



Geometry & Topology

Volume 24 (2020)

Issue 1 (pages 1–532)



GEOMETRY & TOPOLOGY

msp.org/gt

EDITORS

MANAGING EDITORS

Yasha Eliashberg
eliash-gt@math.stanford.edu
Department of Mathematics
Stanford University
Stanford CA 94305-2125
USA

Colin Rourke
gt@maths.warwick.ac.uk
Mathematics Institute
University of Warwick
Coventry CV4 7AL
UK

BOARD OF EDITORS

Mark Behrens	Massachusetts Institute of Technology mbehrens@math.mit.edu	Robion Kirby	University of California, Berkeley kirby@math.berkeley.edu
Martin Bridson	Imperial College, London m.bridson@ic.ac.uk	Frances Kirwan	University of Oxford frances.kirwan@balliol.oxford.ac.uk
Jim Bryan	University of British Columbia jbryan@math.ubc.ca	John Lott	University of California at Berkeley lott@math.berkeley.edu
Dmitri Burago	Pennsylvania State University burago@math.psu.edu	Haynes Miller	Massachusetts Institute of Technology hmr@math.mit.edu
Danny Calegari	California Institute of Technology dannyc@caltech.edu	Shigeyuki Morita	University of Tokyo morita@ms.u-tokyo.ac.jp
Ralph Cohen	Stanford University ralph@math.stanford.edu	John Morgan	Columbia University jm@math.columbia.edu
Tobias Colding	Massachusetts Institute of Technology colding@math.mit.edu	Tom Mrowka	Massachusetts Institute of Technology mrowka@math.mit.edu
Simon Donaldson	Imperial College, London s.donaldson@ic.ac.uk	Walter Neumann	Columbia University neumann@math.columbia.edu
Bill Dwyer	University of Notre Dame dwyer.1@nd.edu	Jean-Pierre Otal	Université d'Orléans jean-pierre.otal@univ-orleans.fr
Benson Farb	University of Chicago farb@math.uchicago.edu	Peter Ozsváth	Columbia University ozsvath@math.columbia.edu
Steve Ferry	Rutgers University sferry@math.rutgers.edu	Leonid Polterovich	Tel Aviv University polterov@post.tau.ac.il
Ron Fintushel	Michigan State University ronfint@math.msu.edu	Peter Teichner	University of California, Berkeley teichner@math.berkeley.edu
Mike Freedman	Microsoft Research michaelf@microsoft.com	Richard Thomas	Imperial College, London richard.thomas@imperial.ac.uk
David Gabai	Princeton University gabai@princeton.edu	Gang Tian	Massachusetts Institute of Technology tian@math.mit.edu
Lothar Götsche	Abdus Salam Int. Centre for Th. Physics gotsche@ictp.trieste.it	Ulrike Tillmann	Oxford University tillmann@maths.ox.ac.uk
Cameron Gordon	University of Texas gordon@math.utexas.edu	Anna Wienhard	Universität Heidelberg wienhard@mathi.uni-heidelberg.de
Jesper Grodal	University of Copenhagen jg@math.ku.dk		

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

The subscription price for 2020 is US \$635/year for the electronic version, and \$910/year (+ \$65, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP. Geometry & Topology is indexed by [Mathematical Reviews](#), [Zentralblatt MATH](#), [Current Mathematical Publications](#) and the [Science Citation Index](#).

Geometry & Topology (ISSN 1465-3060 printed, 1364-0380 electronic) is published 7 times per year and continuously online, by Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840. Periodical rate postage paid at 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/>

© 2020 Mathematical Sciences Publishers

Coalgebraic formal curve spectra and spectral jet spaces

ERIC PETERSON

We import into homotopy theory the algebrogeometric construction of the cotangent space of a geometric point on a scheme. Specializing to the category of spectra local to a Morava K –theory of height d , we show that this can be used to produce a choice-free model of the determinantal sphere as well as an efficient Picard-graded cellular decomposition of $K(\mathbb{Z}_p, d+1)$. Coupling these ideas to work of Westerland, we give a “Snaith’s theorem” for the Iwasawa extension of the $K(d)$ –local sphere.

55N22; 55P20, 55P60

1 Introduction

Much of modern chromatic homotopy theory is underpinned by a process that converts spaces to formal schemes: given an even-periodic cohomology theory E and a space X , if E^*X is suitably nice we can define a formal scheme X_E by the formula

$$X_E := \mathrm{Spf} E^0 X,$$

where the formal topology can be taken to come from the finite subcomplexes of a cellular structure on X . The very definition of complex-orientability is designed so that this construction carries $\mathbb{C}P^\infty$ to the formal affine line:

$$\mathrm{Spf} E^0 \mathbb{C}P^\infty = \{\mathrm{Spec} E^0 \mathbb{C}P^n\}_n \cong \{\mathrm{Spec} E^0[x]/x^{n+1}\}_n =: \hat{\mathbb{A}}^1_{/\mathrm{Spec} E^0}.$$

The H –group structure of $\mathbb{C}P^\infty$ then endows the formal scheme $\mathbb{C}P^\infty_E$ with the structure of a formal Lie group, which history has shown to encode quite a lot of information about E itself. In fact, many familiar values of X are carried to other familiar schemes—for instance, $BU(n)_E$ models the scheme of effective Weil divisors on $\mathbb{C}P^\infty_E$ of rank n , $\mathbb{H}P^\infty_E$ is sent to a formal curve, and the symplectification map is sent to a degree-2 isogeny $\mathbb{C}P^\infty_E \rightarrow \mathbb{H}P^\infty_E$. The reach of this construction is maximized when $E = E_\Gamma$ is taken to be the Morava E –theory associated to a formal group Γ of finite height d over a perfect field k of positive characteristic p . This cohomology theory is designed so that the formal Lie group $\mathbb{C}P^\infty_{E_\Gamma}$ is a versal deformation of Γ , but a great many additional theorems have been proven about its assignment from spaces to formal schemes, as partially tabulated in Figure 1. The most remarkable

X	X_{E_Γ}
one-point space	Lubin–Tate deformation space for Γ
$\mathbb{C}P^\infty$	versal deformation $\tilde{\Gamma}$ of Γ
$BU(n)$	effective Weil divisors of rank n on $\tilde{\Gamma}$
$BU \times \mathbb{Z}$	stable Weil divisors on $\tilde{\Gamma}$
BU	stable Weil divisors on $\tilde{\Gamma}$ of virtual rank 0
$BS^1[p^j]$	the p^j –torsion subgroup $\tilde{\Gamma}[p^j]$
BA^*	the mapping scheme $\text{FORMALGROUPS}(A, \tilde{\Gamma})$
$B^q(S^1)$	the q^{th} exterior power of $\tilde{\Gamma}$
$B^q(S^1[p^j])$	the p^j –torsion subgroup of $\tilde{\Gamma}^{\wedge q}$
BA^* (mod annihilators of $x(a)$)	level- A structures on $\tilde{\Gamma}$
$B\Sigma_n$ (mod transfers)	subgroup divisors of rank n on $\tilde{\Gamma}$
BSU	special stable Weil divisors on $\tilde{\Gamma}$ of virtual rank 0, equivalently, $C_2(\tilde{\Gamma}) := \text{Sym}^2_{\text{Div } \tilde{\Gamma}} \text{Div}_0 \tilde{\Gamma}$
$BU[6, \infty)$	$C_3(\tilde{\Gamma})$
$\mathbb{H}P^\infty$	a formal curve $\bar{\Gamma}$ double-covered by $\tilde{\Gamma}$
BSp	stable Weil divisors on $\bar{\Gamma}$
BO	$\text{Div } \bar{\Gamma} \times_{\text{Div } \bar{\Gamma}[2]} \text{Div } \tilde{\Gamma}[2]$
BSO	$\text{Div } \bar{\Gamma} \times_{\text{Div } \bar{\Gamma}[2]} \text{SDiv } \tilde{\Gamma}[2]$
Spin/SU	$C_2(\tilde{\Gamma})/([a, -a])$
$B\text{Spin}$ (when $\text{ht } \Gamma \leq 2$)	$C_2(\tilde{\Gamma})/([a, b] + [-a, -b])$
$B\text{String}$ (when $\text{ht } \Gamma \leq 2$)	$C_3(\tilde{\Gamma})/([a, b, -a - b])$
$\vdots$	$\vdots$

Figure 1: Assignments from spaces to formal schemes via Morava E –theory.

feature of this table is that when Γ varies and the space X is fixed, the same family of formal schemes appears as output. This inspires us to consider X itself as playing some scheme-like role at the level of homotopy theory and to study algebrogeometric operations on X intrinsically.

The goal of the present work is to use a spectrum-level construction analogous to that of Hopf algebra homology to define the “tangent spectrum” of certain extremely nice spaces X . We first build up to this definition and then justify its utility by demonstrating

that it has the expected behavior in E_Γ –homology. The details of this computation inform us at what homotopical level we expect our construction to behave well: we find that it lives naturally in the category of spectra localized at E_Γ –cohomology, which we abbreviate to the category of Γ –local spectra.¹ With this established, we then use the computational apparatus to extract further interesting constructions within the Γ –local category. The main result of this paper is thus:

Theorem 1 (Definition 2.4.5 and Theorem 3.3.12) *For X any pointed space and Γ any finite-height formal group over a perfect field k , there is a sequence of Γ –local spectra, functorial in X ,*

$$\begin{array}{ccccccc} C_0^\infty & \longrightarrow & C_1^\infty & \longrightarrow & C_2^\infty & \longrightarrow & \cdots \\ \text{fiber} \uparrow & & \text{fiber} \uparrow & & \text{fiber} \uparrow & & \\ C_0^0 & & C_1^1 & & C_2^2 & & \cdots \end{array}$$

with natural equivalences

$$C_0^0 \simeq \mathbb{S}, \quad C_0^\infty \simeq \Sigma_+^\infty X, \quad C_1^\infty \simeq \Sigma^\infty X.$$

If X_{E_Γ} is a formal curve (ie $E_\Gamma^0 X$ is abstractly isomorphic to a univariate formal power series E_Γ^0 –algebra) and $p \gg \text{ht } \Gamma$,² then

$$T_0^*(X_{E_\Gamma}) \cong E_\Gamma^0(C_1^1),$$

where T_0^* indicates the cotangent space at the origin. Moreover, $C_k^k = (C_1^1)^{\wedge k}$ for finite k , and $C_\infty^\infty = *$.

In the case that X_{E_Γ} is a formal curve, we are thus motivated to take C_1^1 as a definition of $T_+ X$, the tangent space of X at its natural pointing $\mathbb{S}^0 \rightarrow \Sigma_+^\infty X$. Drawing from the algebrogeometric interpretation, we refer to the generic procedure as the *jet-space decomposition*.

The known collection of candidate curves to which to apply this result is quite sparse: in the table above, we have already noted that $X = \mathbb{CP}^\infty$ has the required property, as does $B^d S^1$, the top exterior power of $\widetilde{\Gamma}$. In the first case, this procedure recovers the cellular decomposition of $\mathbb{CP}^\infty$ into even spheres, attached along homotopy classes

¹Readers familiar with chromatic homotopy theory will better recognize “ Γ –local spectra” as “ $K(d)$ –local spectra”, where $K(d)$ is Morava’s d^{th} extraordinary K –theory. This localization, therefore, does not truly depend on Γ but rather only on its height d ; see Hopkins and Smith [22, Proposition 1.9].

²This hypothesis comes from our proofs in the appendix and seems likely to be unnecessary for the identification of C_1^1 , where Lemma A.2.8 requires $p > 2$ and $d^2 > 2(p - 1)$. However, it may indeed be necessary for the identification of C_k^k for general k , where the same bound appears in Theorem 3.3.4.

graded by $\Sigma^{-1}(\mathbb{CP}^1)^{\wedge k}$. This second case is particularly interesting because of its coupling to the Γ -local category via the integer d . We will show the following result, which is highly specific to Morava E -theory and to the choice of Γ :

Lemma 2 (Theorem 3.3.9; cf Westerland [41, Proposition 3.21]) *Setting $X = \Sigma_+^\infty B^d S^1$ for $d = \text{ht } \Gamma$, then X_{E_Γ} is a formal curve and $\text{Aut } \Gamma$ acts on the tangent space through the determinant representation.*

Corollary 3 (Corollary 3.3.10) *For $p \gg \text{ht } \Gamma$, the Γ -local spectrum $T_+ \Sigma_+^\infty B^d S^1$ gives a choice-free³ model of $\mathbb{S}^{\text{det}}$, the determinantal sphere of Gross and Hopkins (cf Goerss, Henn, Mahowald and Rezk [15, Remark 2.5] and Hopkins and Gross [19, Theorem 6 and Corollary 3]).*

Corollary 4 (Definition 2.4.5 and Theorem 3.3.12) *For $p \gg \text{ht } \Gamma$, the spectrum $\Sigma_+^\infty B^d S^1$ has a Γ -local Picard-graded cellular decomposition, with one cell of the form $C_k^k \simeq (\mathbb{S}^{\text{det}})^{\wedge k}$ for each $k \geq 0$. This decomposition is **efficient** in the sense that the associated Atiyah–Hirzebruch spectral sequence collapses at E_2 .*

Alternatively, the cellular decomposition of $\mathbb{CP}^\infty$ can also be interpreted as the skeletal filtration for the bar construction $B(\mathbb{C}^\times)$. This, too, admits generalization:

Corollary 5 (Theorem 2.5.6) *There is an A_∞ -ring structure on $(\Sigma^{-1} \mathbb{S}^{\text{det}})_+$ and a Γ -local equivalence*

$$B((\Sigma^{-1} \mathbb{S}^{\text{det}})_+) \xrightarrow{\sim} \Sigma_+^\infty B^d S^1.$$

Additionally, in Section 3.4 we explore the direct connections between these computations and results obtained by Westerland [41].

Notational forewarnings With a fixed formal group Γ in mind, we will often abbreviate the associated Morava K - and E -theories from K_Γ and E_Γ to simply K and E . We will be especially interested in the behavior of limits taken in different categories, and so we will write $\lim_{\mathcal{C}}^\alpha (S_\alpha)$ to denote the limit in the category $\mathcal{C}$ of the α -indexed pro-system S_α .

Acknowledgements The paper covers the original results of my PhD thesis, completed in 2015 at the University of California, Berkeley, under the direction of Constantin Teleman and with the support of the department's NSF grant for the geometry and

³Specifically, all previous constructions of $\mathbb{S}^{\text{det}}$ avail themselves of a generator of $\mathbb{Z}_p^\times$.

topology group, DMS-0838703. The germ of this project came from an unpublished note of Neil Strickland [36], and work on it (in particular, Theorem 3.3.5) began at the University of Illinois, Urbana–Champaign, under Matthew Ando and with the support of UIUC’s NSF grant for graduate students pursuing research projects, DMS-0838434.

This project was boosted considerably by two visits to the Max Planck Institut für Mathematik in Bonn, which was facilitated by Peter Teichner. I benefited significantly from many conversations with Michael Hopkins, with Hal Sadofsky, and especially with Tobias Barthel on the inverse limit spectral sequence presented here. I also very much enjoyed discussing possible extensions of this work with Craig Westerland, whose insights and good humor sharpened and buoyed this whole enterprise (including the publication of this document). Rune Haugseng’s anonymous referee for his paper on Morita categories [16] insisted that he include exactly the ∞ –categorical results I needed to underpin this paper—so, I also extend a hearty thank you to this mysterious benefactor. My own anonymous referee provided many helpful comments that greatly improved the quality of this paper, for which I am very grateful. Finally, I would like to thank Samrita Dhindsa for her love and support throughout.

2 Homotopical coalgebras and comodules

2.1 Motivation: coalgebraic formal schemes

To set the stage, we expand on our discussion from the introduction of treating a space as a scheme-like object, starting with a more careful description of “ X_E ”.

Definition 2.1.1 Let E be an even-periodic cohomology theory, and let X be a CW-space, so that $X = \operatorname{colim}_{\alpha} \{X_{\alpha}\}$ can be written as the colimit over the lattice of inclusions of its compact subspaces X_{α} . Suppose further that the system of finite E^* –algebras $\{E^*X_{\alpha}\}$ is equivalent to a sub-pro-system of even-concentrated finite E^* –algebras $\{E^0X_{\alpha'} \otimes_{E^0} E^*\}$. We then set

$$X_E := \operatorname{Spf} E^0 X := \{\operatorname{Spec} E^0 X_{\alpha'}\}.$$

Remark 2.1.2 [38, Section 8.2, Definition 8.15] A sufficient condition on X to ensure that X_E exists is for H_*X to be torsion-free and even, in which case there is a skeletal structure on X whose subskeleta form a cofinal subsystem of $\{X_{\alpha}\}$ and have the desired evenness property. For instance, this covers the cases of $X = \mathbb{C}P^{\infty}$ and $X = BU$. It does not, however, cover $X = BS^1[p^j]$, to which the definition given here nonetheless applies for $E = E_{\Gamma}$ a Morava E –theory.

Remembering our goal of studying X directly, we would now like to identify a topological system underlying this construction. The compactness of X_α , hence dualizability of the suspension spectrum, gives

$$E^* X_\alpha = \pi_{-*} F(\Sigma_+^\infty X_\alpha, E) = \pi_{-*} (F(\Sigma_+^\infty X_\alpha, \mathbb{S}) \wedge E),$$

from which we are inspired to consider the pro- E_∞ -ring-spectrum

$$\hat{D}X = \{F(\Sigma_+^\infty X_\alpha, \mathbb{S})\}.$$

This pro-spectrum depends only on X and not on E , and the system appearing in [Definition 2.1.1](#) then arises as the homotopy of the base change

$$\eta^* \hat{D}X \simeq \{E \wedge F(\Sigma_+^\infty X_\alpha, \mathbb{S})\},$$

where $\eta: \mathbb{S} \rightarrow E$ is the unit map for the ring spectrum E . Hence, $\hat{D}X$ plays something of a role of a universal object for these constructions.

We now immediately set our eyes on the main theorem. Recall the diagram of A -modules which defines the cotangent space for a geometric point $\mathrm{Spec} A/\mathfrak{m}$ of an affine scheme $\mathrm{Spec} A$:

$$\begin{array}{ccccc} A & \longleftarrow & \mathfrak{m} & \longleftarrow & \mathfrak{m}^2 \\ \downarrow & & \downarrow & & \\ A/\mathfrak{m} & & \mathfrak{m}/\mathfrak{m}^2 = T_0^* \mathrm{Spec} A & & \end{array}$$

Each angle in the diagram is a cokernel sequence of A -modules. Our goal is to lift this diagram of cokernel sequences of modules to a diagram of cofiber sequences of spectra, so that the original diagram is recovered upon applying E -cohomology. A candidate replacement for A itself is the pro-spectrum $\hat{D}X$, and a pointing of the space X induces an augmentation map $\hat{D}X \rightarrow \mathbb{S}$ to the constant pro-spectrum $\mathbb{S}$.

However, working with $\hat{D}X$ is prohibitively complicated: the Spanier–Whitehead duals and pro-system in its definition are designed to mitigate what would be the difference between the cohomology of the limit and the limit of the cohomology of the components, which we may eventually find undesirable and which we would be obligated to thread through our entire discussion. To avoid this conceptual overhead and potential pitfall, we take inspiration from a corresponding purely algebraic situation: for a field k , pro-finite k -algebras are equivalent to ind-finite k -coalgebras by k -linear duality, and *all* k -coalgebras are equivalent to ind-finite ones. Hence, the theory of formal schemes (over $\mathrm{Spec} k$) can be taken to be underpinned by coalgebras instead,⁴ as in the following definition.

⁴This perspective is recorded by Demazure [\[10\]](#) in the case of a field and by Strickland [\[38, Section 4.8\]](#) more generally.

Definition 2.1.3 A coalgebra C engenders a formal scheme $\text{Sch } C$ via the following dual formula for its functor of points:

$$(\text{Sch } C)(T) = \{x \in C \otimes_k T \mid \Delta(x) = x \otimes x, \varepsilon(x) = 1\}.$$

Returning to spectra, we find the following analogies: pro-finite spectra are equivalent to ind-finite spectra by Spanier–Whitehead duality, and spectra themselves are equivalent to ind-finite spectra. Hence, the analogous coalgebraic spectral object to the “profinite ring spectrum” $\hat{D}X$ is $C = \Sigma_+^\infty X$ itself, which we will work with *directly*. We need only firmly record the sense in which C is a coalgebra before we can start doing coalgebraic geometry with it.

2.2 Definitions and constructions

The theory of derived algebraic geometry and spectral schemes is only now coming into view through work of Lurie and others, but the core of *affine* derived algebraic geometry has been laid out for some time now via the theory of structured ring spectra. One of the first hurdles to clear in this arena is to produce a definition of a ring spectrum (ie affine derived scheme) that has an associated ∞ –category of module spectra (ie category of quasicoherent sheaves); such a definition is found in that of an A_∞ –ring spectrum. The precise definition is fairly elaborate, and the reader is better off consulting a textbook for a full account [26, Chapter 4], but the constituent pieces are:

- The ∞ –category SPECTRA and the category Δ of combinatorial simplices.
- A *monoidal structure* on SPECTRA , which can be encoded as a cocartesian fibration $q: \text{SPECTRA}^\otimes \rightarrow \text{ASS}^\otimes$ of operads.
- A particular functor $\text{Cut}: \Delta^{\text{op}} \rightarrow \text{ASS}^\otimes$ to be used to control concatenation of labels for combinatorial simplices.
- A functor $A: \Delta^{\text{op}} \rightarrow \text{SPECTRA}^\otimes$ respecting the pair of functors to $\text{ASS}^\otimes$ and sending “inert” morphisms to inert morphisms. We refer to such morphisms as Δ^{op} –*algebras* in SPECTRA .

Collectively, these amount to a choice of labelings of the combinatorial simplices, identifications $A[n] \simeq A[1]^{\wedge n}$, a unit map $A[0] \rightarrow A[1]$, a multiplication $A[2] \rightarrow A[1]$, and a sea of compatibilities encoding associativity, unitality, and $\wedge$ –composability.

Since we are interested in coassociative coalgebras rather than associative algebras, we tweak this definition only very slightly:

Definition 2.2.1 (cf [16, Definition 4.12] and [26, Proposition 4.1.2.15]) A *co-associative coalgebra* in a monoidal ∞ –category $\mathbf{C}$ is a Δ^{op} –algebra in $\mathbf{C}^{\text{op}}$, ie an associative algebra in the opposite category $\mathbf{C}^{\text{op}}$.

Lemma 2.2.2 *A pointed space X determines an associative algebra object in the monoidal ∞ -category $\mathbf{SPECTRA}^{\mathrm{op}}$ and in its homological Bousfield localization $\mathbf{SPECTRA}_E^{\mathrm{op}}$ for a homology theory E .*

Proof Pointed spaces (using the underlying diagonal map of sets) already determine strictly coassociative algebra objects on the 1-categorical level. By localizing away from the weak equivalences and passing to the associated ∞ -category, we recover an associative algebra object in $\mathbf{SPACES}_*^{\mathrm{op}}$. Because the stabilization functor

$$\Sigma_+^\infty: \mathbf{SPACES}_* \rightarrow \mathbf{SPECTRA}$$

respects the Cartesian monoidal structure on $\mathbf{SPACES}_*$ and the smash monoidal structure on $\mathbf{SPECTRA}$ [26, Propositions 4.8.2.9 and 4.8.2.18] and because the monoidal structure on Γ -local spectra is defined so that the localization functor L_Γ is monoidal, it follows that $L_\Gamma \Sigma_+^\infty X$ is a coaugmented coassociative coalgebra in $\mathbf{SPECTRA}_\Gamma$. $\square$

Our spectral analogue of the first fiber sequence in the diagram defining the algebraic tangent space is thus given by the diagram

$$\begin{array}{ccccc} C & \longrightarrow & M & \longrightarrow & ? \\ \eta \uparrow & & \uparrow & & \\ \mathbb{S} & & ? & & \end{array}$$

where $C = \Sigma_+^\infty X$ is a coalgebraic spectrum and $\eta: \mathbb{S} \rightarrow C$ is the coaugmentation map. In order to construct the second fiber sequence, we would like to interpret this diagram as a sequence of C -comodules, and we would like access to a coalgebraic analogue of the product of ideals. Here, again, we can lean on the already-developed theory of modules for associative algebra spectra: there is a *nonsymmetric* ∞ -operad $\Delta_{/[1]}^{\mathrm{op}}$ [16, Definition 4.2] whose algebras in $\mathbf{C}$ amount to $\mathbf{C}$ -labeled combinatorial simplices with one edge labeled by a fixed spectrum M (the bimodule), all edges indexed below it labeled by a fixed spectrum A (the left algebra), all edges indexed above it labeled by a fixed spectrum B (the right algebra), together with maps describing the associative algebra structures of A and of B , the left action of A on M , the right action of B on M , and compatibilities encoding associativity, unitality, and $\wedge$ -composability.

Definition 2.2.3 ([16, Definition 4.12] and [26, Definition 4.3.1.6]) For a coassociative coalgebra C in a monoidal ∞ -category $\mathbf{C}$, the category of C -comodules is the subcategory of bimodules in $\mathbf{C}^{\mathrm{op}}$ whose left- and right-algebras are equivalent to C .

Even with this technology, a product of ideals is too much to ask: the missing theory of ideals for structured ring spectra remains, to date, a thorn in homotopy theorists’ sides.⁵ However, in the classical case that two A –ideals I and J are *principal* with generators that have *no annihilators*, there is a natural isomorphism $I \otimes_A J \cong I \cdot J$. This includes our primary case of interest, where $A = k[[x]]$ and $I = J = \mathfrak{m} = (x)$ is the ideal of functions vanishing at $x = 0$.⁶ Motivated by this observation, we, again, borrow a definition from Haugseng and Lurie:

Definition 2.2.4 ([16, Definition 4.13 and Lemma 4.14]; cf [26, Definition 4.4.1.1 and Proposition 4.4.1.1]) There is a nonsymmetric ∞ –operad $\Lambda^{\mathrm{op}}_{/[n]}$ which arises as the gluing of n face maps in from $\Delta^{\mathrm{op}}_{/[1]}$ along $(n - 1)$ shared instances of $\Delta^{\mathrm{op}} = \Delta^{\mathrm{op}}_{/[0]}$ in a chain. Accordingly, a $\Lambda^{\mathrm{op}}_{/[n]}$ –algebra in $\mathcal{C}^{\mathrm{op}}$ selects a family of C_{i-1} – C_i –cobimodule objects M_i for $1 \leq i \leq n$: this is a *length n chain of compatible comodules*. A *length n cotensor witness* is then a $\Delta^{\mathrm{op}}_{/[n]}$ –algebra in $\mathcal{C}^{\mathrm{op}}$:

$$\begin{array}{ccc} \Lambda^{\mathrm{op}}_{/[n]} & & \\ \downarrow & \searrow & \\ \Delta^{\mathrm{op}}_{/[n]} & \cdots\cdots\cdots & \mathcal{C} \end{array}$$

In particular, $\Delta^{\mathrm{op}}_{/[n]}$ has $\binom{n}{2}$ face maps $\Delta^{\mathrm{op}}_{/[1]} \rightarrow \Delta^{\mathrm{op}}_{/[n]}$, ie a C_i – C_j –cobimodule for every choice of i and j . A *cotensor product* of the chain is the C_0 – C_n –cobimodule extracted from a witness in this way.

Lemma 2.2.5 ([16, Lemma 4.19 and Corollary 4.20]; cf [26, Theorem 4.4.2.8]) Let $P: \Lambda^{\mathrm{op}}_{/[n]} \rightarrow \mathcal{C}^{\mathrm{op}}$ be a length n chain of compatible cobimodules in $\mathcal{C}$, and let $F: \Delta^{\mathrm{op}}_{/[n]} \rightarrow \mathcal{C}^{\mathrm{op}}$ be a filling of P to a length n tensor witness. The spectrum underlying the cotensor product comodule witnessed by F is necessarily weakly equivalent to the limit of the cobar construction on the compatible cobimodules. Conversely, suppose

⁵The present situation here is a bit complicated. It is known that a ring spectrum can be localized away from any homotopy element, but that quotienting some ring spectra by certain homotopy elements can yield nonassociative or even nonunital spectra (eg respectively, $\mathbb{S}/p$ for odd primes p and $\mathbb{S}/2$). Meanwhile, the intended collection of ideals of $\mathbb{S}$ is considered well-known: the theory of ideals at the level of perfect $\mathbb{S}$ –modules is understood through the nilpotence and periodicity theorems, and the reflection of those results in the homotopy elements of $\mathbb{S}$ is extremely indirect. However, there is an $\mathbb{S}$ –algebra MU which does, in a certain stacky sense, reveal this ideal structure at the level of homotopy elements (but which still does not admit all the ring spectrum quotients that one might naively expect). At any rate, the reader must temper their expectations about the theory being developed here, as it serves our immediate purposes but not much more.

⁶The augmentation ideal of a higher-dimensional power series algebra is *not* principal.

that $n = 2$ and only P is given. If the natural C^{op} -map

$$|X \otimes B(M; S; N) \otimes Y| \rightarrow X \otimes |B(M; S; N)| \otimes Y$$

is an equivalence for all objects X and Y , then there must exist a filler F . $\square$

Remark 2.2.6 Haugseng's filling result is not specific to cotensor products of length 2, and it is functorial as the $\Lambda_{/[n]}^{\text{op}}$ -algebra varies. However, we will be content with the weakened result stated here.

The primary obstacle for us now is the natural equivalence hypothesis in the lemma. In our setting of $C = \text{SPECTRA}$ or $C = \text{SPECTRA}_\Gamma$, totalizations of cosimplicial objects *do not* typically commute with the (local) smash product of (local) spectra. So, while we can use the cobar construction to define a cotensor product and various maps concerning it, essentially no good properties can be deduced for the construction without first manually checking this condition about commuting limits and smash products. For the moment, we content ourselves with constructing the maps, leaving the question of properties for the next section.

Lemma 2.2.7 For a right C -comodule M , there is a natural retraction

$$M \rightarrow M \square_C C \rightarrow M,$$

and dually for a left C -comodule N . $\square$

Lemma 2.2.8 Let M be a C - C -cobimodule equipped with a map $\pi: C \rightarrow M$ of cobimodules. There is then a diagram

$$\begin{array}{ccccc} C & \xrightarrow{\pi} & M & & \\ \downarrow \Delta_C & & \downarrow \psi_R & \searrow \tilde{\Delta}_R & \\ C \square_C C & \xrightarrow{\pi \square_C C} & M \square_C C & \xrightarrow{M \square_C \pi} & M \square_C M \\ \downarrow & & \downarrow & & \\ C & \longrightarrow & M & & \end{array}$$

where the vertical sequences are retractions. Alternatively, there is also a map $\tilde{\Delta}_L$ using the left C -comodule structure of M .

Proof The arrow Δ_C could just as well be called ψ_C , and by this name the top square commutes because π is a homomorphism of cobimodules. The arrow $\tilde{\Delta}_R$ is then defined by either composite. $\square$

Classically, if π is surjective, then Δ_R and Δ_L will agree on points exactly if M is taken to be a coideal. The word “coideal” is too bold in our homotopical context, since it suggests that the quotient is again a coalgebra. We instead adopt the following terminology.

Definition 2.2.9 Let M be a C – C –cobimodule spectrum equipped with a cobimodule map $\pi: C \rightarrow M$. Such a cobimodule spectrum is a *symmetric cobimodule* (for C) when $\tilde{\Delta}_R$ and $\tilde{\Delta}_L$ are homotopic, in which case we will use $\tilde{\Delta}$ as an abbreviation for either.⁷

Lemma 2.2.10 Let $f: D \rightarrow C$ be a map of coalgebras, and suppose that there is a splitting of the cofiber

$$D \xrightarrow{f} C \overset{\pi}{\curvearrowright} \mathrm{cofib} f$$

The cofiber of f , considered as a C – C –cobimodule, is then a symmetric cobimodule.

Proof We enlarge the diagram of Lemma 2.2.8 to include $\tilde{\Delta}_L$ and $\tilde{\Delta}_R$:

We see that the two maps $\tilde{\Delta}_L$ and $\tilde{\Delta}_R$ agree upon precomposition to C . Since $\mathrm{cofib} f$ splits off of C , they agree on $\mathrm{cofib} f$ as well. □

Lemma 2.2.11 For M a symmetric cobimodule admitting cotensor powers $M^{\square_C k}$ and $M^{\square_C (k+1)}$, all of the following maps are homotopic:

$$M^{\square_C k} \xrightarrow{1 \square \dots \square 1 \square \tilde{\Delta} \square 1 \square \dots \square 1} M^{\square_C (k+1)}.$$

⁷This is *not* the robust definition that a party setting out to build up a theory of spectral coalgebraic algebra would choose. They would probably prefer to include coherence conditions on this symmetry. We, however, are only setting out to support Definition 2.2.12.

Proof The definition of the cobar construction gives a homotopy between the following two maps:

$$M \square_C M \xrightarrow[1 \square_C \psi_L]{\psi_R \square_C 1} M \square_C C \square_C M.$$

Using the symmetric cobimodule property of M , we can trade $\tilde{\Delta}_L$ for $\tilde{\Delta}_R$, and hence we can transfer the $\tilde{\Delta}$ to any coordinate. □

We are now poised to properly state our construction:

Definition 2.2.12 Given a pointed coalgebra spectrum $\eta: \mathbb{S} \rightarrow C$, we define C_1^1 by the chain of fiber sequences

$$\begin{array}{ccccc} C & \longrightarrow & M & \xrightarrow{\tilde{\Delta}} & M \square_C M \\ \eta \uparrow & & \uparrow & & \\ \mathbb{S} & & C_1^1 & & \end{array}$$

Remark 2.2.13 If C is merely a pointed coalgebra spectrum, this is as much of the diagram as we can form at this time. After all, we have not yet shown that $M \square_C M$ is a C – C –cobimodule, and hence we cannot guarantee the existence of “ $M \square_C (M \square_C M)$ ”. However, if we start with a space X and follow [Lemma 2.2.2](#), then we can form the sequence of iterated cobar diagrams in the 1–category of spaces

$$X_+ \rightarrow X \rightarrow \Omega(X; X_+; X) \rightarrow \Omega(X; X_+; X; X_+; X) \rightarrow \cdots .$$

Pushing these diagrams forward along Σ^∞ and along the localizer L_Γ then produces inverse diagrams of $(\Gamma$ –local) spectra. Thus, we are at least assured that, for $C = \Sigma_+^\infty X$, the diagrams *determining* the higher cotensor powers (and hence the jet-space decomposition) are well-defined. We are still obligated to justify their utility by computing the value of E_Γ on their limits.

2.3 Computational tools

We would like to justify [Definition 2.2.12](#) above by calculating $E_\Gamma^0 C_1^1$ and checking that it gives the desired cotangent module, and then we would also like to extend it to the full sequence presented in the main theorem. In view of [Lemma 2.2.5](#), this rests directly on having tools available to compute the homology and cohomology of a cotensor product of comodule spectra. With a computational task ahead, it is now convenient to introduce Morava K –theory.

Definition 2.3.1 The coefficient ring $\pi_0 E_\Gamma$ is a power series ring over $\mathbb{W}_p(k)$, itself a complete local ring with maximal ideal generated by p . The *Morava K -theory spectrum*, K_Γ , is the quotient of E_Γ in E_Γ -modules by a generating regular sequence of the maximal ideal of $\pi_0 E_\Gamma$.

As a result, $\pi_* K_\Gamma = k[u^\pm]$ is a graded field. This makes it extremely valuable for computations — especially ones such as ours, where large smash products arise, since K_Γ has Künneth isomorphisms. Remarkably, despite being defined as a quotient, “ K_Γ carries the same information as E_Γ ” in the precise sense that K_Γ -acyclics, homological or cohomological, agree with cohomological E_Γ -acyclics [24, Proposition 2.5]. Hence, any of E_Γ -cohomology, K_Γ -cohomology, or K_Γ -homology are equally good for testing Γ -local equivalences.⁸ We will favor the last option, both for technical reasons (cf [3] and the appendix) and because it covariantly converts the coalgebraic spectrum C into the $(K_\Gamma)_*$ -coalgebra $(K_\Gamma)_* C$, then into the coalgebraic formal scheme $\mathrm{Sch}(K_\Gamma)_* C$. To save on notational overhead, we will often write “ K ” alone when some fixed Γ is understood.

We now turn back to the task at hand: given C - C -cobimodules M and N , we want to study the natural map

$$X \wedge \lim_{\mathcal{C}} \Omega(M; C; N) \wedge Y \rightarrow \lim_{\mathcal{C}} (X \wedge \Omega(M; C; N) \wedge Y)$$

and to find conditions under which it becomes a K -homology equivalence. The presentation of $M \square N$ as the totalization of a cosimplicial object equips it with a coskeletal filtration and hence a spectral sequence approximating its homology. We might hope to use this to gain a foothold on our problem, but the homology of an inverse limit of spectra typically compares poorly with the inverse limit of the homologies of the individual spectra in the system. Remarkably, a result of Sadofsky shows that K -homology does not suffer too badly from this.

Theorem 2.3.2 (cf Theorem A.0.1 and [3]) *Take $p \gg \mathrm{ht} \Gamma$, and let $\{X_\alpha\}_\alpha$ be a pro-system of Γ -local spectra such that $\{K_* X_\alpha\}_\alpha$ is a Mittag-Leffler system of K_* -modules. There is then a convergent spectral sequence of signature*

$$R^* \lim_{\mathrm{COMOD}_{K_* K}}^\alpha \{K_* X_\alpha\}_\alpha \Rightarrow K_* \lim_{\mathrm{SPECTRA}_\Gamma}^\alpha \{X_\alpha\}_\alpha,$$

where the right-derived inverse limit forming the input to the spectral sequence is taken in the category of $K_* K$ -comodules. □

⁸Remarkably, E_Γ -homology does not belong to this list, essentially for the same reason that $\mathbb{Z}_p \otimes_{\mathbb{Z}} \mathbb{Z}_p$ is not p -complete. We will address this further in Definition 3.3.2.

In order to make use of this spectral sequence, we need to compute some of its inputs. The homologies of the finite stages of the filtration are accessible because homology does pass through *finite* limits, and these admit the following uniform description:

Theorem 2.3.3 *Let C be a coalgebra spectrum, M a right C -comodule spectrum, and N a left C -comodule spectrum. Writing F for the fiber of the counit map $\varepsilon: C \rightarrow \mathbb{S}$, there is an n -indexed system of spectral sequences*

$$\left. \begin{aligned} E_{*,*}^1 &\cong K_* M \otimes_{K_*} (K_* F)^{\otimes(*\leq n)} \otimes_{K_*} K_* N, \\ E_{*,(*<n)}^2 &\cong \operatorname{Cotor}_{*,*}^{K_* C}(K_* M, K_* N) \end{aligned} \right\} \Rightarrow K_* \operatorname{Tot}_n \Omega(M; C; N),$$

whose inverse limit in the category of K_* -modules is the spectral sequence

$$\left. \begin{aligned} E_{*,*}^1 &\cong K_* M \otimes_{K_*} (K_* F)^{\otimes*} \otimes_{K_*} K_* N, \\ E_{*,*}^2 &\cong \operatorname{Cotor}_{*,*}^{K_* C}(K_* M, K_* N) \end{aligned} \right\} \Rightarrow \lim_{\operatorname{MODULES}_{K_*}}^n K_* \operatorname{Tot}_n \Omega(M; C; N).$$

The spectral sequences in the system are strongly convergent, and the full spectral sequence is conditionally convergent.

Proof This is an aggregation of several standard results in the literature. Ravenel and Wilson [31, Section 2] provide a convenient summary of the bar spectral sequence, and these spectral sequences arise as its dual. Ravenel [29, Appendix A1] also provides a collection of results on the homological algebra of comodules and in particular gives a definition and lists properties for “Cotor”. Finally, Boardman [6, Theorem 7.1] provides tools for analyzing the convergence of the spectral sequences. $\square$

We now sew these together to analyze Sadofsky’s inverse limit spectral sequence in the case at hand. This is mostly an exercise in elementary homological algebra, so requires substantial bookkeeping but is otherwise straightforward.

Theorem 2.3.4 *Continue to assume $p \gg \operatorname{ht} \Gamma$. Assume that the natural map*

$$K_s \operatorname{Tot}_t \Omega(M; C; N) \rightarrow K_s \operatorname{Tot}_{t-1} \Omega(M; C; N)$$

is injective when $s + t$ is odd and surjective when $s + t$ is even.⁹ There then is an additional convergent spectral sequence

$$R^* \lim_{\operatorname{COMOD}_{K_* K}}^t \{K_* \operatorname{Tot}_t \Omega(M; C; N)\}_t \Rightarrow K_*(M \square_C N),$$

where the derived inverse limit is taken in the category of $K_* K$ -comodules.

⁹In particular, this is satisfied if M , N and C have even-concentrated K -homology.

Proof Throughout, we will consider the degree $s + t$ part of the t -cochains C_{s+t}^t , coboundaries B_{s+t}^t , and cocycles Z_{s+t}^t of the normalized cobar complex

$$\Omega(K_*M; K_*C; K_*N).$$

With an induction beginning at $t = 0$ and $t = 1$, we claim the K -homology of the t^{th} level of the tower (ie the target of the t^{th} partial spectral sequence) is

$$K_s \text{Tot}_t \Omega(M; C; N) = D_{s,t}^1 = \begin{cases} D_{s,t-1}^1 \oplus (C_{(s+t)/2}^t / B_{(s+t)/2}^t) & \text{if } s+t \text{ is even,} \\ D_{s,t-2}^1 \oplus H_{(s+t-1)/2}^{t-1} & \text{if } s+t \text{ is odd,} \end{cases}$$

where $D_{*,*}^1$ denotes the rear of the exact couple of the full spectral sequence. In particular, this formula will show that the inverse limit tower is Mittag-Leffler on K -homology, so that Sadofsky's hypotheses are satisfied. The theorem then follows.

Induction proceeds by considering one triangle in that exact couple:

$$\begin{array}{ccccc} K_s \text{Tot}_{t-1} \Omega(M; C; N) & \leftarrow & K_s \text{Tot}_t \Omega(M; C; N) & & D_{s,t-1}^1 \leftarrow D_{s,t}^1 \\ & \searrow [-1] & \uparrow & = & \searrow [-1] \uparrow \\ & & K_s \Omega^t(M \wedge C^{\wedge t} \wedge N) & & E_{s,t}^1 \end{array}$$

The bottom vertex in the triangle is the K_* -module of cochains, and we take s to be a degree in which $K_{s+t}(M \wedge C \wedge \cdots \wedge C \wedge N)$ is nonvanishing, ie $s + t$ is even. Then, the triangle unrolls into an exact sequence which, using the inductive hypothesis, takes the form

$$\begin{array}{ccccccc} 0 & \longrightarrow & D_{s+1,t}^1 & \longrightarrow & D_{s+1,t-1}^1 & \longrightarrow & E_{s,t}^1 \longrightarrow D_{s,t}^1 \longrightarrow D_{s,t-1}^1 \longrightarrow 0 \\ & & \parallel & & \parallel & & \parallel \\ 0 & \longrightarrow & H_{(s+t)/2}^{t-1} \oplus X & \xrightarrow{i \oplus 1} & \frac{C_{(s+t)/2}^{t-1}}{B_{(s+t)/2}^{t-1}} \oplus X & \xrightarrow{\partial \oplus 0} & C_{(s+t)/2}^t \xrightarrow{\pi \oplus 0} \frac{C_{(s+t)/2}^t}{B_{(s+t)/2}^t} \oplus Y \xrightarrow{0 \oplus 1} Y \longrightarrow 0 \end{array}$$

for some modules X and Y to be determined. We then splice three of these long sequences together to form Figure 2, which is labeled in terms of the cobar cohomology groups in Figure 3. The action lies in the zigzag containing $D_{s,t+1}^1$ through $D_{s,t-2}^1$: if $D_{s,t-1}^1$ and $D_{s,t-2}^1$ are assumed to be as claimed, then the observation involving Y shows that $D_{s,t-1}^1$ is as well, which in turn determines $D_{s,t+1}^1$ as the kernel of the map $D_{s,t}^1 \rightarrow E_{s-1,t+1}^1$, ie as the cocycle subgroup of the cochain group, taken modulo the coboundaries. This proves the inductive claim. $\square$

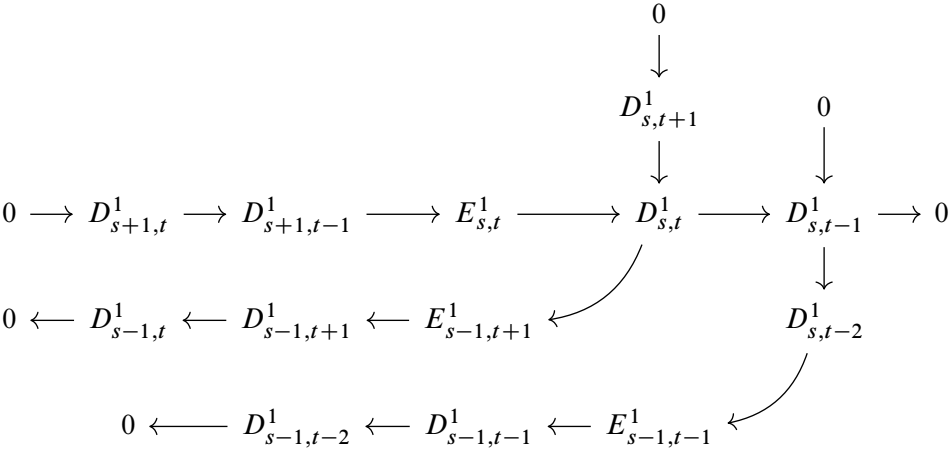


Figure 2: Three interacting exact sequences.

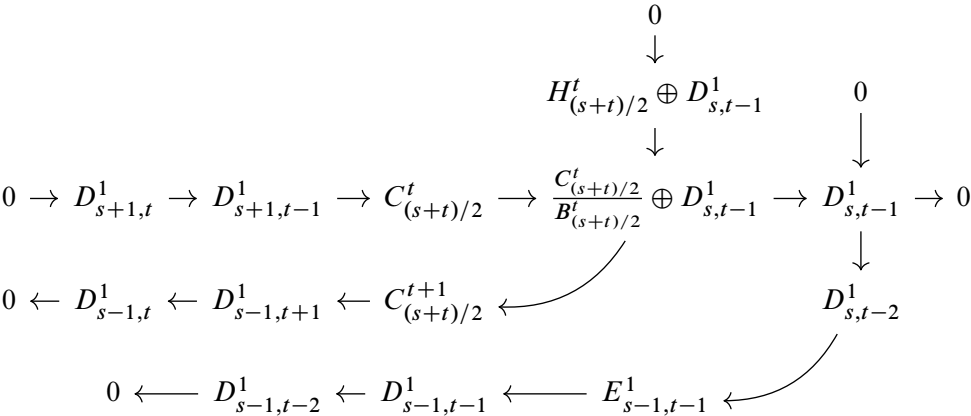


Figure 3: The interacting sequences with cobar complex names.

2.4 Computation in the case of a formal curve

In general, it is very hard to control the spectral sequence of [Theorem 2.3.4](#). Even computing these derived functors is prohibitively complicated in almost any non-degenerate case — for instance, Hopkins (see [\[27, Section 14\]](#) and [\[3\]](#)) has recommended them in an approach to analyzing the chromatic splitting conjecture.¹⁰ This moves us to consider progressively more specialized situations where we can fully determine

¹⁰Computing these derived functors can be compared to computing certain local cohomology groups and to certain “noncontinuous” forms of the group cohomology of the Morava stabilizer group. These are both very difficult invariants.

this spectral sequence, beginning with the case that the coalgebraic formal scheme $\text{Sch } K_0C$ (see [Definition 2.1.3](#)) is a formal variety.

Definition 2.4.1 A k -coalgebra C over a field k will be called a *formal variety* (of dimension n) when, for any augmented nilpotent k -algebra A with maximal ideal $\mathfrak{m}$, there is a natural isomorphism between the group-like elements of $C \otimes A$ and $\mathfrak{m}^{\times n}$.¹¹ A spectral coalgebra C will be called a *formal coalgebraic variety spectrum* (local to Γ) when K_*C is even-concentrated and K_0C is a formal variety in coalgebras. In both settings, we use *formal curve* as a synonym for a formal variety of dimension 1.

For the rest of this section, we take $\eta: \mathbb{S} \rightarrow C$ to be a pointed coalgebraic formal curve spectrum local to Γ , and we write M for the cofiber of η . Supposing that $M^{\square_C j}$ has been shown to support the structure of an even C - C -cobimodule spectrum, we now inductively pursue an even C - C -cobimodule structure on $M^{\square_C(j+1)}$.

Theorem 2.4.2 The system $\{K_*^n \Omega(M; C; M^{\square_C j})\}_n$ is pro-constant.

Proof This system appears as the rear of the exact couple in [Theorem 2.3.4](#), where we showed that it was presented levelwise as direct sums of two sorts of groups: Cotor groups and cochain groups modulo coboundaries. These groups are then linked together by appropriate projections (away from a direct sum factor) and inclusions (of cocycles modulo coboundaries — ie Cotor groups — into cochains modulo coboundaries). Our conclusion will follow from a direct calculation of these Cotor groups.

This problem may be addressed by standard tools in homological algebra and in local cohomology, which is perhaps more visible after taking linear duals and passing to profinite algebras:

$$\text{Cotor}_{K_*C}^{*,*}(K_*M, K_*N) \cong \text{Tor}_{K_*C^\vee}^{*,*}(K_*M^\vee, K_*N^\vee)^\vee.$$

Under our hypotheses, we may choose isomorphisms

$$K_*C^\vee \cong K_*[[x]], \quad K_*M^\vee = \langle x \rangle.$$

This shows $(K_*M)^\vee$ to be a free $(K_*C)^\vee$ -module, hence the higher Tor groups vanish.¹²

¹¹Equivalently, the dual profinite algebra C^* is isomorphic to a power series ring.

¹²For C of higher dimension, these Cotor groups are bounded above, and so we may draw a similar conclusion. However, they fail to be even, which stymies the application of [Theorem 2.3.4](#), which in turn irreparably harms the overall induction. Another avatar of the same computational failure is that the comparison between $K_*M^\vee \otimes_{K_*C^\vee} K_*M^\vee$ and $K_*M^\vee \cdot K_*M^\vee$ supposed in [Section 2.1](#) no longer holds, as $K_*M^\vee$ is no longer cyclic.

We now return to the original problem of pro-constancy. Because so many Cotor groups themselves vanish, we find that the maps in the pro-system are zero except on those nonvanishing factors, where the maps are instead the respective identity morphisms. It follows that, for each fixed s , the tower $D_{s,t}^1$ is pro-constant. $\square$

Corollary 2.4.3 *There is an isomorphism*

$$K_*(M \square_C M^{\square_C j}) \cong K_* M^{\square_{K_* C} (j+1)}.$$

Proof Because the tower is pro-constant, the higher derived inverse limit functors of [Theorem 2.3.4](#) vanish, and hence we need only contend with the higher derived functors of cotensoring itself. These also vanish, again because the maximal ideal in a one-dimensional power series algebra is a free module. $\square$

Corollary 2.4.4 *The spectrum $M^{\square_C (j+1)}$ is a C – C –cobimodule spectrum.*

Proof Tensoring the pro-constant towers of [Theorem 2.4.2](#) with $K_* X$ and $K_* Y$ does not disturb their pro-constancy. It follows that the natural map

$$K_* X \otimes K_* M^{\square_{K_* C} (j+1)} \otimes K_* Y \rightarrow K_* \lim_{\text{SPECTRA } \Gamma}^n \text{Tot}_n(X \wedge \Omega(M; C; M^{\square_C j}) \wedge Y)$$

is an isomorphism, and hence the conditions of [Lemma 2.2.5](#) are satisfied. $\square$

This completes the induction and shows that $M^{\square_C j}$ is a C – C –cobimodule spectrum for all values of j . We use this to justify the following definition.

Definition 2.4.5 Writing C_n^∞ for $M^{\square_C n}$, there are natural projection (or “pinch”) maps $C_n^\infty \rightarrow C_{n'+1}^\infty$ for $n' \geq n$. We define spectra $C_n^{n'}$ by the fiber sequence

$$C_n^{n'} \rightarrow C_n^\infty \rightarrow C_{n'+1}^\infty$$

and the *jet-space decomposition* is the sequential system

$$\begin{array}{ccccccc} C_0^\infty & \longrightarrow & C_1^\infty & \longrightarrow & \cdots & \longrightarrow & C_n^\infty & \longrightarrow & C_{n+1}^\infty & \longrightarrow & \cdots \\ \uparrow & & \uparrow & & & & \uparrow & & \uparrow & & \\ C_0^0 & & C_1^1 & & \cdots & & C_n^n & & C_{n+1}^{n+1} & & \cdots \end{array}$$

Our computational tools have been explicit enough that we can now also deduce that K –cohomology takes the correct value on C_1^1 . By [Corollary 2.4.3](#), we have that $K^0 C_1^1$

indeed belongs to the exact sequence

$$\begin{array}{ccccc} K^0C_1^1 & \longleftarrow & K^0C_1^\infty & \longleftarrow & K^0C_1^\infty \otimes_{K^0C} K^0C_1^\infty \\ & & & \swarrow \mu & \parallel \\ & & & & K^0C_1^\infty \cdot K^0C_1^\infty \end{array}$$

Definition 2.4.6 In the case where C is a coalgebraic formal curve spectrum, we write $T_\eta C$ for the spectrum C_1^1 , so that the isomorphism of Definition 2.4.6 reads as an interchange law

$$K^*T_\eta C = T_{\eta_K}^* C_K.$$

2.5 Koszul duality

We now give an alternative interpretation of the jet-space decomposition,¹³ which we first motivate by example. In Section 3.2 below, we will show for $C = \Sigma_+^\infty \mathbb{CP}^\infty$ that $T_\eta C \simeq \mathbb{S}^2 \simeq \Sigma^\infty \mathbb{CP}^1$ and that the jet-space decomposition of C is into one sphere of each even dimension. This familiar cellular filtration of $\mathbb{CP}^\infty$ also arises from a second avenue that, at a glance, appears completely separate. By appealing to the equivalence

$$\mathbb{CP}^\infty \simeq BU(1),$$

we may use the bar filtration on $BU(1)$ to produce a filtration on $\mathbb{CP}^\infty$ which has filtration quotients $(\Sigma U(1))^{\wedge n} \simeq \mathbb{S}^{2n}$. Moreover, there is a tool for mechanically recognizing this phenomenon: the input $U(1)$ to the functor B may be recovered as the based loopspace

$$U(1) \simeq \Omega \mathbb{CP}^\infty,$$

which is a specific example of the familiar equivalence of categories

$$\text{CONNECTEDSPACES}_* \overset{\Omega}{\underset{B}{\rightleftarrows}} \text{GROUPLIKE}A_\infty\text{SPACES}.$$

Koszul duality (or “co/bar duality”) captures this phenomenon in maximum generality, and the jet-space decomposition may be viewed as a specific instance of it: any coalgebraic formal curve spectrum C can be written as “ BG ” for an A_∞ ring spectrum G extracted from the 1-jets $T_\eta C$.

Our starting point is a second presentation of $T_\eta C$, arising as follows.

¹³The ideas in this subsection will not recur, and an uninterested reader may safely skip ahead.

Lemma 2.5.1 Given a fiber sequence of C – C –cobimodule spectra $M'' \rightarrow M \rightarrow M'$ and a fourth cobimodule spectrum N , the sequence

$$M'' \square_C N \rightarrow M \square_C N \rightarrow M' \square_C N$$

is also a fiber sequence.

Proof Because $-\square_C N$ is defined by taking the inverse limit of a diagram constructed from smashing the input with the fixed spectra $C^{\wedge n} \wedge N$, this functor preserves fiber sequences. $\square$

Corollary 2.5.2 There is a natural equivalence $\mathbb{S} \square_C M \simeq C_1^1$.

Proof By applying [Lemma 2.5.1](#), one can construct the jet-space decomposition by iteratively cotensoring the fiber sequence

$$\begin{array}{ccc} C & \longrightarrow & M \\ \eta \uparrow & & \\ \mathbb{S} & & \end{array}$$

with cotensor powers of M and sewing together the overlapping nodes:

$$\begin{array}{ccccccc} C \rightarrow M = M \square_C C \rightarrow M \square_C M = M^{\square_C 2} \square_C C \rightarrow M^{\square_C 2} \square_C M \rightarrow \dots \\ \eta \uparrow \quad \quad M \square_C \eta \uparrow \quad \quad M^{\square_C 2} \square_C \eta \uparrow \\ \mathbb{S} \quad \quad M \square_C \mathbb{S} \quad \quad M^{\square_C 2} \square_C \mathbb{S} \quad \quad \dots \end{array}$$

In particular, C_1^1 appears as $M \square_C \mathbb{S}$. $\square$

Lemma 2.5.3 When $\mathbb{S} \rightarrow C$ is a pointed coalgebra spectrum such that $\text{Sch } K_* C$ is a formal variety, the spectrum $\mathbb{S} \square_C \mathbb{S}$ has a split filtration of the form

$$\Sigma^{-1} C_1^1 \rightarrow \mathbb{S} \square_C \mathbb{S} \rightarrow \mathbb{S}.$$

Proof Using [Lemma 2.5.1](#), we cotensor the resolution sequence $\mathbb{S} \rightarrow C \rightarrow M$ with the C –comodule $\mathbb{S}$ to get the new fiber sequence

$$\dots \rightarrow \Sigma^{-1} \mathbb{S} \square_C M \rightarrow \mathbb{S} \square_C \mathbb{S} \rightarrow \mathbb{S} \square_C C \rightarrow \mathbb{S} \square_C M \rightarrow \dots.$$

We can use the method of [Theorem 2.3.4](#) to identify $\mathbb{S} \square_C C$ with $\mathbb{S}$, and we can use [Corollary 2.5.2](#) to identify $\mathbb{S} \square_C M$ with C_1^1 . This yields the split fiber sequence

$$\Sigma^{-1} C_1^1 \rightarrow \mathbb{S} \square_C \mathbb{S} \rightarrow \mathbb{S}.$$

$\square$

The object $\mathbb{S} \square_C \mathbb{S} = \Omega(\mathbb{S}; C; \mathbb{S})$ is the subject of Koszul duality. There is the following general result:

Theorem 2.5.4 ([9, Proposition 7.26] and [13, Theorem 2.1.11]) *Let $\mathcal{O}$ be an operad in k –module spectra, where k is some algebra spectrum, and let A be a left module for the operad $\mathcal{O}$ (ie an $\mathcal{O}$ –algebra). The arboreal bar construction $B(k; \mathcal{O}; k)$ provides a co-operad $\mathcal{O}^\vee$ for which $B(A; \mathcal{O}; k)$ is a left comodule, and, dually, the arboreal cobar construction $\Omega(k; \mathcal{O}^\vee; k)$ provides an operad $(\mathcal{O}^\vee)^\vee$ for which $\Omega(C; \mathcal{O}^\vee; k)$ is a left module. Additionally, the operad produced in this way from the coassociative co-operad is the associative operad, and the co-operad produced from the associative operad is the coassociative co-operad.* □

Corollary 2.5.5 *The spectrum $\mathbb{S} \square_C \mathbb{S} = (T_\eta C)_+$ is an associative algebra spectrum.* □

There is always a natural map of algebras

$$\mathrm{Tot} \, \Omega(k; |B(k; A; k)|; k) \rightarrow A,$$

and when $\mathcal{O}$ and A are suitably of finite type (eg $\mathcal{O}$ the associative operad and A a finitely generated polynomial algebra), then this map is an isomorphism (ie co/bar duality is involutive) [9, Proposition 6.4]. It is not clear that such a theorem applies directly here: the category of Γ –local spectra does not have good finiteness properties, nor do power series algebras. Instead, we prove the relevant special case of this family of theorems by hand:

Theorem 2.5.6 *For C a Γ –local coalgebraic formal variety spectrum, the natural map*

$$|B(\mathbb{S}; \mathbb{S} \square_C \mathbb{S}; \mathbb{S})| \rightarrow C$$

is a Γ –local equivalence.

Proof The two coalgebraic formal schemes

$$\mathrm{Sch} \, K_* C, \quad \mathrm{Sch} \, K_* |B(\mathbb{S}; \mathbb{S} \square_C \mathbb{S}; \mathbb{S})|$$

are both formal varieties, and the natural map is both a map of coalgebraic schemes and an isomorphism on tangent spaces. The inverse function theorem for formal varieties thus applies, so that the map is an isomorphism of formal varieties, ie a K_* –equivalence. □

Remark 2.5.7 We can use [Theorem 2.5.6](#) to think of the attaching map H in

$$\begin{array}{ccccc} \Sigma^{-1}C_2^2 & \longrightarrow & C_1^1 & \longrightarrow & C_1^2 \\ \parallel & & \parallel & & \parallel \\ \Sigma^{-1}(T_\eta C)^{\wedge 2} & \xrightarrow{H} & T_\eta C & \longrightarrow & C_1^2 \end{array}$$

as a kind of “Hopf map for $\mathbb{S} \square_C \mathbb{S}$ ”.

3 Projective spaces, formal curves, and Picard elements

In this section, we shift attention from theory-building to examples and computations.

3.1 General features

We continue to assume that C is a coalgebraic formal curve spectrum, ie K^0C is abstractly isomorphic to a 1-dimensional power series ring. In addition to the results of the previous Section, the main consequences of this specialization stem from the following definition and theorem.

Definition 3.1.1 A spectrum L is said to be Γ –*locally invertible* if there is some other spectrum L^{-1} with $L \wedge L^{-1} \simeq \mathbb{S}$ in the Γ –local category. The collection of isomorphism classes of such spectra forms an abelian group, called the *Picard group* of the Γ –local stable category.

Theorem 3.1.2 [21, Theorem 1.3] *A spectrum L is Γ –locally invertible if and only if K_*L is a K_* –line (ie a 1–dimensional K_* –vector space, or a $\otimes$ –invertible object in the category of K_* –vector spaces).* $\square$

Corollary 3.1.3 *When C is a pointed coalgebraic formal curve spectrum, $T_\eta C$ (and, indeed, any stratum C_n^n of the jet-space decomposition) gives an element of the Picard group of the Γ –local stable category.* $\square$

This corollary can be viewed in two ways. First, it can be used to construct elements of the Picard group of the Γ –local stable category by finding examples of pointed coalgebraic formal curve spectra. Second, given such a coalgebraic formal curve spectrum, the jet-space decomposition can be used to give a kind of cellular decomposition of C , by considering the dual sequence

$$\begin{array}{ccccccc}
 * & \longrightarrow & C_0^0 & \longrightarrow & C_0^1 & \longrightarrow & C_0^2 \longrightarrow \dots \\
 \downarrow & & \downarrow & & \downarrow & & \downarrow \\
 C_0^\infty & \equiv & C_0^\infty & \equiv & C_0^\infty & \equiv & C_0^\infty \equiv \dots \\
 \downarrow & & \downarrow & & \downarrow & & \downarrow \\
 C_0^\infty & \longrightarrow & C_1^\infty & \longrightarrow & C_2^\infty & \longrightarrow & C_3^\infty \longrightarrow \dots
 \end{array}$$

where each vertical sequence is a fiber sequence. Using [Corollary 3.1.3](#), and thinking of Picard elements as “generalized spheres”, this gives a Γ –local cellular decomposition of C in terms of Γ –local Picard elements which are perhaps *not* standard spheres.

3.2 Classical projective spaces

We now work through the examples where C is taken to be a classical projective space. In light of [Theorem 3.3.12](#), we will be especially interested in determining the homotopy type of $T_\eta C$.

Let $C = \Sigma_+^\infty \mathbb{C}P^\infty$, with $\eta: \mathbb{S} \rightarrow \Sigma_+^\infty \mathbb{C}P^\infty$ its natural pointing. In this case, we apply the version of [Definition 2.2.12](#) in the global stable category. However, the global stable category is not local for a field spectrum, so we appeal to the auxiliary spectra $K = H\mathbb{Q}$ and $K = H\mathbb{F}_p$ in order to analyze the coskeletal tower. Here, we find

$$H\mathbb{Q}_* T_+ \Sigma_+^\infty \mathbb{C}P^\infty \cong \Sigma^2 \mathbb{Q}, \quad (H\mathbb{F}_p)_* T_+ \Sigma_+^\infty \mathbb{C}P^\infty \cong \Sigma^2 \mathbb{F}_p.$$

It follows from Sullivan’s adelic reconstruction [\[40, Proposition 3.20\]](#) that the integral homology is

$$H\mathbb{Z}_* T_+ \Sigma_+^\infty \mathbb{C}P^\infty \cong \Sigma^2 \mathbb{Z},$$

and since $T_+ \Sigma_+^\infty \mathbb{C}P^\infty$ is a *connective* spectrum,¹⁴ it follows furthermore that its homotopy type is

$$T_+ \Sigma_+^\infty \mathbb{C}P^\infty \simeq \mathbb{S}^2.$$

To verify that the jet-space decomposition records the cellular decomposition of $\mathbb{C}P^\infty$, we transfer to the dual of the jet-space decomposition as in [Section 3.1](#). By connectivity, it again follows that $C_m^n \simeq \mathbb{C}P_m^n$. The global analysis of $C = \Sigma_+^\infty \mathbb{H}P^\infty$ proceeds similarly. We may also perform an identical rational analysis of $K(\mathbb{Q}, 2n)$ to produce $T_+ \Sigma_+^\infty K(\mathbb{Q}, 2n) \simeq \mathbb{S}_{\mathbb{Q}}^{2n}$.

¹⁴Similar considerations may be used to describe the entire Picard group of the stable category; see [\[37, Theorem 2.2\]](#).

We can also address $C = \Sigma_+^\infty \mathbb{R}P^\infty$ and $K = K(\infty) = K_{\widehat{\mathbb{G}}_a} = H\mathbb{F}_2$ (where we use the more general hypothesis available to [Theorem 2.3.4](#), rather than the even-concentration shortcut). By similar reasoning to the complex projective case, performing the tangent space construction in the 2-adic stable category (in fact, because our objects are connective, it suffices to consider the $\widehat{\mathbb{G}}_a$ -local category over $\mathbb{F}_2$) yields

$$T_+ \Sigma_+^\infty \mathbb{R}P^\infty \simeq (\mathbb{S}^1)^\wedge_2,$$

and the jet-space decomposition again recovers the cellular decomposition of $\mathbb{R}P^\infty$.

Finally, it’s worth remarking that the ambient category chosen to perform the tangent space construction is very important. It can neither be too localized nor too delocalized:

- Passing to the $\widehat{\mathbb{G}}_m$ -local category (or, generally, the Γ -local category for Γ defined over k of characteristic 2) factors through 2-completion.¹⁵ Ravenel [\[28, Theorem 9.1\]](#) has shown that for any j there is an equivalence

$$L_{\widehat{\mathbb{G}}_m} \mathbb{R}P_{8j+1}^\infty \simeq L_{\widehat{\mathbb{G}}_m} \mathbb{S}^{-1}.$$

Taking $j = 0$, one sees $L_{\widehat{\mathbb{G}}_m} \mathbb{R}P^\infty \simeq L_{\widehat{\mathbb{G}}_m} \mathbb{S}^{-1}$, so its $K(1)$ -cohomology is not a power series ring, and it is furthermore too small to have the correct Cotor groups. Letting j range, it’s also plain that the bar filtration looks wildly different from the behavior of any sort of expected jet-space decomposition.

- On the other hand, $\mathbb{R}P^\infty$ is p -locally acyclic for $p \neq 2$. It follows that the integral homology of $T_+ \Sigma_+^\infty \mathbb{R}P^\infty$, as computed in the global stable category, will not be a $\mathbb{Z}$ -line.

3.3 Determinantal projective space

We dedicate this subsection to a particularly interesting set of examples, stemming from a calculation of Ravenel and Wilson.

Theorem 3.3.1 (Ravenel–Wilson [\[31\]](#); see also Johnson–Wilson [\[25, Appendix\]](#) and Hopkins–Lurie [\[20, Section 2\]](#)) *Take the ambient prime p to satisfy $p \geq 3$.¹⁶ There is a Hopf ring isomorphism*

$$\bigoplus_{q=0}^\infty K_* B^q S^1[p^j] \cong \bigoplus_{q=0}^\infty (K_* B S^1[p^j])^{\wedge q},$$

¹⁵ . . . but not through $H\mathbb{F}_2$ -localization.
¹⁶ The version of this result due to Hopkins–Lurie does not impose this constraint.

where the target is the free alternating Hopf ring on $K_*BS^1[p^j]$. Moreover, for fixed $q \geq 1$, the system $\{B^1S^1[p^j]_K\}_j$ has the structure of a connected p -divisible group of dimension $\binom{d-1}{q-1}$, where $B^1S^1[p^j]_K$ is the p^j -torsion subgroup. The corresponding formal group is given by $(B^qS^1)_K$. $\square$

Of these examples, the value of $q = 1$ corresponds to the interesting spectrum $\Sigma_+^\infty \mathbb{C}P^\infty$, to which the tools from Section 2 apply. In this other extreme case, $T_+ \Sigma_+^\infty B^dS^1$ has 1-dimensional K -homology, and hence Corollary 3.1.3 and Theorem 3.3.12 can again be applied.¹⁷ In order to determine the Picard element given by Corollary 3.1.3, we need a finer invariant than the mere detection property provided by Theorem 3.1.2. We find what we seek in the $K(n)$ -local version of Morava E -homology:

Definition 3.3.2 ([24, Definition 8.3] and [39, Theorem 12]) The continuous Morava E -homology or Γ -local Morava E -homology functor is defined by

$$E_\Gamma^\vee(X) := \pi_* L_\Gamma(E_\Gamma \wedge X).$$

It is valued in topologized modules over $E_{\Gamma*}$ equipped with a continuous coaction of $E_\Gamma^\vee E_\Gamma$. Equivalently, it can be taken to have values in sheaves on the Lubin–Tate stack for Γ . As with the noncontinuous version, we will abbreviate $E_\Gamma^\vee$ to $E^\vee$ when the fixed formal group Γ need not be emphasized.

Remark 3.3.3 [24, Proposition 8.4(e–f)] The notation $E_\Gamma^\vee$ stems from the following result: if $E_\Gamma^*(X)$ is pro-free and even-concentrated, then

$$E_\Gamma^\vee(X) = (E_\Gamma^*(X))^\vee$$

is its continuous linear dual. It follows that the acyclics for $E_\Gamma^\vee$ -homology agree with the acyclics for K_Γ -homology [24, Proposition 2.5], completing the collection mentioned in Section 2.3. The main point is that there is a Bockstein spectral sequence taking K_Γ -homology as input and converging to $E_\Gamma^\vee$ -homology. If the K_Γ -homology is even-concentrated, this spectral sequence collapses and the $E_\Gamma^\vee$ -homology is pro-free on any basis lifting a basis of the K_Γ -homology.¹⁸

¹⁷It is interesting to revisit the ideas of Section 2.5 in the context of this example: $B^{q-1}S^1$ is the unstable based loop space of B^qS^1 , which has very complicated K -homology, but $\mathbb{S} \square_{\Sigma_+^\infty B^qS^1} \mathbb{S}$ recovers a “delooping” of B^qS^1 in the Γ -local category with simple K -homology. One might also compare this situation with that of $BU(n)$, where the K -homology of $\mathbb{S} \square_{\Sigma_+^\infty BU(n)} \mathbb{S}$ agrees with $K_*U(n)$.

¹⁸This explains the notation: for favorable X , $E_\Gamma^\vee(X)$ is the continuous linear dual of $E_\Gamma^*(X)$.

Theorem 3.3.4 [21, Proposition 7.5] *There is a factorization*

$$\begin{array}{ccccc}
 & & K_\Gamma & & \\
 & \nearrow & & \searrow & \\
 \text{SPECTRA}_\Gamma & \xrightarrow{E_\Gamma^\vee} & \text{MODULES}_{E_{\Gamma^*}, \text{Aut}(\Gamma)} & & \text{MODULES}_{K_{\Gamma^*}} \\
 \uparrow & & \uparrow & & \uparrow \\
 \text{Pic}(\text{SPECTRA}_\Gamma) & \xrightarrow{E_\Gamma^\vee} & \text{LINES}_{E_{\Gamma^*}, \text{Aut}(\Gamma)} & \xrightarrow{/\mathfrak{m}} & \text{LINES}_{K_{\Gamma^*}} \\
 & \searrow & & \nearrow & \\
 & & K_\Gamma & &
 \end{array}$$

Both squares are pullback squares: a Γ -local spectrum is invertible if and only if its continuous E -homology is a line, which is true if and only if its K -homology is a line. Additionally, for $p \gg d$ the map labeled $E_\Gamma^\vee$ in the lower left is an injection on objects.¹⁹ $\square$

We are moved by this theorem to understand the Morava E -homology of Eilenberg–Mac Lane spaces, together with the $\text{Aut } \Gamma$ action. By [Theorem 3.3.1](#) $K_* B^q S^1[p^j]$ is even-concentrated for any choice of q , and hence by [Remark 3.3.3](#) $E^\vee B^q S^1[p^j]$ is a pro-free module. We pursue the Hopf ring structure on these objects, in the hopes that this can be used to extract the $\text{Aut } \Gamma$ action. Taking a cue from the proof of [Remark 3.3.3](#), we begin by considering the short exact sequence of coefficients:

$$0 \rightarrow \bigoplus_{\substack{t_0, \dots, t_{d-1} \geq 0 \\ t_+ = r}} K_* \{[p^{t_0} \cdots u_{d-1}^{t_{d-1}}]\} \rightarrow E_*/\mathfrak{m}^{r+1} \rightarrow E_*/\mathfrak{m}^r \rightarrow 0.$$

Because we know that $E^\vee B^q S^1[p^j]$ is a pro-free (and thus flat [\[24, Theorem A.9\]](#)) E_* -module deforming the original $K_* B^q S^1[p^j]$, we tensor against $E^\vee B^q S^1[p^j]$ to get a new short exact sequence appearing as the top row in [Figure 4](#). We can also build the free alternating Hopf ring $(E^\vee B S^1[p^j])^{\wedge^*}$, and tensoring the above short exact sequence of coefficients with any graded piece of this ring gives the exact sequence (which is not, a priori, left exact) on the bottom row of [Figure 4](#). Finally, the cup product map $(B S^1[p^j])^{\wedge^q} \rightarrow B^q S^1[p^j]$ induces a map on homology

$$(E^\vee B S^1[p^j])^{\otimes t} \rightarrow (E^\vee B S^1[p^j])^{\wedge^q} \xrightarrow{\circ} E^\vee B^q S^1[p^j].$$

Bifactoriality of the tensor product induces the map between these rows by $\circ$ -product.

¹⁹Specifically: $2p - 2 \geq d^2$ and $p \neq 2$.

$$\begin{array}{ccccccc}
 \bigoplus_{\substack{t_0, \dots, t_{d-1} \geq 0 \\ t_+ = r}} (K_* BS^1[p^j])^{\wedge q} \{[p^{t_0} \dots u_{d-1}^{t_{d-1}}]\} & \xrightarrow{(E^\vee BS^1[p^j])^{\wedge q}} & \bigotimes_{E_* / \mathfrak{m}^{r+1}}^{E^\vee BS^1[p^j]} & \xrightarrow{(E^\vee BS^1[p^j])^{\wedge q}} & \bigotimes_{E_* / \mathfrak{m}^r}^{E^\vee BS^1[p^j]} & \rightarrow & 0 \\
 \downarrow & & \downarrow & & \downarrow & & \\
 0 \rightarrow \bigoplus_{\substack{t_0, \dots, t_{d-1} \geq 0 \\ t_+ = r}} (K_* B^q S^1[p^j]) \{[p^{t_0} \dots u_{d-1}^{t_{d-1}}]\} & \xrightarrow{} & \bigotimes_{E_* / \mathfrak{m}^{r+1}}^{E^\vee B^q S^1[p^j]} & \xrightarrow{} & \bigotimes_{E_* / \mathfrak{m}^r}^{E^\vee B^q S^1[p^j]} & \rightarrow & 0
 \end{array}$$

 Figure 4: Diagram of short exact sequences in [Theorem 3.3.5](#).

Theorem 3.3.5 (cf [\[20, Section 3.4\]](#)) *There is an isomorphism of Hopf rings*

$$\bigoplus_{q=0}^{\infty} (E^\vee BS^1[p^j])^{\wedge q} \cong \bigoplus_{q=0}^{\infty} E^\vee B^q S^1[p^j].$$

Proof We perform an induction on r . In the base case of $r = 1$, [Theorem 3.3.1](#) gives a chain of isomorphisms

$$\begin{aligned}
 (E^\vee BS^1[p^j])^{\wedge q} \otimes_{E_*} E_* / \mathfrak{m}^1 &= (K_* BS^1[p^j])^{\wedge q} && \text{(Remark 3.3.3)} \\
 &\cong K_* B^q S^1[p^j] && \text{(Theorem 3.3.1)} \\
 &= E^\vee B^q S^1[p^j] / \mathfrak{m}^1 && \text{(Remark 3.3.3).}
 \end{aligned}$$

This also tells us that the left-hand vertical map in [Figure 4](#) is always an isomorphism. In particular, this map is injective, as is the first nontrivial horizontal map on the second row, so their composite is injective. It follows that the first horizontal map on the first row

$$\bigoplus_{\substack{t_0, \dots, t_{d-1} \geq 0 \\ t_+ = r}} (K_* BS^1[p^j])^{\wedge q} \{[p^{t_0} \dots u_{d-1}^{t_{d-1}}]\} \rightarrow (E^\vee BS^1[p^j] / \mathfrak{m}^{r+1})^{\wedge q}$$

is also injective, and thus that the top sequence is short exact.

Then, assume that $\circ$ -multiplication induces an isomorphism modulo $\mathfrak{m}^r$ for some fixed r , ie that the right-hand vertical map in the above diagram is an isomorphism of modules. As the left-hand and right-hand vertical maps are isomorphisms, the center map must be as well. As t varies, the center maps additionally assemble into a map of graded Hopf rings, and so furthermore induce an isomorphism of such. Induction provides isomorphisms for all r , and the Milnor sequence finishes the argument:

$$\begin{aligned}
E^\vee B^q S^1[p^j] &= \lim_{\text{MODULES}_{E_*}}^r (E^\vee B^q S^1[p^j]/\mathfrak{m}^r) \\
&= \lim_{\text{MODULES}_{E_*}}^r (E^\vee B S^1[p^j]/\mathfrak{m}^r)^{\wedge q} \\
&= (E^\vee B S^1[p^j])^{\wedge q}. \quad \square
\end{aligned}$$

Remark 3.3.6 Given this calculation of the Hopf algebra structure on $E^\vee B^q S^1[p^j]$, one might be led to wonder whether this translates into a presentation of $(B^q S^1[p^j])_E$ as the q^{th} exterior power of the p^j -torsion of $\tilde{\Gamma}$. This turns out to be true, but making sense of the statement is quite complicated: Buchstaber and Lazarev [8] and Goerss [14] showed that Dieudonné theory admits a notion of exterior power, Hedayatzadeh [18; 17] showed that alternating powers of 1-dimensional p -divisible groups can be constructed intrinsically in special cases (and are compatible with the Dieudonné theory model), and Hopkins and Lurie [20, Section 3.5] give another manual (but coordinate-free!) construction relevant to the case at hand.

For the remainder of this section, we will need access to only two values of j :

$$H_q = L_\Gamma \Sigma_+^\infty B^d S^1, \quad T_q = L_\Gamma \Sigma_+^\infty B^d S^1[p].$$

We are now in a position to identify the $\text{Aut } \Gamma$ -representation structure of the E_* -line $E^\vee T_+ H_d$, using the $\text{Aut } \Gamma$ -equivariant map

$$E^\vee T_1^{\times d} \xrightarrow{\circ} E^\vee T_d.$$

It will be helpful to have access to a presentation of the group $\text{Aut } \Gamma$, which arises as the group of units of the algebra $\text{End } \Gamma$. Writing S for the Frobenius endomorphism of Γ , one can take the set $\{1, S, \dots, S^{d-1}\}$ as a $\mathbb{W}(k)$ -basis for $\text{End } \Gamma$; see [29, Theorem A2.2.17].

Lemma 3.3.7 *Let $\beta_1 \in E^\vee H_1$ be an element dual to a coordinate on $\tilde{\Gamma}$. The algebro-geometric cotangent space of $(H_d)_E$ is spanned by the dual to the element*

$$\beta_* = S^0 \beta_1 \otimes S^1 \beta_1 \otimes \cdots \otimes S^{d-1} \beta_1 \in E^\vee H_1^{\times d}$$

pushed forward along the $\circ$ -product map.

Proof In K -theory, this is a rephrasing of Ravenel and Wilson [31, Theorems 5.7.(d) and 9.2.(a)]. We then couple these to the lifting methods used elsewhere in this section to conclude the same result for E -theory. $\square$

Definition 3.3.8 Left multiplication gives a map

$$\mathrm{Aut} \, \Gamma \rightarrow \mathrm{GL}_{\mathbb{W}(k)}(\mathrm{End} \, \Gamma)$$

and postcomposing with the determinant map defines the *determinant representation*

$$\mathrm{Aut} \, \Gamma \rightarrow \mathrm{GL}_{\mathbb{W}(k)}(\mathrm{End} \, \Gamma) \xrightarrow{\det} \mathbb{W}(k)^\times.$$

Theorem 3.3.9 (cf [41, Proposition 3.21]) $\mathrm{Aut} \, \Gamma$ acts on $E^\vee T_+ H_d$ as the *determinant representation*.

Proof The cup product map

$$E^\vee T_1^{\times d} \xrightarrow{\circ} E^\vee T_d$$

is surjective and respects the action of $\mathrm{Aut} \, \Gamma$, as it is induced by a map of spaces. There is a Künneth formula

$$E^\vee T_1^{\times d} \cong (E^\vee T_1)^{\otimes d},$$

and the stabilizer group intertwines with Künneth maps

$$E^\vee A \otimes_{E_*} E^\vee B \xrightarrow{\phi} E^\vee (A \times B) \quad \text{as} \quad g \cdot \phi(a \otimes b) = \phi((g \cdot a) \otimes (g \cdot b)).$$

In turn, an element $g \in \mathrm{Aut} \, \Gamma$ acts on β_* by the formula

$$\begin{aligned} g \cdot (\circ \beta_*) &= \circ(g \cdot \beta_*) = \circ(g \cdot (S^0 \beta_1 \otimes \cdots \otimes S^{d-1} \beta_1)) \\ &= \circ(g \cdot (S^0 \otimes \cdots \otimes S^{d-1}) \beta_1) \\ &= \circ(\det g (S^0 \otimes \cdots \otimes S^{d-1}) \beta_1) \\ &= (\det g)(\circ \beta_*). \end{aligned}$$

□

Corollary 3.3.10 For $p \gg d$, $T_+ H_d$ models the *determinantal sphere* $\mathbb{S}^{\det}$.

Proof Couple the above with the last part of [Theorem 3.3.4](#). □

Remark 3.3.11 $\mathbb{S}^{\det}$ is a familiar sight for chromatic homotopy theorists. Its first appearance was in work of Hopkins and Gross on describing the Γ -local homotopy type of the Brown–Comenetz dualizing spectrum [19, Theorem 6]. It has subsequently played a prominent role in the study of chromatic homotopy theory at the height 2. For instance, taking Γ to be a formal group of height 2, it has been shown to span the rest of the torsion-free part of the Γ -local Picard group [5, Theorem 8.1]. It also has been used to study duality phenomena relating to topological modular forms; see for example work of Behrens [4, Proposition 2.4.1] and of Stojanoska [35, Corollary 13.3].

Still assuming $p \gg d$, [Theorem 3.3.4](#) allows us to determine the Γ -local homotopy type of C_n^n as well. We begin with a general result:

Theorem 3.3.12 For C a coalgebraic formal curve spectrum and $p \gg d$, there is a Γ -local equivalence $C_n^n \simeq (T_\eta^* C)^{\wedge n}$.

Proof We take as inspiration the sequence of maps of k -modules

$$\begin{aligned} \frac{\langle x \rangle}{\langle x^2 \rangle} \otimes_k \frac{\langle x^{n-1} \rangle}{\langle x^n \rangle} &\rightarrow \frac{\langle x \rangle}{\langle x^2 \rangle} \otimes_{k[[x]]} \frac{\langle x^{n-1} \rangle}{\langle x^n \rangle} \\ &\rightarrow \frac{\langle x \rangle \otimes_{k[[x]]} \langle x^{n-1} \rangle}{\langle x^2 \rangle \otimes_{k[[x]]} \langle x^{n-1} \rangle + \langle x \rangle \otimes_{k[[x]]} \langle x^n \rangle} = \frac{\langle x^n \rangle}{\langle x^{n+1} \rangle}. \end{aligned}$$

The coalgebraic spectral analogue of the left-most map is the natural truncation

$$C_1^1 \square_C C_{n-1}^{n-1} = \text{Tot } \Omega(C_1^1; C; C_{n-1}^{n-1}) \rightarrow \text{Tot}^0 \Omega(C_1^1; C; C_{n-1}^{n-1}) = C_1^1 \wedge C_{n-1}^{n-1}.$$

The sum in the denominator complicates the other map, which can no longer be an isomorphism:

$$\begin{aligned} C_n^n &= \text{fiber}(C_n^\infty \rightarrow C_{n+1}^\infty) = \text{fiber}(C_1^\infty \square_C C_{n-1}^\infty \rightarrow C_1^\infty \square_C C_n^\infty) \\ &= C_1^\infty \square_C C_{n-1}^{n-1} \leftarrow C_1^1 \square_C C_{n-1}^{n-1}. \end{aligned}$$

Nonetheless, upon applying continuous E -homology the zigzag

$$E^\vee C_n^n \leftarrow E^\vee(C_1^1 \square_C C_{n-1}^{n-1}) \rightarrow E^\vee(C_1^1 \wedge C_{n-1}^{n-1})$$

determines the $\text{Aut } \Gamma$ -representation structure of $E^\vee C_n^n$ to be that of

$$E^\vee C_1^1 \otimes_{E_*} E^\vee C_{n-1}^{n-1}. \quad \square$$

Corollary 3.3.13 For $C = H_d$ and $p \gg d$, there is an equivalence $C_n^n \simeq (\mathbb{S}^{\det})^{\wedge n}$. $\square$

Remark 3.3.14 It follows from Theorems 3.3.9 and 3.3.12 that there is *no* global finite complex with a map to $B^d S^1$ which in E -cohomology projects to precisely the 1-jets. Namely, $(E_\Gamma)_* B^d S^1$ admits a filtration whose associated-graded is a product of powers of the determinant representation, and any nontrivial map in from a power of the standard representation is obstructed by Schur's lemma.

The reader should compare with the situation with $\mathbb{CP}^\infty$ and ordinary homology, where $\mathbb{CP}^1 \simeq \mathbb{S}^2$ performs this selection of the 1-jets, but selecting the second annulus is obstructed by $\eta \in \pi_1 \mathbb{S}$, available only after coning off $\mathbb{CP}^1$ or *localizing away from 2*. By localizing at Γ , we have enlarged the variety of spheres available and thus given ourselves more tools by which we can carefully select certain individual homology classes in Γ -local spectra.

The rest of the jet-space decomposition gives a remarkable filtration of the Γ –local homotopy type of $B^d S^1$, as though it were a cell complex built out of Picard-graded cells with a simple inductive structure. The global homotopy type $B^d S^1$ of course also comes with a cellular decomposition by global cells — after all, it is presented simplicially by an iterated bar construction — but this information is dramatically more complex. Morally, passing to the Γ –local category has simultaneously enlarged our notion of “cell” and simplified the homotopy type of a complicated space, at once resulting in a simple pattern not globally visible.

3.4 Comparison of $K^{\det}$ with Westerland’s R_d

Prompted by the entirety of Section 3.3, we are moved to consider the following:

Meta-question Given a spectrum X with a relationship to $\mathbb{C}P^\infty_+ =: H_1$, a *determinantal analogue* of X is a Γ –local spectrum with an analogous relationship to

$$\mathbb{X}P^\infty_+ := H_d := L_\Gamma \Sigma^{\infty}_+ B^d S^1.$$

Which classically interesting spectra appearing in chromatic homotopy theory have determinantal analogues? What are they good for?

So far, we have found the following determinantal analogues:

- The subspectra $\mathbb{C}P^n_+ \subseteq \mathbb{C}P^\infty_+$ have analogues $\mathbb{X}P^n_+ \subseteq \mathbb{X}P^\infty_+$.
- The strata $\mathbb{C}P^n_+/\mathbb{C}P^{n-1}_+ \simeq S^{2n}$ have analogues $\mathbb{X}P^n_+/\mathbb{X}P^{n-1}_+ \simeq (S^{\det})^{\wedge n}$.
- The A_∞ ring S^1_+ has an analogue $(\Sigma^{-1}S^{\det})_+$.

Finding determinantal analogues is no easy feat, and so for this section we limit ourselves to searching for direct consequences of those above. In the classical case, the indicated composite

$$\begin{array}{ccccccc} S^0 & \longrightarrow & \mathbb{C}P^1_+ & \longrightarrow & \mathbb{C}P^2_+ & \longrightarrow & \dots \longrightarrow \mathbb{C}P^\infty_+ \\ \parallel & & \updownarrow & & \uparrow & \nearrow \beta & \\ S^0 & & S^2 & & S^4 & & \dots \end{array}$$

is known as the Bott class, and a theorem of Snaith ties this class to complex K –theory:

Theorem 3.4.1 [34] *There is an equivalence*

$$\Sigma^\infty_+ \mathbb{C}P^\infty[\beta^{-1}] \xrightarrow{\cong} KU.$$

Using the determinantal analogue of the cellular decomposition as in

$$\begin{array}{ccccccc} \mathbb{S}^0 & \longrightarrow & \mathbb{X}P^1_+ & \longrightarrow & \mathbb{X}P^2_+ & \longrightarrow & \cdots \longrightarrow \mathbb{X}P^\infty_+ \\ \parallel & & \updownarrow & & \uparrow & \nearrow \beta^{\det} & \\ \mathbb{S}^0 & & \mathbb{S}^{\det} & \nearrow & (\mathbb{S}^{\det})^{\wedge 2} & \cdots & \end{array}$$

we can construct a determinantal analogue of complex K –theory by taking Snaith’s theorem as a definition.

Definition 3.4.2 We set *determinantal K –theory* to be

$$K^{\det} := \mathbb{X}P^{\infty}_+[(\beta^{\det})^{-1}].$$

Remark 3.4.3 When $d = 1$, the spectra H_d and H_1 agree, as do the homotopy classes $\beta^{\det}$ and β , hence $K^{\det}$ agrees with p –adic K –theory $KU_p^\wedge$.

We’re left with the harder part of the meta-question: what use is $K^{\det}$? Toward this end, Craig Westerland has recently considered a closely related spectrum:

Definition 3.4.4 [41, Definition 3.11] Consider the action of $\mathbb{Z}/p^\times$ on T_d by field multiplication. Averaging this action gives rise to an idempotent in K –homology which splits the suspension spectrum as

$$T_d \simeq \bigvee_{k=0}^{p-1} Z^{\wedge k},$$

where Z is a (nonuniquely specified) spectrum with the property $Z^{\wedge p} \simeq Z$. The spectrum R_d is defined by inverting the element of Picard-graded homotopy determined by the composite

$$\alpha: Z \rightarrow \bigvee_{k=0}^{p-1} Z^{\otimes k} \simeq T_d \rightarrow H_d \xrightarrow{\text{Bockstein}} L_{\Gamma} \Sigma_+^\infty B^{d+1} \mathbb{Z}_p.$$

Upon picking a coordinate on $\mathbb{C}P_K^\infty$ and applying K –homology to this composite, one sees that it selects the dual of an induced coordinate on $(B^{d+1} \mathbb{Z}_p)_K$.

Theorem 3.4.5 *There is an equivalence $K^{\det} \simeq R_d$.*

Proof We mimic the style of one of Westerland’s proofs [41, Corollary 3.18], where he shows that inverting the class α above is equivalent to inverting another class ρ , which participates in a Γ –local diagram

$$\begin{array}{ccccc} & & \Sigma_+^\infty B^{d+1} \mathbb{Z}_p & \xrightarrow{\psi^g - g} & \Sigma_+^\infty B^{d+1} \mathbb{Z}_p \\ & \nearrow \rho & \downarrow & & \downarrow \\ \mathbb{S}^{\det} & \xrightarrow{\delta} & \Sigma_+^\infty B^{d+1} \mathbb{Z}_p[\alpha^{-1}] & \xrightarrow{\psi^g - g} & \Sigma_+^\infty B^{d+1} \mathbb{Z}_p[\alpha^{-1}] \end{array}$$

Here g is a generator of $\mathbb{Z}_p^\times$ in ψ^g encodes the natural action of $\mathbb{Z}_p^\times$ on $L_\Gamma \Sigma_+^\infty B^{d+1} \mathbb{Z}_p$ (via the multiplicative action of $\mathbb{Z}_p^\times$ on $\mathbb{Z}_p$), and δ is defined as the indicated push-forward. He constructs isomorphisms [41, Propositions 3.10 and 3.12]

$$\begin{array}{ccc} K_* B^{d+1} \mathbb{Z}_p & \xrightarrow{\cong} & C(\mathbb{Z}_p, K_*) \\ \downarrow & & \downarrow \text{restrict} \\ K_*(\Sigma_+^\infty B^{d+1} \mathbb{Z}_p[\alpha^{-1}]) & \xrightarrow{\cong} & C(\mathbb{Z}_p^\times, K_*) \end{array}$$

which he uses to calculate that the two maps

$$\Sigma_+^\infty B^{d+1} \mathbb{Z}_p[\alpha^{-1}] \xrightarrow{\cong} \Sigma_+^\infty B^{d+1} \mathbb{Z}_p[\alpha^{-1}][\delta^{-1}] \xleftarrow{\cong} \Sigma_+^\infty B^{d+1} \mathbb{Z}_p[\rho^{-1}]$$

are both equivalences, essentially by calculating the images of δ and ρ in the ring $C(\mathbb{Z}_p^\times, K_*)$.

Similarly, we will investigate the two maps

$$\Sigma_+^\infty B^{d+1} \mathbb{Z}_p[(\beta^{\det})^{-1}] \rightarrow \Sigma_+^\infty B^{d+1} \mathbb{Z}_p[\rho^{-1}, (\beta^{\det})^{-1}] \leftarrow \Sigma_+^\infty B^{d+1} \mathbb{Z}_p[\rho^{-1}].$$

Using the identification above of $K_* B^{d+1} \mathbb{Z}_p$ as a ring of functions, $\beta^{\det}$ has been handcrafted so that a generator of $K_* S^{\det}$ is sent by $\beta^{\det}$ according to

$$K_* B^{d+1} \mathbb{Z}_p \xrightarrow{\cong} C(\mathbb{Z}_p, K_*), \quad K_* \beta^{\det} \mapsto (w \mapsto w \pmod{p}).$$

Westerland denotes this element of $C(\mathbb{Z}_p, K_*)$ as f_0 , and he shows that inverting either of α or ρ has the effect of inverting exactly this element [41, proof of Proposition 3.17]. It follows that both of our two maps are equivalences. $\square$

Remark 3.4.6 [41, Section 3.9] Westerland exposes a variety of remarkable features of $K^{\det}$, the grandest of which is the E_∞ -equivalence

$$K^{\det} \simeq E_\Gamma^{hS\mathbb{G}_\Gamma^\pm}.$$

Here Γ is specifically taken over $k = \mathbb{F}_{p^d}$, the group $\mathbb{G}_\Gamma$ is the extension of $\text{Aut } \Gamma$ by $\text{Gal}(\mathbb{F}_{p^d}/\mathbb{F}_p)$, and $S\mathbb{G}_\Gamma^\pm$ is the subgroup of “special” elements [41, Section 2.2] — ie it lies in the fiber sequence

$$1 \rightarrow S\mathbb{G}_\Gamma^\pm \rightarrow \mathbb{G}_\Gamma \xrightarrow{\det^\pm} \mathbb{Z}_p^\times \rightarrow 1,$$

where $\det^\pm$ acts by the determinant on $\text{Aut } \Gamma$ and by $\text{Frob} \mapsto (-1)^{d-1}$ on the Galois component. It follows that there is a short resolution

$$L_\Gamma \mathbb{S} \rightarrow K^{\det} \xrightarrow{\psi-1} K^{\det},$$

where ψ is a certain Adams-type operation inherited from the action of $\mathbb{Z}_p^\times$ on $B^{d+1}\mathbb{Z}_p$ (or, equivalently, the lingering action of $\mathbb{Z}_p^\times$ on the fixed point spectrum); see [41, Proposition 3.14] and [12, Theorem 2]. This spectrum has been investigated before, under the names “Iwasawa extension of the $K(d)$ -local sphere” and “Mahowald’s half-sphere”.

Remark 3.4.7 [41, Corollary 4.21] Westerland also describes a cellular filtration of $\Sigma_+^\infty B^{d+1}\mathbb{Z}_p$ which bears resemblance to the jet-space decomposition described here, but which is accessed by wholly different means. He constructs an analogue for $K^{\det}$ of the classical complex J -homomorphism, and he then checks that the Thom spectrum of the canonical bundle on $B^{d+1}\mathbb{Z}_p$ has the homotopy type of $(\mathbb{S}^{\det})^{-1} \wedge \Sigma^\infty B^{d+1}\mathbb{Z}_p$. This mirrors the following classical fact about projective spaces [1, Proposition 4.3]:

$$\mathrm{Thom}(\mathcal{L} - 1 \downarrow \mathbb{CP}^\infty) \simeq \Sigma^{-2} \Sigma^\infty \mathbb{CP}^\infty,$$

and more generally

$$\mathrm{Thom}(m(\mathcal{L} - 1) \downarrow \mathbb{CP}^{n-m}) \simeq (\mathbb{CP}^1)^{\wedge(-m)} \wedge \mathbb{CP}_m^n.$$

It would be interesting to know if these filtrations coincide.

4 Some open questions

Before closing, we record some research avenues left open by the present work.

4.1 Gross–Hopkins dualizing object

As remarked on in the introduction, the determinantal sphere has previously arisen in connection with the Gross–Hopkins dualizing object $\hat{\mathbb{I}}_{\mathbb{Q}/\mathbb{Z}}$ in the Γ -local category. Our identification of $\mathbb{X}\mathbb{P}^1$ with $\hat{\mathbb{I}}_{\mathbb{Q}/\mathbb{Z}}$ rests on working at a prime p satisfying $p \gg d$, so that we can avail ourselves of Theorem 3.3.4. At small primes, the picture is murkier: we conjecture that $\mathbb{X}\mathbb{P}^1$ disagrees with $\hat{\mathbb{I}}_{\mathbb{Q}/\mathbb{Z}}$, just as the Goerss–Henn–Mahowald–Rezk $\mathbb{S}[\det]$ does, but we do not have a proof that this is so — and we furthermore have neither a proof that $\mathbb{X}\mathbb{P}^1$ agrees with $\mathbb{S}[\det]$ in the small prime regime nor one that it disagrees.²⁰ Understanding the relationship between these spectra has the opportunity either to shed light on Gross–Hopkins duality or to shed light on exotic Picard groups at greater heights, both of which are interesting prospects.

²⁰The reader may also find the recent paper of Barthel, Beaudry, Goerss, and Stojanoska [2] of interest, where they give yet another construction of an object they call $\mathbb{S}(\det)$.

4.2 Chromatic splitting

The chromatic splitting conjecture describes the homotopy groups of such objects as $\mathbb{Q} \otimes L_{K(d)}\mathbb{S}$, and the Gross–Hopkins dualizing object induces a kind of Poincaré duality among these groups. Westerland has noticed that there is a factorization ι' of the fundamental class as in

$$\begin{array}{ccc} \mathbb{S}^{d+1} & \xrightarrow{\iota_{d+1}} & L_{K(d)}\Sigma_+^\infty B^{d+1}\mathbb{Z}_p \\ & \searrow \iota' & \uparrow \beta^{\det} \\ & & \mathbb{X}P^1 \end{array}$$

which at large primes can be interpreted as follows:

$$\begin{aligned} \iota' \in [\mathbb{S}^{d+1}, \mathbb{X}P^1] &= [\mathbb{S}^{d+1}, \Sigma^{d-d^2}\widehat{\mathbb{I}}_{\mathbb{Q}/\mathbb{Z}}] = [\mathbb{S}^{1+d^2}, \widehat{\mathbb{I}}_{\mathbb{Q}/\mathbb{Z}}] \\ &= \mathrm{Hom}(\pi_{-1-d^2}M_d\mathbb{S}^0, \mathbb{Q}_p/\mathbb{Z}_p). \end{aligned}$$

Conjecture 4.2.1 The class ι' is nonzero. It is dual to the top-dimensional class (ie the Poincaré dualizing class) predicted by the rational chromatic splitting conjecture.

A positive resolution of this conjecture could give a foothold on a topological explanation for the chromatic splitting conjecture.

The reader should beware that the validity of this line of questioning is quite shaky at small primes. It is directly affected by the preceding question: it is unclear how to account for the possibility that $S^{\det} \wedge \widehat{\mathbb{I}}_{\mathbb{Q}/\mathbb{Z}}^{-1}$ is a nontrivial exotic Picard element. Furthermore, Beaudry has shown the splitting conjecture to be *false* at a small prime (viz $d = 2$ and $p = 2$), and an adjusted version has not yet been set out.

4.3 Validity at small primes

Several of the utility theorems in this paper required the large prime assumption in the proofs presented here, but it is not clear that this assumption is necessary. In particular, it would be extremely desirable — and not just for the purposes of this paper! — to have a version of [Theorem 2.3.2](#) (cf [Theorem A.0.1](#)) that does not distinguish between the large and small prime cases.

4.4 Analogues of $BU(n)$

The construction of $\mathbb{X}P^\infty$ and $\mathbb{X}P^1$ given here, as well as Westerland’s construction of determinantal analogues of $L_{K(1)}BU$, of $L_{K(1)}MU$, and of a kind of “ J –map”,

may well be a premonition of a much more extensive story of “determinantal homotopy theory”. The most enticing and impactful facet of this would be to find determinantal analogues of the intermediate spaces $BU(n)$, to understand these as classifying some manner of geometric object, and to be able to interpret them as *spaces* in some kind of unstable determinantal context. Several more questions along these lines can be found at the end of Westerland’s work [41, Sections 5.2 and 7].

Appendix The homology of inverse limits of local spectra

In this section we describe a result, attributed to Hal Sadofsky and expressed in a talk by Mike Hopkins [27, Section 14], concerning the homology of inverse limits of certain local systems.²¹ None of the material in this section is original; all of it was known to (at least) Hal Sadofsky and Mike Hopkins, and is of “folk lore” status among the experts who might be interested. Sadofsky’s theorem is stated as follows:

Theorem A.0.1 [32; 33] *Let k be a field spectrum, ie let k be a ring spectrum with k_* a graded field. (In the case that k is a Morava K –theory for Γ , we require $p \gg \text{ht } \Gamma$.) Furthermore let X_α be a pro-system of k –local spectra such that k_*X_α is a Mittag-Leffler system of k_* –modules. There is then a conditionally convergent spectral sequence of signature*

$$R^* \lim_{\text{COMOD}_{k_*k}}^\alpha k_*X_\alpha \Rightarrow k_* \lim_{\text{SPECTRA}}^\alpha X_\alpha,$$

where the derived inverse limit on the left is taken in an appropriate category of comodules.

Remarks A.0.2 (1) There is room for improvement in this result: the author sees no reason to require $p \gg d$ except for his own ineptness. It is quite likely that it holds without this assumption, and a proof of this would strengthen many of the results in this paper.

(2) Locality is *the* essential assumption. For instance, set $k = H\mathbb{Q}$ and consider the system $\{\mathbb{S}^0/p^j\}_j$, with maps the natural projections. The constituent spaces in this system are all rationally acyclic, but the inverse limit is given by the p –adic sphere $(\mathbb{S}^0)_p^\wedge$. Its rational homology is $H\mathbb{Q}_*(\mathbb{S}^0)_p^\wedge \cong \mathbb{Q}_p$, and hence no such convergent spectral sequence can exist. On the other hand, first rationalizing this system produces the trivial system of zero spectra, and thus the rational homology of the system (which is empty) compares well to the rational homology of the inverse limit (which is also

²¹The interested reader can also find related results in a paper of Hovey [23].

empty). Noting that S^0/p^j is also known as the Moore spectrum $M_0(p^j)$, similar systems can also be constructed for any Morava K –theory by employing the generalized Moore spectra of Hopkins and Smith [22, Proposition 5.14].

(3) In the case $k = H\mathbb{Q}$, the algebra of cooperations is trivial, ie $H\mathbb{Q}_*H\mathbb{Q} \cong \mathbb{Q}$. It follows that the categories of $\mathbb{Q}$ –modules and $(H\mathbb{Q}_*H\mathbb{Q})$ –comodules are equivalent, and so Sadofsky’s spectral sequence becomes

$$R^* \lim_{\text{MODULES}_{\mathbb{Q}}}^{\alpha} H\mathbb{Q}_*X_{\alpha} \Rightarrow k_* \lim_{\text{SPECTRA}}^{\alpha} X_{\alpha}.$$

Because of the Mittag-Leffler condition the higher derived inverse limits vanish, and hence we have the isomorphism

$$\lim_{\text{MODULES}_{\mathbb{Q}}}^{\alpha} H\mathbb{Q}_*X_{\alpha} \cong k_* \lim_{\text{SPECTRA}}^{\alpha} X_{\alpha}$$

familiar from the Milnor sequence. In general, we will work to control the first step in this argument (ie the derived inverse limits of *comodules*), while retaining hypotheses so that the second step in the argument (ie the derived inverse limits of *modules*) remains trivial.

(4) For any sequence of spectra (X_{α}) , the inverse system $\{Y_{\alpha} := \prod_{\beta \leq \alpha} X_{\beta}\}$ with maps given by projections is Mittag-Leffler. Then, using the fiber sequence

$$\lim_{\text{SPECTRA}}^{\alpha} X_{\alpha} \rightarrow \prod_{\alpha} X_{\alpha} \rightarrow \prod_{\alpha} X_{\alpha},$$

applying k –homology gives

$$\cdots \rightarrow k_* \lim_{\text{SPECTRA}}^{\alpha} X_{\alpha} \rightarrow k_* \left(\prod_{\alpha} X_{\alpha} \right) \rightarrow k_* \left(\prod_{\alpha} X_{\alpha} \right) \rightarrow \cdots.$$

The middle and right-hand terms of this sequence are calculable by Sadofsky’s theorem, even if the inverse system $\{X_{\alpha}\}$ is itself not Mittag-Leffler, giving some foothold on that case as well.

(5) We remark yet again that computations in the E_2 –page of this spectral sequence are *exceedingly* complex and promise to shed light on some of the most longstanding conjectures in chromatic homotopy theory. Our purposes, however, only concern the behavior of this spectral sequence in a maximally degenerate case.

A.1 Homological algebra for inverse systems of comodules

Before engaging in any algebraic topology, we will first make sense of the homological algebra necessary to study derived inverse limits of pro-systems of comodules. The homological algebra of comodules is well documented elsewhere [29, Appendix 1],

but the homological algebra of *diagrams* of comodules is more scarce. Though we are ultimately interested in the case of a Hopf algebroid (E_*, E_*E) , in this section we will refer to this pair opaquely as (A, Γ) with Γ flat over A . To begin, we recall some classical results.

Lemma A.1.1 [29, Lemma A1.2.1–2] *A comodule Y of the form $Y = \Gamma \otimes_A Y'$ for an A –module Y' is said to be an extended comodule. Such comodules participate in an adjunction²²*

$$\text{MODULES}_A(M, Y') \cong \text{COMOD}_{(A, \Gamma)}(M, Y).$$

Furthermore, if I is an injective A –module, then $\Gamma \otimes_A I$ is an injective Γ –comodule, and hence $\text{COMOD}_{(A, \Gamma)}$ has enough injectives. □

Corollary A.1.2 *If $\{M_\alpha\}_\alpha$ is a pro-system of A –modules and $\{\Gamma \otimes_A M_\alpha\}_\alpha$ is the induced pro-system of extended Γ –comodules, then there is an isomorphism*

$$\Gamma \otimes_A \lim_{\text{MODULES}_A}^\alpha M_\alpha \cong \lim_{\text{COMOD}_\Gamma}^\alpha (\Gamma \otimes_A M_\alpha). \quad \square$$

Now consider the category $\text{PROCOMOD}_{(A, \Gamma)}$ of pro-systems of comodules.

Lemma A.1.3 (Sadofsky) *The category PROCOMOD_Γ has enough injectives. That is, for X a pro-system of (A, Γ) –comodules, there is a levelwise injection to an injective system J (which can additionally be taken to be Mittag-Leffler).*

Proof First choose, for each α in the indexing category, an injection $j'_\alpha: X_\alpha \rightarrow J'_\alpha$ with J'_α an injective comodule. We form a diagram J equipped with a levelwise injection $j: X \rightarrow J$ by setting $J_\alpha = \prod_{\beta \leq \alpha}^n J'_\beta$, with the structure map $J_\alpha^\beta: J_\alpha \rightarrow J_\beta$ given by projection, and with j defined by

$$j_\alpha: X_\alpha \xrightarrow{\prod_{\beta \leq \alpha} j'_\beta \circ X_\beta^\alpha} \prod_{\beta \leq \alpha} J'_\beta = J_\alpha.$$

We now check that this diagram has the relevant lifting property

$$\begin{array}{ccc} X & \xrightarrow{j} & J \\ \downarrow & \nearrow \exists k & \\ Y & & \end{array}$$

²²This adjunction should be thought of as geometrically analogous to the adjunction

$$\text{SETS}(X, Y) \cong G\text{--SETS}(G \times X, Y).$$

whenever the vertical arrow is a levelwise injection. We argue inductively: for an index α without predecessors, the diagram reduces to

$$\begin{array}{ccc} X_\alpha & \xrightarrow{j'_\alpha} & J_\alpha \\ \downarrow & \nearrow \exists k'_\alpha & \\ Y_\alpha & & \end{array}$$

which is precisely the diagram describing the classical injectivity condition. Because J_α was selected to be an injective comodule under X_α , such an extension exists. In the case of a general index α , we have the diagram

$$\begin{array}{ccccc} \prod_{\beta<\alpha} X_\beta & \xrightarrow{\prod_{\beta'\leq\beta<\alpha} j'_{\beta'}\circ X_{\beta'}^\beta} & & & \prod_{\beta<\alpha} J'_\beta \\ & \nwarrow \prod_{\beta<\alpha} X_\beta^\alpha & X_\alpha & \xrightarrow{\prod_{\beta\leq\alpha} j'_\beta\circ X_\beta^\alpha} & \prod_{\beta\leq\alpha} J'_\beta \\ & & \downarrow & \nearrow \text{dashed} & \\ & & Y_\alpha & & \\ \prod_{\beta<\alpha} Y_\beta & \xleftarrow{\prod_{\beta<\alpha} k_\beta} & & & \end{array}$$

The outer triangle exists and commutes by inductive assumption, and the connecting arrows are part of the data of the pro-systems. The dashed map is specified by a pair of morphisms $Y_\alpha \rightarrow \prod_{\beta<\alpha} J'_\beta$ and $Y_\alpha \rightarrow J'_\alpha$. The former arrow is specified by restriction to the outer triangle. For the latter arrow, we use the injectivity of J'_α to select an arrow

$$\begin{array}{ccc} X_\alpha & \xrightarrow{j'_\alpha} & J'_\alpha \\ \downarrow & \nearrow \exists & \\ Y_\alpha & & \end{array}$$

□

Remark A.1.4 It is not true that a diagram whose objects consist of levelwise injective comodules is always an injective object in inverse systems of comodules. The construction above is designed to skirt past questions of compatibility of levelwise lifts.

Because we have enough injectives, the general machinery of homological algebra applies to produce right-derived functors of the left-exact inverse limit functor

$$\lim_{\text{COMOD}\Gamma} : \text{PROCOMOD}\Gamma \rightarrow \text{COMOD}\Gamma.$$

Moreover, there is a cobar complex which performs this computation. To see this, we first remark on a consequence of the adjunction of [Lemma A.1.1](#).

Lemma A.1.5 (Sadofsky) *If M_α is a Mittag-Leffler pro-system of A –modules, then the induced system $M_\alpha \otimes_A \Gamma$ of extended Γ –comodules is flasque.*

Proof We continue to analyze the adjunction of [Lemma A.1.1](#), where it will now be convenient to give these functors names:

$$U: \text{COMOD}_\Gamma \rightleftarrows \text{MODULES}_A : C.$$

Because U is exact, there exists a Grothendieck spectral sequence

$$E_2^{s,t} \cong (R^s \lim_{\text{COMOD}_\Gamma}^\alpha)(R^t C)M_\alpha \Rightarrow R^{s+t}(C \circ \lim_{\text{MODULES}_A}^\alpha)M_\alpha.$$

Additionally, because we have assumed Γ to be flat as an A –module, C is furthermore exact, ie $R^{>0}C = 0$. The spectral sequence then specializes to an isomorphism

$$(R^s \lim_{\text{COMOD}_\Gamma}^\alpha)(M_\alpha \otimes_A \Gamma) \cong R^s(C \circ \lim_{\text{MODULES}_A}^\alpha)M_\alpha.$$

By restricting attention to Mittag-Leffler pro-systems we have ensured that $\lim_{\text{MODULES}_A}^\alpha$ is exact, so that $C \circ \lim_{\text{MODULES}_A}^\alpha$ is the composition of two exact functors and hence has no higher derived functors. Using the isomorphism, we conclude that the higher derived functors of $\lim_{\text{COMOD}_\Gamma}^\alpha$ also vanish on this system, ie the system is flasque. $\square$

Corollary A.1.6 (Sadofsky) *Write $\Omega(\Gamma; \bar{\Gamma}; -)$ for the normalized one-sided cobar cochain complex*

$$\Omega(\Gamma; \bar{\Gamma}; -)[n] = \Gamma \otimes_A \bar{\Gamma}^{\otimes n} \otimes_A -.$$

If X_α is a Mittag-Leffler pro-system of comodules, then

$$R^s \lim_{\text{COMOD}_\Gamma}^\alpha X_\alpha = H^s \lim_{\text{COMOD}_\Gamma}^\alpha \Omega(\Gamma; \bar{\Gamma}; X_\alpha).$$

Proof The cobar complex gives a functorial source of relative injective resolutions of the terms in the pro-system, hence gives a resolution of the pro-system itself. For a fixed index n , the pro-system $\Omega(\Gamma; \bar{\Gamma}; M_\alpha)[n]$ is Mittag-Leffler, hence we may apply [Lemma A.1.5](#) and the theory of flasque resolutions to conclude the indicated formula. $\square$

A.2 Sadofsky’s theorem for finite height Morava K –theories

Now we will discuss a similar theorem communicated to the author by Mike Hopkins [\[27, Section 14\]](#) as a stepping stone toward Sadofsky’s theorem for Morava K –theories.

Theorem A.2.1 (Hopkins) *Let $E(d)$ be a Johnson–Wilson spectrum and $\{X_\alpha\}_\alpha$ be a system of $E(d)$ –local spectra such that $\{E(d)_*X_\alpha\}_\alpha$ is Mittag-Leffler. There is then a convergent spectral sequence of signature*

$$R^* \lim_{\text{COMOD}_{E(d)_*E(d)}}^\alpha E(d)_*X_\alpha \Rightarrow E(d)_* \lim_{\text{SPECTRA}}^\alpha X_\alpha.$$

The proof of this theorem relies on some shorter results, useful in their own right. Our first subgoal is to show that the $E(d)$ –homology of $E(d)$ –modules results in an extended comodule, which gives us access to the limit trick in [Corollary A.1.2](#).

Lemma A.2.2 *For M an $E(d)$ –module spectrum, there is a natural isomorphism*

$$E(d)_*M \cong E(d)_*E(d) \otimes_{E(d)_*} \pi_*M.$$

Proof Since $E(d)$ is an A_∞ –ring spectrum, there is a strongly convergent spectral sequence describing the tensor of a right $E(d)$ –module spectrum N and left $E(d)$ –module spectrum M :

$$\text{Tor}_{*,*}^{E(d)_*}(N, M) \Rightarrow \pi_*(N \wedge_{E(d)} M).$$

Taking $N = E(d) \wedge E(d)$ and noting that $\pi_*N = E(d)_*E(d)$ is a flat right $E(d)_*$ –module, the spectral sequence

$$\text{Tor}_{*,*}^{E(d)_*}(E(d)_*E(d), E(d)_*M) \Rightarrow \pi_*((E(d) \wedge E(d)) \wedge_{E(d)} M)$$

is concentrated on the 0–line. Using the freeness of N , this collapse gives

$$\begin{aligned} E(d)_*E(d) \otimes_{E(d)_*} \pi_*M &\cong \pi_*(E(d) \wedge E(d) \wedge_{E(d)} M) \\ &\cong \pi_*(E(d) \wedge M) = E(d)_*M. \end{aligned} \quad \square$$

Corollary A.2.3 *If $\{M_\alpha\}_\alpha$ is a system of $E(d)$ –module spectra which is Mittag-Leffler on homotopy, then there is an isomorphism*

$$E(d)_* \lim_{\text{MODULES}_{E(d)}}^\alpha M_\alpha \cong \lim_{\text{COMOD}_{E(d)_*E(d)}}^\alpha \{E(d)_*M_\alpha\}.$$

Proof We combine the preceding results in the following sequence:

$$\begin{aligned} E(d)_* \lim_{\text{MODULES}_{E(d)}}^\alpha M_\alpha &\cong E(d)_*E(d) \otimes_{E(d)_*} \pi_* \lim_{\text{MODULES}_{E(d)}}^\alpha M_\alpha && \text{(Lemma A.2.2)} \\ &\cong E(d)_*E(d) \otimes_{E(d)_*} \lim_{\text{MODULES}_{E(d)_*}}^\alpha \{\pi_*M_\alpha\}_\alpha && \text{(Mittag-Leffler)} \end{aligned}$$

$$\begin{aligned}
&\cong \lim_{\text{COMOD}_{E(d)*E(d)}}^{\alpha} (E(d)_*E(d) \otimes_{E(d)_*} \pi_* M_{\alpha}) && \text{(Corollary A.1.2)} \\
&\cong \lim_{\text{COMOD}_{E(d)*E(d)}}^{\alpha} E(d)_* M_{\alpha} && \text{(Lemma A.2.2).} \quad \square
\end{aligned}$$

Our next goal is to find a topological object to which we can apply [Corollary A.1.6](#). This will be a certain Adams-type spectral sequence, and because we have not really used that tool in this paper, we quickly remind the reader of its construction.

Definition A.2.4 For a ring spectrum E and spectrum X , the following diagram defines an E -Adams tower for X :

$$\begin{array}{ccccccc}
X & \longleftarrow & \bar{E} \wedge X & \longleftarrow & \bar{E}^{\wedge 2} \wedge X & \longleftarrow & \bar{E}^{\wedge 3} \wedge X \longleftarrow \dots \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
E \wedge X & & E \wedge \bar{E} \wedge X & & E \wedge \bar{E}^{\wedge 2} \wedge X & & E \wedge \bar{E}^{\wedge 3} \wedge X \dots
\end{array}$$

where $\bar{E}$ participates in the fiber sequence

$$\bar{E} \rightarrow S \xrightarrow{\eta_E} E.$$

In good cases, this spectral sequence converges conditionally to homotopy of the E -nilpotent completion of X [[7](#), Proposition 6.3]. In the case of a smashing localization, the homotopy of the E -nilpotent completion of X agrees with that of the Bousfield E -localization of X [[7](#), Corollary 6.13]. In better cases still, the E^2 -page of the spectral sequence can be identified [[29](#), Theorem 2.2.11] as

$$E_{*,*}^2 \cong \text{Cotor}_{*,*}^{E_*E}(E_*, E_*X).$$

We are now in a position to construct the spectral sequence in Hopkins' inverse limit theorem.

Definition A.2.5 Given an pro-system of spectra X_{α} , we may construct a family of interlocking fiber sequences

$$\begin{array}{ccccccc}
\lim_{\text{SPECTRA}}^{\alpha} X_{\alpha} & \longleftarrow & \lim_{\text{SPECTRA}}^{\alpha} \overline{E(d)} \wedge X_{\alpha} & \longleftarrow & \lim_{\text{SPECTRA}}^{\alpha} \overline{E(d)}^{\wedge 2} \wedge X_{\alpha} & \longleftarrow & \dots \\
\downarrow & & \downarrow & & \downarrow & & \\
\lim_{\text{SPECTRA}}^{\alpha} E(d) \wedge X_{\alpha} & & \lim_{\text{SPECTRA}}^{\alpha} E(d) \wedge \overline{E(d)} \wedge X_{\alpha} & & \lim_{\text{SPECTRA}}^{\alpha} E(d) \wedge \overline{E(d)}^{\wedge 2} \wedge X_{\alpha} & & \dots
\end{array}$$

which upon applying $E(d)$ -homology gives a spectral sequence with E^1 -page

$$E_{*,t}^1 = E(d)_* \lim_{\text{SPECTRA}}^{\alpha} (E(d) \wedge \overline{E(d)}^{\wedge t} \wedge X_{\alpha})$$

and with target $E(d)_* \lim_{\text{SPECTRA}}^\alpha X_\alpha$. If we additionally assume that the pro-system consists of $E(d)$ –local spectra such that the induced system $\{E(d)_* X_\alpha\}_\alpha$ is Mittag-Leffler, we may combine the preceding corollaries to compute

$$\begin{aligned} E(d)_* \lim_{\text{SPECTRA}}^\alpha (E(d) \wedge \overline{E(d)})^{\wedge t} \wedge X_\alpha \\ \cong E(d)_* E(d) \otimes_{E(d)_*} (E(d)_* \overline{E(d)})^{\otimes_{E(d)_*} t} \otimes_{E(d)_*} \lim_{\text{MODULES}_{E(d)_*}}^\alpha E(d)_* X_\alpha \end{aligned}$$

and hence, by [Corollary A.1.6](#),

$$E_{*,*}^2 \cong R^* \lim_{\text{COMOD}_{E(d)_* E(d)}}^\alpha E(d)_* X_\alpha.$$

Lemma A.2.6 *The $E(d)$ –Adams resolution for the $E(d)$ –local sphere $L_{E(d)}\mathbb{S}^0$ is equivalent to a finite-dimensional cosimplicial resolution by $E(d)$ –module spectra. (In particular, the $E(d)$ –Adams spectral sequence has a horizontal vanishing line.)*

Proof We recall the following terminology: a spectrum is E –nilpotent if it belongs to the thick $\otimes$ –ideal generated by E , and E –prenilpotent if it is E –locally equivalent to an E –nilpotent spectrum [\[30, Definition 7.1.6\]](#). Work of Ravenel [\[30, Lemmas 8.3.7 and 8.3.1\]](#) gives an $L_{E(d)}BP$ –prenilpotent finite spectrum F whose ordinary homology is torsion-free. Since finite prenilpotent spectra form a thick subcategory, it follows from their classification [\[22, Theorem 9\]](#) that $\mathbb{S}^0$ is $L_{E(d)}BP$ –prenilpotent. Since $L_{E(d)}BP$ and $E(d)$ share a Bousfield class [\[30, Theorem 7.3.2b and Lemma 8.1.4\]](#), it follows that $L_{E(d)}\mathbb{S}^0$ is thus $E(d)$ –nilpotent and hence has a finite $E(d)$ –Adams resolution. □

Corollary A.2.7 *Every $E(d)$ –local spectrum X has a functorial finite-dimensional cosimplicial resolution by $E(d)$ –module spectra. The length of the resolution is independent of X and dependent only on the prime p and height d .*

Proof Since $L_{E(d)}$ is a smashing localization [\[30, Theorem 7.5.6\]](#), we can smash the finite resolution for $L_{E(d)}\mathbb{S}^0$ guaranteed by [Lemma A.2.6](#) with X . □

Proof of Theorem A.2.1 Having constructed in [Definition A.2.5](#) the relevant spectral sequence, we need only address convergence. [Corollary A.2.7](#) shows that the homotopy inverse system in [Definition A.2.5](#) is weakly equivalent to a finite inverse system. It follows that, upon applying $E(d)$ –homology, the resulting spectral sequence is concentrated in a finite horizontal band and hence is strongly convergent to $E(d)_* \lim_{\text{SPECTRA}}^\alpha X_\alpha$. □

Lemma A.2.8 ([11, footnote 1]; cf [24, Proposition 6.5] and [21, Proposition 7.5]) For $p \gg d$, there exists an $E(d)$ –local Smith–Toda complex $V(d-1)$, which is a finite complex with the property $E(d) \wedge V(d-1) \simeq K(d)$.

Proof We actually demonstrate the existence of $E(d)$ –local Smith–Toda complexes $V(j)$ for $j < d$, proceeding by induction from $V(-1) = \mathbb{S}$. The homology groups $E(d)_*V(j)$ are concentrated in degrees which are multiples of $|v_1| = 2(p-1)$, and so we may also take $p > 2$ and apply Ravenel’s sparseness result [29, Proposition 4.4.2] to conclude that the $E(d)$ –based Adams spectral sequence

$$E_2^{s,t} = \text{Ext}_{E(d)_*E(d)}^{s,t}(E(d)_*V(j), E(d)_*V(j))$$

is concentrated in degrees satisfying $2(p-1) \mid t$, hence that the only nontrivial differentials lie on page indices r satisfying $2(p-1) \mid r-1$. At the same time, the finite resolution length stated qualitatively in Corollary A.2.7 can be shown quantitatively to be d^2 [29, Theorem 6.2.10(b)], so that if $2(p-1) > d^2$ there are then no r satisfying $2(p-1) \mid r-1$ with both the source and target of d_r lying within the t –band $0 \leq t \leq d^2$. It follows that the $E(d)$ –based Adams spectral sequence collapses, hence that the element

$$v_{j+1} \in \text{Ext}_{E(d)_*E(d)}^{|v_{j+1}|,0}(E(d)_*V(j), E(d)_*V(j))$$

survives, and hence that the complex $V(j+1) = V(j)/v_{j+1}$ exists. $\square$

From here, we can conclude Sadofsky’s theorem for $K(d)$.

Proof of Sadofsky’s theorem for $k = K(d)$ Use Lemma A.2.8 to extract an $E(d)$ –local Smith–Toda complex $V(d-1)$. Replace the system $\{X_\alpha\}_\alpha$ by $\{X_\alpha \wedge V(d-1)\}_\alpha$ and apply Sadofsky’s result for $E(d)$ –homology to yield the desired spectral sequence.²³ $\square$

Remark A.2.9 An odd wrinkle of this construction is that the inverse limit of the system $\{K(d)_*X_\alpha\}_\alpha$ is still taken in the category of $E(d)_*E(d)$ –comodules, *not* of $K(d)_*K(d)$ –comodules. However, the $E(d)_*E(d)$ –comodule structure of $K(d)_*X_\alpha$ factors through the Hopf algebroid $(K(d)_*, \Gamma')$, where Γ' is given by

$$\begin{aligned} \Gamma' &= K(d)_* \otimes_{BP_*} BP_* \otimes_{BP_*} K(d)_* = K(d)_*[t_1, t_2, \dots]/(v_d t_j^{p^d} - v_d^{p^j} t_j \mid j > 0) \\ &\subsetneq K(d)_*K(d) = \Gamma' \otimes \Lambda[\tau_1, \dots, \tau_{d-1}]. \end{aligned}$$

The reader should compare these stray τ_* cooperations with the Bockstein operations appearing in the proof of [24, Proposition 8.4(e–f)].

²³For a while, the author thought that because $K(d)$ admits a finite resolution in $E(d)$ –module spectra, this proof would go through at all primes, but it seems to be a dead end.

A.3 Sadofsky’s theorem for ordinary homology with field coefficients

In the case $k = HK$ for a field K of positive characteristic, all of the above constructions can be redone to produce a derived inverse limit spectral sequence for HK –homology. However, our convergence argument fails badly, as the p –complete sphere is no longer finitely resolvable by HK –module spectra — after all, the HK –Adams spectral sequence has both an infinite tower and a vanishing line of slope 1, rather than the horizontal vanishing line present in the $E(d)$ –Adams spectral sequence. It follows that the resulting spectral sequence is merely conditionally convergent, and additional hypotheses on the system $\{X_\alpha\}_\alpha$ are required to do any better. In spite of the lack of topological finiteness, Sadofsky has proven the following theorem whose proof we will not recount:

Theorem A.3.1 [32] *The HK –based inverse limit spectral sequence converges strongly in the case that $R^s \lim_{\text{MODULES}_{A*}}^\alpha H_*(X_\alpha) = 0$ for $s \gg 0$.* $\square$

References

- [1] **M F Atiyah**, *Thom complexes*, Proc. London Math. Soc. 11 (1961) 291–310 [MR](#)
- [2] **T Barthel**, **A Beaudry**, **P G Goerss**, **V Stojanoska**, *Constructing the determinant sphere using a Tate twist*, preprint (2018) [arXiv](#)
- [3] **T Barthel**, **A Beaudry**, **E Peterson**, **H Sadofsky**, *The homology of inverse limits and the algebraic chromatic splitting conjecture*, unpublished (2017)
- [4] **M Behrens**, *A modular description of the $K(2)$ –local sphere at the prime 3*, Topology 45 (2006) 343–402 [MR](#)
- [5] **M Behrens**, *The homotopy groups of $S_{E(2)}$ at $p \geq 5$ revisited*, Adv. Math. 230 (2012) 458–492 [MR](#)
- [6] **J M Boardman**, *Conditionally convergent spectral sequences*, from “Homotopy invariant algebraic structures” (J-P Meyer, J Morava, W S Wilson, editors), Contemp. Math. 239, Amer. Math. Soc., Providence, RI (1999) 49–84 [MR](#)
- [7] **A K Bousfield**, *The localization of spectra with respect to homology*, Topology 18 (1979) 257–281 [MR](#)
- [8] **V Buchstaber**, **A Lazarev**, *Dieudonné modules and p –divisible groups associated with Morava K –theory of Eilenberg–Mac Lane spaces*, Algebr. Geom. Topol. 7 (2007) 529–564 [MR](#)
- [9] **M Ching**, *Bar constructions for topological operads and the Goodwillie derivatives of the identity*, Geom. Topol. 9 (2005) 833–933 [MR](#)

- [10] **M Demazure**, *Lectures on p -divisible groups*, Lecture Notes in Mathematics 302, Springer (1986) [MR](#)
- [11] **ES Devinatz**, *Towards the finiteness of $\pi_* L_{K(n)} S^0$* , Adv. Math. 219 (2008) 1656–1688 [MR](#)
- [12] **ES Devinatz, M J Hopkins**, *Homotopy fixed point spectra for closed subgroups of the Morava stabilizer groups*, Topology 43 (2004) 1–47 [MR](#)
- [13] **V Ginzburg, M Kapranov**, *Koszul duality for operads*, Duke Math. J. 76 (1994) 203–272 [MR](#)
- [14] **P G Goerss**, *Hopfrings, Dieudonné modules, and $E_* \Omega^2 S^3$* , from “Homotopy invariant algebraic structures” (J-P Meyer, J Morava, W S Wilson, editors), Contemp. Math. 239, Amer. Math. Soc., Providence, RI (1999) 115–174 [MR](#)
- [15] **P Goerss, H-W Henn, M Mahowald, C Rezk**, *On Hopkins’ Picard groups for the prime 3 and chromatic level 2*, J. Topol. 8 (2015) 267–294 [MR](#)
- [16] **R Haugseng**, *The higher Morita category of $\mathbb{E}_n$ -algebras*, Geom. Topol. 21 (2017) 1631–1730 [MR](#)
- [17] **SM H Hedayatzadeh**, *Exterior powers of π -divisible modules over fields*, J. Number Theory 138 (2014) 119–174 [MR](#)
- [18] **SM H Hedayatzadeh**, *Exterior powers of Lubin–Tate groups*, J. Théor. Nombres Bordeaux 27 (2015) 77–148 [MR](#)
- [19] **M J Hopkins, B H Gross**, *The rigid analytic period mapping, Lubin–Tate space, and stable homotopy theory*, Bull. Amer. Math. Soc. 30 (1994) 76–86 [MR](#)
- [20] **M J Hopkins, J Lurie**, *Ambidexterity in $K(n)$ -local stable homotopy theory*, preprint (2013) Available at <https://www.math.ias.edu/~lurie/papers/Ambidexterity.pdf>
- [21] **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](#)
- [22] **M J Hopkins, J H Smith**, *Nilpotence and stable homotopy theory, II*, Ann. of Math. 148 (1998) 1–49 [MR](#)
- [23] **M Hovey**, *The generalized homology of products*, Glasg. Math. J. 49 (2007) 1–10 [MR](#)
- [24] **M Hovey, N P Strickland**, *Morava K -theories and localisation*, Mem. Amer. Math. Soc. 666, Amer. Math. Soc., Providence, RI (1999) [MR](#)
- [25] **DC Johnson, W S Wilson**, *The Brown–Peterson homology of elementary p -groups*, Amer. J. Math. 107 (1985) 427–453 [MR](#)
- [26] **J Lurie**, *Higher algebra*, book project (2017) Available at <https://url.msp.org/Lurie-HA>

- [27] **E Peterson**, *Report on E-theory conjectures seminar*, seminar notes (2013) Available at <http://chromotopy.org/latex/misc/mit-ethy.pdf>
- [28] **D C Ravenel**, *Localization with respect to certain periodic homology theories*, Amer. J. Math. 106 (1984) 351–414 [MR](#)
- [29] **D C Ravenel**, *Complex cobordism and stable homotopy groups of spheres*, Pure and Applied Mathematics 121, Academic, Orlando (1986) [MR](#)
- [30] **D C Ravenel**, *Nilpotence and periodicity in stable homotopy theory*, Annals of Mathematics Studies 128, Princeton Univ. Press (1992) [MR](#)
- [31] **D C Ravenel**, **W S Wilson**, *The Morava K-theories of Eilenberg–Mac Lane spaces and the Conner–Floyd conjecture*, Amer. J. Math. 102 (1980) 691–748 [MR](#)
- [32] **H Sadofsky**, *The homology of inverse limits of spectra*, preprint (2001)
- [33] **H Sadofsky**, *Towards the chromatic splitting conjecture*, preprint (2003)
- [34] **V Snaith**, *Localized stable homotopy of some classifying spaces*, Math. Proc. Cambridge Philos. Soc. 89 (1981) 325–330 [MR](#)
- [35] **V Stojanoska**, *Duality for topological modular forms*, Doc. Math. 17 (2012) 271–311 [MR](#)
- [36] **N P Strickland**, *A conjecture about the $K(n)$ -local Picard group*, unpublished
- [37] **N P Strickland**, *On the p -adic interpolation of stable homotopy groups*, from “Adams Memorial Symposium on Algebraic Topology, II” (N Ray, G Walker, editors), London Math. Soc. Lecture Note Ser. 176, Cambridge Univ. Press (1992) 45–54 [MR](#)
- [38] **N P Strickland**, *Formal schemes and formal groups*, from “Homotopy invariant algebraic structures” (J-P Meyer, J Morava, W S Wilson, editors), Contemp. Math. 239, Amer. Math. Soc., Providence, RI (1999) 263–352 [MR](#)
- [39] **N P Strickland**, *Gross–Hopkins duality*, Topology 39 (2000) 1021–1033 [MR](#)
- [40] **D P Sullivan**, *Geometric topology: localization, periodicity and Galois symmetry*, K–Monographs in Mathematics 8, Springer (2005) [MR](#)
- [41] **C Westerland**, *A higher chromatic analogue of the image of J* , Geom. Topol. 21 (2017) 1033–1093 [MR](#)

Department of Mathematics, Harvard University
Cambridge, MA, United States

peterston.eric.c@gmail.com

<http://chromotopy.org>

Proposed: Mark Behrens
Seconded: Stefan Schwede, Haynes R Miller

Received: 29 June 2017
Revised: 8 June 2019

Sutured manifolds and polynomial invariants from higher rank bundles

ALIAKBAR DAEMI

YI XIE

For each integer $N \geq 2$, Mariño and Moore defined generalized Donaldson invariants by the methods of quantum field theory, and made predictions about the values of these invariants. Subsequently, Kronheimer gave a rigorous definition of generalized Donaldson invariants using the moduli spaces of anti-self-dual connections on hermitian vector bundles of rank N . We confirm the predictions of Mariño and Moore for simply connected elliptic surfaces without multiple fibers and certain surfaces of general type in the case that $N = 3$. The primary motivation is to study 3-manifold instanton Floer homologies which are defined by higher rank bundles. In particular, the computation of the generalized Donaldson invariants is exploited to define a Floer homology theory for sutured 3-manifolds.

57M27, 57R57, 57R58

1. Introduction	49
2. Higher rank bundles and polynomial invariants	57
3. Floer homologies for closed 3-manifolds	74
4. Computing polynomial invariants	97
5. Sutured Floer homology	130
6. Gluing theory	142
7. Questions and conjectures	168
Appendix. Invariants of flat connections on $\Sigma(2, 3, 23)$	171
References	173

1 Introduction

Sutured manifolds were introduced by Gabai [34] to study foliations and the Thurston norm of 3-manifolds [82]. A sutured manifold is a pair of a 3-manifold M and an oriented 1-manifold $\alpha \subset M$ which decomposes the boundary of M in an appropriate

way. In [34], Gabai also defines an operation on sutured manifolds, which is called *surface decomposition*. Surface decompositions can be used to simplify sutured manifolds. Foliations of sutured manifolds are also well behaved with respect to surface decompositions. As a result, Gabai was able to construct *taut* foliations for certain families of 3-manifolds in an inductive way.

Floer homological invariants serve as another set of tools for studying topology and geometry of 3-dimensional manifolds. Such invariants were initially constructed for closed and oriented 3-manifolds: $U(N)$ -instanton Floer homology — see Floer [27; 28] and Kronheimer and Mrowka [58]; Heegaard Floer homology — see Ozsváth and Szabó [74]; monopole Floer homology — see Kronheimer and Mrowka [55]; embedded contact homology — see Hutchings [44; 45]. Later, Juhász [46] defined *sutured Floer homology*, a generalization of Heegaard Floer homology to *balanced* sutured 3-manifolds.¹ Subsequently, Kronheimer and Mrowka [56] constructed a sutured version of $U(2)$ -instanton Floer homology and monopole Floer homology, and embedded contact homology was constructed by Colin, Ghiggini, Honda and Hutchings [10; 11] and Kutluhan, Sivek and Taubes [60]. In particular, Kronheimer and Mrowka [57] used sutured $U(2)$ -instanton homology as the main ingredient to establish that Khovanov homology detects the unknot. This invariant was also used to reprove Property P for knots [56], and it lies at the core of a program that hopes to find a computer-free proof of the famous four-color theorem; see Kronheimer and Mrowka [59]. The primary motivation for this article is to extend $U(N)$ -instanton Floer homology to sutured manifolds for higher values of N .

1.1 Motivation

Fix an integer $N \geq 2$, and let K be a knot in an integral homology sphere Y . Also let μ denote an element of the knot group, $\pi_1(Y \setminus K)$, represented by a meridian of K :

Question 1.1 *Does there exist a representation $\varphi: \pi_1(Y \setminus K) \rightarrow \mathrm{SU}(N)$ with non-abelian image such that*

$$(1.2) \quad \varphi(\mu) = c \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \zeta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \zeta^{N-1} \end{bmatrix}$$

where $\zeta = e^{2\pi i/N}$, and $c = e^{\pi i/N}$ or $c = 1$ depending on whether N is even or odd?

¹For the definition of balanced sutured 3-manifolds, see Definition 5.20.

In the case that K is the unknot, the answer to the above question is clearly negative. Note also that if for a knot K , there is a representation to $\mathrm{SU}(N)$ with the mentioned properties, then there is also a desired representation from $\pi_1(Y \setminus K)$ to $\mathrm{SU}(lN)$ for any positive integer l .

Suppose Y is a homotopy sphere² and the answer to [Question 1.1](#) for any nontrivial knot K in Y is positive. A nonabelian representation φ satisfying (1.2) determines a nontrivial representation of $\pi_1(\Sigma_N(K))$ with $\Sigma_N(K)$ being the N -fold cyclic branched cover of Y , branched along K . This verifies the *covering conjecture*, which asserts that $\Sigma_N(K)$, for a nontrivial knot K , is not a homotopy sphere; see Kirby [48, Problem 3.38]. A consequence of the covering conjecture is the *Smith conjecture*, stating that a nontrivial knot is not the fixed-point set of an orientation-preserving homeomorphism $f: S^3 \rightarrow S^3$ of order N ; see Kirby [48, Problem 3.38]. The covering conjecture and the Smith conjecture have both been proved by *geometrization* techniques; see [67].

The sutured $\mathrm{U}(2)$ -instanton homology group SHI_*^2 of Kronheimer and Mrowka [56] can be employed to answer [Question 1.1](#) affirmatively for $N = 2$ (and hence for any even N) and any nontrivial knot K .³ Associated to any knot K , there is a sutured manifold $(M(K), \alpha(K))$, where $M(K)$ is the knot complement and $\alpha(K)$ is the union of two oppositely oriented meridional curves. Kronheimer and Mrowka proved that if the dimension of $\mathrm{SHI}_*^2(M(K), \alpha(K))$ is greater than 1, then there is a nonabelian representation of the knot group of K that satisfies (1.2). Similar to foliations, SHI_*^2 also behaves well with respect to surface decomposition, and one can inductively construct nontrivial elements of $\mathrm{SHI}_*^2(M(K), \alpha(K))$ after simplifying $(M(K), \alpha(K))$ by a series of sutured decompositions. In particular, the dimension of $\mathrm{SHI}_*^2(M(K), \alpha(K))$ is at least two for a nontrivial knot K . It is also shown by Boden and Friedl [5; 29] that if K is a knot with nontrivial Alexander polynomial, then the answer to [Question 1.1](#) is positive for infinitely many values of N . In light of the success of SHI_*^2 in addressing [Question 1.1](#), it is natural to look for the generalization of SHI_*^2 for higher values of N .

The essential device in the definition of the sutured Floer homology group SHI_*^2 is an excision theorem for $\mathrm{U}(2)$ -instanton Floer homology; see Braam and Donaldson [8],

²By the Poincaré conjecture, this is equivalent to saying that $Y = S^3$. However, we are making this assumption to show that our proposed approach does not require the Poincaré conjecture.

³The original notation for sutured $\mathrm{U}(2)$ -instanton homology is SHI_* . Here we use the superscript 2 to indicate that this invariant is the sutured version of $\mathrm{U}(2)$ -instanton homology.

Floer [28] and Kronheimer and Mrowka [56]. The proof of the excision theorem is in turn based on Muñoz's characterization of the structure of a $U(2)$ -instanton Floer homology group associated to the 3-manifold $S^1 \times \Sigma$, where Σ is a Riemann surface [73]. Muñoz's work borrows some results about the cohomology ring of the moduli space of rank 2 stable bundles — see Baranovskii [3], King and Newstead [47], Siebert and Tian [78] and Zagier [86] — which are not available for higher values of the rank.

In the present paper, we establish an excision theorem for $N = 3$ using the relationship between instanton Floer homology and generalizations of Donaldson invariants from Kronheimer [52]. Roughly speaking, there is a $(3+1)$ -dimensional topological quantum field theory which associates $U(N)$ -instanton Floer homology to 3-manifolds and $U(N)$ analogues of Donaldson polynomial invariants to closed 4-manifolds. This relationship between $U(2)$ -instanton Floer homology and polynomial invariants has been extensively used to compute the invariants of 4-manifolds. In this paper, we first use the TQFT structure to compute the $U(3)$ -polynomial invariants of some families of smooth 4-manifolds. Next, we work in the other direction, and use our knowledge of $U(3)$ -polynomial invariants to obtain a better understanding of certain $U(3)$ -Floer homologies. This allows us to prove the excision theorem and define a Floer homology group SHI_*^3 for sutured manifolds in the case that $N = 3$. Computations of generalized polynomial invariants in the physics literature — see Mariño and Moore [64] — suggest that our approach can be also exploited for higher values of N .

1.2 Statement of results

In his groundbreaking work [17], Donaldson defined polynomial invariants for a smooth manifold X using the moduli space of *anti-self-dual* (ASD) connections on X . In his work, X is simply connected, $b^+(X)$ is an integer greater than 1, and the ASD connections are assumed to be defined on an $SU(2)$ -bundle E over X . Although the assumption on $b^+(X)$ is essential, the definition of polynomial invariants was subsequently generalized to the case that X is not simply connected — see Kronheimer and Mrowka [54] — and E is a $U(N)$ -bundle — see Culler [12] and Kronheimer [52]. Polynomial invariants have been extensively studied in the case that $N = 2$. However, there is not much known about these invariants for higher values of N .

For a smooth and connected 4-manifold X , suppose the algebra $\mathbb{A}(X)$ is defined as

$$\mathbb{A}(X) := \text{Sym}^*(H_0(X) \oplus H_2(X)) \otimes \Lambda^*(H_1(X)),$$

where $H_i(X)$ is computed with coefficients in $\mathbb{C}$. Form the tensor product algebra $\mathbb{A}(X)^{\otimes(N-1)}$, and for $\alpha \in H_i(X)$ and $2 \leq r \leq N$, let $\alpha_{(r)}$ be the corresponding element in the $(r-1)^{\text{st}}$ factor of $\mathbb{A}(X)^{\otimes(N-1)}$. In the case that α is the generator of $H_0(X)$, this element of $\mathbb{A}(X)^{\otimes(N-1)}$ is denoted by a_r . We also define a grading on $\mathbb{A}(X)^{\otimes(N-1)}$ such that for $\alpha \in H_i(X)$, the degree of $\alpha_{(r)}$ is equal to $2r - i$. A Hermitian vector bundle E of rank N on X is determined by its first and second Chern classes. Suppose $c_1(E)$ is represented by an embedded surface w in X and $c_2(E)[X] = k$. Then the $U(N)$ -polynomial invariants associated to the bundle E are given by a linear map,⁴

$$D_{X,w,k}^N: \mathbb{A}(X)^{\otimes(N-1)} \rightarrow \mathbb{C}.$$

For $z \in \mathbb{A}(X)^{\otimes(N-1)}$, the complex number $D_{X,w,k}^N(z)$ is nonzero only if

$$(1.3) \quad \deg(z) = 4Nk - 2(N-1)w \cdot w - (N^2 - 1)\frac{1}{2}(\chi(X) + \sigma(X)).$$

Therefore, we will not lose any information if we combine these invariants as

$$(1.4) \quad D_{X,w}^N := \sum_k D_{X,w,k}^N.$$

A substantial part of the present paper is devoted to computing $U(3)$ -polynomial invariants of some families of algebraic surfaces. Our first result in this direction is the following:

Theorem 1 *Suppose X is a K3 surface. Then for any embedded oriented surface w in X and any element $z \in \mathbb{A}(X)^{\otimes 2}$,*

$$(1.5) \quad D_{X,w}^3(a_2^3 z) = 27 D_{X,w}^3(z) \quad \text{and} \quad D_{X,w}^3(a_3 z) = 0.$$

Moreover, if Γ and Λ are two elements of $H_2(X)$, then

$$(1.6) \quad D_{X,w}^3\left(\left(1 + \frac{1}{3}a_2 + \frac{1}{9}a_2^2\right) \cdot e^{\Gamma_{(2)} + \Lambda_{(3)}}\right) = e^{\frac{1}{2}\mathcal{Q}(\Gamma) - \mathcal{Q}(\Lambda)}.$$

In order to clarify the statement of the above theorem, the following remarks are in order. The left-hand side of (1.6) is defined as

$$D_{X,w}^3\left(\left(1 + \frac{1}{3}a_2 + \frac{1}{9}a_2^2\right) \cdot e^{\Gamma_{(2)} + \Lambda_{(3)}}\right) := \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{D_{X,w}^3\left(\left(1 + \frac{1}{3}a_2 + \frac{1}{9}a_2^2\right) \Gamma_{(2)}^i \Lambda_{(3)}^j\right)}{i!j!}.$$

⁴Our definition of $U(N)$ -polynomial invariants slightly differs from Culler's. See Section 2.1 for more details.

Theorem 1 asserts that the above series for a K3 surface is convergent, and the resulting number is equal to $e^{\frac{1}{2}Q(\Gamma)-Q(\Lambda)}$. Here Q denotes the intersection form of X . That is to say, $Q(\Gamma)$ is the algebraic intersection number of Γ with itself. In general, the intersection number of two homology classes Γ and Γ' is denoted by $\Gamma \cdot \Gamma'$. Since (1.6) holds for all choices of Γ and Λ , the dimension formula (1.3) allows us to compute the following polynomial invariants for all choices of nonnegative integers i and j , an integer $k \in \{0, 1, 2\}$, and homology classes Γ and Λ :

$$D_{X,w}^3(a_2^k \Gamma_{(2)}^i \Lambda_{(3)}^j).$$

These invariants determine $D_{X,w}^3(z)$ for all $z \in \mathbb{A}(X)^{\otimes 2}$, because the K3 surface satisfies (1.5) and $b_1(X) = 0$.

Our computation of the invariants of K3 surfaces motivates the following definition: a smooth 4-manifold X with $b^+(X) \geq 2$ and $b^1(X) = 0$ has w -simple type with respect to an embedded surface w if

$$(1.7) \quad D_{X,w}^3(a_2^3 z) = 27 D_{X,w}^3(z) \quad \text{and} \quad D_{X,w}^3(a_3 z) = 0$$

for all $z \in \mathbb{A}(X)^{\otimes 2}$. The 4-manifold X has *simple type* if it has w -simple type with respect to any w in X . As in the case of the K3 surfaces, if X has simple type and the series

$$\widehat{D}_{X,w}(e^{\Gamma_{(2)}+\Lambda_{(3)}}) := D_{X,w}^3\left(\left(1 + \frac{1}{3}a_2 + \frac{1}{9}a_2^2\right) \cdot e^{\Gamma_{(2)}+\Lambda_{(3)}}\right)$$

is convergent for all choices of w and $\Gamma, \Lambda \in H_2(X)$, then these series determine all polynomial invariants of X .

We can extend our calculation for the K3 surfaces to a larger family of complex surfaces. Suppose that $W(m, n)$ is the blowup of $\mathbb{CP}^1 \times \mathbb{CP}^1$ at the $4mn$ singular points of the (complex) curve

$$B := \{p_1, \dots, p_{2m}\} \times \mathbb{CP}^1 \cup \mathbb{CP}^1 \times \{q_1, \dots, q_{2n}\}.$$

Let $\widetilde{B}$ be the proper transform of B , and define $X(m, n)$ to be the branched double cover of $W(m, n)$, branched along the smooth curve $\widetilde{B}$. The horizontal and vertical fibrations of $W(m, n)$ by projective lines lift to two fibrations of $X(m, n)$ whose generic fibers are denoted by f_{m-1} and f_{n-1} . The Riemann surface f_i , for $i \in \{m-1, n-1\}$, has genus i . The complex surface $X(2, 2)$ is a K3 surface. More generally, $X(m, 2)$ is an elliptic surface without multiple fibers, which is usually denoted by $E(m)$; see Gompf and Stipsicz [40].

Theorem 2 *The elliptic surface $E(n)$ has simple type. Moreover, there are rational numbers $\hbar_1$ and $\hbar_2$ independent of n such that for any embedded surfaces w in $E(n)$ and $\Gamma, \Lambda \in H_2(E(n))$, the series $\hat{D}_{E(n),w}(e^{\Gamma(2)+\Lambda(3)})$ is equal to*

$$e^{\frac{1}{2}\mathcal{Q}(\Gamma)-\mathcal{Q}(\Lambda)}\left[\hbar_1 \cosh(\sqrt{3}f \cdot \Gamma) - 2\hbar_2 \cos\left(-\frac{2}{3}\pi w \cdot f + \sqrt{3}f \cdot \Lambda\right)\right]^{n-2},$$

where $f = f_1$ represents an elliptic fiber of $E(n)$. Furthermore, $\hbar_1 + \hbar_2 = \pm 1$ for an appropriate choice of the sign.

The constants $\hbar_1$ and $\hbar_2$ in [Theorem 2](#) are respectively equal to $\frac{2}{3}$ and $\frac{1}{3}$ (work in progress). The set of surfaces $X(m, n)$, as a set of smooth 4-manifolds, is closed with respect to taking *fiber sums*.⁵ For example, we can take the fiber sum of $X(m, n_1)$ and $X(m, n_2)$ along the fiber f_{m-1} , and the resulting 4-manifold is diffeomorphic to $X(m, n_1 + n_2)$. Given two embedded surfaces $\Sigma_1 \subset X(m, n_1)$ and $\Sigma_2 \subset X(m, n_2)$ which intersect a fiber in the same number of points, we can form a surface $\Sigma_1 \# \Sigma_2 \subset X(m, n_1 + n_2)$. Suppose $\mathcal{H}(m, n_1, n_2) \subset H_2(X(m, n_1 + n_2))$ is the space of homology classes generated by homology classes of the surfaces of the form $\Sigma \# \Sigma'$. The following theorem about $X(m, 4)$ is a consequence of [Theorem 4.93](#) about the polynomial invariants of fiber sums. In fact, [Theorem 4.93](#) can be used to obtain similar results about other surfaces in the family $X(m, n)$.

Theorem 3 *For $m \geq 3$, let $w \subset X(m, 4)$ be an embedded surface which has the form $w_1 \# w_2$ for $w_i \subset X(m, 2)$ and $w \cdot f_{m-1} \not\equiv 0 \pmod{3}$. Let K denote the canonical class of $X(m, 4)$. Then there are rational numbers $\hbar_3$ and $\hbar_4$, independent of m , such that for $\Gamma, \Lambda \in \mathcal{H}(m, 2, 2)$ the series $\hat{D}_{X(m,4),w}(e^{\Gamma(2)+\Lambda(3)})$ is convergent and is equal to*

$$e^{\frac{1}{2}\mathcal{Q}(\Gamma)-\mathcal{Q}(\Lambda)}\left[\frac{1}{2}\hbar_1^2\hbar_3^{m-2} \cosh(\sqrt{3}K \cdot \Gamma) + 2\hbar_2^2\hbar_4^{m-2} \cos\left(-\frac{2}{3}\pi w \cdot K + \sqrt{3}K \cdot \Lambda\right)\right],$$

where $\hbar_1$ and $\hbar_2$ are the constants of [Theorem 2](#).

We do not attempt to find the undetermined constants $\hbar_3$ and $\hbar_4$ here. We also believe that $X(m, 4)$ has simple type, and the above theorem holds for any choice of $w \subset X(m, 4)$ and homology classes Γ and Λ . But the current version of the theorem is sufficient for our 3-dimensional applications.

The algebraic surfaces in [Theorems 1, 2 and 3](#) are representatives of surfaces with different possible finite Kodaira dimensions. K3 surfaces, elliptic surface $E(n)$ with $n \geq 3$ and $X(m, 4)$ for $m \geq 3$ have Kodaira dimensions 0, 1 and 2, respectively.

⁵See [Section 3.3](#) for a review of the definition of fiber sums.

Theorem 1 shows that the $U(3)$ –polynomial invariants of a K3 surface associated to homology classes Γ and Λ are determined by the self-intersection of these homology classes. On the other hand, for the $U(3)$ –polynomial invariants of $E(n)$ and $X(m, 4)$ we also need the pairing of Γ and Λ with the canonical class. Recall that the first Chern class of the canonical classes of $E(n)$ and $X(m, 4)$ are represented by $(n - 2)f$ and $(m - 2)f_3 + 2f_{m-1}$, respectively.

In [Section 3](#), we introduce various Floer homology groups associated to the 3–manifolds $S^1 \times \Sigma$, and explain how these vector spaces admit ring structure. We also characterize the vector space structure on these Floer homology groups. [Theorems 2 and 3](#) allow us to obtain further information about the ring structure of these rings. In particular, in [Proposition 5.7](#), which plays an essential role for us, we construct certain elements of Floer homology groups of $S^1 \times \Sigma$ that interact nicely with the ring structure. We use these elements to obtain an excision theorem for $U(3)$ –instanton Floer homology. With the aid of this excision theorem, we construct the promised sutured Floer homology SHI_*^3 , following the approach of Kronheimer and Mrowka [\[56\]](#). This sutured Floer homology group has the following property:

Theorem 4 *For a knot K in a homology sphere Y , suppose that the dimension of $\text{SHI}_*^3(M(K), \alpha(K))$ is greater than 1. Then there is a nonabelian representation of $\pi_1(Y \setminus K)$ into $SU(3)$ that satisfies the holonomy condition [\(1.2\)](#) for $N = 3$.*

This is established by [Corollary 5.34](#). We conjecture that $\dim(\text{SHI}_*^3(M(K), \alpha(K))) > 1$ for any nontrivial knot K in a homology sphere Y such that $Y \setminus K$ is irreducible. This answers [Question 1.1](#) affirmatively for $N = 3$ and any nontrivial knot K in an integral homology sphere Y (without the irreducibility assumption on $Y \setminus K$). We hope to come back to this conjecture elsewhere.

1.3 Outline of contents

[Section 2](#) gives a review of the moduli spaces of anti-self-dual connections on 4–manifolds (possibly with boundary) and $U(N)$ –polynomial invariants. This section also contains a nonvanishing theorem for $U(N)$ –polynomial invariants of algebraic surfaces. The second half of [Section 2](#) discusses how the $U(3)$ –polynomial invariants behave in the presence of negative embedded spheres. In particular, we recall the results of Culler’s thesis [\[12\]](#) about the blowup formula for $U(3)$ –polynomial invariants and discuss how this formula can be simplified for smooth 4–manifolds with simple

type. [Section 3](#) deals with various Floer homology groups which appear in this paper. After giving an exposition of $U(N)$ –instanton Floer homology, we study various Floer homologies of $\Sigma \times S^1$ where Σ is an oriented surface. We also discuss a generalization of $U(N)$ –instanton Floer homology, which is known as Fukaya–Floer homology in the case that $N = 2$.

The Floer homology groups of [Section 3](#) are our main tools in computing $U(3)$ –polynomial invariants of several complex surfaces in [Section 4](#). In particular, the proofs of [Theorems 1, 2 and 3](#) are given in this section. In [Section 4](#), we also study the behavior of $U(3)$ –polynomial invariants with respect to fiber sum. In [Section 5](#), we prove our excision theorem and define the sutured Floer homology group SHI_*^3 . To make the exposition of the paper more comprehensible, we postpone providing proofs for technical results in [Sections 2 and 3](#) until [Section 6](#). These results are proved by gluing theory of the moduli spaces of anti-self-dual connections. [Section 7](#) concerns various questions and conjectures which naturally arise from our work on this paper.

All manifolds in this paper are smooth and oriented. Given such a manifold X , we will write $H_i(X)$ and $H^i(X)$ for the homology and cohomology groups of X with complex coefficients. If we need to work with another coefficient ring R , then we use the notation $H_i(X, R)$ and $H^i(X, R)$. Our main results for this paper concern $U(3)$ –polynomial invariants and $U(3)$ –instanton Floer homologies. However, we believe that our method for the construction of SHI_*^3 should work for arbitrary N . Therefore, we try to state our results for general N , when it is possible.

Acknowledgements We thank Lucas Culler, Simon Donaldson, Peter Kronheimer, Claude LeBrun and Tomasz Mrowka for helpful conversations. We are thankful to Victor Mikhaylov for verifying that our definition of the simple type condition matches the predictions from the physics literature. We also appreciate the referee’s close reading of the paper and helpful comments about the exposition of the paper. We are very grateful to the Simons Center for Geometry and Physics for providing a stimulating environment while we were working on this project.

2 Higher rank bundles and polynomial invariants

2.1 $U(N)$ –polynomial invariants

In this section, we review the definition of $U(N)$ –polynomial invariants of 4–manifolds based on [\[52; 12\]](#). For $N = 2$, there is an extensive literature on the subject (see, for

example, [17; 20; 68; 54]). For higher values of N , these invariants were first defined in [64] by the methods of quantum field theory. A rigorous definition of polynomial invariants for higher rank bundles is given in [52]. As we mentioned earlier, the polynomial invariants of a 4-manifold X are homomorphisms defined on the algebra $\mathbb{A}(X)^{\otimes(N-1)}$. In [52], the polynomial invariants are defined only on the subalgebra

$$\mathbb{A}(X) \otimes 1 \otimes \cdots \otimes 1.$$

Kronheimer's definition was subsequently generalized to the algebra $\mathbb{A}(X)^{\otimes(N-1)}$ in [12]. The construction of Fukaya–Floer homology in Section 6.3 is based on Culler's modification of $U(N)$ –polynomial invariants in [12]. Therefore, we attempt to give enough background on his treatment to motivate the construction of $U(N)$ –Fukaya–Floer homology.

Suppose X is a smooth, closed, oriented and connected 4-manifold, w is an oriented embedded surface in X , and k is an integer. Then there is a $U(N)$ –bundle P , unique up to isomorphism, over X such that $c_1(P) = \text{P.D.}[w]$ and $c_2(P)[X] = k$. An explicit construction of this $U(N)$ –bundle can be given as follows. Suppose $D(w)$ is a regular neighborhood of w in X whose boundary is denoted by $S(w)$. Then we can consider a Hermitian line bundle on $D(w)$ which is trivialized on $S(w)$ and whose relative first Chern class is given by the Thom class of the disc bundle $D(w)$. By extending the trivialization to the complement of $D(w)$, we obtain a Hermitian line bundle L_w where $c_1(L_w) = \text{P.D.}[w]$. The direct sum of L_w and the trivial bundle $\underline{\mathbb{C}}^{N-1}$ defines a $U(N)$ –bundle P_0 on X with $c_2(P_0)[X] = 0$ and $c_1(P_0) = c_1(L_w)$. Next, fix a $U(N)$ –bundle on the 4-dimensional ball D^4 which is trivialized on the boundary and whose relative second Chern class is given by $k \text{ P.D.}[pt]$. Removing a ball from $X \setminus D(w)$ and gluing the above ball gives rise to the same 4-manifold. We can also use the trivializations to glue the $U(N)$ –bundle on D^4 to P_0 and produce a $U(N)$ –bundle P such that $c_1(P) = \text{P.D.}[w]$, $c_2(P)[X] = k$ and the determinant bundle of P is identified with L_w .

A 2-cycle w in a closed 4-manifold is a union of embedded closed surfaces in X . We can apply the construction of the previous paragraph to obtain a Hermitian line bundle L_{w_i} for each connected component w_i of w . Then we can replace L_w in the previous paragraph with the tensor product of the line bundles L_{w_i} and produce a $U(N)$ –bundle P with $c_1(P) = \text{P.D.}[w]$ and $c_2(P)[X] = k$. The *topological energy* of P is defined to be

$$\kappa := k - \frac{N-1}{2N} w \cdot w.$$

Thus the bundle P is determined by the pair (κ, w) up to a canonical isomorphism. We say a closed 2-cycle w in X is coprime to N if there is an embedded oriented surface $\Sigma \subset X$ such that the intersection number $w \cdot \Sigma$ is coprime to N .

Suppose P is a $U(N)$ -bundle on a closed 4-manifold determined by a pair (κ, w) . Fix an integer $l \geq 3$ and an arbitrary smooth connection B_0 on L_w . Let $\mathcal{A}_\kappa(X, w)$ be the space of L^2_l connections on P whose induced connections on $\det(P) = L_w$ are each equal to B_0 . If $\mathfrak{su}(P)$ is the bundle associated to the conjugation action of $U(N)$ on the Lie algebra $\mathfrak{su}(N)$ of $SU(N)$, then $\mathcal{A}_\kappa(X, w)$ is an affine space modeled on the Banach space $L^2_l(X, \mathfrak{su}(P) \otimes \Lambda^1)$. We will also write $\mathcal{G}_\kappa(X, w)$ for the space of L^2_{l+1} automorphisms of P whose fiberwise determinant is equal to 1. Then $\mathcal{G}_\kappa(X, w)$ forms a Banach Lie group with Lie algebra $L^2_{l+1}(X, \mathfrak{su}(P))$. This Lie group acts on $\mathcal{A}_\kappa(X, w)$, and the quotient space is denoted by $\mathcal{B}_\kappa(X, w)$. We will write $[A]$ for an element of $\mathcal{B}_\kappa(X, w)$ represented by a connection A . The center of the Lie group $U(N)$ induces a finite subgroup of $\mathcal{G}_\kappa(X, w)$. If this subgroup is the stabilizer of a connection A , then A is an *irreducible* connection. Otherwise, the connection A is called *reducible*. The space of irreducible connections on P is denoted by $\mathcal{A}_\kappa^*(X, w)$, and we will write $\mathcal{B}_\kappa^*(X, w)$ for the quotient space.

Fix a Riemannian metric on X and let $*$ denote the associated Hodge operator on differential forms of X . Then $*$ defines an involution on the space of 2-forms on X , and 2-forms in the 1-eigenspace (respectively, (-1) -eigenspace) are called *self-dual* (respectively, *anti-self-dual*). A connection $A \in \mathcal{A}_\kappa(X, w)$ is *anti-self-dual* if it satisfies

$$(2.1) \quad F_0^+(A) = 0,$$

where $F_0(A)$ denotes the projection of the curvature of A to the space $\mathfrak{su}(P)$, and $F_0^+(A)$ is the self-dual part of $F_0(A)$. In other words, the connection induced by A on the associated $PU(N)$ -bundle to P has anti-self-dual curvature. Equation (2.1) is invariant with respect to the action of $\mathcal{G}_\kappa(X, w)$ and the quotient space of ASD connections is denoted by $\mathcal{M}_\kappa(X, w)$.

The local behavior of the moduli space $\mathcal{M}_\kappa(X, w)$ around an element $[A]$ is governed by the following elliptic complex, denoted by $\mathcal{D}_A$:

$$(2.2) \quad L^2_{l+1}(X, \mathfrak{su}(P)) \xrightarrow{d_A} L^2_l(X, \mathfrak{su}(P) \otimes \Lambda^1) \xrightarrow{d_A^+} L^2_{l-1}(X, \mathfrak{su}(P) \otimes \Lambda^+),$$

where Λ^+ denotes the bundle of self-dual forms on X . The i^{th} cohomology group of this complex is denoted by H_A^i . The connection A is irreducible if and only if H_A^0 is

trivial. We say A is *regular* if $H^2(A)$ is trivial. If A is an irreducible and regular ASD connection, then in a neighborhood of $[A]$, the moduli space is a smooth manifold of the same dimension as H_A^1 . In this case, the dimension of H_A^1 is given explicitly by

$$(2.3) \quad 4N\kappa - (N^2 - 1)\frac{1}{2}(\chi(X) + \sigma(X)).$$

In general, the index of the elliptic complex $\mathcal{D}_A$ is given by (2.3).

