab+d\bar{\lambda}),(\bar{\lambda}+\bar{\lambda}^2)(e+b)\big).$$

Lemma 4.2 *Throughout the deformation, p_2 is on six bounding bisectors, corresponding to the following group elements (written in word notation; see Section 5):*

$$2,\ \bar{2},\ 3,\ 12,\ \bar{1}\bar{2},\ \bar{1}3.$$

Proof The statement about $G_2^{\pm 1}$ is obvious since p_2 is fixed by G_2. The other four statements all follow from

$$d(p_2, p_0) = d(p_2, (G_2G_1)^{-1}p_0). \tag{19}$$

Indeed,

$$d(p_2, (G_2G_1)^{-1}p_0) = d(p_2, G_2^{-1}G_1^{-1}G_2^{-1}p_0) = d(p_2, G_1G_2p_0),$$

where we have used $G_1p_0 = p_0$ and $(G_1G_2)^3 = \mathrm{id}$. Similarly, using $G_1G_2 = G_2G_3$, we get

$$d(p_2, G_1G_2p_0) = d(p_2, G_2^{-1}G_1G_2p_0) = d(p_2, G_3p_0).$$

Finally, using $G_2 = [G_3, G_1^{-1}]$ we get

$$d(p_2, G_3p_0) = d(p_2, G_2^{-1}G_3p_0) = d(p_2, G_1^{-1}G_3p_0).$$

In order to prove (19), we compute

$$G_1^{-1}G_2^{-1}p_0 = (\bar{b} + \bar{e})\lambda(\bar{b}, a, \bar{\lambda}),$$

and we observe $|(\bar{b} + \bar{e})\lambda| = 1$, so we need only check

$$|\langle p_2, p_0\rangle| = |\langle p_2, X\rangle|,$$

where $X = (\bar{b}, a, \bar{\lambda})$. Now

$$|\langle p_2, p_0\rangle|^2 = |(\lambda + \bar{\lambda}^2)(\bar{e} + \bar{b})|^2 = |1 + \lambda^3|^2 = 2 + \lambda^3 + \bar{\lambda}^3,$$

and

$$\langle p_2, X\rangle = \bar{\lambda}(2 - \lambda^3 - \bar{\lambda}^3 - b^2\lambda),$$

and so,

$$|\langle p_2, X\rangle|^2 = 2 + \lambda^3 + \bar{\lambda}^3. \qquad \square$$

Lemma 4.3 *Through the deformation, $p_{\bar{1}21} = G_1^{-1}p_2$ stays on six bounding bisectors, corresponding to the following group elements (using the word notation introduced in the next section):*

$$2,\ \bar{1}2,\ \bar{1}\bar{2},\ \bar{1}3,\ \bar{1}\bar{1}\bar{2},\ \bar{1}\bar{1}3.$$

Proof The statement follows from Lemma 4.2 by conjugation by G_1^{-1} (which by definition fixes p_0). □

5 Combinatorics of the Ford domain in the unipotent case

In this section, we denote by Γ the image of ρ_0. It is generated by the matrices

$$G_1=\begin{pmatrix}1&1&\frac{1}{2}(-1-\sqrt{7}i)\\0&1&-1\\0&0&1\end{pmatrix},\quad G_3=\begin{pmatrix}1&0&0\\-1&1&0\\\frac{1}{2}(-1+\sqrt{7}i)&1&1\end{pmatrix}.$$

One then sets

$$G_2=[G_3,G_1^{-1}].$$

We will often use word notation in the generating set G_1, G_2, G_3, using bars to denote inverses. For instance, $23\bar{1}3$ denotes $G_2G_3G_1^{-1}G_3$.

We consider the Ford domain centered at the fixed point of G_1, which is p_∞ in the notation of Section 3C, and work in the Siegel half space. We let P denote $\langle G_1\rangle$, and F the corresponding Ford domain. We wish to prove that F is a fundamental domain for the action of the cosets of P in Γ.

We let S denote $\{G_2,G_2^{-1},G_3,G_3^{-1}\}$, and S^P the set of all conjugates of elements of S by powers of G_1. We consider the partial Ford domain D defined in homogeneous coordinates Z by the inequalities

$$|\langle Z,Q\rangle|\le|\langle Z,G(Q)\rangle|$$

for all $G\in S^P$. Clearly $F\subset D$, but we mean to prove:

Theorem 5.1 $F=D$.

The key steps in the proof of Theorem 5.1 will be the following:

- Determine the combinatorics of D.
- Show that the elements in S^P define side-pairing maps for D.
- Verify the hypotheses of the Poincaré polyhedron theorem.

5A Statement of the combinatorics

Clearly D is G_1–invariant, so it is enough to describe the combinatorics of the sides corresponding to $g\in S$, ie $g=G_2,G_3,G_2^{-1},G_3^{-1}$. We will call the corresponding four sides b_1, b_2, b_3 and b_4, respectively, and refer to them as core sides; the corresponding bisectors will be denoted by $\mathcal{B}_1$, $\mathcal{B}_2$, $\mathcal{B}_3$ and $\mathcal{B}_4$. The spinal spheres at infinity of these four bisectors will be denoted by $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$.

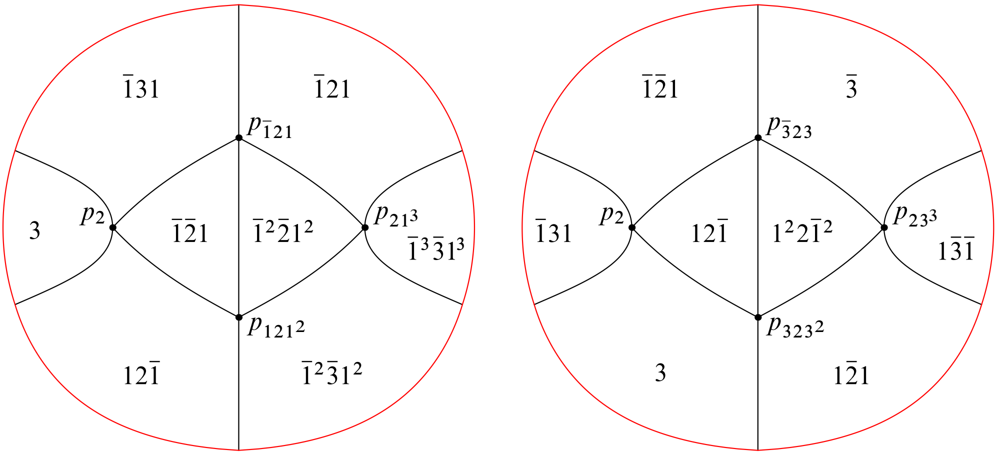


Figure 3: The combinatorics of the face corresponding to G_2 (left) and G_2^{-1} (right); all 2–faces are labeled, except for the boundary at infinity, which is a disk bounded by the most exterior curve (shown in red). We also label the finite vertices, namely, for $w \in \Gamma$, we let p_w denote the isolated fixed point of the group element corresponding to the word w ($1 = G_1$, $2 = G_2$, $3 = G_3$, $\bar{1} = G_1^{-1}$, etc).

We will sometimes index other sides than the four basic sides just described, mostly when describing computations that would unreasonable to perform by hand. We will order them by concatenating sets of four conjugates of the base group elements $2, \bar{2}, 3, \bar{3}$ by different powers of G_1, powers being arranged by increasing values of the absolute values of the exponent (positive powers first). The words corresponding to the first 20 bisectors are given by

$$2, \bar{2}, 3, \bar{3}, \quad 12\bar{1}, 1\bar{2}\bar{1}, 13\bar{1}, 1\bar{3}\bar{1}, \quad \bar{1}21, \bar{1}\bar{2}1, \bar{1}31, \bar{1}\bar{3}1,$$
$$1^22\bar{1}^2, 1^2\bar{2}\bar{1}^2, 1^23\bar{1}^2, 1^2\bar{3}\bar{1}^2, \quad \bar{1}^221^2, \bar{1}^2\bar{2}1^2, \bar{1}^231^2, \bar{1}^2\bar{3}1^2.$$

For example, $\mathcal{B}_5 = G_1(\mathcal{B}_1)$ is the bisector corresponding to $G_1G_2G_1^{-1}$ (or equivalently for G_1G_2, since G_1 fixes the center of our Ford domain), $\mathcal{B}_{10} = G_1^{-1}(\mathcal{B}_2)$ is the bisector for $G_1^{-1}G_2^{-1}G_1$.

We describe their combinatorics in the form of pictures; see Figures 3 and 4. Each picture is drawn in projection from a picture where the bisector is identified with the unit ball in $\mathbb{R}^3$; see Section 3A. Concretely, we use spinal coordinates on 2–faces and parametrize 1–faces by solving equations of the form (3) for one of the variables.

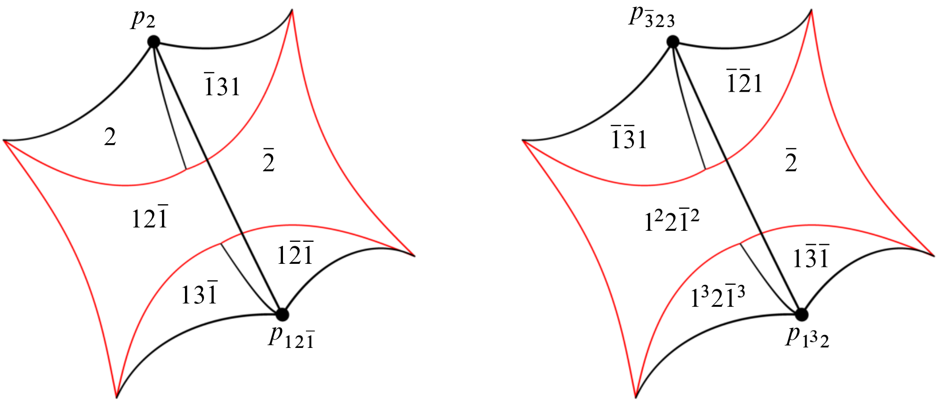


Figure 4: The combinatorics of the face corresponding to G_3 (left) and G_3^{-1} (right)

We also give a list of vertices on the core sides and a list of the bounding bisectors that each vertex lies on; see Tables 1 and 2.

5B Effective local finiteness

The goal of this section is to show that a given face of the Ford domain intersects only finitely many faces. Since the domain is G_1–invariant by construction, we start by normalizing G_1 in a convenient form. We will work in the Siegel half space; see Section 3B.

A natural set of coordinates is obtained by arranging that G_2^2 maps p_∞ to the origin in the Heisenberg group. There is a unique Heisenberg translation that achieves this, given by

$$Q = \begin{pmatrix} 1 & \frac{1}{4}(3 - i\sqrt{7}) & -\frac{1}{2} \\ 0 & 1 & \frac{1}{4}(-3 - i\sqrt{7}) \\ 0 & 0 & 1 \end{pmatrix}.$$

One then gets

$$QG_1Q^{-1} = \begin{pmatrix} 1 & 1 & -\frac{1}{2} \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{and} \qquad QG_2^2Q^{-1} = \begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ 0 & -1 & 0 \\ -2 & 0 & 0 \end{pmatrix}.$$

Of course, one could make the last matrix even simpler by composing with a loxodromic element.

Word	Bounding bisectors	Indices
2	$2, \bar{2}, 3, 12\bar{1}, \bar{1}\bar{2}1, \bar{1}31$	$1, 2, 3, 5, 10, 11$
$\bar{1}21$	$2, \bar{1}21, \bar{1}\bar{2}1, \bar{1}31, \bar{1}^2\bar{2}1^2, \bar{1}^231^2$	$1, 9, 10, 11, 18, 19$
21^3	$2, \bar{1}21, \bar{1}^2\bar{2}1^2, \bar{1}^2\bar{3}1^2, \bar{1}^3\bar{2}1^3, \bar{1}^3\bar{3}1^3$	$1, 9, 18, 20, 26, 28$
121^2	$2, 12\bar{1}, \bar{1}\bar{2}1, \bar{1}\bar{3}1, \bar{1}^2\bar{2}1^2, \bar{1}^2\bar{3}1^2$	$1, 5, 10, 12, 18, 20$

Table 1: Finite vertices on the face for G_2. For each vertex v, we give a word w for an element that fixes precisely v, and we list the words for the bounding bisectors that contain v.

Word	Bounding bisectors	Indices
2	$2, \bar{2}, 3, 12\bar{1}, \bar{1}\bar{2}1, \bar{1}31$	$1, 2, 3, 5, 10, 11$
$\bar{3}23$	$\bar{2}, \bar{3}, 12\bar{1}, \bar{1}\bar{2}1, \bar{1}\bar{3}1, 1^22\bar{1}^2$	$2, 4, 5, 10, 12, 13$
23^3	$\bar{2}, \bar{3}, 1\bar{2}\bar{1}, 1\bar{3}\bar{1}, 1^22\bar{1}^2, 1^32\bar{1}^3$	$2, 4, 6, 8, 13, 21$
323^2	$\bar{2}, 3, 12\bar{1}, 1\bar{2}\bar{1}, 13\bar{1}, 1^22\bar{1}^2$	$2, 3, 5, 6, 7, 13$

Table 2: Finite vertices on the face for G_2^{-1}

We let A_j denote QG_jQ^{-1}. We then have

$$A_2(\infty) = (\alpha, 0), \quad A_2^2(\infty) = (0, 0), \quad A_2^{-1}(\infty) = (-\alpha, 0),$$

where $\alpha = (3 + i\sqrt{7})/(4\sqrt{2})$. Also,

$$A_3(\infty) = \left(-\frac{1}{2\sqrt{2}}, -\frac{\sqrt{7}}{8}\right), \quad A_3^{-1}(\infty) = \left(-\frac{1}{\sqrt{2}}, \frac{\sqrt{7}}{2}\right).$$

The spinal sphere with center $(0, 0)$ and radius r has equation

$$(x^2 + y^2)^2 + t^2 = r^4, \tag{20}$$

so we get a spinal sphere centered at $(a + ib, u)$ by translation:

$$\big((x-a)^2 + (y-b)^2\big)^2 + (t - u - ay - bx)^2 = r^4. \tag{21}$$

By writing out (7), squaring both sides and identifying with (21), one checks that the spheres $\mathcal{S}_1$ and $\mathcal{S}_2$ have radius 1, whereas $\mathcal{S}_3$ and $\mathcal{S}_4$ have radius $2^{-1/4}$. We summarize this information in Table 3.

The action of A_1 on the Heisenberg group is given by

$$(z, t) \mapsto (z - 1, t + \mathfrak{Im}(z)), \tag{22}$$

Sphere	Center	Radius
$\mathcal{S}_1$	$\left(\dfrac{3+i\sqrt{7}}{4\sqrt{2}}, 0\right)$	1
$\mathcal{S}_2$	$\left(-\dfrac{3+i\sqrt{7}}{4\sqrt{2}}, 0\right)$	1
$\mathcal{S}_3$	$\left(-\dfrac{1}{2\sqrt{2}}, -\dfrac{\sqrt{7}}{8}\right)$	$2^{-1/4}$
$\mathcal{S}_4$	$\left(-\dfrac{1}{\sqrt{2}}, \dfrac{\sqrt{7}}{2}\right)$	$2^{-1/4}$

Table 3: Centers and radii of core spinal spheres

and in particular, we get the following:

Proposition 5.2 *The element A_1 preserves every $\mathbb{R}$–circle of the form $(x, 0, t_0)$ with $x \in \mathbb{R}$.*

Recall that $\mathbb{R}$–circles are, by definition, given by the trace at infinity of totally geodesic copies of $H^2_{\mathbb{R}}$ in $H^2_{\mathbb{C}}$. The corresponding real planes in $H^2_{\mathbb{C}}$ are preserved by A_1, and their union is the so-called *invariant fan* of A_1; see [16].

Among all these $\mathbb{R}$–circles, the x axis is somewhat special because of the following:

Proposition 5.3 *The $\mathbb{R}$–plane bounded by the x axis contains the fixed point of G_2.*

Indeed, the fixed point of A_2 is given by

$$V = \left(-\tfrac{1}{2}, 0, 1\right),$$

and for $W = (-x^2 + it, x\sqrt{2}, 1)$, we have

$$\langle V, p_\infty\rangle\langle p_\infty, W\rangle\langle W, V\rangle = -\tfrac{1}{2}(1 + x^2) + it,$$

which is real if and only if $t = 0$.

Note that (22) shows that for any two bisectors $\mathcal{B}_1$ and $\mathcal{B}_2$ not containing p_∞, we have $G_1^k \mathcal{B}_1 \cap \mathcal{B}_2 = \varnothing$ whenever k is large enough. Indeed, it follows from the detailed study of bisector intersection in [14] that, if two bisectors intersect, then the corresponding spinal spheres must intersect.

Moreover, this claim can easily be made effective; ie one can get explicit bounds on how large k needs to be for the above intersection to be empty. If $\mathcal{S}_j = \partial_\infty \mathcal{B}_j$ is contained in a strip $\alpha_j \le x \le \beta_j$, one can simply take $k > \beta_2 - \alpha_1$ or $k < \alpha_2 - \beta_2$. Note that bounds α_j, β_j can be computed fairly easily from the equations of the relevant spinal spheres (see the Table 3 giving the centers and radii). In particular, we get:

Proposition 5.4 *The intersections of the spheres listed below are nonempty only if k lies in the corresponding interval:*

Intersection	Interval	Intersection	Interval
$\mathcal{S}_1 \cap G_1^k \mathcal{S}_1$	$-2 \le k \le 2$	$\mathcal{S}_1 \cap G_1^k \mathcal{S}_2$	$-4 \le k \le 1$
$\mathcal{S}_1 \cap G_1^k \mathcal{S}_3$	$-3 \le k \le 1$	$\mathcal{S}_1 \cap G_1^k \mathcal{S}_4$	$-4 \le k \le 0$
$\mathcal{S}_2 \cap G_1^k \mathcal{S}_2$	$-2 \le k \le 2$	$\mathcal{S}_2 \cap G_1^k \mathcal{S}_3$	$-2 \le k \le 2$
$\mathcal{S}_2 \cap G_1^k \mathcal{S}_4$	$-2 \le k \le 2$	$\mathcal{S}_3 \cap G_1^k \mathcal{S}_3$	$-2 \le k \le 2$
$\mathcal{S}_3 \cap G_1^k \mathcal{S}_4$	$-2 \le k \le 1$	$\mathcal{S}_4 \cap G_1^k \mathcal{S}_4$	$-2 \le k \le 2$

This is not an optimal result, since it takes into account only the variable x and the fact that G_1 translates by one unit in the direction of the x axis. The optimal result is not far from this though; the point of Proposition 5.4 is to get down to a finite list of bounding bisectors intersecting a given one (so that we can use effective computational tools). We will give much more precise information in the next section.

5C Proof of the combinatorics

The techniques we use in order to justify the combinatorics are very similar to the ones explained in detail in [7; 8]. Note that one can think of justifying the combinatorics as a special case of finding the connected components of (many) semialgebraic sets. Indeed, F is clearly semialgebraic, defined by inequalities indexed by $I = \mathbb{N}$:

$$F = \{z \in \mathbb{C}^2 : f_i(z) < 0 \text{ for all } i \in I\}.$$

For convenience, we make the convention that $f_0(z) < 0$ is the defining inequality for the unit ball; in other words,

$$f_0(z) = \langle \tilde{z}, \tilde{z} \rangle,$$

where $\tilde{z} = (z, 1)$. In particular, we consider the boundary at infinity of complex hyperbolic space as a bounding face. All other inequalities have the form $f_j < 0$, where

$$f_j(z) = |\langle \tilde{z}, \tilde{p}_0 \rangle|^2 - |\langle \tilde{z}, \gamma_j \tilde{p}_0 \rangle|^2.$$

The*facets* are of F described by taking some subset $J \subset I$ and replacing the inequalities indexed by elements of J by the corresponding equality:

$$F_J = \{z \in \mathbb{C}^2 : f_j(z) = 0 \text{ for all } j \in J \text{ and } f_i(z) < 0 \text{ for all } i \in I \setminus J\}.$$

The fact that I is infinite will not be a problem because of the results in Section 5B, which imply that our polytope is locally finite.

More generally, we will consider sets of the form

$$F_{J,K} = \{z \in \mathbb{C}^2 : f_j(z) = 0 \text{ for all } j \in J \text{ and } f_i(z) < 0 \text{ for all } i \in K\},$$

where J and K are disjoint. In particular, F_J is the same as $F_{J,I \setminus J}$, and $F_{J,\varnothing}$ is the $|J|$–fold bisector intersection containing F_J.

5C1 Terminology and specification The facets of our polytopes that have dimension k will be called *k–faces*. Moreover, 3–faces will be simply called *sides*, 2–faces will be called *ridges*, 1–faces will be called *edges*, and 0–faces will be called *vertices*.

In terms of computations, it will be important to encode vertices. These can be of two kinds, namely, they can be of the form $F_{A,\varnothing}$ for some A with $|A| = 4$, or they can be singular points of $F_{B,\varnothing}$ with $|B| = 3$. In both cases, they can be obtained by solving a 0–dimensional system (this is the content of Assumption 5.5). For each of them, we encode the vertex by storing a rational univariate representation for the corresponding solution set, and an isolating interval specifying a root of the rational parameter; see Section 5C3.

Note that in the above description, the set A is not unique since a vertex may, in general, lie on more than four bisectors (see the discussion in Section 4B, where we saw examples of vertices lying on at least six bounding bisectors). Moreover, in general, one cannot take A to be just any 4–tuple of bisectors that contain that vertex, since some intersections may not be generic.

We will also need to encode 1–faces. There are two kinds of 1–faces, namely, those that lie in triple bisector intersections (we call these finite 1–faces), and those that lie in the intersection of the sphere at infinity $\partial_\infty H^2_{\mathbb{C}}$ with the closure in $\overline{H}^2_{\mathbb{C}}$ of a bisector intersection (we call these ideal 1–faces, or 1–faces at infinity). Computationally, we make no distinction between these two kinds of 1–faces, since both kinds are given in terms of spinal coordinates for a bisector intersection by an equation that is quadratic in both variables.

We use the term *arc* to mean a subset in $\overline{H}^2_{\mathbb{C}}$ of a triple bisector intersection (or a subset of the trace at infinity of a double bisector intersection) such that

- it is homeomorphic to a closed interval,
- it is parametrized by one of the spinal coordinates, and
- its endpoints are vertices of the polytope, but its interior contains no vertex of the polytope.

Note that a 1–face can always be described as a union of finitely many arcs (but one arc may not suffice: think of a polytope that has a whole Giraud disk as a facet, so that the boundary of that Giraud disk is a 1–face homeomorphic to a circle).

We now expand a little on how to parametrize (pieces of) 1–faces by a single coordinate (we discuss only parametrization by t_1, as the other one is entirely similar). Recall from Section 3A that the relevant defining functions $h(t_1,t_2)$ for triple bisector intersections (or trace at infinity of double bisector intersections) have degree at most two in each variable, so we can write them as

$$a_2(t_1)t_2^2+a_1(t_1)t_2+a_0(t_1),$$

with a_j at most quadratic. With respect to projection onto the first coordinate axis, the curve usually has two branches, given by

$$t_2=\frac{-a_1(t_1)\pm\sqrt{\Delta(t_1)}}{2a_2(t_1)},$$

where

$$\Delta(t_1)=a_1(t_1)^2-4a_2(t_1)a_0(t_1).$$

Specifically, this occurs above intervals of t_1 such that $a_2(t_1)$ does not vanish. Above such an interval, the "top branch" is obtained by taking $+\sqrt{\Delta}$ when $a_2(t_1)>0$, and $-\sqrt{\Delta}$ when $a_2(t_1)>0$. We call the other branch the "bottom branch".

If a_2 is identically zero, then the curve is either empty or consists of a single vertical line (so branches above the t_1 axes are undefined, and there is a single branch with respect to the projection onto the t_2 axis).

If a_2 is not identically zero, it vanishes at one or two points, and above each of these points, one can check whether the curve contains one, two or infinitely many points (one needs to determine whether a_1, a_0 also vanish at these points).

5C2 General procedure The pictures in Section 5A include the statement that each facet is topologically (in fact, piecewise smoothly) a disk with piecewise smooth boundary (with pieces of the boundary corresponding to facets of codimension one higher). This is not at all obvious; one of the difficulties is the fact that the sets F_J are not connected in general, in strong contrast with Dirichlet or Ford domains in the context of constant curvature geometries; see the discussion in [5].

For given J and K, there is an algorithm to decide whether $F_{J,K}$ is empty or not, and furthermore, one can list its connected components (and even produce triangulations). One possible approach to this is the cylindrical algebraic decomposition of semialgebraic sets; see [2], for instance.

The main issue when using such algorithms is that the number of semialgebraic sets to study is extremely large. If F has N faces, in principle, one has to deal with $\binom{N}{k}$ potential facets of codimension k, where $k = 1, 2, 3, 4$, which is a fairly large number of cylindrical decompositions. Rather, we will bypass the cylindrical decomposition and use as much geometric information as we can in order to restrict the number of verifications. Also, rather than using affine coordinates in $\mathbb{C}^2$, we use natural parametrizations for bisector intersections, deduced from spinal coordinates; see Section 3A.

Going back to geometry, the inequality defining complex hyperbolic space in $\mathbb{C}^2$ (which corresponds to f_0) is, of course, a bit different from the other inequalities. In particular, when using the notation $F_{J,K}$, we will always assume one of the index sets J or K contains 0.

If K contains 0, then by definition, $F_{J,K}$ is contained in $H^2_{\mathbb{C}}$; we will denote by $\widehat{F}_{J,K}$ its extension to projective space, namely,

$$\widehat{F}_{J,K} = F_{J,K\setminus\{0\}}.$$

We will also refer to the following set as the trace at infinity of $F_{J,K}$:

$$\partial_\infty F_{J,K} = F_{J\cup\{0\},K\setminus\{0\}}.$$

By $\overline{F}_{J,K}$, we mean the set obtained from the definition of $F_{J,K}$ by replacing $<$ by $\leq$:

$$\overline{F}_{J,K} = \{z \in \mathbb{C}^2 : f_j(z) = 0 \text{ for all } j \in J \text{ and } f_i(z) \leq 0 \text{ for all } i \in K\},$$

which is also

$$\overline{F}_{J,K} = \bigcup_{L\subset K} F_{J\cup L,K\setminus L}.$$

Note that, in general, this is not the closure of $F_{J,K}$ in $\mathbb{C}^2$.

We focus on an algorithm for determining the combinatorics of ridges, or in other words, facets of the form F_J with $|J| = 2$. In most cases, we will also assume $0 \notin J$; ie we study finite facets rather than faces in $\partial_\infty H^2_{\mathbb{C}}$. The algorithm will produce a description of the facets in ∂F_J, so we get a list of the 1– and 0–faces along the way. The 3–faces are easily deduced from the 2–faces.

The basis for our analysis is the following, which follows from the theory of Gröbner bases (see [4], for instance, and also Section 5C3 of the present paper). Let ℓ be a number field.

- There is an algorithm to determine whether a system of n polynomial equations defined over ℓ in n unknowns is 0–dimensional (ie whether there are only finitely many solutions in $\mathbb{C}^n$).

- If the system is indeed 0–dimensional, there is an algorithm to determine the list of solutions; their entries lie in a finite extension $k \supset \ell$. One can also determine the list of rational/real solutions.
- Polynomials with coefficients in ℓ can be evaluated at the solutions of a point with coordinates in k, and one can determine whether the value is positive (or negative or zero).

When such systems have solution sets with unexpectedly high dimension, there is usually a geometric explanation (typically some of the intersecting bisectors share a slice; see [8], for instance). We will not address this issue since it never occurs in the situation of the present paper.

In all situations we will consider here, the field ℓ will be a quadratic number field, and the extension k will have degree at most four over ℓ. This makes all computations very quick (using capabilities of recent computers, and standard implementations of Gröbner bases).

For the rest of the discussion, we make the following assumptions.

Assumption 5.5 (1) For every $L \subset I$ with $|L| = 4$, the dimension of F_L is zero.

(2) For every $J \subset I$ with $|J| = 2$, and every $x \in I$ with $x \notin J$, the restriction g_x of f_x to $F_{J,\varnothing}$ has nondegenerate critical points.

These assumptions are by no means necessary in order to determine the combinatorial structure of $F_{J,K}$, but they will simplify the discussion in several places. Note also that they can be checked efficiently using a computer; in particular, we state

Proposition 5.6 *Let M be the figure eight knot complement. Then the Ford domain of the irreducible boundary unipotent representation $\rho\colon \pi_1(M) \to \mathrm{PU}(2,1)$, centered at the fixed point of the holonomy of any peripheral subgroup, satisfies Assumption 5.5.*

In contrast, the domains that appear in [8] do not satisfy these hypotheses.

The combinatorial description of F_J (ie its connected components and the list of facets adjacent to it) can be obtained by starting from a description of $F_{J,\varnothing}$ and repeatedly studying $F_{J,K\cup\{x\}}$ from $F_{J,K}$, where $x \in I$ is not in $J \cup K$. The latter inductive step is done as follows.

The boundary $\partial F_{J,K}$ can be described as a union of arcs contained in $F_{J\cup\{k\},K\setminus\{k\}}$ for some $k \in K$. For computational purposes, we will always assume that an arc is homeomorphic to a closed interval, that its endpoints are vertices, but none of its interior points are vertices.

Note also that the arcs may not be equal to $F_{J\cup\{k\},K\setminus\{k\}}$, since $F_{J\cup\{x\},\varnothing}$ may have a double point.

For each arc a in $\partial F_{J,K}$ as above, we study the set

$$F_{J\cup\{k,x\}},$$

which, by Assumption 5.5(1), is obtained by solving a 0–dimensional system. Keeping only solutions that lie in a, we get a subdivision of a into connected components of $a\setminus F_{J\cup\{k,x\}}$, and for each such component, we check whether or not it is in $F_{J\cup\{k\},\{x\}}$. If so, it is a component of the boundary of $F_{J,K\cup\{x\}}$.

We then compute the critical points of the restriction to F_J of f_x (this can be done because of Assumption 5.5(2)) and determine whether any such critical point is inside $F_{J,K}$.

Suppose c is in a component $C_{J,K}$ of $F_{J,K}$.

- If $g_x(c)=0$ and c is a saddle point for the restriction g_x of f_x, then a neighborhood of c in $\overline{F}_{J,K\cup\{x\}}$ is the union of two sectors meeting in their apex. $F_{J,K\cup\{x\}}$ will have four boundary arcs in a neighborhood of c. Each such arc will either connect c to another saddle point of g_x, or it will connect it to a vertex in the boundary of $C_{J,K}$. For each such arc, we take a sample point to check whether it is contained in $F_{J\cup\{x\},K}$.

- If $g_x(c)\neq 0$, there could be an isolated component of $F_{J\cup\{x\},K}$ that winds around c. In order to determine whether this happens or not, we consider the slice $t_1=\alpha_1$, and intersect it with $g_x=0$. Recall that this intersection contains either 0, 1 or 2 points (because it is obtained by solving an equation that has degree at most two, which is not identically zero because $g_x(c)\neq 0$). Then there is an isolated component if and only if the intersection consists of precisely two points, and the two intersection points lie in the same connected component of $F_{J,K}$.

Now collecting the boundary arcs with the inside arcs (joining two points that are either saddle or boundary vertices in $F_{J\cap\{k,x\}}$), we get a stratum decomposition for $F_{J,K\cup\{x\}}$.

Moreover, if we make the following assumption, then all components of $F_{J,K\cup\{x\}}$ are topological disks, since their boundary consists of a single component.

Assumption 5.7 (3) The curves $F_{J\cup\{x\},K}$ have no isolated components in $F_{J,K}$.

Once again, in the special case of the Ford domain relevant to the irreducible boundary unipotent rank one, it turns out this hypothesis is satisfied.

5C3 Rational univariate representation We briefly recall what we need about rational univariate representations; for details on this technique, see [24]. Recall that given a 0–dimensional polynomial system

$$\begin{cases} f(t_1,t_2)=0, \\ g(t_1,t_2)=0, \end{cases} \tag{23}$$

with coefficients in the number field ℓ, we can write it as a polynomial system with rational coefficients by using a primitive element for ℓ; the corresponding system has one more variable (which we denote by s), and one more equation (which is the minimal polynomial of a primitive generator for ℓ). We write it in the form

$$\begin{cases} \tilde{f}(t_1,t_2,s)=0, \\ \tilde{g}(t_1,t_2,s)=0, \\ \qquad m(s)=0, \end{cases} \tag{24}$$

where $\tilde{f}$ is obtained from f by expressing its coefficients as polynomials in the primitive element for ℓ. In the cases that interest us, ℓ will be a totally real number field, which we assume from now on.

In this discussion, we consider systems of two equations in two variables (so we get three equations in three variables, counting the extra variable corresponding to the primitive element of the number field), but we could also allow systems that have more equations than the number of variables (the important point is that the ideal generated by the equations should be 0–dimensional).

Now the key point is that there exists a 1–variable polynomial r such that the solutions are parametrized as rational functions of the roots of r. More specifically, there exist polynomials r, p_0, p_1, p_2 and q with integer coefficients such that the solutions of the system can be written in the form

$$s=p_0(u)/q(u), \quad t_1=p_1(u)/q(u), \quad t_2=p_2(u)/q(u), \tag{25}$$

and the latter formula gives a solution of (24) if and only if u is a root of r. Of course, since the minimal polynomial m has several roots in general, this produces more solutions of system (23) than we would like. The solutions of (23) can easily be obtained by sifting the solutions of (24) once we know isolating intervals for the roots of m.