The ASD equation (2.1) can be perturbed by changing the metric on X . *Holonomy perturbations* determine another useful family of perturbations of the ASD equation [16; 80; 27; 19; 52; 58]. By abuse of notation, a solution of the perturbation of the ASD equation by a holonomy perturbation is still called an ASD connection, and the moduli space of the solutions of the perturbed equation is still denoted by $\mathcal{M}_\kappa(X, w)$. Suppose w is coprime to N and $b^+(X) \geq 1$. Then for a generic choice of the metric on X and a small holonomy perturbation the moduli space $\mathcal{M}_\kappa(X, w)$ consists of only irreducible and regular connections [52]. Therefore, the moduli space is a smooth manifold whose dimension is given by (2.3). This manifold is also orientable and in the case that N is odd [52], a canonical choice of orientation can be fixed. If N is even, to fix an orientation of the moduli space, we need an orientation of the determinant line of the elliptic operator

$$d^+ \oplus d^*: L_I^2(X, \Lambda^1) \rightarrow L_{I-1}^2(X, \Lambda^+) \oplus L_{I-1}^2(X).$$

Such an orientation of the determinant line is called a *homology orientation* of X . The $U(N)$ -polynomial invariants of X are given by integrating appropriate cohomology classes on $\mathcal{M}_\kappa(X, w)$.

The pullback of P to the product space $\mathcal{A}_\kappa(X, w) \times X$ admits an action of $\mathcal{G}_\kappa(X, w)$ which lifts the obvious action on the base. The quotient space defines a $PU(N)$ -bundle $\mathbb{P}$ over $\mathcal{B}_\kappa^*(X, w) \times X$, called the *universal bundle* associated to P . In general, $\mathbb{P}$ cannot be lifted to an $SU(N)$ -bundle. However, we can define Chern classes of $\mathbb{P}$ as rational cohomology classes of $\mathcal{B}_\kappa^*(X, w) \times X$, because the rational cohomology groups of the classifying spaces $BPU(N)$ and $BSU(N)$ are isomorphic [7]. With a slight abuse of notation, we will denote the i^{th} Chern class of $\mathbb{P}$ with $c_i(\mathbb{P})$ for $2 \leq i \leq N$.

The slant product of the Chern classes of the universal bundle and the homology classes of X gives rise to cohomology classes of $\mathcal{B}_\kappa^*(X, w)$. This construction can be used to define an algebra homomorphism,

$$(2.4) \quad \mu: \mathbb{A}(X)^{\otimes(N-1)} \rightarrow H^*(\mathcal{B}_\kappa^*(X, w)),$$

which is the unique algebra homomorphism that satisfies the property

(2.5)
$$\mu(\alpha_{(r)}) = (-1)^r c_r(\mathbb{P})/\alpha.$$

Our convention for the definition of μ slightly differs from that of [12] where $(-1)^r$ does not appear in the definition of $\mu(\alpha_{(r)})$.

Let the 2-cycle w be coprime to N , and arrange a metric and a small holonomy perturbation such that the resulting moduli space consists of irreducible and regular points. We will temporarily write $\mathcal{M}_\kappa(X, w, \pi_0)$ for the moduli space to emphasize the dependence on π_0 , which denotes the metric and the holonomy perturbation. If N is even, fix a homology orientation for X . Then $\mathcal{M}_\kappa(X, w, \pi_0)$ can be canonically oriented. Let $d = \dim(\mathcal{M}_\kappa(X, w, \pi_0))$, and fix $z \in \mathbb{A}(X)^{\otimes(N-1)}$ such that $\deg(z) = d$. If the moduli space $\mathcal{M}_\kappa(X, w, \pi_0)$ is compact, then we can evaluate $\mu(z)$ with respect to the fundamental class of $\mathcal{M}_\kappa(X, w, \pi_0)$ and obtain a number $D_{X,w,k}^N(z)$. (Recall that k is the second Chern number of P .) We wish to show that this number does not depend on π_0 , the metric and the holonomy perturbation. Suppose π_1 is another choice of a metric and a small holonomy perturbation avoiding reducible and irregular points. If $b^+(X) \geq 2$, then we can find a path $\{\pi_t\}_{0 \leq t \leq 1}$ of metrics and small holonomy perturbations such that the 1-parameter family of moduli spaces

(2.6)
$$\bigcup_t \mathcal{M}_\kappa(X, w, \pi_t)$$

is a smooth manifold of dimension $d + 1$. Because the class $\mu(z)$ can be pulled back to (2.6), Stokes’ theorem implies that π_0 and π_1 give rise to the same number $D_{X,w,k}^N(z)$, assuming (2.6) is also compact. However, $\mathcal{M}_\kappa(X, w, \pi_0)$ and the 1-parameter family of moduli spaces are not compact in general and we need to pursue a geometric approach to define the evaluation of $\mu(z)$ on $\mathcal{M}_\kappa(X, w, \pi_0)$.

Uhlenbeck compactification of the moduli space $\mathcal{M}_\kappa(X, w)$ compensates for the noncompactness of this space. Suppose $\{[A_n]\}$ is a sequence of the elements of $\mathcal{M}_\kappa(X, w)$. Then there is a multiset $\mathbf{x}$ of m points in X and a connection $A_\infty \in \mathcal{M}_{\kappa-m}(X, w)$ such that, after passing to a subsequence, $(h_n)_*(A_n)$ is L_1^p -convergent to A_∞ on $X \setminus \mathbf{x}$ for any given real number p [52, Proposition 11]. Here h_n is an isomorphism from the $U(N)$ -bundle carrying A_n to the $U(N)$ -bundle carrying A_∞ , which is defined only on $X \setminus \mathbf{x}$ and whose determinant does not depend on n .

Another key input is that $\mathbb{P}$ can be replaced with a Hermitian vector bundle [12]. Consider the standard representation of $SU(N)$ on $\mathbb{C}^N$. The tensor product $SU(N)$ -space $(\mathbb{C}^N)^{\otimes N}$ induces a representation of the group $PU(N)$. Therefore, we can

associate a vector bundle $\mathbb{E}$ of rank N^N to $\mathbb{P}$. We call $\mathbb{E}$ the *universal complex vector bundle*. The Chern class $c_i(\mathbb{P})$ can be written as a polynomial in terms of Chern classes $c_j(\mathbb{E})$ for $2 \leq j \leq i$. For example, $c_2(\mathbb{P})$ is equal to $(1/N^N)c_2(\mathbb{E})$. Therefore, it suffices to define $D_{X,w,k}^N(z)$ for the elements $z \in \mathbb{A}(X)^{\otimes(N-1)}$ such that

$$(2.7) \quad \mu(z) = c_{i_1}(\mathbb{E})/\alpha_1 \cdots c_{i_m}(\mathbb{E})/\alpha_m \quad \text{for } 2 \leq i_j \leq N.$$

In order to define this polynomial invariant, let Σ_i be a submanifold of codimension at least 2 which represents the homology class α_i .

The Chern classes of a vector bundle V of rank r over a manifold M can be represented by stratified subspaces of M . The vector bundle $\text{Hom}(\mathbb{C}^{r-i+1}, V)$ is stratified by rank. If s is a generic section of V , then

$$(2.8) \quad N_i := \{x \in M \mid \text{rank}(s(x)) \leq r - i\}$$

is a stratified subspace of M with strata of even codimension. Therefore, the fundamental class of N_i determines a well-defined homology class which is the Poincaré dual of $c_i(V)$.

In order to define $D_{X,w,k}^N(z)$ for z in (2.7), fix an open neighborhood $\nu(\Sigma_j)$ of the submanifold Σ_j such that the inclusion of this open set in X induces a surjective map of fundamental groups. The unique continuation theorem implies that if the holonomy perturbation in the definition of $\mathcal{M}_\kappa(X, w)$ is small enough, then the restriction of any element of this moduli space to $\nu(\Sigma_j)$ is irreducible [52]. We also assume that $\nu(\Sigma_j)$ and $\nu(\Sigma_k)$ intersect only if Σ_j and Σ_k are embedded surfaces and each point of X lies on at most two such open neighborhoods. Analogous to $\mathbb{E}$, we can form a vector bundle $\mathbb{E}_j$ of rank N^N on $\mathcal{B}^*(\nu(\Sigma_j)) \times \nu(\Sigma_j)$. In order to produce a representative for $c_{i_j}(\mathbb{E}_j)$, we form the bundle $\text{Hom}(\mathbb{C}^{N^N-i_j+1}, \mathbb{E}_j)$, and form a subspace $V_{i_j}(\Sigma_j)$ of $\mathcal{B}^*(\nu(\Sigma_j)) \times \nu(\Sigma_j)$ as in (2.8).

The vector bundles $\mathbb{E}$ and $\mathbb{E}_j$ are related to each other. The pullback of $\mathbb{E}_j$ with respect to the restriction map $r_j: \mathcal{M}_\kappa(X, w) \times \Sigma_j \rightarrow \mathcal{B}^*(\nu(\Sigma_j)) \times \nu(\Sigma_j)$ is the restriction of the bundle $\mathbb{E}$. This suggests that $D_{X,w,k}^N(z)$ can be defined as the signed count of the points in the *cut-down moduli space*

$$\mathcal{N}_\kappa(X, w; z) := \{([A], x_1, \dots, x_m) \in \mathcal{M}_\kappa(X, w) \times \Sigma_1 \times \cdots \times \Sigma_m \mid r_j([A], x_j) \in V_{i_j}(\Sigma_j)\}.$$

For a generic choice of $V_{i_j}(\Sigma_j)$, the cut-down moduli space $\mathcal{N}_\kappa(X, w; z)$ is a compact 0-dimensional space, and the orientation of $\mathcal{M}_\kappa(X, w)$ fixes a sign on each point of this space [52; 12]. The compactness of $\mathcal{N}_\kappa(X, w; z)$ is a consequence of Uhlenbeck

compactification using a standard counting argument [17; 20; 52; 12]. Counting the points of the cut-down with respect to the associated signs defines a number which only depends on the homology class of $\Sigma_1, \dots, \Sigma_m$, and this dependence for each homology class is linear. If $b^+(X) \geq 2$, we can adapt the geometric counting argument to the moduli space (2.6), and show that the invariant $D_{X,w,k}^N$ does not depend on the choice of the metric on X and the holonomy perturbation of the ASD equation.

There is also a standard trick which allows us to define $D_{X,w,k}^N$ in the case that w is not coprime to N . Suppose $\hat{X}$ denotes the blowup of X at an arbitrary point. Suppose also E is the exceptional sphere of $\hat{X}$. Then the cycle $w + E$ is coprime to N . In the non-coprime case, define

$$D_{X,w,k}^N(z) := D_{\hat{X},w+E,k}^N(E_{(2)}^{N-1}z).$$

In the case that w is already coprime, the above definition turns into an identity which is proved in [52].

The invariant $D_{X,w,k}^N(z)$ is defined to be zero if $\deg(z)$ is not equal to the dimension of $\mathcal{M}_k(X, w)$. For a fixed 2-cycle w in X , the moduli spaces of ASD connections appear in different dimensions whose values mod $4N$ are constant. We use (1.4) to define $D_{X,w}^N(z)$, which is called the $U(N)$ -polynomial invariant of (X, w) evaluated at z . This number is nonzero only if

$$\deg(z) \equiv 2(N+1)w \cdot w - (N^2-1)\frac{1}{2}(\chi(X) + \sigma(X)) \pmod{4N}.$$

In particular, if N is an odd integer, the invariant $D_{X,w}^N(z)$ vanishes for the classes z such that $\deg(z)$ is not divisible by 4.

There are relationships among the polynomial invariants of X associated to different 2-cycles. If w and w' are two 2-cycles in X , then

$$(2.9) \quad D_{X,w+Nw'}^N(z) = (-1)^{cw' \cdot w'} D_{X,w}^N(z),$$

where c is zero if N is odd or divisible by 4, and is equal to 1 if N is 2 mod 4. The invariants associated to the 2-cycle w and the same 2-cycle with the reverse orientation, denoted by $-w$, are also related. As explained in [52], there is a diffeomorphism from $\mathcal{M}_k(X, w)$ to $\mathcal{M}_k(X, -w)$. This diffeomorphism is orientation-preserving if N is odd, and changes the orientation by the factor of $(-1)^{w \cdot w}$ if N is even. The diffeomorphism lifts to an antilinear isomorphism of the universal complex vector bundles. Therefore,

$$(2.10) \quad D_{X,-w}^N(z) = (-1)^{(N-1)w \cdot w} D_{X,w}^N(\tau(z)),$$

where $\tau: \mathbb{A}(X)^{\otimes(N-1)} \rightarrow \mathbb{A}(X)^{\otimes(N-1)}$ is the algebra homomorphism that maps $\alpha_{(r)}$ to $(-1)^r \alpha_{(r)}$.

Suppose $\Gamma^2, \dots, \Gamma^N$ are elements of $H_2(X)$ and $z \in \mathbb{A}(X)^{\otimes(N-1)}$. To avoid the convergence issue, we define $D_{X,w}^N(z e^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N})$ in a slightly different way compared to the introduction:

$$D_{X,w}^N(z e^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N}) := \sum_{0 \leq i_2, \dots, i_N < \infty} \frac{D_{X,w}(z (\Gamma_{(2)}^2)^{i_2} \dots (\Gamma_{(N)}^N)^{i_N})}{i_2! \dots i_N!} t_2^{i_2} \dots t_N^{i_N},$$

where t_i is a formal variable. Therefore, the $U(N)$ -series $D_{X,w}^N(z e^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N})$ is an element of the ring of formal power series $\mathbb{C}[[t_2, \dots, t_N]]$. We can also define an element of $\mathbb{C}[[t_2, \dots, t_N]]$ for linear functionals $f_2, \dots, f_N: H_2(X) \rightarrow \mathbb{C}$ and a power series $g(x) = \sum_{i=0}^{\infty} b_i x^i$:

$$(2.11) \quad g(f_2(\Gamma^2) + \dots + f_N(\Gamma^N)) := \sum_{i=0}^{\infty} b_i (f_2(\Gamma^2)t_2 + \dots + f_N(\Gamma^N)t_N)^i.$$

We use a similar convention in the case that f_i are homogeneous polynomials of higher degree (eg the intersection form Q). These conventions allow us to rephrase Theorems 1, 2 and 3 in terms of the identities of the elements of $\mathbb{C}[[t_2, t_3]]$.

Remark 2.12 Suppose X is a 4-manifold with $b^+(X) = 1$. For a generic metric and for small holonomy perturbations the moduli spaces $\mathcal{M}_\kappa(X, w)$ contains only irreducible connections. For each such choice of metric, we can apply the construction of this section to define a $U(N)$ -polynomial invariant $D_{X,w}^N$. However, this polynomial invariant depends on the choice of the metric, because a 1-parameter family of moduli spaces as in (2.6) might have reducible connections. As a result, the space of metrics on X can be divided into chambers such that $D_{X,w}^N$ is constant only inside the interior of each chamber. Such polynomial invariants have been studied for $N = 2$ in [50; 51].

2.2 Cylindrical ends and moduli spaces

One of the important themes of this article is the interplay between gauge theory on 3-manifolds and 4-manifolds. One can see the interaction by considering the analogues of the geometrical objects from the previous subsection on 4-manifolds with boundary. Suppose W is a 4-manifold with boundary Y , and fix a metric which is a product metric in a collar neighborhood of Y . A smooth 2-cycle in W is a properly embedded 2-dimensional submanifold of W . A 2-cycle w in W is a union of smooth 2-cycles in W whose boundary determines a smooth 1-manifold $\gamma \subset Y$. We will say that the

boundary of the pair (W, w) is the pair (Y, γ) . We also form a pair of noncompact manifolds (W^+, w^+) by adding the cylindrical ends $[0, \infty) \times Y$ and $[0, \infty) \times \gamma$ to W and w . As in the previous part, we can associate a $U(N)$ -bundle Q to the pair (Y, γ) . This $U(N)$ -bundle is trivialized on the complement of a regular neighborhood $D(\gamma)$ of γ , and its relative first Chern class on $D(\gamma)$ is given by the Thom class. We will also write L_γ for the determinant bundle of Q . Similar to the 4-dimensional case, let $\mathcal{B}(Y, \gamma)$ be the space of equivalence classes of connections on Q whose determinants are equal to a fixed connection on L_γ . Given $\alpha \in \mathcal{B}(Y, \gamma)$, the stabilizer of a connection representing α is denoted by Γ_α . The element α is irreducible if Γ_α is equal to the center of $SU(N)$. The bundle Q can be extended to a $U(N)$ -bundle P on W using the 2-cycle w . The bundle P also determines in an obvious way a $U(N)$ -bundle on W^+ , which will be also denoted by P .

Fix an element $\alpha \in \mathcal{B}(Y, \gamma)$, and let A_0 and A_1 be two connections on P whose restrictions to the end $[0, \infty) \times Y$ are the pullback of representatives of α . Then we say A_0 and A_1 represent the same *path* if there is an automorphism g of P with determinant 1 such that $g^*(A_1) - A_0$ is a compactly supported 1-form. The equivalence class of a connection under this equivalence relation is called a *path along (W, w) based at α* . The *topological energy* of a path p represented by a connection A is defined by the following Chern–Weil integral:

$$\kappa(p) := \frac{1}{16N\pi^2} \int_{W^+} \text{tr}(\text{ad}(F(A)) \wedge \text{ad}(F(A))).$$

Here $\text{ad}(F(A))$ is regarded as a 2-form with coefficients in $\text{End}(\mathfrak{su}(P))$. The product $\wedge$ in the integrand is induced by the wedge product of differential forms and composition of the elements of $\text{End}(\mathfrak{su}(P))$. The above integral is independent of the chosen connection A and only depends on p . The constant in front of the integral is chosen such that if p is replaced by another path p' based at α , then the energy changes by an integer.

An important special case for us is the cylinder manifold $W = [0, 1] \times Y$. Fix connections $\alpha, \beta \in \mathcal{B}(Y, \gamma)$ and let p be a path along $([0, 1] \times Y, [0, 1] \times \gamma)$ based at α and β on $\{0\} \times Y$ and $\{1\} \times Y$. Then p induces a path, in the ordinary sense, from α to β in $\mathcal{B}(Y, \gamma)$. The topological energy of the path p defines a number whose value, up to an integer, depends only on α and β . Therefore, we can fix $\beta_0 \in \mathcal{B}(Y, \gamma)$ and define a functional $\text{CS}: \mathcal{B}(Y, \gamma) \rightarrow \mathbb{R}/\mathbb{Z}$, called the *Chern–Simons functional*, where $\text{CS}(\alpha)$ is equal to the topological energy of any path from α to β_0 . Since β_0 is chosen arbitrarily, CS is well defined only up to a constant. But in the case that γ is empty, the trivial

connection Θ gives a canonical choice of β_0 . Critical points of the Chern–Simons functional are represented by connections A on Q such that

$$F_0(A) = 0.$$

A critical point $\alpha \in \mathcal{B}(Y, \gamma)$ is called *nondegenerate* if the Hessian of the Chern–Simons functional is nondegenerate at α .

Suppose α is a nondegenerate critical point of the Chern–Simons functional. Suppose also A_0 is a connection on W^+ that represents a path p based at α . Then $\mathcal{A}_p(W, w; \alpha)$ is the space of connections

$$\mathcal{A}_p(W, w; \alpha) := \{A_0 + a \mid a \in L_{l, \delta}^2(W, \mathfrak{su}(P) \otimes \Lambda^1)\},$$

where $l \geq 3$, δ is a small positive number, and $L_{l, \delta}^2(W, \mathfrak{su}(P))$ is a weighted Sobolev space defined as follows. Let t be a function on W^+ that agrees with the cylindrical coordinate on the end of W^+ . For a vector bundle E on W and a positive constant δ , the Banach space $L_{l, \delta}^2(W, E)$ is defined as $e^{-\delta t} L_k^2(W^+, E)$. Suppose $\mathcal{G}_p(W, w; \alpha)$ is also defined as

$$\mathcal{G}_p(W, w; \alpha) := \{g \in \text{Aut}(P) \mid \det(g) = 1, \nabla_{A_0} g \in L_{l, \delta}^2(W, \mathfrak{su}(P) \otimes \Lambda^1)\}.$$

Then $\mathcal{G}_p(W, w; \alpha)$ is a Banach Lie group. This group acts on $\mathcal{A}_p(W, w; \alpha)$ and the quotient space is denoted by $\mathcal{B}_p(W, w; \alpha)$. The space of elements $[A] \in \mathcal{B}_p(W, w; \alpha)$ such that $F_0^+(A) = 0$ forms the *moduli space of ASD connections associated to the path p* . This moduli space is denoted by $\mathcal{M}_p(W, w; \alpha)$.

It is useful to form a framed version of the moduli space of ASD connections. Any gauge transformation in $\mathcal{G}_p(W, w; \alpha)$ is asymptotic to an element of Γ_α . Define the *framed gauge group* $\mathcal{G}_p^0(W, w; \alpha)$ to be the subspace of the elements of $\mathcal{G}_p(W, w; \alpha)$ which are asymptotic to the trivial element of Γ_α . The framed gauge group is also a Banach Lie group and its Lie algebra can be identified with $L_{l+1, \delta}^2(W, \mathfrak{su}(P))$. We write $\tilde{\mathcal{B}}_p(W, w; \alpha)$ for the quotient of $\mathcal{A}_p(W, w; \alpha)$ with respect to the action of $\mathcal{G}_p^0(W, w; \alpha)$. The set of the elements of $\tilde{\mathcal{B}}_p(W, w; \alpha)$ which satisfy the ASD equation is called the *framed (or based) moduli space of ASD connections associated to p* and is denoted by $\tilde{\mathcal{M}}_p(W, w; \alpha)$.

There is an important relationship between the ASD equation and the Chern–Simons functional. We can define an inner product on $\mathcal{B}(Y, \gamma)$ using

$$\langle a, b \rangle := -\frac{1}{16N\pi^2} \int_Y \text{tr}(\text{ad}(a) \wedge * \text{ad}(b)) \quad \text{for } a, b \in \Omega^1(Y, \mathfrak{su}(P)),$$

where tr is defined using the trace on $\text{End}(\mathfrak{su}(N))$. Suppose $\{\alpha(t)\}_{t \in [0,1]}$ is a path in $\mathcal{B}(Y, \gamma)$. This path defines a trajectory of the downward gradient flow of CS with respect to the above metric if it satisfies

$$(2.13) \quad \frac{d\alpha(t)}{dt} = - * F_0(\alpha(t)).$$

The path $\{\alpha(t)\}$ determines a connection A , in temporal gauge, on $[0, 1] \times Y$, and (2.13) is equivalent to the ASD equation $F_0^+(A) = 0$. This relationship between the ASD equation and the Chern–Simons functional allows us to conclude from nondegeneracy of the critical points of CS that the moduli spaces $\widetilde{\mathcal{M}}_p(W, w; \alpha)$ are analytically well behaved.

The local behavior of the framed moduli space $\widetilde{\mathcal{M}}_p(W, w; \alpha)$ around an element $[A]$ is modeled by the following elliptic complex:

$$(2.14) \quad L_{l+1,\delta}^2(W, \mathfrak{su}(P)) \xrightarrow{d_A} L_{l,\delta}^2(W, \Lambda^1 \otimes \mathfrak{su}(P)) \xrightarrow{d_A^+} L_{l-1,\delta}^2(W, \Lambda^+ \otimes \mathfrak{su}(P)).$$

This complex is Fredholm and its homology groups are denoted by H_A^0 , H_A^1 and H_A^2 . Then $H_A^0 = 0$, and the element $[A]$ is called *regular* if H_A^2 is also trivial. In this case, $\widetilde{\mathcal{M}}_p(W, w; \alpha)$ is smooth in a neighborhood of A , and H_A^1 gives a model for the tangent space of the framed moduli space at $[A]$. Therefore, the index of the above complex for a (not necessarily regular) ASD connection A is called the *expected dimension* of $\widetilde{\mathcal{M}}_p(W, w; \alpha)$ and is denoted by $\dim_e(\widetilde{\mathcal{M}}_p(W, w; \alpha))$.

We slightly generalize the above discussion to include the case of 4–dimensional cobordisms. A cobordism $W: Y_0 \rightarrow Y_1$ is a 4–manifold with boundary $\bar{Y}_0 \amalg Y_1$. We also assume that a 2–cycle $w: \gamma_0 \rightarrow \gamma_1$ in W is given, and P is the associated $U(N)$ –bundle. Suppose also that α_0 and α_1 are flat connections on Y_0 and Y_1 and that p is a path along (W, w) from α_0 to α_1 . As before, we assume that α_0 and α_1 are nondegenerate. In this case, W^+ , $\mathcal{B}_p(W, w; \alpha_0, \alpha_1)$ and $\mathcal{M}_p(W, w; \alpha_0, \alpha_1)$ are defined as in the previous case by regarding W as a 4–manifold with boundary. However, there is an alternative elliptic complex that one can associate to the elements of $\mathcal{M}_p(W, w; \alpha_0, \alpha_1)$. Suppose W^+ is the Riemannian 4–manifold given by adding cylindrical ends to W . In this case, we identify the cylindrical end corresponding to the incoming end with $(-\infty, 0] \times Y_0$. As in the previous case, suppose t is a function on W^+ that agrees with the cylindrical coordinates on the ends and define $L_{l,\delta}^2(W, E)$. Therefore, an element of $L_{l,\delta}^2(W, E)$ is allowed to have exponential growth on the incoming end and it is forced to have an exponential decay on the outgoing end.

For $[A] \in \mathcal{B}_p(W, w; \alpha_0, \alpha_1)$, the ASD operator $\mathcal{D}_A$ is defined by

$$(2.15) \quad d_A^* \oplus d_A^+ : L_{l,\delta}^2(W, \Lambda^1 \otimes \mathfrak{su}(P)) \rightarrow L_{l-1,\delta}^2(W, (\Lambda^0 \oplus \Lambda^+) \otimes \mathfrak{su}(P)).$$

The operator $\mathcal{D}_A$ is elliptic, and the excision property of elliptic operators shows that the index of $\mathcal{D}_A$ is additive with respect to composition of cobordisms [19, Section 3.3]. It can be computed explicitly using the Atiyah–Patodi–Singer index theorem [81; 70]:

$$(2.16) \quad \begin{aligned} \text{index}(\mathcal{D}_A) \\ = 4N\kappa(p) - (N^2 - 1) \frac{\chi(W) + \sigma(W)}{2} + \sum_{i=0,1} (-1)^i \frac{h^0(Y_i; \text{ad}_{\alpha_i}) - \rho_{\text{ad}_{\alpha_i}}(Y_i)}{2}, \end{aligned}$$

where $h^0(Y_i; \text{ad}_{\alpha_i})$ denotes the dimension of $H^0(Y_i; \text{ad}_{\alpha_i})$. Moreover, for a flat connection a on a vector bundle V over Y , ρ_a is the Atiyah–Patodi–Singer ρ -invariant of a [2]. As an example, let χ_j be a $U(1)$ -connection on $L(a, b)$ whose holonomy around the standard generator of $\pi_1(L(a, b))$ is equal to $e^{2\pi i j/a}$. It is shown in [2] that

$$(2.17) \quad \rho_{\chi_j}(L(a, b)) = -\frac{4}{a} \sum_{k=1}^{a-1} \cot\left(\frac{\pi k}{a}\right) \cot\left(\frac{\pi k b}{a}\right) \sin^2\left(\frac{\pi k j}{a}\right).$$

Suppose a 4-manifold W is regarded as a cobordism from the empty 3-manifold to the boundary of W . Then $\dim_e(\widetilde{\mathcal{M}}_p(W, w; \alpha))$ is equal to $\text{index}(\mathcal{D}_A) + h^0(\alpha)$ where A is a connection that represents the path p .

In the case of a cylinder $[0, 1] \times Y$, the ASD operator can be used to define a relative $\mathbb{Z}/4N\mathbb{Z}$ -grading on critical points of the Chern–Simons functional associated to a pair (Y, γ) . Fix an arbitrary critical point β_0 of CS, and let α be another critical point of CS. Let the connection A represent an arbitrary path p from α to β_0 . Then the Floer grading of α , denoted by $\deg(\alpha)$, is defined to be $\text{index}(\mathcal{D}_A) \bmod 4N$. Since the choice of β_0 is arbitrary, this grading only gives rise to a relative $\mathbb{Z}/4N\mathbb{Z}$ -grading. In the case that γ is empty, we can make this grading absolute by requiring $\beta_0 = \Theta$. In this case, we can invoke the index formula (2.16) to compute $\deg(\alpha)$ as follows:

$$(2.18) \quad \deg(\alpha) \equiv 4N \cdot \text{CS}(\alpha) - \frac{1}{2}(N^2 - 1) + \frac{1}{2}(h^0(Y; \text{ad}_{\alpha}) - \rho_{\text{ad}_{\alpha}}(Y)) \bmod 4N.$$

Analogous to the case of closed 4-manifolds, we can avoid nonregular points in $\mathcal{M}_p(W, w; \alpha)$ by perturbing the ASD equation. Suppose α is irreducible and nondegenerate. Then there are small holonomy perturbations of the ASD equation, supported in $[0, 1] \times Y \subset W^+$, such that the resulting moduli spaces consist of regular points [58]. Alternatively, if $b^+(X) \geq 2$ and w is coprime to N , then we can arrange for a small

perturbation of the ASD equation such that the moduli spaces $\mathcal{M}_p(W, w; \alpha)$, for arbitrary α and p with $\kappa(p)$ bounded, are regular [52]. In the case that $\mathcal{M}_p(W, w; \alpha)$ consists of only regular points, it is an orientable smooth manifold.

In general, the critical points of the Chern–Simons functional might not be nondegenerate. Suppose that all critical points of the Chern–Simons functional associated to a pair (Y, γ) are irreducible. Then, CS can be also perturbed by appropriate perturbations such that the critical points of the resulting functional are irreducible and nondegenerate [58, Proposition 3.10]. The perturbations used in [58] are also defined in terms of the holonomies of connections on Y . Suppose CS_π is such a perturbation of CS by a small holonomy perturbation and α is a critical point of CS_π . Suppose also (W, w) is a pair whose boundary is equal to (Y, γ) . The negative-gradient flow line of CS_π determines a perturbation of the ASD equation on the end $[0, \infty) \times Y$ of W^+ . This perturbation can be extended to W^+ so that the corresponding moduli space contains only regular points [58]. As in the previous case, the elliptic operator (2.15) can be used to study the local behavior of the moduli spaces. In particular, in the case of a cylinder, the index of $\mathcal{D}_A$ can be used to define a relative $\mathbb{Z}/4N\mathbb{Z}$ -grading on the critical points of CS_π .

2.3 Nonvanishing theorem for algebraic surfaces

For a complex projective surface, the moduli spaces of ASD connections can be identified with the moduli spaces of *stable bundles* with fixed determinant [15]. Stable bundles have been studied extensively using various techniques in algebraic geometry. Thus one can use algebrogeometric methods to extract information about the polynomial invariants of a complex projective surface. For example, the following theorem about nonvanishing of $U(N)$ -polynomial invariants of algebraic surfaces is a generalization of Donaldson's celebrated theorem about $U(2)$ -invariants [17; 20]:

Theorem 2.19 *Suppose X is a complex projective surface with positive geometric genus, h is a hyperplane class (or equivalently a very ample class), and w is a 2-cycle representing c_1 of a holomorphic line bundle L . Suppose also w is coprime to N . Then*

$$D_{X,w}^N(h_{(2)}^d) > 0$$

when d is large enough and

$$d \equiv (N+1)w \cdot w - (N^2-1)\frac{1}{4}(\chi(X) + \sigma(X)) \pmod{2N}.$$

Proof Let $M_\kappa^N(X, L)$ be the moduli space of stable bundles with rank N , determinant L , and energy κ . We first review the proof of the nonvanishing theorem in the rank 2 case. In this case, the key idea is to find a projective embedding of $M_\kappa^2(X, L)$ and then interpret the polynomial invariant $D_{X,w}^2(h_{(2)}^d)$ as a multiple of the degree of the moduli space. The main steps of the proof can be summarized as follows:

- (1) Suppose C is an algebraic curve and $\mathcal{N}(C, d)$ is the moduli space of stable bundles of rank 2 and degree d on C . Then there is a projective embedding of $\mathcal{N}(C, d)$ into a projective space $\mathbb{P}(W)$ [36]. Moreover, this projective embedding is given by the sections of a large power $\mathcal{L}^d$ of the *determinant line bundle* over $\mathcal{N}(C, d)$ [20].
- (2) For any stable bundle $\mathcal{E} \in M_\kappa^2(X, L)$, there is a p_0 such that if $C \subset X$ is a generic curve in the linear system $|\mathcal{O}(p)|$, for $p \geq p_0$, then the restriction of $\mathcal{E}$ to C is also stable [66]. Using this result together with the fact that the moduli space $M_\kappa^2(X, L)$ has finite type, we can find a projective embedding $J: M_\kappa^2(X, L) \rightarrow \mathbb{P}(V)$. This embedding is given by restricting the elements of $M_\kappa^2(X, L)$ to finitely many curves in the linear system $|\mathcal{O}(p)|$ and applying the Gieseker embedding from the first step.
- (3) For κ large enough, the moduli space $M_\kappa^2(X, L)$ is not empty [79; 37].
- (4) The dimension of the irregular part of $M_\kappa^2(X, L)$ is strictly smaller than the virtual dimension of the moduli space, when κ is large enough [17].

By combining these facts, it can be shown that the map J can be chosen such that the degree of the quasiprojective variety $J(M_\kappa(X, L))$ is equal to

$$K^d D_{X,w}^N(h_{(2)}^d)$$

for an appropriate integer K , and for a large enough d . Therefore, the invariant $D_{X,w}^N(h_{(2)}^d)$ is positive. This proof can be generalized to the case that $N > 2$. Steps 1 and 2 are proved for an arbitrary rank in the references mentioned above. The generalization of the third step to the higher rank case is given in [62]. The higher rank version of the fourth fact is also proved in [38]. Then the same arguments as in [17; 20] can be used to realize $D_{X,w}^N(h_{(2)}^d)$ as a multiple of the degree of a quasiprojective variety, which verifies the claim in Theorem 2.19. $\square$

Remark 2.20 In Theorem 2.19, we assume that X is a projective surface. But a similar nonvanishing theorem can be formulated for any Kähler surface, because Kähler surfaces can be deformed into projective surfaces [49] and the $U(N)$ -polynomial invariants only depend on the smooth structure.

2.4 Negative embedded spheres

Motivated by [76], Fintushel and Stern used embedded 2–spheres with negative self-intersection to study polynomial invariants of a 4–manifold X [26; 25] (see also [83]). The same idea is exploited in [12] to obtain information about the properties of the polynomial invariants associated to higher rank bundles. Suppose τ is an embedded sphere in a 4–manifold X which has self-intersection -2 . Fix a 2–cycle w with $w \cdot \tau = 0$, and $z \in \mathbb{A}(\langle \tau \rangle^\perp)^{\otimes 2}$. In general, $\mathbb{A}(V)$ for a vector subspace $V \subseteq H_2(X)$ denotes the subalgebra $\text{Sym}^*(H_0(X) \oplus V) \otimes \Lambda^*(H_1(X))$, and $\langle \tau \rangle^\perp$ is the subspace of $H_2(X)$ consisting of the homology classes orthogonal to τ . The following formulas about the $\text{U}(3)$ –polynomial invariants of X are proved in [12, Section 3.2]:

$$(C_1) \quad D_{X,w}^3(\tau_{(2)}^2 z) = -2 D_{X,w+\tau}^3(z).$$

$$(C_2) \quad D_{X,w}^3(\tau_{(2)}^4 z) = -4 D_{X,w}^3(a_2 \tau_{(2)}^2 z) - 3 D_{X,w}^3(\tau_{(3)}^2 z).$$

$$(C_3) \quad D_{X,w}^3(\tau_{(2)}^3 \tau_{(3)} z) = -3 D_{X,w}^3(a_3 \tau_{(2)}^2 z) - D_{X,w}^3(a_2 \tau_{(2)} \tau_{(3)} z).$$

By a similar approach, we shall prove the following proposition in Section 6.2:

Proposition 2.21 *Suppose X is a smooth 4–manifold with $b^+(X) \geq 2$ and w is a 2–cycle. Suppose also σ is an embedded sphere with self-intersection -3 and $z \in \mathbb{A}(\langle \sigma \rangle^\perp)^{\otimes 2}$.*

(i) *If $w \cdot \sigma \equiv 1 \pmod{3}$, then there is a constant number c such that*

$$D_{X,w}^3\left(\left(-\frac{3}{2}\sigma_{(3)} - \frac{3}{2}\sigma_{(2)}^2 - a_2\right)z\right) = c D_{X,w-\sigma}^3(z).$$

(ii) *If $w \cdot \sigma \equiv 2 \pmod{3}$, then for the same constant c as above,*

$$D_{X,w}^3\left(\left(\frac{3}{2}\sigma_{(3)} - \frac{3}{2}\sigma_{(2)}^2 - a_2\right)z\right) = c D_{X,w+\sigma}^3(z).$$

(iii) *If $w \cdot \sigma \equiv 0 \pmod{3}$, then the following two formulas hold:*

$$\begin{aligned} D_{X,w}^3((\sigma_{(2)}^4 + 4a_2\sigma_{(2)}^2 + 3\sigma_{(3)}^2)z) &= 0, \\ D_{X,w}^3((\sigma_{(2)}^3\sigma_{(3)} + 3a_3\sigma_{(2)}^2 + a_2\sigma_{(2)}\sigma_{(3)})z) &= 0. \end{aligned}$$

In Section 4.2, we will show that the constant c in Proposition 2.21 is equal to -3 .

2.5 Blowup formula for 4–manifolds with simple type

In this subsection, we review the properties of $\text{U}(3)$ –polynomial invariants of blown-up 4–manifolds. This part has the same theme as the previous subsection, because the

exceptional divisor of a blown-up 4-manifold gives rise to a (-1) -sphere. We start with an exposition of the main result in Culler's thesis [12]. Given a 4-manifold X , let $\hat{X}$ denote the blowup of X at one point, which is diffeomorphic to the 4-manifold $X \# \overline{\mathbb{CP}}^2$. We will also denote the exceptional sphere in $\hat{X}$ with E . If w is a 2-cycle in X , then it induces a 2-cycle in $\hat{X}$, which will be also denoted by w . Similarly, we can regard $\mathbb{A}(X)$ as a subalgebra of $\mathbb{A}(\hat{X})$.

Proposition 2.22 *Suppose w is a 2-cycle in X and $z \in \mathbb{A}(X)^{\otimes 2}$. For $0 \leq i \leq 2$ and $0 \leq j \leq 1$, the invariant $D_{\hat{X},w}^3(E_{(2)}^i E_{(3)}^j z)$ is equal to $D_{X,w}^3(z)$ if $i = j = 0$ and it is zero otherwise. The invariant $D_{\hat{X},w+E}^3(E_{(3)} z)$ is equal to $D_{X,w}^3(z)$. Finally,*

$$D_{\hat{X},w+E}^3(E_{(2)}^i z) = \begin{cases} 0 & \text{if } i = 0, 1, 3, \\ D_{X,w}^3(z) & \text{if } i = 2, \\ -D_{X,w}^3(a_2 z) & \text{if } i = 4, \\ D_{X,w}^3(a_3 z) & \text{if } i = 5. \end{cases}$$

Proof As we mentioned in Section 2.1, the identity $D_{\hat{X},w+E}^3(E_{(2)}^2 z) = D_{X,w}^3(z)$ is proved in [52]. Proposition 62 in [12], together with the discussion succeeding it, essentially contains the proof of the remaining identities using a method similar to that in [52]. To be a bit more detailed, one stretches the neck region in $\hat{X}$ between X and $\overline{\mathbb{CP}}^2$, and studies the moduli spaces of ASD connections over $\hat{X}$ in terms of the moduli spaces associated to X and $\overline{\mathbb{CP}}^2$. $\square$

According to this proposition, some of the polynomial invariants of $\hat{X}$ are determined by the invariants of X . The following theorem claims that a similar pattern holds for all invariants of $\hat{X}$. This theorem is essentially proved in [12]. We just slightly expand the results of this thesis using the same methods:

Theorem 2.23 *Suppose w is a 2-cycle in X and $z \in \mathbb{A}(X)^{\otimes 2}$. For nonnegative integers i and j , there are polynomials $B_{i,j}, S_{i,j} \in \mathbb{Q}[a_2, a_3]$ independent of X, w and z which satisfy the following identities:*

$$D_{\hat{X},w}^3(e^{E_{(2)}+E_{(3)}} z) = D_{X,w}^3\left(z \sum_{i,j} B_{i,j}(a_2, a_3) t_2^i t_3^j\right),$$

$$D_{\hat{X},w+E}^3(e^{E_{(2)}+E_{(3)}} z) = D_{X,w}^3\left(z \sum_{i,j} S_{i,j}(a_2, a_3) t_2^i t_3^j\right).$$

Moreover, $B := \sum_{i,j} B_{i,j}(a_2, a_3) t_2^i t_3^j$ and $S := \sum_{i,j} S_{i,j}(a_2, a_3) t_2^i t_3^j$, as power series in the variables t_2 and t_3 with coefficients in the polynomial ring of a_2 and a_3 , are uniquely determined by the initial values given in [Proposition 2.22](#) and the following four PDEs:

$$(2.24) \quad B_{22}B - B_2B_2 = -S \circ \tau \cdot S, \quad S_{22}S - S_2S_2 = -S \circ \tau \cdot B,$$

$$(2.25) \quad B_{2222}B - 4B_{222}B_2 + 3B_{22}B_{22} = -4a_2(B_{22}B - B_2B_2) - 3(B_{33}B - B_3B_3),$$

$$(2.26) \quad B_{2223}B - 3B_{223}B_2 - B_{222}B_3 + 3B_{23}B_{23} \\ = -3a_3(B_{22}B - B_2B_2) - a_2(B_{23}B - B_2B_3).$$

Here τ maps (t_2, t_3) to $(-t_2, -t_3)$. The subscript 2 means taking the partial derivative with respect to t_2 and the subscript 3 should be interpreted similarly.

Using (2.10), the power series S can be used to compute the invariant $D_{X,w-E}^3$.

Sketch of the proof The main tool is a trick due to Fintushel and Stern [26]. Suppose E and E' are the two exceptional spheres in $X \# 2\overline{\mathbb{CP}}^2$. Then the homology class $E - E'$ is represented by a (-2) -sphere. Suppose $\hat{C}_{k,l}$ (respectively, $\hat{C}'_{k,l}$) is the identity that is given by applying (C_1) from [Section 2.4](#) to w (respectively, $w + E + E'$) and $\tilde{z} := (E + E')_{(2)}^k (E + E')_{(3)}^l z$. Similarly, we can derive identities $\bar{C}_{k,l}$ and $\check{C}_{k,l}$ by applying (C_2) and (C_3) to w and $\tilde{z}$ as above. These identities can be used to prove the existence of $B_{i,j}$ and $S_{i,j}$ inductively. To be a bit more detailed, first one shows the existence of $B_{k,0}$ and $S_{k,0}$ inductively using the initial values in [Proposition 2.22](#), $\hat{C}_{k,0}$ and $\hat{C}'_{k,0}$ [12, Proposition 69]. Then another inductive argument with the aid of the identities $\bar{C}_{k,l}$ and $\check{C}_{k,l}$ and [Proposition 2.22](#) shows the existence of $B_{k,l}$ [12, Proposition 73]. Finally, $\hat{C}_{k,l}$, $\hat{C}'_{k,l}$, the initial values of [Proposition 2.22](#) and the fact that $S_{0,1}$ is nonzero imply the existence of $S_{k,l}$ in an inductive way. After proving the existence of the polynomials $B_{k,l}$ and $S_{k,l}$, the identities $\hat{C}_{k,l}$, $\hat{C}'_{k,l}$, $\bar{C}_{k,l}$ and $\check{C}_{k,l}$ can be used to write four differential equations for B and S which are given in (2.24), (2.25) and (2.26). The existence proof shows that these PDEs and the initial values are enough to uniquely determine B and S . $\square$

In general, computing the exact form of the power series B and S (equivalently, solving the PDEs in the statement of [Theorem 2.23](#)) is not straightforward. In the following corollary, we show that for 4-manifolds with simple type, the blowup formula has a simple form:

Theorem 2.27 Suppose (X, w) is a pair of a 4-manifold and a 2-cycle such that X has w -simple type. Suppose also $\hat{X}$ is the blowup of X at one point and E is the exceptional class. Then there are power series $b(t_2, t_3), s(t_2, t_3) \in \mathbb{Q}[[t_2, t_3]]$ such that

$$(2.28) \quad \hat{D}_{\hat{X}, w}(e^{E(2)+E(3)}z) = \hat{D}_{X, w}(z)b(t_2, t_3),$$

$$(2.29) \quad \hat{D}_{\hat{X}, w+E}(e^{E(2)+E(3)}z) = \hat{D}_{X, w}(z)s(t_2, t_3)$$

for $z \in \mathbb{A}(X)^{\otimes 2}$. The power series b and s are given by the formulas

$$(2.30) \quad b(t_2, t_3) = \frac{1}{3}e^{-\frac{1}{2}t_2^2+t_3^2}[\cosh(\sqrt{3}t_2) + 2\cos(\sqrt{3}t_3)],$$

$$(2.31) \quad s(t_2, t_3) = \frac{1}{3}e^{-\frac{1}{2}t_2^2+t_3^2}[\cosh(\sqrt{3}t_2) - \cos(\sqrt{3}t_3) + \sqrt{3}\sin(\sqrt{3}t_3)].$$

Formula (2.30) for the power series $b(t_2, t_3)$ was previously found by means of quantum field theory arguments in [22, Formula 6.22]. When comparing the formulas here and in [22], the reader should note that t_3 needs to be replaced with it_3 because of different conventions in the definition of the μ map.

Proof Evaluating B and S of Theorem 2.23 at $a_2 = 3$ and $a_3 = 0$ produces b and s with the required property. These power series satisfy (2.24), (2.25) and (2.26) with a_2 and a_3 in the last two equations replaced by 3 and 0. In fact, the same proof as the existence proof in Theorem 2.23 shows that b and s are uniquely determined by these equations and the initial values in Proposition 2.22 (with a_2 and a_3 replaced by 3 and 0). The power series (2.30) and (2.31) satisfy the required conditions. Therefore, they are equal to b and s . $\square$

Remark 2.32 Identity (2.10) implies that

$$(2.33) \quad \hat{D}_{\hat{X}, w-E}(e^{E(2)+E(3)}z) = \hat{D}_{X, w}(z)s(t_2, -t_3).$$

3 Floer homologies for closed 3-manifolds

3.1 Admissible pairs

A pair (Y, γ) of a connected closed 3-manifold and an embedded oriented 1-manifold is N -admissible if there is an embedded oriented surface Σ in Y such that the integer $\Sigma \cdot \gamma$ is coprime to N . Suppose Q is the $U(N)$ -bundle associated to the pair (Y, γ) . The admissibility of (Y, γ) is what is called the *nonintegral* condition for the $U(N)$ -bundle Q in [58]. In particular, we can use the construction of [58] and

associate the instanton Floer homology group $I_*^N(Y, \gamma)$ to an N -admissible pair (Y, γ) . Instanton Floer homology can be lifted to a functor from a cobordism category $\text{COB}_{\mathbb{A}}$ to a certain category of vector spaces.

An object of the category $\text{COB}_{\mathbb{A}}$ is an N -admissible pair. A morphism from a pair (Y_0, γ_0) to another pair (Y_1, γ_1) is a triple (W, w, z) , where $W: Y_0 \rightarrow Y_1$ is a cobordism, $w: \gamma_0 \rightarrow \gamma_1$ is a 2-cycle, and $z \in \mathbb{A}(W)^{\otimes(N-1)}$. The composition of two morphisms

$$(W, w, z): (Y_0, \gamma_0) \rightarrow (Y_1, \gamma_1) \quad \text{and} \quad (W', w', z'): (Y_1, \gamma_1) \rightarrow (Y_2, \gamma_2)$$

is equal to $(W' \circ W, w' \circ w, z' \cdot z)$.

Suppose VECTOR_n is the category of relatively $\mathbb{Z}/n\mathbb{Z}$ -graded vector spaces over $\mathbb{C}$. An object of this category is a vector space V with a direct sum decomposition

$$V = \bigoplus_{j \in J} V_j,$$

where $\mathbb{Z}/n\mathbb{Z}$ acts transitively and freely on J . A morphism in this category from $V = \bigoplus_{j \in J} V_j$ to $V' = \bigoplus_{j' \in J'} V'_{j'}$ is a complex linear map $f: V \rightarrow V'$ such that f maps each V_j to a summand $V'_{h(j)}$ of V' such that $h(j+k) = h(j) + k$. Let P-VECTOR_n be the category that has the same objects as VECTOR_n . A morphism in P-VECTOR_n is a vector space homomorphism as above which is well defined only up to a sign. Instanton Floer homology gives a functor $I_*^N: \text{COB}_{\mathbb{A}} \rightarrow \text{P-VECTOR}_{4N}$.

Remark 3.1 The invariant constructed in [58] is more general than the one we described here. In [58], a version of instanton Floer homology is constructed for a triple (Y, γ, K) , where K is a link in Y and γ determines a $U(N)$ -bundle on Y that satisfies a certain nonintegral condition. We need to consider only the case that K is the empty link. On the other hand, in [58], the cobordism maps $I_*^N(W, w, z)$ are defined only in the case that $z = 1$. The more general case is a straightforward generalization and is reviewed below.

For an N -admissible pair (Y, γ) , the vector space $I_*^N(Y, \gamma)$ is defined by applying Morse homological methods to the Chern–Simons functional $\text{CS}: \mathcal{B}(Y, \gamma) \rightarrow \mathbb{R}$. The admissibility condition implies that all critical points of CS are irreducible [58, Proposition 3.1]. Therefore, we can arrange for CS_{π} , a perturbation of the Chern–Simons functional with a small holonomy perturbation, such that all of its critical points are irreducible and nondegenerate [58, Proposition 3.10]. Then CS_{π} has finitely

many critical points. Suppose α and β are two critical points of CS_π and p is a path along $([0, 1] \times Y, [0, 1] \times \gamma)$ based at α and β on $\{0\} \times Y$ and $\{1\} \times Y$. We will write $\mathcal{M}_p(\alpha, \beta)$ for the moduli space of the solutions to the perturbed ASD equation on $\mathbb{R} \times Y$ associated to the path p . Here the perturbation of the ASD equation is induced by the perturbation of the Chern–Simons functional; ie the elements of $\mathcal{M}_p(\alpha, \beta)$ can be regarded as the downward gradient flow lines of CS_π . We can also assume that CS_π is chosen such that the elements of $\mathcal{M}_p(\alpha, \beta)$ for all choices of α , β and p are regular [58, Proposition 3.18]. There is an $\mathbb{R}$ -action on $\mathcal{M}_p(\alpha, \beta)$ given by translation along the $\mathbb{R}$ -factor. The quotient space is denoted by $\check{\mathcal{M}}_p(\alpha, \beta)$. The dimension formula of the previous section implies that

$$\dim(\check{\mathcal{M}}_p(\alpha, \beta)) = \deg(\alpha) - \deg(\beta) - 1 \pmod{4N}.$$

Instanton Floer homology of the pair (Y, γ) is given by the homology of a chain complex $(\mathfrak{C}_*^\pi, d)$ associated to the functional CS_π . The vector space $\mathfrak{C}_*^\pi$ is freely generated by the critical points of CS_π . The differential of a generator α of $\mathfrak{C}_*^\pi$ is also defined as

$$(3.2) \quad d(\alpha) := \sum_{p: \alpha \rightarrow \beta} \# \check{\mathcal{M}}_p(\alpha, \beta) \beta,$$

where the sum is over all paths p such that $\check{\mathcal{M}}_p(\alpha, \beta)$ is 0-dimensional. These 0-dimensional moduli spaces are compact and we orient them as in [58, Section 3.6]. Then $\# \check{\mathcal{M}}_p(\alpha, \beta)$ denotes the signed count of the points in $\check{\mathcal{M}}_p(\alpha, \beta)$. We use the Floer grading $\deg$ to define a relative $\mathbb{Z}/4N\mathbb{Z}$ -grading on $\mathfrak{C}_*^\pi$. Then the differential d decreases this grading by 1. The relatively $\mathbb{Z}/4N\mathbb{Z}$ -graded vector space $\mathbf{I}_*^N(Y, \gamma)$ is defined to be the homology of the chain complex $(\mathfrak{C}_*^\pi, d)$. This homology group is independent of the choice of the Riemannian metric on Y and the perturbation of the Chern–Simons functional.

Suppose the Chern–Simons functional associated to the pair (Y, γ) is *Morse–Bott*. That is to say, the set of critical points of CS is a smooth manifold, and the Hessian of CS is invertible only in the normal direction to the critical manifold. Following the above definition, we need to work with perturbations of the Chern–Simons functional to define $\mathbf{I}_*^N(Y, \gamma)$. However, one can still derive some information about $\mathbf{I}_*^N(Y, \gamma)$ using the unperturbed Chern–Simons functional:

Proposition 3.3 *If the Chern–Simons functional of the pair (Y, γ) is Morse–Bott, then $\dim(\mathbf{I}_*^N(Y, \gamma)) \leq \dim(H_*(\text{crit}(\text{CS})))$.*

Proof This claim can be verified using the standard spectral sequence that starts from the homology of the critical manifold of CS and abuts to the instanton Floer homology of (Y, γ) . $\square$

Suppose $(W, w): (Y_0, \gamma_0) \rightarrow (Y_1, \gamma_1)$ is a cobordism of N -admissible pairs. After fixing Riemannian metrics on Y_0 and Y_1 , we form the noncompact manifold W^+ by adding cylindrical ends to W . Suppose also a perturbation of the Chern–Simons functional for (Y_i, γ_i) is fixed such that we can form the chain complex $(\mathfrak{C}_*^{\pi_i}, d)$ as above. We use a perturbation of the ASD equation on W^+ which is compatible with the chosen perturbations of the Chern–Simons functionals. Given a generator α_i of $\mathfrak{C}_*^{\pi_i}$ and a path $p: \alpha_0 \rightarrow \alpha_1$ along (W, w) , let $\mathcal{M}_p(W, w; \alpha_0, \alpha_1)$ be the moduli space of the corresponding equation. We can pick a perturbation of the ASD equation on W such that this moduli space is a smooth manifold. By fixing a homology orientation of W , we can also assume that all such moduli spaces are oriented [58]. Moreover, if $\mathcal{M}_p(W, w; \alpha_0, \alpha_1)$ is 0-dimensional, then it is compact.

Define a map $\mathfrak{C}(W, w): \mathfrak{C}_*^{\pi_0} \rightarrow \mathfrak{C}_*^{\pi_1}$ by

$$(3.4) \quad \mathfrak{C}(W, w)(\alpha_0) := \sum_{p: \alpha_0 \rightarrow \alpha_1} \# \mathcal{M}_p(W, w; \alpha_0, \alpha_1) \alpha_1,$$

where the sum is over all paths such that $\mathcal{M}_p(W, w; \alpha_0, \alpha_1)$ is 0-dimensional. The term $\# \mathcal{M}_p(W, w; \alpha_0, \alpha_1)$ is equal to the signed count of the points in $\mathcal{M}_p(W, w; \alpha_0, \alpha_1)$. As a matter of fact, (3.4) defines a chain map and the induced map at the level of homology, $I_*^N(W, w, 1): I_*^N(Y_0, \gamma_0) \rightarrow I_*^N(Y_1, \gamma_1)$, determines a morphism of the category P-VECTOR_{4N} . More generally, given $z \in \mathbb{A}(W)^{\otimes(N-1)}$, we can cut down the moduli space $\mathcal{M}_p(W, w; \alpha_0, \alpha_1)$ using z as in Section 2.1, and construct a smooth submanifold $\mathcal{M}_p(W, w; \alpha_0, \alpha_1, z)$ such that

$$(3.5) \quad \dim(\mathcal{M}_p(W, w; \alpha_0, \alpha_1, z)) = \dim(\mathcal{M}_p(W, w; \alpha_0, \alpha_1)) - \deg(z).$$

To be more precise, $\mathcal{M}_p(W, w; \alpha_0, \alpha_1, z)$ is a linear combination of the spaces whose dimensions are given by (3.5). Replacing $\mathcal{M}_p(W, w; \alpha_0, \alpha_1)$ in the right-hand side of (3.4) with $\mathcal{M}_p(W, w; \alpha_0, \alpha_1, z)$ determines a new chain map and the associated map at the level of homology is $I_*^N(W, w, z): I_*^N(Y_0, \gamma_0) \rightarrow I_*^N(Y_1, \gamma_1)$. This map is well defined; namely, it does not depend on the metric on W , the perturbation of the ASD equation and the choices involved in the construction of $\mathcal{M}_p(W, w; \alpha_0, \alpha_1, z)$. Because we fix the central part of connections associated to (W, w) , the cobordism maps only depend on the induced $\text{PU}(N)$ -bundles. In particular, if w is replaced with $w + nw'$, for

a closed 2-cycle w' , then the induced $\mathrm{PU}(N)$ -bundles are the same and they determine the same cobordism maps. This property is the analogue of (2.9) for closed 4-manifolds.

Remark 3.6 A priori it might seem that the cycles in 3-manifolds and 4-dimensional cobordisms only keep track of the first Chern classes of $\mathrm{U}(N)$ -bundles. However, they have strictly more information than these cohomology classes. For example, an element of $H^2(Y, \mathbb{Z})$, for a 3-manifold Y , determines a $\mathrm{U}(N)$ -bundle up to an isomorphism of $\mathrm{U}(N)$ -bundles. But an embedded 1-manifold γ in Y determines a $\mathrm{U}(N)$ -bundle up to a canonical isomorphism. (See [57, Section 4] for more details.) As a manifestation of this issue, suppose cohomology classes α and α' are given on cobordisms $W: Y_0 \rightarrow Y_1$ and $W': Y_1 \rightarrow Y_2$ such that the restrictions of these classes on Y_1 agree with each other. Then there might be an ambiguity in gluing these cohomology classes to construct an element of $H^2(W' \circ W; \mathbb{Z})$. On the other hand, there is no such ambiguity if α and α' are represented by embedded surfaces $w \subset W$ and $w' \subset W'$ such that $w|_{Y_1} = w'|_{Y_1}$.

Remark 3.7 There is not much difficulty in extending the definition of instanton Floer homology to the case of disconnected 3-manifolds. Suppose Y is a disconnected 3-manifold and $\gamma \subset Y$ is a 1-cycle. Then we say (Y, γ) is N -admissible if for each connected component Y_0 of Y , the pair $(Y_0, \gamma \cap Y_0)$ is N -admissible. Then we can repeat the definition analogous to the connected case and construct instanton Floer homology for the pair (Y, γ) . This instanton Floer homology can be computed in terms of the invariants for the connected components:

$$I_*^N(Y, \gamma) := I_*^N(Y_1, \gamma_1) \otimes \cdots \otimes I_*^N(Y_n, \gamma_n).$$

Here the Y_i are connected components of Y and $\gamma_i = \gamma \cap Y_i$. It is also possible to define cobordism maps for a cobordism of pairs (W, w) between two (not necessarily connected) N -admissible pairs and $z \in \mathbb{A}(W)^{\otimes(N-1)}$. However, these maps are not as well behaved with respect to composition as in the previous case. Consider two triples

$$(W, w, z): (Y_0, \gamma_0) \rightarrow (Y_1, \gamma_1) \quad \text{and} \quad (W', w', z'): (Y_1, \gamma_1) \rightarrow (Y_2, \gamma_2).$$

If Y_1 is not connected, what we can say about the cobordism maps is that

$$I_*^N(W' \circ W, w' \circ w, z \cdot z') = c \cdot I_*^N(W', w', z') \circ I_*^N(W, w, z)$$

for some nonzero constant c . The simplest way to fix this issue about functoriality is to work with a variation of the category $\mathrm{P-VECTOR}_{4N}$ where the morphisms are well defined only up to a nonzero scalar. We follow this approach when it is necessary to work with disconnected 3-manifolds.

Remark 3.8 A slightly unsatisfying point about I_*^N is the sign ambiguity in the definition of the cobordism maps. This issue can be avoided in a straightforward way. In the case that N is an even number, we need to change the definition of the category $\text{COB}_{\mathbb{A}}$ slightly. Let $\widetilde{\text{COB}}_{\mathbb{A}}$ have the same objects as $\text{COB}_{\mathbb{A}}$. But a morphism of this new category is a quadruple (W, w, z, o_W) where W, w and z are as before and o_W is a homology orientation for W . Then I_*^N can be lifted to a functor $\widetilde{I}_*^N: \widetilde{\text{COB}}_{\mathbb{A}} \rightarrow \text{VECTOR}_{4N}$. The main point is that initially there is an ambiguity in the orientation of the moduli spaces $\mathcal{M}_p(W, w; \alpha_0, \alpha_1)$ that appear in the definition of the cobordism maps, and a homology orientation of W fixes this ambiguity. In the case that N is odd, there is not such ambiguity and one can readily lift I_*^N to $\widetilde{I}_*^N: \text{COB}_{\mathbb{A}} \rightarrow \text{VECTOR}_{4N}$.

The definition of cobordism maps can be extended to the case that one of the ends is the empty pair. Suppose X is a 4-manifold with boundary Y and w is a properly embedded surface in X such that $\gamma := \partial w = w \cap Y$. Assume that (Y, γ) is an N -admissible pair. Given any element $z \in \mathbb{A}(X)^{\otimes(N-1)}$, we can form an element $D_{X,w}^N(z)$ of $I_*^N(Y, \gamma)$. This construction is the extension of the $U(N)$ -polynomial invariants for closed 4-manifolds in the previous section. Alternatively, (X, w, z) can be regarded as a cobordism from the empty pair to the N -admissible pair (Y, γ) . Although the empty pair is not N -admissible, (3.4) can be used to define $D_{X,w}^N(z)$.

Remark 3.9 Given $z_i \in \mathbb{A}(X_i)^{\otimes(N-1)}$, we can consider the relative elements

$$D_{X_i,w_i}^N(z_i) \in I_*^N(Y, \gamma).$$

Each of these relative elements lies in a graded summand of $I_*^N(Y, \gamma)$. Therefore, the difference $\deg(D_{X_2,w_2}^N(z_2)) - \deg(D_{X_1,w_1}^N(z_1))$ of the relative $\mathbb{Z}/4N\mathbb{Z}$ -gradings is a well-defined number in $\mathbb{Z}/4N\mathbb{Z}$ and is equal to

$$2(N+1)(w_2^2-w_1^2)-(N^2-1)\left(\frac{\chi(X_2)+\sigma(X_2)}{2}-\frac{\chi(X_1)+\sigma(X_1)}{2}\right)-(\deg(z_2)-\deg(z_1)).$$

Note that the term w_i^2 is not well defined and depends on a framing of the 1-cycle γ . However, the difference $w_2^2-w_1^2$ is independent of the framing and the above expression is well defined. A similar formula can be written for the difference between the gradings of two cobordisms with the same ends.

Suppose (X, w) is a cobordism from an N -admissible pair (Y, γ) to the empty pair; namely, X is a 4-manifold whose boundary is identified with $\overline{Y}$, the 3-manifold Y with the reverse orientation. The boundary of the embedded surface w is also identified

with $\bar{\gamma}$. Suppose also $z \in \mathbb{A}(X)^{\otimes(N-1)}$. Similarly, we can construct a functional $D_N^{X,w}(z): I_*^N(Y, \gamma) \rightarrow \mathbb{C}$.

Similar to cobordism maps, $D_{X,w}^N(z)$ and $D_N^{X,w}(z)$ satisfy some functorial properties. For example, if (X, w, z) and (W, w', z') are chosen such that

$$\partial(X, w) = (Y_0, \gamma_0) \quad \text{and} \quad (W, w', z'): (Y_0, \gamma_0) \rightarrow (Y_1, \gamma_1),$$

then

$$D_{W \circ X, w' \circ w}^N(z \cdot z') = I_*^N(W', w', z') \circ D_{X,w}^N(z).$$

A similar property holds for $D_N^{X,w}(z)$. There is also an important relation among these invariants and invariants of closed manifolds from the previous section:

Proposition 3.10 *Suppose Y is a closed and connected 3-manifold, (Y, γ) is an N -admissible pair, and X_1 and X_2 are two smooth 4-manifolds with $\partial X_1 = Y$ and $\partial X_2 = \bar{Y}$. Suppose also that oriented properly embedded surfaces $w_i \subset X_i$ are given such that $\partial w_1 = \gamma$ and $\partial w_2 = \bar{\gamma}$. If $b^+(X_2 \circ X_1) \geq 2$, then for $z_i \in \mathbb{A}(X_i)^{\otimes(N-1)}$,*

$$(3.11) \quad D_{X_2 \circ X_1, w_2 \circ w_1}^N(z_1 \cdot z_2) = D_{X_2, w_2}^{X_2, w_2}(z_2) \circ D_{X_1, w_1}^N(z_1).$$

If $b^+(X_2 \circ X_1) = 1$, then a similar formula holds where the left-hand side of (3.11) is interpreted as the invariant of the chamber associated to metrics with a long neck along Y .

There is another way that the invariants of a closed 4-manifold can be related to the cobordism maps:

Proposition 3.12 *Let (Y, γ) be an N -admissible pair and $(W, w): (Y, \gamma) \rightarrow (Y, \gamma)$ a cobordism of pairs. Let W and Y be connected. Let $\tilde{W}$ be the closed 4-manifold given by gluing the incoming end of W to its outgoing end. Suppose also that $\tilde{w} \subset \tilde{W}$ is defined similarly. If $b^+(\tilde{W}) \geq 2$, then for $z \in \mathbb{A}(W)^{\otimes(N-1)}$,*

$$(3.13) \quad D_{\tilde{W}, \tilde{w}}^N(z) = N \cdot \text{str}(I_*^N(W, w, z)),$$

where str denotes the supertrace of $I_*^N(W, w, z)$.⁶ If $b^+(\tilde{W}) = 1$, then a similar formula holds where the left-hand side of (3.13) is interpreted as the invariant of the chamber associated to metrics with a long neck along Y .

⁶Since the $\mathbb{Z}/4N\mathbb{Z}$ -grading on instanton Floer homology is defined only relatively, there is a sign ambiguity on the right-hand side, even after fixing a homology orientation for W . There is also a sign ambiguity on the left-hand side because we did not fix any homology orientation for $\tilde{W}$. Therefore, (3.13) should be interpreted as an equality up to sign. Although we do not need to do so here, both sign ambiguities can be fixed after fixing a homology orientation for W .

3.2 Floer homology of $S^1 \times \Sigma$

Suppose Σ is the connected oriented Riemann surface of genus g . The 3-manifold $Y_g := \Sigma \times S^1$ and the embedded 1-manifold $\gamma_{g,d} := \{x_1, \dots, x_d\} \times S^1$ determine an N -admissible pair if $(d, N) = 1$. The $U(N)$ -bundle associated to the pair $(Y_g, \gamma_{g,d})$ is the pullback of a $U(N)$ -bundle Q_d of degree d on Σ . Let also L_d denote the determinant of Q_d . Recall that the space $\mathcal{A}(Y_g, \gamma_{g,d})$ is constructed using an auxiliary connection on $L_{\gamma_{g,d}}$. We assume that this connection is the pullback of a connection B_0 on L_d . Similar to the 3-dimensional case, $\mathcal{A}(\Sigma, Q_d)$ is defined to be the space of $U(N)$ -connections on Q_d whose determinants are equal to B_0 . The space $\mathcal{G}_d$ is also defined to be the group of determinant 1 automorphisms of Q_d .

Let $\mathbb{V}_{g,d}^N$ be the vector space $I_*^N(Y_g, \gamma_{g,d})$. The critical points of the Chern–Simons functional for the pair $(Y_g, \gamma_{g,d})$ can be identified with N copies of the space

$$\mathcal{N}_{N,d} := \{A \in \mathcal{A}(\Sigma, Q_d) \mid F_0(A) = 0\} / \mathcal{G}_d.$$

In fact, we can pull back any element of $\mathcal{N}_{N,d}$ to $\Sigma \times [0, 1]$ and then identify the connections on $\Sigma \times \{0\}$ and $\Sigma \times \{1\}$ using an element of $\mathcal{G}_d$ which is induced by a central element of $SU(N)$. For any connection A representing an element of $\mathcal{N}_{N,d}$,

$$(3.14) \quad h^0(\Sigma, \text{ad}_A) = 0, \quad h^1(\Sigma, \text{ad}_A) = (N^2 - 1)(2g - 2) \quad \text{and} \quad h^2(\Sigma, \text{ad}_A) = 0,$$

where $h^i(\Sigma, \text{ad}_A)$ again denotes the dimension of the cohomology group $H^i(\Sigma, \text{ad}_A)$. This implies the space $\mathcal{N}_{N,d}$ is a smooth manifold of dimension $(N^2 - 1)(2g - 2)$ [1]. If α is a critical point of the Chern–Simons functional of the pair $(Y_g, \gamma_{g,d})$ associated to the connection A and a central element of $\mathcal{G}_d$, then the Künneth formula and the identities in (3.14) imply that $H^1(Y_g, \text{ad}_\alpha)$ is isomorphic to $H^1(\Sigma, \text{ad}_A)$. Consequently, the Chern–Simons functional of $(Y_g, \gamma_{g,d})$ is Morse–Bott, and hence $\dim(\mathbb{V}_{g,d}^N) \leq N \dim(H^*(\mathcal{N}_{N,d}))$.

The space $\mathcal{N}_{N,d}$ has been extensively studied in the literature. This space is a Kähler manifold and can be identified with the moduli space of *stable bundles* of rank N and degree d on a Riemann surface of genus g with a fixed determinant. The Poincaré polynomial of this manifold can be computed inductively [42; 14; 1]. Furthermore, a set of generators for the cohomology ring of this space is given [1]. We review a slightly reformulated description of these generators which are more suitable for our purposes here.

Consider the 4-manifold $X_g := \Sigma \times S^2$ and the surface $\chi_g := \{x_1, \dots, x_d\} \times S^2$. The pullback of the elements of $\mathcal{N}_{N,d}$ to X_g are ASD connections associated to

the pair (X_g, χ_g) with $\kappa = 0$. In particular, $\mathcal{N}_{N,d}$ can be regarded as a subset of $\mathcal{B}_0(X_g, \chi_g)$. Consider the subalgebra $\mathbb{A}_g^N := \mathbb{A}(\Sigma)^{\otimes(N-1)}$ of $\mathbb{A}(X_g)^{\otimes(N-1)}$. The μ -map in (2.4) determines a graded algebra homomorphism $\Psi: \mathbb{A}_g^N \rightarrow H^*(\mathcal{N}_{N,d})$. Note that this map is equivariant with respect to $\text{Diff}(\Sigma)$, the group of diffeomorphisms of Σ .

Proposition 3.15 *The map Ψ is surjective. In particular, the cohomology ring of $\mathcal{N}_{N,d}$ is generated by the elements*

$$(3.16) \quad p_r := \Psi(a_r), \quad q_r^j := \Psi(l_{(r)}^j) \quad \text{and} \quad s_r := \Psi(\Sigma_{(r)}),$$

where $2 \leq r \leq N$ and $\{l^j\}_{1 \leq j \leq 2g}$ forms a set of generators for $H_1(\Sigma, \mathbb{Z})$.

The action of $\text{Diff}(\Sigma)$ on $H_1(\Sigma)$ factors through the action of the symplectic group $\text{Sp}(2g)$. Therefore, this proposition implies that the same holds for $H^*(\mathcal{N}_{N,d})$.

Proof In [1], a universal $\text{U}(N)$ -bundle $\mathbb{F}$ is constructed over the product manifold $\mathcal{N}_{N,d} \times \Sigma$ and it is shown that the cohomology ring of $\mathcal{N}_{N,d}$ is generated by the classes

$$(3.17) \quad \tilde{p}_r := c_r(\mathbb{F})/[a], \quad \tilde{q}_r^j := c_r(\mathbb{F})/[l^j] \quad \text{and} \quad \tilde{s}_r := c_r(\mathbb{F})/[\Sigma].$$

The map Ψ is defined similarly using the universal $\text{PU}(N)$ -bundle $\mathbb{P}$ over the space $\mathcal{B}_\kappa(X_g, \chi_g) \times X_g$. The restriction of $\mathbb{P}$ to $\mathcal{N}_{N,d} \times \Sigma \subset \mathcal{B}_\kappa(X_g, \chi_g) \times X_g$ is isomorphic to the $\text{PU}(N)$ -bundle associated to $\mathbb{F}$. Thus the cohomology classes in (3.17) can be identified with the corresponding ones in (3.16) and this verifies the claim. $\square$

The vector space $\mathbb{V}_{g,d}^N$ admits a ring structure which is the analogue of the cup product on $H^*(\mathcal{N}_{N,d})$. Suppose P is the pair of pants cobordism from two copies of S^1 to one copy of S^1 . Then the triple $(\Sigma \times P, \{x_1, \dots, x_d\} \times P, 1)$ defines a map $m: \mathbb{V}_{g,d}^N \otimes \mathbb{V}_{g,d}^N \rightarrow \mathbb{V}_{g,d}^N$. To be more precise, we need to fix a homology orientation in the case that N is even. Suppose $\Delta_g := \Sigma \times D^2$ and $\delta_{g,d} := \{x_1, \dots, x_d\} \times D^2$, where D^2 is the 2-dimensional disc. We fix an arbitrary homology orientation on Δ_g and let $e := D_{\Delta_g, \delta_{g,d}}^N(1) \in \mathbb{V}_{g,d}^N$. We shall see in the proof of Proposition 3.23 that e is nonzero. We choose a homology orientation on $\Sigma \times P$ such that $m(e, e) = e$. Then functoriality of instanton Floer homology can be used to show that m defines a ring structure on $\mathbb{V}_{g,d}^N$ with unit e . We turn the relative $\mathbb{Z}/4N\mathbb{Z}$ -grading on $\mathbb{V}_{g,d}^N$ into an absolute grading by requiring that the unit element have degree 0. With this convention, the multiplication map is $\mathbb{Z}/4N\mathbb{Z}$ -graded; namely, the product of two elements of degree i_1 and i_2 has degree $i_1 + i_2$.

Suppose B is a cylinder, regarded as a cobordism with two circles as the incoming end and empty outgoing end. Then the pair

$$(3.18) \quad \Omega_g := \Sigma \times B \quad \text{and} \quad \omega_g := \{x_1, \dots, x_d\} \times B$$

determines a pairing $\langle \cdot, \cdot \rangle: \mathbb{V}_{g,d}^N \otimes \mathbb{V}_{g,d}^N \rightarrow \mathbb{C}$ which is defined in the following way after we choose an arbitrary homology orientation on Ω_g :

$$\langle \alpha, \alpha' \rangle = D_N^{\Omega_g, \omega_g}(\alpha \otimes \alpha') \quad \text{for } \alpha, \alpha' \in \mathbb{V}_{g,d}^N$$

Proposition 3.19 *The space $\mathbb{V}_{g,d}^N$ as a complex vector space with an action of $\text{Diff}(\Sigma)$ is isomorphic to $H^*(\mathcal{N}_{N,d}(\Sigma))[u]/(u^N - 1)$. In particular, the action of $\text{Diff}(\Sigma)$ on $\mathbb{V}_{g,d}^N$ factors through an action of $\text{Sp}(2g)$*

When $N = 2$, one can show that $\mathbb{V}_{g,d}^N$ is isomorphic to $H^*(\mathcal{N}_{N,d}(\Sigma))[u]/(u^N - 1)$ using the results of [21]. Muñoz [71] gave an alternative proof of this proposition for $N = 2$, and the proof of the general case is based on his approach.

Proof We can define a $\text{Diff}(\Sigma)$ –equivariant algebra homomorphism

$$\Phi: \mathbb{A}_g^N[u]/(u^N - 1) \rightarrow \mathbb{V}_{g,d}^N \quad \text{by} \quad \Phi(u^i z) := D_{\Delta_g, \delta_{g,d} + i\Sigma}^N(z),$$

where on the right-hand side of the defining equation z is regarded as an element of $\mathbb{A}(\Delta_g)^{\otimes(N-1)}$. Let $S: H^*(\mathcal{N}_{N,d}(\Sigma)) \rightarrow \mathbb{A}_g^N$ be a graded linear $\text{Sp}(2g)$ –equivariant right inverse for the map Ψ . Extend Ψ to an algebra homomorphism from $\mathbb{A}_g^N[u]/(u^N - 1)$ to $H^*(\mathcal{N}_{N,d}(\Sigma))[u]/(u^N - 1)$ by requiring that $\Psi(u) = u$. Similarly, we can extend S to $H^*(\mathcal{N}_{N,d}(\Sigma))[u]/(u^N - 1)$. We claim that the $\text{Diff}(\Sigma)$ –equivariant map $\Phi \circ S: H^*(\mathcal{N}_{N,d}(\Sigma))[u]/(u^N - 1) \rightarrow \mathbb{V}_{g,d}^N$ is injective. If the claim does not hold, then there is

$$p = \sum_{m=1}^M z_m u^{i_m} \in \mathbb{A}_g^N[u]/(u^N - 1) \quad \text{with } z_m \in \mathbb{A}_g^N \text{ and } 0 \leq i_m < N$$

such that $\Psi(p) \neq 0$ and $\Phi(p) = 0$. We assume that each z_m is nonzero and lies in one of the graded summands of $\mathbb{A}_g^N$. Furthermore, if $m < n$, then $\deg(z_m) \geq \deg(z_n)$ and equality holds only if $i_m \neq i_n$. Let $z' \in \mathbb{A}_g^N$ be such that

$$(3.20) \quad \deg(z') + \deg(z_1) = (N^2 - 1)(2g - 2)$$

and the cup product of $\Psi(z')$ and $\Psi(z_1)$ is nonzero. By [Proposition 3.10](#), the pairing $\langle \Phi(p), \Phi(u^{N-i_1} z') \rangle$ is equal to

$$(3.21) \quad \sum_{m=1}^M D_{\Sigma \times S^2, w_{g,d} + (N+i_m-i_1)\Sigma}(z' z_m),$$

where the polynomial invariants are computed in the chamber such that the fiber Σ is small. The expression for the dimension in [\(2.3\)](#) shows that the dimension of each component of the moduli space associated to the pair $(\Sigma \times S^2, w_{g,d} + k\Sigma)$ is at least $(N^2 - 1)(2g - 2)$. Moreover, if k is not divisible by N , this dimension is strictly greater than $(N^2 - 1)(2g - 2)$. In the case that $k = 0$ (or divisible by N), the moduli space of dimension $(N^2 - 1)(2g - 2)$ is given by the pullback of the elements of $\mathcal{N}_{N,d}$ to X_g . Therefore, the only nonzero term in [\(3.21\)](#) is $D_{\Sigma \times S^2, w_{g,d}}(z' z_1)$, which is given by evaluating the cohomology class $\mu(z' z_1)$ on the pullback of the elements of $\mathcal{N}_{N,d}$ to X_g . Note that the moduli space is compact in this case and we do not need to use the geometric representatives to evaluate this invariant. Therefore, $D_{\Sigma \times S^2, w_{g,d}}(z' z_1)$ is equal to $\Psi(z') \cup \Psi(z)[\mathcal{N}_{N,d}]$, which is nonzero by assumption. This contradicts the assumption that $\Phi(p) = 0$. Therefore, the map $\Phi \circ S$ is injective. We already know that the dimension of $\mathbb{V}_{g,d}^N$ is not greater than $N \dim(H^*(\mathcal{N}_{N,d}))$. Therefore, $\Phi \circ S$ is a bijection. In particular, the action of $\text{Diff}(\Sigma)$ on $\mathbb{V}_{g,d}^N$ factors through an action of $\text{Sp}(2g)$. $\square$

Corollary 3.22 *The ring $\mathbb{V}_{g,d}^N$ is generated by the elements*

$$\begin{aligned} \epsilon &= D_{\Delta_g, \delta_{g,d} + \Sigma}(1), & \mathfrak{s}_r &= D_{\Delta_g, \delta_{g,d}}(a_r), \\ o_r^j &= D_{\Delta_g, \delta_{g,d}}(l_{(r)}^j), & \rho_r &= D_{\Delta_g, \delta_{g,d}}(\Sigma(r)), \end{aligned}$$

where $2 \leq r \leq N$ and $1 \leq j \leq 2g$. Furthermore, the $\mathbb{Z}/4N\mathbb{Z}$ -gradings of these elements are given by

$$\deg(\epsilon) = 4d, \quad \deg(\mathfrak{s}_r) = -2r, \quad \deg(o_r^j) = -2r + 1, \quad \deg(\rho_r) = -2r + 2.$$

Proof The first part is an immediate consequence of [Proposition 3.19](#). The second part can be also verified easily using [Remark 3.9](#). $\square$

Fix $S: H^*(\mathcal{N}_{N,d}) \rightarrow \mathbb{A}_g^N$ as in the proof of [Proposition 3.19](#). Then we can use the isomorphism $\Phi \circ S$ to pull back the ring structure, the pairing and the $\mathbb{Z}/4N\mathbb{Z}$ -grading of $\mathbb{V}_{g,d}^N$ into $H^*(\mathcal{N}_{N,d})[u]/(u^N - 1)$. Suppose the new multiplication map and the pairing on $H^*(\mathcal{N}_{N,d})[u]/(u^N - 1)$ are also denoted by m and $\langle, \rangle$. We also fix the cohomological $\mathbb{Z}$ -grading on $H^*(\mathcal{N}_{N,d})[u]/(u^N - 1)$, where we set $\deg(u) = 0$. The

cohomological and the $\mathbb{Z}/4N\mathbb{Z}$ -gradings differ by a sign after collapsing into $\mathbb{Z}/4\mathbb{Z}$ -gradings. In the proof of Proposition 3.19, we show that for any element p of degree i in $H^*(\mathcal{N}_{N,d})[u]/(u^N - 1)$, there is an element q of degree $(N^2 - 1)(2g - 2) - i$ with $\langle p, q \rangle \neq 0$. In particular, $\langle \cdot, \cdot \rangle$ defines a nondegenerate pairing.

Suppose p_1 and p_2 are two elements of degrees i_1 and i_2 in $H^*(\mathcal{N}_{N,d})[u]/(u^N - 1)$. Then the product $m(p_1, p_2)$ consists of terms in various gradings. However, there are some constraints on the degrees of these terms. Firstly, they all have the same $\mathbb{Z}/4\mathbb{Z}$ -grading because the multiplication map is graded with respect to the $\mathbb{Z}/4N\mathbb{Z}$ -grading on $\mathbb{V}_{g,d}^N$. Moreover, an argument similar to that of Proposition 3.19 shows that the pairing of $m(p_1, p_2) - p_1 \cup p_2$ and any element of $H^*(\mathcal{N}_{N,d})[u]/(u^N - 1)$ with cohomological degree less than or equal to $(N^2 - 1)(2g - 2) - i_1 - i_2$ is zero. Therefore, the degree of the terms in $m(p_1, p_2)$ are at most $i_1 + i_2$ and the term with the maximal degree is $p_1 \cup p_2$. Therefore, the product m is a deformation of the cup product. We summarize these properties of $\mathbb{V}_{g,d}^N$ in the following proposition:

Proposition 3.23 *The pairing $\langle \cdot, \cdot \rangle$ on $\mathbb{V}_{g,d}^N$ is nondegenerate and the product*

$$m: \mathbb{V}_{g,d}^N \times \mathbb{V}_{g,d}^N \rightarrow \mathbb{V}_{g,d}^N$$

is a deformation of the cup product preserving the $\mathbb{Z}/4\mathbb{Z}$ -grading.

Multiplication with the elements of $\mathbb{V}_{g,d}^N$, constructed in Corollary 3.22, defines a series of operators on $\mathbb{V}_{g,d}^N$. We will use the same notation to denote these operators. The operator ϵ can be alternatively described as the cobordism map associated to the triple $([0, 1] \times Y_g, [0, 1] \times \gamma_{g,d} \cup \Sigma, 1)$. Similarly the remaining operators are cobordism maps associated to triples $([0, 1] \times Y_g, [0, 1] \times \gamma_{g,d}, z)$ for appropriate choices of z . The operators ϵ , $\mathfrak{N}_r$ and ρ_r commute with each other and o_r^j . However, o_r^j and $o_{r'}^{j'}$ anticommute with each other.

In the special case that $g = 1$, the moduli space $\mathcal{N}_{N,d}(\Sigma)$ consists of only one point. Therefore, $\mathbb{V}_{1,d}^N$ has N generators with exactly one generator α_i in degree $4i$ with respect to $\mathbb{Z}/4N\mathbb{Z}$ -grading. In fact, the (nonperturbed) Chern–Simons functional associated to the N -admissible pair $(Y_1, \gamma_{1,d})$ has irreducible and nondegenerate critical points (see [52]). The operator ϵ maps α_i to α_{i+d} . The following proposition characterizes the action of some of the point classes in the case that $d = 1$:

Proposition 3.24 *The operators $\mathfrak{N}_i: \mathbb{V}_{1,1}^N \rightarrow \mathbb{V}_{1,1}^N$ satisfy the identities*

(3.25)
$$\mathfrak{N}_2 = N\epsilon^{-1} \quad \text{and} \quad \mathfrak{N}_{2i-1} = 0.$$

Proof The second identity can be verified easily, because $\deg(\mathfrak{S}_{2i-1})$ is not divisible by 4. The first claim is proved in [85]. Since $\mathfrak{S}_2$ and ϵ commute with each other and $\deg(\mathfrak{S}_2) = -4$, we can conclude that $\mathfrak{S}_2 = c\epsilon^{-1}$. Therefore, we just need to show that $\mathrm{tr}(\mathfrak{S}_2 \circ \epsilon) = N^2$. Using Proposition 3.12, this can be reduced to showing that

$$D_{T^4, T^2 \times \{\mathrm{pt}\} \cup \{\mathrm{pt}\} \times T^2}^N(\mathfrak{S}_2) = N^3,$$

which is established in [85] using properties of stable bundles on abelian varieties. $\square$

In a work in progress, we will give another proof of this proposition which is independent of the results of [85].

3.3 Fukaya–Floer homology

Suppose X_1 and X_2 are 4-manifolds with $\partial X_1 = Y$ and $\partial X_2 = \bar{Y}$, and X is given by gluing these manifolds along their boundaries. Suppose also a 1-cycle $\gamma \subset Y$ and 2-cycles $w_i \subset X_i$ are chosen such that $\partial w_1 = \gamma$ and $\partial w_2 = \bar{\gamma}$. The cycles w_1 and w_2 can be glued to each other to form a 2-cycle $w \subset X$. We also assume that (Y, γ) is an N -admissible pair. Then Floer homology for this N -admissible pair provides a useful device to relate $U(N)$ -polynomial invariants of the form

$$(3.26) \quad D_{X,w}^N(z_1 \cdot z_2) \quad \text{with } z_i \in \mathbb{A}(X_i)^{\otimes(N-1)}$$

to the relative invariants associated to (X_1, w_1, z_1) and (X_2, w_2, z_2) (see Proposition 3.10). As we shall see in the next section, this decomposition theorem for polynomial invariants is a useful tool for computational purposes. However, not all polynomial invariants of (X, w) have the form in (3.26). There are homology classes $\Gamma \in H_2(X)$ such that Γ is not the sum of elements in $H_2(X_1)$ and $H_2(X_2)$. Then, for example, $D_{X,w}^N(\Gamma_{(2)}^i \Gamma_{(3)}^j)$ cannot be expressed in terms of the relative invariants. In this section, we introduce an extension of Floer homology which admits relative invariants for such polynomial invariants. This extension of Floer homology was already constructed in [33; 9] for $N = 2$ and is known as *Fukaya–Floer homology*.

Our extension of Floer homology is a module over a ring R_N . Let $R_{N,j}$ be the polynomial ring over the variables $t_{k,i}$ for $2 \leq k \leq N$ and $1 \leq i \leq j$, modulo the relations $t_{k,i}^2 = 0$:

$$R_{N,j} := \mathbb{C}[t_{k,i} : 2 \leq k \leq N, 1 \leq i \leq j] / (t_{k,i}^2).$$

For $j \geq l \geq 0$, there is an obvious map from $R_{N,j}$ to $R_{N,l}$ which maps $t_{k,i}$ to $t_{k,i}$ when $i \leq l$ and maps $t_{k,i}$ to 0 when $i > l$. The ring R_N is defined to be the inverse

limit of this system of rings. For example, for each $2 \leq k \leq N$, we have an element of R_N as follows:

$$s_k := \sum_{i=1}^{\infty} t_{k,i}.$$

The ring of polynomials $\mathbb{C}[[t_2, \dots, t_N]]$ can be regarded as a subring of R_N by mapping t_k to $s_k \in R_N$. Under this inclusion we have

$$\frac{t_k^l}{l!} \rightarrow \sum_{\substack{S \subset \mathbb{N} \\ |S|=l}} \prod_{i \in S} t_{k,i}.$$

The full version of Fukaya–Floer homology for $N = 2$, whose construction is sketched in [9], is expected to be a module over $\mathbb{C}[[t_2]]$. However, our construction is slightly different and we obtain a module over the ring R_N for general N . This is partly because the definition of polynomial invariants for higher rank bundles slightly differs from the classical definition of $U(2)$ –polynomial invariants. Another reason is that even for $N = 2$, the authors were not able to avoid some analytical difficulties related to the noncompactness of the moduli space of ASD connections and construct a ring over $\mathbb{C}[[t_2]]$.

Consider the N –admissible pair (Y, γ) , and let $L = (l_2, \dots, l_N)$ be an $(N-1)$ –tuple of the elements of $H_1(Y)$. Fukaya–Floer homology associates to (Y, γ, L) an $R_{N,j}$ –module $\mathbb{I}_*^{N,j}(Y, \gamma, L)$ for each nonnegative integer j and an $R_{N,j}$ –module homomorphism $f_j^k: \mathbb{I}_*^{N,j}(Y, \gamma, L) \rightarrow \mathbb{I}_*^{N,k}(Y, \gamma, L)$ for each pair (j, k) of nonnegative integers such that $j \geq k \geq 0$. If $j \geq k \geq l \geq 0$, we require $f_k^l \circ f_j^k = f_j^l$. Fukaya–Floer homology of (Y, γ, L) is the inverse limit of the inverse system $(\{\mathbb{I}_*^{N,j}(Y, \gamma, L)\}_j, \{f_j^k\}_{j \geq k})$. The $R_{N,j}$ –module $\mathbb{I}_*^{N,j}(Y, \gamma, L)$ is the homology of a chain complex $(\mathfrak{C}_*^{N,j}(Y, \gamma, L), d_{N,j})$. In fact, we can arrange for a perturbation CS_{π_j} of the Chern–Simons functional of the N –admissible pair (Y, γ) such that $\mathfrak{C}_*^{N,j}(Y, \gamma, L) = \mathfrak{C}_*^{\pi_j}(Y, \gamma) \otimes_{\mathbb{C}} R_{N,j}$.

The differential $d_{N,j}$ of the Fukaya–Floer chain complex has the form

$$(3.27) \quad d_{N,j}(\alpha) = \sum_{\substack{\bar{S}=(S_2, \dots, S_N) \\ p: \alpha \rightarrow \beta}} h_{\bar{S}}(\alpha, \beta) \left(\prod_{i \in S_k} t_{k,i} \right) \beta,$$

where $S_k \subset \{1, \dots, j\}$ and the path p is chosen such that the dimension of the moduli space $\mathcal{M}_p(\alpha, \beta)$ is equal to

$$2|S_2| + 4|S_3| + \dots + 2(N-1)|S_N| + 1.$$

The constant term of the differential is equal to the differential d of the Floer chain complex. That is to say, if we evaluate all variables $t_{k,i}$ at zero, then we recover d . The definitions of the other terms in (3.27) are discussed in Section 6.3. We extend the Floer grading to $\widetilde{\mathfrak{C}}_*^{N,j}(Y, \gamma, L)$ by requiring that $\deg(t_{k,i}) = 2(k-1)$. Then the differential $d_{N,j}$ has degree -1 .

Suppose (X_1, w_1) is a pair of a 4-manifold and a 2-cycle which fills the N -admissible pair (Y, γ) . Suppose also $z_1 \in \mathbb{A}(X_1)^{\otimes(N-1)}$ and $\Gamma^2, \dots, \Gamma^N$ are properly embedded surfaces in X_1 where $[\partial\Gamma^i] = l_i \subset H_1(Y)$. Then one can associate an element of $\mathbb{I}_*^N(Y, \gamma, L)$ to $(X_1, w_1, z_1, \Gamma^2, \dots, \Gamma^N)$, which is denoted by

$$(3.28) \quad D_{X_1, w_1}^N(z_1 \cdot e^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N}).$$

The element in (3.28) is given by a system of cycles $D_{X_1, w_1}^{N,j}(z_1 \cdot e^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N}) \in \mathfrak{C}_*^{N,j}(Y, \gamma, L)$. We have

$$D_{X_1, w_1}^{N,j}(z_1 \cdot e^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N}) = \sum_{\bar{S}=(S_2, \dots, S_N), \alpha} m_{\bar{S}}(\alpha) \left(\prod_{i \in S_k} t_{k,i} \right) \alpha$$

for appropriate choices of complex numbers $m_{\bar{S}}(\alpha)$. Evaluating all the variables $t_{k,i}$ at zero produces a cycle in $\mathfrak{C}_*^{\pi j}(Y, \gamma)$ which represents the relative invariant $D_{X_1, w_1}^N(z_1)$.

Next, let (X_2, w_2) be a cobordism from (Y, γ) to the empty pair. Suppose that $z_2 \in \mathbb{A}(X_2)^{\otimes(N-1)}$ and $\Lambda_2, \dots, \Lambda_N$ are properly embedded surfaces in X_2 where $[\partial\Lambda^j] = -l_j$. In this case, there is an R_N -linear map from $\mathbb{I}_*^N(Y, \gamma, L)$ to R_N associated to $(X_2, w_2, z_2, \Lambda^2, \dots, \Lambda^N)$, which is denoted by

$$(3.29) \quad D_{X_2, w_2}^{X_2, w_2}(z_2 \cdot e^{\Lambda_{(2)}^2 + \dots + \Lambda_{(N)}^N}).$$

The construction of the element (3.28) and that of the functional (3.29) are given in Section 6.3. We can glue the 4-manifolds (X_1, w_1) and (X_2, w_2) to form a closed pair $(X_2 \circ X_1, w_2 \circ w_1)$. The embedded surfaces Γ^j and Λ^j can also be glued to each other to form a closed embedded surface $\Gamma^j \# \Lambda^j$. Then we have

$$(3.30) \quad D_{X_2 \circ X_1, w_2 \circ w_1}^N(z_1 \cdot z_2 \cdot e^{(\Gamma^2 \# \Lambda^2)_{(2)} + \dots + (\Gamma^N \# \Lambda^N)_{(N)}}) \\ = D_{X_2, w_2}^{X_2, w_2}(z_2 \cdot e^{\Lambda_{(2)}^2 + \dots + \Lambda_{(N)}^N}) \circ D_{X_1, w_1}^N(z_1 \cdot e^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N}).$$

This claim shall be proved as Proposition 6.51 in Section 6.3. A priori, the right-hand side of the above equality is an element of R_N . Part of the claim is that the right-hand side belongs to $\mathbb{C}[[t_2, \dots, t_N]] \subset R_N$ and is equal to the given power series on the left-hand side.

Fukaya–Floer homology is also functorial with respect to cobordisms of N –admissible pairs. Suppose $(W, w): (Y_0, \gamma_0) \rightarrow (Y_1, \gamma_1)$ is such a cobordism. For $2 \leq i \leq N$, suppose also Γ^i is a properly embedded surface in W such that $\Gamma^i \cap Y_j$ represents the homology class $l_i^j \in H_1(Y_j)$. For any $z \in \mathbb{A}(W)$, there is a homomorphism

$$\mathbb{I}_*^N(W, w, ze^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N}): \mathbb{I}_*^N(Y_0, \gamma_0, L_0) \rightarrow \mathbb{I}_*^N(Y_1, \gamma_1, L_1),$$

where $L_j = (l_2^j, \dots, l_N^j)$. This construction is functorial with respect to composition of cobordisms.

Suppose $(Y_g, \gamma_{g,d})$ is the N –admissible pair from Section 3.2 and $L_g = (l_2, \dots, l_N)$ is the $(N-1)$ –tuple of the elements of $H_1(Y_g)$ where l_i is an S^1 fiber. In this article, the main example of Fukaya–Floer homology for us is $\mathbb{I}_*^N(Y_g, \gamma_{g,d}, L_g)$, which is denoted by $\mathbb{I}_{g,d}^N$. Analogously to the previous subsection, we can define a ring homomorphism

$$\widetilde{\Phi}: \mathbb{A}_g^N[u]/(u^N - 1) \rightarrow \mathbb{I}_{g,d}^N \quad \text{by} \quad \widetilde{\Phi}(u^i z) := D_{\Delta_g, \delta_{g,d} + i\Sigma}^N(ze^{D_{(2)}^2 + \dots + D_{(N)}^2}).$$

Recall that D^2 is a 2–dimensional disc, $\Delta_g = \Sigma \times D^2$ and $\delta_{g,d} = \{x_1, \dots, x_d\} \times D^2$. Similarly, we can define a multiplication and a pairing on $\mathbb{I}_{g,d}^N$ by repeating the construction of Section 3.2.

Proposition 3.31 *For any nonzero element $q \in \mathbb{I}_{g,d}^N$, there are $z \in \mathbb{A}_g^N$ and $1 \leq i \leq N$ such that*

$$\langle q, D_{\Delta_g, \delta_{g,d} + i\Sigma}^N(ze^{D_{(2)}^2 + \dots + D_{(N)}^2}) \rangle \neq 0.$$

Proof Suppose q is given by the sequence $\{q_j\}_{j \geq 0}$, where $q_j \in \mathbb{I}_*^{N,j}(Y_g, \gamma_{g,d}, L_g)$. Since q is nonzero, there is a j such that q_j is nonzero. To verify the claim, it suffices to show that there are $z \in \mathbb{A}_g^N$ and $1 \leq i \leq N$ such that

$$\langle q_j, D_{\Delta_g, \delta_{g,d} + i\Sigma}^{N,j}(ze^{D_{(2)}^2 + \dots + D_{(N)}^2}) \rangle \neq 0.$$

Suppose $(\mathfrak{C}_*^{\pi_j}(Y_g, \gamma_{g,d}), d)$ is a Floer chain complex for the pair $(Y_g, \gamma_{g,d})$ such that $\mathfrak{C}_*^{\pi_j}(Y_g, \gamma_{g,d}) \otimes R_{N,j}$ can be used to define $\mathbb{I}_*^{N,j}(Y_g, \gamma_{g,d}, L_g)$. Suppose q_j is represented by the element

$$\sum_{\bar{S} = (S_2, \dots, S_N) \in \mathcal{I}} b_{\bar{S}} \left(\prod_{i \in S_k} t_{k,i} \right) \alpha_{\bar{S}}$$

of this complex, where $\mathcal{I}$ is a set consisting of $(N-1)$ –tuples of finite subsets of $\{1, \dots, j\}$, each $b_{\bar{S}}$ is a nonzero complex number and $\alpha_{\bar{S}} \in \mathfrak{C}_*^{\pi_j}(Y_g, \gamma_{g,d})$. Suppose $\bar{S}_0 = (S_2^0, \dots, S_N^0) \in \mathcal{I}$ is a minimal element with respect to the partial order on $\mathcal{I}$

induced by inclusion. Then $d\alpha_{\bar{S}_0} = 0$, and by changing the representative if necessary, we can assume that $\alpha_{\bar{S}_0}$ is not a boundary in $\mathfrak{C}_*^{\pi_j}(Y_g, \gamma_{g,d})$. Therefore, by Corollary 3.22 and Proposition 3.23, there are $z \in \mathbb{A}_g^N$ and $1 \leq i \leq N$ such that

$$\langle \alpha_{\bar{S}_0}, D_{\Delta_g, \delta_{g,d}+i\Sigma}^N(z) \rangle \neq 0.$$

This implies that the coefficient of $\prod_{2 \leq k \leq N} \prod_{i \in S_k^0} t_{k,i}$ in the pairing

$$\langle q_j, D_{\Delta_g, \delta_{g,d}+i\Sigma}^{N,j}(ze^{D_{(2)}^2+\dots+D_{(N)}^2}) \rangle.$$

is nonzero. □

Example 3.32 In the case that $g = 1$, the Fukaya–Floer homology group of the triple $(Y_g, \gamma_{g,d}, L_g)$ is equal to $R_N^{\oplus N}$. This can be easily seen from the construction of Fukaya–Floer homology in Section 6.3. In fact, this R_N -module is freely generated by the critical points $\alpha_1, \dots, \alpha_N$ of the Chern–Simons functional of $(Y_1 = T \times S^1, \gamma_{1,d})$. Consider the operator

$$(3.33) \quad \tilde{\epsilon} := \mathbb{I}_*^N([0, 1] \times Y_1, [0, 1] \times \gamma_{1,d} + T, e^{([0,1] \times \gamma)_{(2)} + \dots + ([0,1] \times \gamma)_{(N)}}),$$

where γ is an S^1 -fiber of Y_1 . This operator is of order N and has degree $4d$. Moreover, $\alpha_i, \tilde{\epsilon}(\alpha_i), \dots, \tilde{\epsilon}^{N-1}(\alpha_i)$ form a basis for $\mathbb{I}_*^N((Y_g, \gamma_{g,d}, L_g))$ for any i . In particular, the kernel of $\tilde{\epsilon} - 1$ is equal to $R_N \cdot (1 + \tilde{\epsilon} + \dots + \tilde{\epsilon}^{N-1})(\alpha_i)$.

Next, we discuss a prototype for 4-manifolds with boundary Y_g which are of interest to us. Suppose (X_1, Σ) is a pair of a 4-manifold and an embedded surface of genus g with self-intersection 0 such that a regular neighborhood of Σ in X_1 is identified with Δ_g . Suppose (X_2, Σ) is another such pair. As the notation suggests, the embedded surfaces in X_1 and X_2 are identified with each other. Remove regular neighborhoods of Σ in X_1 and X_2 to produce 4-manifolds whose boundaries are Y_g , and then glue the resulting two 4-manifolds along their common boundaries by the orientation-reversing diffeomorphism that maps $(z, x) \in S^1 \times \Sigma$ to $(\bar{z}, x)$. This 4-manifold is denoted by $X_1 \#_{\Sigma} X_2$, and is called the *fiber sum* of X_1 and X_2 along Σ . We will also write X_i° for the complement of a neighborhood of Σ in X_i . Then X_i° can be also regarded as a subspace of $X_1 \#_{\Sigma} X_2$.

Elements of $H_2(X_1)$ and $H_2(X_2)$ can be glued to each other to construct elements of $H_2(X_1 \#_{\Sigma} X_2)$. Suppose $\iota_i: H_2(X_i) \rightarrow \mathbb{C}$ denotes the map that computes the intersection number of an element of $H_2(X_i)$ with Σ . Suppose also $\mathcal{K}$ is the subspace

of the elements $(\Gamma, \Lambda) \in H_2(X_1) \oplus H_2(X_2)$ such that $\iota_1(\Gamma) = \iota_2(\Lambda)$. Then there is a homomorphism $\#: \mathcal{K} \rightarrow H_2(X_1 \#_{\Sigma} X_2)$ with the property that

$$(3.34) \quad j_1^{\#}(\Gamma \# \Lambda) = j_1^{\circ}(\Gamma) \quad \text{and} \quad j_2^{\#}(\Gamma \# \Lambda) = j_2^{\circ}(\Lambda).$$

Here $j_i^{\circ}: H_2(X_i) \rightarrow H_2(X_i^{\circ}, \partial X_i^{\circ})$ is the composition of the map from $H_2(X_i)$ to the relative homology $H_2(X_i, \Delta_g)$ and the excision isomorphism. To abbreviate our notation, from now on, we will write Γ° and Λ° for $j_1^{\circ}(\Gamma)$ and $j_2^{\circ}(\Lambda)$. The maps $j_i^{\#}: H_2(X_1 \#_{\Sigma} X_2) \rightarrow H_2(X_i^{\circ}, \partial X_i^{\circ})$ are also defined similarly.

The homomorphism $\#$ is not uniquely defined and we proceed as follows to fix one such homomorphism. Suppose Γ and Λ are integral homology classes. Then these homology classes can be represented by oriented embedded surfaces, which we denote with the same notation. We can assume that these surfaces are transversal to Σ , and intersect Σ in the same set of points with the same signs. Then there is an obvious way to glue Γ and Λ and to produce an oriented embedded surface in $X_1 \#_{\Sigma} X_2$. The homology class $\Gamma \# \Lambda$ is defined to be the homology of the glued-up surface. We apply this construction to an integral basis of $\mathcal{K}$ and extend it linearly.

We use Poincaré duality to define $\mathcal{L} \subseteq H^2(X_1) \oplus H^2(X_2)$, the counterpart of $\mathcal{K}$, and the gluing map $\#: \mathcal{L} \rightarrow H^2(X_1 \#_{\Sigma} X_2)$. Suppose $(K, L) \in \mathcal{L}$ and $(\Gamma, \Lambda) \in \mathcal{K}$. Then we have the equality of the pairing of cohomology classes with homology classes

$$(K \# L)[\Gamma \# \Lambda] = K[\Gamma] + L[\Lambda].$$

Similarly, we can glue two cycles $w_1 \subset X_1$ and $w_2 \subset X_2$ that intersect Σ transversely in the same set of points with the same signs. The resulting 2-cycle in $X_1 \#_{\Sigma} X_2$ is denoted by $w_1 \# w_2$. We will also write w_i° for the intersection $w_i \cap X_i^{\circ}$.

Proposition 3.35 *For $1 \leq i \leq 4$, suppose X_i is a 4-manifold and T is an embedded surface of genus one in X_i . Suppose also $w_i \subset X_i$ is a 2-cycle such that $w_i \cdot T$ is coprime to N . For each $2 \leq l \leq N$, suppose also Γ_i^l is an element of $H_2(X_i)$ such that $\Gamma_i^j \cdot T = 1$. For $1 \leq i, k \leq 4$, let $D_{i,k}$ be the following element of $\mathbb{C}[[t_2, \dots, t_N]]$:*

$$\sum_{j=1}^N D_{X_i \#_T X_k, w_i \# w_k + jT}^N (e^{(\Gamma_i^2 \# \Gamma_k^2)_{(2)} + \dots + (\Gamma_i^N \# \Gamma_k^N)_{(N)}}).$$

Then

$$D_{1,2} D_{3,4} = D_{1,4} D_{3,2}.$$

Proof The following elements lie in the kernel of the operator $\tilde{\epsilon}$ in (3.33):

$$D_i := \sum_{j=1}^N D_{X_i^\circ, w_i^\circ + jT}^N (e^{(\Gamma_i^2)^\circ_{(2)} + \dots + (\Gamma_i^N)^\circ_{(N)}}).$$

Moreover, (3.30) implies that

$$(3.36) \quad D_{i,k} = \frac{1}{N} \langle D_i, D_k \rangle.$$

Because $D_i \in \ker(\tilde{\epsilon} - 1)$, the claim is a consequence of the description of $\ker(\tilde{\epsilon} - 1)$ in Example 3.32. $\square$

3.4 An $SU(3)$ –instanton Floer homology for $\Sigma(2, 3, 23)$

In Section 3.1, Floer homology is defined for an N –admissible pair (Y, γ) . Then the computation of the polynomial invariants for a closed pair (X, w) which can be decomposed along a copy of (Y, γ) can be reduced to computing relative invariants for each component of $X \setminus Y$ (Proposition 3.10). One wishes to extend the definition of Floer homology so that it can be used in studying polynomial invariants of a pair (X, w) which is decomposed along a nonadmissible pair. However, there is little known in this direction even when $N = 2$. For $N = 2$, the most satisfactory answer is provided in the case that Y is an integral homology sphere and γ is empty [27], which will be denoted by $I_*^2(Y)$. The main reason that one can define $I_*^2(Y)$ for an integral homology sphere is that the only reducible flat connection on Y is the nondegenerate trivial connection. In order to extend $I_*^N(Y)$ to higher values of N , one would face more complicated reducible connections. Due to this complication, there are some difficulties in extending the definition of Floer homology of integral homology spheres to higher values of N . In this section, we make a modest progress in this direction and define $I_*^3(Y)$ for the Brieskorn homology sphere $\Sigma(2, 3, 23)$. Meanwhile, we compute some of the gauge-theoretical invariants for flat connections on $\Sigma(2, 3, 23)$.

Suppose the positive integers a_1 , a_2 and a_3 are pairwise coprime, and $\Sigma(a_1, a_2, a_3)$ is the associated Brieskorn sphere:

$$\Sigma(a_1, a_2, a_3) := \{(z_1, z_2, z_3) \in \mathbb{C}^3 \mid z_1^{a_1} + z_2^{a_2} + z_3^{a_3} = 0, |z_1|^2 + |z_2|^2 + |z_3|^2 = 1\}.$$

This 3–manifold is an integral homology sphere. There is an S^1 –action on this 3–manifold,

$$e^{2\pi i \theta} \cdot (z_1, z_2, z_3) := (e^{2\pi i a_2 a_3 \theta} z_1, e^{2\pi i a_1 a_3 \theta} z_2, e^{2\pi i a_1 a_2 \theta} z_3).$$

This action turns $\Sigma(a_1, a_2, a_3)$ into a Seifert fiber space over S^2 with three exceptional orbits. Complex conjugation on $\mathbb{C}^3$ induces a diffeomorphism of $\Sigma(2, 3, 23)$, which will be also called complex conjugation. There is also a standard presentation of the fundamental group of this 3-manifold, given as

$$(3.37) \quad \pi_1(\Sigma(a_1, a_2, a_3)) = \langle x_1, x_2, x_3, h \mid [h, x_i] = 1, x_i^{a_i} h^{\beta_i} = 1, x_1 x_2 x_3 = 1 \rangle,$$

where β_i is given by the identity

$$\frac{\beta_1}{a_1} + \frac{\beta_2}{a_2} + \frac{\beta_3}{a_3} = \frac{1}{a}.$$

with $a = a_1 a_2 a_3$. The central element h in (3.37) is represented by a generic fiber of the Seifert fibration.

Suppose W is the space $(\Sigma(a_1, a_2, a_3) \times D^2)/S^1$, where the S^1 -action is the product of the Seifert action on $\Sigma(a_1, a_2, a_3)$ and the standard action on D^2 . Alternatively, W is the mapping cylinder of the fibration of $\Sigma(a_1, a_2, a_3)$ over S^2 . This space is an orbifold and has three singular points. Neighborhoods of these singular points are diffeomorphic to cones on the lens spaces $L(a_i, \beta_i)$. Therefore, removing neighborhoods of the orbifold points produces a cobordism W_0 from the union of three lens spaces $L(a_i, \beta_i)$ to $\Sigma(a_1, a_2, a_3)$. We will denote the union of the lens spaces by Y .

The fundamental group of W_0 is equal to $\pi_1(\Sigma(a_1, a_2, a_3))/\langle h \rangle$ and the inclusion of $\Sigma(a_1, a_2, a_3)$ in W_0 induces the quotient map at the level of fundamental groups. Moreover, the induced map from the fundamental group of $L(a_i, \beta_i)$ to the group $\pi_1(\Sigma(a_1, a_2, a_3))/\langle h \rangle$ maps the standard generator of $\pi_1(L(a_i, \beta_i))$ to x_i . The description of the fundamental group implies that the first homology of W_0 is trivial. We also have the short exact sequence

$$0 \rightarrow H^2(W_0, \partial W_0, \mathbb{Z}) \cong \mathbb{Z} \xrightarrow{\iota} H^2(W_0, \mathbb{Z}) \cong \mathbb{Z} \rightarrow H^2(\partial W_0, \mathbb{Z}) \cong \mathbb{Z}/a\mathbb{Z} \rightarrow 0,$$

where the map ι is multiplication by a . The self-intersection pairing, defined on the image of ι , maps a generator of $\text{im}(\iota)$ to $-a$. In particular, $b^+(W_0)$ is equal to 0.

The space $L := \Sigma(a_1, a_2, a_3) \times D^2$ defines an orbifold S^1 -bundle on W . In particular, the restriction of L to W_0 , denoted by L_0 , is a smooth S^1 -bundle. The first Chern class of L_0 is a generator of $H^2(W_0, \mathbb{Z})$. The restriction of this Chern class to $L(a_i, \beta_i)$ is equal to β_i times the standard generator of $H^2(L(a_i, \beta_i), \mathbb{Z})$. In particular, for any complex line bundle on ∂W_0 , there is a k such that the restriction of L_0^k to the boundary is isomorphic to the given line bundle.

The rational cohomology class induced by $c_1(L_0)$ can be lifted to $H^2(W_0, \partial W_0, \mathbb{Q})$. In particular, $c_1(L_0)^2$ is well defined and is equal to $-1/a$. For our purposes, we also fix a connection B_0 on L_0 whose restriction to a neighborhood of ∂W_0 is the pullback of a flat connection on ∂W_0 . In particular, we can assume that the restriction of this connection in a regular neighborhood of $\Sigma(a_1, a_2, a_3)$ is the trivial connection.

Suppose α is a flat $SU(N)$ -connection on $\Sigma(a_1, a_2, a_3)$ whose holonomy around the fiber is central. Then α can be extended as a flat $PU(N)$ -connection A to W_0 . The holonomy of α induces a conjugacy class r_i in $PU(N)$ corresponding to the loop x_i . The class r_i has order a_i and determines a flat $PU(N)$ -connection on $\pi_1(L(a_i, \beta_i))$ which matches the restriction of A to $L(a_i, \beta_i)$. This connection will be also denoted by r_i . As it was pointed in [24], the connection A can be used to compute some of the gauge-theoretical invariants of α :

Proposition 3.38 *Let α and A be given as above. Then $\rho_{\text{ad}(\alpha)}$ is equal to*

$$(3.39) \quad (T_\alpha + 1 - N^2) + \sum_{i=1}^3 \rho_{\text{ad}(r_i)}(L(a_i, \beta_i)),$$

where T_α is the number of trivial summands in the irreducible decomposition of the representation associated to ad_α .

Using the calculations of [2] for lens spaces, (3.39) allows us to compute the ρ -invariant of any connection as above.

Proof According to [2],

$$\rho_{\text{ad}(\alpha)} - \sum_{i=1}^3 \rho_{\text{ad}(r_i)}(L(a_i, \beta_i)) = (N^2 - 1)\sigma(W_0) - \sigma_A(W_0),$$

where $\sigma_A(W_0)$ denotes the signature of the twisted cohomology group $H^2(W_0; \text{ad}(A))$ determined by the flat $PU(N)$ -connection $\text{ad}(A)$. This twisted cohomology group can be decomposed according to the irreducible decomposition of A (or equivalently α). The argument of [24, Lemma 2.6] shows that the contribution of the nontrivial summands is equal to 0. On the other hand, each trivial summand contributes -1 to the sum, because $\sigma(W_0) = -1$. $\square$

The underlying $PU(N)$ -bundle of the connection A on W_0 can be lifted to a $U(N)$ -bundle E . There are also nonnegative integers $k_1, \dots, k_N$ such that the restriction of

the connection

(3.40)
$$B_0^{k_1} \oplus \cdots \oplus B_0^{k_N}$$

to Y , the union of the three lens spaces, is equal to the restriction of A to Y . In particular, $L_0^{k_1} \oplus \cdots \oplus L_0^{k_N}$, the underlying $U(N)$ –bundle of (3.40), has the same restriction as E on Y . Therefore, the determinant of $L_0^{k_1} \oplus \cdots \oplus L_0^{k_N}$ is equal to $\det(E) \otimes L_0^{ak}$ for an appropriate integer k . After replacing k_1 with $k_1 + ak$, the new bundle $L_0^{k_1} \oplus \cdots \oplus L_0^{k_N}$ has the same determinant as E and the connection in (3.40) and A give rise to the same connection after restriction to Y .

Since A and the connection in (3.40) agree on boundary components except on $\Sigma(a_1, a_2, a_3)$, where A gives the connection α and B_0 restricts to the trivial connection, we can conclude that

(3.41)
$$\begin{aligned} \text{CS}(\alpha) &= \mathcal{E}(B_0^{k_1} \oplus \cdots \oplus B_0^{k_N}) - \mathcal{E}(A) \\ &= -\frac{1}{2N} \sum_{i < j} (k_i - k_j)^2 c_1(L)^2 \\ &= \frac{1}{2N} \sum_{i < j} \frac{(k_i - k_j)^2}{a}. \end{aligned}$$

Therefore, this gives us a strategy to compute the Chern–Simons functional of the flat connection α .

Now we focus on $N = 3$ and the Brieskorn homology sphere $\Sigma(2, 3, 23)$. Suppose α is an irreducible connection on $\Sigma(2, 3, 23)$. The irreducibility assumption implies that the holonomy of α along the generic fiber of the Seifert fibration is central. In fact, this central element is equal to the identity [4]. It is also shown in [4] that there are 44 such irreducible representations which are nondegenerate. One can use the method of [4] to find the conjugacy classes of holonomies corresponding to the elements x_1 , x_2 and x_3 of $\pi_1(\Sigma(2, 3, 23))$. Any irreducible flat connection on $\Sigma(2, 3, 23)$ is characterized by its holonomies along x_3 . The possible conjugacy classes for x_3 are listed in Table 4. On the other hand, the holonomies along x_1 and x_2 are conjugate to the diagonal matrices $\text{Diag}(1, -1, -1)$ and $\text{Diag}(1, \zeta, \zeta^2)$, where $\zeta = e^{2\pi i/3}$. Knowledge of these conjugacy classes allows us to apply Proposition 3.38 and (3.41) to compute the ρ –invariants, the Chern–Simons functional and hence the degrees of irreducible flat connections on $\Sigma(2, 3, 23)$:

Proposition 3.42 *There are ten irreducible flat connections of degree 0, five irreducible flat connections of degree 2, nine irreducible flat connections of degree 4, five irreducible flat connections of degree 6, nine irreducible flat connections of degree 8 and six irreducible flat connections of degree 10. There are not any irreducible flat connections of odd degree. The Chern–Simons functional and the ρ -invariants of these irreducible flat connections can be found in Table 4.*

Any reducible flat connection on $\Sigma(2, 3, 23)$ is either trivial or $SU(2)$ -reducible, because this 3-manifold is an integral homology sphere. In particular, nontrivial reducible $SU(3)$ -connections on $\Sigma(2, 3, 23)$ can be regarded as irreducible $SU(2)$ -connections. The results and the methods of [24] can be utilized to study such connections. The holonomy of any nontrivial flat $SU(2)$ -connection along the fiber of the Seifert fibration is the central element $-\text{id}$. The holonomies of this connection along x_1 and x_2 are respectively conjugate to $\text{Diag}(\mathbf{i}, -\mathbf{i})$ and $\text{Diag}(e^{\pi\mathbf{i}/3}, e^{-\pi\mathbf{i}/3})$. As in the $SU(3)$ -case, an irreducible flat $SU(2)$ -connection on $\Sigma(2, 3, 23)$ is determined by the conjugacy class of its holonomy along x_3 . The eigenvalues of the holonomy term along x_3 are equal to $e^{2\pi\mathbf{i}k/23}$ and $e^{-2\pi\mathbf{i}k/23}$ for $2 \leq k \leq 9$:

Proposition 3.43 [24] *There are eight irreducible $SU(2)$ -connections on $\Sigma(2, 3, 23)$. For each degree $2i + 1 \in \mathbb{Z}/8\mathbb{Z}$, there are exactly two such irreducible connections and there is no irreducible connection of even degree.*

Proposition 3.38 and (3.41) give a strategy to compute the Chern–Simons functional and the ρ -invariants of irreducible $SU(2)$ -connections. These computations can be used to verify the second part of the above proposition.

We also need to compute the degrees of flat $SU(2)$ -connections when they are regarded as $SU(3)$ -connections. To distinguish between these connections, we will write $\tilde{\alpha}$ for the $SU(3)$ -connection associated to an $SU(2)$ -connection α . The values of the Chern–Simons functional of α and $\tilde{\alpha}$ are equal to each other. However, the ρ -invariants of these two connections are different because ad_α and $\text{ad}_{\tilde{\alpha}}$ define two different representations of the fundamental group. We cannot use Proposition 3.38 to compute the ρ -invariant of $\tilde{\alpha}$ because this connection does not extend to the cobordism W_0 . In [6], cut-and-paste methods have been utilized to compute the difference $\rho_{\text{ad}_{\tilde{\alpha}}} - \rho_{\text{ad}_\alpha}$ for flat $SU(2)$ -connections on a family of homology spheres which includes $\Sigma(2, 3, 23)$. In particular, the following proposition can be extracted from [6]. The claim in it about the nondegeneracy of reducible flat connections on $\Sigma(2, 3, 23)$ is also proved in [6]. For more details about reducible flat $SU(3)$ -connections on $\Sigma(2, 3, 23)$ see Table 5.

Proposition 3.44 *The eight nontrivial reducible flat connections on $\Sigma(2, 3, 23)$ are nondegenerate. The $SU(3)$ –degrees of these connections are given as follows: there is one connection of degree 1, one connection of degree 3, one connection of degree 5, two connections of degree 7, two connections of degree 9 and one connection of degree 11.*

Define $I_*^3(\Sigma(2, 3, 23))$ to be the complex vector space generated by the irreducible flat connections on $\Sigma(2, 3, 23)$. The significance of this vector space for us is a gluing theorem for the 4–manifolds which split along a copy of $\Sigma(2, 3, 23)$. The proof of the following theorem will be given in [Section 6.1](#). This theorem is analogous to [Proposition 3.10](#) and implies that $I_*^3(\Sigma(2, 3, 23))$ plays a role similar to instanton Floer homology of admissible pairs. Motivated by this, we call $I_*^3(\Sigma(2, 3, 23))$ instanton Floer homology of $\Sigma(2, 3, 23)$.

Proposition 3.45 *Let X_1 and X_2 be two 4–manifolds such that $b^+(X_1), b^+(X_2) \geq 1$, $\partial X_1 = \Sigma(2, 3, 23)$ and $\partial X_2 = \overline{\Sigma(2, 3, 23)}$. Let w_i be a closed 2–cycle in X_i and $z_i \in \mathbb{A}(X_i)^{\otimes 2}$. Assume that*

$$\deg(z_1) \equiv -4w_1^2 - 4(\chi(x_1) + \sigma(X_1)) + 4 \pmod{12}.$$

Then there are

$$D_{X_1, w_1}^3(z_1) \in I_4^3(\Sigma(2, 3, 23)) \quad \text{and} \quad D_3^{X_2, w_2}(z_2): I_*^3(\Sigma(2, 3, 23)) \rightarrow \mathbb{C}$$

such that

$$D_3^{X_2, w_2}(z_2) \circ D_{X_1, w_1}^3(z_1) = D_{X_1 \#_{\Sigma(2, 3, 23)} X_2, w_1 \cup w_2}^3(z_1 \cdot z_2).$$

Moreover, $D_3^{X_2, w_2}(z_2)$ is nonzero only on the terms of degree

$$4w_2^2 + 4(\chi(X_2) + \sigma(X_2)) - 4 + \deg(z_2).$$

4 Computing polynomial invariants

In this section, which is the heart of this paper, we first compute the $U(3)$ –polynomial invariants of $E(n)$. The rank 2 invariants of elliptic surfaces were partially computed in [\[30; 32; 31\]](#) using algebrogeometric techniques. Complete calculations of the $U(2)$ –polynomial invariants of elliptic surfaces are given in [\[53; 25; 63\]](#). In [\[25\]](#) and [\[63\]](#), *Gompf decomposition* of elliptic surfaces plays a key role in computing the invariants.

In the introduction, we also recalled a construction of elliptic surfaces which give rise to a decomposition of $E(n)$ into fiber sum of n copies of $E(1)$. This decomposition of elliptic surfaces can be also exploited to compute some of the $U(2)$ -polynomial invariants [69; 18].

Our method for computing the $U(3)$ -invariants of elliptic surfaces uses the Gompf decomposition, the fiber-sum decomposition and the rich group of the symmetries of elliptic surfaces. In Section 4.1, we first review the Gompf decomposition of elliptic surfaces. Then we show how symmetries of elliptic surfaces give constraints on the structure of their $U(3)$ -invariants. Among the elliptic surfaces $E(n)$, the K3 surface $E(2)$ is special because it has a larger symmetry group, which gives more restrictions on the invariants of this 4-manifold. In particular, most of the $U(3)$ -invariants of this 4-manifold are computed in Section 4.2 using identities (C_1) , (C_2) and (C_3) of Section 2.4. In the same subsection, we extend this computation to all $U(3)$ -invariants of $E(2)$ using the blowup formula, and then we explain what can be said about the invariants of the other surfaces $E(n)$ using the same techniques. In Section 4.3, we complete the computation of the invariants of $E(3)$ using the universal formulas of Proposition 2.21 about (-3) -spheres.

Finally, we consider the general case of elliptic surfaces $E(n)$ with $n \geq 4$. The key ingredient here is Proposition 3.45 about the Floer homology of $\Sigma(2, 3, 23)$. This together with Gompf decomposition allows us to compute all invariants of $E(n)$ by induction on n . The calculations of the invariants of $E(2)$ and $E(3)$ in the previous subsections serve as the basis of the induction. In the last subsection of this section, we also give a general gluing theorem for fiber sums. Similar results for $U(2)$ -invariants are proved in [71]. The proof of the rank 3 case follows the same strategy as in [71].

In the next two sections of the paper, we mainly focus on polynomial invariants and instanton Floer homology in the case $N = 3$. Therefore, we shall drop 3 from our notation I_*^3 , $D_{X,w}^3$, et cetera, when it does not create any confusion. Because we are working with an odd value of N , there is not any sign ambiguity in the definition of $I_*^3(W, w, z)$, $D_{X,w}^3(z)$, et cetera and we do not need to fix a homology orientation for the underlying 4-manifold.

4.1 Structure of the invariants of $E(n)$

A construction of the elliptic surface $E(n)$ was reviewed in the introduction. The simplest 4-manifold in this family, $E(1)$, can be also constructed by blowing up the

projective plane $\mathbb{CP}^2$ at the nine intersection points of two generic cubics. The pencil of cubics generated by the two cubics determines an elliptic fibration of $E(1)$. The nine exceptional divisors give rise to sections of this elliptic fibration, which are embedded spheres with self-intersection -1 . The manifold $E(n)$ is the fiber sum of n copies of $E(1)$ along the fibers of the elliptic fibration. The fibration of $E(1)$ induces an elliptic fibration for $E(n)$, and we will write f for a regular fiber of this fibration. By taking the connected sums of the exceptional sections of $E(1)$, we can form nine disjoint embedded spheres in $E(n)$. These are sections of the elliptic fibration of $E(n)$ and have self-intersection $-n$. We fix one of these sections and we will denote it by σ . When it does not cause any confusion, we will use the same notation to denote the homology and the cohomology classes associated to f and σ .

We can assume that there is a cusp fiber in the elliptic fibration of $E(1)$ by choosing appropriate cubics. This gives a cusp fiber f_0 in the fibration of $E(n)$. A regular neighborhood of $\sigma \cup f_0$ is a 4-manifold with boundary $\Sigma(2, 3, 6n - 1)$, which is called the *Gompf nucleus* and is denoted by $G(n)$ [39]. The intersection form of $G(n)$ is given by

$$\begin{bmatrix} 0 & 1 \\ 1 & -n \end{bmatrix}.$$

The complement of $G(n)$ in $E(n)$, denoted by $B(2, 3, 6n - 1)$, is a *Milnor fiber* and its intersection form is given by

(4.1)
$$n(-E_8) \oplus 2(n - 1) \begin{bmatrix} 0 & 1 \\ 1 & -2 \end{bmatrix}.$$

Here each $-E_8$ summand has a basis of embedded spheres with self-intersection -2 which intersect each other according to $-E_8$. The i^{th} summand of the second type in (4.1) has a basis of an embedded torus g_i with self-intersection 0 and an embedded (-2) -sphere τ_i , where g_i and τ_i intersect each other positively at one point [40].

The 4-manifold $E(2)$, which is a K3 surface, plays a special role in this family. For example, $E(2)$ enjoys a rich group of symmetries. As a manifestation of this fact, we have the following proposition:

Lemma 4.2 *Suppose that e and e' are two nonzero elements of $H^2(E(2), \mathbb{Z})$ with $e \cup e \equiv e' \cup e' \pmod{3}$. Then there is an orientation-preserving diffeomorphism Φ of $E(2)$ such that $\Phi^*(e) \equiv e' \pmod{3}$. In particular, the action of $\text{Diff}(E(2))$ on $H^2(E(2), \mathbb{Z}/3\mathbb{Z})$ has four orbits.*

Proof Let $e_1, e_2 \in H^2(E(2), \mathbb{Z})$ be chosen such that $e_i \cup e_j$ is zero when $i = j$, and is equal to 1 when $i \neq j$. Because the action of $\text{Diff}(E(2))$ on the primitive cohomology classes of a fixed self-intersection is transitive, there is an element Ψ of $\text{Diff}(E(2))$ such that $\Psi^*(e) = m(e_1 + ne_2)$ for $m, n \in \mathbb{Z}$. Therefore $\Psi^*(e) \bmod 3$ is equal to one of the following elements:

$$(4.3) \quad 0, \quad e_1, \quad e_1 - e_2, \quad e_1 + e_2.$$

The nonzero classes in (4.3) can be distinguished from each other by their self-intersection. Thus, there is a $\Psi' \in \text{Diff}(E(2))$ such that $\Psi^*(e) \equiv \Psi'^*(e') \bmod 3$. $\square$

For $1 \leq i \leq 3$, suppose w_i is the 2-cycle in $E(2)$ given by $\sigma - (i-1)f$. Then $w_i^2 \equiv i \bmod 3$. Therefore, these 2-cycles and the empty cycle give representatives for the orbits of the action of $\text{Diff}(E(2))$ on $H^2(E(2), \mathbb{Z}/3\mathbb{Z})$. Alternatively, let w'_i be the union of i elements of the nine disjoint spheres of self-intersection -2 in $E(2)$ that were constructed above. If i is a positive integer and $i \equiv j \bmod 3$ with $1 \leq j \leq 3$, then define $w_i := w_j$ and $w'_i := w'_j$. We also define w_0 and w'_0 to be the empty cycles.

The group of diffeomorphisms of $E(n)$ for $n \geq 3$ is more constrained than that of $E(2)$. For example, any diffeomorphism of $E(n)$ maps the homology class f to $\pm f$. Therefore, we cannot expect that the analogue of Lemma 4.2 holds for an arbitrary n . However, $E(n)$ still has a big diffeomorphism group and we can prove the following weakened version of Lemma 4.2:

Lemma 4.4 Suppose that the integers $n \geq 3$ and $1 \leq i \leq 3$ are given and that $u \in H^2(B(2, 3, 6n-1), \mathbb{Z})$ satisfying $u \cup u \equiv i \bmod 3$ is fixed. Suppose also that $e \in H^2(E(n), \mathbb{Z})$ is such that $e \cup f \equiv 0 \bmod 3$ and $e \cup e \equiv i \bmod 3$. Then there is an element Φ of $\text{Diff}(E(n))$ such that

$$(4.5) \quad \Phi^*(e) \equiv kf + u \bmod 3 \quad \text{or} \quad \Phi^*(e) \equiv kf \bmod 3,$$

where $k = 0$ or 1 . Moreover, the map on $H^2(G(n), \mathbb{Z})$ induced by Φ is $\pm \text{id}$. In particular, the action of $\text{Diff}(E(n))$ on $\langle f \rangle^\perp$ in $H^2(E(n), \mathbb{Z}/3\mathbb{Z})$ has eight orbits.

In the statement of Lemma 4.4 we regard u as an element of $H^2(E(n), \mathbb{Z})$ using the inclusion of $B(2, 3, 6n-1)$ in $E(n)$. Note also that the second case in (4.5) holds only if $i = 3$.

Proof The element e can be written as the sum

$$rf + s\sigma + v,$$

where $v \in H^2(B(2, 3, 6n - 1), \mathbb{Z}) \subset H^2(E(n), \mathbb{Z})$. Because $e \cup f \equiv 0 \pmod{3}$, we know s is divisible by 3. There is also a diffeomorphism of $E(n)$ that maps f to $-f$ and σ to $-\sigma$ [39, Lemma 3.7]. After applying this diffeomorphism if it is necessary, we can assume $e \equiv kf + v \pmod{3}$ with $k = 0$ or 1. Suppose $\text{SO}(H^2(E(n), \mathbb{Z}))$ denotes an element of the special orthogonal group of the lattice $H^2(E(n), \mathbb{Z})$ with respect to the intersection bilinear form. According to [41, Proposition 3.3], there is a spinor norm one element of $\text{SO}(H^2(E(n), \mathbb{Z}))$ that fixes f and σ and maps v to an element of the form $m(\tau_1 + ng_1)$. This element can be realized by a diffeomorphism of $E(n)$ [32, Section 6.2]. This can be used to verify the claim as in Lemma 4.2. $\square$

The eight orbits in Lemma 4.4 can be represented by

$$w_{k,0} = kf \quad \text{and} \quad w_{k,l} = kf + \tau_1 - (l-1)g_1$$

for $k = 0, 1$ and $l = 1, 2, 3$. Alternatively, we can use $w'_{k,l} = kf + \tau_1 + \cdots + \tau_l$. If l is a positive integer and $l \equiv j \pmod{3}$ with $1 \leq j \leq 3$, then let $w_{k,l} = w_{k,j}$ and $w'_{k,l} = w'_{k,j}$.

Consider the $\text{U}(3)$ -polynomial invariant $D^3_{E(n),w}(\Gamma^i_{(2)}\Lambda^j_{(3)})$ for the homology classes Γ and Λ . This polynomial is invariant with respect to the action of the diffeomorphism group of $E(n)$ on (w, Γ, Λ) . Therefore, we can use Lemmas 4.2 and 4.4 to focus on a smaller subset of possible values for w . Since changing $w \pmod{3}$ would not change the polynomial invariant and $H^2(E(n), \mathbb{Z}/3\mathbb{Z})$ is finite, the polynomial $D^3_{E(n),w}(\Gamma^i_{(2)}\Lambda^j_{(3)})$ is invariant with respect to the action of a finite-index subgroup of $\text{Diff}(E(n))$ on (Γ, Λ) . This action of the diffeomorphism group factors through the action of the algebraic group $\text{O}(H_2(E(n)))$. Here the orthogonal group is defined using the intersection form on the complex vector space $H_2(E(n))$. Suppose also that $\text{SO}(H_2(E(n)); f)$ is the subgroup of $\text{O}(H_2(E(n)))$ consisting of the orthogonal transformations that map $f \in H_2(E(n))$ to itself and have determinant 1. As another manifestation of the big diffeomorphism group of $E(n)$, it is shown in [32, Section 6.2] that the image of any finite-index subgroup of $\text{Diff}(E(n))$ in $\text{O}(H_2(E(n)))$ contains an algebraically dense subgroup of $\text{SO}(H_2(E(n)); f)$. Therefore, the polynomial $D^3_{E(n),w}(\Gamma^i_{(2)}\Lambda^j_{(3)})$ is invariant with respect to the action of $\text{SO}(H_2(E(n)); f)$ on (Γ, Λ) . In the case that $n = 2$, one can even replace $\text{SO}(H_2(E(n)); f)$ with $\text{SO}(H_2(E(n)))$.

Lemma 4.6 Suppose (V, Q) is a pair of a complex vector space of dimension greater than 2 and a quadratic form. Suppose also τ_1 and τ_2 are two vectors orthogonal to each other such that $Q(\tau_1)$ and $Q(\tau_2)$ are nonzero. Suppose $P: V \oplus V \rightarrow \mathbb{C}$ is a bihomogeneous polynomial of bidegree (d_1, d_2) that is invariant with respect to the diagonal action of $\text{SO}(V)$ on $V \oplus V$. Then P is determined by its value on $W_0 \oplus W_1$, where W_0 is the span of τ_1 , and W_1 is the span of τ_1 and τ_2 . Moreover, P has the form

$$P(x, y) = \sum_{\substack{i, j, k \geq 0 \\ 2i + j = d_1 \\ j + 2k = d_2}} c_{i, j, k} Q(x)^i Q(x, y)^j Q(y)^k$$

for appropriate constants $c_{i, j, k} \in \mathbb{C}$.

In our application, V will be $H_2(E(2))$, and τ_1 and τ_2 will be two disjoint embedded spheres in $E(2)$.

Proof Suppose $(\alpha, \beta) \in V \oplus V$ is such that $Q(\alpha) \neq 0$. Then the vector β can be written as the sum $\beta_0 + \beta_1$ where β_0 is a multiple of α , and β_1 is orthogonal to α . We also assume that $Q(\beta_1) \neq 0$. It is straightforward to find an element of $\text{SO}(V)$ which maps (α, β) to an element of the form $W_0 \oplus W_1$. The set of vectors (α, β) as above is also dense in $V \oplus V$. Therefore, P is determined by its values on $W_0 \oplus W_1$.

By evaluating P on elements of $W_0 \oplus W_1$, we have

$$P(\lambda\tau_1, \mu_1\tau_1 + \mu_2\tau_2) = \sum m_{a, b, c} \lambda^a \mu_1^b \mu_2^c.$$

There is an element of $\text{SO}(V)$ which maps τ_1 to itself (respectively, $-\tau_1$) and τ_2 to $-\tau_2$ (respectively, itself). Therefore, $m_{a, b, c}$ is nonzero only if $a + b$ and c are even. This implies that there are constants $c_{i, j, k}$ such that

$$P(x, y) = \sum_{\substack{2i + j = d_1 \\ j + 2k = d_2}} c_{i, j, k} Q(x)^i Q(x, y)^j Q(y)^k$$

for $(x, y) \in W_0 \oplus W_1$, where j and k are nonnegative integers. This implies the second part of the lemma because each side of the above equality is invariant with respect to the action of $\text{SO}(V)$. Arguing as in [32, Chapter 6, Lemma 2.2], we can also assume that i only takes nonnegative integers. $\square$

Lemma 4.7 Suppose that (V, Q) is a pair of a complex vector space of dimension greater than 3 and a quadratic form. Fix a vector $f \in V$ with $Q(f) = 0$, and

let the vectors k and τ be chosen such that $Q(k, f)$ and $Q(\tau)$ are nonzero, and $Q(\tau, k) = Q(\tau, f) = 0$. Suppose also $P: V \oplus V \rightarrow \mathbb{C}$ is a polynomial that is invariant with respect to the diagonal action of $\mathrm{SO}(V; f)$. Then $P(x, y)$ is determined by its values on $W_0 \oplus W_1$, where W_0 is the span of the vectors f and k , and W_1 is the span of f, k and τ .

In our application, V, f, k and τ will be $H_2(E(n)), f, \sigma$ and an embedded sphere in $B(2, 3, 6n - 1)$, respectively.

Proof The proof is similar to that of the first part of Lemma 4.6. For a given $(\alpha, \beta) \in V \oplus V$, assume $Q(\alpha, f) \neq 0$. Then the vector β can be uniquely written as $\beta_0 + \beta_1$ where $\beta_0 \in \mathrm{Span}(f, \alpha)$, and β_1 is orthogonal to $\mathrm{Span}(f, \alpha)$. As another assumption, we require that $Q(\beta_1) \neq 0$. These assumptions hold for a dense subset of $V \oplus V$. It can be easily checked that there is an element of $\mathrm{SO}(V; k)$ which maps (α, β) to $W_0 \oplus W_1$. Therefore, $P|_{W_0 \oplus W_1}$ determines P on $V \oplus V$. $\square$

4.2 Invariants of $E(2)$

In this section, we study the $\mathrm{U}(3)$ -polynomial invariants of $E(2)$ and a 2-cycle w . Lemma 4.4 shows that we can assume that the 2-cycle w is either empty or w_i for some $1 \leq i \leq 3$. Equivalently, we can replace w_i with w'_i . Throughout this subsection, we will follow the same notation as in the previous part to denote the surfaces σ, τ_i and g_i embedded in $E(2)$.

Proposition 4.8 *The invariant $D_{E(2), w_i}$ satisfies the identities*

$$(4.9) \quad D_{E(2), w_i} \left(\left(\frac{1}{3} a_2 \right)^j z \right) = D_{E(2), w_{i+j}}(z) \quad \text{and} \quad D_{E(2), w_i}(a_3 z) = 0$$

for $z \in \mathbb{A}(E(2))^{\otimes 2}$. In particular, for $1 \leq i \leq 3$, $E(2)$ has w_i -simple type and $\hat{D}_{E(2), w_i}(e^{\Gamma(2) + \Lambda(3)})$ is independent of the choice of i .

Proof Suppose $N(f)$ is a neighborhood of a regular fiber and X is the complement of $N(f)$. We also identify the boundary of X with $Y_1 = S^1 \times f$. The 4-manifold X contains a copy of $B(2, 3, 11)$. Therefore, we can find two disjoint embedded spheres τ_1 and τ_2 of self-intersection -2 in X . Let S be the subspace of $H_2(X)$ spanned by the vectors τ_1 and τ_2 , and $\mathbb{A}(S)$ be the subalgebra $\mathrm{Sym}^*(H_0(X) \oplus S)$ of $\mathbb{A}(X)$. Then w_i can be decomposed as $w \# w'$, where w (respectively, w') is a 2-cycle in X (respectively, $N(f)$), and w and w' intersect Y_1 in $\gamma := S^1 \times \{\mathrm{pt}\}$. For $z \in \mathbb{A}(S)^{\otimes 2}$,

Proposition 3.24, with the aid of the functoriality properties discussed in [Section 3.1](#), implies that

$$\begin{aligned} D_{E(2),w_i}(a_2^j a_3^k z) &= D^{N(f),w'}(1) \circ I_*(Y_1 \times [0, 1], \gamma \times [0, 1], a_2^j a_3^k) \circ D_{X,w}(z) \\ &= 3^j 0^k D^{N(f),w'}(1) \circ I_*(Y_1 \times [0, 1], \gamma \times [0, 1] - jf, 1) \circ D_{X,w}(z) \\ &= 3^j 0^k D_{E(2),w_i-jf}(z). \end{aligned}$$

This verifies [\(4.9\)](#) for the case $z \in \mathbb{A}(S)^{\otimes 2}$. Using [Lemma 4.6](#), the same claim holds for general z , and as a result $\hat{D}_{E(2),w_i}(e^{\Gamma(2)+\Lambda(3)})$ is equal to

$$D_{E(2),w_1}(e^{\Gamma(2)+\Lambda(3)}) + D_{E(2),w_2}(e^{\Gamma(2)+\Lambda(3)}) + D_{E(2),w_3}(e^{\Gamma(2)+\Lambda(3)}).$$

In particular, $\hat{D}_{E(2),w_i}(e^{\Gamma(2)+\Lambda(3)})$ does not depend on i . □

A similar argument, using [Lemma 4.7](#) instead of [Lemma 4.6](#), proves the following analogous statement for $E(n)$:

Proposition 4.10 *The polynomial invariants $D_{E(n),w_{k,l}}$ for $1 \leq l \leq 3$ satisfy the identities*

$$(4.11) \quad D_{E(n),w_{k,l}}\left(\left(\frac{1}{3}a_2\right)^j z\right) = D_{E(2),w_{k,l+j}}(z) \quad \text{and} \quad D_{E(n),w_{k,l}}(a_3 z) = 0$$

for $z \in \mathbb{A}(E(n))^{\otimes 2}$. In particular, for $1 \leq l \leq 3$, $E(n)$ has $w_{k,l}$ -simple type and $\hat{D}_{E(n),w_{k,l}}(e^{\Gamma(2)+\Lambda(3)})$ is independent of the choice of l .

Proposition 4.12 *For $1 \leq i \leq 3$, we have*

$$(4.13) \quad \hat{D}_{E(2),w_i}(e^{\Gamma(2)+\Lambda(3)}) = e^{\frac{1}{2}Q(\Gamma)-Q(\Lambda)}.$$

The two sides of [\(4.13\)](#) are power series in t_2 and t_3 where the coefficients of $t_2^i t_3^j$ are bihomogeneous polynomials on $H_2(E(2)) \oplus H_2(E(2))$ of degree (i, j) . Identity [\(4.18\)](#) means for each choice of (i, j) these coefficients are equal to each other.

Proof Using [Lemma 4.6](#), we can find a power series $g(r, s, t) \in \mathbb{C}[[r, s, t]]$ such that

$$\hat{D}_{E(2),w_i}(e^{\Gamma(2)+\Lambda(3)}) \cdot e^{-\frac{1}{2}Q(\Gamma)+Q(\Lambda)} = g(Q(\Gamma), Q(\Gamma, \Lambda), Q(\Lambda)).$$

The constant term of g is equal to 1 [\[52\]](#). [Lemma 4.6](#) also implies that it suffices to show that $g(Q(\Gamma), Q(\Gamma, \Lambda), Q(\Lambda))$ is equal to 1 for homology classes Γ and Λ and

2–cycles w_i which are orthogonal to an embedded (-2) –sphere τ . For such choices of Γ and Λ , we have

$$\hat{D}_{E(2), w_i}(e^{\Gamma(2) + \epsilon \tau(2) + \Lambda(3)}) \cdot e^{-\frac{1}{2} \mathcal{Q}(\Gamma) + \epsilon^2 + \mathcal{Q}(\Lambda)} = g(\mathcal{Q}(\Gamma) - 2\epsilon^2, \mathcal{Q}(\Gamma, \Lambda), \mathcal{Q}(\Lambda)).$$

By taking the second derivatives of the above expression with respect to ϵ , evaluating ϵ at 0 and using (C_1) of [Section 2.4](#), we obtain

$$\frac{\partial g}{\partial r}(r, s, t) = 0.$$

Similarly, identities (C_2) and (C_3) imply

$$4 \frac{\partial^2 g}{\partial r^2}(r, s, t) - \frac{\partial g}{\partial t}(r, s, t) = 0 \quad \text{and} \quad 2 \frac{\partial^2 g}{\partial s \partial r}(r, s, t) + \frac{\partial g}{\partial s}(r, s, t) = 0.$$

Therefore, g is equal to the constant power series 1. □

Suppose X is the blowup of $E(2)$ and w is the 2–cycle f in X . If E is the exceptional sphere in X , then [Theorem 2.27](#) can be used to compute the invariants of (X, w) :

$$\hat{D}_{X, w}(e^{\Gamma(2) + \Lambda(3)}) = \frac{1}{3} e^{\frac{1}{2} \mathcal{Q}(\Gamma) - \mathcal{Q}(\Lambda)} (\cosh(\sqrt{3}E \cdot \Gamma) + 2 \cos(\sqrt{3}E \cdot \Lambda)).$$

The homology class $\sigma + E$ can be represented by an embedded (-3) –sphere σ' in X . Fix $\Gamma, \Lambda \in H_2(X)$ which are orthogonal to σ' . Then the above formula can be used to show

$$\begin{aligned} (4.14) \quad \hat{D}_{X, w} \left(\left(-\frac{3}{2} \sigma'_{(3)} - \frac{3}{2} \sigma'^2_{(2)} - a_2 \right) e^{\Gamma(2) + \Lambda(3)} \right) \\ = e^{\frac{1}{2} \mathcal{Q}(\Gamma) - \mathcal{Q}(\Lambda)} (-\sqrt{3} \sin(\sqrt{3}E \cdot \Lambda) + \cos(\sqrt{3}E \cdot \Lambda) - \cosh(\sqrt{3}E \cdot \Gamma)). \end{aligned}$$

By another application of [Theorem 2.27](#) and [Remark 2.32](#), $\hat{D}_{X, w - \sigma'}(e^{\Gamma(2) + \Lambda(3)})$ is equal to

$$(4.15) \quad \frac{1}{3} e^{\frac{1}{2} \mathcal{Q}(\Gamma) - \mathcal{Q}(\Lambda)} (\cosh(\sqrt{3}E \cdot \Gamma) - \cos(\sqrt{3}E \cdot \Lambda) + \sqrt{3} \sin(\sqrt{3}E \cdot \Lambda)).$$

Using [Proposition 2.21](#) and comparing these two identities, we can find the undetermined constant c in [Proposition 2.21](#):

Proposition 4.16 *The constant c is equal to -3 .*

Now we are ready to complete the computation of the invariants of $E(2)$:

Theorem 4.17 *Suppose w is a 2–cycle in $E(2)$. Then $E(2)$ has w –simple type and the $U(3)$ –series of $E(2)$ is given by the formula*

$$(4.18) \quad \hat{D}_{E(2), w}(e^{\Gamma(2) + \Lambda(3)}) = e^{\frac{1}{2} \mathcal{Q}(\Gamma) - \mathcal{Q}(\Lambda)}.$$

Proof In light of [Lemma 4.2](#) and [Proposition 4.12](#), it suffices to consider only the empty 2-cycle w_0 . Let σ' be the embedded (-3) -sphere in $E(2) \# \overline{\mathbb{CP}}^2$ constructed above. Consider the 2-cycle $w' := \sigma$ of $E(2) \# \overline{\mathbb{CP}}^2$ and the element $z := (\sigma_{(2)} - 2E_{(2)})^2 z'$ of $\mathbb{A}(E(2) \# \overline{\mathbb{CP}}^2)^{\otimes 2}$, where $z' \in \mathbb{A}(\langle \sigma \rangle^\perp)^{\otimes 2} \cap \mathbb{A}(E(2))^{\otimes 2}$. By [Proposition 2.22](#), the following invariant of $E(2) \# \overline{\mathbb{CP}}^2$ is equal to $4D_{E(2), w_0}(z')$:

$$(4.19) \quad D_{E(2) \# \overline{\mathbb{CP}}^2, w' - \sigma'}((\sigma_{(2)} - 2E_{(2)})^2 z').$$

Moreover, the first identity of [Proposition 2.21](#) can be used to show (4.19) is equal to $D_{E(2) \# \overline{\mathbb{CP}}^2, w'}(z'')$ for an appropriate choice of $z'' \in \mathbb{A}(E(2) \# \overline{\mathbb{CP}}^2)^{\otimes 2}$. Replacing w' with $w'' := \sigma + \tau_1 + g_1$ shows that

$$4D_{E(2), \tau_1 + g_1}(z') = D_{E(2) \# \overline{\mathbb{CP}}^2, w''}(z'').$$

Since $w' \cdot w' \equiv w'' \cdot w'' \pmod{3}$, the left-hand side of the above identity is equal to $D_{E(2) \# \overline{\mathbb{CP}}^2, w'}(z'')$ by the blowup formula. Therefore, we can deduce that

$$(4.20) \quad D_{E(2), w_0}(z') = D_{E(2), \tau_1 + g_1}(z')$$

for $z' \in \mathbb{A}(\langle \sigma \rangle^\perp)^{\otimes 2}$. As a consequence of [Lemma 4.6](#), Identity (4.20) holds for any choice of z' . In particular, $E(2)$ has simple type with respect to w_0 and (4.18) holds for this choice of w . $\square$

Proposition 4.21 Suppose that $\Gamma, \Lambda \in H_2(B(2, 3, 6n - 1)) \subset H_2(E(n))$ and that $\Gamma', \Lambda' \in H_2(G(n)) \subset H_2(E(n))$. Then

$$(4.22) \quad \hat{D}_{E(n), w_{k,l}}(e^{(\Gamma + \Gamma')_{(2)} + (\Lambda + \Lambda')_{(3)}}) \\ = e^{\frac{1}{2}Q(\Gamma) - Q(\Lambda)} \hat{D}_{E(n), w_{k,3}}(e^{(\Gamma')_{(2)} + (\Lambda')_{(3)}}).$$

Proof The group of orthogonal transformations $\text{SO}(H_2(B(2, 3, 6n - 1)), Q)$, regarded as a subgroup of $\text{SO}(H_2(E(n)))$, acts as the identity on the series $\hat{D}_{E(n), w}(e^{\Gamma_{(2)} + \Lambda_{(3)}})$. This fact can be combined with the argument of [Proposition 4.12](#) to verify (4.22) for $1 \leq l \leq 3$. To finish the proof, it suffices to show that $\hat{D}_{E(n), w_{k,0}}(e^{\Gamma_{(2)} + \Lambda_{(3)}})$ is equal to $\hat{D}_{E(n), w_{k,3}}(e^{\Gamma_{(2)} + \Lambda_{(3)}})$. This also can be achieved with the method of the proof of [Theorem 4.17](#). $\square$

By [Proposition 4.10](#), we already know that $E(n)$ has $w_{k,l}$ -simple type for $1 \leq l \leq 3$. The above proof can be used to show that $E(n)$ has $w_{k,0}$ -simple type, too.

4.3 Invariants of $E(3)$

In this section, we study the $U(3)$ -polynomial invariants of $E(3)$ up to a constant and a sign ambiguity. The following is [Theorem 2](#) from the introduction:

Theorem 4.23 *The 4-manifold $E(3)$ has simple type. There are also real numbers $\hbar_1$ and $\hbar_2$ such that for any 2-cycle w in $E(3)$ and $\Gamma, \Lambda \in H_2(E(3))$, the series $\hat{D}_{E(3),w}(e^{\Gamma(2)+\Lambda(3)})$ is equal to*

$$e^{\frac{1}{2}\mathcal{Q}(\Gamma)-\mathcal{Q}(\Lambda)}(\hbar_1 \cosh(\sqrt{3}f \cdot \Gamma) - 2\hbar_2 \cos(-\frac{2}{3}\pi w \cdot f + \sqrt{3}f \cdot \Lambda)).$$

Furthermore, $\hbar_1 + \hbar_2 = \pm 1$ for an appropriate choice of the sign.

In a work in progress, we show that $\hbar_1 = \frac{2}{3}$ and $\hbar_2 = \frac{1}{3}$. However, we do not need the exact values of these constants in this paper. Later, we only use the fact that $\hbar_1$ and $\hbar_2$ are nonzero. To abbreviate our notation, define

$$(4.24) \quad G(\Gamma, \Lambda, j) := \hbar_1 \cosh(\sqrt{3}f \cdot \Gamma) - 2\hbar_2 \cos(-\frac{2}{3}\pi j + \sqrt{3}f \cdot \Lambda).$$

Proposition 4.25 *Suppose w is a 2-cycle in $E(3)$ with $w \cdot f \not\equiv 0 \pmod{3}$. Then*

$$(4.26) \quad D_{E(3),w}((\frac{1}{3}a_2)^j z) = D_{E(3),w-jf}(z) \quad \text{and} \quad D_{E(3),w}(a_3 z) = 0$$

for $z \in \mathbb{A}(E(3))^{\otimes 2}$. In particular, $E(3)$ has w -simple type. Furthermore, there is a power series $g \in \mathbb{Q}[[t_2, t_3]]$ such that for $\Gamma, \Lambda \in H_2(E(3))$,

$$\hat{D}_{E(3),w}(e^{\Gamma(2)+\Lambda(3)}) = e^{\frac{1}{2}\mathcal{Q}(\Gamma)-\mathcal{Q}(\Lambda)} g(\Gamma \cdot f, \Lambda \cdot f)$$

when $w \cdot f \equiv 1 \pmod{3}$, and

$$\hat{D}_{E(3),w}(e^{\Gamma(2)+\Lambda(3)}) = e^{\frac{1}{2}\mathcal{Q}(\Gamma)-\mathcal{Q}(\Lambda)} g(-\Gamma \cdot f, -\Lambda \cdot f)$$

when $w \cdot f \equiv 2 \pmod{3}$. Furthermore, g is even with respect to the variable t_2 and its constant term is equal to ± 1 .

Proof The 4-manifold $E(3)$ is given as the fiber sum $E(2) \#_f E(1)$. In this proof, let σ_n denote a section of the elliptic fibration of $E(n)$, which is a sphere of self-intersection $-n$. We can assume that $\sigma_3 = \sigma_2 \# \sigma_1$. Firstly, consider the case $w \cdot f \equiv 1 \pmod{3}$. Arguing as in [Lemma 4.4](#), we can assume that $w = w_1 \# \sigma_1$, where w_1 is a 2-cycle in $E(2)$ with $w_1 \cdot f = 1$. Suppose Γ_0 and Λ_0 are two elements of $H_2(E(2))$

such that $\Gamma_0 \cdot f = \Lambda_0 \cdot f = 1$. Then [Proposition 3.35](#) for $X_1 = E(2)$, $X_2 = E(1)$ and $X_3 = X_4 = S^2 \times f$ implies that

$$\begin{aligned} p(t_2, t_3) \sum_{0 \leq j \leq 2} D_{E(2) \#_f E(1), w+jf} (e^{(\Gamma_0 \# \sigma_1)(2) + (\Lambda_0 \# \sigma_1)(3)}) \\ = q(t_2, t_3) \sum_{0 \leq j \leq 2} D_{E(2) \#_f S^2 \times f, w_2+jf} (e^{(\Gamma_0)(2) + (\Lambda_0)(3)}), \end{aligned}$$

where

$$p(t_2, t_3) = \sum_{0 \leq j \leq 2} D_{S^2 \times f, S^2 \times \{\text{pt}\} + jf} (e^{\Delta(2) + \Delta(3)})$$

and

$$q(t_2, t_3) = \sum_{0 \leq j \leq 2} D_{E(1), w_1+jf} (e^{(\sigma_1)(2) + (\sigma_1)(3)}).$$

Note that $b^+(S^2 \times f) = b^+(E(1)) = 1$ and we use the invariants with respect to the metrics that have long necks along f in the above identities. The power series $p(t_2, t_3)$ is invertible, because $p(0, 0) = 1$. Therefore, we can conclude there is a power series $g(t_2, t_3)$ such that

$$(4.27) \quad \sum_{0 \leq j \leq 2} D_{E(3), w+jf} (e^{\Gamma(2) + \Lambda(3)}) = e^{\frac{1}{2}Q(\Gamma) - Q(\Lambda)} g(t_2, t_3)$$

with $\Gamma = \Gamma_0 \# \sigma_1$ and $\Lambda = \Lambda_0 \# \sigma_1$ for Γ_0 and Λ_0 given as above. Then [Lemma 4.7](#) implies that (4.27) holds for arbitrary Γ and Λ with $\Gamma \cdot f = \Lambda \cdot f = 1$. By [Proposition 4.8](#), we can use a similar argument as above to show

$$\sum_{0 \leq j \leq 2} D_{E(3), w+jf} (P(a_2, a_3) e^{\Gamma(2) + \Lambda(3)}) = P(3, 0) e^{\frac{1}{2}Q(\Gamma) - Q(\Lambda)} g(t_2, t_3).$$

In particular, this shows that

$$\begin{aligned} D_{E(3), w} \left(\left(\frac{1}{3} a_2 \right)^j e^{s\Gamma(2) + t\Lambda(3)} \right) &= D_{E(3), w-jf} (e^{\Gamma(2) + \Lambda(3)}), \\ D_{E(3), w} (a_3 e^{s\Gamma(2) + t\Lambda(3)}) &= 0. \end{aligned}$$

The power series g is even with respect to t_2 , because $D(e^{\Gamma(2) + \Lambda(3)})$ is even with respect to t_2 . A similar application of [Proposition 3.35](#) for $X_1 = X_2 = E(1)$ and $X_3 = X_4 = S^2 \times f$ shows that

$$e^{-t_2^2 + 2t_3^2} p(t_2, t_3) = q(t_2, t_3)^2.$$

The constant term of the above equality and the identity $p(0, 0) = 1$ shows that the constant term of g is equal to ± 1 . This fact completes the proof for the case

that $w \cdot f \equiv 1 \pmod{3}$. Using a diffeomorphism of $E(3)$ which maps f to $-f$, we can also treat the case that $w \cdot f \equiv 2 \pmod{3}$. $\square$

To determine the power series g , let σ and σ' be two disjoint sections of the elliptic fibration of $E(3)$. Let also w be chosen such that $w \cdot f = 1$ and $w \cdot \sigma = 2$. Then

$$\begin{aligned} \hat{D}_{E(3),w}(e^{(s\sigma+s'\sigma')_{(2)}+(r\sigma+r'\sigma')_{(3)}}) \\ = e^{-3t_2^2(s^2+s'^2)/2+3t_3^2(r^2+r'^2)} g((s+s')t_2, (r+r')t_3). \end{aligned}$$

By differentiating with respect to s and r , we can conclude that

$$\hat{D}_{E(3),w}(\sigma_{(2)}^i \sigma_{(3)}^j z) = h \frac{d^i}{ds^i} \frac{d^j}{dr^j} \Big|_{s=t=0} e^{-3t_2^2 s^2/2+3t_3^2 r^2} g((s+s')t_2, (r+r')t_3),$$

where $z = e^{s'\sigma'_{(2)}+r'\sigma'_{(3)}}$ and $h = e^{-3t_2^2 s'^2/2+3t_3^2 r'^2}$. By applying these identities to the second formula in [Proposition 2.21](#), we can conclude

$$(4.28) \quad -\frac{1}{2}g_3 + \frac{1}{2}g_{22} - \frac{1}{2}g = g \circ \tau,$$

where g_{22} means the second derivative of the power series $g(t_2, t_3)$ with respect to t_2 , and so on. Moreover, τ maps (t_2, t_3) to $(-t_2, -t_3)$. We can use (4.28) to derive the identity

$$(4.29) \quad \frac{1}{2}(g \circ \tau)_3 + \frac{1}{2}(g \circ \tau)_{22} - \frac{1}{2}g \circ \tau = g.$$

Replacing $g \circ \tau$ in (4.29) with the left-hand side of (4.28) gives rise to the following PDE for g :

$$(4.30) \quad g_{2222} - g_{33} - 2g_{22} - 3g = 0.$$

Next, let w' be a 2-cycle with $w' \cdot f = 1$ and $w' \cdot \sigma = 0$ and consider

$$\hat{D}_{E(3),w'}(e^{(s\sigma+s'\sigma')_{(2)}+(r\sigma+r'\sigma')_{(3)}})$$

instead. With the same argument, the last part of [Proposition 2.21](#) implies that

$$(4.31) \quad g_{2222} - 6g_{22} + 3g_{33} + 9g = 0 \quad \text{and} \quad g_{2223} - 6g_{23} = 0.$$

We can combine (4.30) and the first equation in (4.31) to come up with the following simpler PDE:

$$(4.32) \quad g_{33} - g_{22} + 3g = 0.$$

The second PDE in (4.31) and the fact that g is even with respect to the variable t_2 imply that $g_{st} = p(t) \cosh(\sqrt{6}s)$. The equations (4.30) and (4.32) can be used to write two linear ordinary differential equations for p . It is straightforward to check that the only solution of these ODEs is $p(t) = 0$. Therefore, the power series g has the form $q_1(t) + q_2(s)$. Equation (4.32) can be used to find differential equations for q_1 and q_2 . By solving these ODEs and using the fact that g is even with respect to t_2 , we can conclude

$$(4.33) \quad g(t_2, t_3) = a \cosh(\sqrt{3}t_2) + b \cos(\sqrt{3}t_3) + c \sin(\sqrt{3}t_3).$$

If $g(0, 0) = 1$, then the initial value and (4.28) imply the following constraints on a , b and c , which can be used to prove Theorem 4.23 in the case $w \cdot f \not\equiv 0 \pmod{3}$:

$$a + b = 1 \quad \text{and} \quad a - \frac{1}{2}b - \frac{\sqrt{3}}{2}c = 1.$$

A similar argument can be used in the case that $g(0, 0) = -1$.

Next, let $w \cdot f \equiv 0 \pmod{3}$. Using Lemma 4.4 and Proposition 4.10, it suffices to consider the case that $w = w_{k,1}$ for $k = 0$ or 1 . The following proposition computes the invariants of $E(3)$ for this choice of 2-cycle. In its proof, we use the basis for the homology of $H_2(E(3), \mathbb{Z})$ which is introduced in Section 4.1:

Proposition 4.34 *For any $(\Gamma, \Lambda) \in H_2(E(3)) \oplus H_2(E(3))$,*

$$(4.35) \quad \hat{D}_{E(3), w_{k,1}}(e^{\Gamma(2) + \Lambda(3)}) = e^{\frac{1}{2}\mathcal{Q}(\Gamma) - \mathcal{Q}(\Lambda)} G(\Gamma, \Lambda, 0).$$

Proof Let σ' be a section of the elliptic fibration of $E(3)$ which is disjoint from σ . Then the homology class of σ' is equal to $\sigma + 3f + u$, where $u \in H_2(B(2, 3, 17), \mathbb{Z})$ and $u \cdot u = -6$. Arguing as in Lemma 4.4, there is a diffeomorphism Φ of $E(3)$ that fixes $H_2(G(3))$ and maps u to a linear combination of g_1 and τ_1 . Furthermore, $\Phi_*(u) \equiv 2g_1 - \tau_1 \pmod{3}$. In particular, $\alpha := \Phi(\sigma')$ is a (-3) -sphere with $\alpha \cdot w_{k,1} \equiv 1 \pmod{3}$. Suppose W is the subspace of the elements of $H_2(E(3))$ whose intersection numbers with α are equal to 0. Using Proposition 2.21, the series $\hat{D}_{E(3), w_{k,1}}(e^{\Gamma(2) + \Lambda(3)})$, for any $(\Gamma, \Lambda) \in W \oplus W$, is equal to

$$(4.36) \quad -\frac{1}{3} \hat{D}_{E(3), w_{k,1} + \alpha} \left(\left(-\frac{3}{2} \alpha(3) - \frac{3}{2} \alpha(2)^2 - a_2 \right) e^{\Gamma(2) + \Lambda(3)} \right).$$

Since $(w_{k,1} + \alpha) \cdot f = 1$, we can evaluate (4.36) using our current knowledge of the invariants of $E(3)$. It is straightforward to check that the resulting series is equal to that in (4.35).

The homology classes f , $k := \sigma + \frac{3}{2}f$ and τ_2 satisfy the assumption of [Lemma 4.7](#) for $V = H_2(E(3))$. Suppose W_0 and W_1 are given as in [Lemma 4.7](#), and U is the subset of $W_0 \oplus W_1$ consisting of the pairs that satisfy (4.35). Then U is a Zariski closed subset of $W_0 \oplus W_1$. In order to complete the proof, we shall show that U contains a Euclidean open set in $W_0 \oplus W_1$ and hence $U = W_0 \oplus W_1$. Let $(\Gamma, \Lambda) \in W_0 \oplus W_1$ be given by

$$\Gamma := af + bk \quad \text{and} \quad \Lambda = a'f + b'k + c\tau_2.$$

Consider the homology classes

$$u_1 := \frac{1}{\sqrt{2}}(\tau_1 + \tau_3) \quad \text{and} \quad u_2 := \frac{1}{\sqrt{2}}(\tau_1 - \tau_3),$$

which have nonzero intersection with α . There is an element $A_{t,\theta} \in \text{SO}(H_2(E(n)); f)$ such that

$$A_{t,\theta}(\Gamma) := af + b(k + t^2f + tu_1),$$

$$A_{t,\theta}(\Lambda) := a'f + b'(k + t^2f + tu_1) + c'(\cos(\theta)\tau_2 + \sin(\theta)u_2).$$

If a and a' are close enough to each other and b , b' and c' are close enough to 1, then t and θ can be chosen such that $A_{t,\theta}(\Gamma), A_{t,\theta}(\Lambda) \in W$. Therefore, U contains an open subset of $W_0 \oplus W_1$. $\square$

4.4 Invariants of $E(n)$

In this section, we compute the invariants of the elliptic surface $E(n)$. We start with the simpler case that $w \cdot f \not\equiv 0 \pmod{3}$:

Proposition 4.37 *Suppose w is a 2-cycle in $E(n)$ with $w \cdot f \not\equiv 0 \pmod{3}$. Then $E(n)$ has w -simple type, and for $\Gamma, \Lambda \in H_2(E(n))$,*

$$\hat{D}_{E(n),w}(e^{\Gamma(2)+\Lambda(3)}) = e^{\frac{1}{2}Q(\Gamma)-Q(\Lambda)} G(\Gamma, \Lambda, w \cdot f)^{n-2}.$$

Proof The proof of this proposition is similar to that of [Proposition 4.25](#). Applying [Proposition 3.35](#) for $X_1 = E(n-2)$, $X_2 = E(2)$ and $X_3 = X_4 = E(1)$ gives us enough relations to verify the proposition by induction. $\square$

In the blown-up elliptic surface $E(n) \# \overline{\mathbb{CP}}^2$, there is an embedded sphere with self-intersection $-(n+1)$, given by tubing a section of the elliptic fibration and the exceptional sphere in $\overline{\mathbb{CP}}^2$. Therefore, there is a copy of the Gompf nucleus $G(n+1)$ in $G(n) \# \overline{\mathbb{CP}}^2$. The homology class $E + f$ can be realized by an embedded surface $\bar{E}$

in $M_n := G(n) \# \overline{\mathbb{CP}}^2 \setminus G(n+1)$. The surface $\bar{E}$ determines a generator of $H_2(M_n)$. Similarly, the nucleus $G(n+2)$ can be embedded in $G(n) \# 2\overline{\mathbb{CP}}^2$. Therefore, there are copies of $G(4)$ in the 4-manifolds $E(4)$, $E(3) \# \overline{\mathbb{CP}}^2$ and $E(2) \# 2\overline{\mathbb{CP}}^2$. Let $Z_0 \subset E(4)$, $Z_1 \subset E(3) \# \overline{\mathbb{CP}}^2$ and $Z_2 \subset E(2) \# 2\overline{\mathbb{CP}}^2$ be the complements of $G(4)$ in these manifolds. Then the boundary of Z_i is diffeomorphic to $\Sigma(2, 3, 23)$. It is clear from the inductive construction of $E(n)$ that there is an embedding of Z_i in $E(n-i) \# i\overline{\mathbb{CP}}^2$ for $n \geq 4$. In fact, if $W(n)$ is the fiber sum $E(n-4) \#_f G(4)$, then $E(n-i) \# i\overline{\mathbb{CP}}^2$ is diffeomorphic to $Z_i \#_{\Sigma(2,3,23)} W(n)$.

The 4-manifold Z_i gives rise to elements of $I_*(\Sigma(2, 3, 23))$ as it is explained in [Proposition 3.45](#). Suppose $V_0 \subseteq I_*(\Sigma(2, 3, 23))$ is the vector space generated by the element $D_{Z_0, v_0}(1)$ where v_0 is a 2-cycle in Z_0 with $v_0^2 \equiv 0 \pmod{3}$. Similarly, define V_1 to be the vector space generated by the three elements $D_{Z_1, w}(1)$ where w is one of the following elements, which satisfy $w^2 \equiv 1 \pmod{3}$:

$$v_1 := \bar{E} + \tau_1 - g_1, \quad v_2 := -\bar{E} + \tau_1 - g_1, \quad v_3 := \tau_1.$$

Finally, let V_2 be the subspace of $I_*(\Sigma(2, 3, 23))$ which is generated by the elements of the form $D_{Z_2, w}(1)$ where $w^2 \equiv 2 \pmod{3}$.

Proposition 4.38 *The space V_i is a subspace of $I_4(\Sigma(2, 3, 23))$. Furthermore, the dimension of V_i is at least $\frac{1}{2}(i+2)(i+1)$.*

Proof The first part of the proposition is an immediate consequence of [Proposition 3.45](#). To show that $\dim(V_0) = 1$, let $D_{Z_0, v_0}(1) = 0$. By [Proposition 3.45](#), $D_{E(4), v_0 + w_0}(z)$ vanishes for a 2-cycle w_0 in $G(4)$ and $z \in \mathbb{A}(G(4))^{\otimes 2}$. If w_0 is chosen such that $w_0 \cdot f \not\equiv 0 \pmod{3}$, then [Proposition 4.37](#) asserts that there is a z such that $D_{E(4), v_0 + w_0}(z) \neq 0$, which is a contradiction.

Next, we consider the case that $i = 1$. By [Proposition 3.45](#), a linear relation among the vectors $D_{Z_1, v_l}(1)$ for $1 \leq l \leq 3$ implies that there are constant numbers c_l such that

$$(4.39) \quad \sum_{1 \leq l \leq 3} c_l \hat{D}_{E(3) \# \overline{\mathbb{CP}}^2, v_l + w_0}(e^{\sigma(2) + \sigma(3)}) = 0.$$

Here w_0 is a 2-cycle in $G(4)$ and σ is the embedded (-4) -sphere in $G(4)$. We can use the blowup formula and the results of the previous subsection to evaluate the above power series and to conclude that

$$\sum_{1 \leq l \leq 3} c_l (\cosh(\sqrt{3}t_2) + \zeta^{-(v_l + w_0) \cdot E} e^{i\sqrt{3}t_3} + \zeta^{(v_l + w_0) \cdot E} e^{-i\sqrt{3}t_3}) = 0.$$

Since $v_l \cdot E \equiv -l \pmod{3}$, the constants c_l are equal to zero. Therefore, $\dim(V_1) = 3$. Similar proofs can be applied to the case that $i = 2$. $\square$

Theorem 4.40 Suppose w is a 2-cycle in $E(n)$ and $\Gamma, \Lambda \in H_2(E(n))$. Then

$$(4.41) \quad \widehat{D}_{E(n),w}(e^{\Gamma(2)+\Lambda(3)}) = e^{\frac{1}{2}\mathcal{Q}(\Gamma)-\mathcal{Q}(\Lambda)} G(\Gamma, \Lambda, w \cdot f)^{n-2}.$$

Proof Since the dimension of $I_4(\Sigma(2, 3, 23))$ is equal to 9, Proposition 4.38 implies that there are vectors $\phi_i \in V_i$ such that

$$(4.42) \quad \phi_0 + \phi_1 + \phi_2 = 0.$$

We use this identity to verify (4.41) inductively for the remaining invariants of $E(n)$. Firstly, consider the case that $\phi_0 \neq 0$. Then we can assume that $\phi_0 = D_{Z_0, v_0}(1)$. There are also 2-cycles $w_1, \dots, w_k$ in Z_2 and constants $c_1, c_2, c_3, c'_1, \dots, c'_k$ such that

$$(4.43) \quad \phi_1 = \sum_{l=1}^3 c_l D_{Z_1, v_l}(1) \quad \text{and} \quad \phi_2 = \sum_{j=1}^k c'_j D_{Z_2, w_j}(1).$$

Suppose $\Gamma, \Lambda \in H_2(W(n))$ and w is a 2-cycle in $W(n)$. Then Proposition 3.45 implies that the $U(3)$ -series $\widehat{D}_{E(n), v_0+w}(e^{\Gamma(2)+\Lambda(3)})$ is equal to

$$(4.44) \quad \begin{aligned} & - \sum_l c_l \widehat{D}_{E(n-1)\#2\overline{\mathbb{CP}}^2, v_l+w}(e^{\Gamma(2)+\Lambda(3)}) \\ & - \sum_j c'_j \widehat{D}_{E(n-2)\#2\overline{\mathbb{CP}}^2, w_j+w}(e^{\Gamma(2)+\Lambda(3)}) \\ & = e^{\frac{1}{2}\mathcal{Q}(\Gamma)-\mathcal{Q}(\Lambda)} G(\Gamma, \Lambda, w \cdot f)^{n-4} A(\Gamma, \Lambda), \end{aligned}$$

where the term $A(\Gamma, \Lambda)$ is a linear combination of the six expressions

$$(4.45) \quad \begin{aligned} & \cosh(\sqrt{3}f \cdot \Gamma)^2, \quad \cosh(\sqrt{3}f \cdot \Gamma) \zeta^{\mp w \cdot f} e^{\pm i \sqrt{3}f \cdot \Lambda}, \\ & \zeta^{\mp 2w \cdot f} e^{\pm 2i \sqrt{3}f \cdot \Lambda}, \quad 1, \end{aligned}$$

and the coefficients of this linear combination do not depend on w . To derive (4.44) we use the facts that $\Gamma \cdot E = -\Gamma \cdot f$ and $\Lambda \cdot E = -\Lambda \cdot f$. In the case $w \cdot f \not\equiv 0 \pmod{3}$, $A(\Gamma, \Lambda)$ is equal to $G(\Gamma, \Lambda, w \cdot f)^2$ using Proposition 4.37. This identity holds also in the case that $w \cdot f \equiv 0 \pmod{3}$, because $A(\Gamma, \Lambda)$ is a linear combination of the terms in (4.45) with coefficients which are independent of w . Therefore, for a general 2-cycle w in $W(n)$ and $\Gamma, \Lambda \in H_2(W(n))$,

$$\widehat{D}_{E(n), v_0+w}(e^{\Gamma_2+\Lambda_3}) = e^{\frac{1}{2}\mathcal{Q}(\Gamma)-\mathcal{Q}(\Lambda)} G(\Gamma, \Lambda, w \cdot f)^{n-2}.$$

Then Proposition 4.21 implies $\hat{D}_{E(n),w}(e^{\Gamma_2+\Lambda_3})$ equals $e^{\frac{1}{2}\mathcal{Q}(\Gamma)-\mathcal{Q}(\Lambda)}G(\Gamma, \Lambda, w)^{n-2}$ for any 2-cycle w in $E(n)$ and $\Gamma, \Lambda \in H_2(E(n))$.

Next, assume that $\phi_0 = 0$. We assume that the nonzero vectors ϕ_1 and ϕ_2 are given as in (4.43). Fix an arbitrary 2-cycle w in $W(n+1)$ and homology classes $\Gamma, \Lambda \in H_2(W(n+1))$. Another application of Proposition 3.45 shows that

$$(4.46) \quad \sum_{1 \leq l \leq 3} c_l \hat{D}_{E(n) \# \overline{\mathbb{CP}}^2, v_l + w}(e^{\Gamma(2) + \Lambda(3)})$$

is equal to

$$\sum_j c'_j \hat{D}_{E(n-1) \# 2\overline{\mathbb{CP}}^2, w_j + w}(e^{\Gamma(2) + \Lambda(3)}).$$

By our inductive calculation of the invariants of $E(n)$, the latter expression is equal to

$$e^{\frac{1}{2}\mathcal{Q}(\Gamma)-\mathcal{Q}(\Lambda)}G(\Gamma, \Lambda, w \cdot f)^{n-3}B(\Gamma, \Lambda).$$

Here $B(\Gamma, \Lambda)$ is a linear combination of the six expressions

$$\cosh(\sqrt{3}f \cdot \Gamma)^2, \quad \cosh(\sqrt{3}f \cdot \Gamma)\zeta^{\mp w \cdot f}e^{\pm i\sqrt{3}f \cdot \Lambda}, \quad \zeta^{\mp 2w \cdot f}e^{\pm 2i\sqrt{3}f \cdot \Lambda}, \quad 1.$$

As in the previous case, the coefficients of this linear combination are determined by c'_j and w_j . In particular, they do not depend on w . Therefore, we can determine these coefficients by considering the case that $w \cdot f \not\equiv 0 \pmod{3}$, for which we already computed the invariants. Therefore, $B(\Gamma, \Lambda)$ is equal to

$$G(\Gamma, \Lambda, w \cdot f) \sum_{1 \leq l \leq 3} \frac{1}{3}c_l(\cosh(\sqrt{3}E \cdot \Gamma) + \zeta^{-w \cdot E}e^{i\sqrt{3}E \cdot \Lambda} + \zeta^{w \cdot E}e^{-i\sqrt{3}E \cdot \Lambda}).$$

For an arbitrary 2-cycle $w \subset E(n) \# \overline{\mathbb{CP}}^2$ and $\Gamma, \Lambda \in E(n) \# \overline{\mathbb{CP}}^2$, let $\hat{P}_w(\Gamma, \Lambda)$ be the power series given by subtracting $\hat{D}_w(e^{\Gamma(2) + \Lambda(3)})$ from the power series

$$e^{\frac{1}{2}\mathcal{Q}(\Gamma)-\mathcal{Q}(\Lambda)}G(\Gamma, \Lambda, w \cdot f)^{n-2}\frac{1}{3}(\cosh(\sqrt{3}E \cdot \Gamma) + \zeta^{w \cdot E}e^{i\sqrt{3}E \cdot \Lambda} + \zeta^{-w \cdot E}e^{-i\sqrt{3}E \cdot \Lambda}).$$

Then we can rephrase our conclusion in the form of the identity

$$(4.47) \quad \sum_{1 \leq l \leq 3} c_l \hat{P}_{v_l + w}(\Gamma, \Lambda) = 0$$

where w is a 2-cycle in $W(n+1)$ and $\Gamma, \Lambda \in H_2(W(n+1))$. Suppose $p_w^{i,j}$ is the polynomial on $H_2(E(n) \# \overline{\mathbb{CP}}^2) \oplus H_2(E(n) \# \overline{\mathbb{CP}}^2)$ of bidegree (i, j) determined by the coefficient of $t_2^i t_3^j$ in $\hat{P}_w$. Then $p_w^{i,j}$ can be evaluated at

$$(\Gamma_1, \dots, \Gamma_i; \Lambda_1, \dots, \Lambda_j)$$

for $\Gamma_k, \Lambda_k \in H_2(E(n) \# \overline{\mathbb{CP}}^2)$. By induction on $i + j$, we shall show that $p_w^{i,j}$ vanishes for all possible choices of w . By considering the constant terms of (4.47) for empty w , we have

$$c_1 p_{v_1}^{0,0} + c_2 p_{v_2}^{0,0} + c_3 p_{v_3}^{0,0} = 0.$$

The blowup formula asserts that $p_{v_1}^{0,0} = p_{v_2}^{0,0} = 0$. Therefore, if $c_3 \neq 0$, then $p_{v_3}^{0,0} = 0$. Thus Proposition 4.21 and the blowup formula show that $p_w^{0,0} = 0$ for all 2-cycles w in $E(n) \# \overline{\mathbb{CP}}^2$. If $c_3 = 0$, then by (4.47),

$$c_1 p_{v_1}^{2,0}(\sigma + E, \sigma + E) + c_2 p_{v_2}^{2,0}(\sigma + E, \sigma + E) = 0$$

and

$$c_1 p_{v_1}^{0,1}(\sigma + E) + c_2 p_{v_2}^{0,1}(\sigma + E) = 0.$$

The blowup formula asserts that

$$c_1 p_{v_1-E}^{0,0} + c_2 p_{v_2+E}^{0,0} = 0 \quad \text{and} \quad c_1 p_{v_1-E}^{0,0} - c_2 p_{v_2+E}^{0,0} = 0.$$

Consequently, at least one of the numbers $p_{v_1-E}^{0,0}$ and $p_{v_2+E}^{0,0}$ is zero and we can derive the same conclusion as in the previous case.

Now assume that the polynomial $p_w^{i,j}$ vanishes for $i + j \leq k$ and any 2-cycle in $E(n) \# \overline{\mathbb{CP}}^2$. Fix (i, j) , $\tilde{\Gamma}_1, \dots, \tilde{\Gamma}_i$ and $\tilde{\Lambda}_1, \dots, \tilde{\Lambda}_j$ such that $i + j = k + 1$ and $\tilde{\Gamma}_l$ and $\tilde{\Lambda}_{l'}$ are equal to either $\sigma + E$ or f . Then apply (4.47) to conclude that

$$(4.48) \quad \sum_{1 \leq l \leq 3} c_l p_{v_l}^{i,j}(\tilde{\Gamma}_1, \dots, \tilde{\Gamma}_i; \tilde{\Lambda}_1, \dots, \tilde{\Lambda}_j) = 0.$$

The blowup formula and the induction hypothesis imply that every term of the form $\sigma + E$ can be replaced with σ . Thus, if $c_3 \neq 0$,

$$(4.49) \quad p_{v_1}^{i,j}(\Gamma_1, \dots, \Gamma_i; \Lambda_1, \dots, \Lambda_j) = 0$$

for $\Gamma_k, \Lambda_k \in H_2(G(n))$. Therefore, Proposition 4.21 and the blowup formula allow us to complete the verification of the induction step. If $c_3 = 0$, we can use the analogue of (4.48) for

$$(\sigma + E, \sigma + E, \tilde{\Gamma}_1, \dots, \tilde{\Gamma}_i; \tilde{\Lambda}_1, \dots, \tilde{\Lambda}_j) \quad \text{and} \quad (\tilde{\Gamma}_1, \dots, \tilde{\Gamma}_i; \sigma + E, \tilde{\Lambda}_1, \dots, \tilde{\Lambda}_j)$$

and argue as in the basis of the induction. This completes the proof of the theorem. $\square$

4.5 Gluing 4-manifolds along surfaces of self-intersection zero

In this subsection, we use the calculation of $U(3)$ -polynomial invariants for elliptic surfaces to study invariants of another family of closed 4-manifolds:

Definition 4.50 Suppose that X is a smooth 4-manifold, Σ is an oriented surface of genus $g \geq 1$ embedded in X , w is a 2-cycle in X , and $\mathcal{H}$ is a subspace of $H_2(X)$. Then (X, Σ, w) is *permissible with respect to the subspace $\mathcal{H}$* if the following hold:

- (i) $b^1(X) = 0$ and $b^+(X) > 1$.
- (ii) $\Sigma \cdot \Sigma = 0$.
- (iii) $w \cdot \Sigma \not\equiv 0 \pmod{3}$.
- (iv) Let $z \in \mathbb{A}(\mathcal{H})^{\otimes 2}$, and let u be the 2-cycle $w + l\Sigma$ for $l = 0, 1$ or 2 . Then

$$(4.51) \quad D_{X,u}\left(\left(\frac{1}{3}a_2\right)^3 z\right) = D_{X,u}(z) \quad \text{and} \quad D_{X,u}(a_3 z) = 0.$$

Moreover, there are cohomology classes $K_i \in H^2(X, \mathbb{Z})$ such that K_i is an integral lift of $w_2(TX)$ and $|K_i \cdot \Sigma| \leq 2g - 2$, and for $\Gamma, \Lambda \in \mathcal{H}$, the power series $\hat{D}_{X,u}(e^{\Gamma(2) + \Lambda(3)})$ is equal to

$$(4.52) \quad e^{\frac{1}{2}\mathcal{Q}(\Gamma) - \mathcal{Q}(\Lambda)} \sum_{i,j} c_{ij} \zeta^{-l\Sigma \cdot (\frac{1}{2}(K_i - K_j))} e^{\frac{\sqrt{3}}{2}(K_i + K_j) \cdot \Gamma + \frac{\sqrt{3}}{2}i(K_i - K_j) \cdot \Lambda}.$$

The cohomology class K_i is called a *basic class* of the triple (X, Σ, w) and $c_{i,j}$ is called the coefficient associated to the pair (K_i, K_j) . In the case that $\mathcal{H} = H_2(X)$, we say (X, Σ, w) is permissible.

Example 4.53 The results of Section 4.2 show that the triple $(E(2), f, w)$ forms a permissible triple where f is a fiber in an elliptic fibration of $E(2)$ and w is a 2-cycle such that $w \cdot f \not\equiv 0 \pmod{3}$. In this case, the only basic class in (4.52) is the zero cohomology class. More generally, we can embed a surface of genus g in $E(2)$ whose self-intersection is equal to $2g - 2$. For example, we can construct such an embedded surface by considering the union of g fibers and a section of the fibration, and then smoothing out the intersection points. Let Σ be the proper transform of this surface after blowing up $E(2)$ at $2g - 2$ points on the surface. Let also w be a 2-cycle in $E(2) \# (2g - 2)\overline{\mathbb{CP}}^2$ such that $w \cdot \Sigma \not\equiv 0 \pmod{3}$. Then the blowup formula implies that $(E(2) \# (2g - 2)\overline{\mathbb{CP}}^2, \Sigma, w)$ is a permissible triple. If $E_1, \dots, E_{2g-2}$ are the exceptional classes, then a basic class of this triple has the form $\pm E_1 \pm \dots \pm E_{2g-2}$.

Example 4.54 One can further generalize the previous example by considering a surface Σ with self-intersection 0 embedded in the 4-manifold $E(n) \# k \overline{\mathbb{CP}}^2$. Let w be a 2-cycle in X such that $w \cdot \Sigma \not\equiv 0 \pmod{3}$. Then the blowup formula and [Theorem 4.40](#) can be utilized to verify most requirements of [Definition 4.50](#) for permissibility of (X, Σ, w) . The only missing part is to verify the inequality $|K \cdot \Sigma| \leq 2g - 2$ for basic classes K of X . To check this inequality, note that our basic classes for (X, Σ, w) are the same as $U(2)$ -basic classes for X [[53](#); [25](#); [63](#); [54](#); [26](#)]. Therefore, the desired inequality is a consequence of the adjunction inequality in [[54](#)].

Remark 4.55 Among the conditions in [Definition 4.50](#), (iv) might seem the most mysterious one. The first part of this condition is a weaker version of the simple type condition in (1.7). We expect that if X has simple type in the sense of [[54](#)] and (X, w, Σ) satisfies properties (i), (ii) and (iii) of [Definition 4.50](#), then condition (iv) automatically holds. In fact, we propose a more general conjecture as [Conjecture 7.2](#) in [Section 7](#). This conjecture is the $U(3)$ version of Kronheimer and Mrowka's structure theorem [[54](#)].

Definition 4.56 For $g \geq 1$, the set of all integer pairs (a, b) which satisfy the following two properties is denoted by $\mathcal{C}_g$:

- (i) a and b have the same parity.
- (ii) $|a| + |b| \leq 2g - 2$.

We will write N_g for the number of the elements of $\mathcal{C}_g$, which is equal to $2g(g-1) + 1$.

The following proposition can be verified using the permissible triples provided by [Example 4.53](#):

Proposition 4.57 For any $(a, b) \in \mathcal{C}_g$, there are a permissible triple (X, w, Σ) and basic classes K_i and K_j of the triple such that Σ has genus g and

$$(4.58) \quad a = \frac{1}{2}((K_i + K_j)) \cdot \Sigma \quad \text{and} \quad b = \frac{1}{2}((K_i - K_j)) \cdot \Sigma.$$

For a permissible triple (X, Σ, w) , define

$$D_{X,w,\Sigma}(e^{\Gamma(2)+\Lambda(3)}) := \sum_{0 \leq l \leq 2} D_{X,w+l\Sigma}(e^{\Gamma(2)+\Lambda(3)}).$$

More generally, $D_{X,w,\Sigma}(z)$ is defined to be the sum $D_{X,w+l\Sigma}(z)$ for different values of $0 \leq l \leq 2$. The dimension formula shows that if all terms in z have a fixed degree, then only one of the 2-cycles $w + l\Sigma$ is involved in the definition of $D_{X,w,\Sigma}(z)$.

Lemma 4.59 Suppose (X, w, Σ) is a permissible triple as above and $d_w \in \mathbb{Z}/3\mathbb{Z}$ is defined to be $b^+ + 1 - w \cdot w$. Then $D_{X,w,\Sigma}((\frac{1}{3}a_2)^m e^{\Gamma(2)+\Lambda(3)})$ is equal to

$$\sum c_{ij} \zeta^{l_{i,j}(d_w-m)} e^{\zeta^{2l_{i,j}(\frac{1}{2}\mathcal{Q}(\Gamma)+\frac{\sqrt{3}}{2}\mathbf{i}(K_i-K_j)\cdot\Lambda)+\zeta^{l_{i,j}}(-\mathcal{Q}(\Lambda)+\frac{\sqrt{3}}{2}(K_i+K_j)\cdot\Gamma)},$$

where the inner sum is over all pairs of basic classes (K_i, K_j) . Moreover, $l_{i,j} \in \mathbb{Z}/3\mathbb{Z}$ is equal to $(w \cdot \Sigma)(\frac{1}{2}(K_i - K_j)) \cdot \Sigma$.

Proof For a 4-manifold W with simple type, the power series $D_{X,w}$ can be recovered from $\hat{D}_{X,w}$ in the following way:

$$D_{X,w}((\frac{1}{3}a_2)^m e^{\Gamma(2)+\Lambda(3)}) = \frac{1}{3} \sum_{0 \leq k \leq 2} \zeta^{k(d_w-m)} \hat{D}_{X,w}(e^{\zeta^k \Gamma(2)+\zeta^{2k} \Lambda(3)}).$$

Therefore

$$\begin{aligned} \sum_{0 \leq l \leq 2} D_{X,w+l\Sigma}((\frac{1}{3}a_2)^m e^{\Gamma(2)+\Lambda(3)}) \\ = \frac{1}{3} \sum_{0 \leq k, l \leq 2} \zeta^{k(d_w+l\Sigma-m)} \hat{D}_{X,w+l\Sigma}(e^{\zeta^k \Gamma(2)+\zeta^{2k} \Lambda(3)}). \end{aligned}$$

Then, we can use the permissibility of (X, w, Σ) to rewrite the right-hand side in terms of basic classes. A straightforward simplification gives the desired result. $\square$

Suppose (X, w, Σ) and (X', w', Σ) are two permissible triples such that the embedded surfaces of genus g are identified with each other, and this identification is lifted to the normal bundles. Suppose also w and w' intersect Σ in the same number of points with the same signs. As explained in Section 3.3, we can form the triple $(X \#_{\Sigma} X', w \# w', \Sigma)$. There is also a subspace $\mathcal{K} \subseteq H_2(X) \oplus H_2(X')$ such that there is a map $\#: \mathcal{K} \rightarrow H_2(X \#_{\Sigma} X')$. The main goal of this section is to compute $\hat{D}_{X \#_{\Sigma} X', w \# w'}(e^{\Gamma(2)+\Lambda(3)})$ for $\Gamma, \Lambda \in \text{im}(\#)$ in terms of the invariants of the pairs (X, w) and (X', w') .

The basic idea to achieve this goal is to use the gluing property in (3.30). Therefore, the R_N -module $\mathbb{I}_{g,d}^N$, introduced in Section 3.3, for $N = 3$ and $d = w \cdot \Sigma$, plays a key role in computing the invariants of $(X \#_{\Sigma} X', w \# w')$. In fact, we can replace $\mathbb{I}_{g,d}^3$ with a smaller module. Before giving the definition of this smaller module, we introduce some conventions. From now on, we drop 3 from our notation and denote this Fukaya-Floer homology module with $\mathbb{I}_{g,d}$. Moreover, for a permissible triple (X, w, Σ) the intersection $w \cdot \Sigma$ is denoted by d , unless otherwise is stated.

Let $\tilde{\mathbb{I}}_{g,d} \subset \mathbb{I}_{g,d}$ be the $\mathbb{C}[[t_2, t_3]]$ -module generated by the following relative elements in $\mathbb{I}_{g,d}$:

$$(4.60) \quad D_{X^\circ, w^\circ, \Sigma}(e^{\Gamma_{(2)}^\circ + \Lambda_{(3)}^\circ}) := \sum_{l \in \mathbb{Z}/3\mathbb{Z}} D_{X^\circ, w^\circ + l\Sigma}(e^{\Gamma_{(2)}^\circ + \Lambda_{(3)}^\circ}),$$

where (X, w, Σ) is a permissible triple, and $\Gamma, \Lambda \in H_2(X, \mathbb{Z})$ with $\Gamma \cdot \Sigma = \Lambda \cdot \Sigma = 1$. By (3.30), the pairing of this element with $D_{\Delta_g, \delta_g}(ze^{D_{(2)} + D_{(3)}})$ is equal to the following element of $\mathbb{C}[[t_2, t_3]]$:

$$(4.61) \quad D_{X, w, \Sigma}(ze^{\Gamma_{(2)} + \Lambda_{(3)}}).$$

Suppose $\mathbb{C}[[t_2, t_3]][x, y, z]$ is the ring of polynomials of three variables with coefficients in $\mathbb{C}[[t_2, t_3]]$ and $z = P(a_2, \Sigma_{(2)}, \Sigma_{(3)})$ for $P \in \mathbb{C}[[t_2, t_3]][x, y, z]$. Then Lemma 4.59 shows that the pairing of (4.60) and $D_{\Delta_g, \delta_g}(ze^{D_{(2)} + D_{(3)}})$ is equal to

$$\sum_{(a,b) \in \mathcal{C}_g} \left(\zeta^{dbd_w} P(u(a, b)) \times \sum c_{i,j} e^{\zeta^{2db}(\frac{1}{2}\mathcal{Q}(\Gamma) + \frac{\sqrt{3}}{2}i(K_i - K_j) \cdot \Lambda) + \zeta^{db}(-\mathcal{Q}(\Lambda) + \frac{\sqrt{3}}{2}(K_i + K_j) \cdot \Gamma)} \right),$$

where the inner sum is over all pairs of basic classes (K_i, K_j) such that

$$(4.62) \quad \frac{1}{2}(K_i + K_j) \cdot \Sigma = a \quad \text{and} \quad \frac{1}{2}(K_i - K_j) \cdot \Sigma = b,$$

and

$$(4.63) \quad u(a, b) := (3\zeta^{2db}, \zeta^{db}\sqrt{3}a + \zeta^{2db}t_2, \zeta^{2db}\sqrt{3}ib - 2\zeta^{db}t_3).$$

For $\lambda = (\alpha, \beta) \in \mathcal{C}_g$, fix a polynomial $P_\lambda \in \mathbb{C}[[t_2, t_3]][x, y, z]$ such that

$$(4.64) \quad P_\lambda(u(\alpha, \beta)) = 1 \quad \text{and} \quad P_\lambda(u(a, b)) = 0 \quad \text{for } (a, b) \neq (\alpha, \beta).$$

Define a map $\Phi: \tilde{\mathbb{I}}_{g,d} \rightarrow \mathbb{C}[[t_1, t_2]]^{N_g}$ by

$$\Phi(\eta) := \left\{ \langle \eta, D_{\Delta_g, \delta_g}(P_\lambda(a_2, \Sigma_{(2)}, \Sigma_{(3)})e^{D_{(2)} + D_{(3)}}) \rangle \right\}_{\lambda \in \mathcal{C}_g}.$$

By (3.30), the homomorphism Φ maps the relative element in (4.60) to an element of $\mathbb{C}^{N_g}[[t_2, t_3]]$, whose component corresponding to $\lambda = (a, b)$, denoted by $c_{X, w, \Sigma}^\lambda(\Gamma, \Lambda)$, is equal to

$$(4.65) \quad \zeta^{dbd_w} \sum c_{i,j} e^{\zeta^{2db}(\frac{1}{2}\mathcal{Q}(\Gamma) + \frac{\sqrt{3}}{2}i(K_i - K_j) \cdot \Lambda) + \zeta^{db}(-\mathcal{Q}(\Lambda) + \frac{\sqrt{3}}{2}(K_i + K_j) \cdot \Gamma)},$$

where the sum is over the pairs of basic classes (K_i, K_j) that satisfy (4.62). Let $\mathbb{C}((t_1, t_2))$ denote the field of fractions of $\mathbb{C}[[t_2, t_3]]$. Then Φ induces a map

$$\bar{\Phi}: \tilde{\mathbb{I}}_{g,d} \otimes_{\mathbb{C}[[t_1, t_2]]} \mathbb{C}((t_1, t_2)) \rightarrow \mathbb{C}((t_1, t_2))^{N_g}.$$

Proposition 4.66 *The map Φ is injective. Moreover, $\bar{\Phi}$ is an isomorphism of vector spaces.*

Proof Suppose $\mathcal{I}$ is the ideal in $\mathbb{A}_g^3 \otimes_{\mathbb{C}} \mathbb{C}[[t_2, t_3]]$ generated by a_3 , $\gamma_{(2)}$, $\gamma_{(3)}$ and the elements $P(a_2, \Sigma_{(2)}, \Sigma_{(3)})$, where $P \in \mathbb{C}[[t_2, t_3]][x, y, z]$ is a polynomial evaluating to zero at the points in (4.63). Then the pairing of (4.60) and $D_{\Delta_g, \delta_g}(ze^{D_{(2)}+D_{(3)}})$ vanishes when $z \in \mathcal{I}$. Any element of $\tilde{\mathbb{I}}_{g,d}$ is also invariant with respect to the action of

$$(4.67) \quad \tilde{\epsilon} := \mathbb{I}_*^N([0, 1] \times Y_g, [0, 1] \times \gamma_{g,d} + \Sigma, e^{([0,1] \times \gamma)_{(2)} + \dots + ([0,1] \times \gamma)_{(N)}}).$$

Recall that $Y_g = \Sigma \times S^1$ and γ in (4.67) denotes an S^1 -fiber of Y_g . Any $z \in \mathbb{A}_g^3 \otimes_{\mathbb{C}} \mathbb{C}[[t_2, t_3]]$ can be written as a sum of an element of $\mathcal{I}$ and a $\mathbb{C}[[t_2, t_3]]$ -linear combination of the polynomials $\{P_\lambda\}_{\lambda \in \mathcal{C}_g}$. Therefore, injectivity of Φ is a consequence of Proposition 3.31.

For a given $\lambda_0 \in \mathcal{C}_g$, Proposition 4.57 gives a permissible triple (X, w, Σ) such that the component of the relative element (4.60) corresponding to λ_0 is nonzero. Furthermore, we can change the relative class as in (4.60) by replacing Γ with $\Gamma + s\Sigma$ and Λ with $\Lambda + s\Sigma$. The component of this relative element corresponding to $\lambda = (a, b)$ picks the factor $e^{s(\xi^{2b}\sqrt{3}at_2 + \xi^b\sqrt{3}bt_3 + \xi^b t_2^2 - 2\xi^{2b}t_3^2)}$. Therefore, by taking $\mathbb{C}[[t_1, t_2]]$ -linear combinations of such expressions for different values of s , we can produce elements of $\tilde{\mathbb{I}}_{g,d}$ such that the component corresponding to λ_0 is the only nonzero element. This verifies the second part of the proposition. $\square$

Consider the restriction of the pairing $\langle \cdot, \cdot \rangle$ on $\mathbb{I}_{g,d}$. The above proposition implies that this pairing induces a pairing on $\mathbb{C}((t_1, t_2))^{N_g}$ using the map $\bar{\Phi}$:

$$(4.68) \quad \langle \cdot, \cdot \rangle: \mathbb{C}((t_1, t_2))^{N_g} \times \mathbb{C}((t_1, t_2))^{N_g} \rightarrow \mathbb{C}((t_1, t_2)).$$

Suppose this pairing is given by $\{\xi^{2bd(g-1)}h_{\lambda, \lambda'}^{g,d}\}_{\lambda, \lambda' \in \mathcal{C}_g}$ with respect to the standard basis of $\mathbb{C}((t_1, t_2))^{N_g}$, where b is the second coordinate of λ . The constant $\xi^{2bd(g-1)}$ does not play an important role here. It will be used to obtain a slightly simpler form for our gluing formulas in Proposition 4.71.

Proposition 4.69 *The element $h_{\lambda, \lambda'}^{g,d}(t_2, t_3) \in \mathbb{C}((t_1, t_2))$ is nonzero only if $\lambda = \lambda'$. Furthermore, if $\lambda = (a, b)$ and $|\lambda|_1 := |a| + |b| < 2g - 2$, then $h_{\lambda, \lambda}^{g,d}$ is zero.*

Proof Suppose (X, w, Σ) and (X', w', Σ) are two permissible triples such that w and w' intersect Σ in the same set of points with the same signs. Suppose the homology classes $\Gamma, \Lambda \in H_2(X)$ and $\Gamma', \Lambda' \in H_2(X')$ are chosen such that their intersection with Σ is equal to 1. Then (3.30) asserts that

$$(4.70) \quad D_{X \#_{\Sigma} X', w \# w', \Sigma} (e^{(\Gamma \# \Gamma')_{(2)} + (\Lambda \# \Lambda')_{(3)}}) \\ = \sum_{\lambda, \lambda' \in C_g} \zeta^{2bd(g-1)} h_{\lambda, \lambda'}^{g, d} c_{X, w, \Sigma}^{\lambda}(\Gamma, \Lambda) c_{X', w', \Sigma}^{\lambda'}(\Gamma', \Lambda').$$

Replacing Γ, Γ', Λ and Λ' with $\Gamma + r\Sigma, \Gamma' - r\Sigma, \Lambda + s\Sigma$ and $\Lambda' - s\Sigma$ does not change the left-hand side of the above identity. On the right-hand side, the term corresponding to λ and λ' changes by a factor of the form $e^{r \cdot f(\lambda, \lambda') + s \cdot g(\lambda, \lambda')}$. Here $f(\lambda, \lambda')$ and $g(\lambda, \lambda')$, which can be computed explicitly, are zero if and only if $\lambda = \lambda'$. Therefore, $h_{\lambda, \lambda'}^{g, d}$ has to be nonzero when $\lambda \neq \lambda'$.

Let (X, w, Σ) be the permissible triple of Example 4.53 where Σ has genus $g - 1$. Taking the connected sum of Σ and a homologically trivial torus embedded in a 4-ball produces a permissible triple (X, w, Σ') such that Σ' has genus g . Then $X \#_{\Sigma} X$ can be decomposed as the connected sum of $S^2 \times S^2$ and another 4-manifold with $b^+ > 0$. Theorem 6.14 asserts that for Γ, Λ, Γ' and $\Lambda' \in H_2(X)$, the invariant

$$D_{X \#_{\Sigma'} X, w \# w, \Sigma} (e^{(\Gamma \# \Gamma')_{(2)} + (\Lambda \# \Lambda')_{(3)}})$$

vanishes. If $|\lambda|_1 < 2g - 2$, then we can find Γ and Λ such that $c_{X, w, \Sigma'}^{\lambda}$ is nonzero. Consequently, $h_{\lambda, \lambda}^{g, d}$ is zero for this choice of λ . $\square$

In light of the above proposition, let $h_{a, b}^{g, d}$ be $h_{\lambda, \lambda}^{g, d}$ for $\lambda = (a, b) \in C_g \setminus C_{g-1}$. These are the only nonzero terms among the coefficients of the pairing.

Proposition 4.71 Suppose (X, w, Σ) and (X', w', Σ) are two permissible triples such that w and w' intersect Σ in the same set of points with the same signs. Then

$$(4.72) \quad D_{X \#_{\Sigma} X', w \# w'} \left(\left(\frac{1}{3} a_2 \right)^3 z \right) = D_{X \#_{\Sigma} X', w \# w'}(z) \quad \text{and} \quad D_{X \#_{\Sigma} X', w \# w'}(a_3 z) = 0$$

for $z \in \text{Sym}^*(H_0(X \#_{\Sigma} X') \oplus \text{im}(\#))^{\otimes 2}$. Suppose also that the intersection number of the homology classes $\Gamma, \Lambda \in H_2(X)$ and $\Gamma', \Lambda' \in H_2(X')$ with Σ is 1. Then $\hat{D}_{X \#_{\Sigma} X', w \# w'}(e^{(\Gamma \# \Gamma')_{(2)} + (\Lambda \# \Lambda')_{(3)}})$ is equal to

$$\begin{aligned} & e^{\frac{1}{2} Q(\Gamma \# \Gamma') - Q(\Gamma \# \Gamma')} \\ & \times \sum_{(a, b) \in C_g \setminus C_{g-1}} h_{a, b}^{g, d}(t_2, t_3) \sum c_{ij} d_{i' j'} e^{\left(\frac{\sqrt{3}}{2} (M_{i, i'} + M_{j, j'}) \cdot \Gamma \# \Gamma' + \frac{\sqrt{3}}{2} i (M_{i, i'} - M_{j, j'}) \cdot \Lambda \# \Lambda' \right)}. \end{aligned}$$

For each (a, b) , the second sum is over the pairs of basic classes (K_i, K_j) and $(L_{i'}, L_{j'})$ such that

$$(4.73) \quad \frac{1}{2}(K_i + K_j) \cdot \Sigma = \frac{1}{2}(L_{i'} + L_{j'}) \cdot \Sigma = a, \quad \frac{1}{2}(K_i - K_j) \cdot \Sigma = \frac{1}{2}(L_{i'} - L_{j'}) \cdot \Sigma = b,$$

and $M_{i,i'}$ and $M_{j,j'}$ are respectively equal to $K_i \# L_{i'}$ and $K_j \# L_{j'}$.

Proof The series $D_{X \#_{\Sigma} X', w \# w', \Sigma}(e^{(\Gamma \# \Gamma')_{(2)} + (\Lambda \# \Lambda')_{(3)}})$ can be computed in terms of the cohomology classes $M_{i,i'}$ by substituting (4.65) into (4.70) and applying Proposition 4.69. Then we argue as in Lemma 4.59 to obtain the desired formula for the $U(N)$ -series $\hat{D}_{X \#_{\Sigma} X', w \# w', \Sigma}(e^{(\Gamma \# \Gamma')_{(2)} + (\Lambda \# \Lambda')_{(3)}})$. We can follow a similar strategy to compute $\hat{D}_{X \#_{\Sigma} X', w \# w', \Sigma}(a_2^i a_3^j e^{(\Gamma \# \Gamma')_{(2)} + (\Lambda \# \Lambda')_{(3)}})$ in terms of the classes $M_{i,i'}$. The resulting formulas prove the identities in (4.72). $\square$

The goal of the remaining part of this section is to determine the power series $h_{a,b}^{g,d}(t_2, t_3)$ up to two constants. Firstly, one can obtain a constraint on this power series by changing the orientation of Σ :

$$(4.74) \quad h_{a,b}^{g,2}(t_2, t_3) = h_{-a,-b}^{g,1}(-t_2, -t_3).$$

We shall obtain more constraints by looking at some explicit 4-manifolds. In the case $g = 1$, we can in fact determine $h_{0,0}^{1,d}(t_2, t_3)$ completely using our calculation of the invariants of $E(n)$:

Corollary 4.75 *For $g = 1$, the only nonzero term among the pairing coefficients is*

$$h_{0,0}^{1,d}(t_2, t_3) = (\hbar_1 \cosh(\sqrt{3}t_2) - 2\hbar_2 \cos(-\frac{2}{3}\pi d + \sqrt{3}t_3))^2.$$

In particular, if (X, w, Σ) and (X', w', Σ) are permissible triples such that the genus of Σ is equal to 1, and w and w' intersect Σ in the same set of points with the same signs, then

$$(4.76) \quad \hat{D}_{X \#_{\Sigma} X', w \# w', \Sigma}(e^{(\Gamma \# \Gamma')_{(2)} + (\Lambda \# \Lambda')_{(3)}}) = h_{0,0}^{1,d} \hat{D}_{X,w}(e^{\Gamma_{(2)} + \Lambda_{(3)}}) \hat{D}_{X',w'}(e^{\Gamma'_{(2)} + \Lambda'_{(3)}}).$$

Remark 4.77 Identity (4.76) is a consequence of Proposition 3.35, and it holds even if we only require (i), (ii) and (iii) of Definition 4.50 for the triples (X, w, Σ) and (X', w', Σ) .

The 4-manifold $B(1) := E(1) \# \overline{\mathbb{C}\mathbb{P}}^2$ has two elliptic fibrations with fibers f and f' such that $f \cdot f' = 1$. There are also disjoint embedded spheres E and E' in $B(1)$

such that E (respectively, E') intersects f (respectively, f') positively in one point and is disjoint from f' (respectively, f). The fiber sum of n copies of $B(1)$ along f' produces a 4-manifold $B(n)$ which is diffeomorphic to $E(n) \# n\overline{\mathbb{CP}}^2$ (Figure 1). The exceptional spheres of $B(n)$ are denoted by $E^1, \dots, E^n$, where E^i is given by the exceptional sphere E in the i^{th} summand of $B(n)$.⁷ Furthermore, one can glue copies of f to produce a surface f_n of genus n with self-intersection zero. The homology class of f_n is equal to $[S] - [E^1] - \dots - [E^n]$ for an appropriate surface S of genus n in $E(n)$. Similarly, one can glue copies of E' to produce σ_n , a sphere of self-intersection $-n$ that is disjoint from f_n and the exceptional spheres. The intersection number of σ_n and f' is equal to 1. The following proposition shows that $(B(n), f_n, w)$ is permissible if $w \cdot f \not\equiv 0 \pmod{3}$:

Proposition 4.78 Suppose w is a 2-cycle in $B(n)$ such that $w \cdot f_n \not\equiv 0 \pmod{3}$. For $n \geq 2$, the triple $(B(n), w, f_n)$ is permissible and $\hat{D}_{B(n), w}(e^{\Gamma(2) + \Lambda(3)})$ is given by
$$e^{\frac{1}{2}\mathcal{Q}(\Gamma) - \mathcal{Q}(\Lambda)} G'(\Gamma, \Lambda, w \cdot f')^{n-2} \prod_i \frac{1}{3} (\cosh(\sqrt{3}E^i \cdot \Gamma) + 2\cos(-\frac{2}{3}\pi w \cdot E^i + \sqrt{3}E^i \cdot \Lambda))$$

for $\Gamma, \Lambda \in H_2(B(n))$. Here G' is given by (4.24) after replacing f with f' . In particular, a basic class of $B(n)$ has the form

$$(n - 2k)f' \pm E^1 \pm \dots \pm E^n \quad \text{for } 1 \leq k \leq n - 1.$$

Proof This follows from the blowup formula and Theorem 4.40. □

From the construction, it is clear that there is a diffeomorphism (see Figures 2 and 3)

$$(4.79) \quad \Phi_n: B(n) \#_{f_n} B(n) \rightarrow B(2) \#_{f_2} \dots \#_{f_2} B(2).$$

This diffeomorphism maps f_n and $f'_2 := f' \# f'$ in $B(n) \#_{f_n} B(n)$ to $f'_n := f' \# \dots \# f'$ and f_2 in $B(2) \#_{f_2} \dots \#_{f_2} B(2)$. The sphere $\sigma_n \subset B(n)$ determines two spheres of self-intersection $-n$ in $B(n) \#_{f_n} B(n)$, which are denoted by σ_n^1 and σ_n^2 . The diffeomorphism Φ_n maps σ_n^i to $E_n^i := E^i \# \dots \# E^i$. Therefore, the following elements of $\mathbb{C}[[r_2, s_2, r_3, s_3, t_2, t_3]]$ are equal to each other:

$$(4.80) \quad \hat{D}_{B(n) \#_{f_n} B(n), k f'_2 + l f_n}(e^{(r_2 f_n + s_2 f'_2)(2) + (r_3 f_n + s_3 f'_2)(3)}) \\ = \hat{D}_{B(2) \#_{f_2} \dots \#_{f_2} B(2), k f_2 + l f'_n}(e^{(r_2 f'_n + s_2 f_2)(2) + (r_3 f'_n + s_3 f_2)(3)}).$$

⁷For the special case of exceptional spheres in $B(n)$, we deviate from our previous notation and put i as a superscript.

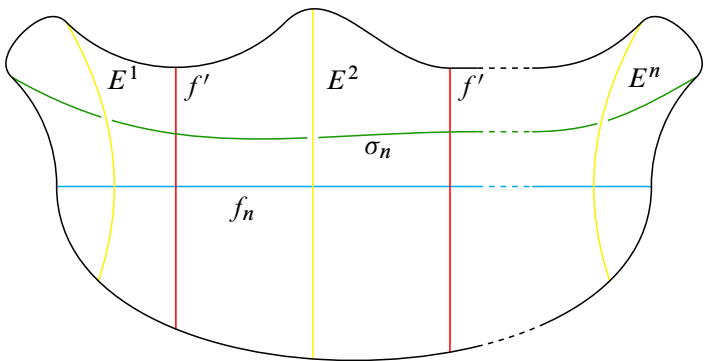


Figure 1: A schematic of $B(n)$: each embedded curve represents an embedded surface and each intersection point of these curves represents an element in the intersection of the corresponding surfaces. Note that E^i and σ_n do not intersect.

Proposition 4.81 Suppose $(a, b) \in \mathcal{C}_2 \setminus \{(0, 0)\}$. Then

(4.82)
$$h_{a,b}^{2,d}(t_2, t_3) = \sum_{(\gamma, \eta) \in \mathcal{C}_2 \setminus \{(0, 0)\}} h_{a,b,\gamma,\eta}^{2,d} e^{\sqrt{3}\gamma t_2 + \sqrt{3}i\eta t_3},$$

where $h_{a,b,\gamma,\eta}^{2,d}$ is a constant number.

Proof The triple $(B(2), df' + 2df_2, f_2)$ is permissible, and Proposition 4.71 can be utilized to compute the following element of $\mathbb{C}[[r_2, r_3, s_2, s_3, t_2, t_3]]$:

$$\hat{D}_{B(2) \#_{f_2} B(2), df'_2 + df_2} (e^{(s_2 f'_2 + r_2 f_2)_{(2)} + (s_3 f'_2 + r_3 f_2)_{(3)}}).$$

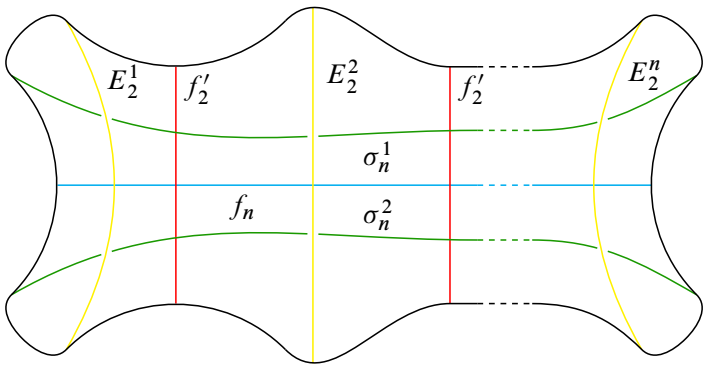


Figure 2: A schematic of $B(n) \#_{f_n} B(n)$.

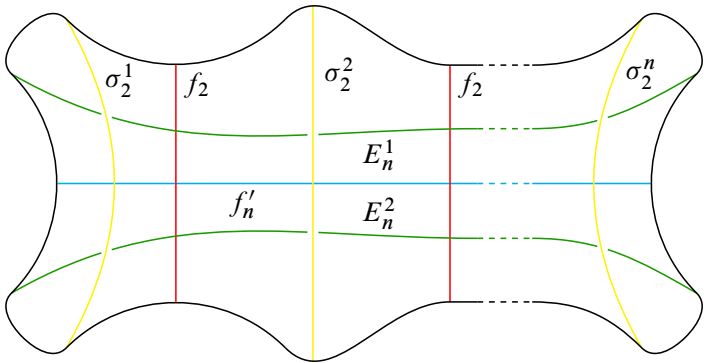


Figure 3: A schematic of $B(2) \#_{f_2} \cdots \#_{f_2} B(2)$.

We can evaluate this series at $t_2 = t_3 = 1$ to obtain a well-defined element of $\mathbb{C}[[r_2, r_3, s_2, s_3]]$. This power series is equal to

(4.83)

$$e^{r_2 s_2 - 2 r_3 s_3} \sum_{(a,b) \in \mathcal{C}_2 \setminus \{(0,0)\}} h_{a,b}^{2,d}(s_2, s_3) e^{\sqrt{3} i a r_2 + \sqrt{3} i b r_3} \left(\sum_{(K_i, K_j)} c_{ij} \right)^2,$$

where the inner sum is over all pairs of basic classes (K_i, K_j) of $(B(2), df' + 2df_2, f_2)$ such that

$$\frac{1}{2}(K_i + K_j) \cdot f_2 = a \quad \text{and} \quad \frac{1}{2}(K_i - K_j) \cdot f_2 = b.$$

For each choice of (a, b) , the inner sum is nonzero. The identity (4.80) shows that the expression (4.83) is invariant with respect to the symmetry of $\mathbb{C}[[r_2, r_3, s_2, s_3]]$ that switches r_2 with s_2 and r_3 with s_3 . This can be used to show that $h_{(a,b)}^{2,d}(t_2, t_3)$ has the form in (4.82). □

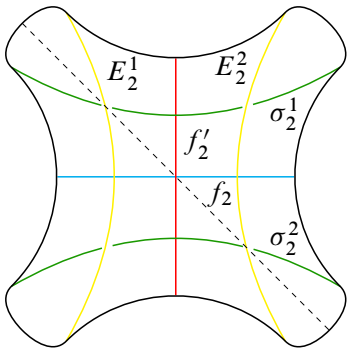


Figure 4: A schematic of $B(2) \#_{f_2} B(2)$: reflection with respect to the dashed line represents the diffeomorphism Φ_2 of this 4-manifold.

Proposition 4.84 The constant numbers $h_{a,b,\gamma,\eta}^{2,d}$ are zero except possibly

$$h_{2,0,2,0}^{2,d}, \quad h_{0,2,0,2}^{2,d}, \quad h_{0,-2,0,-2}^{2,d} \quad \text{and} \quad h_{-2,0,-2,0}^{2,d}.$$

Furthermore, there are real numbers $\hbar_3$ and $\hbar_4$ such that

$$(4.85) \quad h_{2,0,2,0}^{2,d} = h_{-2,0,-2,0}^{2,d} = \hbar_3, \quad h_{0,2,0,2}^{2,d} = \zeta^d \hbar_4 \quad \text{and} \quad h_{0,-2,0,-2}^{2,d} = \zeta^{-d} \hbar_4.$$

Proof Firstly, for the purpose of brevity, let

$$N_{X,w}(\Gamma, \Lambda) := \widehat{D}_{X,w}(e^{\Gamma(2)+\Lambda(3)})e^{\frac{1}{2}\mathcal{Q}(\Gamma)+\mathcal{Q}(\Lambda)}.$$

Proposition 4.81 implies that $N_{B(2)\#_{f_2}B(2),df'_2+d'f_2}(\Gamma, \Lambda)$, for $\Gamma, \Lambda \in \text{im}(\#)$, has the form

$$\sum M_{a,b,\gamma,\eta}^{d,d'}(i, j, i', j')e^{\left(\frac{\sqrt{3}}{2}((K_i\#K_{i'}+K_j\#K_{j'}+2\gamma f_2)\cdot\Gamma)+\frac{\sqrt{3}}{2}i((K_i\#K_{i'}-K_j\#K_{j'}+2\eta f_2)\cdot\Lambda)\right)},$$

where the sum is over the pairs $(a, b), (\gamma, \eta) \in \mathcal{C}_2 \setminus \{(0, 0)\}$ and the basic classes $K_i, K_j, K_{i'}$ and $K_{j'}$ of the permissible triple $(B(2), df' + 2d'f_2, f_2)$ such that

$$\frac{1}{2}(K_i + K_j) \cdot f_2 = \frac{1}{2}(K_{i'} + K_{j'}) \cdot f_2 = a, \quad \frac{1}{2}(K_i - K_j) \cdot f_2 = \frac{1}{2}(K_{i'} - K_{j'}) \cdot f_2 = b.$$

In the above expression, the constant $M_{a,b,\gamma,\eta}^{d,d'}(i, j, i', j')$ is equal to $h_{a,b,\gamma,\eta}^{2,d} c_{ij} c_{i'j'}$, where c_{ij} and $c_{i'j'}$ are the coefficients associated to the pairs of basic classes (K_i, K_j) and $(K_{i'}, K_{j'})$ for the permissible triple $(B(2), df' + 2d'f_2, f_2)$. We need the following elementary lemma:

Lemma 4.86 Suppose V is a vector space and $\{f_i\}_{1 \leq i \leq N}$ is a finite set of distinct complex-valued linear functionals on V . Then the functions $\{e^{f_i}\}_{1 \leq i \leq N}$ are linearly independent over $\mathbb{C}$.

This lemma is an immediate consequence of the existence of a line $l \subseteq V$ such that the restrictions $f_i|_l$ are distinct. We apply this lemma in the case that V is the following subspace of $H_2(B(2)\#_{f_2}B(2))^{\oplus 2}$:

$$(\text{im}(\#) \cap (\Phi_2)_*(\text{im}(\#))) \oplus (\text{im}(\#) \cap (\Phi_2)_*(\text{im}(\#))).$$

Since Φ maps $df'_2 + d'f_2$ to $df_2 + d'f'_2$, we have

$$N_{B(2)\#_{f_2}B(2),df'_2+d'f_2}(\Gamma, \Lambda) = N_{B(2)\#_{f_2}B(2),df_2+d'f'_2}((\Phi_2)_*(\Gamma), (\Phi_2)_*(\Lambda))$$

for $(\Gamma, \Lambda) \in V$. This identity implies that

$$(4.87) \quad \sum M_{a,b,\gamma,\eta}^{d,d'}(i, j, i', j')e^{f_{a,b,\gamma,\eta}^{i,j,i',j'}(\Gamma,\Lambda)} - M_{a,b,\gamma,\eta}^{d',d}(i, j, i', j')e^{g_{a,b,\gamma,\eta}^{i,j,i',j'}(\Gamma,\Lambda)} = 0.$$

	f_2	f'_2	E_2^1	E_2^2	σ_2^1	σ_2^2
$L_1^\delta := E_2^1 + E_2^2 + \delta f_2$	2	δ	$\delta - 2$	$\delta - 2$	0	0
$L_2^\delta := -E_2^1 - E_2^2 + \delta f_2$	-2	δ	$\delta + 2$	$\delta + 2$	0	0
$L_3^\delta := -E_2^1 + E_2^2 + \delta f_2$	0	δ	$\delta + 2$	$\delta - 2$	0	0
$L_4^\delta := E_2^1 - E_2^2 + \delta f_2$	0	δ	$\delta - 2$	$\delta + 2$	0	0
$L_5^\delta := \zeta_1 - \zeta_2 + \delta f_2$	0	δ	δ	δ	0	0
$L_6^\delta := \zeta_2 - \zeta_1 + \delta f_2$	0	δ	δ	δ	0	0
$J_1^\delta := \sigma_2^1 + \sigma_2^2 + \delta f'_2$	δ	2	0	0	$\delta - 2$	$\delta - 2$
$J_2^\delta := -\sigma_2^1 - \sigma_2^2 + \delta f'_2$	δ	-2	0	0	$\delta + 2$	$\delta + 2$
$J_3^\delta := -\sigma_2^1 + \sigma_2^2 + \delta f'_2$	δ	0	0	0	$\delta + 2$	$\delta - 2$
$J_4^\delta := \sigma_2^1 - \sigma_2^2 + \delta f'_2$	δ	0	0	0	$\delta - 2$	$\delta + 2$
$J_5^\delta := \xi_1 - \xi_2 + \delta f'_2$	δ	0	0	0	δ	δ
$J_6^\delta := \xi_2 - \xi_1 + \delta f'_2$	δ	0	0	0	δ	δ

Table 1: Pairing of the cohomology classes L_m^δ and J_m^δ with some elements of $(\text{im}(\#) \cap (\Phi_2)_*(\text{im}(\#)))$.

Here $f_{a,b,\gamma,\eta}^{i,j,i',j'}$ is defined by the pair of cohomology classes

$$\left(\frac{\sqrt{3}}{2} (K_i \# K_{i'} + K_j \# K_{j'} + 2\gamma f_2), \frac{\sqrt{3}}{2} i (K_i \# K_{i'} - K_j \# K_{j'} + 2\eta f_2) \right),$$

and $g_{a,b,\gamma,\eta}^{i,j,i',j'} = \Phi_2^*(f_{a,b,\gamma,\eta}^{i,j,i',j'})$. In the case of $f_{a,b,\gamma,\eta}^{i,j,i',j'}$, this pair is equal to

$$\left(\frac{\sqrt{3}}{2} (L_m^\delta + L_{m'}^{\delta'}), \frac{\sqrt{3}}{2} i (L_m^\delta - L_{m'}^{\delta'}) \right)$$

for an appropriate choice of m and δ , where L_m^δ is as defined in Table 1.

There are similar formulas for $g_{a,b,\gamma,\eta}^{i,j,i',j'}$, where L_m^δ and $L_{m'}^{\delta'}$ are replaced by J_m^δ and $J_{m'}^{\delta'}$. In Table 1, the cohomology classes ζ_1 and ζ_2 are equal to $E^1 \# E^2$ and $E^2 \# E^1$. Moreover, $\xi_i := \Phi_2^*(\zeta_i)$. The table contains evaluations of L_m^δ and J_m^δ at some elements of $\text{im}(\#) \cap (\Phi_2)_*(\text{im}(\#))$. The evaluations of Table 1 show that the only possible identities among the functionals

$$L_i^\delta \quad \text{and} \quad J_i^\delta \quad \text{for } 1 \leq i \leq 4$$

on V are the identities of the elements in the pairs (L_1^2, J_1^2) and (L_2^{-2}, J_2^{-2}) . Therefore, Lemma 4.86 can be used to prove the first part of the proposition. The second part is a

consequence of the remaining information in (4.87), the identity (4.74) and rationality of polynomial invariants. $\square$

Proposition 4.88 *The constant number $\hbar_3$ is nonzero.*

Proof Recall that the 4-manifold $X(m, n)$ is a branched double cover of $W(m, n)$ which is the blowup of $\mathbb{CP}^1 \times \mathbb{CP}^1$ at $4mn$ points. Suppose $\pi: X(m, n) \rightarrow W(m, n)$ is the covering map. Since π does not contract any curve, the pullback of any ample divisor is still ample by the Nakai–Moishezon criterion and projection formula [43]. In particular, the divisor

$$\pi^* \left(\{p\} \times \mathbb{CP}^1 + \mathbb{CP}^1 \times \{q\} - \sum_i E_i \right) = f_{n-1} + f_{m-1} - \sum_{i=1}^{4mn} \pi^*(E_i),$$

where $\{E_i\}_{1 \leq i \leq 4mn}$ is the set of exceptional classes, is ample.

Next we focus on the 4-manifold $X(3, 4)$. Suppose $w_1 = f_3$ and $w_2 = f_3 + f_2$. Then there is a holomorphic line bundle L_i on $X(3, 4)$ such that $c_1(L_i)$ is represented by w_i and w_i can be decomposed as $w_i^1 \# w_i^2$ with respect to the decomposition of $X(3, 4)$ as $E(3) \#_{f_2} E(3)$. Let S denote the ample class $f_2 + f_3 - \sum_{i=1}^{48} \pi^*(E_i)$. According to Theorem 2.19, the coefficient of t_2^k in the series $\hat{D}_{X(3,4),w}(e^{S(2)}) \in \mathbb{C}[[t_2]]$ is positive for large values of k . Since $X(3, 4) = E(3) \#_{f_2} E(3)$ and the homology class S lies in the image of $\#$, we can use Proposition 4.84 to show that

$$\begin{aligned} \hat{D}_{X(3,4),w_i}(e^{S(2)}) \\ = e^{\frac{1}{2}Q(S)} \left[\frac{1}{2} \hbar_1^2 \hbar_3 \cosh(\sqrt{3}(f_3 + 2f_2) \cdot S) + 2\hbar_2^2 \hbar_4 \cos\left(-\frac{2}{3}\pi w_i \cdot (f_3 + 2f_2)\right) \right]. \end{aligned}$$

Therefore

$$2\hat{D}_{X(3,4),w_1}(e^{S(2)}) + \hat{D}_{X(3,4),w_2}(e^{S(2)}) = \frac{3}{2} \hbar_1^2 \hbar_3 e^{\frac{1}{2}Q(S)} \cosh(\sqrt{3}(f_3 + 2f_2) \cdot S).$$

This implies that $\hbar_3$ is a positive number (and $\hbar_1$ is nonzero). $\square$

Proposition 4.89 *The power series $h_{a,b}^{g,d}(t_2, t_3)$ is zero, unless*

$$(a, b) \in \{(\pm(2g-2), 0), (0, \pm(2g-2))\}.$$

Furthermore, we have

$$\begin{aligned} h_{\pm(2g-2),0}^{g,d}(t_2, t_3) &= \hbar_3^{g-1} \left(\frac{2}{\hbar_1} \right)^{2g-4} e^{\pm 2\sqrt{3}t_2}, \\ h_{0,\pm(2g-2)}^{g,d}(t_2, t_3) &= \hbar_4^{g-1} \hbar_2^{4-2g} \zeta^{\pm \frac{2}{3}\pi d} e^{\pm 2\sqrt{3}it_3}. \end{aligned}$$

Proof This theorem can be proved by exploiting the diffeomorphism in (4.79) for $n=g$. Let the 2-cycle w in $B(g) \#_{f_g} B(g)$ be equal to $df'_2 + df_g$. Using Propositions 4.71 and 4.78, we can show that $N_{B(g) \#_{f_g} B(g), w}(\Gamma, \Lambda)$, for $\Gamma, \Lambda \in \mathcal{H} := \text{im}(\#)$, is equal to

$$(4.90) \quad \sum_{(a,b) \in \mathcal{C}_g \setminus \mathcal{C}_{g-1}} h_{a,b}^{g,d}(t_2, t_3) \sum c_{ij} c_{i'j'} e^{(\frac{\sqrt{3}}{2}(M_{i,i'} + M_{j,j'}) \cdot \Gamma + \frac{\sqrt{3}}{2} \mathbf{i}(M_{i,i'} - M_{j,j'}) \cdot \Lambda)},$$

where the second sum is over the pairs of basic classes (K_i, K_j) and $(K_{i'}, K_{j'})$ of the permissible triple $(B(g), df' + 2df_g, f_g)$ such that

$$(4.91) \quad \begin{aligned} \frac{1}{2}(K_i + K_j) \cdot f_g &= \frac{1}{2}(K_{i'} + K_{j'}) \cdot f_g = a, \\ \frac{1}{2}(K_i - K_j) \cdot f_g &= \frac{1}{2}(K_{i'} - K_{j'}) \cdot f_g = b, \end{aligned}$$

$$M_{i,i'} = K_i \# K_{i'} \text{ and } M_{j,j'} = K_j \# K_{j'}.$$

Recall that the 4-manifolds $B(g) \#_{f_g} B(g)$ and $B(2) \#_{f_2} \cdots \#_{f_2} B(2)$ are diffeomorphic to each other using the diffeomorphism Φ_g . Therefore, $N_{B(g) \#_{f_g} B(g), w}(\Gamma, \Lambda)$ can be also computed by regarding $B(g) \#_{f_g} B(g)$ as the fiber sum of g copies of $B(2)$ along surfaces of genus 2. In particular, Propositions 4.81 and 4.84 allow us to obtain the following explicit form for $N_{B(g) \#_{f_g} B(g), w}(\Gamma, \Lambda)$:

$$(4.92) \quad \begin{aligned} \hbar_3^{g-1} \left(\frac{1}{36}\right)^g e^{\sqrt{3}M_g \cdot \Gamma} + \hbar_3^{g-1} \left(\frac{1}{36}\right)^g e^{-\sqrt{3}M_g \cdot \Gamma} + \hbar_4^{g-1} \left(\frac{1}{9}\zeta^d\right)^g e^{\sqrt{3}\mathbf{i}M_g \cdot \Lambda} \\ + \hbar_4^{g-1} \left(\frac{1}{9}\zeta^{-d}\right)^g e^{-\sqrt{3}\mathbf{i}M_g \cdot \Lambda}, \end{aligned}$$

where $M_g = \sigma_g^1 + \sigma_g^2 + 2(g-1)f'_2$. However, this approach works for the homology classes $\Gamma, \Lambda \in \mathcal{H}'$ where $\mathcal{H}'$ is the image of the iterated applications of $\#$ using the decomposition of $B(g) \#_{f_g} B(g)$ as the fiber sum of g copies of $B(2)$. Therefore, (4.90) and (4.92) are equal to each other for $\Gamma, \Lambda \in \mathcal{H} \cap \mathcal{H}'$. Fix $\Gamma, \Lambda \in \mathcal{H} \cap \mathcal{H}'$ and let

$$l := \{\Gamma + sf_g \mid s \in \mathbb{C}\} \quad \text{and} \quad l' := \{\Lambda + sf_g \mid s \in \mathbb{C}\}.$$

Applying Lemma 4.86 to the subspace $l \oplus l'$ of $(\mathcal{H} \cap \mathcal{H}')^{\oplus 2}$ shows that

$$\begin{aligned} h_{\pm(2g-2),0}^{g,d}(t_2, t_3) \left(\frac{1}{36}\right)^g \left(\frac{1}{2}\hbar_1\right)^{2g-4} e^{\pm\sqrt{3}M'_g \cdot \Gamma} &= \hbar_3^{g-1} \left(\frac{1}{36}\right)^g e^{\pm\sqrt{3}M_g \cdot \Gamma}, \\ h_{0,\pm(2g-2)}^{g,d}(t_2, t_3) \left(\frac{1}{9}\right)^g \hbar_2^{2g-4} \zeta^{\mp(2g-2)d} e^{\pm\sqrt{3}\mathbf{i}M'_g \cdot \Lambda} &= \hbar_4^{g-1} \left(\frac{1}{9}\zeta^{\pm d}\right)^g e^{\pm\sqrt{3}\mathbf{i}M_g \cdot \Lambda}, \end{aligned}$$

where $M'_g := (g-2)f'_2 + E_2^1 + \cdots + E_2^g$. Since $\hbar_1$ and $\hbar_2$ are nonzero (work in progress), the above identities prove the second part of the proposition. If $(a, b) \notin \{(\pm(2g-2), 0), (0, \pm(2g-2))\}$, then the same argument shows that

$$h_{a,b}^{g,d}(t_2, t_3) \sum c_{ij} d_{i'j'} e^{(\frac{\sqrt{3}}{2}(M_{i,i'} + M_{j,j'}) \cdot \Gamma + \frac{\sqrt{3}}{2} \mathbf{i}(M_{i,i'} - M_{j,j'}) \cdot \Lambda)} = 0$$

for $\Gamma, \Lambda \in \mathcal{H} \cap \mathcal{H}'$. This sum is over the pairs that satisfy (4.91). Another application of Lemma 4.86 for the homology classes

$$\Gamma = s(E_2^1 \pm \cdots \pm E_2^g) \quad \text{and} \quad \Lambda = s(E_2^1 \pm \cdots \pm E_2^g)$$

in $\mathcal{H} \cap \mathcal{H}'$ shows that $h_{a,b}^{g,d}(t_2, t_3)$ has to be zero. $\square$

The following theorem summarizes our results in this subsection:

Theorem 4.93 *Suppose (X, w, Σ) and (X', w', Σ) are two permissible triples and the genus of Σ is at least 2. Then the triple $(X \#_{\Sigma} X', w \# w', \Sigma)$ is permissible with respect to the image of the map $\#: \mathcal{K} \rightarrow H_2(X \#_{\Sigma} X')$. The basic classes for $(X \#_{\Sigma} X', w \# w', \Sigma)$ are*

$$(4.94) \quad M_{i,i'}^{\gamma} = K_i \# L_{i'} + 2\gamma\Sigma,$$

where K_i and $L_{i'}$ are basic classes of (X, w, Σ) and (X', w', Σ) , $K_i \cdot \Sigma = L_{i'} \cdot \Sigma$, and

$$(K_i \cdot \Sigma, \gamma) \in \{(2g-2, 1), (-(2g-2), -1)\}.$$

For a pair of basic classes $M_{i,i'}^{\gamma} = K_i \# L_{i'} + 2\gamma\Sigma$ and $M_{j,j'}^{\eta} = K_j \# L_{j'} + 2\eta\Sigma$, let

$$\begin{aligned} \frac{1}{2}(K_i + K_j) \cdot \Sigma &= \frac{1}{2}(L_{i'} + L_{j'}) \cdot \Sigma = a, \\ \frac{1}{2}(K_i - K_j) \cdot \Sigma &= \frac{1}{2}(L_{i'} - L_{j'}) \cdot \Sigma = b. \end{aligned}$$

Then the coefficient associated to this pair is equal to $c_{i,j} d_{i',j'} h_{a,b,\gamma,\eta}^{g,d}$, where $c_{i,j}$ is the coefficient associated to (K_i, K_j) for the triple (X, w, Σ) , $d_{i',j'}$ is the coefficient associated to $(L_{i'}, L_{j'})$ for the triple (X', w', Σ) , and $h_{a,b,\gamma,\eta}^{g,d}$ is nonzero in the cases

$$\begin{aligned} h_{(2g-2),0,1,0}^{g,d} &= h_{-(2g-2),0,-1,0}^{g,d} = \hbar_3^{g-1} \left(\frac{2}{\hbar_1} \right)^{2g-4}, \\ h_{0,(2g-2),0,1}^{g,d} &= \frac{\hbar_4^{g-1}}{\hbar_2^{2g-4}} \zeta^d \quad \text{and} \quad h_{0, -(2g-2),0,-1}^{g,d} = \frac{\hbar_4^{g-1}}{\hbar_2^{2g-4}} \zeta^{-d}. \end{aligned}$$

5 Sutured Floer homology

5.1 Eigenvectors

For arbitrary N , we introduce a set of generators of the algebra $\mathbb{V}_{g,d}^N$ in Corollary 3.22. In particular, in the special case that $N = 3$ and $d \equiv 1$ or $2 \pmod{3}$, we have the

following generators of $\mathbb{V}_{g,d}^3$ (which will be denoted by $\mathbb{V}_{g,d}$ from now on):

$$(5.1) \quad \begin{aligned} \epsilon &= D_{\Delta_g, \delta_{g,d} + \Sigma}(1), & \mathfrak{N}_r &= D_{\Delta_g, \delta_{g,d}}(a_r), \\ o_r^j &= D_{\Delta_g, \delta_{g,d}}(l_{(r)}^j), & \rho_r &= D_{\Delta_g, \delta_{g,d}}(\Sigma_{(r)}), \end{aligned}$$

where $r = 2, 3$ and $\{l^j\}_{1 \leq j \leq 2g}$ is a set of generators for $H_1(\Sigma, \mathbb{Z})$. If y is one of the above elements, then the product $m(\cdot, y)$ defines an operator on $\mathbb{V}_{g,d}$, which is still denoted by y . Recall that there is also a pairing on $\mathbb{V}_{g,d}$, which is denoted by $\langle \cdot, \cdot \rangle$.

The operator ϵ is equal to $I_*(Y_g \times [0, 1], \gamma_{g,d} \times [0, 1] + \Sigma, 1)$ and the remaining operators can be described as

$$(5.2) \quad I_*(Y_g \times [0, 1], \gamma_{g,d} \times [0, 1], q),$$

where $q = a_r, l_{(r)}^j$ or $\Sigma_{(r)}$. This alternative description allows us to extend the definition of these operators to arbitrary admissible pairs. Suppose (Y, γ) is a 3–admissible pair, and Σ is an embedded surface of genus g in Y . We also assume that an integral basis $\{l^j\}_{1 \leq j \leq 2g}$ for $H_1(\Sigma)$ is fixed. By replacing $(Y_g, \gamma_{g,d})$ with (Y, γ) , we can define analogues of the operators $\epsilon, \mathfrak{N}_r, o_r^j$ and ρ_r on $I_*(Y, \gamma)$. These operators on $I_*(Y, \gamma)$ are denoted by $\epsilon(\Sigma), \mathfrak{N}_r, o_r^j(\Sigma)$ and $\rho_r(\Sigma)$. When the choice of Σ is clear from our discussion, we drop Σ from our notation for $\epsilon(\Sigma), o_r^j(\Sigma)$ and $\rho_r(\Sigma)$.

Definition 5.3 An element $v \in \mathbb{V}_{g,d}$ is called an *exhaustive eigenvector* if it is a simultaneous eigenvector of the action of the operators in (5.1). An *exhaustive eigenspace* is the set of all exhaustive eigenvectors which have the same eigenvalues with respect to these operators. An exhaustive eigenvector v is called *nondegenerate* if $\langle v, v \rangle \neq 0$.

Remark 5.4 Since $(o_r^j)^2 = 0$, the only eigenvalue of this operator is zero.

Suppose (X, w, Σ) is a permissible triple, Σ is a surface of genus g , and $w \cdot \Sigma = d$. For a pair of basic classes (K_i, K_j) of this triple, let $c_{i,j}$ denote the associated coefficient. Suppose also for a fixed $\lambda = (\alpha, \beta) \in \mathcal{C}_g$ that

$$(5.5) \quad \sum_{(K_i, K_j)} c_{i,j} \neq 0,$$

where the sum is over all pairs of basic classes (K_i, K_j) that satisfy the following equalities for $(a, b) = (\alpha, \beta)$:

$$(5.6) \quad \frac{1}{2}(K_i + K_j) \cdot \Sigma = a \quad \text{and} \quad \frac{1}{2}(K_i - K_j) \cdot \Sigma = b.$$

Recall that P_λ is the polynomial that satisfies (4.64). Suppose $Q_\lambda \in \mathbb{C}[x, y, z]$ is defined as the evaluation of $P_{(a,b)}$ at $t_2 = t_3 = 0$ and consider the following element of $\mathbb{V}_{g,d}$:

$$v_{(\alpha,\beta)} := D_{X^\circ, w^\circ, \Sigma}(Q_\lambda(a_2, \Sigma_{(2)}, \Sigma_{(3)})).$$

Proposition 5.7 *The element $v_{(\alpha,\beta)} \in \mathbb{V}_{g,d}$ is a nonzero exhaustive eigenvector. The eigenvalues of $v_{(\alpha,\beta)}$ with respect to the actions of ϵ , $\mathfrak{N}_2$, $\mathfrak{N}_3$, ρ_2 and ρ_3 are respectively equal to 1, $3\zeta^{2d\beta}$, 0, $\zeta^{d\beta}\sqrt{3}\alpha$ and $\zeta^{2d\beta}\sqrt{3}i\beta$. Furthermore, if $(\alpha, \beta) = (\pm(2g-2), 0)$, then the eigenvector $v_{(\alpha,\beta)}$ is nondegenerate.*

Proof Lemma 4.59 can be used to show that for an arbitrary polynomial $P \in \mathbb{C}[x, y, z]$,

$$(5.8) \quad D_{X,w,\Sigma}(P(a_2, \Sigma_{(2)}, \Sigma_{(3)})) \\ = \sum_{(a,b) \in \mathcal{C}_g} \zeta^{dbd_w} P(3\zeta^{2db}, \zeta^{bd}\sqrt{3}a, \zeta^{2bd}\sqrt{3}ib) \sum_{(K_i, K_j)} c_{i,j},$$

where the inner sum is over all pairs of basic classes (K_i, K_j) that satisfy (5.6) and $c_{i,j}$ is the coefficient associated to the pair (K_i, K_j) . Functorial properties of Floer homology imply that the pairing of $v_{(\alpha,\beta)}$ and $D_{\Delta_g, \delta_{g,d}}(1)$ is equal to $D_{X,w,\Sigma}(Q_\lambda(a_2, \Sigma_{(2)}, \Sigma_{(3)}))$, which is nonzero. Therefore, $v_{(\alpha,\beta)}$ is a nonzero vector.

Suppose y is one of the elements of $\mathbb{V}_{g,d}^3$ in (5.1). Proposition 3.19 and the nondegeneracy of the pairing on $\mathbb{V}_{g,d}$ (Proposition 3.23) imply that $v_{(\alpha,\beta)}$ is an eigenvector of $m(\cdot, y)$ with eigenvalue c_y if and only if

$$(5.9) \quad \langle m(v_{(\alpha,\beta)}, y), D_{\Delta_g, \delta_{g,d}+i\Sigma}(z) \rangle = c_y \langle v_{(\alpha,\beta)}, D_{\Delta_g, \delta_{g,d}+i\Sigma}(z) \rangle$$

for any $z \in \mathbb{A}_g^3$ and $0 \leq i \leq 2$. Another application of the functorial properties of Floer homology shows that both of the expressions in (5.9) can be written as $U(3)$ -invariants of (X, w) . In particular, (5.8) shows that $v_{(\alpha,\beta)}$ is an eigenvector of ϵ , $\mathfrak{N}_2$, ρ_2 and ρ_3 . The vector $v_{(\alpha,\beta)}$ is in the kernel of the operators $\mathfrak{N}_3$ and σ_r^j because X has w -simple type and $b_1(X) = 0$. The pairing $\langle v_{(\alpha,\beta)}, v_{(\alpha,\beta)} \rangle$ can be also computed by Theorem 4.93. Using Proposition 4.88, this number is nonzero for $(\alpha, \beta) = (\pm(2g-2), 0)$. $\square$

Example 4.53 gives a permissible triple such that the condition in (5.5) is satisfied for any $\lambda \in \mathcal{C}_g$. Therefore, for each λ there is an exhaustive eigenvector in $\mathbb{V}_{g,d}$. The condition in (5.5) is not very essential in constructing such an eigenvector. This

condition is used to show that $v_{(\alpha,\beta)}$ is a nonzero element of $\mathbb{V}_{g,d}$. It is possible to replace $v_{(\alpha,\beta)}$ with the element

$$v'_{(\alpha,\beta)} := D_{X^\circ, w^\circ, \Sigma}(Q_\lambda(a_2, \Sigma_{(2)}, \Sigma_{(3)})z),$$

where $z \in \mathbb{A}(X^\circ)^{\otimes 2}$. If the triple (X, w, Σ) has at least one pair of basic classes (K_i, K_j) which satisfy (5.6) for $(a, b) = (\alpha, \beta)$, then z can be chosen such that the above element of $\mathbb{V}_{g,d}$ is nonzero.

Proposition 5.10 *An exhaustive eigenspace is 1-dimensional.*

Proof Suppose $V \subset \mathbb{V}_{g,d}$ is an exhaustive eigenspace, and s_1, s_2, s_3, s_4 and s_5 are respectively the corresponding eigenvalues of $\epsilon, \mathfrak{N}_2, \mathfrak{N}_3, \rho_2$ and ρ_3 . Suppose also that $\mathfrak{I} \subset \mathbb{V}_{g,d}$ is the ideal generated by the elements of the set

$$(5.11) \quad G := \{\epsilon - s_1, \mathfrak{N}_2 - s_2, \mathfrak{N}_3 - s_3, \rho_2 - s_4, \rho_3 - s_5\} \cup \{o_r^j \mid 1 \leq j \leq 2g, 2 \leq r \leq 3\}.$$

Then an element of $\mathfrak{I}$ is the sum of the elements of the form $m(x, y)$ with $x \in G$ and $y \in \mathbb{V}_{g,d}$. For any $v \in V$,

$$\langle v, m(x, y) \rangle = \langle m(v, x), y \rangle = 0.$$

Therefore, V is orthogonal to $\mathfrak{I}$. Since $\mathbb{V}_{g,d}/\mathfrak{I}$ is a 1-dimensional vector space and the pairing is nondegenerate, the dimension of the vector space V is at most 1. $\square$

Lemma 5.12 *Suppose that $v \in \mathbb{V}_{g,d}$ is a nondegenerate exhaustive eigenvector, and s_1, s_2, s_3, s_4 and s_5 are respectively the corresponding eigenvalues of $\epsilon, \mathfrak{N}_2, \mathfrak{N}_3, \rho_2$ and ρ_3 . Then the space*

$$H := \ker_{\text{gen}}(\epsilon - s_1) \cap \ker_{\text{gen}}(\mathfrak{N}_2 - s_2) \cap \ker_{\text{gen}}(\mathfrak{N}_3 - s_3) \cap \ker_{\text{gen}}(\rho_2 - s_4) \cap \ker_{\text{gen}}(\rho_3 - s_5)$$

is 1-dimensional. Here $\ker_{\text{gen}}(T)$, for an operator T , is the union of the kernel of the operators T^k for all values of k .

Proof Suppose the claim does not hold and $v' \in H$ is a vector which is linearly independent of v . Let G be defined as in (5.11). All the operators involved in the definition of H have even degree with respect to the $\mathbb{Z}/2\mathbb{Z}$ -grading of $\mathbb{V}_{g,d}$ induced by the $\mathbb{Z}/12\mathbb{Z}$ -grading. Therefore, H can be decomposed as $H_0 \oplus H_1$ with respect to the $\mathbb{Z}/2\mathbb{Z}$ -grading of $\mathbb{V}_{g,d}$, and we may assume that $v' \in H_i$ for $i = 0$ or 1 . By Proposition 5.10, there is an $x_0 \in G$ such that $m(v', x_0) \neq 0$. Since the restrictions of the elements of G on H are nilpotent, we can also assume, without loss of generality,

that the product of $m(v', x_0)$ and any element $x \in G$ is zero. Therefore, $m(v', x_0) = cv$ for a nonzero complex number c . This implies that

$$\langle m(v', x_0), m(v', x_0) \rangle = \langle cv, cv \rangle \neq 0.$$

On the other hand, we have

$$\langle m(v', x_0), m(v', x_0) \rangle = \pm \langle m(m(v', x_0), x_0), v' \rangle = 0,$$

which is a contradiction. $\square$

Proposition 5.13 [56, Proposition 7.2] Suppose (Y, γ) is a 3-admissible pair and Σ is an embedded surface in Y of genus g such that $\gamma \cdot \Sigma \equiv d \pmod{3}$. If $(s_1, s_2, s_3, s_4, s_5)$ is a simultaneous eigenvalue of the operators $(\epsilon(\Sigma), \mathfrak{N}_2, \mathfrak{N}_3, \rho_2(\Sigma), \rho_3(\Sigma))$, then it is also a simultaneous eigenvalue of the operators $(\epsilon, \mathfrak{N}_2, \mathfrak{N}_3, \rho_2, \rho_3)$ acting on $\mathbb{V}_{g,d}$.

By Proposition 5.7, $v_{2g-2,0}$ is a nondegenerate exhaustive eigenvector of $\mathbb{V}_{g,d}$. Suppose $s_1^g, s_2^g, s_3^g, s_4^g$ and s_5^g denote the corresponding eigenvalues of $\epsilon, \mathfrak{N}_2, \mathfrak{N}_3, \rho_2$ and ρ_3 . Then $s_1^g = 1, s_2^g = 3, s_3^g = 0, s_4^g = \sqrt{3}(2g-2)$ and $s_5^g = 0$. Following [56], we can define a variation of instanton Floer homology:

Definition 5.14 Suppose (Y, γ) is a 3-admissible pair and Σ is an embedded surface in Y of genus g such that $\gamma \cdot \Sigma \equiv d \pmod{3}$. Then $I_*(Y, \gamma|\Sigma)$ is defined as

$$\ker_{\text{gen}}(\epsilon(\Sigma) - s_1^g) \cap \ker_{\text{gen}}(\mathfrak{N}_2 - s_2^g) \cap \ker_{\text{gen}}(\mathfrak{N}_3 - s_3^g) \cap \ker_{\text{gen}}(\rho_2(\Sigma) - s_4^g) \\ \cap \ker_{\text{gen}}(\rho_3(\Sigma) - s_5^g).$$

In this definition, we allow Σ to have more than one connected component. In that case, each connected component Σ' of Σ is required to have genus g and $\gamma \cdot \Sigma' \equiv d \pmod{3}$. In the definition of $I_*(Y, \gamma|\Sigma)$, we include the operators $\epsilon(\Sigma'), \rho_2(\Sigma') - s_4^g$ and $\rho_3(\Sigma') - s_5^g$ for each connected component Σ' of Σ .

Remark 5.15 In the case that $g = 1$, the actions of the operators $\mathfrak{N}_3, \rho_2$ and ρ_3 on $\mathbb{V}_{1,d}$ are trivial, the operator $\mathfrak{N}_2$ is equal to $3\epsilon^{-1}$ and $\epsilon^3 = 1$. Therefore, similar relationships hold among the operators $\epsilon(\Sigma), \mathfrak{N}_2, \mathfrak{N}_3, \rho_2(\Sigma), \rho_3(\Sigma)$ acting on $I_*(Y, \gamma)$, where Σ is a genus one surface in Y . This can be verified in a similar way as in Proposition 5.13. (See [56, Proposition 7.2].) Therefore, $I_*(Y, \gamma|\Sigma)$, for a genus one surface Σ , is simply equal to $\ker(\mathfrak{N}_2 - 3)$.

This variant of instanton Floer homology is also functorial with respect to cobordisms. Suppose $(W, w): (Y_0, \gamma_0) \rightarrow (Y_1, \gamma_1)$ is a cobordism of admissible pairs, $z \in \mathbb{A}(W)^{\otimes 2}$, Σ_i is an embedded oriented and connected surface in Y_i such that $\Sigma_i \cdot \gamma_i \equiv d \pmod{3}$, and Σ_0 and Σ_1 induce the same homology classes of W . More generally, if Σ_1 is disconnected, then each connected component of Σ_1 is required to be homologous to a connected component of Σ_0 inside W . Properties of instanton Floer homology discussed in [Section 3.1](#) imply that $I_*(W, w, z)$ maps $I_*(Y_0, \gamma_0 | \Sigma_0) \subseteq I_*(Y_0, \gamma_0)$ to $I_*(Y_1, \gamma_1 | \Sigma_1) \subseteq I_*(Y_1, \gamma_1)$. Moreover, suppose (X, w) is a cobordism from an admissible pair (Y, γ) to the empty pair and $z \in \mathbb{A}(X)^{\otimes 2}$. Then the restriction of the map $D^{X,w}(z)$ gives rise to a functional on $I_*(Y, \gamma | \Sigma)$, which is denoted with the same notation.

[Lemma 5.12](#) asserts that $I_*(Y_g, \gamma_{g,d} | \Sigma)$ is 1-dimensional. The following pairs from [Section 3.2](#) define cobordisms from two copies of $(Y_g, \gamma_{g,d})$ to the empty pair:

$$(\Omega_g, \omega_{g,d}) \quad \text{and} \quad (\Delta_g \amalg \Delta_g, \delta_{g,d} \amalg \delta_{g,d}).$$

Therefore, they determine two functionals $D^{\Omega_g, \omega_{g,d}}(1)$ and $D^{\Delta_g, \delta_{g,d}}(1) \otimes D^{\Delta_g, \delta_{g,d}}(1)$ on the 1-dimensional vector space $I_*(Y_g, \gamma_{g,d} | \Sigma) \otimes I_*(Y_g, \gamma_{g,d} | \Sigma)$. The nondegeneracy of the exhaustive eigenvector involved in the definition of $I_*(Y_g, \gamma_{g,d} | \Sigma)$ implies that the former functional is nonzero. Therefore, we have the following lemma which provides the distinguishing property of working with a nondegenerate exhaustive eigenspace for us:

Lemma 5.16 *The map*

$$D^{\Delta_g, \delta_{g,d}}(1) \otimes D^{\Delta_g, \delta_{g,d}}(1): I_*(Y_g, \gamma_{g,d} | \Sigma) \otimes I_*(Y_g, \gamma_{g,d} | \Sigma) \rightarrow \mathbb{C}$$

is a multiple of $D^{\Omega_g, \omega_{g,d}}(1): I_*(Y_g, \gamma_{g,d} | \Sigma) \otimes I_*(Y_g, \gamma_{g,d} | \Sigma) \rightarrow \mathbb{C}$.

5.2 Excision and sutured manifolds invariants

Suppose R_1 and R_2 are two embedded surfaces of genus $g \geq 1$ in a 3-manifold Y . Suppose also there is a 1-cycle γ in Y such that $\gamma \cdot R_1 = \gamma \cdot R_2$. We also assume γ intersects R_1 and R_2 transversally, and all the intersection points have the same signs. Fix a diffeomorphism $\phi: R_1 \rightarrow R_2$ such that ϕ maps $\gamma \cap R_1$ to $\gamma \cap R_2$. We cut Y along the surfaces R_1 and R_2 , and then identify the four boundary components of the resulting 3-manifold using ϕ so that the final 3-manifold, $\tilde{Y}$, is an oriented closed 3-manifold with embedded surfaces $\tilde{R}_1$ and $\tilde{R}_2$. Our assumption on ϕ implies that γ

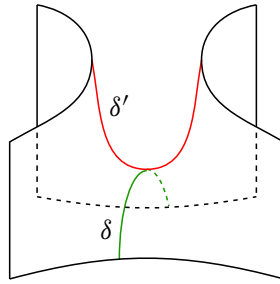


Figure 5: The saddle cobordism $\mathcal{P}$: the union of the two vertical boundary components on the left (respectively, right) is denoted by $\partial_l^v \mathcal{P}$ (respectively, $\partial_r^v \mathcal{P}$).

determines a 1-cycle $\tilde{\gamma}$ in $\tilde{Y}$. We will also write R (respectively, $\tilde{R}$) for the union $R_1 \cup R_2$ (respectively, $\tilde{R}_1 \cup \tilde{R}_2$). Now we are ready to state our excision theorem:

Theorem 5.17 *The following Floer homology groups are isomorphic:*

$$I_*(Y, \gamma | R) = I_*(\tilde{Y}, \tilde{\gamma} | \tilde{R}).$$

Proof This theorem is the analogue of excision theorem for $U(2)$ -instanton Floer homology. The $U(2)$ version of the excision theorem is proved in [28] for $g = 1$ (see also [8]) and in [56] for higher values of g . We follow the same strategy as in the $U(2)$ case to prove the theorem. In particular, the isomorphism between $I_*(Y, \gamma | R)$ and $I_*(\tilde{Y}, \tilde{\gamma} | \tilde{R})$ is induced by a cobordism of pairs $(W, w): (Y, \gamma) \rightarrow (\tilde{Y}, \tilde{\gamma})$. Let Y° be the complement of a regular neighborhood of R in Y . Then the cobordism W is the result of gluing $[0, 1] \times Y^\circ$ to $\mathcal{P} \times R_1$ where $\mathcal{P}$ is the saddle cobordism in Figure 5. The boundary of the 3-manifold Y° is equal to $R_1 \cup \bar{R}_1 \cup R_2 \cup \bar{R}_2$. Then $[0, 1] \times (R_1 \cup \bar{R}_1) \subset [0, 1] \times Y^\circ$ is glued to $\partial_l^v \mathcal{P} \times R_1$ by the identity map and $[0, 1] \times (R_2 \cup \bar{R}_2) \subset [0, 1] \times Y^\circ$ is glued to $\partial_r^v \mathcal{P} \times R_1$ by the map ϕ . (For the definition of $\partial_l^v \mathcal{P}$ and $\partial_r^v \mathcal{P}$ see Figure 5.) The surface cobordism $w: \gamma \rightarrow \tilde{\gamma}$ is also constructed in a similar way. This surface is given by gluing one copy of $\mathcal{P}$ for each intersection point in $\gamma \cap R_1$ to $[0, 1] \times (\gamma \cap Y^\circ)$. Reversing the cobordism (W, w) determines another cobordism $(\bar{W}, \bar{w}): (\tilde{Y}, \tilde{\gamma}) \rightarrow (Y, \gamma)$. To prove the excision theorem, we claim that

$$(5.18) \quad \begin{aligned} I_*(\bar{W}, \bar{w}) \circ I_*(W, w): I_*(Y, \gamma | R) &\rightarrow I_*(Y, \gamma | R), \\ I_*(\bar{W}, \bar{w}) \circ I_*(W, w): I_*(\tilde{Y}, \tilde{\gamma} | \tilde{R}) &\rightarrow I_*(\tilde{Y}, \tilde{\gamma} | \tilde{R}) \end{aligned}$$

are nonzero multiples of the identity map. In the composite cobordism $\bar{W} \circ W: Y \rightarrow Y$, a copy of $\mathcal{P} \times R_1$ is glued to another copy of $\mathcal{P} \times R_1$ with the reverse orientation.

In this part of $\bar{W} \circ W$, the union of the two copies of $\delta \times R_1$ gives rise to a copy of $Y_g := S^1 \times R_1$. The intersection of $\bar{w} \circ w$ with $S^1 \times R_1$ produces a copy of $\gamma_{g,d}$. According to Lemma 5.16 and functoriality of I_* , replacing a neighborhood of $S^1 \times R_1$ with $(\Delta_g \amalg \Delta_g, \delta_{g,d} \amalg \delta_{g,d})$ does not change the map $I_*(\bar{W} \circ W, \bar{w} \circ w)$, up to multiplication by a nonzero constant number. But the resulting cobordism of the pair is the product cobordism $([0, 1] \times Y, [0, 1] \times \gamma)$. Therefore, the first map in (5.18) is a nonzero multiple of the identity map. Replacing δ with δ' and using the same argument proves a similar result for the second map in (5.18). $\square$

The following proposition is the analogue of Corollary 4.8 in [56] and can be proved in a similar way using the excision theorem:

Proposition 5.19 *Let Y be a 3-manifold, γ be a 1-cycle in Y , and $R \subset Y$ be a connected surface of genus $g \geq 1$ such that $\gamma \cdot R \not\equiv 0 \pmod{3}$ and the intersection points of $\gamma \cap R$ are transversal and have the same signs. Let $\tilde{Y}$ be the 3-manifold obtained by cutting Y along R and regluing by an orientation-preserving diffeomorphism $\phi: R \rightarrow R$. Suppose ϕ maps $R \cap \gamma$ to $R \cap \gamma$, and $\tilde{\gamma}$ and $\tilde{R}$ are the induced 1-cycle and the embedded surface in $\tilde{Y}$. Then*

$$I_*(Y, \gamma|R) \cong I_*(\tilde{Y}, \tilde{\gamma}|\tilde{R}).$$

Now we can define invariants for balanced sutured manifolds almost verbatim from [56]. Firstly we recall the definition of balanced sutured manifolds (see [35; 46; 56]):

Definition 5.20 *A sutured manifold (M, α) consists of an oriented 3-manifold M , an oriented closed 1-manifold α in ∂M and a decomposition of ∂M as*

$$(5.21) \quad \partial M = R_+ \cup A(\alpha) \cup R_-.$$

Each connected component of α is called a *suture* and $A(\alpha)$ is the closure of a tubular neighborhood of α . The spaces R_+ and R_- are disjoint and each of them is a union of some of the connected components of $\overline{\partial M} \setminus A(\alpha)$. In particular, each component of ∂R_+ and ∂R_- is a parallel copy of a suture. Suppose R_+ and R_- are oriented with the outward-normal-first convention. Similarly, each component of ∂R_+ (respectively, ∂R_-) inherits an orientation from R_+ (respectively, R_-). This orientation is required to agree (respectively, disagree) with the orientation of the corresponding suture. The sutured manifold (M, α) is *balanced* if neither M nor $R_{\pm}$ has closed components, and $\chi(R_+) = \chi(R_-)$.

Note that the required conditions on R_+ , R_- and $A(\alpha)$ imply that the decomposition (5.21) is unique.

Example 5.22 Suppose $F_{g,k}$ is a surface of genus g with $k \geq 1$ boundary components which is not the 2-dimensional disc. Then $([-1, 1] \times F_{g,k}, \partial F_{g,k} \times \{0\})$ determines a balanced sutured manifold. The decomposition of the boundary of this product sutured manifold is given by

$$(5.23) \quad R_+ = \{1\} \times F_{g,k}, \quad A(\alpha) = [-1, 1] \times \partial F_{g,k}, \quad R_- = \{-1\} \times F_{g,k}.$$

Example 5.24 Suppose Y is a 3-manifold and $K \subset Y$ is a knot. Let $M(K)$ be the complement of a regular neighborhood of K in Y , and $\alpha(K)$ be the union of two oppositely oriented meridional curves on the boundary of $M(K)$. Then $(M(K), \alpha(K))$ determines a balanced sutured manifold. The manifolds R_+ and R_- are homeomorphic to $[0, 1] \times S^1$.

Example 5.25 Suppose that K is a null-homologous knot in Y and S is a Seifert surface for K . We can also associate a sutured manifold $(N(S), \alpha(S))$ to S . The 3-manifold $N(S)$ is defined to be $Y \setminus ((-1, 1) \times \text{int}(S))$, where $(-1, 1) \times S$ is a regular neighborhood of S in Y . The only suture $\alpha(S)$ of this sutured manifold is defined to be $\{0\} \times \partial S$.

Let (M, α) be a balanced sutured manifold such that the number of sutures is greater than one and neither R_+ nor R_- is a union of 2-dimensional discs. We attach the product sutured manifold $([-1, 1] \times F_{0,k}, \{0\} \times \partial F_{0,k})$ to (M, α) where k is the number of sutures of (M, α) . More precisely, we glue M to $[-1, 1] \times F_{0,k}$ by identifying $A(\alpha)$ with $[-1, 1] \times \partial F_{0,k}$ using an orientation-reversing map. The resulting manifold is oriented and has two boundary components,

$$\bar{R}_+ = R_+ \cup \{1\} \times F_{0,k} \quad \text{and} \quad \bar{R}_- = R_- \cup \{-1\} \times F_{0,k}.$$

Since (M, α) is balanced, the oriented surfaces $\bar{R}_+$ and $\bar{R}_-$ have the same positive genus. We choose an orientation-reversing diffeomorphism $\phi: \bar{R}_+ \rightarrow \bar{R}_-$. Identifying $\bar{R}_+$ and $\bar{R}_-$ using the map ϕ determines a closed 3-manifold Y_ϕ , which is called a *closure* of the sutured manifold (M, α) . The 3-manifold Y_ϕ only depends on (M, α) and the choice of the diffeomorphism ϕ . The surface $\bar{R}_+$ also induces an oriented surface in Y_ϕ , which is denoted by $\bar{R}$. We also fix a point y on $F_{0,k}$ and require that ϕ map $(1, y) \in \bar{R}_+$ to $(-1, y) \in \bar{R}_-$. Therefore, the path $[-1, 1] \times \{y\} \subset [-1, 1] \times F_{0,k}$ induces a 1-cycle γ_ϕ in Y_ϕ .

Definition 5.26 The *sutured instanton homology* of the sutured manifold (M, α) is

$$(5.27) \quad \text{SHI}_*(M, \alpha) := I_*(Y_\phi, \gamma_\phi | \bar{R}).$$

If (M, α) has one suture or R_+ (or equivalently R_-) is a union of discs, we replace $F_{0,k}$ with $F_{1,k}$ in the definition of $(Y, \gamma, \bar{R})$ and then extend sutured instanton homology to these sutured manifolds using (5.27).

Proposition 5.19 implies that sutured instanton homology SHI_* is well defined. Notice that our choices imply that the surface $\bar{R}$ in (5.27) has genus at least one. In particular, the required assumption for **Proposition 5.19** is satisfied.

Remark 5.28 In the definition of the closure of a sutured manifold (M, α) , we can replace $F_{0,k}$ with $F_{g,k}$ for an arbitrary g . Then for each choice of g , we can define a sutured Floer homology group as above. Using our excision theorem and the method of [56], we can show that the rank of these sutured Floer homology groups is nondecreasing in g . We expect that these sutured Floer homology groups for various choices of g are isomorphic to each other. However, proving this seems to need a further study of the algebra $\mathbb{V}_{g,1}$. For example, if **Conjecture 7.3** holds, then all of the above versions of sutured Floer homology of (M, α) are isomorphic. We hope to come back to this issue elsewhere.

5.3 Instanton knot homology

Definition 5.29 Given a knot K in a 3-manifold Y , let $(M(K), \alpha(K))$ be the sutured manifold of **Example 5.24**. The $\text{U}(3)$ -*knot homology* of K , denoted by $\text{KHI}_*(Y, K)$, is defined to be $\text{SHI}_*(M(K), \alpha(K))$.

As explained in [56, Lemma 5.2], a closure of $(M(K), \alpha(K))$ can be described as follows. Suppose F is a genus 1 surface with one boundary component, and $c, c' \subset F$ are two oriented nonseparating simple closed curves which intersect in exactly one point and $c \cdot c' = 1$. Let $Z(K)$ be the result of gluing the knot complement $M(K)$ to the product 3-manifold $F \times S^1$ such that the meridian of K is identified with $\{\text{point}\} \times S^1$ and the longitude of K is identified with $\partial F \times \{\text{point}\}$. Suppose also $\gamma(K) \subset Z(K)$ is the 1-cycle given by $c \times \{\text{point}\} \subset F \times S^1$. Then $Z(K)$ is a closure of the sutured manifold $(M(K), \alpha(K))$ and $\gamma(K)$ is the corresponding 1-cycle. The embedded surface $\bar{R}$ is also given by the torus $T = c' \times S^1 \subset F \times S^1$. Consequently,

$$(5.30) \quad \text{KHI}_*(Y, K) = I_*(Z(K), \gamma(K) | T).$$

Next, we characterize the set of the critical points of the Chern–Simons functional associated to the pair $(Z(K), \gamma(K))$:

Proposition 5.31 *For a knot K in a 3-manifold Y , the set of the critical points of the Chern–Simons functional associated to the admissible pair $(Z(K), \gamma(K))$ is a 3-sheeted covering space of*

$$(5.32) \quad \mathcal{R} = \left\{ \rho: \pi_1(Y \setminus K) \rightarrow \mathrm{SU}(3) \mid \rho(\mu) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \zeta & 0 \\ 0 & 0 & \zeta^2 \end{bmatrix} \right\}.$$

Recall that μ is a meridian of K and $\zeta = e^{2\pi i/3}$.

Proof The set of the critical points of the Chern–Simons functional is given by the conjugacy classes of representations $\rho: \pi_1(Z(K) \setminus \gamma(K)) \rightarrow \mathrm{SU}(3)$ such that a meridian of the closed curve $\gamma(K)$ is mapped to ζ . We fix a base point for $Z(K)$ on the torus $c' \times S^1 \subset F \times S^1$. Suppose $J_1 = \rho(c' \times \{\text{point}\})$ and $J_2 = \{\text{point}\} \times S^1$. Since $c' \times S^1$ intersects c in one point, we can assume

$$(5.33) \quad [J_2, J_1] = \zeta \cdot \text{id}.$$

This implies that J_2 and $\zeta^{-1}J_2$ are conjugate to each other. Thus there is a unique representative for the conjugacy class of ρ such that

$$J_1 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad J_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \zeta & 0 \\ 0 & 0 & \zeta^2 \end{bmatrix}.$$

Therefore, the conjugacy class of the representation $\rho|_{\pi_1(F \times S^1 \setminus \gamma(K))}$ is uniquely determined by $J_3 \in \mathrm{SU}(3)$, which is equal to the image of ρ for a parallel copy of c . Since J_3 has to commute with J_2 it is a diagonal matrix. Therefore, the restriction of the above representative of ρ to the knot group $\pi_1(M(K))$ determines an element of $\mathcal{R}$. Furthermore, ρ maps the longitude of K to $[J_3, J_1]$. Now the claim can be easily verified, because the map that sends a diagonal matrix J_3 to $[J_3, J_1]$ is 3-to-1. $\square$

Corollary 5.34 *Suppose K is a knot in a homology sphere Y . If $\dim(\mathrm{KHI}(Y, K)) > 1$, then there exists a nonabelian representation $\rho: \pi_1(Y \setminus K) \rightarrow \mathrm{SU}(3)$ such that the image of the meridian is conjugate to*

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \zeta & 0 \\ 0 & 0 & \zeta^2 \end{bmatrix}.$$

Proof Suppose there is not a representation with this property. Then the only critical points of the Chern–Simons functional of $(Z(K), \gamma(K))$ are the three flat connections induced by the abelian representation in (5.32). Since these critical points are non-degenerate,⁸ $I_*(Z(K), \gamma(K))$ is the homology of a chain complex which has three generators. The order-three map ϵ has degree 4 with respect to the $\mathbb{Z}/12\mathbb{Z}$ grading. Thus the three eigenspaces of this operator have the same dimensions. Therefore, $\ker(\epsilon - 1)$ has to be at most 1-dimensional, which is a contradiction. $\square$

Proposition 5.35 *Let K be a null-homologous knot in Y and S be a Seifert surface of genus $g \geq 1$ for K . Then $\dim(\text{KHI}(Y, K)) \geq 2 \dim(\text{SHI}_*(N(S), \alpha(S)))$, where $(N(S), \alpha(S))$ is the sutured manifold of Example 5.25.*

Proof According to (5.30), the $U(3)$ -knot homology is equal to $I_*(Z(K), \gamma(K)|T)$. In order to form a closure of $(N(S), \alpha(S))$ and to define $\text{SHI}_*(N(S), \alpha(S))$, we first glue $[-1, 1] \times F_{1,1}$ along the suture $\alpha(S)$. In this case, $\bar{R}^\pm$ are two copies of $S \cup F_{1,1}$. If we identify $\bar{R}^\pm$ in the obvious way, then the resulting space is again the 3-manifold $Z(K)$. Let $\gamma(S)$ and $\bar{R}(S)$ be the resulting 1-cycle and the surface in the closure $Z(K)$. The 1-cycle $\gamma(S)$ is a copy of $S^1 \times \{\text{point}\} \subset S^1 \times F_{1,1}$. The surface $\bar{R}(S)$ has genus $g + 1$ and is given by gluing the Seifert surface S to $F_{1,1}$. Arguing as in [56, Proposition 7.9] and using our excision theorem, we can show that

$$I_*(Z(K), \gamma(K)|T) = I_*(Z(K), \gamma(K) + \gamma(S)|T)$$

and

$$I_*(Z(S), \gamma(S)|\bar{R}(S)) = I_*(Z(S), \gamma(K) + \gamma(S)|\bar{R}(S)).$$

Remark 5.15 implies that the homology group $I_*(Z(K), \gamma(K) + \gamma(S)|T)$ is equal to

$$(5.36) \quad I_*(Z(K), \gamma(K) + \gamma(S)) \cap \ker(\mathfrak{S}_2 - 3).$$

We can further decompose the vector space in (5.36) using the eigenvalues of the operators $\rho_2(\bar{R}(S))$. In particular, the vector spaces

$$(5.37) \quad I_*(Z(K), \gamma(K) + \gamma(S)) \cap \ker(\mathfrak{S}_2 - 3) \cap \ker_{\text{gen}}(\rho_2(\bar{R}(S)) - 2\sqrt{3}g)$$

and

$$(5.38) \quad I_*(Z(K), \gamma(K) + \gamma(S)) \cap \ker(\mathfrak{S}_2 - 3) \cap \ker_{\text{gen}}(\rho_2(\bar{R}(S)) + 2\sqrt{3}g)$$

⁸This is a consequence of the fact that the Alexander polynomial of a knot K in an integral homology sphere does not have a root which is a third root of unity.

are two distinct summands of $I_*(Z(K), \gamma(K) + \gamma(S)|T)$. Since the operator $\bar{R}(S)$ has degree 2 with respect to the $\mathbb{Z}/12\mathbb{Z}$ -grading of $I_*(Z(K), \gamma(K) + \gamma(S)|T)$, the vector spaces in (5.37) and (5.38) have the same dimension. It is also clear from the definition that (5.37) contains $I_*(Z(S), \gamma(K) + \gamma(S)|\bar{R}(S))$, which verifies the claim of this proposition. $\square$

Conjecture 5.39 *If $Y \setminus K$ is irreducible, then $\dim(\text{SHI}_*(N(S), \alpha(S))) \geq 1$.*

If $Y \setminus K$ is irreducible, then the sutured manifold $(N(S), \alpha(S))$ is *taut*. Kronheimer and Mrowka proved a nonvanishing theorem for the $U(2)$ -sutured Floer homology SHI_*^2 of taut sutured manifolds [56]. As explained in Section 7, we expect that a similar nonvanishing theorem holds for our version of sutured Floer homology.

If Conjecture 5.39 holds, then the answer to Question 1.1 is positive for $N = 3$ and for any nontrivial knot in an integral homology sphere Y . Using the prime decomposition theorem for 3-manifolds, one can decompose Y as a connected sum $Y_1 \# Y_2$ such that $K \subset Y_1$ and $Y_1 \setminus K$ is irreducible. With the aid of Corollary 5.34 and Proposition 5.35, we can conclude from Conjecture 5.39 that the answer to Question 1.1 for $N = 3$ and the pair (Y_1, K) is positive. This clearly would imply the claim for (Y, K) .

6 Gluing theory

6.1 Moduli spaces on manifolds with long neck

Suppose that Y is a connected 3-manifold and $\gamma \subset Y$ is a cycle. We do not assume that (Y, γ) is N -admissible. However, we assume that the critical points of the (possibly perturbed) Chern–Simons functional of (Y, γ) are nondegenerate. Suppose also that (X, w) is a pair with boundary (Y, γ) . As explained in Section 2.2, we can form moduli spaces $\mathcal{M}_p(X, w)$ and their framed counterparts $\widetilde{\mathcal{M}}_p(X, w)$ by working with perturbations of the ASD equation. The Uhlenbeck compactness theorem of Section 2.1 has an analogue for 4-manifolds with cylindrical ends. A proof of this result for $N = 2$ is given in [19, Chapter 5], and it can be extended to higher rank by combining the arguments of [19] and [52]:

Theorem 6.1 *Suppose $\{[A_i]\}_{i \in \mathbb{N}}$ is a sequence of connections in the moduli space $\widetilde{\mathcal{M}}_p(X, w)$. Then there is an element $([B], [C_1], \dots, [C_k])$ of the space*

$$(6.2) \quad \widetilde{\mathcal{M}}_{p_0}(X, w; \alpha_0) \times_{\Gamma_{\alpha_0}} \widetilde{\mathcal{M}}_{p_1}(\alpha_0, \alpha_1) \times_{\Gamma_{\alpha_1}} \cdots \times_{\Gamma_{\alpha_{k-1}}} \widetilde{\mathcal{M}}_{p_k}(\alpha_{k-1}, \alpha_k)$$

and an element $(\mathbf{x}, y_1, \dots, y_k)$ of the space

$$(6.3) \quad (X^+)^{m_0}/S_{m_0} \times (Y \times \mathbb{R})^{m_1}/S_{m_1} \times \cdots \times (Y \times \mathbb{R})^{m_k}/S_{m_k}$$

for appropriate nonnegative integers m_i such that $\{[A_i]\}_{i \in \mathbb{N}}$, after passing to a sequence, is **weakly chain convergent** to $(([B], \mathbf{x}), ([C_1], y_1), \dots, ([C_k], y_k))$. Furthermore,

$$(6.4) \quad \kappa(p) = \kappa(p_0) + \kappa(p_1) + \cdots + \kappa(p_k) + m_0 + m_1 + \cdots + m_k.$$

Note that $\tilde{\mathcal{M}}_q(\alpha, \beta)$ is the moduli space of framed connections associated to a path $q: \alpha \rightarrow \beta$ over $(Y \times \mathbb{R}, \gamma \times \mathbb{R})$. This moduli space is equipped with an action of $\Gamma_\alpha \times \Gamma_\beta$. The weak chain convergence of $\{[A_i]\}_{i \in \mathbb{N}}$ to $(([B], \mathbf{x}), ([C_1], y_1), \dots, ([C_k], y_k))$ means that the following holds [19]: the sequence $\{[A_i]\}_{i \in \mathbb{N}}$, after choosing appropriate gauge representatives, is L_1^P -convergent to B on compact sets of $X \setminus \mathbf{x}$. Moreover, there is a sequence of real numbers

$$t_i^1 := 0 < t_i^2 < \cdots < t_i^k$$

with $\lim_i t_i^{j+1} - t_i^j = \infty$ such that the translation of $A_i|_{Y \times [0, \infty)}$ by the constant t_i^j is L_1^P -convergent to C_j on compact sets of $Y \times \mathbb{R} \setminus y_j$. Identity (6.4) implies that

$$(6.5) \quad \text{index}(\mathcal{D}_{A_i}) \geq \text{index}(\mathcal{D}_{B_1}) + \text{index}(\mathcal{D}_{C_1}) + \cdots + \text{index}(\mathcal{D}_{C_k})$$

and the two sides of the inequality are also equal to each other mod $4N$. In (6.5), equality holds if and only if the integers m_j are all zero, and in this case, L_1^P convergence on compact subspaces can be improved to C^∞ convergence on compact subspaces.

Remark 6.6 Theorem 6.1 can be extended to the case that W has more than one boundary component in an obvious way. In this case, we need to fix a chain of the elements of moduli spaces for each boundary component.

There is another compactness theorem we need to review, in which we stretch a 4-manifold along a neck and consider the associated moduli spaces. Suppose (X_1, w_1) and (X_2, w_2) are pairs whose boundaries are (Y, γ) and $(\bar{Y}, \bar{\gamma})$, respectively. Then we can glue these pairs to form (X, w) . We also fix Riemannian metrics on the X_i which are product metrics in neighborhoods of their boundaries associated to a fixed metric on Y . Let X^T be the Riemannian manifold diffeomorphic to X which has an isometric copy of $Y \times (-T, T)$ and is such that $X^T \setminus Y \times (-T, T)$ is isometric to the disjoint union of X_1 and X_2 . The proof of the following theorem is similar to that of Theorem 6.1.

Theorem 6.7 Suppose A_i is a connection on the moduli space $\mathcal{M}_\kappa(X^{T_i}, w)$ such that $\lim_{i \rightarrow \infty} T_i = \infty$. Then there is an element $([B_1], [C_1], \dots, [C_k], [B_2])$ of

$$(6.8) \quad \widetilde{\mathcal{M}}_p(X_1, w_1; \alpha_0) \times_{\Gamma_{\alpha_0}} \widetilde{\mathcal{M}}_{p_1}(\alpha_0, \alpha_1) \times_{\Gamma_{\alpha_1}} \cdots \times_{\Gamma_{\alpha_{k-1}}} \widetilde{\mathcal{M}}_{p_k}(\alpha_{k-1}, \alpha_k) \\ \times_{\Gamma_{\alpha_k}} \widetilde{\mathcal{M}}_{p'}(X_2, w_2; \alpha_k)$$

and an element $(\mathbf{x}_1, \mathbf{y}_1, \dots, \mathbf{y}_k, \mathbf{x}_2)$ of the space

$$(6.9) \quad (X_1^+)^{m_0}/S_{m_0} \times (Y \times \mathbb{R})^{m_1}/S_{m_1} \times \cdots \times (Y \times \mathbb{R})^{m_k}/S_{m_k} \times (X_2^+)^{m_{k+1}}/S_{m_{k+1}}$$

such that $\{[A_i]\}_{i \in \mathbb{N}}$ is weakly chain convergent to

$$(([B_1]\mathbf{x}_1), ([C_1], \mathbf{y}_1), \dots, ([C_k], \mathbf{y}_k), ([B_2], \mathbf{x}_2)).$$

Moreover, we have

$$(6.10) \quad \kappa = \kappa(p) + \kappa(p_1) + \cdots + \kappa(p_k) + \kappa(p') + m_0 + m_1 + \cdots + m_k + m_{k+1}.$$

The following gluing theorem can be regarded as an inverse to [Theorem 6.7](#). There are various places in the literature where similar gluing theorems are discussed [[81](#); [19](#); [55](#)]. [Theorem 6.11](#) can be proved with similar strategies (see eg [[19](#), Theorem 4.17 and Section 4.7.1]):

Theorem 6.11 Let (X_i, w_i) be given as above. For $1 \leq i \leq k$, let $p_i: \alpha_{i-1} \rightarrow \alpha_i$ be a path along $(Y \times \mathbb{R}, w \times \mathbb{R})$, and let $\tilde{N}_i$ be a compact $(\Gamma_{\alpha_{i-1}} \times \Gamma_{\alpha_i})$ -invariant subspace of $\widetilde{\mathcal{M}}_{p_i}(\alpha_{i-1}, \alpha_i)$ which consists of regular points. Suppose we are also given two other compact spaces

$$\tilde{N}_0 \subset \widetilde{\mathcal{M}}_p(X_1, w_1; \alpha_0) \quad \text{and} \quad \tilde{N}_{k+1} \subset \widetilde{\mathcal{M}}_{p'}(X_2, w_2; \alpha_k)$$

which contain only regular points and are respectively invariant with respect to the action of Γ_{α_0} and Γ_{α_k} . Here we assume that the perturbations of the ASD equation on the ends of X_1 and X_2 are induced by a fixed perturbation of the Chern–Simons functional of (Y, γ) . Then there is a space $\tilde{U}_i$ containing $\tilde{N}_i$ which is an open subset of the relevant moduli space and is invariant with respect to the action of the relevant group. Moreover, for large enough values of T , there is a **gluing map**

$$\Phi_T: \tilde{U}_0 \times_{\Gamma_{\alpha_0}} \tilde{U}_1 \times_{\Gamma_{\alpha_1}} \cdots \times_{\Gamma_{\alpha_k}} \tilde{U}_{k+1} \rightarrow \mathcal{M}_\kappa(X^T, w),$$

where

$$(6.12) \quad \kappa = \kappa(p) + \kappa(p_1) + \cdots + \kappa(p_k) + \kappa(p').$$

The gluing map is a diffeomorphism into its image and satisfies the following properties:

For any fixed element a in the domain of Φ_T , the sequence $\Phi_T(a)$ is chain convergent to a , as T goes to infinity. Moreover, if A_i is a connection on the moduli space $\mathcal{M}_\kappa(X^{T_i}, w)$ such that $\lim_{i \rightarrow \infty} T_i = \infty$ and the sequence $\{A_i\}$ is chain convergent to an element of

$$\tilde{N}_0 \times_{\Gamma_{\alpha_0}} \tilde{N}_1 \times_{\Gamma_{\alpha_1}} \cdots \times_{\Gamma_{\alpha_k}} \tilde{N}_{k+1}$$

then A_i lies in the image of the gluing map for large enough values of i .

Remark 6.13 Theorems 6.7 and 6.11 are strong enough to study the cut-down moduli spaces on a 4-manifold with a long neck. To demonstrate this in the context of an example, suppose (X_i, w_i) is as above and $b^+(X_i) \geq 1$. Let Σ be an embedded surface in X_2 . Let $\nu(\Sigma) \subset X_2$ be an open neighborhood of Σ such that the inclusion of $\nu(\Sigma)$ in X_2 induces a surjection of fundamental groups. Let κ be chosen such that $\mathcal{M}_\kappa(X^T, w)$ has expected dimension two. We make the simplifying assumption that all critical points of the Chern–Simons functional on (Y, γ) are irreducible and nondegenerate, and all the moduli spaces on $Y \times \mathbb{R}$ are regular. By choosing a generic metric and a small compactly supported perturbation, we can assume that the moduli spaces of the form $\mathcal{M}_p(X_1, w_1; \alpha)$ with expected dimension at most 0 contain only irreducible and regular elements [52, Lemma 24]. Similarly, we can arrange for a metric, a small compactly supported perturbation on X_2 , and a geometric representative $V_2(\Sigma) \subset \mathcal{B}^*(\nu(\Sigma)) \times \nu(\Sigma)$ such that:

- (i) The moduli spaces of the form $\mathcal{M}_p(X_2, w_2; \alpha)$ with dimension at most two consist of irreducible and regular solutions.
- (ii) The map $r: \mathcal{M}_p(X_2, w_2; \alpha) \times \Sigma \rightarrow \mathcal{B}^*(\nu(\Sigma)) \times \nu(\Sigma)$, for any moduli space $\mathcal{M}_p(X_2, w_2; \alpha)$ of dimension at most two, is transversal to $V_2(\Sigma)$. Suppose $\mathcal{N}_p(X_1, w_1; \alpha, \Sigma)$ denotes the cut-down moduli space.

The chosen holonomy perturbations on X_1 and X_2 induce a holonomy perturbation on X^T for large values of T . Theorem 6.1 implies that for large enough values of T , the space $\mathcal{M}_\kappa(X^T, w) \times \Sigma$ is also cut-down transversely by $V_2(\Sigma)$ and the resulting space is compact. Furthermore, the elements in the cut-down space $\mathcal{N}_\kappa(X^T, w; \Sigma)$ are in correspondence with the elements of the space

$$\bigcup_{\kappa(p_1)+\kappa(p_2)=\kappa} \mathcal{M}_{p_1}(X_1, w_1; \alpha, \Sigma) \times \mathcal{N}_{p_2}(X_2, w_2; \alpha).$$

In the following, we use a similar strategy to study the cut-down moduli spaces on 4-manifolds with long necks, without going into details.

The following vanishing theorem is a standard application of the above theorems [17; 20; 68]:

Theorem 6.14 Suppose X_1 and X_2 are two 4-manifolds with $b^+(X_i) \geq 1$. Then for any 2-cycle w in the connected sum $X_1 \# X_2$ and any $z \in \mathbb{A}(X_1 \# X_2)^{\otimes 2}$, the number $D_{X_1 \# X_2, w}(z)$ is equal to zero.

Proof We can assume that $w = w_1 \cup w_2$ and $z = z_1 \cdot z_2$, where w_i is a 2-cycle in X_i and $z_i \in \mathbb{A}(X_i)^{\otimes 2}$. By replacing X_i , w_i and z_i with $X_i \# \overline{\mathbb{CP}}^2$, $w_i \cup E_i$ and $z_i \cdot (E_i)_{(2)}^2$, we can also assume that w_i is coprime to N . Here E_i is the exceptional class in $X_i \# \overline{\mathbb{CP}}^2$. We fix a Riemannian metric on $X_1 \# X_2$ whose restriction to the connected sum region is isometric to the product metric $[-T, T] \times S^3$, where T is a large constant and S^3 has the standard metric. Using the standard metric on S^3 allows us to ensure that all the framed moduli spaces on the cylinder $\mathbb{R} \times S^3$ are regular. We fix a holonomy perturbation of the ASD equation on each X_i such that the perturbation is supported outside of a neighborhood of the connected sum region, and the cut-down moduli spaces $\mathcal{N}_p(X_i, w_i; \Theta, z_i)$ of expected dimension at most zero are regular. Here Θ is the trivial connection, which is the only flat connection on S^3 . Now we can use Theorem 6.11 to conclude that the 0-dimensional moduli space $\mathcal{N}_\kappa(X_1 \# X_2, w_1 \cup w_2; z_1 \cdot z_2)$ is empty when T is large enough. $\square$

Next, we utilize the gluing and the compactness theorems to prove Proposition 3.45 on connected sums along $\Sigma(2, 3, 23)$. Recall that in addition to the trivial connections, there are 44 irreducible and eight $SU(2)$ -reducible connections on the trivial $SU(3)$ -bundle over $\Sigma(2, 3, 23)$. All these connections are nondegenerate. We choose a perturbation of the Chern–Simons functional of $\Sigma(2, 3, 23)$ (and the empty cycle) such that the critical points of the perturbed functional are the same as those of the Chern–Simons functional and the following assumption about the regularity of the elements of $[A] \in \mathcal{M}_p(\Sigma(2, 3, 23); \alpha, \beta)$ holds: if A is irreducible, then A is regular, and if A is reducible and induced by an $SU(2)$ -connection, then A is regular as a solution of the (perturbed) ASD equation for $SU(2)$ -connections [19; 58].

Suppose X_1 is a 4-manifold with $b^+(X_1) \geq 1$ and $\partial X_1 = \Sigma(2, 3, 23)$. Suppose also that w_1 is a closed 2-cycle in X_1 which is coprime to 3. Fix a metric with cylindrical ends on X_1 and a small holonomy perturbation of the ASD equation which is compatible with the perturbation of the Chern–Simons functional. Let $\mathcal{M}_p(X_1, w_1; \alpha)$ be the moduli space of solutions to a perturbation of the ASD equation associated to the pair (X_1, w_1) and the path p :

Proposition 6.15 *Suppose a positive integer n_0 is given. There is a metric and a holonomy perturbation of the ASD equation on X_1 such that the following holds: Let p be a path along (X_1, w_1) based at a flat connection α on $\Sigma(2, 3, 23)$ such that the index of the elements of $\mathcal{M}_p(X_1, w_1; \alpha)$ is at most n_0 . Then the moduli space $\mathcal{M}_p(X_1, w_1; \alpha)$ consists of regular solutions and does not have any reducible connections. Moreover, suppose $z_1 \in \mathbb{A}(X)^{\otimes 2}$ is such that*

$$d_{X_1, w_1} := -4w_1^2 - 4(\chi(X_1) + \sigma(X_1)) - 4 \equiv \deg(z_1) + 4 \pmod{12}.$$

Suppose also that the expected dimension of $\mathcal{N}_p(X_1, w_1; \alpha, z_1)$ is zero. Then there is a geometric representative for z_1 such that the cut-down moduli space $\mathcal{N}_p(X_1, w_1; \alpha, z_1)$ is compact.

Note that $\mathcal{N}_p(X_1, w_1; \alpha, z_1)$ might be a linear combination of different spaces. Then compactness of this space is defined to be the compactness of all the involved spaces in the linear combination. In what follows, we gloss over this point about the nature of the spaces $\mathcal{N}_p(X_1, w_1; \alpha, z_1)$.

Proof The arguments of [52] can be used to show that the metric and the perturbation can be chosen such that if p is a path as in the statement of the proposition, then $\mathcal{M}_p(X_1, w_1; \alpha)$ is regular and does not contain any reducible solutions. We can also assume that $V(z_1)$ is chosen such that all moduli spaces $\mathcal{N}_p(X_1, w_1; \alpha)$ with expected dimension at most $n_0 - \deg(z_1)$ are cut down transversely. Next, let the dimension of $\mathcal{M}_p(X_1, w_1; \alpha)$ be equal to $\deg(z_1)$, and $\{A_i\}_i$ be a sequence of connections in the cut-down moduli space $\mathcal{N}_p(X_1, w_1; \alpha, z_1)$. By Theorem 6.1, this sequence converges to an element $(([B], \mathbf{x}), ([C_1], \mathbf{x}), \dots, ([C_k], \mathbf{y}_k))$, where $[B] \in \widetilde{\mathcal{N}}_{p_0}(X_1, w_1; \alpha_0, z_1)$, $[C_i] \in \widetilde{\mathcal{M}}_{p_i}(\alpha_{i-1}, \alpha_i)$, $\alpha_k = \alpha$, and

$$(6.16) \quad \deg(z_1) \geq \text{index}(\mathcal{D}_B) + \text{index}(\mathcal{D}_{C_1}) + \dots + \text{index}(\mathcal{D}_{C_k}),$$

with equality holding if and only if the multisets $\mathbf{x}, \mathbf{y}_1, \dots, \mathbf{y}_k$ are empty. The main point of the rest of the argument is to show that instantons C_i with nonpositive values of $\text{index}(\mathcal{D}_{C_i})$ do not appear in the limit points of $\mathcal{N}_p(X_1, w_1; \alpha, z_1)$.

We first show that $\text{index}(\mathcal{D}_B)$ in (6.16) is bounded above. If C_i is irreducible, then $\text{index}(\mathcal{D}_{C_i})$ is positive. In the case that C_i is reducible, we cannot guarantee that $\text{index}(\mathcal{D}_{C_i})$ is positive. However, the index of $\mathcal{D}_{C_i}$ as an $\text{SU}(2)$ -connection is positive. Using Table 5, it is straightforward to check that for a reducible connection $C_i \in \widetilde{\mathcal{M}}_{p_i}(\mathbb{R} \times \Sigma(2, 3, 23); \alpha_{i-1}, \alpha_i)$ with nonpositive $\text{index}(\mathcal{D}_{C_i})$, the flat connection α_i

has to be β_8 , and α_{i-1} is equal to either the trivial connection Θ or the $SU(2)$ -connection β_2 . In the first case, $\text{index}(\mathcal{D}_{C_i})$ is equal to -3 and in the latter case $\text{index}(\mathcal{D}_{C_i})$ is equal to -2 . Suppose

$$(6.17) \quad \{i_1, \dots, i_l\} \subseteq \{0, \dots, k-1\}$$

is the set of indices such that $\alpha_{i_j} = \beta_8$. We have

$$(6.18) \quad \text{index}(\mathcal{D}_{C_{i_{j+1}}}) + \text{index}(\mathcal{D}_{C_{i_{j+2}}}) + \dots + \text{index}(\mathcal{D}_{C_{i_{j+1}}}) \equiv 0 \pmod{12}$$

for $1 \leq j \leq k-1$. In (6.18), the last term is at least -3 and the other terms are positive. This shows that the sum in (6.18) is nonnegative. We can use this to conclude that the sum $\text{index}(\mathcal{D}_{C_1}) + \dots + \text{index}(\mathcal{D}_{C_k})$ in (6.16) is at least -3 . That is to say, $\text{index}(\mathcal{D}_B)$ is not greater than $\deg(z_1) + 3$. By increasing the value of n_0 if necessary, we can assume that $\mathcal{N}_{p_0}(X_1, w_1; \alpha_0, z_1)$, which contains B , is cut down transversely. In particular, $\text{index}(\mathcal{D}_B) \geq \deg(z_1)$.

If $l \geq 1$, then the index formula (2.16) shows

$$\text{index}(\mathcal{D}_B) + \text{index}(\mathcal{D}_{C_1}) + \dots + \text{index}(\mathcal{D}_{C_{i_1}}) \equiv d_{X_1, w_1} - 3 \equiv \deg(z_1) + 1 \pmod{12}.$$

In the above expression, the first term on the left-hand side is not less than $\deg(z_1)$, the last term is not less than -3 , and the remaining terms are positive. Therefore, the sum on the left-hand side is not less than $\deg(z_1) + 1$. This also shows that the right-hand side of (6.16) is at least $\deg(z_1) + 1$, which is a contradiction, and as a result, $l = 0$. Therefore, in (6.16), $\text{index}(\mathcal{D}_{C_i})$ is always positive. This also implies that $k = 0$ and $\mathbf{x}$ is empty. Consequently, $\mathcal{N}_p(X_1, w_1; \alpha, z_1)$ is compact. $\square$

Form the moduli spaces $\mathcal{N}_p(X_1, w_1; \alpha, z_1)$ with the perturbations from Proposition 6.15, and define the following element of $I_4(\Sigma(2, 3, 23))$:

$$(6.19) \quad D_{X_1, w_1}(z_1) := \sum \# \mathcal{N}_p(X_1, w_1; \alpha, z_1) \cdot \alpha,$$

where the sum is over all irreducible connections α in $\Sigma(2, 3, 23)$ and the paths p such that

$$(6.20) \quad \mathcal{N}_p(X_1, w_1; \alpha, z_1)$$

is 0-dimensional. Here we follow the standard conventions to orient (6.20) [58; 52]. Since $N = 3$, we do not need a homology orientation of (X_1, w_1) to fix a sign for $D_{X_1, w_1}(z_1)$.

Next, let X_2 be a 4-manifold with $b^+(X_2) \geq 1$ and $\partial X_2 = \overline{\Sigma(2, 3, 23)}$. Let also w_2 be a closed 2-cycle in X_2 which is coprime to 3. Fix a metric with cylindrical ends on X_2 and a small holonomy perturbation of the ASD equation which is compatible with the perturbation of the Chern–Simons functional:

Proposition 6.21 *Suppose a positive integer n_0 is given. There is a metric and a holonomy perturbation of the ASD equation on X_2 such that the following holds: Let p be a path along (X_2, w_2) based at a flat connection α on $\Sigma(2, 3, 23)$ such that the index of the elements of $\mathcal{M}_p(X_2, w_2; \alpha)$ is at most n_0 . Then the moduli space $\mathcal{M}_p(X_2, w_2; \alpha)$ consists of regular solutions and does not have any reducible connections. Moreover, suppose $z_2 \in \mathbb{A}(X)^{\otimes 2}$ is chosen such that $\deg(z_2)$ is divisible by 4. Suppose also the expected dimension of $\mathcal{N}_p(X_2, w_2; \alpha, z_2)$ is zero. Then there is a geometric representative for z_2 such that the cut-down moduli space $\mathcal{N}_p(X_2, w_2; \alpha, z_2)$ is compact.*

The proof of this proposition is analogous to that of Proposition 6.15, and we leave it to the reader. The cut-down moduli spaces in Proposition 6.21 can be used to define the following functional on $I_*(\Sigma(2, 3, 23))$:

$$(6.22) \quad D^{X_2, w_2}(z_2)(\alpha) := \#\mathcal{N}_p(X_2, w_2; \alpha, z_2),$$

where α is an irreducible connection on $\Sigma(2, 3, 23)$ and the path p is chosen such that $\mathcal{N}_p(X_2, w_2; \alpha, z_2)$ is 0-dimensional. The index formula shows that this space is 0-dimensional only if

$$(6.23) \quad \deg(\alpha) \equiv 4w_2^2 + 4(\chi(X_2) + \sigma(X_2)) - 4 + \deg(z_2) \pmod{12}.$$

Thus, $D^{X_2, w_2}(z_2)$ is nonzero only on the elements of $I_*(\Sigma(2, 3, 23))$ that satisfy (6.23).

Proposition 6.24 *For $i = 1, 2$, suppose (X_i, w_i) and z_i are as in Propositions 6.15 and 6.21. Then*

$$(6.25) \quad D_{X_1 \circ X_2, w_1 \cup w_2}(z_1 \cdot z_2) = D^{X_2, w_2}(z_2) \circ D_{X_1, w_1}(z_1).$$

Proof Use Propositions 6.15 and 6.21 to fix metrics on X_1 and X_2 . Let also $X_1 \circ X_2$ be equipped with a Riemannian metric which is compatible with the metrics on X_1 and X_2 and has a long neck along $\Sigma(2, 3, 23)$. By slightly modifying the holonomy perturbations provided by Propositions 6.15 and 6.21, we can assume that the perturbation of X_i on the complement of a compact subset of X_i^+ is induced by the chosen perturbation of the Chern–Simons functional of $\Sigma(2, 3, 23)$ and the claims of these propositions about

the moduli spaces of the form $\mathcal{N}_p(X_1, w_1; \alpha, z_1)$ and $\mathcal{N}_p(X_2, w_2; \alpha, z_2)$ still hold. The perturbations of the ASD equations on X_1 and X_2 induce a perturbation of the ASD equation on $X_1 \circ X_2$. By applying [Theorem 6.11](#) and arguments similar to those in the proof of [Proposition 6.15](#), we can conclude that, for a long enough neck along $\Sigma(2, 3, 23)$, we have the following diffeomorphism of 0-dimensional moduli spaces:

$$\mathcal{N}_\kappa(X_1 \circ X_2, w_1 \cup w_2, z_1 \cdot z_2) \cong \bigcup_{\alpha} \mathcal{N}_{p_1}(X_1, w_1; \alpha, z_1) \times \mathcal{N}_{p_2}(X_2, w_2; \alpha, z_2).$$

Here α is an irreducible connection on $\Sigma(2, 3, 23)$, and κ , p_1 and p_2 are chosen such that all the above cut-down moduli spaces are 0-dimensional. Standard arguments show that the above diffeomorphism is compatible with respect to the orientation of the involved moduli spaces. This diffeomorphism implies the claim in [\(6.25\)](#). $\square$

This proposition essentially proves [Proposition 3.45](#) from [Section 3.4](#). We only need to extend the above proposition to the case that w_1 and w_2 are not necessarily coprime to 3 and $\deg(z_2)$ is not divisible by 4. The assumption on the w_i can be removed by the blowing-up trick. The dimension of the moduli space of rank 3 instantons on the closed 4-manifold $X_1 \circ X_2$ is always divisible by 4. Therefore, if we define $D^{X_2, w_2}(z_2) = 0$ in the case that $\deg(z_2) \not\equiv 0 \pmod{4}$, then the above theorem still holds.

6.2 Gluing theory for negative embedded spheres

In this subsection, we give a proof of [Proposition 2.21](#) based on the techniques which are discussed in [Section 6.1](#). Suppose X is a smooth 4-manifold with $b^+(X) \geq 2$. Suppose also σ is an embedded sphere in X with $\sigma \cdot \sigma = -3$. A tubular neighborhood of σ , denoted by Z , is a disc bundle over σ with Euler class -3 and boundary $Y = L(3, -1)$. Thus X can be split as $Z \circ X_1$, where X_1 is the closure of the complement of Z . We will also write w_0 for a fiber of the disc bundle Z . Fix the orientation on w_0 such that w_0 intersects σ negatively in one point.

Fix a Riemannian metric with a cylindrical end on Z . Let also L be the complex line bundle over Z associated to the 2-cycle w_0 . Since $b^+(Z) = b^1(Z) = 0$, there is a unique ASD connection on L with finite energy. This connection, denoted by B , is asymptotic to a flat connection χ which maps the generator of $\pi_1(Y)$ to $\zeta = e^{2\pi i/3}$. Consider the ASD $U(N)$ -connection $A := B^{k_1} \oplus \cdots \oplus B^{k_N}$ which is asymptotic to the flat connection $\alpha := \chi^{k_1} \oplus \cdots \oplus \chi^{k_N}$. The first Chern class of the underlying $U(N)$ -connection is equal to $(k_1 + \cdots + k_N) \text{P.D.}[w_0]$. The topological energy of A

is given by the formula

$$\kappa(A) = \frac{1}{12N} \sum_{1 \leq i, j \leq N} (k_i - k_j)^2.$$

Therefore, we can use (2.16) to compute the index of $\mathcal{D}_A$. In particular, if $|k_i - k_j| \leq 3$ for all i and j , then (2.16) shows [12] that

$$\text{index}(\mathcal{D}_A) = 1 - N^2 + \sum_{1 \leq i, j \leq N} |k_i - k_j|.$$

Table 2 consists of various choices of $\text{U}(3)$ –connections. For each connection A , the first Chern class of the underlying bundle of A is equal to $k \text{ P.D.}[w_0]$, where k is also given in Table 2.

A	α	k	$\text{index } \mathcal{D}_A$	$\dim \widetilde{\mathcal{M}}(A)$	Γ_α
$B^2 \oplus 1 \oplus 1$	$\chi^2 \oplus 1 \oplus 1$	2	0	4	$\text{S}(\text{U}(1) \times \text{U}(2))$
$B \oplus B \oplus 1$	$\chi \oplus \chi \oplus 1$	2	−4	0	$\text{S}(\text{U}(2) \times \text{U}(1))$
$B^{-1} \oplus B^2 \oplus B$	$\chi^2 \oplus \chi^2 \oplus \chi$	2	4	8	$\text{S}(\text{U}(2) \times \text{U}(1))$
$B \oplus B^{-1} \oplus B^{-1}$	$\chi \oplus \chi^2 \oplus \chi^2$	−1	0	4	$\text{S}(\text{U}(1) \times \text{U}(2))$
$B^{-1} \oplus 1 \oplus 1$	$\chi^2 \oplus 1 \oplus 1$	−1	−4	0	$\text{S}(\text{U}(1) \times \text{U}(2))$
$B^{-2} \oplus B \oplus 1$	$\chi \oplus \chi \oplus 1$	−1	4	8	$\text{S}(\text{U}(2) \times \text{U}(1))$

Table 2

We will write $\mathcal{M}(A)$ (respectively, $\widetilde{\mathcal{M}}(A)$) for the moduli space (respectively, the framed moduli space) of connections on Z corresponding to the path which is represented by A . From now on, we assume that an anti-self-dual metric with positive scalar curvature is fixed on Z [61].⁹

Proposition 6.26 *For each A in Table 2, A has the minimal topological energy among all ASD connections with the same limiting flat connection as A and the same c_1 .*

Proof Let A' be an ASD connection with a smaller topological energy and the same limiting flat connection and c_1 as A . The difference $\dim(\widetilde{\mathcal{M}}(A)) - \dim(\widetilde{\mathcal{M}}(A'))$ is at least 12. On the other hand, any such connection is regular because of the choice of

⁹We use such metrics to ensure regularity of moduli spaces of ASD connections with respect to these metrics. The construction of [61] provides asymptotically cylindrical ASD metrics rather than cylindrical metrics. However, that is enough for our purposes because we can pick a cylindrical approximation to the these metrics and then argue as in [13, Corollary 5.13].

the metric. Therefore, the dimension of $\widetilde{\mathcal{M}}(A')$ is at least $\dim(\Gamma_\alpha) - \dim(\Gamma_{A'})$. This can be used to rule out the existence of A' . $\square$

We study the moduli spaces $\mathcal{M}(A)$ for various choices of A . The same argument as in [Proposition 6.26](#) shows that if $A = B \oplus B \oplus 1$, then the moduli space $\mathcal{M}(A)$ contains only the completely reducible connection A . Next, we turn to the case that $A = B^2 \oplus 1 \oplus 1$. An argument analogous to that in [Proposition 6.26](#) shows that there are three types of connections in this space:

- finitely many irreducible connections;
- the complete reducible connection $A = B^2 \oplus 1 \oplus 1$;
- reducible connections of the form $R \oplus 1$, where R is an irreducible $\mathrm{U}(2)$ -connection in the 1-dimensional space $\mathcal{M}(B^2 \oplus 1)$.

Proposition 6.27 *The unique component of $\mathcal{M}(B^2 \oplus 1)$ containing $B^2 \oplus 1$ is a half-line $[0, \infty)$. All the other components are either circles or copies of $\mathbb{R}$ consisting of only irreducible connections. The corresponding components of $\widetilde{\mathcal{M}}(B^2 \oplus 1 \oplus 1)$ are $\mathbb{C}^2$, $S^3 \times S^1$ and $S^3 \times \mathbb{R}$.*

Proof The proof of the first part is straightforward. For the second part, note that the orbit of the connection $R \oplus 1$ in the framed moduli space, for an irreducible connection R , is $\Gamma_\alpha / \Gamma_{R \oplus 1} = \mathrm{S}(\mathrm{U}(1) \times \mathrm{U}(2)) / \mathrm{S}(\mathrm{U}(1)_2 \times \mathrm{U}(1)) = S^3$, where $\mathrm{U}(1)_2 \times \mathrm{U}(1)$ denotes 3×3 diagonal matrices where the first two diagonal entries are equal to each other. $\square$

Now we are ready to prove the first part of [Proposition 2.21](#):

Proposition 6.28 *Suppose X and σ are as above and $z \in \mathbb{A}(\langle \sigma \rangle^\perp)^{\otimes 2}$. Suppose also that w is a 2-cycle in X such that $w \cdot \sigma \equiv 1 \pmod{3}$. Then there is a constant c such that*

$$\mathrm{D}_{X,w}^3 \left(\left(-\frac{3}{2} \sigma_{(3)} - \frac{3}{2} \sigma_{(2)}^2 - a_2 \right) z \right) = c \mathrm{D}_{X,w-\sigma}^3(z).$$

Proof For simplicity of exposition, we assume that $z = 1$. A similar proof works in the more general case. Equip X with a Riemannian metric that has a neck of length T along the lens space Y , and denote the resulting Riemannian manifold by X^T . We first study the 4-dimensional moduli spaces of the form $\mathcal{M}_{k_0}(X^T, w)$ for large values of T . Let α_1 and α_2 denote the flat connections $\chi^2 \oplus 1 \oplus 1$ and $\chi \oplus \chi \oplus 1$. We write G_i for the stabilizer group Γ_{α_i} . We also write $\mathcal{R}_1$ for $\widetilde{\mathcal{M}}(B^2 \oplus 1 \oplus 1)$ and $\mathcal{R}_2$ for $\widetilde{\mathcal{M}}(B \oplus B \oplus 1)$. Clearly $\mathcal{R}_i$ is a G_i -manifold. By [Theorem 6.11](#) and

the above description of the low-dimensional moduli spaces on Z , the moduli space $\mathcal{M}_{\kappa_0}(X^T, w)$ can be covered with two open sets $\mathcal{U}$ and $\mathcal{V}$ of the form

$$\mathcal{U} = \widetilde{\mathcal{M}}_{p_1}(X_1, w \cap X_1; \alpha_1) \times_{G_1} \mathcal{R}_1 \quad \text{and} \quad \mathcal{V} = \widetilde{\mathcal{M}}_{p_2}(X_1, w \cap X_1; \alpha_2) \times_{G_2} \mathcal{R}_2,$$

where the framed moduli spaces $\widetilde{\mathcal{M}}_{p_1}(X_1, w \cap X_1; \alpha_1)$ and $\widetilde{\mathcal{M}}_{p_2}(X_1, w \cap X_1; \alpha_2)$ are respectively of dimensions 4 and 8. By choosing a generic metric on X_1 and a small holonomy perturbation of the ASD equation supported in X_1 , we can assume the action of Γ_{α_i} on $\widetilde{\mathcal{M}}_{p_i}(X_1, w \cap X_1; \alpha_i)$ is free. In particular, there are G_i equivariant maps

$$(6.29) \quad \widetilde{f}_i: \widetilde{\mathcal{M}}_{p_i}(X_1, w \cap X_1; \alpha_i) \rightarrow EG_i,$$

where EG_i is the universal principal G_i -bundle over the classifying space BG_i . Every connected component of $\widetilde{\mathcal{M}}_{p_1}(X_1, w \cap X_1; \alpha_1)$ can be identified with G_1 and we can assume that the map $\widetilde{f}_1$ is the same on different connected components of $\widetilde{\mathcal{M}}_{p_1}(X_1, w \cap X_1; \alpha_1)$.

Gluing theory also gives a description of the intersection $\mathcal{U} \cap \mathcal{V}$ for large values of T . Let $\mathcal{N}$ be the $(G_1 \times G_2)$ -manifold $\widetilde{\mathcal{M}}_q(\alpha_1, \alpha_2)$, where the path q is chosen such that this space is 8-dimensional. Since the orbit of any irreducible connection with respect to the action of $G_1 \times G_2$ and translations is at least 9-dimensional, $\mathcal{N}$ does not contain any irreducible connections. Thus $\mathcal{N}$ consists of ASD connections of the form $R \oplus 1$ on $\mathbb{R} \times L(3, 1)$, where R is asymptotic to $\chi^2 \oplus 1$ and $\chi \oplus \chi$ on the two ends. After dividing by the action of translations, there are finitely many choices for the connection R . Thus, this manifold, as a $(G_1 \times G_2)$ -space, is the union of finitely many spaces of the form

$$(6.30) \quad \mathbb{R} \times \frac{G_1 \times G_2}{S(U(1)_2 \times U(1))}.$$

In particular, the action of G_i on this space is free. The G_1 -space $\mathcal{N} \times_{G_2} \mathcal{R}_2$ can be identified with the end of $\mathcal{R}_1$. The G_2 -space $\widetilde{\mathcal{M}}_{p_1}(X_1, w \cap X_1; \alpha_1) \times_{G_1} \mathcal{N}$ can be identified with the end of $\widetilde{\mathcal{M}}_{p_2}(X_1, w \cap X_1; \alpha_2)$. Using these identifications, the overlap $\mathcal{U} \cap \mathcal{V}$ is identified with the space

$$(6.31) \quad \widetilde{\mathcal{M}}_{p_1}(X_1, w \cap X_1; \alpha_1) \times_{G_1} \mathcal{N} \times_{G_2} \mathcal{R}_2.$$

In summary, $\mathcal{M}_{\kappa_0}(X^T, w)$ is the pushout of the following diagram:

$$(6.32) \quad \begin{array}{ccc} \widetilde{\mathcal{M}}_{p_1}(X_1, w \cap X_1; \alpha_1) \times_{G_1} \mathcal{N} \times_{G_2} \mathcal{R}_2 & \hookrightarrow & \widetilde{\mathcal{M}}_{p_1}(X_1, w \cap X_1; \alpha_1) \times_{G_1} \mathcal{R}_1 \\ \downarrow & \lrcorner & \downarrow \\ \widetilde{\mathcal{M}}_{p_2}(X_1, w \cap X_1; \alpha_2) \times_{G_2} \mathcal{R}_2 & \longrightarrow & \mathcal{M}_{\kappa_0}(X^T, w) \end{array}$$

We can form a universal version of the above diagram as follows:

$$(6.33) \quad \begin{array}{ccc} EG_1 \times_{G_1} \mathcal{N} \times_{G_2} \mathcal{R}_2 & \xrightarrow{h_2} & EG_1 \times_{G_1} \mathcal{R}_1 \\ \downarrow h_1 & \lrcorner & \downarrow \\ EG_2 \times_{G_2} \mathcal{R}_2 & \longrightarrow & M \end{array}$$

The free action of G_2 on $EG_1 \times_{G_1} \mathcal{N}$ gives a G_2 -equivariant map

$$f: EG_1 \times_{G_1} \mathcal{N} \rightarrow EG_2.$$

The vertical map h_1 is induced by f . The horizontal map h_2 is induced by the inclusion of $\mathcal{N} \times_{G_2} \mathcal{R}_2$ into $\mathcal{R}_1$. The space M is then the pushout of h_1 and h_2 . The maps $\tilde{f}_1$ and $\tilde{f}_2$ determine the maps

$$f_i: \tilde{\mathcal{M}}_{p_i}(X_1, w \cap X_1; \alpha_i) \times_{G_i} \mathcal{R}_i \rightarrow EG_i \times_{G_i} \mathcal{R}_i$$

and

$$g: \tilde{\mathcal{M}}_{p_1}(X_1, w \cap X_1; \alpha_1) \times_{G_1} \mathcal{N} \times_{G_2} \mathcal{R}_2 \rightarrow EG_1 \times_{G_1} \mathcal{N} \times_{G_2} \mathcal{R}_2.$$

The maps $\tilde{f}_1$ and $\tilde{f}_2$ can be chosen such that f_1 , f_2 and g give rise to a map from diagram (6.32) to diagram (6.33). We will write F for the induced map from $\mathcal{M}_{\kappa_0}(X^T, w)$ to M .

There is a similar description of the universal bundle over the space $\mathcal{M}_{\kappa_0}(X^T, w) \times \sigma$. Suppose q_1 is the path over Z determined by the connection $B^2 \oplus 1 \oplus 1$. We can form an equivariant universal bundle $\tilde{\mathbb{P}}_1$ over $\tilde{\mathcal{B}}_{q_1}(Z, w \cap Z; \alpha) \times \sigma$ similar to the universal bundles in Section 2.1. There is an action of G_1 on $\tilde{\mathbb{P}}_1$ which lifts the obvious action of this group on $\tilde{\mathcal{B}}_{q_1}(Z, w \cap Z; \alpha) \times \sigma$. This bundle induces a PU(3)-bundle over $\mathcal{R}_1 \times \sigma$ with an action of G_1 , which we still denote by $\tilde{\mathbb{P}}_1$. Similarly, we can construct a PU(3)-bundle $\tilde{\mathbb{P}}_2$ over $\mathcal{R}_2 \times \sigma$ with an action of G_2 . Since $\mathcal{R}_2$ consists of only the class of the connection $B \oplus B \oplus 1$, the bundle $\tilde{\mathbb{P}}_2$ has the form [12, Proposition 46]

$$(6.34) \quad \tilde{\mathbb{P}}_2 = F \boxtimes L \oplus P,$$

where F and P are the standard U(2)- and U(1)-representations of G_2 , and L is the line bundle defined at the beginning of the subsection. To be more precise, (6.34) gives a lift of $\tilde{\mathbb{P}}_2$ to a U(3)-bundle with an action of G_2 . The restriction of $\tilde{\mathbb{P}}_1$, as a PU(3)-bundle with an action of G_1 , to the subset $\mathcal{N} \times_{G_2} \mathcal{R}_2 \times \sigma$ of $\mathcal{R}_1 \times \sigma$ can be identified

with $\mathcal{N} \times_{G_2} \widetilde{\mathbb{P}}_2$. Thus, $\widetilde{\mathbb{P}}_1$ and $\widetilde{\mathbb{P}}_2$ give rise to the following diagram of $\mathrm{PU}(3)$ –bundles:

(6.35)

$$\begin{array}{ccc} \mathbb{P}_3 := EG_1 \times_{G_1} \mathcal{N} \times_{G_2} \widetilde{\mathbb{P}}_2 & \xrightarrow{\bar{h}_2} & EG_1 \times_{G_1} \widetilde{\mathbb{P}}_1 \\ \downarrow \bar{h}_1 & \lrcorner & \downarrow \\ EG_2 \times_{G_2} \widetilde{\mathbb{P}}_2 & \longrightarrow & \bar{\mathbb{P}} \end{array}$$

The $\mathrm{PU}(3)$ –bundle $\bar{\mathbb{P}}$ over $M \times \sigma$ is the pushout of the above diagram. The pullback of $\bar{\mathbb{P}}$ with respect to the map $(F, \mathrm{id}): \mathcal{M}_{\kappa_0}(X^T, w) \times \sigma \rightarrow M \times \sigma$ is equal to the universal bundle.

Suppose $\varphi \in H^4(M, \mathbb{Q})$ is defined to be

$$\tfrac{3}{2}c_3(\bar{\mathbb{P}})/\sigma - \tfrac{3}{2}(c_2(\bar{\mathbb{P}})/\sigma)^2 - c_2(\bar{\mathbb{P}})/x.$$

We claim the restriction of φ to the subspace $EG_2 \times_{G_2} \mathcal{R}_2$ vanishes. To see this, let

$$\mathfrak{p} := c_1(P) \quad \text{and} \quad \mathfrak{f} := c_2(F) \in H_{G_2}^*(\mathcal{R}_2) = H^*(BG_2)$$

be the equivariant Chern classes of the bundles P and F in the description of $\widetilde{\mathbb{P}}_2$ in (6.34). Then it is easy to verify our claim from the following identities:

$$\begin{aligned} c_2(EG_2 \times_{G_2} \widetilde{\mathbb{P}}_2)/\sigma &= -\mathfrak{p}, & c_3(EG_2 \times_{G_2} \widetilde{\mathbb{P}}_2)/\sigma &= \tfrac{1}{3}\mathfrak{p}^2 + \tfrac{2}{3}\mathfrak{f}, \\ c_2(EG_2 \times_{G_2} \widetilde{\mathbb{P}}_2)/x &= -\mathfrak{p}^2 + \mathfrak{f}. \end{aligned}$$

By (6.30), the space $EG_1 \times_{G_1} \mathcal{N} \times_{G_2} \mathcal{R}_2$ is homotopy equivalent to finitely many copies of

$$EG_1/S(\mathrm{U}(1)_2 \times \mathrm{U}(1)) \cong BS(\mathrm{U}(1)_2 \times \mathrm{U}(1)) \cong B(\mathrm{U}(1) \times \mathrm{U}(1))$$

whose odd-degree cohomology groups vanish. Therefore, we can apply the Mayer-Vietoris exact sequence for the pushout diagram in (6.33) to conclude that there is a unique choice of a relative cohomology class

$$\psi \in H^4(EG_1 \times_{G_1} \mathcal{R}_1, EG_1 \times_{G_1} \mathcal{N} \times_{G_2} \mathcal{R}_2)$$

such that φ is equal to the image of ψ in $H^4(M, \mathbb{Q})$. This discussion shows that the pairing of $F^*(\varphi)$ and the fundamental class of the 4–dimensional moduli space $\mathcal{M}_{\kappa_0}(X^T, w)$ is equal to the pairing of $f_1^*(\psi)$ and the relative fundamental class of $\widetilde{\mathcal{M}}_{p_1}(X_1, w \cap X_1; \alpha_1) \times_{G_1} \mathcal{R}_1$. The description of the moduli spaces and the map $\widetilde{f}_1$ shows that there is a constant c such that the latter pairing is equal to

$$c \cdot \#\mathcal{M}_{p_1}(X_1, w \cap X_1; \alpha_1).$$

Since $\widetilde{\mathcal{M}}(B^{-1} \oplus 1 \oplus 1)$ consists of a single point, we have

(6.36)
$$\begin{aligned} \mathcal{M}_{p_1}(X_1, w \cap X_1; \alpha_1) &= \widetilde{\mathcal{M}}_{p_1}(X_1, w \cap X_1; \alpha_1) \times_{\Gamma_{\alpha_1}} \widetilde{\mathcal{M}}(B^{-1} \oplus 1 \oplus 1) \\ &= \mathcal{M}_{\kappa_1}(X^T, w - \sigma). \end{aligned}$$

Here κ_1 is chosen such that the space $\mathcal{M}_{\kappa_1}(X^T, w - \sigma)$ is 0-dimensional. We use Theorem 6.11 to conclude the second identity for large enough values of T . The identities in (6.36) allow us to verify the desired claim. $\square$

The second part of Proposition 2.21 can be proved by applying the first part to the (-3) -sphere σ with the reverse orientation. Next, we turn to the proof of the last part of Proposition 2.21. As in the previous case, we need to study some low-dimensional moduli spaces over Z . Table 3 consists of various choices of $U(3)$ -connections on Z with vanishing c_1 . The proof of Proposition 6.26 shows that each connection A in this table has the minimal energy among all ASD connections with the same limiting flat connection and vanishing c_1 .

A	α	index $\mathcal{D}_A$	$\dim \widetilde{\mathcal{M}}(A)$	Γ_α
$1 \oplus 1 \oplus 1$	$1 \oplus 1 \oplus 1$	-8	0	$SU(3)$
$B \oplus B^{-1} \oplus 1$	$\chi \oplus \chi^2 \oplus 1$	0	2	$S(U(1) \times U(1) \times U(1))$
$B \oplus B \oplus B^{-2}$	$\chi \oplus \chi \oplus \chi$	4	12	$SU(3)$
$B^{-1} \oplus B^{-1} \oplus B^2$	$\chi^2 \oplus \chi^2 \oplus \chi^2$	4	12	$SU(3)$

Table 3

For $A = B \oplus B^{-1} \oplus 1$, the moduli space $\mathcal{M}(A)$ contains three types of connections:

- finitely many irreducible connections;
- the completely reducible connection $A = B \oplus B^{-1} \oplus 1$;
- reducible connections of the form $R \oplus 1$, where R is an irreducible $SU(2)$ -connection in the 1-dimensional space $\mathcal{M}(B \oplus B^{-1})$.

Proposition 6.37 *Each of the connected components of $\mathcal{M}(B \oplus B^{-1})$ which contains $B \oplus B^{-1}$ is a half-line $[0, \infty)$. All the other components are either circles or copies of $\mathbb{R}$ consisting of only irreducible connections. The corresponding components of $\widetilde{\mathcal{M}}(A)$ are $\mathbb{C}$, $S^1 \times S^1$ and $S^1 \times \mathbb{R}$, and the action of $\Gamma_\alpha = S^1$ is standard.*

Proposition 6.38 Suppose X and σ are as above and $z \in \mathbb{A}((\sigma)^\perp)^{\otimes 2}$. Suppose also w is a 2-cycle in X_1 . Then

$$D_{X,w}^3((\sigma_{(2)}^4 + 4a_2\sigma_{(2)}^2 + 3\sigma_{(3)}^2)z) = 0$$

and

$$D_{X,w}^3((\sigma_{(2)}^3\sigma_{(3)} + 3a_3\sigma_{(2)}^2 + a_2\sigma_{(2)}\sigma_{(3)})z) = 0.$$

Proof Similar to the proof of Proposition 6.28, we may assume that $z = 1$. We need to study the 8-dimensional moduli space for the first identity and the 10-dimensional moduli space for the second identity. Suppose κ_0 is a constant number such that the expected dimension of the moduli space $\mathcal{M}_{\kappa_0}(X^T, w)$ is 8 or 10. This moduli space is compact. Moreover, Theorem 6.11 and the above description of the low-dimensional moduli spaces on Z show that $\mathcal{M}_{\kappa_0}(X^T, w)$ can be given as the pushout of the diagram

$$(6.39) \quad \begin{array}{ccc} \widetilde{\mathcal{M}}_{p_1}(X_1, w; \alpha_1) \times_{G_1} \mathcal{N} \times_{G_2} \mathcal{R}_2 & \hookrightarrow & \widetilde{\mathcal{M}}_{p_1}(X_1, w; \alpha_1) \times_{G_1} \mathcal{R}_1 \\ \downarrow & \lrcorner & \downarrow \\ \widetilde{\mathcal{M}}_{p_2}(X_1, w; \alpha_2) \times_{G_2} \mathcal{R}_2 & \longrightarrow & \mathcal{M}_{\kappa_0}(X^T, w) \end{array}$$

where $\alpha_1 = \chi \oplus \chi^2 \oplus 1$, $\alpha_2 = 1 \oplus 1 \oplus 1$, $G_i = \Gamma_{\alpha_i}$, $\mathcal{R}_1 = \widetilde{\mathcal{M}}(B \oplus B^{-1} \oplus 1)$, $\mathcal{R}_2 = \widetilde{\mathcal{M}}(1 \oplus 1 \oplus 1)$, $\mathcal{N} = \widetilde{\mathcal{M}}_q(\alpha_1, \alpha_2)$ and the path q is chosen such that the index of any element in $\mathcal{N}$ is equal to 0. As in the previous case, we can form the universal version of the above diagram as

$$(6.40) \quad \begin{array}{ccc} EG_1 \times_{G_1} \mathcal{N} \times_{G_2} \mathcal{R}_2 & \xrightarrow{h_2} & EG_1 \times_{G_1} \mathcal{R}_1 \\ \downarrow h_1 & \lrcorner & \downarrow \\ EG_2 \times_{G_2} \mathcal{R}_2 & \longrightarrow & M \end{array}$$

with a map $F: \mathcal{M}_{\kappa_0}(X^T, w) \rightarrow M$.