Note that even though all the equations relevant to this paper have coefficients in a fixed number field (namely $\ell=\mathbb{Q}(\sqrt{7})$), the vertices usually have entries in a larger number field (namely the field generated by a given root of the rational parametrizing polynomial r).

Note also that the solutions lie in a subfield $L \subset \mathbb{C}$ if and only if the corresponding root u of r lies in L. In particular, if we want to find *real* solutions of the system, we can restrict to studying *real* roots of r, which can be specified by isolating intervals.

Using a rational univariate representation for the vertices provides a convenient set of methods that allow us to

(i) find the list of faces that contain a given vertex;

(ii) for each bounding bisector not containing a vertex, check which side the vertex is in;

(iii) check if two vertices are the same;

(iv) check whether a given vertex is inside a given arc;

(v) if two vertices in $F_{J \cup \{x\}, \varnothing}$ are given, check whether these two vertices are joined by an arc in $F_{J \cup \{x\}, \varnothing}$.

Items (i) and (ii) are very simple because all our equations are defined over a given ℓ. Given a polynomial $h(t_1, t_2) = \widetilde{h}(t_1, t_2, s)$, we start by substituting the parametrization (25) in $\widetilde{h}$, replacing u by the appropriate interval of values of the rational parameter. If the corresponding interval does not contain 0, we know the sign of h at that vertex.

Otherwise, we keep the exact parametrization (25) and get a rational function in u that represents h at the solutions of (24), and we check whether it vanishes at the appropriate root of r. This corresponds to checking whether our favorite root of the rational parametrizing polynomial r is also a root of another given polynomial with integer coefficients (namely the numerator of the above rational function); this can be done by computing their greatest common divisor, and isolating its real roots.

If the rational function does not vanish, we compute a more precise interval for the value of $\widetilde{h}$, and refine precision until the interval does not contain 0. Of course, in all generality, this may require such high precision that it would exhaust the system memory, but this does not seem to happen for the verifications that appear in this paper, at least for our implementation on standard modern computers.

We now sketch how to implement item (iii). Suppose we are given two rational parametrizations

$$\begin{aligned} s &= p_0(u)/q(u), \quad & t_1 &= p_1(u)/q(u), \quad & t_2 &= p_2(u)/q(u), \\ s &= a_0(v)/b(v), \quad & t_1 &= a_1(v)/b(v), \quad & t_2 &= a_2(v)/b(v), \end{aligned}$$

where u (resp. v) is to be taken to be a specific root of $r(u)$ (resp. $c(v)$). Equality corresponds to verifying whether $p_1(u)b(v) - q(u)a_1(v)$ (resp. $p_2(u)b(v) - q(u)a_2(v)$)

vanishes at the corresponding roots. If the rational parameters were the same, this would simply amount to computing a greatest common divisor, but in general, the parameters from both rational representations are different.

One way to handle this is to solve the system

$$\begin{cases} p_1(u)b(v) - q(u)a_1(v) = 0, \\ r(u) = 0, \\ c(v) = 0, \end{cases}$$

which can be done using a rational univariate representation once again. The result then follows from sifting solutions and keeping only those that give the right root for u and v, and checking whether the sift gives a solution of not.

In order to explain how to check (iv), we need to describe in more detail how we encode arcs. We will assume

- that every arc is parametrized by one of the spinal coordinates (this can always be achieved, perhaps after subdividing certain arcs if necessary),
- that the endpoints of every arc are vertices (parametrized by a rational univariate representation, as discussed above), and
- that there are no vertices strictly inside any arc.

Then, in order to check whether a given vertex is inside an arc parametrized by t_1, we need to compare its t_1 value with the t_1 values of the endpoints of the arc. This amounts to checking the sign of an expression of the form

$$p_1(u)/q(u) - a_1(v)/b(v),$$

where u (resp. v) is a specific root of r (resp. c). This is the same as the test that occurs in item (iii).

If the vertex t_1 value is between the t_1–values of the endpoints of the arc, we still need to check whether it is in the correct arc.

5C4 Sample computations We explicitly determine some sets F_J with $|J| = 2$, in order to illustrate the phenomena that can occur when applying the algorithm from the previous section. The general scheme to parametrize $F_{J,\varnothing}$ is explained in [7], for instance.

When $0 \notin J = \{j,k\}$, we distinguish two basic cases, depending on whether p_0, p_j and p_k are in a common complex line. This happens if and only if some/any lifts $\widetilde{p}_j \in \mathbb{C}^3$ are linearly dependent. In that case, the bisectors $F_{\{j\}}$ and $F_{\{k\}}$ have the same complex spine, and their intersection is either empty or a complex line (this never happens in the Ford domains studied in this paper).

Otherwise, $F_{J,\varnothing}$ can be parametrized by vectors of the form

$$(\overline{z}_1 p_0 - p_j) \boxtimes (\overline{z}_2 p_0 - p_k) = z_1 p_{k0} + z_2 p_{0j} + p_{jk},$$

with $|z_1| = |z_2| = 1$, and where p_{mn} denotes $p_m \boxtimes p_n$; see Section 3A.

Valid pairs (z_1, z_2) in the Clifford torus $|z_1| = |z_2| = 1$ are given by pairs with

$$\langle z_1 p_{k0} + z_2 p_{0j} + p_{jk},\ z_1 p_{k0} + z_2 p_{0j} + p_{jk}\rangle < 0,$$

which can be rewritten as

$$\mathfrak{Re}(\mu_0(z_1) z_2) = \nu_0(z_1),$$

for μ_0 and ν_0 affine in $z_1, \overline{z}_1$.

In terms of the notations of Section 5C2, the restriction g_0 of f_0 to $F_{J,\varnothing}$ is given by

$$g_0(z_1, z_2) = \mathfrak{Re}(\mu_0(z_1) z_2) - \nu_0(z_1)).$$

In order to draw pictures, we will sometimes use log-coordinates (t_1, t_2) for $F_{J,\varnothing}$, and we write, for $j = 1, 2$,

$$z_j = \exp(2\pi i t_j).$$

Given $l \notin J$, we already mentioned in Section 3A how to write the restriction g_l of f_l to F_J. Note that $\langle p_{k0}, p_0\rangle = \langle p_{0j}, p_0\rangle = 0$, so the equation $f_x = 0$ reads

$$|\langle p_{jk}, p_0\rangle| = |\langle z_1 p_{k0} + z_2 p_{0j} + p_{jk}, p_l\rangle|,$$

which again can be written in the form

$$\mathfrak{Re}(\mu_l(z_1) z_2) = \nu_l(z_1).$$

In order to compute the critical points of the restriction to $|z_1| = |z_2| = 1$ of a function $h(z_1, \overline{z}_1, z_2, \overline{z}_2)$, we search for points where

$$\frac{\partial h}{\partial z_1} z_1 - \frac{\partial h}{\partial \overline{z}_1} \overline{z}_1 = 0 \quad \text{and} \quad \frac{\partial h}{\partial z_2} z_2 - \frac{\partial h}{\partial \overline{z}_2} \overline{z}_2 = 0.$$

Gröbner bases for the corresponding systems tell us whether these critical points are nondegenerate (see Assumption 5.7), and if so, we can compute them fairly explicitly, ie describe their coordinates as roots of explicit polynomials (in particular, they can be computed to arbitrary precision).

Proposition 5.8 *Let $J = \{1, 2\}$. Then F_J is empty, and $\overline{F}_J$ is a singleton, given by $F_{\{1,2,3,5,10,11\}}$.*

The singleton in the proposition is $\{p_2\}$, for p_2 as in Lemma 4.2. It follows from the proposition that p_2 lies precisely on six bounding bisectors (Lemma 4.2 only showed that it was on at least six, listed in Tables 1 and 2).

Proof For $J = \{1,2\}$, we get

$$\mu_0(z_1) = -2 - \overline{z}_1, \quad \nu_0(z_1) = -3 + z_1 + \overline{z}_1.$$

The discriminant

$$|\mu|^2 - \nu^2 = -6 + 16\,\mathfrak{Re}\, z_1 - 2\,\mathfrak{Re}\, z_1^2$$

vanishes for precisely four complex values of z_1, which are the roots of

$$z_1^4 - 8z_1^3 + 6z_1^2 - 8z_1 + 1. \tag{26}$$

Since we know $F_{J,\{0\}}$ is connected [14, Theorem 9.2.6], we know that at most two of these roots lie on the unit circle. In fact, $z_1 = z_2 = 1$ gives a point in $F_{J,\{0\}}$, so $F_{J,\{0\}}$ is nonempty; hence there must be two (complex conjugate) roots on the unit circle. Indeed, these roots have argument $2\pi t$ with $t = \pm 0.20682703\ldots$.

A more satisfactory way to check that the polynomial (26) has precisely two roots on the unit circle is to split $z_1 = x_1 + iy_1$ into its real and imaginary parts (this gives a general method that does not rely on geometric arguments).

Indeed, z_1 is a root of (26) if and only if (x_1, y_1) is a solution of the system $-6 + 16x_1 - 2x_1^2 + 2y_1^2 = 0$, $x_1^2 + y_1^2 = 1$. These equations imply that $x_1 = 2 \pm \sqrt{3}$, and then

$$y_1^2 = 2 - 4x_1,$$

which is positive only for $x_1 = 2 - \sqrt{3}$, and then we get $y_1 = \pm\sqrt{4\sqrt{3} - 6}$.

In order to run the algorithm from the preceding section, we write the restriction g_3 of f_3 to $F_{J,\varnothing}$, which is given by

$$-3 + 2\,\mathfrak{Re}\Big(\frac{1 - i\sqrt{7}}{2} z_1 + \frac{5 - i\sqrt{7}}{2} z_2 + \frac{-3 + i\sqrt{7}}{2} z_1 \overline{z}_2\Big).$$

Gröbner basis calculations show the system $g_0(z) = g_3(z) = |z_1|^2 - 1 = |z_2|^2 - 1 = 0$ has precisely two solutions, given in log-coordinates by

$$(-0.20418699\ldots, -0.03294828\ldots), \quad (0.15576880\ldots, -0.07655953\ldots).$$

Once again, the most convenient way to use Gröbner bases is to work with four variables x_1, y_1, x_2, y_2 given by real and imaginary parts of z_1 and z_2 (with extra equations $x_j^2 + y_j^2 = 1$).

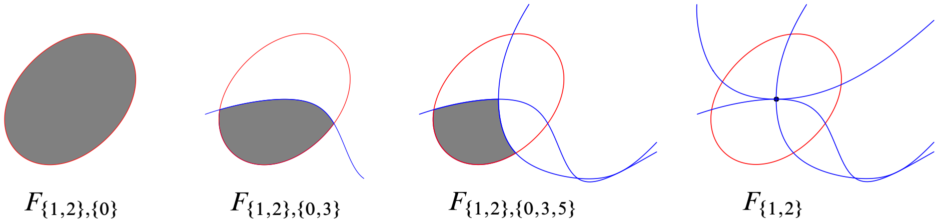


Figure 5: Steps of the algorithm to determine $F_{\{1,2\}}$

The combinatorics of $F_{J,K}$ for $K=\{0,3\}$ are illustrated in Figure 5 (middle left). It is a disk with two boundary arcs, given by $F_{\{1,2,0\},\{3\}}$ and $F_{\{1,2,3\},\{0\}}$.

As the next element to include in K, we choose 5 rather than 4, in order to shorten the discussion slightly. The curve $F_{\{1,2,5\},\varnothing}$ intersects $F_{\{1,2,0\},\varnothing}$ two points, given in log-coordinates by

$$(0.04600543\dots, 0.20593006\dots), \quad (0.05483483\dots, -0.17019919\dots).$$

Only the second one is inside the arc $F_{\{1,2,0\},\{3\}}$.

The curve $F_{\{1,2,5\},\varnothing}$ intersects $F_{\{1,2,3\},\varnothing}$ in five points (z_1,z_2), given by

$$(1,1),\quad (i,-i),\quad (-i,i),\quad \left(\frac{9+5i\sqrt{7}}{16},\frac{-3+i\sqrt{7}}{4}\right),\quad \left(\frac{-3+i\sqrt{7}}{4},\frac{1-3i\sqrt{7}}{8}\right),$$

only one of which is in $F_{\{1,2,3\},\{0\}}$, namely $(1,1)$.

Now $F_{\{1,2\},\{0,3,5\}}$ has three boundary arcs, given by $F_{\{1,2,0\},\{3,5\}}$, $F_{\{1,2,3\},\{0,5\}}$ and $F_{\{1,2,5\},\{0,3\}}$; see Figure 5 (middle right).

Next, we include 10 in K. The curve $F_{\{1,2,10\},\varnothing}$ intersects $F_{\{1,2,0\},\varnothing}$ in two points, none of which is in $F_{\{1,2,0\},\{3,5\}}$. Hence the arc $F_{\{1,2,0\},\{3,5\}}$ is either completely inside or completely outside $F_{\{1,2,0\},\{3,5,10\}}$. One easily checks that it is outside, by taking a sample point.

The curve $F_{\{1,2,10\},\varnothing}$ intersects $F_{\{1,2,3\},\varnothing}$ in five points, and none of these is in $F_{\{1,2,3\},\{0,5\}}$. The arc $F_{\{1,2,3\},\{0,5\}}$ is either completely inside or completely outside $F_{\{1,2,3\},\{0,5,10\}}$, and a sample point shows it is outside.

Similarly, the curve $F_{\{1,2,10\},\varnothing}$ intersects $F_{\{1,2,5\},\varnothing}$ in six points, none of which is in $F_{\{1,2,5\},\{0,3\}}$, and the arc $F_{\{1,2,5\},\{0,3\}}$ is completely outside $F_{\{1,2,3\},\{0,5,10\}}$.

This implies that $F_{\{1,2\}}$ is empty; see Figure 5 (far right).

Finally, consider the intersection of $F_{\{1,2,10\},\varnothing}$ with the three vertices of $F_{\{1,2\},\{0,3,5\}}$. One easily checks that the only intersection is the point with complex spinal coordinates given by $(1,1)$, and this point is indeed a vertex of F. It is in homogeneous coordinates in $\mathbb{C}^3$ given by

$$\Big(\frac{3-i\sqrt{7}}{2},-2,-\frac{3-i\sqrt{7}}{2}\Big),$$

and it is on precisely six bounding bisectors (by construction it is on $\mathcal{B}_1$ and $\mathcal{B}_2$, and it is also in $\mathcal{B}_3$, $\mathcal{B}_5$, $\mathcal{B}_{10}$ and $\mathcal{B}_{11}$). In terms of the notation of Section 5C2, this point is

$$F_{\{1,2,3,5,10,11\}}.$$

In fact one easily checks that this point is the fixed point of G_2 (which, by definition of the bounding bisectors, is obviously in $\mathcal{B}_1\cap\mathcal{B}_2$). □

Remark 5.9 (1) Throughout the proof of Proposition 5.8, we have ignored the issue of critical points. In principle, at each stage, we may have missed some isolated components of the curves $F_{\{1,2,k\},\varnothing}$; if this were the case, the set $F_{\{1,2\}}$ would still be contained in the set which we just described. Hence it must be empty anyway.

(2) The curves $F_{\{1,2,10\},\varnothing}$ and $F_{\{1,2,3\},\varnothing}$ are, in fact, tangent at $(1,1)$, which is a vertex of F. We shall come back to this point later, when discussing stability of the combinatorics of F under deformations.

Proposition 5.10 *$F_{\{1,3\}}$ is combinatorially a triangle, with three boundary arcs given by $F_{\{1,3,0\}}$, $F_{\{1,3,5\}}$ and $F_{\{1,3,11\}}$, and three vertices given by $F_{\{0,1,3,5\}}$, $F_{\{0,1,3,11\}}$ and $F_{\{1,2,3,5,10,11\}}$.*

Note that this triangle appears in Figure 3 (left) and 4 (left), it is the intersection of the bounding bisectors $\mathcal{B}_1$ and $\mathcal{B}_3$ corresponding to G_2 and G_3, respectively. The edges in $H^2_{\mathbb{C}}$ are on $\mathcal{B}_5$, which corresponds to $G_1G_2G_1^{-1}$, and $\mathcal{B}_{11}$, which corresponds to $G_1^{-1}G_3G_1$.

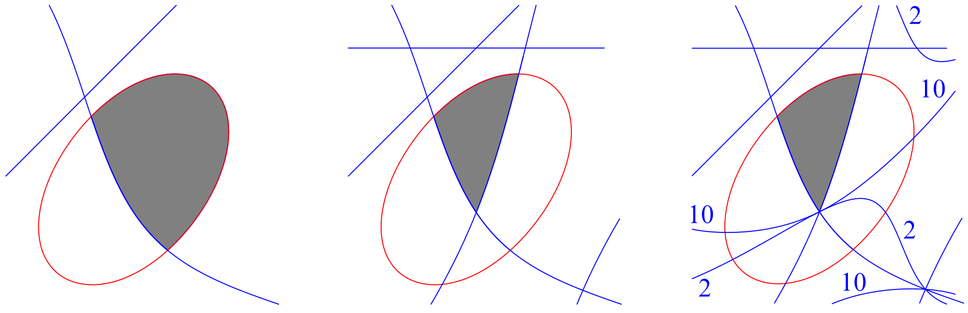
Proof As in the argument for $F_{\{1,2\}}$, we study $F_{J,K}$ for increasing sets K, freely choosing the order we use to increase K. We describe an efficient way to get down to $F_{\{1,3\}}$ in the form of a picture; see Figure 6.

We start by studying $F_{\{1,3\},\{5\}}$. Note that the curve $F_{\{1,3,5\},\varnothing}$ has two double points. These points can be obtained by writing the equation $g_5=0$ as

$$\mathfrak{Re}(\mu(z_1)z_2)=\nu(z_1),$$

where

$$\mu(z_1)=\frac{3+i\sqrt{7}}{2}-\bar{z}_1,\quad \nu(z_1)=1-\mathfrak{Re}\Big(\frac{3+i\sqrt{7}}{2}z_1\Big).$$



$F_{\{1,3\},\{0,5\}}$ $F_{\{1,3\},\{0,5,11\}}$ $F_{\{1,3\},\{0,5,11,2,10\}}$

Figure 6: Steps of the algorithm to determine $F_{\{1,3\}}$

The discriminant $|\mu(z_1)|^2 - \nu(z_1)^2$ is given by

$$2 + \mathfrak{Re}\Big(\frac{-1-3i\sqrt{7}}{4} z_1^2\Big),$$

which vanishes for $z_1 = \pm\frac{1}{4}(3 - i\sqrt{7})$. Plugging this back into the equation $g_5 = 0$ gives $z_2 = \mp\frac{1}{4}(3 - i\sqrt{7})$. One easily checks that $g_0(z_1, z_2) > 0$ for these two double points, ie they lie outside complex hyperbolic space.

One checks that $F_{\{1,3,5\},\varnothing}$ intersects $F_{\{1,3,0\},\varnothing}$ in precisely two points (and these intersections are transverse), so we get two arcs in the boundary of $F_{\{1,3\},\{0,5\}}$, namely $F_{\{1,3,5\},\{0\}}$ and $F_{\{1,3,0\},\{5\}}$; see Figure 6 (left).

In principle, there could be an extra arc in $F_{\{1,3,5\},\{0\}}$, not intersecting $F_{\{1,3,0\},\varnothing}$, so we compute critical points of g_5. They are given by the solutions of the system

$$\begin{cases} \mathfrak{Im}\Big(\Big(\overline{z}_2 + \dfrac{3+i\sqrt{7}}{2}\Big)z_1\Big) = 0, \\ \mathfrak{Im}\Big(\Big(\overline{z}_1 + \dfrac{3+i\sqrt{7}}{2}\Big)z_2\Big) = 0, \end{cases}$$

that satisfy $|z_1| = |z_2| = 1$.

There are four such critical points, and they have the form $(\pm\alpha, \pm\alpha)$, where $\alpha = \frac{1}{4}(3 - i\sqrt{7})$ (of course this list includes the double points computed before). The corresponding points are outside F; in fact, $g_0(\pm\alpha, \pm\alpha) > 0$.

A similar analysis justifies the middle part of Figure 6, ie that $\overline{F}_{\{1,3\},\{0,5,11\}}$ is combinatorially a triangle (with one side on $\partial_\infty H^2_{\mathbb{C}}$).

We sketch how to justify that $F_{\{1,3\}} = F_{\{1,3\},\{0,5,11\}}$. For $k = 2$ and $k = 10$, the curve $F_{\{1,3,k\},\varnothing}$ actually goes through a vertex of $F_{\{1,3\}} = F_{\{1,3\},\{0,5,11\}}$; if $k \neq 0, 2, 5, 10, 11$, then $F_{\{1,3,k\},\varnothing}$ does not intersect even $\overline{F}_{\{1,3\},\{0,5,11\}}$.

We start by studying $F_{\{1,3,0\},\varnothing} \cap F_{\{1,3,2\},\varnothing}$. In order to use standard root isolation methods, we use real equations in x_1, y_1, x_2, y_2. Computing a Gröbner basis for the ideal generated by g_0, g_3, $x_1^2+y_1^2-1$ and $x_2^2+y_2^2-1$, we see that it contains

$$39-840\sqrt{7}y_2+4088y_2^2+608y_2^3\sqrt{7}-9152y_2^4+1024y_2^5\sqrt{7}+7168y_2^6,$$

which has precisely two real roots, given approximately by $y_2^{(1)}=0.01815877\ldots$ and $y_2^{(2)}=0.65602473\ldots$.

The Gröbner basis also gives an expression for x_1, y_1 and x_2 in terms of y_2, namely

$$\begin{aligned}x_1&=\tfrac{1}{14725}\big(-4943+16836\sqrt{7}y_2-142640y_2^2\\&\qquad\qquad+53184y_2^3\sqrt{7}+72128y_2^4-75264y_2^5\sqrt{7}\big),\\ y_1&=\tfrac{1}{14725}\big(5058\sqrt{7}+45888y_2-112560y_2^2\sqrt{7}\\&\qquad\qquad+309472y_2^3+74432y_2^4\sqrt{7}-422912y_2^5\big),\\ x_2&=\tfrac{1}{19}\big(20-21\sqrt{7}y_2+16y_2^2+32y_2^3\sqrt{7}\big).\end{aligned}$$

Substituting either value $y_2^{(j)}$ gives two points $a^{(j)}=(x_1^{(j)},y_1^{(j)},x_2^{(j)},y_2^{(j)})$ for $j=1,2$, and we claim that $g_5(a^{(1)})>0$ and $g_{11}(a^{(2)})>0$. Clearly this can be checked by simple interval arithmetic, in fact

$$g_5(a^{(1)})=3.80716606\ldots,\quad g_{11}(a^{(2)})=3.94518313\ldots.$$

The analysis of $F_{\{1,3,5\},\varnothing}\cap F_{\{1,3,2\},\varnothing}$ is in a sense simpler, because all the solutions to the corresponding system are defined over $\mathbb{Q}(i,\sqrt{7})$. The system has precisely five solutions, given by

$$\Big(i,\frac{1+\sqrt{7}}{4}+i\frac{1-\sqrt{7}}{4}\Big),\quad\Big(-i,\frac{1-\sqrt{7}}{4}-i\frac{1+\sqrt{7}}{4}\Big),$$
$$\Big(\frac{-3+i\sqrt{7}}{4},\frac{3-i\sqrt{7}}{4}\Big),\quad\Big(\frac{9+5i\sqrt{7}}{16},-\frac{9+5i\sqrt{7}}{16}\Big),\quad\Big(1,\frac{3+i\sqrt{7}}{4}\Big).$$

Only one of these solutions satisfies $g_0\le 0$, namely the last one (in other words, only one intersection point lies $\overline{H}_{\mathbb{C}}^2$).

Note that we already found one point in $F_{\{1,3,2\},\varnothing}\cap F_{\{1,3,5\},\varnothing}$, namely the fixed point of G_2; see the proof of Proposition 5.8.

Similarly, one verifies that $F_{\{1,3,2\},\varnothing}\cap F_{\{1,3,11\},\varnothing}$ contains precisely six points, only one of which gives a point in (the closure of) complex hyperbolic space.

Once again, since we already know one point in this intersection (namely the fixed point of G_2), we get that the intersection of $F_{\{1,3,2\},\varnothing}$ with $\partial F_{\{0,1,3,5,11\},\varnothing}$ consists of precisely one point. This implies that $\partial F_{\{0,1,3,5,11\},\varnothing}$ is either completely inside or completely outside $\partial F_{\{0,1,3,5,11,2\},\varnothing}$. It is easy to check that it is inside by testing a sample point (for instance one of the other vertices of the triangle $\partial F_{\{0,1,3,5,11\},\varnothing}$).

We now show that $F_{\{1,3,2\},\varnothing}$ does not intersect $F_{\{0,1,3,5,11\},\varnothing}$ by computing the critical points of g_2. There are six critical points, given by

$$\Bigl(-1,-\frac{1+3i\sqrt{7}}{8}\Bigr),\quad \Bigl(\frac{3-i\sqrt{7}}{4},\frac{3-i\sqrt{7}}{4}\Bigr),\quad \Bigl(\pm\frac{1+i\sqrt{7}}{\sqrt{8}},\pm\frac{1-i\sqrt{7}}{\sqrt{8}}\Bigr),$$

and one easily checks that none of them is inside $F_{\{0,1,3,5,11\},\varnothing}$. In particular, we get that the minimum value of g_3 on $\overline{F}_{\{0,1,3,5,11\},\varnothing}$ is 0, and it is realized precisely at one vertex (namely the fixed point of G_2).

In other words, we get $F_{\{0,1,3,5,11\},\varnothing} = F_{\{0,1,2,3,5,11\},\varnothing}$; ie including the inequality $g_2 < 0$ at this stage has no effect. An entirely similar computation shows that $F_{\{0,1,2,3,5,11\},\varnothing} = F_{\{0,1,2,3,5,10,11\},\varnothing}$.

For all $k \neq 0,1,2,3,5,10,11$, we have that $F_{\{0,1,3,k\},\varnothing}$ does not intersect even the closure $\overline{F}_{\{0,1,2,3,5,11\},\varnothing}$; one can use arguments as above using interval arithmetic. □

Similar arguments allow us to handle the detailed study of all the polygons that appear on Figures 3 and 4.

Proposition 5.11 *$F_{\{1,4\},\varnothing}$ is a Giraud disk, which is entirely contained in the exterior of $\mathcal{B}_5$. In particular, $F_{\{1,4\}}$ is empty.*

Proof We will prove that $F_{\{5\},\varnothing}$ does not intersect the Giraud torus $\widehat{F}_{\{1,4\},\varnothing}$. In order to see this, we use complex spinal coordinates and write $g_5(z_1,z_2)$ for the restriction of f_5 to the Clifford torus $|z_1| = |z_2| = 1$.

One computes explicitly that

$$g_5(z_1,z_2) = 4 + 2\,\mathfrak{Re}\Bigl(\frac{1+i\sqrt{7}}{2}z_1\overline{z}_2\Bigr).$$

This is clearly always positive when $|z_1| = |z_2| = 1$.

In other words, the Giraud torus $\widehat{F}_{\{1,4\},\varnothing}$ is entirely outside F. □

Proposition 5.12 *$F_{\{1,6\},\varnothing}$ is empty. The Giraud torus $\widehat{F}_{\{1,6\},\varnothing}$ is completely outside complex hyperbolic space; in other words, the bisectors $\mathcal{B}_1$ and $\mathcal{B}_6$ are disjoint.*

Proof We write the equation of $F_{\{0,1,6\},\varnothing}$ in spinal coordinates for the Giraud torus $F_{\{1,6\},\varnothing}$, which reads

$$g_0(z_1,z_2) = 18 - 2\,\mathfrak{Re}(4(z_1+z_2) + z_1\overline{z}_2).$$

Clearly this is nonnegative when $|z_1| = |z_2| = 1$, and in that case, it is zero if and only if $z_1 = z_2 = 1$.

In other words, $\widehat{\mathcal{B}}_1$ and $\widehat{\mathcal{B}}_2$ intersect in a point in $\overline{H}^2_{\mathbb{C}}$. Note that this point is not in the closure of F; in fact, it is strictly outside the half spaces bounded by $\mathcal{B}_2$, $\mathcal{B}_3$, $\mathcal{B}_5$, $\mathcal{B}_7$ and $\mathcal{B}_{11}$. □

Proposition 5.13 *$F_{\{3,8\}}$ is empty. The Giraud torus $F_{\{3,8\},\varnothing}$ contains a disk in $H^2_{\mathbb{C}}$, but $\overline{F}_{\{3,8\},\{2,6\}}$ is empty.*

Proof The proof is actually very similar to that of Proposition 5.10, but since the corresponding set is empty, we go through some of the details.

The curve $F_{\{3,8,2\},\varnothing}$ intersects $F_{\{3,8,0\},\varnothing}$ in precisely two points, and cuts out a disk in the Giraud disk $F_{\{3,8\},\varnothing}$, so that $F_{\{3,8\},\{0,2\}}$ is a disk with only two boundary arcs.

One then easily verifies that $F_{\{3,8,6\},\varnothing}$ does not intersect $\overline{F}_{\{3,8\},\{0,2\}}$, so $F_{\{3,8\},\{0,2,6\}}$ is either equal to $F_{\{3,8\},\{0,2\}}$ or is empty (one needs to check critical points in order to verify this).

By taking a sample point z and checking $f_6(z) > 0$, one gets that $F_{\{3,8\},\{0,2,6\}}$ is empty. □

The study of $\mathcal{B}_1 \cap \mathcal{B}_k$ for various values of k is similar to one of the previous few propositions; we list the relevant arguments in Table 4. When the proof is similar to Proposition 5.11, the indices l listed in brackets indicate that $\mathcal{B}_1 \cap \mathcal{B}_k$ is entirely outside the half space bounded by $\mathcal{B}_l$.

The corresponding list of arguments used to study of $\mathcal{B}_3 \cap \mathcal{B}_k$ for various values of k in Table 5.

Note that the arguments for $\mathcal{B}_2$ (resp. $\mathcal{B}_4$) are, of course, almost the same as those for $\mathcal{B}_1$ (resp. $\mathcal{B}_3$), since the corresponding faces are actually paired by G_2 (resp. G_3).

5C5 Genericity In order to study deformations ρ_t of the boundary unipotent representation $\rho_0\colon \pi_1(M) \to \mathrm{PU}(2,1)$, we will need more information that just the combinatorics.

We will determine the nontransverse bisector intersections and prove that they remain nontransverse in the family of Ford domains for groups in the 1–parameter family where the unipotent generator becomes twist parabolic.

The next proposition follows from the restrictive character of bounding bisectors, namely that they are all covertical (because they define faces of a Ford domain).

Proposition	Indices
5.4	8, 14–16, 21–25, 29–33, 35
5.8	2, 12, 19, 26
5.10	3, 5, 9, 10, 11, 18, 20, 28
5.11	4[5, 10], 7[3], 13[2, 5, 10], 17[9], 27[9, 18], 36[17, 28, 34]
5.12	6, 34

Table 4: Indices where the arguments of each proposition apply to study $\mathcal{B}_1 \cap \mathcal{B}_k$

Proposition	Indices
5.4	16, 17, 22–36
5.8	10, 13
5.10	1, 2, 5, 6, 7, 11
5.11	9[11], 14[7], 15[7], 18[1,10], 19[11], 20[7], 21[6,13]
5.12	4, 12
5.13	8

Table 5: Indices where the arguments of each proposition apply to study $\mathcal{B}_3 \cap \mathcal{B}_k$

Proposition 5.14 *Let $J=\{j,k\}$ with $j \neq k$. Then the intersection $F_{\{j\},\varnothing} \cap F_{\{k\},\varnothing} = F_{J,\varnothing}$ is transverse at every point of $F_{J,\varnothing}$.*

The analogous statement is not true when $|J| \geq 3$, since $F_{J,\varnothing}$ can have singular points; see Figure 5, for instance. This will not be bothersome in the context of our polyhedron F because of the following:

Proposition 5.15 *Suppose $|J| = 3$ and F_J is nonempty. Then the corresponding intersection of three bisectors (or two bisectors and $\partial_\infty H^2_{\mathbb{C}}$) is transverse at every point of F_J.*

Proof This follows from the fact that double points of $F_{J,\varnothing}$ occur only away from the face F_J. Indeed, one can easily locate these double points by the techniques explained in Section 5C4, and check that they are outside F by using interval arithmetic. □

The situation near vertices is slightly more subtle, mainly because our group contains some torsion elements; hence one expects the intersections to be nongeneric near the fixed points of those torsion elements.

We will check possible tangencies between 1–faces intersecting at each vertex. More generally, for each $j \neq k$, we will study tangencies between all the curves of the form $F_{\{j,k,l\},\varnothing}$ for $l \neq j,k$ that occur at a vertex of F.