There are also $\text{PU}(3)$ -bundles $\widetilde{\mathbb{P}}_i$ on $\mathcal{R}_i$ with an action of G_i . These bundles give rise to a pushout diagram of bundles

$$(6.41) \quad \begin{array}{ccc} \mathbb{P}_3 := EG_1 \times_{G_1} \mathcal{N} \times_{G_2} \widetilde{\mathbb{P}}_2 & \xrightarrow{\bar{h}_2} & EG_1 \times_{G_1} \widetilde{\mathbb{P}}_1 \\ \downarrow \bar{h}_1 & \lrcorner & \downarrow \\ EG_2 \times_{G_2} \widetilde{\mathbb{P}}_2 & \longrightarrow & \bar{\mathbb{P}} \end{array}$$

such that $(F, \text{id})^*(\bar{\mathbb{P}})$ is equal to the universal bundle on $\mathcal{M}_{k_0}(X^T, w) \times \sigma$. Let $\varphi_1 \in H^8(M, \mathbb{Q})$ and $\varphi_2 \in H^{10}(M, \mathbb{Q})$ be defined by

$$\begin{aligned}\varphi_1 &:= (c_2(\bar{\mathbb{P}})/\sigma)^4 + 4(c_2(\bar{\mathbb{P}})/x) \cdot (c_2(\bar{\mathbb{P}})/\sigma)^2 + 3(c_3(\bar{\mathbb{P}})/\sigma)^2, \\ \varphi_2 &:= (c_2(\bar{\mathbb{P}})/\sigma)^3 \cdot (c_3(\bar{\mathbb{P}})/\sigma) + 3(c_3(\bar{\mathbb{P}})/x) \cdot (c_2(\bar{\mathbb{P}})/\sigma)^2 \\ &\quad + (c_2(\bar{\mathbb{P}})/x) \cdot (c_2(\bar{\mathbb{P}})/\sigma) \cdot (c_3(\bar{\mathbb{P}})/\sigma).\end{aligned}$$

The restrictions of cohomology classes φ_1 and φ_2 to the open sets $EG_i \times_{G_i} \mathcal{R}_i$ vanish. We show this claim for the connected component $EG_1 \times_{G_1} \mathbb{C}$ of $EG_1 \times_{G_1} \mathcal{R}_1$. The restriction of the bundle $\widetilde{\mathbb{P}}_1$ to the G_1 -manifold $\mathbb{C}$ is given by [12, Proposition 46]

$$P \boxtimes L \oplus Q \boxtimes L^{-1} \oplus R \boxtimes \underline{\mathbb{C}},$$

where P , Q and R are the G_1 -equivariant line bundles over $\mathbb{C}$ associated to the three standard 1-dimensional representations of G_1 . Note that $P \otimes Q \otimes R$ is the trivial bundle with the trivial action of G_1 . Suppose $p, q \in H_{G_1}^*(\mathbb{C})$ denote the equivariant first Chern classes of the bundles P and Q . Then

$$\begin{aligned}c_2(EG_1 \times_{G_1} \widetilde{\mathbb{P}}_1)/\sigma &= p - q, & c_3(EG_1 \times_{G_1} \widetilde{\mathbb{P}}_1)/\sigma &= q^2 - p^2, \\ c_2(EG_1 \times_{G_1} \widetilde{\mathbb{P}}_1)/x &= -p^2 - q^2 - pq, & c_3(EG_1 \times_{G_1} \widetilde{\mathbb{P}}_1)/x &= -pq(p + q).\end{aligned}$$

These identities can be used to show that the restrictions of φ_1 and φ_2 to $EG_1 \times_{G_1} \mathbb{C}$ are equal to zero. Using similar arguments, it is even easier to show that the restrictions of these two cohomology classes to $EG_2 \times_{G_2} \mathcal{R}_2$ and the remaining connected components of $EG_1 \times_{G_1} \mathcal{R}_1$ are equal to zero. Since the cohomology groups of $EG_1 \times_{G_1} \mathcal{N} \times_{G_2} \mathcal{R}_2$ in odd degrees vanish, the Mayer-Vietoris exact sequence for diagram (6.40) implies that φ_1 and φ_2 are both equal to zero. In particular, this verifies the claim of this proposition. $\square$

6.3 Gluing theory for Fukaya–Floer homology

One of the primary goals of this subsection is to define the Fukaya–Floer homology for an N -admissible pair (Y, γ) and an $(N-1)$ -tuple $L = (l_2, \dots, l_N)$ of the elements of $H_1(Y)$. As it is mentioned in Section 3.3, we shall construct a chain complex $(\mathfrak{C}_*^{N,j}(Y, \gamma, L), d_{N,j})$ over $R_{N,j}$, for each nonnegative integer number j . We also define a chain map

$$F_j^k: \mathfrak{C}_*^{N,j}(Y, \gamma, L) \rightarrow \mathfrak{C}_*^{N,k}(Y, \gamma, L) \quad \text{for } j \geq k \geq 0$$

such that F_j^k is a homomorphism of $R_{N,j}$ -modules, and for a triple of integers satisfying $j \geq k \geq l \geq 0$, the map $F_k^l \circ F_j^k$ is chain homotopy equivalent to F_j^l .

Let $\mathbb{I}_*^{N,j}(Y, \gamma, L)$ be the homology of the chain complex $(\mathfrak{C}_*^{N,j}(Y, \gamma, L), d_{N,j})$ and $f_j^k: \mathbb{I}_*^{N,j}(Y, \gamma, L) \rightarrow \mathbb{I}_*^{N,k}(Y, \gamma, L)$ be the map induced by F_j^k . Then the Fukaya–Floer homology group $\mathbb{I}_*^N(Y, \gamma, L)$ is defined to be the inverse limit of the inverse system $(\{\mathbb{I}_*^{N,j}(Y, \gamma, L)\}_j, \{f_j^k\}_{j \geq k})$.

For simplicity of exposition, we assume that $l_3 = \cdots = l_N = 0$ and l_2 is an integral homology class. Later we shall explain how the definition should be adapted to the arbitrary case. For each $i \in \mathbb{N}$, let η_i be an oriented closed curve that represents l_2 . Let also $\nu(\eta_i)$ be a regular neighborhood of η_i such that the inclusion of $\nu(\eta_i)$ into Y induces a surjective map at the level of fundamental groups. Furthermore, we assume that the open sets $\nu(\eta_i)$ are disjoint.

For a fixed integer $j \geq 0$, let CS_{π_j} be a perturbation of the Chern–Simons functional associated to (Y, γ) such that all critical points of CS_{π_j} are irreducible and nondegenerate, and all moduli spaces $\mathcal{M}_p(\alpha, \beta)$ with dimension at most $2j+1$ consist of regular points. If i is a nonnegative integer number not greater than j and $\dim(\mathcal{M}_p(\alpha, \beta)) \leq 2j+1$, then we may also assume that the restriction of any element in $\mathcal{M}_p(\alpha, \beta)$ to the open subspace $\nu(\eta_i) \times (0, 1)$ is irreducible.¹⁰ We define

$$(6.42) \quad \mathfrak{C}_*^{N,j}(Y, \gamma, L) := \mathfrak{C}_*^{\pi_j}(Y, \gamma) \otimes R_{N,j}.$$

Next, we need to define the differential $d_{N,j}$.

For any path p between two critical points α and β of CS_{π_j} and $S = \{i_1, \dots, i_k\} \subset \{1, \dots, j\}$, we have the diagonal $\mathbb{R}$ -action $\{\sigma_t\}_{t \in \mathbb{R}}$ on the space

$$(6.43) \quad \mathcal{M}_p(\alpha, \beta) \times (\eta_{i_1} \times \mathbb{R}) \times \cdots \times (\eta_{i_k} \times \mathbb{R}).$$

Here the action of $t \in \mathbb{R}$ maps $(x, r) \in \eta_i \times \mathbb{R}$ to $(x, r - t)$. If p is not the constant path or S is not the empty set, then this action is free and we will write $\check{\mathcal{M}}_p(\alpha, \beta; S)$ for the quotient space. Otherwise, $\check{\mathcal{M}}_p(\alpha, \beta; S)$ is defined to be empty. In the case that S is empty, the quotient space is the space $\check{\mathcal{M}}_p(\alpha, \beta)$. If $S' \subset S$, then there is an obvious projection map from $\check{\mathcal{M}}_p(\alpha, \beta; S)$ to $\check{\mathcal{M}}_p(\alpha, \beta; S')$, which is denoted by $\pi_{S \rightarrow S'}$.

We can form a partial compactification of $\check{\mathcal{M}}_p(\alpha, \beta; S)$ by defining

$$\check{\mathcal{M}}_p^+(\alpha, \beta; S) := \bigcup_k \bigcup_{\{(p_i, S_i)\}_{1 \leq i \leq k}} \prod_{i=1}^k \check{\mathcal{M}}_{p_i}(\alpha_{i-1}, \alpha_i; S_i),$$

¹⁰This is a consequence of the unique continuation property of (nonperturbed) ASD connections and the Uhlenbeck compactness theorem. For more details see [52].

where $S_i \subset S$ and $p_i: \alpha_{i-1} \rightarrow \alpha_i$ is a path between two critical points of CS_{π_j} such that the sets S_i are disjoint, their union is equal to S , and the composition $p_1 \circ \cdots \circ p_k$ is equal to p . Note that the set S_i may be empty. A sequence $u_l \in \check{\mathcal{M}}_p(\alpha, \beta; S)$ is chain convergent to $u_\infty = (u_\infty^1, \dots, u_\infty^k) \in \prod_{i=1}^k \check{\mathcal{M}}_{p_i}(\alpha_{i-1}, \alpha_i; S_i)$ if there is a sequence of k -tuples of real numbers $(t_l^1, \dots, t_l^k)$ such that

$$t_l^1 < \cdots < t_l^k, \quad \lim_{l \rightarrow \infty} t_l^{m+1} - t_l^m = \infty \quad \text{for all } 1 \leq m \leq k-1$$

and

$$\sigma_{t_l^m} \circ \pi_{S \rightarrow S_m}(u_l) \xrightarrow{C_{\text{loc}}^\infty} u_\infty^m.$$

We use this notion of convergence to define a topology on $\check{\mathcal{M}}_p^+(\alpha, \beta; S)$.

Remark 6.44 Due to the possibility of bubbling off instantons, the space $\check{\mathcal{M}}_p^+(\alpha, \beta; S)$ is not necessarily compact.

Example 6.45 Suppose $A \in \mathcal{M}_p(\alpha, \beta)$, where $p: \alpha \rightarrow \beta$ is a nontrivial path between two critical points of the perturbed Chern–Simons functional. Suppose also $x \in \eta_1$ is fixed. Then $u_i = [A, (x, i)]$ defines a sequence of elements of $\check{\mathcal{M}}_p(\alpha, \beta; S)$ where $S = \{1\}$. This sequence is convergent to (u_∞^1, u_∞^2) :

$$u_\infty^1 := [A] \in \check{\mathcal{M}}_p(\alpha, \beta) \quad \text{and} \quad u_\infty^2 := [\pi^*(\beta), (x, 0)] \in \check{\mathcal{M}}_q(\beta, \beta; S).$$

Here q is the constant path from β to β , and $\pi^*(\beta)$ is the pullback of the connection β to $\mathbb{R} \times Y$.

For $S = \{i_1, \dots, i_k\}$, define the map

$$(6.46) \quad \Phi_S: \check{\mathcal{M}}_p(\alpha, \beta; S) \rightarrow \prod_{i \in S} \mathcal{B}^*(v(\eta_i) \times (0, 1)) \times \eta_i$$

by

$$\begin{aligned} \Phi_S([A, (x_{i_1}, t_{i_1}), \dots, (x_{i_k}, t_{i_k})]) \\ = ((T_{t_{i_1}}^* A|_{v(\eta_{i_1}) \times (0, 1)}, x_{i_1}), \dots, (T_{t_{i_k}}^* A|_{v(\eta_{i_k}) \times (0, 1)}, x_{i_k})), \end{aligned}$$

where $T_{t_i}: \mathbb{R} \times Y \rightarrow \mathbb{R} \times Y$ denotes the translation that maps (x, t) to $(x, t + t_i)$. This map extends to $\check{\mathcal{M}}_p^+(\alpha, \beta; S)$ in the obvious way and the extension is continuous. As in Section 2.1, we can form a universal $\text{PU}(N)$ -bundle $\mathbb{P}_i$ and an associated $\text{SU}(N^N)$ vector bundle $\mathbb{E}_i$ on $\mathcal{B}^*(v(\eta_i) \times (0, 1)) \times \eta_i$. We need to find a subspace of $\mathcal{B}^*(v(\eta_i) \times (0, 1)) \times \eta_i$ representing $c_2(\mathbb{P}_i)$ or equivalently $(1/N^N)c_2(\mathbb{E}_i)$, and use the

inverse image of this representative to cut down the moduli space $\check{\mathcal{M}}_p(\alpha, \beta; S)$. We can proceed as in [Section 2.1](#). The rank stratification of the vector bundle $\text{Hom}(\mathbb{C}^{N^N-1}, \mathbb{E}_i)$ determines a codimension-four representative $V_2(\eta_i)$ for the second Chern class of the universal bundle. For all choices of α , β , S and p , we may assume that Φ_S in [\(6.46\)](#) is transversal to

$$V_2(\eta_{i_1}) \times \cdots \times V_2(\eta_{i_k}).$$

Let $\check{\mathcal{N}}_p(\alpha, \beta; S)$ be the inverse image of the above space by the map Φ_S . Then the closure of $\check{\mathcal{N}}_p(\alpha, \beta; S)$ in $\check{\mathcal{M}}_p^+(\alpha, \beta; S)$, denoted by $\check{\mathcal{N}}_p^+(\alpha, \beta; S)$, is also a stratified space with smooth strata of the form

$$\prod_{i=1}^k \check{\mathcal{N}}_{p_i}(\alpha_{i-1}, \alpha_i; S_i).$$

Lemma 6.47 *If the dimension of $\check{\mathcal{N}}_p^+(\alpha, \beta; S)$ is at most 1, then this space is compact.*

Proof This is a consequence of [Remark 6.6](#) and the standard dimension counting argument used for the definition of the polynomial invariants for closed 4-manifolds. $\square$

Now we are in a position to define the differential $d_{N,j}$. Firstly consider the operator

$$d_S: \mathfrak{C}_*^{\pi_j}(Y, \gamma) \rightarrow \mathfrak{C}_*^{\pi_j}(Y, \gamma), \quad d_S(\alpha) = \sum_{p: \alpha \rightarrow \beta} \# \check{\mathcal{N}}_p^+(\alpha, \beta; S) \cdot \beta,$$

where the sum is over all paths p from α to other critical points of CS_{π_j} such that the dimension of the moduli space $\mathcal{M}_p(\alpha, \beta)$ is equal to $2|S| + 1$. The cut-down moduli space $\check{\mathcal{N}}_p^+(\alpha, \beta; S)$ is 0-dimensional and hence it is compact by [Lemma 6.47](#). Finally, the Fukaya–Floer differential $d_{N,j}$ is defined as

$$d_{N,j} := \sum_{S \subset \mathbb{N}^+} \frac{\prod_{i \in S} t_{2,i}}{N^{N|S|}} d_S.$$

The term $N^{N|S|}$ appears in the above expression due to the relationship between $c_2(\mathbb{P}_i)$ and $c_2(\mathbb{E}_i)$.

Proposition 6.48 *The map $d_{N,j}$ defines a differential; ie $d_{N,j}^2 = 0$.*

Proof For critical points α and β of CS_{π_j} and $S \subset \mathbb{N}$, let $h_S(\alpha, \beta)$ be a number that satisfies the identity

$$d_{N,j}^2 \alpha = \sum_{\beta, S} \frac{m_S(\alpha, \beta)}{N^{N|S|}} \left(\prod_{i \in S} t_{2,i} \right) \beta.$$

Fix α , β and $S \subseteq \{1, \dots, k\}$, and suppose $p: \alpha \rightarrow \beta$ is a path such that $\dim(\mathcal{M}_p(\alpha, \beta))$ is equal to $2|S| + 2$. Then the term $m_S(\alpha, \beta)$ is equal to

$$m_S(\alpha, \beta) = \sum_{\substack{p_1: \alpha \rightarrow \gamma \\ p_2: \gamma \rightarrow \beta}} \sum_{S_1, S_2} \# \check{\mathcal{N}}_{p_1}^+(\alpha, \gamma; S_1) \cdot \# \check{\mathcal{N}}_{p_2}^+(\gamma, \beta; S_2)$$

with $p_1 \circ p_2 = p$, $S_1 \cup S_2 = S$, S_1 and S_2 disjoint,

$$\dim(\mathcal{M}_{p_1}(\alpha, \gamma)) = 2|S_1| + 1 \quad \text{and} \quad \dim(\mathcal{M}_{p_2}(\gamma, \beta)) = 2|S_2| + 1.$$

Therefore, $m_S(\alpha, \beta)$ is equal to the signed count of the boundary points of the compact 1-manifold $\check{\mathcal{N}}_p^+(\alpha, \beta; S)$. This implies that $m_S(\alpha, \beta)$ is equal to zero. $\square$

The above discussion can be generalized in a straightforward way to the case of an arbitrary $(N-1)$ -tuple $(l_2, \dots, l_N)$. For each l_k , we choose a sequence of disjoint representatives $\{\eta_{k,i}\}_{i \in \mathbb{N}}$ and the open neighborhoods $v(\eta_{k,i})$. We keep assuming that $\eta_{k,i}$ is an oriented simple closed curve in Y , the open sets $v(\eta_{k,i})$ are disjoint and the inclusion of $v(\eta_{k,i})$ in Y induces a surjective map at the level of fundamental groups. In a more general case that l_i is a homology class with complex coefficients, we need to consider the straightforward generalization that $\eta_{k,i}$ is a linear combination of disjoint closed curves. Then for each $(N-1)$ -tuple $\bar{S} = (S_2, \dots, S_N)$ of subsets of $\{1, 2, \dots, j\}$, we can form the moduli space $\check{\mathcal{M}}_p(\alpha, \beta; \bar{S})$ and its partial compactification $\check{\mathcal{M}}_p^+(\alpha, \beta; \bar{S})$. As in the previous case, each stratum of the partial compactification is a product of moduli spaces of the form $\check{\mathcal{M}}_p(\alpha, \beta; \bar{S})$. The next step is to cut down $\check{\mathcal{M}}_p^+(\alpha, \beta; \bar{S})$ with divisors $V_k(\eta_{k,i})$ for $i \in S_k$. As in the case of closed 4-manifolds, we can use the auxiliary complex vector bundles of rank N^N associated to the universal $\text{PU}(N)$ -bundle and construct geometric representatives for $V_k(\eta_{k,i})$. The desired representatives are linear combinations of codimension $2k$ stratified spaces with even-dimensional strata. We will write $\check{\mathcal{N}}_p^+(\alpha, \beta; \bar{S})$ for the cut-down moduli space. A generic choice of these divisors allow us to obtain a cut-down moduli space. This space is compact when its dimension is less than or equal to 1.

The 0-dimensional cut-down moduli spaces $\check{\mathcal{N}}_p^+(\alpha, \beta; \bar{S})$ can be used to define an operator $d_{\bar{S}}$ acting on $\mathfrak{C}_*^{\pi_j}(Y, \gamma)$:

$$d_{\bar{S}}\alpha = \sum_{p: \alpha \rightarrow \beta} \# \check{\mathcal{N}}_p^+(\alpha, \beta; \bar{S}) \cdot \beta.$$

We combine these operators to form

$$d_{N,j} := \sum_{\bar{S}=(S_2, \dots, S_N)} \left(\prod_{i \in S_k} t_{k,i} \right) d_{\bar{S}}.$$

As in the previous case, the 1-dimensional moduli spaces can be used to show that $d_{N,j}^2 = 0$.

Before giving the definition of the maps F_j^k , we study the functorial properties of the chain complex $(\mathfrak{C}_*^{N,j}(Y, \gamma, L), d_{N,j})$. Let (X, w) be a pair whose boundary is (Y, γ) . Suppose that $\Gamma^2, \dots, \Gamma^N$ are properly embedded surfaces in X such that $[\partial\Gamma^i] = l_i \in H_1(Y)$. For $z \in \mathbb{A}(X)^{N-1}$, we want to define the relative invariant

$$D_{X,w}^{N,j}(ze^{\sum_i \Gamma_{(i)}^i}) \in \mathbb{I}_*^{N,j}(Y, \gamma, L),$$

where $L = (l_2, \dots, l_N)$. This is similar to the definition of the differential of the Fukaya–Floer chain complex. For simplicity of exposition, we assume that $\Gamma^3 = \dots = \Gamma^N = 0$, and $z = 1$. As in the case of the definition of the differential in Fukaya–Floer homology, the more general case is just slightly different. In the manifold X^+ with a cylindrical end, let $\{\Sigma_i\}_{i \in \mathbb{N}}$ be a sequence of surfaces which are given by perturbing the surface Γ^2 . We assume that these surfaces intersect generically and their intersections with $\mathbb{R}^{\geq 0} \times Y$ are equal to the disjoint product surfaces $\{\eta_i \times \mathbb{R}^{\geq 0}\}_{i \in \mathbb{N}}$. We choose a holonomy perturbation of the ASD equation on X^+ , compatible with the chosen perturbation CS_{π_j} of the Chern–Simons functional of Y , such that all moduli spaces $\mathcal{M}_p(X, w; \beta)$ consists of regular points. Given a subset $S = \{i_1, \dots, i_k\} \subset \{1, \dots, j\}$ and a path p along (X, w) based at the critical point β of CS_{π_j} , consider the space

$$(6.49) \quad \mathcal{M}_p(X, w; \beta, S) := \mathcal{M}_p(X, w; \beta) \times \Sigma_{i_1} \times \dots \times \Sigma_{i_k}.$$

For $S' \subset S$, the obvious projection map from $\mathcal{M}_p(X, w; \beta, S)$ to $\mathcal{M}_p(X, w; \beta, S')$ is denoted by $\pi_{S \rightarrow S'}$. There is also a partially defined translation map for $\mathcal{M}_p(X, w; \beta, S)$. Suppose $\mathcal{M}_p^{\text{cyl}}(X, w; \beta, S)$ denotes the following subset of $\mathcal{M}_p(X, w; \beta, S)$:

$$\mathcal{M}_p(X, w; \beta) \times (\eta_{j_1} \times \mathbb{R}^{\geq 0}) \times \dots \times (\eta_{j_k} \times \mathbb{R}^{\geq 0}).$$

For $u = ([A], (x_1, t_1), \dots, (x_k, t_k)) \in \mathcal{M}_p^{\text{cyl}}(X, w; \beta, S)$ define $\sigma_t(u)$ to be the element

$$([T_t^*(A|_{Y \times \mathbb{R}^{\geq 0}})], (x_1, t_1 - t), \dots, (x_k, t_k - t)).$$

Note that the connection $T_t^*(A|_{Y \times \mathbb{R}^{\geq 0}})$ is partially defined on the cylinder $Y \times \mathbb{R}$.

As in the cylinder case, we form a partial compactification of $\mathcal{M}_p(X, w; \beta, S)$ given by

$$\mathcal{M}_p^+(X, w; \beta, S) := \bigcup_k \bigcup_{\{(p_i, S_i)\}_{1 \leq i \leq k}} \mathcal{M}_{p_1}(X, w, \alpha_1, S_1) \times \prod_{i=2}^k \check{\mathcal{M}}_{p_i}(\alpha_{i-1}, \alpha_i; S_i),$$

where $\alpha_1, \dots, \alpha_k$ are critical points of CS_{π_j} , $\alpha_k = \beta$, p_1 is a path along (X, w) , $p_i: \alpha_{i-1} \rightarrow \alpha_i$ is a path along the cylinder, $p = p_1 \circ \dots \circ p_k$, the sets S_i are disjoint and their union is equal to S . As before, S_i may be empty. In a little more detail, a sequence $u_l \in \mathcal{M}_p(X, w; \beta, S)$ is chain convergent to $u_\infty = (u_\infty^1, \dots, u_\infty^k) \in \mathcal{M}_{p_1}(X, w, \alpha_1, S_1) \times \prod_{i=2}^k \check{\mathcal{M}}_{p_i}(\alpha_{i-1}, \alpha_i; S_i)$ if there is a sequence of $(k-1)$ -tuples of real numbers $(t_l^2, \dots, t_l^k)$ such that

$$t_l^1 := 0 < t_l^2 < \dots < t_l^k, \quad \lim_{l \rightarrow \infty} t_l^{m+1} - t_l^m = \infty \quad \text{for all } 1 \leq m \leq k-1$$

and

$$\pi_{S \rightarrow S_1}(u_l) \xrightarrow{C_{\text{loc}}^\infty} u_\infty^1, \quad \sigma_{t_l^i} \circ \pi_{S \rightarrow S_i}(u_l) \xrightarrow{C_{\text{loc}}^\infty} u_\infty^i \quad \text{for } i \geq 2.$$

Here part of the assumption is that $\sigma_{t_l^i} \circ \pi_{S \rightarrow S_i}(u_l)$ is well defined for $i \geq 2$. That is to say, $\pi_{S \rightarrow S_i}(u_l) \in \mathcal{M}_p^{\text{cyl}}(X, w; \beta, S_i)$.

We momentarily assume that $S = \{1\}$. Suppose $v(\Sigma_1)$ is an open neighborhood of Σ_1 such that the inclusion map of $v(\Sigma_1)$ induces a surjective map. Suppose also that $\mathcal{B}^{**}(v(\Sigma_1))$ is the set of connections on $v(\Sigma_1)$ whose restrictions to the sets of the form $v(\eta_1) \times (t-1, t)$ are irreducible. We can assume that the perturbation π_j is small enough that the restriction map $r_1: \mathcal{M}_p(X, w; \beta, S) \rightarrow \mathcal{B}^{**}(v(\Sigma_1)) \times \Sigma_1$ is well defined.¹¹ We can form a universal $\text{PU}(N)$ -bundle $\mathbb{P}_1$ and the associated $\text{SU}(N^N)$ -bundle $\mathbb{E}_1$ on $\mathcal{B}^{**}(v(\Sigma_1)) \times \Sigma_1$. Our goal is to define a geometric representative for $c_2(\mathbb{P}_1)$ or equivalently $(1/N^N)c_2(\mathbb{E}_1)$ which is compatible with our choice of the geometric representative in the case of cylinders. Note that $\mathcal{M}_p(X, w; \beta, S)$ is the union of the two sets

$$B_1 = \mathcal{M}_p(X, w; \beta) \times (\eta_1 \times [1, \infty)) \quad \text{and} \quad B_2 = \mathcal{M}_p(X, w; \beta) \times (\Sigma_1 \setminus (\eta_1 \times (2, \infty))).$$

The map $r_1|_{B_1}$ can be composed with the map

$$F: \mathcal{B}^{**}(v(\Sigma_1)) \times (\eta_1 \times [1, \infty)) \rightarrow \mathcal{B}^*(v(\eta_1) \times (0, 1)) \times \eta_1, \\ F([A], (x, t)) = ([A|_{v(\Sigma) \times (t-1, t)}], x).$$

Therefore, we can choose the geometric representative $V_2(\Sigma_1) \subset \mathcal{B}^{**}(v(\Sigma_1)) \times \Sigma_1$ for $c_2(\mathbb{E}_1)$ such that

$$V_2(\Sigma_1) \cap (\mathcal{B}^{**}(v(\Sigma_1)) \times (\eta_1 \times [1, \infty))) = F^{-1}(V_2(\eta_1)).$$

Arguing as in [52], we can also assume that $V_2(\Sigma_1)$ is transversal to the map r_1 for all choices of the path p . In this process, first we slightly modify $V_2(\eta)$ to make

¹¹This is again a consequence of unique continuation and the Uhlenbeck compactness theorem [52].

the map $r_1|_{B_1}$ transversal. Then we extend $F^{-1}(V_2(\eta_1))$ to $\mathcal{B}^{**}(v(\Sigma_1)) \times \Sigma_1$ so that $r_1|_{B_2}$ is also transversal. We will write $\mathcal{N}_p(X, w; \beta, S)$ for the cut-down moduli space $r_1^{-1}(V_2(\Sigma_1))$. A similar construction can be used to define $\mathcal{N}_p(X, w; \beta, S)$ in the case that S has more than one element. The closure of $\mathcal{N}_p(X, w; \beta, S)$ in $\mathcal{M}_p^+(X, w; \beta, S)$, denoted by $\mathcal{N}_p^+(X, w; \beta, S)$, is also a stratified space with smooth strata of the form

$$\mathcal{N}_{p_1}(X, w, \alpha_1, S_1) \times \prod_{i=2}^k \check{\mathcal{N}}_{p_i}(\alpha_{i-1}, \alpha_i; S_i).$$

As in the cylinder case, the cut-down moduli space $\mathcal{N}_p^+(X, w; \beta, S)$ is compact when its dimension is at most one. The relative invariant $D_{X,w}^N(e^{\Gamma(2)})$ is defined using 0-dimensional moduli spaces as below:

$$(6.50) \quad D_{X,w}^{N,j}(e^{\Gamma(2)}) := \sum_{S \subset \{1, \dots, j\}} \frac{\prod_{i \in S} t_{2,i}}{N^{|S|}} (\#\mathcal{N}_p^+(X, w; \beta, S)) \cdot \beta \in \mathfrak{C}_*^{N,j}(Y, \gamma, L).$$

Following the proof of [Proposition 6.48](#), we can show that (6.50) determines a cycle in $\mathfrak{C}_*^{N,j}(Y, \gamma, L)$. We will denote the corresponding element in $\mathbb{I}_*^{N,j}(Y, \gamma, L)$ with the same notation.

The definition of the relative element in (6.50) can be extended to similar situations. For example, suppose (Y_0, γ_0) and (Y_1, γ_1) are two N -admissible pairs and $L_i = (l_i^2, \dots, l_i^N)$ is an $(N-1)$ -tuple of the elements in $H_1(Y_i)$. Suppose also (W, w, z) is a morphism from (Y_0, γ_0) to (Y_1, γ_1) , and Γ^j is a properly embedded surface in W such that $[\partial_i \Gamma^j] = l_i^j$. Then there is a chain map,

$$\mathfrak{C}^{N,j}(W, w, ze^{\Gamma(2)} + \dots + \Gamma_{(N)}^N): \mathfrak{C}_*^{N,j}(Y_0, \gamma_0, L_0) \rightarrow \mathfrak{C}_*^{N,j}(Y_1, \gamma_1, L_1),$$

which induces a map at the level of homology,

$$\mathbb{I}_*^{N,j}(W, w, ze^{\Gamma(2)} + \dots + \Gamma_{(N)}^N): \mathbb{I}_*^{N,j}(Y_0, \gamma_0, L_0) \rightarrow \mathbb{I}_*^{N,j}(Y_1, \gamma_1, L_1).$$

Alternatively, if (Y, γ, L) is as above, (X, w) is a 4-manifold whose boundary is equal to $(\bar{Y}, \bar{\gamma})$, and Γ^j is an embedded surface such that $[\partial \Gamma^j] = -l_j$, then we have an $R_{N,j}$ -linear map

$$D_{N,j}^{X,w}(ze^{\Gamma(2)} + \dots + \Gamma_{(N)}^N): \mathbb{I}_*^{N,j}(Y, \gamma, L) \rightarrow R_{N,j}.$$

Now we are ready to define the maps F_j^k for a triple (Y, γ, L) . For a pair of nonnegative integers j and k with $j \geq k$, we have chosen perturbations CS_{π_j} and CS_{π_k} of the Chern–Simons functional of the pair (Y, γ) . Since $j \geq k$, the functional CS_{π_j}

satisfies all the properties that we required for CS_{π_k} . In particular, the chain complex $(\mathfrak{C}_{*}^{N,j}(Y, \gamma, L) \otimes R_{N,k}, d_{N,j})$ gives an alternative chain complex to define $\mathbb{I}_{*}^{N,k}(Y, \gamma, L)$. The functoriality mentioned in the previous paragraph implies that there is a chain map

$$\mathfrak{C}^{N,j}([0, 1] \times Y, [0, 1] \times \gamma, e^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N}): \mathfrak{C}_{*}^{\pi_j}(Y, \gamma) \otimes R_{N,k} \rightarrow \mathfrak{C}_{*}^{\pi_k}(Y, \gamma) \otimes R_{N,k},$$

where Γ^i is the surface $[0, 1] \times \eta_{i,1}$. The map F_j^k is the composition of the above map and the obvious map from $\mathfrak{C}_{*}^{\pi_j}(Y, \gamma) \otimes R_{N,j}$ to $\mathfrak{C}_{*}^{\pi_j}(Y, \gamma) \otimes R_{N,k}$. A standard argument shows that $F_k^l \circ F_j^k$ is chain homotopy equivalent to F_j^l for any triple $j \geq k \geq l \geq 0$. Therefore, the maps f_j^k , induced by the maps F_j^k , have the required properties for an inverse system. This completes the definition of Fukaya–Floer homology $\mathbb{I}_{*}^N(Y, \gamma, L)$ for the triple. It is also standard to show that this R_N –module is independent of the choices that were made. The homomorphism f_j^k maps $D_{X,w}^{N,j}(ze^{\sum_i \Gamma_{(i)}^i}) \in \mathbb{I}_{*}^{N,j}(Y, \gamma, L)$ to $D_{X,w}^{N,k}(ze^{\sum_i \Gamma_{(i)}^i}) \in \mathbb{I}_{*}^{N,k}(Y, \gamma, L)$. Thus we have an induced element of $\mathbb{I}_{*}^N(Y, \gamma, L)$, which we denote by $D_{X,w}^N(ze^{\sum_i \Gamma_{(i)}^i})$. Similarly, we can use the construction of the previous paragraph to define $\mathbb{I}_{*}^N(W, w, ze^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N})$ and $D_N^{X,w}(ze^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N})$.

Now we are ready to give a proof of (3.30). For the convenience of the reader, we restate the claim as the following proposition:

Proposition 6.51 *Let (X_1, w_1) be a pair whose boundary is equal to an admissible pair (Y, γ) . Let (X_2, w_2) be another pair whose boundary is equal to $(\bar{Y}, \bar{\gamma})$. Let $z_1 \in \mathbb{A}(X_1)^{\otimes(N-1)}$ and $z_2 \in \mathbb{A}(X_2)^{\otimes(N-1)}$. Let Γ^j be a properly embedded surface in X_1 with boundary l_j . Let Λ^j be a properly embedded surface in X_2 whose boundary is equal to l_j with the reverse orientation. Then we can form the closed 4–manifold $X_2 \circ X_1$ and the 2–cycle $w_2 \circ w_1$. The embedded surfaces Γ^j and Λ^j can be glued to each other along their boundary to form a closed surface $\Gamma^j \# \Lambda^j$. Then the invariant*

$$(6.52) \quad D_{X_2 \circ X_1, w_2 \circ w_1}^N(z_1 \cdot z_2 \cdot e^{(\Gamma^2 \# \Lambda^2)_{(2)} + \dots + (\Gamma^N \# \Lambda^N)_{(N)}}) \in \mathbb{C}[[t_2, \dots, t_N]] \subset R_N$$

of the closed 4–manifold $X_2 \circ X_1$ is equal to

$$D_N^{X_2, w_2}(z_2 \cdot e^{\Lambda_{(2)}^2 + \dots + \Lambda_{(N)}^N}) \circ D_{X_1, w_1}^N(z_1 \cdot e^{\Gamma_{(2)}^2 + \dots + \Gamma_{(N)}^N}).$$

Proof We make simplifying assumptions as before: assume $z_1 = z_2 = 1$ and $\Gamma^3, \dots, \Gamma^N$ and $\Lambda^3, \dots, \Lambda^N$ are empty. Choose two series of properly embedded surfaces $\{\Sigma_i\}_{i \in \mathbb{N}} \subset X_1$ and $\{T_i\}_{i \in \mathbb{N}} \subset X_2$ such that Σ_i (respectively, T_i) is given by perturbing Γ^2 (respectively, Λ^2). We also assume that $\partial \Sigma_i$ (respectively, ∂T_i) is

equal to η_i (respectively, η_i with the reverse orientation), the curves η_i are disjoint and the embedded surfaces Σ_i and T_i intersect generically. We fix metrics on X_1 and X_2 which are product metrics in a neighborhood of their boundaries corresponding to a fixed metric on Y . For each nonnegative integer j , we prove that the image of the element in (6.52) in $R_{N,j}$ is equal to

$$D_{N,j}^{X_2,w_2}(e^{\Lambda^2_{(2)}}) \circ D_{X_1,w_1}^{N,j}(e^{\Gamma^2_{(2)}}).$$

To achieve this goal, we proceed as before to define open sets $v(\eta_i) \subset Y$, $v(\Sigma_i) \subset X_1^+$ and $v(T_i) \subset X_2^+$ and the geometric representatives

$$\begin{aligned} V_2(\eta_i) &\subset \mathcal{B}^*(v(\eta_i) \times (0, 1)) \times \eta_i, \\ V_2(\Sigma_i) &\subset \mathcal{B}^{**}(v(\Sigma_i)) \times \Sigma_i, \\ V_2(T_i) &\subset \mathcal{B}^{**}(v(T_i)) \times T_i \end{aligned}$$

suitable for the definition of the differential $d_{N,j}$ of the Fukaya–Floer chain complex and the relative elements $D_{X_1,w_1}^{N,j}(e^{\Gamma^2_{(2)}})$ and $D_{N,j}^{X_2,w_2}(e^{\Lambda^2_{(2)}})$.

Suppose X^T is the metric on $X_2 \circ X_1$ induced by the metrics on X_1 and X_2 with a neck of length T along Y . Suppose $w \subset X^T$ is the 2–cycle induced by w_1 and w_2 , and $R_i^T \subset X^T$ is the embedded surface induced by the surfaces Σ_i and T_i . The 4–manifold X can be decomposed into the three pieces

$$X_1, \quad X_2 \quad \text{and} \quad Y \times \left[-\frac{1}{2}T, \frac{1}{2}T\right].$$

We assume that X_1 and X_2 are disjoint and that they intersect $Y \times \left[-\frac{1}{2}T, \frac{1}{2}T\right]$ in $Y \times \left[-\frac{1}{2}T, -\frac{1}{2}T + 2\right]$ and $Y \times \left[\frac{1}{2}T - 2, \frac{1}{2}T\right]$, respectively. The Riemann surface R_i^T can be decomposed into a union of the three sets

$$\Sigma_i^c \subset X_1, \quad T_i^c \subset X_2 \quad \text{and} \quad \eta_i \times \left[-\frac{1}{2}T + 1, \frac{1}{2}T - 1\right] \subset Y \times \left[-\frac{1}{2}T + 1, \frac{1}{2}T - 1\right].$$

Suppose $\mathcal{B}_\kappa^{**}(X^T, w)$ is the subset of $\mathcal{B}_\kappa^*(X, w)$ which consists of connections whose restriction to any set of the following form is irreducible:

$$v(\eta_i) \times (t - 1, t), \quad -T + 1 < t < T.$$

Suppose κ is chosen such that the dimension of the moduli space $\mathcal{M}_\kappa(X^T, w)$ is at most $2j$. Then unique continuation and gluing theory show that for large enough values of T , the moduli space $\mathcal{M}_\kappa(X^T, w)$ is a subset of $\mathcal{B}_\kappa^{**}(X^T, w)$. Similar to the case of 4–manifolds with cylindrical ends, $V_2(\eta_i)$, $V_2(\Sigma_i)$ and $V_2(T_i)$ can be used to define a geometric representative $V(R_i^T) \subset \mathcal{B}_\kappa^{**}(X^T, w) \times v(R_i^T)$. Another application

of gluing theory shows that for large enough values of T , these divisors determine a transversal cut of $\mathcal{M}_\kappa(X^T, w) \times R_{i_1}^T \times \cdots \times R_{i_k}^T$ for any set $S = \{i_1, \dots, i_k\} \subset \{1, \dots, j\}$. In the case that the cut-down moduli space is zero-dimensional, it can be identified, for large values of T , with the set

$$\bigcup_{p_1, p_2, S_1, S_2} \mathcal{N}_{p_1}(X_1, w_1; \alpha, S_1) \times \mathcal{N}_{p_2}(X_2, w_2; \alpha, S_2),$$

where each p_i is a loop along (X_i, w_i) based at the connection α such that $\kappa = \kappa(p_1) + \kappa(p_2)$. Moreover, S_1 and S_2 are disjoint sets with $S = S_1 \cup S_2$. These geometric results for different choices of S can be translated to the algebraic identity

(6.53)
$$\begin{aligned} D_{N,j}^{X_2,w_2}(e^{\Lambda^2_{(2)}}) \circ D_{X_1,w_1}^{N,j}(e^{\Gamma^2_{(2)}}) &= \sum_{S \subset \{1,\dots,j\}} \left(\prod_{i \in S} t_{2,i} \right) D_{X^T,w}^N \left(\prod_{i \in S} (R_i^T)_{(2)} \right) \\ &= \sum_{S \subset \{1,\dots,j\}} \left(\prod_{i \in S} t_{2,i} \right) D_{X^T,w}^N ((\Gamma^2 \# \Lambda^2)_{(2)}^{|S|}). \end{aligned}$$

In the second equality, we used the fact that R_j^T represents the homology class of $\Gamma^2 \# \Lambda^2$. The term in (6.53) is equal to the image of $D_{X^T,w}^N(e^{(\Gamma^2 \# \Lambda^2)_{(2)}})$ in $R_{N,j}$. $\square$

7 Questions and conjectures

In this section, we propose some questions and conjectures for future directions. This section is divided into two parts: the first subsection is concerned with the polynomial invariants of 4-manifolds. In the second part, we discuss some conjectures related to the algebra $\mathbb{V}_{g,d}^N$.

7.1 Structure of polynomial invariants and 4-manifolds with simple type

In Section 2.5, the definition of the simple type property of 4-manifolds uses $U(3)$ -polynomial invariants. As we pointed out earlier, the definition is motivated by Kronheimer and Mrowka’s simple type property, defined using $U(2)$ -polynomial invariants [54]. There is another version of simple type property defined by Seiberg–Witten invariants. It is unknown whether there are examples of smooth 4-manifolds with $b^+ \geq 2$ which do not have Kronheimer–Mrowka simple type or Seiberg–Witten simple type. It is also shown in [23] that many 4-manifolds with Seiberg–Witten simple type have Kronheimer–Mrowka simple type. Therefore, it is natural to ask whether

there is any relationship among $U(3)$ –simple type, Kronheimer–Mrowka simple type and Seiberg–Witten simple type. A more challenging question would be to investigate whether there is a 4–manifold with $b^+ \geq 2$ which does not have $U(3)$ –simple type. A more approachable question is the following:

Question 7.1 *What is the analogue of the simple type condition with respect to $U(N)$ –polynomial invariants?*

As in the $U(2)$ and the $U(3)$ case, the simple condition has to be formulated in terms of point classes. In light of [Proposition 3.24](#), it is plausible that one of the required conditions is

$$D_{X,w}^N(a_2^N z) = N^N D_{X,w}^N(z).$$

For $N = 2, 3$, the blowup formula for $U(N)$ simple type manifolds has a simpler form [\[26; 12\]](#). One might hope that the same holds for higher values of N and follow this direction to gain more insights into the correct definition for the simple type condition. The physics literature [\[64; 22\]](#) suggests that the blowup formula for an arbitrary N is related to function theory on a hyperelliptic curve with coefficients in $\mathbb{Q}[a_2, \dots, a_N]$. Evaluation of $(a_2, \dots, a_N)$ determines a hyperelliptic curve on the complex numbers, and the simple type condition is related to the evaluations that produce a fully degenerate curve.

The relationship between $U(2)$ –polynomial invariants and Seiberg–Witten invariants goes beyond the simple type conditions. In [\[54\]](#), Kronheimer and Mrowka prove that $U(2)$ –polynomial invariants are completely determined by a finite set of cohomology classes (known as Kronheimer–Mrowka basic classes) and a set of rational numbers, one for each basic class. In [\[84\]](#), Witten argues that basic classes and corresponding rational numbers can be determined in terms of Seiberg–Witten invariants. To be more detailed, recall that for each spin^c structure $\mathfrak{s}$ on a 4–manifold X (satisfying appropriate conditions, such as $b^+(X) \geq 2$) there is a Seiberg–Witten invariant $\text{SW}_X(\mathfrak{s})$. Then Witten’s conjecture states that any basic class of X is equal to $c_1(S_{\mathfrak{s}}^+)$, where $\mathfrak{s}$ is a spin^c structure with nonzero $\text{SW}_X(\mathfrak{s})$ and $S_{\mathfrak{s}}^+$ is the half-spin bundle associated to $\mathfrak{s}$. Witten’s conjecture is generalized to higher values of N in [\[64\]](#). The calculations of $U(3)$ –polynomial invariants in this paper agree with the Moore–Mariño conjecture in [\[64\]](#) and can be exploited to fix the undetermined constants in this conjecture. In particular, a modified version of the Moore–Mariño conjecture runs as follows:

Conjecture 7.2 *Let X be a four-manifold which has $U(3)$ –simple type. Let $\{K_i\}$ be the set of Kronheimer–Mrowka basic classes. Then the $U(3)$ –series of X has the form*

$$\begin{aligned} \hat{D}_{X,w}(e^{\Gamma(2)+\Lambda(3)}) \\ = e^{\frac{1}{2}Q(\Gamma)-Q(\Lambda)} \sum_{i,j} c_{ij} \zeta^{-w \cdot (\frac{1}{2}(K_i-K_j))} e^{\frac{\sqrt{3}}{2}(K_i+K_j) \cdot \Gamma + \frac{\sqrt{3}}{2}i(K_i-K_j) \cdot \Lambda}, \end{aligned}$$

where $c_{i,j}$ is given as

$$2^{\chi+\frac{3}{2}\sigma+\frac{1}{2}K_i \cdot K_j} 3^{2+\frac{7}{4}\chi+\frac{11}{4}\sigma} SW_X(s_i) SW_X(s_j).$$

Here s_i is chosen such that the associated basic class is equal to K_i .

7.2 The algebra $\mathbb{V}_{g,d}^N$

In Section 5.1, a list of simultaneous eigenvectors for the operators ϵ , $\mathfrak{N}_2$, $\mathfrak{N}_3$, $\rho_{(2)}$ and $\rho_{(3)}$, acting on $\mathbb{V}_{g,d}^3$, is constructed. We also showed that there are at least two nondegenerate simultaneous eigenvectors, which form the essential ingredient to prove the excision theorem in Section 5.2. However, we do not know whether our approach produces all simultaneous eigenvectors:

Conjecture 7.3 *Suppose $V_{g,d}^N \subset \mathbb{V}_{g,d}^N$ is the set of vectors which are invariant with respect to the action of ϵ . Then for $N = 3$, any simultaneous eigenvector of the operators acting on $V_{g,d}^N$ that are induced by $\mathfrak{N}_2$, $\mathfrak{N}_3$, $\rho_{(2)}$ and $\rho_{(3)}$ has the form $(3\zeta^{2d\beta}, 0, \zeta^{d\beta} \sqrt{3}\alpha, \zeta^{2d\beta} \sqrt{3}i\beta)$ with $(\alpha, \beta) \in \mathbb{C}_g$.*

For $N = 2$, there are three operators ϵ , $\mathfrak{N}_2$ and $\rho_{(2)}$. In this case, Muñoz obtains a complete understanding of the action of ϵ , $\mathfrak{N}_2$ and $\rho_{(2)}$ in [73]. In particular, his results show that the simultaneous eigenvectors of $\mathfrak{N}_2$ and $\rho_{(2)}$, acting on $V_{g,d}^2$, have the form

$$((-1)^r 2, \pm 2r i^{r+1}), \quad 0 \leq r \leq g-1.$$

All of these eigenvalues can be produced using the method of Proposition 5.7. Therefore, the analogue of Conjecture 7.3 holds for $N = 2$. Muñoz’s method of understanding the action of ϵ , $\mathfrak{N}_2$ and $\rho_{(2)}$ is based on the characterization of the ring structure of the cohomology ring $\mathcal{N}_{2,d}(\Sigma_g)$ in [86; 47; 78; 3], which is not available for higher values of N .

If Conjecture 7.3 holds, then we can use the method of [56] and show that $\text{SHI}_*(M, \alpha)$ is nonzero for a *taut balanced sutured manifold* [46; 56]. This nonvanishing result can be used to show that Conjecture 5.39 holds. We can also use this to show that

$\mathrm{KHI}_*(K)$ detects the genus of K . Thus the answer to [Question 1.1](#) for a nontrivial knot and $N = 3$ is positive. In fact, in order to make this series of conclusions, we need the following weaker version of [Conjecture 7.3](#):

Conjecture 7.4 *If (x, y) is a pair of simultaneous eigenvalues for $(\rho_{(2)}, \rho_{(3)})$ in $\mathbb{V}_{g,d}$, then $|x| + |y| \leq \sqrt{3}(2g - 2)$.*

There is also a symplectic analogue of the algebra $\mathbb{V}_{g,d}^N$. The manifold $\mathcal{N}_{N,d}(\Sigma_g)$ is Kähler and the associated Gromov-Witten invariants can be used to define the *quantum cohomology* ring $\mathrm{QH}^*(\mathcal{N}_{N,d})$ [75; 65]. The underlying vector space of $\mathrm{QH}^*(\mathcal{N}_{N,d})$ is $H^*(\mathcal{N}_{N,d})$ and the ring structure is also a deformation of the cup product. Thus, it has similar structure to $V_{g,d}^N = \ker(\epsilon - 1)$, and it is natural to make the following conjecture:

Conjecture 7.5 *The ring $V_{g,d}^N$ is isomorphic to $\mathrm{QH}^*(\mathcal{N}_{N,d})$.*

This conjecture for $N = 2$ is proved by Muñoz [72] using the characterization of the cohomology ring $H^*(\mathcal{N}_{N,d})$ in [86; 47; 78; 3]. The $N = 2$ special case of this conjecture was also proved using an adiabatic limit argument in [77].

Appendix Invariants of flat connections on $\Sigma(2, 3, 23)$

There are 44 irreducible flat $\mathrm{SU}(3)$ -connections on $\Sigma(2, 3, 23)$ [4]. These flat connections are determined by their holonomies along the loop x_3 in the standard presentation of the fundamental group of $\Sigma(2, 3, 23)$ (see (3.37)). For each flat connection, the conjugacy class of this holonomy is determined by its eigenvalues which have the form $e^{2\pi i k/23}$, $e^{2\pi i l/23}$ and $e^{2\pi i m/23}$. The possible values of $\{k, l, m\}$ are given in [Table 4](#).

The complex conjugation diffeomorphism of $\Sigma(2, 3, 23)$ maps a flat connection with the associated triple $\{k, l, m\}$ to the flat connection with the complementary associated triple $\{23 - k, 23 - l, 23 - m\}$. If both these triples give the same flat connection, we denote this connection with α_j for an appropriate choice of the integer j . Otherwise, the resulting connections are denoted by α_j^1 and α_j^2 .

The gauge-theoretical invariants of these flat connections are also given in [Table 4](#).

There are eight nontrivial flat $\mathrm{SU}(2)$ -connections on $\Sigma(2, 3, 23)$. As in the irreducible case, these connections are determined by the conjugacy class of their holonomies along x_3 . For each $2 \leq k \leq 9$, there is a unique flat $\mathrm{SU}(2)$ -connection on $\Sigma(2, 3, 23)$ where the eigenvalues of holonomy along x_3 are equal to $e^{2\pi i k/23}$ and $e^{-2\pi i k/23}$. We

holonomy							holonomy						
α	k	l	m	$\text{CS}(\alpha)$	ρ_{ad_α}	deg	α	k	l	m	$\text{CS}(\alpha)$	ρ_{ad_α}	deg
α_1	0	4	19	$\frac{1}{138}$	$-\frac{364}{23}$	4	α_{16}^1	2	4	17	$\frac{19}{138}$	$-\frac{476}{23}$	8
α_2	0	5	18	$\frac{49}{138}$	$-\frac{540}{23}$	0	α_{16}^2	6	19	21	$\frac{19}{138}$	$-\frac{476}{23}$	8
α_3	0	6	17	$\frac{31}{138}$	$-\frac{520}{23}$	10	α_{17}^1	2	5	16	$\frac{1}{138}$	$-\frac{456}{23}$	6
α_4	0	7	16	$\frac{85}{138}$	$-\frac{488}{23}$	2	α_{17}^2	7	18	21	$\frac{1}{138}$	$-\frac{456}{23}$	6
α_5	0	8	15	$\frac{73}{138}$	$-\frac{444}{23}$	0	α_{18}^1	2	6	15	$\frac{55}{138}$	$-\frac{516}{23}$	0
α_6	0	9	14	$\frac{133}{138}$	$-\frac{572}{23}$	8	α_{18}^2	8	17	21	$\frac{55}{138}$	$-\frac{516}{23}$	0
α_7	0	10	13	$\frac{127}{138}$	$-\frac{504}{23}$	6	α_{19}^1	2	7	14	$\frac{43}{138}$	$-\frac{472}{23}$	10
α_8	0	11	12	$\frac{55}{138}$	$-\frac{424}{23}$	10	α_{19}^2	9	16	21	$\frac{43}{138}$	$-\frac{472}{23}$	10
α_9^1	1	4	18	$-\frac{43}{138}$	$-\frac{472}{23}$	10	α_{20}^1	2	8	13	$\frac{103}{138}$	$-\frac{508}{23}$	4
α_9^2	5	19	22				α_{20}^2	10	15	21	$\frac{103}{138}$	$-\frac{508}{23}$	
α_{10}^1	1	5	17	$\frac{127}{138}$	$-\frac{412}{23}$	4	α_{21}^1	2	9	12	$\frac{97}{138}$	$-\frac{440}{23}$	2
α_{10}^2	6	18	22	$\frac{127}{138}$	$-\frac{412}{23}$		α_{21}^2	11	14	21	$\frac{97}{138}$	$-\frac{440}{23}$	
α_{11}^1	1	6	16	$\frac{7}{138}$	$-\frac{524}{23}$	8	α_{22}^1	3	4	16	$\frac{67}{138}$	$-\frac{468}{23}$	0
α_{11}^2	7	17	22	$\frac{7}{138}$	$-\frac{524}{23}$		α_{22}^2	7	19	20	$\frac{67}{138}$	$-\frac{468}{23}$	
α_{12}^1	1	7	15	$\frac{97}{138}$	$-\frac{532}{23}$	4	α_{23}^1	3	5	15	$\frac{85}{138}$	$-\frac{488}{23}$	2
α_{12}^2	8	16	22	$\frac{97}{138}$	$-\frac{532}{23}$		α_{23}^2	8	18	20	$\frac{85}{138}$	$-\frac{488}{23}$	
α_{13}^1	1	8	14	$\frac{121}{138}$	$-\frac{528}{23}$	6	α_{24}^1	3	6	14	$\frac{37}{138}$	$-\frac{404}{23}$	8
α_{13}^2	9	15	22	$\frac{121}{138}$	$-\frac{528}{23}$		α_{24}^2	9	17	20	$\frac{37}{138}$	$-\frac{404}{23}$	
α_{14}^1	1	9	13	$\frac{79}{138}$	$-\frac{420}{23}$	0	α_{25}^1	3	7	13	$\frac{61}{138}$	$-\frac{492}{23}$	0
α_{14}^2	10	14	22	$\frac{79}{138}$	$-\frac{420}{23}$		α_{25}^2	10	16	20	$\frac{61}{138}$	$-\frac{492}{23}$	
α_{15}^1	1	10	12	$\frac{109}{138}$	$-\frac{484}{23}$	4	α_{26}^1	3	8	12	$\frac{19}{138}$	$-\frac{476}{23}$	8
α_{15}^2	11	13	22	$\frac{109}{138}$	$-\frac{484}{23}$		α_{26}^2	11	15	20	$\frac{19}{138}$	$-\frac{476}{23}$	

Table 4: Holonomies and gauge-theoretical invariants of irreducible flat SU(3)–connections on $\Sigma(2, 3, 23)$.

will write β_k for this connection. The gauge-theoretical invariants of these connections are given in Table 5. In this table, $\tilde{\alpha}$ denotes the reducible SU(3)–connection associated to an SU(2)–connection α .

α	$\text{CS}(\alpha)$	ρ_{ad_α}	$\rho_{\text{ad}_{\tilde{\alpha}}}$	$\deg \alpha$	$\deg \tilde{\alpha}$
β_2	$\frac{1}{552}$	$-\frac{343}{69}$	$-\frac{206}{23}$	1	1
β_3	$\frac{169}{552}$	$-\frac{559}{69}$	$-\frac{406}{23}$	5	9
β_4	$\frac{73}{552}$	$-\frac{475}{69}$	$-\frac{410}{23}$	3	7
β_5	$\frac{265}{552}$	$-\frac{643}{69}$	$-\frac{402}{23}$	7	11

α	$\text{CS}(\alpha)$	ρ_{ad_α}	$\rho_{\text{ad}_{\tilde{\alpha}}}$	$\deg \alpha$	$\deg \tilde{\alpha}$
β_6	$\frac{193}{552}$	$-\frac{511}{69}$	$-\frac{382}{23}$	5	9
β_7	$\frac{409}{552}$	$-\frac{631}{69}$	$-\frac{534}{23}$	1	5
β_8	$\frac{361}{552}$	$-\frac{451}{69}$	$-\frac{490}{23}$	7	3
β_9	$\frac{49}{552}$	$-\frac{523}{69}$	$-\frac{434}{23}$	3	7

Table 5: Gauge-theoretical invariants of reducible flat $\text{SU}(3)$ -connections on $\Sigma(2, 3, 23)$.

References

[1] **MF Atiyah, R Bott**, *The Yang–Mills equations over Riemann surfaces*, Philos. Trans. Roy. Soc. London Ser. A 308 (1983) 523–615 [MR](#)

[2] **MF Atiyah, V K Patodi, IM Singer**, *Spectral asymmetry and Riemannian geometry, II*, Math. Proc. Cambridge Philos. Soc. 78 (1975) 405–432 [MR](#)

[3] **V Y Baranovskii**, *The cohomology ring of the moduli space of stable bundles with odd determinant*, Izv. Ross. Akad. Nauk Ser. Mat. 58 (1994) 204–210 [MR](#) In Russian; translated in [Russ. Acad. Sci. Izv. Math.](#) 45 (1995) 207–213

[4] **H U Boden**, *Unitary representations of Brieskorn spheres*, Duke Math. J. 75 (1994) 193–220 [MR](#)

[5] **H U Boden, S Friedl**, *Metabelian $\text{SL}(n, \mathbb{C})$ representations of knot groups*, Pacific J. Math. 238 (2008) 7–25 [MR](#)

[6] **H U Boden, C M Herald, P A Kirk, E P Klassen**, *Gauge theoretic invariants of Dehn surgeries on knots*, Geom. Topol. 5 (2001) 143–226 [MR](#)

[7] **A Borel**, *Topology of Lie groups and characteristic classes*, Bull. Amer. Math. Soc. 61 (1955) 397–432 [MR](#)

[8] **P J Braam, S K Donaldson**, *Floer’s work on instanton homology, knots and surgery*, from “The Floer memorial volume” (H Hofer, C H Taubes, A Weinstein, E Zehnder, editors), Progr. Math. 133, Birkhäuser, Basel (1995) 195–256 [MR](#)

[9] **P J Braam, S K Donaldson**, *Fukaya–Floer homology and gluing formulae for polynomial invariants*, from “The Floer memorial volume” (H Hofer, C H Taubes, A Weinstein, E Zehnder, editors), Progr. Math. 133, Birkhäuser, Basel (1995) 257–281 [MR](#)

[10] **V Colin, P Ghiggini, K Honda**, *Embedded contact homology and open book decompositions*, preprint (2010) [arXiv](#)

[11] **V Colin, P Ghiggini, K Honda, M Hutchings**, *Sutures and contact homology, I*, Geom. Topol. 15 (2011) 1749–1842 [MR](#)

- [12] **LH Culler**, *The blowup formula for higher rank Donaldson invariants*, PhD thesis, Massachusetts Institute of Technology (2014) [MR](#) Available at <https://search.proquest.com/docview/1566428730>
- [13] **L Culler, A Daemi, Y Xie**, *Surgery, polygons and $SU(N)$ -Floer Homology*, preprint (2017) [arXiv](#)
- [14] **U V Desale, S Ramanan**, *Poincaré polynomials of the variety of stable bundles*, Math. Ann. 216 (1975) 233–244 [MR](#)
- [15] **SK Donaldson**, *Anti self-dual Yang–Mills connections over complex algebraic surfaces and stable vector bundles*, Proc. London Math. Soc. 50 (1985) 1–26 [MR](#)
- [16] **SK Donaldson**, *The orientation of Yang–Mills moduli spaces and 4-manifold topology*, J. Differential Geom. 26 (1987) 397–428 [MR](#)
- [17] **SK Donaldson**, *Polynomial invariants for smooth four-manifolds*, Topology 29 (1990) 257–315 [MR](#)
- [18] **SK Donaldson**, *Floer homology and algebraic geometry*, from “Vector bundles in algebraic geometry” (NJ Hitchin, PE Newstead, WM Oxbury, editors), London Math. Soc. Lecture Note Ser. 208, Cambridge Univ. Press (1995) 119–138 [MR](#)
- [19] **SK Donaldson**, *Floer homology groups in Yang–Mills theory*, Cambridge Tracts in Mathematics 147, Cambridge Univ. Press (2002) [MR](#)
- [20] **SK Donaldson, PB Kronheimer**, *The geometry of four-manifolds*, Clarendon, New York (1990) [MR](#)
- [21] **S Dostoglou, D A Salamon**, *Self-dual instantons and holomorphic curves*, Ann. of Math. 139 (1994) 581–640 [MR](#)
- [22] **JD Edelstein, M Gómez-Reino, M Mariño**, *Blowup formulae in Donaldson–Witten theory and integrable hierarchies*, Adv. Theor. Math. Phys. 4 (2000) 503–543 [MR](#)
- [23] **PMN Feehan, TG Leness**, *Witten’s conjecture for many four-manifolds of simple type*, J. Eur. Math. Soc. 17 (2015) 899–923 [MR](#)
- [24] **R Fintushel, R J Stern**, *Instanton homology of Seifert fibred homology three spheres*, Proc. London Math. Soc. 61 (1990) 109–137 [MR](#)
- [25] **R Fintushel, R J Stern**, *Donaldson invariants of 4-manifolds with simple type*, J. Differential Geom. 42 (1995) 577–633 [MR](#)
- [26] **R Fintushel, R J Stern**, *The blowup formula for Donaldson invariants*, Ann. of Math. 143 (1996) 529–546 [MR](#)
- [27] **A Floer**, *An instanton-invariant for 3-manifolds*, Comm. Math. Phys. 118 (1988) 215–240 [MR](#)
- [28] **A Floer**, *Instanton homology and Dehn surgery*, from “The Floer memorial volume” (H Hofer, CH Taubes, A Weinstein, E Zehnder, editors), Progr. Math. 133, Birkhäuser, Basel (1995) 77–97 [MR](#)

- [29] **S Friedl**, *Eta invariants as sliceness obstructions and their relation to Casson–Gordon invariants*, *Algebr. Geom. Topol.* 4 (2004) 893–934 [MR](#)
- [30] **R Friedman**, *Rank two vector bundles over regular elliptic surfaces*, *Invent. Math.* 96 (1989) 283–332 [MR](#)
- [31] **R Friedman**, *Vector bundles and $SO(3)$ –invariants for elliptic surfaces*, *J. Amer. Math. Soc.* 8 (1995) 29–139 [MR](#)
- [32] **R Friedman, J W Morgan**, *Smooth four-manifolds and complex surfaces*, *Ergeb. Math. Grenzgeb.* 27, Springer (1994) [MR](#)
- [33] **K Fukaya**, *Floer homology for oriented 3–manifolds*, from “Aspects of low-dimensional manifolds” (Y Matsumoto, S Morita, editors), *Adv. Stud. Pure Math.* 20, Kinokuniya, Tokyo (1992) 1–92 [MR](#)
- [34] **D Gabai**, *Foliations and the topology of 3–manifolds*, *J. Differential Geom.* 18 (1983) 445–503 [MR](#)
- [35] **D Gabai**, *Foliations and the topology of 3–manifolds*, *Bull. Amer. Math. Soc.* 8 (1983) 77–80 [MR](#)
- [36] **D Gieseker**, *Geometric invariant theory and applications to moduli problems*, from “Invariant theory” (F Gherardelli, editor), *Lecture Notes in Math.* 996, Springer (1983) 45–73 [MR](#)
- [37] **D Gieseker**, *A construction of stable bundles on an algebraic surface*, *J. Differential Geom.* 27 (1988) 137–154 [MR](#)
- [38] **D Gieseker, J Li**, *Moduli of high rank vector bundles over surfaces*, *J. Amer. Math. Soc.* 9 (1996) 107–151 [MR](#)
- [39] **R E Gompf**, *Nuclei of elliptic surfaces*, *Topology* 30 (1991) 479–511 [MR](#)
- [40] **R E Gompf, A I Stipsicz**, *4–manifolds and Kirby calculus*, *Graduate Studies in Mathematics* 20, Amer. Math. Soc., Providence, RI (1999) [MR](#)
- [41] **V Gritsenko, K Hulek, G K Sankaran**, *Abelianisation of orthogonal groups and the fundamental group of modular varieties*, *J. Algebra* 322 (2009) 463–478 [MR](#)
- [42] **G Harder, M S Narasimhan**, *On the cohomology groups of moduli spaces of vector bundles on curves*, *Math. Ann.* 212 (1974/75) 215–248 [MR](#)
- [43] **R Hartshorne**, *Algebraic geometry*, *Graduate Texts in Mathematics* 52, Springer (1977) [MR](#)
- [44] **M Hutchings**, *Embedded contact homology and its applications*, from “Proceedings of the International Congress of Mathematicians, II” (R Bhatia, A Pal, G Rangarajan, V Srinivas, M Vanninathan, P Gastesi, editors), Hindustan Book Agency, New Delhi (2010) 1022–1041 [MR](#)

- [45] **M Hutchings**, *Lecture notes on embedded contact homology*, from “Contact and symplectic topology” (F Bourgeois, V Colin, A Stipsicz, editors), Bolyai Soc. Math. Stud. 26, János Bolyai Math. Soc., Budapest (2014) 389–484 [MR](#)
- [46] **A Juhász**, *Holomorphic discs and sutured manifolds*, Algebr. Geom. Topol. 6 (2006) 1429–1457 [MR](#)
- [47] **A D King, P E Newstead**, *On the cohomology ring of the moduli space of rank 2 vector bundles on a curve*, Topology 37 (1998) 407–418 [MR](#)
- [48] **R C Kirby, editor**, *Problems in low-dimensional topology*, from “Geometric topology, II” (W H Kazez, editor), AMS/IP Stud. Adv. Math. 2, Amer. Math. Soc., Providence, RI (1997) 35–473 [MR](#)
- [49] **K Kodaira**, *On compact complex analytic surfaces, I*, Ann. of Math. 71 (1960) 111–152 [MR](#)
- [50] **D Kotschick**, *SO(3)–invariants for 4–manifolds with $b_2^+ = 1$* , Proc. London Math. Soc. 63 (1991) 426–448 [MR](#)
- [51] **D Kotschick, J W Morgan**, *SO(3)–invariants for 4–manifolds with $b_2^+ = 1$, II*, J. Differential Geom. 39 (1994) 433–456 [MR](#)
- [52] **P B Kronheimer**, *Four-manifold invariants from higher-rank bundles*, J. Differential Geom. 70 (2005) 59–112 [MR](#)
- [53] **P B Kronheimer, T S Mrowka**, *Recurrence relations and asymptotics for four-manifold invariants*, Bull. Amer. Math. Soc. 30 (1994) 215–221 [MR](#)
- [54] **P B Kronheimer, T S Mrowka**, *Embedded surfaces and the structure of Donaldson’s polynomial invariants*, J. Differential Geom. 41 (1995) 573–734 [MR](#)
- [55] **P Kronheimer, T Mrowka**, *Monopoles and three-manifolds*, New Mathematical Monographs 10, Cambridge Univ. Press (2007) [MR](#)
- [56] **P Kronheimer, T Mrowka**, *Knots, sutures, and excision*, J. Differential Geom. 84 (2010) 301–364 [MR](#)
- [57] **P B Kronheimer, T S Mrowka**, *Khovanov homology is an unknot-detector*, Publ. Math. Inst. Hautes Études Sci. 113 (2011) 97–208 [MR](#)
- [58] **P B Kronheimer, T S Mrowka**, *Knot homology groups from instantons*, J. Topol. 4 (2011) 835–918 [MR](#)
- [59] **P B Kronheimer, T S Mrowka**, *Tait colorings, and an instanton homology for webs and foams*, J. Eur. Math. Soc. 21 (2019) 55–119 [MR](#)
- [60] **C Kutluhan, S Sivek, C H Taubes**, *Sutured ECH is a natural invariant*, preprint (2013) [arXiv](#)
- [61] **C LeBrun**, *Explicit self-dual metrics on $\mathbb{CP}_2 \# \cdots \# \mathbb{CP}_2$* , J. Differential Geom. 34 (1991) 223–253 [MR](#)

- [62] **W-P Li, Z Qin**, *Stable vector bundles on algebraic surfaces*, Trans. Amer. Math. Soc. 345 (1994) 833–852 [MR](#)
- [63] **P Lisca**, *Computations of instanton invariants using Donaldson–Floer theory*, Invent. Math. 119 (1995) 347–359 [MR](#)
- [64] **M Mariño, G Moore**, *The Donaldson–Witten function for gauge groups of rank larger than one*, Comm. Math. Phys. 199 (1998) 25–69 [MR](#)
- [65] **D McDuff, D Salamon**, *J–holomorphic curves and symplectic topology*, 2nd edition, American Mathematical Society Colloquium Publications 52, Amer. Math. Soc., Providence, RI (2012) [MR](#)
- [66] **V B Mehta, A Ramanathan**, *Restriction of stable sheaves and representations of the fundamental group*, Invent. Math. 77 (1984) 163–172 [MR](#)
- [67] **J W Morgan, H Bass** (editors), *The Smith conjecture*, Pure Appl. Math. 112, Academic, Orlando (1984) [MR](#)
- [68] **J W Morgan, T S Mrowka**, *A note on Donaldson’s polynomial invariants*, Int. Math. Res. Not. 1992 (1992) 223–230 [MR](#)
- [69] **J W Morgan, T S Mrowka**, *The smooth classification of elliptic surfaces*, from “Geometry, topology, & physics” (S-T Yau, editor), Conf. Proc. Lecture Notes Geom. Topology 4, International, Cambridge, MA (1995) 246–292 [MR](#)
- [70] **J W Morgan, T Mrowka, D Ruberman**, *The L^2 –moduli space and a vanishing theorem for Donaldson polynomial invariants*, Monographs in Geometry and Topology 2, International, Cambridge, MA (1994) [MR](#)
- [71] **V Muñoz**, *Gluing formulae for Donaldson invariants for connected sums along surfaces*, Asian J. Math. 1 (1997) 785–800 [MR](#)
- [72] **V Muñoz**, *Quantum cohomology of the moduli space of stable bundles over a Riemann surface*, Duke Math. J. 98 (1999) 525–540 [MR](#)
- [73] **V Muñoz**, *Ring structure of the Floer cohomology of $\Sigma \times \mathbb{S}^1$* , Topology 38 (1999) 517–528 [MR](#)
- [74] **P Ozsváth, Z Szabó**, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. 159 (2004) 1027–1158 [MR](#)
- [75] **Y Ruan, G Tian**, *A mathematical theory of quantum cohomology*, J. Differential Geom. 42 (1995) 259–367 [MR](#)
- [76] **D Ruberman**, *Relations among Donaldson invariants arising from negative 2–spheres and tori*, Duke Math. J. 83 (1996) 645–660 [MR](#)
- [77] **D A Salamon**, *Quantum products for mapping tori and the Atiyah–Floer conjecture*, from “Northern California Symplectic Geometry Seminar” (Y Eliashberg, D Fuchs, T Ratiu, A Weinstein, editors), Amer. Math. Soc. Transl. Ser. 2 196, Amer. Math. Soc., Providence, RI (1999) 199–235 [MR](#)

- [78] **B Siebert, G Tian**, *Recursive relations for the cohomology ring of moduli spaces of stable bundles*, Turkish J. Math. 19 (1995) 131–144 [MR](#)
- [79] **CH Taubes**, *Self-dual connections on 4-manifolds with indefinite intersection matrix*, J. Differential Geom. 19 (1984) 517–560 [MR](#)
- [80] **CH Taubes**, *Casson’s invariant and gauge theory*, J. Differential Geom. 31 (1990) 547–599 [MR](#)
- [81] **CH Taubes**, *L^2 moduli spaces on 4-manifolds with cylindrical ends*, Monographs in Geometry and Topology 1, International, Cambridge, MA (1993) [MR](#)
- [82] **WP Thurston**, *A norm for the homology of 3-manifolds*, Mem. Amer. Math. Soc. 339, Amer. Math. Soc., Providence, RI (1986) 99–130 [MR](#)
- [83] **W Wieczorek**, *The Donaldson invariant and embedded 2-spheres*, J. Reine Angew. Math. 489 (1997) 15–51 [MR](#)
- [84] **E Witten**, *Monopoles and four-manifolds*, Math. Res. Lett. 1 (1994) 769–796 [MR](#)
- [85] **Y Xie**, *On the framed singular instanton Floer homology from higher rank bundles*, PhD thesis, Harvard University (2016) [MR](#) Available at <https://tinyurl.com/Xie-thesis>
- [86] **D Zagier**, *On the cohomology of moduli spaces of rank two vector bundles over curves*, from “The moduli space of curves” (R Dijkgraaf, C Faber, G van der Geer, editors), Progr. Math. 129, Birkhäuser, Boston (1995) 533–563 [MR](#)

*Simons Center for Geometry and Physics, State University of New York
Stony Brook, NY, United States*

*Beijing International Center for Mathematical Research, Peking University
Beijing, China*

adaemi@scgp.stonybrook.edu, yixie@pku.edu.cn

Proposed: Simon Donaldson

Received: 4 March 2018

Seconded: Ciprian Manolescu, Peter Ozsváth

Revised: 25 March 2019

Contact handles, duality, and sutured Floer homology

ANDRÁS JUHÁSZ

IAN ZEMKE

We give an explicit construction of the Honda–Kazez–Matić gluing maps in terms of contact handles. We use this to prove a duality result for turning a sutured manifold cobordism around and to compute the trace in the sutured Floer TQFT. We also show that the decorated link cobordism maps on the hat version of link Floer homology defined by the first author via sutured manifold cobordisms and by the second author via elementary cobordisms agree.

57R58; 57M27, 57R17

1. Introduction	179
2. 1–handle and 3–handle maps and triangle maps; compound stabilizations	184
3. Contact cell decompositions and the gluing map	194
4. Contact handles and partial open book decompositions	217
5. Properties of the gluing map	224
6. Turning around cobordisms of sutured manifolds and duality	237
7. Triangle cobordisms	247
8. Trace and cotrace cobordisms	283
9. Equivalence of two link cobordism map constructions	291
References	306

1 Introduction

The purpose of this paper is to provide an explicit construction of the Honda–Kazez–Matić gluing map [12] in terms of contact handles, and use this to prove several results about the sutured Floer TQFT defined by the first author [15]. Additionally, we show that the decorated link cobordism maps on the hat version of link Floer homology defined via sutured manifold cobordisms by the first author [15] and the maps defined using elementary link cobordisms by the second author [30] agree.

1.1 The contact gluing map

Sutured manifolds were introduced by Gabai [5] to construct taut foliations on 3-manifolds, and are also ubiquitous in contact topology. In this paper, a sutured manifold is a pair (M, γ) , where M is a compact oriented 3-manifold with boundary, and the set of sutures $\gamma \subseteq \partial M$ is an oriented 1-manifold that divides ∂M into subsurfaces $R_+(\gamma)$ and $R_-(\gamma)$ that meet along γ . For example, if M carries a contact structure such that ∂M is convex with dividing set γ , then (M, γ) is a sutured manifold.

We say that (M, γ) is balanced if M has no closed components, each component of M contains a suture, and $\chi(R_+(\gamma)) = \chi(R_-(\gamma))$. Sutured Floer homology, defined by the first author [14], assigns an $\mathbb{F}_2$ -vector space $\text{SFH}(M, \gamma)$ to a balanced sutured manifold (M, γ) . It is a common extension of the hat version of Heegaard Floer homology of closed 3-manifolds and link Floer homology, both due to Ozsváth and Szabó [24; 23; 26], to 3-manifolds with boundary.

Let (M, γ) and (M', γ') be sutured manifolds such that $M \subseteq \text{int}(M')$. Given a contact structure ξ on $M' \setminus \text{int}(M)$ such that $\partial M \cup \partial M'$ is convex with dividing set $\gamma \cup \gamma'$, Honda, Kazez, and Matić [12] define a gluing map

$$\Phi_\xi: \text{SFH}(-M, -\gamma) \rightarrow \text{SFH}(-M', -\gamma')$$

using partial open book decompositions that satisfy a “contact compatibility” condition near the boundary. However, the contact compatibility condition makes working with and computing the gluing map impractical. In the first part of this paper, we give a new definition of the contact gluing map based on contact handle attachments, prove invariance via contact cell decompositions, and show that our map agrees with the Honda–Kazez–Matić gluing map. In particular, this allows us to give a simple diagrammatic description of the gluing map for a single contact handle attachment. For a precise statement about the gluing map associated to a contact handle attachment, see Proposition 5.6. Contact handles were introduced by Giroux [6]; see Definition 5.5.

We now describe the map C_{h^i} that we assign to attaching a contact i -handle h^i for $i \in \{0, 1, 2, 3\}$. Let (Σ, α, β) be a diagram of (M, γ) ; then $(\bar{\Sigma}, \alpha, \beta)$ is a diagram of $(-M, -\gamma)$. Attaching a contact 0-handle corresponds to taking the disjoint union of Σ with a disk. A contact 1-handle corresponds to attaching a 1-handle to $\partial \Sigma$; see Figure 1. Adding a disk or a 1-handle to $\partial \Sigma$ does not change the sutured Floer complex, and we define C_{h^i} to be the tautological map on intersection points. A contact 2-handle is attached to ∂M along a curve l that intersects γ in two points. Let $\lambda_\pm \subseteq \Sigma$

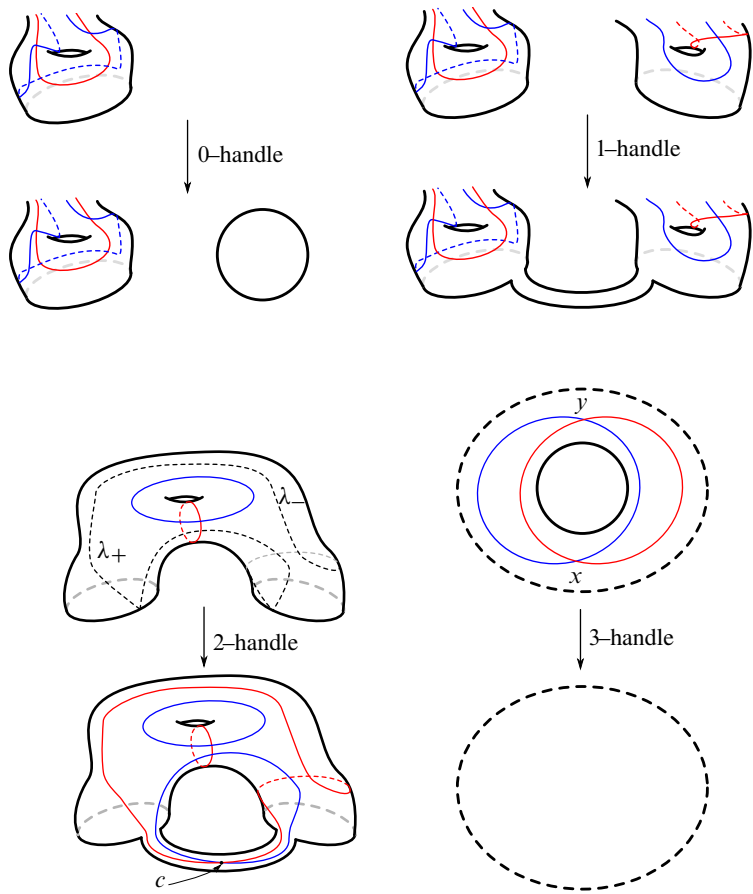


Figure 1: The diagrams used in the definition of the contact handle maps.

be the properly embedded arc corresponding to $l \cap R_{\pm}(\gamma)$. As in Figure 1, we glue a 1-handle H to Σ along $\partial\Sigma$, and add a curve α to α and a curve β to β that intersect in H in a single point c and are such that $\alpha \cap \Sigma = \lambda_-$ and $\beta \cap \Sigma = \lambda_+$. Then, given a generator $x \in \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}$, we let $C_{h^2}(x) = x \times \{c\}$. Finally, suppose that we attach a contact 3-handle h^3 along an S^2 component S of ∂M containing the suture $\gamma_S = \gamma \cap S$, giving rise to the sutured manifold (M', γ') . We then choose a diagram where γ_S is encircled by a curve $\alpha \in \alpha$ and a curve $\beta \in \beta$ such that $\alpha \cap \beta = \{x, y\}$ and such that there are no other α or β curves between α and γ_S or β and γ_S . Let Σ' be the result of gluing a disk to Σ along γ_S . Then $(\Sigma', \alpha \setminus \{\alpha\}, \beta \setminus \{\beta\})$ is a diagram of (M', γ') ; see Figure 1. We let $C_{h^3}(x \times \{x\}) = 0$ and $C_{h^3}(x \times \{y\}) = x$, where $\mu(x \times \{x\}, x \times \{y\}) = 1$ in $(\bar{\Sigma}, \alpha, \beta)$.

Note that Zarev [28] has also defined a type of gluing map in sutured Floer homology, corresponding to a convex decomposition. Combining this with the EH invariant of Honda, Kazez, and Matić [13], one can define a map for gluing a contact structure to a sutured manifold. Zarev conjectured that this map agrees with the Honda–Kazez–Matić gluing map, though we will not address Zarev’s construction in this paper.

1.2 The sutured Floer TQFT

The first author [15] defined the category of balanced sutured manifolds and sutured manifold cobordisms, and extended SFH to a functor on this category. A sutured manifold cobordism from (M_0, γ_0) to (M_1, γ_1) is a triple $\mathcal{W} = (W, Z, [\xi])$, where W is a 4-manifold with boundary and corners, Z is a codimension-0 compact submanifold of ∂W such that $\partial W \setminus \text{int}(Z) = -M_0 \sqcup M_1$, and $[\xi]$ is a certain equivalence class of a contact structure ξ on Z such that ∂Z is convex with dividing set $\gamma_0 \sqcup \gamma_1$. The sutured cobordism map

$$F_{\mathcal{W}}: \text{SFH}(M_0, \gamma_0) \rightarrow \text{SFH}(M_1, \gamma_1)$$

is the composition of the contact gluing map for $-\xi$ and 4-dimensional handle maps.

Let $\xi_{I \times \partial M}$ denote the I -invariant contact structure on $-I \times \partial M$ that induces the dividing set γ on ∂M . Consider the trace cobordism

$$\Lambda_{(M, \gamma)} = (I \times M, \xi_{I \times \partial M})$$

from $(M, \gamma) \sqcup (-M, \gamma)$ to $\emptyset$, and the cotrace cobordism

$$V_{(M, \gamma)} = (I \times M, \xi_{I \times \partial M})$$

from $\emptyset$ to $(-M, \gamma) \sqcup (M, \gamma)$. In Theorem 8.1, we answer [15, Conjecture 11.13] positively:

Theorem 1.1 *The trace cobordism $\Lambda_{(M, \gamma)}$ induces the canonical trace map*

$$\text{tr}: \text{SFH}(M, \gamma) \otimes \text{SFH}(-M, \gamma) \rightarrow \mathbb{F}_2,$$

obtained by evaluating cohomology on homology. The cotrace cobordism $V_{(M, \gamma)}$ induces the canonical cotrace map

$$\text{cotr}: \mathbb{F}_2 \rightarrow \text{SFH}(-M, \gamma) \otimes \text{SFH}(M, \gamma).$$

The proof relies on the following deep technical result, which is Theorem 7.1. Before stating it, recall that, given a sutured triple diagram $(\Sigma, \alpha, \beta, \gamma)$, we can associate

to it a sutured manifold cobordism $\mathcal{W}_{\alpha,\beta,\gamma}$ from $(M_{\alpha,\beta}, \gamma_{\alpha,\beta}) \sqcup (M_{\beta,\gamma}, \gamma_{\beta,\gamma})$ to $(M_{\alpha,\gamma}, \gamma_{\alpha,\gamma})$.

Theorem 1.2 *Let $\mathcal{T} = (\Sigma, \alpha, \beta, \gamma)$ be an admissible balanced sutured triple diagram. Then the cobordism map*

$$F_{\mathcal{W}_{\alpha,\beta,\gamma}}: \mathrm{CF}(\Sigma, \alpha, \beta) \otimes \mathrm{CF}(\Sigma, \beta, \gamma) \rightarrow \mathrm{CF}(\Sigma, \alpha, \gamma)$$

is chain homotopic to the map $F_{\alpha,\beta,\gamma}$ defined in [15, Definition 5.13] that counts holomorphic triangles on the triple diagram $\mathcal{T}$.

One can obtain from [Theorem 1.1](#) a positive answer to [15, Question 11.9]:

Theorem 1.3 *If $\mathcal{W}: (M, \gamma) \rightarrow (M', \gamma')$ is a balanced cobordism of sutured manifolds, and $\mathcal{W}^\vee$ is the cobordism obtained by turning around $\mathcal{W}$, then*

$$F_{\mathcal{W}^\vee} = (F_{\mathcal{W}})^\vee,$$

with respect to the trace pairing.

We also give a self-contained proof of this result, without invoking [Theorem 1.1](#). As a special case, we obtain that the decorated link cobordism maps $F_{\mathcal{X}}^J$ of the first author satisfy an analogous duality property when we turn a decorated link cobordism $\mathcal{X}$ around; see Juhász and Marengon [17, Section 5.7]. Indeed, these maps are defined by assigning a sutured manifold cobordism to a decorated link cobordism, and applying the SFH functor.

The second author [30] later gave a different construction of link Floer cobordism maps $F_{\mathcal{X}}^Z$ by composing maps defined for elementary link cobordisms, and showing independence of the decomposition. Note that this construction makes sense for all versions of link Floer homology, not just the hat version. In the last section, we prove that the two maps agree:

Theorem 1.4 *Given a decorated link cobordism $\mathcal{X}$, we have $F_{\mathcal{X}}^J = \hat{F}_{\mathcal{X}}^Z$.*

A key technical lemma that we use throughout the paper gives a simple formula for the naturality map for a compound stabilization operation on a sutured diagram (called a $(k, 0)$ – or $(0, l)$ –stabilization by Juhász, Thurston, and Zemke [18]), which consists of a simple stabilization, followed by handle sliding some α –curves over the new α –curve, or some β –curves over the new β –curve; see [Proposition 2.2](#).

In an upcoming paper, we will use [Theorem 1.1](#) to compute the effect of a generalization of the Fintushel–Stern knot surgery operation using a self-concordance of a knot, called

concordance surgery by Akbulut [1, Section 2], on the Ozsváth–Szabó 4–manifold invariant. The formula involves the graded Lefschetz number of the concordance map on knot Floer homology. In another work, we will apply [Theorem 1.1](#) to compute the invariant due to Juhász and Marengon [16] of a slice disk obtained by the deformation–spinning construction of Litherland [19]. Hence, we will show that this invariant can effectively distinguish different slice disks of a knot, answering [16, Question 1.4].

1.3 Notation and conventions

Throughout this paper, if A and B are smooth manifolds, then we write $A \cong B$ if A and B are diffeomorphic. Given a submanifold A of C , we write $N(A)$ for a regular neighborhood of A in C . We denote Heegaard diagrams by $\mathcal{H}$, and handle decompositions by H . If M is an oriented n –manifold, then we will denote the same manifold with its orientation reversed by $\overline{M}$ when n is even, and by $-M$ when n is odd. The closure of a set X is $\text{cl}(X)$. If ξ is a co-oriented 2–plane field on M , we will write $-\xi$ for the co-oriented 2–plane field obtained by reversing the co-orientation of ξ .

We orient the boundary of a manifold using the “outward normal first” convention. To be consistent with this convention, all our cobordisms go from left to right.

Acknowledgements We would like to thank Ko Honda and Jacob Rasmussen for helpful discussions on the contact gluing maps.

Juhász was supported by a Royal Society Research Fellowship, and Zemke by an NSF Postdoctoral Research Fellowship (DMS-1703685). This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 674978).

2 1–handle and 3–handle maps and triangle maps; compound stabilizations

In this section, we describe several results about the interactions between holomorphic triangles and 1–handle, 3–handle, and stabilization maps. These will be used in later sections.

2.1 1–handle and 3–handle maps

Let (Σ, α, β) be an admissible sutured diagram, and let $p_1, p_2 \in \Sigma$ be a pair of points that are both in components of $\Sigma \setminus (\alpha \cup \beta)$ that intersect $\partial \Sigma$. We construct the admissible

sutured diagram $(\Sigma', \alpha \cup \{\alpha_0\}, \beta \cup \{\beta_0\})$ by removing disks centered at p_1 and p_2 , and adding an annulus A connecting the boundaries of the disks. Furthermore, α_0 and β_0 are homologically nontrivial curves in A that intersect transversely at two points $\theta_{\alpha_0, \beta_0}^+$ and $\theta_{\alpha_0, \beta_0}^-$ such that $\theta_{\alpha_0, \beta_0}^+$ has the larger relative Maslov grading. The first author [15, Section 7] defined the 1–handle map

$$F_1^{\alpha_0, \beta_0}: \text{CF}(\Sigma, \alpha, \beta) \rightarrow \text{CF}(\Sigma', \alpha \cup \{\alpha_0\}, \beta \cup \{\beta_0\}), \quad x \mapsto x \times \theta_{\alpha_0, \beta_0}^+,$$

and the 3–handle map

$$\begin{aligned} F_3^{\alpha_0, \beta_0}: \text{CF}(\Sigma', \alpha \cup \{\alpha_0\}, \beta \cup \{\beta_0\}) &\rightarrow \text{CF}(\Sigma, \alpha, \beta), \\ x \times \theta_{\alpha_0, \beta_0}^+ &\mapsto 0 \quad \text{and} \quad x \times \theta_{\alpha_0, \beta_0}^- \mapsto x. \end{aligned}$$

If $\mathcal{T} = (\Sigma, \alpha, \beta, \gamma)$ is an admissible sutured triple, then it induces a holomorphic triangle map

$$F_{\mathcal{T}}: \text{CF}(\Sigma, \alpha, \beta) \otimes \text{CF}(\Sigma, \beta, \gamma) \rightarrow \text{CF}(\Sigma, \alpha, \gamma).$$

If $p_1, p_2 \in \Sigma \setminus (\alpha \cup \beta \cup \gamma)$ are distinct points that are in components of $\Sigma \setminus (\alpha \cup \beta \cup \gamma)$ that intersect $\partial \Sigma$, then we can similarly form the admissible Heegaard triple

$$\mathcal{T}' := (\Sigma', \alpha' = \alpha \cup \{\alpha_0\}, \beta' = \beta \cup \{\beta_0\}, \gamma' = \gamma \cup \{\gamma_0\}),$$

where Σ' is obtained by adding a 1–handle A with feet at p_1 and p_2 and three new curves, α_0 , β_0 , and γ_0 , that are homologically nontrivial in A and pairwise intersect

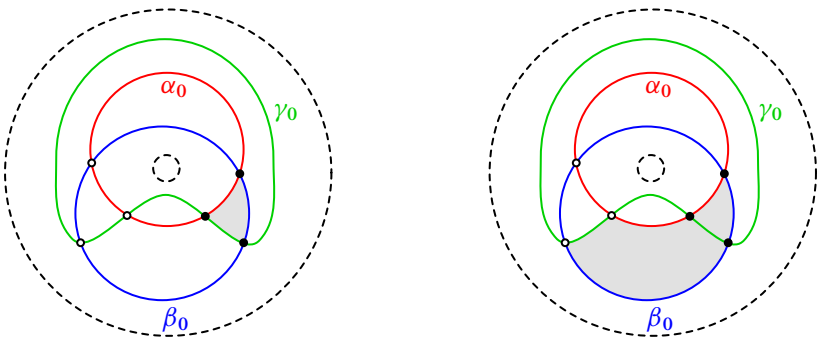


Figure 2: The annulus A is bounded by the two dashed circles. The intersection points $\theta_{\alpha_0, \beta_0}^+$, $\theta_{\beta_0, \gamma_0}^+$, and $\theta_{\alpha_0, \gamma_0}^+$ are marked by solid circles, and $\theta_{\alpha_0, \beta_0}^-$, $\theta_{\beta_0, \gamma_0}^-$, and $\theta_{\alpha_0, \gamma_0}^-$ by empty circles. On the left, the only index 0 triangle connecting $\theta_{\alpha_0, \beta_0}^+$, $\theta_{\beta_0, \gamma_0}^+$, and $\theta_{\alpha_0, \gamma_0}^+$ is shaded. On the right, the only index 0 triangle connecting $\theta_{\alpha_0, \beta_0}^+$, $\theta_{\beta_0, \gamma_0}^-$, and $\theta_{\alpha_0, \gamma_0}^-$ is shaded.

in two points; see Figure 2. If x and y are sets of attaching curves on some Heegaard surface S , then we denote the diagram (S, x, y) by $\mathcal{H}_{x,y}$.

Proposition 2.1 *With the above notation, the following diagrams are commutative:*

$$\begin{array}{ccc}
 \mathrm{CF}(\mathcal{H}_{\alpha,\beta}) \otimes \mathrm{CF}(\mathcal{H}_{\beta,\gamma}) & \xrightarrow{F_{\mathcal{T}}} & \mathrm{CF}(\mathcal{H}_{\alpha,\gamma}) \\
 \downarrow F_1^{\alpha_0,\beta_0} \otimes F_1^{\beta_0,\gamma_0} & & \downarrow F_1^{\alpha_0,\gamma_0} \\
 \mathrm{CF}(\mathcal{H}_{\alpha',\beta'}) \otimes \mathrm{CF}(\mathcal{H}_{\beta',\gamma'}) & \xrightarrow{F_{\mathcal{T}'}} & \mathrm{CF}(\mathcal{H}_{\alpha',\gamma'})
 \end{array}$$

$$\begin{array}{ccccc}
 \mathrm{CF}(\mathcal{H}_{\alpha,\beta}) \otimes \mathrm{CF}(\mathcal{H}_{\beta',\gamma'}) & \xrightarrow{\mathrm{id}_{\mathrm{CF}(\mathcal{H}_{\alpha,\beta})} \otimes F_3^{\beta_0,\gamma_0}} & \mathrm{CF}(\mathcal{H}_{\alpha,\beta}) \otimes \mathrm{CF}(\mathcal{H}_{\beta,\gamma}) & \xrightarrow{F_{\mathcal{T}}} & \mathrm{CF}(\mathcal{H}_{\alpha,\gamma}) \\
 \downarrow F_1^{\alpha_0,\beta_0} \otimes \mathrm{id}_{\mathrm{CF}(\mathcal{H}_{\beta',\gamma'})} & & & \nearrow & \\
 \mathrm{CF}(\mathcal{H}_{\alpha',\beta'}) \otimes \mathrm{CF}(\mathcal{H}_{\beta',\gamma'}) & \xrightarrow{F_{\mathcal{T}'}} & \mathrm{CF}(\mathcal{H}_{\alpha',\gamma'}) - F_3^{\alpha_0,\gamma_0} & &
 \end{array}$$

$$\begin{array}{ccccc}
 \mathrm{CF}(\mathcal{H}_{\alpha',\beta'}) \otimes \mathrm{CF}(\mathcal{H}_{\beta,\gamma}) & \xrightarrow{F_3^{\alpha_0,\beta_0} \otimes \mathrm{id}_{\mathrm{CF}(\mathcal{H}_{\beta,\gamma})}} & \mathrm{CF}(\mathcal{H}_{\alpha,\beta}) \otimes \mathrm{CF}(\mathcal{H}_{\beta,\gamma}) & \xrightarrow{F_{\mathcal{T}}} & \mathrm{CF}(\mathcal{H}_{\alpha,\gamma}) \\
 \downarrow \mathrm{id}_{\mathrm{CF}(\mathcal{H}_{\alpha',\beta'})} \otimes F_1^{\beta_0,\gamma_0} & & & \nearrow & \\
 \mathrm{CF}(\mathcal{H}_{\alpha',\beta'}) \otimes \mathrm{CF}(\mathcal{H}_{\beta',\gamma'}) & \xrightarrow{F_{\mathcal{T}'}} & \mathrm{CF}(\mathcal{H}_{\alpha',\gamma'}) - F_3^{\alpha_0,\gamma_0} & &
 \end{array}$$

Proof Consider the first diagram. The assumption that the points p_1 and p_2 are in components of $\Sigma \setminus (\alpha \cup \beta \cup \gamma)$ that intersect $\partial\Sigma$ allows one to reduce the claim to the model computation

$$(2-1) \quad F_{\alpha_0,\beta_0,\gamma_0}(\theta_{\alpha_0,\beta_0}^+ \otimes \theta_{\beta_0,\gamma_0}^+) = \theta_{\alpha_0,\gamma_0}^+$$

in the annulus A , which was established in the proof of [15, Theorem 7.6]; see the left-hand side of Figure 2.

We now show the claim for the second diagram. Let $x \in \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}$ and $y \in \mathbb{T}_{\beta} \cap \mathbb{T}_{\gamma}$. Then

$$F_{\mathcal{T}} \circ (\mathrm{id}_{\mathrm{CF}(\mathcal{H}_{\alpha,\beta})} \otimes F_3^{\beta_0,\gamma_0})(x \otimes (y \times \theta_{\beta_0,\gamma_0}^-)) = F_{\mathcal{T}}(x \otimes y).$$

On the other hand,

$$\begin{aligned}
 F_3^{\alpha_0,\gamma_0} \circ F_{\mathcal{T}'} \circ (F_1^{\alpha_0,\beta_0} \otimes \mathrm{id}_{\mathrm{CF}(\mathcal{H}_{\beta',\gamma'})})(x \otimes (y \times \theta_{\beta_0,\gamma_0}^-)) \\
 = F_3^{\alpha_0,\gamma_0}(F_{\mathcal{T}}(x, y) \times F_{\alpha_0,\beta_0,\gamma_0}(\theta_{\alpha_0,\beta_0}^+ \otimes \theta_{\beta_0,\gamma_0}^-)).
 \end{aligned}$$

Hence, commutativity for the generator $x \otimes (y \times \theta_{\beta_0,\gamma_0}^-)$ follows from

$$F_{\alpha_0,\beta_0,\gamma_0}(\theta_{\alpha_0,\beta_0}^+ \otimes \theta_{\beta_0,\gamma_0}^-) = \theta_{\alpha_0,\gamma_0}^-,$$

which can be shown similarly to (2-1); see the right-hand side of Figure 2. Note that there is a unique index 0 pseudoholomorphic triangle in A connecting $\theta_{\alpha_0, \beta_0}^+$, $\theta_{\beta_0, \gamma_0}^-$, and $\theta_{\alpha_0, \gamma_0}^-$, and there is none connecting $\theta_{\alpha_0, \beta_0}^+$, $\theta_{\beta_0, \gamma_0}^-$, and $\theta_{\alpha_0, \gamma_0}^+$.

On a generator of the form $x \otimes (y \times \theta_{\beta_0, \gamma_0}^+)$, we have

$$F_{\mathcal{T}} \circ (\text{id}_{\text{CF}(\mathcal{H}_{\alpha, \beta})} \otimes F_3^{\beta_0, \gamma_0})(x \otimes (y \times \theta_{\beta_0, \gamma_0}^+)) = 0,$$

and, using (2-1),

$$\begin{aligned} F_3^{\alpha_0, \gamma_0} \circ F_{\mathcal{T}'} \circ (F_1^{\alpha_0, \beta_0} \otimes \text{id}_{\text{CF}(\mathcal{H}_{\beta', \gamma'})})(x \otimes (y \times \theta_{\beta_0, \gamma_0}^+)) \\ = F_3^{\alpha_0, \gamma_0}(F_{\mathcal{T}}(x, y) \times F_{\alpha_0, \beta_0, \gamma_0}(\theta_{\alpha_0, \beta_0}^+ \otimes \theta_{\beta_0, \gamma_0}^+)) \\ = F_3^{\alpha_0, \gamma_0}(F_{\mathcal{T}}(x, y) \times \theta_{\alpha_0, \gamma_0}^+) \\ = 0. \end{aligned}$$

This establishes commutativity of the second diagram. Commutativity of the third diagram is analogous. $\square$

2.2 Compound stabilization

In this section, we describe an elaboration of the usual stabilization operation on Heegaard diagrams. Suppose that $\mathcal{H} = (\Sigma, \alpha, \beta)$ is an admissible sutured diagram, and that λ is an embedded path on Σ between two distinct points on $\partial\Sigma$ that avoids the α curves. We define the *compound stabilization* of $\mathcal{H}$ along λ , as follows.

First, construct a surface Σ' by pushing λ into the sutured compression body U_{α} , and add a tube that is the boundary of a regular neighborhood of λ . Let α_0 be a longitude of the tube, concatenated with a portion of the curve λ on Σ . Furthermore, let β_0 be a meridian of the tube. The curve α_0 may intersect other β curves; however, β_0 intersects only α_0 . The construction is shown in Figure 3. Let us denote by $\mathcal{H}'$ the Heegaard diagram $(\Sigma', \alpha \cup \{\alpha_0\}, \beta \cup \{\beta_0\})$. This is an instance of a $(k, 0)$ -stabilization, using the terminology of [18, Definition 6.26], where $k = |\alpha_0 \cap \beta|$. If λ avoids β , then we can perform an analogous operation, with the roles of α and β swapped, which is an instance of a $(0, l)$ -stabilization. We also call this a compound stabilization. In the opposite direction, we say that $\mathcal{H}$ is obtained from $\mathcal{H}'$ by a *compound destabilization*.

We denote the unique intersection point of α_0 and β_0 by c_{α_0, β_0} . There is a map

$$\sigma^{\alpha_0, \beta_0}: \text{SFH}(\mathcal{H}) \rightarrow \text{SFH}(\mathcal{H}'),$$

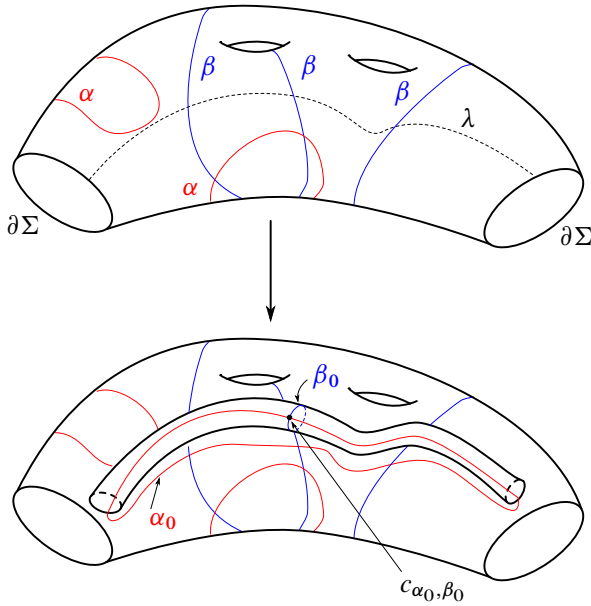


Figure 3: An example of the compound stabilization operation along a path λ .

defined by

$$\sigma^{\alpha_0, \beta_0}(\mathbf{x}) = \mathbf{x} \times c_{\alpha_0, \beta_0},$$

which is a chain isomorphism since the tube is added near $\partial\Sigma$.

On the other hand, there is also a naturality map $\Psi_{\mathcal{H} \rightarrow \mathcal{H}'}$. One would expect these to be equal. Indeed, we prove the following (compare [12, Proposition 3.7]):

Proposition 2.2 *The compound stabilization map $\sigma^{\alpha_0, \beta_0}$ is chain homotopic to the naturality map $\Psi_{\mathcal{H} \rightarrow \mathcal{H}'}$.*

One strategy to prove the above theorem would be to handle slide all the β curves that intersect α_0 across β_0 . The map from naturality induced by these handle slides can be computed by counting holomorphic triangles. To prove Proposition 2.2, one could analyze how holomorphic triangles degenerate as one stretches two necks (one on each end of the tube we are adding). While this can be done, we will give a somewhat indirect argument that avoids performing a neck-stretching argument.

Suppose that $\mathcal{T} = (\Sigma, \alpha, \beta, \beta')$ is an admissible sutured triple with a path λ from $\partial\Sigma$ to itself that does not intersect any α curves. Then we can perform the compound

stabilization procedure on $(\Sigma, \alpha, \beta, \beta')$ to obtain a Heegaard triple

$$\mathcal{T}' = (\Sigma', \alpha \cup \{\alpha_0\}, \beta \cup \{\beta_0\}, \beta' \cup \{\beta'_0\}),$$

where $(\Sigma', \alpha \cup \{\alpha_0\}, \beta \cup \{\beta_0\})$ is the compound stabilization of (Σ, α, β) along λ . Furthermore, the curve β'_0 is isotopic to β_0 and $|\beta_0 \cap \beta'_0| = 2$, while $|\alpha_0 \cap \beta_0| = |\alpha_0 \cap \beta'_0| = 1$. An example is shown in Figure 4. Let $\theta_{\beta_0, \beta'_0}^+$ be the point of $\beta_0 \cap \beta'_0$ with the higher relative Maslov grading, and write $\alpha_0 \cap \beta_0 = \{c_{\alpha_0, \beta_0}\}$ and $\alpha_0 \cap \beta'_0 = \{c_{\alpha_0, \beta'_0}\}$.

Lemma 2.3 *If $\mathcal{T} = (\Sigma, \alpha, \beta, \beta')$ is an admissible sutured triple and*

$$\mathcal{T}' = (\Sigma', \alpha \cup \{\alpha_0\}, \beta \cup \{\beta_0\}, \beta' \cup \{\beta'_0\})$$

is a compound stabilization of $\mathcal{T}$, as described in the previous paragraph, then

$$F_{\mathcal{T}'}(\mathbf{x} \times c_{\alpha_0, \beta_0}, \mathbf{y} \times \theta_{\beta_0, \beta'_0}^+) = F_{\mathcal{T}}(\mathbf{x}, \mathbf{y}) \times c_{\alpha_0, \beta'_0}.$$

Proof Since the tube is added near $\partial\Sigma$, the result is obtained by a model computation inside the tube. This is shown in Figure 4. $\square$

Remark 2.4 Despite the notation, the triangle map computation of Lemma 2.3 does not assume that the curves β and β' appearing in the triple $\mathcal{T}$ are related by a sequence of handle slides or isotopies. However, we will only need the result for examples where that is the case.

Analogously, we need to consider moves of the α_0 curve appearing in a compound stabilization. To this end, suppose that $\mathcal{T} = (\Sigma, \alpha', \alpha, \beta)$ is a sutured triple with two paths, λ and λ' , from $\partial\Sigma$ to itself such that λ and λ' have the same endpoints and disjoint interiors. Furthermore, suppose that λ avoids α and λ' avoids α' . We can construct a compound stabilization of the triple $(\Sigma, \alpha', \alpha, \beta)$ to obtain

$$\mathcal{T}' = (\Sigma', \alpha' \cup \{\alpha'_0\}, \alpha \cup \{\alpha_0\}, \beta \cup \{\beta_0\}),$$

where $(\Sigma', \alpha \cup \{\alpha_0\}, \beta \cup \{\beta_0\})$ is the compound stabilization of (Σ, α, β) along λ . Furthermore, the curve α'_0 is a concatenation of a portion of the curve λ' on Σ with a longitude of the tube $\Sigma' \setminus \Sigma$ such that $|\alpha_0 \cap \alpha'_0| = 2$ and $|\alpha_0 \cap \beta_0| = |\alpha'_0 \cap \beta_0| = 1$. In the tube, α_0 , α'_0 , and β_0 are configured as in Figure 5. Let $\theta_{\alpha'_0, \alpha_0}^+$ be the point of $\alpha'_0 \cap \alpha_0$ with the larger relative Maslov grading, and write $\alpha_0 \cap \beta_0 = \{c_{\alpha_0, \beta_0}\}$ and $\alpha'_0 \cap \beta_0 = \{c_{\alpha'_0, \beta_0}\}$.

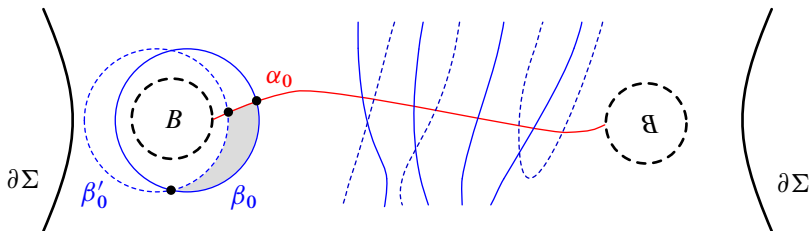


Figure 4: A compound stabilization of a sutured triple, and the model computation of Lemma 2.3. The α curves are shown as red solid lines, the β curves are shown as blue solid lines, and the β' curves are shown as blue dashed lines. The circles marked B are identified.

Lemma 2.5 *If $\mathcal{T} = (\Sigma, \alpha', \alpha, \beta)$ is an admissible sutured triple and*

$$\mathcal{T}' = (\Sigma', \alpha' \cup \{\alpha'_0\}, \alpha \cup \{\alpha_0\}, \beta \cup \{\beta_0\})$$

is a compound stabilization, as described in the previous paragraph, then

$$F_{\mathcal{T}'}(x \times \theta_{\alpha'_0, \alpha_0}^+, y \times c_{\alpha_0, \beta_0}) = F_{\mathcal{T}}(x, y) \times c_{\alpha'_0, \beta_0}.$$

Proof As before, since the ends of the tube $\Sigma' \setminus \Sigma$ are near $\partial\Sigma$, we obtain constraints on the multiplicities of any homology class of triangles which has holomorphic representatives. An easy model computation shows that triangles with representatives have homology class $\psi \sqcup \psi_0$, where ψ is a homology class on $(\Sigma, \alpha', \alpha, \beta)$ and ψ_0 is a homology class supported entirely on the tube. The appropriate model computation is shown in Figure 5. □

Using the above two lemmas, we now prove Proposition 2.2.

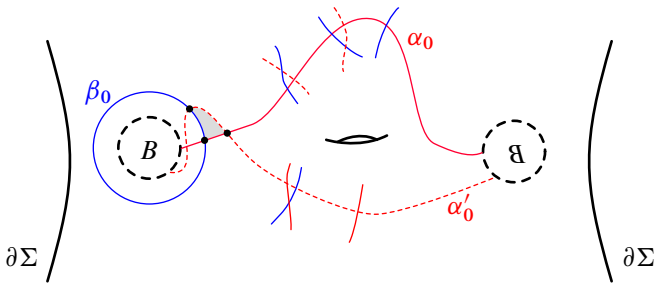


Figure 5: The model computation of Lemma 2.5. The α' curves are shown as dashed red, the α curves are shown as solid red, and the β curves are shown as solid blue.

Proof of Proposition 2.2 Let $\mathcal{H}' = (\Sigma', \alpha \cup \{\alpha_0\}, \beta \cup \{\beta_0\})$ denote a compound stabilization of (Σ, α, β) using a path λ with ends on $\partial\Sigma$, and let D be the tube attached. Let $\hat{\mathcal{H}} = (\hat{\Sigma}, \alpha \cup \{\alpha_1\}, \beta \cup \{\beta_1\})$ be another compound stabilization of $\mathcal{H}$, along a path that is parallel to λ . Write B for the attached tube. Let $\hat{\mathcal{H}}' = (\hat{\Sigma}', \alpha \cup \{\alpha_0, \alpha_1\}, \beta \cup \{\beta_0, \beta_1\})$ denote the two-fold compound stabilization of Σ along both paths. Write $\{c_0\} = \alpha_0 \cap \beta_0$ and $\{c_1\} = \alpha_1 \cap \beta_1$.

We claim that

$$(2-2) \quad \Psi_{\mathcal{H}' \rightarrow \hat{\mathcal{H}}'}(\mathbf{x} \times c_0) = \Psi_{\mathcal{H} \rightarrow \hat{\mathcal{H}}}(\mathbf{x}) \times c_0.$$

To see this, we note that a sequence of diagrams from $\mathcal{H}'$ to $\hat{\mathcal{H}}'$ can be constructed by starting with $\mathcal{H}'$, performing a simple stabilization near the boundary, and then moving one foot of the new tube along Σ , parallel to α_0 . At various points, we will have to handle slide a β curve across β_1 . It is not obvious what the holomorphic triangle count will be for each handle slide. However, by the holomorphic triangle count from Lemma 2.3, it is unchanged by the presence of the compound stabilization along α_0 . In particular, the triangles counted by going from $\mathcal{H}'$ to $\hat{\mathcal{H}}'$ are the same as the ones counted in the analogous sequence of diagrams from $\mathcal{H}$ to $\hat{\mathcal{H}}$, so (2-2) follows.

We now consider the path of Heegaard diagrams from $\hat{\mathcal{H}}'$ to $\hat{\mathcal{H}}$ shown in Figure 6. The diagram $\hat{\mathcal{H}}''$ is obtained by handle sliding β_1 over β_0 . We let β'_1 denote the curve resulting from this handle slide. The diagram $\hat{\mathcal{H}}'''$ is obtained by isotoping the Heegaard surface by sliding the foot of the tube marked D inside β'_1 over the tube marked B , carrying α_0 and β_0 along, and then handle sliding α_0 over α_1 , giving rise to a new curve α'_0 . Note that $\hat{\mathcal{H}}'''$ is a simple stabilization of $\hat{\mathcal{H}}$ (in [18], this type of stabilization was also referred to as a $(0, 0)$ -stabilization), and hence there is a destabilization from $\hat{\mathcal{H}}'''$ to $\hat{\mathcal{H}}$.

Using the presence of the boundary $\partial\Sigma$ to simplify the computation, one can see that the only holomorphic triangles contributing to the change of diagrams map $\Psi_{\hat{\mathcal{H}}' \rightarrow \hat{\mathcal{H}}''}$ have homology class $\psi \sqcup \psi_0$, where ψ is a holomorphic triangle on an unstabilized Heegaard triple $(\Sigma, \alpha, \beta, \beta')$, where β' is a small isotopy of β , and ψ_0 is the homology class shown in Figure 7. Using an additional triangle map to move the β' back to β (and only isotoping β_0 and β'_1 a small amount), which can be analyzed similarly, we have that

$$(2-3) \quad \Psi_{\hat{\mathcal{H}}' \rightarrow \hat{\mathcal{H}}''}(\mathbf{x} \times c_0 \times c_1) = \mathbf{x} \times c_0 \times c'_1,$$

where $c'_1 = \alpha_1 \cap \beta'_1$.

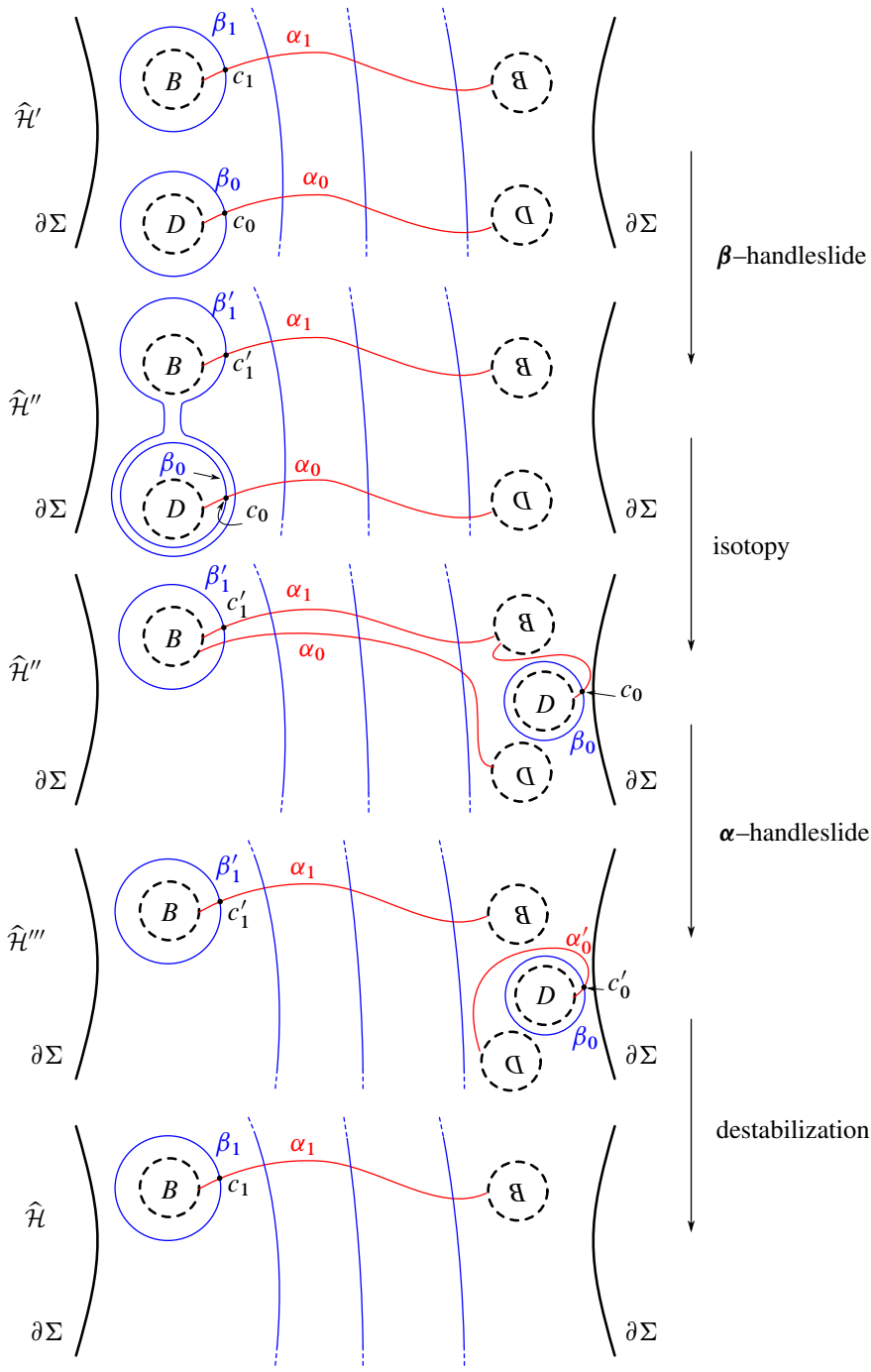


Figure 6: A sequence of Heegaard diagrams from $\hat{\mathcal{H}}'$ to $\hat{\mathcal{H}}$.

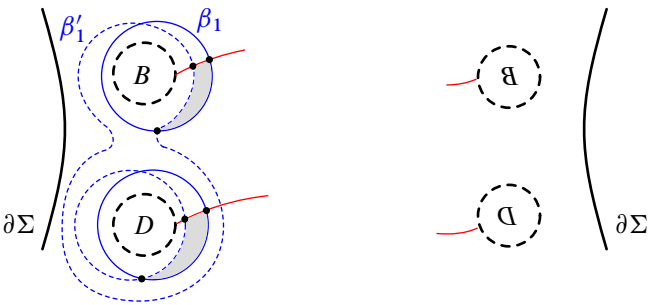


Figure 7: Computing $\Psi_{\widehat{\mathcal{H}}' \rightarrow \widehat{\mathcal{H}}''}$.

By Lemma 2.5, we have that

(2-4)
$$\Psi_{\widehat{\mathcal{H}}'' \rightarrow \widehat{\mathcal{H}}'''}(\mathbf{x} \times c_0 \times c'_1) = \mathbf{x} \times c'_0 \times c'_1,$$

where $c'_0 = \alpha'_0 \cap \beta_0$. Equations (2-3) and (2-4) imply that

(2-5)
$$\Psi_{\widehat{\mathcal{H}}' \rightarrow \widehat{\mathcal{H}}'''}(\mathbf{x} \times c_0 \times c_1) = \mathbf{x} \times c'_0 \times c'_1.$$

The diagrams $\widehat{\mathcal{H}}'''$ and $\widehat{\mathcal{H}}$ are related by a destabilization of the curves α'_0 and β_0 . Hence

(2-6)
$$\Psi_{\widehat{\mathcal{H}}''' \rightarrow \widehat{\mathcal{H}}}(\mathbf{x} \times c'_0 \times c'_1) = \mathbf{x} \times c_1.$$

In $\widehat{\mathcal{H}}$, the only α -curve that intersects β_1 is α_1 , and $\alpha_1 \cap \beta_1 = \{c_1\}$, hence every generator in $\widehat{\mathcal{H}}$ is of the form $\mathbf{y} \times c_1$ for some $\mathbf{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$. So, there are constants $c_{\mathbf{x}, \mathbf{y}} \in \mathbb{F}_2$ such that

$$\Psi_{\mathcal{H} \rightarrow \widehat{\mathcal{H}}}(\mathbf{x}) = \sum_{\mathbf{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta} c_{\mathbf{x}, \mathbf{y}} (\mathbf{y} \times c_1).$$

If we substitute this into (2-2), we get that

$$\Psi_{\mathcal{H}' \rightarrow \widehat{\mathcal{H}}'}(\mathbf{x} \times c_0) = \sum_{\mathbf{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta} c_{\mathbf{x}, \mathbf{y}} (\mathbf{y} \times c_0 \times c_1).$$

Together with (2-5) and (2-6), we arrive at the equality

$$\begin{aligned} \Psi_{\mathcal{H}' \rightarrow \widehat{\mathcal{H}}}(\mathbf{x} \times c_0) &= (\Psi_{\widehat{\mathcal{H}}''' \rightarrow \widehat{\mathcal{H}}} \circ \Psi_{\widehat{\mathcal{H}}' \rightarrow \widehat{\mathcal{H}}'''} \circ \Psi_{\mathcal{H}' \rightarrow \widehat{\mathcal{H}}'})(\mathbf{x} \times c_0) \\ &= \sum_{\mathbf{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta} c_{\mathbf{x}, \mathbf{y}} (\mathbf{y} \times c_1) = \Psi_{\mathcal{H} \rightarrow \widehat{\mathcal{H}}}(\mathbf{x}). \end{aligned}$$

Hence, we obtain that

$$\Psi_{\mathcal{H}' \rightarrow \mathcal{H}}(\mathbf{x} \times c_0) = (\Psi_{\widehat{\mathcal{H}} \rightarrow \mathcal{H}} \circ \Psi_{\mathcal{H}' \rightarrow \widehat{\mathcal{H}}})(\mathbf{x} \times c_0) = (\Psi_{\widehat{\mathcal{H}} \rightarrow \mathcal{H}} \circ \Psi_{\mathcal{H} \rightarrow \widehat{\mathcal{H}}})(\mathbf{x}) = \mathbf{x}.$$

We note that the last equality follows from naturality. Hence $\Psi_{\mathcal{H}' \rightarrow \mathcal{H}}(\mathbf{x} \times c_0) = \mathbf{x}$, completing the proof of [Proposition 2.2](#). $\square$

3 Contact cell decompositions and the gluing map

In this section, we give a definition of the contact gluing map using contact cell decompositions, and prove invariance. The construction is similar to the one due to Honda, Kazez, and Matić [\[12\]](#). On a formal level, the gluing map is described as follows. Suppose that (M, γ) is a sutured submanifold of (M', γ') and that ξ is a co-oriented contact structure on $M' \setminus \text{int}(M)$ such that ∂M is a convex surface with dividing set γ and $\partial M'$ is a convex surface with dividing set γ' . Note that this implies that a contact vector field positively transverse to ∂M lies on the positive side of ξ along $R_-(M, \gamma)$, and on the negative side of ξ along $R_+(M, \gamma)$. In this situation, there is an induced map

$$\Phi_\xi: \text{SFH}(-M, -\gamma) \rightarrow \text{SFH}(-M', -\gamma'),$$

called the gluing map.

3.1 Sutured cell decompositions

In order to discuss cell decompositions of contact 3-manifolds, we need the following notion of cell decomposition for surfaces with divides:

Definition 3.1 Let F be a closed, orientable surface, and $\gamma \subseteq F$ a dividing set. A *sutured cell decomposition* of (F, γ) consists of the following:

- **Fattened 0-cells** A collection of pairwise disjoint disks $B_1, \dots, B_n \subseteq F$ such that each $B_i \cap \gamma$ is an arc and each component of γ intersects some B_i .
- **1-cells** A collection of pairwise disjoint, properly embedded arcs

$$\lambda_1, \dots, \lambda_m \subseteq F \setminus \bigcup_{i=1}^n \text{int}(B_i)$$

disjoint from γ such that each component of

$$F \setminus (B_1 \cup \dots \cup B_n \cup \lambda_1 \cup \dots \cup \lambda_m)$$

intersects γ nontrivially and is homeomorphic to an open disk.

A sutured cell decomposition of a torus with two parallel divides is illustrated in [Figure 8](#).

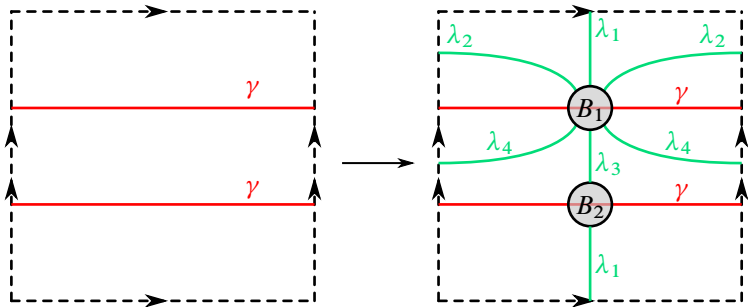


Figure 8: A torus F with two parallel divides γ (left) and a sutured cell decomposition (right).

Remark 3.2 Let $F_0 \subseteq F$ be the closure of a connected component of $F \setminus (B_1 \cup \dots \cup B_n \cup \lambda_1 \cup \dots \cup \lambda_m)$.

Since each component of γ intersects some B_i , the dividing set $\gamma \cap F_0$ contains no closed curves. Hence, according to Giroux’s criterion [10, Theorem 3.5], if F is a convex surface in the contact manifold (M, ξ) , and $\partial B_1 \cup \dots \cup \partial B_n \cup \lambda_1 \cup \dots \cup \lambda_m$ is a Legendrian graph, then F_0 has a tight neighborhood in M .

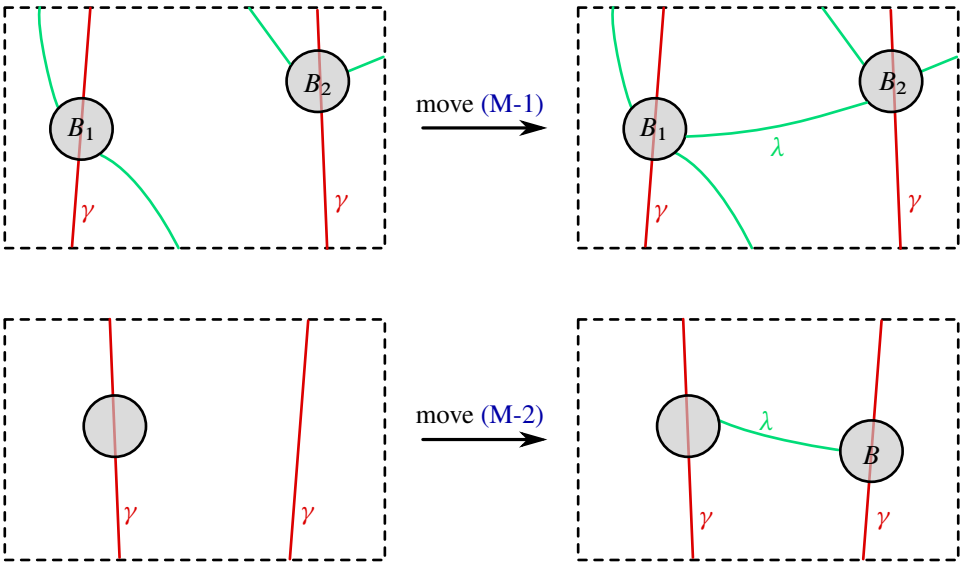


Figure 9: Moves (M-1) and (M-2) between sutured cell decompositions of (F, γ) . The gray disks are the fattened 0-cells, and the green arcs marked λ are the Legendrians. The red arcs marked γ form the dividing set.

We now describe two moves between sutured cell decompositions $\mathcal{D}$ and $\mathcal{D}'$ of a surface F with dividing set γ :

- (M-1) $\mathcal{D}'$ is obtained from $\mathcal{D}$ by adding or removing a 1-cell $\lambda \subseteq F \setminus \gamma$ that has both ends on boundaries of fattened 0-cells in $\mathcal{D}$, but is otherwise disjoint from the 0- and 1-cells of $\mathcal{D}$.
- (M-2) $\mathcal{D}'$ is obtained from $\mathcal{D}$ by adding or removing a fattened 0-cell B and a 1-cell λ that connects B to another 0-cell in $\mathcal{D}$, but is otherwise disjoint from the 0- and 1-cells of $\mathcal{D}$.

Moves (M-1) and (M-2) are illustrated in Figure 9.

Lemma 3.3 *If F is a closed, orientable surface and $\gamma \subseteq F$ is a dividing set, then any two sutured cell decompositions of (F, γ) can be connected by a sequence of moves (M-1) and (M-2).*

To prove Lemma 3.3, it is convenient to consider the following notion of cell decomposition for surfaces with divides which is more general than Definition 3.1:

Definition 3.4 *A generalized sutured cell decomposition $\mathcal{D}^\bullet$ of a surface F with divides γ consists of the following:*

- **Fattened 0-cells** A collection of pairwise disjoint disks $B_1, \dots, B_n \subseteq F$ such that each $B_i \cap \gamma$ is an arc and each component of γ intersects some B_i .
- **1-cells** A collection of pairwise disjoint, properly embedded arcs

$$\lambda_1, \dots, \lambda_m \subseteq F \setminus \bigcup_{i=1}^n \text{int}(B_i)$$

which are transverse to γ such that each component of

$$F \setminus (B_1 \cup \dots \cup B_n \cup \lambda_1 \cup \dots \cup \lambda_m)$$

is homeomorphic to an open disk (which may be disjoint from γ).

- **Splitting arcs** A collection of oriented, properly embedded, pairwise disjoint arcs

$$c_1, \dots, c_l \subseteq \bigcup_{i=1}^n B_i,$$

disjoint from each of the λ_i , such that $|c_i \cap \gamma| = 1$ for all i and such that each component of the closure of $F \setminus (B_1 \cup \dots \cup B_n \cup \lambda_1 \cup \dots \cup \lambda_m)$ which is disjoint from γ intersects the terminal endpoint of at least one c_i .

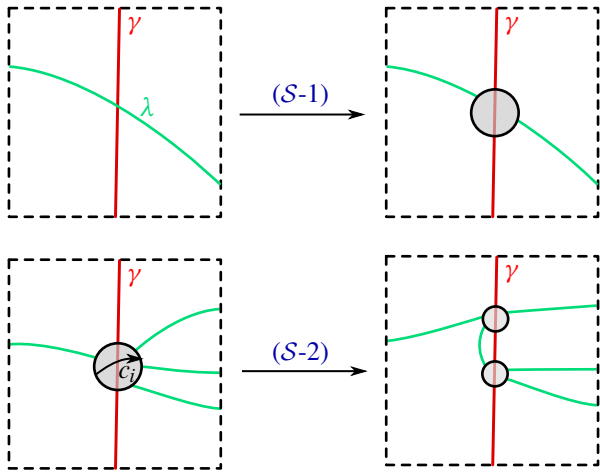


Figure 10: Splitting a generalized sutured cell decomposition $\mathcal{D}^\bullet$. On the left is $\mathcal{D}^\bullet$, and on the right is the split, $\mathcal{S}(\mathcal{D}^\bullet)$.

Note that a sutured cell decomposition can be viewed as a generalized sutured cell decomposition with no splitting arcs. In the other direction, if $\mathcal{D}^\bullet$ is a generalized sutured cell decomposition of a surface F with dividing set γ , we can define the *split* of $\mathcal{D}^\bullet$, denoted by $\mathcal{S}(\mathcal{D}^\bullet)$, as the (genuine) sutured cell decomposition obtained by performing the following two modifications to $\mathcal{D}^\bullet$:

- (S-1) A small, fattened 0-cell is added at each intersection point between a 1-cell λ_i and the dividing set γ .
- (S-2) The fattened 0-cells are split in half along the arcs $c_1, \dots, c_l$. A new 1-cell is added for each of the arcs c_i , as shown in Figure 10. The new 1-cell is on the initial side of c_i , with respect to the orientation of c_i .

Proof of Lemma 3.3 Suppose $\mathcal{D}_1$ and $\mathcal{D}_2$ are two sutured cell decompositions of (F, γ) . By move (M-2), we can increase the number of fattened 0-cells in both, and hence we can assume that they have the same number of fattened 0-cells. It is not hard to see that we can isotope or rescale a fattened 0-cell by a sequence of moves (M-1) and (M-2). The procedure is illustrated in Figure 11. So we can assume that $\mathcal{D}_1$ and $\mathcal{D}_2$ have the same fattened 0-cells.

We can view the fattened 0-cells as 0-handles of the surface, and the 1-cells $\lambda_1, \dots, \lambda_n$ as the cores of 1-handles. Therefore, a generalized sutured cell decomposition with

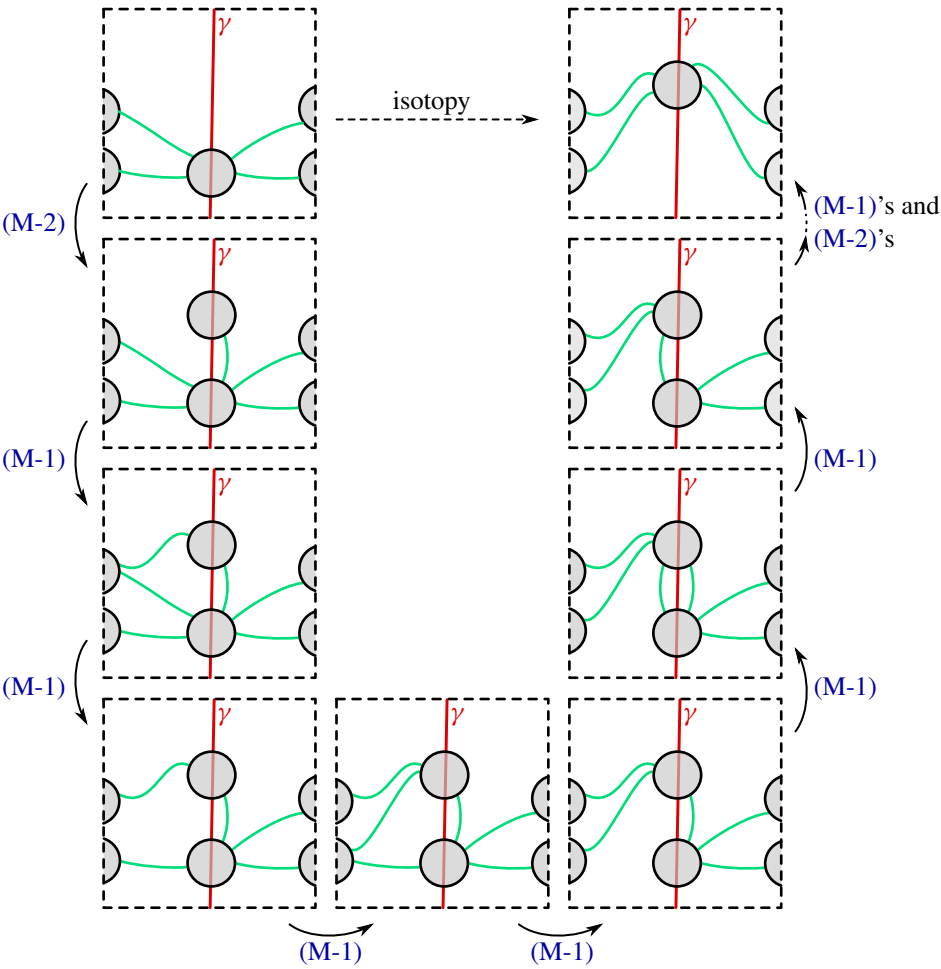


Figure 11: Isotoping a fattened 0-cell via a sequence of moves (M-1) and (M-2).

fattened 0-cells $B_1, \dots, B_n$ corresponds to a Morse function

$$f: \Sigma \setminus \text{int}(B_1 \cup \dots \cup B_n) \rightarrow [0, \infty)$$

that has no index 0 critical points, achieves its minimal value of 0 on $\partial B_1 \cup \dots \cup \partial B_n$, and is such that the union of the descending manifolds of the index 1 critical points is $\lambda_1 \cup \dots \cup \lambda_m$. The collection of splitting arcs $c_1, \dots, c_l$ is some additional combinatorial data, which can always be constructed, given such a Morse function. Two such Morse functions can always be connected by a path of smooth functions with no index 0 critical points relative to $\partial B_1 \cup \dots \cup \partial B_n$ (see [Lemma 7.6](#) for a proof of a closely

related result that adapts to our present setting). Consequently, two generalized sutured cell decompositions can be connected by the following moves:

- (gM-1) Adding or removing a splitting arc disjoint from all the other splitting arcs and 1-cells.
- (gM-2) Isotoping a 1-cell.
- (gM-3) Adding or removing a 1-cell.
- (gM-4) Arc sliding a 1-cell λ_i across another 1-cell λ_j .

We can view any sutured cell decomposition as a generalized sutured cell decomposition with no splitting arcs, and with no intersections between the 1-cells and the dividing set. Hence, to prove the main claim, it suffices to show that if two generalized sutured cell decompositions $\mathcal{D}_1^\bullet$ and $\mathcal{D}_2^\bullet$ differ by moves (gM-1)–(gM-4), then their splits $\mathcal{S}(\mathcal{D}_1^\bullet)$ and $\mathcal{S}(\mathcal{D}_2^\bullet)$ differ by a sequence of moves (M-1) and (M-2).

We first address move (gM-1). We illustrate in Figure 12 an example of how to connect the splits of two generalized sutured cell decompositions that differ by move (gM-1).

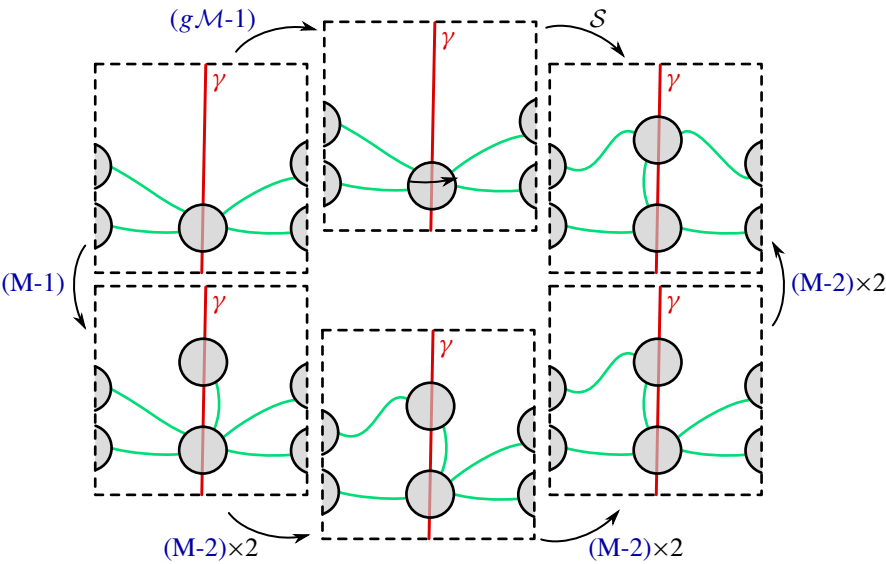


Figure 12: Connecting $\mathcal{S}(\mathcal{D}_1^\bullet)$ and $\mathcal{S}(\mathcal{D}_2^\bullet)$ by a sequence of moves (M-1) and (M-2) when $\mathcal{D}_1^\bullet$ and $\mathcal{D}_2^\bullet$ differ by move (gM-1).

Next, we consider move (gM-2), isotopies of a 1-cell λ_i . Moves of type (gM-2) can be further broken down into three subtypes:

- (1) Isotopies supported outside a neighborhood of γ .
- (2) Isotopies supported in a neighborhood of γ that are fixed on the circles ∂B_j and which either create or cancel a pair of intersection points between γ and λ_i .
- (3) Isotopies supported in a neighborhood of a single ∂B_i that isotope an end of a 1-cell λ_i across γ .

Isotopies of type (1) can be addressed using a manipulation similar to the one shown in Figure 11, so we leave this case to the reader. In case (2), a sequence of moves (M-1) and (M-2) suffice; an example is shown in Figure 13. We leave case (3) to the reader, since the manipulation is similar.

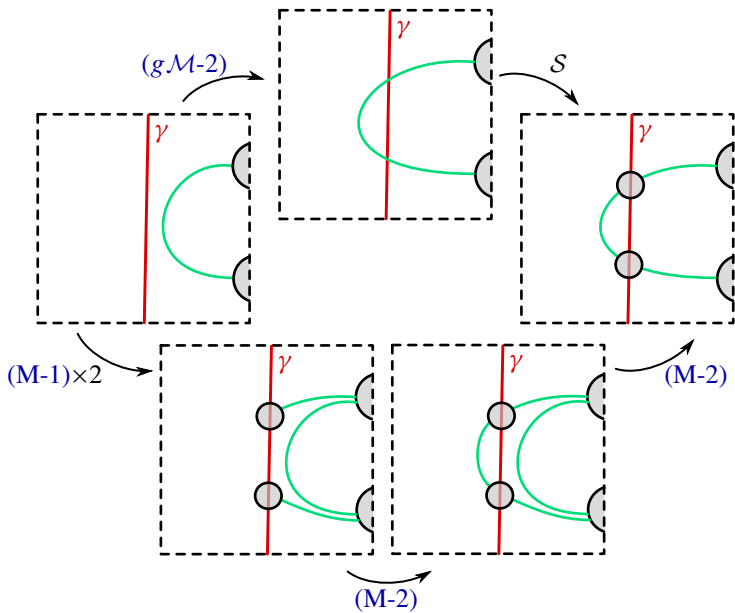


Figure 13: An example of connecting $S(\mathcal{D}_1^\bullet)$ and $S(\mathcal{D}_2^\bullet)$ by a sequence of moves (M-1) and (M-2) when $\mathcal{D}_1^\bullet$ and $\mathcal{D}_2^\bullet$ differ by move $(gM-2)$.

Sufficiency of moves (M-1) and (M-2) to connect splits of generalized sutured cell decompositions differing by moves $(gM-3)$ and $(gM-4)$ is proven in an analogous fashion. We leave the argument as an easy exercise for the reader. $\square$

3.2 Contact cell decompositions

In this section, we describe some background on contact cell decompositions. The technical content is due to Honda, Kazez, and Matić [13, Section 1.1] and Giroux [8].

Definition 3.5 Suppose that (M, γ) is a sutured submanifold of (M', γ') and ξ is a contact structure on $Z := M' \setminus \text{int}(M)$ such that ∂Z is convex with dividing set $\gamma \cup \gamma'$. A *contact cell decomposition* of (Z, ξ) consists of the following data:

- (1) A nonvanishing contact vector field v defined on a neighborhood of ∂Z in Z and transverse to ∂Z such that it induces the dividing set $\gamma \cup \gamma'$. The flow of v induces a diffeomorphism of $\partial Z \times I$ with a collar neighborhood of ∂Z in Z . Under this diffeomorphism, v corresponds to $\partial/\partial t$, the boundary ∂M is identified with $\partial M \times \{0\}$, and $\partial M'$ is identified with $\partial M' \times \{1\}$. We let $v = v|_{\partial M \times I}$ and $v' = v|_{\partial M' \times I}$.
- (2) “Barrier” surfaces $S \subseteq \partial M \times (0, 1)$ and $S' \subseteq \partial M' \times (0, 1)$ in Z that are isotopic to ∂M and $\partial M'$, respectively, and are transverse to v . Write N and N' for the collar neighborhoods of ∂M and $\partial M'$ in Z that are bounded by S and S' , respectively, and set $Z' = Z \setminus \text{int}(N \cup N')$.
- (3) A Legendrian graph $\Gamma \subseteq Z'$ that intersects $\partial Z'$ transversely in a finite collection of points along the dividing set of $\partial Z'$ with respect to the vector field v . Furthermore, Γ is tangent to the vector field v near $\partial Z'$.
- (4) A choice of regular neighborhood $N(\Gamma)$ of Γ such that $\xi|_{N(\Gamma)}$ is tight and $\partial N(\Gamma) \setminus \partial Z'$ is a convex surface. Furthermore, $N(\Gamma) \cap \partial Z'$ is a collection of disks D with Legendrian boundary such that $\text{tb}(\partial D) = -1$. We assume that the edge rounding procedure of Honda [10, Section 3.3.2] has been performed so that $N(\Gamma)$ meets $\partial Z'$ tangentially along the Legendrian unknots forming $\partial N(\Gamma)$.
- (5) A collection of convex 2-cells $D_1, \dots, D_n$ inside $Z' \setminus \text{int}(N(\Gamma))$ with Legendrian boundary on $\partial(Z' \setminus \text{int}(N(\Gamma)))$ and $\text{tb}(\partial D_i) = -1$.

Furthermore, the following hold:

- (a) The complement of $N(\Gamma) \cup D_1 \cup \dots \cup D_n$ in Z' is a finite collection of topological 3-balls, and ξ is tight on each.
- (b) The disks in $N(\Gamma) \cap \partial Z'$ and the Legendrian arcs $\partial D_i \cap \partial Z'$ induce a sutured cell decomposition of $\partial Z'$ with dividing set $\{x \in \partial Z' : v_x \in \xi_x\}$ (Definition 3.1).

Remark 3.6 Given surfaces S and S' in Z and a transverse contact vector field v defined in a neighborhood of ∂Z that satisfy (1) and (2), it is not always possible to construct a contact cell decomposition of Z with S and S' as barrier surfaces. For example, the characteristic foliations on S and S' may obstruct the existence of sutured cell decompositions (Definition 3.1) such that the arcs λ_i and the curves ∂B_i are Legendrian.

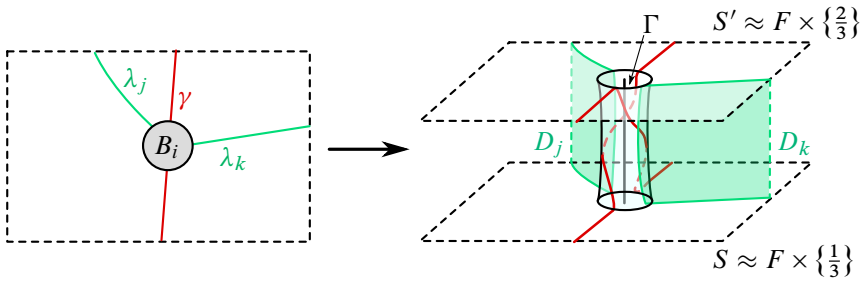


Figure 14: The left shows a sutured cell decomposition of a surface F with dividing set γ . The right shows the induced product contact cell decomposition of an I -invariant contact structure on $F \times I$ with dividing set γ from [Example 3.7](#). The barrier surfaces are S and S' , which are C^0 close to $F \times \{\frac{1}{3}\}$ and $F \times \{\frac{2}{3}\}$, respectively. The fattened 0-cell B_i is shown on the left in gray. The dividing set γ is shown in red. The Legendrian arcs λ_j and λ_k , as well as the corresponding convex 2-cells D_j and D_k , are shown in green.

We now describe an important example of contact cell decompositions:

Example 3.7 Suppose that (M, γ) is a sutured submanifold of (M', γ') and ξ is a compatible contact structure on $Z = M' \setminus \text{int}(M)$. Furthermore, suppose that (Z, ξ) is contactomorphic to $F \times I$ with an I -invariant contact structure for a surface $F \cong \partial M \cong \partial M'$. Suppose that the disks $B_1, \dots, B_n \subseteq F$ and the (not necessarily Legendrian) properly embedded arcs $\lambda_1, \dots, \lambda_m \subseteq F \setminus \text{int}(B_1 \cup \dots \cup B_n)$ give a sutured cell decomposition of $F \times \{0\}$, as in [Definition 3.1](#).

There is an induced contact cell decomposition of Z , called the *product contact cell decomposition*, which we describe presently. For an illustration, see [Figure 14](#). Let $G \subseteq F$ be the graph

$$G = \partial B_1 \cup \dots \cup \partial B_n \cup \lambda_1 \cup \dots \cup \lambda_m;$$

this graph is not necessarily Legendrian. However, using Giroux's flexibility theorem [\[6, Proposition 3.6\]](#), we can find a surface $S \subseteq F \times I$ that is the image of a C^0 -small isotopy of $F \times \{\frac{1}{3}\}$ that is transverse to $\partial/\partial t$ and such that the image of $G \times \{\frac{1}{3}\}$ is Legendrian. We let S' be the translation of S by $\frac{1}{3}$ in the $\partial/\partial t$ direction. We use S and S' for the barrier surfaces of our contact cell decomposition. We will also write $B_1, \dots, B_n$ and $\lambda_1, \dots, \lambda_m$ for the images on S of the corresponding cells on $F \times \{\frac{1}{3}\}$, under the chosen C^0 -small isotopy. Let N and N' be the collars of ∂M and $\partial M'$ bounded by S and S' , respectively, and write $Z' = Z \setminus \text{int}(N \cup N')$.

For every $i \in \{1, \dots, n\}$, choose a point $p_i \in \text{int}(\gamma \cap B_i)$. We define the Legendrian graph Γ as

$$\bigcup_{i=1}^n (\{p_i\} \times I) \cap Z',$$

and set $N(\Gamma) = \bigcup_{i=1}^n (B_i \times I) \cap Z'$. The contact 2-cells of the decomposition are defined to be $D_i := (\lambda_i \times I) \cap Z'$. Using Giroux's flexibility theorem on $\partial N(\Gamma)$, we can assume that each D_i has Legendrian boundary. The disks D_i have $\text{tb}(\partial D_i) = -1$ since their boundaries intersect the dividing set exactly twice. If C is a component of $S \setminus (B_1 \cup \dots \cup B_n \cup \lambda_1 \cup \dots \cup \lambda_m)$, then $(C \times I) \cap Z'$ is a tight contact ball in the complement of $N(\Gamma) \cup D_1 \cup \dots \cup D_n$ in Z' .

To show invariance of the gluing map, we need to describe how two contact cell decompositions are related. We have the following; cf [13, Theorem 1.2]:

Proposition 3.8 *Suppose that (M, γ) is a sutured submanifold of (M', γ') , and ξ is a contact structure on $Z = M' \setminus \text{int}(M)$ such that ∂Z is convex with dividing set $\gamma \cup \gamma'$. If $\mathcal{C}_1$ and $\mathcal{C}_2$ are contact cell decompositions of Z , then there is a sequence $\mathcal{C}_{(1)}, \dots, \mathcal{C}_{(\ell)}$ of contact cell decompositions with $\mathcal{C}_1 = \mathcal{C}_{(1)}$ and $\mathcal{C}_2 = \mathcal{C}_{(\ell)}$ such that $\mathcal{C}_{(i+1)}$ is obtained from $\mathcal{C}_{(i)}$ by one of the following moves, or its inverse:*

- (C-1) **Isotopy** Replacing a contact cell decomposition $\mathcal{C}$ with a contact cell decomposition of the form $\phi(\mathcal{C})$, where $\phi \in \text{Diff}(Z)$ (not necessarily a contactomorphism) fixes ∂Z pointwise and is isotopic to the identity relative to ∂Z .
- (C-2) **Index 0/1 cancellation** Subdividing a Legendrian edge of Γ or adding a Legendrian edge that has one endpoint on Γ , but which is otherwise disjoint from Γ and the 2-cells of $\mathcal{C}$.
- (C-3) **Index 1/2 cancellation** Adding a Legendrian edge λ to Γ and adding a convex 2-cell D with $\text{tb}(\partial D) = -1$ such that the interiors of λ and D are disjoint from all the other cells, and $\partial D = c' \cup c''$, where c' is a Legendrian arc on a neighborhood of λ , and $c'' \subseteq \partial(Z' \setminus \text{int } N(\Gamma))$ is a Legendrian arc that is disjoint from the dividing set on $\partial(Z' \setminus \text{int } N(\Gamma))$.
- (C-4) **Index 2/3 cancellation** Adding a convex disk D with $\partial D \subseteq \partial(Z' \setminus \text{int } N(\Gamma))$ and $\text{tb}(\partial D) = -1$ disjoint from the other cells.

Proof The result follows from an adaptation of the subdivision techniques due to Giroux [8] and Honda, Kazez, and Matić [13, Theorem 1.2]. Giroux's technique

involves the following move: Replace a convex 2-cell D with a pair of convex 2-cells D_1 and D_2 with $\text{tb}(\partial D_i) = -1$ that meet along a Legendrian arc λ such that $D_1 \cup D_2 = D$ and $\lambda \subseteq D$ is an arc that intersects the dividing set of D transversely at a single point.

Lemma 3.9 *The above subdivision of a 2-cell can be achieved using moves (C-1) through (C-4).*

Proof Let D' be a parallel copy of D , intersecting D only along ∂D . Let $D'_1 \subseteq D'$, $D'_2 \subseteq D'$, and $\lambda' \subseteq D'$ be the images of D_1 , D_2 , and λ , respectively. We can add D'_1 and λ' to the decomposition using move (C-3). Then we can add D'_2 to the decomposition using move (C-4). Then we remove D from the decomposition using the inverse of move (C-4). Finally, we use move (C-1) to move D'_1 , D'_2 , and λ' into the positions of D_1 , D_2 , and λ , while preserving all the other cells. $\square$

Step 1 *Using move (C-1), we can change the vector field v_1 of $\mathcal{C}_1$ to the vector field v_2 of $\mathcal{C}_2$.*

We will focus on changing the contact vector field $v_1 = v_1|_{\partial M \times I}$ to the vector field $v_2 = v_2|_{\partial M \times I}$, as changing $v'_1 = v_1|_{\partial M' \times I}$ to the vector field $v'_2 = v_2|_{\partial M' \times I}$ is analogous. The idea is that the space of germs of contact vector fields defined on open neighborhoods of ∂M that induce the dividing set γ is convex and hence contractible. There is an open neighborhood of ∂M where the convex combinations $v_t = (2-t)v_1 + (t-1)v_2$ are nonvanishing for $t \in [1, 2]$. Then v_t is transverse to ∂M for every $t \in [1, 2]$. We can define an isotopy ψ_t of a neighborhood of ∂M , by flowing a point $p \in M$ backward along v_1 until it hits ∂M (say, at a time $-T(p)$), then flowing along v_t for time $T(p)$. This yields a 1-parameter family of contactomorphic embeddings of a neighborhood U of ∂M into Z . We can extend this to a family of contactomorphisms ψ_t of all of M for $t \in [1, 2]$ by writing ψ_t as the integral of a time-dependent contact vector field X_t (ie X_t is a contact vector field for each t) defined over $\partial M \times I$ and then extending X_t to all of Z using time-dependent contact Hamiltonians that vanish outside a neighborhood of ∂M . Using move (C-1), we can isotope S_1 into U using move (C-1) and then push $\mathcal{C}_1$ forward under ψ_1 .

Step 2 *If $\mathcal{C}_1$ and $\mathcal{C}_2$ are contact cell decompositions with the same vector fields v and v' , then we can achieve that $\mathcal{C}_1$ and $\mathcal{C}_2$ have the same barrier surfaces with identical induced sutured cell decompositions using moves (C-1) through (C-4).*

Let $\mathcal{D}_i$ and $\mathcal{D}'_i$ be the sutured cell decompositions induced on the barrier surfaces S_i and S'_i of the contact cell decomposition $\mathcal{C}_i$ for $i \in \{1, 2\}$. Write $\pi: \partial M \times I \rightarrow \partial M$

and $\pi': \partial M' \times I \rightarrow \partial M'$ for the projections. Note that $\pi(\mathcal{D}_1)$ and $\pi(\mathcal{D}_2)$ are sutured cell decompositions of ∂M , and $\pi'(\mathcal{D}'_1)$ and $\pi'(\mathcal{D}'_2)$ are sutured cell decompositions of $\partial M'$. Let us focus on the barrier surfaces S_1 and S_2 near ∂M , and their sutured cell decompositions $\mathcal{D}_1$ and $\mathcal{D}_2$. By Lemma 3.3, it follows that $\pi(\mathcal{D}_1)$ and $\pi(\mathcal{D}_2)$ can be connected by moves (M-1) and (M-2).

We will show that, using moves (C-1)–(C-4), we can connect $\mathcal{C}_1$ to a contact cell decomposition $\hat{\mathcal{C}}_2$ with the same contact vector field as $\mathcal{C}_1$ and $\mathcal{C}_2$, whose induced sutured cell decomposition of ∂M is C^0 close to $\pi(\mathcal{D}_2)$. It is sufficient to show the claim when $\pi(\mathcal{D}_2)$ is obtained from $\pi(\mathcal{D}_1)$ by a single instance of move (M-1) or (M-2).

Firstly, we can construct two contact automorphisms φ_1 and φ_2 of (Z, ξ) that are isotopic to the identity relative to ∂Z such that $\varphi_1(\mathcal{C}_1)$ and $\varphi_2(\mathcal{C}_2)$ are cell decompositions with Legendrian graphs $\varphi_1(\Gamma_1)$ and $\varphi_2(\Gamma_2)$, respectively, that intersect $\partial M \times I$ and $\partial M' \times I$ along arcs of the form $\{p\} \times [a, 1]$ or $\{p'\} \times [0, a]$ for various $p \in \gamma$, $p' \in \gamma'$, and $a > 0$. Furthermore, by applying move (C-1), we may assume that each component of the intersection of every 2-cell with $\partial Z \times I$ is C^0 close to a set of the form $(\lambda \times I) \cap Z'$ for some arc λ in ∂Z .

Consider first the case when $\pi(\mathcal{D}_2)$ is obtained from $\pi(\mathcal{D}_1)$ by adding a 1-cell, as in move (M-1). Write $\lambda \subseteq \partial M$ for the new 1-cell, which is added to the complement of the other 1-cells of $\pi(\mathcal{D}_1)$. To construct $\hat{\mathcal{C}}_2$, we add a new Legendrian edge e to the graph Γ_1 of $\mathcal{C}_1$, as well as a new 2-cell c , as follows. Let $\lambda_1 \subseteq S_1$ denote the preimage of λ under $\pi|_{S_1}$. Using Legendrian realization that only changes S_1 in the $\partial/\partial t$ -direction, we may assume that the resulting arc $\lambda_1 \subseteq S_1$ is Legendrian, and projects C^0 close to λ under π . Note that the isotopy of S_1 used for Legendrian realization can be chosen to fix the 0- and 1-cells of $\mathcal{D}_1$, and hence can be realized by an instance of move (C-1) according to the following result:

Lemma 3.10 *Let $\mathcal{C}$ be a contact cell decomposition with barrier surface S and induced sutured cell decomposition $\mathcal{D}$ of S . Then any isotopy of S through surfaces that are transverse to $\nu = \partial/\partial t$ that fixes the 0- and 1-cells of $\mathcal{D}$ can be achieved using move (C-1). The same holds for S' , ν' , and $\mathcal{D}'$.*

Proof As above, we can assume that the graph Γ is a union of Legendrians of the form $\{p\} \times [a, 1]$ for various $p \in \gamma$ and $0 < a < 1$, and the 2-cells are small perturbations of sets of the form $(\lambda \times I) \cap Z'$ for arcs $\lambda \subseteq \partial M$. Suppose that $f: S \rightarrow \partial M \times I$ is an embedding that is the identity on the 0- and 1-cells of $\mathcal{D}$ and whose image is transverse to ν . We can extend f to an automorphism of Z that is the identity outside $\partial M \times (0, 1)$.

It is straightforward to see that $f(\mathcal{C})$ is a contact cell decomposition, and that $\mathcal{C}$ and $f(\mathcal{C})$ are related by move (C-1). $\square$

Let $\varepsilon > 0$ be small, and let

$$\theta: S_1 \times (-\varepsilon, \varepsilon) \rightarrow \partial M \times I$$

denote the map induced by the flow of $v = \partial/\partial t$, ie translation in the I -direction. We obtain the new edge e by extending $\theta(\lambda_1 \times \{\varepsilon\})$ into $N(\Gamma_1)$. The new 2-cell c is the intersection of $\theta(\lambda_1 \times [0, \varepsilon])$ with the complement of $N(\Gamma_1 \cup e)$. A schematic is shown in Figure 15. Adding e and c to $\mathcal{C}_1$ is an instance of move (C-3).

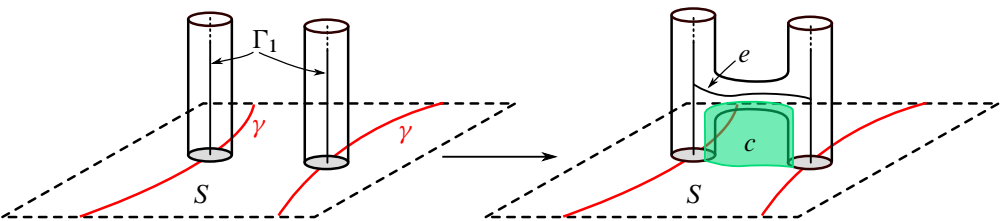


Figure 15: Adding a new edge e and a new 2-cell c to a contact cell decomposition to realize move (M-1) on the induced sutured cell decomposition of ∂M . Shown on the left is a neighborhood of the Legendrian graph Γ_1 , as well as the barrier surface S . Shown on the right is a neighborhood of $\Gamma_1 \cup e$, as well as the new 2-cell c .

In the opposite direction, suppose that $\pi(\mathcal{D}_2)$ is obtained from $\pi(\mathcal{D}_1)$ by removing a 1-cell λ , as in move (M-1). Let c be the 2-cell of $\mathcal{C}_1$ corresponding to $\lambda_1 = (\pi|_{S_1})^{-1}(\lambda)$. Note that the 2-cells of $\pi(\mathcal{D}_1)$ neighboring λ are disjoint, since otherwise there would be a component of the complement of $\pi(\mathcal{D}_2)$ that were not a disk. We subdivide $\mathcal{C}_1$ away from $S \cup S'$ such that the 3-cells of $\mathcal{C}_1$ corresponding to the 2-cells of $\mathcal{D}_1$ on the two sides of λ_1 become distinct. We then subdivide c along $e := \theta(\lambda_1 \times \{\varepsilon\})$ for ε small using Lemma 3.9, and cancel the 2-cell $\theta(\lambda_1 \times [0, \varepsilon]) \setminus N(\Gamma_1 \cup e)$ with one of the neighboring 3-cells using move (C-4).

Next, we consider the case when $\pi(\mathcal{D}_2)$ is obtained from $\pi(\mathcal{D}_1)$ by adding a fattened 0-cell B and a 1-cell λ , as in move (M-2). To construct the new contact cell decomposition $\hat{\mathcal{C}}_2$, we add two new Legendrian edges e_1 and e_2 to $\mathcal{C}_1$ that meet at a valence 2 vertex, as well as a new 2-cell c , as follows. Let $\lambda_1 \subseteq S_1$ denote the Legendrian realization of the preimage under $\pi|_{S_1}$ of an extension of the Legendrian λ to a point $p \in \text{int}(B)$. The Legendrian realization can be achieved using move (C-1) according

to [Lemma 3.10](#). We define the edge e_1 to be $\theta(\lambda_1 \times \{\varepsilon\})$ for some small $\varepsilon > 0$ and e_2 as $\theta(\{p\} \times [0, \varepsilon])$. We let the 2-cell c be $\theta(\lambda_1 \times [0, \varepsilon]) \setminus N(\Gamma_1 \cup e_1 \cup e_2)$. Note that adding e_1 , e_2 , and c to $\mathcal{C}_1$ can be achieved by moves [\(C-2\)](#) and [\(C-3\)](#). Writing $\hat{\mathcal{D}}_2$ for the sutured cell decomposition of S induced by $\hat{\mathcal{C}}_2$, we note that $\pi(\hat{\mathcal{D}}_2)$ is C^0 close to the one obtained from $\pi(\mathcal{D}_1)$ by move [\(M-2\)](#).

In the opposite direction, suppose that $\pi(\mathcal{D}_2)$ is obtained from $\pi(\mathcal{D}_1)$ by removing a fattened 0-cell B and a 1-cell λ , as in move [\(M-2\)](#). Then we split the 2-cell of $\mathcal{C}_1$ corresponding to $\lambda_1 = (\pi|_{S_1})^{-1}(\lambda)$ along $e_1 := \theta(\lambda_1 \times \{\varepsilon\})$, and cancel the 2-cell $\theta(\lambda_1 \times [0, \varepsilon]) \setminus N(\Gamma_1 \cup e_1)$ and the 1-cell $\theta(\{p\} \times [0, \varepsilon])$ for $p = B \cap \Gamma_1$ using move [\(C-3\)](#).

Having obtained $\hat{\mathcal{C}}_2$ as above, we can construct an isotopy of M that is supported in a neighborhood of ∂M and moves $\hat{\mathcal{C}}_2$ to a contact cell decomposition $\mathcal{C}_2''$ that shares the same barrier surface and induced sutured cell decomposition of ∂M as $\mathcal{C}_2$. It follows that $\hat{\mathcal{C}}_2$ and $\mathcal{C}_2''$ are related by move [\(C-1\)](#). This completes Step 2.

Step 3 Suppose $\mathcal{C}_1$ and $\mathcal{C}_2$ are contact cell decompositions with the same barrier surfaces S and S' and the same contact vector fields v and v' such that $\mathcal{C}_1$ and $\mathcal{C}_2$ induce the same sutured cell decompositions $\mathcal{D}$ and $\mathcal{D}'$ of S and S' and such that the Legendrian 1-skeletons of $\mathcal{C}_1$ and $\mathcal{C}_2$ intersect S and S' at the same points. Then $\mathcal{C}_1$ and $\mathcal{C}_2$ can be connected by moves [\(C-1\)](#) through [\(C-4\)](#).

The proof follows from the strategy of [\[13, Theorem 1.2\]](#), which we summarize using our present notation. As in Step 2, we can assume that the graph Γ is a union of Legendrians of the form $\{p\} \times [a, 1]$ or $\{p'\} \times [0, a]$ for various $p \in \gamma$, $p' \in \gamma'$, and $0 < a < 1$, and the 2-cells are small perturbations of sets of the form $(\lambda \times I) \cap Z'$ for arcs $\lambda \subseteq \partial Z$.

Using [Lemma 3.9](#), we subdivide each 2-cell D of $\mathcal{C}_1$ and $\mathcal{C}_2$ that intersects $\partial Z \times I$ by adding a Legendrian arc obtained by perturbing $D \cap (\partial M \times \{1\} \cup \partial M' \times \{0\})$ along v relative to its endpoints. Furthermore, using move [\(C-4\)](#), we add a new convex 2-cell D with Legendrian boundary and $\text{tb}(\partial D) = -1$ near $\partial M \times \{1\}$ for each 2-cell of the sutured cell decomposition $\mathcal{D}$ and near $\partial M \times \{0\}$ for each 2-cell of $\mathcal{D}'$. Then we apply the subdivision procedure of Giroux [\[8\]](#) away from $\partial Z \times I$. $\square$

3.3 Contact handle maps

We first recall the definition of contact handles in dimension 3 due to Giroux [\[6\]](#); see also Ozbagci [\[21\]](#).

Definition 3.11 For $k \in \{0, 1, 2, 3\}$, a 3-dimensional contact handle of index k attached to a sutured manifold (M, γ) is a tight contact ball (B_0, ξ_0) with convex boundary (possibly with corners) that is attached via a map $\phi: S \rightarrow \partial M$ for some subset $S \subseteq \partial B_0$. Furthermore, the dividing set of ξ_0 on S is mapped into γ under ϕ , and we have the following requirements, depending on the index:

Index 0 $B_0 = D^3$ has no corners, and $S = \emptyset$. The dividing set on ∂B_0 is a single circle.

Index 1 As a manifold with corners, B_0 is $I \times D^2$, and $S = \partial I \times D^2$. The dividing set on $\partial I \times D^2$ consists of one arc on each component. The dividing set on $I \times \partial D^2$ consists of two arcs, each connecting the two components of $\partial I \times \partial D^2$.

Index 2 As a manifold with corners, B_0 is $D^2 \times I$ and $S = \partial D^2 \times I$. The dividing set is the same as on a contact 1-handle.

Index 3 $B_0 = D^3$ has no corners, and $S = \partial D^3$. The dividing set on ∂B_0 is a single circle.