Proposition 5.16 *Let p be an ideal vertex of F, ie a vertex in $\partial_\infty H^2_{\mathbb{C}}$. Then there are precisely three bounding bisectors $\mathcal{B}_i$, $\mathcal{B}_j$ and $\mathcal{B}_k$ meeting at p (where $i,j,k>0$). The intersection of the four hypersurfaces in $\mathbb{C}^2$ given by the three extors, $\widehat{\mathcal{B}}_i$, $\widehat{\mathcal{B}}_j$ and $\widehat{\mathcal{B}}_k$, and $\partial_\infty H^2_{\mathbb{C}}$ is transverse; in particular, none of the four incident 1–faces are tangent at p.*

Note that the ideal 1–faces are drawn in red on Figures 3 and 4, so the vertices on the red curves are the ideal ones. The indices (i,j,k) that appear in the proposition, ie the bounding bisectors that contain a given ideal vertex, can be read off Figure 7. For example, $(1,3,5)$, $(1,3,11)$, $(1,9,11)$, $\dots$ are triples of indices that correspond to ideal vertices.

Proof We treat the example of $F_{\{0,1,3,5\}}$, the other ones being entirely similar. The parametrization of the Giraud disk $F_{\{1,3\},\{0\}}$ was already explained in Section 5C4.

The relevant vertex satisfies

$$\begin{aligned} x_1 &= 0.80979557\dots, & y_1 &= -0.58671213\dots, \\ x_2 &= -0.53336432\dots, & y_2 &= 0.84588562\dots. \end{aligned} \tag{27}$$

We write the equations of the bisectors in affine coordinates for complex hyperbolic space corresponding to the spinal coordinates, ie such that (z_1,z_2) corresponds to

$$p_{13}+z_1p_{30}+z_2p_{01},$$

where p_{jk} denotes, as before, the box product $p_j \boxtimes p_k$.

In these coordinates, $\mathcal{B}_1$ is given by $|z_1| = x_1^2+y_1^2 = 1$ and $\mathcal{B}_3$ is given by $|z_2| = x_2^2+y_2^2 = 1$; of course, other bisectors have more complicated equations.

The equation of the boundary of the ball is

$$2-\sqrt{7}y_2-4x_1-x_2-y_2\sqrt{7}x_1+x_2\sqrt{7}y_1+2x_1^2+2y_1^2+x_2^2+y_2^2-x_2x_1-y_2y_1=0,$$

and the equation for $\mathcal{B}_5$ is given by

$$3(x_1+x_2)-\sqrt{7}(y_1+y_2)-2x_2x_1-2y_2y_1-x_1^2-y_1^2-x_2^2-y_2^2=0.$$

One then computes the gradient of the left hand side of each of these four equations, and checks that they are linearly independent at the point from (27) (this is readily done using interval arithmetic). □

Proposition 5.17 *There are precisely six bounding bisectors containing p_2, indexed by 1, 2, 3, 5, 10, 11. The pairwise and 3–fold intersections of these six bisectors are all transverse, but some 4–fold are not, namely $\{1,2,3,10\}$, $\{1,2,5,11\}$ and $\{3,5,10,11\}$.*

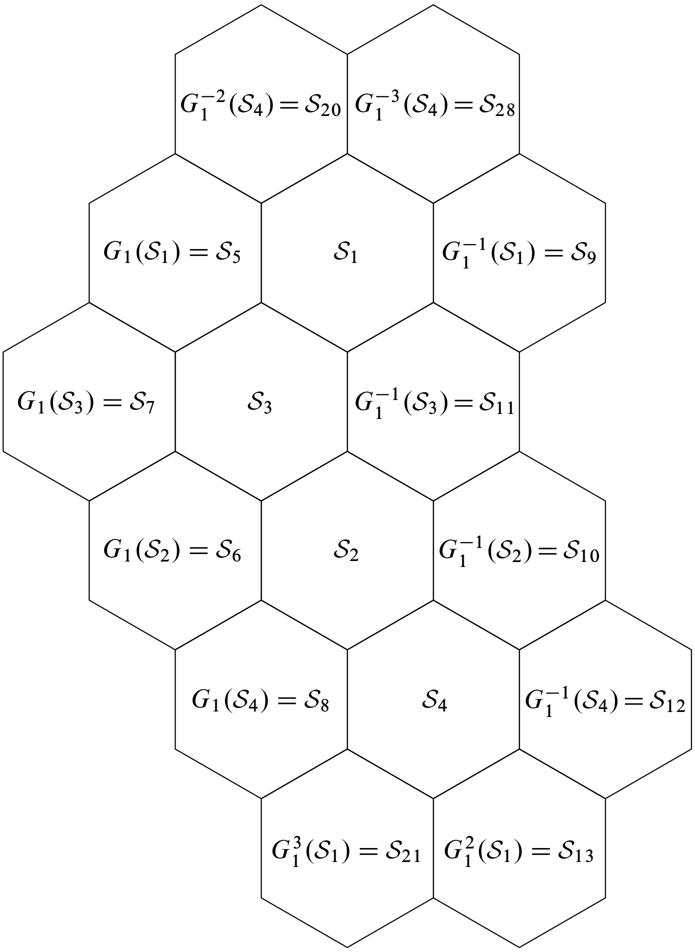

$G_1^{-2}(\mathcal{S}_4)=\mathcal{S}_{20}$
$G_1^{-3}(\mathcal{S}_4)=\mathcal{S}_{28}$
$G_1(\mathcal{S}_1)=\mathcal{S}_5$
$\mathcal{S}_1$
$G_1^{-1}(\mathcal{S}_1)=\mathcal{S}_9$
$G_1(\mathcal{S}_3)=\mathcal{S}_7$
$\mathcal{S}_3$
$G_1^{-1}(\mathcal{S}_3)=\mathcal{S}_{11}$
$G_1(\mathcal{S}_2)=\mathcal{S}_6$
$\mathcal{S}_2$
$G_1^{-1}(\mathcal{S}_2)=\mathcal{S}_{10}$
$G_1(\mathcal{S}_4)=\mathcal{S}_8$
$\mathcal{S}_4$
$G_1^{-1}(\mathcal{S}_4)=\mathcal{S}_{12}$
$G_1^{3}(\mathcal{S}_1)=\mathcal{S}_{21}$
$G_1^{2}(\mathcal{S}_1)=\mathcal{S}_{13}$

Figure 7: The combinatorics at infinity of the fundamental domain, near the faces for $G_2^{\pm}$ and $G_3^{\pm}$, which are representatives of all faces modulo the action of G_1

The precise list of bisectors that contain this vertex were already justified in Section 4B; see Lemma 4.2 and Proposition 5.8. The point of Proposition 5.17 is to give precise information about transversality. Recall from Section 4B that p_2 is, by definition, the isolated fixed point of G_2, and the bisectors $\mathcal{B}_1$, $\mathcal{B}_2$, $\mathcal{B}_3$, $\mathcal{B}_5$, $\mathcal{B}_{10}$ and $\mathcal{B}_{11}$ are the bounding bisectors corresponding to the group elements G_2, G_2^{-1}, G_3, G_1G_2, $G_1^{-1}G_2^{-1}$ and $G_1^{-1}G_3$, respectively; see Section 5A.

Proof We work in spinal coordinates for $\mathcal{B}_1\cap\mathcal{B}_3$, and as in the preceding proof, we use $z_j=x_j+iy_j$, $j=1,2$ as global coordinates on $H^2_{\mathbb{C}}$. The point p_2 is given by $z_1=1$, $z_2=\frac{1}{4}(3+i\sqrt{7})$.

Vector	Tangent to	Exit in $+$ direction	Exit in $-$ direction
u_1	$1,2,3,10$	11	5
u_2	$1,2,5,11$	3	10
u_3	$3,5,10,11$	1	2

Table 6: Each direction tangent vector u_k to a nontranverse quadruple intersection at p_2 exits the polyhedron; in the last two columns we list the two half spaces it exits (transversely) in the $\pm u_k$ direction.

The equations of the six bisectors are as follows:

$$\begin{aligned}
1:&\quad 4-4(x_1^2+y_1^2)=0,\\
2:&\quad 2+x_1+2x_2+(y_1-2y_2)\sqrt{7}+(x_1y_2-x_2y_1)\sqrt{7}\\
&\qquad\qquad +3(x_1x_2+y_1y_2)-(x_1^2+y_1^2)-4(x_2^2+y_2^2)=0,\\
3:&\quad 4-4(x_2^2+y_2^2)=0,\\
5:&\quad 3(x_1+x_2)-\sqrt{7}(y_1+y_2)-2(x_2x_1+y_2y_1)-(x_1^2+y_1^2)-(x_2^2+y_2^2)=0,\\
10:&\quad 2-4(x_1-x_2)+4(x_2x_1+y_2y_1)-2(x_1^2+y_1^2)-2(x_2^2+y_2^2)=0,\\
11:&\quad 3-2x_2+3x_1+\sqrt{7}y_1+3(x_1x_2+y_1y_2)+(x_2y_1-y_2x_1)\sqrt{7}\\
&\qquad\qquad -4(x_1^2+y_1^2)-(x_2^2+y_2^2)=0.
\end{aligned}$$

One computes the gradients at the point $x_1=1,\ y_1=0,\ x_2=3/4,\ y_2=\sqrt{7}/4$, which are given by

$$\begin{aligned}
v_1&=(-8,0,0,0), & v_5&=(-1/2,-3\sqrt{7}/2,-1/2,-3\sqrt{7}/2),\\
v_2&=(3,\sqrt{7},-1,-3\sqrt{7}), & v_{10}&=(-5,\sqrt{7},5,-\sqrt{7}),\\
v_3&=(0,0,-6,-2\sqrt{7}), & v_{11}&=(-9/2,5\sqrt{7}/2,-1/2,-3\sqrt{7}/2),
\end{aligned}$$

and the claim of the proposition follows from explicit rank computations.

The tangent vectors to the intersection are given by

$$\begin{aligned}
u_1&=(0,8/3,-\sqrt{7}/3,1),\\
u_2&=(0,0,-3\sqrt{7},1),\\
u_3&=(-2\sqrt{7}/3,-2/3,-\sqrt{7}/3,1),
\end{aligned}$$

and one easily checks that any curve tangent to these vectors must exit the polyhedron in a transverse fashion, more specifically, the exited bisectors are given in Table 6. □

6 Side pairings

6A Faces paired by G_2

We now justify the fact that G_2^{-1} defines an isometry between the faces for G_2 and G_2^{-1}. On the level of 2–faces, this follows from the proposition below.

Proposition 6.1 *The isometry G_2^{-1} maps*

(1) G_3p_0 *to* $G_1^{-1}G_3p_0$;

(2) $G_1^{-3}G_3^{-1}p_0$ *to* $G_1G_3^{-1}p_0$;

(3) $G_1^{-1}G_3p_0$ *to* $G_1^{-1}G_2^{-1}p_0$;

(4) $G_1^{-1}G_2p_0$ *to* $G_3^{-1}p_0$;

(5) $G_1^{-2}G_3^{-1}p_0$ *to* $G_1G_2^{-1}p_0$;

(6) $G_1G_2p_0$ *to* G_3p_0;

(7) $G_1^{-1}G_2^{-1}p_0$ *to* $G_1G_2p_0$;

(8) $G_1^{-2}G_2^{-1}p_0$ *to* $G_1^2G_2p_0$.

Proof We show a slightly stronger statement; namely, in order to show $G_2^{-1}gp_0 = hp_0$, we will exhibit $h^{-1}G_2^{-1}g$ as an explicit power of G_1.

The result follows from the presentation of the group (strictly speaking, they only depend on the relations we know to hold, not on the fact that this really gives a presentation). For the sake of brevity, we use word notation.

(1) $\bar{3}1\bar{2}3 = \bar{2}\bar{1}21\bar{2}\bar{2}12 = \bar{2}\bar{1}\bar{2}\bar{1}\bar{2} = 1$;

(2) $3\bar{1}\bar{2}\bar{1}^3\bar{3} = \bar{2}12 \cdot 121121 \cdot \bar{1} \cdot 121121 \cdot 2 = \bar{2}(12121)(12121)1212 = \bar{2}^4\bar{1} = \bar{1}$;

(3) $21\bar{2}\bar{1}3 = 21\bar{2}\bar{1}\bar{2}12 = \bar{1}$;

(4) $3\bar{2}\bar{1}2 = \mathrm{id}$;

(5) $2\bar{1}\bar{2}\bar{1}^2\bar{3} = 2(\bar{1}\bar{2}\bar{1})^2 2 = 2(121)2 = \bar{1}$;

(6) $\bar{3}\bar{2}12 = \mathrm{id}$;

(7) $\bar{2}\bar{1}\bar{2}\bar{1}\bar{2} = 1$;

(8) $\bar{2}\bar{1}^2\bar{2}\bar{1}^2\bar{2} = 1^2$. □

On the level of vertices, we have

- $G_2^{-1} p_2 = p_2$;
- $G_2^{-1} p_{\bar{1}21} = p_{\bar{3}23}$;
- $G_2^{-1} p_{21^3} = p_{23^3}$;
- $G_2^{-1} p_{121^2} = p_{323^2} = p_{12\bar{1}}$.

6B Faces paired by G_3

The corresponding statement about the side-pairing map for the other two base faces is the following.

Proposition 6.2 *The isometry G_3^{-1} maps*

(1) $G_2 p_0$ *to* $G_1^{-1} G_3^{-1} p_0$;

(2) $G_2^{-1} p_0$ *to* $G_2^{-1} p_0$;

(3) $G_1 G_2 p_0$ *to* $G_1^2 G_2 p_0$;

(4) $G_1 G_2^{-1} p_0$ *to* $G_1 G_3^{-1} p_0$;

(5) $G_1 G_3 p_0$ *to* $G_1^3 G_2 p_0$;

(6) $G_1^{-1} G_3 p_0$ *to* $G_1^{-1} G_2^{-1} p_0$.

Proof The method of proof is identical to that of Proposition 6.1.

(1) $31\bar{3}2 = \bar{2}1212\bar{1}\bar{2}\bar{2} = \bar{2}(\bar{2}\bar{1}\bar{2})^2\bar{2} = 1$;

(2) $2\bar{3}\bar{2} = 2\bar{2}\bar{1} = 1$;

(3) $\bar{2}\bar{1}^2\bar{3}12 = \bar{2}\bar{1}(212)^2 = \bar{2}\bar{1}\bar{2}\bar{1}\bar{2} = 1$;

(4) $3\bar{1}\bar{3}\bar{1}\bar{2} = \mathrm{id}$;

(5) $\bar{2}\bar{1}^3\bar{3}13 = \bar{2}\bar{1}\bar{1}\cdot\bar{1}\bar{2}\bar{1}\cdot 212\cdot 2212 = \bar{2}\bar{1}(212)^2 = \bar{2}\bar{1}\bar{2}\bar{1}\bar{2} = 1$;

(6) $21\bar{3}\bar{1}3 = 21\bar{2}\bar{1}2\bar{1}\bar{2}12 = 21\bar{2}\bar{1}\bar{2}\bar{2}\bar{2}\bar{1}\bar{2}12 = (21212)^3 = \bar{1}^2$. □

On the level of vertices, we have

- $G_3^{-1} p_2 = p_{\bar{3}23}$;
- $G_3^{-1} p_{12\bar{1}} = p_{1^32}$.

The last equality holds because

$$\bar{3}12\bar{1}3 = \bar{2}\bar{1}212\bar{1}\bar{2}12 = 1(212)^3112 = 1^32.$$

Ridge cycle	Relation
$2\cap 3 \xrightarrow{\bar{2}} \bar{1}31\cap\bar{2} \xrightarrow{\bar{1}\bar{3}1} \bar{3}\cap\bar{1}\bar{3}1 \xrightarrow{3} 2\cap 3$	$2=[3,\bar{1}]$
$2\cap\bar{1}^3\bar{3}1^3 \xrightarrow{\bar{2}} 1\bar{3}\bar{1}\cap\bar{2} \xrightarrow{13\bar{1}} 3\cap 13\bar{1} \xrightarrow{\bar{3}} 1^32\bar{1}^3\cap\bar{3}$	$\bar{1}^3\bar{3}131\bar{1}\bar{2}=\mathrm{id}$
$2\cap\bar{1}31 \xrightarrow{\bar{2}} \bar{1}\bar{2}1\cap\bar{2} \xrightarrow{\bar{1}21} \bar{1}^3\bar{3}1^3\cap\bar{1}21 \xrightarrow{\bar{1}^331^3} \bar{1}^221^2\cap\bar{1}^331^3$	$\bar{1}31^221\bar{2}=\mathrm{id}$
$2\cap\bar{1}21 \xrightarrow{\bar{2}} \bar{3}\cap\bar{2} \xrightarrow{3} \bar{2}\cap 3 \xrightarrow{2} 12\bar{1}\cap 2$	$12=23$
$2\cap\bar{1}\bar{2}1 \xrightarrow{\bar{2}} 12\bar{1}\cap\bar{2}$	$(12)^3$
$2\cap\bar{1}^2\bar{2}1^2 \xrightarrow{\bar{2}} 1^22\bar{1}^2\cap\bar{2}$	$(121)^3$

Table 7: Ridge cycles and the corresponding relations in the group

7 Ridge cycles

Because of Giraud's theorem, the ridge cycles automatically satisfy the hypotheses of the Poincaré polyhedron theorem. In particular, we get the following:

Theorem 7.1 *D is a fundamental domain for the action of cosets of $\langle G_1\rangle$ in Γ. In particular, $D=F$ (see Theorem 5.1).*

Every ridge cycle is equivalent to one of the cycles listed in Table 7 (equivalent means that we allow shifting within the cycle, and also conjugation by a power of G_1). We list the cycle until we come back to the image of the initial ridge under a power G_1^k (in that case, we close up the cycle by G_1^{-k}).

Using the relations

$$12=23,\quad (12)^3=(121)^3=\mathrm{id},$$

the other relations give $2^4=\mathrm{id}$. Indeed, $\bar{1}^3\bar{3}131\bar{2}=\mathrm{id}$ gives

$$\mathrm{id}=\bar{1}^2\bar{3}13\bar{1}\bar{2}\bar{1}=\bar{1}^2\bar{2}\bar{1}2\cdot 1\bar{2}12\cdot\bar{1}\bar{2}\bar{1}=\bar{1}(121)^221\bar{2}12(121)^2=21\bar{2}^3121=21(\bar{2}^4)\bar{1}\bar{2}.$$

It is easy to check that the above set of relations is actually *equivalent* to

$$12=23,\quad (12)^3=(121)^3=2^4=\mathrm{id}.$$

We summarize the above discussion in the following:

Theorem 7.2 *The group Γ has a presentation given by*

$$\big\langle G_1,G_2,G_3\mid G_2=[G_3,G_1^{-1}],\ G_1G_2=G_2G_3,\ G_2^4=\mathrm{id},\ (G_1G_2)3=\mathrm{id},\ (G_2G_1G_2)^3=\mathrm{id}\big\rangle.$$

8 Topology of the manifold at infinity

In this section, we prove that $\Gamma \setminus \Omega$ is indeed homeomorphic to the figure eight knot complement. This was already proved in [7] using a very different fundamental domain for the action of the group.

We write F for the Ford domain for Γ, E for $\partial_\infty F$, and C for ∂E. By construction, F, E and C are all G_1–invariant.

We will use Heisenberg coordinates (z,t) for $\partial H^2_{\mathbb{C}} \setminus \{p_\infty\}$; see Section 5B. In these coordinates, the action of G_1 is given by

$$G_1(z,t) = (z-1, t + \mathfrak{Im}(z)). \tag{28}$$

It follows from the results in Section 5A that C is tiled by hexagons, and that there are four orbits of these hexagons under the action of G_1. We need a bit more information about the identifications on these hexagons, namely, we need

- the incidence relations between various hexagons, and
- the identifications on C given by side-pairing maps.

The incidence relations follow immediately from the results in Section 5A, which are summarized in Figure 7.

The union U of the four hexagons labeled 1, 2, 3, 4 is embedded in C, and the action of G_1 induces identifications on ∂U. We denote by $\sim$ the corresponding equivalence relation on U; it is easy to check that $U/\sim$ is a torus.

We get the following result.

Proposition 8.1 *C is an unknotted topological cylinder, and E is the region exterior to C.*

Proof It follows from the fact that C is invariant under the action of G_1 that it is an unknotted cylinder in $\mathbb{C} \times \mathbb{R}$ (it is a $\mathbb{Z}$–covering of $C/\langle G_1 \rangle$). In fact, the real axis gives a core curve for the solid cylinder bounded by C. In view of G_1–invariance, it is enough to check that the interval [0, 1] on the x axis is outside E. This is readily checked; in fact, this interval is actually completely inside the spinal sphere $\mathcal{S}_1$. □

The identifications in C come from side pairings, which are described in Section 6. Figures 3 and 4 contain a list of vertices, which are uniquely determined by the list of faces they are on (in fact they are on precisely three bisectors).

For instance, there is a vertex on $b_1 \cap b_3 \cap G_1(b_1)$. By Proposition 6.1, G_2^{-1} maps this to the vertex on $b_2 \cap b_3 \cap G_1^{-1} b_3$. The vertex on $b_1 \cap b_3 \cap G_1^{-1}(b_3)$ is mapped to the vertex on $b_2 \cap G_1^{-1}(b_2) \cap G_1^{-1} b_3$. The image of these two points determine the image of the entire hexagon on b_1 (in Figure 7, the map flips the orientation of the hexagon).

By doing similar verifications, one checks that the identification pattern on the hexagons on $\mathcal{S}_1, \dots, \mathcal{S}_4$ is the same as the one for the Ford domain of the holonomy of the real hyperbolic structure on the figure eight knot complement, see Figure 2.

Now since the exterior of C is homeomorphic to $C \times [0, +\infty[$ (in a G_1–equivariant way), we get:

Corollary 8.2 *$\Gamma \setminus E$ is homeomorphic to the figure eight knot complement.*

9 Stability of the combinatorics

The first remark is that distinct bounding bisectors for the Ford domain for the unipotent solution are never cospinal, and as a consequence, the intersections $\hat{\gamma}_1 \cap \hat{\gamma}_2$ are uniquely determined by the triple p_0, $\gamma_1 p_0$, $\gamma_2 p_0$. Of course, this property will hold for all values of the twist parameter of G_1.

Now every point of an open 2–face is on precisely two bounding bisectors, and that intersection is transverse. In other words, every open 2–face will survive in small perturbations.

A similar remark holds for 1–faces, namely, no 1–face of the Ford domain for the boundary unipotent case is contained in a geodesic. In fact, every point on an open 1–face is on precisely three bounding bisectors, and these intersect transversely as well.

The only issue is to analyze vertices. There is nothing to check for the ideal vertices since they are defined as the intersection of four hypersurfaces (three bounding bisectors and the boundary of the ball) that intersect transversely.

The finite vertices are on more than four bounding bisectors, but they are also fixed by elliptic elements in the group. In fact, we already justified that they stayed on the same bisectors for small deformations; see Section 4B, more specifically, Lemmas 4.2 and 4.3. The transversality statement of Proposition 5.17 will remain true for small perturbations as well.

This implies that the combinatorics stay stable in small deformations.

10 Stability of the side pairing

Let $F^{(0)}$ be the Ford domain for the boundary unipotent group, and $F^{(t)}$ the one for the twist parabolic group corresponding to parameter t.

The proof that $F^{(0)}$ has side-pairings relies on the determination of the precise combinatorics, and also of the group relations. By the previous section, the combinatorics are stable, and by Proposition 4.1, the relations hold throughout the deformation. The proof of Propositions 6.1 and 6.2 then shows that $F^{(t)}$ has side-pairings, at least for small values of t.

The verification that the Ford domain for the boundary unipotent group satisfies the hypotheses of the Poincaré polyhedron theorem is given in Section 7. Since all intersections of bounding bisectors are Giraud disks, the cycle condition is a direct consequence of the existence of pairings.

Let Γ_t denote the image of ρ_t. We now get:

Theorem 10.1 *There exists a $\delta > 0$ such that whenever $|t| < \delta$, Γ_t is discrete with nonempty domain of discontinuity, its manifold at infinity is homeomorphic to the figure eight knot complement, and it has the presentation*

$$\big\langle G_1, G_2, G_3 \mid G_2 = [G_3, G_1^{-1}],\ G_1G_2 = G_2G_3,\ G_2^4 = \mathrm{id},\ (G_1G_2)^3 = \mathrm{id},\ (G_2G_1G_2)^3 = \mathrm{id}\big\rangle.$$

References

[1] **S Anan'in**, **C H Grossi**, **N Gusevskii**, *Complex hyperbolic structures on disc bundles over surfaces*, Int. Math. Res. Not. 2011 (2011) 4295–4375 MR

[2] **S Basu**, **R Pollack**, **M-F Roy**, *Algorithms in real algebraic geometry*, 2nd edition, Algorithms and Computation in Mathematics 10, Springer (2006) MR

[3] **N Bergeron**, **E Falbel**, **A Guilloux**, *Tetrahedra of flags, volume and homology of* $\mathrm{SL}(3)$, Geom. Topol. 18 (2014) 1911–1971 MR

[4] **H Cohen**, *A course in computational algebraic number theory*, Graduate Texts in Mathematics 138, Springer (1993) MR

[5] **M Deraux**, *Deforming the $\mathbb{R}$–Fuchsian $(4,4,4)$–triangle group into a lattice*, Topology 45 (2006) 989–1020 MR

[6] **M Deraux**, *On spherical CR uniformization of 3–manifolds*, Exp. Math. 24 (2015) 355–370 MR

[7] **M Deraux**, **E Falbel**, *Complex hyperbolic geometry of the figure-eight knot*, Geom. Topol. 19 (2015) 237–293 MR

[8] **M Deraux**, **J R Parker**, **J Paupert**, *New non-arithmetic complex hyperbolic lattices*, Invent. Math. 203 (2016) 681–771 MR

[9] **E Falbel**, *A spherical CR structure on the complement of the figure eight knot with discrete holonomy*, J. Differential Geom. 79 (2008) 69–110 MR

[10] **E Falbel**, **A Guilloux**, **P-V Koseleff**, **F Rouillier**, **M Thistlethwaite**, *Character varieties for* $\mathrm{SL}(3,\mathbb{C})$*: the figure eight knot*, Exp. Math. 25 (2016) 219–235 MR

[11] **E Falbel**, **P-V Koseleff**, **F Rouillier**, *Representations of fundamental groups of* 3*–manifolds into* $\mathrm{PGL}(3,\mathbb{C})$*: exact computations in low complexity*, Geom. Dedicata 177 (2015) 229–255 MR

[12] **G Giraud**, *Sur certaines fonctions automorphes de deux variables*, Ann. Sci. École Norm. Sup. 38 (1921) 43–164 MR

[13] **W M Goldman**, *Conformally flat manifolds with nilpotent holonomy and the uniformization problem for* 3*–manifolds*, Trans. Amer. Math. Soc. 278 (1983) 573–583 MR

[14] **W M Goldman**, *Complex hyperbolic geometry*, Clarendon, New York (1999) MR

[15] **W M Goldman**, **M Kapovich**, **B Leeb**, *Complex hyperbolic manifolds homotopy equivalent to a Riemann surface*, Comm. Anal. Geom. 9 (2001) 61–95 MR

[16] **W M Goldman**, **J R Parker**, *Dirichlet polyhedra for dihedral groups acting on complex hyperbolic space*, J. Geom. Anal. 2 (1992) 517–554 MR

[17] **M Heusener**, **V Muñoz**, **J Porti**, *The* $\mathrm{SL}(3,\mathbb{C})$*–character variety of the figure eight knot*, preprint (2015) arXiv

[18] **Y Kamishima**, **T Tsuboi**, *CR-structures on Seifert manifolds*, Invent. Math. 104 (1991) 149–163 MR

[19] **J R Parker**, *Complex hyperbolic Kleinian groups*, to appear, Cambridge University Press

[20] **J R Parker**, **I D Platis**, *Open sets of maximal dimension in complex hyperbolic quasi-Fuchsian space*, J. Differential Geom. 73 (2006) 319–350 MR

[21] **J R Parker**, **P Will**, *A complex hyperbolic Riley slice*, preprint (2015) arXiv

[22] **R Riley**, *A quadratic parabolic group*, Math. Proc. Cambridge Philos. Soc. 77 (1975) 281–288 MR

[23] **R Riley**, *Nonabelian representations of* 2*–bridge knot groups*, Quart. J. Math. Oxford Ser. 35 (1984) 191–208 MR

[24] **F Rouillier**, *Solving zero-dimensional systems through the rational univariate representation*, Appl. Algebra Engrg. Comm. Comput. 9 (1999) 433–461 MR

[25] **R E Schwartz**, *Degenerating the complex hyperbolic ideal triangle groups*, Acta Math. 186 (2001) 105–154 MR

[26] **R E Schwartz**, *Complex hyperbolic triangle groups*, from "Proceedings of the International Congress of Mathematicians, Vol II" (T Li, editor), Higher Ed. Press, Beijing (2002) 339–349 MR

[27] **R E Schwartz**, *Real hyperbolic on the outside, complex hyperbolic on the inside*, Invent. Math. 151 (2003) 221–295 MR

[28] **R E Schwartz**, *Spherical CR geometry and Dehn surgery*, Annals of Mathematics Studies 165, Princeton University Press (2007) MR

[29] **D Toledo**, *Representations of surface groups in complex hyperbolic space*, J. Differential Geom. 29 (1989) 125–133 MR

[30] **P Will**, *The punctured torus and Lagrangian triangle groups in* $\mathrm{PU}(2,1)$, J. Reine Angew. Math. 602 (2007) 95–121 MR

Institut Fourier, Université Grenoble Alpes,
100 rue des maths, 38610 Gières, France

deraux@ujf-grenoble.fr

Proposed: Walter Neumann
Seconded: Jesper Grodal, Dmitri Burago

Received: 9 September 2015
Revised: 15 February 2016

Geometry & Topology 20 (2016) 3623–3673

Concordance maps in knot Floer homology

ANDRÁS JUHÁSZ
MARCO MARENGON

We show that a decorated knot concordance $\mathcal{C}$ from K to K' induces a homomorphism $F_{\mathcal{C}}$ on knot Floer homology that preserves the Alexander and Maslov gradings. Furthermore, it induces a morphism of the spectral sequences to $\widehat{\mathrm{HF}}(S^3) \cong \mathbb{Z}_2$ that agrees with $F_{\mathcal{C}}$ on the E^1 page and is the identity on the E^∞ page. It follows that $F_{\mathcal{C}}$ is nonvanishing on $\widehat{\mathrm{HFK}}_0(K, \tau(K))$. We also obtain an invariant of slice disks in homology 4–balls bounding S^3.

If $\mathcal{C}$ is invertible, then $F_{\mathcal{C}}$ is injective, hence

$$\dim \widehat{\mathrm{HFK}}_j(K, i) \le \dim \widehat{\mathrm{HFK}}_j(K', i)$$

for every $i, j \in \mathbb{Z}$. This implies an unpublished result of Ruberman that if there is an invertible concordance from the knot K to K', then $g(K) \le g(K')$, where g denotes the Seifert genus. Furthermore, if $g(K) = g(K')$ and K' is fibred, then so is K.

57M27, 57R58

1 Introduction

Knot Floer homology was introduced independently by Ozsváth and Szabó [28] and Rasmussen [31], and the first author [16] defined maps induced on it by decorated knot cobordisms. Given a knot K in S^3, its knot Floer homology with $\mathbb{Z}_2$ coefficients is a finite dimensional bigraded $\mathbb{Z}_2$–vector space

$$\bigoplus_{i,j \in \mathbb{Z}} \widehat{\mathrm{HFK}}_j(K, i),$$

well-defined up to isomorphism, where i is called the Alexander grading and j is the homological grading. The Euler characteristic of $\widehat{\mathrm{HFK}}_*(K, i)$ is the i^{th} coefficient of the symmetrized Alexander polynomial of K, and hence knot Floer homology can be viewed as a categorification of the Alexander polynomial. First, we recall [16, Definition 4.1].

Definition 1.1 For $i \in \{0, 1\}$, let Y_i be a connected, oriented 3–manifold, and let L_i be a nonempty link in Y_i. Then a *link cobordism* from (Y_0, L_0) to (Y_1, L_1) is a pair (X, F), where

Published: 21 December 2016 DOI: 10.2140/gt.2016.20.3623

(1) X is a connected, oriented cobordism from Y_0 to Y_1,

(2) F is a properly embedded, compact, orientable surface in X, and

(3) $\partial F = L_0 \cup L_1$.

Knots K_0 and K_1 in S^3 are said to be *concordant* if there is a cobordism (X, F) from (S^3, K_0) to (S^3, K_1) such that $X = S^3 \times I$ and F is diffeomorphic to $S^1 \times I$. In this case, we call (X, F) a *concordance* from K_0 to K_1. In this paper, we also allow more general concordances where X is a cobordism from S^3 to S^3 such that $H_1(X) = H_2(X) = 0$.

In this paper, a *decorated knot* is a pair (K, P) such that K is a knot, P is a pair of points in K, and we are given a decomposition of K into compact 1–manifolds $R_+(P)$ and $R_-(P)$ such that $R_+(P) \cap R_-(P) = P$. Given decorated knots (K_0, P_0) and (K_1, P_1) in S^3, a *decorated concordance* from (K_0, P_0) to (K_1, P_1) is a triple (X, F, σ) such that (X, F) is a concordance from K_0 to K_1, and σ consists of two disjoint, properly embedded arcs in F, one connecting $R_+(K_0)$ and $R_+(K_1)$, the other $R_-(K_0)$ and $R_-(K_1)$.

Dylan Thurston and the first author [17] showed that knot Floer homology is natural for decorated knots, and Sarkar [35] proved that moving the basepoints P around the knot induces a nontrivial automorphism in many cases. Hence only decorated concordances induce maps on knot Floer homology.

Recall from [28, Lemma 3.6] that for every decorated knot (K, P) in S^3, there is a corresponding spectral sequence

$$\widehat{\mathrm{HFK}}(K, P) \implies \widehat{\mathrm{HF}}(S^3) \cong \mathbb{Z}_2.$$

Given an admissible doubly pointed Heegaard diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, w, z)$ for (K, P), the singly pointed diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, w)$ represents (S^3, w), and z gives rise to the knot filtration on $\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, w)$. The spectral sequence arises from this filtered complex. The E^0 page is the associated graded complex $\widehat{\mathrm{CFK}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, w, z)$, whose homology is $\widehat{\mathrm{HFK}}(K, P)$, the E^1 page. The spectral sequence limits to the homology of $\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, w)$, which is $\widehat{\mathrm{HF}}(S^3) \cong \mathbb{Z}_2$. The filtration level of the generator of $\mathbb{Z}_2$ in the E^∞ page is the Ozsváth–Szabó τ invariant [26], denoted by $\tau(K)$.

The main result of this paper is that a decorated concordance $\mathcal{C}$ induces a nonvanishing homomorphism $F_{\mathcal{C}}$ on knot Floer homology that preserves the Alexander and homological gradings, and also induces a morphism of the corresponding spectral sequences. The map $F_{\mathcal{C}}$ is functorial and depends only on the decorated concordance $\mathcal{C}$, while the chain map $f_{\mathcal{C}}$ (or even its filtered homotopy type) need not be functorial, and it can depend on auxiliary data other than $\mathcal{C}$.

Theorem 1.2 *Let (K_0, P_0) and (K_1, P_1) be decorated knots in S^3. Let $\mathcal{C} = (X, F, \sigma)$ be a decorated concordance between them such that $H_1(X) = H_2(X) = 0$. Then*

$$F_{\mathcal{C}}(\widehat{\mathrm{HFK}}_j(K_0, P_0, i)) \le \widehat{\mathrm{HFK}}_j(K_1, P_1, i)$$

for every $i, j \in \mathbb{Z}$.

Furthermore, given an admissible diagram $(\Sigma_r, \boldsymbol{\alpha}_r, \boldsymbol{\beta}_r, w_r, z_r)$ of (K_r, P_r) for r in $\{0, 1\}$, there is a filtered chain map

$$f_{\mathcal{C}}\colon \widehat{\mathrm{CF}}(\Sigma_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, w_0) \to \widehat{\mathrm{CF}}(\Sigma_1, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, w_1)$$

of homological degree zero such that the induced morphism of spectral sequences agrees with $F_{\mathcal{C}}$ on the E^1 page and with $\mathrm{Id}_{\mathbb{Z}_2}$ on the total homology and on the E^∞ page.

Note that the fact that the map induced by a filtered map f on the total homology is an isomorphism in general does not imply that the map f^∞ induced between the E^∞ pages is also an isomorphism. As an example, consider a complex $C \cong \mathbb{Z}_2$ in filtration level one, and a complex $\overline{C} \cong \mathbb{Z}_2$ in filtration level zero. If $f\colon C \to \overline{C}$ is an isomorphism, then $H(f)$ is an isomorphism but f^∞ is not.

In the case of the filtered map $f_{\mathcal{C}}$ induced by a decorated concordance $\mathcal{C}$, the fact that $f_{\mathcal{C}}^\infty$ is an isomorphism follows from the fact that $\tau(K_0) = \tau(K_1)$, which was shown by Ozsváth and Szabó [26, Theorem 1.1]. An alternative proof of this can be given by observing that a decorated concordance gives filtered maps both ways that induce isomorphisms on the total homology, as in the proofs of Theorem 1 in Rasmussen [32] and Theorem 3.4 in Sarkar [34].

The invariant $\tau(K)$ can also be defined as the smallest Alexander grading of an element of $\widehat{\mathrm{HFK}}(K, P)$ that represents a cycle on each page of the spectral sequence, and whose homology class in the E^∞ page is 1. We denote the set of such elements by $A_1(K)$. Then we have the following nonvanishing result for the knot concordance maps:

Corollary 1.3 *Let (K_0, P_0) and (K_1, P_1) be decorated knots in S^3, and suppose that $\mathcal{C} = (X, F, \sigma)$ is a decorated concordance between them. Let $\tau = \tau(K_0) = \tau(K_1)$. Then, the map*

$$F_{\mathcal{C}}\colon \widehat{\mathrm{HFK}}_0(K_0, P_0, \tau) \to \widehat{\mathrm{HFK}}_0(K_1, P_1, \tau)$$

is nonzero, and $F_{\mathcal{C}}(A_1(K_0)) \subseteq A_1(K_1)$.

In fact, for any decorated knot (K, P) in S^3, we shall see that

$$A_1'(K) := A_1(K) \cap \widehat{\mathrm{HFK}}_0(K, P, \tau(K)) \neq \varnothing,$$

and the map $F_{\mathcal{C}}\colon A_1'(K_0) \to A_1'(K_1)$ is nonzero.

Let B be an integral homology 4–ball with boundary S^3. Suppose that $S \subset B$ is a slice disk for the decorated knot (K, P) in S^3. If we remove a ball from B about a point of S, we obtain a concordance $\mathcal{C}(S)$ from the unknot U to K. By Lemma 3.11, the element

$$t_{S,P} := F_{\mathcal{C}(S)}(1) \in \widehat{\mathrm{HFK}}_0(K, P, 0)$$

is independent of what decoration we choose on $\mathcal{C}(S)$. It is nonzero by Corollary 1.3, and is an invariant of the surface S up to isotopy in B fixing K.

Question 1.4 Can $t_{S,P}$ distinguish different slice disks? More precisely, is there a decorated knot (K, P) in S^3 that has two different slice disks S and S' in D^4 such that $t_{S,P} \neq t_{S',P}$?

Note that, given different decorations P and P' on K, the basepoint moving map of Sarkar [35] takes $t_{S,P}$ to $t_{S,P'}$, so the answer is independent of the choice of basepoints.

We can use the above viewpoint to refine the approach of Freedman, Gompf, Morrison and Walker [6] for disproving the smooth 4–dimensional Poincaré conjecture (SPC4). Suppose that we are given a counterexample to SPC4 with no 3–handles and a single 4–handle. Removing the 4–handle, we obtain an exotic 4–ball B with boundary homeomorphic to S^3. The belt circles of the 2–handles give a link $L \subset \partial B$, and the cocores of the 2–handles give a collection of disks $C \subset B$ with boundary L. If we band sum the components of L in some way, we obtain a knot $K \subset \partial B$, together with a disk $D \subset B$ obtained from C. Hence D induces an element $t_{D,P} \in \widehat{\mathrm{HFK}}(K, P)$ for any decoration P. If $t_{D,P} \neq t_{S,P}$ for S an arbitrary slice disk of K, then this implies that B is indeed exotic.

The approach of Freedman et al only works if K is not slice in the standard 4–ball, but it is in the homotopy 4–ball B. By the work of Ozsváth and Szabó [26, Theorem 1.1], the τ invariant vanishes if K bounds a disk in a homotopy ball, and so does Rasmussen's s invariant according to Kronheimer and Mrowka [19], so neither can be used for the above purpose. We could use any other theory equipped with knot concordance maps in manifolds homeomorphic to $S^3 \times I$. However, note that the Khovanov homology concordance maps of Jacobsson [12] are only defined when the ambient manifold is diffeomorphic to $S^3 \times I$.

A knot is called doubly slice if it is a hyperplane cross-section of an unknotted S^2 in S^4. Motivated by a question of Fox [5] asking which knots are doubly slice, Sumners [38] introduced the notion of invertible knot cobordisms. In his terminology, cobordism stands for concordance; we use the latter for clarity.

Definition 1.5 Let K_0 and K_1 be knots in S^3. We say that a concordance $(S^3 \times I, F)$ from K_0 to K_1 is *invertible* if there is a concordance $(S^3 \times I, F')$ from K_1 to K_0 such that the composition of $(S^3 \times I, F)$ and $(S^3 \times I, F')$ from K_0 to K_0 is equivalent to the trivial cobordism. We write $K_0 \le K_1$ if there is an invertible cobordism from K_0 to K_1.

In other words, F is invertible if and only if $(S^3 \times I, F)$ has a left inverse in the cobordism category of links. A knot K is doubly slice if and only if $U \le K$. The relation $\le$ is a partial order on the set of knots in S^3, which follows from Silver and Whitten [36], as we shall explain later.

Theorem 1.6 *If there is an invertible concordance from K_0 to K_1, then*

$$\dim \widehat{\mathrm{HFK}}_j(K_0, i) \le \dim \widehat{\mathrm{HFK}}_j(K_1, i)$$

for every $i, j \in \mathbb{Z}$.

This provides an obstruction to the existence of an invertible concordance from K_0 to K_1. According to the work of Manolescu, Ozsváth and Sarkar [23], knot Floer homology is algorithmically computable, and Baldwin and Gillam [3] used this algorithm to compute it for knots with at most 12 crossings.

For a knot K in S^3, we denote its Seifert genus by $g(K)$. Ozsváth and Szabó [27] proved that knot Floer homology detects the genus of a knot, in the sense that

$$g(K) = \max\{i \in \mathbb{Z} : \widehat{\mathrm{HFK}}_*(K, i) \ne 0\}.$$

For a simpler proof of this fact, see Ni [25]. Furthermore, knot Floer homology also detects fibredness of knots, as $\dim \widehat{\mathrm{HFK}}_*(K, g(K)) = 1$ if and only if K is fibred. This was shown by Ghiggini [8] in the genus one case, and by Ni [25] and the first author [14; 15] in the general case. These two results, together with Theorem 1.6, immediately imply the following unpublished result of Ruberman.

Corollary 1.7 *The function g is monotonic with respect to the partial order $\le$ induced by invertible concordance. More concretely, if there is an invertible concordance from K_0 to K_1, then $g(K_0) \le g(K_1)$. Furthermore, if K_1 is fibred and $g(K_0)$ is equal to $g(K_1)$, then K_0 is also fibred.*

We now outline a more elementary proof of these results communicated to us by Ruberman, and which does not use the assumption $g(K_0) = g(K_1)$ for the second statement. Also see the proof of Silver and Whitten [36, Proposition 3.7] and the paragraph following it.

Proof Let F be an invertible concordance from K_0 to K_1 with inverse F'. Then there is a diffeomorphism $d\colon S^3 \times I \to S^3 \times I$ such that $d(F' \circ F) = K_0 \times I$ and $d|_{S^3 \times \partial I}$ is the identity. Let $i\colon S^3 \to S^3 \times I$ be the embedding $i(x) = \bigl(x, \frac{1}{2}\bigr)$, and let $p\colon S^3 \times I \to S^3$ be the projection. Then the composition

$$f = p \circ d \circ i\colon S^3 \to S^3$$

maps K_1 to K_0 such that $f^{-1}(K_0) = K_1$. We can isotope d such that $d\bigl(K_1 \times \{\frac{1}{2}\}\bigr)$ becomes transverse to the I–fibration of $K_0 \times I$, and hence $f|_{K_1}$ is an embedding with image K_0. If S is a minimal genus Seifert surface for K_1, then $f|_S$ satisfies the conditions of [7, Corollary 6.23], hence there exists a Seifert surface T of $K_0 = f(K_1)$ such that $g(T) \le g(S)$. It follows that $g(K_0) \le g(K_1)$. Recall that [7, Corollary 6.23] is a deep generalization of Dehn's lemma to higher genus surfaces due to Gabai. It states that if M is a compact oriented 3–manifold, S a compact oriented surface with connected boundary, and $f\colon S \to M$ a map such that $f|_{\partial S}$ is an embedding and $f^{-1}(f(\partial S)) = \partial S$, then there exists an embedded surface T in M such that $\partial T = f(\partial S)$ and $g(T) \le g(S)$.

Let $E(K_i)$ denote the exterior of the knot K_i for $i \in \{0, 1\}$. Then

$$f|_{E(K_1)}\colon E(K_1) \to E(K_0)$$

is a degree-one map as it is an orientation-preserving diffeomorphism between the boundary tori. Hence, by Rong [33, Lemma 1.2], it induces a surjection on the fundamental groups, and also on the commutator subgroups. If K_1 is fibred, then the commutator subgroup $\pi_1(E(K_1))'$ is finitely generated, hence $\pi_1(E(K_0))'$ is also finitely generated, so K_0 is fibred by a result of Stallings [37]. □

Let K and K' be knots in S^3 such that there is an epimorphism $\pi_1(E(K)) \to \pi_1(E(K'))$ preserving peripheral structure. By Silver and Whitten [36], this induces a partial order $\succeq$ on the set of knots. For example, if there is a degree-one map

$$(E(K), \partial E(K)) \to (E(K'), \partial E(K')),$$

in particular if $K \ge K'$, then $K \succeq K'$. Notice that this implies that $\ge$ is also a partial order. Based on the above proof and Theorem 1.6, it is natural to ask whether $K \succeq K'$ also implies that

$$\dim \widehat{\mathrm{HFK}}_*(K, i) \ge \dim \widehat{\mathrm{HFK}}_*(K', i) \tag{1-1}$$

for every $i \in \mathbb{Z}$. Note that this would imply [36, Conjecture 3.6] claiming that, if $K \succeq K'$, then $g(K) \ge g(K')$. Compare this with Karakurt and Lidman [18, Conjecture 9.4], which claims that if $f\colon Y \to Y'$ is a nonzero-degree map between integer homology spheres, then $\dim \widehat{\mathrm{HF}}(Y) \geq \dim \widehat{\mathrm{HF}}(Y')$. However, inequality (1-1) turns out to be false due to the following example constructed by Jennifer Hom.

Example 1.8 Let $K = (T_{2,3})_{2,3}$ be the $(2,3)$–cable of the right-handed trefoil $T_{2,3}$, and let $K' = T_{2,3}$. Then $K \succeq K'$. In fact, there is a degree-one map

$$(E(K), \partial E(K)) \to (E(K'), \partial E(K')).$$

Indeed, let $T \subset E(K)$ be the boundary of the solid torus used in the satellite construction for K. Then the exterior of T is $E(K')$, hence fibred over S^1. If we collapse the fibres to disks, we obtain a degree-one map from the exterior of T to $D^2 \times S^1$, and hence from $E(K)$ to $E(K')$. But both K and K' are determined by their Alexander polynomials, K' because it is alternating, and K by the work of Hedden [9, Theorem 1.0.6]. The symmetrized Alexander polynomial of K is

$$t^3 - t^2 + 1 - t^{-2} + t^{-3},$$

while the symmetrized Alexander polynomial of K' is $t - 1 + t^{-1}$. So $\widehat{\mathrm{HFK}}(K, 1) = 0$ and $\widehat{\mathrm{HFK}}(K', 1) = \mathbb{Z}_2$, violating inequality (1-1).

In light of this, we propose the following weaker question.

Question 1.9 Suppose that $K \succeq K'$. Then is it true that

$$\dim \widehat{\mathrm{HFK}}(K) \geq \dim \widehat{\mathrm{HFK}}(K')?$$

The paper is organized as follows: In Section 2, we review sutured manifold cobordisms and the maps induced by them on sutured Floer homology. In Section 3, we define the knot concordance maps, show that they preserve the Alexander grading (Proposition 3.10), and prove Theorem 1.6. Section 4 gives a brief overview of spectral sequences arising from a filtered complex. In Section 5, we show that, on the chain level, a knot concordance map can be represented by a chain map that preserves the Alexander filtration (Theorem 5.4) and therefore induces a morphism of spectral sequences (Theorem 5.5); this is precisely the second part of Theorem 1.2. Corollary 1.3 follows from Corollary 5.7. Finally, we prove in Section 6 that the knot concordance maps preserve the homological grading, which concludes the proof of Theorem 1.2.

Acknowledgements We would like to thank Daniel Ruberman for pointing out a more elementary proof of Corollary 1.7, and Ciprian Manolescu for drawing our attention to the grading shift formula in [22] and for his comments on an earlier version of this paper. We are also grateful to Jennifer Hom for Example 1.8. Finally, we would like to thank the referee for the invaluable suggestions.

Juhász was supported by a Royal Society Research Fellowship. Marengon was supported by an EPSRC Doctoral Training Award.

2 Cobordisms of sutured manifolds

In this section, we briefly review sutured manifold cobordisms, and the maps they induce on sutured Floer homology, as defined by the first author [16].

2A Sutured manifolds and sutured cobordisms

Definition 2.1 [7, Definition 2.6] A *sutured manifold* is a compact oriented 3–manifold M with boundary together with a set $\gamma \subseteq \partial M$ of pairwise disjoint annuli $A(\gamma)$ and tori $T(\gamma)$. Furthermore, the interior of each component of $A(\gamma)$ contains a homologically nontrivial oriented simple closed curve, called a *suture*. We denote the set of sutures by $s(\gamma)$.

Finally, every component of $R(\gamma) = \partial M \setminus \mathrm{Int}(\gamma)$ is oriented such that $\partial R(\gamma)$ is coherent with the sutures. Let $R_+(\gamma)$ (or $R_-(\gamma)$) denote the components of $R(\gamma)$ whose normal vectors points out of (into) M.

Definition 2.2 [13, Definition 2.2] We say that a sutured manifold (M, γ) is *balanced* if M has no closed components, $\chi(R_+(\gamma))$ is equal to $\chi(R_-(\gamma))$, and the map $\pi_0(A(\gamma)) \to \pi_0(\partial M)$ is surjective.

From now on, we only consider sutured manifolds where $T(\gamma) = \varnothing$, and view γ as a "thickened" oriented 1–manifold. So we often do not distinguish between γ and $s(\gamma)$; it shall be clear from the context which one we mean.

Definition 2.3 [16, Definition 2.3] Let (M, γ) be a sutured manifold, and suppose that ξ_0 and ξ_1 are contact structures on M such that ∂M is a convex surface with dividing set γ with respect to both ξ_0 and ξ_1. Then we say that ξ_0 and ξ_1 are *equivalent* if there is a 1–parameter family $\{\xi_t : t \in I\}$ of contact structures such that ∂M is convex with dividing set γ with respect to ξ_t for every $t \in I$. In this case, we write $\xi_0 \sim \xi_1$, and we denote by $[\xi]$ the equivalence class of the contact structure ξ.

Definition 2.4 [16, Definitions 2.4 and 2.14] Let (M_0, γ_0) and (M_1, γ_1) be sutured manifolds. A *cobordism* from (M_0, γ_0) to (M_1, γ_1) is a triple $\mathcal{W} = (W, Z, [\xi])$, where

- W is a compact oriented 4–manifold with boundary,

- $Z \subseteq \partial W$ is a compact, codimension-0 submanifold with boundary (viewed within ∂W), such that $\partial W \setminus \operatorname{Int}(Z) = -M_0 \sqcup M_1$, and we view Z as a sutured manifold with sutures $\gamma_0 \cup \gamma_1$,
- ξ is a positive contact structure on Z such that ∂Z is a convex surface with dividing set γ_i on ∂M_i for $i \in \{0, 1\}$.

Finally, a cobordism is called *balanced* if both (M_0, γ_0) and (M_1, γ_1) are balanced.

In this paper, we will only consider balanced sutured manifolds and balanced cobordisms.

Definition 2.5 [16, Definition 2.7] We call two cobordisms $\mathcal{W} = (W, Z, [\xi])$ and $\mathcal{W}' = (W', Z', [\xi'])$ from (M_0, γ_0) to (M_1, γ_1) *equivalent* if there is an orientation-preserving diffeomorphism $\varphi\colon W \to W'$ such that $d(Z) = Z'$, $d_*(\xi) = \xi'$ and $d|_{M_0 \cup M_1} = \mathrm{Id}$.

Definition 2.6 [16, Definition 10.4] A cobordism $\mathcal{W} = (W, Z, [\xi])$ from (M_0, γ_0) to (N, γ_1) is a *boundary cobordism* if W is balanced, N is parallel to $M_0 \cup (-Z)$, and we are also given a deformation retraction $r\colon W \times [0, 1] \to M_0 \cup (-Z)$ such that $r_0|_W = \mathrm{Id}_W$ and $r_1|_N$ is an orientation-preserving diffeomorphism from N to $M_0 \cup (-Z)$.

Definition 2.7 [16, Definition 5.1] We say that a cobordism $\mathcal{W} = (W, Z, [\xi])$ from (M_0, γ_0) to (M_1, γ_1) is *special* if

1. $\mathcal{W}$ is balanced,
2. $\partial M_0 = \partial M_1$, and $Z = \partial M_0 \times I$ is the trivial cobordism between them,
3. ξ is an I–invariant contact structure on Z such that each $\partial M_0 \times \{t\}$ is a convex surface with dividing set $\gamma_0 \times \{t\}$ for every $t \in I$ with respect to the contact vector field $\partial/\partial t$.

In particular, it follows from (3) that $\gamma_0 = \gamma_1$.

Remark 2.8 Every sutured cobordism can be seen as the composition of a boundary cobordism and a special cobordism; see [16, Definition 10.1]. Let $\mathcal{W} = (W, Z, [\xi])$ be a balanced cobordism from (M_0, γ_0) to (M_1, γ_1). Let (N, γ_1) be the sutured manifold $(M_0 \cup (-Z), \gamma_1)$. Then we can think of the cobordism $\mathcal{W}$ as a composition $\mathcal{W}^s \circ \mathcal{W}^b$, where $\mathcal{W}^b$ is a boundary cobordism from (M_0, γ_0) to (N, γ_1) and $\mathcal{W}^s$ is a special cobordism from (N, γ_1) to (M_1, γ_1).

2B Relative Spin^c structures

Definition 2.9 [16, Definition 3.1] Given a sutured manifold (M,γ), we say that a vector field v defined on a subset of M containing ∂M is *admissible* if it is nowhere vanishing, it points into M along $R_-(\gamma)$, it points out of M along $R_+(\gamma)$, and $v|_\gamma$ is tangent to ∂M and either points into $R_+(\gamma)$ or is positively tangent to γ (we think of ∂M as a smooth surface, and of γ as a 1–manifold).

Let v and w be admissible vector fields on M. We say that v and w are *homologous*, and we write $v \sim w$, if there is a collection of balls $B \subseteq M$, one in each component of M, such that v and w are homotopic on $M \setminus B$ through admissible vector fields. Then $\operatorname{Spin}^c(M,\gamma)$ is the set of homology classes of admissible vector fields on M.

If (M,γ) is balanced, $\operatorname{Spin}^c(M,\gamma)$ is an affine space over $H^2(M,\partial M)$. Throughout this paper, we will denote relative Spin^c structures by $\mathfrak{s}^\circ$, to distinguish them from ordinary Spin^c structures on oriented 3–manifolds, usually denoted by $\mathfrak{s}$.

Remark 2.10 Let v_0 be a fixed vector field on ∂M arising as $v|_{\partial M}$ for some admissible vector field v on M. We define $\operatorname{Spin}^c_{v_0}(M,\gamma)$ as the set of nowhere vanishing vector fields on M that restrict to v_0 on ∂M, up to isotopy through such vector fields relative to ∂M in the complement of a collection of balls. Since the space of all possible v_0 is contractible, $\operatorname{Spin}^c_{v_0}(M,\gamma)$ can be canonically identified with $\operatorname{Spin}^c(M,\gamma)$. This was the approach taken in [13].

Definition 2.11 [16, Definition 3.2] Let (M,γ) be a sutured manifold. We say that an oriented 2–plane field ξ defined on a subset of M containing ∂M is *admissible* if there exists a Riemannian metric g on M such that $\xi^{\perp_g}$ is an admissible vector field. If ξ is defined on the whole manifold M, we write

$$\mathfrak{s}^\circ_\xi = [\xi^{\perp_g}] \in \operatorname{Spin}^c(M,\gamma).$$

This is independent of the choice of g since the space of metrics g for which $\xi^{\perp_g}$ is an admissible vector field is convex.

We now recall the notion of relative Spin^c structures on sutured cobordisms. If J is an almost complex structure on a 4–manifold W and H is a 3–dimensional submanifold, then there is a 2–plane field induced on H called the *field of complex tangencies* along H; see [16, Lemma 3.4].

Definition 2.12 [16, Definition 3.5] Suppose that $\mathcal{W} = (W, Z, [\xi])$ is a cobordism from the sutured manifold (M_0,γ_0) to (M_1,γ_1). We say that an almost complex structure J defined on a subset of W containing ∂Z is *admissible* if the field of complex tangencies on M_i (defined on a subset of M_i containing ∂M_i) is admissible in (M_i, γ_i) for $i \in \{0, 1\}$, and the field ξ_J of complex tangencies on Z (defined on a subset of Z containing ∂Z) is admissible in $(Z, \gamma_0 \cup \gamma_1)$.

A *relative* Spin^c *structure* on $\mathcal{W}$ is a homology class of pairs (J, P), where

- $P \subseteq \operatorname{Int}(W)$ is a finite collection of points,
- J is an admissible almost complex structure defined over $W \setminus P$,
- if ξ_J is the field of complex tangencies along Z, then $\mathfrak{s}^\circ_\xi = \mathfrak{s}^\circ_{\xi_J}$.

We say that (J, P) and (J', P') are *homologous* if there exists a compact 1–manifold $C \subseteq W \setminus \partial Z$ such that $P, P' \subseteq C$; furthermore, $J|_{W \setminus C}$ and $J'|_{W \setminus C}$ are isotopic through admissible almost complex structures. We denote by $\operatorname{Spin}^c(\mathcal{W})$ the set of relative Spin^c structures over $\mathcal{W}$.

Remark 2.13 As in the case of sutured manifolds, we will denote relative Spin^c structures on sutured cobordisms by $\mathfrak{s}^\circ$, in order to distinguish them from ordinary Spin^c structures on oriented 4–manifolds, which we denote by $\mathfrak{s}$, in analogy with the case of oriented 3–manifolds.

Remark 2.14 $\operatorname{Spin}^c(\mathcal{W})$ is an affine space over

$$\ker\big(H^2(W, \partial Z) \to H^2(Z, \partial Z)\big).$$

There are restriction maps

$$\operatorname{Spin}^c(W) \to \operatorname{Spin}^c(M_i, \gamma_i)$$

for $i \in \{0, 1\}$.

2C Sutured Floer homology

The first author [13] associated an $\mathbb{F}_2$–vector space $\operatorname{SFH}(M, \gamma)$ to each balanced sutured manifold (M, γ), called the *sutured Floer homology* of (M, γ). It splits along the relative Spin^c structures on (M, γ):

$$\operatorname{SFH}(M, \gamma) = \bigoplus_{\mathfrak{s}^\circ \in \operatorname{Spin}^c(M, \gamma)} \operatorname{SFH}(M, \gamma, \mathfrak{s}^\circ).$$

Each vector space $\operatorname{SFH}(M, \gamma, \mathfrak{s}^\circ)$ is an invariant of the sutured manifold together with the relative Spin^c structure. Sutured Floer homology is a common generalization of Heegaard Floer homology of closed oriented 3–manifolds [29] and knot Floer homology [28; 31].

The first author proved [16] that a balanced cobordism $\mathcal{W}$ from (M_0,γ_0) to (M_1,γ_1) induces a homomorphism

$$F_{\mathcal{W}}\colon \mathrm{SFH}(M_0,\gamma_0)\to\mathrm{SFH}(M_1,\gamma_1).$$

If $\mathcal{W}$ is endowed with a relative Spin^c structure $\mathfrak{s}^\circ$, then we also have a map

$$F_{\mathcal{W},\mathfrak{s}^\circ}\colon \mathrm{SFH}(M_0,\gamma_0,\mathfrak{s}^\circ|_{M_0})\to\mathrm{SFH}(M_1,\gamma_1,\mathfrak{s}^\circ|_{M_1}).$$

Let **BSut** denote the category of balanced sutured manifolds and equivalence classes of cobordisms, whereas $\mathbf{Vect}_{\mathbb{F}_2}$ denotes the category of vector spaces over $\mathbb{F}_2$.

Theorem 2.15 [16, Theorem 11.12] SFH *defines a functor* $\mathbf{BSut}\to\mathbf{Vect}_{\mathbb{F}_2}$, *which is a* $(3{+}1)$*–dimensional TQFT in the sense of* [2] *and* [4].

We conclude this section by outlining the construction of the cobordism map associated to a balanced cobordism. Let $\mathcal{W}=(W,Z,[\xi])$ be a balanced cobordism from (M_0,γ_0) to (M_1,γ_1), and suppose that every component Z_0 of Z intersects M_1 (this last hypothesis can actually be dropped; see [16, Section 10]). According to Remark 2.8, we can view $\mathcal{W}$ as the composition of a boundary cobordism $\mathcal{W}^b$ from (M_0,γ_0) to (N,γ_1) and a special cobordism $\mathcal{W}^s$ from (N,γ_1) to (M_1,γ_1). Using the *contact gluing map* defined by Honda, Kazez and Matić [11], the first author [16, Section 9] constructed a map

$$F_{\mathcal{W}^b}\colon \mathrm{SFH}(M_0,\gamma_0)\to\mathrm{SFH}(N,\gamma_1)$$

associated to the special cobordism $\mathcal{W}^b$.

The special cobordism $\mathcal{W}^s$ also induces a map: Choose a decomposition of $\mathcal{W}^s$ as $\mathcal{W}_3\circ\mathcal{W}_2\circ\mathcal{W}_1$, where $\mathcal{W}_i$ is the trace of i–handle attachments. The first author [16] defined a map $F_{\mathcal{W}_i}$ associated to each cobordism $\mathcal{W}_i$, and the map associated to $\mathcal{W}^s$ is defined as

$$F_{\mathcal{W}^s}=F_{\mathcal{W}_3}\circ F_{\mathcal{W}_2}\circ F_{\mathcal{W}_1}\colon \mathrm{SFH}(N,\gamma_1)\to\mathrm{SFH}(M_1,\gamma_1).$$

Finally, the cobordism map $F_{\mathcal{W}}$ is the composition $F_{\mathcal{W}^s}\circ F_{\mathcal{W}^b}$, which is independent of all the choices made.

All cobordism maps above admit refinements $F_{\mathcal{W},\mathfrak{s}^\circ}$ along relative Spin^c structures. The map $F_{\mathcal{W}}$ can be recovered from the maps $F_{\mathcal{W},\mathfrak{s}^\circ}$ for all Spin^c structures [16, Definition 10.9 and Proposition 10.11], and the Spin^c cobordism maps satisfy a type of composition law [16, Theorem 11.3].

3 Knot concordance maps

In [16], the first author constructed maps induced on knot Floer homology by decorated link cobordisms. We recall the necessary definitions, starting with reviewing the real blowup procedure.

Definition 3.1 Suppose that M is a smooth manifold, and let $L \subset M$ be a properly embedded submanifold. For every $p \in L$, let $N_pL = T_pM/T_pL$ be the fibre of the normal bundle of L over p, and let $UN_pL = (N_pL \setminus \{0\})/\mathbb{R}_+$ be the fibre of the unit normal bundle of L over p. Then the *(spherical) blowup* of M along L, denoted by $\mathrm{Bl}_L(M)$, is a manifold with boundary obtained from M by replacing each point $p \in L$ by UN_pL. There is a natural projection $\mathrm{Bl}_L(M) \to M$. For further details, see Arone and Kankaanrinta [1].

We now review decorated links, required to define knot Floer homology functorially. The following is [16, Definition 4.4].

Definition 3.2 A *decorated link* is a triple (Y, L, P), where L is a nonempty link in the connected oriented 3–manifold Y, and $P \subset L$ is a finite set of points. We require that for every component L_0 of L, the number $|L_0 \cap P|$ is positive and even. Furthermore, we are given a decomposition of L into compact 1–manifolds $R_+(P)$ and $R_-(P)$ such that $R_+(P) \cap R_-(P) = P$.

We can canonically assign a balanced sutured manifold $Y(L, P) = (M, \gamma)$ to every decorated link (Y, L, P), as follows. Let $M = \mathrm{Bl}_L(Y)$ and $\gamma = \bigcup_{p \in P} UN_pL$. Furthermore,

$$R_\pm(\gamma) := \bigcup_{x \in R_\pm(P)} UN_xL,$$

oriented as $\pm\partial M$, and we orient γ as $\partial R_+(\gamma)$.

The following is [16, Definiton 4.2].

Definition 3.3 A *surface with divides* (S, σ) is a compact orientable surface S, possibly with boundary, together with a properly embedded 1–manifold σ that divides S into two compact subsurfaces that meet along σ.

We are now ready to define decorated link cobordisms. The following is [16, Definition 4.5].

Definition 3.4 We say that the triple $\mathcal{X} = (X, F, \sigma)$ is a *decorated link cobordism* from (Y_0, L_0, P_0) to (Y_1, L_1, P_1) if

(1) (X, F) is a link cobordism from (Y_0, L_0) to (Y_1, L_1),

(2) (F, σ) is a surface with divides such that the map

$$\pi_0(\partial\sigma) \to \pi_0((L_0 \setminus P_0) \cup (L_1 \setminus P_1))$$

is a bijection,

(3) we can orient each component R of $F \setminus \sigma$ such that whenever $\partial\overline{R}$ crosses a point of P_0, it goes from $R_+(P_0)$ to $R_-(P_0)$, and whenever it crosses a point of P_1, it goes from $R_-(P_1)$ to $R_+(P_1)$,

(4) if F_0 is a closed component of F, then $\sigma \cap F_0 \neq \varnothing$.