In Figure 16, we have drawn the dividing sets on contact handles. Note that, for handles of index 1 and 2, the curves $\partial I \times \partial D^2$ and $\partial D^2 \times \partial I$, respectively, are not Legendrian.

3.3.1 Contact 0-handle map Adding a contact 0-handle h^0 amounts to adding a copy of the product sutured manifold $(D^2 \times [-1, 1], S^1 \times \{0\})$ to (M, γ) . Noting that

$$\text{SFH}(-D^2 \times [-1, 1], -S^1 \times \{0\}) \cong \mathbb{F}_2,$$

the contact 0-handle map is simply the tautological map

$$\begin{aligned} C_{h^0}: \text{SFH}(-M, -\gamma) &\xrightarrow{\cong} \text{SFH}(-M, -\gamma) \otimes \mathbb{F}_2 \\ &\xrightarrow{\cong} \text{SFH}(-M, -\gamma) \otimes \text{SFH}(-D^2 \times [-1, 1], -S^1 \times \{0\}) \\ &\xrightarrow{\cong} \text{SFH}((-M, -\gamma) \sqcup (-D^2 \times [-1, 1], -S^1 \times \{0\})). \end{aligned}$$

On the chain level, this map is given as follows. Choose a diagram $\mathcal{H} = (\bar{\Sigma}, \alpha, \beta)$ of $(-M, -\gamma)$. Since $\mathcal{H}_0 = (\bar{D}^2, \emptyset, \emptyset)$ is a diagram of $(-D^2 \times [-1, 1], -S^1 \times \{0\})$, the disjoint union $\mathcal{H} \sqcup \mathcal{H}_0$ is a diagram of $(-M, -\gamma) \sqcup (-D^2 \times [-1, 1], -S^1 \times \{0\})$. Let $i: \bar{\Sigma} \rightarrow \bar{\Sigma} \sqcup \bar{D}^2$ be the inclusion. Then, for $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$, we set $C_{h^0}(\mathbf{x}) = (i(x_1), \dots, i(x_d))$. This clearly induces C_{h^0} on homology.

3.3.2 Contact 1-handle map A contact 1-handle h^1 determines two points along the sutures. If (Σ, α, β) is a Heegaard surface for (M, γ) , glue a strip to $\partial \Sigma$ where the

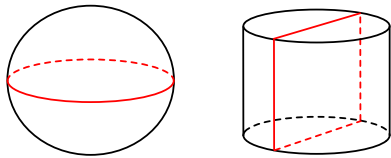


Figure 16: Contact handles. On the left is a picture of a contact 0–handle or 3–handle. On the right is a contact 1–handle or 2–handle.

feet of the 1–handle are attached; see Figure 17. Add no new α or β curves. The map on complexes is the tautological map induced by the inclusion of Heegaard surfaces. Let us call this map C_{h^1} . This is an isomorphism of chain complexes, since the domains of the curves counted by the boundary map on $CF(\Sigma, \alpha, \beta)$ have coefficient zero along $\partial\Sigma$, where the strip is attached; see [14, Lemma 9.13].

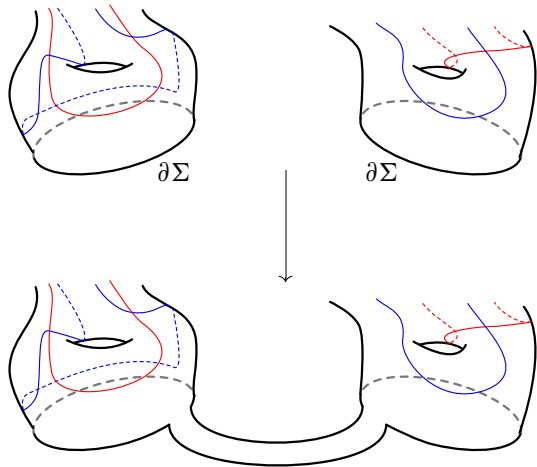


Figure 17: A contact 1–handle, on the level of diagrams.

Lemma 3.12 *The map C_{h^1} is natural; ie it commutes with change of diagrams maps.*

Proof The proof is straightforward, since the change of diagrams maps are either tautological (stabilization, isotopy) or involve counting holomorphic curves that do not intersect $\partial\Sigma \times \text{Sym}^{k-1}(\Sigma) \subseteq \text{Sym}^k(\Sigma)$ (continuation maps for changes of the almost complex structure, triangle maps for changes of the α and β curves). $\square$

3.3.3 Contact 2–handle map We now define a map for contact 2–handles. Let (Σ, α, β) be an admissible diagram of (M, γ) . Suppose that l is the curve on ∂M along which we add a 2–handle h^2 , and l intersects γ (the sutures) exactly twice.

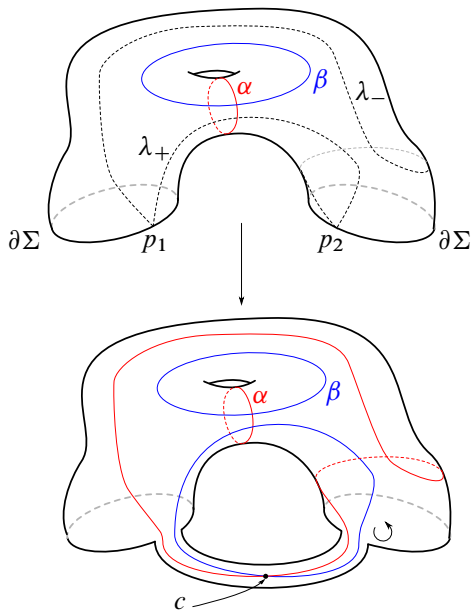


Figure 18: The contact 2–handle map. The orientation of $\bar{\Sigma}$ is shown on the bottom of the surface.

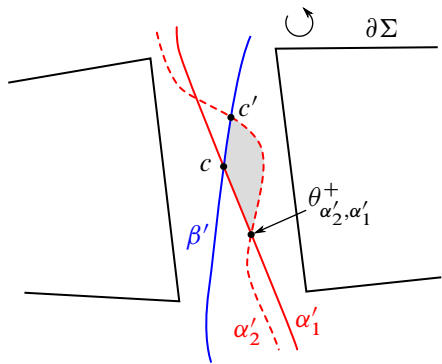


Figure 19: The local computation required to show that the contact 2–handle map is independent of handle slides of the α' curve used in the definition of the contact 2–handle map. A local computation using the vertex multiplicities (identical to the one in [13, Lemma 3.5]) shows that any holomorphic triangle with vertices at c and $\theta_{\alpha'_2, \alpha'_1}^+$ also has a vertex at c' , and has domain equal to the shaded region. The orientation of $\bar{\Sigma}$ is shown.

Let $p_1, p_2 \in \gamma$ be the two points of intersection and $l_{\pm} = l \cap R_{\pm}(\gamma)$. We denote the result of the 2–handle attachment by (M', γ') .

We now construct a diagram of (M', γ') . Choose a sutured Morse function f and gradient-like vector field v on (M, γ) that induce (Σ, α, β) , and follow the flow of v from l_+ and l_- onto Σ . Writing λ_+ and λ_- for the resulting arcs on Σ , we note that λ_+ avoids the β curves and λ_- avoids the α curves. Hence, we can form a diagram $(\Sigma', \alpha \cup \{\alpha'\}, \beta \cup \{\beta'\})$, as in Figure 18. The surface Σ' is obtained by adding a band B to the boundary of Σ at p_1 and p_2 . The curve α' is obtained by concatenating the curve λ_- with an arc in the band. The curve β' is obtained by concatenating the curve λ_+ with an arc in the band. We assume that α' and β' intersect in a single point in the band (and possibly other places outside the band). Furthermore, we assume that the intersection point of α' and β' has the configuration shown in Figure 18; ie locally, it looks like there is a holomorphic disk on $(\bar{\Sigma}', \alpha \cup \{\alpha'\}, \beta \cup \{\beta'\})$ going towards the intersection point that does not intersect $\partial \Sigma'$.

To see that $(\Sigma', \alpha \cup \{\alpha'\}, \beta \cup \{\beta'\})$ is indeed a diagram of (M', γ') , let

$$J = [-1, 1] \times \{0\} \subseteq D^2 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}.$$

Consider the arc $a = J \times \{\frac{1}{2}\}$ in the 2-handle $D^2 \times I \subseteq M'$ connecting p_1 and p_2 . Furthermore, consider the neighborhood

$$N(a) = (([-1, 1] \times (-\varepsilon, \varepsilon)) \cap D^2) \times I$$

of a , where $\varepsilon \in (0, 1)$. Then $N(a) \cap R_{\pm}(\gamma')$ is a regular neighborhood of

$$\gamma' \cap (D^2 \times I) = J \times \{0, 1\},$$

and $N(a)$ is a product 1-handle. On the sutured diagram level, attaching $N(a)$ corresponds to adding the band $B = J \times I$. Then $(D^2 \times I) \setminus N(a)$ is the disjoint union of two 3-dimensional 2-handles, attached to $(M \cup N(a), \gamma')$ along $l_{\pm} \cup s_{\pm}$, where $s_{\pm}$ is a properly embedded arc in $N(a) \cap R_{\pm}(\gamma')$. These attaching curves correspond to α' and β' on the Heegaard surface Σ' . The diagram $(\Sigma', \alpha \cup \{\alpha'\}, \beta \cup \{\beta'\})$ is also admissible.

The contact 2-handle map

$$C_{h^2}: \text{SFH}(\bar{\Sigma}, \alpha, \beta) \rightarrow \text{SFH}(\bar{\Sigma}', \alpha \cup \{\alpha'\}, \beta \cup \{\beta'\})$$

is defined for $x \in \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}$ by the formula

$$C_{h^2}(x) := x \times c,$$

where $c \in \alpha' \cap \beta'$ is the intersection point in the band.

Lemma 3.13 (1) *The map C_{h^2} is a chain map.*

(2) *The map C_{h^2} is independent of the choices made in the construction, ie the choice of diagram (Σ, α, β) and the choice of arcs λ_+ and λ_- obtained by projecting the two arcs l_+ and l_- onto Σ .*

Proof The claim that C_{h^2} is a chain map is straightforward. We wish to show that $\partial \circ C_{h^2} = C_{h^2} \circ \partial$, and hence we need to check that $\partial(x \times c) = \partial x \times c$. To this end, we note that any disk counted by $\partial(x \times c)$ must have zero multiplicity around c , since the boundary $\partial\Sigma$ is nearby. Hence, the homology class of the disk is equal to a class on $(\bar{\Sigma}, \alpha, \beta)$ with the constant class at c added.

We now show that C_{h^2} is independent of the choices made in the construction (ie commutes with the change of diagrams maps, appropriately). We note that there are two sources of ambiguity: the curves λ_+ and λ_- in Σ , and the diagram (Σ, α, β) . Let us address the ambiguity of λ_+ and λ_- . Since they are gotten by flowing curves in ∂M under the gradient-like vector field of a Morse function, any two choices of λ_+ are related by handle slides over β curves (as well as any isotopy of λ_+ , relative $\partial\Sigma$, such that it never intersects any β curves). Similarly, any two choices of λ_- are related by handle slides over the α curves and isotopies within $\Sigma \setminus \alpha$. To see that the change of diagrams map commutes with C_{h^2} , one realizes the handle slide and isotopy maps as triangle maps, and uses the local computation at the end of the proof of [13, Lemma 3.5], which is illustrated by [13, Figure 8]. For the reader's convenience, we have reproduced the relevant picture in Figure 19, since we will use the local computation later also. $\square$

3.3.4 Contact 3-handle maps We will now define the contact 3-handle map. Let (M', γ') be the result of gluing a contact 3-handle h^3 to the balanced sutured manifold (M, γ) , and suppose that (M', γ') is also balanced. The 3-handle is attached to a 2-sphere component $S \subseteq \partial M$ such that $\gamma_S = \gamma \cap S$ is a single closed curve. One defines the contact 3-handle map C_{h^3} to be equal to the composition of the 4-dimensional 3-handle map obtained by surgering $(-M, -\gamma)$ along the 2-sphere obtained by pushing S into $\text{int}(M)$, followed by the inverse of the contact 0-handle map corresponding to removing the resulting copy of $(-D^2 \times [-1, 1], -S^1 \times \{0\})$.

On the diagram level, we choose a diagram (Σ, α, β) of (M, γ) such that there are curves $\alpha \in \alpha$ and $\beta \in \beta$ that are parallel to $\gamma_S \subseteq \partial\Sigma$ and intersect transversely in two points x and y , and there are no other α or β curves between α and γ_S or β and γ_S . Let Σ' be the result of gluing a disk to Σ along γ_S . Then $(\Sigma', \alpha \setminus \{\alpha\}, \beta \setminus \{\beta\})$ is a

diagram of (M', γ') . Suppose that x has larger relative grading than y in $(\bar{\Sigma}, \alpha, \beta)$. For $\mathbf{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$, we set $C_{h^3}(\mathbf{x} \times \{x\}) = 0$ and $C_{h^3}(\mathbf{x} \times \{y\}) = \mathbf{x}$.

3.4 Definition of the contact gluing map

We now define the gluing map, in terms of a contact cell decomposition (Definition 3.5). Suppose that (M, γ) and (M', γ') are balanced sutured manifolds, (M, γ) is a sutured submanifold of (M', γ') , and ξ is contact structure on $Z = M' \setminus \text{int}(M)$ that has dividing set $\gamma \sqcup \gamma'$. Let $\mathcal{C}$ be a contact cell decomposition of Z . Let ν and ν' be the chosen contact vector fields, and S and S' the chosen barrier surfaces near ∂M and $\partial M'$, respectively. Recall that we write N and N' for the collar neighborhoods of ∂M and $\partial M'$ bounded by S and S' , respectively. Let $\Gamma \subseteq Z \setminus \text{int}(N \cup N')$ denote the Legendrian 1-skeleton and $D_1, \dots, D_n$ the convex 2-cells.

A neighborhood of each vertex of Γ that is not contained in S or S' determines a contact 0-handle. A neighborhood of each edge of Γ is a contact 1-handle. A neighborhood of each convex 2-cell D_i is a contact 2-handle. Finally, after removing neighborhoods of the graph Γ and the disks D_i , we are left with a collection of tight contact 3-balls that we view as a collection of contact 3-handles. Write $h_1, \dots, h_n$ for these handles, ordered so that their indices are nondecreasing.

Let γ_0 be the dividing set on S with respect to ν and ξ , and write γ'_0 for the dividing set on S' with respect to ν' and ξ . The flow of ν induces a diffeomorphism

$$\psi^\nu: (M, \gamma) \rightarrow (M \cup N, \gamma_0)$$

that is well defined up to isotopy. Note that $(N', \gamma'_0 \cup \gamma')$ is a balanced sutured manifold, and ξ induces the dividing set $\gamma'_0 \cup \gamma'$ on $\partial N'$. There is a canonical isomorphism

$$\text{SFH}(-M \sqcup -N', -\gamma \cup -\gamma'_0 \cup -\gamma') \cong \text{SFH}(-M, -\gamma) \otimes \text{SFH}(-N', -\gamma'_0 \cup -\gamma').$$

Hence, we can define a map

$$\Phi_{\sqcup N'}: \text{SFH}(-M, -\gamma) \rightarrow \text{SFH}(-M \sqcup -N', -\gamma \cup -\gamma'_0 \cup -\gamma')$$

via the formula

$$\Phi_{\sqcup N'}(\mathbf{x}) = \mathbf{x} \otimes \text{EH}(N', \gamma'_0 \cup \gamma', \xi|_{N'}),$$

where $\text{EH}(N', \gamma'_0 \cup \gamma', \xi|_{N'}) \in \text{SFH}(-N', -\gamma'_0 \cup -\gamma')$ is the contact invariant of $\xi|_{N'}$, constructed using a partial open book decomposition, as defined by Honda, Kazez, and Matić [13]. We describe the invariant $\text{EH}(N', \gamma'_0 \cup \gamma', \xi|_{N'})$ in more detail in Section 4,

when we prove some properties of the gluing map. We will show in [Lemma 4.3](#) that $\Phi_{\sqcup N'}$ can be written as a composition of contact handle maps.

We now define the contact gluing map as the composition

$$(3-1) \quad \Phi_{\xi, \mathcal{C}} := C_{h_n} \circ \cdots \circ C_{h_1} \circ (\psi_*^v \otimes \text{id}) \circ \Phi_{\sqcup N'},$$

where C_{h_i} is the map induced by the contact handle h_i , as in [Section 3.3](#).

3.5 Invariance of the contact gluing map

In this section, we prove the following:

Theorem 3.14 *The contact gluing map $\Phi_{\xi, \mathcal{C}}$ is independent of the contact cell decomposition $\mathcal{C}$.*

As a first step, the following lemma is helpful:

Lemma 3.15 *If h_1 and h_2 are two contact handles (of arbitrary index) that are disjoint and are attached to (M, γ) , then*

$$C_{h_1} \circ C_{h_2} = C_{h_2} \circ C_{h_1}.$$

Proof The proof follows by analyzing the formulas for the two maps. If one of h_1 and h_2 is a 0-handle or a 1-handle, then the statement is obvious. Let us consider the case when h_1 and h_2 are disjoint 2-handles. Let us recall how the maps C_{h_i} are defined for $i \in \{1, 2\}$. We first attach a band B_i to the boundary of a sutured diagram for $(-M, -\gamma)$. The band B_i is attached where the attaching circle of h_i intersects $\gamma \subseteq \partial M$. We pick curves α_i and β_i , according to the attaching circle of h_i , that intersect at a single point c_i in B_i . The map C_{h_i} is defined by $C_{h_i}(\mathbf{x}) = \mathbf{x} \times c_i$. Since the handles h_1 and h_2 are disjoint, the bands B_1 and B_2 can be chosen to be disjoint, and the curves α_i and β_i can be assumed to not intersect the band B_j when $i \neq j$. Hence, it follows that we can use the curves α_1 , β_1 , α_2 , and β_2 to compute both compositions $C_{h_2} \circ C_{h_1}$ and $C_{h_1} \circ C_{h_2}$, and since the formulas for the two compositions clearly agree, we conclude that $C_{h_2} \circ C_{h_1} = C_{h_1} \circ C_{h_2}$. In a similar manner, one can show that the same formula holds if at least one of h_1 and h_2 is a 3-handle. $\square$

Proof of Theorem 3.14 Independence of the relative ordering of cells of the same index follows from [Lemma 3.15](#), so it is sufficient to check invariance under the moves in [Proposition 3.8](#).

First consider the case when $\mathcal{C}_1$ and $\mathcal{C}_2$ differ by move (C-1) (isotopy). The maps $\Phi_{\xi, \mathcal{C}_1}$ and $\Phi_{\xi, \mathcal{C}_2}$ differ by postcomposition with a map $\phi_*: \text{SFH}(M', \gamma') \rightarrow \text{SFH}(M', \gamma')$

for a diffeomorphism $\phi: M' \rightarrow M'$ that is isotopic to the identity relative to $\partial M'$. By naturality of sutured Floer homology [18, Theorem 1.9], the map ϕ_* is the identity.

We now consider move (C-2) (index 0/1 cell cancellation). We need to check that subdividing an edge of Γ , or adding a Legendrian edge λ to Γ that meets Γ at a single vertex and intersects none of the other cells does not change the map. Let us first consider subdivision. The contact 0-handle map adds a disk to the Heegaard surface with no new α or β curves, and the contact 1-handle map attaches a band to the boundary of the Heegaard surface with no new α or β curves. When we subdivide a Legendrian edge into two Legendrian edges that meet at a single vertex, the contact handle map changes by first adding a disk to S (the 0-handle map), followed by two bands, each of which has one foot on the new disk, and another foot at a foot of the original band. Clearly, the induced maps agree. Invariance under adding a Legendrian edge λ that intersects Γ at a single vertex follows similarly.

We now consider the case when $\mathcal{C}_1$ and $\mathcal{C}_2$ differ by move (C-3) (index 1/2 contact cell cancellation). The proof is a model computation, which we now describe. The computation is summarized in Figure 20. Note that the contact cell decompositions $\mathcal{C}_1$ and $\mathcal{C}_2$ have the same barrier surfaces S and S' near ∂M and $\partial M'$, respectively. Recall that we write $N, N' \subseteq Z$ for the neighborhoods of ∂M and $\partial M'$ that are bounded by S and S' , and $Z' := Z \setminus \text{int}(N \cup N')$. By definition of an index 1/2 contact cell cancellation, the graph Γ_2 is formed by adding a Legendrian edge λ to Γ_1 , and a new 2-cell D is added that intersects $\partial(Z' \setminus \text{int } N(\Gamma_1))$ along an arc that does not intersect the dividing set.

To demonstrate, let us assume that the new 2-cell D intersects $\partial(Z' \setminus \text{int } N(\Gamma_1))$ along R_+ . If (Σ, α, β) is a sutured diagram for $M \cup N \cup N(\Gamma_1) \cup N'$, the effect of adding the 1-handle h^1 corresponding to λ is to attach a band B to the boundary of $\bar{\Sigma}$. The attaching circle of the 2-handle h^2 corresponding to D intersects the dividing set of $\partial N(\lambda)$ at two points. On the level of diagrams, the effect is to attach a band B' to B , and add new curves α_0 and β_0 , as in Figure 20. Note that the effect of adding B and B' to $\bar{\Sigma}$ is to attach a tube to two points along the interior of $\bar{\Sigma}$ near $\partial \bar{\Sigma}$, and the curve β_0 is a meridian of the tube, while the curve α_0 is the concatenation of a longitude of the tube with a path on $\bar{\Sigma}$ between the two ends of the tube. The curves α_0 and β_0 intersect at a single point c , and the composition of the contact 1-handle and 2-handle maps is

$$(C_{h^2} \circ C_{h^1})(x) = x \times c.$$

This is the compound stabilization map from Section 2.2. By Proposition 2.2, this is equal to the transition map from the naturality of sutured Floer homology.

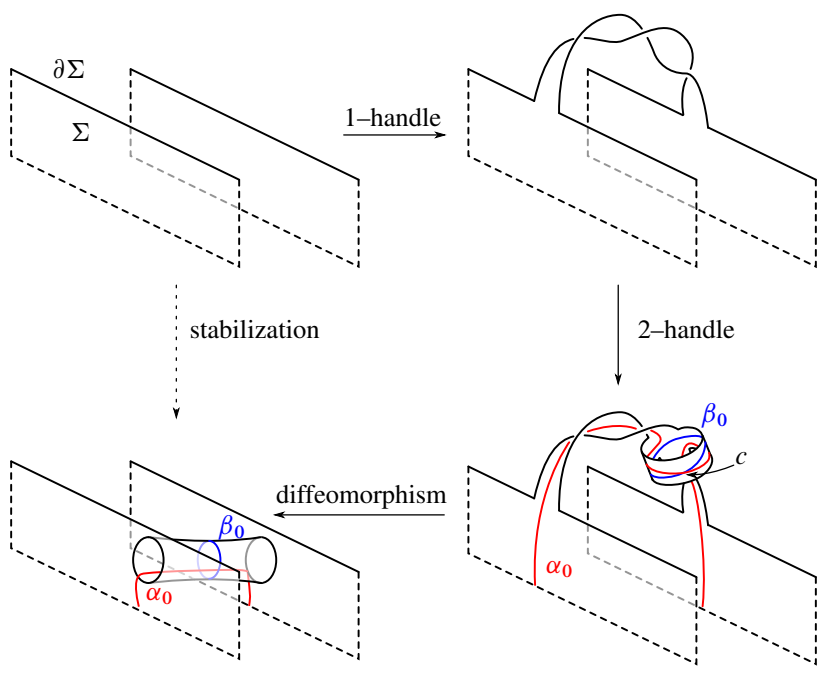


Figure 20: The composition of the contact 1–handle map followed by a contact 2–handle map, for a canceling pair of contact 1– and 2–cells. The composition is the compound stabilization map described in Section 2.2.

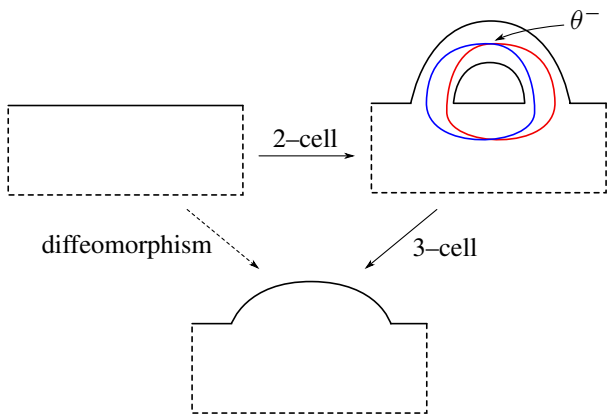


Figure 21: The composition of a contact 2–handle map followed by a contact 3–handle map, for a canceling pair of contact 2– and 3–cells.

Finally, we consider invariance under move (C-4) (index 2/3 contact cell cancellation). In this move, we add a convex 2-cell D to the decomposition that has Legendrian boundary and $\text{tb}(\partial D) = -1$ so that ∂D intersects the dividing set on $\partial(Z' \setminus \text{int } N(\Gamma))$ exactly twice. By definition of a contact cell decomposition, the disk D cuts one of the contact 3-cells in our decomposition into two contact 3-cells. Hence, the effect on the contact gluing map from (3-1) is to insert a contact 2-handle map C_{h^2} , followed by a contact 3-handle map C_{h^3} . The composition is easily seen to be a diffeomorphism map, as demonstrated in Figure 21.

Indeed, for notational simplicity, assume that h^2 is attached to (M, γ) . Let (M', γ') be the result of attaching h^2 to (M, γ) , and (M'', γ'') the result of attaching h^3 to (M', γ') . Given a diagram (Σ, α, β) of (M, γ) , we get a diagram of (M', γ') by attaching a band B to a component of $\partial \Sigma$, and add curves α' to α and β' to β parallel to the suture γ_S along which h^3 is attached and such that $\alpha' \cap \beta' \cap B = \{c\}$. For $x \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$, we have $C_{h^2}(x) = x \times c$. We can choose α' and β' so that $|\alpha' \cap \beta'| = 2$, and we write $\alpha' \cap \beta' = \{\theta^+, \theta^-\}$, where θ^+ and θ^- are distinguished by the relative Maslov grading. By construction, on $(\bar{\Sigma}, \alpha \cup \{\alpha'\}, \beta \cup \{\beta'\})$ there are two bigons connecting the two points of $\alpha' \cap \beta'$, which we can use to determine that c has lower relative grading, so $c = \theta^-$. It follows that $(C_{h^3} \circ C_{h^2})(x) = C_{h^3}(x \times c) = x$. $\square$

Remark 3.16 It is possible to directly show invariance of the contact gluing map under subdividing a contact 2-cell into two contact 2-cells that meet along a Legendrian arc. Indeed, the topological manipulation described in the proof of Lemma 3.9 gives a recipe for doing so. Nonetheless, the model computations required to show invariance of the gluing map under moves (C-1)–(C-4) are simpler.

4 Contact handles and partial open book decompositions

4.1 Partial open book decompositions and sutured Heegaard diagrams

In order to prove basic properties of the gluing map, and to compare our construction to the one due to Honda, Kazez, and Matić [12], we need the following definition:

Definition 4.1 [13, Section 1.2] A *partial open book decomposition* is a triple (S, P, h) consisting of a compact, connected, oriented surface S with nonempty boundary, a compact subsurface $P \subseteq S$ (the *page*) such that S is obtained from $\text{cl}(S \setminus P)$ by successively attaching 1-handles, and a smooth embedding $h: P \rightarrow S$ (the *monodromy*) such that $h|_{\partial S \cap P} = \text{id}$.

An abstract (nonembedded) partial open book decomposition (S, P, h) defines a sutured manifold (M, γ) via the formula

$$M = (S \times [0, \frac{1}{2}] \cup P \times [\frac{1}{2}, 1]) / \sim_h,$$

where $\sim_h$ is the equivalence relation defined as

$$\begin{aligned} (x, t) \sim_h (x, t') & \quad \text{if } x \in \partial S \text{ and } t, t' \in [0, \frac{1}{2}], \\ (x, t) \sim_h (x, t') & \quad \text{if } x \in \partial P \cap \partial S \text{ and } t, t' \in [\frac{1}{2}, 1], \\ (x, 1) \sim_h (h(x), 0) & \quad \text{if } x \in P. \end{aligned}$$

The manifold M contains the properly embedded surface

$$\Sigma := (S \times \{\frac{1}{4}\}) \cup (P \times \{\frac{3}{4}\}).$$

The curves $\gamma = \partial M \cap \Sigma$ divide ∂M into two subsurfaces that meet along γ . Furthermore, (M, γ) is a balanced sutured manifold and Σ is a sutured Heegaard surface for (M, γ) . Using our orientation conventions, $R_+(M)$ deformation retracts onto $(S \setminus P) \times \{\frac{1}{2}\}$ and $R_-(M)$ deformation retracts onto $(S \setminus h(P)) \times \{0\}$.

Definition 4.2 If (S, P, h) is a partial open book decomposition, a *basis of arcs for P in S* is a collection of pairwise disjoint, properly embedded arcs $\mathbf{a} = \{a_1, \dots, a_k\}$ on P with ends on $P \cap \partial S$ such that $P \setminus (a_1 \cup \dots \cup a_k)$ deformation retracts onto $\partial P \setminus \partial S$ (or, equivalently, $S \setminus (a_1 \cup \dots \cup a_k)$ deformation retracts onto $\text{cl}(S \setminus P)$).

Given a basis of arcs for P in S , we can construct attaching curves α and β on Σ that make (Σ, α, β) a Heegaard diagram for (M, γ) . Let b_i be an isotopic copy of a_i obtained by moving the ends of a_i along $\partial \Sigma$ in the positive direction such that $a_i \cap b_j = \delta_{ij}$. We then set

$$\alpha_i = (a_i \times \{\frac{1}{4}\}) \cup (a_i \times \{\frac{3}{4}\}) \quad \text{and} \quad \beta_i = (b_i \times \{\frac{3}{4}\}) \cup (h(b_i) \times \{\frac{1}{4}\}).$$

A partial open book decomposition (S, P, h) determines a unique contact structure ξ on (M, γ) , up to equivalence, as follows; see [2, Proposition 1.2]. Let

$$(4-1) \quad U_1 := (S \times [0, \frac{1}{2}]) / \sim_1 \quad \text{and} \quad U_2 := (P \times [\frac{1}{2}, 1]) / \sim_2,$$

where $\sim_1$ and $\sim_2$ are the relations defined by

$$(x, t) \sim_1 (x, t') \quad \text{if } x \in \partial S, \quad \text{and} \quad (x, t) \sim_2 (x, t') \quad \text{if } x \in \partial P \cap \partial S.$$

Then $\xi|_{U_1}$ is the unique tight contact structure on U_1 with dividing set $\partial S \times \{\frac{1}{4}\}$, and $\xi|_{U_2}$ is the unique tight contact structure on U_2 with dividing set $\partial P \times \{\frac{3}{4}\}$. Hence,

we say that (S, P, h) is a partial open book decomposition of the contact 3-manifold (M, γ, ξ) if we are given a contactomorphism between the contact 3-manifold defined by (S, P, h) and (M, γ, ξ) .

Given a partial open book decomposition (S, P, h) of (M, γ, ξ) and a basis of arcs $\{a_1, \dots, a_k\}$, let $x_i = (a_i \times \{\frac{3}{4}\}) \cap (b_i \times \{\frac{3}{4}\})$ for $i \in \{1, \dots, k\}$ and $\mathbf{x}_\xi = x_1 \times \dots \times x_k$. Then Honda, Kazez, and Matić [13] showed that $\mathbf{x}_\xi$ is a cycle whose homology class $\text{EH}(M, \gamma, \xi) \in \text{SFH}(-M, -\gamma)$ is independent of the choice of partial open book and basis of arcs and is hence an invariant of ξ .

4.2 Partial open books and contact handles

Contact handle decompositions were defined by Giroux [6]. In this section, we describe some useful relations between contact handle decompositions and the Honda–Kazez–Matić definition of a partial open book; compare [2].

Given a partial open book decomposition (S, P, h) for the contact sutured manifold (M, γ, ξ) , we can naturally construct a contact handle decomposition of (M, γ) with no 3-handles, as follows. Recall that the manifold M is obtained by gluing the contact handlebodies U_1 and U_2 together, as described in Section 4.1, along a portion of their boundaries, using the map h .

We start by constructing a contact handle decomposition of U_1 consisting of only 0-handles and 1-handles. Such a description is obtained by giving the surface S a decomposition into 2-dimensional 0-handles and 1-handles. Next, we extend this decomposition to a contact handle decomposition of all of M . To do this, pick a basis of arcs $\mathbf{a}$ for P in S . The closed curves

$$l_i := (a_i \times \{\frac{1}{2}\}) \cup (h(a_i) \times \{0\})$$

bound disks in U_2 that intersect the dividing set $\partial S \times \{0\}$ of ∂U_1 exactly twice. By perturbing U_1 and U_2 slightly inside M , we can assume that the curves l_i are Legendrian. As the curves l_i each intersect the dividing set of ∂U_1 exactly twice, it follows that they bound convex disks with $\text{tb} = -1$, and that neighborhoods of these convex disks are contact 2-handles. Furthermore, after attaching these contact 2-handles, we obtain the sutured manifold M .

In the opposite direction, given a contact handle decomposition $\mathbf{H}$ of (M, γ, ξ) with handles ordered with nondecreasing index and no 3-handles, viewed as a cobordism from $\emptyset$ to ∂M , we can construct a partial open book decomposition (S, P, h) , as follows. The handlebody U_1 is the union of the 0- and 1-handles. If γ_1 is the dividing

set of ξ on ∂U_1 , then (U_1, γ_1) is diffeomorphic to the product sutured manifold $((S \times [0, \frac{1}{2}])/\sim, \partial S \times \{\frac{1}{4}\})$ where $S := R_+(\gamma_1) \subseteq \partial U_1$. Let U_2 be the union of the 2–handles, and let γ_2 be the dividing set of ξ on ∂U_2 . Then (U_2, γ_2) is a product sutured manifold of the form $((P \times [\frac{1}{2}, 1])/\sim, \partial P \times \{\frac{3}{4}\})$ for $P := U_2 \cap S$. We finally set h to be $\pi_1 \circ i_2: P \rightarrow S$, where $\pi_1: S \times [0, \frac{1}{2}] \rightarrow S$ is the projection and

$$i_2: P \rightarrow P \times \{1\} \subseteq R_-(\gamma_1) \approx S \times \{0\}$$

is the canonical embedding.

Lemma 4.3 *Let $h_1, \dots, h_n$ be the handles of a contact handle decomposition $\mathbf{H}$ of (M, γ, ξ) , ordered so that their indices are nonincreasing. If there are no 3–handles, then*

$$\text{EH}(M, \gamma, \xi) = (C_{h_1} \circ \dots \circ C_{h_n})(1) \in \text{SFH}(-M, -\gamma),$$

where $1 \in \mathbb{F}_2 \cong \text{SFH}(\emptyset)$.

Proof Let (S, P, h) be the partial open book corresponding to $\mathbf{H}$, as above. Suppose that $h_1, \dots, h_k$ are the 2–handles. On the level of diagrams, $C_{h_{k+1}} \circ \dots \circ C_{h_n}$ corresponds to adding a disk for each 0–handle, and a band for each 1–handle. The union of these is $S \times \{\frac{1}{4}\}$. Adding a 2–handle h_i for $i \in \{1, \dots, k\}$ corresponds to attaching a band B_i to S , and adding curves α_i and β_i . Then $B_1 \cup \dots \cup B_k = P \times \{\frac{3}{4}\}$, and α_i and β_i are obtained from the basis of arcs $\{a_1, \dots, a_k\}$ as described in Section 4.1, where a_i is the core of B_i . Let $x_i = \alpha_i \cap \beta_i \cap B_i$ for $i \in \{1, \dots, k\}$. Then the element $(C_{h_1} \circ \dots \circ C_{h_n})(1) = x_1 \times \dots \times x_k$ tautologically agrees with the cycle $\mathbf{x}_\xi$ representing $\text{EH}(M, \gamma, \xi)$, as defined by Honda, Kazez, and Matić [13] for the partial open book decomposition (S, P, h) and the basis of arcs $\{a_1, \dots, a_k\}$. $\square$

We now describe the effect of attaching a single contact handle on the level of partial open books. We begin with the effect of attaching a contact 1–handle:

Lemma 4.4 *Suppose that we obtain the contact sutured manifold (M', γ', ξ') from (M, γ, ξ) by attaching a contact 1–handle h^1 . Let (S, P, h) be a partial open book decomposition for the contact structure ξ on (M, γ) . A partial open book decomposition (S', P', h') for the contact structure ξ' on (M', γ') can be obtained by setting*

- (1) $S' = S \cup B$, where B is a band attached to ∂S ,
- (2) $P' = P$, and
- (3) $h' = \iota_S \circ h$, where $\iota_S: S \rightarrow S'$ is the embedding.

Proof First, we isotope the partial open book (S, P, h) in M such that the attaching sphere of h^1 lies in $\partial S \setminus P$. We set $S' = S \cup B$, where the band $B \subseteq h^1$ has boundary equal to the dividing set on h^1 . Then

$$(S' \times [0, \frac{1}{2}] \cup P \times [\frac{1}{2}, 1]) / \sim_{h'} = (S \times [0, \frac{1}{2}] \cup P \times [\frac{1}{2}, 1]) / \sim_{h'} \cup (B \times [0, \frac{1}{2}]) / \sim_{h'}.$$

Furthermore,

$$(S \times [0, \frac{1}{2}] \cup P \times [\frac{1}{2}, 1]) / \sim_{h'} = (S \times [0, \frac{1}{2}] \cup P \times [\frac{1}{2}, 1]) / \sim_h = M,$$

$$\text{and } (B \times [0, \frac{1}{2}]) / \sim_{h'} = h^1. \quad \square$$

We now consider the effect of attaching a contact 2-handle h^2 to (M, γ) . Let ξ' denote the contact structure on $M \cup h^2$ with dividing set γ' , obtained by gluing the tight contact structure ξ_2 on h^2 to ξ . Let $p_1, p_2 \in \gamma$ denote the two points of intersection of the attaching circle of h^2 with γ . Furthermore, the attaching circle of h^2 consists of a path l_+ in $R_+(\gamma)$ from p_1 to p_2 , concatenated with the reverse of a path l_- in $R_-(\gamma)$ from p_1 to p_2 . We can isotope the partial open book decomposition (S, P, h) so that p_1 and p_2 lie in $\partial S \setminus P$. Since we have identifications

$$S \setminus P \cong R_+(M) \quad \text{and} \quad S \setminus h(P) \cong R_-(M),$$

we can view l_+ as a path λ_+ in $S \setminus P$, and l_- as a path λ_- in $S \setminus h(P)$. Using the above notation, we are now prepared to describe a partial open book decomposition for the contact structure ξ' on $(M \cup h^2, \gamma')$.

Lemma 4.5 *Suppose h^2 is a contact 2-handle attached to (M, γ, ξ) , and let ξ' be the resulting contact structure on $(M \cup h^2, \gamma')$. Given a partial open book decomposition (S, P, h) for ξ on (M, γ) , a partial open book decomposition for ξ' on $(M \cup h^2, \gamma')$ is given by (S', P', h') , where*

- (1) $S' = S$,
- (2) $P' = P \cup N(\lambda_+)$,
- (3) $h'|_P = h$, and h' maps $N(\lambda_+)$ to $N(\lambda_-)$.

Proof Using (4-1), we write M as the union of the subsets $U_1 = (S \times [0, \frac{1}{2}]) / \sim_1$ and $U_2 = (P \times [\frac{1}{2}, 1]) / \sim_2$. We note that

$$(P' \times [\frac{1}{2}, 1]) / \sim_2 = (P \times [\frac{1}{2}, 1]) / \sim_2 \cup (N(\lambda_+) \times [\frac{1}{2}, 1]) / \sim_2.$$

Let $U'_2 = (P' \times [\frac{1}{2}, 1])/\sim_2$ and $h^2 = (N(\lambda_+) \times [\frac{1}{2}, 1])/\sim_2$. Then $(U_1 \cup U'_2)/\sim_{h'}$ is obtained from $(U_1 \cup U_2)/\sim_h$ by attaching h^2 along a neighborhood of the circle $(\lambda_+ \times \{\frac{1}{2}\}) \cup (\lambda_- \times \{0\})$. By definition of a partial open book decomposition, it follows that (S', P', h') is a partial open book for ξ' on $(M \cup h^2, \gamma')$. $\square$

4.3 Positive stabilizations and contact handle cancellations

Honda, Kazez, and Matić [13] extended the notion of positive stabilizations to partial open book decompositions, adapting Giroux's construction [8] for open books of closed manifolds. In this section, as an instructive example, we show how to interpret their construction in terms of canceling pairs of contact handles.

We begin with Honda, Kazez, and Matić's definition of a positive stabilization of a partial open book:

Definition 4.6 Suppose that (S, P, h) is a partial open book decomposition of the contact 3-manifold (M, γ, ξ) , and suppose that c is a properly embedded arc on S . The arc c is allowed to intersect P ; however, we require $\partial c \subseteq \partial S \setminus P$. The *positive stabilization* (S', P', h') of (S, P, h) along the arc c is defined as follows. Let $S' := S \cup B$, where B is a band that we attach along $\partial c \subseteq \partial S$. Furthermore, let $P' := P \cup B$ be the new page. Let $\tau \subseteq S'$ be the curve obtained by concatenating the arc c with a core of B . The new partial monodromy map $h': P' \rightarrow S'$ is defined as

$$h' := R_\tau \circ (h \cup \text{id}_B),$$

where R_τ is a right-handed Dehn twist along τ , with respect to the orientation of S .

We now wish to relate positive stabilizations to canceling pairs of contact handles. As described in Section 4.2, the partial open book decomposition (S, P, h) , together with a choice of handle decomposition of the surface S into 0-handles and 1-handles, as well as a basis of arcs $\mathbf{a}$ for P , determines a contact handle decomposition of (M, γ) . We now show that the contact handle decomposition arising from a positive stabilization (S', P', h') of (S, P, h) can be obtained from a contact handle decomposition arising from (S, P, h) by inserting a pair of canceling index 1 and 2 contact handles.

Lemma 4.7 Suppose that (S, P, h) is a partial open book decomposition of (M, γ, ξ) and that (S', P', h') is a positive stabilization of (S, P, h) along a properly embedded arc $c \subseteq S$. Let U_1 and U_2 be the tight contact handlebodies defined in (4-1) associated to (S, P, h) , whose union is M . Consider a contact handle decomposition $\mathbf{H}$

of (M, γ, ξ) arising from the partial open book (S, P, h) as in Section 4.2. Then a handle decomposition $\mathbf{H}'$ arising from the positive stabilization (S', P', h') can be obtained from $\mathbf{H}$ by adding a pair of canceling index 1 and 2 contact handles between U_1 and U_2 .

Proof We attach a pair of canceling contact 1– and 2–handles to U_1 . Let h^1 denote a 1–handle, attached with feet along $\partial c \times \{\frac{1}{4}\}$. We then attach a contact 2–handle h^2 with attaching circle equal to the concatenation of $c \times \{0\} \subseteq \partial U_1$ with a longitude of the 1–handle h^1 ; see the middle row of Figure 22. The contact manifolds U_1 and $U_1 \cup h^1 \cup h^2$ are equivalent, since h^1 and h^2 are canceling contact handles.

Note that $(S, \emptyset, \emptyset)$ is a partial open book decomposition for U_1 . By Lemma 4.4, if B denotes a band attached at $\partial c \subseteq \partial S$, then the surface $(S \cup B, \emptyset, \emptyset)$ is a partial open book decomposition for $U_1 \cup h^1$.

Let $h^2_1, \dots, h^2_n$ denote the contact 2–handles obtained by picking an arc basis of P . To build M from $U_1 \cup h^1$, we first attach the contact 2–handle h^2 , followed by $h^2_1, \dots, h^2_n$.

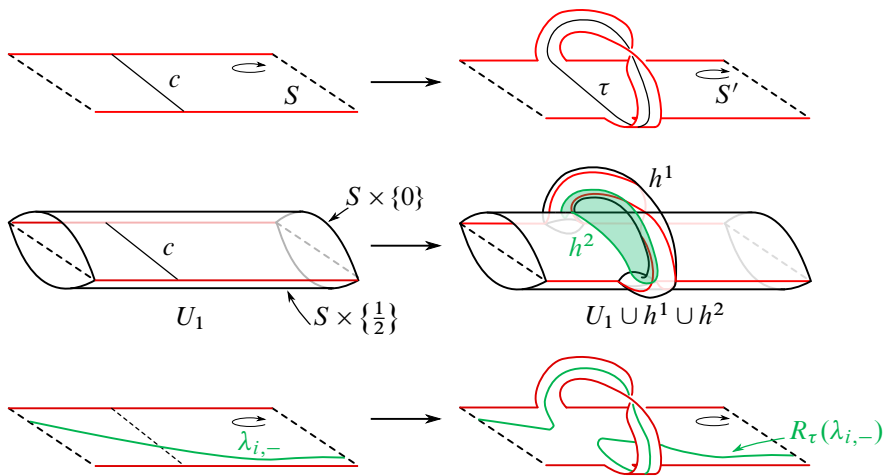


Figure 22: Positively stabilizing a partial open book is the same as inserting a pair of canceling contact 1– and 2–handles. On the top row, we show a schematic of the new partial open book. In the middle row, we show the contact handlebody $U_1 = S \times [0, \frac{1}{2}] / \sim_1$ (left), as well as the equivalent contact handlebody $U_1 \cup h^1 \cup h^2$ (right). On the bottom row, we show that the attaching circle of any contact 2–handle in the previous decomposition is changed by a positive Dehn twist along τ . Note that the Dehn twist looks negative; however, the picture is of $(S \times [0, \frac{1}{2}]) / \sim_1$ turned “upside down”, because we are drawing the images of the contact 2–handles on $R_-(U_1) = S \times \{0\}$.

We can determine the effect of this on the partial open book using [Lemma 4.5](#). The partial monodromy map after attaching h^2 (but before attaching the other 2–handles) is determined by the attaching circle of the 2–handle h^2 . Upon examining [Figure 22](#), the partial monodromy map is the composition of the inclusion of B into S , followed by a right-handed Dehn twist along the curve τ obtained by concatenating c with the core of B .

We now need to compute the effect on the partial open book of attaching $h_1^2, \dots, h_n^2$. Since the contact 2–handles $h_1^2, \dots, h_n^2$ were obtained by picking a basis of arcs of P , it follows that the attaching circle of the 2–handle h_i^2 is equal to the concatenation of properly embedded arcs $\lambda_{+,i} \times \{\frac{1}{2}\} \subseteq S \times \{\frac{1}{2}\}$ and $\lambda_{-,i} \times \{0\} \subseteq S \times \{0\}$, where $\lambda_{+,i}, \lambda_{-,i} \subseteq S$.

The monodromy map h is determined up to isotopy relative to $\partial P \cap \partial S$ by requiring $\lambda_{+,i}$ to be sent to $\lambda_{-,i}$ by h . It is clear that, after attaching h^1 and h^2 , the attaching arcs $\lambda_{-,i}$ must be modified if they intersect the arc c on S , since now they are attached on top of h^1 and h^2 . Indeed, by examining [Figure 22](#), we see that the effect is to replace each $\lambda_{-,i}$ by $R_\tau(\lambda_{-,i})$. It follows that the new partial diffeomorphism map is simply

$$R_\tau \circ (h \cup \text{id}_B),$$

completing the proof. □

5 Properties of the gluing map

5.1 The gluing map for I –invariant contact structures

In this section, we prove that gluing on a copy of $\partial M \times I$ induces the identity map, in a sense that we now describe. Suppose that (M, γ) is a sutured submanifold of (M', γ') and ξ is a contact structure on $M' \setminus \text{int}(M)$. Furthermore, suppose that there is a Morse function f on $M' \setminus \text{int}(M)$ such that $f|_{\partial M} \equiv 0$, $f|_{\partial M'} \equiv 1$, f has no critical points, and there is a contact vector field ν such that the derivative satisfies $\nu(f) > 0$. Furthermore, suppose that the dividing set of ξ on $\partial M \cup \partial M'$, with respect to ν , is equal to $\gamma \cup \gamma'$. The vector field ν induces a diffeomorphism

$$\phi^\nu: (-M, -\gamma) \rightarrow (-M', -\gamma'),$$

which is well defined up to isotopy, relative to ∂M . The induced diffeomorphism map

$$\phi_*^\nu: \text{SFH}(-M, -\gamma) \rightarrow \text{SFH}(-M', -\gamma')$$

has a simple description. If $(\bar{\Sigma}, \alpha, \beta)$ is a sutured Heegaard diagram for $(-M, -\gamma)$, then we can construct a Heegaard diagram for $(-M', -\gamma')$ as

$$(\bar{\Sigma} \cup \bar{A}, \alpha, \beta),$$

where A is the *characteristic surface* of ξ , with respect to ν ; ie

$$A := \{p \in M' \setminus \text{int}(M) : \nu_p \in \xi_p\}.$$

The surface A is a collection of annuli. With respect to these two diagrams, the diffeomorphism map takes the form

$$\phi_*^\nu(x) = x.$$

Our gluing map satisfies the following analogue of [12, Theorem 6.1].

Proposition 5.1 *Suppose that (M, γ) is a sutured submanifold of (M', γ') and ξ is a contact structure on $M' \setminus \text{int}(M)$. Furthermore, suppose that there is a Morse function f on $M' \setminus \text{int}(M)$ such that $f|_{\partial M} \equiv 0$, $f|_{\partial M'} \equiv 1$, and there is contact vector field ν such that $\nu(f) > 0$. Under the above assumptions, the contact gluing map Φ_ξ satisfies*

$$\Phi_\xi = \phi_*^\nu,$$

where $\phi^\nu: (-M, -\gamma) \rightarrow (-M', -\gamma')$ is the diffeomorphism described above.

Before we begin with the proof, we need the following definition regarding sutured cell decompositions of surfaces:

Definition 5.2 The sutured cell decompositions $\mathcal{D} = (B_1, \dots, B_n, \lambda_1, \dots, \lambda_m)$ and $\mathcal{D}^* = (B'_1, \dots, B'_n, \lambda'_1, \dots, \lambda'_m)$ of the surface with divides (F, γ) are *dual* if the following hold:

- (1) $B_i \cap B'_j = \emptyset$ for all i and j .
- (2) Each component of $F \setminus (B_1 \cup \dots \cup B_n \cup \lambda_1 \cup \dots \cup \lambda_m)$ intersects γ in a single arc and contains a single fattened 0-cell B'_k . The same statement holds with the roles of $\mathcal{D}$ and $\mathcal{D}^*$ reversed.
- (3) $|\lambda_i \cap \lambda'_j| = \delta_{ij}$, where δ_{ij} denotes the Kronecker delta.

A sutured cell decomposition $\mathcal{D}$ that admits a dual is called *dualizable*.

As an example, the sutured cell decomposition of a torus with two parallel sutures shown in Figure 8 is dualizable.

Remark 5.3 Not all sutured cell decompositions $\mathcal{D}$ admit dual cell decompositions. For example, if the intersection of a component of $F \setminus (B_1 \cup \cdots \cup B_n \cup \lambda_1 \cup \cdots \cup \lambda_m)$ with γ is disconnected, then $\mathcal{D}$ is not dualizable.

Remark 5.4 Every surface with divides (F, γ) such that $\pi_0(\gamma) \rightarrow \pi_0(F)$ is surjective admits a dualizable cell decomposition. To construct one, we pick sets of arcs A_0 and A_1 along γ such that the arcs in $A_0 \cup A_1$ are pairwise disjoint and A_0 and A_1 each contain exactly one arc in every component of γ . We can then view $R_+(\gamma)$ as a cobordism with product boundary from A_0 to A_1 . By picking a Morse function on $R_+(\gamma)$ which is minimized along A_0 , maximized along A_1 , and which only has index 1 critical points, we get a collection of arcs $\lambda_1, \dots, \lambda_k$ (the stable manifolds of the index 1 critical points) with boundary on A_0 which cut $R_+(\gamma)$ into a collection of disks, each of which contains exactly one component of A_1 . By picking a similar Morse function on $R_-(\gamma)$, we obtain another collection of arcs $\lambda_{k+1}, \dots, \lambda_m$ in $R_-(\gamma)$ with boundary on A_0 that cut $R_-(\gamma)$ into disks, each of which contains exactly one arc of A_1 . We get a sutured cell decomposition with a fattened 0-cell B_i along each arc of A_0 , and we use the collection of arcs $\lambda_1, \dots, \lambda_m$. The 0-cells $B_1, \dots, B_n$ and the arcs $\lambda_1, \dots, \lambda_m$ determine a handle decomposition of F , and the dual sutured cell decomposition is obtained by using the dual handle decomposition obtained by turning the Morse functions upside down.

Proof of Proposition 5.1 Let $\mathcal{D}$ and $\mathcal{D}^*$ be dual sutured cell decompositions of $(\partial M, \gamma)$. Let $\mathcal{C}$ be the product contact cell decomposition of $M' \setminus \text{int}(M)$ constructed from $\mathcal{D}$ as in Example 3.7, with barrier surfaces S and S' . Write N and N' for the collar neighborhoods of ∂M and $\partial M'$ in $M' \setminus \text{int}(M)$ that are bounded by S and S' , respectively. We denote the dividing set on S' by γ'_0 .

For each fattened 0-cell B of $\mathcal{D}$, there is a contact 1-handle h_B of $\mathcal{C}$. For each arc λ of $\mathcal{D}$, there is a contact 2-handle h_λ of $\mathcal{C}$. For each 2-cell c of $\mathcal{D}$, there is a contact 3-handle h_c of $\mathcal{C}$. Let $h_1, \dots, h_n$ be an enumeration of these handles with nondecreasing index. Using (3-1), the gluing map is defined as

$$(5-1) \quad \Phi_\xi(\mathbf{x}) := (C_{h_n} \circ \cdots \circ C_{h_1})(\psi_*^v(\mathbf{x}) \otimes \text{EH}(N', \gamma' \cup \gamma'_0, \xi|_{N'})).$$

The element $\text{EH}(N', \gamma' \cup \gamma'_0, \xi|_{N'})$ is defined using a partial open book decomposition of $(N', \gamma' \cup \gamma'_0, \xi|_{N'})$. In Section 4.2, we described how a contact handle decomposition of N' with no 3-handles, viewed as a cobordism from $\emptyset$ to $\partial N'$, can be used to construct a partial open book. By turning around our construction of a contact cell

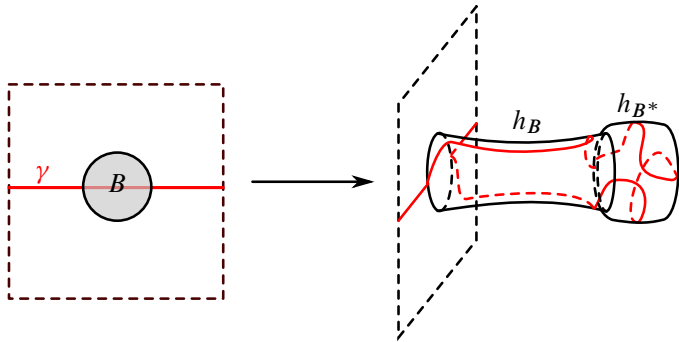


Figure 23: A fattened 0-cell B of a sutured cell decomposition $\mathcal{D}$ induces a pair of canceling index 0 and 1 contact handles. We label the handles as h_B and h_{B^*} . Here B^* denotes the 2-cell of $\mathcal{D}^*$ that is dual to B . A portion of the sutured cell decomposition $\mathcal{D}$ is shown on the left, and the corresponding contact handles are shown on the right.

decomposition $\mathcal{C}$ from $\mathcal{D}$, we can also construct a contact handle decomposition of N' from a sutured cell decomposition of $(\partial M, \gamma)$; however, the indices of the corresponding handles will be different. For our argument to work, we will actually consider the handle decomposition $\mathbf{H}'$ of N' induced by the dual sutured cell decomposition $\mathcal{D}^*$. For each fattened 0-cell B' of $\mathcal{D}^*$, there is a contact 2-handle $h_{B'}$ of $\mathbf{H}'$. For each arc λ' of $\mathcal{D}^*$, there is a contact 1-handle $h_{\lambda'}$ of $\mathbf{H}'$. For each 2-cell c' of $\mathcal{D}^*$, there is a contact 0-handle $h_{c'}$ of $\mathbf{H}'$.

By the above construction, there is a correspondence between the k -handles of $\mathbf{H}'$ and the $(2-k)$ -cells of $\mathcal{D}^*$, which, in turn, correspond to the k -cells of $\mathcal{D}$. Finally, there is a correspondence between the k -cells of $\mathcal{D}$ and the $(k+1)$ -handles of $\mathcal{C}$. Combining these, we get a correspondence between the k -handles of $\mathbf{H}'$ and the $(k+1)$ -handles of $\mathcal{C}$. Let $h'_1, \dots, h'_n$ be an enumeration of the contact handles of $\mathbf{H}'$ such that h'_i corresponds to h_i under the above correspondence. By [Lemma 4.3](#),

$$\mathrm{EH}(N', \gamma' \cup \gamma'_0, \xi) = (C_{h'_n} \circ \dots \circ C_{h'_1})(1),$$

where 1 is the generator of $\mathrm{SFH}(\emptyset) \cong \mathbb{F}_2$. Using [Lemma 3.15](#), we can commute contact handle maps for disjoint contact handles, so we can rearrange (5-1) as

(5-2)
$$\Phi_\xi(x) = ((C_{h_n} \circ C_{h'_n}) \circ \dots \circ (C_{h_1} \circ C_{h'_1}))(\psi_*^v(x)).$$

The contact handles h_i and h'_i do not always form a canceling pair in the sense of [Proposition 3.8](#). However, they are close enough to a canceling pair to allow us to

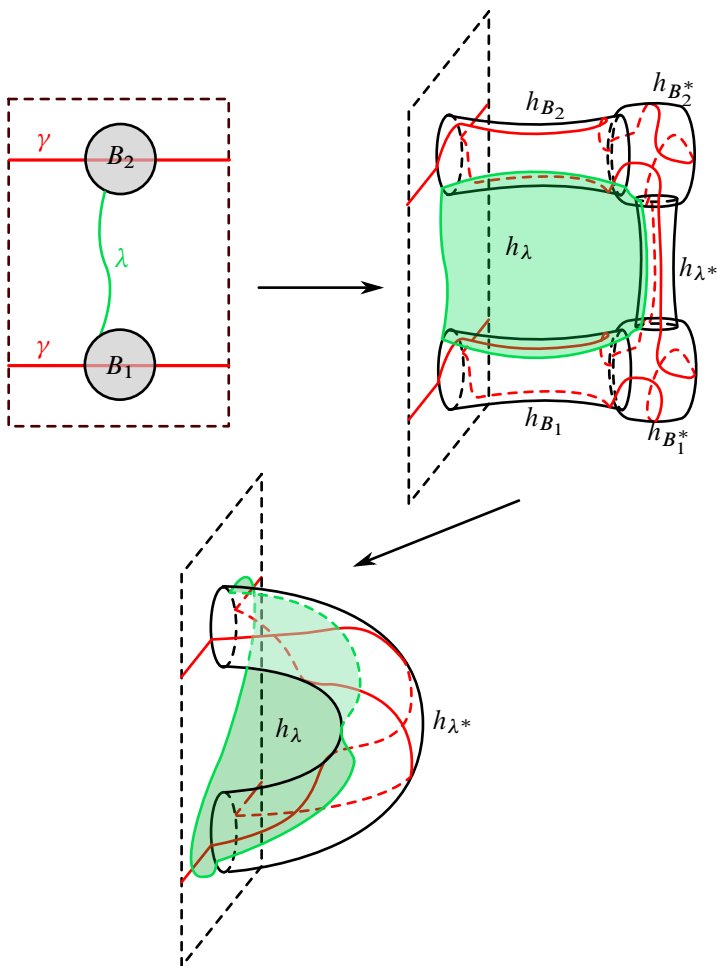


Figure 24: A canceling pair of index 1 and 2 contact handles induced by a 1–cell λ of $\mathcal{D}$, and the dual 1–cell λ^* of $\mathcal{D}^*$. The first picture shows the sutured cell decomposition $\mathcal{D}$ near λ . The second shows the contact handles associated to the cells B_1 , B_2 , B_1^* , B_2^* , λ , and λ^* . The third picture is obtained by canceling h_{B_i} against $h_{B_i^*}$ for $i \in \{1, 2\}$. After this cancellation and a small isotopy of h_λ , the handles h_λ and h_{λ^*} form a canceling pair of contact handles.

reduce the above composition to the diffeomorphism map ψ_*^ν by performing a sequence of handle cancellations and isotopies, as we now describe.

Let us first consider the case when h_i and h'_i correspond to a 0–cell B of $\mathcal{D}$. Under the previously described correspondence, there is a 2–cell B^* of $\mathcal{D}^*$ that B corresponds

to. Using the previous notation, we have $h_i = h_B$ and $h'_i = h_{B^*}$. Then h_B and h_{B^*} form a pair of canceling index 0 and 1 contact handles; see Figure 23. Hence $C_{h_i} \circ C_{h'_i}$ is a diffeomorphism map by the computation in the proof of Theorem 3.14.

We now consider the case when h_i and h'_i correspond to a 1-cell λ of $\mathcal{D}$. Let λ^* denote the corresponding dual arc of $\mathcal{D}^*$. Let B_1 and B_2 be the fattened 0-cells of $\mathcal{D}$ at $\partial\lambda$ (note that we do not exclude the possibility that $B_1 = B_2$). Let h_{B_1} and h_{B_2} denote the corresponding 1-handles of $\mathcal{C}$. The arc λ corresponds to a 2-handle h_λ of $\mathcal{C}$. Let B_1^* and B_2^* denote the 2-cells of $\mathcal{D}^*$ that correspond to B_1 and B_2 , respectively. The 2-cells B_1^* and B_2^* correspond to 0-handles $h_{B_1^*}$ and $h_{B_2^*}$ of $\mathbf{H}'$. The dual arc λ^* of $\mathcal{D}^*$ induces a contact 1-handle of $\mathbf{H}'$. As described above, the handles h_{B_i} and $h_{B_i^*}$ form a canceling pair of index 1 and 2 contact handles. After canceling these two handles, the handles h_λ and h_{λ^*} do not quite form a canceling pair of index 1 and 2 handles in the sense of Proposition 3.8, because the attaching circle of the 2-handle h_λ does not intersect the dividing set along h_{λ^*} . Instead, it intersects the dividing set near the feet of the 1-handle h_{λ^*} . This is shown in Figure 24. After performing an isotopy of h_λ , the handles h_λ and h_{λ^*} form a canceling pair of index 1 and 2 contact handles. Hence, the composition $C_{h_i} \circ C_{h'_i}$ induces a diffeomorphism map, by the computation in the proof Theorem 3.14.

Finally, we consider the case when h_i and h'_i correspond to a 2-cell c of $\mathcal{D}$. Corresponding to c , there is a fattened 0-cell c^* of $\mathcal{D}^*$. The 2-cell c induces a 3-handle $h_c = h_i$ of $\mathcal{C}$, as well as a 2-handle $h_{c^*} = h'_i$ in $\mathbf{H}'$. The handles h_c and h_{c^*} form a canceling pair of index 2 and 3 contact handles. Hence, the composition $C_{h_c} \circ C_{h_{c^*}}$ is equal to a diffeomorphism map, by the same computation as in the proof of Theorem 3.14.

We have shown that $C_{h_i} \circ C_{h'_i}$ is equal to a diffeomorphism map for every $i \in \{1, \dots, n\}$, induced by canceling the handles h_i and h'_i that are stacked horizontally in the v -direction. It follows from (5-2) that Φ_ξ is equal to the diffeomorphism map ϕ_*^v . $\square$

5.2 Morse-type contact handles

In Section 3.3, we defined maps for gluing a contact handle h onto the boundary of a sutured manifold (M, γ) . We note that M is *not* a sutured submanifold of $M \cup h$, so the Honda–Kazez–Matić framework does not assign a gluing map to the inclusion $M \subseteq M \cup h$. Nonetheless, there is a natural notion of contact handle that fits into the Honda–Kazez–Matić TQFT framework:

Definition 5.5 Suppose that (M, γ) is a sutured submanifold of (M', γ') , and ξ is a contact structure on $Z = M' \setminus \text{int}(M)$ with dividing set $\gamma \cup \gamma'$. We say that (Z, ξ) is a *Morse-type contact handle of index k* if there is a contact vector field ν on Z that points into Z on ∂M and out of Z on $\partial M'$ and a decomposition $Z = Z_0 \cup h$ such that

- (1) Z_0 is diffeomorphic to $\partial M \times I$,
- (2) ν is nonvanishing on Z_0 , points into Z_0 on $\partial M \times \{0\}$ and out of Z_0 on $\partial M \times \{1\}$, and each flow line of ν is an arc from $\partial M \times \{0\}$ to $\partial M \times \{1\}$,
- (3) h is a topological 3-ball with piecewise smooth boundary, and ξ is tight on h .

Furthermore, h is a contact k -handle attached to $M \cup Z_0$, as in [Definition 3.11](#), with corners smoothed.

If (M, γ) is a sutured submanifold of (M', γ') , and $(Z, \xi) = (M' \setminus \text{int}(M), \xi)$ is a Morse-type contact handle of index k attached to (M, γ) , then we call a choice of contact vector field ν and decomposition $Z = Z_0 \cup h$ a *parametrization* of (Z, ξ) . Given a parametrization of (Z, ξ) , there is a natural candidate for the contact gluing map Φ_ξ , namely

$$C_h \circ \phi_*^{\nu|_{Z_0}}: \text{SFH}(-M, -\gamma) \rightarrow \text{SFH}(-M', -\gamma').$$

In the above expression,

$$\phi_*^{\nu|_{Z_0}}: \text{SFH}(-M, -\gamma) \rightarrow \text{SFH}(-M \cup -Z_0, -\gamma_0)$$

is the diffeomorphism map induced by the vector field ν , as discussed in [Section 5.1](#), where γ_0 is the dividing set of ξ on $\partial(M \cup Z_0)$. Furthermore,

$$C_h: \text{SFH}(-M \cup -Z_0, -\gamma_0) \rightarrow \text{SFH}(-M', -\gamma')$$

is the contact handle map, as defined in [Section 3.3](#). Indeed, we will prove the following:

Proposition 5.6 Suppose (M, γ) is a sutured submanifold of (M', γ') , and $(Z, \xi) = (M' \setminus \text{int}(M), \xi)$ is a Morse-type contact handle, with a parametrizing contact vector field ν and decomposition $Z = Z_0 \cup h$. Then the contact gluing map Φ_ξ is equal to the composition $C_h \circ \phi_*^{\nu|_{Z_0}}$.

Proof The proof is essentially the same for all handle indices, so for definiteness we will focus on 2-handles. By assumption, the contact vector field ν is nonvanishing on Z_0 , and on a collar neighborhood of $\partial M'$. Let $D \subseteq Z$ be a core of h . Pick

an incoming barrier surface $S \subseteq Z_0$. Extend D down into Z_0 so that $\partial D \subseteq S$ is Legendrian with $\text{tb}(\partial D) = -1$. Then we can perturb D while fixing ∂D so that it becomes convex. Let N denote the collar of ∂M bounded by S , and let

$$\widetilde{Z} := \text{cl}(Z \setminus (N \cup N(D))).$$

We can pick $N(D)$ such that ν is nonvanishing on $\widetilde{Z}$. Using the flow of ν , one can construct a Morse function f on $\widetilde{Z}$ that is 0 on $\partial \widetilde{Z} \setminus \partial M'$ and 1 on $\partial M'$ and such that $\nu(f) > 0$.

The image of $\partial N(D) \setminus S$ in $\partial M'$ under the flow of ν consists of two disks, D_1 and D_2 . Pick a dualizable sutured cell decomposition $\mathcal{D}$ of $(\partial M', \gamma')$ with no 0-cells or 1-cells that intersect D_1 or D_2 . Let $B_1, \dots, B_m$ be the 0-cells of $\mathcal{D}$, and let $\lambda_1, \dots, \lambda_n$ be the 1-cells. Write $c_1, \dots, c_k$ for the 2-cells. Adapting [Example 3.7](#), after performing a C^0 -small isotopy of S , and picking a barrier surface S' that bounds a collar neighborhood N' of $\partial M'$, we can construct a contact cell decomposition $\mathcal{C}$ of Z that has barrier surfaces S and S' . The contact cell decomposition $\mathcal{C}$ has no 0-cells, one 1-cell for each 0-cell $B_1, \dots, B_m$ of $\mathcal{D}$, one 2-cell for each 1-cell $\lambda_1, \dots, \lambda_n$ of $\mathcal{D}$, one 3-cell for each 2-cell $c_1, \dots, c_k$ of $\mathcal{D}$, and also the 2-cell D . Let us write $h_1, \dots, h_n$ for the handles induced by $\mathcal{D}$, and write h_D for $N(D)$, viewed as a contact 2-handle. By definition

(5-3)
$$\Phi_\xi(\mathbf{x}) = (C_{h_n} \circ \dots \circ C_{h_1} \circ C_{h_D})(\phi_*^{v|N}(\mathbf{x}) \otimes \text{EH}(N', \xi|_{N'})).$$

A dual sutured cell decomposition $\mathcal{D}^*$ of $(\partial M', \gamma')$ gives rise to a contact handle decomposition $\mathbf{H}'$ of $(N', \xi|_{N'})$ starting at the empty sutured manifold. We can compute $\text{EH}(N', \xi|_{N'})$ by applying [Lemma 4.3](#) to $\mathbf{H}'$. Exactly as in the proof of [Proposition 5.1](#), the handles of $\mathbf{H}'$ cancel $h_1, \dots, h_n$ pairwise, and we can reduce [\(5-3\)](#) to

$$(\phi_*^{v|\widetilde{Z}} \circ C_{h_D} \circ \phi_*^{v|N})(\mathbf{x}).$$

Given the description of the diffeomorphism maps $\phi_*^{v|\widetilde{Z}}$ and $\phi_*^{v|N}$ from [Section 5.1](#), the above expression is clearly equal to $C_h \circ \phi_*^{v|Z_0}$. □

5.3 Functoriality of the gluing map

We now show that the gluing map defined in this paper satisfies the functoriality property of the Honda–Kazez–Matić construction [\[12, Proposition 6.2\]](#). This property will be useful when we prove the equivalence of our construction with the Honda–Kazez–Matić construction.

Proposition 5.7 *Suppose that we have a chain of sutured submanifolds*

$$(M, \gamma) \subseteq (M', \gamma') \subseteq (M'', \gamma''),$$

as well as a contact structure ξ on $M'' \setminus \text{int}(M)$ such that ∂M , $\partial M'$, and $\partial M''$ are convex with dividing sets γ , γ' , and γ'' , respectively. Writing ξ' for $\xi|_{M' \setminus \text{int}(M)}$ and ξ'' for $\xi|_{M'' \setminus \text{int}(M')}$, we have

$$\Phi_\xi = \Phi_{\xi''} \circ \Phi_{\xi'}.$$

Proof Define $Z' := M' \setminus \text{int}(M)$ and $Z'' := M'' \setminus \text{int}(M')$. Let $\mathcal{C}'$ and $\mathcal{C}''$ be contact cell decompositions of (Z', ξ') and (Z'', ξ'') . Let us write $h'_1, \dots, h'_n$ for the contact handles of $\mathcal{C}'$, and $h''_1, \dots, h''_m$ for the contact handles of $\mathcal{C}''$, ordered so that their indices are nondecreasing. Let v' and v'' denote the contact vector fields chosen on the incoming ends of Z' and Z'' , let N_1 and N_2 denote the incoming layers of Z' and Z'' , and let N'_1 and N'_2 denote the outgoing layers, respectively, as described in Definition 3.5. By definition, the composition $\Phi_{\xi''} \circ \Phi_{\xi'}$ is equal to

$$(5-4) \quad (C_{h''_m} \circ \dots \circ C_{h''_1}) \circ \Phi_{\sqcup N'_2} \circ \phi_*^{v''|_{N_2}} \circ (C_{h'_n} \circ \dots \circ C_{h'_1}) \circ \Phi_{\sqcup N'_1} \circ \phi_*^{v'|_{N_1}}.$$

The map $\Phi_{\sqcup N'_i}$ is given by tensoring with $\text{EH}(N'_i, \xi|_{N'_i})$ for $i \in \{1, 2\}$. As in Lemma 4.3, the element $\Phi_{\sqcup N'_1}$ can be written as a composition of contact handle maps $C_{h_1} \circ \dots \circ C_{h_\ell}$, for a sequence of contact 0-, 1-, and 2-handles $h_1, \dots, h_\ell$. Hence, the composition in (5-4) can be written as

$$(5-5) \quad (C_{h''_m} \circ \dots \circ C_{h''_1}) \circ \Phi_{\sqcup N'_2} \circ \phi_*^{v''|_{N_2}} \circ (C_{h'_n} \circ \dots \circ C_{h'_1}) \circ (C_{h_\ell} \circ \dots \circ C_{h_1}) \circ \phi_*^{v'|_{N_1}}.$$

By picking v'' and N_2 appropriately, we can assume that the diffeomorphism

$$\phi^{v''|_{N_2}}: M' \rightarrow M' \cup N_2$$

is a contactomorphism on all of Z' and is the identity on ∂M . Write $\bar{h}_k = \phi^{v''|_{N_2}}(h_k)$ and $\bar{h}'_k = \phi^{v''|_{N_2}}(h'_k)$. Using the diffeomorphism invariance of the contact handle maps, we can rewrite (5-5) as

$$(5-6) \quad (C_{h''_m} \circ \dots \circ C_{h''_1}) \circ \Phi_{\sqcup N'_2} \circ (C_{\bar{h}'_n} \circ \dots \circ C_{\bar{h}'_1}) \circ (C_{\bar{h}_\ell} \circ \dots \circ C_{\bar{h}_1}) \circ \phi_*^{v'|_{N_1}}.$$

The map $\Phi_{\sqcup N'_2}$ can be commuted with all the contact handle maps to the right of it by Lemma 3.15. After possibly isotoping some of the remaining contact handles, we can apply Lemma 3.15 and reorder the handles so that they are attached with nondecreasing index. Furthermore, after isotoping some of the handles, we can assume that the handles in the above composition are induced by a contact cell decomposition (ie the 0-handles

and 1–handles are induced by a Legendrian graph, and the 2–handles are induced by a sequence of convex disks with $\text{tb} = -1$ attached to a neighborhood of the graph and $\partial((Z' \cup Z'') \setminus \text{int}(N_1 \cup N_2'))$. It follows that (5-6) is equal to $\Phi_{\xi, \mathcal{C}}$ for some contact cell decomposition $\mathcal{C}$ of (Z, ξ) , completing the proof. $\square$

5.4 Equivalence with the Honda–Kazez–Matić construction

In this section, we prove that our construction of the gluing map from Section 3.4 is equivalent to the original construction due to Honda, Kazez, and Matić [12]. We will write Φ_{ξ}^{HKM} for the map defined using their construction.

Theorem 5.8 *Suppose (M, γ) is a sutured submanifold of (M', γ') with no isolated component and that ξ is a contact structure on $M' \setminus \text{int}(M)$ with convex boundary and dividing set $\gamma \cup \gamma'$. Then the Honda–Kazez–Matić gluing map Φ_{ξ}^{HKM} is equal to the gluing map Φ_{ξ} we defined in Section 3.4.*

Proof Using the composition law for both constructions of the gluing map (Proposition 5.7 and [12, Proposition 6.2]), it is sufficient to show the claim when $M' \setminus \text{int}(M)$ consists of a single Morse-type contact handle of index 0, 1, or 2.

For a Morse-type index 0 handle, the claim is straightforward. Write $M' \setminus \text{int}(M)$ as $Z_0 \cup h^0$ where $Z_0 \cong (I \times \partial M)$ and h^0 is a 3–ball. The contact structure ξ is the union of the I –invariant contact structure on Z_0 and the unique tight contact structure on h^0 . Write γ'_0 for the suture on $M \cup Z_0$, and suppose that ν is a parametrizing contact vector field on Z_0 .

Under the identification

$$\text{SFH}(M \cup Z_0 \cup h^0, \gamma'_0 \cup \gamma_0) \cong \text{SFH}(M \cup Z_0, \gamma'_0) \otimes \text{SFH}(h^0, \gamma_0),$$

where γ_0 consists of a single suture on h^0 , both gluing maps

$$\Phi_{\xi}, \Phi_{\xi}^{\text{HKM}}: \text{SFH}(M, \gamma) \rightarrow \text{SFH}(M \cup Z_0, \gamma'_0) \otimes \text{SFH}(h^0, \gamma_0)$$

take the form

$$x \mapsto \phi_*^{\nu}(x) \otimes \text{EH}(h^0, \gamma_0, \xi_0) = \phi_*^{\nu}(x) \otimes 1,$$

where 1 denotes the generator of $\text{SFH}(h^0, \gamma_0) \cong \mathbb{F}_2$. For Φ_{ξ}^{HKM} , this follows from [12, Proposition 6.1], and for our map Φ_{ξ} , from Proposition 5.6.

Before we consider index 1 and 2 Morse-type contact handles, we must first give a more detailed description of the construction of Honda, Kazez, and Matić [12]. The

definition of the map Φ_{ξ}^{HKM} uses the description of partial open books from [13]. A contact 3-manifold with convex boundary is called *product disk decomposable* if it is a union of handlebodies, and it contains a collection of pairwise disjoint compressing disks, each intersecting the dividing set in two points, whose complement is a union of standard contact balls. Given a contact sutured manifold (M, γ, ξ) , one picks a properly embedded Legendrian graph $K \subseteq M$ that intersects ∂M along a collection of univalent vertices in γ such that $M \setminus \text{int}(N(K))$ is product disk decomposable. Here $N(K)$ denotes a standard contact neighborhood, which is also product disk decomposable. It follows that $M \setminus \text{int}(N(K))$ is contactomorphic to $(S \times I)/\sim_1$ for a compact surface S with boundary, and $N(K)$ is contactomorphic to $(P \times I)/\sim_2$ for a compact surface with boundary P (for the definition of $\sim_1$ and $\sim_2$, see Section 4.1). Here P has piecewise smooth boundary whose edges can naturally be divided into two types: those that intersect $\partial N(K) \setminus \partial M$, and those that intersect ∂M . By $(P \times I)/\sim_2$, we mean the space obtained by quotienting out the I direction along the edges that are contained in $\partial N(K)$. Since $(P \times I)/\sim_2$ meets $(S \times I)/\sim_1$ along $\partial N(K) \setminus \partial M$, the surfaces $P \times \{0\}$ and $P \times \{1\}$ give two embeddings of P into S . Using the projection of $S \times I$ onto S , we identify $P \times \{0\} \subseteq S \times \{1\}$ with a subset of S , for which we also write P . The surface $P \times \{1\}$ then gives another smooth embedding $h: P \rightarrow S$, which is the monodromy map.

The Honda–Kazez–Matić map is easiest to define if one picks a contact structure ζ on (M, γ) such that ∂M is convex with dividing set γ . To define the map Φ_{ξ}^{HKM} , one picks Legendrian graphs $K \subseteq M$ and $K' \subseteq M' \setminus \text{int}(M)$ whose complements are product disk decomposable. After modifying K' in a neighborhood of ∂M , one extends K to a Legendrian graph on all of M' whose complement is product disk decomposable. Furthermore, outside a small neighborhood of ∂M , the extension agrees with K and K' . Let us write $\bar{K}$ for this Legendrian graph. The graph K is also required to satisfy a *contact compatibility* condition near ∂M , described in [12], though the specific form is not important for our present argument. The graph $K \subseteq M$ induces a partial open book (S, P, h) for (M, γ, ζ) , which induces a diagram (Σ, α, β) for (M, γ) . The graph $\bar{K}$ induces a partial open book (S', P', h') for (M', γ') , which gives rise to the diagram $(\Sigma', \alpha', \beta')$. Furthermore $\Sigma' = \Sigma \cup \Sigma''$, $\alpha' = \alpha \cup \alpha''$, and $\beta' = \beta \cup \beta''$, for a surface Σ'' and a collection of curves α'' and β'' on Σ' . There is a canonical intersection point $x_{\xi} \in \mathbb{T}_{\alpha''} \cap \mathbb{T}_{\beta''}$, and the map Φ_{ξ}^{HKM} on the chain level is defined by the formula

$$\Phi_{\xi}^{\text{HKM}}(x) = x \times x_{\xi}.$$

We now consider a Morse-type contact 1–handle addition. In this case,

$$(M' \setminus \text{int}(M), \xi) \cong (Z_0 \cup h_1, \xi_0 \cup \xi_1),$$

where (Z_0, ξ_0) is an I –invariant contact structure on $I \times \partial M$, and (h^1, ξ_1) is a contact 1–handle. Let $\bar{K} \subseteq M \cup Z_0$ denote an extension that can be used to compute the gluing map. We note that if $K' \subseteq Z_0$ is a graph whose complement is product disk decomposable, then the complement of K' in $Z_0 \cup h^1$ is also product disk decomposable. It follows that we can use the same graph

$$\bar{K} \subseteq M \cup Z_0 \subseteq M \cup Z_0 \cup h^1$$

to compute the gluing map for $Z_0 \cup h^1$. Write (S, P, h) for the partial open book of (M, γ) induced by K . Write (S'_0, P'_0, h'_0) for the partial open book of $(M \cup Z_0, \gamma'_0)$ induced by $\bar{K} \subseteq M \cup Z_0$ and (S', P', h') for the partial open book of $(M \cup Z_0 \cup h^1, \gamma')$ induced by $\bar{K}$.

Because $(M \cup Z_0 \cup h^1) \setminus N(\bar{K})$ is obtained by attaching a contact 1–handle to $(M \cup Z_0) \setminus N(\bar{K})$, we can apply [Lemma 4.4](#) to see that the partial open book (S', P', h') is obtained from (S'_0, P'_0, h'_0) by attaching a 1–handle to S'_0 along $\partial S'_0 \setminus \partial P'_0$, and setting $P' = P'_0$ and $h' = \iota_{S'_0} \circ h'_0$, where $\iota_{S'_0}: S'_0 \rightarrow S'$ is the embedding. The same basis of arcs $\mathbf{a}'$ for P'_0 in S'_0 can be used for P' in S' , which we assume extends a basis of arcs $\mathbf{a}$ for P in S . Let (Σ, α, β) , $(\Sigma'_0, \alpha'_0, \beta'_0)$, and $(\Sigma', \alpha', \beta')$ be the diagrams induced by (S, P, h) , (S'_0, P'_0, h'_0) , and (S', P', h') with the bases $\mathbf{a}$, $\mathbf{a}'$, and $\mathbf{a}'$, respectively.

The Heegaard surface Σ' is thus obtained by attaching a 1–handle along the boundary of Σ'_0 . Notice that, since $\alpha'_0 = \alpha'$ and $\beta'_0 = \beta'$, and Σ' is obtained from Σ'_0 by attaching a band along $\partial \Sigma'_0$, the groups $\text{SFH}(\Sigma'_0, \alpha'_0, \beta'_0)$ and $\text{SFH}(\Sigma', \alpha', \beta')$ are naturally isomorphic, and the isomorphism is given by the contact 1–handle map defined in this paper. Furthermore, we observe that

$$(5-7) \quad \Phi_{\xi_0 \cup \xi_1}^{\text{HKM}} = C_{h^1} \circ \Phi_{\xi_0}^{\text{HKM}}.$$

By [\[12, Proposition 6.1\]](#), the above expression is equal to $C_{h^1} \circ \phi_*^v$, where

$$\phi_*^v: \text{SFH}(\Sigma, \alpha, \beta) \rightarrow \text{SFH}(\Sigma'_0, \alpha'_0, \beta'_0)$$

is the composition of the tautological map induced by a diffeomorphism and the transition maps induced by naturality. By [Proposition 5.6](#), we see that the expression in (5-7) agrees with the definition of $\Phi_{\xi_0 \cup \xi_1}$ in this paper.

The argument when

$$(M' \setminus \operatorname{int}(M), \xi) = (Z_0 \cup h^2, \xi_0 \cup \xi_2)$$

is a Morse-type contact 2–handle is similar. Suppose that $K \subseteq M$ is a Legendrian graph such that $M \setminus \operatorname{int}(N(K))$ is product disk decomposable and induces a partial open book that satisfies the contact compatibility condition near ∂M . We then let $\bar{K}_0$ denote a Legendrian extension to $M \cup Z_0$ whose complement is product disk decomposable and which can be used to compute the map $\Phi_{\xi_0}^{\text{HKM}}$ for gluing (Z_0, ξ_0) to M . We can define an extension $\bar{K}$ of K into all of $M \cup Z_0 \cup h^2$ by setting

$$\bar{K} := \bar{K}_0 \cup c,$$

where c is a Legendrian cocore of the 2–handle h^2 . We note that $(M \cup Z_0 \cup h^2) \setminus \operatorname{int}(N(\bar{K}))$ is product disk decomposable. Let (S, P, h) , (S'_0, P'_0, h'_0) , and (S', P', h') denote the partial open books induced by $K \subseteq (M, \gamma)$, $\bar{K}_0 \subseteq (M \cup Z_0, \gamma'_0)$, and $\bar{K} \subseteq (M \cup Z_0 \cup h^2, \gamma')$, respectively. Let $\pi: S \times I \rightarrow S$ be the projection. Writing the 2–handle h^2 as $(B \times I)/\sim_2$, where B is a square, we observe that P' is obtained by adding the 1–handle $\pi(B \times \{0\})$ to P'_0 . The monodromy is extended to P' by mapping $\pi(B \times \{0\}) \subseteq S$ to $\pi(B \times \{1\}) \subseteq S$.

We start with a basis of arcs $\mathbf{a}$ for $P \subseteq S$, and extend $\mathbf{a}$ to a basis $\mathbf{a}'_0$ for $P'_0 \subseteq S'_0$. A basis of arcs $\mathbf{a}'$ for $P' \subseteq S'$ can then be obtained from $\mathbf{a}'_0$ by adding a new arc $\mathbf{a}'$ which is a cocore of the band $\pi(B \times \{0\})$. Write $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$, $(\Sigma'_0, \boldsymbol{\alpha}'_0, \boldsymbol{\beta}'_0)$, and $(\Sigma', \boldsymbol{\alpha}', \boldsymbol{\beta}')$ for the diagrams obtained from the partial open books (S, P, h) , (S'_0, P'_0, h'_0) , and (S', P', h') , with bases $\mathbf{a}$, $\mathbf{a}'_0$, and $\mathbf{a}'$, respectively. Also, let us write $\boldsymbol{\alpha}'_0 = \boldsymbol{\alpha} \cup \boldsymbol{\alpha}''_0$ and $\boldsymbol{\beta}'_0 = \boldsymbol{\beta} \cup \boldsymbol{\beta}''_0$, where $\boldsymbol{\alpha}''_0$ and $\boldsymbol{\beta}''_0$ are the curves induced by the arcs in $\mathbf{a}'_0 \setminus \mathbf{a}$, and $\mathbf{x}_{\xi_0}$ for the canonical intersection point in $\mathbb{T}_{\boldsymbol{\alpha}''_0} \cap \mathbb{T}_{\boldsymbol{\beta}''_0}$. Finally, write $\boldsymbol{\alpha}'$ and $\boldsymbol{\beta}'$ for the curves induced by the new basis arc $\mathbf{a}'$, and write $\mathbf{c}'$ for the canonical intersection point of $\boldsymbol{\alpha}' \cap \boldsymbol{\beta}'$. The map $\Phi_{\xi_0 \cup \xi_2}^{\text{HKM}}$ on $\mathbf{x} \in \mathbb{T}_{\boldsymbol{\alpha}} \cap \mathbb{T}_{\boldsymbol{\beta}}$ is defined by the formula

$$\Phi_{\xi_0 \cup \xi_2}^{\text{HKM}}(\mathbf{x}) = \mathbf{x} \times \mathbf{x}_{\xi_0} \times \mathbf{c}'.$$

Noting that $\Phi_{\xi_0}^{\text{HKM}}(\mathbf{x}) = \mathbf{x} \times \mathbf{x}_{\xi_0}$, we see that

$$\Phi_{\xi_0 \cup \xi_2}^{\text{HKM}}(\mathbf{x}) = (C_{h^2} \circ \Phi_{\xi_0}^{\text{HKM}})(\mathbf{x}).$$

By the same argument as for contact 1–handles, this is equal to $(C_{h^2} \circ \phi_*^{\mathcal{V}})(\mathbf{x}) = \Phi_{\xi_0 \cup \xi_2}(\mathbf{x})$, completing the proof. □

6 Turning around cobordisms of sutured manifolds and duality

In this section, we compute the effect of turning around a cobordism of sutured manifolds, proving [Theorem 1.3](#).

6.1 The canonical trace pairing

As described in [\[4, Proposition 2.14\]](#) and [\[15, Section 11.2\]](#), there is duality between $\mathrm{SFH}(M, \gamma)$ and $\mathrm{SFH}(-M, \gamma)$. If (Σ, α, β) is a diagram for (M, γ) , then (Σ, β, α) is a diagram for $(-M, \gamma)$. Since $\mathbb{T}_\alpha \cap \mathbb{T}_\beta$ is equal to $\mathbb{T}_\beta \cap \mathbb{T}_\alpha$, we can define a map

$$(6-1) \quad \mathrm{tr}: \mathrm{CF}(\Sigma, \alpha, \beta) \otimes \mathrm{CF}(\Sigma, \beta, \alpha) \rightarrow \mathbb{F}_2$$

by the formula

$$\mathrm{tr}(\mathbf{x} \otimes \mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{x} = \mathbf{y}, \\ 0 & \text{otherwise.} \end{cases}$$

It is straightforward to see that tr is a chain map, because J -holomorphic discs on (Σ, α, β) from $\mathbf{x}$ to $\mathbf{y}$ are in bijection with J -holomorphic discs on (Σ, β, α) from $\mathbf{y}$ to $\mathbf{x}$. Note that the trace pairing gives a natural isomorphism

$$\mathrm{CF}(\Sigma, \beta, \alpha) \cong \mathrm{CF}(\Sigma, \alpha, \beta)^\vee := \mathrm{Hom}_{\mathbb{F}_2}(\mathrm{CF}(\Sigma, \alpha, \beta), \mathbb{F}_2).$$

In particular, tr is the usual pairing between homology and cohomology.

In the opposite direction, there is the cotrace map,

$$\mathrm{cotr}: \mathbb{F}_2 \rightarrow \mathrm{CF}(\Sigma, \alpha, \beta) \otimes \mathrm{CF}(\Sigma, \beta, \alpha).$$

The cotrace map is defined by the formula

$$\mathrm{cotr}(1) = \sum_{\mathbf{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta} \mathbf{x} \otimes \mathbf{x}.$$

We note that if V is a finite-dimensional vector space over $\mathbb{F}_2$, then there are canonical isomorphisms

$$\mathrm{Hom}_{\mathbb{F}_2}(V, V) \cong \mathrm{Hom}_{\mathbb{F}_2}(V \otimes V^\vee, \mathbb{F}_2) \cong \mathrm{Hom}_{\mathbb{F}_2}(\mathbb{F}_2, V^\vee \otimes V).$$

Under these isomorphisms, the tr and cotr maps are identified with $\mathrm{id}_V \in \mathrm{Hom}_{\mathbb{F}_2}(V, V)$.

In [\[15\]](#), the first author defined a pairing

$$\langle \cdot, \cdot \rangle: \mathrm{SFH}(M, \gamma) \otimes \mathrm{SFH}(-M, -\gamma) \rightarrow \mathbb{F}_2,$$

which at first glance appears to have a different domain than the trace map defined in (6-1). However, there is a canonical isomorphism

$$(6-2) \quad \text{SFH}(-M, \gamma) \cong \text{SFH}(-M, -\gamma),$$

which we describe presently. The diagram (Σ, β, α) represents $(-M, \gamma)$, while $(-\Sigma, \alpha, \beta)$ represents $(-M, -\gamma)$. If $\phi \in \pi_2(x, y)$ is a homology class of disks for (Σ, β, α) , then there is a uniquely determined homology class $\bar{\phi} \in \pi_2(x, y)$ for $(-\Sigma, \alpha, \beta)$. Furthermore, by precomposing a holomorphic disk $u: \mathbb{D} \rightarrow \text{Sym}^n(\Sigma)$ with the unique antiholomorphic involution of $\mathbb{D}$ that fixes $\pm i \in \partial\mathbb{D}$, we obtain a bijection

$$\mathcal{M}_J(\phi)/\mathbb{R} \cong \mathcal{M}_{-J}(\bar{\phi})/\mathbb{R},$$

establishing the isomorphism in (6-2). We note that tr agrees with $\langle \cdot, \cdot \rangle$ under the isomorphism in (6-2).

6.2 Sutured manifold cobordisms and the induced maps

In this section, we review the definition of sutured manifold cobordisms, special cobordisms, boundary cobordisms, and the construction of the sutured cobordism maps. We finally give a simpler definition of the cobordism maps using our gluing map from Section 3.4. The following is [15, Definition 2.3].

Definition 6.1 The contact structures ξ_0 and ξ_1 on the sutured manifold (M, γ) are *equivalent* if they can be connected by a 1-parameter family $\{\xi_t : t \in I\}$ of contact structures on (M, γ) such that ∂M is convex with dividing set γ for each ξ_t . In this case, we write $\xi_1 \sim \xi_2$, and denote the equivalence class of ξ by $[\xi]$.

Sutured manifold cobordisms were defined in [15, Definition 2.4].

Definition 6.2 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 $\mathcal{W} = (W, Z, [\xi])$ such that

- W is a compact, oriented 4-manifold with boundary and corners,
- Z is a codimension-0 submanifold of ∂W , and $\partial W \setminus \text{int}(Z) = -M_0 \sqcup M_1$,
- ξ is a positive contact structure on $(Z, \gamma_0 \cup \gamma_1)$.

Note that equivalent contact structures can have different characteristic foliations on ∂M , which gives us enough flexibility to compose cobordisms. The sutured manifold

cobordism $\mathcal{W}$ is *balanced* if (M_0, γ_0) and (M_1, γ_1) are balanced sutured manifolds. Furthermore, we say that Z_0 is an *isolated component* of Z if $Z_0 \cap M_1 = \emptyset$. The following is [15, Definition 5.1]:

Definition 6.3 The cobordism $\mathcal{W} = (W, Z, [\xi])$ from (M_0, γ_0) to (M_1, γ_1) is *special* if

- $\mathcal{W}$ is balanced,
- $\partial M_0 = \partial M_1$, and $Z = -I \times \partial M_0$ is the trivial cobordism between them,
- ξ is an I -invariant contact structure on Z such that $\{t\} \times \partial M_0$ is convex with dividing set $\{t\} \times \gamma_0$ for every $t \in I$, with respect to the contact vector field $\partial/\partial t$.

Given a special cobordism $\mathcal{W}$ from (M_0, γ_0) to (M_1, γ_1) , we define the map

$$F_{\mathcal{W}}: \text{SFH}(M_0, \gamma_0) \rightarrow \text{SFH}(M_1, \gamma_1)$$

by composing maps associated to 4-dimensional handle attachments along the interior of M_0 ; see [15, Section 8]. The following is equivalent to [15, Definition 10.4].

Definition 6.4 A sutured cobordism $(W, Z, [\xi])$ from (M, γ) to (M', γ') is called a *boundary cobordism* if $M \subseteq \text{Int}(M')$, $W = I \times M'/\sim$, where $(t, x) \sim (t', x)$ for every $x \in \partial M'$ and $t \in I$, and ξ is a contact structure on $Z = -\{0\} \times (M' \setminus \text{int}(M))$ inducing the sutures $\{0\} \times \gamma$ on $\{0\} \times \partial M$ and $\{0\} \times \gamma'$ on $\{0\} \times \partial M'$.

If $\mathcal{W} = (W, Z, [\xi])$ is a boundary cobordism from (M, γ) to (M', γ') , then we can view $(-M, -\gamma)$ as a sutured submanifold of $(-M', -\gamma')$, and $-\xi$ is a positive contact structure on $Z = -M' \setminus \text{int}(-M)$ with dividing set $-\gamma_0 \cup -\gamma_1$. If Z has no isolated components, then the map $F_{\mathcal{W}}: \text{SFH}(M, \gamma) \rightarrow \text{SFH}(M', \gamma')$ is defined as the Honda–Kazez–Matić gluing map $\Phi_{-\xi}$.

Every balanced cobordism $\mathcal{W} = (W, Z, [\xi])$ from (M_0, γ_0) to (M_1, γ_1) can uniquely be written as a composition $\mathcal{W}^s \circ \mathcal{W}^b$, where

$$\mathcal{W}^b = (I \times (M_0 \cup -Z)/\sim, \{0\} \times Z, [\xi])$$

is a boundary cobordism from (M_0, γ_0) to $(M_0 \cup -Z, \gamma_1)$. Furthermore,

$$\mathcal{W}^s = (W, -I \times \partial M_1, [\eta])$$

is a special cobordism from $(M_0 \cup -Z, \gamma_1)$ to (M_1, γ_1) , where $-I \times \partial M_1$ is a collar of ∂M_1 in Z , and η is an I -invariant contact structure with dividing set $\{t\} \times \gamma_1$

on $\{t\} \times \partial M_1$ for every $t \in I$, with respect to $\partial/\partial t$. We call $\mathcal{W}^s$ the special part, and $\mathcal{W}^b$ the boundary part, of $\mathcal{W}$. If Z has no isolated components, then the cobordism map $F_{\mathcal{W}}$ is defined as $F_{\mathcal{W}^s} \circ F_{\mathcal{W}^b}$.

According to [15, Definition 10.1], in the general case, we choose a standard contact ball $B_0 \subseteq \text{int}(Z_0)$ with convex boundary and dividing set δ_0 in each isolated component Z_0 of Z . We write (B, δ) for the union of the balls (B_0, δ_0) , and consider the cobordism $\mathcal{W}' = (W, Z', [\xi'])$ from (M_0, γ_0) to $(M_1, \gamma_1) \sqcup (B, \delta)$, where $Z' = Z \setminus \text{int}(B)$ and $\xi' = \xi|_{Z'}$. Since Z' has no isolated components and

$$\text{SFH}((M_1, \gamma_1) \sqcup (B, \delta)) \cong \text{SFH}(M_1, \gamma_1),$$

we can define $F_{\mathcal{W}} := F_{\mathcal{W}'}$. This is independent of the choice of B .

The gluing map that we defined in Section 3.4 also assigns maps to contact 3–handles, and hence $\Phi_{-\xi}$ makes sense even if Z has isolated components whenever $M_0 \cup -Z$ has no closed components, giving rise to an alternative definition of $F_{\mathcal{W}^b}$. We now show that the two constructions agree.

Proposition 6.5 *Let $\mathcal{W} = (W, Z, \xi)$ be a sutured manifold cobordism from (M_0, γ_0) to (M_1, γ_1) , possibly with Z having isolated components, but such that $M_0 \cup -Z$ has no closed components. Then*

$$F_{\mathcal{W}} = F_{\mathcal{W}^s} \circ \Phi_{-\xi},$$

where $F_{\mathcal{W}}$ is the cobordism map defined in [15, Definition 10.1] using the Honda–Kazez–Matić gluing map, and $\Phi_{-\xi}$ is the gluing map from Section 3.4.

Proof By definition, $F_{\mathcal{W}} = i \circ F_{\mathcal{W}'} = i \circ F_{(\mathcal{W}')^s} \circ F_{(\mathcal{W}')^b}$, where

$$\mathcal{W}' = (W, Z', [\xi']): (M, \gamma) \rightarrow (M', \gamma') \sqcup (B, \delta)$$

is the cobordism defined above, and

$$i: \text{SFH}((M_1, \gamma_1) \sqcup (B, \delta)) \xrightarrow{\sim} \text{SFH}(M_1, \gamma_1)$$

is the canonical isomorphism. As Z' has no isolated components, $F_{(\mathcal{W}')^b}$ is the Honda–Kazez–Matić gluing map $\Phi_{-\xi'}$, which agrees with our gluing map from Section 3.4 by Theorem 5.8.

On the other hand, $\Phi_{-\xi} = \Phi_{-\xi_B} \circ \Phi_{-\xi'}$, where $\xi_B = \xi|_B$. Hence, it suffices to show

$$i \circ F_{(\mathcal{W}')^s} = F_{\mathcal{W}^s} \circ \Phi_{-\xi_B}.$$

Let $\mathcal{W}_B$ be the special cobordism from $(M_0 \cup -Z', \gamma_1 \cup \delta)$ to $(M_0 \cup -Z, \gamma_1) \sqcup (B, \delta)$ obtained by pushing ∂B slightly into Z' , and attaching a 4-dimensional 3-handle along each component. Then

$$(\mathcal{W}')^s = (\mathcal{W}^s \sqcup \text{id}_{(B, \delta)}) \circ \mathcal{W}_B.$$

By definition, the contact 3-handle map $\Phi_{-\xi_B}$ is equal to $j \circ F_{\mathcal{W}_B}$, where

$$j: \text{SFH}((M_0 \cup -Z, \gamma_1) \sqcup (B, \delta)) \xrightarrow{\sim} \text{SFH}(M_0 \cup -Z, \gamma_1)$$

is the canonical isomorphism; see [Section 3.3.4](#). Hence, it suffices to show that

$$i \circ F_{\mathcal{W}^s \sqcup \text{id}_{(B, \delta)}} = F_{\mathcal{W}^s} \circ j.$$

This holds since $F_{\mathcal{W}^s \sqcup \text{id}_{(B, \delta)}} = F_{\mathcal{W}^s} \otimes \text{id}_{\text{SFH}(B, \delta)}$ and $\text{SFH}(B, \delta) \cong \mathbb{F}_2$. $\square$

6.3 Turning around sutured cobordisms

In this section, we use our contact handle maps to prove a first result about duality in sutured Floer homology. If

$$\mathcal{W} = (W, Z, [\xi]): (M, \gamma) \rightarrow (M', \gamma')$$

is a cobordism of sutured manifolds, then we can form the cobordism

$$\mathcal{W}^\vee := (W, Z, [\xi]): (-M', \gamma') \rightarrow (-M, \gamma)$$

by reversing which ends of W are viewed as incoming or outgoing. The main result of this section is the following:

Theorem 6.6 *If $\mathcal{W}: (M, \gamma) \rightarrow (M', \gamma')$ is a balanced cobordism of sutured manifolds, and $\mathcal{W}^\vee$ is the cobordism obtained by turning around $\mathcal{W}$, then*

$$F_{\mathcal{W}^\vee} = (F_{\mathcal{W}})^\vee,$$

with respect to the trace pairing from [Section 6.1](#).

Suppose that (M, γ) is a sutured submanifold of (M', γ') , and let ξ be a contact structure on $Z = -M' \setminus \text{int}(M)$ with dividing set $\gamma \cup \gamma'$ on the convex surface ∂Z . Consider the boundary cobordism

$$\mathcal{W} := (I \times M' / \sim, \{0\} \times Z, [\xi])$$

from (M, γ) to (M', γ') . Then $\mathcal{W}^\vee$ is a sutured cobordism from $(-M', \gamma')$ to $(-M, \gamma)$. In general, $\mathcal{W}^\vee$ will be neither a special cobordism nor a boundary cobordism.

It is the product of a boundary cobordism

$$(\mathcal{W}^\vee)^b: (-M', \gamma') \rightarrow (-M' \cup -Z, \gamma)$$

and a special cobordism

$$(\mathcal{W}^\vee)^s: (-M' \cup -Z, \gamma) \rightarrow (-M, \gamma).$$

The 4-manifold underlying $(\mathcal{W}^\vee)^s$ is also $M' \times I / \sim$. We need the following topological description of the special cobordism $(\mathcal{W}^\vee)^s$:

Lemma 6.7 *Suppose that (M, γ) is a sutured submanifold of (M', γ') and*

$$\mathcal{W}^\vee := (I \times M' / \sim, \{0\} \times Z, [\xi]): (-M', \gamma') \rightarrow (-M, \gamma)$$

is the dual of the corresponding boundary cobordism, as described above, and suppose that (Z, ξ) has a contact handle decomposition relative to ∂M with an associated Morse function $f: Z \rightarrow I$. Then the special part $(\mathcal{W}^\vee)^s$ of $\mathcal{W}^\vee$ has a Morse function F whose critical points are in bijective correspondence with the critical points of f . Furthermore, if p is a critical point of f and p' is the associated critical point of F , then

$$\text{ind}_{p'}(F) = 4 - \text{ind}_p(f).$$

The intersection of the descending manifold of a critical point of F with

$$-M' \cup_{\partial M'} -Z = -M \cup_{\partial M} Z \cup_{\partial M'} -Z$$

is equal to the union of the ascending flow lines of the corresponding critical point of f in Z , together with their images in $-Z$.

Proof We first define an auxiliary function

$$G: (I \times [-1, 2]) / (I \times \{2\}) \rightarrow I,$$

where $I = [0, 1]$. We require G to satisfy the following:

- $G(t, s) = t$ for $s = -1$.
- $\nabla G \neq 0$ for all (t, s) .
- If $(t, s) \in I \times I$, then $(\partial G / \partial t)(t, s) = 0$ if and only if $t = \frac{1}{2}$.
- If $(t, s) \in \{\frac{1}{2}\} \times I$, then $(\partial G / \partial s) < 0$.
- $G|_{\{0\} \times [-1, 2]} \equiv G|_{I \times \{2\}} \equiv G|_{\{1\} \times [0, 2]} \equiv 0$.

The graph of such a function $G(t, s)$ is shown in [Figure 25](#).

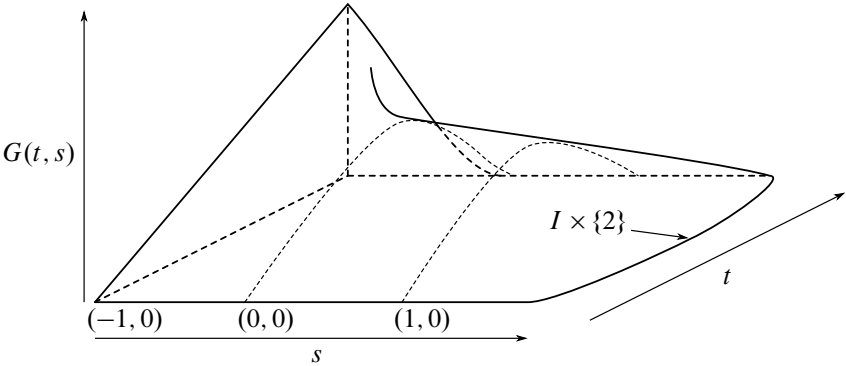


Figure 25: An example of a function $G(t, s): (I \times [-1, 2]) / (I \times \{2\}) \rightarrow I$ considered in Lemma 6.7.

Let us view M' as

$$M' \cong M \cup (\partial M \times [-1, 0]) \cup -Z \cup (\partial M' \times [1, 2]),$$

and extend f over all of M' so that

- $f|_M \equiv -1$,
- $f|_{\partial M \times [-1, 0]}(x, s) = s$, and
- $f|_{\partial M' \times [1, 2]}(x, s) = s$.

We then consider the function $F: I \times M' / \sim \rightarrow I$ given by

$$F(t, x) := G(t, f(x)).$$

It is then straightforward to verify that F has the stated properties. □

Remark 6.8 For an index 2 critical point of f , the attaching sphere of the corresponding critical point of F will be a knot K . The framing of K depends on some auxiliary choices, such as a choice of Riemannian metric, and the precise choice of G . However, up to isotopy, the framing is determined uniquely by the property that the framing of $K \cap Z$ is the mirror of the framing of $K \cap (-Z)$.

Proof of Theorem 6.6 The claim was shown for special cobordisms in [15, Theorem 11.8]. Hence, it only remains to verify it for boundary cobordisms. By the composition law for the gluing map, it is sufficient to prove the claim when the boundary cobordism $\mathcal{W} = (W, Z, [\xi])$ is formed by adding a single contact k -handle for $k \in \{0, 1, 2, 3\}$. The cobordism map

$$F_{\mathcal{W}}: \text{SFH}(M, \gamma) \rightarrow \text{SFH}(M', \gamma')$$

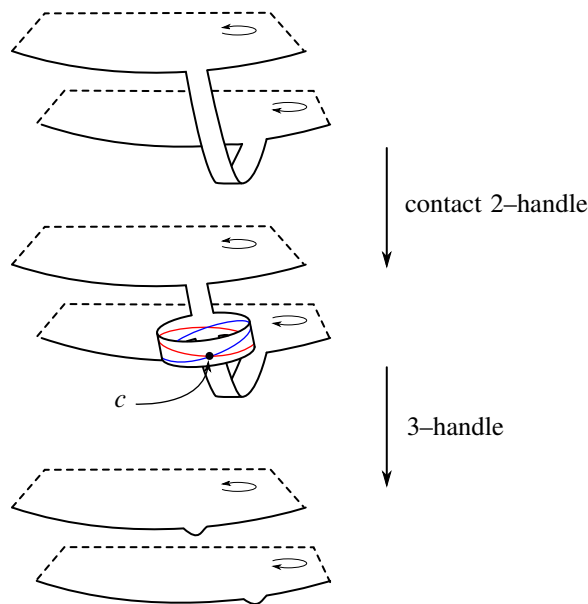


Figure 26: Computing the cobordism map for a turned-around contact 1-handle. Orientations on the Heegaard surface are shown.

is the gluing map $\Phi_{-\xi}$ corresponding to the sutured submanifold $(-M, -\gamma)$ of $(-M', -\gamma')$.

For a contact 0-handle, the map $\Phi_{-\xi}$ is the tautological one from $\text{SFH}(M, \gamma)$ to

$$\text{SFH}(M', \gamma') \cong \text{SFH}(M, \gamma) \otimes \text{SFH}(D^2 \times [-1, 1], S^1 \times \{0\}) \cong \text{SFH}(M, \gamma) \otimes \mathbb{F}_2.$$

On the other hand, consider $\mathcal{W}^\vee$ from $(-M', \gamma')$ to $(-M, \gamma)$. Here Z has an isolated component Z_0 corresponding to the contact 0-handle; ie $Z_0 \cap M = \emptyset$. Hence, by [15, Definition 10.1], the map $F_{\mathcal{W}^\vee}$ is defined by removing a standard contact ball B with connected dividing set δ on ∂B from the interior of Z_0 , and adding (B, δ) to $(-M, \gamma)$. The resulting cobordism is a product from $(-M', \gamma') = (-M, \gamma) \sqcup (B, \delta)$ to $(-M', \gamma')$. Hence $F_{\mathcal{W}^\vee}$ is $\text{id}_{\text{SFH}(-M', \gamma')}$, followed by the canonical identification of $\text{SFH}(-M', \gamma')$ with $\text{SFH}(-M, \gamma)$, which is $\Phi_{-\xi}^\vee$.

Now suppose that $\mathcal{W}$ is formed by adding a contact 1-handle to (M, γ) . In this case, $F_{\mathcal{W}} = \Phi_{-\xi}$, which is obtained by adding a strip to the boundary of a Heegaard diagram for $(-M, -\gamma)$. By Lemma 6.7, the cobordism $\mathcal{W}^\vee$ is obtained by gluing $(Z, -\xi)$ along $\partial M'$ to $M' = M \cup_{\partial M} -Z$, then attaching a 4-dimensional 3-handle. Note that we attach Z along $\partial M'$, not ∂M , so it becomes a contact 2-handle. The 4-dimensional

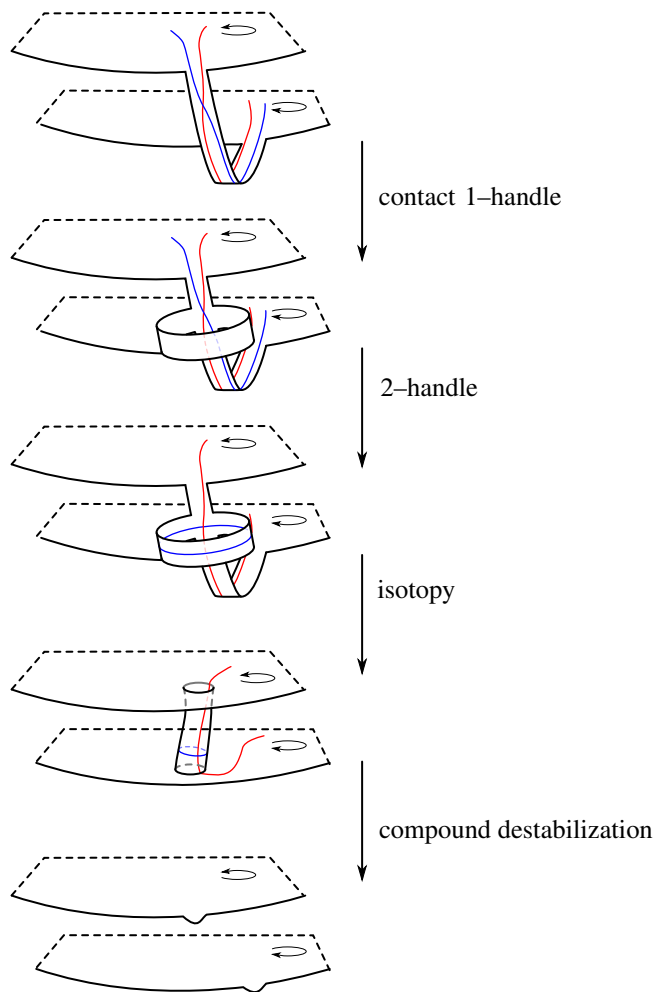


Figure 27: Computing the cobordism map for a turned-around contact 2-handle.

3-handle is attached to a 2-sphere in $Z \cup_{\partial M'} -Z$. As in [Lemma 6.7](#), the 2-sphere is the union of the ascending flow lines of f in Z (ie the cocore of Z , viewed as a contact 1-handle attached to $-M$) together with its image in the copy of $-Z$ that we glue onto $-M'$. Diagrammatically, this is shown in [Figure 26](#). An easy model computation shows that this is equal to the dual of the contact 1-handle map.

We now consider a sutured cobordism $\mathcal{W}$ formed by a contact 2-handle attachment to the sutured manifold (M, γ) . In this case, the dual cobordism $\mathcal{W}^\vee$ is formed by gluing $(Z, -\xi)$ to $M' = M \cup -Z$ along $\partial M'$ and then attaching a 4-dimensional

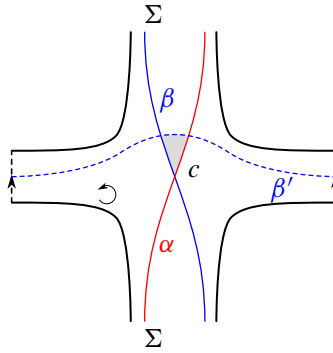


Figure 28: The triangle map computation for the 4-dimensional 2-handle map in the cobordism map for a turned-around contact 2-handle. The orientation of the surface is shown. The two dashed lines with arrows on the left and right are identified. The only homology classes of triangles that have a vertex at c have multiplicity 1 in the shaded region, and zero in the other regions shown.

2-handle. Gluing $(Z, -\xi)$ to M' is now a contact 1-handle attachment. As described in Lemma 6.7, the knot that we attach a 2-handle along is given as the union of the cocore of Z (viewed as a contact 2-handle) as well as its image in $-Z$. Let $\mathcal{H}'$ be an admissible diagram of $(-M', \gamma')$. After adding the contact 1-handle and the 4-dimensional 2-handle, we get a diagram that is a compound stabilization of $\mathcal{H}'$. After performing a compound destabilization, we get back to $\mathcal{H}'$. An easy triangle map computation for the 2-handle map shows that the composition of the triangle map and the compound destabilization map is dual to the contact 2-handle map. That the compound stabilization map agrees with the map from naturality is shown in Proposition 2.2. A schematic for the turned-around contact 2-handle cobordism is shown in Figure 27. The triangle map is shown in more detail in Figure 28.

Finally, we consider the case when $\mathcal{W}$ is formed by adding a contact 3-handle. In this case, Z has an isolated component, and the cobordism map $F_{\mathcal{W}}$ is obtained by removing a standard contact ball from the 3-ball we are adding. Then the map is computed as a trivial gluing map, followed by a 4-dimensional 3-handle map. The 4-dimensional 3-handle is attached along an embedded 2-sphere that is the push-off of the boundary 2-sphere on which we are attaching the contact 3-handle. The dual cobordism $\mathcal{W}^\vee$ is formed by adding a contact 0-handle, followed by a 4-dimensional 1-handle. Hence, the claim that $F_{\mathcal{W}}^\vee = F_{\mathcal{W}^\vee}$ follows from the fact that the 4-dimensional 1-handle and 3-handle maps are dual to each other, as in [25, Theorem 3.5]. $\square$

7 Triangle cobordisms

If $\mathcal{T} = (\Sigma, \alpha, \beta, \gamma)$ is a balanced sutured triple diagram, then there is a natural sutured cobordism

$$\mathcal{W}_{\alpha, \beta, \gamma} = (W_{\alpha, \beta, \gamma}, Z_{\alpha, \beta, \gamma}, \xi_{\alpha, \beta, \gamma}),$$

as in [15, Section 5]. We note, however, that the construction of the contact structure $\xi_{\alpha, \beta, \gamma}$ in [15, Section 5] was incorrect, as it involved gluing contact structures along annuli whose boundaries did not intersect the dividing set. In this section, we will provide a different description, which we will take as the definition.

Before describing the 4-manifold $W_{\alpha, \beta, \gamma}$, we establish some notation. For $\tau \in \{\alpha, \beta, \gamma\}$, let U_τ be the sutured compression body obtained from $\Sigma \times I$ by attaching 3-dimensional 2-handles along $\tau \times \{0\} \subseteq \Sigma \times \{0\}$. We view Σ as being embedded into ∂U_τ as $\Sigma \times \{1\}$. Similarly, we denote by R_τ the result of surgering $\Sigma \times \{0\}$ along $\tau \times \{0\}$. Using this notation,

$$\partial U_\tau = \Sigma \cup (\partial \Sigma \times I) \cup \bar{R}_\tau,$$

where $\bar{R}_\tau$ is the surface R_τ with the opposite orientation.

For $\tau, \tau' \in \{\alpha, \beta, \gamma\}$, consider the 3-manifold

$$M_{\tau, \tau'} := U_\tau \cup_\Sigma -U_{\tau'},$$

and let $R_{\tau, \tau'}$ denote the surface

$$\bar{R}_\tau \cup_{\partial \Sigma} R_{\tau'}.$$

Here, we write $-U_{\tau'}$ for the 3-manifold $U_{\tau'}$ with the opposite orientation. Note that, using this orientation convention, we have

$$\partial M_{\tau, \tau'} = R_{\tau, \tau'}.$$

The oriented 1-manifold $\partial \Sigma$ has a natural embedding into $\partial M_{\tau, \tau'}$, which we denote by $\gamma_{\tau, \tau'}$, and $(M_{\tau, \tau'}, \gamma_{\tau, \tau'})$ is a sutured manifold with diagram (Σ, τ, τ') .

Let Δ be a regular triangle in $\mathbb{C}$ with edges e_α , e_β , and e_γ appearing clockwise, and give Δ the complex orientation; see Figure 29. We define the 4-manifold

$$W_{\alpha, \beta, \gamma} := (\Delta \times \Sigma) \cup (e_\alpha \times U_\alpha) \cup (e_\beta \times U_\beta) \cup (e_\gamma \times U_\gamma) / \sim,$$

where $\sim$ denotes gluing $\Delta \times \Sigma$ to $e_\tau \times U_\tau$ along $e_\tau \times \Sigma$ for $\tau \in \{\alpha, \beta, \gamma\}$. Also, let

$$Z_{\alpha, \beta, \gamma} := (\Delta \times \partial \Sigma) \cup (e_\alpha \times R_\alpha) \cup (e_\beta \times R_\beta) \cup (e_\gamma \times R_\gamma) \subseteq \partial W_{\alpha, \beta, \gamma},$$

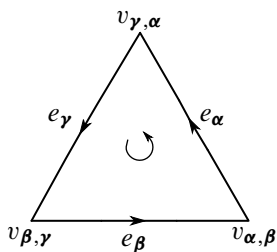


Figure 29: A triangle Δ , oriented as in the construction of $W_{\alpha,\beta,\gamma}$. The boundary orientations of the edges e_α , e_β , and e_γ are also shown.

which we orient as the boundary of $W_{\alpha,\beta,\gamma}$. We then have the decomposition

$$\partial W_{\alpha,\beta,\gamma} = -M_{\alpha,\beta} \cup -M_{\beta,\gamma} \cup M_{\alpha,\gamma} \cup Z_{\alpha,\beta,\gamma}.$$

Note that $M_{\alpha,\gamma} = -M_{\gamma,\alpha}$. Using our orientation conventions,

$$\partial Z_{\alpha,\beta,\gamma} = R_{\alpha,\beta} \sqcup R_{\beta,\gamma} \sqcup \bar{R}_{\alpha,\gamma}.$$

There is a natural collection of sutures $\kappa_{\alpha,\beta,\gamma}$ on $\partial Z_{\alpha,\beta,\gamma}$, defined as

$$\kappa_{\alpha,\beta,\gamma} := \{v_{\alpha,\beta}, v_{\beta,\gamma}, v_{\gamma,\alpha}\} \times \partial \Sigma = \gamma_{\alpha,\beta} \cup \gamma_{\beta,\gamma} \cup \gamma_{\gamma,\alpha},$$

where $v_{\sigma,\tau}$ is the vertex of Δ between e_σ and e_τ . Then

$$R_+(\kappa_{\alpha,\beta,\gamma}) = (\{v_{\alpha,\beta}\} \times R_\beta) \cup (\{v_{\beta,\gamma}\} \times R_\gamma) \cup (\{v_{\gamma,\alpha}\} \times R_\alpha).$$

There is a natural contact structure $\xi_{\alpha,\beta,\gamma}$ on $Z_{\alpha,\beta,\gamma}$ with dividing set $\kappa_{\alpha,\beta,\gamma}$, which is positive for the orientation of $Z_{\alpha,\beta,\gamma}$ we have described. We delay our description of $\xi_{\alpha,\beta,\gamma}$ until [Section 7.1](#). The triple $\mathcal{W}_{\alpha,\beta,\gamma} = (W_{\alpha,\beta,\gamma}, Z_{\alpha,\beta,\gamma}, \xi_{\alpha,\beta,\gamma})$ is a sutured manifold cobordism from $(M_{\alpha,\beta}, \gamma_{\alpha,\beta}) \sqcup (M_{\beta,\gamma}, \gamma_{\beta,\gamma})$ to $(M_{\alpha,\gamma}, \gamma_{\alpha,\gamma})$. The main goal of this section is to prove the following:

Theorem 7.1 *Let $(\Sigma, \alpha, \beta, \gamma)$ be an admissible balanced sutured triple diagram. Then the cobordism map*

$$F_{\mathcal{W}_{\alpha,\beta,\gamma}}: \text{CF}(\Sigma, \alpha, \beta) \otimes \text{CF}(\Sigma, \beta, \gamma) \rightarrow \text{CF}(\Sigma, \alpha, \gamma)$$

is chain homotopic to the map $F_{\alpha,\beta,\gamma}$ defined in [\[15, Definition 5.13\]](#) that counts holomorphic triangles on the triple diagram $(\Sigma, \alpha, \beta, \gamma)$.

The proof occupies the remainder of this section. The first step is to define the contact structure $\xi_{\alpha,\beta,\gamma}$ in detail and describe some useful properties. Next, we introduce some

special diagrams called “doubled diagrams” and “weakly conjugated diagrams” that appear when computing the cobordism map for $\mathcal{W}_{\alpha,\beta,\gamma}$. We then compute convenient formulas for the contact gluing map for $(Z_{\alpha,\beta,\gamma}, \xi_{\alpha,\beta,\gamma})$, and for the 4-dimensional handle attachments of the cobordism $\mathcal{W}_{\alpha,\beta,\gamma}$. Finally, we put all the pieces together, using associativity of the triangle maps and some other relations to show that the cobordism map $F_{\mathcal{W}_{\alpha,\beta,\gamma}}$ is chain homotopic to the triangle map $F_{\alpha,\beta,\gamma}$.

7.1 The contact structure $\xi_{\alpha,\beta,\gamma}$ on $Z_{\alpha,\beta,\gamma}$

The sutured manifold $(Z_{\alpha,\beta,\gamma}, \kappa_{\alpha,\beta,\gamma})$ has a natural contact structure $\xi_{\alpha,\beta,\gamma}$ that we define by decomposing $Z_{\alpha,\beta,\gamma}$ along convex surfaces. The construction generalizes to the case of sutured multid diagrams $(\Sigma, \eta_1, \dots, \eta_n)$ for arbitrary $n \geq 1$, though in this paper, we will only need it for $n \in \{1, 2, 3\}$.

Let P_n be an n -gon, viewed as the complex unit disk with boundary divided into n arcs labeled $e_{\eta_1}, \dots, e_{\eta_n}$ clockwise. We write v_{η_{i-1}, η_i} for the terminal endpoint of e_{η_i} for $i \in \{1, \dots, n\}$, where $\eta_0 := \eta_n$. Let

$$(7-1) \quad Z_{\eta_1, \dots, \eta_n} = (\Delta \times \partial\Sigma) \cup (e_{\eta_1} \times R_{\eta_1}) \cup \dots \cup (e_{\eta_n} \times R_{\eta_n})$$

with sutures

$$\kappa_{\eta_1, \dots, \eta_n} = \{v_{\eta_1, \eta_2}, \dots, v_{\eta_n, \eta_1}\} \times \partial\Sigma = \gamma_{\eta_1, \eta_2} \cup \dots \cup \gamma_{\eta_n, \eta_1}$$

and

$$R_+(\kappa_{\eta_1, \dots, \eta_n}) = (\{v_{\eta_1, \eta_2}\} \times R_{\eta_2}) \cup \dots \cup (\{v_{\eta_n, \eta_1}\} \times R_{\eta_1}).$$

We now define the contact structure $\xi_{\eta_1, \dots, \eta_n}$ on $(Z_{\eta_1, \dots, \eta_n}, \kappa_{\eta_1, \dots, \eta_n})$. Let us write Z_0 for $P_n \times \partial\Sigma$, which is a union of $|\partial\Sigma|$ solid tori. Let γ_0 consist of two parallel longitudinal sutures on each component of ∂Z_0 such that their projections to P_n wind positively around ∂P_n with respect to the orientation of P_n (ie they wind counterclockwise around $\partial P_n \subseteq \mathbb{C}$). Since (Z_0, γ_0) is product disc decomposable, it admits a unique tight contact structure ξ_0 , up to equivalence, which has ∂Z_0 as a convex surface with dividing set γ_0 .

For $\tau \in \{\eta_1, \dots, \eta_n\}$, we write (Z_τ, γ_τ) to denote the product sutured manifold $(e_\tau \times R_\tau, \{m_\tau\} \times \partial R_\tau)$, where m_τ is the midpoint of e_τ . The sutured manifold (Z_τ, γ_τ) is product disc decomposable, and hence admits a unique tight contact structure ξ_τ with dividing set γ_τ , up to equivalence. Let s_τ denote a small translation of the suture γ_τ for $\tau \in \{\eta_1, \dots, \eta_n\}$ such that each component of s_τ intersects the corresponding

component of γ_τ transversely at exactly two points. Let $N(s_\tau) \subseteq \partial Z_\tau$ denote a small regular neighborhood of s_τ that intersects γ_τ along two arcs. Using Giroux's Legendrian realization principle, we may assume that each $\partial N(s_\tau)$ is Legendrian.

We now describe how the subsurfaces $N(s_\tau) \subseteq Z_\tau$ are glued to Z_0 . By picking γ_0 appropriately, we may assume that each component of $\{m_\tau\} \times \partial \Sigma$ intersects γ_0 transversely at two points. For each $\tau \in \{\eta_1, \dots, \eta_n\}$, we pick a small neighborhood $N(\{m_\tau\} \times \partial \Sigma) \subseteq \partial Z_0$ such that each component of $N(\{m_\tau\} \times \partial \Sigma)$ intersects γ_0 along two arcs. Using Legendrian realization, we may assume that $\partial N(\{m_\tau\} \times \partial \Sigma)$ is Legendrian.

We glue (Z_τ, ξ_τ) for every $\tau \in \{\eta_1, \dots, \eta_n\}$ to (Z_0, ξ_0) by identifying $N(s_\tau)$ and $N(\{m_\tau\} \times \partial \Sigma)$. We let $\xi_{\eta_1, \dots, \eta_n}$ be the resulting contact structure. After rounding the Legendrian corners, the contact structure $\xi_{\eta_1, \dots, \eta_n}$ has dividing set isotopic to $\kappa_{\eta_1, \dots, \eta_n}$. This is shown in Figure 30 for $n = 3$ and $(\eta_1, \eta_2, \eta_3) = (\alpha, \beta, \gamma)$. As $N(s_\tau)$ is unique up to isotopy, $\xi_{\eta_1, \dots, \eta_n}$ is well defined up to equivalence.

The following will be useful throughout the paper:

Lemma 7.2 *If (Σ, α, β) is a sutured Heegaard diagram, then the sutured manifold $(Z_{\alpha, \beta}, \kappa_{\alpha, \beta})$ is diffeomorphic to*

$$(I \times R_{\alpha, \beta}, (-\{0\} \times \gamma_{\alpha, \beta}) \cup (\{1\} \times \gamma_{\alpha, \beta})),$$

and the contact structure $\xi_{\alpha, \beta}$ is isotopic to the I -invariant contact structure on $I \times R_{\alpha, \beta}$ with dividing set $\{t\} \times \gamma_{\alpha, \beta}$ on $\{t\} \times R_{\alpha, \beta}$ for every $t \in I$.

Proof By construction, $Z_{\alpha, \beta}$ is obtained by gluing $e_\alpha \times R_\alpha$ and $e_\beta \times R_\beta$ to $P_2 \times \partial \Sigma$. This is diffeomorphic to gluing $I \times R_\alpha$ to $I \times R_\beta$ by identifying $(t, x) \in I \times \partial \Sigma$ with $(1-t, x) \in I \times \partial \Sigma$, where $\partial R_\alpha = \partial R_\beta = \partial \Sigma$. Hence $Z_{\alpha, \beta}$ is diffeomorphic to $I \times (\bar{R}_\alpha \cup R_\beta) = I \times R_{\alpha, \beta}$. Under this diffeomorphism, the sutures $\kappa_{\alpha, \beta} = \gamma_{\alpha, \beta} \cup \gamma_{\beta, \alpha}$ are mapped to $(-\{0\} \times \gamma_{\alpha, \beta}) \cup (\{1\} \times \gamma_{\alpha, \beta})$, since $\gamma_{\beta, \alpha} = -\gamma_{\alpha, \beta}$. In particular, $R_+(\kappa_{\alpha, \beta})$ is identified with $(\{0\} \times R_\alpha) \cup (\{1\} \times R_\beta)$.

To see that the contact structure $\xi_{\alpha, \beta}$ is isotopic to the I -invariant contact structure in the statement, we will construct a decomposition of the latter along convex annuli that cuts $I \times R_{\alpha, \beta}$ into the disjoint union of the three contact manifolds (Z_α, ξ_α) , (Z_β, ξ_β) , and (Z_0, ξ_0) . Let $s_1, s_2 \subseteq R_{\alpha, \beta}$ be two disjoint curves that are both small translates of the dividing set $\gamma_{\alpha, \beta} = \partial \Sigma$. We assume that s_1 and s_2 each intersect $\gamma_{\alpha, \beta}$ transversely at two points. Using Legendrian realization, we can assume that both s_1 and s_2 are Legendrian.

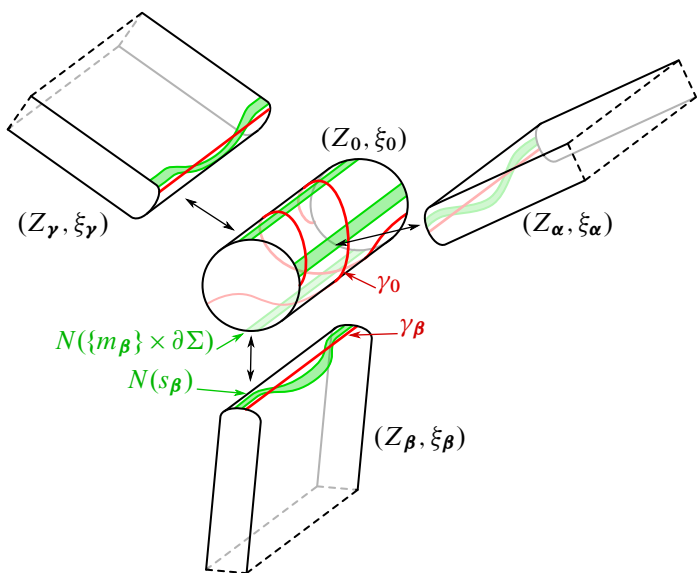


Figure 30: Constructing the contact structure $\xi_{\alpha,\beta,\gamma}$ on $Z_{\alpha,\beta,\gamma}$ by gluing along convex surfaces. We glue (Z_α, ξ_α) , (Z_β, ξ_β) , and (Z_γ, ξ_γ) to (Z_0, ξ_0) along the green strips, which have Legendrian boundary. The faces on the front side are identified with the faces on the back.

We will cut $I \times R_{\alpha,\beta}$ along the two annuli $I \times s_1$ and $I \times s_2$. We note that since s_1 and s_2 are Legendrians, the characteristic foliations on $I \times s_1$ and $I \times s_2$ are simple to describe. They consist of the horizontal leaves $\{t\} \times s_i$ for all $t \in I$, as well as two vertical lines of singularities along $I \times \{p\}$ for $p \in s_i \cap \gamma_{\alpha,\beta}$. This is shown in Figure 31. We note that the characteristic foliation satisfies the Poincaré–Bendixson property (the limit set of a flow line consists of a singular point, a periodic orbit, or a finite union of singular points and connecting orbits) and has no closed orbits or retrograde saddle connections. So, by the work of Giroux [7], the surfaces $I \times s_i$ are

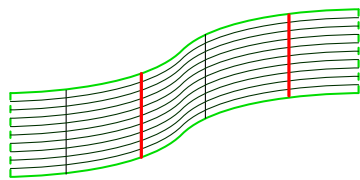


Figure 31: The characteristic foliation (thin black) and the dividing set (thick red) on the annulus $I \times s_i \subseteq I \times R_{\alpha,\beta}$. The vertical black lines consist of singularities. The left- and right-hand sides of the surface are identified to form an annulus.

convex. Furthermore, the dividing set on $I \times s_i$ consists of two curves of the form $I \times \{q\}$, where $q \in s_i \setminus \gamma_{\alpha,\beta}$. See also [3, Example 2.24].

After rounding the Legendrian corners that appear when we cut along $I \times s_i$, we obtain the disjoint union of the sutured manifolds (Z_α, ξ_α) , (Z_β, ξ_β) , and (Z_0, γ_0) . Furthermore, the contact structures obtained on the three pieces are isotopic to ξ_α , ξ_β , and ξ_0 , since they are tight by Giroux's criterion [10, Theorem 3.5], and ξ_α , ξ_β , and ξ_0 are the unique tight contact structures, by definition. A picture of the convex decomposition of $I \times R_{\alpha,\beta}$ along $I \times s_1$ and $I \times s_2$ is shown in Figure 32. $\square$

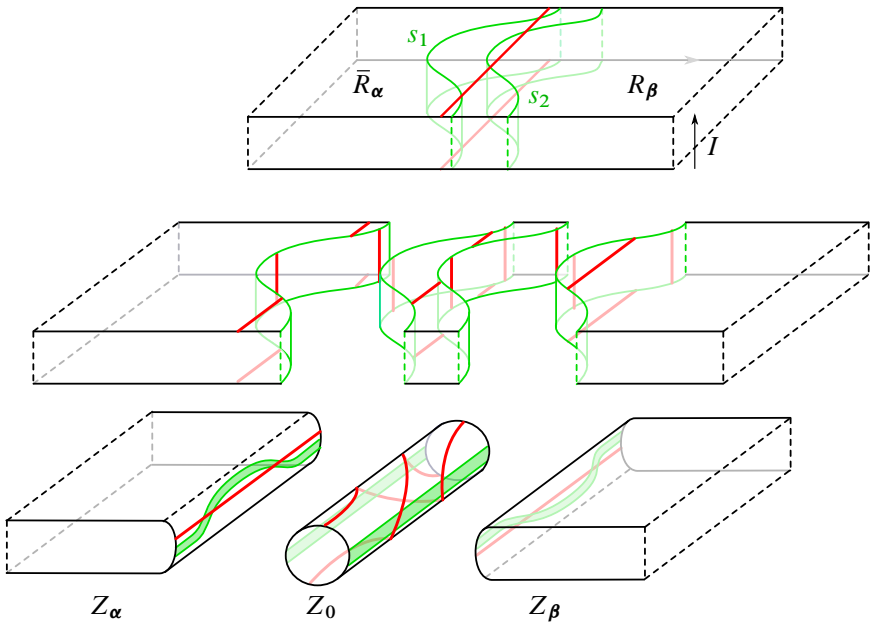


Figure 32: Decomposing $I \times R_{\alpha,\beta}$ (top) along two convex annuli $I \times s_1$ and $I \times s_2$, we obtain the middle picture. After rounding the Legendrian corners, we obtain the bottom picture, which is the disjoint union of Z_α , Z_β , and Z_0 . The front and back sides of each picture are identified.

Finally, we need one additional description of the contact structure $\xi_{\alpha,\beta,\gamma}$ on $Z_{\alpha,\beta,\gamma}$, in terms of gluing $(Z_{\alpha,\beta}, \xi_{\alpha,\beta})$ and $(Z_{\beta,\gamma}, \xi_{\beta,\gamma})$ together. By Lemma 7.2, the contact manifolds $(Z_{\alpha,\beta}, \xi_{\alpha,\beta})$ and $(Z_{\beta,\gamma}, \xi_{\beta,\gamma})$ are contactomorphic to $I \times R_{\alpha,\beta}$ and $I \times R_{\beta,\gamma}$, respectively, with I -invariant contact structures. We pick automorphisms of each of the surfaces $R_{\alpha,\beta}$ and $R_{\beta,\gamma}$, supported in neighborhoods of $\gamma_{\alpha,\beta}$ and $\gamma_{\beta,\gamma}$, that perturb the dividing sets $\gamma_{\alpha,\beta}$ and $\gamma_{\beta,\gamma}$ slightly. Write $s_{\alpha,\beta} \subseteq R_{\alpha,\beta}$ and $s_{\beta,\gamma} \subseteq R_{\beta,\gamma}$ for the images of $\gamma_{\alpha,\beta}$ and $\gamma_{\beta,\gamma}$, respectively. We assume that $s_{\alpha,\beta}$

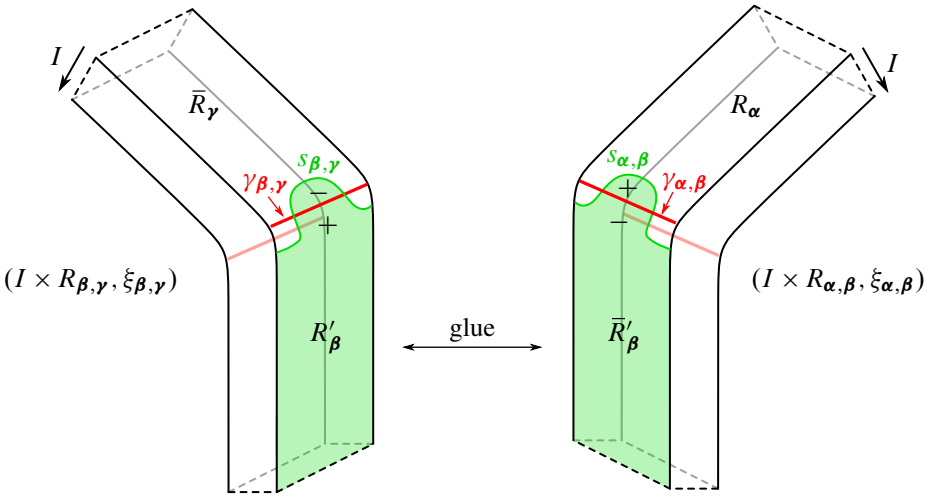


Figure 33: Gluing $(I \times R_{\alpha,\beta}, \xi_{\alpha,\beta})$ and $(I \times R_{\beta,\gamma}, \xi_{\beta,\gamma})$ together to obtain $(Z_{\alpha,\beta,\gamma}, \xi_{\alpha,\beta,\gamma})$. We glue along the green shaded subsurfaces labeled R'_β and $\bar{R}'_\beta$. The dividing sets $\gamma_{\alpha,\beta}$ and $\gamma_{\beta,\gamma}$ are shown in red. The curves $s_{\alpha,\beta}$ and $s_{\beta,\gamma}$ are small perturbations of the dividing sets and are Legendrian.

and $\gamma_{\alpha,\beta}$ intersect transversely at two points, and similarly for $s_{\beta,\gamma}$ and $\gamma_{\beta,\gamma}$. Let $\bar{R}'_\beta \subseteq \{0\} \times R_{\alpha,\beta}$, oriented as the boundary of $I \times R_{\alpha,\beta}$, denote the image of $\{0\} \times \bar{R}_\beta \subseteq \{0\} \times R_{\alpha,\beta}$, and let $R'_\beta \subseteq \{0\} \times R_{\beta,\gamma}$ denote the image of $\{0\} \times R_\beta \subseteq \{0\} \times R_{\beta,\gamma}$. Using the Legendrian realization principle, we may assume that $s_{\alpha,\beta}$ and $s_{\beta,\gamma}$ are Legendrian. The following description of $(Z_{\alpha,\beta,\gamma}, \xi_{\alpha,\beta,\gamma})$ will be useful for our purposes:

Lemma 7.3 *The contact structure $(Z_{\alpha,\beta,\gamma}, \xi_{\alpha,\beta,\gamma})$ is equivalent to the one obtained by gluing $(I \times R_{\alpha,\beta}, \xi_{\alpha,\beta})$ and $(I \times R_{\beta,\gamma}, \xi_{\beta,\gamma})$ together along the surfaces R'_β and $\bar{R}'_\beta$ (which are convex with Legendrian boundary) described above; see [Figure 33](#).*

Proof As in [Lemma 7.2](#), we work backwards, and provide a convex decomposition of $(Z_{\alpha,\beta,\gamma}, \xi_{\alpha,\beta,\gamma})$ into the disjoint union of $(I \times R_{\alpha,\beta}, \xi_{\alpha,\beta})$ and $(I \times R_{\beta,\gamma}, \xi_{\beta,\gamma})$. By definition, $Z_{\alpha,\beta,\gamma}$ is the union of Z_α , Z_β , Z_γ , and Z_0 . Let us view Z_β as $I \times R_\beta$ with I -invariant contact structure, and Legendrian corners along $\{0, 1\} \times \partial R_\beta$, as on the right side of the middle level of [Figure 32](#). We start with the surface $S_\beta := \{\frac{1}{2}\} \times R_\beta$, which is convex with Legendrian boundary. We can view ∂S_β as a Legendrian arc on ∂Z_0 that intersects the dividing set on ∂Z_0 at two points. There is an annulus $S_0 \subseteq Z_0$ with one boundary component on ∂S_β , and another boundary component on ∂Z_0 between Z_γ and Z_α . Furthermore, after perturbing the surface S_0 to be

convex, it cuts (Z_0, ξ_0) into two copies of (Z_0, ξ_0) . To see this, we consider (Z_0, ξ_0) as the I -invariant contact manifold (with Legendrian corners) in the center of the middle of Figure 32. In this picture, an example of an annulus cutting (Z_0, ξ_0) into two copies of (Z_0, ξ_0) would be the intersection of Z_0 with the slice $\{\frac{1}{2}\} \times R_{\alpha, \beta}$. We let our decomposing surface S be the union $S_\beta \cup S_0$. This is shown schematically in Figure 34. When we cut along S , we get two components, one of which is $Z_\alpha \cup Z_0 \cup Z_\beta$, and one of which is $Z_\beta \cup Z_0 \cup Z_\gamma$. Lemma 7.2 identifies these latter two contact manifolds with $I \times R_{\alpha, \beta}$ and $I \times R_{\beta, \gamma}$ with I -invariant contact structures. $\square$

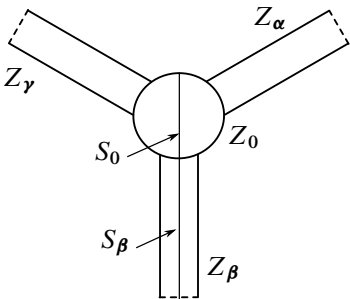


Figure 34: Decomposing $Z_{\alpha, \beta, \gamma}$ along the convex surface $S = S_\beta \cup S_0$ into $Z_\alpha \cup Z_0 \cup Z_\beta$ and $Z_\beta \cup Z_0 \cup Z_\gamma$.

7.2 A handle decomposition of $(W_{\alpha, \beta, \gamma})^s$

Recall that $\mathcal{W}_{\alpha, \beta, \gamma} = (W_{\alpha, \beta, \gamma}, Z_{\alpha, \beta, \gamma}, \xi_{\alpha, \beta, \gamma})$ is a sutured manifold cobordism from $(M_{\alpha, \beta}, \gamma_{\alpha, \beta}) \sqcup (M_{\beta, \gamma}, \gamma_{\beta, \gamma})$ to $(M_{\alpha, \gamma}, \gamma_{\alpha, \gamma})$. The boundary cobordism $(\mathcal{W}_{\alpha, \beta, \gamma})^b$ corresponds to gluing $(Z_{\alpha, \beta, \gamma}, -\xi_{\alpha, \beta, \gamma})$ to

$$(-M_{\alpha, \beta}, -\gamma_{\alpha, \beta}) \sqcup (-M_{\beta, \gamma}, -\gamma_{\beta, \gamma}).$$

In light of Lemma 7.3, we can topologically view this as gluing $\bar{R}_\beta \subseteq -\partial M_{\alpha, \beta}$ to $R_\beta \subseteq -\partial M_{\beta, \gamma}$. Hence, the special cobordism $(W_{\alpha, \beta, \gamma})^s$ goes from $(M_{\alpha, \beta} \cup_{R_\beta} M_{\beta, \gamma}, \gamma_{\alpha, \gamma})$ to $(M_{\alpha, \gamma}, \gamma_{\alpha, \gamma})$. In this section, we give a topological description of the special cobordism $(W_{\alpha, \beta, \gamma})^s$ in terms of 4-dimensional handle attachments.

A handle decomposition of $(W_{\alpha, \beta, \gamma})^s$ can be constructed from a sutured Morse function on U_β , as we now describe. Let $f_\beta: U_\beta \rightarrow I$ be a Morse function induced by the diagram (Σ, β) ; ie we view U_β as a collection of 2-handles glued to $\Sigma \times I$ along $\beta \times \{0\} \subseteq \Sigma \times \{0\}$. Furthermore, we pick f_β such that $f_\beta^{-1}(1) = \Sigma \times \{1\}$, $f_\beta^{-1}(0) = R_\beta$, and $f_\beta(y, t) = t$ for $y \in \partial \Sigma$. We also assume that f_β has $|\beta|$ index 1 critical points, whose ascending manifolds intersect $\Sigma \times \{1\}$ along the β curves, and that f_β has no

other critical points. For a curve $\beta_i \in \beta$, let $\lambda_i \subseteq U_\beta$ denote the descending manifold of the critical point of f_β corresponding to β_i . We have the following topological description of $W_{\alpha,\beta,\gamma}$:

Lemma 7.4 *The special cobordism $(W_{\alpha,\beta,\gamma})^s$ from $(M_{\alpha,\beta} \cup_{R_\beta} M_{\beta,\gamma}, \gamma_{\alpha,\gamma})$ to $(M_{\alpha,\gamma}, \gamma_{\alpha,\gamma})$ consists of $|\beta|$ 2-handle attachments. The 2-handles are attached along the link formed by concatenating the arcs $\lambda_i \subseteq U_\beta$ (which have boundary on $\bar{R}_\beta$) with their reflections in $-U_\beta$. The framing on this link is determined by picking an arbitrary framing on λ_i and concatenating it with the mirrored framing on the image of λ_i in $-U_\beta$.*

Proof This can be proven similarly to Lemma 6.7, by using a Morse function built from f_β as well as another auxiliary function. We leave the details to the reader. $\square$

7.3 Arc slides and bases of arcs

In this section, we describe some basic topological facts about arc decompositions of surfaces with boundary.

Definition 7.5 Suppose that Σ is a compact, orientable surface with no closed components and $\mathcal{I} \subseteq \partial\Sigma$ is a collection of subarcs of $\partial\Sigma$ such that each component of $\partial\Sigma$ contains at least one component of $\mathcal{I}$.

- (1) We say that $\mathbf{a} = \{a_1, \dots, a_n\}$ is an *arc basis* for $(\Sigma, \mathcal{I})$ if $\mathbf{a}$ is a set of pairwise disjoint, properly embedded arcs with boundary on $\mathcal{I}$ that form a basis of $H_1(\Sigma, \mathcal{I})$ (or, equivalently, $\mathbf{a}$ cuts Σ into a collection of closed disks, each containing exactly one component of $\partial\Sigma \setminus \mathcal{I}$).
- (2) We say an arc a'_i is formed by an *allowable arc slide* of a_i across a_j if there is an embedded hexagon $P_6 \subseteq \Sigma$ whose boundary consists of a_i , a_j , a'_i , as well as three subarcs of $\mathcal{I}$, and is otherwise disjoint from the arcs $a_1, \dots, a_n$; see Figure 35.

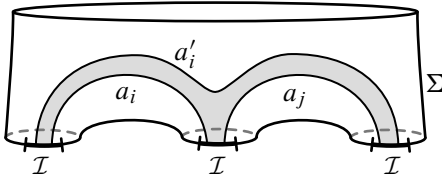


Figure 35: An allowable arc slide of a_i across a_j .

The reader may compare the following to [13, Lemma 3.3]:

Lemma 7.6 *Suppose Σ is a compact, orientable surface with boundary and no closed components and $\mathcal{I}$ is a collection of pairwise disjoint, closed subintervals of $\partial\Sigma$ such that each boundary component of $\partial\Sigma$ contains at least one component of $\mathcal{I}$. If $\mathbf{a}$ and $\mathbf{a}'$ are two arc bases of $(\Sigma, \mathcal{I})$, then $\mathbf{a}$ can be obtained from $\mathbf{a}'$ by a sequence of allowable arc slides.*

Proof Pick another collection of closed, pairwise disjoint subarcs $\mathcal{I}' \subseteq \text{int}(\partial\Sigma \setminus \mathcal{I})$ such that each component of $\partial\Sigma \setminus \mathcal{I}$ contains exactly one component of $\mathcal{I}'$. We view Σ as a cobordism from $\mathcal{I}$ to $\mathcal{I}'$, with horizontal boundary $\mathcal{I} \cup \mathcal{I}'$, and vertical boundary $\partial\Sigma \setminus \text{int}(\mathcal{I} \cup \mathcal{I}')$. We can view each basis of arcs $\mathbf{a}$ as corresponding to a Morse function f and gradient-like vector field v on Σ such that f has only index 1 critical points with stable manifolds $\mathbf{a}$. Furthermore, f is minimized along $\mathcal{I}$, maximized along $\mathcal{I}'$, and the components of $\partial\Sigma \setminus \text{int}(\mathcal{I} \cup \mathcal{I}')$ are flow lines of v .

More generally, we can consider pairs (f, v) of Morse functions f with gradient-like vector fields v such that f is minimized on $\mathcal{I}$, maximized on $\mathcal{I}'$, has no critical points along $\partial\Sigma$, but is allowed to have critical points of index 0, 1, and 2. Assuming (f, v) is also Morse–Smale, we can construct a graph $\Gamma \subseteq \Sigma$ whose vertices are the index 0 critical points of f and the points of $\mathcal{I}$ that flow to index 1 critical points along v and whose edges are the stable manifolds of the index 1 critical points. The graph Γ intersects $\partial\Sigma$ along a collection of valence 1 vertices in $\mathcal{I}$.

With this in mind, we say a graph $\Gamma \subseteq \Sigma$ is a *decomposing graph* of $(\Sigma, \mathcal{I})$ if Γ intersects $\partial\Sigma$ in a collection of valence 1 vertices in $\mathcal{I}$, and the interior of each component of $\Sigma \setminus \Gamma$ is homeomorphic to an open 2–ball that contains at most one component of $\partial\Sigma \setminus \mathcal{I}$. Note that, for a decomposing graph of an orientable surface, Definition 7.5 modifies easily to describe an arc slide of two edges of Γ that meet at a vertex in $\text{int}(\Sigma)$ and are consecutive with respect to the cyclic ordering of the edges adjacent to the vertex.

By interpreting decomposing graphs in terms of Morse functions and gradient-like vector fields, and considering the bifurcations occurring in generic 1–parameter families of smooth functions, it follows that any two decomposing graphs can be connected by a sequence of the following moves and their inverses:

(G-1) **Index 0/1 births** The addition of a vertex $v \in \text{int}(\Sigma) \setminus \Gamma$ as well as a new edge e connecting v to an existing vertex of $\Gamma \setminus \partial\Sigma$ or to a point in $\mathcal{I} \setminus \Gamma$.

- (G-2) **Index 1/2 births** The addition of a new edge to Γ whose interior is contained in $\Sigma \setminus \Gamma$ and which has both endpoints on the same vertex of $\Gamma \setminus \partial\Sigma$.
- (G-3) **Arc slides** An allowable arc slide of two adjacent edges along $\mathcal{I}$ or an arc slide of a pair of edges that meet at a vertex of $\Gamma \cap \text{int}(\Sigma)$ and are consecutive with respect to the cyclic ordering of the edges adjacent to the vertex.

Note that if an index 1/2 birth occurs in a generic 1-parameter family of Morse functions and gradient-like vector fields, then the corresponding decomposing graph Γ changes by adding an edge e to a component B of $\Sigma \setminus \Gamma$ with both endpoints on $\partial B \setminus (\partial\Sigma \setminus \mathcal{I})$. By assumption, B is an open 2-ball and $\partial B \cap (\partial\Sigma \setminus \mathcal{I})$ is either empty or a single arc. We can add e using moves (G-1)–(G-3). Indeed, using move (G-1), we add an edge e' to Γ with one endpoint along $\partial B \cap \mathcal{I}$ on the subarc of $\partial B \setminus (\partial\Sigma \setminus \mathcal{I})$ between the endpoints of e , and one endpoint at a new interior vertex $v \in B$. We then add a loop edge e'' at v using move (G-2). We slide the endpoints of e'' along the two sides of e' to $\mathcal{I}$, and finally slide them to the endpoints of e using move (G-3).

If $\Gamma = (\Gamma_1, \dots, \Gamma_n)$ is a sequence of decomposing graphs such that consecutive terms differ up to isotopy by an instance of moves (G-1)–(G-3) and their inverses, then we say that Γ is a *Cerf sequence*. We define $n_1(\Gamma)$ to be the number of moves (G-1), and $n_2(\Gamma)$ to be the number of moves (G-2) that occur in Γ . Note that if Γ_1 and Γ_n are arc bases, then $n_1(\Gamma) = n_2(\Gamma) = 0$ if and only if Γ consists only of allowable arc slides.

We will now show that if $\Gamma = (\Gamma_1, \dots, \Gamma_n)$ is a Cerf sequence of decomposing graphs connecting two arc bases $\mathbf{a}$ and $\mathbf{a}'$, and $n_1(\Gamma) > 0$, then there is a Cerf sequence Γ' connecting $\mathbf{a}$ and $\mathbf{a}'$ with

$$(7-2) \quad n_1(\Gamma') = n_1(\Gamma) - 1 \quad \text{and} \quad n_2(\Gamma') = n_2(\Gamma).$$

Such a sequence Γ' will be constructed by contacting certain edges of the Γ_i .

We say that an edge e of Γ_i is *special* if ∂e consists of two distinct points, at least one of which is contained in $\text{int}(\Sigma)$. It is easy to see that if Γ is a decomposing graph of $(\Sigma, \mathcal{I})$ with at least one vertex in $\text{int}(\Sigma)$, then Γ has at least one special edge. If e is a special edge of Γ , we construct a new decomposing graph $C_e(\Gamma)$ of Σ , which we call the *contraction* of Γ along e . Outside a small neighborhood of e , we define the graph $C_e(\Gamma)$ to be Γ . If e has two vertices in $\text{int}(\Sigma)$, then $C_e(\Gamma)$ is formed by collapsing e and its two endpoints into a single vertex. If instead e has one endpoint on $\partial\Sigma$, then $C_e(\Gamma)$ is formed by deleting e and performing an isotopy of the edges previously incident to e until they hit $\partial\Sigma$. The contraction operation is shown Figure 36.

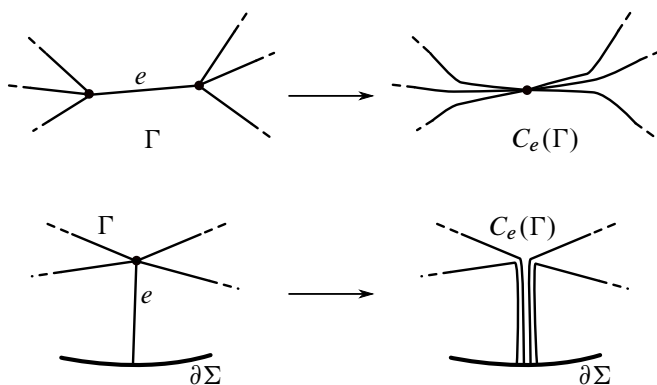


Figure 36: Contracting a decomposing graph along a special edge. On the top is when e has boundary equal to two vertices in $\text{int}(\Sigma)$. On the bottom is the case when ∂e has a vertex on $\partial\Sigma$.

Now suppose Γ is a Cerf sequence connecting $\mathbf{a}$ and $\mathbf{a}'$ with $n_1(\Gamma) > 0$. Let k be the first index where Γ_k is obtained from Γ_{k-1} by move (G-1) (an index 0/1 birth). We construct a sequence of special edges $e_k, \dots, e_l$ in $\Gamma_k, \dots, \Gamma_l$, respectively, for some $l \leq n$. We define the first special edge, e_k , to be the edge which is added to Γ_{k-1} to form Γ_k . We construct the remaining special edges e_i recursively. Supposing e_i has already been chosen, we define e_{i+1} as follows, depending on which move is used to form Γ_{i+1} from Γ_i :

- (1) **Γ_{i+1} is obtained by move (G-1) or its inverse** If e_i is deleted by this move, we declare e_i to be the final special edge in our sequence and set $l = i$. If e_i is unchanged, we set $e_{i+1} = e_i$.
- (2) **Γ_{i+1} is obtained by move (G-2) or its inverse** Since e_i is special, it cannot be deleted by this move. Noting this, we set $e_{i+1} = e_i$.
- (3) **Γ_{i+1} is obtained by move (G-3)** There are two subcases:
 - (a) **An edge e is slid over another edge, and $e \neq e_i$** Noting that e_i remains special, we set $e_{i+1} = e_i$.
 - (b) **The edge e_i is slid over another edge e** Write e'_i for the edge obtained by arc sliding e_i over e . If e'_i is not special, then it's easy to check that e must be special, and we set $e_{i+1} = e$. If e'_i is special, we set $e_{i+1} = e'_i$.

We now demonstrate that consecutive terms in the sequence

$$\Gamma' := (\Gamma_1, \dots, \Gamma_{k-1}, C_{e_k}(\Gamma_k), \dots, C_{e_l}(\Gamma_l), \Gamma_{l+1}, \dots, \Gamma_n)$$

are related by either a single instance of moves (G-1) or (G-2) or their inverses, by

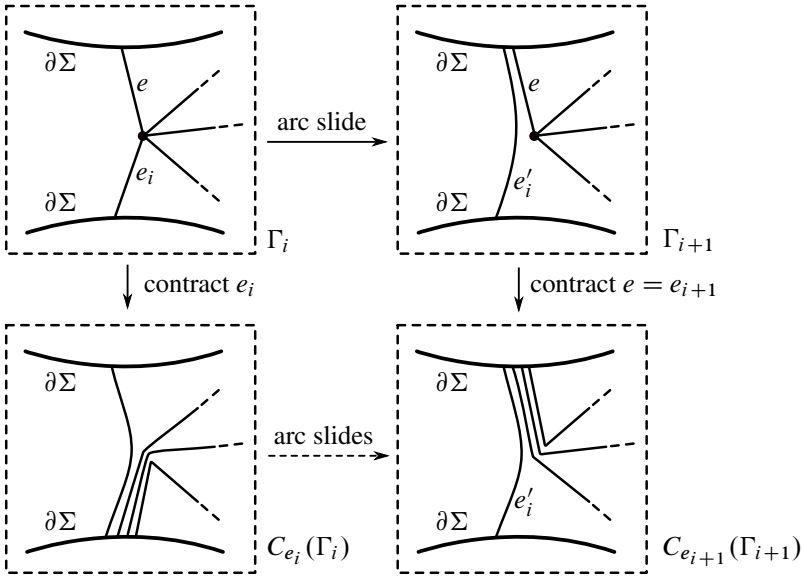


Figure 37: An illustration of the fact that $C_{e_{i+1}}(\Gamma_{i+1})$ is obtained from $C_{e_i}(\Gamma_i)$ by a sequence of arc slides when Γ_{i+1} is obtained from Γ_i by an arc slide of the special edge e_i across another edge e , and the resulting arc e'_i is no longer special.

multiple instances of move (G-3), or by no moves. Furthermore, we will show that Γ' satisfies (7-2).

We first consider the case when Γ_{i+1} is obtained from Γ_i by adding or removing an edge e via an instance of moves (G-1) or (G-2) or their inverses. If $e = e_i$ is removed, then, by definition, e_i is the last special edge (ie $i = l$), and clearly $C_{e_i}(\Gamma_i) = \Gamma_{i+1}$. If $e \neq e_i$, then $C_{e_{i+1}}(\Gamma_{i+1})$ is also obtained from $C_{e_i}(\Gamma_i)$ by adding or removing the edge e via a move of the same type as the one relating Γ_i and Γ_{i+1} .

We now consider the case when Γ_{i+1} is obtained from Γ_i by an instance of move (G-3). As indicated above, we break the argument into two cases. Case (1) occurs when we slide an edge e over another edge e' , and $e \neq e_i$. Since e_i is unchanged, it remains special. If $e_i \neq e'$, then it is straightforward to see that $C_{e_{i+1}}(\Gamma_{i+1})$ is also obtained from $C_{e_i}(\Gamma_i)$ by an arc slide. If $e_i = e'$, then, in fact, $C_{e_{i+1}}(\Gamma_{i+1})$ and $C_{e_i}(\Gamma_i)$ are isotopic.

We finally consider the most complicated case, Case (2), in which Γ_{i+1} is obtained by sliding e_i over another edge e . We write e'_i for the edge resulting from arc sliding e_i over e . There are two further subcases, depending on whether e'_i is special or not.

Suppose first that e'_i is not special, so $e_{i+1} = e$. By definition, either e'_i has both endpoints on $\partial\Sigma$, or e'_i has both endpoints on the same vertex. We claim that in both cases $C_{e_{i+1}}(\Gamma_{i+1})$ differs from $C_{e_i}(\Gamma_i)$ by a sequence of arc slides. See Figure 37 for an illustration when e'_i has both endpoints on $\partial\Sigma$. A similar argument holds when e'_i has both endpoints on the same vertex of Γ_{i+1} .

The other subcase of (2) occurs when e'_i is special. Write $\{v_1, v_2\} = \partial e_i$ and $\{v_2, v_3\} = \partial e$. Since e_i and e'_i are special, it follows that both $v_1 \neq v_2$ and $v_1 \neq v_3$. The argument can be further divided into the cases that $v_2 = v_3$ and $v_2 \neq v_3$. In both cases, a manipulation similar to the one shown in Figure 37 shows that $C_{e_{i+1}}(\Gamma_{i+1})$ is obtained from $C_{e_i}(\Gamma_i)$ by a sequence of allowable arc slides.

By the above, for $i \in \{k, \dots, l-1\}$, the graph $C_{e_{i+1}}(\Gamma_{i+1})$ is obtained from $C_{e_i}(\Gamma_i)$ using the same type of moves as the one relating Γ_{i+1} and Γ_i , and exactly one move is used in case of a move of type (G-1). Noting that $\Gamma_{k-1} = C_{e_k}(\Gamma_k)$ and $C_{e_l}(\Gamma_l) = \Gamma_{l+1}$, Equation (7-2) is immediate.

Repeating this procedure, starting with a fixed Γ , we can construct a Cerf sequence Γ'' such that

$$n_1(\Gamma'') = 0 \quad \text{and} \quad n_2(\Gamma'') = n_2(\Gamma).$$

By turning Morse functions corresponding to each decomposing graph upside down (switching the roles of the index 0 and 2 critical points) and repeating the above procedure to the induced dual decomposing graphs, we obtain a sequence of decomposing graphs from $\mathbf{a}$ to $\mathbf{a}'$ which differ only by a sequence of allowable arc slides, completing the proof. $\square$

7.4 Doubled diagrams

Suppose that $\mathcal{H} = (\Sigma, \alpha, \beta)$ is a sutured Heegaard diagram for the balanced sutured manifold (M, γ) . There is a special way of constructing a new diagram of (M, γ) from $\mathcal{H}$ which will be important for computing the cobordism map $F_{W_{\alpha, \beta, \gamma}}$. Let us first pick two collections of subintervals $\mathcal{I}_0$ and $\mathcal{I}_1$ of $\partial\Sigma$ such that each component of $\partial\Sigma$ contains exactly one subinterval from $\mathcal{I}_0$ and one subinterval from $\mathcal{I}_1$ that are disjoint.

We now define the *Heegaard surface that is doubled along R_β* to be

$$D_\beta(\Sigma) := \Sigma \natural_{\mathcal{I}_0} \bar{\Sigma} \natural_{\mathcal{I}_1} R_\beta,$$

where $\natural$ denotes the boundary connected sum operation. Here $\bar{\Sigma}$ denotes a push-off of Σ into U_β , with the opposite orientation.

Before we define the attaching curves, let us first describe how $D_{\beta}(\Sigma)$ is embedded in $M_{\alpha, \beta}$. A schematic is shown in Figure 38. Strictly speaking, we have changed the sutures. An isotopy supported in a neighborhood of the original sutures moves the new sutures to the original sutures.