Finally, we recall how to associate a sutured manifold cobordism complementary to a decorated link cobordism. For this purpose, we first discuss S^1–invariant contact structures on circle bundles; see also [16, Section 4]. Let $\pi\colon M \to F$ be a principal circle bundle over a compact oriented surface F. An S^1–invariant contact structure ξ on M determines a diving set σ on the base F, by requiring that $x \in \sigma$ if and only if ξ is tangent to $\pi^{-1}(x)$, and a splitting of F as $R_+(\sigma) \cup R_-(\sigma)$. The image of any local section of π is a convex surface with dividing set projecting onto σ. According to Lutz [21] and Honda [10, Theorem 2.11 and Section 4], given a dividing set σ on F that intersects each component of F nontrivially and divides F into subsurfaces $R_+(\sigma)$ and $R_-(\sigma)$, there is a unique S^1–invariant contact structure ξ_σ on M, up to isotopy, such that the dividing set associated to ξ_σ is exactly σ, the coorientation of ξ_σ induces the splitting $R_\pm(\sigma)$, and the boundary ∂M is a convex.

The following is [16, Definition 4.9].

Definition 3.5 Let (X, F, σ) be a decorated link cobordism from the decorated link (Y_0, L_0, P_0) to (Y_1, L_1, P_1). We define the sutured cobordism $\mathcal{W} = \mathcal{W}(X, F, \sigma)$ as follows. Choose an arbitrary splitting of F into $R_+(\sigma)$ and $R_-(\sigma)$ such that $R_+(\sigma) \cap R_-(\sigma) = \sigma$, and orient F such that $\partial R_+(\sigma)$ (with $R_+(\sigma)$ oriented as a subsurface of F) crosses P_0 from $R_+(P_0)$ to $R_-(P_0)$ and P_1 from $R_-(P_1)$ to $R_+(P_1)$. Then $\mathcal{W}$ is defined to be the triple $(W, Z, [\xi])$, where $W = \mathrm{Bl}_F(X)$ and $Z = UNF$, oriented as a submanifold of ∂W, finally $\xi = \xi_\sigma$ is an S^1–invariant contact structure with dividing set σ on F and convex boundary ∂Z with dividing set projecting to $P_0 \cup P_1$.

The contact vector fields with respect to which a local section of $UNF \to F$ and ∂Z are transverse are different, so they can project to different subsets of $L_0 \cup L_1$. Specifically, the dividing set for ∂Z projects to $P_0 \cup P_1$, while $\partial\sigma$ is disjoint from $P_0 \cup P_1$.

Notice that if F does not have any closed component, then it deformation retracts onto a 1–dimensional CW complex, and therefore any S^1–bundle on it has a section, hence is trivial if the bundle is orientable. In particular, $UNF \approx F \times S^1$.

In the present paper, we only consider decorated links (Y, L, P) where $Y = S^3$, the link L has a single component, and $|P| = 2$. Hence, we drop Y from the notation and only write (K, P) for such a decorated knot.

Definition 3.6 A *decorated concordance* is a decorated link cobordism (X, F, σ) such that

(1) X is an integer homology $S^3 \times I$ with boundary $(-S^3) \sqcup S^3$,

(2) the surface F is an annulus, and

(3) σ consists of two arcs connecting the two components of ∂F.

If $X = S^3 \times I$, we drop X from the notation and only write (F, σ).

Lemma 3.7 *Let X be an oriented cobordism from S^3 to S^3. Then X has the same homology and cohomology as $S^3 \times I$ if and only if $H_1(X) = H_2(X) = 0$.*

Proof The "only if" part is obvious. So suppose that $H_1(X) = H_2(X) = 0$. Then let $\overline{X}$ be the closed 4–manifold obtained by gluing two 4–balls to ∂X. We denote by $B \subset X$ the union of these 4–balls. Then, for $i \in \{1, 2\}$, we have

$$0 = H_i(X) \cong H^{4-i}(X, \partial X) \cong H^{4-i}(\overline{X}, B) \cong H^{4-i}(\overline{X}).$$

Here, the first isomorphism follows from Poincaré–Lefschetz duality, the second from excision, and the third from the cohomological long exact sequence of the pair $(\overline{X}, B)$. So $H^2(\overline{X}) = H^3(\overline{X}) = 0$, hence

$$H_1(\overline{X}) \cong H^3(\overline{X}) = 0 \quad \text{and} \quad H^1(\overline{X}) = \mathrm{Hom}(H_1(\overline{X}), \mathbb{Z}) = 0.$$

As $\overline{X}$ has the same integral cohomology is S^4, after removing two balls, X has the same integral homology and cohomology as $S^3 \times I$. □

It follows from [16, Proposition 4.10] that a decorated concordance $\mathcal{C} = (X, F, \sigma)$ from (K_0, P_0) to (K_1, P_1) induces a homomorphism

$$F_{\mathcal{C}} \colon \widehat{\mathrm{HFK}}(K_0, P_0) \to \widehat{\mathrm{HFK}}(K_1, P_1),$$

where $\widehat{\mathrm{HFK}}(K_i, P_i)$ are the natural knot Floer homology groups defined in [17]. Indeed, $\mathcal{W} = \mathcal{W}(X, F, \sigma)$ is a cobordism from the sutured manifold $S^3(K_0, P_0)$ to $S^3(K_1, P_1)$, and hence induces a homomorphism

$$F_{\mathcal{W}} \colon \mathrm{SFH}(S^3(K_0, P_0)) \to \mathrm{SFH}(S^3(K_1, P_1)).$$

But $\mathrm{SFH}(S^3(K_0, P_0)) \cong \widehat{\mathrm{HFK}}(K_0, P_0)$ and $\mathrm{SFH}(S^3(K_1, P_1)) \cong \widehat{\mathrm{HFK}}(K_1, P_1)$ tautologically. This assignment is functorial under composition of link cobordisms.

3A Relative Spin^c structures and knot concordances

In the case of knot concordances, the relative Spin^c structures behave nicely, as explained in this section.

Lemma 3.8 *Suppose $\mathcal{C} = (X, F, \sigma)$ is a decorated concordance from (K_0, P_0) to (K_1, P_1). If $(M_i, \gamma_i) = S^3(K_i, P_i)$ is the balanced sutured manifold complementary to (K_i, P_i) for $i \in \{0, 1\}$, and $\mathcal{W} = \mathcal{W}(\mathcal{C}) = (W, Z, [\xi])$ is the sutured manifold cobordism from (M_0, γ_0) to (M_1, γ_1) complementary to $\mathcal{C}$, then*

$$F_{\mathcal{W}} = \bigoplus_{\mathfrak{s}^\circ \in \mathrm{Spin}^c(\mathcal{W})} F_{\mathcal{W}, \mathfrak{s}^\circ}. \tag{3-1}$$

Furthermore, $\mathrm{Spin}^c(\mathcal{W})$ *is an affine space over $H^2(W, Z) \cong \mathbb{Z}$, and the restriction maps*

$$r_i \colon \mathrm{Spin}^c(\mathcal{W}) \to \mathrm{Spin}^c(M_i, \gamma_i)$$

are isomorphisms for $i \in \{0, 1\}$.

Proof As in Remark 2.8, we write $\mathcal{W} = \mathcal{W}^s \circ \mathcal{W}^b$, where $\mathcal{W}^b$ is a boundary cobordism from (M_0, γ_0) to (N, γ_1), where $N = M_0 \cup (-Z)$, and $\mathcal{W}^s$ is a special cobordism from (N, γ_1) to (M_1, γ_1). As Z is a product, N is diffeomorphic to the knot complement $M_0 \approx S^3 \setminus N(K_0)$, and hence $H_2(N) = 0$. So, by [16, Remark 10.10] and [16, Proposition 10.11],

$$F_{\mathcal{W}} = \bigoplus_{\mathfrak{s}^\circ \in \mathrm{Spin}^c(\mathcal{W})} F_{\mathcal{W}, \mathfrak{s}^\circ}.$$

As $H^k(Z, \partial M_1) = 0$ for $k \in \{1, 2\}$, we can apply [16, Lemma 3.7] to conclude that

$$\mathrm{Spin}^c(\mathcal{W}) \cong H^2(W, \partial M_1).$$

Of course, $H^2(W, \partial M_1) \cong H^2(W, \partial M_0) \cong H^2(W, Z)$. By excision, we have that $H^2(W, Z) \cong H^2(X, N(F))$, where $N(F)$ is a regular neighbourhood of F. From the long exact sequence of the pair $(X, N(F))$ and the fact that $H^1(X) = H^2(X) = 0$, and since $H^1(N(F)) \cong H^1(S^1) \cong \mathbb{Z}$, we obtain that $H^2(X, N(F)) \cong \mathbb{Z}$.

The restriction maps

$$r_i \colon \mathrm{Spin}^c(\mathcal{W}) \to \mathrm{Spin}^c(M_i, \gamma_i)$$

for $i \in \{0,1\}$ are modelled on the restriction maps $H^2(W, \partial M_i) \to H^2(M_i, \partial M_i)$ for $i \in \{0,1\}$. From the long exact sequence of the triple $(W, M_i, \partial M_i)$, the sequence

$$H^2(W, M_i) \to H^2(W, \partial M_i) \to H^2(M_i, \partial M_i) \to H^3(W, M_i) \tag{3-2}$$

is exact. Now consider the relative Mayer–Vietoris sequence of the pairs (W, M_i) and $(N(F), N(K_i))$, whose union is $(X, \partial_i X)$, where $\partial_i X \approx S^3$ is the ingoing boundary component of X when $i = 0$ and is the outgoing boundary component when $i = 1$:

$$H^k(X, \partial_i X) \to H^k(W, M_i) \oplus H^k(N(F), N(K_i)) \to H^k(Z, \partial M_i).$$

Here, $H^k(X, \partial_i X) \cong H^k(S^3 \times I, S^3 \times \{0\}) = 0$, and the last term is zero as Z deformation retracts onto ∂M_i. Consequently, $H^k(W, M_i) = 0$ for every k, and by the exact sequence (3-2), this means that the restriction maps r_i are isomorphisms for $i \in \{0,1\}$. □

In the following lemma, v_0 denotes any fixed vector field on a balanced sutured manifold (M, γ) obtained by restricting an admissible vector field to ∂M; see Definition 2.9 and Remark 2.10.

Lemma 3.9 *Let $\mathcal{C} = (X, F, \sigma)$ be a knot concordance from (K_0, P_0) to (K_1, P_1). As in Lemma 3.8, let $(M_i, \gamma_i) = S^3(K_i, P_i)$ for $i \in \{0,1\}$, and let*

$$\mathcal{W} = \mathcal{W}(\mathcal{C}) = (W, Z, [\xi]).$$

For $i \in \{0,1\}$, let S_i be a Seifert surface for K_i, and let t_i be the trivialization of $v_0^\perp$ given by a vector field tangent to ∂M_i in the meridional direction. Then, for any relative Spin^c structure $\mathfrak{s}^\circ \in \mathrm{Spin}^c(\mathcal{W})$,

$$\langle c_1(r_0(\mathfrak{s}^\circ), t_0), [S_0] \rangle = \langle c_1(r_1(\mathfrak{s}^\circ), t_1), [S_1] \rangle, \tag{3-3}$$

where r_0 and r_1 are the restriction maps in Lemma 3.8.

From Lemma 3.9, we can already deduce the following proposition, which can be seen as a first step towards the proof of Theorem 1.2.

Proposition 3.10 *If $\mathcal{C}$ is a decorated concordance between two knots (K_0, P_0) and (K_1, P_1), then the map induced between the knot Floer homologies preserves the Alexander grading; that is,*

$$F_{\mathcal{C}}(\widehat{\mathrm{HFK}}(K_0, P_0, i)) \le \widehat{\mathrm{HFK}}(K_1, P_1, i)$$

for every $i \in \mathbb{Z}$.

Proof We use the same notation as in Lemmas 3.8 and 3.9. It follows from Lemma 3.8 that the map $F_{\mathcal{C}} = F_{\mathcal{W}}$ splits as the sum of the maps $F_{\mathcal{W},\mathfrak{s}^\circ}$ for $\mathfrak{s}^\circ \in \mathrm{Spin}^c(\mathcal{W})$; see (3-1). It is therefore sufficient to check that, for every relative Spin^c structure $\mathfrak{s}^\circ \in \mathrm{Spin}^c(\mathcal{W})$, the map

$$F_{\mathcal{W},\mathfrak{s}^\circ}\colon \mathrm{SFH}(M_0,\gamma_0,\mathfrak{s}^\circ|_{M_0}) \to \mathrm{SFH}(M_1,\gamma_1,\mathfrak{s}^\circ|_{M_1})$$

preserves the Alexander grading.

According to the proof of [14, Theorem 1.5] on page 333, if t_i is the trivialization of $v_0^\perp$ given by a vector field tangent to ∂M_i in the meridional direction, then

$$\mathrm{SFH}(M_i,\gamma_i,\mathfrak{s}^\circ) = \widehat{\mathrm{HFK}}\big(K_i, P_i, -\tfrac{1}{2}\langle c_1(\mathfrak{s}^\circ,t_i),[S_i]\rangle\big),$$

where S_i is a Seifert surface of K_i for $i \in \{0,1\}$. The result now follows from Lemma 3.9, which states that

$$\langle c_1(\mathfrak{s}^\circ|_{M_0},t_0),[S_0]\rangle = \langle c_1(\mathfrak{s}^\circ|_{M_1},t_1),[S_1]\rangle. \qquad \square$$

Proof of Lemma 3.9 Choose an admissible almost complex structure J on $W \setminus P$ whose homology class is $\mathfrak{s}^\circ$, where $P \subset \mathrm{Int}(W)$ is a finite set of points, as in Definition 2.12. Let ξ_J be the field of complex tangencies of J along Z. Then, by definition, $\mathfrak{s}^\circ_\xi = \mathfrak{s}^\circ_{\xi_J}$. In fact, we can choose J such that $\xi_J = \xi$. Choose a trivialization of the normal S^1–bundle of F whose total space is Z. If we identify F with $S^1 \times I$ such that σ maps to $P_0 \times I$ for $P_0 = \sigma \cap K_0$, then this identification, together with the above trivialization, induces a diffeomorphism $d\colon Z \to S^1 \times S^1 \times I$, where the first factor is the fibre direction, and such that ξ is mapped to an I–invariant contact structure with dividing set $S^1 \times P_0 \times \{a\}$ on $S^1 \times S^1 \times \{a\}$ for every $a \in I$, and $\{\theta\} \times P_0 \times I$ on $\{\theta\} \times S^1 \times I$ for every $\theta \in S^1$. Hence, we can perturb the 2–plane field ξ such that it is always tangent to the second S^1 factor, ie the longitudinal direction. So we can choose J such that ξ_J is also invariant in the σ direction, and it contains the longitude direction. If v is a nowhere zero section of ξ_J tangent to the longitude direction, then — under a homotopy of $\xi_J|_{\partial M_i}$ to $v_0^\perp$ through admissible 2–plane fields — the vector field $v|_{\partial M_0}$ represents a trivialization τ_0 that corresponds to t_0 and $v|_{\partial M_1}$ represents a trivialization τ_1 that corresponds to t_1.

The 2–plane field ξ_J, together with the trivialization given by v, gives a complex 1–dimensional subbundle of $(TW|_Z, J)$ together with a trivialization. The complement of ξ_J is also trivial, canonically trivialized by its intersection with TZ, which then gives rise to a trivialization τ of $TW|_Z$. As J is defined over the 3–skeleton of W, it makes sense to talk about the relative Chern class $c_1(TW, J, \tau) \in H^2(W, Z)$. If ξ_J^i denotes the field of complex tangencies of J along M_i, then the complement of ξ_J^i is a trivial bundle (trivialized by its intersection with TM_i), so

$$c_1(\xi_J^i, \tau_i) = c_1(TW|_{M_i}, J, \tau) = c_1(TW, J, \tau)|_{M_i},$$

where the second equality follows from the naturality of Chern classes. By construction, ξ_J^i represents $\mathfrak{s}_i^\circ$.

Recall that S_i is a Seifert surface of K_i for $i \in \{0,1\}$. Note that $H_2(W,Z) \cong \mathbb{Z}$, and that there is a bilinear intersection pairing

$$H_2(W,Z) \otimes H_2(W, M_0 \cup M_1) \to \mathbb{Z}.$$

Consider the cycle $m = S^1 \times \{\mathrm{pt}\} \times I$ in $C_2(W, M_0 \cup M_1)$. As both S_0 and S_1 intersect m once positively, they both represent the generator of $H_2(W,Z) \cong \mathbb{Z}$. Hence

$$\langle c_1(\mathfrak{s}_0^\circ, \tau_0), [S_0] \rangle = \langle c_1(\mathcal{W}, J, \tau), [S_0] \rangle = \langle c_1(TW, J, \tau), [S_1] \rangle = \langle c_1(\mathfrak{s}_1^\circ, \tau_1), [S_1] \rangle,$$

and (3-3) follows as we saw that τ_0 corresponds to t_0 and τ_1 corresponds to t_1. □

As a consequence of Proposition 3.10, we can prove Theorem 1.6.

Proof of Theorem 1.6 Suppose that F is an invertible concordance from K_0 to K_1. Choose an arbitrary pair of points P_0 on K_0 and P_1 on K_1, making them into decorated knots, and an arbitrary pair of arcs σ on F making F into a decorated concordance from (K_0, P_0) to (K_1, P_1). Let F' be the inverse of F, and choose a decoration σ' on it such that (F', σ') is a decorated concordance from (K_1, P_1) to (K_0, P_0). As the composition of F and F' is equivalent to the trivial cobordism $K_0 \times I$ from K_0 to K_0, we can choose σ' such that the composition of $\mathcal{C} = (F, \sigma)$ and $\mathcal{C}' = (F', \sigma')$ is equivalent to the product decorated cobordism $(K_0 \times I, P \times I)$, where $P = \sigma \cap K_0$ is a pair of points. By the functoriality of $F_{\mathcal{C}}$ and the fact that a product cobordism induces the identity map,

$$F_{\mathcal{C}'} \circ F_{\mathcal{C}} = \mathrm{Id}_{\widehat{\mathrm{HFK}}(K_0, P_0)},$$

and so $F_{\mathcal{C}}$ is injective. We shall see in Section 6 that $F_{\mathcal{C}}$ preserves the homological grading. Hence Proposition 3.10 implies that

$$\dim \widehat{\mathrm{HFK}}_j(K_0, P_0, i) \le \dim \widehat{\mathrm{HFK}}_j(K_1, P_1, i)$$

for every $i, j \in \mathbb{Z}$. Up to isomorphism, $\widehat{\mathrm{HFK}}_j(K_i, P_i)$ is independent of the choice of P_i, and the result follows. □

We shall see in Section 6 that the concordance maps also preserve the homological grading. Then we have the following.

Lemma 3.11 *Suppose $\mathcal{C} = (X, F, \sigma)$ is a decorated concordance from (K_0, P_0) to (K_1, P_1). If K_0 is the unknot U, then the element*

$$F_{\mathcal{C}}(1) \in \widehat{\mathrm{HFK}}_0(K_1, P_1, 0)$$

is independent of the decorations σ and P_0, where $1 \in \widehat{\mathrm{HFK}}(K_0, P_0) \cong \mathbb{Z}_2$.

Proof Suppose that σ' is another decoration with the same endpoints as σ, let $\mathcal{C}' = (X, F, \sigma')$, and define

$$k = [\sigma' - \sigma] \in H_1(F) \cong \mathbb{Z}.$$

Consider the decorated concordance $\mathcal{C}_k = (S^3 \times I, U \times I, \sigma_k)$, where σ_k spirals around k times. Then $\mathcal{C}' = \mathcal{C} \circ \mathcal{C}_k$. As $\widehat{\mathrm{HFK}}(U) \cong \mathbb{Z}_2$, we have $F_{\mathcal{C}_k} = \mathrm{Id}_{\mathbb{Z}_2}$. By the functoriality of the knot concordance maps, we obtain that $F_{\mathcal{C}'} = F_{\mathcal{C}}$. Since $\widehat{\mathrm{HFK}}(U) \cong \mathbb{Z}_2$ has no nontrivial automorphisms, it does not matter how we choose the markings P_0. □

4 Filtered complexes and spectral sequences

In this section, we briefly recall the definitions and properties of spectral sequences that we need. We mainly refer to the book of McCleary [24]. The spectral sequences we are interested in arise from filtered chain complexes, so we focus on this case only.

Definition 4.1 A *filtered chain complex* is a chain complex $\left(C = \bigoplus_{k \in \mathbb{Z}} C_k, \partial\right)$, such that $\partial C_k \subseteq C_{k-1}$, with a nested sequence of subcomplexes

$$\cdots \subseteq \mathcal{F}_{p-1}C \subseteq \mathcal{F}_p C \subseteq \mathcal{F}_{p+1}C \subseteq \cdots$$

such that $\bigcup_{p \in \mathbb{Z}} \mathcal{F}_p C = C$ and $\partial(\mathcal{F}_p C) \subseteq \mathcal{F}_p C$.

We say that the filtered chain complex is *bounded* if there are integers $a \le b$ such that

$$\{0\} = \mathcal{F}_a C \subseteq \cdots \subseteq \mathcal{F}_b C = C.$$

We obtain a spectral sequence from a filtered chain complex as follows; see [24, Proof of Theorem 2.6].

Definition 4.2 For $p, q, r \in \mathbb{Z}$, we define

$$\begin{aligned}
Z^r_{p,q} &= \mathcal{F}_p C_{p+q} \cap \partial^{-1}(\mathcal{F}_{p-r} C_{p+q-1}),\\
B^r_{p,q} &= \mathcal{F}_p C_{p+q} \cap \partial(\mathcal{F}_{p+r} C_{p+q+1}),\\
Z^\infty_{p,q} &= \mathcal{F}_p C_{p+q} \cap \ker \partial,\\
B^\infty_{p,q} &= \mathcal{F}_p C_{p+q} \cap \operatorname{im} \partial.
\end{aligned}$$

For $0 \leq r \leq \infty$, the *r–page* (or *r–term*) is the complex $\bigl(E^r = \bigoplus_{p,q\in\mathbb{Z}} E^r_{p,q}, \partial^r\bigr)$, where

$$E^r_{p,q} = \frac{Z^r_{p,q}}{Z^{r-1}_{p-1,q+1} + B^{r-1}_{p,q}},$$

and the differential

$$\partial^r\colon E^r_{p,q} \to E^r_{p-r,q+r-1}$$

is induced by the differential ∂ on the complex C.

Sometimes we only focus on the p grading. In such cases, we drop q from the notation, and write $E^r_p = \bigoplus_{q\in\mathbb{Z}} E^r_{p,q}$. For the following, see [24, Proof of Theorem 2.6].

Theorem 4.3 *The pages $\{(E^r, \partial^r)\}$ induced by a filtered chain complex form a spectral sequence in the sense of [24, Definition 2.2]; ie*

$$E^{r+1}_{p,q} = H_{p,q}(E^r_{*,*}, \partial^r) := \frac{\ker(\partial^r|_{E^r_{p,q}})}{\operatorname{im}(\partial^r|_{E^r_{p+r,q-r+1}})}.$$

If the filtration is bounded, then there is a canonical isomorphism

$$E^\infty_{p,q} \cong \frac{\mathcal{F}_p(H_{p+q}(C))}{\mathcal{F}_{p-1}(H_{p+q}(C))},$$

where the filtration on the total homology $H(C) = \bigoplus_{k\in\mathbb{Z}} H_k(C)$ is the one induced from C:

$$\mathcal{F}_p(H(C)) := \operatorname{im}\bigl(H(\mathcal{F}_pC, \partial|_{\mathcal{F}_pC}) \to H(C,\partial)\bigr).$$

Remark 4.4 Notice that $E^0_{p,q}$ is the graded module

$$\frac{\mathcal{F}_pC_{p+q}}{\mathcal{F}_{p-1}C_{p+q}}$$

associated with the filtration. The page $E^1_{p,q}$ is the homology $H_q(E^0_{p,*}, \partial^0)$ of the associated graded module with the induced differential.

4A Morphisms of spectral sequences

According to McCleary [24], we have the following.

Definition 4.5 Let (E^r, ∂^r) and $(\bar{E}^r, \bar{\partial}^r)$ be spectral sequences. A *morphism of spectral sequences* is a sequence of module homomorphisms $f^r\colon E^r_{*,*} \to \bar{E}^r_{*,*}$ for $r \in \mathbb{N}$, of bidegree $(0,0)$, such that f^r commutes with the differentials; that is, $f^r \circ \partial^r = \bar{\partial}^r \circ f^r$, and each f^{r+1} is induced by f^r on homology; ie f^{r+1} is the composite

$$f^{r+1}\colon E^{r+1}_{*,*} \cong H(E^r_{*,*}, \partial^r) \xrightarrow{H(f^r)} H(\bar{E}^r_{*,*}, \bar{\partial}^r) \cong \bar{E}^{r+1}_{*,*}.$$

Remark 4.6 Let $f\colon C \to \bar{C}$ be a map of filtered complexes of homological degree zero; ie

- $f(C_k) \subseteq \bar{C}_k$,
- $f \circ \partial = \bar{\partial} \circ f$,
- $f(\mathcal{F}_p C) \subseteq \mathcal{F}_p \bar{C}$.

Then f induces a morphism between the spectral sequences associated to C and $\bar{C}$.

Remark 4.7 If (E^r, ∂^r) and $(\bar{E}^r, \bar{\partial}^r)$ are bounded spectral sequences, $\{f^r\colon E^r \to \bar{E}^r\}$ is a morphism of spectral sequences, and f^∞ is nonzero on $E^\infty_{p,q}$, then f^r is nonzero on $E^r_{p,q}$ for all $r \in \mathbb{N}$.

4B The τ invariant

In this subsection, we recall the definition and few properties of the Ozsváth–Szabó τ invariant, and we discuss it in a slightly more general setting.

Definition 4.8 If C is a nonacyclic bounded filtered complex over $\mathbb{F}_2$, we define

$$\tau(C) := \min\{p \in \mathbb{Z} : H(\mathcal{F}_p C) \to H(C) \text{ is nontrivial}\}.$$

Definition 4.8 generalizes the Ozsváth–Szabó τ invariant in the sense that, if $C = \widehat{\mathrm{CF}}(\mathcal{H})$ for some Heegaard diagram for a decorated knot (K, P), then $\tau(C) = \tau(K)$.

Remark 4.9 An alternative definition of $\tau(C)$ is given by the following property:

$$E^\infty_p(C) \begin{cases} = 0 & \text{if } p < \tau(C), \\ \neq 0 & \text{if } p = \tau(C). \end{cases}$$

Furthermore, if the total homology $H(C) = \mathbb{F}_2$, then

$$E^\infty_p(C) \begin{cases} = 0 & \text{if } p \neq \tau(C), \\ \neq 0 & \text{if } p = \tau(C). \end{cases}$$

We conclude the section with a technical lemma that we will use to prove that a decorated concordance induces a nontrivial map between the E^∞ pages of the spectral sequences arising from the knot filtrations.

Lemma 4.10 *Let $f\colon C \to \overline{C}$ be a filtered map of degree zero between nonacyclic bounded filtered complexes over $\mathbb{F}_2$ such that*

(1) $H(C) \cong \mathbb{F}_2$ *and* $H(\overline{C}) \cong \mathbb{F}_2$,

(2) $\tau(C) = \tau(\overline{C})$, *and*

(3) $H(f)\colon H(C) \to H(\overline{C})$ *is an isomorphism.*

Then $E_\tau^\infty(C) \cong \mathbb{F}_2$ and $E_\tau^\infty(\overline{C}) \cong \mathbb{F}_2$, and the map $f^\infty\colon E_\tau^\infty(C) \to E_\tau^\infty(\overline{C})$ is also an isomorphism.

Proof Since (1) and (2) hold, by Theorem 4.3 and Definition 4.8, there are canonical isomorphisms

$$E_\tau^\infty(C) \cong H(C) \cong \mathbb{F}_2 \quad \text{and} \quad E_\tau^\infty(\overline{C}) \cong H(\overline{C}) \cong \mathbb{F}_2.$$

The commutativity of the following diagram concludes the proof:

$$\begin{tikzcd}
E_\tau^\infty(C) \arrow[r, "f^\infty"] \arrow[d, "\wr"] & E_\tau^\infty(\overline{C}) \arrow[d, "\wr"] \\
H(C) \arrow[r, "\simeq", "H(f)"'] & H(\overline{C})
\end{tikzcd}$$

□

5 Concordance maps preserve the knot filtration

5A The knot filtration

Let K be a null-homologous knot in a closed oriented 3–manifold Y. Ozsváth and Szabó [28], and independently Rasmussen [31], proved that K gives rise to a filtration of the Heegaard Floer chain complex $\widehat{\mathrm{CF}}(Y)$, well-defined up to filtered chain homotopy equivalence, called the knot filtration. Such a filtration can be defined in terms of the Alexander grading; see also [28, Section 2.3].

Definition 5.1 Let S be a Seifert surface for the knot K, and let $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, w, z)$ be a doubly pointed Heegaard diagram for K, as defined by Ozsváth and Szabó [28]. Given a generator $\boldsymbol{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$, its S–Alexander grading is

$$\mathcal{A}_S(\boldsymbol{x}) = \tfrac{1}{2}\langle c_1(\underline{\mathfrak{s}}(\boldsymbol{x})), [S]\rangle,$$

where $\underline{\mathfrak{s}}(\boldsymbol{x})$ is the Spin^c structure on $Y_0(K)$ extending $\mathfrak{s}(\boldsymbol{x}) \in \mathrm{Spin}^c(Y)$. We denote the corresponding filtration by $\mathcal{F}_S$.

Remark 5.2 Consider the sutured manifold $Y(K) = (M, \gamma)$ complementary to K. As in the proof of [14, Theorem 1.5] on page 333, let t be the trivialization of $v_0^\perp$ given by a vector field tangent to ∂M in the meridional direction. Then

$$\mathcal{A}_S(\boldsymbol{x}) = \tfrac{1}{2}\langle c_1(\mathfrak{s}^\circ(\boldsymbol{x}), t), [S]\rangle,$$

where $\mathfrak{s}^\circ(\boldsymbol{x})$ now denotes an element of $\mathrm{Spin}^c(M, \gamma)$.

If Y is a rational homology 3–sphere, all Seifert surfaces of K are homologous in the knot exterior, so the Alexander grading does not depend on S, and we simply denote it by $\mathcal{A}(\boldsymbol{x})$, and the filtration by $\mathcal{F}(\boldsymbol{x})$.

The following lemma describes how the relative Alexander grading can be read off the Heegaard diagram; see [28, Lemma 2.5] and [31, page 25].

Lemma 5.3 *Let $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, w, z)$ be a Heegaard diagram for a null-homologous knot K in a 3–manifold Y, and let S be a Seifert surface for K. If $\phi \in \pi_2(\boldsymbol{x}, \boldsymbol{y})$, then*

$$n_z(\phi) - n_w(\phi) = \mathcal{A}_S(\boldsymbol{x}) - \mathcal{A}_S(\boldsymbol{y}).$$

5B Knot filtration and concordances

Our aim is to prove that the knot filtration is preserved by the chain maps induced by concordances.

Theorem 5.4 *Let $\mathcal{C}$ be a decorated concordance from (K_0, P_0) to (K_1, P_1), and let $(\Sigma_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, w_i, z_i)$ be a doubly pointed diagram representing (K_i, P_i) for $i \in \{0, 1\}$. Then there is a chain map*

$$f_{\mathcal{C}}\colon \widehat{\mathrm{CF}}(\Sigma_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, w_0, z_0) \to \widehat{\mathrm{CF}}(\Sigma_1, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, w_1, z_1)$$

preserving the knot filtration; ie for every generator $\boldsymbol{x} \in \mathbb{T}_{\alpha_0} \cap \mathbb{T}_{\beta_0}$,

$$\mathcal{A}(f_{\mathcal{C}}(\boldsymbol{x})) \le \mathcal{A}(\boldsymbol{x}),$$

such that $f_{\mathcal{C}}$ induces the identity of $\widehat{\mathrm{HF}}(S^3)$ on the total homology, and $F_{\mathcal{C}}$ on the homology of the associated graded complexes.

Theorem 5.4 yields a morphism of spectral sequences in the sense of Definition 4.5, hence we have the following corollary.

Theorem 5.5 *Suppose that $\mathcal{C}$ is a decorated concordance from (K_0, P_0) to (K_1, P_1). Then there is a morphism of spectral sequences from $\widehat{\mathrm{HFK}}(K_0, P_0) \Longrightarrow \widehat{\mathrm{HF}}(S^3)$ to $\widehat{\mathrm{HFK}}(K_1, P_1) \Longrightarrow \widehat{\mathrm{HF}}(S^3)$ such that the map induced on the E^1 page is $F_{\mathcal{C}}$, and the map induced on the E^∞ page is $\mathrm{Id}_{\widehat{\mathrm{HF}}(S^3)}$.*

Proof Suppose that $\mathcal{C} = (X, F, \sigma)$. Since $H_1(X) = H_2(X) = 0$, it follows from the work of Ozsváth and Szabó [26, Theorem 1.1] that $\tau(K_0) = \tau(K_1)$. Indeed, the knot $K = K_0 \# \overline{K_1}$ bounds a disk in a homology 4–ball W with boundary S^3, and hence $\tau(K) = \tau(K_0) - \tau(K_1) = 0$ by [26, Theorem 1.1]. By Theorem 5.4, we have a filtered map $f_{\mathcal{C}}$ that induces an isomorphism on the total homology. We can therefore apply Lemma 4.10 to conclude that the map induced on the E^∞ page is also an isomorphism. □

Definition 5.6 We say that an element $x \in \widehat{\mathrm{HFK}}(K, P)$ *survives* the spectral sequence to $\widehat{\mathrm{HF}}(S^3) \cong \mathbb{Z}_2$ if there is a sequence of cycles $x_i \in E^i$ for $i \geq 1$ such that $x_1 = x$ and $x_{i+1} = [x_i]$; we denote the set of such elements by $A(K)$. Furthermore, we have a partition $A(K) = A_0(K) \cup A_1(K)$, where $A_j(K)$ consists of those elements for which $x_i = j \in \mathbb{Z}_2$ for i sufficiently large (note that the spectral sequence is bounded).

The subset $A_0(K)$ is a linear subspace of $A(K)$, and $A_1(K)$ is an affine translate of $A_0(K)$. Each of the sets $A(K)$, $A_0(K)$ and $A_1(K)$ is a knot invariant.

It follows from the definition of the Ozsváth–Szabó τ invariant [26] that

$$A_1(K) \cap \widehat{\mathrm{HFK}}(K, i) \begin{cases} = \varnothing & \text{if } i \neq \tau(K), \\ \neq \varnothing & \text{if } i = \tau(K). \end{cases} \tag{5-1}$$

If $a \in A_1(K)$, let a_0 denote the homogeneous component of a in homological grading zero. It is straightforward to check that a_0 survives the spectral sequence. Since the homological grading on $\widehat{\mathrm{CFK}}$ is inherited from the one on $\widehat{\mathrm{CF}}$, and since the homological grading of $1 \in \widehat{\mathrm{HF}}(S^3)$ is zero, it follows that $a_0 \in A_1(K)$. Combined with (5-1), this implies that

$$A_1'(K) := A_1(K) \cap \widehat{\mathrm{HFK}}_0(K, \tau(K)) \neq \varnothing. \tag{5-2}$$

Notice that $A_1'(K)$ is also a knot invariant.

The following result is a straightforward consequence of Theorem 5.5, Proposition 3.10 and (5-2), and implies Corollary 1.3 of the introduction.

Corollary 5.7 *Suppose $\mathcal{C} = (X, F, \sigma)$ is a decorated concordance from (K_0, P_0) to (K_1, P_1), and let $\tau = \tau(K_0) = \tau(K_1)$. Then, for $j \in \{0, 1\}$,*

$$F_{\mathcal{C}}(A_j(K_0)) \subseteq A_j(K_1)$$

and hence it is nonzero from $\widehat{\mathrm{HFK}}_0(K_0, P_0, \tau)$ to $\widehat{\mathrm{HFK}}_0(K_1, P_1, \tau)$.

Proof The fact that $F_{\mathcal{C}}(A_j(K_0)) \subseteq A_j(K_1)$ follows from Theorem 5.5. In Section 6, we shall see that $F_{\mathcal{C}}$ preserves the homological grading. Then, by Proposition 3.10, $F_{\mathcal{C}}$ maps $\widehat{\mathrm{HFK}}_0(K_0, P_0, \tau)$ to $\widehat{\mathrm{HFK}}_0(K_1, P_1, \tau)$. So we only need to prove that this map is nonzero.

By (5-2), we have $A'_1(K_0) \neq \varnothing$; let $x \in A'_1(K_0)$. Then, by the previous paragraph,

$$F_{\mathcal{C}}(x) \in A_1(K_1) \cap \widehat{\mathrm{HFK}}_0(K_1, \tau) = A'_1(K_1),$$

hence $F_{\mathcal{C}}(x) \neq 0$. □

We now turn to the proof of Theorem 5.4, which will take the rest of this section.

5C Triviality of the gluing map

Given a sutured manifold cobordism $\mathcal{W} = (W, Z, [\xi])$ from (M_0, γ_0) to (M_1, γ_1), the map

$$F_{\mathcal{W}}\colon \mathrm{SFH}(M_0, \gamma_0) \to \mathrm{SFH}(M_1, \gamma_1)$$

is the composition $F_{\mathcal{W}^s} \circ \Phi_{-\xi}$, where

$$\Phi_{-\xi}\colon \mathrm{SFH}(M_0, \gamma_0) \to \mathrm{SFH}(N, \gamma_1)$$

is the gluing map given by Honda, Kazez and Matić [11] for the sutured submanifold $(-M_0, -\gamma_0)$ of $(-N, -\gamma_1)$ with $N = M_0 \cup (-Z)$, and $F_{\mathcal{W}^s}$ is a "surgery map" corresponding to handles attached along the *interior* of the sutured manifold N. The cobordism $\mathcal{W}^s$ is a *special cobordism*, meaning its vertical part is a product and the contact structure on it is I–invariant.

If $\mathcal{C} = (X, F, \sigma)$ is a decorated concordance from (K_0, P_0) to (K_1, P_1), let $\mathcal{W} = \mathcal{W}(\mathcal{C})$ be the complementary sutured manifold cobordism from $S^3(K_0, P_0) = (M_0, \gamma_0)$ to $S^3(K_1, P_1) = (M_1, \gamma_1)$. Let $T^2 \times I$ be a collar neighbourhood of ∂M_0 such that $T^2 \times \{1\}$ is identified with ∂M_0. Since the dividing set on F consists of two arcs connecting the two components of ∂F, there is a diffeomorphism $d\colon T^2 \times I \to Z$ such that $\xi' = d^*(\xi)$ is an I–invariant contact structure on $T^2 \times I$, and hence induces the trivial gluing map by [11, Theorem 6.1]. More precisely, if we write $M'_0 = M_0 \setminus (T^2 \times I)$ and γ'_0 for the projection of γ_0 to $T^2 \times \{0\}$, then there is a diffeomorphism $\varphi\colon (M'_0, \gamma'_0) \to (M_0, \gamma_0)$ supported in a neighbourhood of $T^2 \times \{0\}$ such that

$$\Phi_{-\xi'} = \varphi_*\colon \mathrm{SFH}(M'_0, \gamma'_0) \to \mathrm{SFH}(M_0, \gamma_0).$$

Let $D\colon M_0 \to N$ be the diffeomorphism that agrees with φ on M_0' and with d on $T^2 \times I$, smoothed along $T^2 \times \{0\}$. By the diffeomorphism invariance of the gluing construction, the diagram

$$\begin{tikzcd}
\mathrm{SFH}(M_0', \gamma_0') \arrow[r, "\varphi_*"] \arrow[d, "\Phi_{-\xi'}"] & \mathrm{SFH}(M_0, \gamma_0) \arrow[d, "\Phi_{-\xi}"] \\
\mathrm{SFH}(M_0, \gamma_0) \arrow[r, "D_*"] & \mathrm{SFH}(N, \gamma)
\end{tikzcd}$$

is commutative, hence $\Phi_{-\xi} = D_*$.

We now show that D_* preserves the Alexander grading on the chain level. If we glue $D^2 \times S^1$ to N along ∂N such that the meridian is glued to a suture in $s(\gamma_1)$, we obtain a 3–manifold Y diffeomorphic to S^3, and the image of $\{0\} \times S^1$ is a knot K' in Y. We can canonically extend D to a diffeomorphism from (S^3, K_0) to (Y, K'). Given a knot diagram $\mathcal{H}_0 = (\Sigma_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, w_0, z_0)$ for (S^3, K_0), its image $D(\mathcal{H}_0)$ is a diagram of (Y, K'). Given a Seifert surface S of K_0 and a generator $\boldsymbol{x} \in \mathbb{T}_{\alpha_0} \cap \mathbb{T}_{\beta_0}$, the image $D(S)$ is a Seifert surface of K', and $D(\boldsymbol{x})$ satisfies

$$\langle c_1(\mathfrak{s}^\circ(\boldsymbol{x}), t), [S]\rangle = \big\langle c_1\big(\mathfrak{s}^\circ(D(\boldsymbol{x})), D_*(t)\big), [D(S)]\big\rangle.$$

As $D(\gamma_0) = \gamma_1$, the trivialization $D_*(t)$ points in the meridional direction for K', and it follows that $\mathcal{A}(\boldsymbol{x}) = \mathcal{A}(D(\boldsymbol{x}))$. It is apparent from the above discussion that we can identify (S^3, K_0) and (Y, K') via D, so from now on we will think of $\mathcal{W}$ as a special cobordism from (S^3, K_0) to (S^3, K_1).

5D Notation

In this subsection, we fix the notation for the rest of the paper. Recall that (K_0, P_0) and (K_1, P_1) denote two decorated knots in S^3, and that we have a decorated concordance $\mathcal{C} = (X, F, \sigma)$ from (K_0, P_0) to (K_1, P_1).

We denote by $\mathcal{W} = (W, Z, [\xi])$ the sutured cobordism $\mathcal{W}(\mathcal{C})$ associated to the knot concordance $\mathcal{C}$. It follows from the discussion in Section 5C that $\mathcal{W}$ can be thought of as a special cobordism. The 4–manifold W can be obtained by attaching to $M_0 \times I$ along the interior of $M_0 \times \{1\}$ a sequence of 4–dimensional 1–handles, followed by 2–handles, and finally 3–handles. We denote the number of i–handles by c_i for $i \in \{1, 2, 3\}$, and often write p for c_1 and ℓ for c_2. We split the cobordism $\mathcal{W}$ into three parts $\mathcal{W}_1$, $\mathcal{W}_2$ and $\mathcal{W}_3$, in such a way that $\mathcal{W}_i = (W_i, Z_i, [\xi_i])$ is a cobordism from (M_{i-1}, γ_{i-1}) to (M_i, γ_i), and is the trace of the i–handle attachments; see the left-hand side of Figure 1. Notice that $(M_0, \gamma_0) = S^3(K_0, P_0)$ and $(M_3, \gamma_3) = S^3(K_1, P_1)$ by construction.

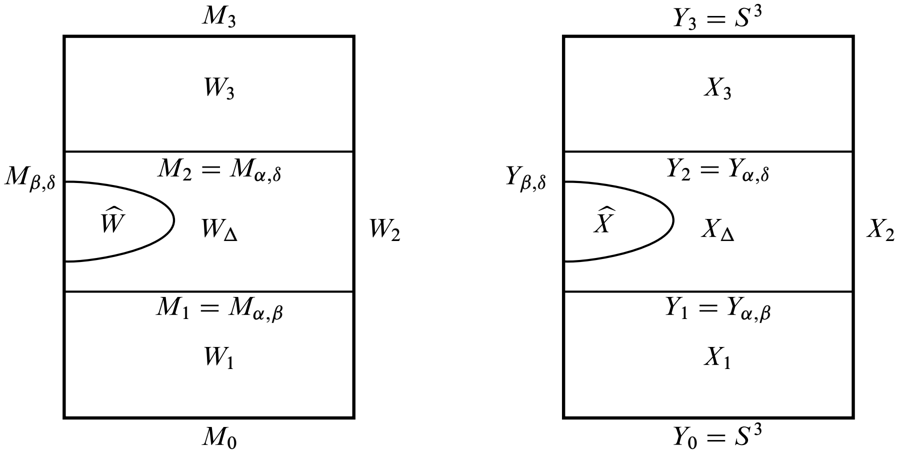


Figure 1: The left-hand side shows the sutured cobordism $\mathcal{W} = (W, Z, [\xi])$, and how we split it into different pieces. The picture on the right-hand side shows the cobordism of 3–manifolds X, and the corresponding decomposition into smaller cobordisms.

In order to represent sutured manifolds, we use Heegaard diagrams with basepoints. If $w, z \in \Sigma \setminus (\boldsymbol{\alpha} \cup \boldsymbol{\beta})$, the Heegaard diagram $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, w, z)$ represents the complement of a knot in a 3–manifold. In order to recover the sutured Heegaard diagram as originally defined by the first author [13], one should remove a small disk around each basepoint.

Let $\mathcal{T} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, w, z)$ be a doubly pointed triple diagram for the cobordism $\mathcal{W}_2$ (see Section 5H), where $d = |\boldsymbol{\alpha}| = |\boldsymbol{\beta}| = |\boldsymbol{\delta}|$. Furthermore, suppose that the 2–handles are attached along an ℓ–component framed link $\mathbb{L}$. We further split the manifold W_2 into two pieces according to [16, Proposition 6.6]: The piece $\mathcal{W}_{\alpha,\beta,\delta} = (W_\Delta, Z_\Delta, \xi_\Delta)$ denotes the sutured manifold cobordism obtained from the triangle construction in [16, Sections 5 and 6], while $\mathcal{W}_\beta(\mathbb{L}) = (\widehat{W}, \widehat{Z}, \widehat{\xi})$ is a sutured manifold cobordism from

$$(R_+(\gamma_1), \partial R_+(\gamma_1) \times I) \# \left(\mathop{\#}_{i=1}^{d-\ell} (S^2 \times S^1) \right)$$

to $\varnothing$. The horizontal boundary of $\widehat{W}$ is the sutured manifold $M_{\beta,\delta}$, defined by the diagram $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\delta}, w, z)$. By analogy, we also use the notation $M_{\alpha,\beta} \cong (M_1, \gamma_1)$ and $M_{\alpha,\delta} \cong (M_2, \gamma_2)$.

We can fill in the vertical boundary of the sutured cobordism $\mathcal{W}$ by gluing $D^2 \times S^1 \times I$ along $S^1 \times S^1 \times I$ to Z such that $S^1 \times \{(1,0)\}$ is glued to a meridian of K_0 to obtain cobordisms of closed 3–manifolds rather than knot complements. In terms of

Heegaard diagrams, this amounts to forgetting the z basepoints. We denote the closed 3–manifolds by the letter Y rather than M. As for the cobordisms, we use the letter X instead of the letter W. See the right-hand side of Figure 1.

Lastly, let $S_0 \subseteq M_0$ and $S_3 \subseteq M_3$ be Seifert surfaces for K_0 and K_1, respectively. Since (M_1, γ_1) is obtained from (M_0, γ_0) by taking connected sums with copies of $S^1 \times S^2$, the surface S_0 also defines a surface $S_1 \subseteq M_1$, which is contained in the M_0 summand of M_1. Analogously, the Seifert surface S_3 induces a Seifert surface $S_2 \subseteq M_2$.

5E Definition of the chain map $f_{\mathcal{C}}$

We now define the chain map $f_{\mathcal{C}}$. Given an admissible doubly pointed diagram $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, w, z)$ for a decorated knot (Y, K, P), we denote by $\widehat{\mathrm{CF}}(\mathcal{H})$ the Heegaard Floer chain complex that counts disks avoiding w and filtered by z. Its homology is $\widehat{\mathrm{HF}}(Y, w)$, while the homology of the associated graded complex $\widehat{\mathrm{CFK}}(\mathcal{H})$ is $\widehat{\mathrm{HFK}}(Y, K, P)$.

Suppose that the 1–handles are attached along p framed pairs of points $\mathbb{P} \subset M_0$. Pick an admissible diagram $\mathcal{H}^0$ of (M_0, γ) subordinate to $\mathbb{P}$, and let

$$f_{\mathcal{H}^0,\mathbb{P}}\colon \widehat{\mathrm{CF}}(\mathcal{H}^0) \to \widehat{\mathrm{CF}}(\mathcal{H}^0_{\mathbb{P}})$$

be the 1–handle map defined in [16, Definition 7.5]. The 2–handles are attached along an ℓ–component framed link $\mathbb{L} \subset M_1$. Choose an admissible diagram $\mathcal{H}^1$ subordinate to $\mathbb{L}$, and let

$$f_{\mathcal{H}^1,\mathbb{L}}\colon \widehat{\mathrm{CF}}(\mathcal{H}^1) \to \widehat{\mathrm{CF}}(\mathcal{H}^1_{\mathbb{L}})$$

be the 2–handle map defined in [16, Definition 6.8], on the chain level. This map counts triangles that avoid w but might pass through z. Finally, let $\mathcal{H}^2$ be an admissible diagram of (M_2, γ) subordinate to framed spheres $\mathbb{S} \subset M_2$ corresponding to the 3–handles. The corresponding 3–handle map

$$f_{\mathcal{H}^2,\mathbb{S}}\colon \widehat{\mathrm{CF}}(\mathcal{H}^2) \to \widehat{\mathrm{CF}}(\mathcal{H}^2_{\mathbb{S}})$$

was introduced in [16, Definition 7.8].

Given admissible diagrams $\mathcal{H}$ and $\mathcal{H}'$ of a sutured manifold (M, γ), we refer the reader to [16, Section 5.2] for the definition of the canonical isomorphism

$$F_{\mathcal{H},\mathcal{H}'}\colon \mathrm{SFH}(\mathcal{H}) \to \mathrm{SFH}(\mathcal{H}').$$

We can obtain a chain level representative by connecting $\mathcal{H}$ and $\mathcal{H}'$ through a sequence of ambient isotopies, (de)stabilizations, and equivalences of the attaching sets. If (M, γ)

is complementary to a knot (Y, K), we can view this as a sequence of moves on knot diagrams. Each induces a chain homotopy equivalence on $\widehat{\mathrm{CF}}$ preserving the knot filtration according to [28; 31], and induces an isomorphism both on the homology of the whole complex (isomorphic to $\widehat{\mathrm{HF}}(Y)$), and the homology of the associated graded complex (isomorphic to $\widehat{\mathrm{HFK}}(Y, K)$). Note that the triangle maps corresponding to changing the attaching curves do not pass over w but might cross z, so they are in fact naturality maps for the closed 3–manifold and *not* the knot. We proved in [17] that the maps on the homology are independent of the sequence of moves connecting $\mathcal{H}$ and $\mathcal{H}'$. We write $f_{\mathcal{H},\mathcal{H}'}$ for the chain level representative of $F_{\mathcal{H},\mathcal{H}'}$ described above. With the above notation in place, we set

$$f_{\mathcal{C}} := f_{\mathcal{H}^2,\mathbb{S}} \circ f_{\mathcal{H}^1_{\mathbb{L}},\mathcal{H}^2} \circ f_{\mathcal{H}^1,\mathbb{L}} \circ f_{\mathcal{H}^0_{\mathbb{P}},\mathcal{H}^1} \circ f_{\mathcal{H}^0,\mathbb{P}},$$

from $\widehat{\mathrm{CF}}(\mathcal{H}^0)$ to $\widehat{\mathrm{CF}}(\mathcal{H}^2_{\mathbb{S}})$. Note that each of the diagrams involved in the above formula can be viewed as a knot diagram after gluing disks along $s(\gamma)$ that do not change during the cobordism, so we can distinguish z and w throughout. If we are given diagrams $\mathcal{H}$ of (M_0, γ_0) and $\mathcal{H}'$ of (M_3, γ_3), then we have to pre- and postcompose the above map $f_{\mathcal{C}}$ with $f_{\mathcal{H}^2_{\mathbb{S}},\mathcal{H}'}$ and $f_{\mathcal{H},\mathcal{H}^0}$.

We split the proof of Theorem 5.4 into a number of steps, and we prove that for each $\mathcal{W}_i$ the knot filtration is preserved.

5F 1– and 3–handles

First, consider the case of the 1–handle attachments along the framed pairs of points $\mathbb{P} \subset \mathrm{Int}(M_0)$. As in Section 5D, we write $\mathcal{W}_1 := \mathcal{W}(\mathbb{P})$ for the trace of the surgery along $\mathbb{P}$; this is a cobordism from (M_0, γ_0) to (M_1, γ_1). Recall [16, Section 7] that there is an isomorphism $\mathrm{Spin}^c(\mathcal{W}_1) \cong \mathrm{Spin}^c(M_0, \gamma_0)$. Furthermore, a Spin^c structure $\mathfrak{s}^\circ \in \mathrm{Spin}^c(M_1, \gamma_1)$ extends to $\mathcal{W}_1$ if and only if $c_1(\mathfrak{s}^\circ)$ vanishes on the belt spheres of all the 1–handles. Given $\mathfrak{s}^\circ \in \mathrm{Spin}^c(\mathcal{W})$, we write $\mathfrak{s}^\circ_0$ for its restriction to (M_0, γ_0), and $\mathfrak{s}^\circ_1$ for its restriction to (M_1, γ_1).

Lemma 5.8 *Let* $\mathfrak{s}^\circ_0 \in \mathrm{Spin}^c(M_0, \gamma_0)$, *and let* $\mathfrak{s}^\circ_1 \in \mathrm{Spin}^c(M_1, \gamma_1)$ *denote the corresponding* Spin^c *structure. Then*

$$\langle c_1(\mathfrak{s}^\circ_0, t), [S_0] \rangle = \langle c_1(\mathfrak{s}^\circ_1, t), [S_1] \rangle.$$

Proof This is a consequence of the naturality of the first Chern class and the fact that both S_0 and S_1 are actually contained in $M_0 \setminus N(\mathbb{P})$. We can suppose that S_0 is properly embedded in $M_0 \setminus N(\mathbb{P})$. By definition, S_1 is a surface contained in $M_0 \setminus N(\mathbb{P}) \subseteq M_1$ that is isotopic to S_0 in $M_0 \setminus N(\mathbb{P})$.

Since S_0 and S_1 are isotopic in $M_0 \setminus N(\mathbb{P})$ and $\mathfrak{s}_1^\circ|_{M_0 \setminus N(\mathbb{P})} = \mathfrak{s}_0^\circ|_{M_0 \setminus N(\mathbb{P})}$, by the naturality of the first Chern class

$$\begin{aligned}\langle c_1(\mathfrak{s}_1^\circ, t), [S_1]\rangle &= \langle c_1(\mathfrak{s}_1^\circ|_{M_0\setminus N(\mathbb{P})}, t), [S_1]\rangle \\ &= \langle c_1(\mathfrak{s}_0^\circ|_{M_0\setminus N(\mathbb{P})}, t), [S_0]\rangle \\ &= \langle c_1(\mathfrak{s}_0^\circ, t), [S_0]\rangle.\end{aligned}$$

Notice that the trivialization t of the vector field v_0 on $\partial M_0 = \partial M_1$ does not change because the boundary is left unaffected by the surgery. □

Remark 5.9 Since $c_1(\mathfrak{s}_1^\circ, t)$ vanishes on the belt spheres of the 1–handles, the above result also holds for an arbitrary Seifert surface S_1.

Corollary 5.10 *The map $f_{\mathcal{H}^0,\mathbb{P}}\colon \widehat{\mathrm{CF}}(\mathcal{H}^0) \to \widehat{\mathrm{CF}}(\mathcal{H}^0_{\mathbb{P}})$ preserves the Alexander grading (see Definition 5.1) with respect to arbitrary Seifert surfaces S_0 and S_1; ie*

$$\mathcal{A}_{S_1}(f_{\mathcal{H}^0,\mathbb{P}}(\boldsymbol{x})) = \mathcal{A}_{S_0}(\boldsymbol{x})$$

for any $\boldsymbol{x} \in \mathbb{T}_{\alpha^0} \cap \mathbb{T}_{\beta^0}$, where $\mathcal{H}^0 = (\Sigma^0, \boldsymbol{\alpha}^0, \boldsymbol{\beta}^0, w^0, z^0)$.

Proof This is a straightforward consequence of Lemma 5.8, Remark 5.9, and the fact that the relative Spin^c structure induced by $\mathfrak{s}^\circ(\boldsymbol{x})$ on (M_1, γ) is exactly $\mathfrak{s}^\circ(f_{\mathcal{H}^0,\mathbb{P}}(\boldsymbol{x}))$. □

A dual reasoning gives the following results for the map $f_{\mathcal{H}^2,\mathbb{S}}$, which are analogous to Lemma 5.8 and Corollary 5.10.

Lemma 5.11 *Let $\mathfrak{s}_3^\circ \in \mathrm{Spin}^c(M_3, \gamma_3)$, and let $\mathfrak{s}_2^\circ \in \mathrm{Spin}^c(M_2, \gamma_2)$ denote the corresponding Spin^c structure. Then*

$$\langle c_1(\mathfrak{s}_2^\circ, t), [S_2]\rangle = \langle c_1(\mathfrak{s}_3^\circ, t), [S_3]\rangle.$$

Corollary 5.12 *The map $f_{\mathcal{H}^2,\mathbb{S}}\colon \widehat{\mathrm{CF}}(\mathcal{H}^2) \to \widehat{\mathrm{CF}}(\mathcal{H}^2_{\mathbb{S}})$ preserves the Alexander grading with respect to arbitrary Seifert surfaces S_2 and S_3; ie*

$$\mathcal{A}_{S_3}(f_{\mathcal{H}^2,\mathbb{S}}(\boldsymbol{x})) = \mathcal{A}_{S_2}(\boldsymbol{x})$$

for any $\boldsymbol{x} \in \mathbb{T}_{\alpha^2} \cap \mathbb{T}_{\beta^2}$ such that $f_{\mathcal{H}^2,\mathbb{S}}(\boldsymbol{x}) \neq 0$, where $\mathcal{H}^2 = (\Sigma^2, \boldsymbol{\alpha}^2, \boldsymbol{\beta}^2, w^2, z^2)$.

5G 2–handles

The proof that the Alexander grading is preserved under the attachment of the 2–handles is less straightforward than in the case of 1–handles and 3–handles.

Lemma 5.13 *Let $\mathcal{C}$ be a decorated concordance from (K_0, P_0) to (K_1, P_1). With the notation of Section 5D, let $\mathcal{W}_2$ denote the 2–handle cobordism from (M_1, γ_1) to (M_2, γ_2) obtained by surgery along a framed link $\mathbb{L}$, and let S_1 and S_2 be corresponding Seifert surfaces. Then there is an admissible doubly pointed triple diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, w, z)$ subordinate to a bouquet for $\mathbb{L}$ as follows: If $\boldsymbol{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ is such that $\mathfrak{s}(\boldsymbol{x}) \in \mathrm{Spin}^c(Y_{\alpha,\beta})$ extends to X_1, then for any $\boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\delta$ that appears with nonzero coefficient in $f_{\mathcal{H}^1,\mathbb{L}}(\boldsymbol{x})$, and such that $\mathfrak{s}(\boldsymbol{y}) \in \mathrm{Spin}^c(Y_{\alpha,\delta})$ extends to X_3, we have*

$$\mathcal{F}_{S_2}(\boldsymbol{y}) \le \mathcal{F}_{S_1}(\boldsymbol{x}).$$

Moreover, if ψ is a holomorphic triangle connecting $\boldsymbol{x}$, θ (the top-graded generator of $\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\delta}, w, z)$), and $\boldsymbol{y}$ that does not cross w, then

$$\mathcal{F}_{S_2}(\boldsymbol{y}) = \mathcal{F}_{S_1}(\boldsymbol{x}) - n_z(\psi). \tag{5-3}$$

Notice that, in Lemma 5.13, we consider ordinary Spin^c structures rather than relative ones. Recall that relative Spin^c structures are defined for sutured cobordisms, which we denote by the letter $\mathcal{W}$, while ordinary Spin^c structures are defined for cobordisms of 3–manifolds, which we denote by the letter X; see Figure 1.

Idea of the proof Consider an admissible Heegaard triple diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$ subordinate to a bouquet for a framed link $\mathbb{L}$, as explained in [16, Section 6]. Suppose that $\boldsymbol{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ is such that $\mathfrak{s}(\boldsymbol{x}) \in \mathrm{Spin}^c(Y_{\alpha,\beta})$ extends to X_1. Let $\theta \in \mathbb{T}_\beta \cap \mathbb{T}_\delta$ be the top-graded generator of $\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\delta})$, and let $\boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\delta$ be such that $\mathfrak{s}(\boldsymbol{y}) \in \mathrm{Spin}^c(Y_{\alpha,\delta})$ extends to X_3. Given a holomorphic triangle $\psi \in \pi_2(\boldsymbol{x}, \theta, \boldsymbol{y})$, let

$$c = \mathcal{A}_{S_2}(\boldsymbol{y}) - \mathcal{A}_{S_1}(\boldsymbol{x}) + n_z(\psi) - n_w(\psi).$$

First, we prove that c is independent of ψ, $\boldsymbol{x}$ and $\boldsymbol{y}$. If $\psi_1, \psi_2 \in \pi_2(\boldsymbol{x}, \theta, \boldsymbol{y})$, then the domain $\mathcal{D}(\psi_1) - \mathcal{D}(\psi_2)$ is triply periodic. If we prove that, for every triply periodic domain D, we have

$$n_z(D) - n_w(D) = 0,$$

then c is independent of ψ. For this reason, the next subsection is devoted to the study of triply periodic domains in the setting of Lemma 5.13.

Given two different intersection points $\boldsymbol{x}' \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ and $\boldsymbol{y}' \in \mathbb{T}_\alpha \cap \mathbb{T}_\delta$ such that $\mathfrak{s}(\boldsymbol{x}') \in \mathrm{Spin}^c(Y_{\alpha,\beta})$ extends to X_1 and $\mathfrak{s}(\boldsymbol{y}') \in \mathrm{Spin}^c(Y_{\alpha,\delta})$ extends to X_3, there are domains $D_{\boldsymbol{x}}$ connecting $\boldsymbol{x}$ with $\boldsymbol{x}'$ and $D_{\boldsymbol{y}}$ connecting $\boldsymbol{y}$ with $\boldsymbol{y}'$ that do not pass through w (but might have nontrivial multiplicities at z). Adding these domains to $D(\psi)$, we get a triangle domain connecting $\boldsymbol{x}'$, θ and $\boldsymbol{y}'$ with the same c by Lemma 5.3.

Then we show that $c = 0$ by isotoping $\boldsymbol{\alpha}$ to obtain a diagram where such $\boldsymbol{x}$, $\boldsymbol{y}$ and ψ as above exist, and invoke Lemma 3.9. Finally, if ψ appears in the surgery map $f_{\mathcal{H}^1,\mathbb{L}}(\boldsymbol{x})$, then $n_w(\psi) = 0$ and it has a pseudoholomorphic representative, so $n_z(\psi) \geq 0$. Consequently, $A_{S_2}(\boldsymbol{y}) \leq A_{S_1}(\boldsymbol{x})$, as desired. □

We now explain the missing details in the above outline.

5H Triply periodic domains

The following argument was motivated by the work of Manolescu and Ozsváth [22].

Definition 5.14 A *doubly pointed triple Heegaard diagram* is a tuple

$$\mathcal{T} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, w, z),$$

where Σ is a closed, oriented surface, and there is an integer $d \geq 0$ such that the sets $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ all consist of d pairwise disjoint simple closed curves in $\Sigma \setminus \{w, z\}$ that are linearly independent in $H_1(\Sigma \setminus \{w, z\})$.

We denote by $Y_{\alpha,\beta}$, $Y_{\alpha,\delta}$ and $Y_{\beta,\delta}$ the 3–manifolds represented by the Heegaard diagrams $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$, $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\delta})$ and $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\delta})$, respectively.

Definition 5.15 Let $\mathcal{T} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, w, z)$ be a doubly pointed triple Heegaard diagram. Let $D_1, \dots, D_l$ denote the closures of the components of $\Sigma \setminus (\boldsymbol{\alpha} \cup \boldsymbol{\beta} \cup \boldsymbol{\delta})$. Then the set of *domains* in $\mathcal{T}$ is

$$D(\mathcal{T}) = \mathbb{Z}\langle D_1, \dots, D_l \rangle.$$

We denote by $n_z(\mathcal{D})$ (respectively $n_w(\mathcal{D})$) the multiplicity of a domain $\mathcal{D} \in D(\mathcal{T})$ in the region D_i that contains z (respectively w).

A *triply periodic domain* is an element $\mathcal{P} \in D(\mathcal{T})$ such that $\partial\mathcal{P}$ is a $\mathbb{Z}$–linear combination of curves in $\boldsymbol{\alpha} \cup \boldsymbol{\beta} \cup \boldsymbol{\delta}$. We denote the set of triply periodic domains by $\Pi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta}}$.

A *doubly periodic domain* is an element $\mathcal{P} \in D(\mathcal{T})$ such that $\partial\mathcal{P}$ is a $\mathbb{Z}$–linear combination of curves either in $\boldsymbol{\alpha} \cup \boldsymbol{\beta}$, or in $\boldsymbol{\beta} \cup \boldsymbol{\delta}$, or in $\boldsymbol{\alpha} \cup \boldsymbol{\delta}$. We denote the set of the three types of doubly periodic domains by $\Pi_{\alpha,\beta}$, $\Pi_{\alpha,\delta}$ and $\Pi_{\beta,\delta}$, respectively.

The following result states that every triply periodic domain in the diagram describing the surgery map for $\mathcal{W}_2$ can be written as a sum of doubly periodic domains.

Proposition 5.16 *Let $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$ denote a Heegaard diagram associated to the cobordism X_2. Then*

$$\Pi_{\alpha,\beta,\delta} = \Pi_{\alpha,\beta} + \Pi_{\alpha,\delta} + \Pi_{\beta,\delta}.$$

Given a triple diagram associated to a surgery on an ℓ–component link $\mathbb{L}$, one can construct a 4–manifold $X_\triangle$ as in [30, Section 2.2]; see [16, Section 5] for the analogous construction in the sutured setting. The 3–manifolds $Y_{\alpha,\beta}$, $Y_{\alpha,\delta}$ and $Y_{\beta,\delta}$, defined by the Heegaard diagrams $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$, $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\delta})$ and $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\delta})$, respectively, naturally sit in $\partial X_\triangle$. The cobordism X_2 corresponding to the attachment of the 2–handles is obtained by gluing the 4–manifold $\widehat{X} = \natural_{i=1}^{\ell}(S^1 \times D^3)$ to $X_\triangle$ along $Y_{\beta,\delta} \cong \#_{i=1}^{\ell}(S^1 \times S^2)$.