We now describe compressing curves on $D_{\beta}(\Sigma)$. First, pick a collection of pairwise disjoint arcs on Σ with boundary on $\mathcal{I}_0$ that form a basis of $H_1(\Sigma, \mathcal{I}_0)$. One then doubles these across $\mathcal{I}_0$ and obtains a collection of curves $\Delta_{\Sigma} \subseteq \Sigma \natural_{\mathcal{I}_0} \bar{\Sigma} \subseteq D_{\beta}(\Sigma)$. Similarly, one picks a collection of pairwise disjoint arcs on R_{β} (or equivalently, a collection of arcs on Σ that avoid the β curves, up to handle slides across β) with boundary on $\mathcal{I}_1$ that form a basis of $H_1(R_{\beta}, \mathcal{I}_1)$. Doubling these curves across $\mathcal{I}_1$ yields closed curves $\delta_{\beta} \subseteq \bar{\Sigma} \natural_{\mathcal{I}_1} R_{\beta} \subseteq D_{\beta}(\Sigma)$.

Write $\bar{\beta}$ for the images of the β curves on $\bar{\Sigma}$. Note that the α , $\bar{\beta}$, and δ_{β} curves are all disjoint. Since (M, γ) is balanced, $|\alpha| = |\beta|$, and an easy computation shows that

$$|\alpha| + |\beta| + |\delta_{\beta}| = |\Delta_{\Sigma}|.$$

The doubled Heegaard diagram is now defined as

$$D_{\beta}(\mathcal{H}) = (D_{\beta}(\Sigma), D_{\beta}(\alpha), \Delta_{\Sigma}) = (\Sigma \natural_{\mathcal{I}_0} \bar{\Sigma} \natural_{\mathcal{I}_1} R_{\beta}, \alpha \cup \bar{\beta} \cup \delta_{\beta}, \Delta_{\Sigma}).$$

Remark 7.7 There is some asymmetry in the above construction, since we took Σ and connected it along $\mathcal{I}_0$ to a surface $\bar{\Sigma} \natural_{\mathcal{I}_1} R_{\beta}$ that was in the U_{β} handlebody. We could instead connect Σ along $\mathcal{I}_0$ to the surface $R_{\alpha} \natural_{\mathcal{I}_1} \bar{\Sigma}$ in the U_{α} handlebody and construct analogous attaching curves Δ_{Σ} and $\delta_{\alpha} \cup \bar{\alpha} \cup \beta$ for the Heegaard surface $D_{\alpha}(\Sigma) := R_{\alpha} \natural_{\mathcal{I}_1} \bar{\Sigma} \natural_{\mathcal{I}_0} \Sigma$. We will write

$$D_{\alpha}(\mathcal{H}) = (D_{\alpha}(\Sigma), \Delta_{\Sigma}, \delta_{\alpha} \cup \bar{\alpha} \cup \beta)$$

for this Heegaard diagram of (M, γ) . If there is any ambiguity, we will call the Heegaard diagram $D_{\beta}(\mathcal{H})$ the β -double and $D_{\alpha}(\mathcal{H})$ the α -double.

7.5 Weakly conjugated diagrams

Given a Heegaard diagram $(\Sigma, \alpha, \beta, w)$ of a based 3-manifold, we can consider the conjugate diagram $(\bar{\Sigma}, \beta, \alpha, w)$ that represents the same based 3-manifold. This was described by Ozsváth and Szabó [24] and was explored further by Hendricks and Manolescu [9]. Given a sutured diagram (Σ, α, β) for (M, γ) , one can consider the sutured diagram $(\bar{\Sigma}, \beta, \alpha)$; however, this is now a diagram for $(M, -\gamma)$. So, unlike in

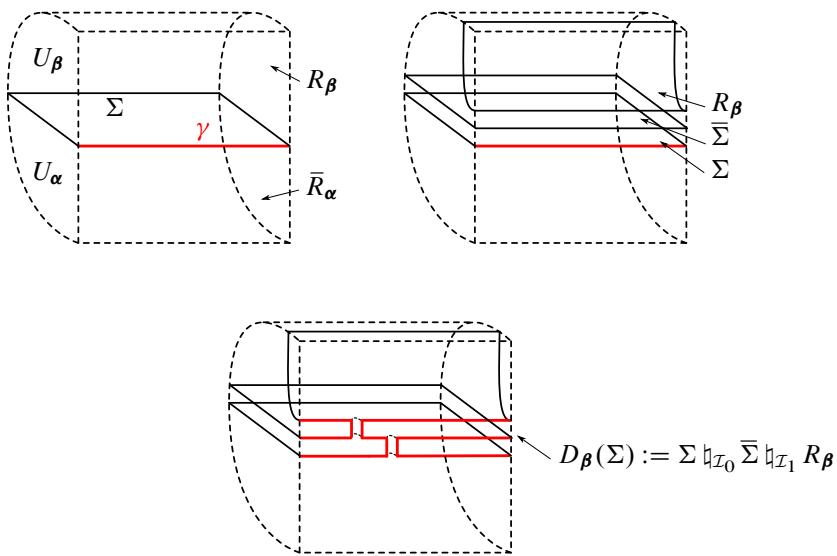


Figure 38: The doubled Heegaard surface $D_{\beta}(\Sigma)$. A neighborhood of a portion of the sutures γ in M is shown. The sutures are drawn in red. Strictly speaking, the new sutures $\partial D_{\beta}(\Sigma)$ are different from γ ; however, an isotopy supported in a neighborhood of γ moves $\partial D_{\beta}(\Sigma)$ back to γ .

the case of closed 3-manifolds, this operation does not induce a conjugation action on $\text{SFH}(M, \gamma)$.

Nonetheless, a similar diagrammatic manipulation appears when we compute the cobordism map for $\mathcal{W}_{\alpha, \beta, \gamma}$. In analogy to the terminology for the conjugation action

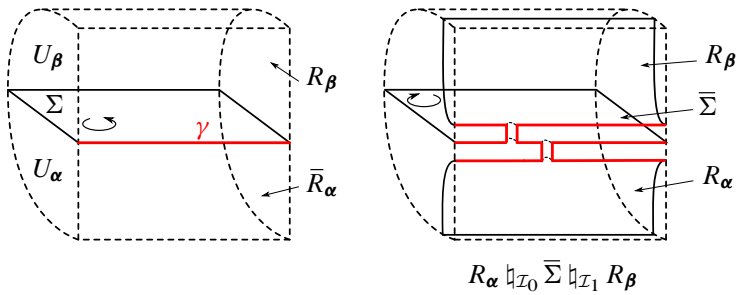


Figure 39: The weakly conjugated Heegaard surface $C(\Sigma)$. A neighborhood of a portion of the sutures is shown. The sutures are drawn in red. An isotopy supported in a neighborhood of the sutures moves the boundary of the new Heegaard surface, as we have drawn it, to the position of $\partial \Sigma$, the boundary of the old Heegaard surface.

on Heegaard diagrams for closed 3-manifolds, we will say that the diagrams that appear are *weakly conjugated*. We describe the construction of weakly conjugated diagrams in this section.

As with the doubled diagrams, we pick collections of subintervals $\mathcal{I}_0$ and $\mathcal{I}_1$ in $\partial\Sigma$ such that each component of $\partial\Sigma$ contains exactly one subinterval from $\mathcal{I}_0$ and from $\mathcal{I}_1$ that are disjoint. We can then form the weakly conjugated Heegaard surface

$$C(\Sigma) := R_\alpha \natural_{\mathcal{I}_0} \bar{\Sigma} \natural_{\mathcal{I}_1} R_\beta.$$

This is shown in Figure 39. As described, $\partial C(\Sigma)$ is different from $\partial\Sigma$, but an isotopy supported in a neighborhood of $\partial\Sigma$ moves $\partial C(\Sigma)$ to $\partial\Sigma$.

We now describe compressing curves on $C(\Sigma)$. Note that α and β still bound compressing disks on Σ . As curves on $\bar{\Sigma}$, we denote them by $\bar{\alpha}$ and $\bar{\beta}$. However, these are not complete collections of compressing curves, as we have increased the genus of the Heegaard surface by attaching R_α and R_β . Hence, we pick a collection of pairwise disjoint arcs on R_α that form a basis of $H_1(R_\alpha, \mathcal{I}_0)$ and double them across $\mathcal{I}_0$ to get a collection of curves δ_α on $R_\alpha \natural_{\mathcal{I}_0} \bar{\Sigma}$. Similarly, we pick a collection of pairwise disjoint arcs on R_β that form a basis of $H_1(R_\beta, \mathcal{I}_1)$ and double them across $\mathcal{I}_1$ to get a collection of curves δ_β on $\bar{\Sigma} \natural_{\mathcal{I}_1} R_\beta$. We define the weakly conjugated Heegaard diagram of $\mathcal{H}$ to be

$$C(\mathcal{H}) = (C(\Sigma), C(\beta), C(\alpha)) := (R_\alpha \natural_{\mathcal{I}_0} \bar{\Sigma} \natural_{\mathcal{I}_1} R_\beta, \bar{\beta} \cup \delta_\beta, \bar{\alpha} \cup \delta_\alpha).$$

Then $C(\mathcal{H})$ is also a diagram of (M, γ) .

7.6 The change of diagrams map from $D_\beta(\mathcal{H})$ to $\mathcal{H}$

In this section, we prove a relatively simple formula for the change of diagrams map from the β -double of a diagram $D_\beta(\mathcal{H})$ to the original diagram $\mathcal{H} = (\Sigma, \alpha, \beta)$. Recall that

$$D_\beta(\mathcal{H}) = (\Sigma \natural_{\mathcal{I}_0} \bar{\Sigma} \natural_{\mathcal{I}_1} R_\beta, \alpha \cup \bar{\beta} \cup \delta_\beta, \Delta_\Sigma).$$

We will write $\beta' := \bar{\beta} \cup \delta_\beta$. Note that $(D_\beta(\Sigma), \Delta_\Sigma, \beta \cup \beta')$ is the α -double of the diagram (Σ, β, β) (see Remark 7.7), which represents

$$(M_{\beta, \beta}, \gamma_{\beta, \beta}),$$

which we note is the sutured manifold obtained by gluing two copies of the sutured handlebody U_β together along Σ .

Lemma 7.8 *There is a relatively graded isomorphism*

$$\mathrm{SFH}(M_{\beta, \beta}, \gamma_{\beta, \beta}) \cong \bigotimes_{i=1}^{|\beta|} ((\mathbb{F}_2)_{\frac{1}{2}} \oplus (\mathbb{F}_2)_{-\frac{1}{2}}).$$

In particular, there is a top-graded element $[\Theta_{\Delta_{\Sigma, \beta \cup \beta'}}] \in \mathrm{SFH}(D_{\beta}(\mathcal{H}))$.

Proof Let the γ curves be small Hamiltonian translates of the β curves such that $|\beta_i \cap \gamma_j| = 2\delta_{ij}$. Then (Σ, β, γ) is an admissible diagram for $(M_{\beta, \beta}, \gamma_{\beta, \beta})$ whose sutured Floer complex contains $2^{|\beta|}$ generators. Furthermore, since every component of $\Sigma \setminus \beta$ contains a component of $\partial\Sigma$, it is straightforward to see that the only homology classes of disks on (Σ, β, γ) have domains which are supported in the small bigons between β_i and γ_i . From these considerations, the only index 1 holomorphic disks are the classes whose domain consists of a bigon connecting the higher graded point of $\beta_i \cap \gamma_i$ to the lower graded point. Hence, modulo 2, the differential on $\mathrm{CF}(\Sigma, \beta, \gamma)$ vanishes, and as a relatively graded group, we have

$$\mathrm{SFH}(\Sigma, \beta, \gamma) = \mathrm{CF}(\Sigma, \beta, \gamma) \cong \bigotimes_{i=1}^{|\beta|} ((\mathbb{F}_2)_{\frac{1}{2}} \oplus (\mathbb{F}_2)_{-\frac{1}{2}}),$$

completing the proof. □

Using the holomorphic triangle map and the class $[\Theta_{\Delta_{\Sigma, \beta \cup \beta'}}]$, we construct a map

$$F_2: \mathrm{SFH}(D_{\beta}(\Sigma), \alpha \cup \beta', \Delta_{\Sigma}) \rightarrow \mathrm{SFH}(D_{\beta}(\Sigma), \alpha \cup \beta', \beta \cup \beta')$$

using the formula

$$F_2 := F_{\alpha \cup \beta', \Delta_{\Sigma, \beta \cup \beta'}}(-, \Theta_{\Delta_{\Sigma, \beta \cup \beta'}}).$$

We can also define a 3-handle map

$$F_3 := F_3^{\beta', \beta'}: \mathrm{SFH}(D_{\beta}(\Sigma), \alpha \cup \beta', \beta \cup \beta') \rightarrow \mathrm{SFH}(\Sigma, \alpha, \beta)$$

(here, the second copy of β' is a small Hamiltonian translate of β' , though we omit this from the notation).

Lemma 7.9 *The composition $F_3 \circ F_2$ is chain homotopic to the change of diagrams map from $\mathrm{CF}(D_{\beta}(\mathcal{H}))$ to $\mathrm{CF}(\mathcal{H})$.*

Proof The idea is simple: The map F_2 can be interpreted as a composition of 2-handle maps, and the map F_3 can be interpreted as a composition of 3-handle maps, for a collection of 4-dimensional 2-handles and 3-handles that topologically cancel. Hence

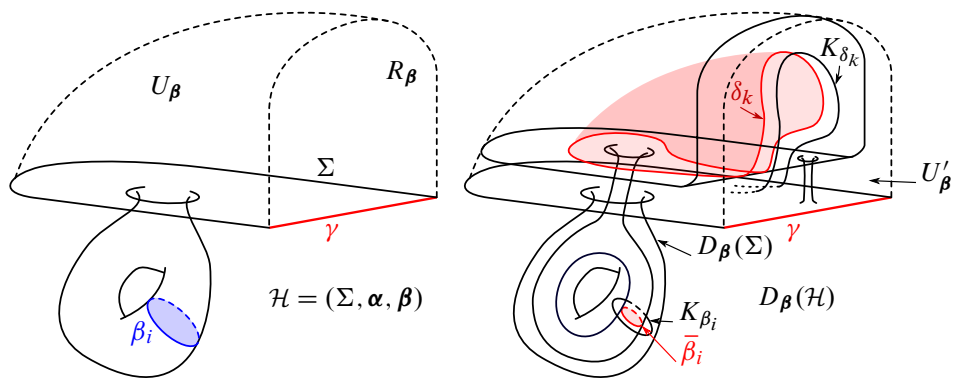


Figure 40: Constructing the link $\mathbb{L} \subseteq U_\beta$. The left shows a part of the handlebody U_β for a diagram (Σ, α, β) , together with a curve $\beta_i \in \beta$. The right depicts the β -double $D_\beta(\mathcal{H})$ of $\mathcal{H}$. The corresponding curve $\bar{\beta}_i \in \bar{\beta}$ and a curve $\delta_k \in \delta_\beta$ are shown, as well as the corresponding components K_{β_i} and a portion of K_{δ_k} of $\mathbb{L}$. The shaded regions denote compressing disks bounded by the curves β_i , $\bar{\beta}_i$, and δ_k .

their composition represents the transition map from naturality, by the well-definedness of the sutured cobordism maps [15].

We now explain the technical details. Let us first describe the framed link along which surgery induces F_2 . Let U'_β denote the β -handlebody of the diagram $D_\beta(\mathcal{H})$ (ie Δ_Σ bounds compressing disks in U'_β). For each curve in β and each curve in δ_β , we will construct a component of $\mathbb{L}$. For a curve $\beta_i \in \beta$, there is a knot $K_{\beta_i} \subseteq U'_\beta$, obtained by pushing β_i into U'_β slightly. We choose the framing of K_{β_i} to be parallel to Σ . Similarly, given $\delta_k \in \delta_\beta$, we can construct a framed knot K_{δ_k} by pushing δ_k into U'_β , and take the framing induced by the tangent space of the surface $D_\beta(\Sigma)$. The construction is illustrated in Figure 40. We define the framed link

$$\mathbb{L} := \bigcup_{\tau \in \beta \cup \delta_\beta} K_\tau.$$

The map F_3 is clearly the 3-handle map for a collection of $|\beta'|$ 2-spheres in $M(\mathbb{L})$ that topologically cancel the link $\mathbb{L}$. In Lemma 7.10, we will show that the triple $(D_\beta(\Sigma), \alpha \cup \beta', \Delta_\Sigma, \beta \cup \beta')$ can be related to a triple subordinate to a bouquet for $\mathbb{L}$ by a sequence of handle slides and isotopies of the Δ_Σ and $\beta \cup \beta'$ curves. Assuming this topological fact for the moment, we now show that this implies that F_2 is the 2-handle map for surgery on $\mathbb{L}$. Suppose $(D_\beta(\Sigma), \alpha \cup \beta', \xi, \zeta)$ is a Heegaard triple which is subordinate to a bouquet for $\mathbb{L}$, and ξ and ζ are related to Δ_Σ and $\beta \cup \beta'$,

respectively, by a sequence of handle slides and isotopies. To show that F_2 is the 2-handle map, it is sufficient to show that

$$(7-3) \quad F_{\alpha \cup \beta', \Delta_\Sigma, \beta \cup \beta'}(-, \Theta_{\Delta_\Sigma, \beta \cup \beta'}) \simeq \Psi_{\alpha \cup \beta'}^{\zeta \rightarrow \beta \cup \beta'} \circ F_{\alpha \cup \beta', \xi, \zeta}(\Psi_{\alpha \cup \beta'}^{\Delta_\Sigma \rightarrow \xi}(-), \Theta_{\xi, \zeta}),$$

the left-hand side being F_2 , and the right-hand side being the definition of the 2-handle map precomposed with the transition map for changing Δ_Σ to ξ , and postcomposed with the transition map for changing ζ to $\beta \cup \beta'$. We compute

$$\begin{aligned} (7-4) \quad & \Psi_{\alpha \cup \beta'}^{\zeta \rightarrow \beta \cup \beta'} \circ F_{\alpha \cup \beta', \xi, \zeta}(\Psi_{\alpha \cup \beta'}^{\Delta_\Sigma \rightarrow \xi}(-), \Theta_{\xi, \zeta}) \\ & \simeq \Psi_{\alpha \cup \beta'}^{\zeta \rightarrow \beta \cup \beta'} \circ F_{\alpha \cup \beta', \Delta_\Sigma, \zeta}(-, \Psi_{\xi \rightarrow \Delta_\Sigma}^\zeta(\Theta_{\xi, \zeta})) \\ & \simeq F_{\alpha \cup \beta', \Delta_\Sigma, \zeta}(-, \Psi_{\Delta_\Sigma}^{\zeta \rightarrow \beta \cup \beta'}(\Psi_{\xi \rightarrow \Delta_\Sigma}^\zeta(\Theta_{\xi, \zeta}))). \end{aligned}$$

The first chain homotopy follows from associativity for the 4-tuple $(\alpha \cup \beta', \Delta_\Sigma, \xi, \zeta)$, and the second follows from associativity applied to the 4-tuple $(\alpha \cup \beta', \Delta_\Sigma, \zeta, \beta \cup \beta')$. Using the fact that the top-graded generator is well defined on the level of homology, as in Lemma 7.8, it follows that $\Psi_{\Delta_\Sigma}^{\zeta \rightarrow \beta \cup \beta'}(\Psi_{\xi \rightarrow \Delta_\Sigma}^\zeta(\Theta_{\xi, \zeta}))$ and $\Theta_{\Delta_\Sigma, \beta \cup \beta'}$ differ by a boundary, so the last line of (7-4) is chain homotopic to F_2 , establishing (7-3).

Using the fact that F_2 is chain homotopic to the 2-handle map for $\mathbb{L}$, it follows that the composition $F_3 \circ F_2$ is chain homotopic to the map from naturality. $\square$

To finish the proof of Lemma 7.9, it is sufficient to show the following:

Lemma 7.10 *The Heegaard triple $(D_\beta(\Sigma), \alpha \cup \beta', \Delta_\Sigma, \beta \cup \beta')$ is related to a triple subordinate to the framed link $\mathbb{L}$ by a sequence of handle slides and isotopies of the Δ_Σ and $\beta \cup \beta'$ curves.*

Proof We recall that the Δ_Σ curves were constructed by picking a set of arcs $s = \{s_1, \dots, s_n\}$ forming a basis of $H_1(\Sigma, \mathcal{I}_0)$ and then doubling the arcs of s across $\mathcal{I}_0$ onto $\bar{\Sigma}$ to obtain a collection of $n := 2(g(\Sigma) + |\partial\Sigma| - |\Sigma|)$ closed curves. By Lemma 7.6, any two such bases s of $H_1(\Sigma, \mathcal{I}_0)$ can be connected by a sequence of allowable arc slides. An allowable arc slide of two arcs in s induces a handle slide of the corresponding doubled curves in Δ_Σ . Consequently, we can assume that Δ_Σ is constructed using any convenient basis of $H_1(\Sigma, \mathcal{I}_0)$.

The curves δ_β are obtained by picking a set of

$$k := \text{rk } H_1(R_\beta, \mathcal{I}_1) = 2(g(R_\beta) + |\partial\Sigma| - |R_\beta|)$$

arcs $d_1, \dots, d_k$ which form a basis of $H_1(R_\beta, \mathcal{I}_1)$. Noting that we view R_β as the result of surgering Σ on β , we can assume that Δ_Σ and δ_β are constructed using

arcs s and d satisfying the following:

- (1) For each $\beta \in \boldsymbol{\beta}$, there is a pair of arcs $s, s' \in s$ such that $|s \cap \beta| = 1$, and s is disjoint from all the other β and d curves. The curve s' is obtained by taking β and isotoping a neighborhood of the point $\beta \cap s$ along s until it intersects $\partial \Sigma$. In particular, s' is disjoint from $d \cup \boldsymbol{\beta}$. Furthermore, we can handle slide β across its image $\bar{\beta}$ on $\bar{\Sigma}$ along the arc s to obtain a curve isotopic to $s' \cup \bar{s}'$, where $\bar{s}'$ is the image of s' on $\bar{\Sigma}$.
- (2) For each $d \in \boldsymbol{d}$, there is a corresponding arc $s \in s$ satisfying $|s \cap d| = 1$. Furthermore, s is disjoint from $(\boldsymbol{d} \setminus \{d\}) \cup \boldsymbol{\beta}$.

See Figure 41 for an example of the arcs s and d , and the resulting attaching curves Δ_Σ and δ_β .

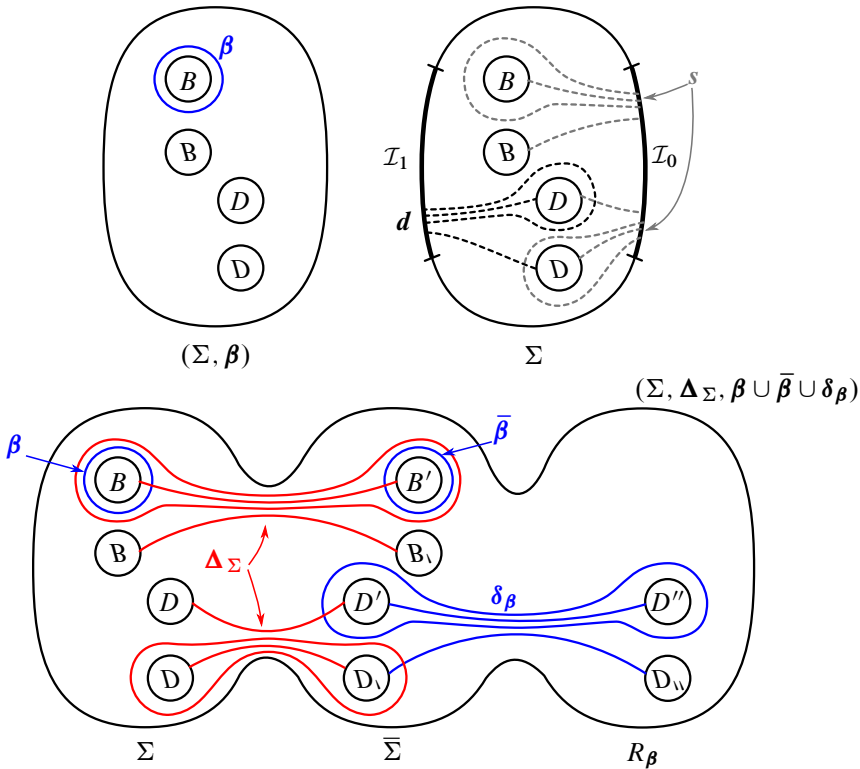


Figure 41: On the top left is an example of a monodisk $(\Sigma, \boldsymbol{\beta})$ with $|\boldsymbol{\beta}| = 1$, $g(\Sigma) = 2$, and $|\partial \Sigma| = 1$. On the top right are the arcs s used to form Δ_Σ , as well as the arcs d used to form δ_β . On the bottom is the diagram $(D_\beta(\Sigma), \Delta_\Sigma, \boldsymbol{\beta} \cup \bar{\boldsymbol{\beta}}')$.

We handle slide each $\beta \in \boldsymbol{\beta}$ across the corresponding $\bar{\beta} \in \bar{\boldsymbol{\beta}}$ along the arcs s to form the curve β^H . Let us call the resulting set of curves $\boldsymbol{\beta}^H$. We note that each curve in $\bar{\boldsymbol{\beta}} \cup \delta_{\boldsymbol{\beta}}$ is a longitude of a component of $\mathbb{L}$, and each curve in Δ_{Σ} is either a meridian or a small Hamiltonian isotopy of a curve in $\boldsymbol{\beta}^H$. It follows that the triple

$$(D_{\boldsymbol{\beta}}(\Sigma), \alpha \cup \bar{\boldsymbol{\beta}} \cup \delta_{\boldsymbol{\beta}}, \Delta_{\Sigma}, \boldsymbol{\beta}^H \cup \bar{\boldsymbol{\beta}} \cup \delta_{\boldsymbol{\beta}})$$

is subordinate to a bouquet for $\mathbb{L}$. □

7.7 The change of diagrams map from $C(\mathcal{H})$ to $D_{\boldsymbol{\beta}}(\mathcal{H})$

Analogously to the formula in Lemma 7.9, we can describe the change of diagrams map from $C(\mathcal{H})$ to $D_{\boldsymbol{\beta}}(\mathcal{H})$ in concrete terms.

Pick a collection of pairwise disjoint arcs on R_{α} with endpoints in $\mathcal{I}_0$ that form a basis of $H_1(R_{\alpha}, \mathcal{I}_0)$. These induce arcs on Σ which are disjoint from α and well defined up to handle slides across α ; let $\Delta_{\alpha} \subseteq \Sigma \setminus \mathcal{I}_0$ be the curves formed by doubling them across $\mathcal{I}_0$. By construction, the curves in Δ_{α} are disjoint from α and $\bar{\alpha}$. Also, let δ_{α} denote the images of the curves Δ_{α} on $R_{\alpha} \setminus \mathcal{I}_0$. Let $\alpha' := \bar{\alpha} \cup \Delta_{\alpha}$ and $\boldsymbol{\beta}' := \bar{\boldsymbol{\beta}} \cup \delta_{\boldsymbol{\beta}}$.

Recall the surface $D_{\boldsymbol{\beta}}(\Sigma)$ is $\Sigma \setminus \mathcal{I}_0 \setminus \mathcal{I}_1 R_{\boldsymbol{\beta}}$ and $C(\Sigma)$ is defined as $R_{\alpha} \setminus \mathcal{I}_0 \setminus \mathcal{I}_1 R_{\boldsymbol{\beta}}$. Since surgering the surface Σ along the α curves yields R_{α} , we see that there is a 1-handle map

$$G_1 := F_1^{\alpha, \alpha'}: \text{CF}(C(\mathcal{H})) = \text{CF}(C(\Sigma), \boldsymbol{\beta}', \bar{\alpha} \cup \delta_{\alpha}) \rightarrow \text{CF}(D_{\boldsymbol{\beta}}(\Sigma), \alpha \cup \boldsymbol{\beta}', \alpha \cup \alpha').$$

Lemma 7.11 *The diagram $(D_{\boldsymbol{\beta}}(\Sigma), \alpha \cup \alpha', \Delta_{\Sigma})$ represents the sutured manifold*

$$(S^1 \times S^2)^{\#n}(|\partial\Sigma|) \setminus \mathcal{I}_1 (R_{\boldsymbol{\beta}} \times I, \partial R_{\boldsymbol{\beta}} \times I)$$

for some n , where $(S^1 \times S^2)^{\#n}(|\partial\Sigma|)$ denotes the sutured manifold obtained by removing $|\partial\Sigma|$ balls from $(S^1 \times S^2)^{\#n}$ and adding a connected suture to each boundary component. Furthermore, $\setminus \mathcal{I}_1$ denotes the boundary connected sum taken along $\mathcal{I}_1$ by adding $|\mathcal{I}_1|$ product 1-handles.

Proof All the attaching curves are disjoint from $R_{\boldsymbol{\beta}}$. If we cut $R_{\boldsymbol{\beta}}$ off of $D_{\boldsymbol{\beta}}(\Sigma)$, we are left with the diagram $(\Sigma \setminus \mathcal{I}_0 \setminus \bar{\Sigma}, \alpha \cup \alpha', \Delta_{\Sigma})$. Recall that $\alpha' = \bar{\alpha} \cup \Delta_{\alpha}$, where Δ_{α} is obtained by choosing a basis of arcs for $H_1(R_{\alpha}, \mathcal{I}_0)$ and doubling the induced curves on Σ onto $\Sigma \setminus \mathcal{I}_0 \setminus \bar{\Sigma}$. We can assume that Δ_{Σ} is constructed by starting with small Hamiltonian isotopes of the curves in Δ_{α} and then adjoining an additional $2|\alpha|$ curves that are disjoint from Δ_{α} . Then the curves in Δ_{α} together with their isotopes

in Δ_Σ determine $|\Delta_\alpha|$ embedded 2–spheres, which may or may not be separating. After surgering these out, we are left with the double of a diagram for a disjoint union of some number of copies of $(S^1 \times S^2)^{\#l}(m)$ for various l and m . It follows that $(D_\beta(\Sigma), \alpha \cup \alpha', \Delta_\Sigma)$ represents

$$(S^1 \times S^2)^{\#n}(|\partial\Sigma|) \downarrow_{\mathbb{Z}_1} (R_\beta \times I, \partial R_\beta \times I)$$

for some n . □

As a consequence of [Lemma 7.11](#), there is a top-graded generator

$$\Theta_{\alpha \cup \alpha', \Delta_\Sigma} \in \text{SFH}(D_\beta(\Sigma), \alpha \cup \alpha', \Delta_\Sigma),$$

and hence we can also define a triangle map

$$G_2 := F_{\alpha \cup \beta', \alpha \cup \alpha', \Delta_\Sigma}(-, \Theta_{\alpha \cup \alpha', \Delta_\Sigma}).$$

Lemma 7.12 *The composition $G_2 \circ G_1$ is chain homotopic to the change of diagrams map from $C(\mathcal{H})$ to $D_\beta(\mathcal{H})$.*

Proof The proof is similar to the proof of [Lemma 7.9](#). We can interpret the composition $G_2 \circ G_1$ as the cobordism map for a canceling collection of $|\alpha|$ pairs of 4–dimensional 1– and 2–handles. We now describe the attaching spheres of the 1–handles and 2–handles (see also [\[29, Section 7.2\]](#) for a detailed account of a similar topological manipulation in the setting of closed 3–manifolds). Let $D_1, \dots, D_{|\alpha|}$ denote compressing disks attached along the curves $\alpha \subseteq \Sigma$, where $\Sigma \subseteq M$ denotes the original Heegaard surface.

The two feet of a 1–handle are obtained by pushing off the center point of the disk D_i in both normal directions. The canceling 2–handle for this 1–handle is attached along the core of the 1–handle concatenated with an arc connecting the two feet that intersects the center of D_i transversely. By adapting the proof of [Lemma 7.10](#), it is not hard to see that, after a sequence of handle slides, the triple $(D_\beta(\Sigma), \alpha \cup \beta', \alpha \cup \alpha', \Delta_\Sigma)$ becomes subordinate to the link described above. □

7.8 A diagram for $M_{\alpha, \beta} \cup_{R_\beta} M_{\beta, \gamma}$ and a formula for the special cobordism map $(W_{\alpha, \beta, \gamma})^s$

Let $\mathcal{T} = (\Sigma, \alpha, \beta, \gamma)$ be an admissible sutured triple diagram. We now describe a diagram $A(\mathcal{T})$ for the sutured manifold $(M_{\alpha, \beta} \cup_{R_\beta} M_{\beta, \gamma}, \gamma_{\alpha, \gamma})$, which has sutures along ∂R_β , where the two manifolds are glued together.

Analogous to the doubled diagram, we let $\mathcal{I}_0, \mathcal{I}_1 \subseteq \partial\Sigma$ be disjoint collections of subintervals such that each component of $\partial\Sigma$ contains exactly one subinterval from $\mathcal{I}_0$ and $\mathcal{I}_1$. We form the diagram

$$A(\mathcal{T}) = (D_{\gamma}(\Sigma), D_{\gamma}(\alpha), A(\beta)) = (\Sigma \natural_{\mathcal{I}_0} \bar{\Sigma} \natural_{\mathcal{I}_1} R_{\gamma}, \alpha \cup \bar{\gamma} \cup \delta_{\gamma}, \beta \cup \bar{\beta} \cup \Delta_{\beta}),$$

where the component Σ of $D_{\gamma}(\Sigma)$ is embedded in $M_{\alpha, \beta}$ and the component $\bar{\Sigma}$ in $M_{\beta, \gamma}$. We call $A(\mathcal{T})$ the *amalgamation of $\mathcal{T}$ along R_{β}* . Here $\delta_{\gamma} \subseteq \bar{\Sigma} \natural_{\mathcal{I}_1} R_{\gamma}$ is obtained by doubling a collection of arcs forming a basis of $H_1(R_{\gamma}, \mathcal{I}_1)$. Similarly, $\Delta_{\beta} \subseteq \Sigma \natural_{\mathcal{I}_0} \bar{\Sigma}$ is obtained by choosing a basis of arcs in $H_1(R_{\beta}, \mathcal{I}_0)$ and doubling a lift of them to Σ across $\mathcal{I}_0$.

Note that the doubling construction can be viewed as an instance of amalgamation, in the sense that

$$D_{\beta}(\Sigma, \alpha, \beta) = A(\Sigma, \alpha, \emptyset, \beta).$$

Here we interpret $R_{\emptyset}$ as Σ , so that Δ_{β} is the collection Δ_{Σ} defined in the construction of a doubled diagram.

Lemma 7.13 *The diagram $A(\mathcal{T})$ is a diagram for $(M_{\alpha, \beta} \cup_{R_{\beta}} M_{\beta, \gamma}, \gamma_{\alpha, \gamma})$.*

Proof The diagram $A(\mathcal{T})$ is obtained by first replacing R_{β} with Σ in

$$C(\mathcal{H}_{\beta, \gamma}) = (R_{\beta} \natural_{\mathcal{I}_0} \bar{\Sigma} \natural_{\mathcal{I}_1} R_{\gamma}, \bar{\gamma} \cup \delta_{\gamma}, \bar{\beta} \cup \delta_{\beta}),$$

after which δ_{β} becomes Δ_{β} . As compressing Σ along β gives R_{β} , if we add β , the resulting diagram

$$(\Sigma \natural_{\mathcal{I}_0} \bar{\Sigma} \natural_{\mathcal{I}_1} R_{\gamma}, \bar{\gamma} \cup \delta_{\gamma}, \beta \cup \bar{\beta} \cup \Delta_{\beta})$$

represents $(-U_{\beta} \cup_{R_{\beta}} U_{\beta} \cup_{\Sigma} -U_{\gamma}, \partial\Sigma)$. Finally, we add α , which amounts to gluing U_{α} to $-U_{\beta} \cup_{R_{\beta}} U_{\beta} \cup_{\Sigma} -U_{\gamma}$ by identifying $\Sigma \subseteq U_{\alpha}$ with $\bar{\Sigma} \subseteq \partial(-U_{\beta})$. $\square$

By Lemma 7.4, the special cobordism

$$(W_{\alpha, \beta, \gamma})^S: (M_{\alpha, \beta} \cup_{R_{\beta}} M_{\beta, \gamma}, \gamma_{\alpha, \gamma}) \rightarrow (M_{\alpha, \gamma}, \gamma_{\alpha, \gamma})$$

is a 2-handle cobordism, for surgery on a framed link $\mathbb{L} \subseteq M_{\alpha, \beta} \cup_{R_{\beta}} M_{\beta, \gamma}$. We recall the description of the framed link $\mathbb{L}$. One takes a Morse function f_{β} on U_{β} that is 0 on $\bar{R}_{\beta}$ and 1 on Σ and has β as the intersection of the ascending manifolds of the critical points of f_{β} with Σ . Then the descending manifolds of the critical points of f_{β} determine a collection of $|\beta|$ properly embedded arcs $\lambda_i \subseteq U_{\beta}$ that have both ends on R_{β} . The link $\mathbb{L}$ is obtained by taking the union of the arcs $\lambda_i \subseteq U_{\beta} \subseteq M_{\beta, \gamma}$, together with their images in $-U_{\beta} \subseteq M_{\alpha, \beta}$.

Let $\Delta_\Sigma \subseteq \Sigma \natural_{\mathcal{I}_0} \bar{\Sigma}$ be a collection of curves obtained by picking arcs forming a basis of $H_1(\Sigma, \mathcal{I}_0)$ and then doubling them across $\mathcal{I}_0$. One can assume that the chosen basis of $H_1(\Sigma, \mathcal{I}_0)$ extends the lift of the basis of $H_1(R_\beta, \mathcal{I}_0)$ to Σ we chose in the construction of the Δ_β curves, so that $\Delta_\beta \subseteq \Delta_\Sigma$, though this is not essential.

We note that the diagram

$$(D_\gamma(\Sigma), A(\beta), \Delta_\Sigma) = (\Sigma \natural_{\mathcal{I}_0} \bar{\Sigma} \natural_{\mathcal{I}_1} R_\gamma, \beta \cup \bar{\beta} \cup \Delta_\beta, \Delta_\Sigma)$$

represents

$$(S^1 \times S^2)^{\#n}(|\partial\Sigma|) \natural_{\mathcal{I}_1} (I \times R_\gamma, I \times \partial R_\gamma)$$

for some n by Lemma 7.11. Consequently, there is a top-graded generator

$$\Theta_{A(\beta), \Delta_\Sigma} \in \text{SFH}(D_\gamma(\Sigma), A(\beta), \Delta_\Sigma) = \text{SFH}(\Sigma \natural_{\mathcal{I}_0} \bar{\Sigma} \natural_{\mathcal{I}_1} R_\gamma, \beta \cup \bar{\beta} \cup \Delta_\beta, \Delta_\Sigma).$$

Note that $(D_\gamma(\Sigma), D_\gamma(\alpha), \Delta_\Sigma)$ is a doubled diagram for $(M_{\alpha, \gamma}, \gamma_{\alpha, \gamma})$. We have the following:

Lemma 7.14 *The special cobordism map*

$$F_{(W_{\alpha, \beta, \gamma})^s}: \text{CF}(M_{\alpha, \beta} \cup_{R_\beta} M_{\beta, \gamma}, \gamma_{\alpha, \gamma}) \rightarrow \text{CF}(M_{\alpha, \gamma}, \gamma_{\alpha, \gamma})$$

is chain homotopic to the triangle map

$$F_{D_\gamma(\alpha), A(\beta), \Delta_\Sigma}(-, \Theta_{A(\beta), \Delta_\Sigma}).$$

Proof By Lemma 7.4, the special cobordism $(W_{\alpha, \beta, \gamma})^s$ is a 2-handle cobordism, for a framed link $\mathbb{L} \subseteq -U_\beta \cup_{R_\beta} U_\beta$ described above. By adapting the proof of Lemma 7.10, we see that the triple

$$(D_\gamma(\Sigma), D_\gamma(\alpha), A(\beta), \Delta_\Sigma)$$

can be related by a sequence of handle slides and isotopies to a triple that is subordinate to a bouquet for $\mathbb{L}$. Thus, $F_{(W_{\alpha, \beta, \gamma})^s}$ is chain homotopic to the triangle map above. $\square$

7.9 A holomorphic triangle description of the gluing map

Let $\mathcal{T} = (\Sigma, \alpha, \beta, \gamma)$ be a sutured Heegaard triple. In this section, we will present a natural candidate for the map for gluing

$$(Z_{\alpha, \beta, \gamma}, -\xi_{\alpha, \beta, \gamma})$$

to

$$(-M_{\alpha, \beta}, -\gamma_{\alpha, \beta}) \sqcup (-M_{\beta, \gamma}, -\gamma_{\beta, \gamma})$$

and prove that it is indeed the gluing map.

Let

$$C(\mathcal{H}_{\beta, \gamma}) = (R_{\beta} \natural \bar{\Sigma} \natural R_{\gamma}, \bar{\gamma} \cup \delta_{\gamma}, \bar{\beta} \cup \delta_{\beta})$$

be a weak conjugate of $\mathcal{H}_{\beta, \gamma} = (\Sigma, \beta, \gamma)$, as described in [Section 7.5](#). We now construct our candidate

$$\Phi: \mathrm{CF}(\mathcal{H}_{\alpha, \beta}) \otimes \mathrm{CF}(C(\mathcal{H}_{\beta, \gamma})) \rightarrow \mathrm{CF}(A(\mathcal{T}))$$

for the gluing map. Note that the domain of Φ is $\mathrm{CF}(M_{\alpha, \beta}, \gamma_{\alpha, \beta}) \otimes \mathrm{CF}(M_{\beta, \gamma}, \gamma_{\beta, \gamma})$, and its range is $\mathrm{CF}(M_{\alpha, \beta} \cup_{R_{\beta}} M_{\beta, \gamma}, \gamma_{\alpha, \gamma})$.

The definition of the map Φ is formally similar to the map for connected sums due to Ozsváth and Szabó [\[23\]](#). We will call Φ the *amalgamation map*, since its image is in the Floer homology of the amalgamated diagram from the previous section. We also remark that the Ozsváth–Szabó maps that inspire the construction of Φ were called the *intertwining maps* in [\[29\]](#), where they played a similar role as in our present context.

Note that there is a 1–handle map

$$F_1^{\bar{\gamma} \cup \delta_{\gamma}, \bar{\gamma} \cup \delta_{\gamma}}: \mathrm{CF}(\mathcal{H}_{\alpha, \beta}) \rightarrow \mathrm{CF}(D_{\gamma}(\Sigma), D_{\gamma}(\alpha), D_{\gamma}(\beta)),$$

where $D_{\gamma}(\alpha) = \alpha \cup \bar{\gamma} \cup \delta_{\gamma}$ and $D_{\gamma}(\beta) = \beta \cup \bar{\gamma} \cup \delta_{\gamma}$.

Recall that the underlying Heegaard surface of $C(\mathcal{H}_{\beta, \gamma})$ is $R_{\beta} \natural \bar{\Sigma} \natural R_{\gamma}$, and the Heegaard surface $D_{\gamma}(\Sigma)$ is $\Sigma \natural \bar{\Sigma} \natural R_{\gamma}$. Because the surface R_{β} is the result of surgering Σ along the β curves, it follows that there is a 1–handle map

$$F_1^{\beta, \beta}: \mathrm{CF}(C(\mathcal{H}_{\beta, \gamma})) \rightarrow \mathrm{CF}(D_{\gamma}(\Sigma), D_{\gamma}(\beta), A(\beta)),$$

obtained by adding the β curves back into $R_{\beta} \subseteq C(\Sigma)$. Note that, after we add these 1–handles, turning R_{β} into Σ , the curves δ_{β} become Δ_{β} .

Finally, we define

$$\Phi := F_{D_{\gamma}(\alpha), D_{\gamma}(\beta), A(\beta)} \circ (F_1^{\bar{\gamma} \cup \delta_{\gamma}, \bar{\gamma} \cup \delta_{\gamma}} \otimes F_1^{\beta, \beta}).$$

A key ingredient in our analysis of the sutured cobordism $\mathcal{W}_{\alpha, \beta, \gamma}$ is the following:

Proposition 7.15 *The amalgamation map*

$$\Phi: \mathrm{SFH}(M_{\alpha, \beta}, \gamma_{\alpha, \beta}) \otimes \mathrm{SFH}(M_{\beta, \gamma}, \gamma_{\beta, \gamma}) \rightarrow \mathrm{SFH}(M_{\alpha, \beta} \cup_{R_{\beta}} M_{\beta, \gamma}, \gamma_{\alpha, \gamma})$$

defined above is chain homotopic to the contact gluing map for gluing $(Z_{\alpha, \beta, \gamma}, -\xi_{\alpha, \beta, \gamma})$ to

$$(-M_{\alpha, \beta}, -\gamma_{\alpha, \beta}) \sqcup (-M_{\beta, \gamma}, -\gamma_{\beta, \gamma}).$$

Proof First, we describe a contact handle decomposition of $(Z_{\alpha,\beta,\gamma}, -\xi_{\alpha,\beta,\gamma})$, relative to $R_{\alpha,\beta} \sqcup R_{\beta,\gamma}$. We pick a collection of subintervals $\mathcal{I}_0 \subseteq \partial R_{\beta}$ such that each component of ∂R_{β} contains exactly one subinterval. We pick an arc basis $\lambda_1, \dots, \lambda_n$ for $(R_{\beta}, \mathcal{I}_0)$. We claim that a contact handle decomposition of $(Z_{\alpha,\beta,\gamma}, \xi_{\alpha,\beta,\gamma})$, relative to $R_{\alpha,\beta} \sqcup R_{\beta,\gamma}$, can be constructed as follows:

- (1) The contact 1–handles are the components of $N(I \times \mathcal{I}_0)$ (in particular, we have one 1–handle for each component of ∂R_{β}).
- (2) The contact 2–handles are $N(I \times \lambda_i)$ for $i \in \{1, \dots, n\}$.

The 1–handles are simply added with feet along the corresponding subintervals of $\mathcal{I}_0 \subseteq R_{\beta} \subseteq \partial M_{\alpha,\beta}$ and $\bar{\mathcal{I}}_0 \subseteq \bar{R}_{\beta} \subseteq \partial M_{\beta,\gamma}$. The 2–handles are attached along curves obtained by concatenating an arc λ_i on R_{β} with its mirror on $\bar{R}_{\beta}$.

To show that the above description determines a contact handle decomposition, we use [Lemma 7.3](#), which allows us to decompose $Z_{\alpha,\beta,\gamma}$ along the convex surface obtained by extending $\{\frac{1}{2}\} \times R_{\beta} \subseteq Z_{\beta}$ across the solid tori Z_0 to get a decomposing surface which is diffeomorphic to R_{β} . See [Figures 33](#) and [34](#) for schematics of the decomposing surface. Phrased another way, we can write $(Z_{\alpha,\beta,\gamma}, -\xi_{\alpha,\beta,\gamma})$ as $(I \times R_{\alpha,\beta}, -\xi_{\alpha,\beta})$ and $(I \times R_{\beta,\gamma}, -\xi_{\beta,\gamma})$ glued together along two subsurfaces $\bar{R}'_{\beta} \subseteq \{0\} \times R_{\alpha,\beta}$ and $R'_{\beta} \subseteq \{0\} \times R_{\beta,\gamma}$ that are small perturbations of $\bar{R}_{\beta} \subseteq \{0\} \times \bar{R}_{\alpha,\beta}$ and $R_{\beta} \subseteq \{0\} \times R_{\beta,\gamma}$, respectively. We pick R'_{β} and $\bar{R}'_{\beta}$ such that $\mathcal{I}_0 \subseteq R'_{\beta}$ and $\bar{\mathcal{I}}_0 \subseteq \bar{R}'_{\beta}$. We identify R'_{β} and $\bar{R}'_{\beta}$ with their images on $\partial M_{\alpha,\beta}$ and $\partial M_{\beta,\gamma}$, respectively.

The above description implies that the cores of the 1–handles in our contact handle decomposition are Legendrian, and the attaching circles of the 2–handles cross the dividing set exactly twice. Using Legendrian realization after attaching the contact 1–handles, we can assume the attaching curves of the 2–handles are Legendrian, and hence have $\text{tb} = -1$. It follows that the above description determines a contact handle decomposition of $Z_{\alpha,\beta,\gamma}$, starting at $R_{\alpha,\beta} \sqcup R_{\beta,\gamma}$ and ending at $R_{\alpha,\gamma}$. Using the above contact handle decomposition of $(Z_{\alpha,\beta,\gamma}, -\xi_{\alpha,\beta,\gamma})$, we can now give a description of the gluing map associated to $(Z_{\alpha,\beta,\gamma}, -\xi_{\alpha,\beta,\gamma})$ on the level of Heegaard diagrams.

The contact 1–handle maps have a simple description in terms of diagrams; one simply adds a band with ends at the feet of the 1–handle. There are two dividing arcs on the boundary of the band, and they are distinguished: One intersects $I \times R_{\beta} \subseteq Z_{\alpha,\beta,\gamma}$ nontrivially, while the other is disjoint from $I \times R_{\beta}$; see [Figure 42](#). Let us call the edge that intersects $I \times R_{\beta}$ the “interior” dividing arc of the 1–handle. The other edge

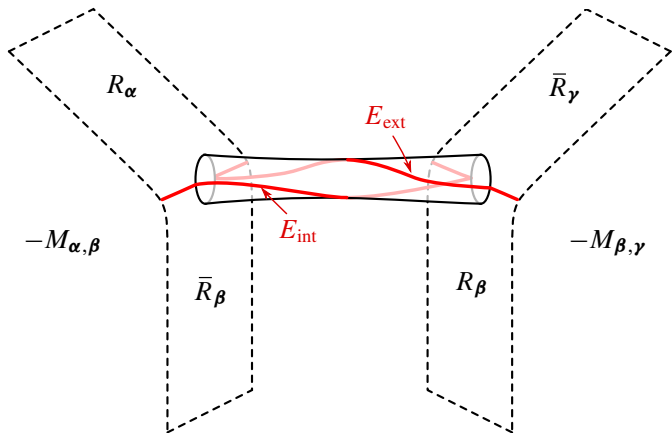


Figure 42: A contact 1–handle added to $-M_{\alpha,\beta} \sqcup -M_{\beta,\gamma}$. On the 1–handle, the interior edge of the dividing set (in E_{int}) is contained in $I \times R_\beta$, and the exterior edge (in E_{ext}) is disjoint from $I \times R_\beta$.

we call the “exterior” dividing arc. We will write E_{int} for the interior arcs, and E_{ext} for the exterior arcs.

Note that the attaching circles of the 2–handles only intersect the dividing set along the contact 1–handles, and only along the interior arcs E_{int} . Hence, on the level of diagrams, the contact 2–handle map involves adding a band to E_{int} and also adding an α –curve and a β –curve.

We will write Σ_0 for the new portion of the Heegaard surface which is added by the contact 1–handles and 2–handles. We will shortly see that Σ_0 can be identified with $\bar{R}_\beta$; see (7-5). We call the collection of β –curves that we add with the contact 2–handles $\bar{\delta}_\beta$ (the notation will be justified below; see (7-5)), and we call the α –curves that we add ϵ .

By construction, the diagram for the domain of the amalgamation map is the disjoint union of $\mathcal{H}_{\alpha,\beta} = (\Sigma, \alpha, \beta)$ and a weak conjugate

$$C(\mathcal{H}_{\beta,\gamma}) = (R_\beta \natural \bar{\Sigma} \natural R_\gamma, \bar{\gamma} \cup \delta_\gamma, \bar{\beta} \cup \delta_\beta)$$

of $\mathcal{H}_{\beta,\gamma} = (\Sigma, \beta, \gamma)$. After adding all the 1–handles and 2–handles, we obtain the Heegaard diagram

$$G(\mathcal{T}) = (\Sigma(\mathcal{T}), \alpha(\mathcal{T}), \beta(\mathcal{T})) := (\Sigma \natural \Sigma_0 \natural R_\beta \natural \bar{\Sigma} \natural R_\gamma, \alpha \cup \epsilon \cup \bar{\gamma} \cup \delta_\gamma, \beta \cup \bar{\delta}_\beta \cup \bar{\beta} \cup \delta_\beta)$$

for $M_{\alpha,\beta} \cup_{R_\beta} M_{\beta,\gamma}$. Note that the diagram $G(\mathcal{T})$ is similar to, but not equal to, the Heegaard diagram

$$A(\mathcal{T}) = (\Sigma \natural \bar{\Sigma} \natural R_\gamma, \alpha \cup \bar{\gamma} \cup \delta_\gamma, \beta \cup \bar{\beta} \cup \Delta_\beta)$$

of $M_{\alpha,\beta} \cup_{R_\beta} M_{\beta,\gamma}$ that appears in the target of the map in Proposition 7.15. We now describe the curves δ_β , ϵ , and $\bar{\delta}_\beta$ and the surface Σ_0 more explicitly. For each contact 1-handle, there is a corresponding 0-handle of the surface Σ_0 . For each contact 2-handle, we attach a 1-handle to Σ_0 , along the interior edges E_{int} of the portion of Σ_0 built when attaching the contact 1-handles. For each contact 2-handle, we also add a curve in ϵ and a curve in $\bar{\delta}_\beta$.

Let $c_i \subseteq \Sigma_0$ be the core of the band of the contact 2-handle associated to the arc λ_i on R_β . We extend the curves c_i across the bands of the 1-handles so that they have both ends on E_{ext} ; see the top left and top right of Figure 43. If we isotope each c_i near E_{ext} so that its ends are in $\mathcal{I}_0$ and then double it across $\mathcal{I}_0$ onto $R_\beta \subseteq -M_{\beta,\gamma}$, in a sense specified in (7-6), then we get the ϵ curves. If we isotope each c_i near E_{ext} in the opposite direction until its ends are in $\bar{\mathcal{I}}_0$, and then we double it across $\bar{\mathcal{I}}_0$ onto $\Sigma \subseteq -M_{\alpha,\beta}$, in a sense specified in (7-7), then we get the $\bar{\delta}_\beta$ curves. Examples of the curves $\bar{\delta}_\beta$ and ϵ on $\bar{R}_\beta$ are shown in Figure 43.

Let $\mathcal{I}_1$ denote the collection of subarcs of ∂R_β that are used to connect R_β to $\bar{\Sigma}$ in the weakly conjugated diagram $C(\mathcal{H}_{\beta,\gamma})$. Since the curves ϵ were added by the contact 2-handles, and the 2-handles have attaching circles constructed using the basis of arcs $\lambda_1, \dots, \lambda_n$ for $(R_\beta, \mathcal{I}_0)$, the arcs $\epsilon \cap R_\beta$ cut R_β into a collection of disks, each of which contains a single component of $\mathcal{I}_1$. Similarly, since $\epsilon \cap \Sigma_0$ are the cores of the 1-handles used to build Σ_0 , the arcs $\epsilon \cap \Sigma_0$ cut Σ_0 into a collection of disks, each of which contains exactly one arc of $\bar{\mathcal{I}}_0$ along its boundary. Consequently, there is an orientation-reversing diffeomorphism

$$(7-5) \quad \phi: \Sigma_0 \rightarrow R_\beta,$$

specified up to isotopy by the property that

$$(7-6) \quad \phi(\epsilon \cap \Sigma_0) = \epsilon \cap R_\beta$$

and also that $\mathcal{I}_0 \subseteq \Sigma_0$ is mapped to $\mathcal{I}_0 \subseteq R_\beta$, and $\bar{\mathcal{I}}_0 \subseteq \Sigma_0$ is mapped to $\mathcal{I}_1 \subseteq R_\beta$.

Since we have freedom to assume that the curves δ_β are constructed with any convenient basis of $H_1(R_\beta, \mathcal{I}_1)$, we can assume that they are constructed using the images of the

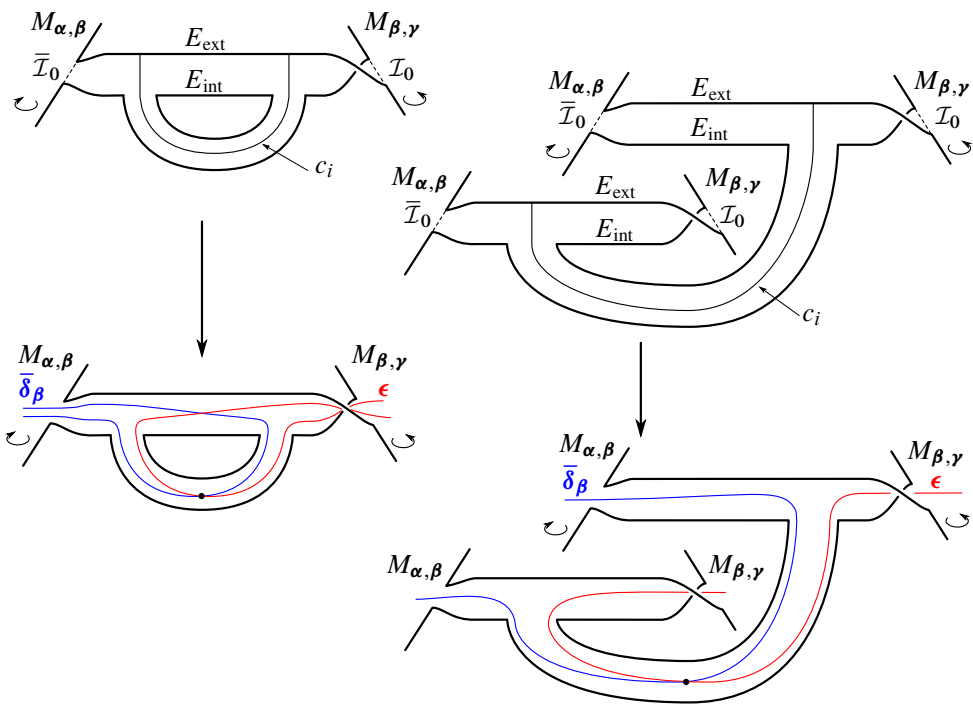


Figure 43: The effect on the Heegaard diagram of attaching a single contact 2–handle in our decomposition of $Z_{\alpha,\beta,\gamma}$. The regions shown consist of a band for a contact 2–handle with both ends attached to a single contact 1–handle (left) or a pair of contact 1–handles (right). Viewing the band from the contact 2–handle as a 1–handle added to the Heegaard surface, the core c_i is shown in the top row. Isotoping the ends of c_i in a neighborhood of $E_{\text{ext}} \cup \mathcal{I}_0$ so they lie in $\mathcal{I}_0$, and then doubling across $\mathcal{I}_0$, we get ϵ . Isotoping the ends of c_i in a neighborhood of $E_{\text{ext}} \cup \bar{\mathcal{I}}_0$ so they lie in $\bar{\mathcal{I}}_0$, we get $\bar{\delta}_\beta$. The orientations of the Heegaard surfaces for $M_{\alpha,\beta}$ and $M_{\beta,\gamma}$ are shown.

arcs $\bar{\delta}_\beta \cap \Sigma_0$ under ϕ . Consequently, we assume

(7-7)
$$\phi(\bar{\delta}_\beta \cap \Sigma_0) = \delta_\beta \cap R_\beta.$$

In light of (7-5), we will henceforth write $\bar{R}_\beta$ for Σ_0 .

The next step to understanding the gluing map is to destabilize the region $\bar{R}_\beta \natural R_\beta$. Unfortunately, the curves $\bar{\delta}_\beta$ and ϵ are not suitable for this (even if we use the compound destabilization operation from Section 2.2). In order to present the curves in a reasonable manner, we need to do some handle slides amongst the δ_β , $\bar{\delta}_\beta$, and ϵ curves, as we now describe.

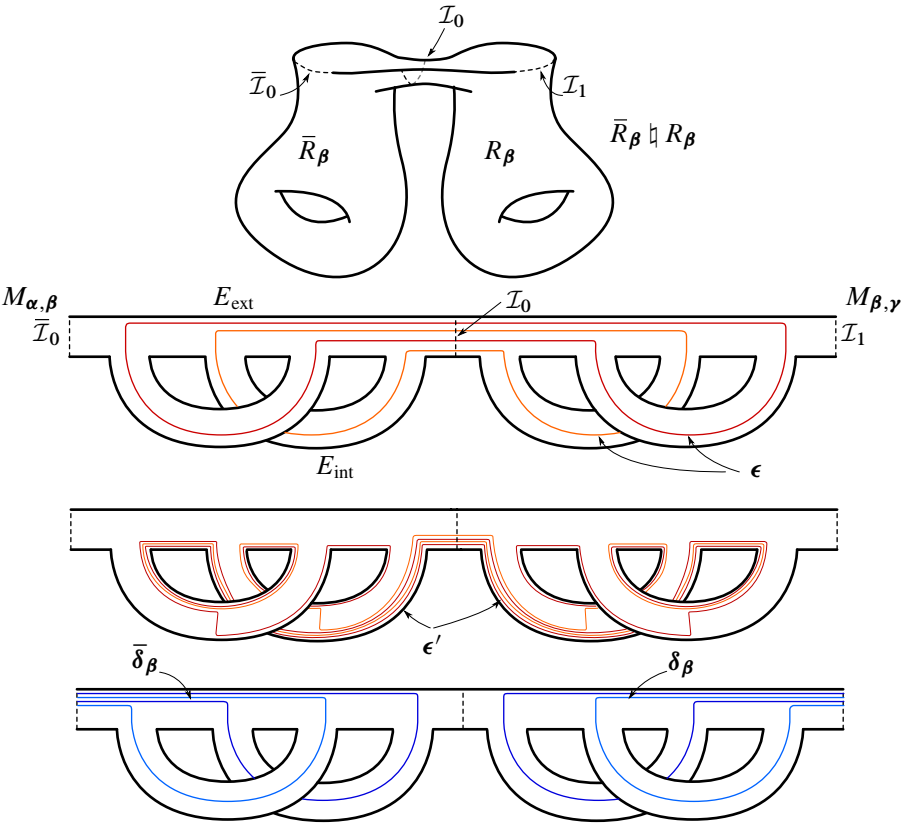


Figure 44: The portion of the Heegaard surface identified with $\bar{R}_\beta \natural R_\beta$. In this case, R_β is a genus 1 surface with one boundary component. The curves ϵ are shown in the second row. The curves ϵ' are shown in the third. The curves δ_β and $\bar{\delta}_\beta$ are shown in the last row.

We perform two moves. First, we modify the ϵ curves, as follows. Recall that we obtained $\bar{\delta}_\beta$ and ϵ by isotoping the cores of the 2-dimensional 1-handles near the E_{ext} boundary arcs of the contact 1-handles. Let us write c_i^* for the cocore of the handle with core c_i . By construction, $|c_i^* \cap c_j| = \delta_{ij}$.

Since the bands associated to the contact 2-handles are all attached to E_{int} , we can subdivide each component of $\partial \bar{R}_\beta$ into four subarcs, which we label as $\mathcal{I}_0$, E_{ext} , $\bar{\mathcal{I}}_0$, or E_{int} , in a way which is compatible with the analogous designation of the boundaries of the contact 1-handle bands.

We now isotope each c_i^* in a neighborhood of $E_{\text{int}} \cup \mathcal{I}_0$ such that its ends lie in $\mathcal{I}_0$. Let ϵ' denote the closed curves obtained by doubling the resulting curves across $\mathcal{I}_0$

onto R_β . Note that we perform this isotopy after *all* the 2–handle bands have been added (not after just the corresponding 2–handle has been attached). An example is shown in [Figure 44](#).

Since

$$|(\delta_\beta)_i \cap \epsilon'_j| = |(\bar{\delta}_\beta)_i \cap \epsilon'_j| = \delta_{ij},$$

we can handle slide δ_β over $\bar{\delta}_\beta$ along ϵ' in such a way that the resulting curves $\hat{\delta}_\beta$ do not intersect the ϵ' curves. With this configuration, we note that ϵ' intersects only $\bar{\delta}_\beta$, and, furthermore, the two sets of curves come in pairs. A sequence of destabilizations can then be used to surger out the ϵ' curves, while removing the $\bar{\delta}_\beta$ curves. Once we do this, we are left with the diagram $A(\mathcal{T})$ of $M_{\alpha,\beta} \cup_{R_\beta} M_{\beta,\gamma}$ described in [Section 7.8](#).

Note that ϵ and ϵ' are both obtained by picking a set of arcs which form a basis of $H_1(\bar{R}_\beta, \mathcal{I}_0)$ and then doubling them onto $\bar{R}_\beta \natural R_\beta$. By [Lemma 7.6](#), two such collections of arcs can be connected by a sequence of allowable arc slides ([Definition 7.5](#)). An allowable arc slide with respect to $\mathcal{I}_0$ induces a handle slide of the corresponding curves obtained by doubling across $\mathcal{I}_0$. Consequently, it follows that ϵ and ϵ' can be connected by a sequence of handle slides on $\bar{R}_\beta \natural R_\beta$, though the particular sequence of handle slides is not of importance for us.

We now describe the effect of these Heegaard moves on the sutured Floer complexes and relate them to the desired triangle map formula. We write

$$C^{\epsilon, \bar{\delta}_\beta}: \text{CF}(\mathcal{H}_{\alpha,\beta}) \otimes \text{CF}(C(\mathcal{H}_{\beta,\gamma})) \rightarrow \text{CF}(G(\mathcal{T}))$$

for the contact gluing map, obtained by adding the surface $\bar{R}_\beta$ to $\Sigma \sqcup (R_\beta \natural \bar{\Sigma} \natural R_\gamma)$ to get $\Sigma(\mathcal{T}) = \Sigma \natural \bar{R}_\beta \natural R_\beta \natural \bar{\Sigma} \natural R_\gamma$ and then adding in the curves ϵ and $\bar{\delta}_\beta$. If $x \in \text{CF}(\mathcal{H}_{\alpha,\beta})$ and $y \in \text{CF}(C(\mathcal{H}_{\beta,\gamma}))$, the map $C^{\epsilon, \bar{\delta}_\beta}$ is defined by the formula $x \times y \mapsto x \times c \times y$, where c is the distinguished intersection point of $\mathbb{T}_\epsilon \cap \mathbb{T}_{\bar{\delta}_\beta}$.

On the other hand, we can precompose the map $C^{\epsilon, \bar{\delta}_\beta}$ with a transition map for a small isotopy of the curves β on $\mathcal{H}_{\alpha,\beta}$ and of the curves $\bar{\gamma} \cup \delta_\gamma$ on $C(\mathcal{H}_{\beta,\gamma})$. By a small abuse of notation, we also write β and $\bar{\gamma} \cup \delta_\gamma$ for the translates. Hence

$$(7-8) \quad \Phi_{-\xi}(x \times y) := C^{\epsilon, \bar{\delta}_\beta} \circ (\Psi_\alpha^{\beta \rightarrow \beta}(x) \otimes \Psi_{\bar{\gamma} \cup \delta_\gamma}^{\bar{\beta} \cup \delta_\beta}(y)).$$

The transition maps in (7-8) can be computed using holomorphic triangle maps. We note that, even though we are doing a move of the curves β on $\mathcal{H}_{\alpha,\beta}$ and a move of

the curves $\bar{\gamma} \cup \delta_\gamma$ on $C(\mathcal{H}_{\beta, \gamma})$, since the two diagrams are disjoint, we can compute the holomorphic triangles simultaneously to arrive at the formula

$$(7-9) \quad \Phi_{-\xi}(x \times y) = C^{\epsilon, \bar{\delta}_\beta} \circ F_{\alpha \cup \bar{\gamma} \cup \delta_\gamma, \beta \cup \bar{\gamma} \cup \delta_\gamma, \beta \cup \bar{\beta} \cup \delta_\beta}(x \times \Theta_{\bar{\gamma} \cup \delta_\gamma, \bar{\gamma} \cup \delta_\gamma}, y \times \Theta_{\beta, \beta}).$$

Using the local computation shown in Figure 19, we can commute the map $C^{\epsilon, \bar{\delta}_\beta}$ to the right of the triangle map. When we do this, the top-graded generators $\Theta_{\bar{\gamma} \cup \delta_\gamma, \bar{\gamma} \cup \delta_\gamma}$ and $\Theta_{\beta, \beta}$ can be rewritten as the images of sequences of 1-handle maps. Hence, we can write the right-hand side of (7-9) as

$$(7-10) \quad F_{\alpha \cup \epsilon \cup \bar{\gamma} \cup \delta_\gamma, \beta \cup \epsilon \cup \bar{\gamma} \cup \delta_\gamma, \beta \cup \bar{\delta}_\beta \cup \bar{\beta} \cup \delta_\beta}(F_1^{\epsilon \cup \bar{\gamma} \cup \delta_\gamma, \epsilon \cup \bar{\gamma} \cup \delta_\gamma}(x), F_1^{\beta, \beta}(C^{\epsilon, \bar{\delta}_\beta}(y))).$$

Replacing the right-hand side of (7-9) with (7-10), we streamline the notation by writing

$$(7-11) \quad \Phi_{-\xi} = F_{\alpha(\mathcal{T}), \beta \cup \gamma', \beta(\mathcal{T})} \circ (F_1^{\gamma', \gamma'} \otimes (F_1^{\beta, \beta} \circ C^{\epsilon, \bar{\delta}_\beta})),$$

where

$$\alpha(\mathcal{T}) := \alpha \cup \epsilon \cup \bar{\gamma} \cup \delta_\gamma, \quad \beta(\mathcal{T}) := \beta \cup \bar{\delta}_\beta \cup \bar{\beta} \cup \delta_\beta, \quad \text{and} \quad \gamma' := \epsilon \cup \bar{\gamma} \cup \delta_\gamma.$$

In the above expression, the contact handle map $C^{\epsilon, \bar{\delta}_\beta}$ is defined by adding the surface $R_\beta \natural \bar{R}_\beta$ to $R_\beta \natural \bar{\Sigma} \natural R_\gamma$ to get $R_\beta \natural \bar{R}_\beta \natural R_\beta \natural \bar{\Sigma} \natural R_\gamma$ and then adding in the curves ϵ and $\bar{\delta}_\beta$. On the level of complexes, it is defined by the formula $y \mapsto c \times y$, where c is the distinguished intersection point of $\mathbb{T}_\epsilon \cap \mathbb{T}_{\bar{\delta}_\beta}$. The copy of R_β that is added should be thought of as Σ surgered along the β curves. Note that $C^{\epsilon, \bar{\delta}_\beta}$ can be described by attaching contact 0-handles and 1-handles to add R_β and then attaching contact 2-handles to add $\bar{R}_\beta$ and the curves ϵ and $\bar{\delta}_\beta$.

We now change ϵ to ϵ' . Define

$$\gamma'' := \epsilon' \cup \bar{\gamma} \cup \delta_\gamma.$$

Applying associativity to the 4-tuple $(\alpha \cup \gamma'', \alpha(\mathcal{T}), \beta \cup \gamma', \beta(\mathcal{T}))$, we conclude from (7-11) that

$$\begin{aligned} (7-12) \quad & \Psi_{\alpha(\mathcal{T}) \rightarrow \alpha \cup \gamma''}^{\beta(\mathcal{T})} \circ \Phi_{-\xi} \\ & \simeq \Psi_{\alpha(\mathcal{T}) \rightarrow \alpha \cup \gamma''}^{\beta(\mathcal{T})} \circ F_{\alpha(\mathcal{T}), \beta \cup \gamma', \beta(\mathcal{T})} \circ (F_1^{\gamma', \gamma'} \otimes (F_1^{\beta, \beta} \circ C^{\epsilon, \bar{\delta}_\beta})) \\ & \simeq F_{\alpha \cup \gamma'', \beta \cup \gamma', \beta(\mathcal{T})} \circ ((\Psi_{\alpha(\mathcal{T}) \rightarrow \alpha \cup \gamma''}^{\beta \cup \gamma'} \circ F_1^{\gamma', \gamma'}) \otimes (F_1^{\beta, \beta} \circ C^{\epsilon, \bar{\delta}_\beta})). \end{aligned}$$

Applying naturality of sutured Floer homology as well as associativity for the 4-tuple $(\alpha \cup \gamma'', \beta \cup \gamma', \beta \cup \gamma'', \beta(\mathcal{T}))$ to (7-12), we conclude that

$$\begin{aligned}
 (7-13) \quad & F_{\alpha \cup \gamma'', \beta \cup \gamma', \beta(\mathcal{T})} \circ ((\Psi_{\alpha(\mathcal{T}) \rightarrow \alpha \cup \gamma''}^{\beta \cup \gamma'} \circ F_1^{\gamma', \gamma'}) \otimes (F_1^{\beta, \beta} \circ C^{\epsilon, \bar{\delta}_\beta})) \\
 & \simeq F_{\alpha \cup \gamma'', \beta \cup \gamma', \beta(\mathcal{T})} \circ ((\text{id} \circ \Psi_{\alpha(\mathcal{T}) \rightarrow \alpha \cup \gamma''}^{\beta \cup \gamma'} \circ F_1^{\gamma', \gamma'}) \otimes (F_1^{\beta, \beta} \circ C^{\epsilon, \bar{\delta}_\beta})) \\
 & \simeq F_{\alpha \cup \gamma'', \beta \cup \gamma', \beta(\mathcal{T})} \\
 & \quad \circ \left(((\Psi_{\alpha \cup \gamma''}^{\beta \cup \gamma'' \rightarrow \beta \cup \gamma'} \circ \Psi_{\alpha \cup \gamma''}^{\beta \cup \gamma' \rightarrow \beta \cup \gamma''}) \circ \Psi_{\alpha(\mathcal{T}) \rightarrow \alpha \cup \gamma''}^{\beta \cup \gamma'} \circ F_1^{\gamma', \gamma'}) \right. \\
 & \quad \left. \otimes (F_1^{\beta, \beta} \circ C^{\epsilon, \bar{\delta}_\beta}) \right) \\
 & \simeq F_{\alpha \cup \gamma'', \beta \cup \gamma'', \beta(\mathcal{T})} \\
 & \quad \circ ((\Psi_{\alpha(\mathcal{T}) \rightarrow \alpha \cup \gamma''}^{\beta \cup \gamma' \rightarrow \beta \cup \gamma''} \circ F_1^{\gamma', \gamma'}) \otimes (\Psi_{\beta \cup \gamma' \rightarrow \beta \cup \gamma''}^{\beta(\mathcal{T})} \circ F_1^{\beta, \beta} \circ C^{\epsilon, \bar{\delta}_\beta})).
 \end{aligned}$$

Combining (7-11), (7-12), and (7-13), we conclude

$$\begin{aligned}
 & \Psi_{\alpha(\mathcal{T}) \rightarrow \alpha \cup \gamma''}^{\beta(\mathcal{T})} \circ \Phi_{-\xi} \\
 & \simeq F_{\alpha \cup \gamma'', \beta \cup \gamma'', \beta(\mathcal{T})} \circ ((\Psi_{\alpha(\mathcal{T}) \rightarrow \alpha \cup \gamma''}^{\beta \cup \gamma' \rightarrow \beta \cup \gamma''} \circ F_1^{\gamma', \gamma'}) \otimes (\Psi_{\beta \cup \gamma' \rightarrow \beta \cup \gamma''}^{\beta(\mathcal{T})} \circ F_1^{\beta, \beta} \circ C^{\epsilon, \bar{\delta}_\beta})).
 \end{aligned}$$

We claim that

$$(7-14) \quad \Psi_{\alpha(\mathcal{T}) \rightarrow \alpha \cup \gamma''}^{\beta \cup \gamma' \rightarrow \beta \cup \gamma''} \circ F_1^{\gamma', \gamma'} \simeq F_1^{\gamma'', \gamma''}.$$

To establish (7-14), first note that the domain of both maps is $\text{SFH}(\Sigma, \alpha, \beta)$. The 1-handle maps $F_1^{\gamma', \gamma'}$ and $F_1^{\gamma'', \gamma''}$ are defined by taking the boundary connected sum of Σ with $\Sigma' := \bar{R}_\beta \natural R_\beta \natural \bar{\Sigma} \natural R_\gamma$. The boundary connected sum operation yields $|\partial \Sigma|$ arcs on $\Sigma(\mathcal{T}) = \Sigma \natural \Sigma'$ that separate Σ from Σ' and intersect none of the attaching curves in α , β , γ' , or γ'' . Hence, the change of diagrams map appearing on the left-hand side of (7-14) involves only counting holomorphic triangles with image on the disjoint union of Σ and Σ' . Since there is a unique top-graded generator of $\text{SFH}(\Sigma', \gamma', \gamma')$, and γ'' is obtained from γ' by a sequence of handle slides, that generator will be preserved by the change of diagrams map from $\text{SFH}(\Sigma', \gamma', \gamma')$ to $\text{SFH}(\Sigma', \gamma'', \gamma'')$. Hence, the 1-handle map will be preserved. Equation (7-14) now follows.

We obtain that the gluing map is chain homotopic to

$$F_{\alpha \cup \gamma'', \beta \cup \gamma'', \beta(\mathcal{T})} \circ (F_1^{\gamma'', \gamma''} \otimes (\Psi_{\beta \cup \gamma' \rightarrow \beta \cup \gamma''}^{\beta(\mathcal{T})} \circ F_1^{\beta, \beta} \circ C^{\epsilon, \bar{\delta}_\beta})).$$

Note that, on the diagram $(\Sigma(\mathcal{T}), \beta \cup \gamma'', \beta(\mathcal{T}))$, the $\bar{\delta}_\beta$ curves each have only one intersection point, which occurs with an ϵ' curve. The ϵ' curves still intersect both δ_β and $\bar{\delta}_\beta$. Further, each ϵ' curve intersects exactly one $\bar{\delta}_\beta$ curve. Hence, we

can consider the compound stabilization map $\sigma^{\epsilon', \bar{\delta}_\beta}$ defined in [Section 2.2](#). It agrees with the map from naturality by [Proposition 2.2](#).

We claim that

$$\begin{aligned} \Psi_{\beta \cup \gamma' \rightarrow \beta \cup \gamma''}^{\beta(\mathcal{T})} \circ F_1^{\beta, \beta} \circ C^{\epsilon', \bar{\delta}_\beta} \\ \simeq F_1^{\beta, \beta} \circ \sigma^{\epsilon', \bar{\delta}_\beta}: \text{SFH}(C(\mathcal{H}_{\beta, \gamma})) \rightarrow \text{SFH}(\Sigma(\mathcal{T}), \beta \cup \gamma'', \beta(\mathcal{T})). \end{aligned}$$

By [Proposition 2.1](#), it suffices to show the claim with the β curves surgered out and with no 1-handle maps (note that it is easy to show that on Σ , there is a path from each β curve to $\partial\Sigma$ that avoids $\bar{\delta}_\beta \cap \Sigma$, so the hypotheses of the previously mentioned proposition are satisfied). Therefore, it suffices to show that

$$\sigma^{\epsilon', \bar{\delta}_\beta} \simeq \Psi_{\gamma' \rightarrow \gamma''}^{\bar{\delta}_\beta \cup \bar{\beta} \cup \delta_\beta} \circ C^{\epsilon', \bar{\delta}_\beta},$$

or, equivalently, that

$$\text{id} \simeq (\sigma^{\epsilon', \bar{\delta}_\beta})^{-1} \circ \Psi_{\gamma' \rightarrow \gamma''}^{\bar{\delta}_\beta \cup \bar{\beta} \cup \delta_\beta} \circ C^{\epsilon', \bar{\delta}_\beta}.$$

However, this holds because of the functoriality of the gluing map, ie because the right-hand side represents the map induced by gluing a trivial a copy of $I \times R'_\beta$ to $-M_{\beta, \gamma}$.

Hence, we conclude that the gluing map $\Phi_{-\xi}$ is chain homotopic to

$$F_{\alpha \cup \gamma'', \beta \cup \gamma'', \beta(\mathcal{T})} \circ (F_1^{\gamma'', \gamma''} \otimes (F_1^{\beta, \beta} \circ \sigma^{\epsilon', \bar{\delta}_\beta})).$$

We still need to handle slide the δ_β over $\bar{\delta}_\beta$ to become $\hat{\delta}_\beta$ and then compound destabilize. That is, the gluing map is chain homotopic to

$$(\sigma^{\epsilon', \bar{\delta}_\beta})^{-1} \circ \Psi_{\alpha \cup \gamma''}^{\beta(\mathcal{T}) \rightarrow \beta(\mathcal{T})'} \circ F_{\alpha \cup \gamma'', \beta \cup \gamma'', \beta(\mathcal{T})} \circ (F_1^{\gamma'', \gamma''} \otimes (F_1^{\beta, \beta} \circ \sigma^{\epsilon', \bar{\delta}_\beta})),$$

where $\beta(\mathcal{T})'$ is equal to $\beta \cup \bar{\delta}_\beta \cup \bar{\beta} \cup \hat{\delta}_\beta$. By associativity applied to the 4-tuple $(\alpha \cup \gamma'', \beta \cup \gamma'', \beta(\mathcal{T}), \beta(\mathcal{T})')$, this is chain homotopic to

$$(\sigma^{\epsilon', \bar{\delta}_\beta})^{-1} \circ (F_{\alpha \cup \gamma'', \beta \cup \gamma'', \beta(\mathcal{T})'} \circ (F_1^{\gamma'', \gamma''} \otimes (\Psi_{\beta \cup \gamma''}^{\beta(\mathcal{T}) \rightarrow \beta(\mathcal{T})'} \circ F_1^{\beta, \beta} \circ \sigma^{\epsilon', \bar{\delta}_\beta}))).$$

By a simple generalization of [Lemma 2.3](#) to deal with multiple compound stabilizations of the Heegaard triple simultaneously, this now becomes

$$F_{D_\gamma(\alpha), D_\gamma(\beta), A(\beta)} \circ (F_1^{C(\gamma), C(\gamma)} \otimes ((\sigma^{\epsilon', \bar{\delta}_\beta})^{-1} \circ \Psi_{\beta \cup \gamma''}^{\beta(\mathcal{T}) \rightarrow \beta(\mathcal{T})'} \circ F_1^{\beta, \beta} \circ \sigma^{\epsilon', \bar{\delta}_\beta})),$$

where $C(\gamma) = \bar{\gamma} \cup \delta_\gamma$. Applying the triangle counts from [Proposition 2.1](#) to the map $F_1^{\beta, \beta}$ for 1-handles added near the boundary and also commuting the map $F_1^{\beta, \beta}$

with the destabilization map, this becomes

$$F_{D_{\gamma}(\alpha), D_{\gamma}(\beta), A(\beta)} \circ (F_1^{C(\gamma), C(\gamma)} \otimes (F_1^{\beta, \beta} \circ (\sigma^{\epsilon', \bar{\delta}_{\beta}})^{-1} \circ \Psi_{\gamma''}^{\beta(\mathcal{T}) \setminus \beta \rightarrow \beta(\mathcal{T})' \setminus \beta} \circ \sigma^{\epsilon', \bar{\delta}_{\beta}})).$$

We claim that

$$(\sigma^{\epsilon', \bar{\delta}_{\beta}})^{-1} \circ \Psi_{\gamma''}^{\beta(\mathcal{T}) \setminus \beta \rightarrow \beta(\mathcal{T})' \setminus \beta} \circ \sigma^{\epsilon', \bar{\delta}_{\beta}} \simeq \text{id},$$

which follows simply from naturality, as it is a loop in the space of Heegaard diagrams. We have now arrived at our desired formula, concluding the proof of [Proposition 7.15](#). $\square$

7.10 Computation of the triangle cobordism map

We now prove that the sutured cobordism map for $\mathcal{W}_{\alpha, \beta, \gamma}$ is chain homotopic to the map that counts holomorphic triangles on a single Heegaard triple, by using the formula from the previous section for the contact gluing map.

Theorem 7.16 *If $(\Sigma, \alpha, \beta, \gamma)$ is a sutured Heegaard triple, then the sutured cobordism map*

$$F_{\mathcal{W}_{\alpha, \beta, \gamma}}: \text{CF}(\Sigma, \alpha, \beta) \otimes \text{CF}(\Sigma, \beta, \gamma) \rightarrow \text{CF}(\Sigma, \alpha, \gamma)$$

is chain homotopic to the map that counts holomorphic triangles on the triple $(\Sigma, \alpha, \beta, \gamma)$.

Proof We first compose with the change of diagrams map $\text{id} \otimes \Psi_{\mathcal{H}_{\beta, \gamma} \rightarrow C(\mathcal{H}_{\beta, \gamma})}$. The next step is to use the gluing map to glue the two copies of R_{β} together. Then we perform surgery on a $|\beta|$ -component framed link. See [Section 7.8](#) for a description of the triple diagram

$$(D_{\gamma}(\Sigma), D_{\gamma}(\alpha), A(\beta), \Delta_{\Sigma})$$

used for computing the 2-handle cobordism map. This yields (omitting writing the first change of diagrams map)

$$F_{D_{\gamma}(\alpha), A(\beta), \Delta_{\Sigma}} \circ ((F_{D_{\gamma}(\alpha), D_{\gamma}(\beta), A(\beta)} \circ (F_1^{\bar{\gamma} \cup \delta_{\gamma}, \bar{\gamma} \cup \delta_{\gamma}} \otimes F_1^{\beta, \beta})) \otimes \Theta_{A(\beta), \Delta_{\Sigma}}).$$

We now use associativity applied to the 4-tuple $(D_{\gamma}(\alpha), D_{\gamma}(\beta), A(\beta), \Delta_{\Sigma})$ to see that this is chain homotopic to

$$F_{D_{\gamma}(\alpha), D_{\gamma}(\beta), \Delta_{\Sigma}} \circ (F_1^{\bar{\gamma} \cup \delta_{\gamma}, \bar{\gamma} \cup \delta_{\gamma}} \otimes (F_{D_{\gamma}(\beta), A(\beta), \Delta_{\Sigma}} \circ (F_1^{\beta, \beta} \otimes \Theta_{A(\beta), \Delta_{\Sigma}}))).$$

Note that

$$\Psi := F_{D_{\gamma}(\beta), A(\beta), \Delta_{\Sigma}} \circ (F_1^{\beta, \beta} \otimes \Theta_{A(\beta), \Delta_{\Sigma}})$$

is the change of diagrams map $\Psi_{C(\mathcal{H}_{\beta,\gamma}) \rightarrow D_\gamma(\mathcal{H}_{\beta,\gamma})}$, from a weakly conjugated diagram to a doubled diagram, by [Lemma 7.12](#). Thus, the cobordism map is

$$(7-15) \quad F_{D_\gamma(\alpha), D_\gamma(\beta), \Delta_\Sigma} \circ (F_1^{\bar{\gamma} \cup \delta_\gamma, \bar{\gamma} \cup \delta_\gamma} \otimes \Psi).$$

The range of this map is not $\text{SFH}(\mathcal{H}_{\alpha,\gamma})$ but rather a double of this diagram along R_γ , so we must compose with the change of diagrams map from $D_\gamma(\mathcal{H}_{\alpha,\gamma})$ to $\mathcal{H}_{\alpha,\gamma}$, which is

$$F_3^{\bar{\gamma} \cup \delta_\gamma, \bar{\gamma} \cup \delta_\gamma} \circ (F_{D_\gamma(\alpha), \Delta_\Sigma, D_\gamma(\gamma)} \circ (- \otimes \Theta_{\Delta_\Sigma, D_\gamma(\gamma)})),$$

by [Lemma 7.9](#). By postcomposing (7-15) with this expression, and applying associativity of the 4-tuple $(D_\gamma(\alpha), D_\gamma(\beta), \Delta_\Sigma, D_\gamma(\gamma))$, we conclude that the composition is chain homotopic to

$$F_3^{\bar{\gamma} \cup \delta_\gamma, \bar{\gamma} \cup \delta_\gamma} \circ (F_{D_\gamma(\alpha), D_\gamma(\beta), D_\gamma(\gamma)} \circ (F_1^{\bar{\gamma} \cup \delta_\gamma, \bar{\gamma} \cup \delta_\gamma} \otimes (F_{D_\gamma(\beta), \Delta_\Sigma, D_\gamma(\gamma)} \circ (\Psi \otimes \Theta_{\Delta_\Sigma, D_\gamma(\gamma)})))).$$

Using the 3-handle and triangle map computation from [Proposition 2.1](#) (note that it is an easy exercise to verify that the hypotheses of that proposition are satisfied), this is equal to

$$F_{\alpha, \beta, \gamma} \circ (- \otimes F_3^{\bar{\gamma} \cup \delta_\gamma, \bar{\gamma} \cup \delta_\gamma} \circ (F_{D_\gamma(\beta), \Delta_\Sigma, D_\gamma(\gamma)} \circ (\Psi \otimes \Theta_{\Delta_\Sigma, D_\gamma(\gamma)}))).$$

We note that

$$F_3^{\bar{\gamma} \cup \delta_\gamma, \bar{\gamma} \cup \delta_\gamma} \circ (F_{D_\gamma(\beta), \Delta_\Sigma, D_\gamma(\gamma)} \circ (\Psi \otimes \Theta_{\Delta_\Sigma, D_\gamma(\gamma)}))$$

is just a change of diagrams map. Writing Ψ' for the compositions of all the three change of diagrams maps (the initial one from $\mathcal{H}_{\beta,\gamma}$ to a weakly conjugated diagram $C(\mathcal{H}_{\beta,\gamma})$, which we have been omitting writing, and then the last two, going back to $\mathcal{H}_{\beta,\gamma}$ through a doubled diagram), the composition is thus equal to

$$F_{\alpha, \beta, \gamma}(- \otimes \Psi') \simeq F_{\alpha, \beta, \gamma}(-, -),$$

since $\Psi' \simeq \text{id}_{(\Sigma, \beta, \gamma)}$ by naturality. □

8 Trace and cotrace cobordisms

Let $\xi_{I \times \partial M}$ denote the I -invariant contact structure on $-I \times \partial M$ that induces the dividing set γ on ∂M . In this section, we consider the trace cobordism

$$\Lambda_{(M, \gamma)} = (I \times M, -I \times \partial M, [\xi_{I \times \partial M}])$$

from $(M, \gamma) \sqcup (-M, \gamma)$ to $\emptyset$, and the cotrace cobordism

$$V_{(M, \gamma)} = (I \times M, -I \times \partial M, [\xi_{I \times \partial M}])$$

from $\emptyset$ to $(-M, \gamma) \sqcup (M, \gamma)$. The main result of this section is the following:

Theorem 8.1 *The trace cobordism $\Lambda_{(M, \gamma)}: (M, \gamma) \sqcup (-M, \gamma) \rightarrow \emptyset$ induces the canonical trace map*

$$\text{tr}: \text{SFH}(M, \gamma) \otimes \text{SFH}(-M, \gamma) \rightarrow \mathbb{F}_2.$$

The cotrace cobordism $V_{(M, \gamma)}: \emptyset \rightarrow (-M, \gamma) \sqcup (M, \gamma)$ induces the canonical cotrace map

$$\text{cotr}: \mathbb{F}_2 \rightarrow \text{SFH}(-M, \gamma) \otimes \text{SFH}(M, \gamma).$$

To prove [Theorem 8.1](#), we first need a convenient topological description of the trace cobordism. Suppose that (Σ, α, β) is a diagram for (M, γ) . We will write $(M_{\alpha, \alpha}, \gamma_{\alpha, \alpha})$ for the sutured manifold constructed from the diagram (Σ, α, α) , and we note that $(M_{\alpha, \alpha}, \gamma_{\alpha, \alpha})$ can be obtained by surgering $(I \times R_{\alpha}, I \times \partial R_{\alpha})$ along k 0–spheres. We consider the triangular sutured manifold cobordism

$$\mathcal{W}_{\alpha, \beta, \alpha} = (W_{\alpha, \beta, \alpha}, Z_{\alpha, \beta, \alpha}, [\xi_{\alpha, \beta, \alpha}])$$

from $(M, \gamma) \sqcup (-M, \gamma)$ to $(M_{\alpha, \alpha}, \gamma_{\alpha, \alpha})$, defined in [Section 7](#).

There is also a cobordism $\mathcal{W}_{\alpha} = (W_{\alpha}, Z_{\alpha}, [\xi_{\alpha}])$ from $(M_{\alpha, \alpha}, \gamma_{\alpha, \alpha})$ to $\emptyset$. The 4–manifold W_{α} is defined as

$$W_{\alpha} = (P_1 \times \Sigma) \cup_{e_{\alpha} \times \Sigma} (e_{\alpha} \times U_{\alpha}).$$

Here P_1 denotes a monogon, viewed as having a single boundary edge e_{α} .

Lemma 8.2 *The cobordism $\Lambda_{(M, \gamma)}$ is equivalent to the composition $\mathcal{W}_{\alpha} \circ \mathcal{W}_{\alpha, \beta, \alpha}$.*

Proof It follows from [\[15, Proposition 6.6\]](#) that $W_{\alpha} \cup W_{\alpha, \beta, \alpha}$ is diffeomorphic to $I \times M_{\alpha, \beta}$ and, furthermore, that $Z_{\alpha, \beta, \alpha} \cup Z_{\alpha}$ is diffeomorphic to $-I \times \partial M$. The fact that ξ_{α} and $\xi_{\alpha, \beta, \alpha}$ glue up to give $\xi_{I \times \partial M}$ can be proven by adapting [Lemma 7.3](#). Schematically, the decomposition of $\Lambda_{(M, \gamma)}$ into $\mathcal{W}_{\alpha, \beta, \alpha}$ and $\mathcal{W}_{\alpha}$ is shown in [Figure 45](#). $\square$

Let α' denote be a small Hamiltonian translate of α . Note that, of course, $(M_{\alpha,\alpha}, \gamma_{\alpha,\alpha})$ and $(M_{\alpha,\alpha'}, \gamma_{\alpha,\alpha'})$ are homeomorphic. It follows from [Theorem 7.1](#) and the composition law that the map

$$F_{\Lambda_{(M,\gamma)}}: CF(\Sigma, \alpha, \beta) \otimes CF(\Sigma, \beta, \alpha) \rightarrow \mathbb{F}_2$$

is equal to the composition

(8-1)
$$F_{\mathcal{W}_\alpha} \circ F_{\alpha,\beta,\alpha'} \circ (\text{id}_{CF(\Sigma,\alpha,\beta)} \otimes \Psi_{(\Sigma,\beta,\alpha) \rightarrow (\Sigma,\beta,\alpha')}),$$

where $F_{\alpha,\beta,\alpha'}$ is the map that counts holomorphic triangles on $(\Sigma, \alpha, \beta, \alpha')$.

It remains to compute the cobordism map for $\mathcal{W}_\alpha$. Note that, on $(\Sigma, \alpha, \alpha')$, there is a canonical bottom-graded intersection point $\Theta_{\alpha,\alpha'}^-$.

Proposition 8.3 *The cobordism $\mathcal{W}_\alpha$ from $(M_{\alpha,\alpha'}, \gamma_{\alpha,\alpha'})$ to $\emptyset$ induces the map*

$$x \mapsto \begin{cases} 1 & \text{if } x = \Theta_{\alpha,\alpha'}^-, \\ 0 & \text{otherwise.} \end{cases}$$

Before we prove the above result, we need to find a convenient Morse function for the cobordism $\mathcal{W}_\alpha$. As a first step, we prove the following:

Lemma 8.4 *Let U_α be the sutured compression body that is formed by attaching 3-dimensional 2-handles to $I \times \Sigma$ along the curves $\{0\} \times \alpha$. After rounding corners, we can view U_α as a (nonsutured) handlebody of genus $|\alpha| - \chi(R_\alpha) + 1$ and boundary*

$$\partial U_\alpha = (\{1\} \times \Sigma) \cup \bar{R}_\alpha,$$

where R_α is the surface obtained by surgering Σ along the α curves. Furthermore, a (possibly overcomplete) set of compressing disks for U_α can be obtained by taking $|\alpha|$ compressing disks D_α with boundary $\{1\} \times \alpha$ for $\alpha \in \alpha$, as well as disks of the form $D_{c_i^*} := I \times c_i^*$ for pairwise disjoint, embedded arcs $c_1^*, \dots, c_{b_1(R_\alpha)}^*$ in Σ that avoid the α curves and form a basis of $H_1(R_\alpha, \partial R_\alpha)$. These cut U_α into $b_0(R_\alpha)$ 3-balls.

Remark 8.5 A basis of arcs $c_1^*, \dots, c_{b_1(R_\alpha)}^*$ can be obtained by picking a Morse function $g: R_\alpha \rightarrow I$ (viewing R_α as a cobordism from $\emptyset$ to ∂R_α) that has $b_0(R_\alpha)$ local minima and $b_1(R_\alpha)$ index 1 critical points. The Morse function g determines a handle decomposition of R_α with $b_0(R_\alpha)$ 0-handles (ie disks) and $b_1(R_\alpha)$ 1-handles. The arcs $c_1^*, \dots, c_{b_1(R_\alpha)}^*$ can be taken as the cocores of the 1-handles in this decomposition.

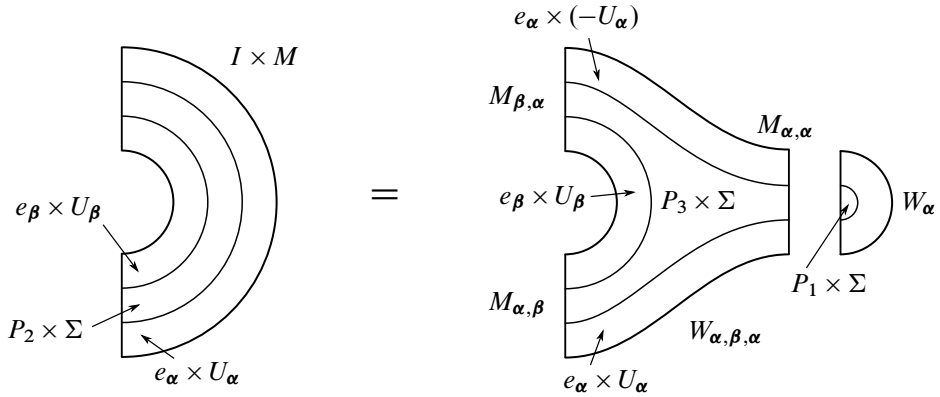


Figure 45: The decomposition $I \times M \cong W_{\alpha, \beta, \alpha} \cup_{M_{\alpha, \alpha}} W_{\alpha}$. It is convenient to view the polygons P_1 , P_2 , and P_3 as having fattened vertices.

Proof of Lemma 8.4 View U_{α} as a cobordism from $\overline{\partial U}_{\alpha} = (\{1\} \times \overline{\Sigma}) \cup R_{\alpha}$ to the empty set. The α curves determine 3-dimensional 2-handles in U_{α} . After attaching these 2-handles, the remaining cobordism is homeomorphic to $I \times R_{\alpha}$ (with corners smoothed), viewed as a cobordism from $R_{\alpha} \cup_{\partial R_{\alpha}} \overline{R}_{\alpha}$ to $\emptyset$. In such a way, we reduce the argument to the case when there are no α curves, where it is straightforward. $\square$

We need an additional Morse theory argument:

Lemma 8.6 Suppose that U_{α} is the sutured compression body induced by the sutured monodisk (Σ, α) . Then $I \times U_{\alpha}$ can be viewed, after smoothing corners, as a (nonsutured) cobordism from the closed manifold $U_{\alpha} \cup_{\partial U_{\alpha}} -U_{\alpha}$ to $\emptyset$. Furthermore, there is a Morse function $F: I \times U_{\alpha} \rightarrow I$ such that

- $F^{-1}(0) = U_{\alpha} \cup_{\partial U_{\alpha}} -U_{\alpha}$,
- F has no index 0, 1, or 2 critical points,
- F has $|\alpha| + b_1(R_{\alpha})$ index 3 critical points, and
- F has $b_0(R_{\alpha})$ index 4 critical points.

The attaching spheres of the 3-handles are obtained taking the union of the disks D_{α_i} and $D_{c_i^*}$ (defined in Lemma 8.4) in U_{α} , together with their images in $-U_{\alpha}$.

Proof A model for the 4-manifold obtained by rounding the corners of $I \times U_{\alpha}$ can be taken to be $X := I \times U_{\alpha} / \sim$, where $(t, x) \sim (t', x)$ if $x \in \partial U_{\alpha}$. Using Lemma 8.4, we can construct a Morse function $f: U_{\alpha} \rightarrow I$ with $f^{-1}(0) = \partial U_{\alpha}$ such that f has no

index 0 or 1 critical points, $|\alpha| + b_1(R_\alpha)$ index 2 critical points, and $b_0(R_\alpha)$ index 3 critical points. Furthermore, the descending manifolds of the index 2 critical points in U_α are the disks D_{α_i} and $D_{c_i^*}$.

To construct a Morse function on X with the stated critical points, the argument is now a modification of [Lemma 6.7](#). More precisely, we will construct an auxiliary function

$$G: (I \times I)/(I \times \{0\}) \rightarrow I.$$

We view $(I \times I)/(I \times \{0\})$ as having the same smooth structure at the point $I \times \{0\}$ as the upper half-plane; see [Figure 46](#). Furthermore, we assume that

- $G|_{I \times \{0\}} \equiv G|_{\{0\} \times I} \equiv G|_{\{1\} \times I} \equiv 0$,
- $\partial G / \partial s > 0$ in $(0, 1) \times I$,
- $(\partial G / \partial t)(t, s) = 0$ if and only if $(t, s) \in \{\frac{1}{2}\} \times I$.

An example of such a function G is shown in [Figure 46](#). We then consider the function $F: X \rightarrow I$ defined by

$$F(t, y) = G(t, f(y)).$$

It is straightforward to verify that F on $I \times U_\alpha$ with its corners rounded is Morse and has critical points with attaching spheres as stated. □

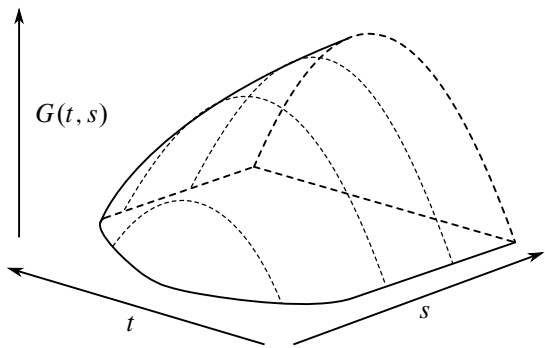


Figure 46: An example of an auxiliary function $G: (I \times I)/(I \times \{0\}) \rightarrow I$ in [Lemma 8.6](#).

We are now ready to prove [Proposition 8.3](#):

Proof of [Proposition 8.3](#) The sutured cobordism

$$\mathcal{W}_\alpha = (W_\alpha, Z_\alpha, [\xi_\alpha])$$

from $(M_{\alpha, \alpha'}, \gamma_{\alpha, \alpha'})$ to $\emptyset$ is the composition of the boundary cobordism $\mathcal{W}_\alpha^b$ obtained

by gluing $Z_\alpha := I \times R_\alpha$ to $\partial M_{\alpha,\alpha'} \cong R_\alpha \cup_{\partial R_\alpha} \bar{R}_\alpha$ and the special cobordism $\mathcal{W}_\alpha^s$ (between closed 3-manifolds) diffeomorphic to $I \times U_\alpha$ from $U_\alpha \cup -U_\alpha$ to $\emptyset$. We use [Lemma 8.6](#) to give a handle decomposition of the cobordism $\mathcal{W}_\alpha^s$ consisting of $|\alpha| - \chi(R_\alpha) + 1$ 3-handles and $b_0(R_\alpha)$ 4-handles.

Note that this description of the cobordism $\mathcal{W}_\alpha$ does not quite allow us to compute the cobordism map $F_{\mathcal{W}_\alpha}$, because when we glue $Z_\alpha = I \times R_\alpha$ to $M_{\alpha,\alpha'}$, we obtain a closed (ie nonsutured manifold). The necessary modification is to instead remove $b_0(R_\alpha)$ 3-balls from Z_α , to obtain a sutured manifold cobordism from $(M_{\alpha,\alpha'}, \gamma_{\alpha,\alpha'})$ to $\bigsqcup_{i=1}^{b_0(R_\alpha)} (B^3, \gamma_0)$ (where $\gamma_0 \subseteq \partial B^3$ is a simple closed curve), and then compose with the natural isomorphism

$$\bigotimes_{i=1}^{b_0(R_\alpha)} \text{SFH}(B^3, \gamma_0) \cong \mathbb{F}_2.$$

Let us write Z'_α for Z_α with $b_0(R_\alpha)$ 3-balls removed and $\mathcal{W}'$ for the induced special cobordism from $(M_{\alpha,\alpha'} \cup Z'_\alpha, \bigsqcup_{i=1}^{b_0(R_\alpha)} \gamma_0)$ to $\bigsqcup_{i=1}^{b_0(R_\alpha)} (B^3, \gamma_0)$. Note that $\mathcal{W}'$ can be given a handle decomposition that is the same as the handle decomposition for $\mathcal{W}_\alpha^s$ with the 4-handles removed.

We write $S_\alpha \subseteq M_{\alpha,\alpha'} \cup Z'_\alpha$ for the 2-spheres obtained by doubling the compressing disk D_α for $\alpha \in \alpha$ and $S_{c_i^*} \subseteq M_{\alpha,\alpha'} \cup Z'_\alpha$ for the 2-spheres obtained by doubling the compressing disk $D_{c_i^*} \subseteq U_\alpha$. We recall that the curves c_i^* were obtained by picking a Morse function on R_α that had $b_0(R_\alpha)$ index 0 critical points and $b_1(R_\alpha)$ index 1 critical points. The index 1 critical points each determine a 2-dimensional 1-handle, added to the handles of index 0. The curves c_i^* are then the cocores of these handles.

Note that, by the composition law for sutured cobordisms, we can commute the 3-handle maps for the spheres S_α with the contact gluing map for gluing in Z'_α . As the composition of the 3-handle maps for the spheres S_α has the same formula as the one in the statement of the proposition, under the identification

$$\text{SFH}(I \times R_\alpha, I \times \partial R_\alpha) \cong \mathbb{F}_2,$$

we thus reduce to the case when there are no α curves, ie when (Σ, α) is the diagram $(R_\alpha, \emptyset)$.

To see the claim when there are no α curves, we note that the cobordism map is obtained by first gluing Z'_α to the product sutured manifold $(I \times R_\alpha, I \times \partial R_\alpha)$ and then attaching $b_1(R_\alpha)$ 4-dimensional 3-handles. The contact manifold Z'_α is obtained by attaching $b_1(R_\alpha)$ contact 2-handles. If c_i^* denotes the arc on R_α from [Lemma 8.4](#),

obtained as the cocore of a handle decomposition of R_α , as above, then the attaching 1-spheres s_i for the contact 2-handles forming Z'_α are given by

$$s_i = (\{0, 1\} \times c_i) \cup (I \times \partial c_i).$$

Note that s_i bounds the disk $D_{c_i}^*$, described in Lemma 8.4, which is the descending manifold of the Morse function f , constructed on U_α in the proof of Lemma 8.6. The 2-sphere $S_{c_i}^*$ along which we attach the $b_1(R_\alpha)$ 3-handles are then equal to the union of $D_{c_i}^*$, together with the core of the corresponding contact 2-handle. Now an easy model computation shows that the composition of these $b_1(R_\alpha)$ contact 2-handle maps and the $b_1(R_\alpha)$ 3-handle maps sends $1 \in \text{SFH}(I \times R_\alpha, I \times \partial R_\alpha)$ to $1 \in \bigotimes_{i=1}^{b_0(R_\alpha)} \text{SFH}(B^3, \gamma_0)$. This model computation is shown in Figures 47 and 48.

Using the composition law for sutured cobordisms, the formula for the cobordism map $F_{\mathcal{W}_\alpha}$ now follows. $\square$

We can now prove the main theorem of this section:

Proof of Theorem 8.1 Let us consider the trace cobordism map. Recall from (8-1) that $\Lambda_{(M, \gamma)}$ can be written as the composition of $\mathcal{W}_\alpha$ with $\mathcal{W}_{\alpha, \beta, \alpha}$. Noting that $\mathcal{W}_{\alpha, \beta, \alpha}$ is equivalent to $\mathcal{W}_{\alpha, \beta, \alpha'}$ (where α' is a small Hamiltonian translate of α) and using the formula for $F_{\mathcal{W}_\alpha}$ from Proposition 8.3, we know that

$$F_{\Lambda_{(M, \gamma)}}: \text{CF}(\Sigma, \alpha, \beta) \otimes \text{CF}(\Sigma, \beta, \alpha) \rightarrow \mathbb{F}_2$$

takes the form

$$F_{\Lambda_{(M, \gamma)}}(x \otimes y) = \langle F_{\alpha, \beta, \alpha'}(x, \Psi_{\beta}^{\alpha \rightarrow \alpha'}(y)), \Theta_{\alpha, \alpha'}^- \rangle,$$

where

$$\langle z, z' \rangle := \begin{cases} 1 & \text{if } z = z', \\ 0 & \text{otherwise.} \end{cases}$$

Note, however, that the triangle map $F_{\alpha, \beta, \alpha'}$ counts the same triangles as the triangle map $F_{\alpha', \alpha, \beta}$ and that $\Theta_{\alpha, \alpha'}^- \in \text{CF}(\Sigma, \alpha, \alpha')$ is the same intersection point as $\Theta_{\alpha', \alpha}^+ \in \text{CF}(\Sigma, \alpha', \alpha)$. Hence

$$\langle F_{\alpha, \beta, \alpha'}(x, \Psi_{\beta}^{\alpha \rightarrow \alpha'}(y)), \Theta_{\alpha, \alpha'}^- \rangle = \langle F_{\alpha', \alpha, \beta}(\Theta_{\alpha', \alpha}^+(x), \Psi_{\beta}^{\alpha \rightarrow \alpha'}(y)) \rangle.$$

However, $F_{\alpha', \alpha, \beta}(\Theta_{\alpha', \alpha}^+, -)$ is the transition map $\Psi_{\alpha \rightarrow \alpha'}^\beta(-)$, from naturality. Hence the cobordism map becomes simply the composition

$$\langle \Psi_{\alpha \rightarrow \alpha'}^\beta(x), \Psi_{\beta}^{\alpha \rightarrow \alpha'}(y) \rangle,$$

which is easily seen to be $\langle \Psi_{\alpha' \rightarrow \alpha}^\beta \circ \Psi_{\alpha \rightarrow \alpha'}^\beta(x), y \rangle = \langle x, y \rangle$. The formula for the

cotrace cobordism map $F_{V(M,\gamma)}$ follows from the above formula for the trace cobordism map $F_{\Lambda(M,\gamma)}$, combined with [Theorem 6.6](#). □

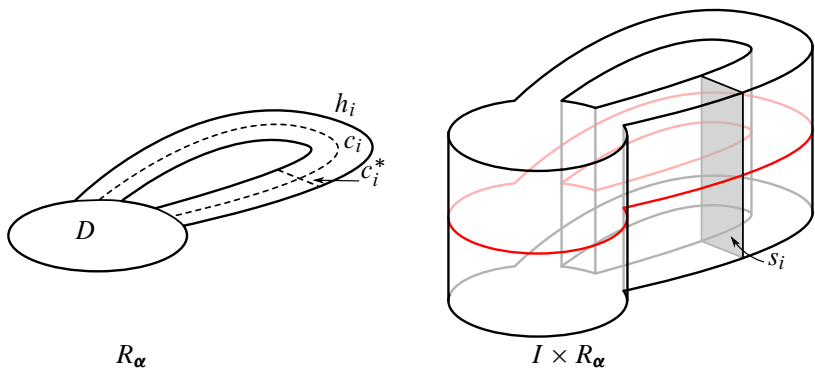


Figure 47: R_α and $I \times R_\alpha$. On the left, a 1-handle h_i from a handle decomposition of R_α is shown, together with the core c_i and the cocore c_i^* . On the right is the product manifold $(I \times R_\alpha, I \times \partial R_\alpha)$, together with the closed curve s_i , along which we attach a contact 2-handle. The red lines on the right indicate the sutures of $I \times R_\alpha$.

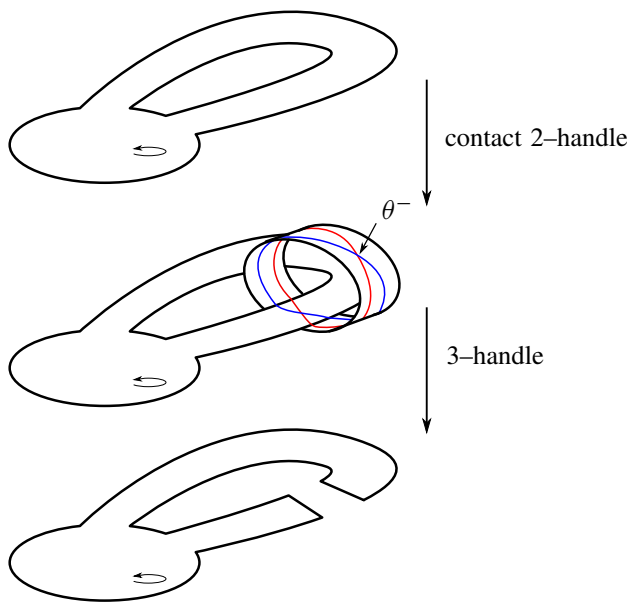


Figure 48: The model computation from [Proposition 8.3](#). The contact 2-handle map takes the form $1 \mapsto \theta^-$, and the 3-handle map sends θ^- to 1.

9 Equivalence of two link cobordism map constructions

In this section, we describe an application of the techniques of this paper to link Floer homology.

9.1 Background on link Floer homology

Knot Floer homology is an invariant of knots embedded in 3-manifolds constructed by Ozsváth and Szabó [22] and independently by Rasmussen [27]. Link Floer homology is a generalization to links, constructed by Ozsváth and Szabó [26].

Definition 9.1 A *multibased link* $\mathbb{L} = (L, \mathbf{w}, \mathbf{z})$ in a 3-manifold Y is an oriented link $L \subseteq Y$, together with two disjoint collections of basepoints $\mathbf{w}, \mathbf{z} \subseteq L$, such that

- (1) each component of L has at least two basepoints,
- (2) the basepoints along a link component of L alternate between $\mathbf{w}$ and $\mathbf{z}$ as one traverses the link.

To a multibased link $\mathbb{L}$ in Y , link Floer homology associates an $\mathbb{F}_2$ -module

$$\widehat{\mathrm{HFL}}(Y, \mathbb{L}).$$

To construct the modules, one picks a Heegaard diagram for the pair $(Y, \mathbb{L})$, in the following sense:

Definition 9.2 A *Heegaard diagram* $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \mathbf{z})$ for an oriented multibased link $(Y, (L, \mathbf{w}, \mathbf{z}))$ is a tuple satisfying the following:

- (1) $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is a Heegaard diagram for Y such that $Y \setminus \Sigma$ is the union of two handlebodies $U_{\boldsymbol{\alpha}}$ and $U_{\boldsymbol{\beta}}$ that meet along Σ .
- (2) $\Sigma \cap L = \mathbf{w} \cup \mathbf{z}$.
- (3) Each component of $\Sigma \setminus \boldsymbol{\alpha}$ and $\Sigma \setminus \boldsymbol{\beta}$ contains exactly one $\mathbf{w}$ basepoint and one $\mathbf{z}$ basepoint.
- (4) $L \cap U_{\boldsymbol{\alpha}}$ is isotopic in $U_{\boldsymbol{\alpha}}$, relative to $\partial(L \cap U_{\boldsymbol{\alpha}})$, to a collection of arcs in $\Sigma \setminus \boldsymbol{\alpha}$. Similarly, $L \cap U_{\boldsymbol{\beta}}$ is isotopic in $U_{\boldsymbol{\beta}}$, relative to $\partial(L \cap U_{\boldsymbol{\beta}})$, to a collection of arcs in $\Sigma \setminus \boldsymbol{\beta}$.
- (5) The link L intersects Σ positively at the $\mathbf{z}$ basepoints and negatively at the $\mathbf{w}$ basepoints.

A somewhat more concise way of defining a Heegaard diagram for a multibased link is as a diagram for the sutured manifold $Y(\mathbb{L})$, obtained by removing a neighborhood of the link L from Y and adding sutures to $\partial(Y \setminus N(L))$ that are positively oriented meridians of L over the $\mathbf{w}$ basepoints and negatively oriented meridians of L over the $\mathbf{z}$ basepoints. Link Floer homology, as defined by Ozsváth and Szabó [26], is easily seen to satisfy the isomorphism

$$\widehat{\mathrm{HFL}}(Y, \mathbb{L}) \cong \mathrm{SFH}(Y(\mathbb{L})).$$

If $(\Sigma, \alpha, \beta, \mathbf{w}, \mathbf{z})$ is a diagram for $(Y, \mathbb{L})$, then the link Floer complex

$$\widehat{\mathrm{CFL}}(\Sigma, \alpha, \beta, \mathbf{w}, \mathbf{z})$$

is generated over $\mathbb{F}_2$ by intersection points $\mathbf{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$, and the differential counts Maslov index 1 pseudoholomorphic discs that go over none of the $\mathbf{w}$ or $\mathbf{z}$ basepoints.

For links in S^3 or null-homologous links in $(S^1 \times S^2)^{\#n}$, there is some additional structure on $\widehat{\mathrm{HFL}}(S^3, \mathbb{L})$. If $(\Sigma, \alpha, \beta, \mathbf{w}, \mathbf{z})$ is a diagram for a link in S^3 , one can define three relative gradings, $\mathrm{gr}_\mathbf{w}$, $\mathrm{gr}_\mathbf{z}$, and A , on $\widehat{\mathrm{CFL}}(\Sigma, \alpha, \beta, \mathbf{w}, \mathbf{z})$. If $\mathbf{x}$ and $\mathbf{y}$ are two intersection points, then the gradings $\mathrm{gr}_\mathbf{w}$ and $\mathrm{gr}_\mathbf{z}$ are defined by picking a homology class $\phi \in \pi_2(\mathbf{x}, \mathbf{y})$ and setting

$$\mathrm{gr}_\mathbf{w}(\mathbf{x}, \mathbf{y}) := \mu(\phi) - 2n_\mathbf{w}(\phi) \quad \text{and} \quad \mathrm{gr}_\mathbf{z}(\mathbf{x}, \mathbf{y}) := \mu(\phi) - 2n_\mathbf{z}(\phi),$$

where $n_\mathbf{w}(\phi)$ and $n_\mathbf{z}(\phi)$ denote the sum of the multiplicities of ϕ over the $\mathbf{w}$ or $\mathbf{z}$ basepoints, respectively. It is easy to see that the formulas for $\mathrm{gr}_\mathbf{w}$ and $\mathrm{gr}_\mathbf{z}$ are independent of the choice of homology class ϕ . An absolute lift of the relative grading $\mathrm{gr}_\mathbf{w}$ can be fixed by requiring that $\widehat{\mathrm{HF}}(\Sigma, \alpha, \beta, \mathbf{w})$, which is isomorphic as a relatively graded group to $\bigotimes^{|\mathbf{w}|-1}((\mathbb{F}_2)_{-\frac{1}{2}} \oplus (\mathbb{F}_2)_{\frac{1}{2}})$, have top-graded generator in grading $\frac{1}{2}(|\mathbf{w}| - 1)$. An absolute lift of the grading $\mathrm{gr}_\mathbf{z}$ can be specified similarly. Finally, the Alexander grading A can be defined as

$$A := \frac{1}{2}(\mathrm{gr}_\mathbf{w} - \mathrm{gr}_\mathbf{z}).$$

9.2 The link Floer homology TQFT

In [15], the first author provided a construction of cobordism maps for decorated link cobordisms. The construction used the following notion of cobordism between multibased links:

Definition 9.3 Let Y_1 and Y_2 be 3-manifolds containing multibased links $\mathbb{L}_1 = (L_1, \mathbf{w}_1, \mathbf{z}_1)$ and $\mathbb{L}_2 = (L_2, \mathbf{w}_2, \mathbf{z}_2)$, respectively. A *decorated link cobordism* from $(Y_1, \mathbb{L}_1)$ to $(Y_2, \mathbb{L}_2)$ is a triple $(X, S, \mathcal{A})$, where

- (1) X is an oriented cobordism from Y_1 to Y_2 ,
- (2) S is a properly embedded oriented surface in X with $\partial S = -L_1 \cup L_2$, and
- (3) $\mathcal{A}$ is a properly embedded 1-manifold in S that divides S into two subsurfaces $S_{\mathbf{w}}$ and $S_{\mathbf{z}}$ that meet along $\mathcal{A}$ such that $\mathbf{w}_1, \mathbf{w}_2 \subseteq S_{\mathbf{w}}$ and $\mathbf{z}_1, \mathbf{z}_2 \subseteq S_{\mathbf{z}}$.

Note that this definition is slightly different from [15, Definition 4.5] and follows [30]. The equivalence of the two definitions is explained in [17, Section 2.3].

If $\mathcal{X} = (X, S, \mathcal{A})$ from $(Y_1, \mathbb{L}_1)$ to $(Y_2, \mathbb{L}_2)$ is a decorated link cobordism, then there is a well-defined cobordism $\mathcal{W}(\mathcal{X}) = (W, Z, [\xi])$ of sutured manifolds from $Y(\mathbb{L}_1)$ to $Y(\mathbb{L}_2)$, as we now describe. The 4-manifold W is defined by the formula

$$W := X \setminus N(S),$$

where $N(S)$ denotes a regular neighborhood of S , viewed as the unit normal disk bundle of S . The set Z is defined as the unit normal circle bundle of S , oriented as a submanifold of ∂W . Since S is an oriented surface, Z is a principal S^1 -bundle over S . According to Lutz [20] and Honda [111], the dividing set $\mathcal{A}$ uniquely determines an S^1 -invariant contact structure on Z with dividing set $\mathcal{A}$ on S , up to isotopy. The contact structure ξ is defined to be this S^1 -invariant contact structure. The link cobordism map

$$F_{\mathcal{X}}^J: \widehat{\text{HFL}}(Y_1, \mathbb{L}_1) \rightarrow \widehat{\text{HFL}}(Y_2, \mathbb{L}_2)$$

is defined to be the sutured cobordism map

$$F_{\mathcal{W}(\mathcal{X})}: \text{SFH}(Y_1(\mathbb{L}_1)) = \widehat{\text{HFL}}(Y_1, \mathbb{L}_1) \rightarrow \text{SFH}(Y_2(\mathbb{L}_2)) = \widehat{\text{HFL}}(Y_2, \mathbb{L}_2).$$

The second author [30] constructed another link cobordism map $F_{\mathcal{X}, \mathfrak{s}}^Z$, where $\mathfrak{s} \in \text{Spin}^c(X)$ is a Spin^c structure on X , that did not use the Honda–Kazez–Matić gluing map. It instead was defined by writing a link cobordism as a composition of elementary link cobordisms. The map $F_{\mathcal{X}, \mathfrak{s}}^Z$ is defined on a more general version of link Floer homology than $F_{\mathcal{X}}^J$, though it induces a map $\widehat{F}_{\mathcal{X}, \mathfrak{s}}^Z$ on the hat version. Let

$$\widehat{F}_{\mathcal{X}}^Z := \sum_{\mathfrak{s} \in \text{Spin}^c(X)} \widehat{F}_{\mathcal{X}, \mathfrak{s}}^Z.$$

It is not obvious that the maps $F_{\mathcal{X}}^J$ and $\widehat{F}_{\mathcal{X}}^Z$ agree.

The first author and Marengon [17] made some steps towards computing the maps $F_{\mathcal{X}}^J$ when $\mathcal{X}$ is an elementary link cobordism, though most of the computational results from [17] are still in terms of the Honda–Kazez–Matić gluing map, and hence it is challenging to directly compare the maps $F_{\mathcal{X}}^J$ and $\widehat{F}_{\mathcal{X}}^Z$. Nonetheless, combining several results from [17] with the results of this paper, we are able to prove the following:

Theorem 9.4 *Given a decorated link cobordism $\mathcal{X}$, we have $F_{\mathcal{X}}^J = \widehat{F}_{\mathcal{X}}^Z$.*

9.3 Elementary link cobordisms

In this section, we provide the following definition:

Definition 9.5 A decorated link cobordism $\mathcal{X} = (X, S, \mathcal{A})$: $(Y_1, \mathbb{L}_1) \rightarrow (Y_2, \mathbb{L}_2)$ is an *elementary link cobordism* if one of the following is satisfied:

(1) **Identity cobordism**

$$(Y_1, \mathbb{L}_1) = (Y_2, \mathbb{L}_2) = (Y, \mathbb{L}) \quad \text{and} \quad (X, S, \mathcal{A}) = (I \times Y, I \times L, I \times p),$$

where $p \subseteq L$ consists of exactly one point in each component of $L \setminus (w \cup z)$.

(2) **1–, 2–, or 3–handle attachment** The cobordism $(X, S, \mathcal{A})$ is obtained by attaching a 4–dimensional 1–, 2–, or 3–handle, with framed attaching sphere disjoint from $\mathbb{L}_1$.

(3) **0– or 4–handle attachment** The cobordism $\mathcal{X}$ is obtained by attaching a 4–dimensional 0–handle or 4–handle, viewed as a smooth 4–ball, that contains a standard disk intersecting the boundary of the 4–ball in an unknot, with dividing set consisting of a single arc on the disk.

(4) **Saddle cobordism** The cobordism $\mathcal{X}$ has underlying 4–manifold $X = I \times Y$ and a surface S such that projection to I induces a Morse function that has a single index 1 critical point that occurs in a w –region or a z –region and such that the dividing arcs all travel from $\mathbb{L}_1$ to $\mathbb{L}_2$.

(5) **Stabilization cobordism** The cobordism $\mathcal{X}$ has underlying 4–manifold $X = I \times Y$ and surface $S = I \times L$. Furthermore, exactly one arc of $\mathcal{A}$ goes from $\mathbb{L}_1$ to $\mathbb{L}_1$ or from $\mathbb{L}_2$ to $\mathbb{L}_2$. All other arcs are of the form $I \times \{p\}$ for various $p \in L$. A stabilization cobordism is *positive* if it adds two basepoints and is *negative* if it removes two basepoints.

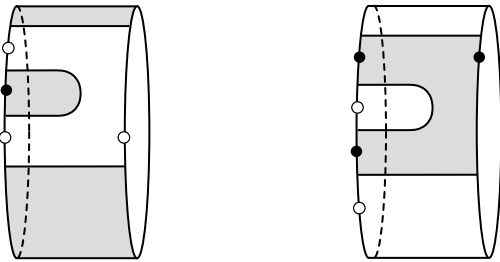


Figure 49: Two examples of stabilization cobordisms. The two surfaces are each of the form $I \times L$ and sit inside $I \times Y$. The shaded regions are the w regions, and the unshaded regions are the z regions.

A schematic of a saddle cobordism can be found in [Figure 55](#). Examples of stabilization cobordisms are shown in [Figure 49](#).

It is straightforward to see that an arbitrary link cobordism $\mathcal{X}$ such that $\pi_0(S) \rightarrow \pi_0(X)$ is a surjection can be decomposed into a sequence of link cobordisms that are each diffeomorphic to one of the elementary link cobordisms in the above list. We remark that the above list is overcomplete, in the following sense:

Remark 9.6 An elementary positive stabilization cobordism can be written as a composition of a 0–handle cobordism (adding an unknot with two basepoints) followed by a 1–handle, followed by a saddle cobordism. Similarly an elementary negative stabilization cobordism can be written as a composition of a saddle cobordism, followed by a 3–handle and a 4–handle.

9.4 Link triple diagrams and contact structures

Definition 9.7 A link triple diagram

$$(\Sigma, \alpha, \beta, \gamma, w, z)$$

is a Heegaard triple $(\Sigma, \alpha, \beta, \gamma)$ with $2(n - g(\Sigma) + 1)$ basepoints $w \cup z \subseteq \Sigma$, where $n = |\alpha| = |\beta| = |\gamma|$, such that each component of $\Sigma \setminus \tau$ for $\tau \in \{\alpha, \beta, \gamma\}$ is planar and contains exactly one w basepoint and exactly one z basepoint.

Given a link triple diagram $\mathcal{T} = (\Sigma, \alpha, \beta, \gamma, w, z)$, we can construct a decorated link cobordism

$$\mathcal{X}_{\mathcal{T}} = (X_{\mathcal{T}}, S_{\mathcal{T}}, \mathcal{A}_{\mathcal{T}}),$$

as follows. The 4-manifold $X_{\mathcal{T}}$ is constructed as in [25], by the formula

$$X_{\mathcal{T}} := (\Delta \times \Sigma) \cup (e_{\alpha} \times U_{\alpha}) \cup (e_{\beta} \times U_{\beta}) \cup (e_{\gamma} \times U_{\gamma}) / \sim.$$

Here, the handlebody U_{τ} for $\tau \in \{\alpha, \beta, \gamma\}$ is obtained by attaching 3-dimensional 2-handles to $\Sigma \times I$ along $\tau \times \{0\}$ and filling in the resulting sphere boundary components with 3-dimensional 3-handles.

We obtain the surface $S_{\mathcal{T}}$ as follows. Pick Morse functions f_{α} , f_{β} , and f_{γ} on U_{α} , U_{β} , and U_{γ} that induce the attaching curves α , β , and γ , respectively. By concatenating the ascending flow lines passing through the basepoints in $\mathbf{w}$ and $\mathbf{z}$, we get a collection of $|\mathbf{w}|$ arcs K_{α} , K_{β} , and K_{γ} in each of U_{α} , U_{β} , and U_{γ} , respectively. Each arc has exactly one endpoint in $\mathbf{w}$ and one endpoint in $\mathbf{z}$, and we orient it from $\mathbf{w}$ to $\mathbf{z}$. Then the surface $S_{\mathcal{T}}$ is defined as

$$S_{\mathcal{T}} := (\Delta \times (-\mathbf{w} \cup \mathbf{z})) \cup (e_{\alpha} \times K_{\alpha}) \cup (e_{\beta} \times K_{\beta}) \cup (e_{\gamma} \times K_{\gamma}).$$

Finally, we describe the dividing set $\mathcal{A}_{\mathcal{T}}$ on $S_{\mathcal{T}}$. For $\tau \in \{\alpha, \beta, \gamma\}$, let $p_{\tau} \subseteq K_{\tau}$ denote a collection of points obtained by picking a single point in each arc of K_{τ} . Then

$$\mathcal{A}_{\mathcal{T}} := (e_{\alpha} \times p_{\alpha}) \cup (e_{\beta} \times p_{\beta}) \cup (e_{\gamma} \times p_{\gamma})$$

is a dividing set on $S_{\mathcal{T}}$.

Given a link triple diagram $\mathcal{T} = (\Sigma, \alpha, \beta, \gamma, \mathbf{w}, \mathbf{z})$, we naturally obtain a sutured Heegaard triple $\mathcal{T}_0 = (\Sigma_0, \alpha, \beta, \gamma)$, where $\Sigma_0 = \Sigma \setminus N(\mathbf{w} \cup \mathbf{z})$. Let $\mathcal{W}_{\mathcal{T}_0} = (W_{\mathcal{T}_0}, Z_{\mathcal{T}_0}, [\xi_{\mathcal{T}_0}])$ denote the associated sutured cobordism.

Proposition 9.8 *Let $\mathcal{T} = (\Sigma, \alpha, \beta, \gamma, \mathbf{w}, \mathbf{z})$ be a link triple diagram, and $\mathcal{T}_0 = (\Sigma_0, \alpha, \beta, \gamma)$ the corresponding sutured triple diagram. Then $Z_{\mathcal{T}_0}$ is the unit normal circle bundle of $S_{\mathcal{T}}$, and there is a projection map $\pi: Z_{\mathcal{T}_0} \rightarrow S_{\mathcal{T}}$ with fiber S^1 . The contact structure $\xi_{\mathcal{T}_0}$ on $Z_{\mathcal{T}_0}$ described in Section 7 is equivalent to the S^1 -invariant contact structure on $Z_{\mathcal{T}_0}$ with respect to the dividing set $\mathcal{A}_{\mathcal{T}}$ and the projection map π .*

Proof Let us write ξ_{S^1} for the S^1 -invariant contact structure on the unit normal circle bundle $\text{SN}(S_{\mathcal{T}})$ of $S_{\mathcal{T}}$. The proof of the proposition will be to describe a convex decomposition of $(\text{SN}(S_{\mathcal{T}}), \xi_{S^1})$ into the disjoint union of the contact manifolds (Z_0, ξ_0) , $(Z_{\alpha}, \xi_{\alpha})$, (Z_{β}, ξ_{β}) , and $(Z_{\gamma}, \xi_{\gamma})$, whose union is $(Z_{\mathcal{T}_0}, \xi_{\mathcal{T}_0})$ by definition. Note that $S_{\mathcal{T}}$ has no closed components, so $\text{SN}(S_{\mathcal{T}})$ is diffeomorphic to $S_{\mathcal{T}} \times S^1$, with the map π given by projection onto the first factor.

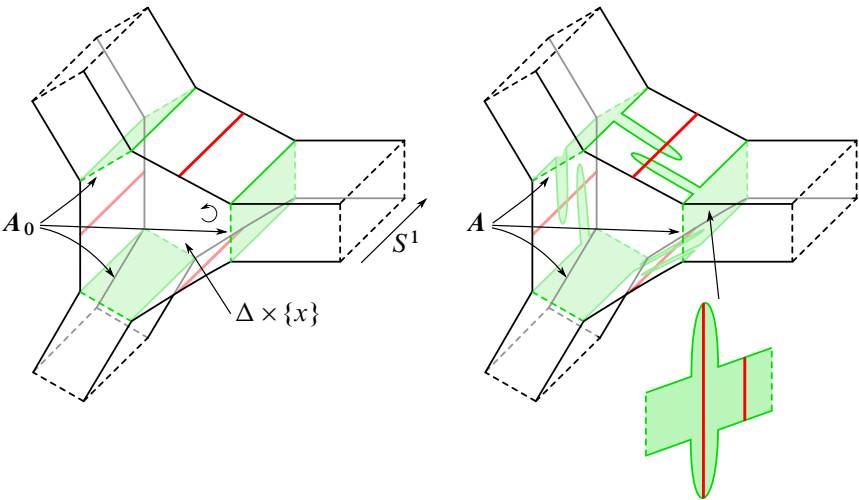


Figure 50: The annuli A_0 in $S_{\mathcal{T}} \times S^1$ (left) and the annuli A obtained by performing finger moves along the boundaries toward the dividing set on $\partial S_{\mathcal{T}} \times S^1$ (right). We prove that the dividing set on each annulus in A is as shown in the bottom. The hexagonal region in the middle corresponds to a component of $\Delta \times \{x\}$ for a basepoint $x \in w \cup z$. The orientation of $\Delta \times \{x\}$ is shown, and we take the product orientation on $S_{\mathcal{T}} \times S^1$ in the picture.

We will decompose $\text{SN}(S_{\mathcal{T}}) \approx S_{\mathcal{T}} \times S^1$ along $3|\partial \Sigma_0|$ convex annuli. To construct the annuli, it is convenient to view Δ as a smooth 2-disk and view the edges e_{α} , e_{β} , and e_{γ} as being closed, disjoint subintervals of $\partial \Delta$. We let A_0 denote the union of annuli of the form $e_{\tau} \times \{w\} \times S^1$ and $e_{\tau} \times \{z\} \times S^1$ inside $S_{\mathcal{T}} \times S^1$ for $\tau \in \{\alpha, \beta, \gamma\}$. Note that we cannot decompose along the annuli in A_0 , since their boundaries are disjoint from the dividing set, and hence we cannot use Legendrian realization to ensure that ∂A_0 is Legendrian and A_0 is convex. Instead, we perform a finger move along each boundary component of each annulus in A_0 until it intersects the dividing set. Let A denote the resulting collection of annuli. The configuration of the annuli A_0 and A is shown in [Figure 50](#).

We can perturb the annuli in A so that they are convex with Legendrian boundary (note this requires using Legendrian realization along $\partial S_{\mathcal{T}} \times S^1$, and hence involves replacing ξ_{S^1} with an equivalent contact structure that is no longer S^1 -invariant near the boundary).

We now claim that the dividing sets on the annuli in A are as in the bottom of [Figure 50](#), consisting of two arcs that go from one boundary component of the annulus to the

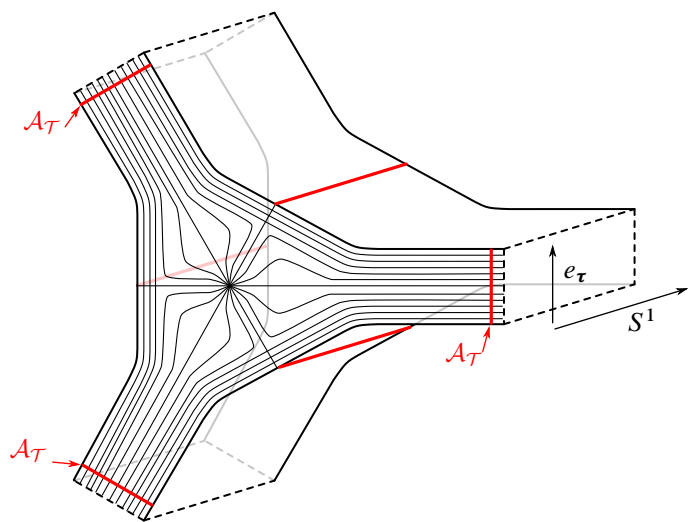


Figure 51: The characteristic foliation of $S_{\mathcal{T}} \times \{p\}$ of an S^1 -invariant contact 1-form on $S_{\mathcal{T}} \times S^1$ with dividing set $\mathcal{A}_{\mathcal{T}}$.

other and which do not wind around the annulus. To see this, we first claim that, for an appropriately chosen S^1 -invariant contact 1-form, we can embed a neighborhood of each annulus inside an e_{τ} -invariant contact structure on $e_{\tau} \times I \times S^1$ that has dividing set on $(\partial e_{\tau}) \times I \times S^1$ equal to $(\partial e_{\tau}) \times \{\frac{1}{2}\} \times S^1$. To do this, we recall that an S^1 -invariant contact 1-form on $S_{\mathcal{T}} \times S^1$ can be defined as $\beta + f \cdot d\theta$, where β is a 1-form on $S_{\mathcal{T}}$ and $f: S_{\mathcal{T}} \rightarrow \mathbb{R}$ is a function that is zero exactly on $\mathcal{A}_{\mathcal{T}}$. It is straightforward to write down conditions for such a 1-form to be a contact form on $S_{\mathcal{T}} \times S^1$. Since the contact form $\beta + f \cdot d\theta$ is essentially determined by the characteristic foliation $\ker \beta$ on $S_{\mathcal{T}} \times \{p\}$, we will focus on constructing a singular foliation with the appropriate dividing set that is invariant under translation along a nonvanishing vector field (thought

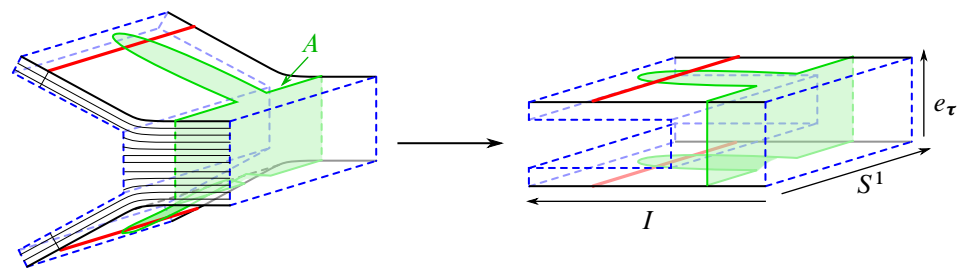


Figure 52: A neighborhood of an annulus $A \in \mathcal{A}$ in $S_{\mathcal{T}} \times S^1$ which embeds into an e_{τ} -invariant contact structure on $e_{\tau} \times I \times S^1$.

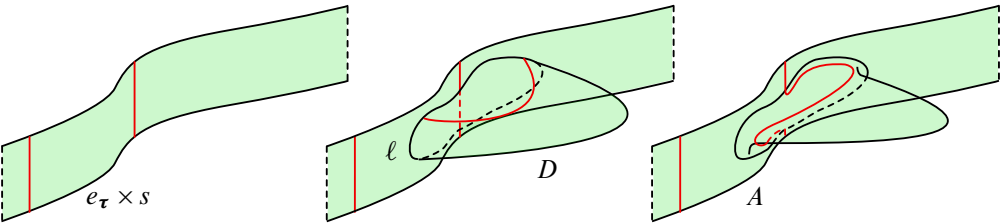


Figure 53: Attaching a convex disk to move the annulus $e_{\tau} \times s \subseteq e_{\tau} \times I \times S^1$ to the position of an annulus $A \in \mathcal{A}$. On the left is (a C^0 –small perturbation of) the annulus $e_{\tau} \times s$. In the middle is a convex disk with Legendrian boundary ℓ with $\text{tb}(\ell) = -1$. On the right is a convex annulus that we can take to be A , obtained by edge rounding along ℓ .

of as $\partial/\partial e_{\tau}$) in a neighborhood of each annulus in $\mathcal{A}$. It is an elementary, though somewhat tedious, exercise to explicitly construct an appropriate contact 1–form by picking β and f appropriately, so we will leave that step to the reader. An example of an appropriately chosen characteristic foliation on $S_{\mathcal{T}} \times \{p\}$ is shown in Figure 51.

We can choose a neighborhood of an annulus in $\mathcal{A}$ that embeds into an e_{τ} –invariant contact structure on $e_{\tau} \times I \times S^1$ as in Figure 52. Note that, inside $e_{\tau} \times I \times S^1$, the annulus A pictured on the right of Figure 52 is isotopic, relative to its boundary, to a surface of the form $e_{\tau} \times s$, for a Legendrian s in $(\partial e_{\tau}) \times I \times S^1$. The characteristic

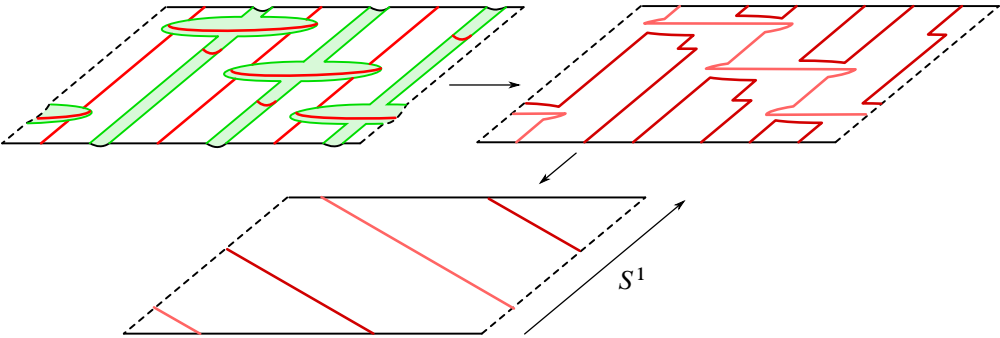


Figure 54: Rounding corners after cutting $S_{\mathcal{T}} \times S^1$ along the annuli in $\mathcal{A}$. Shown is the dividing set on the boundary of a solid torus component of $(S_{\mathcal{T}} \times S^1) \setminus \mathcal{A}$ corresponding to $\Delta \times \{x\} \times S^1$ for a basepoint $x \in \mathbf{w} \cup \mathbf{z}$. On the top left, we show the dividing before rounding the Legendrian corners. On the top right, we show the result of rounding corners. On the bottom, we show the result of isotoping the dividing set. We view Z_0 as being “below” the surface shown.

foliation and dividing set on $e_{\tau} \times s$ are the same as on the annulus shown in Figure 31. Namely, the characteristic foliation consists of horizontal leaves lying in $\{t\} \times s$, as well as two vertical singular sets of the form $e_{\tau} \times \{p\}$, for two points p in s . The dividing set on $e_{\tau} \times s$ consists of two vertical arcs, as well.

Note that if we could show that A and $e_{\tau} \times s$ were isotopic through convex surfaces, we would be done, since the dividing set on A would be isotopic to the one on $e_{\tau} \times s$, which is what we are trying to show. This is somewhat geometrically hard to prove, so we argue as follows. Perturb $e_{\tau} \times s$ slightly so that there is a Legendrian loop ℓ intersecting one of the dividing arcs twice. This does not change the isotopy type of the dividing set. Let $D_0 \subseteq e_{\tau} \times s$ be the disk bounded by ℓ . Now take a convex disk D in $e_{\tau} \times I \times S^1$ with $D \cap (e_{\tau} \times s) = \ell$. The dividing set on D consists of a single arc, since we can assume D lies in a tight contact ball. We can replace $e_{\tau} \times s$ by $((e_{\tau} \times s) \setminus D_0) \cup D$ and then round along the Legendrian corner ℓ . When we do this, we do not change the isotopy type of the dividing set, and we can move $e_{\tau} \times s$ to A via a sequence of such moves. The move is shown in Figure 53. This establishes that the annulus A has a dividing set isotopic to the one shown in Figure 50.

Having determined the dividing sets along the convex annuli in A , we can cut along the annuli in A , then round the Legendrian corners. The dividing set on the sutured manifold corresponding to $\Delta \times \partial \Sigma$ is shown in Figure 54. As a sutured manifold, this is the same as (Z_0, γ_0) . Similarly, it is easy to see that rounding the Legendrian corners on the other 3 pieces yields the sutured manifolds $(Z_{\alpha}, \gamma_{\alpha})$, $(Z_{\beta}, \gamma_{\beta})$, and $(Z_{\gamma}, \gamma_{\gamma})$. It follows that $\text{SN}(S_{\mathcal{T}}) \approx Z_{\mathcal{T}_0}$. On the other hand, we note that ξ_{S^1} is tight, since the dividing set on $S_{\mathcal{T}}$ contains no contractible components. It follows that ξ_{S^1} restricts to tight contact structures on (Z_0, γ_0) , $(Z_{\alpha}, \gamma_{\alpha})$, $(Z_{\beta}, \gamma_{\beta})$, and $(Z_{\gamma}, \gamma_{\gamma})$. However, each of these four sutured manifolds are product disk decomposable, so, up to equivalence, there is a unique tight contact structure on each one. Hence the restrictions of ξ_{S^1} are equivalent to ξ_0 , ξ_{α} , ξ_{β} , and ξ_{γ} , respectively. Since $\xi_{\alpha, \beta, \gamma}$ is constructed by gluing together ξ_0 , ξ_{α} , ξ_{β} , and ξ_{γ} , it follows that ξ_{S^1} and $\xi_{\alpha, \beta, \gamma}$ are equivalent on $Z_{\mathcal{T}_0}$. $\square$

9.5 Saddle cobordisms and link triple diagrams

In this section, we review the construction of saddle cobordism maps [30, Section 6]. As an important step towards proving Theorem 9.4, we show that the maps from [30] and [15] agree for such cobordisms.

Definition 9.9 Suppose that Y is a 3-manifold containing an oriented multibased link $\mathbb{L} = (L, \mathbf{w}, \mathbf{z})$. We say that an oriented square $B \subseteq Y$ is a β -band for a multibased link $\mathbb{L}$ in Y if B is smoothly embedded in Y , it is identified with $[-1, 1] \times [-1, 1]$, and

- (1) $B \cap L = [-1, 1] \times \{-1, 1\}$,
- (2) the boundary orientation of B agrees with the orientation of $-L$,
- (3) $B \cap (\mathbf{w} \cup \mathbf{z}) = \emptyset$, and both ends of B are in regions of $L \setminus (\mathbf{w} \cup \mathbf{z})$ that go from $\mathbf{z}$ to $\mathbf{w}$.

Note that if B is a β -band for the link $\mathbb{L}$ in Y , then there is a well-defined multibased link

$$\mathbb{L}(B) = (L(B), \mathbf{w}, \mathbf{z})$$

obtained by band surgery on B . The following is [30, Definition 6.4]:

Definition 9.10 The link triple diagram

$$(\Sigma, \alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_n, \beta'_1, \dots, \beta'_n, \mathbf{w}, \mathbf{z})$$

is subordinate to the β -band B if the following hold:

- (1) $(\Sigma_0, \alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_{n-1})$, where $\Sigma_0 = \Sigma \setminus N(\mathbf{w} \cup \mathbf{z})$, is a diagram for the sutured manifold $Y(\mathbb{L}) \setminus N(B)$.
- (2) $(\Sigma, \alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_n, \mathbf{w}, \mathbf{z})$ is a diagram for $(Y, \mathbb{L})$.
- (3) $\beta'_1, \dots, \beta'_{n-1}$ are small Hamiltonian translates of the curves $\beta_1, \dots, \beta_{n-1}$.
- (4) The curve β'_n is induced by the band B , and $(\Sigma, \alpha_1, \dots, \alpha_n, \beta'_1, \dots, \beta'_n, \mathbf{w}, \mathbf{z})$ is a diagram for $(Y, \mathbb{L}(B))$.

Notice that, if we ignore the basepoints $\mathbf{w}$ and $\mathbf{z}$, then the curve β'_n is related to β_n by a sequence of handle slides and isotopies. It follows that the 4-manifold $X_{\alpha, \beta, \beta'}$ induced by a Heegaard triple subordinate to a β -band is diffeomorphic to $I \times Y$ with a neighborhood of $\{\frac{1}{2}\} \times U_{\beta}$ removed.

The band B induces a saddle cobordism $S(B) \subseteq I \times Y$ from L to $L(B)$. The surface $S(B)$ is obtained by rounding the corners of the surface

$$S_0(B) := ([0, \tfrac{1}{2}] \times L) \cup (\{\tfrac{1}{2}\} \times B) \cup ([\tfrac{1}{2}, 1] \times L(B)).$$

We note that $S(B)$ has two natural choices of dividing sets. To construct them, pick points $\mathbf{p} \subseteq L \setminus (\mathbf{w} \cup \mathbf{z})$ and $\mathbf{q} \subseteq L(B) \setminus (\mathbf{w} \cup \mathbf{z})$ such that the following hold:

- (1) Each component of $L \setminus (\mathbf{w} \cup \mathbf{z})$ contains exactly one point of $\mathbf{p}$, and each component of $L(B) \setminus (\mathbf{w} \cup \mathbf{z})$ contains exactly one point of $\mathbf{q}$.
- (2) If $C \subseteq L \setminus (\mathbf{w} \cup \mathbf{z})$ is a component disjoint from B , then $\mathbf{p} \cap C = \mathbf{q} \cap C$.
- (3) If C is a component of $L \setminus (\mathbf{w} \cup \mathbf{z})$ that intersects B , and $\{p\} = C \cap \mathbf{p}$, then $p \in B$. Similarly, if C' is a component of $L(B) \setminus (\mathbf{w} \cup \mathbf{z})$ that intersects B , and $\{q\} = C' \cap \mathbf{q}$, then $q \in B$.

A dividing $\mathcal{A}_0$ on $S(B) \setminus (\{\frac{1}{2}\} \times B)$ is then specified by the equation

$$\mathcal{A}_0 := ([0, \tfrac{1}{2}] \times \mathbf{p}) \cup ([\tfrac{1}{2}, 1] \times \mathbf{q}).$$

There are two natural ways to extend the dividing set $\mathcal{A}_0$ to B . We write $\mathcal{A}_{\mathbf{w}}$ and $\mathcal{A}_{\mathbf{z}}$ for the two possible extensions, as shown in Figure 55.

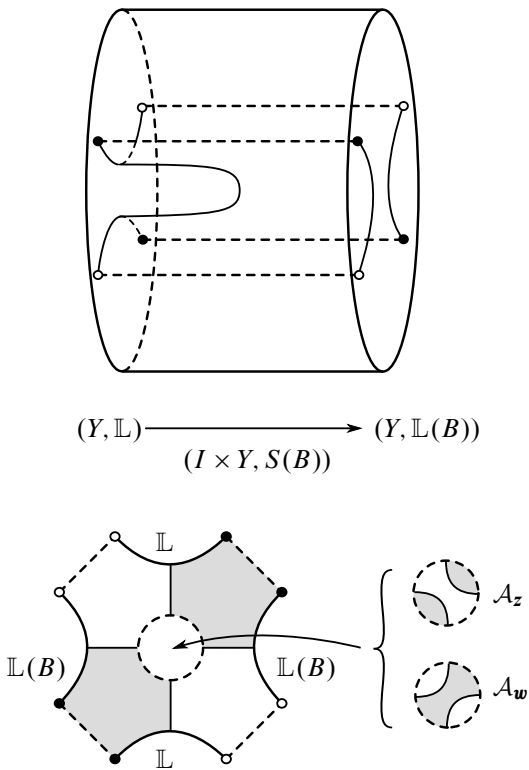


Figure 55: A portion of the surface $S(B) \subseteq I \times Y$, as well as the two dividing sets $\mathcal{A}_{\mathbf{w}}$ and $\mathcal{A}_{\mathbf{z}}$ on $S(B)$. The $\mathbf{w}$ basepoints are shown as solid dots, while the $\mathbf{z}$ basepoints are open dots. The $\mathbf{w}$ regions are shown as shaded, the $\mathbf{z}$ regions are unshaded.

Lemma 9.11 *If $(\Sigma, \alpha, \beta, \beta', w, z)$ is a triple subordinate to a β -band B , then*

$$(\Sigma, \beta, \beta', w, z)$$

represents an unlink $\mathbb{U}$ in $(S^1 \times S^2)^{\#g(\Sigma)}$, where all components of $\mathbb{U}$ have two basepoints, except for one component that has four basepoints. With respect to each of the Maslov gradings gr_w and gr_z , there is a top-graded generator of $\widehat{\text{HFL}}(\Sigma, \beta, \beta', w, z)$, for which we write $\Theta_{\beta, \beta'}^w$ and $\Theta_{\beta, \beta'}^z$.

Proof For $i \in \{1, \dots, n-1\}$, the curves β_i and β'_i are Hamiltonian translates of each other and hence determine a 2-sphere in the manifold $U_\beta \cup_\Sigma U_{\beta'}$. After surgering all of these out, we are left with $|w| - 2$ copies of S^3 , each containing a doubly based unknot, as well as one copy of S^3 with an unknot containing four basepoints. After surgering all of these 2-spheres out, it is easy to see $\widehat{\text{HFL}}$ is generated by two elements, one of which is in $(\text{gr}_w, \text{gr}_z)$ -grading $(+\frac{1}{2}, -\frac{1}{2})$ and one of which is in grading $(-\frac{1}{2}, +\frac{1}{2})$. The effect of undoing the surgeries we did on the 2-spheres corresponds to adding back in a collection of 1-handles, which clearly preserves the property of having a top gr_w -graded element and a top gr_z -graded element. $\square$

Lemma 9.12 *Consider the link cobordism $\mathcal{X} = (X, S, \mathcal{A})$ from the empty set to an unlink in $(S^1 \times S^2)^{\#k}$ constructed by setting X to be a 4-dimensional genus k handlebody, S to be a collection of n standardly embedded disks in X intersecting $\partial X \cong (S^1 \times S^2)^{\#k}$ in n disjoint unknots, and letting $\mathcal{A}$ consist of a single arc on each component of S except for one component of S where $\mathcal{A}$ consists of two arcs. Then*

$$F_{\mathcal{X}}^J(1) = \widehat{F}_{\mathcal{X}}^Z(1) = \begin{cases} \Theta^w & \text{if } \chi(S_w) = \chi(S_z) + 1, \\ \Theta^z & \text{if } \chi(S_w) = \chi(S_z) - 1, \end{cases}$$

where 1 denotes the generator of $\widehat{\text{HFL}}(\emptyset, \emptyset) \cong \mathbb{F}_2$.

Proof We decompose the cobordism $\mathcal{X}$ into a composition of elementary link cobordisms (Definition 9.5). Write $\mathcal{X} = \mathcal{X}_4 \circ \mathcal{X}_3 \circ \mathcal{X}_2 \circ \mathcal{X}_1$, where

- $\mathcal{X}_1$ is a 0-handle cobordism which adds a doubly based unknot in S^3 ;
- $\mathcal{X}_2$ is a stabilization cobordism which adds two basepoints to the unknot in S^3 ;
- $\mathcal{X}_3$ consists of $(n-1)$ 0-handles, each adding a doubly based unknot;
- $\mathcal{X}_4$ consists of $(k+n-1)$ 1-handles.

If $\mathbb{U} \subseteq S^3$ is a doubly based unknot, then $\widehat{\text{HFL}}(S^3, \mathbb{U}) \cong \mathbb{F}_2$. Noting that the 0–handle maps are nonzero, since they can be canceled topologically, using multiplicativity of link Floer homology under disjoint unions, it follows that

$$F^J_{\mathcal{X}_i} = \widehat{F}^Z_{\mathcal{X}_i}$$

for $i = 1, 3$.

If $\mathbb{U}'$ is an unknot in S^3 with four basepoints, then, by Lemma 9.11, we have $\widehat{\text{HFL}}(S^3, \mathbb{U}') \cong \mathbb{F}_2 \oplus \mathbb{F}_2$. Furthermore, $\widehat{\text{HFL}}(S^3, \mathbb{U}')$ is generated by two elements, $\Theta^{\mathbf{w}}$ and $\Theta^{\mathbf{z}}$, which are distinguished by grading. The element $\Theta^{\mathbf{w}}$ has $(\text{gr}_{\mathbf{w}}, \text{gr}_{\mathbf{z}})$ bigrading $(+\frac{1}{2}, -\frac{1}{2})$, while $\Theta^{\mathbf{z}}$ has bigrading $(-\frac{1}{2}, +\frac{1}{2})$.

It follows immediately from the definition of the maps in [30] that

$$\widehat{F}^Z_{\mathcal{X}_2}(1) = \begin{cases} \Theta^{\mathbf{w}} & \text{if } \chi(S_{\mathbf{w}}) = \chi(S_{\mathbf{z}}) + 1, \\ \Theta^{\mathbf{z}} & \text{if } \chi(S_{\mathbf{w}}) = \chi(S_{\mathbf{z}}) - 1, \end{cases}$$

since the cobordism map for a stabilization cobordism is defined using the quasistabilization map [30, Section 3.2]. On the other hand, a straightforward functoriality argument shows that $F^J_{\mathcal{X}_2}$ must be nonzero. By [17, Theorem 5.18], it follows that $F^J_{\mathcal{X}_2}(1)$ has the same $\text{gr}_{\mathbf{w}}$ and $\text{gr}_{\mathbf{z}}$ grading as $\widehat{F}^Z_{\mathcal{X}_2}(1)$. Since $\widehat{\text{HFL}}(S^3, \mathbb{U}')$ is a 2–dimensional vector space over $\mathbb{F}_2$, it follows that $F^J_{\mathcal{X}_2} = \widehat{F}^Z_{\mathcal{X}_2}$.

Finally, the two cobordism map constructions use the same 4–dimensional handle attachment maps, so

$$F^J_{\mathcal{X}_4} = \widehat{F}^Z_{\mathcal{X}_4}.$$

Furthermore, as in Lemma 9.11, the 1–handle maps preserve the top-graded elements $\Theta^{\mathbf{w}}$ and $\Theta^{\mathbf{z}}$. Composing all the maps, the claim now follows. $\square$

In [30], the link cobordism maps for

$$\mathcal{X}_{\mathbf{w}} = (I \times Y, S(B), \mathcal{A}_{\mathbf{w}}) \quad \text{and} \quad \mathcal{X}_{\mathbf{z}} = (I \times Y, S(B), \mathcal{A}_{\mathbf{z}})$$

are defined to be

$$(9-1) \quad \widehat{F}^Z_{\mathcal{X}_{\mathbf{w}}}(-) := F_{\alpha, \beta, \beta'}(- \otimes \Theta^{\mathbf{z}}_{\beta, \beta'}) \quad \text{and} \quad \widehat{F}^Z_{\mathcal{X}_{\mathbf{z}}}(-) := F_{\alpha, \beta, \beta'}(- \otimes \Theta^{\mathbf{w}}_{\beta, \beta'}).$$

Lemma 9.13 *For the decorated saddle cobordisms $\mathcal{X}_{\mathbf{w}}$ and $\mathcal{X}_{\mathbf{z}}$, defined above, we have $F^J_{\mathcal{X}_{\mathbf{w}}} = \widehat{F}^Z_{\mathcal{X}_{\mathbf{w}}}$ and $F^J_{\mathcal{X}_{\mathbf{z}}} = \widehat{F}^Z_{\mathcal{X}_{\mathbf{z}}}$.*

Proof The key observation is that if $(\Sigma, \alpha, \beta, \beta', w, z)$ is a Heegaard triple subordinate to a β -band, then the 4-manifold $X_{\alpha, \beta, \beta'}$ is equal to $I \times Y$ with a neighborhood of $\{\frac{1}{2}\} \times U_{\beta}$ removed, and the surface with divides

$$(S_{\alpha, \beta, \beta'}, \mathcal{A}_{\alpha, \beta, \beta'}) = (S(B) \setminus (\{\frac{1}{2}\} \times B), \mathcal{A}_0).$$

There is a cobordism $X_{\beta}: \emptyset \rightarrow Y_{\beta, \beta'}$ consisting of 0-handles and 1-handles. Inside X_{β} , there is a surface S_0 that consists of $|w| - 1$ disks. We note that $|w| - 2$ of the disks have boundary equal to a doubly based unknot in $Y_{\beta, \beta'}$, but one disk has boundary equal to an unknot with four basepoints. It is clear that if we fill in the $Y_{\beta, \beta'}$ boundary of the link cobordism $(X_{\alpha, \beta, \beta'}, S_{\alpha, \beta, \beta'})$ with (X_{β}, S_0) , then we obtain the (undecorated) link cobordism $(I \times Y, S(B))$. On the other hand, as described above, there are two natural dividing sets on S_0 , shown in Figure 55. Define

$$\mathcal{A}_{w,0} = S_0 \cap \mathcal{A}_w \quad \text{and} \quad \mathcal{A}_{z,0} = S_0 \cap \mathcal{A}_z.$$

Using Lemma 9.12, we have that

$$F_{X_{\beta}, S_0, \mathcal{A}_{w,0}}^J(1) = \Theta_{\beta, \beta'}^z \quad \text{and} \quad F_{X_{\beta}, S_0, \mathcal{A}_{z,0}}^J(1) = \Theta_{\beta, \beta'}^w,$$

as maps from $\widehat{\text{HFL}}(\emptyset, \emptyset) \cong \mathbb{F}_2$ to $\widehat{\text{HFL}}(\Sigma, \beta, \beta', w, z)$.

Using Theorem 7.1 and Proposition 9.8, we know that the sutured link cobordism map $F_{\mathcal{X}_{\alpha, \beta, \beta'}}^J$ for the decorated link cobordism

$$\mathcal{X}_{\alpha, \beta, \beta'} = (X_{\alpha, \beta, \beta'}, S_{\alpha, \beta, \beta'}, \mathcal{A}_{\alpha, \beta, \beta'})$$

is the map $F_{\alpha, \beta, \beta'}$ that counts holomorphic triangles on the Heegaard triple $(\Sigma, \alpha, \beta, \beta')$. It follows from the composition law that

$$F_{\mathcal{X}_z}^J(-) = F_{\alpha, \beta, \beta'}(- \otimes F_{X_{\beta}, S_0, \mathcal{A}_{z,0}}(1)) = F_{\alpha, \beta, \beta'}(- \otimes \Theta_{\beta, \beta'}^w) = \hat{F}_{\mathcal{X}_z}^Z(-).$$

The result for $F_{\mathcal{X}_w}^J$ follows similarly. □

9.6 Proof of Theorem 9.4

We can now prove that the link cobordism map constructions from [15] and [30] agree:

Proof of Theorem 9.4 Given an arbitrary decorated link cobordism $\mathcal{X} = (X, S, \mathcal{A})$, it is easy to see that one can decompose $\mathcal{X}$ into a sequence of link cobordisms which are each diffeomorphic to an elementary link cobordism (Definition 9.5). Using the composition law, it remains to verify the claim for each type of elementary link

cobordism. The maps obviously agree for identity cobordisms, and 4–dimensional 0–, 1–, 2–, 3–, or 4–handle cobordisms. By [Lemma 9.13](#), they agree for decorated saddle cobordisms. Finally, as in [Remark 9.6](#), a stabilization cobordism can be decomposed into a composition of the other elementary cobordisms, and hence, having established the claim for the other types of elementary cobordisms, it follows as well for stabilization cobordisms. $\square$

References

- [1] **S Akbulut**, *Variations on Fintushel–Stern knot surgery on 4–manifolds*, Turkish J. Math. 26 (2002) 81–92 [MR](#)
- [2] **T Eteġü, B Özbaġcı**, *Partial open book decompositions and the contact class in sutured Floer homology*, Turkish J. Math. 33 (2009) 295–312 [MR](#)
- [3] **J Etnyre**, *Convex surfaces in contact geometry: class notes*, course notes (2004) Available at <http://people.math.gatech.edu/~etnyre/preprints/papers/surfaces.pdf>
- [4] **S Friedl, A Juhász, J Rasmussen**, *The decategorification of sutured Floer homology*, J. Topol. 4 (2011) 431–478 [MR](#)
- [5] **D Gabai**, *Foliations and the topology of 3–manifolds*, J. Differential Geom. 18 (1983) 445–503 [MR](#)
- [6] **E Giroux**, *Convexit  en topologie de contact*, Comment. Math. Helv. 66 (1991) 637–677 [MR](#)
- [7] **E Giroux**, *Structures de contact en dimension trois et bifurcations des feuilletages de surfaces*, Invent. Math. 141 (2000) 615–689 [MR](#)
- [8] **E Giroux**, *G om trie de contact: de la dimension trois vers les dimensions sup rieures*, from “Proceedings of the International Congress of Mathematicians, II” (T Li, editor), Higher Ed., Beijing (2002) 405–414 [MR](#)
- [9] **K Hendricks, C Manolescu**, *Involutive Heegaard Floer homology*, Duke Math. J. 166 (2017) 1211–1299 [MR](#)
- [10] **K Honda**, *On the classification of tight contact structures, I*, Geom. Topol. 4 (2000) 309–368 [MR](#)
- [11] **K Honda**, *On the classification of tight contact structures, II*, J. Differential Geom. 55 (2000) 83–143 [MR](#)
- [12] **K Honda, WH Kazez, G Mati **, *Contact structures, sutured Floer homology and TQFT*, preprint (2008) [arXiv](#)
- [13] **K Honda, WH Kazez, G Mati **, *The contact invariant in sutured Floer homology*, Invent. Math. 176 (2009) 637–676 [MR](#)
- [14] **A Juhász**, *Holomorphic discs and sutured manifolds*, Algebr. Geom. Topol. 6 (2006) 1429–1457 [MR](#)

- [15] **A Juhász**, *Cobordisms of sutured manifolds and the functoriality of link Floer homology*, Adv. Math. 299 (2016) 940–1038 [MR](#)
- [16] **A Juhász**, **M Marengon**, *Concordance maps in knot Floer homology*, Geom. Topol. 20 (2016) 3623–3673 [MR](#)
- [17] **A Juhász**, **M Marengon**, *Computing cobordism maps in link Floer homology and the reduced Khovanov TQFT*, Selecta Math. 24 (2018) 1315–1390 [MR](#)
- [18] **A Juhász**, **D P Thurston**, **I Zemke**, *Naturality and mapping class groups in Heegaard Floer homology*, preprint (2012) [arXiv](#) To appear under Mem. Amer. Math. Soc.
- [19] **R A Litherland**, *Deforming twist-spun knots*, Trans. Amer. Math. Soc. 250 (1979) 311–331 [MR](#)
- [20] **R Lutz**, *Structures de contact sur les fibrés principaux en cercles de dimension trois*, Ann. Inst. Fourier (Grenoble) 27 (1977) 1–15 [MR](#)
- [21] **B Ozbagci**, *Contact handle decompositions*, Topology Appl. 158 (2011) 718–727 [MR](#)
- [22] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and knot invariants*, Adv. Math. 186 (2004) 58–116 [MR](#)
- [23] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and three-manifold invariants: properties and applications*, Ann. of Math. 159 (2004) 1159–1245 [MR](#)
- [24] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. 159 (2004) 1027–1158 [MR](#)
- [25] **P Ozsváth**, **Z Szabó**, *Holomorphic triangles and invariants for smooth four-manifolds*, Adv. Math. 202 (2006) 326–400 [MR](#)
- [26] **P Ozsváth**, **Z Szabó**, *Holomorphic disks, link invariants and the multi-variable Alexander polynomial*, Algebr. Geom. Topol. 8 (2008) 615–692 [MR](#)
- [27] **J A Rasmussen**, *Floer homology and knot complements*, PhD thesis, Harvard University (2003) [MR](#) [arXiv](#)
- [28] **R Zarev**, *Joining and gluing sutured Floer homology*, preprint (2010) [arXiv](#)
- [29] **I Zemke**, *Duality and mapping tori in Heegaard Floer homology*, preprint (2018) [arXiv](#)
- [30] **I Zemke**, *Link cobordisms and functoriality in link Floer homology*, J. Topol. 12 (2019) 94–220 [MR](#)

Mathematical Institute, University of Oxford
Oxford, United Kingdom

Department of Mathematics, Princeton University
Princeton, NJ, United States

juhasza@maths.ox.ac.uk, izemke@math.princeton.edu

Proposed: András I Stipsicz

Seconded: Ciprian Manolescu, Peter Ozsváth

Received: 7 April 2018

Revised: 30 March 2019

Conical metrics on Riemann surfaces

I: The compactified configuration space and regularity

RAFE MAZZEO

XUWEN ZHU

We introduce a compactification of the space of simple positive divisors on a Riemann surface, as well as a compactification of the universal family of punctured surfaces above this space. These are real manifolds with corners. We then study the space of constant curvature metrics on this Riemann surface with prescribed conical singularities at these divisors. Our interest here is in the local deformation for these metrics, and in particular the behavior of these families as conic points coalesce. We prove a sharp regularity theorem for this phenomenon in the regime where these metrics are known to exist. This setting will be used in a subsequent paper to study the space of spherical conic metrics with large cone angles, where the existence theory is still incomplete. Of independent interest is how setting up the analysis on these compactified configuration spaces provides a good framework for analyzing “confluent families” of regular singular, ie conic, elliptic differential operators.

53C25; 58H15

1. Introduction	309
2. Resolution of point configurations	313
3. Geometry of merging cone points	334
4. Preliminary analysis	339
5. Flat conical metrics	344
6. Hyperbolic conic metrics	348
7. Spherical metrics: cone angles less than 2π	366
Appendix. b -fibrations and polyhomogeneity	368
References	371

1 Introduction

This paper is a sequel to [16], by the first author and Weiss, concerning the space of metrics with constant curvature and prescribed conic singularities on a compact

Riemann surface M . In that paper a careful analysis was made of the deformation theory for such metrics provided all the cone angles are less than 2π . This assumption simplifies both the analytic and the geometric considerations considerably. Consider the space of tuples $(c, p, \vec{\beta}, K, A)$, where c is a conformal structure on M , p a collection of k distinct points, $\vec{\beta}$ a k -tuple of parameters prescribing the cone angles at the points p_j with each $\beta_j \in (0, 1)$ (this corresponds to all cone angles lying in $(0, 2\pi)$), and constant K specifying the Gauss curvature and $A > 0$ specifying the area, all subject to the requirement forced by Gauss–Bonnet that

$$(1) \quad \chi(M, \vec{\beta}) := \chi(M) + \sum_{j=1}^k (\beta_j - 1) = \frac{1}{2\pi} KA.$$

It is known through the work of several authors that for each such tuple there exists a unique metric on M which has constant curvature K , area A and conical singularities at the points p_j with cone angle $2\pi\beta_j$. There is a caveat when $K > 0$ and $k > 2$, which states that in this case an extra condition is needed on the cone angles, namely that they satisfy the so-called Troyanov condition

$$(2) \quad \min\{2, 2\beta_j\} + k - \chi(M) > \sum_{i=1}^k \beta_i, \quad j = 1, \dots, k,$$

which is trivial when restricted to $\vec{\beta} \in (0, 1)^k$ except when $M = S^2$. The main result of [16] states that the Teichmüller space $\mathcal{T}_{y,k}^{\text{conic}}$ of all such solutions modulo the space of diffeomorphisms of M isotopic to the identity is a smooth manifold.

It is known that the situation becomes much more complicated when some or all of the β_j are greater than 1, at least in the case that $K > 0$. One classical inspiration to study this case is when each $\beta_j \in \mathbb{N}$, ie all cone angles are integer multiples of 2π , in which case examples are easily obtained as ramified covers over other compact surfaces with metrics of constant curvature. Existence and uniqueness of spaces with arbitrary cone angles and curvature $K \leq 0$ subject to (1) is relatively easy; see McOwen [17]. Much more recent is the dramatic breakthrough by Mondello and Panov [19], which establishes, through beautiful and purely geometric reasoning, necessary and sufficient conditions on the possible set of values $\vec{\beta}$ for which there exists a metric with constant curvature 1 (a spherical metric) on S^2 with these prescribed cone angle parameters.

This last-cited paper leaves open some fundamental questions. The one which interests us here is to describe the space of points p and cone angle parameters $\vec{\beta}$ for which

there exist spherical metrics with this data prescribing the conic singularities. The answer is complicated and (at least to our understanding) not completely explicit. An initial hope might be to show that the space of all solutions (mod diffeomorphisms) is a smooth manifold. From this one might then further try to apply various techniques from geometric analysis to count solutions. Unfortunately, for spherical cone metrics with cone angles greater than 2π , this space fails to be smooth on certain subvarieties. One of our goals, which will be addressed in a sequel to this paper, is to understand this failure more precisely. Briefly, however, the key observation is that if g is a spherical metric for which the deformation theory is obstructed, it is possible to consider this solution in a larger moduli space where the deformation theory is unobstructed. This broader setting consists of letting certain of the cone points p_j split into clusters of cone points with smaller angles. This is an analytic manifestation of one of the important steps in the geometric arguments of Mondello and Panov [19].

The analysis needed to carry this out turns out to be somewhat complicated and requires the development of some machinery which will occupy a significant part of this paper. We regard this machinery of independent interest, and expect that it may be a useful tool in studying various other analytic problems involving geometric objects which are singular or otherwise distinguished at families of points which can cluster. We mention in particular the study of solutions of the two-dimensional vortex equation on a Riemann surface, as well as the study of analytic constructions related to holomorphic quadratic differentials in relationship to the Hitchin moduli space.

1.1 Outline of results

This paper has two main parts. In the first part, [Section 2](#), we develop these general ideas, which involve the construction of a resolution via real blowup of the configuration space of k points on M and of the universal family of marked surfaces over this blown-up configuration space. Similar constructions are classical in algebraic geometry if one uses complex blowups, but our use of real blowups and other C^∞ methods here lead to spaces which are compact manifolds with corners which encode the different modes of clustering of these k points. This construction is closely related to other recent work, notably the ongoing work of Kottke and Singer [12] on the compactification of the moduli space of monopoles in $\mathbb{R}^3$. We describe the construction of the extended configuration space $\mathcal{E}_k$ (the base manifold) in [Section 2.1](#), and the resolution of the universal family $\mathcal{C}_k$ (the total space) in [Section 2.2](#). We then give the two simplest examples when $k = 2$ or 3 in [Sections 2.3](#) and [2.4](#). In [Section 2.5](#) we give a description

of the combinatorial structure of the boundary faces for the generic k -point case, and in particular show that we obtain a b -fibration.

In the second part we consider the space of metrics with constant curvature and prescribed conic singularities; in this paper we restrict attention to flat and hyperbolic metrics with no angle constraints, and spherical metrics with cone angles less than 2π . Our main theorem here is a new regularity result, which we give a sketch below, and refer to Theorems 16, 29 and 32 for the precise statements.

Theorem *The family of hyperbolic or flat metrics with conic singularities with arbitrary cone angle, or spherical conic surfaces with cone angles less than 2π and satisfying the Troyanov constraint, lifts to be polyhomogeneous, a natural generalization of smoothness, on the compactified universal family of curves $\mathcal{C}_k$ over this extended configuration space $\mathcal{E}_k$.*

In Section 3 we set up the geometric process of merging cone points, described both locally and globally. In Section 4 we recall some facts on analysis of conic elliptic operators. Sections 5–7 give the proof of the main theorem in the flat, hyperbolic and spherical cases. In each of the three cases we study solutions to a family of singular elliptic PDEs on the new space $\mathcal{C}_k$ constructed above. For the flat case in Section 5, the proof is done by a direct computation and we show that the solutions given by Green's functions are polyhomogeneous. The proof for the nonzero curvature cases are more involved. In Section 6.1 we prove the result when two cone points merge, which involves first constructing approximate solutions to arbitrarily high order, followed by using the maximum principle to get the exact solution, and finally using commutator argument to show conormality and polyhomogeneity. In Section 6.2 the case where more cone points merge is proved by a similar argument but with a more involved process in constructing the approximate solutions. In Section 7 the spherical case is proved in a similar way, except that the maximum principle no longer holds and is replaced by invertibility of the linearized operator.

These results are first steps in our program to understand the entire moduli space of constant curvature conical metrics on surfaces. The explanation of the extended configuration family, which is the setting for this regularity theory, is already of interest, and its definition is vindicated by our main regularity theorem. In a second paper we will employ this machinery to understand features of the moduli space of spherical cone metrics where the cone angles are greater than 2π . Our eventual goal is to understand

the stratified nature of these moduli spaces in sufficient detail that we can produce a count of solutions. We also hope to reach a better correspondence between the classical results and tools used to study these problems and the ones developed here.

Acknowledgements The authors are happy to acknowledge useful conversations with Misha Kapovich, Richard Melrose and Michael Singer. Mazzeo was supported by the NSF grant DMS-1608223. Zhu was supported by the NSF grant DMS-1905398.

2 Resolution of point configurations

The first part of this paper focuses on a rather intricate geometric construction, which is a resolution via real blowup of the configuration space of k points on a compact Riemann surface M , as well as the resolution of the universal family over this space.

To be more specific, let $\mathcal{D}_k(M)$ denote the space of nonnegative divisors on M of total degree k . Thus a point of $\mathcal{D}_k(M)$ consists of an ordered k -tuple of not necessarily distinct points $p_1, \dots, p_k \in M$. Although it is more common to study this using algebrogeometric ideas, we take a decidedly real and $\mathcal{C}^\infty$ approach. Away from coincidences where two or more of the p_j are the same, $\mathcal{D}_k(M)$ is a copy of M^k with all the partial diagonals removed. Of course, Σ_k acts freely on this open set. Our first goal is to define a *real* compactification $\mathcal{E}_k(M)$ of this open dense set in $\mathcal{D}_k(M)$, which we call the extended configuration space. This compactification is a manifold with corners, which comes equipped with a blowdown map $\beta: \mathcal{E}_k \rightarrow M^k$. We next consider the product $\mathcal{E}_k(M) \times M$; this is a trivial bundle which has a tautological multivalued section σ : if $q \in \mathcal{E}_k(M)$ and $\mathfrak{p} = \beta(q)$, then $\sigma(\mathfrak{p})$ is the divisor $\mathfrak{p}$ considered as a subset of M . We shall define a resolution of this object, again as a manifold with corners, using a suitable blowup of the graph of σ ; this is called the extended configuration family and denoted by $\mathcal{C}_k(M)$. This is not quite a fibration over $\mathcal{E}_k(M)$ since certain fibers are “broken”; instead it is a slightly more general type of map called a b -fibration, a natural extension of the notion of fibrations to the category of manifolds with corners. See the [appendix](#) for a general discussion about b -fibrations and manifolds with corners.

2.1 The extended configuration space $\mathcal{E}_k$

Our first goal is to define a good compactification for $\mathcal{D}_k^s$, the space of all “simple” divisors, defined in (3) below. We begin with some notation. Suppose first that

$\mathcal{I} \subset \{1, \dots, k\}$ is an index set with $|\mathcal{I}| \geq 2$. The $\mathcal{I}^{\text{th}}$ partial diagonal is the subset

$$\Delta_{\mathcal{I}} = \{p \in \mathcal{D}_k : p_i = p_j \text{ for all } i, j \in \mathcal{I}\}.$$

There is a reverse partial order of diagonals corresponding to the inclusion of index sets,

$$\mathcal{I} \subset \mathcal{J} \iff \Delta_{\mathcal{I}} \supset \Delta_{\mathcal{J}}.$$

The union of the two index sets is defined in the usual sense. If $\mathcal{I}$ and $\mathcal{J}$ have at least one common element, one can identify the partial diagonal corresponding to their union as the intersection of their diagonals:

$$\Delta_{\mathcal{I} \cup \mathcal{J}} = \Delta_{\mathcal{I}} \cap \Delta_{\mathcal{J}} \quad \text{if } \mathcal{I} \cap \mathcal{J} \neq \emptyset.$$

The assumption of nonempty intersection guarantees that the intersection of two diagonals is still a diagonal. Otherwise, if $\mathcal{I} \cap \mathcal{J} = \emptyset$, there is a strict inclusion

$$\Delta_{\mathcal{I} \cup \mathcal{J}} \subsetneq \Delta_{\mathcal{I}} \cap \Delta_{\mathcal{J}}.$$

On the other hand, when $|\mathcal{I} \cap \mathcal{J}| \geq 2$, then $\Delta_{\mathcal{I} \cap \mathcal{J}}$ is the smallest diagonal containing both $\Delta_{\mathcal{I}}$ and $\Delta_{\mathcal{J}}$.

We also let

$$\Delta_{\mathcal{I}}^0 = \Delta_{\mathcal{I}} \setminus \left(\bigcup_{\mathcal{J} \supsetneq \mathcal{I}} \Delta_{\mathcal{J}} \right).$$

It is then clear that the ensemble $\{\Delta_{\mathcal{I}}^0\}$ is a stratification of $M^k = \mathcal{D}_k$; the dense open stratum equals

(3)
$$\mathcal{D}_k^s = M^k \setminus \left(\bigcup_{\mathcal{I}} \Delta_{\mathcal{I}}^0 \right).$$

To resolve the point collisions, we resolve all the partial diagonals; this is done by blowing up the diagonals iteratively in order of decreasing index set. In other words, if $\mathcal{I} \subsetneq \mathcal{J}$, then $\Delta_{\mathcal{J}}$ is blown up before $\Delta_{\mathcal{I}}$. There is still some freedom in the order of blowup, since the inclusion of index sets only provides a partial order. Below we show that the final space is well defined and does not depend on which specific order we blow up.

Let X be a manifold with corners, containing two p -submanifolds, Y_1 and Y_2 . (A p -submanifold $Y \subset X$ is defined to be a submanifold for which some neighborhood $\mathcal{U} \supset Y$ is diffeomorphic as a manifold with corners to the normal bundle NY .) The iterated blowup $[X; Y_1; Y_2]$ is the manifold with corners obtained as follows. First blow

up Y_1 in X to obtain a space $[X; Y_1]$. Now lift $Y_2 \setminus (Y_1 \cap Y_2)$ to this space and take its closure. Finally, take the blowup of this lift in $[X; Y_1]$. In general the resulting space depends on the order in which these blowups are taken; the reverse order may result in a nondiffeomorphic space. There are two special situations where the order does not matter: the first is if $Y_1 \subset Y_2$, and the second is if Y_1 and Y_2 meet transversely so that their normal bundles are disjoint (away from the zero section). Similarly, when there are more p -submanifolds $\{Y_i\}_{i=1}^k$, $[X; Y_1; \dots; Y_k]$ is well defined if the following is true: for any Y_i and Y_j , either $Y_i \subset Y_j$, or Y_i and Y_j are transversal.

In the prescription for blowing up the partially ordered sequence of partial diagonals, we are blowing up these partial diagonals by inclusion, ie we always blow up the “smaller” submanifolds first. However, we must check that the second criterion about transversality is satisfied.

Lemma 1 *Let $\mathcal{I}$ and $\mathcal{J}$ be any two index sets. Suppose that $X_{\mathcal{I}, \mathcal{J}}$ is the manifold with corners obtained by blowing up all the partial diagonals $\Delta_{\mathcal{K}}$ in M^k for which $\mathcal{K} \supset \mathcal{I}$ and $\mathcal{K} \supset \mathcal{J}$. Then the lifts of $\Delta_{\mathcal{I}}$ and $\Delta_{\mathcal{J}}$ are transverse in $X_{\mathcal{I}, \mathcal{J}}$.*

Proof When $\mathcal{I} \cap \mathcal{J} = \emptyset$, we can choose the complex coordinates $s_1, \dots, s_k$ so that $\Delta_{\mathcal{I}} = \{s_1 = \dots = s_p = 0\}$ and $\Delta_{\mathcal{J}} = \{s_{p+1} = \dots = s_q = 0\}$. Clearly then $\Delta_{\mathcal{I}}$ and $\Delta_{\mathcal{J}}$ intersect transversely.

On the other hand, if $\mathcal{I} \cap \mathcal{J} \neq \emptyset$, then $\mathcal{I}' = \mathcal{I} \setminus (\mathcal{I} \cap \mathcal{J})$ and $\mathcal{J}' = \mathcal{J} \setminus (\mathcal{I} \cap \mathcal{J})$ are disjoint. Choosing coordinates so that $\{s_1 = \dots = s_p = s_{\ell+1} = \dots = s_k = 0\}$ on $\Delta_{\mathcal{I}}$ and $\{s_{p+1} = \dots = s_q = s_{\ell+1} = \dots = s_k = 0\}$ on $\Delta_{\mathcal{J}}$, it is not hard to check that the lifts of these submanifolds to the blowup around the set $\{s_1 = \dots = s_q = s_{\ell+1} = \dots = s_k = 0\}$ are disjoint, hence transverse by default. $\square$

Using this lemma, we may now proceed through this sequence of blowups to obtain the extended (ordered) configuration space

$$(4) \quad \mathcal{E}_k = \left[\mathcal{D}_k; \bigcup_{\mathcal{I}} \Delta_{\mathcal{I}} \right].$$

One consequence of this operation is that the action of symmetric group Σ_k on $\mathcal{D}_k$ is resolved.

Proposition 2 *The symmetric group Σ_k acts freely on $\mathcal{E}_k$.*

Proof The fixed points of Σ_k on $\mathcal{D}_k$ are precisely the partial diagonals and, moreover, the isotropy group at $\Delta_{\mathcal{I}}$ is a subgroup of the isotropy group at $\Delta_{\mathcal{J}}$ when $\Delta_{\mathcal{I}} \supset \Delta_{\mathcal{J}}$,

or equivalently when $\mathcal{I} \subset \mathcal{J}$. Thus our iterative blowup corresponds to the blowup which resolves this group action, defined by blowing up the fixed point sets ordered by reverse isotropy type inclusion, and it is not hard to check in this case that the isotropy groups of the lifted group action are all trivial. This is a special case of a more general iterated blowup considered by Albin and Melrose [1], which resolves a general Lie group action. □

The space $\mathcal{E}_k$ appears rather complicated at first glance, but the combinatorial structure of its faces mirrors the partially ordered set of subsets $\{\mathcal{I}\}$ of $\{1, \dots, k\}$. For each element $\mathcal{I}$ of this set, there is a boundary hypersurface $F_{\mathcal{I}}$ of $\mathcal{E}_k$ generated by blowing up $\Delta_{\mathcal{I}}$. We also denote by $\rho_{\mathcal{I}}$ the boundary-defining function for this face. Identifying the interior of $\mathcal{E}_k$ with the nonsingular part of M^k away from all the diagonals, we see that $\rho_{\mathcal{I}}$ provides a measurement of the radius of a cluster of $|\mathcal{I}|$ coalescing points.

2.2 The extended configuration family $\mathcal{C}_k$

We next consider the universal family over $\mathcal{E}_k$. This is a space $\mathcal{C}_k$ equipped with a b -fibration

$$\widehat{\beta}: \mathcal{C}_k \rightarrow \mathcal{E}_k$$

such that for each $\mathfrak{p} \in \mathcal{E}_k^{\text{reg}} (\simeq \mathcal{D}_k^s)$, the fiber $\widehat{\beta}^{-1}(\mathfrak{p})$ is the surface M blown up at the points of $\mathfrak{p}$. We point out that $\widehat{\beta}$ is different than the map β discussed earlier, which is the blowdown $\mathcal{E}_k \rightarrow M^k$. In the following, it is often simpler to refer to points $\mathfrak{p}$ on $\mathcal{E}_k$; these are, however, elements of the compactified configuration space, so the actual divisor, or k -tuple of points on M , is really the image under β of this point. In any case, with this understanding, if $\mathfrak{p}$ lies in one of the boundary faces of $\mathcal{E}_k$, then the fiber $\widehat{\beta}^{-1}(\mathfrak{p})$ is a union of surfaces with boundary which encode the various ways the corresponding cluster of points can come together.

To define this universal family, we begin with the trivial fibration $\mathcal{E}_k \times M \rightarrow \mathcal{E}_k$, and let z be a generic point on the fiber, which we may as well assume is a local holomorphic coordinate there. We wish to resolve the graph of the canonical “section” σ of this bundle,

$$\{(\mathfrak{p}, z) \in \mathcal{E}_k \times M : z \in \sigma(\mathfrak{p})\}.$$

Since σ is multivalued, we must first blow up the crossing loci, which are contained in the graphs of σ over the faces of $\mathcal{E}_k$. More specifically, if $\mathfrak{p}$ lies in a face $F_{\mathcal{I}}$, we write $\sigma^{\mathcal{I}}(\mathfrak{p})$ for the corresponding “coincidence point” $p_{i_1} = \dots = p_{i_r}$ if $\mathcal{I} = (i_1, \dots, i_r)$ (so

$\sigma(\mathfrak{p})$ has r copies of this point and $k - r$ other points), and then define the coincidence set

(5)

$$F_{\mathcal{I}}^{\sigma} = \{\rho_{\mathcal{I}} = 0, z = \sigma^{\mathcal{I}}(\mathfrak{p})\}.$$

The space $\mathcal{C}_k$ may now be defined by iteratively blowing up this collection of submanifolds with respect to the partial order on index sets, culminating at the last step in the blowups of the nonsingular parts of the graph of σ , ie the submanifolds $F_i^{\sigma} = \{z = p_i\}$ for $i = 1, \dots, k$, where $\mathfrak{p}$ does not lie in any partial diagonal. Altogether,

(6)

$$\mathcal{C}_k = [\mathcal{E}_k \times M; \{F_{\mathcal{I}}^{\sigma}\}].$$

The following lemma shows, just as for $\mathcal{E}_k$, that the blowup is well defined.

Lemma 3 *The lifts of $F_{\mathcal{I}}^{\sigma}$ and $F_{\mathcal{J}}^{\sigma}$ are transverse after $F_{\mathcal{I} \cup \mathcal{J}}^{\sigma}$ has been blown up. In particular, the lifts of F_i^{σ} and F_j^{σ} do not meet when $i \neq j$.*

Proof As before, this follows from the fact that

$$F_{\mathcal{I}}^{\sigma} \cap F_{\mathcal{J}}^{\sigma} = F_{\mathcal{I} \cup \mathcal{J}}^{\sigma}$$

when $\mathcal{I} \cap \mathcal{J} \neq \emptyset$, while $F_{\mathcal{I}}^{\sigma}$ and $F_{\mathcal{J}}^{\sigma}$ are transverse away from $F_{\mathcal{I} \cup \mathcal{J}}^{\sigma}$ when $\mathcal{I} \cap \mathcal{J} = \emptyset$. The last assertion is obvious. □

2.3 The simplest case, $k = 2$

The description of the boundary faces of $\mathcal{E}_k$ and $\mathcal{C}_k$ is somewhat complicated and at first glance confusing, so to warm up, we present the cases $k = 2$ and 3 in some detail since it is possible to see what is going on without too much work then.

The space of ordered divisors $\mathcal{D}_2$ is simply M^2 , and there is a single diagonal $\Delta_{12} = \{p_1 = p_2\}$, hence

$$\mathcal{E}_2 = [\mathcal{D}_2; \Delta_{12}].$$

Here and below we keep the subscript 12 to foreshadow the general case. From local coordinates (z_1, z_2) near $(p_0, p_0) \in \Delta$, we determine the center of mass $\zeta = \frac{1}{2}(z_1 + z_2)$ and displacement $w = \frac{1}{2}(z_1 - z_2)$, so that

(7)

$$z_1 = \zeta + w, \quad z_2 = \zeta - w.$$

The blowup amounts to setting $w = \rho_{12}e^{i\theta}$ with $\theta \in [0, 2\pi]$ and adding the face $\rho_{12} = 0$.

The front face F_{12} is then a possibly nontrivial circle bundle over the diagonal. Indeed, it is the unit normal bundle of the diagonal in M^2 , and hence has Euler characteristic equal to $\chi(M)$. In any case, we have coordinates (θ, ζ) on F_{12} and a full set of coordinates $(\rho_{12}, \theta, \zeta)$ near this face in the blowup.

The symmetric group Σ_2 interchanges the two coordinates (z_1, z_2) , and hence sends $\zeta \mapsto \zeta$ and $w \mapsto -w$. In local coordinates, $(\rho_{12}, \theta, \zeta) \mapsto (\rho_{12}, \pi + \theta, \zeta)$, and it is easy to see that this is a free action.

The extended configuration family is now obtained from the product $\mathcal{E}_2 \times M$ by blowing up in succession the two submanifolds

$$F_{12}^\sigma = \{(\rho_{12} = 0, \theta, \zeta, z) : z = \zeta\} \subset F_{12} \times M \quad \text{and} \quad F_1^\sigma \cup F_2^\sigma = \{(\mathbf{p}, \sigma(\mathbf{p})) : \mathbf{p} \in \mathcal{D}_2\}.$$

For the first of these blowups, introduce spherical coordinates (R_{12}, Ω) around the codimension three submanifold $\{\rho_{12} = 0, \zeta = z\}$, so $R_{12} \geq 0$ and $\Omega \in \mathbb{S}_+^2$. We write

$$(8) \quad \Omega = (\rho_{12}, z - \zeta)/R_{12} = (\sin \omega, \cos \omega e^{i\phi}),$$

where $\omega \in [0, \frac{\pi}{2}]$ and $\phi \in [0, 2\pi]$, and so

$$(9) \quad \rho_{12} = R_{12} \sin \omega, \quad z = \zeta + R_{12} \cos \omega e^{i\phi}.$$

We also set $z - \zeta = r e^{i\phi}$.

The face created by this blowup, which we call $\mathfrak{C}_{12}$, is the total space of a fibration $\pi_{12}: \mathfrak{C}_{12} \rightarrow F_{12}$, with each fiber a copy of $\mathbb{S}_+^2$. The preimage of a point $(0, \theta, \zeta) \in F_{12}$ is the union of two manifolds with boundary: the first is the blowup $[M, \{\zeta\}]$ of M around the point ζ , and the second is $\mathbb{S}_+^2$. These meet along their common boundary, which is a circle. From (9),

$$(10) \quad \pi_{12}^{-1}(\rho_{12}) = A R_{12} \omega,$$

where A is a strictly positive smooth function. The significance of this computation is that the lift of the defining function for F_{12} equals the product of defining functions for the fiber M blown up at ζ and the half-sphere, up to a nonvanishing smooth factor. That is, the boundary-defining functions satisfy the b -fibration condition.

We now turn to the second blowup. Consider the graph of σ ,

$$(11) \quad \begin{aligned} F_1^\sigma &= \{z = z_1\} = \{z = \zeta + \rho_{12} e^{i\theta}\} = \{\omega = \frac{\pi}{4}, \phi = \theta\}, \\ F_2^\sigma &= \{z = z_2\} = \{z = \zeta - \rho_{12} e^{i\theta}\} = \{\omega = \frac{\pi}{4}, \phi = \theta + \pi\}. \end{aligned}$$

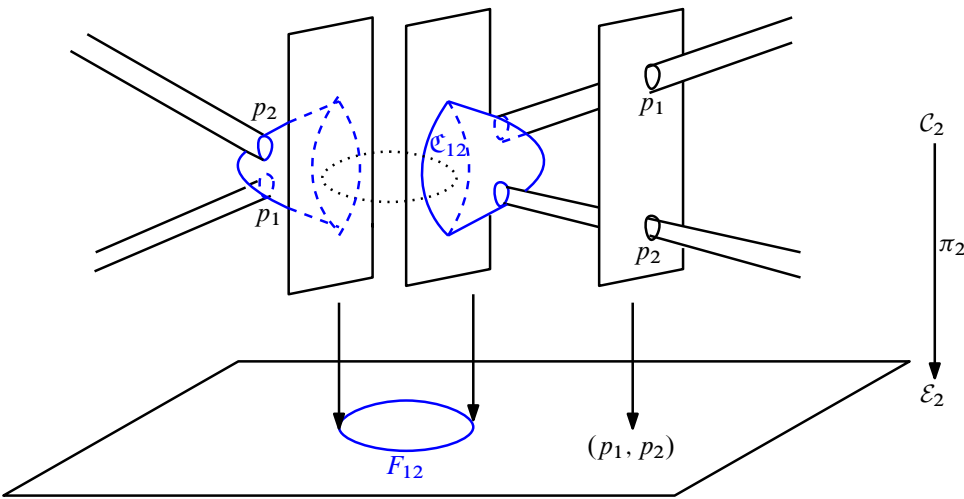


Figure 1: The singular fibration of $\mathcal{C}_2 \rightarrow \mathcal{E}_2$. Here we removed the center of mass ζ , and the coordinate in the base is w ; see (7). The boundary face in the base, F_{12} , is parametrized by θ such that $w = \rho_{12}e^{i\theta}$. When θ goes from 0 to π , the two points p_1 and p_2 on the fiber interchange.

These two components intersect $\mathfrak{C}_{12}$ at two disjoint copies of $\mathbb{S}^1$. Their intersection with each $\pi_{12}^{-1}(0, \theta, \zeta)$ consists of two points on each $\mathbb{S}^2_+$ fiber. In other words, the boundaries of these components are each circles, and there is one point of each of these circles in each $\mathbb{S}^2_+$ fiber. These intersection points are given explicitly in (11). Observe that as θ goes from 0 to π , these two points interchange. The final configuration space is equal to

$$\mathcal{C}_2 := [\mathcal{E}_2 \times M; F_{12}^\sigma; F_1^\sigma \cup F_2^\sigma].$$

Note finally that since each F_i^σ projects surjectively to $\mathcal{E}_2$, the relation describing the pullback of boundary-defining functions stays the same as for the space before this last blowup.

This space is equipped with a b -fibration

$$\pi: \mathcal{C}_2 \rightarrow \mathcal{E}_2.$$

Over a regular point $\mathfrak{p} \notin \Delta_{12}$, the preimage $\pi^{-1}(\mathfrak{p})$ is a copy of M blown up at the two points of $\mathfrak{p}$. On the other hand, the preimage $\pi^{-1}(0, \theta, \zeta)$ of a point on F_{12} is the union of M blown up at the single point ζ and the half-sphere $\mathbb{S}^2_+$ blown up at two points, $[\mathbb{S}^2_+; \{\omega = \frac{\pi}{4}, \phi = \theta\} \cup \{\omega = \frac{\pi}{4}, \phi = \theta + \pi\}]$. See Figure 1 for an illustration of the fibration.

2.4 The case $k = 3$

The next case illustrates the iterative nature of the general construction.

As before, we work in local coordinates $(z_1, z_2, z_3) \in M^3$ near a point (p_0, p_0, p_0) in Δ_{123} .

The initial blowup of the central diagonal Δ_{123} in $\mathcal{D}_3$ results in the space $[\mathcal{D}_3; \Delta_{123}]$. Just as for $\mathcal{E}_2$, this is a manifold with boundary; its boundary, or front face at this first step, is called F_{123} and is a sphere bundle over Δ_{123} , but now with three-dimensional spherical fibers. In choosing coordinates, it seems to be more convenient to break the symmetry by using classical Jacobi coordinates (introduced originally to study the N -body problem in celestial mechanics). Thus we define a center of mass ζ in the first two variables as well as two displacement variables,

$$\zeta = \frac{1}{2}(z_1 + z_2), \quad w_1 = \frac{1}{2}(z_1 - z_2), \quad w_2 = z_3 - \frac{1}{2}(z_1 + z_2).$$

In these coordinates,

$$\Delta_{123} = \{\zeta + w_1 = \zeta - w_1 = \zeta + w_2\} = \{w_1 = w_2 = 0\}.$$

The resolution is the blowup of the origin in $\mathbb{C}_{w_1 w_2}$, which is captured by spherical coordinates in the fibers of the normal bundle:

$$N_\zeta \Delta_{123} \cong \mathbb{R}^4_{w_1, w_2} \ni \rho_{123} \Theta, \quad (\rho_{123}, \Theta) \in \mathbb{R}^+ \times \mathbb{S}^3.$$

To proceed, write

$$\Theta = (e^{i\phi_1} \cos \theta, e^{i\phi_2} \sin \theta), \quad (\theta, \phi_1, \phi_2) \in [0, \frac{\pi}{2}] \times [0, 2\pi]^2,$$

corresponding to the fact that $\mathbb{S}^3$ is a join of two circles, so that

(12)
$$w_1 = \rho_{123} \cos \theta e^{i\phi_1}, \quad w_2 = \rho_{123} \sin \theta e^{i\phi_2}.$$

These coordinates are singular at $\theta = 0, \frac{\pi}{2}$. To remedy this near $\theta = \frac{\pi}{2}$, for example, we use instead the projective coordinate $\tilde{w}_1 = w_1/\rho_{123}$ along with ϕ_2 .

In these new coordinates, the lifts of the partial diagonals in a neighborhood of Δ_{123} have the form

(13)
$$\begin{aligned} \Delta_{12} &= \{z_1 = z_2\} = \{w_1 = 0\} = \{\theta = \frac{\pi}{2}\} = \{\tilde{w}_1 = 0\}, \\ \Delta_{13} &= \{z_1 = z_3\} = \{w_1 = w_2\} = \{\theta = \frac{\pi}{4}, \phi_1 = \phi_2\}, \\ \Delta_{23} &= \{z_2 = z_3\} = \{w_1 = -w_2\} = \{\theta = \frac{\pi}{4}, \phi_1 = -\phi_2\}. \end{aligned}$$

The additional expression for the lift of Δ_{12} is included because $\theta = \frac{\pi}{2}$ is a singular locus for the (θ, ϕ_1, ϕ_2) coordinate system. Each of these lifts is locally a product $\mathbb{S}^1 \times \mathbb{R}^+$ near $\mathbb{S}^3 \times \{0\}$, and their intersections with each fiber of F_{123} are three disjoint circles. We then blow these up in any order to obtain

$$\mathcal{E}_3 = [\mathcal{D}_3; \Delta_{123}; \Delta_{12} \cup \Delta_{13} \cup \Delta_{23}].$$

There are three new front faces, F_{ij} , each with a boundary-defining function ρ_{ij} , and each (locally) diffeomorphic to $\mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{R}^+$. The intersection $F_{ij} \cap F_{123}$ is a torus $\mathbb{S}^1 \times \mathbb{S}^1$.

Let us illustrate this geometry near $F_{12} \cap F_{123}$ in coordinates. From (12) and (13), $\phi_2 \in \mathbb{S}^1$ and $\rho_{123} \in \mathbb{R}^+$ parametrize Δ_{12} , while we set $\tilde{w}_1 := w_1/\rho_{123}$ as a coordinate for the normal bundle. Blowing up at $\tilde{w}_1 = 0$ amounts to writing $\tilde{w}_1 = \rho_{12}e^{i\theta_{12}}$. This is of course really the same as setting $\rho_{12} = \cos \theta$ and $\theta_{12} = \phi_1$, but the new monikers have been introduced to conform with the general notation when $k > 3$. Altogether, we have a full set of local coordinates

$$(14) \quad (\zeta, \rho_{12}, \theta_{12}, \phi_2, \rho_{123}) \in \mathbb{R}^2 \times \mathbb{R}^+ \times \mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{R}^+.$$

These are related to the original coordinates by

$$(15) \quad z_1 = \zeta + \rho_{123}\rho_{12}e^{i\theta_{12}}, \quad z_2 = \zeta - \rho_{123}\rho_{12}e^{i\theta_{12}}, \quad z_3 = \zeta + \rho_{123}\sqrt{1 - \rho_{12}^2}e^{i\phi_2}.$$

Now consider the configuration space $\mathcal{E}_3 \times M$. Near $(F_{12} \cap F_{123}) \times M$ we extend the local coordinates (14) to

$$(16) \quad (\zeta, \rho_{12}, \theta_{12}, \phi_2, \rho_{123}, z) \in \mathbb{R}^2 \times \mathbb{R}^+ \times \mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{R}^+ \times \mathbb{R}^2,$$

where z is a local coordinate on M . The coordinate ζ is sometimes omitted below. The first step is to resolve the “central” coincidence set — cf (5) —

$$F_{123}^\sigma = \{z - \zeta = \rho_{123} = 0\},$$

which we do by introducing spherical coordinates $(z - \zeta, \rho_{123}) = R_{123}\Omega_{123}$, where $R_{123} \geq 0$ and

$$(17) \quad \mathbb{S}_+^2 \ni \Omega_{123} = (\cos \omega e^{i\phi}, \sin \omega), \quad \omega \in [0, \frac{\pi}{2}], \phi \in \mathbb{S}^1.$$

This introduces the front face $\mathfrak{C}_{123} = \{R_{123} = 0\}$. Coordinates at this stage are

$$(\zeta, R_{123}, \Omega_{123}, \rho_{12}, \theta_{12}, \phi_2) \in \mathbb{R}^2 \times \mathbb{R}^+ \times \mathbb{S}_+^2 \times \mathbb{R}^+ \times \mathbb{S}^1 \times \mathbb{S}^1.$$

Note also that

$$(18) \quad \begin{aligned} z_1 - \zeta &= -(z_2 - \zeta) = R_{123} \sin \omega \rho_{12} e^{i\theta_{12}}, \\ z_3 - \zeta &= R_{123} \sin \omega \sqrt{1 - \rho_{12}^2} e^{i\phi_2}, \\ z - \zeta &= R_{123} \cos \omega e^{i\phi}. \end{aligned}$$

As before, the coordinates (ω, ϕ) become degenerate at $\omega = \frac{\pi}{2}$, so in a neighborhood of this locus we replace these by the projective coordinate $\hat{z} = (z - \zeta)/R_{123}$.

Now fix a point $q = (\rho_{12}, \theta_{12}, \phi_2)$ on the front face F_{123} in the base $\mathcal{E}_3$. The fiber above q after the blowup above is a union of $\mathbb{R}^2$ blown up at the origin (or, more globally, the surface M blown up at the point ζ) and a half-sphere, parametrized by the coordinates (ω, ϕ) (or $\hat{z}$) above. These meet along the circle $\{\omega = 0\}$. If π_3 is the blowdown map, then, analogous to (10),

$$(19) \quad \pi_3^* \rho_{123} = R_{123} \sin \omega = A R_{123} \omega$$

for some smooth nonvanishing function A .

The partial coincidence sets F_{ij}^σ intersect the fiber over q only in the interior of the hemisphere. In local coordinates,

$$(20) \quad \begin{aligned} F_{12}^\sigma &= \{z = z_1 = z_2\} = \{\rho_{12} = 0, \omega = \frac{\pi}{2}\} = \{\tilde{w}_1 = 0, \hat{z} = 0\}, \\ F_{13}^\sigma &= \{z = z_1 = z_3\} = \{\rho_{12} = \frac{\sqrt{2}}{2}, \tan \omega = \sqrt{2}, \theta_{12} = \phi_2 = \phi\}, \\ F_{23}^\sigma &= \{z = z_2 = z_3\} = \{\rho_{12} = \frac{1}{2}\sqrt{2}, \tan \omega = \sqrt{2}, \theta_{12} + \pi = \phi_2 = \phi\}. \end{aligned}$$

The final expression for the lift of F_{12}^σ is included because of the degeneracy of the other coordinate system at $\rho_{12} = 0$, $\omega = \frac{1}{2}\pi$.

Now blow up F_{12}^σ . This introduces a new front face $\mathfrak{C}_{12}$, the fibers of which (for each (ζ, q)) are hemispheres “on top” of the previous hemispheres. Carrying out the analogous blowup for the two other partial diagonals as well leads to the collection of front faces $\mathfrak{C}_{ij}$, each diffeomorphic to $\mathbb{R}^+ \times \mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}_+^2$. Each $\mathfrak{C}_{ij}$ fibers over the front face F_{ij} in the base $\mathcal{E}_3$ by projecting off the final $\mathbb{S}_+^2$. Again, this is very similar to the two-point case.

Let $\Omega_{12} \in \mathbb{S}_+^2$ be the variable in the fiber of $\mathfrak{C}_{12}$ and R_{12} the corresponding boundary-defining function. Thus

$$(\cos \omega e^{i\phi}, \rho_{12}) = R_{12} \Omega_{12}$$

and

$$(21) \quad \Omega_{12} = (\cos \omega_{12} e^{i\phi_{12}}, \sin \omega_{12}), \quad \omega_{12} \in [0, \frac{1}{2}\pi], \phi_{12} \in \mathbb{S}^1.$$

Coordinates near $\mathfrak{C}_{12}$ are then

$$(\zeta, R_{123}, R_{12}, \Omega_{12}, \theta_{12}, \phi_2) \in \mathbb{R}^2 \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{S}_+^2 \times \mathbb{S}^1 \times \mathbb{S}^1,$$

and these translate to the original coordinates by

$$\begin{aligned} z_1 - \zeta &= -(z_2 - \zeta) = R_{123} \sqrt{1 - (R_{12} \cos \omega_{12})^2} R_{12} \sin \omega_{12} e^{i\theta_{12}}, \\ (22) \quad z_3 - \zeta &= R_{123} \sqrt{1 - (R_{12} \cos \omega_{12})^2} \sqrt{1 - (R_{12} \sin \omega_{12})^2} e^{i\phi_2}, \\ z - \zeta &= R_{123} R_{12} \cos \omega_{12} e^{i\phi_{12}}. \end{aligned}$$

The boundary-defining function relation near the corner $\mathfrak{C}_{123} \cap \mathfrak{C}_{12}$ is given by

$$(23) \quad \pi_3^* \rho_{123} = R_{123} \sqrt{1 - (R_{12} \cos \omega_{12})^2} \sim R_{123}, \quad \pi_3^* \rho_{12} = R_{12} \sin \omega_{12} \sim R_{12} \omega_{12}.$$

We conclude by blowing up the lifts of the three submanifolds $F_i^\sigma = \{z = z_i\}$. In a generic fiber over F_{123} (ie away from all the partial diagonals), F_i^σ meets the front face $\mathfrak{C}_{123}$ on the hemisphere at three distinct points. Using the coordinates (18),

$$\begin{aligned} (24) \quad F_1^\sigma &= \{\cot \omega = \rho_{12}, \theta_{12} = \phi\}, \\ F_2^\sigma &= \{\cot \omega = \rho_{12}, \theta_{12} = \phi + \pi\}, \\ F_3^\sigma &= \{\cot \omega = \sqrt{1 - \rho_{12}^2}, \phi_2 = \phi\}. \end{aligned}$$

On approach to the resolved partial diagonals, two of the points z_i converge to one another on M . However, their lifts converge to distinct points on a hemispherical fiber of the innermost front face. In other words, F_i^σ and F_j^σ both intersect $\mathfrak{C}_{ij}$ while, if k is the third value distinct from i and j , then F_k^σ does not. In terms of the coordinates (22), we have

$$\begin{aligned} (25) \quad F_1^\sigma &= \{\omega_{12} = \frac{\pi}{4}, \theta_{12} = \phi_{12}\}, \\ F_2^\sigma &= \{\omega_{12} = \frac{\pi}{4}, \theta_{12} = \phi_{12} + \pi\}, \\ F_3^\sigma &= \{\cot \omega = \sqrt{1 - \rho_{12}^2}, \phi_2 = \phi_{12}\}. \end{aligned}$$

The roles of z_1 , z_2 and z_3 are interchangeable in this whole discussion, even though the Jacobi coordinates w_1 and w_2 break this symmetry. it is also not hard to check that the symmetric group Σ_3 acts freely on $\mathcal{C}_3$.

As this case makes clear, the geometry illustrated in Figure 2 requires compound singular coordinate transformations, which quickly become quite involved. The optimistic interpretation is that the compound asymptotics of solutions to natural elliptic operators

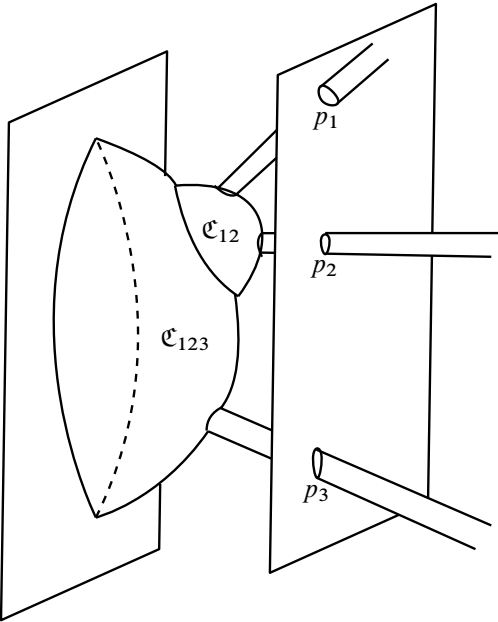


Figure 2: One of the singular fibers in $\mathcal{C}_3$, where two of the points collide faster than the third one.

which appear later in this paper are captured entirely by the intricate but still quite comprehensible geometry of this iterated blowup.

2.5 Boundary faces of $\mathcal{E}_k$ and $\mathcal{C}_k$

The previous discussion suggests that while it is possible to write out the iterated polar coordinate systems corresponding to iterated blowups, these become prohibitively complicated after a few steps. The special cases above should be used to gain intuition about the general case. In this section we undertake a more systematic study of the boundary faces and corners of $\mathcal{E}_k$ and $\mathcal{C}_k$. The goal is to explain some general features of the geometry of these boundary faces, which is done through a combination of invariant and local coordinate reasoning.

Faces of $\mathcal{E}_k$ As per our initial discussion in [Section 2.1](#), the boundary hypersurfaces $F_{\mathcal{I}}$ of $\mathcal{E}_k$ are in bijective correspondence with subsets $\mathcal{I} \subset \{1, \dots, k\}$ by way of the associated partial diagonals $\Delta_{\mathcal{I}}$. The set of all such subsets $\{\mathcal{I}\}$ is ordered by inclusion; thus $\{1, \dots, k\}$ is the unique maximal element and the minimal elements are the singletons $\{j\}$. The associated directed graph $\mathcal{G}_k$ has a vertex for each $\mathcal{I}$ and

we choose the direction on the edges to point from $\mathcal{I}$ to $\mathcal{J}$ when $\mathcal{I} \supset \mathcal{J}$. This “flips” the inclusion ordering, so we should think of $\{1, \dots, k\}$ as the bottom vertex (or root) of the tree, with the edges flowing upward.

We first consider the interiors of each face $F_{\mathcal{I}}$. Write $\Delta_{\mathcal{I}}^0$ for the open dense subset of elements in $\Delta_{\mathcal{I}}$ where *only* the points p_i for $i \in \mathcal{I}$ coincide. Fixing $\mathcal{I}$, let $M_{\mathcal{I}}^k$ be the subset of M^k where we remove all partial diagonals $\Delta_{\mathcal{J}}^0$ with $\mathcal{J} \cap \mathcal{I} \neq \emptyset$. The blowup

$$[M_{\mathcal{I}}^k; \Delta_{\mathcal{I}}^0]$$

can be regarded as a relatively open subset of $\mathcal{E}_k$. If $|\mathcal{I}| = \ell$, then the normal bundle of $\Delta_{\mathcal{I}}^0$ in $M_{\mathcal{I}}^k$ is naturally identified with $(TM)^{\oplus(\ell-1)}$. This means that the new front face coming from the blowup of $\Delta_{\mathcal{I}}^0$ is a bundle over a dense open subset in $M^{k-\ell+1} \cong \Delta_{\mathcal{I}}$ with fiber $\mathbb{S}^{2\ell-3}$.

Reorder the indices so that $\mathcal{I} = \{1, \dots, \ell\}$ and write $z' = (z_1, \dots, z_{\ell})$ and $z'' = (z_{\ell+1}, \dots, z_k)$; thus $z'' \in M^{k-\ell}$ and $\Delta_{\mathcal{I}}$ is the complete diagonal in the z' sub-system. Subdivide z' further, writing it as the sum of a center of mass $\zeta^{\mathcal{I}}(z')$ and $w' = (w_1, \dots, w_{\ell}) \in \mathbb{C}^{|\mathcal{I}|}$, where $\sum_{i \in \mathcal{I}} w_i = 0$, so w' lies in a space of real dimension $2|\mathcal{I}| - 2$. The blowup affects only the w' coordinates, and we see in this coordinate description that the new front face is locally a fibration over $\mathbb{R}^{2k-2\ell+2}$ with fiber $\mathbb{S}^{2\ell-3}$.

We now turn to the more difficult task of understanding the structure of each boundary hypersurface of $\mathcal{E}_k$ as a manifold with corners. By the iterative nature of these spaces, it is sufficient to focus on the innermost face $F_{1\dots k}$, which for convenience we denote by $F_{\max}$ below. We also write $\mathcal{I}_{\max} = \{1, \dots, k\}$.

Proposition 4 *The boundary faces and corners of $F_{\max}$ are in bijective correspondence with the connected trees $T \subset \mathcal{G}_k$ such that any minimal vertex $\mathcal{I} \in T$ has $|\mathcal{I}| \geq 2$.*

The correspondence is that if Z is any corner of this principal face, then the vertices of the associated tree $T = T_Z$ are the subsets $\mathcal{I}$ for which the face created by blowing up $\Delta_{\mathcal{I}}$ contains Z in its closure.

To get a feel for this, consider a few examples. First, a boundary hypersurface of $F_{\max}$ is the intersection of precisely two boundary hypersurfaces of $\mathcal{E}_k$: $F_{\max}$ and some other $F_{\mathcal{I}}$. The corresponding tree has two vertices, $\mathcal{I}_{\max}$ and $\mathcal{I}$, connected by an edge. Slightly more generally, if $\Delta_{\max} \subset \Delta_{\mathcal{I}_1} \subset \Delta_{\mathcal{I}_2} \subset \dots \subset \Delta_{\mathcal{I}_{\ell}}$, then $Z =$

$F_{\max} \cap F_{\mathcal{I}_1} \cap \cdots \cap F_{\mathcal{I}_\ell}$ is a (nonempty) corner, and T_Z is the tree $\mathcal{I}_{\max} \rightarrow \mathcal{I}_1 \rightarrow \cdots \rightarrow \mathcal{I}_\ell$. On the other hand, if $\mathcal{I}$ and $\mathcal{J}$ are disjoint, then the lifts of $\Delta_{\mathcal{I}}$ and $\Delta_{\mathcal{J}}$ to $[M^k; \Delta_{\max}]$ intersect transversely, so the blowups of these two partial diagonals may be taken in either order, and the intersection $F_{\max} \cap F_{\mathcal{I}} \cap F_{\mathcal{J}}$ is a corner of codimension two in $F_{\max}$. The corresponding tree has root $\mathcal{I}_{\max}$ connected to $\mathcal{I}$ and $\mathcal{J}$, but these two vertices are at the same level and not connected to one another. There is a similar description for any collection $\mathcal{I}_1, \dots, \mathcal{I}_s$ of disjoint subsets.

To prove Proposition 4, we recapitulate the construction of $\mathcal{E}_k$, emphasizing what happens at each stage of the iterative process. A basic principle here is that the lifts of the partial diagonals at each stage are “ p -submanifolds”. Recall that if X is a manifold with corners, then a submanifold Y is called a p -submanifold if, near every point $q \in Y$, there is a local coordinate system $(x, y) \in (\mathbb{R}^+)^{\ell} \times \mathbb{R}^m$ for X such that Y is given by setting some number of the x_i and y_j to 0. In other words, locally, Y is a product inside of X . Denote by $M^k(r)$ the space obtained after blowing up all partial diagonals $\Delta_{\mathcal{I}}$ with $|\mathcal{I}| \geq k - r$. Also, let $F(r)$ denote the central front face in $M^k(r)$, so $F(k) = F_{\max}$.

At the initial step, $M^k(0) = [M^k; \Delta_{\max}]$ and $F(0)$ is an $\mathbb{S}^{2k-3}$ bundle over $M \cong \Delta_{\max}$. The partial diagonals $\Delta_{\mathcal{I}}$ with $|\mathcal{I}| = k - 1$ lift to $M^k(0)$ to a disjoint collection of four-dimensional submanifolds with boundary; these intersect each $\mathbb{S}^{2k-3}$ fiber of $F(0)$ in k disjoint copies of $\mathbb{S}^1$. The space $M^k(1)$ is obtained by blowing up these lifted partial diagonals, and doing so yields a manifold with corners of codimension two. The new boundary hypersurface has $\mathbb{S}^{2k-5}$ fibers over a (disconnected) four-dimensional base. The codimension two corner of $M^k(1)$ is the boundary of $F(1)$; it has k components, each of which is a bundle with fiber $\mathbb{S}^1 \times \mathbb{S}^{2k-5}$ over $\Delta_{\max}$.

Continuing on, the partial diagonals $\Delta_{\mathcal{I}}$ with $|\mathcal{I}| = k - 2$ lift to p -submanifolds in $M^k(1)$. These lifts intersect each $\mathbb{S}^{2k-5}$ fiber in the new boundary face (over the $(k-1)$ -fold diagonals) in copies of $\mathbb{S}^1$, and intersect the fibers of $F(1)$ in a copy of $\mathbb{S}^3$. The important observation is that even in low dimension, ie $k = 4$, these 3-spheres in $\mathbb{S}^{2k-3}$ do not intersect in $M^k(1)$ by virtue of the fact that we have already blown up their intersection loci, the union of $\mathbb{S}^1$. When we blow up this set of lifted partial diagonals, the new boundary hypersurface of $F(2)$ is fibered by copies of $[\mathbb{S}^3; \bigsqcup \mathbb{S}^1] \times \mathbb{S}^{2k-7}$, and the corners of this boundary face, which are now corners of codimension 3 in $M^k(2)$, have fibers equal to $\mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^{2k-7}$.

As we proceed further in this construction, the thing to note is that at each stage, the lift of each $\Delta_{\mathcal{I}}$ with $|\mathcal{I}| = k - r - 1$ to $M^k(r)$ is a p -submanifold. Furthermore, if $\mathcal{I}$

and $\mathcal{J}$ are two subsets of size $k - r - 1$, then either $\mathcal{I} \cap \mathcal{J}$ share an element in common, or the intersection is empty. In the first case, it is straightforward to check that their lifts to $M(r)$ are in fact disjoint. In the second case, these two lifted diagonals intersect transversely, and hence it is not necessary to blow up their intersection since we can blow these up in either order to obtain the same result.

Now let us return to Proposition 4. As already described, given a corner Z of $F_{\max}$, we may associate to it a subgraph $T_Z \subset \mathcal{G}_k$. Clearly T is connected since all branches lead to the root $\mathcal{I}_{\max}$ (since $Z \subset F_{\max}$ in particular). Next, suppose that $\mathcal{I} \cap \mathcal{J} \neq \emptyset$, but neither one contains the other. Then the lifts of $\Delta_{\mathcal{I}}$ and $\Delta_{\mathcal{J}}$ would already be separated in $M(r)$, where $r = k - |\mathcal{I} \cup \mathcal{J}|$, which is where $\Delta_{\mathcal{I} \cup \mathcal{J}}$ is blown up, hence $F_{\mathcal{I}} \cap F_{\mathcal{J}} = \emptyset$. This shows that it is impossible for there to exist $\mathcal{I}_1, \mathcal{I}_2 \supset \mathcal{K}$ with $\mathcal{I}_1 \not\supset \mathcal{I}_2$ and $\mathcal{I}_2 \not\supset \mathcal{I}_1$. This proves that T_Z is a tree.

Conversely, if T is any connected tree in $\mathcal{G}_k$ emanating from $\mathcal{I}_{\max}$ (and which does not terminate at a node with $|\mathcal{I}| = 1$), then we must show that

$$Z = \bigcap_{\mathcal{I} \in T} F_{\mathcal{I}}$$

is nonempty. Since T is a tree, if $\mathcal{I}$ and $\mathcal{J}$ lie on the same branch, then one of these is a proper subset of the other, while if they lie on different branches, then they are disjoint. We can divide T into branches B_i , and it follows from the earlier discussion that the intersection of the $F_{\mathcal{I}}$ along a branch B_i is a nonempty corner Z_i ; on the other hand, if the branch B_2 is rooted at some vertex of another branch B_1 , then we can reduce back to the case of one node splitting in two as in the previous paragraph to see that the corresponding corners Z_1 and Z_2 intersect transversely in a nontrivial subset. We conclude using induction on the number of branches. This proves the result.

Faces of $\mathcal{C}_k$ We wish to carry out a similar analysis of the faces of $\mathcal{C}_k$. As before, we proceed inductively, so it suffices to analyze the structure of the central face $\mathfrak{C}_{\max}$.

Recall that the construction of $\mathcal{C}_k$ involves iteratively blowing up the coincidence sets $F_{\mathcal{I}}^{\sigma}$ in $\mathcal{E}_k \times M$ defined in (5). Let $\mathcal{C}_k(r)$ denote the space obtained by blowing up all such coincidence sets with $|\mathcal{I}| = j$ and $k \geq j \geq k - r$, in order of decreasing cardinality, and also write $\mathfrak{C}(r)$ for the central front face of $\mathcal{C}_k(r)$.

If $\rho_{\max}$ is a defining function for $F_{\max}$ in $\mathcal{E}_k$, then $\mathcal{C}_k(0)$ is the blowup of the set $\{\rho_{\max} = 0, z = \sigma^{\max}(\mathfrak{p})\}$. The front face $\mathfrak{C}(0)$ is fibered over $F_{\max}$ with fiber a half-sphere $\mathbb{S}^2_+$. The lifts of the coincidence sets $F_{\mathcal{I}}^{\sigma}$ intersect $\mathfrak{C}(0)$ as (usually mutually

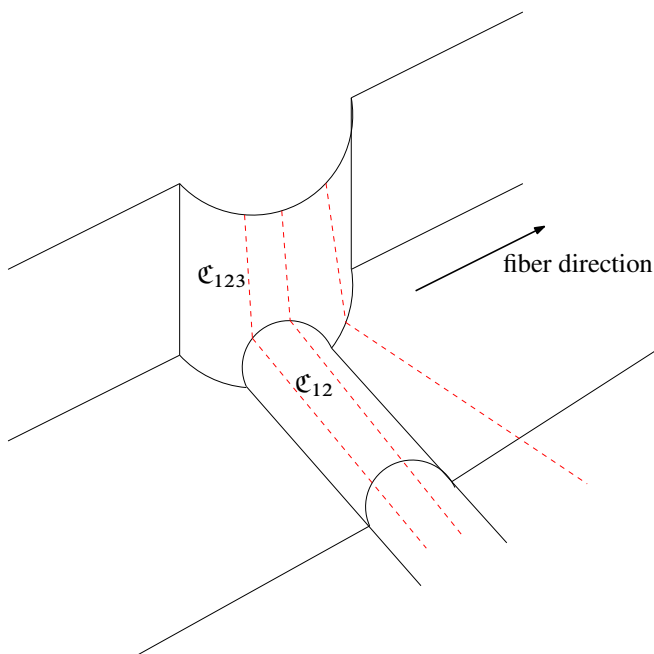


Figure 3: The intersection of front faces $\mathfrak{C}_{123}$ and $\mathfrak{C}_{12}$, where the dotted red lines are where the singleton coincidence sets F_i^σ for $i = 1, 2, 3$ meet the front faces.

intersecting) p -submanifolds. We denote the faces created by blowing up $F_{\mathcal{I}}^\sigma$ by $\mathfrak{C}_{\mathcal{I}}$; with a slight abuse of notation, we often do not distinguish between this face at some intermediate step of the construction and at the final stage in $\mathcal{C}_k$.

For expository purposes we jump immediately to the final step and consider the “generic” region, ie the preimage of the interior of $\mathcal{E}_k$. The only coincidence sets in this region are those with $|\mathcal{I}| = 1$, which means that the blowups in this region are unaffected by any other blowups. These (singleton) coincidence sets F_i^σ are of codimension two in $\mathcal{E}_k \times M$, so the faces $\mathfrak{C}_i$ are fibered by copies of $\mathbb{S}^1$. In particular, the intersection of $\mathfrak{C}_i$ with $\mathfrak{C}_{\max}$ intersects each $\mathbb{S}^2_+$ fiber in an $\mathbb{S}^1$. In other words, the portion of $\mathfrak{C}_{\max}$ over the interior of $F_{\max}$ (ie the principal front face of $\mathcal{C}_k$ away from all the faces lying over partial diagonals) is fibered by copies of $\mathbb{S}^2_+$ blown up at k distinct points, the locations of which are determined by the corresponding point in $F_{\max}$ (the “directions of approach” of the coalescing cluster of k points).

Now return to the construction in the proper order, and consider the passage from $\mathcal{C}_k(0)$ to $\mathcal{C}_k(1)$. This involves blowing up the coincidence sets $F_{\mathcal{I}}^\sigma$ with $|\mathcal{I}| = k - 1$. Over

the interior of the faces $F_{\mathcal{I}} \times M$ in $\mathcal{E}_k \times M$, the picture is analogous to the blowup at the principal front face in $\mathcal{E}_{k-1} \times M$: indeed, the corresponding point in M^k is, up to reordering, of the form $(p, \dots, p, p_k)$. As $\rho_{\max} \rightarrow 0$, the points p and p_k coalesce. The coincidence set $F_{\mathcal{I}}^\sigma$ is a p -submanifold which intersects the codimension two corner $(F_{\max} \cap F_{\mathcal{I}}) \times M$. When it is blown up, the new face is a bundle with $\mathbb{S}_+^2$ fibers over $F_{\mathcal{I}}$, uniformly to this intersection.

The new feature is that the fibers of $\mathcal{C}_k(1) \rightarrow \mathcal{E}_k$ over the corner $F_{\max} \cap F_{\mathcal{I}}$ are each a “tower” of hemispheres of height two, ie two copies of $\mathbb{S}_+^2$, the second one attached along its boundary to the circle created by blowing up a point in the first $\mathbb{S}_+^2$. The submanifold F_k^σ (corresponding to the singleton set $\{k\} = \{1, \dots, k\} \setminus \mathcal{I}$) intersects each of these fibers at a point of $\mathbb{S}_+^2$ away from this second hemisphere. On the other hand, the other F_i^σ (away from the boundaries of the corner $F_{\max} \cap F_{\mathcal{I}}$) intersect the $\mathbb{S}_+^2$ fibers of the face over $F_{\mathcal{I}}$ in $k-1$ distinct points, so after blowing these up, the fibers are hemispheres blown up at $k-1$ points. At the intersection with $F_{\max}$, the fibers are each a tower of two hemispheres, the inner one blown up at $k-1$ distinct points and the outer one at one additional point. All of this has been illustrated earlier in [Figure 3](#) for the case $k=3$.

The rest of the construction follows the same pattern. When we perform a blow up in $\mathcal{C}_k(r)$ of a coincidence set $F_{\mathcal{I}}^\sigma$ for some $\mathcal{I}$ with $|\mathcal{I}| = k-r-1$, having previously blown up all coincidence sets $F_{\mathcal{J}}^\sigma$ with $|\mathcal{J}| > k-r-1$, then the interior of this face, ie the portion lying over the interior of $F_{\mathcal{I}}$, is again fibered by copies of $\mathbb{S}_+^2$, and in this region the blowups of the sets F_i^σ for $i \in \mathcal{I}$ produce hemisphere fibers blown up at $|\mathcal{I}|$ distinct points. These fibers intersect the fibers of the previous faces $\mathfrak{C}_{\mathcal{J}}$ in similar ways, creating a new level in the tower of hemispheres over those corners.

We now state the final result, which is a description of all boundary faces and corners of $\mathfrak{C}_{\max}$. We have already described this face over the interior of $F_{\max}$: it is a fibration with each fiber a copy of $\mathbb{S}_+^2$ blown up at k distinct points. More generally, at any corner Z of $F_{\max}$, consider the preimage $\mathfrak{C}_{\max}(Z)$, ie the portion of the boundary of $\mathfrak{C}_{\max}$ lying over Z . This is a tower of hemispheres, each blown up at a set of points, so that altogether k points in this entire tower are blown up. [Figure 4](#) illustrates a typical scenario when $k=5$.

We can classify the towers of hemispheres which arise in this way. Recall first that associated to the corner Z is a tree T in the power set $\mathcal{G}_k$. We claim that the half-spheres in this tower correspond precisely to the nodes in T . Indeed, T consists of

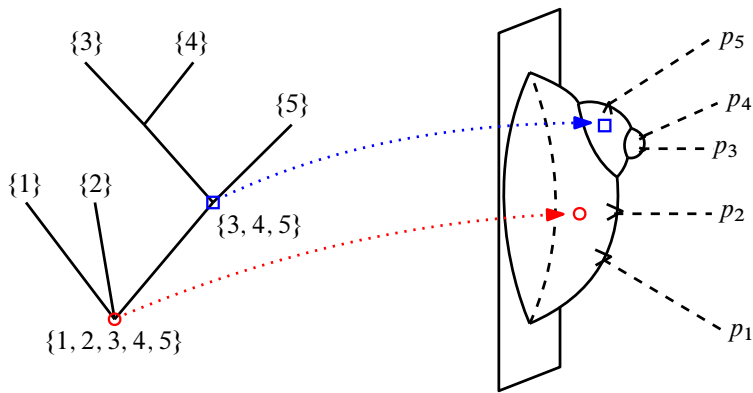


Figure 4: A tree structure encodes the clustering of bubbles.

all multi-indices $\mathcal{I}$ with $|\mathcal{I}| > 1$ such that $F_{\mathcal{I}} \supset Z$. We use induction on the height of the tree. The cases of height 0 and 1 were described earlier. The same argument applies when we pass from the space obtained up to height r as we take the blowups corresponding to the nodes in T at height $r + 1$. This shows that there is a half-sphere corresponding to each node of T . Conversely, the blowups required to construct this tower of half-spheres corresponds exactly to the sequence of blowups in this induction. In summary, we have:

Proposition 5 *The boundary faces and corners of $\mathfrak{C}_{\max}$ in $\mathcal{C}_k$ are in bijective correspondence with pairs $(T, \mathcal{I})$, where, to each corner $\mathfrak{Z}$ of $\mathfrak{C}_{\max}$, T is the tree associated to the corner Z of $F_{\max}$ under $\mathfrak{Z}$ and $\mathcal{I}$ is a node of T .*

We also consider *augmented trees*, which are simply trees as before, but now allowing the terminal nodes to consist of single-element sets. These nodes correspond to the faces $\mathfrak{C}_j$ of $\mathcal{C}_k$, but in particular to the boundary components of the hemispheres in the penultimate faces.

Cluster decompositions It is helpful to understand both $\mathcal{E}_k$ and $\mathcal{C}_k$ using the intuition that each point $\mathbf{p} \in M^k$ and $(\mathbf{p}, z) \in M^k \times M$ has a neighborhood $\mathcal{U}$ in which there is a well-defined decomposition of the points into clusters $\mathcal{Q}_1, \dots, \mathcal{Q}_{k'}$ for some $k' \leq k$. Each cluster $\mathcal{Q}_i$ has a center of mass ξ_i , and the k' -tuple $(\xi_1, \dots, \xi_{k'})$ determines a divisor, as do the points in each $\mathcal{Q}_i$. This cluster decomposition picture changes as the points move around, and the corners of $\mathcal{E}_k$ and $\mathcal{C}_k$ quantify precisely where the clusters trade points.

We say a bit more about this now. Recall the blowdown map $\beta: \mathcal{E}_k \rightarrow M^k$. Fix $\epsilon > 0$ and a collection of disjoint index sets $\mathcal{I}_* = \{\mathcal{I}_j\}$ for $j = 1, \dots, \ell$ such that $\bigcup \mathcal{I}_j = \{1, \dots, k\}$. Now define the open sets

(26) $\mathcal{U}_{\mathcal{I}_*, \epsilon} = \beta^* \{p \in M^k : d(p_{i'}, p_{j'}) > \epsilon \text{ if } i' \in \mathcal{I}_i \text{ and } j' \in \mathcal{I}_j \text{ for some } i \neq j\}.$

(The precise distance function used here is not important.) Thus in each $\mathcal{U}_{\mathcal{I}_*, \epsilon}$, coalescing occurs only within each cluster. Furthermore, any point $q \in \mathcal{E}_k \setminus F_{\max}$ lies in one of these open sets. Indeed, $\beta(q) = p$ lies in some corner Z of $\mathcal{E}_k$. Take the corresponding tree T . Note that since $q \notin F_{\max}$, the lowest element of T is some $\mathcal{I} \neq \mathcal{I}_{\max}$. Denote by $\mathcal{I}_1, \dots, \mathcal{I}_{\ell'}$ the vertices at the other extreme, ie the highest elements of T . If $\bigcup_{j=1}^{\ell'} \mathcal{I}_j \neq \mathcal{I}_{\max}$, set $\ell = \ell' + 1$ and define $\mathcal{I}_\ell = \mathcal{I}_{\max} \setminus \bigcup_{j=1}^{\ell'} \mathcal{I}_j$. These correspond to the “free” points which are not in any larger cluster. Otherwise, let $\ell = \ell'$.

Lemma 6 *There exists $\epsilon > 0$ such that $q \in U_{\mathcal{I}_*, \epsilon}$.*

Proof By the definition of T , $\beta(q) = p$ lies in the intersection of diagonals $\{\Delta_{\mathcal{I}_j}\}_{j=1}^{\ell-1}$. We must first choose ϵ so that the ϵ -ball around p in M^k does not intersect any other partial diagonals. Within this ball, clustering only happens amongst the points p_i with i lying in a single index set $\mathcal{I}_j$. Since $q \notin F_{\max}$, there are at least two clusters, ie $\ell \geq 2$, so we can suppose that 2ϵ is smaller than the minimal distance between the various clusters. It is then easy to see that p is contained in the open set $U_{\mathcal{I}_*, \epsilon}$. □

These cluster decompositions allow us to describe some further useful structure of the spaces $\mathcal{E}_k$ and $\mathcal{C}_k$.

Lemma 7 (1) *For any $q \in \mathcal{E}_k \setminus F_{\max}$ and associated neighborhood $\mathcal{U}_{\mathcal{I}_*, \epsilon}$, there exists a smaller neighborhood $\mathcal{U} \subset U_{\mathcal{I}_*, \epsilon}$ of q which is a product of neighborhoods $\mathcal{U}_i$ in $\mathcal{E}_{k(i)}$ for smaller values of $k(i)$. More specifically,*

$$\mathcal{U} = \prod_{j=1}^{\ell} \mathcal{U}_j, \quad \text{where } \mathcal{U}_j \subset \mathcal{E}_{|\mathcal{I}_j|}.$$

(2) *For $q \in \mathcal{E}_k$ and $\mathcal{U}$ a product neighborhood of q as in (1), there exist ℓ open sets $\mathcal{V}_i \subset \mathcal{C}_k$ with $\bigcup_{i=1}^{\ell} \mathcal{V}_i = \hat{\beta}^{-1}(\mathcal{U})$ such that in each $\mathcal{V}_i$, the fibration $\hat{\beta}$ is a product form*

$$(\hat{\beta}_{|\mathcal{I}_i|}) \colon \tilde{\mathcal{V}}_i \rightarrow \mathcal{U}_i) \times \prod_{j \neq i} \mathcal{U}_j,$$

where $\tilde{\mathcal{V}}_i \subset \mathcal{C}_{|\mathcal{I}_i|}$.

(3) Consider any point $q \in \mathcal{C}_k$, and let $(T, \mathcal{J})$ be the data encoding the hemisphere tower on which it sits. We assume that q lies in the interior of the maximal hemisphere of this tower. Let $\mathcal{J}' = \mathcal{J} \setminus \bigcup_{\mathcal{J} \supset \mathcal{I} \in T} \mathcal{I}$ be the set of free points in $\mathcal{J}$. If $\mathcal{J}' \neq \emptyset$, then there exists a neighborhood of z with product decomposition

$$\widehat{\beta}^{-1}(\mathcal{U}_{\mathcal{I}_*, \epsilon}) \supset \mathcal{V}_{\mathcal{C}} \times \mathcal{V}_{\mathcal{E}}, \quad \mathcal{V}_{\mathcal{C}} \subset \mathcal{C}_{|\mathcal{J}'|}, \mathcal{V}_{\mathcal{E}} \subset \mathcal{E}_{k-|\mathcal{J}'|}.$$

(4) If there are no free points, ie $\mathcal{J}' = \emptyset$, then there is a decomposition

$$q \in \mathcal{V}_{\mathcal{C}} \times \mathcal{V}_{\mathcal{E}}, \quad \mathcal{V}_{\mathcal{C}} \subset M, \mathcal{V}_{\mathcal{E}} \subset \mathcal{E}_k.$$

Proof For (1), the decomposition separates points into independent clusters. If $q \in \mathcal{U}_{\mathcal{I}_*, \epsilon}$, then it has a neighborhood which does not intersect any $F_{\mathcal{J}}$ except when $\mathcal{J} \subset \mathcal{I}_i$ for some i . All the possible blowups in this neighborhood occur within each cluster, so one can write $q = (q_1, \dots, q_{\ell})$ where $q_j \in E_{\mathcal{I}_j}$. The conclusion is now clear.

For (2), the clusters in $\mathcal{U}_{*, \epsilon}$ are separated by distance at least ϵ , so this decomposition exists. Restricting to each $\mathcal{V}_j$, then, only p_i for $i \in \mathcal{I}_j$ are included, so locally the map $\mathcal{C}_k \rightarrow \mathcal{E}_k$ splits in such a way that the lift of projection $(\mathfrak{p}, z) \rightarrow \mathfrak{p}$ restricts to the lift of $\{(\mathfrak{p}_{\mathcal{I}_j}, z) \rightarrow \mathfrak{p}_{\mathcal{I}_j}\} \times \mathfrak{p}_{\{1, \dots, k\} \setminus \mathcal{I}_j}$.

It is possible to refine the decomposition in (2) further when moving deeper into the tree. To prove (3) and (4), take a point in $\mathcal{C}_k$ lying above $(\mathfrak{p}, z)$ and let $\mathcal{C}_{\mathcal{J}}$ be the boundary face, some point on the interior of which projects to $(\mathfrak{p}, z)$. By the definition of T , the only boundary faces intersecting $\mathcal{C}_{\mathcal{J}}$ in this fiber correspond to the vertices $\mathcal{I} \supset \mathcal{J}$. Since z lies in the interior of a boundary face, it has a neighborhood which does not intersect any of these $\mathcal{C}_{\mathcal{I}}$. If $\mathcal{J}' \neq \emptyset$, there are $|\mathcal{J}'|$ free points contained in this region, hence only $\mathcal{C}_i$ for $i \in \mathcal{J}'$ intersect this neighborhood. There are no other free points in this neighborhood, and it contains no other boundary faces. In the first case, write $\mathfrak{p} = (\mathfrak{p}_{\mathcal{J}'}, \mathfrak{p}_{\mathcal{J}''})$ with $\mathcal{J}'' = \{1, \dots, k\} \setminus \mathcal{J}'$. Only the coincidence set $F_{\mathcal{J}'}^{\sigma}$, ie where $z = p_i$ with $i \in \mathcal{J}'$, intersects this neighborhood. Thus it is given by a neighborhood of $\mathcal{C}_{|\mathcal{J}'|}$ which does not intersect any partial diagonals. The other $\mathfrak{p}_{\mathcal{J}''} \in \mathcal{E}_{k-|\mathcal{J}'|}$ fill out the remaining base variables. For (4), there are no free points in this neighborhood, hence the fiber is just a neighborhood of M . Using the same reasoning as above yields the product decomposition. □

***b*-fibrations** The final and crucial fact we need is the following result:

Proposition 8 *The natural projection $\widehat{\beta}: \mathcal{C}_k \rightarrow \mathcal{E}_k$ is a *b*-fibration.*

We review the definition of b -fibrations in the [appendix](#), and their central importance in the theory of manifolds with corners. In particular, this proposition will be crucial in proving the fine regularity results for the families of fiberwise constant curvature metrics later in this paper.

Proof The proof is by induction on k . The result is obvious when $k = 1$, since in that case $\mathcal{E}_1 = M$ and $\mathcal{C}_1 = [M \times M; \Delta_{12}]$, and a map where the range is a manifold without boundary or corners is trivially a b -fibration.

Note that when $k = 2$, we have already written the explicit relationship between boundary-defining functions — see [\(10\)](#) — which proves the result in that case as well.

Now suppose that the assertion is true for any $\ell < k$. The first goal is to show that $\hat{\beta}$ is a b -map, which means that the pullback of any boundary-defining function $\rho_{\mathcal{I}}$ for a face $F_{\mathcal{I}} \subset \mathcal{E}_k$ is a product of boundary-defining functions in $\mathcal{C}_k$ (up to a nonvanishing smooth factor). If $q \in \mathcal{C}_k$ lies on the interior of some boundary face, then by [\(3\)](#) and [\(4\)](#) of the previous lemma, we can replace $\hat{\beta}$ by the product of maps

(27)
$$(\hat{\beta}_{|\mathcal{J}'|}: \mathcal{C}_{|\mathcal{J}'|} \rightarrow \mathcal{E}_{|\mathcal{J}'|}) \times \mathcal{E}_{k-|\mathcal{J}'|}.$$

When $|\mathcal{J}'| = k$, the tree T associated to (p, z) has precisely one node, and only the face $F_{\max}$ contains $\hat{\beta}(q)$. Furthermore, in this case, q lies in the interior of the boundary face $\mathfrak{C}_{\max}$, so just as for the cases $k = 2$ and 3 , we have

$$\hat{\beta}^* \rho_{\max} = AR_{\max} w,$$

where $R_{\max}$ is the boundary-defining function for $\mathfrak{C}_{\max}$ and w is a defining function for the boundaries of M blown up at k points. If $|\mathcal{J}'| < k$, then $\hat{\beta}_{|\mathcal{J}'|}$ in [\(27\)](#) is a b -fibration by the inductive hypothesis, so the boundary-defining functions $\rho_{\mathcal{I}}$ with $\mathcal{I} \subset \mathcal{J}'$ pull back to products of boundary-defining functions in $\mathcal{C}_{|\mathcal{J}'|}$ (up to a nonvanishing smooth function), while other boundary-defining functions $\rho_{\mathcal{I}}$ with $\mathcal{I} \not\subset \mathcal{J}'$ pull back trivially.

On the other hand, suppose q lies on the boundary of a boundary face, so there is an associated tree T and node $\mathcal{J}$, and let $\mathcal{I} \supset \mathcal{J}$ be the node directly over $\mathcal{J}$. (If there is no node above $\mathcal{J}$, then in the following we regard $R_{\mathcal{I}} = w$ as a boundary-defining function for the surface M blown up at the appropriate number of points.) By [\(1\)](#) of the previous lemma, there is a local product decomposition such that the points in $\mathcal{I}$ are separated from all others, and locally $\hat{\beta}$ splits as

$$(\hat{\beta}_{|\mathcal{I}|}: \mathcal{C}_{|\mathcal{I}|} \rightarrow \mathcal{E}_{|\mathcal{I}|}) \times \mathcal{E}_{k-|\mathcal{I}|}.$$

Furthermore,

(28) $\hat{\beta}^* \rho_{\mathcal{J}} = AR_{\mathcal{J}}w_{\mathcal{I}}, \quad \hat{\beta}^* \rho_{\mathcal{J}'} = R_{\mathcal{J}'} \quad \text{if } \mathcal{J}' \neq \mathcal{J},$

where $R_{\mathcal{J}}$ is the boundary-defining function of $\mathfrak{C}_{\mathcal{J}}$ and $w_{\mathcal{I}}$ is the boundary-defining function of $\mathfrak{C}_{\mathcal{I}}$ in that corner. All other boundary-defining functions pull back trivially by virtue of this product decomposition.

The remaining issue is to show that $\hat{\beta}$ does not map any one of the boundary hypersurfaces $\mathfrak{C}_{\mathcal{I}}$ of $\mathcal{C}_k$ to corners of $\mathcal{E}_k$. This is clear from the construction since $\mathfrak{C}_{\mathcal{I}}$ maps to the boundary hypersurface $F_{\mathcal{I}}$.

These facts together prove that $\hat{\beta}$ is a b -fibration. □

3 Geometry of merging cone points

3.1 Review of existing results

We study constant curvature metrics with conical singularities, which is defined by the following: a smooth metric on $M \setminus \mathfrak{p}$ with constant curvature, and near the punctures the metric is asymptotically conical, that is, there exist local coordinates such that the metric is given by

$$e^{\phi} |z|^{2(\beta-1)} |dz|^2$$

with ϕ being smooth. There is also a geodesic coordinate description given by

$$g = \begin{cases} d\tau^2 + \beta^2 \tau^2 d\phi^2, & K = 0, \\ d\tau^2 + \beta^2 \sinh^2 \tau d\phi^2, & K = -1, \\ d\tau^2 + \beta^2 \sin^2 \tau d\phi^2, & K = 1. \end{cases}$$

In particular, in each case with curvature of different signs, asymptotically it is always given by the flat metric.

The central problem to study is: given $(c, \mathfrak{p}, \vec{\beta}, K, A)$ as the conical data, does there exist a constant curvature conical metric with this data? Is the metric unique? The study of this singular uniformization problem has a long history and has been very active recently. When the curvature K is nonpositive, the conclusion is relatively straightforward. By the results of McOwen [17] and Luo and Tian [14], for any fixed $(c, \mathfrak{p}, \vec{\beta}, K, A)$ that satisfy the Gauss–Bonnet formula and $K \leq 0$, there exists a unique constant curvature conical metric prescribed by the tuple. When $K > 0$, the situation is complicated depending on the cone angles. When all the cone angles are smaller

than 2π , by the results of Troyanov [22] and Luo and Tian [14], there is a unique constant curvature metric when the cone angles are in the “Troyanov region” (2). And when all the cone angles are less than 2π , there is a moduli space structure, obtained by the first author and Weiss [16].

When some of the cone angles are bigger than 2π , unlike the cases above, there is no uniform result. When M is a sphere, there are some results depending on the number of cone points. When $k = 2$, Troyanov [21] gave the results. When $k = 3$, the characterization via complex analytical methods was given by Eremenko [8] and Umehara and Yamada [23]. When $k = 4$ with symmetry, complex analysis techniques can also be applied, see Eremenko, Gabrielov and Tarasov [10]. When the genus of M is greater than 0, there are some general existing results by Carlotto and Malchiodi [3; 4] and Bartolucci, De Marchis and Malchiodi [2]. In [20] some new families of metrics are constructed by relating to Strebel differentials. We also mention here some related problems, including Toda systems and mean field equations; see Chen and Lin [5; 6] and references therein. In particular, the recent result by Lee, Lin, Yang and Zhang [13] considering the singularity formation of two-point collision is morally related to the cone points merging behavior studied in this paper.

Recently, the breakthrough of Mondello and Panov [19] shows the necessary condition of existence on a sphere by the holonomy condition

$$(29) \quad d_1(\vec{\beta} - \vec{1}, \mathbb{Z}_{\text{odd}}) \geq 1.$$

They also showed that when the strict inequality holds (“noncoaxial” situation) there exists at least one such metric. The recent results by Dey [7], Kapovich [11] and Eremenko [9] determined the necessary condition of existence when the equality holds in (29).

When trying to extend the result of [16] to get a smooth manifold structure in this case, we found that there are obstructions, reflected in the fact that the linearized operator fails to be surjective on some subvarieties. A key component in the construction of Mondello and Panov is the splitting of cone points, which turns out to be the key to resolve the analytic obstruction. We are going to describe the geometry of this process (splitting, or equivalently merging, of cone points) below.

3.2 Local geometry of merging cone points

In the next three sections, we consider the behavior of constant curvature metrics with some of the cone points merging together (or, equivalently, when a cone splits into

several cones). The cases we are going to study in this paper include all hyperbolic and flat metrics, and spherical metrics with angles less than 2π .

We first describe this process locally by looking at the family of metrics, parametrized by $t \in [0, \epsilon)$,

$$(30) \quad g(t) = |z - p_1(t)|^{2(\beta_1-1)} |z - p_2(t)|^{2(\beta_2-1)} |dz|^2,$$

where $p_1(t)$ and $p_2(t)$ are smooth coordinates that parametrize two moving points on M . We also require $p_1(0) = p_2(0)$. When $t \neq 0$, it gives a metric with two separate cone points with angles $2\pi\beta_1$ and $2\pi\beta_2$. And the distance between the two cone points decreases as t goes to 0. Eventually, in the limit $t = 0$, the metric is given by

$$|z|^{2(\beta_1+\beta_2-2)} |dz|^2,$$

which is a conic metric with a single cone angle $2\pi(\beta_1 + \beta_2 - 1)$ if $\beta_1 + \beta_2 - 1 > 0$.

This process of merging two cone points can be generalized to multiple points. After merging j points each with angle $\{2\pi\beta_i\}_{i=1}^j$, the angle we get is given by the defect formula

$$(31) \quad 2\pi\beta_0 := 2\pi \left(\sum_{i=1}^j \beta_i - (j-1) \right).$$

One thing to notice from the defect formula is that not all conic points can be merged; it only happens when the “admissible condition” below is satisfied, otherwise there is an obstruction to produce a new conic points. When $\beta < 0$, $|z|^{2(\beta-1)} |dz|^2$ is no longer conical, and we get some open ends, the form of which depend on the curvature constant. Therefore, we define the following condition for merging:

Definition 9 We say a set of cone angles $\{\beta_i\}_{i \in \mathcal{I}}$ is *admissible* if

$$(32) \quad \sum_{i \in \mathcal{I}} \beta_i > |\mathcal{I}| - 1.$$

In particular, this implies that when $k = 2$, the two cone points must satisfy $\beta_1 + \beta_2 > 1$.

3.3 Global geometry

The metrics we are going to consider in this paper are the following:

- (33) • flat or hyperbolic conical metrics with $\vec{\beta} \in \mathbb{R}_+^k$ such that the Gauss–Bonnet formula (1) is satisfied; or
- spherical conical metrics with $\vec{\beta} \in (0, 1)^k$ satisfying either (2) if $k \geq 3$ or $\text{genus}(M) > 0$, or $\beta_1 = \beta_2$ if $k = 2$ and $M = \mathbb{S}^2$.

From [22; 14; 17], there is a unique constant curvature metric for each of the configuration $(\mathfrak{p}, \vec{\beta})$ with $\vec{\beta}$ satisfying one of the conditions above. Now, considering the family of metrics with varying cone points, we would like to understand the uniform behavior of those metrics when some of cone points merge together. Using the defect formula we can see that

$$\beta_0 - 1 = \sum_{i=1}^j (\beta_i - 1).$$

Together with the Gauss–Bonnet formula, this implies that the curvature remains the same constant in this merging process.

Because of the restriction of cone angles defined above, not all cones can be merged to produce new cones. Moreover, the Troyanov constraint for spherical metrics is not preserved in this merging process. Therefore, for a given set of cone angles $\vec{\beta}$, the fiber conical metrics might potentially only be defined in a subset of $\mathcal{C}_k$. And we define the following admissible region:

Definition 10 For fixed $\vec{\beta} = \{\beta_i\}_{i=1}^k$ satisfying one of the conditions in (33), the admissible extended configuration space $\mathcal{E}_{k,\vec{\beta}}$ is defined to be the union of configurations in $\mathcal{E}_k$ with which there exists a constant conical curvature metric on M , ie

(34) $\mathcal{E}_{k,\vec{\beta}} := \mathcal{E}_k^0 \cup \bigcup_{\substack{\sum_{i \in \mathcal{I}_j} \beta_i > |\mathcal{I}_j| - 1 \\ \text{for all } j}} \left\{ \mathfrak{p} \in \bigcap_{j=1}^l F_{\mathcal{I}_j} : \text{there exists a constant curvature metric} \right.$
 $\left. \text{with configuration } (\mathfrak{p}, \{\sum_{i \in \mathcal{I}_j} (\beta_i - 1) + 1\}_{j=1}^l) \right\}.$

The admissible extended configuration family $\mathcal{C}_{k,\beta}$ is defined in a similar way by only considering the admissible combinations, or equivalently $\mathcal{C}_k \supset \mathcal{C}_{k,\vec{\beta}} = \pi_k^{-1}(\mathcal{E}_{k,\vec{\beta}})$.

Flat and hyperbolic cases The metrics we consider will be any $\vec{\beta} \in \mathbb{R}_+^k$ such that the Gauss–Bonnet formula (1) is satisfied. In particular, while the Gauss–Bonnet constraint

$$\sum_{i=1}^k (\beta_i - 1) \leq \chi(M)$$

gives an upper bound of β_i , we do not require the cone angles to be uniformly small.

The admissible extended configuration space is relatively easy in this case. As long as the merging cone angles are admissible in the sense of Definition 9, there exists a flat

(or hyperbolic, depending on the angle combination) conical metric after merging. In particular, in this case we have

$$\mathcal{E}_{k,\vec{\beta}} = \mathcal{E}_k^0 \cup \bigcup_{\sum_{i \in \mathcal{I}} \beta_i > |\mathcal{I}| - 1} F_{\mathcal{I}}$$

and

$$\mathcal{C}_{k,\vec{\beta}} = \mathcal{C}_k^0 \cup \bigcup_{\sum_{i \in \mathcal{I}} \beta_i > |\mathcal{I}| - 1} \mathfrak{C}_{\mathcal{I}}.$$

When $k = 2$ and $M = \mathbb{S}^2$, there are neither flat nor hyperbolic conical metrics on M . When $k \geq 3$ or the genus of M is greater than 0, $\mathcal{C}_{k,\vec{\beta}} \setminus \mathcal{C}_k^0$ is nonempty in general. In particular, when the genus of M and the cone angles are all sufficiently large, all directions of merging will be allowed and in that situation $\mathcal{C}_{k,\vec{\beta}} = \mathcal{C}_k$.

The spherical case In this case all the cone angles are assumed to be less than 2π , hence $\vec{\beta} \in (0, 1)^k$. Notice that by the relation (31), the cone angle obtained after merging, denoted by $2\pi\beta_0$, is also less than 2π . Therefore, during this merging process we stay inside the class defined in (33).

The extra rigidity in the spherical case of (33) and the fact that merging does not preserve this constraint make $\mathcal{E}_{k,\vec{\beta}}$ and $\mathcal{C}_{k,\vec{\beta}}$ much more complicated to describe, compared to the previous case. We illustrate a few cases here, keeping in mind that the football case ($M = \mathbb{S}^2$, $k = 2$ and $\beta_1 = \beta_2$) is special for a reason that will be made clear in Section 7.

We start with $M = \mathbb{S}^2$. From [21], there is no “tear-drop” metric on $\mathbb{S}^2$ with only one conical point. Hence there is no admissible merging on $\mathcal{C}_2$, which implies $\mathcal{C}_{2,\vec{\beta}} = \mathcal{C}_2^0$. On a sphere with three conical points and assuming $0 < \beta_1 \leq \beta_2 \leq \beta_3 < 1$, the Troyanov condition is given by

$$(35) \quad 2\beta_1 + 1 > \sum_{i=1}^3 \beta_i, \quad \text{or} \quad \beta_1 > \beta_2 + \beta_3 - 1.$$

We show that it cannot merge to a football. Since all cone angles are less than 2π , the merging process decreases the angle strictly, ie $\beta_i + \beta_j - 1 < \min\{\beta_i, \beta_j\}$. Hence the only feasible way to achieve a football would be merging the two bigger angles, and this gives

$$\beta_2 + \beta_3 - 1 = \beta_1,$$

which contradicts (35). Since the three points cannot be simultaneously merged into one point either, we have $\mathcal{C}_{3,\vec{\beta}} = \mathcal{C}_3^0$.

When $k \geq 4$, depending on different $\vec{\beta}$, the situation can be very different. Assuming $0 < \beta_1 \leq \beta_2 \leq \beta_3 \leq \beta_4 < 1$ satisfying the Troyanov condition

$$2\beta_1 + 2 > \sum_{i=1}^4 \beta_i,$$

if $1 < \beta_1 + \beta_4 = \beta_2 + \beta_3$ (which can be achieved for example by taking $\vec{\beta} = (\frac{1}{2}, \frac{2}{3}, \frac{2}{3}, \frac{5}{6})$), then by merging the two groups $\{\beta_1, \beta_4\}$ and $\{\beta_2, \beta_3\}$ simultaneously, we get a football; however, since we cannot split a football to get an admissible 3-point configuration, one cannot merge $\{\beta_1, \beta_4\}$ nor $\{\beta_2, \beta_3\}$ without the other group. That is, $\mathfrak{C}_{14} \cap \mathfrak{C}_{23} \subset \mathcal{C}_{4, \vec{\beta}}$ but $\mathfrak{C}_{14}^0 \cup \mathfrak{C}_{23}^0 \not\subset \mathcal{C}_{4, \vec{\beta}}$. However, it is possible to merge $\{\beta_2, \beta_4\}$ or $\{\beta_3, \beta_4\}$ in most of those cases (as in the example given above), hence $\mathfrak{C}_{24}^0 \cup \mathfrak{C}_{34}^0 \subset \mathcal{C}_{4, \vec{\beta}}$. In contrast, if $\beta_1 + \beta_4 \neq \beta_2 + \beta_3$ then one cannot get a football but merging into a 3-point configuration is still possible.

When the genus of M is greater than 1, the description of $\mathcal{C}_{k, \vec{\beta}}$ still depends on $\vec{\beta}$ and the Troyanov condition; however, since we will not get any football configuration in this case, it is analytically the same as the flat or hyperbolic cases.

4 Preliminary analysis

Our approach to the study of the geometric problems described in the last section involves the analysis of conic elliptic operators on spaces with isolated conic singularities, as employed already in [16]. The new feature here is that we study families of such operators on spaces with coalescing conic singularities.

4.1 b -vector fields on M and conic elliptic operators

Let M be a manifold with isolated conic singularities at the collection of points $\mathfrak{p} = \{p_1, \dots, p_k\}$, and denote by $\widehat{M} = [M; \mathfrak{p}]$ the blowup of M at these points. Thus $\widehat{M}$ is a manifold with k boundary components; when $\dim M = 2$, each boundary component is diffeomorphic to a circle. Choose local polar coordinates (r, θ) near any p_j , so r is a boundary-defining function for the boundary face created by blowing up p_j and θ is a set of local coordinates on that face, eg θ is the angular coordinate if $\partial \widehat{M}$ is a union of circles. We recall the space of b -vector fields, which consists of all smooth vector fields on $\widehat{M}$ which are tangent to the boundary. In these local coordinates,

$$\mathcal{V}_b = \mathcal{C}^\infty\text{-span}\{r\partial_r, \partial_\theta\}.$$

A differential operator is called a b -operator if it can be written locally as a finite sum of products of elements of $\mathcal{V}_b$,

$$L = \sum_{j+|\alpha|\leq m} a_{j\alpha}(r, \theta)(r\partial_r)^j \partial_\theta^\alpha$$

(here we continue to think of θ as possibly multidimensional and α a multi-index). This operator is called b -elliptic if its b -symbol is nonvanishing,

$${}^b\sigma_m(L) = \sum_{j+|\alpha|=m} a_{j\alpha}(r, \theta)\rho^j \eta^\alpha \neq 0 \quad \text{for } (\rho, \eta) \neq (0, 0).$$

Finally, we say that L is an elliptic conic operator if $L = r^{-m}A$ where A is an elliptic b -operator of order m . We are primarily concerned with elliptic conic operators of order 2, in particular the Laplacian on a surface with isolated conic singularities.

Now suppose that X is a more general manifold with corners, ie any point $q \in X$ has a neighborhood $\mathcal{U}$ diffeomorphic to a neighborhood of the origin in an orthant $\mathbb{R}_+^k \times \mathbb{R}^{n-k}$ (in which case q lies on a codimension k corner). As before we define the space of b -vector fields on X to consist of the smooth vector fields which are tangent to all boundary faces. In local coordinates $(x_1, \dots, x_k, y_1, \dots, y_{n-k})$ near a codimension k corner, with all $x_j \geq 0$, we have that

$$\mathcal{V}_b(X) = \mathcal{C}^\infty\text{-span}\{x_j \partial_{x_j}, \partial_{y_\ell}\}.$$

Our main examples of manifolds with corners here, of course, are the extended configuration spaces $\mathcal{E}_k$ and the extended configuration families $\mathcal{C}_k$. We do not consider here b -differential operators on general manifolds with corners. Instead, as motivated by our problem, the fibers $\pi^{-1}(q) \subset \mathcal{C}_k$ are unions of two-dimensional surfaces with boundary, possibly “tied” along their boundaries, and we consider the families of elliptic conic operators on these fibers, parametrized by $q \in \mathcal{E}_k$. Nonetheless, it still turns out to be important to consider b -vector fields on the entire space $\mathcal{C}_k$. In doing so, it is convenient to organize the boundary hypersurfaces into three types, corresponding to the boundary faces $\mathfrak{C}_\mathcal{I}$ which resolve point coincidences, the faces $\mathfrak{C}_i$ corresponding to individual conic points, and the faces corresponding to the blowups of each fiber M_p . We denote the boundary-defining functions for these by $\rho_\mathcal{I}$, ρ_i and ρ , respectively.

4.2 b -Hölder spaces on M and mapping properties of conic elliptic operators

Conic elliptic operators act naturally between weighted Sobolev and Hölder spaces defined relative to differentiations by the vector fields in $\mathcal{V}_b(M)$. Our ultimate problem

is nonlinear, so we define only the b -Hölder spaces and state the key mapping properties on these. For simplicity, we restrict attention to the two-dimensional case.

Definition 11 (weighted b -Hölder spaces) For any function defined on M_p , define the seminorm $[u]_{b;\delta}$ in the usual way away from a neighborhood of the boundary faces, while in each such neighborhood, in local polar coordinates, we set

$$[u]_{b;\delta} := \sup_{0 < \bar{r} < r_0} \sup_{\bar{r} \leq r, r' \leq 2\bar{r}} \frac{|u(r, \theta) - u(r', \theta')| \bar{r}^\delta}{|(r, \theta) - (r', \theta')|^\delta}.$$

Then $\mathcal{C}_b^{0,\delta}(M)$ consists of the functions u which are bounded and for which $[u]_{b;\delta} < \infty$.

Next, for any $\ell \in \mathbb{N}$, define $\mathcal{C}_b^{\ell,\delta}(M)$ to consist of all functions u such that $V_1 \cdots V_j u \in \mathcal{C}_b^{0,\delta}(M)$ for every $j \leq \ell$ and $V_i \in \mathcal{V}_b(M)$.

Finally, $r^\mu \mathcal{C}_b^{k,\delta}(M) = \{u = r^\mu v : v \in \mathcal{C}_b^{\ell,\delta}(M)\}$.

It is immediate from these definitions that if L is a second-order conic elliptic operator, then, for every $\ell \geq 2$,

(36)
$$L: r^\mu \mathcal{C}_b^{\ell,\delta}(M) \rightarrow r^{\mu-2} \mathcal{C}_b^{\ell-2,\delta}(M)$$

is bounded.

It can happen that this map does not have closed range for certain values of μ . Indeed, μ is called an indicial root of L if there exists some function $\phi(\theta)$ such that $L(r^\mu \phi(\theta)) = \mathcal{O}(r^{\mu-1})$. The expected order of decay or blowup is $r^{\mu-2}$, so μ is an indicial root only if there is some leading-order cancellation. It is not hard to check that if μ is an indicial root, then an appropriate sequence of cutoffs of $r^\mu \phi(\theta)$ can be constructed to show that (36) does not have closed range. This is explained at length in [16]. The following is the basic Fredholm result, proved in [15], but see also [16].

Proposition 12 If μ is not an indicial root, then (36) is Fredholm.

If $L = \Delta + V$ where $V \in \mathcal{C}_b^{0,\delta}$ for example (or even just $V \in r^{-2+\epsilon} \mathcal{C}_b^{0,\delta}$ for any $\epsilon > 0$), then, at a conic point p with cone angle $2\pi\beta$, the indicial roots consist of the set of values j/β for $j \in \mathbb{Z}$. The coefficient $\phi(\theta)$ corresponding to the indicial root j/β can be any linear combination of $\sin j\theta$ and $\cos j\theta$.

We also consider $L = \Delta + V$, when $V \in \mathcal{C}_b^{\ell,\delta}$, as an unbounded operator

(37)
$$L: \mathcal{C}_b^{\ell+2,\delta}(M) \rightarrow \mathcal{C}_b^{\ell,\delta}(M).$$

We then seek to characterize its domain, ie the (nonclosed) subspace

$$\mathcal{D}_b^{\ell,\delta}(L) = \{u \in \mathcal{C}_b^{\ell+2,\delta}(M) : Lu \in \mathcal{C}_b^{\ell,\delta}(M)\}.$$

This is called the Hölder Friedrichs domain for L .

Proposition 13 [16] *The space $\mathcal{D}_b^{\ell,\delta}(L)$ consists of functions $u \in \mathcal{C}_b^{\ell+2,\delta}(M)$ such that, near each conic point p ,*

$$u = a_0 + \sum_{j=1}^{N(\beta)} (a_{j1} \cos j\theta + a_{j2} \sin j\theta) r^{j/\beta} + \tilde{u}$$

for some constants a_{j1} and a_{j2} , where $N(\beta)$ is the largest value N such that $N/\beta < 2$, and $\tilde{u} \in r^2 \mathcal{C}_b^{\ell+2,\delta}(M)$.

4.3 Families of conic elliptic operators

The previous subsection reviews a few standard results about conic elliptic operators on surfaces. Our interest is in families of such operators, particularly as the conic points coalesce. In particular, suppose L_p is such a family where the cone points are located at some simple divisor $p \in \mathcal{D}_k^s$. A key difficulty in extending the theory for individual operators to families is that the function space on which L_p acts vary with p . We use the geometric machinery developed above to handle this.

More specifically, we first consider weighted Hölder spaces on $\mathcal{C}_k$ and the restrictions of these spaces to the fibers $\pi^{-1}(q)$ for $q \in \mathcal{E}_k$, then define the appropriate families of weighted Hölder spaces on which we may describe extensions of the mapping properties.

In the following, let $\vec{v}$ be a weight vector, with components indexed by the hypersurfaces of $\mathcal{C}_k$. We then define in the obvious way the weighted Hölder spaces $r^{\vec{v}} \mathcal{C}_b^{\ell,\delta}(\mathcal{C}_k)$. To make sense of the restrictions of these spaces to each fiber, we need an easy result.

Lemma 14 *The restriction of $r^{\vec{v}} \mathcal{C}_b^{\ell,\delta}(\mathcal{C}_k)$ to each fiber $M_p = \pi^{-1}(p)$ is precisely the weighted space $r^{\vec{v}} \mathcal{C}_b^{\ell,\delta}(M_p)$, where (abusing notation slightly), the weight vector $\vec{v}$ here has components indexed by the boundary components of M_p .*

Proof Observe first that the boundary faces of M_p are the components of the intersection $(\bigcup_{\mathcal{I}} \mathcal{C}_{\mathcal{I}} \cup \bigcup_i \mathcal{C}_i) \cap M_p$.

The fact that the restriction of $\mathcal{C}_b^{0,\delta}(\mathcal{C}_k)$ to M_p equals $\mathcal{C}_b^{0,\delta}(M_p)$ is straightforward from the definitions since the boundary-defining functions for the faces of $\mathcal{C}_k$ restrict

to the boundary-defining functions for the faces of each fiber, and there are coordinates tangent to the faces of C_k which also restrict to the θ coordinates on each fiber.

Next, the weight functions restrict naturally as well. Thus we must finally show that $C_b^{\ell,\delta}(C_k)$ restricts to $C_b^{\ell,\delta}(M_p)$. For this, note that if $V \in \mathcal{V}_b(M_p)$, then there is an extension of V to $\hat{V} \in \mathcal{V}_b(C_k)$. Thus, if $u \in C_b^{\ell,\delta}(C_k)$ and $V_i \in \mathcal{V}_b(M_p)$, and if $\hat{V}_i$ are the lifts for $i \leq \ell$, then $\hat{V}_1 \cdots \hat{V}_\ell u \in C_b^{0,\delta}(C_k)$, and the restriction of this expression is just $V_1 \cdots V_\ell u$, which by the first step lies in $C_b^{0,\delta}(M_p)$. $\square$

Next, for any fiber M_p in C_k , if $\vec{\beta}$ is the set of cone angles, then we construct the Friedrichs–Hölder domain by including at each p_j the terms with local expressions $r^{j/\beta_i} \phi_j(\theta)$ for $0 \leq j < 2\beta_i$.

Definition 15 If $\vec{\beta}$ is fixed, then the fiberwise Friedrichs–Hölder domain associated to a family of conic metrics g_p is given by

$$\mathcal{D}_{\text{Fr}}^{\ell,\delta}(C_k) = \{u \in C_b^{\ell,\delta}(C_k) : \Delta_{g_p}(u|_{M_p}) \in C_b^{\ell,\delta}(M_p), \, p \in \mathcal{E}_k\}.$$

It is clear that $\mathcal{D}_{\text{Fr}}^{\ell,\delta}(C_k)$ varies smoothly with $p \in \mathcal{E}_k$ over the regular fibers (where all the p_i are distinct). We may proceed just as in [16] to obtain that

$$(38) \qquad u = a_0 + \sum_{j=1}^{[2\beta]} r^{k/\beta} \phi_j(\theta) + \tilde{u}, \quad \tilde{u} \in r^2 C_b^{\ell+2,\delta},$$

where as before $\phi_j = a_{j1} \cos j\theta + a_{j2} \sin j\theta$. In this free region, smoothness follows from the smoothness of the boundary-defining functions with respect to p .

When p approaches a face $F_{\mathcal{I}}$ of $\mathcal{E}_k$, the aggregate cone angle is $\beta = \sum_{i \in \mathcal{I}} (\beta_i - 1) + 1$, and functions in $\mathcal{D}_{\text{Fr}}^{\ell,\delta}(C_k)$ have the form

$$(39) \qquad u = a_0 f_0(w) + \sum_{j=1}^{[2\beta]} \rho^{j/\beta} f_j(w) + \tilde{u}, \quad \tilde{u} \in \rho^2 C_b^{\ell+2,\delta}(C_k);$$

here $\rho = \rho_{\mathcal{I}}$ is the boundary-defining function for the half-sphere $\mathfrak{C}_{\mathcal{I}}$ and the $f_j(w)$ for $j = 0, \dots, [2\beta]$ are functions on $\mathfrak{C}_{\mathcal{I}}$ such that each $\rho^{j/\beta} f_j(w)$ is (formally) annihilated by the rescaled operator $\rho^{2\beta} \Delta_{g_p}$ at $\mathfrak{C}_{\mathcal{I}}$. On this front face, the conic points are all separated. Therefore, functions in the Friedrichs–Hölder domain annihilated by $\rho^{2\beta} \Delta_{g_p}$ are as in (38), where each cone points on this face has the obvious cone angle extended from the interior of C_k . This means that functions on fibers M_p near this face extend smoothly to this face.

5 Flat conical metrics

We now begin our analysis of the space of constant curvature conic metrics by studying the simplest case, when the problem is linear. Thus we fix a closed surface M and a set of cone angles $\vec{\beta}$ such that

$$(40) \quad \chi(M, \vec{\beta}) := \chi(M) + \sum_{j=1}^k (\beta_j - 1) = 0.$$

It is standard that if (40) is satisfied, then for each marked conformal structure (c, p) there exists a unique flat conic metric with area 1 and cone angle $2\pi\beta_i$ at p_i . Our goal in this section is to show that this family of flat conic metrics is polyhomogeneous on $\mathcal{C}_k$.

This is a local regularity theorem, so we fix a smooth family $g_0(c)$ of smooth constant curvature metrics on M representing a neighborhood in the space of (unmarked) conformal structures. For each $p \in \mathcal{D}_k^s$, consider the linear problem

$$\Delta_{g_0(c)} G = 2\pi \sum_{i=0}^k (\beta_i - 1) \delta_{p_i}.$$

Then

$$-\int_M K = -2\pi \chi(M) = \int_M 2\pi \sum (\beta_i - 1) \delta_{p_i};$$

we see that the Liouville equation

$$\Delta_{g_0(c)} u + K_{g_0(c)} = 0$$

has a solution $u = G$ which is unique if we require that $\int_M G = 0$. This solution G is essentially the Green function for Δ_{g_0} . It clearly depends smoothly on c , $p \in \mathcal{D}_k^s$ and $z \in M \setminus p$, and near each p_i has the form

$$G \sim (\beta_i - 1) \log|z| + \tilde{G}_i,$$

where each $\tilde{G}_i$ is $\mathcal{C}^\infty$ in a neighborhood of p_i . We then define

$$(41) \quad g_0(c, p, \vec{\beta}) = e^{2G} g_0(c).$$

Each of these metrics is flat and, because of the asymptotic structure of G , has a conic singularity with cone angle $2\pi\beta_i$ at p_i . This family of metrics is smooth when all the points p_i are distinct.

Theorem 16 Fix $\vec{\beta}$ satisfying (40); then the family of flat metrics $g_0(\mathfrak{c}, \mathfrak{p}, \vec{\beta})$ extends to a polyhomogeneous family of fiber metrics on $\mathcal{C}_k$.

Because $g_0(\mathfrak{c})$ is smooth, it suffices to show that the scalar function G extends to be polyhomogeneous on $\mathcal{C}_k$. Note that by the remarks above, G is $\mathcal{C}^\infty$ on the interior of the extended configuration family, so our task is to examine its behavior near each of the boundary faces and corners of $\mathcal{C}_k$. In other words, we must prove that there exists an index family $\{E_{\mathcal{I}}\}$ such that

$$G \sim \sum_{(j,\ell) \in \mathcal{E}_{\mathcal{I}}} (\rho_{\mathcal{I}})^j (\log \rho_{\mathcal{I}})^\ell a_{j\ell}(w_{\mathcal{I}}),$$

where $w_{\mathcal{I}}$ are variables in the interior of each $\mathfrak{C}_{\mathcal{I}}$ and each $a_{j\ell}$ is polyhomogeneous with index family $\{E_{\mathcal{J}}\}_{\mathcal{J} \neq \mathcal{I}}$. Note that polyhomogeneity of G near the simplest faces $\mathfrak{C}_i$ is obvious. We also suppress the smooth dependence of G on $\mathfrak{c}$.

5.1 The case of two cone points

The proof of Theorem 16 is by induction on k . We begin with the proof when $k = 2$.

Proposition 17 When $k = 2$, $G(z, \mathfrak{p})$ is polyhomogeneous on $\mathcal{C}_2$.

Proof Suppose that the two points p_1 and p_2 converge at the point p_{12} , which we may as well assume is fixed and is the center of mass of these two points. Referring to the local coordinates in Section 2.3, we may as well restrict to a slice where $\zeta = \zeta_0 = 0$. Then

(42)
$$\begin{aligned} G(z, \mathfrak{p}) &= (\beta_1 - 1) \log|z - w| + (\beta_2 - 1) \log|z + w| \\ &= (\beta_1 - 1) \log|re^{i\phi} - \rho e^{i\theta}| + (\beta_2 - 1) \log|re^{i\phi} + \rho e^{i\theta}|. \end{aligned}$$

By (8), $r = R_{12} \cos \omega$, $\rho = R_{12} \sin \omega$, so

(43)
$$\begin{aligned} G(z, \mathfrak{p}) &= (\beta_{12} - 1) \log R_{12} + (\beta_1 - 1) \log|\cos \omega e^{i(\phi-\theta)} - \sin \omega| \\ &\quad + (\beta_2 - 1) \log|\cos \omega e^{i(\phi-\theta)} + \sin \omega|. \end{aligned}$$

Here, and later in this paper, we set

(44)
$$\beta_{12} = \beta_1 + \beta_2 - 1,$$

ie $2\pi\beta_{12}$ is the cone angle which results when two cone points with cone angles $2\pi\beta_1$ and $2\pi\beta_2$ merge. The expression in (43) is certainly polyhomogeneous as $R_{12} \rightarrow 0$

away from $\omega = 0$ (the corner, where $\mathfrak{C}_{12}$ meets $[M; \{p_{12}\}]$) and the points where $e^{i(\phi-\theta)} = \pm 1$. To understand behavior near the corner, write $s = \rho/r = \tan \omega$, so that when $\omega < \frac{\pi}{4}$, say, and recalling that $r = R_{12} \cos \omega$, we have

(45)
$$G(z, \mathfrak{p}) = (\beta_{12} - 1) \log r + \frac{1}{2}(\beta_1 - 1) \log(1 - 2s \cos(\theta - \phi) + s^2) \\ + \frac{1}{2}(\beta_2 - 1) \log(1 + 2s \cos(\theta - \phi) + s^2).$$

The second and third terms on the right are smooth and vanish at $s = 0$. Note that there is an apparent asymmetry in the indices 1 and 2 here; however, when the points p_1 and p_2 are switched, the angle θ changes to $\theta + \pi$, so this expression is actually symmetric after all. Finally, near $\omega = \frac{\pi}{4}$ and $\theta = \phi$, for example, write $\tan \omega = 1 + \sigma$, so that $\cos \omega = 1/\sqrt{2 + 2\sigma + \sigma^2}$. Then

$$G(z, \mathfrak{p}) = (\beta_{12} - 1) \log(R_{12}/\sqrt{2 + 2\sigma + \sigma^2}) + (\beta_1 - 1) \log|e^{i(\theta-\phi)} - 1 - \sigma| \\ + (\beta_2 - 1) \log|e^{i(\theta-\phi)} + 1 + \sigma|,$$

and this is obviously polyhomogeneous around the face $\mathfrak{C}_1$ created by blowing up $\sigma = 0$. The argument is the same near $\mathfrak{C}_2$.

The assertion about polyhomogeneity of G on $\mathcal{C}_2$ is now proved. □

5.2 Inductive proof for higher k

Proof of Theorem 16 Suppose that the result has been proven for $\mathcal{C}_j$ for any $j < k$. Without loss of generality, we can restrict to the slice with the fixed center of mass $\zeta = \zeta_0 = 0$.

We first consider the case that is away from the central diagonal $\mathfrak{C}_{1\dots k}$, that is, at most $k - 1$ points can merge together. This is the case for example when the configuration $\vec{\beta} \in \mathbb{R}^k$ is such that $\sum_{i=1}^k (\beta_i - 1) \leq 1$. Then we can cover $\mathcal{E}_{k,\vec{\beta}}$ by open sets $\{U_{\mathcal{I}^*,\epsilon}\}$ defined in (26). That is, the only possible merging happens within the subclusters, and the distance between any clusters is bounded away from 0. From Lemma 7, $U_{\mathcal{I}^*,\epsilon}$ locally has a product structure, identified with an open subset $\prod_{j=1}^\ell U_{\mathcal{I}_j} \subset \prod_{j=1}^\ell \mathcal{E}_{|\mathcal{I}_j|}$. The total space fibers over $U_{\mathcal{I}^*,\epsilon}$, and is given locally by a product of fibrations. In this case, the conformal factor can be written as a sum

$$G = \sum_{j=1}^\ell \left(\sum_{i \in \mathcal{I}_j} (\beta_i - 1) \log|z - z_i| \right).$$

Since $\{z_i\}_{i \in \mathcal{I}_j}$ is bounded away from any other clusters $\mathfrak{p}_{\mathcal{I}_{j'}}$, for $j' \neq j$, the term $\sum_{i \in \mathcal{I}_j} (\beta_i - 1) \log|z - z_i|$ is only singular near V_i as defined in Lemma 7. By induction,

this term is polyhomogeneous on $\mathcal{C}_{|\mathcal{I}_j|}$, hence is polyhomogeneous on $U_{\mathcal{I}^*, \epsilon}$. The same argument can be applied to other terms. By considering all the open covers, we get polyhomogeneity of G on $\mathcal{C}_k$ in this case.

Now we consider the behavior near the central face, and all the k points can merge together. We now prove that away from all subdiagonals, G is polyhomogeneous near $\mathfrak{C}_{1\dots k}^0$. In this region we write $p_i = z_i + \zeta$ and assume that the center of mass $\zeta = 0$. Then, writing $(z, z_1, \dots, z_k) = R_{1\dots k}\Omega$ and $\Omega = (\Omega_0, \dots, \Omega_k)$, we have

$$G = \sum_{i=1}^k (\beta_i - 1) \log|z - z_i| = \left(\sum_{i=1}^k \beta_i - k \right) \log R_{1\dots k} + \sum_{i=1}^k (\beta_i - 1) \log|\Omega_0 - \Omega_i|$$

and, since z remains bounded away from the subdiagonals, only the term $\log R_{1\dots k}$ is singular here and this is obviously polyhomogeneous on the interior of $\mathfrak{C}_{1\dots k}$. And each term $(\beta_i - 1) \log|\Omega_0 - \Omega_i|$ is singular only near $\mathfrak{C}_i$ and is polyhomogeneous.

Near the outer boundary of $\mathfrak{C}_{1\dots k}$, set $z = re^{i\phi}$ and $w_i = z_i/r$. Then

$$G = \left(\sum_{i=1}^k \beta_i - k \right) \log r + \sum_{i=1}^k (\beta_i - 1) \log|e^{i\phi} - w_i|.$$

Notice that all the faces $\mathfrak{C}_i$ occur along the submanifolds $\{z = z_i\} \subset \{|w_i| = 1\}$, so, provided we stay away from these submanifolds, only the first term $(\sum_{i=1}^k \beta_i - k) \log r$ is singular, and it is polyhomogeneous. At the principal diagonal $\mathfrak{C}_i$, however, $w_i = e^{i\theta_i}$ and the additional singular term is $\log|e^{i\phi} - e^{i\theta_i}|$, which is polyhomogeneous there.

Finally, if $\mathfrak{p}$ is near any one of the partial diagonals, including near their intersection with $\mathfrak{C}_{1\dots k}$, then it is in a neighborhood of some intersection of front faces $\{\mathfrak{C}_{\mathcal{I}_j}\}_{j=1}^\ell$, where each $\mathcal{I}_j$ is a proper subset of $\{1, \dots, k\}$, and the $\mathcal{I}_j$ have no elements in common. The resolution ensures that the faces $\mathfrak{C}_{\mathcal{I}_i}$ and $\mathfrak{C}_{\mathcal{I}_j}$ are disjoint, so we can once again factor out the defining function $R_{1\dots k}$ and separate out the indices i which do not lie in any of the $\mathcal{I}_j$, and write

$$(46) \quad G = \left(\sum_{i=1}^k \beta_i - k \right) \log R_{1\dots k} + \sum_{i \notin \bigcup_j \mathcal{I}_j} (\beta_i - 1) \log|w - w_i| + \sum_{j=1}^\ell f_j,$$

where $w = z/R_{1\dots k}$ and $w_i = z_i/R_{1\dots k}$. Here f_j is the rescaled factor

$$f_j = \sum_{i \in \mathcal{I}_j} (\beta_i - 1) \log|w - w_i| = \sum_{i \in \mathcal{I}_j} (\beta_i - 1) \log R_{\mathcal{I}_j} + \sum_{i \in \mathcal{I}_j} (\beta_i - 1) \log|\Omega_0^{\mathcal{I}_j} - \Omega_i^{\mathcal{I}_j}|,$$

where $R_{\mathcal{I}_j}$ is the boundary-defining function for $\mathfrak{C}_{\mathcal{I}_j}$ and the coordinates over this face are given by $(w, w_i)_{i \in \mathcal{I}_j} = R_{\mathcal{I}_j}(\Omega_0^{\mathcal{I}_j}, \Omega_i^{\mathcal{I}_j})$. By induction, each rescaled factor f_j is up to a smooth summand the Green function near $\mathcal{I}_j$ and hence is polyhomogeneous near the collection of faces which constitute the resolution near this cluster. This behavior is uniform as $R_{1\dots k} \rightarrow 0$.

It is perhaps wise to illustrate this induction for $\mathcal{C}_3$. In this case, near $\mathfrak{C}_{12} \cap \mathfrak{C}_{123}$ we can write

$$G = (\beta_1 - 1) \log|z - \epsilon_1(1 + \epsilon_2)| + (\beta_2 - 1) \log|z - \epsilon_1(1 - \epsilon_2)| + (\beta_3 - 1) \log|z + \epsilon_1|.$$

We wish to examine the region $R_{123} \rightarrow 0$ and $R_{12} \rightarrow 0$, and in this region, $\epsilon_1 \sim R_{123}$ and $\epsilon_2 \sim R_{12}$. Therefore, using $w_{12} = (w_3 - 1)/\epsilon_2$ as a coordinate on $\mathfrak{C}_{12}$,

(47)

$$G = \left(\sum_{i=1}^3 \beta_i - 3\right) \log \epsilon_1 + \left(\sum_{i=1}^2 \beta_i - 2\right) \log \epsilon_2 + (\beta_1 - 1) \log|w_{12} - 1| \\ + (\beta_2 - 1) \log|w_{12} + 1| + (\beta_3 - 1) \log|w_3 + 1|.$$

Since $w_3 \sim 1$ here, this is polyhomogeneous. □

6 Hyperbolic conic metrics

We next turn to the analytic description of the space of hyperbolic cone metrics which, as explained earlier, exist whenever

$$\chi(M) + \sum_{i=1}^k (\beta_i - 1) < 0.$$

The problem is now genuinely nonlinear and the proof of polyhomogeneity correspondingly more difficult. Indeed, the proof is directly inductive on the number of cone points. We now explain the strategy, which requires several steps.

For the case $k = 2$, we construct a family of background metrics which is hyperbolic away from the merging points and flat near these points, with a transitional region in between. Let ρ be the degeneration parameter which measures the distance to the fiber where the points coincide. We then solve for the expansion of the conformal factor iteratively on $M_{p'}$ then on $\mathfrak{C}_{12}$. This way we construct approximate solutions to arbitrarily high order of ρ . We then solve away the error in the exact curvature equation on each fiber using the maximum principle. Finally, we use a commutator argument to show conormality and polyhomogeneity. Once the theorem has been established for

$k = 2$, we follow an inductive procedure to construct families of background metrics in the general case with the same properties, and once again solve away the error terms and show polyhomogeneity.

The case $k = 2$ already contains essentially all of the substantial difficulties, so this case is presented in careful detail.

6.1 The case of two merging cone points

Consider a family of simple divisors $\mathbf{p}$ which converge to a point $\mathbf{q} \in F_{12} \subset \mathcal{E}_k$. We may as well assume that $p_3, \dots, p_k$ remain fixed, but p_1 and p_2 merge at a point p_{12} which, for simplicity, we assume is the center of mass of p_1 and p_2 and also remains fixed. We write $\mathbf{p}'$ for the (simple) $(k-1)$ -tuple $(p_{12}, p_3, \dots, p_k)$. We are working locally near $\mathbf{q}$, and this point is far from any of the other partial diagonals, so we use the local coordinates on $\mathcal{E}_2$ and $\mathcal{C}_2$, $\rho = \rho_{12}$ and $R = R_{12}$, and for simplicity we set $\theta_{12} = 0$, which amounts to fixing the direction through which p_1 and p_2 approach one another. If β_i are the cone parameters at p_i , then, as noted earlier, β_1 and β_2 determine the limiting cone parameter $\beta_{12} = \beta_1 + \beta_2 - 1$ at p_{12} . In order for the two points to merge, it is necessary that

$$\beta_1 + \beta_2 > 1 \iff \beta_{12} > 0.$$

The fiber $\pi^{-1}(\mathbf{q}) \subset \mathcal{C}_k$ consists of two surfaces with boundary, $M_{\mathbf{p}'} = [M; \{\mathbf{p}'\}]$ (the surface M blown up at the points in $\mathbf{p}'$) and the face $\mathcal{C}_{12}$, and these meet along a common circle.

The initial metric We now construct a family of metrics on M with k conic singularities at the family of divisors $\mathbf{p}$ above, which extends as a smooth family of fiberwise metrics on $\mathcal{C}_k$. This family is obtained locally near the fiber $\pi^{-1}(\mathbf{q})$ by gluing the fixed hyperbolic metric $h_{0,\mathbf{p}'}$ with conic singularities at $\mathbf{p}'$, with cone parameters $\beta_{12}, \beta_3, \dots, \beta_k$, to the degenerating family of flat metrics $g_{0,\mathbf{p}}^{\text{fl}}$ in (41). To do this, define

$$(48) \quad g_{0,\mathbf{p}} = \chi g_{0,\mathbf{p}}^{\text{fl}} + (1 - \chi)h_{0,\mathbf{p}'},$$

where

$$\chi(z, \mathbf{p}) = \begin{cases} 1 & \text{if } \rho < \bar{\rho} \text{ and } |z| < 2\bar{\rho}, \\ 0 & \text{if } \rho > 2\bar{\rho} \text{ or } |z| > 4\bar{\rho}, \end{cases}$$

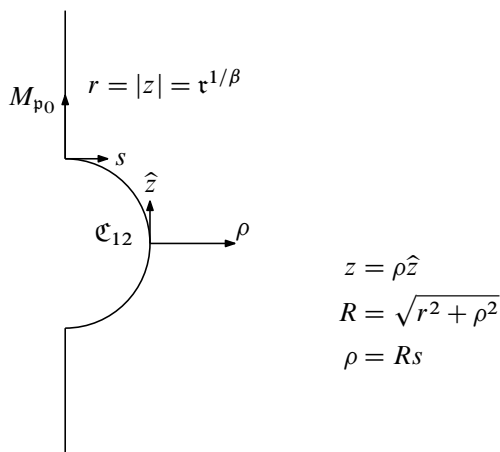


Figure 5: Coordinates used in the computation of expansion near the singular fiber $\mathfrak{C}_{12} \cup M_{p'}$.

is a smooth nonnegative cutoff function for some small $\bar{\rho} > 0$. We usually drop p and p' from the subscripts for simplicity, and also write

$$K_{0,\rho} = \begin{cases} 0 & \text{if } \rho < \bar{\rho} \text{ and } |z| < 2\bar{\rho}, \\ -1 & \text{if } \rho > 2\bar{\rho} \text{ or } |z| > 4\bar{\rho}, \end{cases}$$

for the curvature of the metrics in this family.

Our goal is to obtain precise analytic control of the solution to the conformal curvature equation

$$(49) \quad \Delta_{g_{0,p}} u + e^{2u} + K_{0,\rho} = 0$$

as $\rho \rightarrow 0$. In the neighborhood where $\chi = 1$, (49) becomes

$$(50) \quad \Delta_{g_{0,\rho}} u + e^{2u} = 0.$$

Here $\Delta_{g_{0,p}}$ is the Laplace–Beltrami operator with nonnegative spectrum.

Our goal is to prove that the solution u to (49) is polyhomogeneous on $\mathcal{E}_k$ near the point $p' \in \mathcal{C}_k$ with $z = 0$. To do this we construct the solution anew (even though its existence is guaranteed by standard barrier arguments) by first constructing an approximate solution which satisfies (49) to any fixed arbitrarily high order as $\rho \rightarrow 0$, and then correcting this to the exact solution using an analytic construction which guarantees that this additional correction term also vanishes to that same high order.

The specifics of the first part of this are that we construct the entire Taylor series for u along each of the two faces $M_{p'}$ and $\mathfrak{C}_{12}$. These series expansions are related to one another and must satisfy a set of matching conditions along the corner where these faces intersect. The fact that we can correct any finite part of this Taylor series to an exact solution with a term which vanishes to that order means that these series represent the true expansions for this exact solution.

Expansion at $M_{p'} \cap \mathfrak{C}_{12}$ within $M_{p'}$ Recall the coordinates $z = re^{i\phi}$ and ρ near $\mathfrak{C}_{12}$ in $\mathcal{E}_k$ (as before, an angular coordinate is suppressed), and set

$$r = |z| \quad \text{and} \quad s = \rho/r.$$

Thus $s = 0$ defines the surface $M_{p'}$ while $r = 0$ defines $\mathfrak{C}_{12}$. There is a freedom in the conformal coordinate z on $M_{p'}$ by holomorphic reparametrization, and we fix this below in Lemma 18. We also define the coordinate

$$\mathfrak{r} = \frac{1}{\beta_{12}} |z|^{\beta_{12}}$$

on $M_{p'}$, which is the radial distance function for the background flat conic metric. It follows from (45) that, near the corner $\mathfrak{r} = s = 0$,

$$g_{0,p} = \alpha(s, \mathfrak{r}, \phi, \theta) (d\mathfrak{r}^2 + \beta_{12}^2 \mathfrak{r}^2 d\phi^2),$$

where $\alpha(s, \mathfrak{r})$ is polyhomogeneous with $\alpha(0, \mathfrak{r}) = 1$ near $\mathfrak{r} = 0$. Our first result is that by choosing the coordinate z carefully, the expansion for α has a particularly simple form; it simultaneously also gives the first term in the expansion of u at $M_{p'}$.

Lemma 18 *There is a unique bounded solution $u'_0 \in \bigcap_{m \geq 0} C_b^{m,\delta}(M_{p'})$ to the restriction of (49) to $M_{p'}$. This solution is polyhomogeneous as $\mathfrak{r} \rightarrow 0$, and if this defining function is chosen appropriately, then*

$$u'_0 \sim \sum_{j \in \mathbb{N}_0} a_j \mathfrak{r}^{2j}.$$

Proof All of this except the last assertion, ie the existence and polyhomogeneous regularity, is contained in [17; 16]. For simplicity set $\beta = \beta_{12}$. Existence and uniqueness of the solution is proved in [17] by constructing bounded sub- and supersolutions, and this method also leads to the uniqueness of u'_0 amongst bounded solutions. The regularity theorem is proved in [16]. First, (scale-invariant) local elliptic regularity shows that $u'_0 \in C_b^{m,\delta}$ for every $m \geq 0$. The refined regularity theorem in Section 3

of that paper states that u'_0 is polyhomogeneous, with $u'_0 \sim \sum_{\ell,j \geq 0} a_{\ell j}(\phi) \tau^{\ell/\beta + j}$ as $\tau \rightarrow 0$, where $a_{\ell j}$ is linear in $\cos \ell \phi$ and $\sin \ell \phi$.

It remains to prove the last assertion: that all $a_{\ell j} = 0$ when $\ell \neq 0$ in some choice of coordinates. To this end, recall that near a cone point, there are geodesic coordinates $(\tilde{r}, \theta)$ in terms of which the hyperbolic metric takes the canonical polar form

(51)
$$g = d\tilde{r}^2 + \beta^2 \sinh^2 \tilde{r} d\phi^2.$$

On the other hand, there is a local holomorphic coordinate z centered at the conical point for which, in the region near p_{12} in which it is flat, $g_{0,p} = |z|^{2(\beta-1)} |dz|^2$, so that

$$g = e^{2u'_0} |z|^{2(\beta-1)} |dz|^2$$

there. Since $\frac{1}{\beta} |z|^\beta = \tau$, we have

$$d\tilde{r}^2 + \beta^2 \sinh^2 \tilde{r} d\phi^2 = e^{2u'_0(z)} |z|^{2(\beta-1)} |dz|^2 = e^{2u'_0(z)} (d\tau^2 + \beta^2 \tau^2 d\phi^2).$$

This gives

$$\frac{d\tilde{r}}{d\tau} = e^{u'_0}, \quad \sinh \tilde{r} = e^{u'_0} \tau,$$

or, equivalently,

$$\frac{d\tilde{r}}{\sinh \tilde{r}} = \frac{d\tau}{\tau},$$

and hence

$$\tanh \frac{1}{2} \tilde{r} = c \tau.$$

We finally scale z so that $c = \frac{1}{2}$.

We have now shown that

$$\tilde{r} \sim \tau \left(1 + \sum_{j=1}^\infty \tilde{a}_j \tau^{2j} \right),$$

ie $\tilde{r}$ is an odd function of τ , so that $e^{u'_0} = \tau^{-1} \sinh \tilde{r}$ is even in τ and equals 1 when $\tau = 0$. We conclude that u'_0 is even and vanishes at $\tau = 0$, hence

(52)
$$u'_0 \sim \sum_{j=1}^\infty \tilde{a}_{0j} \tau^{2j} = \sum_{j \geq 1} a_{0j} \tau^{2j\beta},$$

where each a_{0j} is constant. □

Expansion at $M_{p'}$ We next turn to the complete expansion of u at the face $M_{p'}$. The defining function for this face is s , and we know a priori that u is smooth in s at $s = 0$, hence

$$(53) \quad u \sim \sum_{j=0}^{\infty} \tilde{u}'_j(r, \phi) s^j.$$

Our goal is to compute these coefficients $\tilde{u}'_j$, and, more specifically, to understand their expansions as $r \rightarrow 0$. In doing this, it is more convenient to write (53) as an expansion in ρ since the Laplacian on the fibers commutes with ρ . Recall that near the corner $M_{p'} \cap \mathfrak{C}_{12}$, we have $\rho = R \sin \omega$, so if we set $s = \sin \omega$, then $s = \rho/R$. Furthermore, along $s = 0$, we can take $R = r$. Therefore,

$$u \sim \sum_{j=0}^{\infty} \rho^j u'_j, \quad \text{where } u'_j = r^{-j} \tilde{u}'_j.$$

For simplicity of notation, we assume here $2k\beta \notin \mathbb{N}$ for any $k \in \mathbb{N}$. The other case has no essential difference except the notation. In particular, the expansions below are the same, and for some values $\ell \in \mathbb{N}$, the coefficients of r^ℓ appear in more than one place because of the coincidence $\ell' + 2k\beta = \ell'$.

Proposition 19 *As $s \rightarrow 0$, there is an expansion*

$$(54) \quad u \sim \tilde{u}'_0 + \sum_{j=1}^{\infty} s^j \tilde{u}'_j,$$

where, in terms of the functions u'_j for $j \geq 1$,

$$(55) \quad u'_j \sim \sum_{\ell \in \mathbb{N}} r^\ell a_{j\ell 0}(\phi) + \sum_{\ell, k \in \mathbb{N}, \ell \geq 0, k \geq 1} r^{\ell+2k\beta} a_{j\ell k}(\phi),$$

or, equivalently,

$$(56) \quad \tilde{u}'_j \sim \sum_{\ell \in \mathbb{N}} r^{j+\ell} a_{j\ell 0}(\phi) + \sum_{\ell, k \in \mathbb{N}, \ell \geq 0, k \geq 1} r^{j+\ell+2k\beta} a_{j\ell k}(\phi)$$

Here $a_{j\ell 0}$ are trigonometric polynomials of pure degree ℓ , ie linear combinations of $\cos(\ell\phi)$ and $\sin(\ell\phi)$, while $a_{j\ell k}$ for $\ell \geq 0$ and $k \geq 1$ are trigonometric polynomials of degree at most ℓ . In particular,

$$(57) \quad \tilde{u}'_1 \sim \sum_{\ell \in \mathbb{N}} r^{1+\ell} a_{1\ell 0}(\phi) + \sum_{\ell, k \in \mathbb{N}, \ell \geq 0, k \geq 1} r^{1+\ell+2k\beta} a_{1\ell k}(\phi).$$

And $a_{1\ell k}$ is a linear combination of $\cos(\ell\phi)$ and $\sin(\ell\phi)$ for any $\ell \geq 0$ and $k \geq 0$.

Proof Clearly u is C^∞ up to $M_{p'}$ away from $z = 0$, and thus has an expansion in nonnegative integer powers of ρ . Expand $K_{g_{0,p}} = \sum_{j=0}^{\infty} \rho^j K_j$ in (49) and insert a formal series expansion for u , as in the statement of this theorem. Notice that $K_j \equiv 0$ near $\mathfrak{C}_{12}$. Since $\Delta_{g_{0,p}}$ commutes with ρ away from $\mathfrak{C}_{12}$, we obtain a recursive set of equations which successively determine all of the u'_j . The first of these is the curvature equation

$$\Delta_{g_{0,p}} u'_0 + e^{2u'_0} + K_0 = 0$$

on $M_{p'}$. By the previous lemma, u'_0 has an expansion involving only the powers $r^{2k\beta}$ for $k \in \mathbb{N}$.

The equation for u'_j for $j \geq 1$ is

$$(58) \quad e^{-2u'_0} \Delta_{g_{0,p}} u'_j + 2u'_j = -e^{-2u'_0} (K_j + (e^{2u^{(j-1)}} - 1 - 2u^{(j-1)})_j),$$

where $u^{(j-1)} = \sum_{i=0}^{j-1} \rho^i u'_i$ and the notation $(w)_j$ means that we take the coefficient of ρ^j in the expansion of w . This is the Laplacian with positive spectrum, so this equation always admits a solution on $M_{p'}$. Using the special coordinates (r, ϕ) from Lemma 18 near $r = 0$, this equation reduces in a neighborhood of $r = 0$ to

$$(59) \quad (\Delta_{2\beta} + 2e^{2u'_0})u'_j = -(e^{2u^{(j-1)}} - 1 - 2u^{(j-1)})_j,$$

where $\Delta_{2\beta}$ is the Laplacian for the conic metric $|z|^{2(\beta-1)}|dz|^2$ where $z = re^{i\phi}$ (which gives the local form of $g_{0,p}$). After multiplying by $r^{2\beta}$, the equation above can be rewritten as

$$(60) \quad ((r\partial_r)^2 + \partial_\phi^2)u'_j + 2r^{2\beta}e^{2u'_0}u'_j = -r^{2\beta}(e^{2u^{(j-1)}} - 1 - 2u^{(j-1)})_j.$$

The bounded formal solutions (ie solutions to leading order) of

$$((r\partial_r)^2 + \partial_\phi^2)u'_j + 2r^{2\beta}e^{2u'_0}u'_j = 0$$

are $r^\ell q_\ell(\phi)$ for $\ell \in \mathbb{N}$, where q is a trigonometric polynomial of pure degree ℓ . The nonnegative indicial roots are $\ell = 0, 1, 2, \dots$.

For $j = 1$, (59) becomes

$$(61) \quad ((r\partial_r)^2 + \partial_\phi^2)u'_1 + 2r^{2\beta}e^{2u'_0}u'_1 = -r^{2\beta}K_1;$$

note here $(e^{2u^{(0)}} - 1 - 2u^{(0)})_1 = (e^{2u'_0} - 1 - 2u'_0)_1 = 0$. Noting that K_1 vanishes near $\mathfrak{C}_{12}$ and $e^{2u'_0} \sim 1 + \sum_{k \geq 1} a_k r^{2k\beta}$, the solution of this equation has an expansion

of the form

$$(62) \quad u'_1 \sim \sum_{\ell \in \mathbb{N}} r^\ell a_{1\ell 0}(\phi) + \sum_{\ell, k \in \mathbb{N}, \ell \geq 0, k \geq 1} r^{\ell+2k\beta} a_{1\ell k}(\phi).$$

Here the terms $a_{1\ell 0}$ come from indicial roots, and are formally undetermined near $r = 0$, but of course are fixed because u'_1 solves a global equation on $M_{p'}$. And each $a_{1\ell 0}$ is a linear combination of $\cos(\ell\phi)$ and $\sin(\ell\phi)$. All the other terms arise by matching coefficients on the two sides of this equation. In particular, because of the multiplication by $r^{2\beta}$, the leading term in the second sum is given by $r^{2\beta}$ and there are no log terms. And each $a_{1\ell k}(\phi)$ is of pure degree ℓ .

We now prove by induction that the expansion of u'_j for $j > 1$ is as in (55). The guiding principle in all of this is that the right-hand side of (60) does not contain any indicial term of the linear operator. In particular, the right-hand side has an expansion where terms are given by $r^{2\beta+\ell+2k\beta} a_{\ell k}$ for $\ell \geq 0$ and $k \geq 0$ and its coefficient $a_{\ell k}$ is a trigonometric polynomial of degree at most ℓ .

The equation for u'_2 is

$$(63) \quad ((r\partial_r)^2 + \partial_\phi^2)u'_2 + 2r^{2\beta} e^{2u'_0} u'_2 = -r^{2\beta} (K_2 + 2(u'_1)^2),$$

and this right-hand side has an expansion

$$(64) \quad r^{2\beta} \left(\sum_{\ell \geq 0} \sum_{\ell'=0}^{\ell} (c_{\ell\ell'} \cos(\ell'\phi) + d_{\ell\ell'} \sin(\ell'\phi)) r^\ell + \sum_{\ell \geq 0, k \geq 1} \sum_{\ell'=0}^{\ell} (c_{\ell k \ell'} \cos(\ell'\phi) + d_{\ell k \ell'} \sin(\ell'\phi)) r^{\ell+2k\beta} \right).$$

Indeed, the expansion of $r^{2\beta} (u'_1)^2$ contains terms

$$(65) \quad r^{\ell+2\beta} q_\ell(\phi) \quad \text{and} \quad r^{\ell+2(k+1)\beta} p_{\ell k}(\phi) \quad \text{for } k \geq 0.$$

The coefficients q_ℓ and $p_{\ell k}$ are finite sums of finite products of trigonometric polynomials $\prod q_j$, with $\deg q_j = \ell_j$ and $\sum \ell_j = \ell$. Even if each q_j is pure, this product usually includes all lower degrees as well. Therefore all q_ℓ and $p_{\ell k}$ are trigonometric polynomials of degree at most ℓ . Since K_2 vanishes identically near $r = 0$, we can see that the right-hand side does not contain any indicial terms $r^\ell e^{i\ell\phi}$ because of the $r^{2\beta}$ shift, which implies that the solution u'_2 does not contain any $\log r$ terms. Solving

(63) term by term gives

(66)
$$u_2' \sim \sum_{\ell \in \mathbb{N}} r^\ell a_{2\ell 0}(\phi) + \sum_{\ell, k \in \mathbb{N}, \ell \geq 0, k \geq 1} r^{\ell+2k\beta} a_{2\ell k}(\phi).$$

Here the first term contains indicial roots where $a_{2\ell 0}(\phi)$ is a linear combination of $\cos(\ell\phi)$ and $\sin(\ell\phi)$, while the second term comes from matching coefficients on two sides and $a_{2\ell k}$ is of degree at most ℓ .

Now suppose (55) is true for u_i' for $i < j$. Then the terms on the right-hand side of (58) are linear combinations of terms $\prod_{\sum i=j} u_i'$. By tracking the terms in (55), we obtain that the right-hand side is a linear combination of terms in (65). Applying the same guiding principle, we see that u_j' has an expansion as in (55). By induction, this concludes the proof of the proposition. □

Expansion at $\mathfrak{C}_{12}$ We next consider the expansion at $\mathfrak{C}_{12}$. Unlike the preceding construction at $M_{p'}$, some terms in this expansion can only be determined once we take into account their compatibility with the previous expansion. Write

(67)
$$u \sim \sum_{\alpha \in \mathcal{E}} R^\alpha \tilde{u}''_\alpha(s, \phi)$$

near this face, where $\mathcal{E}$ is an index set which is determined in the course of the argument below; see (75). The double prime indicates that the terms are coefficients in the expansion near $\mathfrak{C}_{12}$. As usual, $R = \sqrt{\rho^2 + r^2}$ and $R \sin \omega = \rho$, and, as before, we set $s = \sin \omega$. (This is a good coordinate away from the pole of this hemisphere.) For many purposes it is simpler to use the projective coordinates $\hat{z} = z/\rho$ and ρ , which are valid on the interior of $\mathfrak{C}_{12}$; $|\hat{z}| \rightarrow \infty$ at the outer boundary of this face, so $s \sim 1/|\hat{z}|$, and ρ is only a defining function for this face away from its outer boundary. See Figure 5 for an illustration of the coordinates.

In these projective coordinates, still writing $\beta = \beta_{12}$,

(68)
$$g_{0,p} = \rho^{2\beta} e^{2\tilde{u}} |d\hat{z}|^2, \quad \hat{u} = (\beta_1 - 1) \log|\hat{z} - 1| + (\beta_2 - 1) \log|\hat{z} + 1| + \tilde{u},$$

where $\tilde{u}$ is harmonic as a function of $\hat{z}$, so in particular is smooth across the singular points $\hat{z} = \pm 1$ in this face. It is also bounded in a neighborhood of $\mathfrak{C}_{12}$, so in fact its restriction to $\mathfrak{C}_{12}$ must be constant. In other words, $\rho^{-2\beta} g_{0,p}$ restricts to a flat metric $\hat{g}$ on the interior of $\mathfrak{C}_{12}$ which has two conic singularities at $w = \pm 1$ and is asymptotic to the large end of a cone with cone angle $2\pi\beta$ as $\hat{z} \rightarrow \infty$.

Now write $R^\alpha \tilde{u}''_\alpha = \rho^\alpha s^{-\alpha} \tilde{u}''_\alpha = \rho^\alpha u''_\alpha$ where $u''_\alpha(s, \phi) = s^{-\alpha} \tilde{u}''_\alpha$. Using that the fiber Laplacian commutes with fiber variable ρ , ie $[\Delta_{g_{0,p}}, \rho] = 0$, the curvature equation thus leads to equations for each of these coefficients,

$$(69) \quad \hat{\Delta} u''_\alpha = -2u''_{\alpha-2\beta} - (e^{2u^{(\alpha-2\beta)}} - 1 - 2u^{(\alpha-2\beta)})_{\alpha-2\beta},$$

where, here and below, we denote by $\hat{\Delta}$ the Laplacian for the flat conic metric $\hat{g} = \rho^{-2\beta} g_{0,p}$. Also, analogous to our previous notation, $u^{(\alpha-2\beta)}$ denotes the sum of all terms $\rho^{\alpha'} u''_{\alpha'}$ with $\alpha' < \alpha - 2\beta$, and $(\cdot)_{\alpha-2\beta}$ indicates the coefficient of $\rho^{\alpha-2\beta}$ in the expansion in parentheses. The downward shift by 2β occurs because of the factor $\rho^{-2\beta}$ on $\Delta_{g_{0,p}}$.

We now analyze these coefficients. The shift of exponents here motivates the fact that we carry this out for α in the succession of ranges

$$2\ell\beta < \alpha < 2(\ell+1)\beta, \quad \ell = 0, 1, 2, \dots;$$

the endpoints $2\ell\beta$ are handled separately.

The case $\alpha \in (0, 2\beta)$

Lemma 20 *The only terms in (67) with $\alpha \in (0, 2\beta)$ are those for which $\alpha \in \{1, 2, \dots, [2\beta]\}$. And the term u''_α is determined by $\{a_{(\alpha-j)j0}(\phi) : 0 \leq j \leq \alpha-1\}$ in (55).*

Proof Write the fiberwise Laplacian $\Delta_{g_{0,p}}$ locally as $\rho^{-2\beta} \hat{\Delta}$, where $\hat{\Delta}$ is the Laplacian for the flat conical metric $\hat{g} = e^{2\hat{u}} |d\hat{z}|^2$. Inserting (67) into the equation, and recalling that $\hat{\Delta}$ commutes with ρ , we obtain that $\hat{\Delta} u''_\alpha = 0$ for $\alpha < 2\beta$, so u''_α is harmonic with respect to $e^{2\hat{u}} |d\hat{z}|^2$, and hence also with respect to $|d\hat{z}|^2$.

The fact that the original coefficient $\tilde{u}''_\alpha$ is bounded as $s \rightarrow 0$, ie $\hat{z} \rightarrow \infty$, means that $u''_\alpha = \tilde{u}''_\alpha s^{-\alpha}$ grows at most like $s^{-\alpha} \sim |\hat{z}|^\alpha$. This means that it is a harmonic polynomial $p_\alpha(\hat{z})$ of degree less than or equal to α . Now, if α were not an integer, then $s^\alpha p_\alpha(\hat{z})$ would be a sum of terms, each vanishing at a nonintegral rate as $s \rightarrow 0$. This is impossible since u is smooth in s at $s = 0$. Hence the only allowable exponents $\alpha < 2\beta$ are nonnegative integers. When $\alpha = 0$, the only possibility for u''_α would be a degree 0 harmonic polynomial, hence a constant, which leads to the constant term $c\rho^0 r^0$ in the expansion, and by the choice of u'_0 we know that this constant must vanish. Therefore we can assume $1 \leq \alpha < 2\beta$.

Write $u''_\alpha = p_\alpha = \sum_{j=0}^\alpha p_{\alpha j}$ where each $p_{\alpha j}(\hat{z})$ is a harmonic polynomial of degree j . The term $p_{\alpha j}$ corresponds to a term of growth $\rho^\alpha s^{-j} a_j$ where a_j is a linear combination of $\cos \ell \phi$ and $\sin \ell \phi$. Compatibility at the corner means that this must match the coefficient of $s^{-j+\alpha} r^\alpha$, that is, $a_{(\alpha-j)j0}(\phi)$ in (55). And as we observed there, these coefficients indeed have pure degree j . In particular, when $j = \alpha$, this corresponds to a term $\rho^0 r^\alpha$ which vanishes from the expansion of u'_0 . Therefore $0 \leq j \leq \alpha - 1$, and there is a unique homogeneous harmonic polynomial $p_{\alpha j}$ which satisfies this boundary condition. This determines u_α for any integer $\alpha < 2\beta$. $\square$

The case $\alpha = 2\beta$

Lemma 21 When $\alpha = 2\beta$, u''_α is determined by a_{01} in (52).

Proof Essentially the same calculation as above yields that

$$\hat{\Delta} u''_{2\beta} + 1 = 0, \quad u''_{2\beta} = \tilde{u}''_{2\beta} s^{-2\beta}.$$

Here

$$\hat{\Delta} = e^{-2U} \Delta_{\hat{z}}, \quad e^{2U} = e^{(2((\beta_1-1)\log|\hat{z}-1| + (\beta_2-1)\log|\hat{z}+1|))},$$

where $\Delta_{\hat{z}}$ is the Laplacian for $|dz|^2$. Noting that $e^{2U} \sim c|\hat{z}|^{2\beta-2}$ as $|\hat{z}| \rightarrow \infty$ for some constant $c \neq 0$, it follows from well-known existence theory for the Laplacian on asymptotically conic manifolds that there exists a solution $u''_{2\beta}$ to this equation which asymptotic to $A|\hat{z}|^{2\beta}$ for some constant A . This solution is unique up to harmonic polynomials of degree strictly less than 2β . However, as before, any such harmonic polynomial would lead to a term in the expansion of $\tilde{u}''_{2\beta} = u''_{2\beta} s^{2\beta}$ which is not smooth at $s = 0$, and this is impossible. In addition, $R^{2\beta} u''_{2\beta} s^{2\beta} \rightarrow AR^{2\beta}$ as $s \rightarrow 0$, so A must equal the constant a_{01} in (52). This determines $u''_{2\beta}$ uniquely. $\square$

We have now explained all coefficients for the initial part of our index set,

$$\mathcal{E}_{2\beta} := \{\alpha \in \mathcal{E} : \alpha \leq 2\beta\} = \{1, 2, \dots, [2\beta], 2\beta\}.$$

The case $2\beta < \alpha < 4\beta$ Now we consider the cases for $\alpha \in (2\beta, 4\beta)$.

Lemma 22 When $\alpha \in (2\beta, 4\beta)$, the index set $\mathcal{E} \cap (2\beta, 4\beta)$ is given by

$$(70) \quad \{2\beta < \alpha < 4\beta : \alpha - 2\beta \in \mathbb{N} \text{ or } \alpha \in \mathbb{N}\}.$$

When $\alpha = \ell + 2\beta$ for some $\ell \in \mathbb{N}$, u''_α is determined by $\{a_{(\ell-j)j1} : 0 \leq j \leq \ell - 1\}$ in (57). When $\alpha \in \mathbb{N}$, u''_α is determined by $\{a_{(\alpha-j)j0}(\phi) : 0 \leq j \leq \alpha - 1\}$.

Proof When $\alpha < 4\beta$, based on whether the right-hand side of (69) is trivial, there are two cases. The first is when $\alpha - 2\beta = \ell \in \{1, 2, \dots, [2\beta]\}$. In this case we get the inhomogeneous equation

$$(71) \quad \hat{\Delta} u''_{\alpha} = -2u''_{\ell} - (e^{2u^{(\ell)}} - 1 - 2u^{(\ell)})_{\ell}$$

where, as before, $(\cdot)_{\ell}$ is the coefficient of ρ^{ℓ} in the expansion of the expression in parentheses. The right-hand side is a linear combination of terms

$$(72) \quad u''_{j_1} \cdots u''_{j_k}$$

where the sum is over all partitions $(j_1, \dots, j_k)$ with $\sum_{i=1}^k j_i = \ell$. Recall that each u''_{j_i} is a sum of harmonic polynomials in $\hat{z}$ of degrees strictly less than j_i . When there is only one term in the sum of (72), ie $k = 1$, then $u''_{j_1} = u''_{\ell}$ is a sum of terms $p_{\ell r}$ for $\ell - 1 \geq r \geq 0$, where each $p_{\ell r}$ is a harmonic polynomial of pure degree r , and, in particular, near the “infinity” $s = 0$ the angular coefficient of $s^{-(\ell-1)}$ is a linear combination of $\cos((\ell-1)\phi)$ and $\sin((\ell-1)\phi)$. On the other hand, when there are at least factors in the summand, ie $k \geq 2$, then each of these products is a homogeneous polynomial of degree $j \leq \ell - 2$, and near infinity the angular coefficients is a trigonometric polynomial of degree no more than j . To combine these two situations, the right-hand side is given by

$$s^{-(\ell-1)} c_{\ell-1}(\phi) + \sum_{0 \leq i \leq \ell-2} s^{-i} d_i(\phi),$$

where $c_{\ell-1}$ is of pure degree $\ell - 1$ and d_i is mixed of degree at most i .

The solution u''_{α} is the sum of an inhomogeneous term $\sum_{i=0}^{\ell-1} q_i$ and potential homogeneous terms. Here each inhomogeneous term $q_i(\hat{z})$ solves away the s^{-i} term in the above expansion, hence $q_i \sim A_i(\phi) s^{-i-2\beta}$. For the top degree $i = \ell - 1$, $A_{\ell-1}$ is a linear combination of $\cos((\ell-1)\phi)$ and $\sin((\ell-1)\phi)$. And for the lower degrees $i < \ell - 1$, A_i is a trigonometric polynomial of degree at most i . Those coefficients are matched at the corner, since the term q_i would lead to $\rho^{\ell-i} r^{i+2\beta} A_i$, so A_i is given by $a_{(\ell-i)i1}$ in (55). In particular, when $i = \ell - 1$, $A_{\ell-1}$ is matched by $a_{1(\ell-1)1}$, which is indeed of pure degree $\ell - 1$, while for $i \leq \ell - 2$, A_i is given by $a_{(\ell-i)i1}$, which is a trigonometric polynomial of degree at most i from Proposition 19.

Regarding the potential homogeneous terms, u''_{α} is unique up to addition by harmonic polynomials of degree strictly less than $\alpha - 2\beta$. However this would give a term in

$u_\alpha = u''_\alpha s^\alpha$ which is not smooth. By the same reasoning as in the case $\alpha = 2\beta$, we have shown that u''_α is uniquely determined by coefficients listed above.

The other case for the equation (71) is given by the homogeneous equation $\hat{\Delta}u''_\alpha = 0$. By the same reasoning as before, α must be an integer, and $u''_\alpha = \sum p_{\alpha j}(\hat{z})$ where each $p_{\alpha j}$ is a harmonic polynomial of degree $j \leq \alpha - 1$. And for each j , the boundary asymptotic of the term $\rho^\alpha p_{\alpha j}(\hat{z})$ is given by $\rho^{\alpha-j} R^j A_{\alpha j}(\phi)$, which is a linear combination of $\sin j\phi$ and $\cos j\phi$, hence is matched by the coefficient $a_{(\alpha-j)j0}(\phi)$. $\square$

The case $\alpha = 4\beta$

Lemma 23 When $\alpha = 4\beta$, u''_α is determined by a_{02} in (52).

Proof When $\alpha = 4\beta$, the term $R^{4\beta}u_{4\beta}$ solves $\hat{\Delta}u''_{4\beta} = 2u''_{2\beta} \sim A|\hat{z}|^{2\beta}$. Using the same argument as for $u''_{2\beta}$, $u''_{4\beta}$ is unique and asymptotic to $B|\hat{z}|^{4\beta}$, where B is given by the constant a_{02} in (52). $\square$

Hence we have shown

$$(73) \quad \mathcal{E}_{4\beta} = \{2\beta < \alpha < 4\beta : \alpha - 2\beta \in \mathbb{N} \text{ or } \alpha \in \mathbb{N}\} \cup \{4\beta\}.$$

The case $2(n-1)\beta < \alpha \leq 2n\beta$ Iteratively we can repeat the argument for $\alpha \in (2(n-1)\beta, 2n\beta]$.

Lemma 24 The index set $\mathcal{E}_{2n\beta}$ is given by

$$\{\alpha \in (2(n-1)\beta, 2n\beta] : \alpha = j + 2k\beta\}.$$

For $\alpha = j + 2k\beta$ with $j, k \geq 1$, u''_α is determined by $\{a_{(j-\ell)\ell k} : 0 \leq \ell \leq j-1\}$ in (55). When $\alpha = 2n\beta$, u''_α is determined by a_{0n} in (52). When $\alpha \in \mathbb{N}$, u''_α is determined by $\{a_{(\alpha-j)j0} : 0 \leq j \leq \alpha-1\}$ in (55).

Proof As before there are two cases: $\alpha - 2\beta = \sum \alpha'$ for some $\alpha' \in \mathcal{E}$, which by induction means $\alpha = j + 2k\beta$ with $n \geq k \geq 1$ and $j \geq 0$, and $k = n$ if and only if $j = 0$; or $\alpha \in \mathbb{N}$. Note that the endpoint $\alpha = 2n\beta$ is included in the first case.

In the first case, assuming $j > 0$ hence $k \leq n-1$, then u''_α solves an inhomogeneous equation (71), where the right-hand side is a sum of terms

$$(74) \quad u''_{\alpha_1} \cdots u''_{\alpha_m}, \quad \sum_i \alpha_i = \alpha - 2\beta = j + 2(k-1)\beta.$$

By induction, each term u''_{α_i} where $\alpha_i = j_i + 2k_i\beta$ is a sum of terms $\sum_{\ell=0}^{j_i-1} A_\ell s^{-\ell-2k_i\beta}$ such that $A_{j_i-1}(\phi)$ is of pure degree $j_i - 1$ and each other A_ℓ is mixed of degree at most ℓ . Therefore, as discussed before, the terms in (74) are characterized in two categories:

- (a) $m = 1$, ie $\alpha_1 = j + 2(k - 1)\beta$, in which case the first term is given by $A_{j-1}(\phi)s^{-(j-1)-2(k-1)\beta}$, where A_{j-1} is of pure degree $j - 1$, while the rest of the terms would combine with (b);
- (b) $m \geq 2$, in which case the product contains terms $A_\ell s^{-\ell-2(k-1)\beta}$, where $\ell \leq j-2$, and each A_ℓ is mixed of degree at most ℓ .

For the same reason as before, u''_α has a unique solution which is asymptotically given by $\sum_{\ell=0}^{j-1} A_\ell s^{-\ell-2k\beta}$, and each term leads $A_\ell \rho^{j-\ell} R^{\ell+2k\beta}$, and each $A_\ell(\phi)$ is determined by coefficients $a_{(j-\ell)\ell k}$ in (55). In particular, only when $\ell = j - 1$, $a_{(j-\ell)\ell k}$ is of pure degree ℓ , while for $\ell \leq j - 2$, $a_{(j-\ell)\ell k}$ is mixed of degree ℓ . So it is matched.

On the other hand, if in the first case $j = 0$ and $k = n$, then $\alpha = 2n\beta$. Then u''_α satisfies the same equation (71) with the special requirement that all the α_i in (74) are of the form $2j\beta$. Then by induction the right-hand side is given by $A|\hat{z}|^{2(n-1)\beta}$, hence $u''_{2n\beta}$ is unique and asymptotic to $B|\hat{z}|^{2n\beta}$, which is matched by coefficient a_{0n} .

In the second case ($\alpha \in \mathbb{N}$), u''_α solves the homogeneous equation $\hat{\Delta}u''_\alpha = 0$ and is a combination of harmonic polynomials of degree $j < \alpha$, and by the same argument as before, each term is determined by $\{a_{(\alpha-j)j0} : 0 \leq j \leq \alpha - 1\}$. $\square$

With the discussion above, we have:

Proposition 25 *The index set $\mathcal{E}$ is given by*

$$(75) \quad \mathcal{E} = \{j + 2k\beta : j, k \in \mathbb{N}\}.$$

For each term $R^\alpha \tilde{u}''_\alpha$ with $\alpha = j + 2k\beta$, $\tilde{u}''_\alpha(s, \phi)$ is smooth up to $s = 0$ and asymptotically given by a sum of terms with growth $\{s^\ell : \ell \in \mathbb{N}, \ell \leq j\}$. That is, for the solution near the corner there is a product-type expansion

$$u \sim \sum_{\alpha=j+2k\beta} \sum_{\ell=0}^{j-1} R^\alpha s^\ell u_{\alpha\ell}(\phi).$$

The approximate solution

Lemma 26 *There exists a polyhomogeneous function*

$$\tilde{u} \sim \sum_{\alpha=j+2k\beta \in \mathcal{E}, \ell \leq j} R^\alpha s^\ell u_{\alpha\ell}(\phi) + \mathcal{O}(\rho^{N+\epsilon})$$

which satisfies

$$(76) \quad \Delta_{g_{0,p}} \tilde{u} + e^{2\tilde{u}} + K_{g_{0,p}} = \mathcal{O}(\rho^N)$$

for any $N \geq 0$ as $\rho = Rs \rightarrow 0$.

Proof We take $\tilde{u}$ to be a Borel sum of the formal polyhomogeneous series constructed above; (76) is then obvious. $\square$

We now write

$$(77) \quad \tilde{g}_0 = e^{2\tilde{u}} g_{0,p}.$$

Correction to an exact solution The final step is to correct the approximate solution to an exact one by solving

$$\Delta_{\tilde{g}_0} v + e^{2v} + K_{\tilde{g}_0} = 0$$

for each ρ , or, equivalently,

$$(78) \quad \Delta_{\tilde{g}_0} v + 2v = -(K_{\tilde{g}_0} + 1) - (e^{2v} - 1 - 2v).$$

Of course, this solution is already known to exist and be unique, but the method here will show that it is polyhomogeneous on $\mathcal{C}_k$. Indeed, we find a solution to (78) satisfying $|v| \leq C\rho^N$ for any fixed N ; these are all the same by uniqueness, of course. However, in this way we can estimate the b -derivatives of this solution up to that order, and hence, since N is arbitrary, to all orders.

For convenience, write $f = -(K_{\tilde{g}_0} + 1)$; we have arranged that f is smooth and vanishes to all orders at $\rho = 0$. We also set $Q(v) = -(e^{2v} - 1 - 2v)$.

Proposition 27 *For each $N > 0$ and $0 < \rho < \epsilon$, there is a unique bounded solution v to (78) such that $|v| \leq C_0 \rho^N$ for some constant C_0 .*

Proof By the maximum principle, if $(\Delta + 2)w = h$, then $\sup |w| \leq \frac{1}{2} \sup |h|$. For a given choice of N , there exists a constant C_0 such that $|f| \leq C_0 \rho^N$ for all sufficiently small ρ . There is also a constant C_1 such that $|Q(v)| \leq C_1 |v|^2$ when $|v| \leq 1$. Now

define the sequence v_j by $v_0 = 0$ and

$$(\Delta_{\tilde{g}_0} + 2)v_{j+1} = f + Q(v_j).$$

By the remarks above, $\sup |v_{j+1}| \leq \frac{1}{2}C_0\rho^N + \frac{1}{2}C_1 \sup |v_j|^2$. Assume inductively that $\sup |v_j| \leq A_j = C_0\rho^N$. Then

$$\sup |v_{j+1}| \leq \frac{1}{2}C_0\rho^N + \frac{1}{2}C_1A_j^2 \leq \frac{1}{2}(C_0\rho^N + C_1C_0^2\rho^{2N}) = C_0\rho^N \left(\frac{1}{2} + \frac{1}{2}C_1C_0\rho^N\right),$$

which we can make less than $C_0\rho^N$ by choosing $\rho < (C_1C_0)^{-1/N}$.

It now follows by standard theory that the v_j converge to a solution v which satisfies $|v| \leq C_0\rho^N$, and the maximum principle shows that this solution is unique, and in particular independent of N . □

Polyhomogeneity of the solution In order to prove that the solution $u = \tilde{u} + v$ is polyhomogeneous, it suffices to prove that v is conormal of order N , ie that $W_1 \cdots W_\ell v = \mathcal{O}(\rho^N)$ for any $W_j \in \mathcal{V}_b(C_k)$ and for any ℓ .

Observe that the problem localizes near $\mathfrak{C}_{12}$ since the polyhomogeneity of u in all regions where conic points are not coalescing was proved in [16]. Now apply a single b -vector field W to (78). This gives

$$(\Delta_{\tilde{g}_0} + 2)Wv = Wf + 2W\tilde{u}\Delta_{\tilde{g}_0}v + WQ(v).$$

This uses that $\Delta_{\tilde{g}_0} = e^{-2\tilde{u}}\Delta_{g_{0,p}}$. Now $|Wf| \leq C\rho^N$ and $|\Delta_{\tilde{g}_0}v| \leq |-2v + f + Q(v)| \leq C\rho^N$, while $|W\tilde{u}| \leq C$. Finally, we can write $WQ(v) = \lambda Wv$ where λ is a smooth function which is also bounded by $C\rho^N$. Absorbing this last term on the left-hand side perturbs the constant 2 by a small amount, so we can bound Wv by $C'\rho^N$ by the maximum principle.

This argument can obviously be iterated any number of times, which shows that v is indeed conormal of order N .

We have proved:

Proposition 28 *When $k = 2$, or more generally near the locus in $\mathcal{C}_k$ where precisely two conic points collide, the solution u to (49) is polyhomogeneous.*

6.2 Hyperbolic metrics with an arbitrary number of merging cone points

The construction for the case of two merging cone points initiates and provides the pattern for an inductive argument to prove the corresponding regularity result for the solution u when an arbitrary number of points coalesce. The construction in that simpler

case was given in sufficient detail that the steps below for $k > 2$ are straightforward generalizations.