Lemma 5.17 [29, Propositions 2.15 and 8.3] *Given a pointed triple Heegaard diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, z)$, there are isomorphisms*

$$\pi_{\alpha,\beta}\colon \Pi_{\alpha,\beta} \xrightarrow{\cong} \mathbb{Z} \oplus H_2(Y_{\alpha,\beta}) \quad \text{and} \quad \pi_{\alpha,\beta,\delta}\colon \Pi_{\alpha,\beta,\delta} \xrightarrow{\cong} \mathbb{Z} \oplus H_2(X_\triangle).$$

In both cases, the projection onto the $\mathbb{Z}$ summand is given by n_z.

Lemma 5.18 *Given a pointed triple Heegaard diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, z)$, the isomorphisms from Lemma 5.17 fit into the commutative diagram*

$$\begin{tikzcd}
\Pi_{\alpha,\beta} \arrow[r, "\pi_{\alpha,\beta}"] \arrow[d] & \mathbb{Z} \oplus H_2(Y_{\alpha,\beta}) \arrow[d, "\mathrm{Id}_{\mathbb{Z}} \oplus i_*"] \\
\Pi_{\alpha,\beta,\delta} \arrow[r, "\pi_{\alpha,\beta,\delta}"] & \mathbb{Z} \oplus H_2(X_\triangle)
\end{tikzcd}$$

where $i\colon Y_{\alpha,\beta} \to X_\triangle$ is the embedding.

Proof Let $\mathcal{P}$ be a doubly periodic domain in $\Pi_{\alpha,\beta}$. By construction, the 2–chain in $X_\triangle$ associated to $\mathcal{P}$ — thought of as a triply periodic domain — is homotopic, hence homologous to $i_*(H(\mathcal{P}))$, where $H(\mathcal{P})$ is the 2–chain in $Y_{\alpha,\beta}$ obtained by capping off the boundary of the doubly periodic domain $\mathcal{P}$. Therefore, the projections onto the second summand commute. The projections onto the $\mathbb{Z}$ summands commute because in both cases they are obtained by taking n_z. □

Proof of Proposition 5.16 By Lemmas 5.17 and 5.18, it is sufficient to prove that the map

$$\chi\colon H_2(Y_{\alpha,\beta}) \oplus H_2(Y_{\alpha,\delta}) \oplus H_2(Y_{\beta,\delta}) \to H_2(X_\triangle)$$

is surjective.

From the long exact sequence associated to the pair $(X_\triangle, Y_{\alpha,\beta} \sqcup Y_{\alpha,\delta} \sqcup Y_{\beta,\delta})$, we see that the map χ is surjective if and only if

$$\varphi\colon H_2(X_\triangle, Y_{\alpha,\beta} \sqcup Y_{\alpha,\delta} \sqcup Y_{\beta,\delta}) \to H_1(Y_{\alpha,\beta} \sqcup Y_{\alpha,\delta} \sqcup Y_{\beta,\delta})$$

is injective. From the inclusion of pairs

$$i_{\alpha,\beta,\delta}\colon (X_\triangle, Y_{\alpha,\beta} \sqcup Y_{\alpha,\delta} \sqcup Y_{\beta,\delta}) \hookrightarrow (X, X_1 \sqcup X_3 \sqcup \widehat{X}),$$

we obtain the commutativity of the following diagram:

$$\begin{tikzcd}
H_2(X_\triangle, Y_{\alpha,\beta} \sqcup Y_{\alpha,\delta} \sqcup Y_{\beta,\delta}) \arrow[r, "\varphi"] \arrow[d, "(i_{\alpha,\beta,\delta})_*"', "\wr"] & H_1(Y_{\alpha,\beta}) \oplus H_1(Y_{\alpha,\delta}) \oplus H_1(Y_{\beta,\delta}) \arrow[d, "(i_{\alpha,\beta})_* \oplus (i_{\alpha,\delta})_* \oplus (i_{\beta,\delta})_*"] \\
H_2(X, X_1 \sqcup X_3 \sqcup \widehat{X}) \arrow[r, "\simeq", "\widetilde{\varphi}"'] & H_1(X_1) \oplus H_1(X_3) \oplus H_1(\widehat{X})
\end{tikzcd}$$

where $i_{\alpha,\beta}$, $i_{\alpha,\delta}$ and $i_{\beta,\delta}$ are the restrictions of $i_{\alpha,\beta,\delta}$ to $Y_{\alpha,\beta}$, $Y_{\alpha,\delta}$ and $Y_{\beta,\delta}$, respectively. The map $(i_{\alpha,\beta,\delta})_*$ is an isomorphism by excision. The fact that $\widetilde{\varphi}$ is an isomorphism follows from the long exact sequence in homology associated with the pair $(X, X_1 \sqcup X_3 \sqcup \widehat{X})$, together with the fact that $H_2(X) = H_1(X) = 0$.

The commutativity of the above diagram implies that the map φ is injective, and therefore concludes the proof of the proposition. □

Remark 5.19 The important condition in Proposition 5.16 is that the map

$$\rho\colon H_2(X_1) \oplus H_2(X_3) \oplus H_2(\widehat{X}) \to H_2(X)$$

is surjective, which is obviously true as $H_2(X) = 0$. The surjectivity of ρ is equivalent to the injectivity of $\widetilde{\varphi}$, which implies the injectivity of φ.

In Proposition 5.16, we saw that, in the case of a triple diagram describing the 2–handle attachments in the cobordism X, every triply periodic domain can be expressed as a sum of doubly periodic domains. We now analyze the doubly periodic domains.

Proposition 5.20 *Consider a null-homologous knot K in a 3–manifold Y. Given a doubly pointed Heegaard diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, w, z)$ for (Y, K), every periodic domain $\mathcal{P}$ satisfies*

$$n_z(\mathcal{P}) - n_w(\mathcal{P}) = 0.$$

Proof Let $H(\mathcal{P}) \in C_2(Y)$ be the 2–cycle obtained by capping off the boundary of $\mathcal{P}$ with the cores of the 3–dimensional 2–handles attached to $\Sigma \times I$ along $\boldsymbol{\alpha} \times \{0\}$ and $\boldsymbol{\beta} \times \{1\}$. Then $n_z(\mathcal{P}) - n_w(\mathcal{P})$ is precisely the algebraic intersection number of $H(\mathcal{P})$ and K, which is zero as K is null-homologous. □

5I Representing homology classes

Let $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$ be a Heegaard diagram for a 3–manifold Y. It is straightforward to see that any element of $H_1(Y)$ can be represented by a 1–cycle in Σ. In this subsection, we strengthen this result for the case of concordances in the following sense.

Lemma 5.21 *Choose an arbitrary handle decomposition of the cobordism X from S^3 to S^3, and let X_2 denote the trace of the 2–handle attachments. Suppose that $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, w, z)$ is a doubly pointed triple Heegaard diagram subordinate to a bouquet for a link $\mathbb{L}$ that defines X_2. Then the map*

$$i\colon H_1(\Sigma) \to H_1(Y_{\alpha,\beta}) \oplus H_1(Y_{\alpha,\delta}),$$

induced by the inclusions $\Sigma \hookrightarrow Y_{\alpha,\beta}$ and $\Sigma \hookrightarrow Y_{\alpha,\delta}$, is surjective.

In other words, given any two classes in the first homologies of $Y_{\alpha,\beta}$ and $Y_{\alpha,\delta}$, there is a 1–cycle in Σ that represents both simultaneously.

Proof Consider the following short exact sequence of abelian groups:

$$0 \to \frac{H_1(\Sigma)}{\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle \cap \langle \boldsymbol{\alpha}, \boldsymbol{\delta} \rangle} \to \frac{H_1(\Sigma)}{\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle} \oplus \frac{H_1(\Sigma)}{\langle \boldsymbol{\alpha}, \boldsymbol{\delta} \rangle} \to \frac{H_1(\Sigma)}{\langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta} \rangle} \to 0.$$

The middle term is isomorphic to $H_1(Y_{\boldsymbol{\alpha},\boldsymbol{\beta}}) \oplus H_1(Y_{\boldsymbol{\alpha},\boldsymbol{\delta}})$, and the last term is isomorphic to $H_1(X_\triangle)$, where $X_\triangle$ is the 4–manifold obtained by the triangle construction; see [29, Proposition 8.2]. The short exact sequence above can then be rewritten as

$$0 \to \frac{H_1(\Sigma)}{\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle \cap \langle \boldsymbol{\alpha}, \boldsymbol{\delta} \rangle} \xrightarrow{f} H_1(Y_{\boldsymbol{\alpha},\boldsymbol{\beta}}) \oplus H_1(Y_{\boldsymbol{\alpha},\boldsymbol{\delta}}) \xrightarrow{g} H_1(X_\triangle) \to 0.$$

If we prove that $H_1(X_\triangle) = 0$, then by exactness we have that the map f is surjective. So the map i in the statement of the lemma is surjective too, because it is obtained by composing the following two maps:

$$H_1(\Sigma) \to \frac{H_1(\Sigma)}{\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle \cap \langle \boldsymbol{\alpha}, \boldsymbol{\delta} \rangle} \xrightarrow{f} H_1(Y_{\alpha,\beta}) \oplus H_1(Y_{\alpha,\delta}).$$

Therefore, we only need to prove that $H_1(X_\Delta)=0$. For this purpose, consider the Mayer–Vietoris long exact sequence associated to the decomposition $X=A\cup B$, where $A=X_\Delta$ and $B=X_1\sqcup X_3\sqcup \widehat{X}$. A portion of the long exact sequence is

$$H_2(X)\to H_1(A\cap B)\xrightarrow{\iota} H_1(A)\oplus H_1(B)\to H_1(X).$$

Since X has trivial first and second homology groups, by exactness the map ι gives an isomorphism

$$H_1(Y_{\alpha,\beta})\oplus H_1(Y_{\alpha,\delta})\oplus H_1(Y_{\beta,\delta})\xrightarrow{\sim} H_1(X_\Delta)\oplus H_1(X_1)\oplus H_1(X_3)\oplus H_1(\widehat{X}). \tag{5-4}$$

If c_k denotes the number of k–handles in the decomposition of the cobordism X and $d=|\boldsymbol{\alpha}|$, then it is straightforward to check that

$$\begin{aligned}H_1(Y_{\alpha,\beta})&\cong H_1(X_1)\cong \mathbb{Z}^{c_1},\\ H_1(Y_{\beta,\delta})&\cong H_1(\widehat{X})\cong \mathbb{Z}^{d-c_2},\\ H_1(Y_{\alpha,\delta})&\cong H_1(X_3)\cong \mathbb{Z}^{c_3}.\end{aligned}$$

It now follows from (5-4) that $H_1(X_\Delta)=0$, which concludes the proof. □

5J Proof of Lemma 5.13

The cobordism $\mathcal{W}_2$ can be represented via surgery on a framed ℓ–component link $\mathbb{L}$. Let $\mathcal{T}=(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},w,z)$ be a doubly pointed triple Heegaard diagram subordinate to a bouquet for the framed link $\mathbb{L}$. As in [16, Section 6], we suppose $d=|\boldsymbol{\alpha}|=|\boldsymbol{\beta}|=|\boldsymbol{\delta}|$ and that the curve δ_i is an isotopic translate of β_i for $i\in\{\ell+1,\dots,d\}$.

Following notation established in Section 5D and in Figure 1, let $Y_{\alpha,\beta}$, $Y_{\alpha,\delta}$ and $Y_{\beta,\delta}$ denote the closed manifolds associated to the Heegaard diagrams $(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta})$, $(\Sigma,\boldsymbol{\alpha},\boldsymbol{\delta})$, and $(\Sigma,\boldsymbol{\beta},\boldsymbol{\delta})$, respectively. Each of these closed manifolds contains a knot, defined by the basepoints w and z. We denote the knot exteriors — thought of as sutured manifolds — by $M_{\alpha,\beta}$, $M_{\alpha,\delta}$ and $M_{\beta,\delta}$. We let γ denote the sutures of all three sutured manifolds.

Let $\mathfrak{s}$ be the unique Spin^c structure on X. By definition, $\mathfrak{s}|_{X_\Delta}$ is the unique Spin^c structure on X_Δ that extends to the whole cobordism X. Suppose that $\boldsymbol{x}\in\mathbb{T}_\alpha\cap\mathbb{T}_\beta$ and $\boldsymbol{y}\in\mathbb{T}_\alpha\cap\mathbb{T}_\delta$ are such that $\mathfrak{s}(\boldsymbol{x})=\mathfrak{s}|_{Y_{\alpha,\beta}}$ and $\mathfrak{s}(\boldsymbol{y})=\mathfrak{s}|_{Y_{\alpha,\delta}}$. Let $\theta\in\mathbb{T}_\beta\cap\mathbb{T}_\delta$ denote the top-graded generator. Consider a Whitney triangle $\psi\in\pi_2(\boldsymbol{x},\theta,\boldsymbol{y})$, possibly crossing the basepoints z and w, and let

$$c=\mathcal{A}_{S_2}(\boldsymbol{y})-\mathcal{A}_{S_1}(\boldsymbol{x})+n_z(\psi)-n_w(\psi). \tag{5-5}$$

Our aim is to show that $c=0$. First, we show that c is independent of the triangle ψ in $\pi_2(\boldsymbol{x},\theta,\boldsymbol{y})$ for fixed $\boldsymbol{x}$ and $\boldsymbol{y}$. Indeed, let $\psi_1,\psi_2\in\pi_2(\boldsymbol{x},\theta,\boldsymbol{y})$. The domain

$\mathcal{P} = \mathcal{D}(\psi_1) - \mathcal{D}(\psi_2)$ is triply periodic. By Proposition 5.16, $\mathcal{P}$ can be expressed as the sum of three doubly periodic domains $\mathcal{P}_{\alpha,\beta}$, $\mathcal{P}_{\beta,\delta}$ and $\mathcal{P}_{\alpha,\delta}$.

Since $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$ is subordinate to a bouquet for $\mathbb{L}$, the diagrams $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, w, z)$, $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\delta}, w, z)$ and $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\delta}, w, z)$ each define a null-homologous knot in a connected sum of a number of copies of $S^1 \times S^2$. Hence, by Proposition 5.20, $n_z(P) = n_w(P)$ for every $P \in \{\mathcal{P}_{\alpha,\beta}, \mathcal{P}_{\beta,\delta}, \mathcal{P}_{\alpha,\delta}\}$. So $n_z(\mathcal{P}) = n_w(\mathcal{P})$, and

$$n_z(\psi_1) - n_w(\psi_1) = n_z(\psi_2) - n_w(\psi_2).$$

Therefore, c is independent of the triangle ψ for fixed $\boldsymbol{x}$ and $\boldsymbol{y}$; see (5-5).

To check that c is independent of $\boldsymbol{x}$, we consider another generator $\boldsymbol{x}'$ such that $\mathfrak{s}(\boldsymbol{x}') = \mathfrak{s}|_{Y_{\alpha,\beta}} = \mathfrak{s}(\boldsymbol{x})$. Since $\boldsymbol{x}$ and $\boldsymbol{x}'$ represent the same Spin^c structure, there is a Whitney disk $\phi \in \pi_2(\boldsymbol{x}', \boldsymbol{x})$ (that possibly crosses the basepoints w and z). If $\psi \in \pi_2(\boldsymbol{x}, \theta, \boldsymbol{y})$, then $\phi \# \psi \in \pi_2(\boldsymbol{x}', \theta, \boldsymbol{y})$. By Lemma 5.3, the number c defined in (5-5) is the same for ψ and $\phi \# \psi$, so c does not depend on $\boldsymbol{x}$. A similar reasoning also proves that c is independent of $\boldsymbol{y}$.

What remains to prove is that $c = 0$. We do this by constructing a Whitney triangle ψ for which $c = 0$.

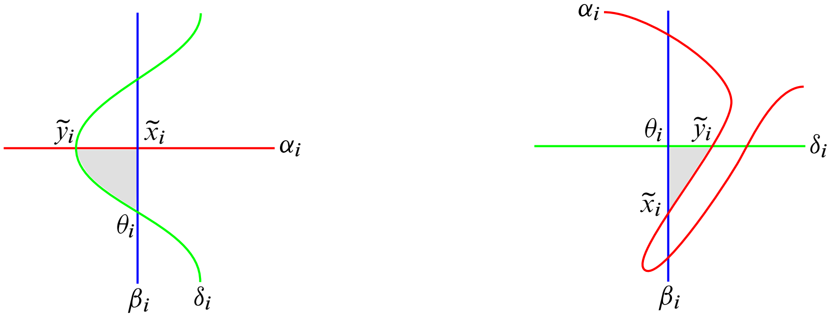


Figure 2: This shows the domain of the Whitney triangle $\tilde{\psi}$. The curves β_i and δ_i, for $i \in \{\ell+1, \dots, d\}$, are small isotopic translates of each other, and — after isotoping α_i — we can find a "small" triangle bounded by α_i, β_i and δ_i, shown shaded on the left. For $i \in \{1, \dots, \ell\}$, after applying finger moves to the α–curves, we can assume that there is a triangle, shown shaded on the right. The sum of all these triangles is the domain of $\tilde{\psi}$.

By isotoping the α–curves, we can create intersection points $\tilde{\boldsymbol{x}}$ in $\mathbb{T}_\alpha \cap \mathbb{T}_\beta$ and $\tilde{\boldsymbol{y}}$ in $\mathbb{T}_\alpha \cap \mathbb{T}_\delta$ such that there is a "small" triangle $\tilde{\psi} \in \pi_2(\tilde{\boldsymbol{x}}, \theta, \tilde{\boldsymbol{y}})$. The domain of $\tilde{\psi}$ is shown in Figure 2. For each $i \in \{\ell+1, \dots, d\}$, we isotope α_i — pushing the other α–curves alongside — until it intersects both δ_i and β_i near θ_i, and consider the shaded triangle shown on the left-hand side of the figure. For each $i \in \{1, \dots, \ell\}$, after some finger moves on the α–curves — again, pushing the other α–curves along — we can assume that there is a small triangle near each intersection point $\theta_i = \beta_i \cap \delta_i$, as shown shaded on the right-hand side of the figure. The sum of all these small triangles is the domain of the Whitney triangle $\widetilde{\psi}$. We denote the generators connected by $\widetilde{\psi}$ by $\widetilde{\boldsymbol{x}} \in \mathbb{T}_{\alpha,\beta}$ and $\widetilde{\boldsymbol{y}} \in \mathbb{T}_{\alpha,\delta}$; ie $\widetilde{\psi} \in \pi_2(\widetilde{\boldsymbol{x}}, \theta, \widetilde{\boldsymbol{y}})$.

The Whitney triangle $\widetilde{\psi}$ satisfies $n_z(\widetilde{\psi}) = n_w(\widetilde{\psi}) = 0$, but the constant c is not necessarily defined for it, because $\mathfrak{s}(\widetilde{\boldsymbol{x}})$ and $\mathfrak{s}(\widetilde{\boldsymbol{y}})$ might not coincide with $\mathfrak{s}|_{Y_{\alpha,\beta}}$ and $\mathfrak{s}|_{Y_{\alpha,\delta}}$, respectively, where $\mathfrak{s} \in \operatorname{Spin}^c(X)$ is the unique Spin^c structure; see (5-5). The next lemma proves that we can replace $\widetilde{\psi}$ with a Whitney triangle ψ for which the constant c is defined.

Lemma 5.22 *We can further isotope the α–curves so that there is a Whitney triangle ψ in $\pi_2(\boldsymbol{x}, \theta, \boldsymbol{y})$ satisfying*

- $n_z(\psi) = n_w(\psi) = 0$,
- $\mathfrak{s}(\boldsymbol{x}) = \mathfrak{s}|_{Y_{\alpha,\beta}}$ *and* $\mathfrak{s}(\boldsymbol{y}) = \mathfrak{s}|_{Y_{\alpha,\delta}}$.

Proof Given generators $\boldsymbol{x}' = (x_1', \dots, x_d')$ and $\boldsymbol{x}'' = (x_1'', \dots, x_d'')$ in $\mathbb{T}_\alpha \cap \mathbb{T}_\beta$, Ozsváth and Szabó associate to them [29, Definition 2.11] a class $\varepsilon(\boldsymbol{x}', \boldsymbol{x}'') \in H_1(Y_{\alpha,\beta})$. Choose 1–chains $a \subset \boldsymbol{\alpha}$ and $b \subset \boldsymbol{\beta}$ such that

$$\partial a = \partial b = x_1'' + \dots + x_d'' - x_1' - \dots - x_d'.$$

Then $a - b$ represents an element of $H_1(\Sigma)$ whose image in $H_1(Y_{\alpha,\beta})$ under the inclusion map is $\varepsilon(\boldsymbol{x}', \boldsymbol{x}'')$. Ozsváth and Szabó proved [29, Lemma 2.19] that

$$\mathfrak{s}(\boldsymbol{x}'') - \mathfrak{s}(\boldsymbol{x}') = \operatorname{PD}(\varepsilon(\boldsymbol{x}', \boldsymbol{x}'')). \tag{5-6}$$

Consider the Whitney triangle $\widetilde{\psi} \in \pi_2(\widetilde{\boldsymbol{x}}, \theta, \widetilde{\boldsymbol{y}})$ defined above, and whose domain is shown in Figure 2. Its domain is the disjoint union of d triangles $\widetilde{T}_1, \dots, \widetilde{T}_d$.

We define the homology classes $h_1 \in H_1(Y_{\alpha,\beta})$ and $h_2 \in H_1(Y_{\alpha,\delta})$ as

$$h_1 = \operatorname{PD}(\mathfrak{s}|_{Y_{\alpha,\beta}} - \mathfrak{s}(\widetilde{\boldsymbol{x}})), \tag{5-7a}$$

$$h_2 = \operatorname{PD}(\mathfrak{s}|_{Y_{\alpha,\delta}} - \mathfrak{s}(\widetilde{\boldsymbol{y}})), \tag{5-7b}$$

where $\mathfrak{s}$ is the unique Spin^c structure on X. By Lemma 5.21, there is a homology class $h \in H_1(\Sigma)$ such that $i(h) = (h_1, h_2)$; ie h represents h_1 in $H_1(Y_{\alpha,\beta})$ and h_2 in $H_1(Y_{\alpha,\delta})$. We can represent h as $m\lambda$, where λ is a simple closed curve on Σ that satisfies the following conditions:

- λ intersects the triangle $\widetilde{T}_1$ as on the left-hand side of Figure 3,

- λ is disjoint from all the triangles $\widetilde{T}_2, \dots, \widetilde{T}_d$, and
- λ is disjoint from the basepoints z and w.

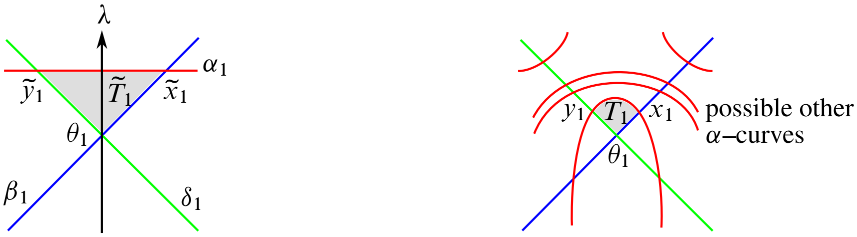


Figure 3: The pictures above show how to modify the Whitney triangle $\widetilde{\psi}$ defined in Figure 2 to obtain a Whitney triangle ψ satisfying the requirements of Lemma 5.22. The picture on the left shows the loop λ near the triangle $\widetilde{T}_1$. The picture on the right shows the new triangle T_1 in the triple Heegaard diagram obtained after performing a finger move on the α–curves along λ.

If we perform a finger move on the α–curves along the loop $m\lambda$, the result will look like the right-hand side of Figure 3. If x_1 and y_1 are as on the right-hand side of Figure 3, we define $\boldsymbol{x} = (x_1, \widetilde{x}_2, \dots, \widetilde{x}_d)$ and $\boldsymbol{y} = (y_1, \widetilde{y}_2, \dots, \widetilde{y}_d)$. Notice that, by construction,

$$\varepsilon(\widetilde{\boldsymbol{x}}, \boldsymbol{x}) = h_1 \quad \text{and} \quad \varepsilon(\widetilde{\boldsymbol{y}}, \boldsymbol{y}) = h_2. \tag{5-8}$$

Let ψ be a Whitney triangle with domain $T_1 \sqcup \widetilde{T}_2 \sqcup \dots \sqcup \widetilde{T}_d$, where T_1 is the shaded triangle on the right-hand side of Figure 3. By construction, $n_z(\psi) = n_w(\psi) = 0$. Furthermore, by (5-6), (5-8) and (5-7), we have

$$\begin{aligned}\mathfrak{s}(\boldsymbol{x}) &= \mathfrak{s}(\widetilde{\boldsymbol{x}}) + \mathrm{PD}(\varepsilon(\widetilde{\boldsymbol{x}}, \boldsymbol{x})) \\ &= \mathfrak{s}(\widetilde{\boldsymbol{x}}) + \mathrm{PD}(h_1) \\ &= \mathfrak{s}(\widetilde{\boldsymbol{x}}) + (\mathfrak{s}|_{Y_{\boldsymbol{\alpha},\boldsymbol{\beta}}} - \mathfrak{s}(\widetilde{\boldsymbol{x}})) = \mathfrak{s}|_{Y_{\alpha,\beta}}.\end{aligned}$$

Analogously, we have $\mathfrak{s}(\boldsymbol{y}) = \mathfrak{s}|_{Y_{\alpha,\delta}}$. □

Before showing that $c = 0$ for the triangle $\psi \in \pi_2(\boldsymbol{x}, \theta, \boldsymbol{y})$ constructed above, we prove that the relative Spin^c structure $\mathfrak{s}^\circ(\psi) \in \mathrm{Spin}^c(\mathcal{W}_2)$ extends to a relative Spin^c structure on $\mathcal{W}$.

Recall that $Y_1 = Y_{\alpha,\beta}$ is obtained from Y_0 by performing surgery along some framed 0–spheres. The belt circles of the 1–handles involved give rise to embedded 2–spheres $O_1, \dots, O_p \subset Y_1$. Similarly, $Y_2 = Y_{\alpha,\delta}$ is obtained from Y_3 by surgery along some framed 0–spheres, giving rise to embedded spheres $O'_1, \dots, O'_s \subset Y_2$. In Lemma 5.22, we achieved that $\mathfrak{s}(\boldsymbol{x}) = \mathfrak{s}|_{Y_{\alpha,\beta}}$ and $\mathfrak{s}(\boldsymbol{y}) = \mathfrak{s}|_{Y_{\alpha,\delta}}$. This implies that $\mathfrak{s}(\boldsymbol{x})$ extends to X_1, or equivalently, that $\langle c_1(\mathfrak{s}(\boldsymbol{x})), [O_i]\rangle = 0$ for every $i \in \{1, \dots, p\}$. Similarly, $\langle c_1(\mathfrak{s}(\boldsymbol{y})), [O'_j]\rangle = 0$ for every $j \in \{1, \dots, s\}$. However

$$\langle c_1(\mathfrak{s}(\boldsymbol{x})), [O_i]\rangle = \langle c_1(\mathfrak{s}^\circ(\boldsymbol{x})), [O_i]\rangle = \langle c_1(\mathfrak{s}^\circ(\boldsymbol{x}), t), [O_i]\rangle,$$

as $\mathfrak{s}^\circ(\boldsymbol{x})$ and $\mathfrak{s}(\boldsymbol{x})$ are represented by the same vector field on $M_1 \subset Y_1$. Since M_0 is obtained from M_1 by compressing the 2–spheres $O_1, \dots, O_p$, the equality

$$\langle c_1(\mathfrak{s}^\circ(\boldsymbol{x}), t), [O_i]\rangle = 0$$

implies $\mathfrak{s}^\circ(\boldsymbol{x})$ extends to $\mathfrak{s}_1^\circ \in \mathrm{Spin}^c(\mathcal{W}_1)$. Similarly, $\mathfrak{s}^\circ(\boldsymbol{y})$ extends to $\mathfrak{s}_3^\circ \in \mathrm{Spin}^c(\mathcal{W}_3)$. The Mayer–Vietoris sequence now implies that there is a Spin^c structure $\mathfrak{s}^\circ \in \mathrm{Spin}^c(\mathcal{W})$ such that $\mathfrak{s}^\circ|_{\mathcal{W}_1} = \mathfrak{s}_1^\circ$, $\mathfrak{s}^\circ|_{\mathcal{W}_2} = \mathfrak{s}^\circ(\psi)$ and $\mathfrak{s}^\circ|_{\mathcal{W}_3} = \mathfrak{s}_3^\circ$.

We are now ready to prove that, for the Whitney triangle ψ constructed above, $c = 0$. Recall that, by definition,

$$c = \mathcal{A}_{S_2}(\boldsymbol{y}) - \mathcal{A}_{S_1}(\boldsymbol{x}) = \langle c_2(\mathfrak{s}^\circ(\boldsymbol{y}), t), [S_2]\rangle - \langle c_1(\mathfrak{s}^\circ(\boldsymbol{x}), t), [S_1]\rangle;$$

see (5-5), Definition 5.1 and Remark 5.2. Since ψ is a Whitney triangle connecting $\boldsymbol{x}$, θ and $\boldsymbol{y}$, we have that $\mathfrak{s}^\circ(\psi)|_{M_1} = \mathfrak{s}^\circ(\boldsymbol{x})$ and $\mathfrak{s}^\circ(\psi)|_{M_2} = \mathfrak{s}^\circ(\boldsymbol{y})$, and therefore

$$c = \langle c_1(\mathfrak{s}^\circ(\psi), t), [S_2]\rangle - \langle c_1(\mathfrak{s}^\circ(\psi), t), [S_1]\rangle.$$

Notice that we can omit the restrictions of the (relative) Spin^c structures by the naturality of Chern classes.

Now the relative Spin^c structure $\mathfrak{s}^\circ(\psi)$ extends to some relative Spin^c structure $\mathfrak{s}^\circ \in \mathrm{Spin}^c(\mathcal{W})$. Then, by Lemmas 5.8 and 5.11, we have

$$c = \langle c_1(\mathfrak{s}^\circ, t), [S_3]\rangle - \langle c_1(\mathfrak{s}^\circ, t), [S_0]\rangle.$$

From Lemma 3.9, it finally follows that $c = 0$.

We can now conclude the proof of Lemma 5.13. By (5-5), for any Whitney triangle ψ in $\pi_2(\boldsymbol{x}, \theta, \boldsymbol{y})$, where $\boldsymbol{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ and $\boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\delta$ are such that $\mathfrak{s}(\boldsymbol{x})$ and $\mathfrak{s}(\boldsymbol{y})$ extend to X_1 and X_3, respectively, we have

$$\mathcal{A}_{S_2}(\boldsymbol{y}) - \mathcal{A}_{S_1}(\boldsymbol{x}) + n_z(\psi) - n_w(\psi) = 0.$$

If ψ contributes to the surgery map $f_{\mathcal{H}^1,\mathbb{L}}(\boldsymbol{x})$, then $n_w(\psi) = 0$, and it has a pseudo-holomorphic representative, so $n_z(\psi) \geq 0$. Consequently, $A_{S_2}(\boldsymbol{y}) \leq A_{S_1}(\boldsymbol{x})$, as desired. □

5K Naturality maps

Recall from [17] that, given two admissible Heegaard diagrams $\mathcal{H}$ and $\mathcal{H}'$ for the same 3–manifold Y, there is a naturality map

$$f_{\mathcal{H},\mathcal{H}'}\colon \widehat{\mathrm{CF}}(\mathcal{H}) \to \widehat{\mathrm{CF}}(\mathcal{H}'),$$

which is the composition of maps associated to isotopies of the attaching sets, handleslides, (de)stabilisations, and diffeomorphisms of the Heegaard surface isotopic to the identity in Y. On the homology, it induces an isomorphism

$$F_{\mathcal{H},\mathcal{H}'}\colon \widehat{\mathrm{HF}}(\mathcal{H}) \to \widehat{\mathrm{HF}}(\mathcal{H}')$$

that is independent of the sequence of Heegaard moves.

In our case, $\mathcal{H}$ and $\mathcal{H}'$ are doubly pointed Heegaard diagrams, which define the same decorated knot (Y, K, P). Together with Dylan Thurston, the first author proved [17, Proposition 2.37] that $\mathcal{H}$ and $\mathcal{H}'$ can be connected by a sequence of Heegaard moves *that do not cross the basepoints w and z*. If we forget about the z basepoint, this sequence induces the naturality map $f_{\mathcal{H},\mathcal{H}'}\colon \widehat{\mathrm{CF}}(\mathcal{H}) \to \widehat{\mathrm{CF}}(\mathcal{H}')$ above. As we explained, the z basepoints on $\mathcal{H}$ and $\mathcal{H}'$ induce filtrations on $\widehat{\mathrm{CF}}(\mathcal{H})$ and $\widehat{\mathrm{CF}}(\mathcal{H}')$. It follows from the work of Ozsváth and Szabó [28] and Rasmussen [31] that, if $f_{\mathcal{H},\mathcal{H}'}$ is the map associated to either an isotopy, a handleslide, a (de)stabilization, or a diffeomorphism of the Heegaard surface isotopic to the identity in Y, then it preserves the knot filtration. If $f_{\mathcal{H},\mathcal{H}'}$ is an isotopy map or a handleslide map, then the map induced on the E^1 page is the corresponding naturality map $F_{\mathcal{H},\mathcal{H}'}$ on $\widehat{\mathrm{HFK}}$; ie it is the map obtained by counting all holomorphic triangles that do not cross z. If $f_{\mathcal{H},\mathcal{H}'}$ is a (de)stabilization or diffeomorphism map, then it is an isomorphism of filtered complexes.

As the above result is only outlined in the works of Ozsváth and Szabó [28] and Rasmussen [31], we provide a bit more detail. With the techniques of this paper, we can prove the following analogue of Lemma 5.13.

Lemma 5.23 *Let K be a null-homologous knot in $Y = \#_{i=1}^{p}(S^1 \times S^2)$. Choose a Seifert surface S for K. Suppose that $\mathcal{H}$ and $\mathcal{H}'$ are admissible doubly pointed Heegaard diagrams for (Y, K, P) that only differ by an isotopy or a handleslide.*

Given an admissible doubly pointed triple diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, w, z)$ for the Heegaard move $\mathcal{H} \to \mathcal{H}'$, if $\boldsymbol{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$, then for any $\boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\delta$ that has nontrivial coefficient in the expansion of $f_{\mathcal{H},\mathcal{H}'}(\boldsymbol{x})$, we have that

$$\mathcal{F}_S(\boldsymbol{y}) \le \mathcal{F}_S(\boldsymbol{x}).$$

Furthermore, if ψ is a holomorphic triangle connecting $\boldsymbol{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$, $\theta \in \mathbb{T}_\beta \cap \mathbb{T}_\delta$ (the top-dimensional generator of $\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\delta}, w, z)$) and $\boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\delta$ that does not cross w, then

$$\mathcal{F}_S(\boldsymbol{y}) = \mathcal{F}_S(\boldsymbol{x}) - n_z(\psi). \tag{5-9}$$

Remark 5.24 The (de)stabilization and diffeomorphism maps do not appear in the statement of Lemma 5.23 because they are not triangle maps. They are already isomorphisms at the level of filtered chain complexes.

Idea of the proof As in the proof of Lemma 5.13, we let

$$c = \mathcal{A}_S(\boldsymbol{y}) - \mathcal{A}_S(\boldsymbol{x}) - n_w(\psi) + n_z(\psi),$$

and prove that this is independent of ψ, $\boldsymbol{x}$ and $\boldsymbol{y}$. The main differences from the proof of Lemma 5.13 are the following:

Triply periodic domains We closely follow the proof of Proposition 5.16. In this case, $X \cong Y \times I$, the boundary of the 4–manifold X_Δ consists of $Y \sqcup Y \sqcup Y_{\boldsymbol{\beta},\boldsymbol{\delta}}$, and the cobordisms X_1 and X_3 are replaced by identity cobordisms $Y \times I$. Finally, the proof of the injectivity of φ follows from the surjectivity of the map

$$\rho\colon H_2(Y \times I) \oplus H_2(Y \times I) \oplus H_2(\widehat{X}) \to H_2(X),$$

as noted in Remark 5.19.

Doubly periodic domains One can use Proposition 5.20 for the two copies of Y and for $Y_{\boldsymbol{\beta},\boldsymbol{\delta}}$.

Proving that $\boldsymbol{c = 0}$ This is easier than in the case of the 2–handle maps, because we already know that the naturality map preserves the graded Euler characteristic, and this forces the grading shift c to be 0. Also, as X_1 and X_3 are products, Spin^c structures automatically extend to them, hence we do not need to isotope the α–curves. □

5L Proof of Theorem 5.4

We are now ready to prove Theorem 5.4. In the proof we use the notation introduced in Section 5D, and we assume that the gluing map is the identity map, as explained in Section 5C.

Suppose that $\boldsymbol{x}$ is a generator of $\widehat{\mathrm{CF}}(\mathcal{H}^0)$ such that $f_{\mathcal{C}}(\boldsymbol{x}) \neq 0$. Let $\boldsymbol{y}$ be a generator of $\widehat{\mathrm{CF}}(\mathcal{H}^2_{\mathbb{S}})$ that appears in the expression of $f_{\mathcal{C}}(\boldsymbol{x})$ with nonzero coefficient. Then there exist generators $\boldsymbol{x}' \in \widehat{\mathrm{CF}}(\mathcal{H}^0_{\mathbb{P}})$, $\boldsymbol{x}'' \in \widehat{\mathrm{CF}}(\mathcal{H}^1)$, $\boldsymbol{y}'' \in \widehat{\mathrm{CF}}(\mathcal{H}^1_{\mathbb{L}})$ and $\boldsymbol{y}' \in \widehat{\mathrm{CF}}(\mathcal{H}^2)$ that appear with nonzero coefficient in $f_{\mathcal{H}^0,\mathbb{P}}(\boldsymbol{x})$, $f_{\mathcal{H}^0_{\mathbb{P}},\mathcal{H}^1}(\boldsymbol{x}')$, $f_{\mathcal{H}^1,\mathbb{L}}(\boldsymbol{x}'')$ and $f_{\mathcal{H}^1_{\mathbb{L}},\mathcal{H}^2}(\boldsymbol{y}'')$, respectively, and such that y appears with nonzero coefficient in $f_{\mathcal{H}^2,\mathbb{S}}(y')$. Notice that, by construction, $\mathfrak{s}(x'')$ extends to X_1 and $\mathfrak{s}(y'')$ extends to X_3.

By Lemma 5.23, we know that the naturality maps preserve the knot filtration, and by Corollaries 5.10 and 5.12 so do the maps $f_{\mathcal{H}^0,\mathbb{P}}$ and $f_{\mathcal{H}^2,\mathbb{S}}$. Finally, Lemma 5.13 proves that $\mathcal{F}_{S_2}(y'') \le \mathcal{F}_{S_1}(x'')$. By putting all these together, we obtain that

$$\mathcal{F}_{S_3}(y) = \mathcal{F}_{S_2}(y') \le \mathcal{F}_{S_2}(y'') \le \mathcal{F}_{S_1}(x'') \le \mathcal{F}_{S_1}(x') = \mathcal{F}_{S_0}(x). \tag{5-10}$$

Thus $f_{\mathcal{C}}$ is a map of filtered complexes and so, by Remark 4.6, it induces a morphism of spectral sequences.

Furthermore, each of the maps $f_{\mathcal{H}^0,\mathbb{P}}$, $f_{\mathcal{H}^0_{\mathbb{P}},\mathcal{H}^1}$, $f_{\mathcal{H}^1,\mathbb{L}}$, $f_{\mathcal{H}^1_{\mathbb{L}},\mathcal{H}^2}$ and $f_{\mathcal{H}^2,\mathbb{S}}$ is a map of filtered complexes. The map induced by $f_{\mathcal{C}}$ on the E^1 page is the composition of the maps induced by each of the above maps on the E^1 page.

We now consider the case when the inequalities in (5-10) are all equalities. Lemmas 5.8 and 5.11 imply that the maps induced by $f_{\mathcal{H}^0,\mathbb{P}}$ and $f_{\mathcal{H}^2,\mathbb{S}}$ on the E^1 page are the 1– and 3–handle maps for $\widehat{\mathrm{HFK}}$. As for the 2–handle map $f_{\mathcal{H}^1,\mathbb{L}}$, by (5-3) in Lemma 5.13, we have that $\mathcal{F}(y'') = \mathcal{F}(x'')$ if and only if there is a pseudoholomorphic triangle ψ connecting x'', θ and y'' such that $n_w(\psi) = n_z(\psi) = 0$, and in this case all such holomorphic triangles satisfy this equality. Hence, the map induced by $f_{\mathcal{H}^1,\mathbb{L}}$ on the E^1 page is the 2–handle map for $\widehat{\mathrm{HFK}}$. Finally, it follows from the discussion in Section 5K that the maps induced on the E^1 page by the naturality maps for $\widehat{\mathrm{CF}}$ are the naturality maps for $\widehat{\mathrm{HFK}}$. Alternatively, one can use (5-9) in Lemma 5.23 and argue in the same way as for the 2–handle maps.

This immediately implies that the map induced by $f_{\mathcal{C}}$ on the E^1 page is obtained by counting (for the naturality maps and the 2–handle map) the pseudoholomorphic triangles that do not cross w and z, and so it is $F_{\mathcal{C}}$.

On the other hand, the map induced by $f_{\mathcal{C}}$ on the total homology is given by counting all holomorphic triangles that do not cross w but might cross z. This is precisely the map $\widehat{F}_X\colon \widehat{\mathrm{HF}}(S^3) \to \widehat{\mathrm{HF}}(S^3)$ induced by the cobordism X. Because $H_1(X) = H_2(X) = 0$, we have $\widehat{F}_X = \mathrm{Id}_{\widehat{\mathrm{HF}}(S^3)}$ by [26, Lemma 3.4]. □

6 Concordance maps preserve the homological grading

In this section, we show that concordance maps also behave well with respect to another grading of $\widehat{\mathrm{CF}}$, namely the homological grading.

Let $\mathcal{H}$ be an admissible pointed Heegaard diagram for the closed, connected, oriented, based 3–manifold (Y, w), together with a Spin^c structure $\mathfrak{s} \in \mathrm{Spin}^c(Y)$ such that

$c_1(\mathfrak{s}) \in H^2(Y)$ is torsion. Ozsváth and Szabó [29, Section 4] showed that $\widehat{\mathrm{CF}}(\mathcal{H},\mathfrak{s})$ admits a relative $\mathbb{Z}$–grading. For generators $\boldsymbol{x},\boldsymbol{y} \in \widehat{\mathrm{CF}}(\mathcal{H},\mathfrak{s})$ and $\phi \in \pi_2(\boldsymbol{x},\boldsymbol{y})$, we have

$$\mathrm{gr}(\boldsymbol{x},\boldsymbol{y}) = \mu(\phi) - 2n_w(\phi). \tag{6-1}$$

They showed [30, Theorem 7.1] that this can be lifted to an absolute $\mathbb{Q}$–grading $\widetilde{\mathrm{gr}}$, in the sense that $\mathrm{gr}(\boldsymbol{x},\boldsymbol{y}) = \widetilde{\mathrm{gr}}(\boldsymbol{x}) - \widetilde{\mathrm{gr}}(\boldsymbol{y})$. Such a grading is called the *Maslov grading* or *homological grading*.

Example 6.1 If $Y = S^3$ with its unique Spin^c structure $\mathfrak{s}_0$, and if $\mathcal{H}$ is a Heegaard diagram of Y, then on $\widehat{\mathrm{CF}}(\mathcal{H},\mathfrak{s}_0)$ the absolute $\mathbb{Q}$–grading is actually an absolute $\mathbb{Z}$–grading. The generator of $\widehat{\mathrm{HF}}(S^3,\mathfrak{s}_0) \cong \mathbb{Z}_2$ is homogeneous of grading zero.

More generally, if $Y = \#_{i=1}^k(S^1 \times S^2)$ with Heegaard diagram $\mathcal{H}$, and $\mathfrak{s}_0 \in \mathrm{Spin}^c(Y)$ is such that $c_1(\mathfrak{s}_0) = 0$, then $\widetilde{\mathrm{gr}}$ is an absolute $\mathbb{Z}$–grading on $\widehat{\mathrm{CF}}(\mathcal{H},\mathfrak{s}_0)$.

The main result of this section is the following.

Theorem 6.2 *Let $\mathcal{C}$ be a decorated concordance from (S^3,K_0,P_0) to (S^3,K_1,P_1), and let $\mathcal{H}_i$ be an admissible doubly pointed diagram of (S^3,K_i,P_i) for $i \in \{0,1\}$. Then, the chain map*

$$f_{\mathcal{C}}\colon \widehat{\mathrm{CF}}(\mathcal{H}_0) \to \widehat{\mathrm{CF}}(\mathcal{H}_1)$$

preserves the absolute homological grading; that is, if $x \in \widehat{\mathrm{CF}}(\mathcal{H}_0)$ is $\widetilde{\mathrm{gr}}$–homogeneous, so is $f_{\mathcal{C}}(x)$, and if $f_{\mathcal{C}}(x) \neq 0$, then

$$\widetilde{\mathrm{gr}}(f_{\mathcal{C}}(x)) = \widetilde{\mathrm{gr}}(x).$$

Remark 6.3 Notice that the statement of Theorem 6.2 is stronger than the fact that $f_{\mathcal{C}}$ preserves the Maslov filtration. We actually claim that the Maslov grading is not decreased by $f_{\mathcal{C}}$.

Idea of the proof We proceed similarly to the proof of Theorem 5.4, and use the notation from Section 5D and Figure 1. As the diffeomorphism D constructed in Section 5C induces a homomorphism D_* that preserves the homological grading, we can assume the gluing map is trivial and we are dealing with a special cobordism.

First, we prove that, in the right Spin^c structure, the maps $f_{\mathcal{H}^0,\mathbb{P}}$, $f_{\mathcal{H}^0_{\mathbb{P}},\mathcal{H}^1}$, $f_{\mathcal{H}^1,\mathbb{L}}$, $f_{\mathcal{H}^1_{\mathbb{L}},\mathcal{H}^2}$ and $f_{\mathcal{H}^2,\mathbb{S}}$ each preserve the relative Maslov grading gr. This is only implicit in the work of Ozsváth and Szabó [30], so we provide more detail. Then we show that the absolute grading shift of $f_{\mathcal{C}}$, which is the composition of all the above maps, is zero.

For the 1– and 3–handle maps $f_{\mathcal{H}^0,\mathbb{P}}$ and $f_{\mathcal{H}^2,\mathbb{S}}$, it is straightforward to check that the relative Maslov grading is preserved using (6-1) above.

Now consider the 2–handle map $f_{\mathcal{H}^1,\mathbb{L}}$. Let $(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},w,z)$ be an admissible triple Heegaard diagram subordinate to a bouquet for $\mathbb{L}$. For generators $\boldsymbol{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ and $\boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\delta$ such that $\mathfrak{s}(\boldsymbol{x}) = \mathfrak{s}|_{Y_{\alpha,\beta}}$ and $\mathfrak{s}(\boldsymbol{y}) = \mathfrak{s}|_{Y_{\alpha,\delta}}$, where $\mathfrak{s}$ denotes the unique Spin^c structure on X, and for every Whitney triangle $\psi \in \pi_2(\boldsymbol{x},\theta,\boldsymbol{y})$, we let

$$d = \widetilde{\mathrm{gr}}(\boldsymbol{y}) - \widetilde{\mathrm{gr}}(\boldsymbol{x}) + \mu(\psi) - 2n_w(\psi).$$

We show that d is independent of ψ, $\boldsymbol{x}$ and $\boldsymbol{y}$. Since the triangles ψ contributing to $f_{\mathcal{H}^1,\mathbb{L}}$ have $\mu(\psi) = 0$ and $n_w(\psi) = 0$, it follows that the absolute grading is shifted by d, so the relative grading is preserved.

We already know from the work of Ozsváth and Szabó [29] that the naturality maps $f_{\mathcal{H}^0_{\mathbb{P}},\mathcal{H}^1}$ and $f_{\mathcal{H}^1_{\mathbb{L}},\mathcal{H}^2}$ preserve the relative homological grading gr. Alternatively, this can also be shown using the techniques of Section 5K.

Finally, $f_{\mathcal{C}}$, which is the composition of all the above maps, preserves the relative homological grading, or equivalently, it shifts the absolute homological grading by some constant e. This implies that, for every $r \in \mathbb{N}$, the map $E^r(f_{\mathcal{C}})$ shifts the homological grading by the same constant e independent of r. Since we know that the map in total homology is $\mathrm{Id}_{\widehat{\mathrm{HF}}(S^3)}$ and preserves the absolute grading by [26, Lemma 3.4], it immediately follows that $e = 0$. □

The rest of this section is devoted to filling in the details of the above outline.

6A Spin^c structures

Let $\mathfrak{s}$ be the unique Spin^c structure on X. Then

$$f_{\mathcal{C}} = f_{\mathcal{C},\mathfrak{s}} = f_{\mathcal{H}^2,\mathbb{S},\mathfrak{s}} \circ \cdots \circ f_{\mathcal{H}^0,\mathbb{P},\mathfrak{s}},$$

where the restrictions of $\mathfrak{s}$ are omitted for the sake of clarity.

So it suffices to consider the above maps in the Spin^c structure $\mathfrak{s}$. In the rest of the section, we will focus on the maps $f_{\mathcal{H}^2,\mathbb{S},\mathfrak{s}},\dots,f_{\mathcal{H}^0,\mathbb{P},\mathfrak{s}}$, and for simplicity, we will denote the restrictions of $\mathfrak{s}$ by the same letter.

6B 1– and 3–handles

The 1–handle map $f_{\mathcal{H}^0,\mathbb{P},\mathfrak{s}}$ satisfies the following.

Lemma 6.4 *Let $x'', \tilde{x}'' \in \widehat{\mathrm{CF}}(\mathcal{H}^0, \mathfrak{s})$ be generators. Then*

$$\mathrm{gr}(x'', \tilde{x}'') = \mathrm{gr}(f_{\mathcal{H}^0,\mathbb{P},\mathfrak{s}}(x''), f_{\mathcal{H}^0,\mathbb{P},\mathfrak{s}}(\tilde{x}''));$$

ie the relative homological grading is preserved under the 1–handle map.

Proof Let $\phi \in \pi_2(x'', \tilde{x}'')$. Then the domain of ϕ also represents a Whitney disk between $f_{\mathcal{H}^0,\mathbb{P}}(x'')$ and $f_{\mathcal{H}^0,\mathbb{P}}(\tilde{x}'')$ in the Heegaard diagram $\mathcal{H}^0_{\mathbb{P}}$ that we also denote by ϕ. By (6-1), we have

$$\mathrm{gr}(x'', \tilde{x}'') = \mu(\phi) - 2n_w(\phi) = \mathrm{gr}(f_{\mathcal{H}^0,\mathbb{P},\mathfrak{s}}(x''), f_{\mathcal{H}^0,\mathbb{P},\mathfrak{s}}(\tilde{x}'')). \qquad \square$$

A dual argument gives the following result for the 3–handle map $f_{\mathcal{H}^2,\mathbb{S},\mathfrak{s}}$.

Lemma 6.5 *Let $y', \tilde{y}' \in \widehat{\mathrm{CF}}(\mathcal{H}^2, \mathfrak{s})$ be generators such that $f_{\mathcal{H}^2,\mathbb{S},\mathfrak{s}}(y') \neq 0$ and $f_{\mathcal{H}^2,\mathbb{S},\mathfrak{s}}(\tilde{y}') \neq 0$. Then*

$$\mathrm{gr}(y', \tilde{y}') = \mathrm{gr}(f_{\mathcal{H}^2,\mathbb{S},\mathfrak{s}}(y'), f_{\mathcal{H}^2,\mathbb{S},\mathfrak{s}}(\tilde{y}'));$$

ie the relative homological grading is preserved under the 3–handle map.

6C 2–handles

For 2–handles, we have the following.

Lemma 6.6 *Let $x, \tilde{x} \in \widehat{\mathrm{CF}}(\mathcal{H}^1)$ be generators such that $\mathfrak{s}(x) = \mathfrak{s}(\tilde{x}) = \mathfrak{s}$. Then $f_{\mathcal{H}^1,\mathbb{L},\mathfrak{s}}(x)$ and $f_{\mathcal{H}^1,\mathbb{L},\mathfrak{s}}(\tilde{x})$ are $\widetilde{\mathrm{gr}}$–homogeneous, and if they are nonzero, then*

$$\mathrm{gr}(x, \tilde{x}) = \mathrm{gr}(f_{\mathcal{H}^1,\mathbb{L},\mathfrak{s}}(x), f_{\mathcal{H}^1,\mathbb{L},\mathfrak{s}}(\tilde{x})).$$

Proof For $x \in \widehat{\mathrm{CF}}(\mathcal{H}^1)$, $y \in \widehat{\mathrm{CF}}(\mathcal{H}^1_{\mathbb{L}})$ and $\psi \in \pi_2(x, \theta, y)$ such that $\mathfrak{s}(\psi) = \mathfrak{s}$, let

$$d = \widetilde{\mathrm{gr}}(y) - \widetilde{\mathrm{gr}}(x) + \mu(\psi) - 2n_w(\psi). \tag{6-2}$$

First, we check that d is independent of ψ. As in the proof of Lemma 5.13, it suffices to show that, for every triply periodic domain $\mathcal{P}$,

$$\mu(\mathcal{P}) = 2n_w(\mathcal{P}). \tag{6-3}$$

Since, by Proposition 5.16, every triply periodic domain is the sum of doubly periodic domains, it is sufficient to prove (6-3) in the case of doubly periodic domains in Heegaard diagrams of $Y_{\alpha,\beta}$, $Y_{\alpha,\delta}$ and $Y_{\beta,\delta}$.

Consider, for example, $Y_{\alpha,\beta}$ and $z \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$, with a periodic domain $\mathcal{P} \in \Pi_{\alpha,\beta}$ based at z. As $\mathfrak{s}(z)$ extends to the cobordism X_1, we see that $c_1(\mathfrak{s}(z))$ vanishes on the belt spheres of the 1–handles. Furthermore, since $H(\mathcal{P}) \in H_2(Y)$ is a linear combination of the belt spheres, we obtain that

$$\langle c_1(\mathfrak{s}(z)), H(\mathcal{P})\rangle = 0.$$

By the work of Ozsváth and Szabó [29, Theorem 4.9] and Lipshitz [20, Lemma 4.10],

$$\mu(\mathcal{P}) = \langle c_1(\mathfrak{s}(z)), H(\mathcal{P})\rangle + 2n_w(\mathcal{P}),$$

and the result follows. This proves that d is independent of ψ.

Next, we check that d is independent of $\boldsymbol{x}$ and $\boldsymbol{y}$. Let $\tilde{\boldsymbol{x}}$ be another generator of $\widehat{\mathrm{CF}}(\mathcal{H}^1)$ such that $\mathfrak{s}(\tilde{\boldsymbol{x}}) = \mathfrak{s}$. Then there is a Whitney disk $\phi \in \pi_2(\tilde{\boldsymbol{x}}, \boldsymbol{x})$, hence $\phi \# \psi \in \pi_2(\tilde{\boldsymbol{x}}, \theta, \boldsymbol{y})$. Then, by (6-1),

$$\begin{aligned} d &= \widetilde{\mathrm{gr}}(\boldsymbol{y}) - \widetilde{\mathrm{gr}}(\boldsymbol{x}) + \mu(\psi) - 2n_w(\psi) \\ &= \widetilde{\mathrm{gr}}(\boldsymbol{y}) - \widetilde{\mathrm{gr}}(\boldsymbol{x}) + \mu(\psi) - 2n_w(\psi) + (\widetilde{\mathrm{gr}}(\boldsymbol{x}) - \widetilde{\mathrm{gr}}(\tilde{\boldsymbol{x}}) + \mu(\phi) - 2n_w(\phi)) \\ &= \widetilde{\mathrm{gr}}(\boldsymbol{y}) - \widetilde{\mathrm{gr}}(\tilde{\boldsymbol{x}}) + \mu(\phi \# \psi) - 2n_w(\phi \# \psi). \end{aligned}$$

Thus, d is independent of $\boldsymbol{x}$. An analogous argument shows independence of $\boldsymbol{y}$.

Finally, all the holomorphic triangles that appear in the definition of the map $f_{\mathcal{H}^1,\mathbb{L},\mathfrak{s}}$ satisfy $\mu(\psi) = 0$ and $n_w(\psi) = 0$. Then, it follows from (6-2) that $f_{\mathcal{H}^1,\mathbb{L},\mathfrak{s}}$ increases the absolute grading $\widetilde{\mathrm{gr}}$ by d. In particular, it preserves the relative grading gr. □

6D Naturality maps

We already know from the work of Ozsváth and Szabó [29] that the naturality maps preserve the Maslov grading. Alternatively, one can prove that the handleslide and isotopy maps preserve the Maslov grading using the techniques of Lemma 6.6. The (de)stabilization maps are already isomorphisms on the chain level.

6E Proof of Theorem 6.2

As explained in Section 6A,

$$f_{\mathcal{C}} = f_{\mathcal{H}^2,\mathbb{S},\mathfrak{s}} \circ f_{\mathcal{H}^1_{\mathbb{L}},\mathcal{H}^2,\mathfrak{s}} \circ f_{\mathcal{H}^1,\mathbb{L},\mathfrak{s}} \circ f_{\mathcal{H}^0_{\mathbb{P}},\mathcal{H}^1,\mathfrak{s}} \circ f_{\mathcal{H}^0,\mathbb{P},\mathfrak{s}}.$$

All the above maps preserve the relative Maslov grading by Lemmas 6.4, 6.5 and 6.6, so $f_{\mathcal{C}}$ shifts the absolute Maslov grading by some constant e. It follows that the maps induced between the spectral sequences $E^r(f_{\mathcal{C}})$ shift the absolute Maslov grading by the same constant e. On the other hand, the map in total homology is $\mathrm{Id}_{\widehat{\mathrm{HF}}(S^3)}$, which is homogeneous of degree 0, so we obtain that $e = 0$. □

References

[1] **G Arone**, **M Kankaanrinta**, *On the functoriality of the blow-up construction*, Bull. Belg. Math. Soc. Simon Stevin 17 (2010) 821–832 MR

[2] **M Atiyah**, *Topological quantum field theories*, Inst. Hautes Études Sci. Publ. Math. 68 (1988) 175–186 MR

[3] **J A Baldwin**, **W D Gillam**, *Computations of Heegaard–Floer knot homology*, J. Knot Theory Ramifications 21 (2012) art. id. 1250075, 65 pages MR

[4] **C Blanchet**, **V Turaev**, *Axiomatic approach to topological quantum field theory*, from "Encyclopedia of Mathematical Physics" (J-P Françoise, G L Naber, T S Tsun, editors), Academic Press, Oxford (2006) 232 – 234

[5] **R H Fox**, *Some problems in knot theory*, from "Topology of 3–manifolds and related topics", Prentice-Hall, Englewood Cliffs, NJ (1962) 168–176 MR

[6] **M Freedman**, **R Gompf**, **S Morrison**, **K Walker**, *Man and machine thinking about the smooth 4–dimensional Poincaré conjecture*, Quantum Topol. 1 (2010) 171–208 MR

[7] **D Gabai**, *Foliations and the topology of* 3*–manifolds*, J. Differential Geom. 18 (1983) 445–503 MR

[8] **P Ghiggini**, *Knot Floer homology detects genus-one fibred knots*, Amer. J. Math. 130 (2008) 1151–1169 MR

[9] **M Hedden**, *On knot Floer homology and cabling*, PhD thesis, Columbia University (2005) Available at `http://search.proquest.com/docview/305015665`

[10] **K Honda**, *On the classification of tight contact structures, II*, J. Differential Geom. 55 (2000) 83–143 MR

[11] **K Honda**, **W Kazez**, **G Matić**, *Contact structures, sutured Floer homology and TQFT*, preprint (2008) arXiv

[12] **M Jacobsson**, *An invariant of link cobordisms from Khovanov homology*, Algebr. Geom. Topol. 4 (2004) 1211–1251 MR

[13] **A Juhász**, *Holomorphic discs and sutured manifolds*, Algebr. Geom. Topol. 6 (2006) 1429–1457 MR

[14] **A Juhász**, *Floer homology and surface decompositions*, Geom. Topol. 12 (2008) 299–350 MR

[15] **A Juhász**, *The sutured Floer homology polytope*, Geom. Topol. 14 (2010) 1303–1354 MR

[16] **A Juhász**, *Cobordisms of sutured manifolds and the functoriality of link Floer homology*, Adv. Math. 299 (2016) 940–1038 MR

[17] **A Juhász**, **D Thurston**, *Naturality and mapping class groups in Heegaard Floer homology*, preprint (2012) arXiv

[18] **Ç Karakurt**, **T Lidman**, *Rank inequalities for the Heegaard Floer homology of Seifert homology spheres*, Trans. Amer. Math. Soc. 367 (2015) 7291–7322 MR

[19] **P B Kronheimer**, **T S Mrowka**, *Gauge theory and Rasmussen's invariant*, J. Topol. 6 (2013) 659–674 MR

[20] **R Lipshitz**, *A cylindrical reformulation of Heegaard Floer homology*, Geom. Topol. 10 (2006) 955–1097 MR

[21] **R Lutz**, *Structures de contact sur les fibrés principaux en cercles de dimension trois*, Ann. Inst. Fourier (Grenoble) 27 (1977) 1–15 MR

[22] **C Manolescu**, **P Ozsváth**, *On the Khovanov and knot Floer homologies of quasi-alternating links*, from "Proceedings of Gökova Geometry–Topology Conference 2007" (S Akbulut, T Önder, R J Stern, editors), GGT, Gökova (2008) 60–81 MR

[23] **C Manolescu**, **P Ozsváth**, **S Sarkar**, *A combinatorial description of knot Floer homology*, Ann. of Math. 169 (2009) 633–660 MR

[24] **J McCleary**, *A user's guide to spectral sequences*, 2nd edition, Cambridge Studies in Advanced Mathematics 58, Cambridge University Press (2001) MR

[25] **Y Ni**, *Knot Floer homology detects fibred knots*, Invent. Math. 170 (2007) 577–608 MR Erratum in 170 (2009) 235–238

[26] **P Ozsváth**, **Z Szabó**, *Knot Floer homology and the four-ball genus*, Geom. Topol. 7 (2003) 615–639 MR

[27] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and genus bounds*, Geom. Topol. 8 (2004) 311–334 MR

[28] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and knot invariants*, Adv. Math. 186 (2004) 58–116 MR

[29] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. 159 (2004) 1027–1158 MR

[30] **P Ozsváth**, **Z Szabó**, *Holomorphic triangles and invariants for smooth four-manifolds*, Adv. Math. 202 (2006) 326–400 MR

[31] **J A Rasmussen**, *Floer homology and knot complements*, PhD thesis, Harvard University (2003) MR Available at `http://search.proquest.com/docview/305332635`

[32] **J Rasmussen**, *Khovanov homology and the slice genus*, Invent. Math. 182 (2010) 419–447 MR

[33] **Y W Rong**, *Degree one maps between geometric* 3*–manifolds*, Trans. Amer. Math. Soc. 332 (1992) 411–436 MR

[34] **S Sarkar**, *Grid diagrams and the Ozsváth–Szabó tau-invariant*, Math. Res. Lett. 18 (2011) 1239–1257 MR

[35] **S Sarkar**, *Moving basepoints and the induced automorphisms of link Floer homology*, Algebr. Geom. Topol. 15 (2015) 2479–2515 MR

[36] **D S Silver**, **W Whitten**, *Knot group epimorphisms*, J. Knot Theory Ramifications 15 (2006) 153–166 MR

[37] **J Stallings**, *On fibering certain 3–manifolds*, from "Topology of 3–manifolds and related topics", Prentice-Hall, Englewood Cliffs, NJ (1962) 95–100 MR

[38] **D W Sumners**, *Invertible knot cobordisms*, Comment. Math. Helv. 46 (1971) 240–256 MR

Mathematical Institute, University of Oxford
Andrew Wiles Building, Radcliffe Observatory Quarter
Woodstock Road, Oxford, OX2 6GG, United Kingdom

Department of Mathematics, Imperial College London
180 Queen's Gate, London, SW7 2AZ, United Kingdom

juhasza@maths.ox.ac.uk, m.marengon13@imperial.ac.uk

http://www.maths.ox.ac.uk/people/andras.juhasz,
http://www.imperial.ac.uk/people/m.marengon13

Proposed: Ciprian Manolescu Received: 18 September 2015
Seconded: Peter Teichner, Ronald Stern Revised: 25 January 2016

Guidelines for Authors

Submitting a paper to Geometry & Topology

Papers must be submitted using the upload page at the GT website. You will need to choose a suitable editor from the list of editors' interests and to supply MSC codes.

Preparing your article for Geometry & Topology

The normal language used by the journal is English. Articles written in other languages are acceptable, provided your chosen editor is comfortable with the language and you supply an additional English version of the abstract.

At the time of submission you need only supply a PDF file. Once accepted for publication, the paper must be supplied in LaTeX, preferably using the journal's class file. More information on preparing articles in LaTeX for publication in GT is available on the GT website.

`arXiv` papers

GT papers are published simultaneously on the `arXiv`. If your paper has previously been deposited there, we shall need the `arXiv` password at acceptance time, to submit the published version to the `arXiv`.

References

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited at least once in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used, but will be converted to the house style (see a current issue for examples).

Figures

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple, Mathematica or XFig. Many drawing tools such as Adobe Illustrator and Aldus FreeHand also produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, check with production by sending an email to gt@msp.warwick.ac.uk.

Proofs

Page proofs will be made available to authors (or to the designated corresponding author) in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

GEOMETRY & TOPOLOGY

Volume 20 Issue 6 (pages 3057–3673) 2016

1465-3060(2016)20:6;1-U