Theorem 29 Fix $k \geq 2$ and $\vec{\beta}$. Then the family of fiberwise conic hyperbolic metrics $g_{p,\beta}$ is polyhomogeneous on C_k .

Proof We begin as before, writing $g_{p,\beta} = e^{2u} g_{0,p}$, where $g_{0,p}$ is the metric which is flat in a neighborhood of the coalescing points. We have already proved that this flat metric is polyhomogeneous, so it suffices to prove that u is polyhomogeneous on C_k .

As in Section 2.5, for any $q \in C_k$, there is a (nonaugmented) tree T and a terminal node on that tree encoding the chain of faces leading to q , where if $\mathcal{I}$ is the node corresponding to q , then q lies in the interior of $\mathfrak{C}_{\mathcal{I}}$. The argument below is an induction on the depth N of the node, ie the height of the tree from the root up to this node.

We have essentially already given the construction of an approximate solution when $N = 1$. More precisely, in the extension of the approximate solution from the fiber M_p to the face $\mathfrak{C}_{12}$, that face is blown up at the two points where the incidence sets F_1^σ and F_2^σ meet $\mathfrak{C}_{12}$. When $k > 2$ (but we are still considering the case $N = 1$), the corresponding construction involves blowing up $\mathfrak{C}_{12\dots k}$ at the k distinct points where the F_j^σ meet this face. The induced metric on $\mathfrak{C}_{12\dots k}$ is flat, with k conic singularities at these intersection points, and a complete conic structure near the outer boundary. The solvability of the sequence of equations to determine the expansion at this face proceeds exactly as before.

Suppose now that we have described how to carry the construction out for all trees of height strictly less than N . Let T be a tree with height N and $\mathcal{I}$ a terminal node (so that only singleton incident sets F_j^σ for $j \in \mathcal{I}$ intersect it). There is a maximal ascending chain $\mathcal{I} = \mathcal{I}_N \subset \dots \subset \mathcal{I}_1$, and by induction we have constructed the full series expansion for the approximate solution near each of the faces $\mathfrak{C}_{\mathcal{I}_\ell}$ for $\ell \leq N-1$. In particular, there is a complete series at the penultimate face $\mathfrak{C}_{\mathcal{I}_{N-1}}$, which is a hemisphere blown up at τ points, where τ is the number of terminal vertices emanating from the vertex $\mathcal{I}_{N-1}$ in T . This face carries a flat metric g_{N-1} , which has incomplete conic singularities at these τ interior boundaries as well as a complete conic structure at its outer boundary.

We now choose the terms in the series expansion at $\mathfrak{C}_{\mathcal{I}_N}$. This process is again almost identical to the one for $k = 2$. To simplify notation, drop subscripts, and let r and s denote the radial variables to $\mathfrak{C}_{\mathcal{I}_N} \cap \mathfrak{C}_{\mathcal{I}_{N-1}}$ in $\mathfrak{C}_{\mathcal{I}_{N-1}}$ and $\mathfrak{C}_{\mathcal{I}_N}$, respectively. No analogue of Lemma 18 is necessary here since $g_{N-1} = dr^2 + \beta_N^2 r^2 d\theta^2$ near

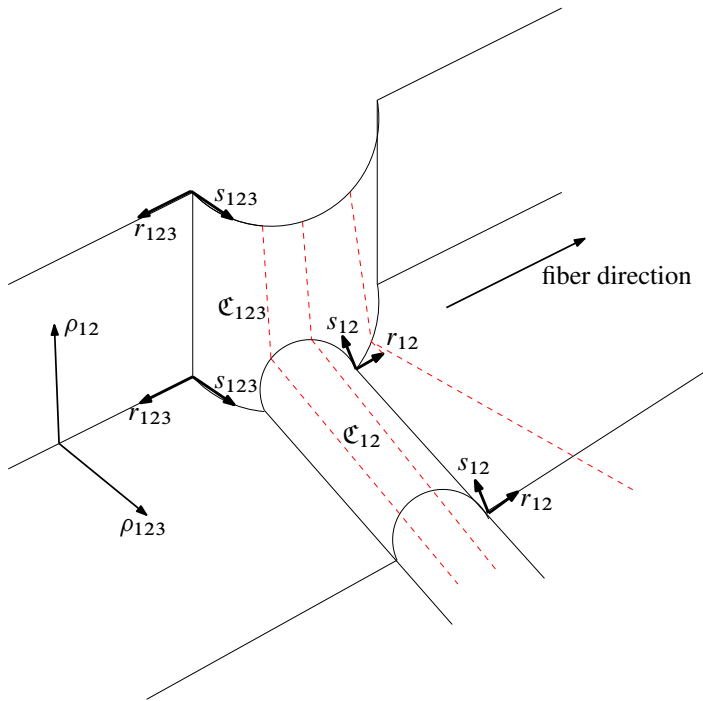


Figure 6: Boundary faces and defining functions of $\mathcal{C}_3$.

this boundary. The approximate solution u has an expansion $u \sim \sum s^j u'_j$ as in Proposition 19. These coefficients are then used to determine the boundary values for the coefficients in the expansion $u \sim \sum R^\alpha u''_\alpha$ as $s \rightarrow 0$; here R is the radial variable to $\mathcal{C}_{\mathcal{I}_N}$ (which restricts to $\mathcal{C}_{\mathcal{I}_{N-1}}$ as r). Rewrite this expansion as $u \sim \sum \rho^\alpha \tilde{u}''_\alpha$ (so we can commute the Laplacian past the powers of ρ). In terms of the defining functions R_j for the intermediate faces $\mathcal{C}_{\mathcal{I}_j}$, where $R_N = R$, $R_{N-1} = s$ and R_0 is equal to a defining function for $M_{p'}$, we can take $\rho = R_0 \cdots R_N$, so that

$$u''_\alpha = \tilde{u}''_\alpha (R_0 \cdots R_{N-1})^\alpha.$$

However, on $\mathcal{C}_{\mathcal{I}_N}$, $R_0, \dots, R_{N-2}$ are all constant, so more simply $u''_\alpha = \tilde{u}''_\alpha s^\alpha$. This puts us in the same situation as before, where we must show that $\tilde{u}''_\alpha = 0$ unless $\alpha \in \mathbb{N}$, and in that case, $\tilde{u}''_\alpha$ is a polynomial of order α , so that u''_α is smooth up to $\mathcal{C}_{\mathcal{I}_{N-1}}$. The calculations and arguments here are just as before.

This paragraph does not conceal any delicate points; the admittedly intricate discussion carried out to extend the expansion from $M_{p'}$ to $\mathcal{C}_{12}$ adapts directly to the extension from $\mathcal{C}_{\mathcal{I}_{N-1}}$ to $\mathcal{C}_{\mathcal{I}_N}$.

Now take a Borel sum of this multiseriess at all the faces of $\mathcal{C}_k$ to obtain a function $\tilde{u}$ for which

$$\Delta_{g_{0,p}} \tilde{u} + K_{0,p} + e^{2\tilde{u}} = \mathcal{O}(\rho^\ell) \quad \text{for all } \ell \geq 0.$$

Writing $\tilde{g}_0 = e^{2\tilde{u}} g_{0,p}$, we then define v uniquely by

$$\Delta_{\tilde{g}_0} v + K_{\tilde{g}_0} + e^{2v} = 0$$

with $|v| \leq C_\ell \rho^\ell$ for any ℓ . This shows that the solution u is polyhomogeneous on $\mathcal{C}_k$. $\square$

7 Spherical metrics: cone angles less than 2π

We conclude this paper by extending Theorems 16 and 29 to the spherical case. As we have explained earlier, the existence theory for spherical cone metrics is completely understood only when all cone angles are less than 2π , so we restrict ourselves to that case here. In a sequel to this paper we investigate this problem for spherical metrics with large cone angles. The results in that case are considerably more intricate and of a slightly different nature than the considerations here.

First note that if $\vec{\beta} \in (0, 1)^k$, then the cone angle produced by merging any subset of these must still have cone angle less than 2π :

$$2\pi \left(\sum_{i \in \mathcal{I}} (\beta_i - 1) + 1 \right) \in (0, 2\pi) \quad \text{for all } \mathcal{I} \subset \{1, \dots, k\}.$$

We also recall that a (spherical) football is the spherical suspension of a circle of length $2\pi\beta$; it is a surface of genus zero with two antipodal conic points, each with angle $2\pi\beta$. This angle may be any positive number.

The main regularity theorem and its proof are very similar to those in the hyperbolic case. Indeed, the only real difference is that it is no longer immediately obvious that the linearized operator is invertible, but, fortunately, this is not the case.

Lemma 30 [16, Proposition 13] *If g is a spherical cone metric on the sphere with all cone angles less than 2π , then the first nonzero eigenvalue for the Friedrichs extension of its Laplacian is always strictly greater than 2, unless (S^2, g) is a football, in which case this eigenvalue is exactly equal to 2.*

In our second paper the primary focus will be on the extensions of this analysis when the Friedrichs extension of the Laplace operator contains 2 in its spectrum.

Remark We shall restrict here to the case $k \geq 3$ and to the fibers of $\mathcal{C}_k$ lying in the dense open subset $\mathcal{C}'_k$ which do not lie above the preimage in $\mathcal{E}_k$ of any of the sets

- (i) the complete diagonal $\Delta_{1\dots k}$, or
- (ii) the intersection of any two partial diagonals $\Delta_{\mathcal{I}} \cap \Delta_{\mathcal{J}}$ with $\mathcal{I} \cap \mathcal{J} = \emptyset$ and $\mathcal{I} \cup \mathcal{J} = \{1, \dots, k\}$ (we include the case where either $|\mathcal{I}| = 1$ or $|\mathcal{J}| = 1$).

These two cases correspond to the degenerations where the k points merge into either one or two points.

The basic existence and uniqueness result in this setting is well known:

Lemma 31 [22; 14] *If $k \geq 3$ and all cone angles are less than 2π , then there exists a unique spherical cone metric on M provided $\chi(M, \vec{\beta}) > 0$, and when $M = \mathbb{S}^2$, the cone angles also satisfy the Troyanov condition*

$$\beta_j - 1 > \sum_{i \neq j} (\beta_i - 1) \quad \text{for each } j.$$

If $k = 2$ and $M = \mathbb{S}^2$, there is a spherical metric if and only if $\beta_1 = \beta_2$, and this metric is unique up to conformal dilation.

Theorem 32 *Let $k \geq 3$, restrict to the open dense subset $\mathcal{C}'_k$ of the configuration family and fix any k -tuple $\vec{\beta}$ of cone angle parameters lying in the Troyanov region. Then the family of fiberwise spherical metrics with these fixed cone angles is polyhomogeneous on $\mathcal{C}'_k$.*

Proof As in Theorem 29, it suffices to work locally near any point q in a corner $\bigcap_{i=1}^{\ell} \mathfrak{C}_{\mathcal{I}_\ell}$, where $\mathcal{I}_1 \supset \dots \supset \mathcal{I}_\ell$ are the nodes associated to q in the associated tree.

Start with a model metric $g_{0,p}$, defined as in (48), which is obtained by gluing the family of (lifted) flat conic metrics to the family of spherical metrics away from the coalescence locus:

$$g_{0,p} = \chi g_{0,p}^{\text{fl}} + (1 - \chi) g_{p'}^{\text{sp}}.$$

Next construct approximate solutions u_N using exactly the same argument as before. Note that we must use Lemma 31 when determining the expansion on the original surface, and replace $\sinh \tilde{r}$ by $\sin \tilde{r}$ in (51). Otherwise, the steps are carried out in exactly the same way. Altogether, taking a Borel sum of the resulting series, we obtain an approximate solution $\tilde{u}$.

We next obtain the correction term v , which solves

$$\Delta_{\tilde{g}_0} v + e^{2v} + K_{\tilde{g}_0} = 0, \quad \tilde{g}_0 = e^{2\tilde{u}} g_{0,p}.$$

By assumption, the linearization of this equation, $\Delta_{\tilde{g}_0} - 2$, is invertible on the singular fiber, hence we may use [Lemma 30](#). Using continuity of the eigenvalues, $\Delta_{\tilde{g}_0} - 2$ remains invertible on nearby fibers. Hence by the implicit function theorem there exists a solution v which depends smoothly on all parameters.

The arguments of [Propositions 27](#) and [28](#) now show that $|v| \leq C_0 \rho^\ell$ for any $\ell \in \mathbb{N}$, with similar estimates for all b -derivatives, so v vanishes to infinite order at these boundaries. This shows that $u = \tilde{u} + v$, and hence $g_p = e^{2u} g_{0,p}$, is polyhomogeneous. $\square$

Appendix b -fibrations and polyhomogeneity

In this appendix we recall some basic facts about manifold with corners, b -fibrations and polyhomogeneous functions. For details we refer to [\[18\]](#) or [\[15, Appendix\]](#).

Definition 33 A space X is called a *manifold with corners* of dimension n if at each point there exists a nonnegative integer k such that X is modeled diffeomorphically near that point by a neighborhood of the origin in the product $(\mathbb{R}^+)^k \times \mathbb{R}^{n-k}$. The boundary faces and corners of X correspond to the boundaries and corners in these local representations. We list the boundary hypersurfaces of X as $\{X_i\}_{i=1}^J$. The corners of codimension ℓ are the submanifolds $X_{i_1} \cap \cdots \cap X_{i_\ell}$. We require that all boundary faces and corners be embedded; this is merely a simplifying assumption to avoid talking about special cases. However, with this assumption we can choose for each i a smooth function ρ_i which is positive on $X \setminus X_i$ and which vanishes simply at X_i . This is called a boundary-defining function for that face. If $s = (s_1, \dots, s_J) \in \mathbb{C}^J$, then $\rho^s := \rho_1^{s_1} \cdots \rho_J^{s_J}$.

The spaces $\mathcal{E}_k$ and $\mathcal{C}_k$ constructed in [Section 2](#) are manifolds with boundaries, with boundary faces $\{\mathcal{C}_{\mathcal{I}}\}$ and $\{F_{\mathcal{I}}\}$, where in each case the index $\mathcal{I}$ ranges over all subsets of $\{1, \dots, k\}$. We do not introduce a special notation for the corners.

There is a whole ecosystem of geometric and analytic objects naturally defined on a manifold with corners. One key object is the following:

Definition 34 The space of b -vector fields $\mathcal{V}_b(X)$ on X is the space of all smooth vector fields on X which are tangent to all boundaries. The b -tangent bundle ${}^bTX$ is a canonical bundle whose full space of smooth sections is precisely $\mathcal{V}_b(X)$. Its dual is the b -cotangent bundle ${}^bT^*X$.

The local coordinate expression of a general b -vector field is given in [Section 4.1](#).

We next come to the first of two most natural replacements for the space of smooth functions on X .

Definition 35 (conormality) The bounded conormal functions on X is the space

$$(79) \quad \mathcal{A}^0(X) := \{u : V_1 \cdots V_\ell u \in L^\infty(X) \text{ for all } V_i \in \mathcal{V}_b \text{ and } \ell \in \mathbb{N}\}.$$

For any pair of multi-indices $s \in \mathbb{C}^J$ and $p \in \mathbb{N}_0^J$, we also define

$$(80) \quad \mathcal{A}^{s,p}(X) := \rho^s (\log \rho)^p \mathcal{A}^0(X).$$

Finally,

$$(81) \quad \mathcal{A}^*(X) := \bigcup_{s,p} \mathcal{A}^{s,p}(X).$$

Note that any element of $\mathcal{A}^*(X)$ is smooth in the interior of X .

There is a subclass of $\mathcal{A}^*(X)$ which is more useful in practice: the space of polyhomogeneous functions. These are associated to an index family:

Definition 36 An index set $\mathcal{E}$ is a discrete subset $\{(s_j, p_j) \in \mathbb{C} \times \mathbb{N}_0\}$ such that

$$(82) \quad (s_j, p_j) \in E \quad \text{and} \quad |(s_j, p_j)| \rightarrow \infty \implies \Re(s_j) \rightarrow \infty.$$

Now suppose that $\mathcal{E} = \{E_1, \dots, E_J\}$ is a collection of index sets, one for each boundary face of X ; we call this an index family. The space of *polyhomogeneous* functions with index family $\mathcal{E}$ (or more simply $\mathcal{E}$ -smooth functions) on X , $\mathcal{A}_{\text{phg}}^\mathcal{E}(X)$, consists of the elements of $\mathcal{A}^*(X)$ with asymptotic expansions with exponents given by the elements of $\mathcal{E}$, ie $u \in \mathcal{A}_{\text{phg}}^\mathcal{E}$ if

$$u \sim \sum_{(s,p) \in \mathcal{E}} \rho^s (\log \rho)^p u_{s,p},$$

where each $u_{s,p}$ is a smooth function the relevant corner of X . This expansion is meant in the classical sense, and has a product type at the corners; it is a tractable replacement for a Taylor expansion at the faces and corners.

We write $u \in \mathcal{A}_{\text{phg}}^*(X)$ when the index set is not specified (or is obvious from the context).

It follows readily from this definition, and is useful in applications, to note that if $u \in \mathcal{A}_{\text{phg}}$, then for each i , any coefficient $u_{s,p}^{(i)}$, which is simply the coefficient of $\rho_i^{s_i} (\log \rho_i)^{p_i}$ in the expansion at X_i , is itself a polyhomogeneous function on X_i with index family $\mathcal{E}^{(i)}$ obtained by omitting $\mathcal{E}_i$.

There is a convenient criterion for polyhomogeneity.

Proposition 37 *If $u \in \mathcal{A}^*(X)$ and, for every $N > 0$,*

$$(83) \quad u - \sum_{\substack{\Re s < N \\ (s,p) \in \mathcal{E}}} \rho^s (\log \rho)^p u_{s,p} \in \rho^{N-1} \mathcal{A}^0(X),$$

then $u \in \mathcal{A}_{\text{phg}}^{\mathcal{E}}(X)$.

We also define a distinguished class of mappings between manifolds with corners.

Definition 38 Let X and Y be manifolds with corners, with corresponding sets of boundary-defining functions $\{r_i\}$ and $\{\rho_j\}$, respectively, associated to the enumerations of boundary faces $\{X_i\}_{i \in \mathcal{I}}$ and $\{Y_j\}_{j \in \mathcal{J}}$. A map $f: X \rightarrow Y$ is called a *b-fibration* if the following conditions are satisfied:

b-map For any index $j \in \mathcal{J}$, the pullback of the corresponding boundary-defining function ρ_j is a smooth nonvanishing multiple of the product of boundary-defining functions of X , ie

$$f^*(\rho_j) = h \prod_{i=1}^I r_i^{e(i,j)}, \quad h > 0, e(i,j) \in \mathbb{N}.$$

The exponent set $e(i,j)$ is called the lifting matrix of f .

b-submersion At each boundary point $p \in \partial X$, the map ${}^b f_*: T_p X \rightarrow T_{f(p)} Y$ is surjective.

b-fibration The lifting matrix $(e(i,j))$ has the property that for each i there is at most one j such that $e(i,j) \neq 0$. In other words, no boundary hypersurface of X is mapped into a corner of Y .

References

- [1] **P Albin, R Melrose**, *Resolution of smooth group actions*, from “Spectral theory and geometric analysis” (M Braverman, L Friedlander, T Kappeler, P Kuchment, P Topalov, J Weitsman, editors), Contemp. Math. 535, Amer. Math. Soc., Providence, RI (2011) 1–26 [MR](#)
- [2] **D Bartolucci, F De Marchis, A Malchiodi**, *Supercritical conformal metrics on surfaces with conical singularities*, Int. Math. Res. Not. 2011 (2011) 5625–5643 [MR](#)
- [3] **A Carlotto, A Malchiodi**, *A class of existence results for the singular Liouville equation*, C. R. Math. Acad. Sci. Paris 349 (2011) 161–166 [MR](#)
- [4] **A Carlotto, A Malchiodi**, *Weighted barycentric sets and singular Liouville equations on compact surfaces*, J. Funct. Anal. 262 (2012) 409–450 [MR](#)
- [5] **C-C Chen, C-S Lin**, *Topological degree for a mean field equation on Riemann surfaces*, Comm. Pure Appl. Math. 56 (2003) 1667–1727 [MR](#)
- [6] **C-C Chen, C-S Lin**, *Mean field equation of Liouville type with singular data: topological degree*, Comm. Pure Appl. Math. 68 (2015) 887–947 [MR](#)
- [7] **S Dey**, *Spherical metrics with conical singularities on 2–spheres*, Geom. Dedicata 196 (2018) 53–61 [MR](#)
- [8] **A Eremenko**, *Metrics of positive curvature with conic singularities on the sphere*, Proc. Amer. Math. Soc. 132 (2004) 3349–3355 [MR](#)
- [9] **A Eremenko**, *Co-axial monodromy*, preprint (2017) [arXiv](#)
- [10] **A Eremenko, A Gabrielov, V Tarasov**, *Metrics with conic singularities and spherical polygons*, Illinois J. Math. 58 (2014) 739–755 [MR](#)
- [11] **M Kapovich**, *Branched covers between spheres and polygonal inequalities in simplicial trees*, preprint (2017) Available at <https://www.math.ucdavis.edu/~kapovich/EPR/covers.pdf>
- [12] **C Kottke, M Singer**, *Partial compactification of monopoles and metric asymptotics*, preprint (2015) [arXiv](#)
- [13] **Y Lee, C-S Lin, W Yang, L Zhang**, *Degree counting for Toda system with simple singularity: one point blow up*, J. Differential Equations 268 (2020) 2163–2209 [MR](#)
- [14] **F Luo, G Tian**, *Liouville equation and spherical convex polytopes*, Proc. Amer. Math. Soc. 116 (1992) 1119–1129 [MR](#)
- [15] **R Mazzeo**, *Elliptic theory of differential edge operators, I*, Comm. Partial Differential Equations 16 (1991) 1615–1664 [MR](#)
- [16] **R Mazzeo, H Weiss**, *Teichmüller theory for conic surfaces*, from “Geometry, analysis and probability” (J-B Bost, H Hofer, F Labourie, Y Le Jan, X Ma, editors), Progr. Math. 310, Springer (2017) 127–164 [MR](#)

- [17] **RC McOwen**, *Point singularities and conformal metrics on Riemann surfaces*, Proc. Amer. Math. Soc. 103 (1988) 222–224 [MR](#)
- [18] **R Melrose**, *Differential analysis on manifolds with corners*, book project (1996) Available at <http://www-math.mit.edu/~rbm/book.html>
- [19] **G Mondello, D Panov**, *Spherical metrics with conical singularities on a 2-sphere: angle constraints*, Int. Math. Res. Not. 2016 (2016) 4937–4995 [MR](#)
- [20] **J Song, Y Chen, B Li, B Xu**, *Drawing cone spherical metrics via Strebel differentials*, Int. Math. Res. Not. (online publication May 2018) art. id. rny103
- [21] **M Troyanov**, *Metrics of constant curvature on a sphere with two conical singularities*, from “Differential geometry” (FJ Carreras, O Gil-Medrano, A M Naveira, editors), Lecture Notes in Math. 1410, Springer (1989) 296–306 [MR](#)
- [22] **M Troyanov**, *Prescribing curvature on compact surfaces with conical singularities*, Trans. Amer. Math. Soc. 324 (1991) 793–821 [MR](#)
- [23] **M Umehara, K Yamada**, *Metrics of constant curvature 1 with three conical singularities on the 2-sphere*, Illinois J. Math. 44 (2000) 72–94 [MR](#)

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

Department of Mathematics, UC Berkeley
Berkeley, CA, United States

rafemazzeo@gmail.com, xuwenzhu@math.berkeley.edu

Proposed: Gang Tian

Seconded: Simon Donaldson, Tobias H Colding

Received: 17 June 2018

Accepted: 24 June 2019

$\mathrm{GL}_2\mathbb{R}$ –invariant measures in marked strata: generic marked points, Earle–Kra for strata and illumination

PAUL APISA

We show that nontrivial $\mathrm{GL}(2, \mathbb{R})$ –invariant point markings for translation surfaces in strata of abelian differentials exist only when the translation surface belongs to a hyperelliptic component. As an application, we establish constraints on sections of the universal curve restricted to orbifold covers of subvarieties of the moduli space of Riemann surfaces that contain a Teichmüller disk. We also solve the finite blocking problem for generic translation surfaces.

30F60, 32G15, 37F30

1 Introduction

Let $\mathcal{QM}_g$ be the moduli space of quadratic differentials on genus g Riemann surfaces. The space admits a $\mathrm{GL}(2, \mathbb{R})$ action generated by Teichmüller geodesic flow and complex scalar multiplication. The space $\mathcal{QM}_g$ also admits a $\mathrm{GL}(2, \mathbb{R})$ –invariant stratification given by specifying the number of zeros and poles of the quadratic differentials and their orders of vanishing.

A point of $\mathcal{QM}_g$ will be denoted by (X, q) , where X is a Riemann surface and q is a quadratic differential. A point (X, q) will be called generic if its $\mathrm{GL}(2, \mathbb{R})$ orbit is dense in the connected component of the stratum containing it. We are interested in studying collections of points P on quadratic differentials (X, q) . We will always assume that P does not contain any zeros or poles of q . Say that P is generic if the orbit closure of $(X, q; P)$ is as big as possible, ie

$$\dim_{\mathbb{C}} \overline{\mathrm{GL}(2, \mathbb{R}) \cdot (X, q; P)} = |P| + \dim_{\mathbb{C}} \overline{\mathrm{GL}(2, \mathbb{R}) \cdot (X, q)}.$$

A nongeneric point is called a periodic point.

A translation surface, which we will denote by (X, ω) , is a Riemann surface X together with a holomorphic one-form ω . The one-form determines a quadratic differential by taking its square. The $\mathrm{GL}(2, \mathbb{R})$ action on the space of quadratic differentials preserves

the locus of translation surfaces. We will now characterize the nongeneric collections of points on generic translation surfaces.

Theorem 1.1 *If (X, ω) is a generic translation surface of genus at least two and P is a nongeneric collection of points then (X, ω) belongs to a hyperelliptic component of a stratum and P contains either a Weierstrass point or two points exchanged by the hyperelliptic involution.*

Remark 1.2 In Apisa and Wright [3], it is shown by separate methods that if (X, q) is a generic quadratic differential that does not belong to a stratum of abelian differentials then the conclusion of Theorem 1.1 holds. Neither result implies the other.

Holomorphic sections of the universal curve In Hubbard [14], it was shown that holomorphic sections of the universal curve over $\text{Teich}_{g,n}$ — the Teichmüller space of genus g Riemann surfaces with n punctures — exist only when $(g, n) = (2, 0)$, in which case the sections mark fixed points of the hyperelliptic involution. Earle and Kra [6] generalized this result by allowing the sections to mark punctured points. They showed that in this more general setting, the only new sections that could arise were in genus two and are given by taking a punctured point and either marking it or its image under the hyperelliptic involution. The following result gives a constraint on holomorphic sections of finite orbifold covers of the universal curve that explicitly identifies a finite collection of points that may be marked by holomorphic sections.

Let C be a subvariety of $\mathcal{M}_{g,n}$ and let Γ be a torsionfree finite-index subgroup of the mapping class group $\text{Mod}_{g,n}$. Let $C(\Gamma)$ be the preimage of C on $\text{Teich}_{g,n}/\Gamma$.

Theorem 1.3 *If C contains the Teichmüller disk generated by the quadratic differential (X, q) , then any holomorphic section of the universal curve over $C(\Gamma)$ marks a point p on X that is a pole, zero or periodic point of (X, q) .*

Remark 1.4 By Eskin, Filip and Wright [7, Theorem 1.5], there are finitely many periodic points on (X, q) if and only if its holonomy double cover is not a torus cover. Therefore, for such quadratic differentials, Theorem 1.3 shows that a holomorphic section of the universal curve can only mark finitely many points and explicitly describes which points may be marked.

Two applications of Theorems 1.1 and 1.3 Apart from $\mathcal{M}_{g,n}$ itself, other examples of algebrogeometrically interesting varieties that contain a Teichmüller disk are the theta-null divisor (see Müller [27] and Grushevsky and Zakharov [13]), the antiramification

locus (see Farkas and Verra [10]) and the Weierstrass divisor (see Cukierman [5]). More examples are listed in Mullane [26] and the Kodaira dimensions of many such loci are computed in Gendron [12]. Each of these examples is the projection of a nonhyperelliptic component $\mathcal{H}$ of some stratum of abelian differentials to $\mathcal{M}_{g,n}$. Therefore, the following is immediate from Theorems 1.1 and 1.3.

Corollary 1.5 *If $\mathcal{C}$ is any of the preceding loci and X is a Riemann surface in $\mathcal{C}$, then there is an abelian differential ω such that (X, ω) belongs to $\mathcal{H}$ and the only points on X that may be marked by a holomorphic section of the universal curve defined on $\mathcal{C}$ are zeros of ω .*

Remark 1.6 In particular, this corollary recovers a classical fact that for any finite-index torsionfree subgroup Γ of the mapping class group, there are no holomorphic sections of the universal curve defined over $\mathrm{Teich}_{g,n}/\Gamma$ unless $g = 2$.

Another application of Theorem 1.3 is to sections of the universal curve defined over Hilbert modular surfaces in genus two. In [20; 21; 22; 23], McMullen showed that the closures of Teichmüller disks generated by an abelian differential on a genus two Riemann surface is $\mathcal{M}_2$, is contained in a locus of torus covers, or is a Hilbert modular surface. The periodic points on a torus cover (X, ω) are exactly the preimages of torsion points on the torus. It was shown in Apisa [1] that for any other abelian differential (X, ω) in genus two—except those whose orbit closures project to E_5 , ie the Hilbert modular surface parametrizing curves whose Jacobian admits real multiplication by the maximal order in $\mathbb{Q}[\sqrt{5}]$ —that the only periodic points are Weierstrass points.

Corollary 1.7 *The only holomorphic sections of the universal curve defined over genus two Hilbert modular surfaces (excluding E_5) mark Weierstrass points or zeros of the eigenform.*

Remark 1.8 Up to the hyperelliptic involution there is one additional point that may be marked on the universal curve above E_5 and it is described in Kumar and Mukamel [16].

Finite blocking on translation surfaces One motivation for studying marked points on translation surfaces is rational billiards. Given a rational billiard table and two points p and q one may ask whether there is a billiard trajectory from p to q . More ambitiously, one may seek to ascertain whether there is a finite collection of points

that block all shots from p to q . To study rational billiards, one often applies the unfolding construction of Katok and Zemlyakov [30] to turn the billiard table into a translation surface where billiard trajectories correspond to straight lines on the translation surface. The previously posed problem then becomes the finite blocking problem: given two points p and q on a translation surface is there a finite collection of points that intersects every straight line from p to q ? We will continue to assume that neither p nor q coincide with a singularity of the flat metric.

Theorem 1.9 *A generic translation surface contains a pair of finitely blocked points if and only if it belongs to a hyperelliptic component, in which case the pair consists of a point and its image under the hyperelliptic involution.*

Proof of Theorem 1.9 given Theorem 1.1 By Apisa and Wright [3, Theorem 3.15] and Theorem 1.1, finitely blocked pairs of points only occur when the generic translation surfaces belongs to a hyperelliptic component; and in this case, the pair of points consists of either two Weierstrass points or two points exchanged by the hyperelliptic involution and in both cases the finite blocking set is the collection of Weierstrass points.

Since the collection of translation surfaces represented by strictly convex $2n$ -gons with opposite sides identified is open, nonempty and $\mathrm{GL}(2, \mathbb{R})$ -invariant in hyperelliptic components of strata, its complement consists of nongeneric translation surfaces. Thus, the generic translation surface may be represented by a strictly convex $2n$ -gon with opposite sides identified. The Weierstrass points are the midpoints of the polygon and its edges (and the vertices when n is even). By convexity a Weierstrass point is at most finitely blocked from itself and so the pairs of points finitely blocked from each other are exactly the ones containing a point and its image under the hyperelliptic involution. $\square$

For more on the finite blocking problem, see Lelièvre, Monteil and Weiss [17]. For applications to billiards, see Mirzakhani and Wright [25, Sections 6–7], Apisa and Wright [3, Section 3] and Apisa [1].

Organization The proof of Theorem 1.3 is independent of the rest of the paper and appears in Section 3. The outline of the proof of Theorem 1.1 is given in Section 6 and reduced to two more technical results that are established in the subsequent two sections. The two main tools used in the proof of Theorem 1.1 are the construction of generic horizontally and vertically periodic translation surfaces in every component of every stratum of holomorphic one-forms in Section 4 and results in Section 5 that constrain the positions of periodic points using cylinders.

Acknowledgement The author thanks Alex Eskin for suggesting the problem and thanks Alex Eskin, Alex Wright and Curt McMullen for their insightful comments. He thanks Ronen Mukamel for suggesting the parallel between the main theorems and the work of Hubbard and of Earle and Kra. He thanks Matt Bainbridge for suggesting the connection to the problem of classifying holomorphic sections. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1144082 and upon work supported by the National Science Foundation under Award No. 1803625. The author gratefully acknowledges their support.

2 Background

Let $\mathcal{M}_{g,n}$ denote the moduli space of connected orientable genus g surfaces with n punctures. Let $\Omega\mathcal{M}_{g,n}$ be the moduli space of holomorphic one-forms (equivalently abelian differentials) on Riemann surfaces of genus g with n punctures. Notice that a collection of polygons in $\mathbb{C}$ — with sides identified by translation to form a genus g surface with n punctures — produces an element of $\Omega\mathcal{M}_{g,n}$ by defining a one-form that is equal to dz on each polygon, where z is the coordinate on $\mathbb{C}$. Such a surface is called a translation surface. Every element of $\Omega\mathcal{M}_{g,n}$ may be written as a translation surface. If X is a Riemann surface and ω a holomorphic one-form, then the flat metric on (X, ω) refers to the metric induced on X from its presentation as a translation surface.

As mentioned in the introduction, $\Omega\mathcal{M}_g$ admits a $\mathrm{GL}(2, \mathbb{R})$ action and a $\mathrm{GL}(2, \mathbb{R})$ -invariant stratification. In terms of the polygonal presentation of translation surfaces, the $\mathrm{GL}(2, \mathbb{R})$ action may be described as the action induced by the $\mathrm{GL}(2, \mathbb{R})$ action on the polygons after making the identification of $\mathbb{C}$ with $\mathbb{R}^2$ via $x + iy \rightarrow (x, y)$.

Recall that the stratification of $\Omega\mathcal{M}_g$ was given by fixing the orders and multiplicities of the zeros of the one-forms. Given a one-form ω on a Riemann surface X we will let Σ denote its zero set. In the polygonal presentation, the points of Σ are the cone points of the flat metric (ie the points where the circumference of a circle of radius ϵ is not $2\pi\epsilon$ for sufficiently small ϵ). Local coordinates on strata are given by period coordinates, which is a coordinate system that records the periods of a basis of relative homology $H_1(X, \Sigma; \mathbb{Z})$. It follows that the tangent space at (X, ω) of a stratum of abelian differentials containing (X, ω) is $H^1(X, \Sigma; \mathbb{C})$.

If $|\Sigma| = 1$, then the change of coordinates between period coordinate charts is given by changing one homology basis with another and hence is an element of $\mathrm{Sp}(2g, \mathbb{Z})$.

More generally, when $|\Sigma|$ is arbitrary, the change of coordinates for period coordinates remains a constant volume-preserving linear function. This endows strata of $\Omega\mathcal{M}_g$ with a linear structure and a well-defined Lebesgue measure; see Zorich [31] for more details.

Lebesgue measure on strata of $\Omega\mathcal{M}_g$, also called Masur–Veech measure, induces a finite $\mathrm{SL}(2, \mathbb{R})$ –invariant measure on $\mathcal{U}$, the locus of unit-area translation surfaces. This measure is ergodic with respect to the $\mathrm{SL}(2, \mathbb{R})$ action on $\mathcal{U}$ by Masur [18] and Veech [28]. Notice that it is necessary to restrict to $\mathcal{U}$ so that the total mass of the Masur–Veech measure is finite. We must also replace the $\mathrm{GL}(2, \mathbb{R})$ action by an $\mathrm{SL}(2, \mathbb{R})$ action in order to preserve the set $\mathcal{U}$. One might now inquire whether there is a classification of ergodic $\mathrm{SL}(2, \mathbb{R})$ –invariant measures on $\mathcal{U}$.

2.1 Affine invariant submanifolds and equivariant maps between them

Definition 2.1 An affine invariant submanifold is a closed subset of a stratum of $\Omega\mathcal{M}_g$, which, in local period coordinates, can be expressed as the vanishing locus of a collection of real linear equations.

The arguments of Masur and Veech mentioned in the previous paragraph show that Lebesgue measure on affine invariant submanifolds restricted to $\mathcal{U}$ produces a finite ergodic $\mathrm{SL}(2, \mathbb{R})$ –invariant measure. Eskin and Mirzakhani [8] showed that this is the only way to produce such measures on $\mathcal{U}$. Using this result, Eskin, Mirzakhani and Mohammadi [9] showed that the $\mathrm{GL}(2, \mathbb{R})$ orbit closure of any holomorphic one-form in a stratum is an affine invariant submanifold.

Definition 2.2 Given an affine invariant submanifold $\mathcal{M}$, let $\mu_{\mathcal{M}}$ be Lebesgue measure on $\mathcal{M} \cap \mathcal{U}$. Given a $\mathrm{GL}(2, \mathbb{R})$ –equivariant measurable map $f: \mathcal{M} \rightarrow \mathcal{N}$ between two affine invariant submanifolds, define the pushforward of $\mathcal{M}$, denoted by $f_*\mathcal{M}$, to be the affine invariant submanifold represented by the ergodic $\mathrm{SL}(2, \mathbb{R})$ –invariant measure given by $f_*\mu_{\mathcal{M}}$.

Notice that $f(\mathcal{M})$ and $f_*\mathcal{M}$ coincide up to sets of measure zero. Specifically, $f_*\mathcal{M} \setminus f(\mathcal{M})$ has $f_*\mu_{\mathcal{M}}$ –measure zero since its preimage lies in the complement of $\mathrm{supp}(\mu_{\mathcal{M}}) = \mathcal{M} \cap \mathcal{U}$. Conversely, $f(\mathcal{M}) \setminus f_*\mathcal{M}$ lies outside the support of $f_*\mu_{\mathcal{M}}$ and hence has $f_*\mu_{\mathcal{M}}$ –measure zero. Thus, its preimage on $\mathcal{M}$ has measure zero.

Example 2.3 Let $\mathcal{M}_{g,n}$ be the moduli space of genus g surfaces with n labeled marked points. Consider the affine invariant submanifold $\mathcal{M}$ in $\Omega\mathcal{M}_{2,2}$ consisting of holomorphic one-forms with a double zero and with two distinct marked points,

marking the two punctures, that are exchanged by the hyperelliptic involution. Let $\mathcal{H}$ be the stratum of $\Omega\mathcal{M}_{2,1}$ consisting of abelian differentials with one double zero and one distinct marked point that records the location of the puncture. There is a forgetful map $f: \mathcal{M} \rightarrow \mathcal{H}$ that forgets the second marked point. Notice that $f_*\mathcal{M}$ is the collection of all holomorphic one-forms with a double zero and with a marked point that does not coincide with the double zero. However, $f(\mathcal{M})$ is the collection of all holomorphic one-forms with a double zero and with a marked point that does not coincide with any Weierstrass point. This illustrates how $f_*\mathcal{M}$ and $f(\mathcal{M})$ may differ up to sets of measure zero.

Finally we note the following.

Theorem 2.4 (Filip [11]) *Affine invariant submanifolds are algebraic subvarieties of $\Omega\mathcal{M}_g$.*

2.2 The tangent space of an affine invariant submanifold

As shown above, the tangent space at (X, ω) of a stratum of abelian differentials containing (X, ω) is $H^1(X, \Sigma; \mathbb{C})$. Therefore, the tangent space of affine invariant submanifolds containing a point (X, ω) is a subspace of $H^1(X, \Sigma; \mathbb{C})$. Since affine invariant submanifolds are locally linear, the tangent subspace is locally constant in period coordinates.

Let $p: H^1(X, \Sigma; \mathbb{C}) \rightarrow H^1(X; \mathbb{C})$ be the projection from relative to absolute cohomology. Notice that the kernel of the projection has size at most $|\Sigma| - 1$.

Theorem 2.5 (Avila, Eskin and Möller [4]) *If $\mathcal{M}$ is an affine invariant submanifold containing a translation surface (X, ω) , then $p(T_{(X, \omega)}\mathcal{M})$ is complex symplectic.*

2.3 Cylinder deformations

Definition 2.6 A cylinder c in a translation surface (X, ω) is a maximal isometrically embedded Euclidean cylinder that contains no points of Σ in its interior (the isometric embedding is with respect to the flat metric on (X, ω)). The cylinder is foliated by isotopic curves — whose homotopy class γ_c is called the core curve — that generate the fundamental group of the cylinder. The distance between the two boundaries of c is called the height, h_c , of the cylinder. The condition that a cylinder is maximal just means that h_c is as big as possible. The modulus of a cylinder is defined to be $h_c / \left| \int_{\gamma_c} \omega \right|$. Notice that there are two choices (each of which is an inverse of the other)

for the core curve of c . Whenever a collection of parallel cylinders is specified we will tacitly assume that their core curves all have periods whose arguments, when taken as complex numbers, are identical.

Masur showed in [19] that every translation surface contains a cylinder. A fundamental tool in the sequel will be cylinder deformations.

Definition 2.7 Suppose that (X, ω) is a translation surface in an affine invariant submanifold $\mathcal{M}$. Two cylinders C_1 and C_2 on (X, ω) will be said to be $\mathcal{M}$ -equivalent if C_1 and C_2 are parallel at all surfaces in an open neighborhood of (X, ω) in $\mathcal{M}$. A maximal collection of equivalent cylinders on (X, ω) will be called an $\mathcal{M}$ -equivalence class. Given a collection $\mathcal{C}$ of cylinders whose periods have argument $e^{i\theta}$, the standard shear, $\sigma_{\mathcal{C}}$, and the standard dilation, $a_{\mathcal{C}}$, are the cohomology classes

$$\sigma_{\mathcal{C}} := e^{i\theta} \sum_{c \in \mathcal{C}} h_c \gamma_c^* \quad \text{and} \quad a_{\mathcal{C}} := i e^{i\theta} \sum_{c \in \mathcal{C}} h_c \gamma_c^*.$$

These cohomology classes belong to $H^1(X, \Sigma; \mathbb{C})$, where Σ is the zero set of ω . Recall that this cohomology group is the tangent space at (X, ω) of the stratum containing (X, ω) . Since this stratum is linear we can travel along the line in the direction of the $\sigma_{\mathcal{C}}$ (resp. $a_{\mathcal{C}}$). In the case that the cylinders in $\mathcal{C}$ are horizontal, this line corresponds to the path given by applying the matrix $\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ (resp. $\begin{pmatrix} 1 & 0 \\ 0 & 1+t \end{pmatrix}$), for $t \in \mathbb{R}$, to the cylinders in $\mathcal{C}$ while fixing the complement of $\mathcal{C}$. This justifies the terminology “standard shear” (resp. dilation). If $\mathcal{C}$ is a nonhorizontal collection of cylinders whose core curves are oriented so that their periods have argument $e^{i\theta}$, then flowing along $\sigma_{\mathcal{C}}$ (resp. $a_{\mathcal{C}}$) corresponds to shearing (resp. dilating the heights of) the cylinders in $\mathcal{C}$.

Example 2.8 Let $\mathcal{H}$ be a connected component of a stratum of abelian differentials. Let C_1 and C_2 be two parallel cylinders on a translation surface (X, ω) in $\mathcal{H}$ with core curves γ_1 and γ_2 respectively. Suppose without loss of generality that C_1 and C_2 are horizontal. If γ_1 and γ_2 are homologous, then $\int_{\gamma_1} \eta = \int_{\gamma_2} \eta$ for any nearby translation surface $(Y, \eta) \in \mathcal{H}$ and so the two cylinders are $\mathcal{H}$ -parallel. Conversely, if γ_1 and γ_2 are not homologous then there is a simple closed curve γ whose oriented intersection number with γ_1 is 1 and whose oriented intersection number with γ_2 is zero. Since γ^* belongs to $H^1(X, \Sigma; \mathbb{C})$, which is the tangent space to $\mathcal{H}$, we may travel along a line in the $i \cdot \gamma^*$ direction. This keeps γ_2 horizontal while making γ_1 nonhorizontal. This example shows that two cylinders are $\mathcal{H}$ -equivalent if and only if their core curves are homologous.

Theorem 2.9 (Wright [29, Corollary 3.4]) Suppose that (X, ω) is a horizontally periodic translation surface whose $\mathrm{GL}(2, \mathbb{R})$ orbit closure is $\mathcal{M}$. Let $(C_1, \dots, C_n)$ be an enumeration of the horizontal cylinders and suppose that cylinder C_i has modulus m_i and core curve (oriented from left to right) γ_i of length c_i for $i = 1, \dots, n$. Let $W \subseteq \mathbb{Q}^n$ be the subset of rational homogeneous linear relations that the moduli $(m_i)_{i=1}^n$ satisfy; ie $w \in W$ if and only if $w \cdot m = 0$. If $(v_i)_{i=1}^n \in \mathbb{C}^n$ belongs to $W^\perp$, then

$$\sum_{i=1}^n c_i v_i \gamma_i^* \in T_{(X, \omega)} \mathcal{M}.$$

Theorem 2.10 [29, Lemma 4.11] If $\mathcal{C}$ is an $\mathcal{M}$ -equivalence class of horizontal cylinders on a translation surface (X, ω) contained in an affine invariant submanifold $\mathcal{M}$ then the standard shear and the standard dilation are both contained in $T_{(X, \omega)} \mathcal{M}$.

2.4 The boundary of an affine invariant submanifold

In Mirzakhani and Wright [24], a partial compactification of any affine invariant submanifold is defined as follows. A sequence (X_n, ω_n) of translation surfaces in a component of a stratum $\mathcal{H}$ converges to (Y, η) — where Y is a compact Riemann surface with finitely many connected components and η is a holomorphic one-form on each component together with a collection of marked points (which are considered part of the zero set Σ of η by thinking of the marked points as zeros of order zero) — if there is a nested collection of open neighborhoods $U_n \subseteq Y$ such that:

- (1) $\bigcap_n U_n = \Sigma$.
- (2) There are maps $g_n: Y \setminus U_n \rightarrow X_n$ that are diffeomorphisms onto their image and such that $g_n^* \omega_n$ converges to ω in the compact open topology.
- (3) The injectivity radius of all points $p \in X_n$ — ie the supremum of the radius of balls (in the flat metric) centered at p that are embedded and do not contain points of Σ — not in the image of g_n goes to zero uniformly in n .

Given an affine invariant submanifold $\mathcal{M}$, let $\overline{\mathcal{M}}^{\mathrm{MW}}$ denote its partial compactification, which is given by $\mathcal{M}$ together with the union of all its limit points (as defined above). The $\mathrm{GL}(2, \mathbb{R})$ action on $\mathcal{M}$ extends continuously to $\overline{\mathcal{M}}^{\mathrm{MW}}$.

Theorem 2.11 (Mirzakhani and Wright [24, Corollary 1.2]) Suppose that $\mathcal{M}$ is an affine invariant submanifold and (X_n, ω_n) is a sequence of points in $\mathcal{M}$ that converges to a (possibly disconnected) translation surface (Y, η) in the boundary of $\mathcal{M}$. The orbit closure of any component of (Y, η) has dimension smaller than $\dim \mathcal{M}$.

2.5 Marked points on affine invariant submanifolds

Finally, the fundamental object of study in the sequel will be the following.

Definition 2.12 Let $\mathcal{M}$ be an affine invariant submanifold in a connected component of a stratum of abelian differentials $\mathcal{H}$. Fix a positive integer n . Let $\mathcal{H}(0^n)$ be the collection of translation surfaces in $\mathcal{H}$ together with n distinct marked points that do not coincide with zeros of the underlying abelian differential. The space $\mathcal{H}(0^n)$ is called a component of a marked stratum of abelian differentials. Let $\pi: \mathcal{H}(0^n) \rightarrow \mathcal{H}$ be the forgetful map. Define $\mathcal{M}(0^n)$ to be the preimage of $\mathcal{M}$ under π .

The forgetful map π extends to an equivariant map $\overline{\mathcal{M}(0^n)}^{\text{MW}} \rightarrow \overline{\mathcal{M}}^{\text{MW}}$ by forgetting the marked points. Let $\overline{\mathcal{M}(0^n)}$ be the preimage of $\mathcal{M}$ in $\overline{\mathcal{M}(0^n)}^{\text{MW}}$ under the forgetful map. Notice that $\overline{\mathcal{M}(0^n)}$ is the partial compactification of $\mathcal{M}(0^n)$, where marked points are allowed to coincide with each other and with points of Σ .

Extending this definition to quadratic differentials introduces a minor complication because the Mirzakhani–Wright compactification is specifically defined for affine invariant submanifolds in strata of abelian differentials. Nevertheless, every quadratic differential has a canonical branched cover—called the holonomy double cover—so that the pullback of the quadratic differential under this cover is the square of an abelian differential.

Definition 2.13 Let $\mathcal{M}$ be an affine invariant submanifold in a connected component of a stratum of quadratic differentials $\mathcal{Q}$. Let $\widetilde{\mathcal{Q}}$ denote the collection of holonomy double covers of quadratic differentials in $\mathcal{Q}$, which is itself an affine invariant submanifold. On each surface in $\widetilde{\mathcal{Q}}$ we use the term “holonomy involution” to refer to the involution whose quotient gives the holonomy double cover.

As before, fix a positive integer n and let $\mathcal{Q}(0^n)$ denote the collection of n distinct marked points that do not coincide with zeros or poles of the quadratic differential. The locus of holonomy double covers $\widetilde{\mathcal{Q}(0^n)}$ is the $\text{GL}(2, \mathbb{R})$ -invariant subset of $\widetilde{\mathcal{Q}(0^{2n})}$ consisting of surfaces in $\widetilde{\mathcal{Q}}$ together with $2n$ distinct points that do not coincide with zeros and which project to exactly n points under the holonomy involution. Let $\mathcal{C}$ be the closure of $\widetilde{\mathcal{Q}(0^n)}$ in $\widetilde{\mathcal{Q}(0^{2n})}$ (this uses [Definition 2.12](#) with $\widetilde{\mathcal{Q}}$ as the affine invariant submanifold). Define the points in $\overline{\mathcal{Q}(0^n)}$ to be the set of marked surfaces in $\mathcal{C}$ after quotienting by the holonomy involution. This creates a bijection between $\mathcal{C}$ and $\overline{\mathcal{Q}(0^n)}$. We give $\overline{\mathcal{Q}(0^n)}$ the topology that makes this bijection a homeomorphism.

The forgetful map from $\mathcal{Q}(0^n)$ to $\mathcal{Q}$ extends to one between $\overline{\mathcal{Q}(0^n)}$ and $\mathcal{Q}$. Let $\overline{\mathcal{M}(0^n)}$ (resp. $\mathcal{M}(0^n)$) be the preimage of $\mathcal{M}$ under the forgetful map whose domain is $\overline{\mathcal{Q}(0^n)}$ (resp. $\mathcal{Q}(0^n)$). The fibers of the forgetful map with domain $\overline{\mathcal{M}(0^n)}$ are all collections of n marked points, where the points are now allowed to coincide with each other, zeros and poles.

Lemma 2.14 *Let $\mathcal{M}$ be an affine invariant submanifold and let $\mathcal{N}$ be an affine invariant submanifold in $\mathcal{M}(0)$. Let $\pi: \mathcal{M}(0) \rightarrow \mathcal{M}$ be the map that forgets the marked points. If $\pi_*\mathcal{N} = \mathcal{M}$ then $\pi(\mathcal{N})$ is an open dense subset of $\mathcal{M}$ and the fiber over a generic point in $\mathcal{M}$ is nonempty. Moreover, π is open.*

Proof Since all fibers of the forgetful map $\pi: \overline{\mathcal{M}(0)} \rightarrow \mathcal{M}$ are compact, π is proper and its image is closed. The image of π has full measure in $\pi_*\mathcal{N}$, which is $\mathcal{M}$ by assumption. If $\mathcal{N} = \mathcal{M}(0)$ then the result is immediate, so suppose instead that $\mathcal{N}$ and $\mathcal{M}$ have the same dimension. Notice that all surfaces in $\overline{\mathcal{M}(0)}$ are connected by construction. As a reminder, $\overline{\mathcal{M}(0)}$ is the partial compactification defined in Definition 2.12, which is a subset of the Mirzakhani–Wright partial compactification. Therefore, by Mirzakhani and Wright [24, Corollary 1.2] — see our Theorem 2.11 — the components of the boundary of $\overline{\mathcal{N}}^{\mathrm{MW}} \cap \overline{\mathcal{M}(0)}$ have strictly smaller dimension than $\mathcal{M}$ and hence their pushforward $\mathcal{C}$ cannot be $\mathcal{M}$ by Sard’s theorem. Therefore, $\pi(\mathcal{N})$ contains the complement of $\mathcal{C}$, which is an open dense set in $\mathcal{M}$. Since $\pi: \mathcal{N} \rightarrow \mathcal{M}$ is a finite holomorphic map between equidimensional varieties (by Filip [11]; see our Theorem 2.4), it is open. $\square$

Definition 2.15 If $\mathcal{M}$ is an affine invariant submanifold then let $\mathcal{M}^{\mathrm{ord}}(0^n)$ be the finite cover of $\mathcal{M}(0^n)$ where the marked points are labeled. Let $\pi_k: \mathcal{M}(0^n) \rightarrow \mathcal{M}(0^{n-1})$ be the map that forgets the k^{th} marked point for $k \in \{1, \dots, n\}$.

Lemma 2.16 *Let $\mathcal{N}$ be an affine invariant submanifold in $\mathcal{M}(0^n)$ that pushes forward to $\mathcal{M}$ under the map π that forgets all marked points. If $(X, \omega; P)$ is generic in $\mathcal{M}(0^n)$, then there is an open neighborhood U of $(X, \omega; P)$ in $\mathcal{N}$ such that $\pi(U)$ is an open set around (X, ω) in $\mathcal{M}$.*

Proof Without loss of generality, we will work on $\mathcal{M}^{\mathrm{ord}}(0^n)$. Since each π_k restricted to $\mathcal{N}$ is open by Lemma 2.14 and since π is equal to $\pi_1 \circ \dots \circ \pi_n$, its restriction to $\mathcal{N}$ is open as well. $\square$

3 Holomorphic sections over varieties containing a Teichmüller disk and proof of Theorem 1.3

Throughout this section we will make the following assumption.

Assumption 3.1 Let Γ be a torsionfree finite-index subgroup of the mapping class group. Let C be a complex analytic subvariety of $\text{Teich}_{g,n}/\Gamma$, where $\text{Teich}_{g,n}$ is the Teichmüller space of a closed genus g surface with n punctures, with $3g - 3 + n > 0$. Let $\pi: \mathcal{C}_{g,n} \rightarrow \text{Teich}_{g,n}/\Gamma$ be the universal curve. Suppose additionally that C contains a Teichmüller disk generated by the quadratic differential (X, q) , where X is a Riemann surface and q a quadratic differential on X . Suppose that $\mathcal{Q}$ is the stratum of quadratic differentials to which (X, q) belongs and let $\mathcal{M}$ be the orbit closure of (X, q) in $\mathcal{Q}$.

Lemma 3.2 *Every holomorphic section of the universal curve over C induces a $\text{GL}(2, \mathbb{R})$ -equivariant section of the forgetful map from $\overline{\mathcal{M}(0)}$ to $\mathcal{M}$.*

Proof Let $s: C \rightarrow \mathcal{C}_{g,n}$ be a holomorphic section of the restriction of π to $\pi^{-1}(C)$. Let $\iota: \mathbb{D} \rightarrow C$ be the inclusion of the Teichmüller disk into C . The inclusion is an isometry with respect to the underlying Kobayashi hyperbolic metrics. Since $\iota = \pi \circ s$ and π and s are contractions in the Kobayashi hyperbolic metrics, it follows that s restricted to the embedded Teichmüller disk in C is an isometry in the Kobayashi metrics. Therefore, $s(\iota(\mathbb{D}))$ is a Teichmüller disk in $\mathcal{C}_{g,n}$.

Sufficiently close Riemann surfaces X_1 and X_2 contained in $\iota(\mathbb{D})$ are joined by a dilatation-minimizing homeomorphism given by a geodesic in $\iota(\mathbb{D})$. By Teichmüller's theorem, this homeomorphism is unique up to pre- and post-composition with a conformal automorphism. However, since Γ is torsionfree there are no such automorphisms in Γ that fix X_1 or X_2 . If γ is the geodesic from X_1 to X_2 in $\iota(\mathbb{D})$, then $s(\gamma)$ is a geodesic of the same length in $\mathcal{C}_{g,n}$ and hence corresponds to a homeomorphism with the same dilatation. By uniqueness this path must correspond to the same homeomorphism and so if Teichmüller geodesic flow along (X_1, q_1) produces the geodesic γ , Teichmüller geodesic flow along $(s(X), q)$ produces $s(\gamma)$.

Let $\tilde{s}: \mathcal{M} \rightarrow \overline{\mathcal{M}(0)}$ be the section of the forgetful map from $\overline{\mathcal{M}(0)}$ to $\mathcal{M}$ given by sending a quadratic differential (X, q) to $(s(X), q)$. The argument above shows that this map is $\text{GL}(2, \mathbb{R})$ -equivariant on Teichmüller disks since it is equivariant under complex scalar multiplication and Teichmüller geodesic flow. Since $\mathcal{M}$ is foliated by $\text{GL}(2, \mathbb{R})$ -invariant Teichmüller disks the claim follows. $\square$

Remark 3.3 The proof of [Lemma 3.2](#) only uses the hypothesis that Γ is torsionfree, not that it is finite-index.

Proof of Theorem 1.3 Let s be a holomorphic section of the restriction of π to $\pi^{-1}(C)$. Since it is a continuous section, it follows that $s(C)$, and hence $\tilde{s}(\mathcal{M})$, is closed. By [Lemma 3.2](#), $\tilde{s}(\mathcal{M})$ is closed and $\mathrm{GL}(2, \mathbb{R})$ -invariant and therefore it is an affine invariant submanifold by Eskin, Mirzakhani and Mohammadi [\[9\]](#). Notice that this application of Eskin, Mirzakhani and Mohammadi uses the fact that Γ is finite-index since this implies that the Lebesgue measure of the collection of unit-area half-translation surfaces in $\mathcal{M}$ is finite. Since $\tilde{s}(\mathcal{M})$ is an affine invariant submanifold that does not coincide with $\mathcal{M}(0)$, it follows that the point that s marks above X is a periodic point, zero or pole of (X, q) . $\square$

Remark 3.4 The same proof shows that measurable equivariant sections of the forgetful map from $\overline{\mathcal{M}}(0)$ to $\mathcal{M}$ only mark periodic points, zeros or poles. In the measurable setting, the section is used to push forward Lebesgue measure to a measure on $\overline{\mathcal{M}}(0)$, which must be Lebesgue measure on an affine invariant submanifold by Eskin and Mirzakhani [\[8\]](#). The details are omitted.

4 Explicit translation surfaces in every component of every stratum

We will construct explicit generic translation surfaces in each connected component of every stratum of abelian differentials. The connected components were classified in [\[15\]](#).

Theorem 4.1 (Kontsevich and Zorich [\[15\]](#)) *All strata are connected except:*

- For $g > 3$, $\mathcal{H}(2g - 2)$ has three connected components characterized by odd spin, even spin and hyperellipticity.
- For odd $g > 3$, $\mathcal{H}(g - 1, g - 1)$ has three connected components characterized by odd spin, even spin and hyperellipticity.
- For even $g > 3$, $\mathcal{H}(g - 1, g - 1)$ has two connected components characterized by hyperellipticity and nonhyperellipticity.
- For $g > 3$, $\mathcal{H}(2k_1, \dots, 2k_n)$ has two connected components characterized by odd and even spin (excluding the case $\mathcal{H}(g - 1, g - 1)$ for odd $g > 3$, which, as mentioned above, has three components).
- $\mathcal{H}(4)$ and $\mathcal{H}(2, 2)$ each have two connected components, a hyperelliptic and an odd one.

To distinguish which connected component of a stratum a specific translation surface belongs to, we will use the following criterion.

Theorem 4.2 [15, Corollary 2] *Let $\mathcal{H}$ be a stratum of abelian differentials. For each connected component C of the minimal stratum there is a unique component of $\mathcal{H}$ that contains C in its closure.*

We are now in a position to create horizontally and vertically periodic translation surfaces in each component of each stratum. First, we establish a convention:

Convention for figures We will often use polygons, all of whose edges will be vertical or horizontal, to represent translation surfaces using the following two conventions. The edge of a polygon will mean a line segment in the boundary of the polygon that connects two vertices and has no vertex in its interior.

- (1) The intersection of a dotted line and an edge is a vertex of the polygon.
- (2) If a pair of unmarked vertical (resp. horizontal) edges contain interior points that can be connected by a horizontal (resp. vertical) line that lies in the interior of the polygon then they are identified.

Under this convention the two translation surfaces in [Figure 1](#) are identical.

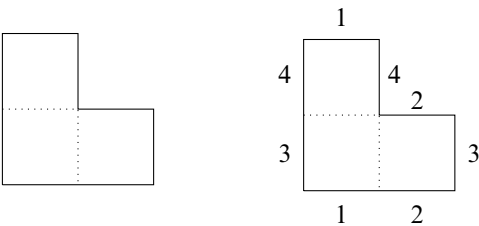


Figure 1: Equivalent representations of the same translation surface.

Using separatrix diagrams, Kontsevich and Zorich produce surfaces that belong to each component of the minimal stratum; see [15, Figure 4]. The surfaces are represented as translation surfaces in [Figure 2](#).

Proposition 4.3 (genericity criterion) *Suppose that (X, ω) is a translation surface in a component $\mathcal{H}$ of a stratum of abelian differentials of genus g surfaces with n zeros and no marked points. If (X, ω) has $g + n - 1$ horizontal cylinders whose moduli satisfy no rational linear relation, then (X, ω) is generic.*

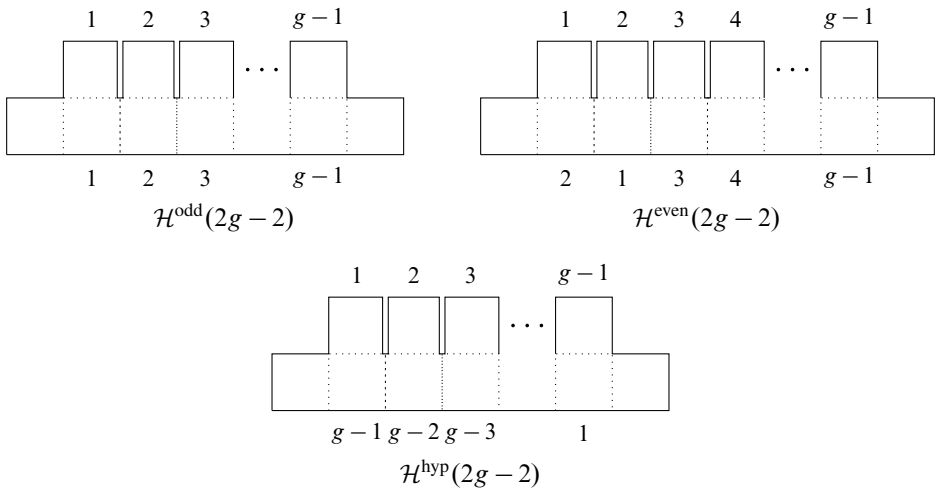


Figure 2: Surfaces in each component of the minimal stratum.

Proof Let $\mathcal{M}$ be the orbit closure of (X, ω) . It suffices to show that $\mathcal{M}$ coincides with a component of $\mathcal{H}$. By Wright [29, Corollary 3.4] — see our Theorem 2.9 — since the moduli satisfy no rational linear relation the tangent space of $\mathcal{M}$ at (X, ω) includes $\{\gamma_1^*, \dots, \gamma_{g+n-1}^*\}$, where the γ_i are core curves oriented from left to right of the horizontal cylinders and γ_i^* denotes the dual cohomology class under the intersection pairing. Recall that the identification of $T_{(X, \omega)}\mathcal{M}$ with a subspace of $H^1(X, \Sigma; \mathbb{C})$ for Σ the zero set of ω was described in Section 2.2. The dual cohomology classes span a complex vector space of dimension $g + n - 1$.

Let $p: T_{(X, \omega)}\mathcal{M} \rightarrow H^1(X, \mathbb{C})$ be the projection from relative to absolute cohomology. By Avila, Eskin and Möller [4] — see our Theorem 2.5 — the image of the projection is a complex symplectic vector space. The kernel of the projection has (complex) dimension at most $n - 1$. Since the projection of $\{\gamma_1^*, \dots, \gamma_{g+n-1}^*\}$ spans an isotropic subspace, which has dimension at most g , it follows that the kernel of p has dimension exactly $n - 1$ and that the projection of $\{\gamma_1^*, \dots, \gamma_{g+n-1}^*\}$ spans a Lagrangian subspace. Since $p(T_{(X, \omega)}\mathcal{M})$ is complex symplectic it follows that p is a surjection with maximal-dimensional kernel. It follows that $T_{(X, \omega)}\mathcal{M}$ is isomorphic to $H^1(X, \Sigma; \mathbb{C})$ and hence that $\mathcal{M}$ has full dimension. Since $\mathcal{M}$ is open and closed it must coincide with a component of $\mathcal{H}$. $\square$

Generic surfaces in $\mathcal{H}^{\text{hyp}}(2g - 2)$ and $\mathcal{H}^{\text{hyp}}(g - 1, g - 1)$ It is straightforward to verify that the translation surfaces in Figure 3 are in the indicated components; see

for example [2, Section 2]. The genericity criterion (Proposition 4.3) implies that the translation surfaces are generic provided that all moduli of horizontal cylinders satisfy no rational linear relation.

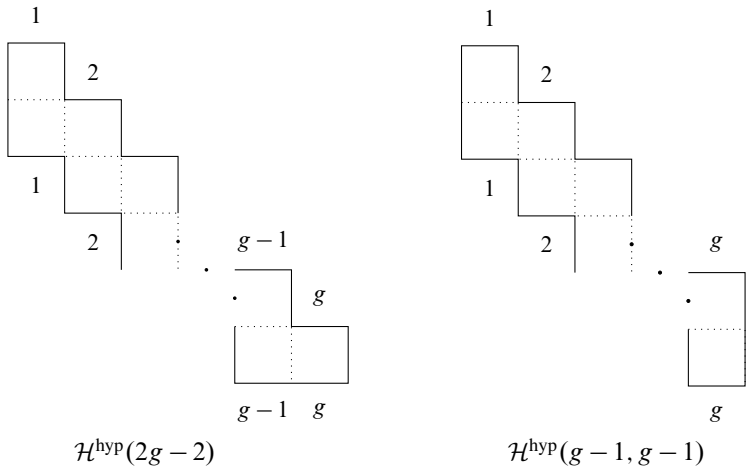


Figure 3: Hyperelliptic translation surfaces.

Generic surfaces in even components of strata Proceed as follows to find a surface in $\mathcal{H} := \mathcal{H}^{\text{even}}(2k_1, \dots, 2k_n)$. Start with the surface in Figure 4, set $g = 1 + \sum_i k_i$ and $S := \{a_0, a_{k_1}, a_{k_1+k_2}, \dots, a_{k_1+\dots+k_{n-1}}\}$. Then collapse every saddle connection in $\{a_1, \dots, a_{g-2}\} \setminus S$. The resulting surface belongs to the even component since further collapsing the saddle connections in $S \setminus \{a_0\}$ gives a path in the stratum whose endpoint is a surface in the even minimal component. (Specifically it will be the surface in the top right of Figure 2 after we perform a half Dehn-twist in horizontal cylinders C_1 and C_2 , which has the effect of changing the labels from the order $(1, 2)$ (resp. $(3, 4)$) on the top boundary of C_1 (resp. C_2) to $(2, 1)$ (resp. $(4, 3)$).)

We want to specify a translation surface (X, ω) in $\mathcal{H}$ constructed by the process described in the preceding paragraph. To do so, we must first specify the lengths of the saddle connections in Figure 4. Let $(\alpha_i)_{i=0}^{g-2} \cup (x_i)_{i=1}^{g+1} \cup (y_i)_{i=0}^{g-1}$ be a collection of positive real numbers with the property that all rational linear relations satisfied by these numbers are generated by the relations $x_1 = x_4$ and $x_2 = x_3$. Now we will set the horizontal saddle connections labeled by an integer $i \in \{1, \dots, g+1\}$ to have length x_i , the heights of the cylinders C_i (resp. C) to have length y_i for $i > 0$ (resp. length y_0) and the saddle connections labeled a_i to have length 0 if they were collapsed by the procedure described in the previous paragraph and α_i otherwise.

By construction the moduli of the vertical (resp. horizontal) cylinders of (X, ω) satisfy no rational linear relation and so by the genericity criterion (Proposition 4.3), (X, ω) is generic.

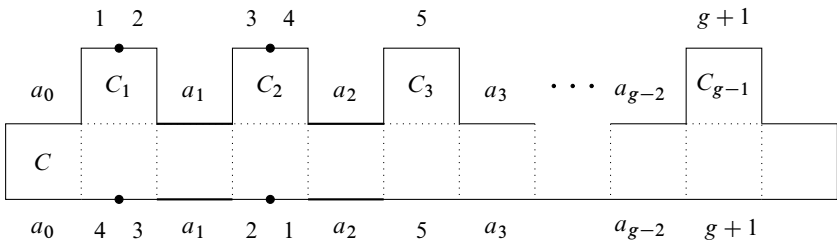


Figure 4: $\mathcal{H}^{\text{even}}(2, \dots, 2)$ with labeled horizontal cylinders.

The following lemma will be needed in the final two sections in order to apply the main technical result of the paper, Lemma 5.5. A reader who seeks to understand the motivation for proving the following result is encouraged to read the statement of Lemma 5.5 first.

Lemma 4.4 *If $g > 3$ then on (X, ω) there are at least two $\mathcal{H}$ -equivalence classes of vertical cylinders $\mathcal{D}_1$ and $\mathcal{D}_2$ such that $C \cap \mathcal{D}_i$ is connected for $i = 1, 2$ and such that the complement of $\mathcal{D}_1 \cup \mathcal{D}_2$ in C consists of two disjoint rectangles of different horizontal lengths.*

Remark 4.5 In the special case that the complement of $\mathcal{D}_1 \cup \mathcal{D}_2$ in C is connected, we will take one rectangle to be an “empty rectangle” whose horizontal length is zero.

Remark 4.6 The restriction that $g > 3$ is natural since when $g = 3$ the even components of strata coincide with hyperelliptic components.

Proof Let D_i be the unique vertical cylinder passing through C_i for $i \in \{2, \dots, g+1\}$. Let D_1 (resp. D_2) be the vertical cylinder that passes through saddle connection 1 (resp. 2). Finally, let A_i be the vertical cylinder passing through the saddle connection labeled a_i (provided that this saddle connection has not been collapsed in the construction of (X, ω)).

By Example 2.8, two cylinders are $\mathcal{H}$ -equivalent if and only if their core curves are homologous. This shows that the $\mathcal{H}$ -equivalence classes of vertical cylinders are

$$\{D_1, D_2\}, \quad \{D_3\}, \quad \dots, \quad \{D_{g-1}\}, \quad \{A_1\}, \quad \{A_0, A_2, \dots, A_{g-2}\}.$$

If $g > 4$, then let $\mathcal{D}_1 := \{D_3\}$ and $\mathcal{D}_2 = \{D_4\}$. These are two $\mathcal{H}$ -equivalence classes that have connected intersection with C . The complement of $\mathcal{D}_1 \cup \mathcal{D}_2$ in C consists of two rectangles, one of which has length α_3 or zero (depending on whether or not saddle connection a_3 was collapsed) and the other of which is a positive real number that is not rationally linearly related to α_3 by construction. Suppose now that $g = 4$. Suppose first that the horizontal saddle connection labeled a_1 is uncollapsed. Set $\mathcal{D}_1 = \{A_1\}$ and $\mathcal{D}_2 = \{D_3\}$. The complement of $\mathcal{D}_1 \cup \mathcal{D}_2$ in C is a disjoint union of two rectangles. These rectangles have horizontal lengths $|a_2| + x_1 + x_2$ and $|a_0| + x_1 + x_2$, where $|\cdot|$ denotes flat length. These two lengths are distinct by construction. Suppose now that a_1 is collapsed. Set $\mathcal{D}_1 = \{D_1, D_2\}$ and $\mathcal{D}_2 = \{D_3\}$. Again the complement of $\mathcal{D}_1 \cup \mathcal{D}_2$ in C is a disjoint union of two rectangles whose lengths are $|a_0|$ and $|a_2|$, which again are distinct by construction. $\square$

Generic surfaces in remaining components To find generic surfaces in all other connected components of the remaining strata we glue together copies of the surfaces in Figure 5 along the horizontal cylinders that intersect all vertical cylinders. By the genericity criterion (Proposition 4.3) whenever the vertical cylinders have rationally unrelated moduli this surface is generic.

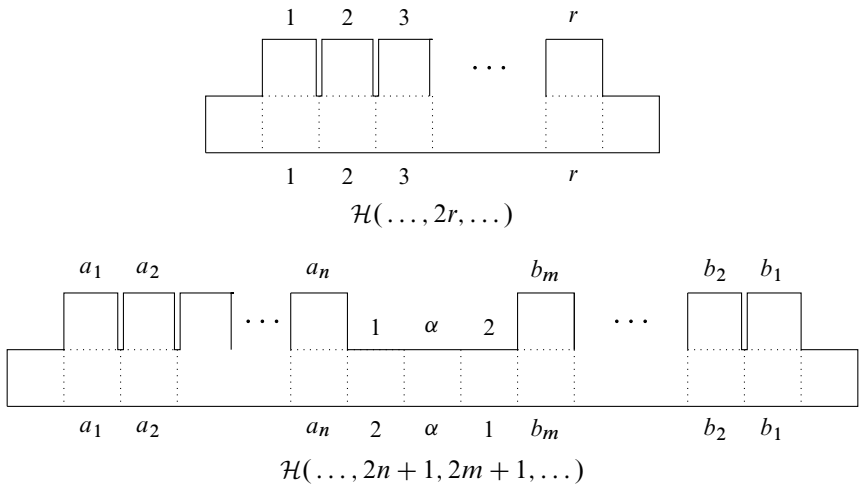


Figure 5: Surfaces in $\mathcal{H}^{\text{odd}}$, $\mathcal{H}^{\text{nonhyp}}$ and connected strata.

Definition 4.7 Each of the nonhyperelliptic surfaces constructed in this subsection contain a horizontal cylinder that intersects every vertical cylinder. This cylinder will be called the central horizontal cylinder.

Lemma 4.8 *Let (X, ω) be a surface constructed by the above process. Let C be the central horizontal cylinder. There are $\mathcal{H}$ -equivalence classes of cylinders $\mathcal{D}_1$ and $\mathcal{D}_2$ that contain one cylinder intersecting C exactly once. The complement of $\mathcal{D}_1$ and $\mathcal{D}_2$ in C consists of two disjoint rectangles whose horizontal lengths are distinct for generic choices of lengths of the horizontal saddle connections.*

Proof Suppose first that the (X, ω) has two zeros of even order. Then the surface contains two copies of the surface in Figure 5, top. It is sufficient to take the two vertical cylinders that pass through the horizontal saddle connection labeled 1. Similarly, if there are four zeros of odd order we have two copies of the surface in Figure 5, bottom, and we take the two vertical cylinders that pass through the horizontal saddle connection α .

If there is one zero of even order and two zeros of odd order, then we have a surface like the one in Figure 5, top, and one like the one in Figure 5, bottom, and we take the vertical cylinder passing through the saddle connection labeled 1 on the first and the one passing through the horizontal saddle connection labeled α on the second.

If there is only one zero, then it is of even order and we take the two vertical cylinders that pass through 1 and 2. If there are only two zeros and both are of odd order then we take the vertical cylinder passing through α and the vertical cylinder passing through a_1 or b_1 (whichever exists).

As in Lemma 4.4, two cylinders are $\mathcal{H}$ -equivalent if and only if their core curves are homologous. All cylinders selected in this proof have no distinct $\mathcal{H}$ -equivalent cylinders. (To see this, note that the constructed surface is made up of vertically periodic subsurfaces as in Figure 5. On each subsurface no two vertical cylinders have homologous core curves. Moreover, if two vertical cylinders have homologous core curves then they must belong to the same subsurface.)

All cylinders selected in this proof intersect C exactly once, and therefore it is immediate that $C \setminus (\mathcal{D}_1 \cup \mathcal{D}_2)$ is a union of two disjoint rectangles. By construction, one of these rectangles completely contains a vertical cylinder. By altering the height (ie horizontal length) of this cylinder we see that the two rectangles do not have generically identical horizontal lengths. $\square$

5 Marked points in cylinders

In this section we will prove results about marked points and cylinders that form the technical core of the paper. We will make the following standing assumption:

Assumption 5.1 $\mathcal{M}$ is an affine invariant submanifold and (X, ω) is a translation surface whose $\mathrm{GL}(2, \mathbb{R})$ orbit closure is $\mathcal{M}$.

Lemma 5.2 *Let P be a collection of distinct points on (X, ω) and suppose that $\mathcal{M}'$ is the orbit closure of $(X, \omega; P)$. If $\mathcal{C}$ is an $\mathcal{M}$ -equivalence class of cylinders on (X, ω) , then $\mathcal{C}'$ is an $\mathcal{M}'$ -equivalence class, where $\mathcal{C}'$ contains the cylinders in $\mathcal{C}$ divided into subcylinders by the points in P .*

Proof First we will show that any two cylinders in $\mathcal{C}'$ are $\mathcal{M}'$ equivalent. Let C_i be two cylinders in $\mathcal{C}'$ and let γ_i be their core curves for $i = 1, 2$. By assumption, there is a neighborhood U of (X, ω) in $\mathcal{M}$ on which γ_1 and γ_2 are collinear. Let U' be a preimage of the U in $\mathcal{M}'$ on which C_1 and C_2 persist as cylinders. In this neighborhood, γ_1 and γ_2 must remain collinear and hence $\mathcal{M}'$ -equivalent.

It remains to show that if C_1 and C_2 are $\mathcal{M}'$ -equivalent cylinders with core curves γ_1 and γ_2 , then γ_1 and γ_2 must be collinear on a neighborhood of (X, ω) in $\mathcal{M}$. By Lemma 2.16 there is a neighborhood of $(X, \omega; P)$ in $\mathcal{M}'$ that projects to an open neighborhood of (X, ω) in $\mathcal{M}$. Let U' be such a neighborhood of $(X, \omega; P)$ in $\mathcal{M}'$ where γ_1 and γ_2 are collinear, and let U be its image under the forgetful map. Since γ_1 and γ_2 are collinear on U' they are collinear on U and hence are homotopic to core curves of $\mathcal{M}$ -equivalent cylinders. $\square$

Definition 5.3 Given a cylinder C in a translation surface we say that the height of the cylinder is the distance between the two boundaries in the flat metric. Suppose that p is a marked point contained in a cylinder C . Let h_C be the height of the cylinder, and let h_p be the distance from the point to one of the two boundary curves of C . We say that p lies at rational height in C if the ratio h_p/h_C is rational.

Lemma 5.4 (rational height lemma) *Let $\mathcal{C}$ be an $\mathcal{M}$ -equivalence class of cylinders. Suppose that any two cylinders in $\mathcal{C}$ have a rational ratio of moduli. If a periodic point belongs to the interior of a cylinder in $\mathcal{C}$ then it lies at rational height.*

Proof Let p be a periodic point contained in the interior of a cylinder in $\mathcal{C}$. Let $\mathcal{M}'$ be the orbit closure of $(X, \omega; p)$. Let $\mathcal{C}'$ be the collection of subcylinders on $(X, \omega; p)$ into which $\mathcal{C}$ is divided. By Lemma 5.2, $\mathcal{C}'$ is an $\mathcal{M}'$ equivalence class. Let $\sigma_{\mathcal{C}'}$ be the standard shear on $\mathcal{C}'$. Since the cylinders in $\mathcal{C}$ have a rational ratio of moduli, the flow along σ_C is periodic. Suppose, for a contradiction, that p does not have rational height. In this case, the flow along $\sigma_{\mathcal{C}'}$ is not periodic and so the orbit closure

of $(X, \omega; p)$ contains $(X, \omega; q)$, where q is any point in C along the core curve of C that intersects p .

Let γ_1 and γ_2 be the two core curves of the cylinders into which p divides C . Since p may be moved along the core curve of C while fixing all cylinders in (X, ω) , it follows that the tangent space of $\mathcal{M}'$ at $(X, \omega; p)$ contains the deformation $\gamma_1^* \smile \gamma_2^*$. Let U' be a neighborhood as in Lemma 2.16 of $(X, \omega; p)$ in $\mathcal{M}'$ on which the cylinders in $\mathcal{C}$ persist. This neighborhood projects to a neighborhood U of (X, ω) in $\mathcal{M}$. Since the tangent space contains the deformation $\gamma_1^* \smile \gamma_2^*$ the fiber of the projection from $\mathcal{M}'$ to $\mathcal{M}$ that forgets marked points has real dimension at least one. Therefore, the dimension of $\mathcal{M}'$ is strictly larger than the dimension of $\mathcal{M}$, which contradicts the assumption that p is a periodic point. $\square$

Lemma 5.5 *Let (X, ω) be a generic translation surface in an affine invariant submanifold $\mathcal{M}$. Let C be a horizontal cylinder, and let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two vertical distinct $\mathcal{M}$ -equivalence classes of cylinders such that:*

- (1) *The intersection of $\mathcal{D}_i$ with the interior of C is connected and nonempty for $i = 1, 2$.*
- (2) *Any cylinder in the $\mathcal{M}$ -equivalence class $\mathcal{C}$ of C has a modulus that is an integer multiple of the modulus of C .*

If p is an $\mathcal{M}$ -periodic point in the interior of C , then up to relabeling $\mathcal{D}_1$ and $\mathcal{D}_2$, the point p is at the center of the rectangle given by the intersection of $\mathcal{D}_1$ and C . Furthermore, removing $\mathcal{D}_1$ and $\mathcal{D}_2$ divides C into two rectangles of equal size.

Proof Let $\mathcal{M}'$ be the orbit closure of $(X, \omega; p)$. Let $\mathcal{C}$ be the collection of horizontal cylinders equivalent to C on (X, ω) .

For simplicity we begin by applying an element of $\mathrm{GL}(2, \mathbb{R})$ so that C has unit height. By the rational height lemma (Lemma 5.4) the marked point p lies at rational height h in C . Since we have normalized the height of C to be one this means that h is rational; in particular suppose that $h = n/m$, where n and m are coprime positive integers.

Part 1 We may assume that p belongs to $\mathcal{D}_1$ Suppose not. By Lemma 5.2, since p is not contained in $\mathcal{D}_1$, the standard shear $a_{\mathcal{D}_1}$ is tangent to $\mathcal{M}'$. Traveling in the $a_{\mathcal{D}_1}$ direction in $\mathcal{M}'$ from $(X, \omega; p)$ widens $\mathcal{D}_1$ while fixing the part of the translation surface (and marked point) in the complement $\mathcal{D}_1$. Travel in this direction until the intersection of $\mathcal{D}_1$ and C accounts for at least $(m-1)/m$ proportion of the area of C . Let (Y, η) be the new translation surface.

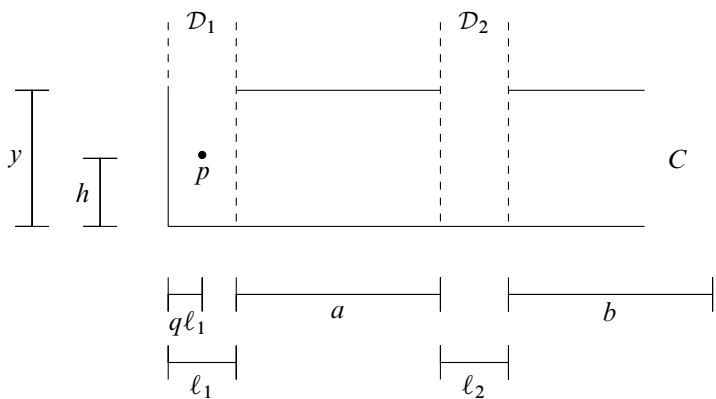


Figure 6: [Lemma 5.5](#) shows that $a = b$ and that, after scaling so C has unit height, $q = h = \frac{1}{2}$.

Part 1a (Y, η) may be taken to be generic in $\mathcal{M}$ We formed (Y, η) by traveling in the $a_{\mathcal{D}_1}$ direction from (X, ω) in $\mathcal{M}$. Let ℓ be the segment joining (X, ω) to (Y, η) in $\mathcal{M}$. Each proper affine invariant submanifold (of which there are only countably many by Eskin, Mirzakhani and Mohammadi [9]) contained in $\mathcal{M}$ intersects ℓ in a closed set. If a neighborhood U of (Y, η) had the property that every point in $U \cap \ell$ was contained in some proper affine invariant submanifold, then by the Baire category theorem there would be a proper affine invariant submanifold $\mathcal{N}$ such that $\mathcal{N} \cap \ell$ had interior in $U \cap \ell$. Since affine invariant submanifolds are linear this would imply that all of ℓ was contained in $\mathcal{N}$, which contradicts the fact that ℓ contains a translation surface (X, ω) that is generic in $\mathcal{M}$. Therefore, we may assume that the point (Y, η) was chosen to be generic in $\mathcal{M}$.

Part 1b The hypotheses of the lemma continue to hold on (Y, η) and the fiber of $\mathcal{M}'$ over (Y, η) contains $(Y, \eta; p)$ where p belongs to $\mathcal{D}_1$ Traveling in the $a_{\mathcal{D}_1}$ direction keeps p fixed in the complement of $\mathcal{D}_1 \cup \mathcal{D}_2$ and keeps the heights of the cylinders in C constant. By Wright [29, Theorem 1.9], the ratio of lengths of core curves of cylinders in C are constant, and so the condition on moduli of cylinders in C continues to hold.

Letting C' be the collection of cylinders on $(Y, \eta; p)$ that project to cylinders in C we have that C' is an $\mathcal{M}'$ -equivalence class by [Lemma 5.2](#). Travel from $(Y, \eta; p)$ in the $\sigma_{C'}$ direction until one complete Dehn twist has been performed in C . The resulting unmarked translation surface is (Y, η) since all cylinders in C have moduli that are integer multiples of the modulus of C . Since $\mathcal{D}_1 \cap C$ is connected and accounts for

at least $(m-1)/m$ of the area of C , it follows that the marked point p now belongs to $\mathcal{D}_1$ and we have shown that $(Y, \eta; p)$ belongs to $\mathcal{M}'$.

Part 1c If the conclusion of the lemma holds on (Y, η) , it does so on (X, ω) also

To pass from (X, ω) to (Y, η) , we first traveled along $a_{\mathcal{D}_1}$ and then traveled along $\sigma_{C'}$ to perform one Dehn twist in C . If the conclusion of the lemma holds on (Y, η) , then traveling in the opposite direction along $\sigma_{C'}$ to perform the opposite Dehn twist in C moves p to the midpoint of the rectangle $\mathcal{D}_2 \cap C$. Let $\mathcal{D}'_1$ be the collection of subcylinders into which the cylinders in $\mathcal{D}_1$ are divided by p . Traveling along $a_{\mathcal{D}'_1}$ back to (X, ω) keeps the complement of $\mathcal{D}_1$ fixed and so the conclusion of the lemma held on $(X, \omega; p)$, as desired. Therefore, we may suppose without loss of generality (by replacing (X, ω) with (Y, η)) that p belongs to $\mathcal{D}_1$.

Part 2 Determining the position of p Suppose now that p belongs to $\mathcal{D}_1$. As before, $a_{\mathcal{D}_2}$ is tangent to $\mathcal{M}'$. Travel from $(X, \omega; p)$ in the $a_{\mathcal{D}_2}$ direction in $\mathcal{M}'$ until the intersection of $\mathcal{D}_2$ and C accounts for at least $(m-1)/m$ proportion of the area of C . Without loss of generality, we may replace $(X, \omega; p)$ with the resulting marked translation surface.

Let ℓ_i be the horizontal length of the rectangle $C \cap \mathcal{D}_i$ for $i = 1, 2$. The complement of $\mathcal{D}_1 \cup \mathcal{D}_2$ in C is two disjoint rectangles. Let a (resp. b) be the horizontal length of the rectangle to the right (resp. left) of $\mathcal{D}_1 \cap C$; see Figure 6. Let ℓ be the length of the core curve of C . Let $q \in [0, 1]$ be chosen so that p is a distance of $q\ell_1$ from the left boundary of $\mathcal{D}_1 \cap C$. Let $\mathcal{D}'_1$ be the collection of subcylinders into which p divides $\mathcal{D}_1$ on $(X, \omega; p)$.

Travel in the $a_{\mathcal{D}'_1}$ direction from $(X, \omega; p)$ so that the length of the core curve of C increases by s and then travel in the $\sigma_{C'}$ direction to perform exactly one Dehn twist in C (we will assume that the core curves of horizontal cylinders are oriented from left to right; this specifies that we are applying the matrix $\begin{pmatrix} 1 & \mathrm{Mod}(C) \\ 0 & 1 \end{pmatrix}^{-1}$) to the cylinders in $\mathcal{C}'$, where $\mathrm{Mod}(C)$ denote the modulus of C). The distance of the marked point from the left-hand boundary of $\mathcal{D}_2 \cap C$ is

$$h(\ell + s) - (1 - q)(\ell_1 + s) - a.$$

Traveling back along the $a_{\mathcal{D}'_1}$ direction returns to the unmarked surface (X, ω) while leaving the position of the marked point fixed in the complement of $\mathcal{D}_1$. Since p is a periodic point, the fiber of the forgetful map from $\mathcal{M}'$ to $\mathcal{M}$ over (X, ω) is finite.

Therefore, $h(\ell + s) - (1 - q)(\ell_1 + s) - a$ is constant as a function of s . In other words, $h = (1 - q)$. If we sheared C in the other direction we would have by symmetry that $h = q$ and so $q = h = \frac{1}{2}$. By symmetry, after shearing the marked point into $\mathcal{D}_2$ the distance from the left-hand boundary of $\mathcal{D}_2 \cap C$ is $\frac{1}{2}\ell_2$, ie

$$\frac{1}{2}(\ell - \ell_1) - a = \frac{1}{2}\ell_2.$$

Since $\ell = \ell_1 + a + \ell_2 + b$ we see that $a = b$, as desired. $\square$

6 Proof of Theorem 1.1

Throughout this section, we make the following assumption.

Assumption 6.1 Let $\mathcal{M}$ be a nonempty affine invariant submanifold in $\mathcal{H}(0^n)$, where $\mathcal{H}$ is an unmarked stratum of abelian differentials. Suppose that $\mathcal{M}$ contains marked points on a translation surface (X, ω) that is generic in $\mathcal{H}$ and that is one of the surfaces constructed in Section 4. Finally, suppose after passing to a finite cover that the marked points are labeled as $\{p_1, \dots, p_n\}$. Let $\pi_k: \mathcal{H}^{\text{ord}}(0^n) \rightarrow \mathcal{H}^{\text{ord}}(0^{n-1})$ be the map that forgets the k^{th} marked point for $k \in \{1, \dots, n\}$.

Theorem 6.2 *Periodic points exist on (X, ω) if and only if $\mathcal{H}$ is hyperelliptic, in which case they are Weierstrass points.*

Theorem 6.3 *If $n \geq 2$ and $(\pi_k)_*\mathcal{M} = \mathcal{H}^{\text{ord}}(0^{n-1})$ for all $k \in \{1, \dots, n\}$ then $\mathcal{H}$ is hyperelliptic, $n = 2$, and the fiber of $\mathcal{M}$ over (X, ω) contains all pairs of distinct points exchanged by the hyperelliptic involution.*

If $\mathcal{H}$ is hyperelliptic, then given a pair $\{i, j\}$ of integers in $\{1, \dots, n\}$, let $\mathcal{H}_{ij}$ denote the subset of $\mathcal{H}(0^n)$ where p_i and p_j are exchanged by the hyperelliptic involution. If $i = j$, this will mean that p_i is a fixed point of the hyperelliptic involution. We will prove the following strengthening of Theorem 1.1.

Theorem 6.4 *If $\mathcal{M}$ is as in Assumption 6.1 then either $\mathcal{M} = \mathcal{H}(0^n)$ or the stratum $\mathcal{H}$ is hyperelliptic and there is a subset S of pairs of integers in $\{1, \dots, n\}$ such that $\mathcal{M} = \bigcap_{\{i,j\} \in S} \mathcal{H}_{ij}$.*

Proof of Theorem 6.4 given Theorem 6.2 and Theorem 6.3 Fix $\mathcal{H}$ and proceed by induction on n . The $n = 1$ case is Theorem 6.2. Suppose now that $n > 1$ and that $\mathcal{M} \neq \mathcal{H}(0^n)$.

Suppose first that for some $k \in \{1, \dots, n\}$, $(\pi_k)_*\mathcal{M}$ has dimension equal to $\dim \mathcal{M} - 1$. Suppose without loss of generality after relabeling that $k = 1$. By the induction hypothesis, $\mathcal{H}$ is hyperelliptic and $(\pi_1)_*\mathcal{M} = \bigcap_{\{i,j\} \in S} \mathcal{H}_{ij}(0^{n-1})$, where S is some subset of pairs of integers in $\{2, \dots, n\}$. It follows that $\mathcal{M}$ is contained in $\bigcap_{\{i,j\} \in S} \mathcal{H}_{ij}(0^n)$ and therefore coincides with it since both manifolds are connected, closed and of the same dimension.

Suppose now that for all $k \in \{1, \dots, n\}$, $(\pi_k)_*\mathcal{M}$ has the same dimension as $\mathcal{M}$ and that it does not coincide with $\mathcal{H}(0^{n-1})$. By the induction hypothesis, for each k , there is a subset S_k of pairs of integers in $\{1, \dots, n\} \setminus \{k\}$ such that $(\pi_k)_*\mathcal{M} = \bigcap_{\{i,j\} \in S_k} \mathcal{H}_{ij}$. The number of elements of S_k is exactly the codimension of $(\pi_k)_*\mathcal{M}$ in $\mathcal{H}(0^{n-1})$. Suppose without loss of generality that $\{1, \ell_1\}$ is contained in S_{ℓ_2} . Then this pair cannot be contained in S_1 and so $S_1 \cup S_{\ell_2}$ contains at least as many elements as the codimension of $\mathcal{M}$ in $\mathcal{H}(0^n)$. Hence, $\mathcal{M} = \bigcap_{\{i,j\} \in S_1 \cup S_{\ell_2}} \mathcal{H}_{ij}(0^n)$ since both manifolds are connected, closed and of the same dimension.

The only case that remains is when $(\pi_k)_*\mathcal{M} = \mathcal{H}^{\text{ord}}(0^{n-1})$ for all $k \in \{1, \dots, n\}$, which follows from [Theorem 6.3](#). $\square$

Proof of Theorem 1.1 Let (Y, η) be a generic translation surface in a component $\mathcal{H}$ of a stratum of abelian differentials. Let P be a nongeneric collection of n points on (Y, η) and let $\mathcal{M}$ be the orbit closure of $(Y, \eta; P)$ in $\mathcal{H}(0^n)$. By assumption, $\mathcal{M} \neq \mathcal{H}(0^n)$. By [Lemma 2.14](#), for a generic surface (X, ω) as in [Assumption 6.1](#), there is a collection of points Q — which are necessarily nongeneric — such that $(X, \omega; Q) \in \mathcal{M}$. By [Theorem 6.4](#), it follows that $\mathcal{H}$ is hyperelliptic and that there is a subset S of pairs of integers in $\{1, \dots, n\}$ such that $\mathcal{M} = \bigcap_{\{i,j\} \in S} \mathcal{H}_{ij}$. Since P is nongeneric it follows that it must either contain a Weierstrass point or two points exchanged by the hyperelliptic involution. $\square$

7 Proof of Theorem 6.3

[Assumption 6.1](#) will remain in effect for this section. Assume too that $\mathcal{M}$ is a proper affine invariant submanifold in $\mathcal{H}^{\text{ord}}(0^n)$ where $n \geq 2$ and suppose that $(\pi_k)_*\mathcal{M} = \mathcal{H}^{\text{ord}}(0^{n-1})$ for all $k \in \{1, \dots, n\}$. If $\mathcal{H}$ is nonhyperelliptic then let C be the central horizontal cylinder and if $\mathcal{H}$ is hyperelliptic let C be any horizontal cylinder that intersects two vertical cylinders. By [Lemmas 4.4](#) and [4.8](#), there are two $\mathcal{H}$ -equivalence classes of vertical cylinders $\mathcal{D}_1$ and $\mathcal{D}_2$ whose intersection with C is connected and nonempty.

Lemma 7.1 *There are two marked points. If one marked point lies in a cylinder in $\mathcal{D}_1$ or $\mathcal{D}_2$, the other one lies in that cylinder as well.*

Proof Let $P = \{p_1, \dots, p_{n-1}\}$ be a collection of $n - 1$ points where p_1 lies in $\mathcal{D}_1$ and divides the cylinders in $\mathcal{D}_1$ into subcylinders whose moduli admit no rational homogeneous linear relation. Suppose too that the points $p_2, \dots, p_{n-1}$ belong to $\mathcal{D}_2$ and divide it into vertical subcylinders whose moduli also admit no rational homogeneous linear relation. The genericity criterion (Proposition 4.3) implies that $(X, \omega; P)$ is generic in $\mathcal{H}(0^{n-1})$. By Lemma 2.14, there is a point p_n such that $(X, \omega; P \cup \{p_n\})$ belongs to $\mathcal{M}$.

Let $\mathcal{D}$ be an equivalence class in $\{\mathcal{D}_1, \mathcal{D}_2\}$ that does not contain p_n . Since $\mathcal{D}$ is its own equivalence class in $\mathcal{H}$ and since it is divided into subcylinders whose moduli admit no rational homogeneous linear relation, it follows that each subcylinder in $\mathcal{D}$ may be sheared while fixing the rest of the surface. Phrased differently, in the fiber of $\mathcal{M}$ over (X, ω) any marked point in $\mathcal{D}$ may be moved freely while fixing all other points in P . However, if p_k is a point (for $k \in \{1, \dots, n - 1\}$) that belongs to $\mathcal{D}$, then the fiber of $\pi_k: \mathcal{M} \rightarrow \mathcal{H}^{\text{ord}}(0^{n-1})$ is one-dimensional and, by assumption, $(\pi_k)_* \mathcal{M} = \mathcal{H}^{\text{ord}}(0^{n-1})$. This implies that $\mathcal{M} = \mathcal{H}^{\text{ord}}(0^n)$, which is a contradiction. Therefore, there are no points belonging to $\mathcal{D}$ and so $n = 2$.

For the final statement, suppose, for a contradiction, that $\{p_1, p_2\}$ is a fiber of $\mathcal{M}$ over (X, ω) under the map that forgets marked points and suppose too that p_1 belongs to $\mathcal{D}$ for $\mathcal{D} \in \{\mathcal{D}_1, \mathcal{D}_2\}$, but that p_2 does not. Since the map π_2 that forgets the second point is open by Lemma 2.16, there is a nearby surface $(X, \omega; p'_1, p'_2)$ in $\mathcal{M}$ where p'_1 divides the cylinders in $\mathcal{D}$ into subcylinders whose moduli admit no rational homogeneous linear relation and p'_2 remains outside of $\mathcal{D}$. This contradicts the previous paragraph. $\square$

Because $\pi_2: \mathcal{M} \rightarrow \mathcal{H}(0)$ is a finite holomorphic map, we see that given a point $(X, \omega; p_1, p_2) \in \mathcal{M}$ we move p_1 to a new point p'_1 (at least locally) and there will be a unique nearby point $(X, \omega; p'_1, p'_2) \in \mathcal{M}$. Since the equations that define the affine invariant submanifold $\mathcal{M}$ have real coefficients, if p_1 moves horizontally (resp. vertically), so does p_2 .

Now suppose without loss of generality that p_1 lies at irrational height in $\mathcal{D}_1$ and at irrational height in C . Let p_2 be a point such that $(X, \omega; p_1, p_2) \in \mathcal{M}$. By Lemma 7.1, p_2 must also lie in the interior $\mathcal{D}_1 \cap C$, which is a rectangle. Moving p_1 to the left we see that p_1 reaches the left boundary of $\mathcal{D}_1 \cap C$ at the same moment that p_2 reaches

the vertical boundary. Now reversing direction and moving p_1 to the right we see again that p_1 and p_2 arrive at the vertical boundary of $\mathcal{D}_1$ at the same moment. This implies either that one point lies above the other and both move at the same speed in the same direction (horizontally) or that both points at some point were on opposite boundaries of $\mathcal{D}_1$ and move in opposite directions (horizontally) at the same speed. The first case cannot occur, since if it does we may simply shear the central horizontal cylinder and find two marked points that do not lie above each other and that still move in the same direction at the same speed. The same argument applied to the vertical direction shows that when one point moves at unit speed in the v direction, the other point moves at unit speed in the $-v$ direction. Coupled with the fact that the points arrive at the $\mathcal{D}_1 \cap C$ boundary at the same times we have that when $\mathcal{H}$ is hyperelliptic the two points are exchanged by the hyperelliptic involution.

We will now show that $\mathcal{H}$ must be hyperelliptic. The present situation is pictured in Figure 7. If $\mathcal{H}$ is not hyperelliptic then we may ensure that $a < b$ (by Lemmas 4.4 and 4.8). If p_2 moves to the right at unit speed, then p_1 moves to the left at unit speed and hence p_2 arrives in the interior of $\mathcal{D}_2$ before p_1 , contradicting Lemma 7.1.

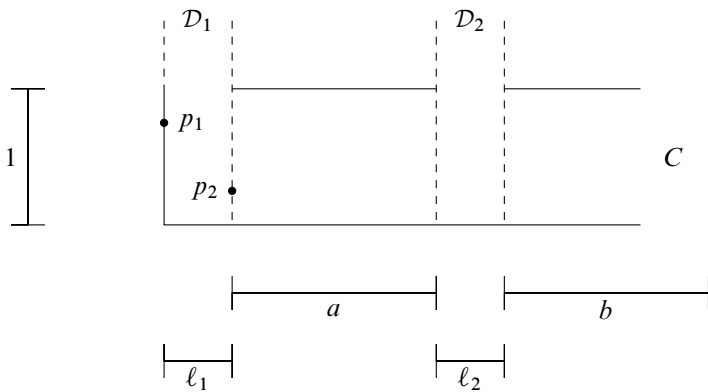


Figure 7: Two marked points in the central horizontal cylinder.

8 Proof of Theorem 6.2

Throughout this section we will make the following assumption.

Assumption 8.1 Let $\mathcal{H}$ be a stratum of abelian differentials and let (X, ω) be a generic translation surface in $\mathcal{H}$ constructed in Section 4. Let $\mathcal{M}$ be an affine invariant

submanifold properly contained in $\mathcal{H}(0)$. By Lemma 2.14, the fiber in $\mathcal{M}$ over (X, ω) is nonempty and any point p in the fiber is a periodic point.

Definition 8.2 A cylinder C is called $\mathcal{H}$ -free if $\{C\}$ is an $\mathcal{H}$ -equivalence class. This is equivalent to the condition that there is no other parallel cylinder C with a core curve that is homologous to the core curve of C .

Proposition 8.3 The periodic points on $\mathcal{H}^{\text{hyp}}(2g-2)$ and $\mathcal{H}^{\text{hyp}}(g-1, g-1)$ are exactly the Weierstrass points.

Proof Let $\mathcal{H}$ be either $\mathcal{H}^{\text{hyp}}(2g-2)$ or $\mathcal{H}^{\text{hyp}}(g-1, g-1)$. The surface (X, ω) is pictured again in Figure 8.

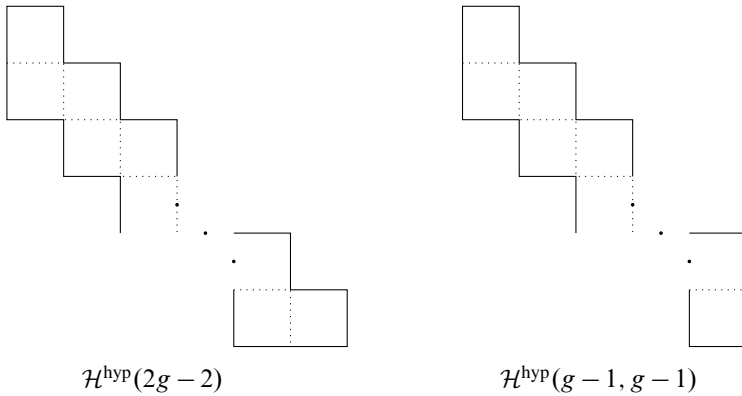


Figure 8: Hyperelliptic translation surfaces.

Each horizontal and vertical cylinder is $\mathcal{H}$ -free. By Lemma 5.5 if p is a periodic point in (X, ω) that lies in the interior of a horizontal cylinder that intersects two vertical cylinders, it is automatically a Weierstrass point. The same holds if it lies in the interior of a vertical cylinder that intersects two horizontal ones. We may therefore assume (up to exchanging each instance of the word “horizontal” for the word “vertical” and vice versa) that p lies in the interior of a horizontal cylinder that intersects only one vertical cylinder and on the boundary of a vertical cylinder. By Lemma 5.2, we may shear this cylinder and remain in $\mathcal{M}$. Shearing the horizontal cylinder so as to perform one complete Dehn twist moves the periodic point into the interior of a vertical cylinder that intersects two horizontal cylinders and so we have that p is a Weierstrass point by Lemma 5.5. $\square$

Assumption 8.4 Assume now that $\mathcal{H}$ is nonhyperelliptic and let C be the central horizontal cylinder in (X, ω) .

Proposition 8.5 A periodic point on (X, ω) must lie on the boundary of C and in the interior of a vertical cylinder that is not $\mathcal{H}$ -free.

Proof Let p be a periodic point in (X, ω) . By Lemmas 4.4, 4.8 and 5.5 the periodic point cannot lie in the interior of C . We will proceed now by cases based on the containment of p in vertical cylinders.

Case 1 p is contained in a vertical cylinder V that is $\mathcal{H}$ -free, is contained in C and only intersects the core curve of C once In this case, Lemma 5.2 implies that we may shear V so as to perform one complete Dehn twist and remain in $\mathcal{M}$. This moves p to a periodic point in the interior of C , which is a contradiction.

Case 2 p is contained in a vertical cylinder V that is $\mathcal{H}$ -free and is contained in C By the previous case, V must intersect the core curve of C at least twice. By construction of the surfaces in Section 4 the situation must be as depicted in Figure 9. The marked point is then contained in a $\mathcal{H}$ -free cylinder (drawn in dashed lines). Using Lemma 5.2 to shear this cylinder to perform one complete Dehn twist we see that p may be moved to a periodic point in the interior of C , which is a contradiction.

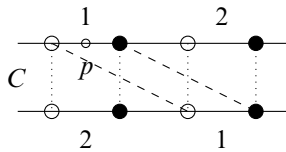


Figure 9: The translation surface in Case 2.

Case 3 p is contained in a $\mathcal{H}$ -free vertical cylinder V that is not contained in C By construction of the surfaces in Section 4, V only intersects two horizontal cylinders, C and H , and core curves intersect exactly once; see Figure 10. Notice that C is $\mathcal{H}$ -free since by construction there is a vertical cylinder V' contained in it. The core curve of V' cannot intersect any other horizontal cylinder and so C is not homologous to (equivalently $\mathcal{H}$ -equivalent to) any other horizontal cylinder. Similarly, H is $\mathcal{H}$ -free since it is not equivalent to C and since applying the matrix $\begin{pmatrix} 1 & 1 \\ t & 0 \end{pmatrix}$, for small t , to V makes H and C nonhorizontal while fixing all other horizontal cylinders (two equivalent cylinders must remain parallel in a neighborhood of the original surface).

Applying [Lemma 5.5](#), with $\mathcal{D}_1 = \{C\}$ and $\mathcal{D}_2 = \{H\}$ we see that if p lies in the interior of V , then there must be a periodic point in C , which is a contradiction. If p does not lie in the interior of V , then by [Lemma 5.2](#) we may shear H to perform one complete Dehn twist while fixing the remainder of the translation surface and remaining in $\mathcal{M}$. This shear moves p to the interior of V and so we are done.

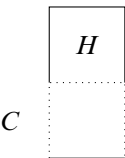


Figure 10: The vertical cylinder V .

Case 4 p is contained in a vertical cylinder that is not contained in C and that is not $\mathcal{H}$ -free By construction of the surfaces in [Section 4](#), $\mathcal{H}$ is an even component of a stratum of abelian differentials. Moreover, either p is contained on the boundary of a vertical cylinder, as in [Figure 11](#), or is contained in one of the cylinders passing through the saddle connections labeled $\{1, \dots, 4\}$ in [Figure 11](#).

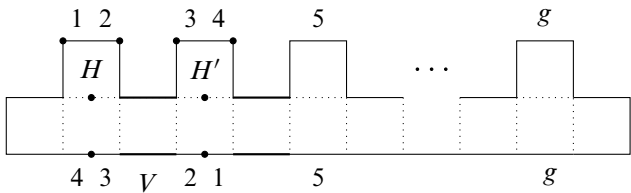


Figure 11: Points on the boundary of C and on the boundary of a vertical cylinder.

Let H and H' be the indicated horizontal cylinders in [Figure 11](#), which are $\mathcal{H}$ -free. Suppose without loss of generality that p is contained in the horizontal cylinder H or its boundary. By [Lemma 5.2](#) we may shear the cylinders H and H' to arrive at the surface in [Figure 12](#). Let D be the diagonal cylinder with dashed boundary that passes through the horizontal saddle connection labeled 1. By [Lemma 5.2](#) we may shear H if necessary to perform one complete Dehn twist and move p into the interior of D while remaining in $\mathcal{M}$. By [Lemma 5.5](#)—where D is intersected by the equivalence classes $\{H\}$ and $\{C\}$ —it follows that there is a periodic point contained in the interior of C . By [Lemma 5.2](#), we may shear H and H' while fixing the remainder of the translation surface to return to (X, ω) with a periodic point p in the interior of C , a contradiction.

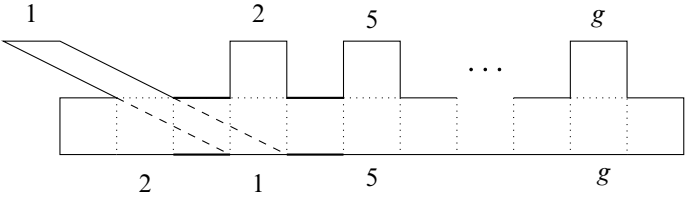


Figure 12: Moving potentially periodic points into the interior of C .

Therefore, p lies on the boundary of C and is contained in the interior of a vertical cylinder which is not $\mathcal{H}$ -free. $\square$

Corollary 8.6 *The only component of $\mathcal{H}(2g-2)$ with periodic points is $\mathcal{H}^{\mathrm{hyp}}(2g-2)$.*

Proof Let $\mathcal{H}$ be a minimal nonhyperelliptic stratum of abelian differentials. If $\mathcal{M}$ is an affine invariant submanifold properly contained in $\mathcal{H}(0)$ that pushes forward to $\mathcal{H}$ under the forgetful map, then its fiber over (X, ω) is nonempty by Lemma 2.14. If $(X, \omega; p)$ is an element of the fiber then p is contained in a vertical cylinder that is contained in C and that is not $\mathcal{H}$ -free. By construction of the surfaces in Section 4 there are no such cylinders and so we are done. $\square$

Proof of Theorem 6.2 Let $\mathcal{H}$ be a component of a stratum of abelian differentials with at least two zeros. Proceed by induction on $\dim_{\mathbb{C}} \mathcal{H}$. The result has already been established for hyperelliptic components (Proposition 8.3), which establishes the base case and allows us to assume without loss of generality that $\mathcal{H}$ is not a hyperelliptic component. Assume that p is a periodic point on (X, ω) . By Proposition 8.5, p is contained on the boundary of C and in a vertical cylinder V that is contained in C and that is not $\mathcal{H}$ -free.

Case 1 C contains an $\mathcal{H}$ -free vertical cylinder W By construction of the surfaces in Section 4, there are only two types of $\mathcal{H}$ -free vertical cylinders contained in C , which are pictured in Figure 13.

By Lemma 5.2, if (X, ω) contains an $\mathcal{H}$ -free cylinder W we may travel in $\mathcal{M}$ in the direction of the standard shear a_W to shrink the horizontal cross curve of W until it vanishes. This degenerates $\mathcal{M}$ to the boundary of $\mathcal{H}(0)$. Notice that in both cases this degeneration causes two distinct zeros to collide on the boundary, but, by considering Euler characteristic, no genus is lost. By Mirzakhani and Wright [24, Corollary 1.2], the resulting translation surface has an orbit closure of strictly smaller dimension than $\mathcal{M}$.

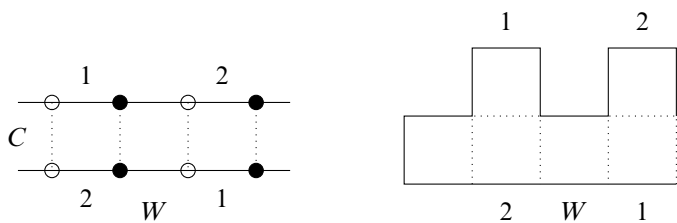


Figure 13: The two types of $\mathcal{H}$ -free cylinders W in C on a subsurface of (X, ω) in components with two zeros of odd order (left) and a subsurface of (X, ω) in even components (right).

By the genericity criterion (Proposition 4.3), the boundary translation surface remains generic in the component of the stratum to which it belongs, which necessarily has complex dimension one less than $\mathcal{M}$. Therefore, p remains a periodic point.

By the induction hypothesis, p is a Weierstrass point and the boundary translation surface belongs to a hyperelliptic component. In particular, p must lie halfway across V on the boundary of C , dividing V into two subcylinders of equal modulus, which we will call V_1 and V_2 . By construction, on $(X, \omega; p)$ the only rational linear homogeneous equation that holds on moduli of cylinders equivalent to V is that the moduli of V_1 and V_2 are equal. By Wright [29]—in particular Theorems 2.9 and 2.10—and Lemma 5.2, V may be sheared on $(X, \omega; p)$ while remaining in $\mathcal{M}$ and fixing the remainder of the translation surface. Shearing so as to perform one complete Dehn twist moves p into the interior of C on (X, ω) , which contradicts Proposition 8.5.

Notice that as a corollary of this step, $\mathcal{H}$ does not contain two zeros of odd order.

Case 2 The surface belongs to an even component Let H and H' be the horizontal cylinders labeled in Figure 11. By Lemma 5.2, the result of shearing them while fixing the remainder of the surface remains in $\mathcal{M}$. Therefore, we shear them to find the surface in $\mathcal{M}$ depicted in Figure 14, which contains a vertical cylinder that contains H and H' and that passes through them exactly once.

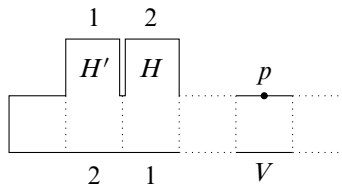


Figure 14: A translation surface in the even component.

By [Lemma 5.2](#), applying the standard dilation $a_{\{H\}}$ to H causes its vertical cross curve to vanish and passes to a surface on the boundary of $\mathcal{M}$. When the cross curve vanishes, the boundary translation surface (shown in [Figure 15](#)) has the zero of order $2k + 2$ (where k is a positive integer) on the boundary of H split into two zeros, one of order 1 and one of order $2k - 1$.

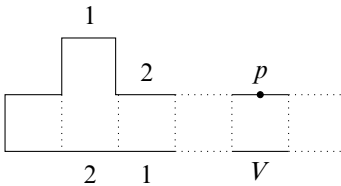


Figure 15: The boundary translation surface (Y, η) .

Let $\mathcal{M}'$ be the orbit closure of $(Y, \eta; p)$. By the genericity criterion ([Proposition 4.3](#)), (Y, η) remains generic. By Mirzakhani and Wright [[24](#), Corollary 1.2], the dimension of $\mathcal{M}'$ must be strictly smaller than the dimension of $\mathcal{M}$. It follows that p is a periodic point on (Y, η) . However, no such points exist on translation surfaces of genus greater than two with a zero of odd order. Thus, (X, ω) cannot belong to an even component.

Case 3 The surface does not belong to an even component The marked point is contained in one of the two configurations shown in [Figure 16](#), where V is adjacent on the right to a vertical cylinder W that contains a horizontal cylinder H (both shown in the figure).

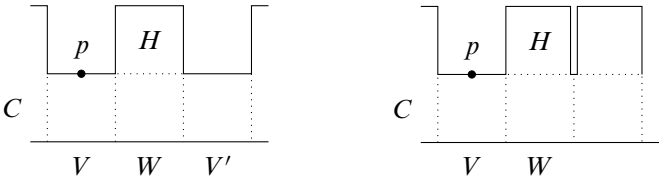


Figure 16: Two possible configurations: the periodic point borders an order 2 zero (left) and the periodic point borders a zero of order 4 or more (right).

By [Lemma 5.2](#), applying the standard dilation $a_{\{H\}}$ to H causes its vertical cross curve to vanish and passes to a surface on the boundary of $\mathcal{M}$. The underlying translation surface moves from $\mathcal{H}(2k, \dots)$ to $\mathcal{H}(0, 2k - 2, \dots)$ (see [Figure 17](#)).

Let $(Y, \eta; \{p\} \cup Q)$ be the boundary translation surface where Q are the marked points that arise in the degeneration. Let $\mathcal{M}'$ be the orbit closure of $(Y, \eta; \{p\} \cup Q)$. The

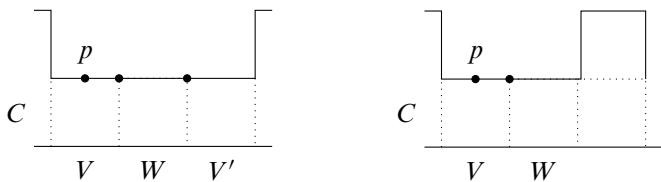


Figure 17: The result of collapsing H : the surface collapses to $\mathcal{H}(0, 0, \dots)$ (left) and the surfaces collapses to $\mathcal{H}(0, 2k - 2, \dots)$ (right).

genericity criterion (Proposition 4.3) implies that $(Y, \eta; Q)$ is generic in the stratum $\mathcal{H}'$ that contains it. By Mirzakhani and Wright [24, Corollary 1.2], $\mathcal{M}'$ is an affine invariant submanifold that is properly contained in $\mathcal{H}'(0)$. The induction hypothesis implies that (Y, η) belongs to a hyperelliptic component and that either p is a Weierstrass point or is exchanged with a point in Q under the hyperelliptic involution.

Consider first the configuration in Figure 17, right. Since $V \cup W$ must be fixed by the hyperelliptic involution and since W may be made arbitrarily long horizontally we see that p is neither a Weierstrass point nor a point exchanged under the hyperelliptic involution with a point in Q .

Consider now the configuration in Figure 17, left, and let V_a and V_b be the left and right subcylinders that p splits V into. We see that p must be exchanged under the hyperelliptic involution with the rightmost point in Q and so V_a and V' have identical moduli. Repeating the argument with the vertical cylinder W' that V borders on the left shows that (X, ω) must contain the subsurface shown in Figure 18 and satisfy the property that the modulus of V' is the same as the modulus of V_a and the modulus of V'' the same as the modulus of V_b .

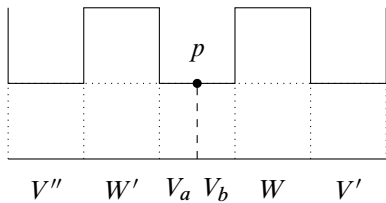


Figure 18: The surface $(X, \omega; p)$.

Letting $\text{Mod}(D)$ denote the modulus of a cylinder D we see that on this surface,

$$\text{Mod}(V) = \text{Mod}(V_a) + \text{Mod}(V_b) = \text{Mod}(V') + \text{Mod}(V''),$$

which is a rational linear homogeneous relation on moduli satisfied by vertical cylinders on (X, ω) , which contradicts the fact that (X, ω) is constructed so as to prevent the moduli of vertical cylinders from satisfying such a relation. $\square$

References

- [1] **P Apisa**, *Periodic points in genus two: holomorphic sections over Hilbert modular varieties, Teichmüller dynamics, and billiards*, preprint (2017) [arXiv](#)
- [2] **P Apisa**, *$\mathrm{GL}_2\mathbb{R}$ orbit closures in hyperelliptic components of strata*, *Duke Math. J.* 167 (2018) 679–742 [MR](#)
- [3] **P Apisa, A Wright**, *Marked points on translation surfaces*, preprint (2017) [arXiv](#)
- [4] **A Avila, A Eskin, M Möller**, *Symplectic and isometric $\mathrm{SL}(2, \mathbb{R})$ -invariant subbundles of the Hodge bundle*, *J. Reine Angew. Math.* 732 (2017) 1–20 [MR](#)
- [5] **F Cukierman**, *Families of Weierstrass points*, *Duke Math. J.* 58 (1989) 317–346 [MR](#)
- [6] **C J Earle, I Kra**, *On sections of some holomorphic families of closed Riemann surfaces*, *Acta Math.* 137 (1976) 49–79 [MR](#)
- [7] **A Eskin, S Filip, A Wright**, *The algebraic hull of the Kontsevich–Zorich cocycle*, *Ann. of Math.* 188 (2018) 281–313 [MR](#)
- [8] **A Eskin, M Mirzakhani**, *Invariant and stationary measures for the $\mathrm{SL}(2, \mathbb{R})$ action on moduli space*, *Publ. Math. Inst. Hautes Études Sci.* 127 (2018) 95–324 [MR](#)
- [9] **A Eskin, M Mirzakhani, A Mohammadi**, *Isolation, equidistribution, and orbit closures for the $\mathrm{SL}(2, \mathbb{R})$ action on moduli space*, *Ann. of Math.* 182 (2015) 673–721 [MR](#)
- [10] **G Farkas, A Verra**, *The universal theta divisor over the moduli space of curves*, *J. Math. Pures Appl.* 100 (2013) 591–605 [MR](#)
- [11] **S Filip**, *Splitting mixed Hodge structures over affine invariant manifolds*, *Ann. of Math.* 183 (2016) 681–713 [MR](#)
- [12] **Q Gendron**, *The Deligne–Mumford and the incidence variety compactifications of the strata of $\Omega\mathcal{M}_g$* , *Ann. Inst. Fourier (Grenoble)* 68 (2018) 1169–1240 [MR](#)
- [13] **S Grushevsky, D Zakharov**, *The double ramification cycle and the theta divisor*, *Proc. Amer. Math. Soc.* 142 (2014) 4053–4064 [MR](#)
- [14] **J Hubbard**, *Sur la non-existence de sections analytiques à la courbe universelle de Teichmüller*, *C. R. Acad. Sci. Paris Sér. A-B* 274 (1972) 978–979 [MR](#)
- [15] **M Kontsevich, A Zorich**, *Connected components of the moduli spaces of Abelian differentials with prescribed singularities*, *Invent. Math.* 153 (2003) 631–678 [MR](#)
- [16] **A Kumar, R E Mukamel**, *Real multiplication through explicit correspondences*, *LMS J. Comput. Math.* 19 (2016) 29–42 [MR](#)

- [17] **S Lelièvre, T Monteil, B Weiss**, *Everything is illuminated*, Geom. Topol. 20 (2016) 1737–1762 [MR](#)
- [18] **H Masur**, *Interval exchange transformations and measured foliations*, Ann. of Math. 115 (1982) 169–200 [MR](#)
- [19] **H Masur**, *Closed trajectories for quadratic differentials with an application to billiards*, Duke Math. J. 53 (1986) 307–314 [MR](#)
- [20] **C T McMullen**, *Billiards and Teichmüller curves on Hilbert modular surfaces*, J. Amer. Math. Soc. 16 (2003) 857–885 [MR](#)
- [21] **C T McMullen**, *Teichmüller curves in genus two: discriminant and spin*, Math. Ann. 333 (2005) 87–130 [MR](#)
- [22] **C T McMullen**, *Teichmüller curves in genus two: torsion divisors and ratios of sines*, Invent. Math. 165 (2006) 651–672 [MR](#)
- [23] **C T McMullen**, *Dynamics of $SL_2(\mathbb{R})$ over moduli space in genus two*, Ann. of Math. 165 (2007) 397–456 [MR](#)
- [24] **M Mirzakhani, A Wright**, *The boundary of an affine invariant submanifold*, Invent. Math. 209 (2017) 927–984 [MR](#)
- [25] **M Mirzakhani, A Wright**, *Full-rank affine invariant submanifolds*, Duke Math. J. 167 (2018) 1–40 [MR](#)
- [26] **S Mullane**, *On the effective cone of $\overline{\mathcal{M}}_{g,n}$* , Adv. Math. 320 (2017) 500–519 [MR](#)
- [27] **F Müller**, *The pullback of a theta divisor to $\overline{\mathcal{M}}_{g,n}$* , Math. Nachr. 286 (2013) 1255–1266 [MR](#)
- [28] **W A Veech**, *Gauss measures for transformations on the space of interval exchange maps*, Ann. of Math. 115 (1982) 201–242 [MR](#)
- [29] **A Wright**, *Cylinder deformations in orbit closures of translation surfaces*, Geom. Topol. 19 (2015) 413–438 [MR](#)
- [30] **A N Zemlyakov, A B Katok**, *Topological transitivity of billiards in polygons*, Mat. Zametki 18 (1975) 291–300 [MR](#) In Russian; translated in *Math. Notes* 18 (1975) 760–764
- [31] **A Zorich**, *Flat surfaces*, from “Frontiers in number theory, physics, and geometry, I” (P Cartier, B Julia, P Moussa, P Vanhove, editors), Springer (2006) 437–583 [MR](#)

Department of Mathematics, University of Chicago
Chicago, IL, United States

Current address: Department of Mathematics, Yale University
New Haven, CT, United States

paul.apisa@gmail.com

Proposed: Benson Farb

Seconded: Anna Wienhard, Jean-Pierre Otal

Received: 6 September 2018

Revised: 29 March 2019

Recognition of being fibered for compact 3–manifolds

ANDREI JAIKIN-ZAPIRAIN

Let M be a compact orientable aspherical 3–manifold. We show that if the profinite completion of $\pi_1(M)$ is isomorphic to the profinite completion of a free-by-cyclic group or to the profinite completion of a surface-by-cyclic group, then M fibers over the circle with compact fiber.

57M27; 20E18, 20J05, 57M05

1 Introduction

Let Γ be a finitely generated residually finite group. It is natural to ask what properties of Γ are determined by the set of its finite quotients. Recall that two finitely generated groups have the same collection of finite quotients if and only if their profinite completions are isomorphic (see the discussion around Theorem 2.2 of A Reid [15]). Thus, our previous question can be reformulated as follows:

Question 1 *Let Γ_1 and Γ_2 be two finitely generated groups. Assume that they have isomorphic profinite completions. What group-theoretic properties are shared by Γ_1 and Γ_2 ?*

There are many recent developments around this question. We recommend the reader to consult the survey of Reid [15], where all these new results are described.

In this paper we consider a particular case of [Question 1](#), where the groups Γ_1 and Γ_2 are the fundamental groups of 3–manifolds. This case has been also studied actively in recent years (for more details, see [\[15, Section 4\]](#)).

Throughout this paper, all manifolds are connected and aspherical. We allow 3–manifolds to have nonempty boundary, but we assume that there are no spherical components.

By a free-by-cyclic group we mean a group $F \rtimes \mathbb{Z}$, where F is a finitely generated free group, and a surface-by-cyclic group is a group $S \rtimes \mathbb{Z}$, where S is the fundamental group of a compact closed orientable surface. Our result shows that the profinite

completion of the fundamental group of a 3–manifold determines whether the manifold fibers over the circle with compact fiber.

Theorem 1.1 *Let M be a compact orientable 3–manifold. Assume that the profinite completion of $\pi_1(M)$ is isomorphic to the profinite completion of a free-by-cyclic group or to the profinite completion of a surface-by-cyclic group. Then M fibers over the circle with compact fiber.*

Under some additional conditions, this result has been proved by Bridson and Reid [5], Boileau and Friedl [4] and Bridson, Reid and Wilton [6]. Our proof and all the previous proofs use in an essential way results of I Agol [1; 2] and P Przytycki and D Wise [14; 18] on separability of 3–manifold groups.

Corollary 1.2 *Let M and N be two compact orientable 3–manifolds such that $\widehat{\pi_1(M)} \cong \widehat{\pi_1(N)}$. Then M fibers over the circle if and only if N does.*

Let us explain briefly the strategy of the proof of [Theorem 1.1](#). If Γ is a finitely generated group, then we say that a nontrivial class $\phi \in H^1(\Gamma, \mathbb{Z}) = \text{Hom}(\Gamma, \mathbb{Z})$ is *fibred* if $\ker \phi$ is finitely generated. If $\Gamma = \pi_1(M)$ is the fundamental group of a compact orientable 3–manifold M , then, by a well-known result of J Stallings [17], fibred classes in $H^1(\pi_1(M), \mathbb{Z})$ are in one-to-one correspondence with fibrations of M over the circle with compact fiber.

Now, let us assume that M satisfies the conditions of [Theorem 1.1](#). First we have to find a useful criterion for a class $\phi \in H^1(\pi_1(M), \mathbb{Z})$ to be fibred. We start with a criterion proved by S Friedl and S Vidussi ([Proposition 3.3](#)). In [Section 3.1](#) we explain the notion of cohomological goodness introduced by J-P Serre [16]. Using that the fundamental group of a 3–manifold is cohomologically good we reformulate [Proposition 3.3](#) in terms of the profinite completion of $\pi_1(M)$ and obtain [Corollary 3.4](#). It tells us that $\phi \in H^1(\pi_1(M), \mathbb{Z})$ is fibred if certain homology groups of some subgroups of $\widehat{\pi_1(M)}$ are finite. The latter property can be understood in terms of certain rank functions, which we study in [Section 2](#). In [Section 4](#) we prove a proposition that together with [Corollary 2.2](#) allows us to find a desired class. All this is explained again in more detail in [Section 5](#).

Acknowledgements

This paper is partially supported by the grants MTM2017-82690-P and MTM2014-53810-C2-01 of the Spanish MINECO, the grant PRX16/00179 of the mobility program

“Salvador de Madariaga” of the Spanish MECD and by the ICMAT Severo Ochoa project SEV-2015-0554.

This paper was written while the author was visiting the Mathematical Institute of the University of Oxford. I would like to thank everyone involved for their fine hospitality.

I would like to thank Pavel Zalesskii for bringing this problem to my attention and to Martin Bridson, Alan Reid and Henry Wilton for very useful discussions. I would also like to thank Diego López-Álvarez and an anonymous referee whose remarks and suggestions substantially improved the article.

2 Rank functions

Let R be a commutative domain and let $Q(R)$ denote its field of fractions. Given a matrix A over R we denote by $\text{rk}_R(A)$ the $Q(R)$ -rank of A . Similarly, if $\phi: M_1 \rightarrow M_2$ is a homomorphism of finitely generated free left R -modules, we denote by $\text{rk}_R(\phi)$ the $Q(R)$ -rank of the induced map $Q(R) \otimes_R M_1 \rightarrow Q(R) \otimes_R M_2$.

We may extend the previous notation in the following way. Let S be a ring and assume that S_0 is a subring of S such that $S \cong (S_0)^k$ for some $k \geq 1$ as a left S_0 -module. If $\alpha: S_0 \rightarrow R$ is a homomorphism of rings, then α induces a structure of right S_0 -module on R . Take $A \in \text{Mat}_{n \times m}(S)$ and consider the homomorphism of finitely generated free left R -modules

$$\phi_\alpha^A: (R \otimes_{S_0} S)^n \rightarrow (R \otimes_{S_0} S)^m, \quad (x_1, \dots, x_n) \mapsto (x_1, \dots, x_n)A.$$

If R is a commutative domain, then we put $\text{rk}_\alpha(A) = \text{rk}_R(\phi_\alpha^A)$.

Let us describe two concrete examples of rank functions that will appear in this paper. Let Γ be a residually finite group and H a subgroup of finite index. Put $S = \mathbb{Z}[\Gamma]$ and $S_0 = \mathbb{Z}[H]$. It is clear that S is a left free S_0 -module. We fix a prime p once and for all throughout the paper.

(A) Let K be a normal subgroup of Γ such that Γ/K is abelian and torsion-free. Let $\alpha_1: \mathbb{Z}[H] \rightarrow \mathbb{F}_p[H/H \cap K]$ be the canonical map. We define

$$\text{rk}_{H,K}(A) = \frac{\text{rk}_{\alpha_1}(A)}{|\Gamma:H|}, \quad \text{where } A \text{ is a matrix over } \mathbb{Z}[\Gamma].$$

(B) For a profinite group G and a profinite ring T we denote by $T[[G]]$ the completed group algebra

$$T[[G]] = \varprojlim_{U \trianglelefteq_o G} T[G/U].$$

Now let G be a profinite completion of Γ . We identify Γ with its canonical image in G . Let U be the closure of H in G and let N be a closed normal subgroup of G such that G/N is an abelian and torsion-free pro- p group. Let $\alpha_2: \mathbb{Z}[H] \rightarrow \mathbb{F}_p[[U/U \cap N]]$ be the canonical map. In this context we define

$$\mathrm{rk}_{H,N}(A) = \frac{\mathrm{rk}_{\alpha_2}(A)}{|G:U|}, \quad \text{where } A \text{ is a matrix over } \mathbb{Z}[\Gamma].$$

This rank function coincides with the previous one in the case $K = \Gamma \cap N$. Indeed, for every matrix A over $\mathbb{Z}[\Gamma]$ we have the equality

$$\mathrm{rk}_{\alpha_1}(A) = \mathrm{rk}_{\alpha_2}(A)$$

because α_2 is the composition of α_1 and the canonical embedding of $\mathbb{F}_p[H/H \cap K]$ into $\mathbb{F}_p[[U/U \cap N]]$. Thus, we obtain that

$$(1) \quad \mathrm{rk}_{H,\Gamma \cap N}(A) = \mathrm{rk}_{H,N}(A).$$

The rank functions $\mathrm{rk}_{H,K}$ and $\mathrm{rk}_{H,N}$ are examples of Sylvester matrix rank functions on $\mathbb{Z}[\Gamma]$. The interested reader can find more information about them in [11]. We need the following elementary lemma about rank functions:

Lemma 2.1 *Let S be a ring and assume that S_0 is a subring of S such that S is a finitely generated free left S_0 -module. Let $\pi: R_1 \rightarrow R_2$ be a homomorphism between two commutative domains and $\alpha_1: S_0 \rightarrow R_1$ a homomorphism. We put $\alpha_2 = \pi \circ \alpha_1$. Then for any matrix A over S we have*

$$\mathrm{rk}_{\alpha_1}(A) \geq \mathrm{rk}_{\alpha_2}(A).$$

Proof For simplicity of exposition, let us assume that $A \in S$ (that is, A is a 1-by-1 matrix).

Let $\{v_1, \dots, v_k\}$ be a basis of S as a left S_0 -module. Consider the matrix associated to ϕ_{Id}^A with respect to $\{v_1, \dots, v_k\}$. It is a k -by- k matrix $B = (b_{ij})$ over S_0 defined by means of

$$v_i A = \sum_{j=1}^k b_{ij} v_j.$$

Let $B_1 = \alpha_1(B)$ and $B_2 = \alpha_2(B)$, and observe that $B_2 = \pi(B_1)$. Thus, $\mathrm{rk}_{R_1}(B_1) \geq \mathrm{rk}_{R_2}(B_2)$. Therefore, we obtain that

$$\mathrm{rk}_{\alpha_1}(A) = \mathrm{rk}_{R_1}(B_1) \geq \mathrm{rk}_{R_2}(B_2) = \mathrm{rk}_{\alpha_2}(A). \quad \square$$

We will apply the lemma in the following situation:

Corollary 2.2 *Let Γ be a residually finite group and let $G = \hat{\Gamma}$. Let N be a closed normal subgroup of G such that $G/N \cong \mathbb{Z}_p$ and let K be a normal subgroup of Γ such that $\Gamma/K \cong \mathbb{Z}$. Assume that $\Gamma \cap N \leq K$. Then, for every matrix A over $\mathbb{Z}[\Gamma]$ and every subgroup H of Γ of finite index,*

$$\mathrm{rk}_{H,N}(A) \geq \mathrm{rk}_{H,K}(A).$$

Proof Using (1), we obtain that $\mathrm{rk}_{H,\Gamma \cap N}(A) = \mathrm{rk}_{H,N}(A)$. Since $\Gamma \cap N \leq K$, the previous lemma implies that $\mathrm{rk}_{H,\Gamma \cap N}(A) \geq \mathrm{rk}_{H,K}(A)$. Hence we are done. $\square$

3 Preliminaries on 3-manifolds

3.1 Cohomologically good groups

Let Γ be a group and $\hat{\Gamma}$ its profinite completion. The group Γ is called *cohomologically good* if the homomorphism of cohomology groups

$$i^n(M): H^n(\hat{\Gamma}, M) \rightarrow H^n(\Gamma, M)$$

induced by the natural homomorphism $i: \Gamma \rightarrow \hat{\Gamma}$ is an isomorphism for every $\mathbb{Z}[\Gamma]$ -module M having a finite number of elements. This notion was introduced by Serre (see [16, Section I.2.6]) and has been studied in several papers (see for example [10; 9; 13]).

It is known that free, surface, free-by-cyclic, surface-by-cyclic and 3-manifold groups are cohomologically good. The result on 3-manifold groups (see for example [7; 9, Remark 5.14; 3]) uses the recent advances in the theory of 3-manifolds, including the solution of the virtual Haken conjecture by Agol [2]. We want to point out that under the hypothesis of Theorem 1.1 that the profinite completion of $\pi_1(M)$ is isomorphic to the profinite completion of a free-by-cyclic group or to the profinite completion of a surface-by-cyclic group, the cohomological goodness of $\pi_1(M)$ can be proved directly as in [9, Corollary 5.4].

In the case where Γ is an FP_∞ -group there is an alternative way to define the cohomological goodness.

Proposition 3.1 *Let Γ be an FP_∞ -group and let*

$$\dots \rightarrow \mathbb{Z}[\Gamma]^{d_i} \xrightarrow{\phi_i} \mathbb{Z}[\Gamma]^{d_{i-1}} \rightarrow \dots \rightarrow \mathbb{Z}[\Gamma]^{d_1} \xrightarrow{\phi_1} \mathbb{Z}[\Gamma] \rightarrow \mathbb{Z} \rightarrow 0$$

be a resolution of the trivial $\mathbb{Z}[\Gamma]$ -module $\mathbb{Z}$. Then Γ is cohomologically good if and only if the induced sequence

$$\dots \rightarrow \widehat{\mathbb{Z}}[\widehat{\Gamma}]^{d_i} \xrightarrow{\widehat{\phi}_i} \widehat{\mathbb{Z}}[\widehat{\Gamma}]^{d_{i-1}} \rightarrow \dots \rightarrow \widehat{\mathbb{Z}}[\widehat{\Gamma}]^{d_1} \xrightarrow{\widehat{\phi}_1} \widehat{\mathbb{Z}}[\widehat{\Gamma}] \rightarrow \widehat{\mathbb{Z}} \rightarrow 0$$

is exact.

Proof The “if” part is clear. Let us show the “only if” part.

By [16, page 15], since Γ is cohomologically good we have that for every $n \geq 1$, every prime p , every subgroup H of Γ of finite index and every $\alpha \in H^k(H, \mathbb{F}_p)$, there exists another subgroup T of finite index in H such that the restriction of α to T vanishes in $H^k(T, \mathbb{F}_p)$. Hence the direct limit of $H^k(H, \mathbb{F}_p)$ (when H runs over all subgroups of finite index of Γ) is equal to zero, and so the inverse limit of $H_k(H, \mathbb{F}_p)$ is also equal to zero. Now we can apply [12, Theorem 2.5] and conclude that $H_k(\Gamma, \mathbb{Z}_p[\widehat{\Gamma}]) = 0$ for each p , and therefore $H_k(\Gamma, \widehat{\mathbb{Z}}[\widehat{\Gamma}]) = 0$. This implies the exactness of the second sequence in the statement of the proposition. $\square$

3.2 Profinite completions of 3-manifold groups

In the following proposition we show that the hypotheses of Theorem 1.1 impose several restrictions on the boundary of M .

Proposition 3.2 *Let M be a compact orientable 3-manifold and let Γ be a free-by-cyclic group or a surface-by-cyclic group. Assume that $\widehat{\pi_1(M)} \cong \widehat{\Gamma}$. Then the boundary of M is a union of incompressible tori or it is empty.*

Proof If Γ is a free-by-cyclic group, the proposition follows from [5, Corollary 4.3]. If Γ is a surface-by-cyclic group, then $\widehat{\pi_1(M)}$ is of cohomological dimension 3. Since $\pi_1(M)$ is cohomologically good, $\pi_1(M)$ is of cohomological dimension 3 as well, and so M does not have boundary. $\square$

3.3 A criterion for fibering

Let $\phi: \Gamma \rightarrow \mathbb{Z}$ be a nontrivial homomorphism and let H be a subgroup of finite index in Γ . We denote by H_ϕ the intersection $H \cap \ker \phi$. We have the following criterion of fibering, stated in a slightly different way in a paper of Friedl and Vidussi [8]:

Proposition 3.3 *Let M be a compact orientable 3-manifold with toroidal or empty boundary and let $\phi: \pi_1(M) \rightarrow \mathbb{Z}$ be a nontrivial map. Then the class ϕ is fibered if and only if for every normal subgroup H of $\pi_1(M)$ of finite index the group $H_1(H_\phi, \mathbb{F}_p)$ is finite.*

Proof The “only if” part is clear, since if ϕ fibers, $\ker \phi$ is finitely generated.

Let H be a normal subgroup of $\pi_1(M)$ of finite index and assume that $H_1(H_\phi, \mathbb{F}_p)$ is finite. Hence $H_1(\pi_1(M), \mathbb{F}_p[\pi_1(M)/H_\phi]) \cong H_1(H_\phi, \mathbb{F}_p)$ is also finite. Put $L = \mathbb{F}_p[\pi_1(M)/H]$. The tensor product

$$\tilde{L} = \mathbb{F}_p[\pi_1(M)/\ker \phi] \otimes_{\mathbb{F}_p} L$$

is a finitely generated right $\pi_1(M)$ -module (the elements of $\pi_1(M)$ act diagonally on L_ϕ). The i^{th} twisted Alexander polynomial $\Delta_{M,\phi,i}^L$ is the order of the $\mathbb{F}_p[\pi_1(M)/\ker \phi]$ -module $H_i(\pi_1(M), \tilde{L})$ (see [4] for details). Observe that $\tilde{L}$ is a direct sum of a finite number of copies of $\mathbb{F}_p[\pi_1(M)/H_\phi]$. Therefore, $H_1(\pi_1(M), \tilde{L})$ is also finite. Hence

$$\Delta_{M,\phi,1}^{\mathbb{F}_p[\pi_1(M)/H]} \neq 0.$$

Thus, we can apply [8, Theorem 1.1] and obtain the “if” part of the proposition. $\square$

Let Γ be a finitely generated group. A map $\phi: \Gamma \rightarrow \mathbb{Z}$ induces a homomorphism $\phi_k: \hat{\Gamma} \rightarrow \mathbb{Z}_p/p^k\mathbb{Z}_p$. We put

$$\hat{\phi} = \lim_k \phi_k: \hat{\Gamma} \rightarrow \mathbb{Z}_p$$

and say that $\hat{\phi}$ is associated to ϕ .

Let G be a profinite group, U an open subgroup and $\varphi: G \rightarrow \mathbb{Z}_p$ a nontrivial homomorphism. We denote by U_φ the intersection $U \cap \ker \varphi$. Note that $U/U_\varphi \cong \mathbb{Z}_p$. Now we present a profinite analogue of Proposition 3.3.

Corollary 3.4 *Let M be a compact orientable 3-manifold with toroidal or empty boundary and let $\phi: \pi_1(M) \rightarrow \mathbb{Z}$ be a nontrivial map. Denote by $\hat{\phi}: \widehat{\pi_1(M)} \rightarrow \mathbb{Z}_p$ the homomorphism associated to ϕ . Then the class ϕ is fibered if and only if for every normal open subgroup U of $\widehat{\pi_1(M)}$ the group $H_1(U_{\hat{\phi}}, \mathbb{F}_p)$ is finite.*

Proof The “only if” part is clear. Let us prove the “if” part.

We write a projective resolution of $\mathbb{Z}$ as a left $\mathbb{Z}[\pi_1(M)]$ -module

$$\mathbb{Z}[\pi_1(M)]^r \xrightarrow{m_B} \mathbb{Z}[\pi_1(M)]^d \rightarrow \mathbb{Z}[\pi_1(M)] \rightarrow \mathbb{Z} \rightarrow 0,$$

where m_B is multiplication by a matrix B . Since, $\pi_1(M)$ is cohomologically good, using Proposition 3.1, we obtain an exact sequence

$$\widehat{\mathbb{Z}[\pi_1(M)]}^r \xrightarrow{m_B} \widehat{\mathbb{Z}[\pi_1(M)]}^d \rightarrow \widehat{\mathbb{Z}[\pi_1(M)]} \rightarrow \widehat{\mathbb{Z}} \rightarrow 0.$$

Let H be a normal subgroup of $\pi_1(M)$ of finite index and let U be the closure of H in $\widehat{\pi_1(M)}$. Then U is a normal open subgroup of $\widehat{\pi_1(M)}$. Note that

$$H_1(U_{\widehat{\phi}}, \mathbb{F}_p) \cong H_1(\widehat{\pi_1(M)}, \mathbb{F}_p[\widehat{\pi_1(M)}/U_{\widehat{\phi}}]).$$

Thus, since $H_1(U_{\widehat{\phi}}, \mathbb{F}_p)$ is finite, we obtain that $\text{rk}_{H, \ker \widehat{\phi}}(B) = d - 1$. Now, observe that, by (1),

$$\text{rk}_{H, \ker \phi}(B) = \text{rk}_{H, \ker \widehat{\phi}}(B) = d - 1.$$

Therefore,

$$H_1(H_{\phi}, \mathbb{F}_p) \cong H_1(\pi_1(M), \mathbb{F}_p[\pi_1(M)/H_{\phi}])$$

is also finite. Hence we can apply Proposition 3.3 and obtain the “if” part of the corollary. □

4 Finding a fibered class

The following proposition will be a key result that will help us to find a fibered class in Theorem 1.1.

Proposition 4.1 *Let $V = \mathbb{Z}_p^k$ and let $I \cong J \cong \mathbb{Z}^k$ be two dense subgroups of V . Put $V^* = \text{Hom}(V, \mathbb{Z}_p)$. For $\varphi \in V^*$ we denote by $\varphi|_I$ the restriction of φ to I . Then, for every $\psi \in V^*$ such that $\psi(I) \leq \mathbb{Z}$, there exists an element $\varphi \in V^*$ such that $\varphi(J) \leq \mathbb{Z}$ and $\ker \varphi|_I \leq \ker \psi|_I$.*

Proof Let $\{v_1, \dots, v_k\}$ and $\{u_1, \dots, u_k\}$ be $\mathbb{Z}$ -bases of I and J , respectively. Define elements $v_1^*, \dots, v_k^*, u_1^*, \dots, u_k^*$ of V^* by means of

$$v_i^*(v_j) = u_i^*(u_j) = \delta_{ij}.$$

Since $\{v_1, \dots, v_k\}$ and $\{u_1, \dots, u_k\}$ are $\mathbb{Z}_p$ -bases of V , $\{v_1^*, \dots, v_k^*\}$ and $\{u_1^*, \dots, u_k^*\}$ are $\mathbb{Z}_p$ -bases of V^* . Hence there exists a matrix $A \in \text{GL}_k(\mathbb{Z}_p)$ such that

(2)
$$(u_1^*, \dots, u_k^*) = (v_1^*, \dots, v_k^*)A.$$

Write A as

$$(3) \quad A = \sum_{i=1}^s A_i z_i,$$

where $A_i \in \text{Mat}_k(\mathbb{Z})$ and $\{z_1, \dots, z_s\} \subset \mathbb{Z}_p$ are $\mathbb{Z}$ -linearly independent.

Let $R = \mathbb{Q}[t_1, \dots, t_s]$ be a polynomial ring. We put $B = \sum_{i=1}^s A_i t_i \in \text{Mat}_k(R)$. Since $A = \sum_{i=1}^s A_i z_i$ is invertible over $\mathbb{Z}_p$, $\det A \neq 0$. Hence $\det B \neq 0$ as well and, as a consequence, there are $q_1, \dots, q_s \in \mathbb{Z}$ such that

$$\det \left(\sum_{i=1}^s A_i q_i \right) \neq 0.$$

Let us put

$$(4) \quad C = \sum_{i=1}^s A_i q_i.$$

We can express ψ as

$$(5) \quad \psi = \alpha_1 v_1^* + \dots + \alpha_k v_k^*, \quad \text{where } \alpha_i \in \mathbb{Z}.$$

Since C is invertible over $\mathbb{Q}$, there are $\beta_1, \dots, \beta_k \in \mathbb{Z}$ and $0 \neq n \in \mathbb{Z}$ such that

$$(6) \quad C \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = n \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}.$$

Now we set

$$\begin{aligned} \varphi &= \sum_{i=1}^k \beta_i u_i^* = (v_1^*, \dots, v_k^*) A \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} && \text{(by (2))} \\ &= \sum_{j=1}^s z_j (v_1^*, \dots, v_k^*) A_j \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} && \text{(by (3))}, \end{aligned}$$

and define

$$(7) \quad \varphi_j = (v_1^*, \dots, v_k^*) A_j \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \in \mathbb{Z} v_1^* + \dots + \mathbb{Z} v_k^*.$$

Then, since $\{z_1, \dots, z_s\}$ are $\mathbb{Z}$ -linearly independent,

$$\ker \varphi|_I = \ker(z_1(\varphi_1)|_I + \dots + z_s(\varphi_s)|_I) = \bigcap_{i=1}^s \ker(\varphi_j)|_I.$$

Thus, we obtain

$$\begin{aligned}
 \ker \varphi|_I &\leq \ker \left(\sum_{j=1}^s q_j(\varphi_j)|_I \right) = \ker \left((v_1^*, \dots, v_k^*) C \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} |_I \right) \quad (\text{by (7) and (4)}) \\
 &= \ker \left((v_1^*, \dots, v_k^*) \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix} |_I \right) \quad (\text{by (6)}) \\
 &= \ker \psi|_I \quad (\text{by (5)}). \quad \square
 \end{aligned}$$

5 Proof of Theorem 1.1

In this section we prove [Theorem 1.1](#). Let $\Gamma = S \rtimes \mathbb{Z}$ be a free-by-cyclic or a surface-by-cyclic group. Then $\hat{\Gamma} \cong \hat{S} \rtimes \hat{\mathbb{Z}}$. We write a projective resolution of $\mathbb{Z}$ as a left $\mathbb{Z}[\Gamma]$ -module,

$$\mathbb{Z}[\Gamma]^r \xrightarrow{m_A} \mathbb{Z}[\Gamma]^d \rightarrow \mathbb{Z}[\Gamma] \rightarrow \mathbb{Z} \rightarrow 0,$$

where m_A is multiplication by a matrix A . We will need the following lemma:

Lemma 5.1 *Let $\varphi: \hat{\Gamma} \rightarrow \mathbb{Z}_p$ be a nontrivial homomorphism and let H be a subgroup of Γ of finite index. Denote by U the closure of H in $\hat{\Gamma}$. Then $H_1(U_\varphi, \mathbb{F}_p)$ is finite if and only if $\text{rk}_{H, \ker \varphi}(A) = d - 1$.*

Proof We may use the same argument as in the proof of [Corollary 3.4](#), because Γ is cohomologically good. $\square$

Let M be a compact orientable 3-manifold and assume that $\widehat{\pi_1(M)}$ and $\hat{\Gamma}$ are isomorphic. By [Proposition 3.2](#), M has toroidal or empty boundary.

Denote by $\alpha: \widehat{\pi_1(M)} \rightarrow \hat{\Gamma}$ an isomorphism between $\widehat{\pi_1(M)}$ and $\hat{\Gamma}$. In order to show that $\pi_1(M)$ is fibered we want to use [Corollary 3.4](#). To do so we are going to work in $\hat{\Gamma}$ and then transport the obtained information from $\hat{\Gamma}$ to $\widehat{\pi_1(M)}$ using α^{-1} .

Let $f: \Gamma \rightarrow \mathbb{Z}$ be such that $\ker f = S$ and let $\hat{f}: \hat{\Gamma} \rightarrow \mathbb{Z}_p$ be the map associated to f . Denote by V the maximal torsion-free abelian pro- p quotient of $\hat{\Gamma}$. Then $V \cong \mathbb{Z}_p^k$ for some $k \geq 1$. Let I and J be the images in V of Γ and $\alpha(\pi_1(M))$, respectively.

Applying [Proposition 4.1](#) for $\psi = \hat{f}$, we obtain that there exists $\varphi: \hat{\Gamma} \rightarrow \mathbb{Z}_p$ such that

$$\varphi(\alpha(\pi_1(M))) \leq \mathbb{Z} \quad \text{and} \quad \ker \varphi \cap \Gamma \leq \ker f.$$

Let U be an open normal subgroup of $\hat{\Gamma}$. Put $H = \Gamma \cap U$. Observe that U is equal to the closure of H in $\hat{\Gamma}$. By [Corollary 2.2](#),

$$\mathrm{rk}_{H, \ker \varphi}(A) \geq \mathrm{rk}_{H, \ker f}(A) = d - 1.$$

Hence, by [Lemma 5.1](#), $H_1(U_\varphi, \mathbb{F}_p)$ is finite.

Now, we put

$$\phi = \varphi \circ \alpha|_{\pi_1(M)} \in H^1(\pi_1(M), \mathbb{Z}).$$

Let $\hat{\phi}: \widehat{\pi_1(M)} \rightarrow \mathbb{Z}_p$ be associated to ϕ . Then we have that $\hat{\phi} = \varphi \circ \alpha$.

Thus, we obtain that

$$H_1(U_{\hat{\phi}}, \mathbb{F}_p) \cong H_1(\alpha(U_{\hat{\phi}}), \mathbb{F}_p) = H_1(\alpha(U)_\varphi, \mathbb{F}_p)$$

is finite for every normal open subgroup U of $\widehat{\pi_1(M)}$. Therefore, applying [Corollary 3.4](#), we obtain that ϕ is fibered. This finishes the proof of [Theorem 1.1](#).

References

- [1] **I Agol**, *Criteria for virtual fibering*, J. Topol. 1 (2008) 269–284 [MR](#)
- [2] **I Agol**, *The virtual Haken conjecture*, Doc. Math. 18 (2013) 1045–1087 [MR](#)
- [3] **M Aschenbrenner, S Friedl, H Wilton**, *3-Manifold groups*, Eur. Math. Soc., Zürich (2015) [MR](#)
- [4] **M Boileau, S Friedl**, *The profinite completion of 3-manifold groups, fiberedness and the Thurston norm*, preprint (2015) [arXiv](#)
- [5] **M Bridson, A Reid**, *Profinite rigidity, fibering, and the figure-eight knot*, preprint (2015) [arXiv](#)
- [6] **MR Bridson, A W Reid, H Wilton**, *Profinite rigidity and surface bundles over the circle*, Bull. Lond. Math. Soc. 49 (2017) 831–841 [MR](#)
- [7] **W Cavendish**, *Finite-sheeted covering spaces and solenoids over 3-manifolds*, PhD thesis, Princeton University (2012) [MR](#) <https://search.proquest.com/docview/1040698746>
- [8] **S Friedl, S Vidussi**, *A vanishing theorem for twisted Alexander polynomials with applications to symplectic 4-manifolds*, J. Eur. Math. Soc. 15 (2013) 2027–2041 [MR](#)
- [9] **F Grunewald, A Jaikin-Zapirain, A G S Pinto, P A Zalesskii**, *Normal subgroups of profinite groups of non-negative deficiency*, J. Pure Appl. Algebra 218 (2014) 804–828 [MR](#)

- [10] **F Grunewald, A Jaikin-Zapirain, P A Zalesskii**, *Cohomological goodness and the profinite completion of Bianchi groups*, Duke Math. J. 144 (2008) 53–72 [MR](#)
- [11] **A Jaikin-Zapirain**, *L^2 -Betti numbers and their analogues in positive characteristic*, from “Groups St Andrews 2017 in Birmingham” (C M Campbell, C W Parker, M R Quick, E F Robertson, C M Roney-Dougal, editors), London Math. Soc. Lecture Note Ser. 455, Cambridge Univ. Press (2019) 346–405 [MR](#)
- [12] **D H Kochloukova, P A Zalesskii**, *Profinite and pro- p completions of Poincaré duality groups of dimension 3*, Trans. Amer. Math. Soc. 360 (2008) 1927–1949 [MR](#)
- [13] **K Lorensen**, *Groups with the same cohomology as their profinite completions*, J. Algebra 320 (2008) 1704–1722 [MR](#)
- [14] **P Przytycki, D T Wise**, *Separability of embedded surfaces in 3-manifolds*, Compos. Math. 150 (2014) 1623–1630 [MR](#)
- [15] **A W Reid**, *Profinite rigidity*, from “Proceedings of the International Congress of Mathematicians” (B Sirakov, P N de Souza, M Viana, editors), volume 2, World Sci., Hackensack, NJ (2018) 1211–1234 [MR](#)
- [16] **J-P Serre**, *Galois cohomology*, Springer (2002) [MR](#)
- [17] **J Stallings**, *On fibering certain 3-manifolds*, from “Topology of 3-manifolds and related topics” (J M K Fort, editor), Prentice-Hall, Englewood Cliffs, NJ (1962) 95–100 [MR](#)
- [18] **D T Wise**, *Research announcement: the structure of groups with a quasiconvex hierarchy*, Electron. Res. Announc. Math. Sci. 16 (2009) 44–55 [MR](#)

Departamento de Matemáticas, University Autónoma de Madrid and Instituto de Ciencias Matemáticas
 Madrid, Spain

andrei.jaikin@uam.es

<http://matematicas.uam.es/~andrei.jaikin>

Proposed: Ian Agol

Received: 27 September 2018

Seconded: Martin Bridson, Cameron Gordon

Revised: 11 March 2019

Edge stabilization in the homology of graph braid groups

BYUNG HEE AN

GABRIEL C DRUMMOND-COLE

BEN KNUDSEN

We introduce a novel type of stabilization map on the configuration spaces of a graph which increases the number of particles occupying an edge. There is an induced action on homology by the polynomial ring generated by the set of edges, and we show that this homology module is finitely generated. An analogue of classical homological and representation stability for manifolds, this result implies eventual polynomial growth of Betti numbers. We calculate the exact degree of this polynomial, in particular verifying an upper bound conjectured by Ramos. Because the action arises from a family of continuous maps, it lifts to an action at the level of singular chains which contains strictly more information than the homology-level action. We show that the resulting differential graded module is almost never formal over the ring of edges.

13D40, 20F36, 55R80; 05C40

1 Introduction

Configuration spaces of manifolds have numerous applications in algebraic topology and homotopy theory; see Arnold [3], McDuff [22], and Cohen, Lada, and May [6] for a few notable examples. More recently, there has been a growing swell of interest in configuration spaces of *graphs*. For a graph Γ — which is to say a finite, one-dimensional cell complex — we write

$$B_k(\Gamma) = \{(x_1, \dots, x_k) \in \Gamma^k : x_i \neq x_j \text{ if } i \neq j\} / \Sigma_k$$

for the k^{th} unordered *configuration space* of Γ . First considered by Ghrist [11] from the point of view of robotics and motion planning, these spaces were shown by Abrams [1, Corollary 3.11] to be aspherical for Γ connected, and thus they classify their respective fundamental groups, the graph braid groups of Γ . Much effort has been dedicated to understanding the homological, geometric, and combinatorial properties of these groups by a number of different researchers and groups. A nonexhaustive list includes work of An, Drummond-Cole, and Knudsen [2], Chetthi and Lütgehetmann [4], Farley and Sabalka [8; 9], Harrison, Keating, Robbins, and Sawicki [12], Kim, Ko, and

Park [13], Ko and Park [16], Lütgehetmann [18], Lütgehetmann and Recio-Mitter [19], Maciążek and Sawicki [20; 21], Ramos [24], Sabalka [26], and Świątkowski [27].

1.1 Stability phenomena

Configuration spaces of different cardinalities relate to one another in a variety of ways, and it is often simpler to study the graded space $B(-) = \bigsqcup_{k \geq 0} B_k(-)$. For example, if M is a manifold with nonempty boundary, there is a *stabilization map*

$$\sigma: B(M) \rightarrow B(M)$$

which inserts a new particle near the boundary. This map increases the cardinality k by 1, and for sufficiently large k , a theorem of McDuff [22, Theorem 1.2] asserts that the induced map $H_i(B_k(M)) \rightarrow H_i(B_{k+1}(M))$ is an isomorphism. One says that the configuration spaces of such manifolds exhibit *homological stability*.

One can define such a map on the configuration spaces of a graph, inserting a new particle at the end of a dangling edge, but stability almost never occurs. In order to remedy this defect, we draw inspiration from a related situation in which homological stability fails, namely that of the *ordered* configuration spaces $\text{Conf}_k(M)$. Although the Betti numbers of these spaces do not stabilize, the symmetric group representations occurring in homology do. This phenomenon of *representation stability*, as formalized by Church, Ellenberg, and Farb, may be summarized in the statement [5, Theorem 6.2.1] that the collection $\{H_*(\text{Conf}_k(M))\}_{k \geq 0}$ is a *finitely generated module* over a certain combinatorial category.

In light of these results, a philosophy has emerged that a good notion of homological stability in a given context should be the property of finite generation with respect to some naturally occurring action. For example, in hindsight, one interprets the theorem of McDuff mentioned above as a statement about the action of the polynomial ring $\mathbb{Z}[\sigma_*]$.

1.2 Edge stabilization

In this paper, we investigate a new stability phenomenon in the homology $H_*(B(\Gamma))$, which takes the form of an action by the polynomial ring generated by the edges of Γ . We prove the following analogue of the homological and representation stability enjoyed by configuration spaces of manifolds.

Theorem 1.1 *Let Γ be a graph with set of edges E . For any $i \geq 0$, the $\mathbb{Z}[E]$ -module $H_i(B(\Gamma); \mathbb{Z})$ is finitely generated.*

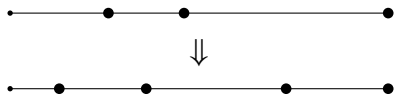


Figure 1: Edge stabilization.

In fact, since finite generation and presentation are equivalent over Noetherian rings, the module is finitely presented. By a theorem of Hilbert, [Theorem 1.1](#) implies eventual polynomial growth of Betti numbers. Our next result gives the exact degree of this polynomial in terms of a certain connectivity invariant Δ_Γ^i , which is roughly the largest number of connected components of edges obtainable by removing i essential vertices from Γ ; see [Definition 3.1](#) for a precise definition. In stating the result, we assume that Γ is not a discrete graph, for which the question is trivial.

Theorem 1.2 *Let $\mathbb{F}$ be a field and Γ a graph with at least one edge. For k sufficiently large, $\dim_{\mathbb{F}} H_i(B_k(\Gamma); \mathbb{F})$ coincides with a polynomial in k of degree $\Delta_\Gamma^i - 1$.*

This module structure, first introduced and exploited as a purely algebraic phenomenon by Ramos in [\[24, Section 3.2\]](#) for trees and by the authors of the present paper [\[2, Section 4.2\]](#) in general, has a simple topological origin. The *edge stabilization map*

$$\sigma_e: B(\Gamma) \rightarrow B(\Gamma)$$

acts by replacing the subconfiguration of particles on e with the collection of averages of consecutive particles or endpoints; see [Figure 1](#) and [Section 2.2](#).

1.3 Edge formality

Because the various edge stabilization maps commute on the nose — not merely up to homotopy — the action of $\mathbb{Z}[E]$ on homology arises from a $\mathbb{Z}[E]$ –action on singular chains. This chain level structure is surprisingly rich; indeed, we show that it almost always carries strictly more information than the action on homology.

Theorem 1.3 *Let Γ be a graph with set of edges E and R a commutative ring. The singular chain complex of $B(\Gamma)$ with coefficients in R is formal as a differential bigraded $R[E]$ –module if and only if Γ is a small graph.*

Here we declare a graph to be *small* if, after smoothing as many bivalent vertices as possible, no vertex has three distinct edges; see [Definition 2.3](#). Small graphs are very primitive objects; indeed, each connected component of a small graph is an isolated vertex, an interval I , a cycle C , a lollipop L , a figure-eight 8 , or a handcuff H . See [Figure 2](#).

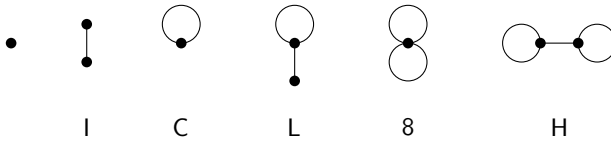


Figure 2: Homeomorphism types of connected small graphs.

This result is a measure of the inherent complexity of the family of graph braid groups of Γ . It also provides an explanation for this complexity, as all nonformal behavior arises from variations on the following simple example. The configuration space of two points in a star graph with three edges is homotopy equivalent to a circle. The fundamental class of this circle is represented by a cycle where the two points orbit one another by taking turns passing through the central vertex. This cycle is decomposable over the ring of edges, but the class itself is indecomposable, and nonformality follows.

1.4 Techniques and previous work

We prove our theorems by appealing to a small and explicit chain complex $S(\Gamma)$ with an action of $\mathbb{Z}[E]$ — see [Theorem 2.10](#) — which functorially computes the homology and nonfunctorially models the singular chains of $B(\Gamma)$, both as $\mathbb{Z}[E]$ –modules. In the proof of [Theorem 1.2](#), we exploit a family of spectral sequences computing the homology of this complex; see [Remark 2.14](#) and [Section 3](#). These spectral sequences are an expansion of the “vertex explosion” technique introduced by the authors in [\[2, Definition 5.14\]](#), and we expect them to prove useful in future computations.

The Świątkowski complex $S(\Gamma)$ first arose [\[27, Section 1\]](#) as the cellular chain complex of a cubical deformation retract of $B(\Gamma)$ and was later derived [\[2, Section 4.2\]](#) independently by the authors of this paper from ideas involved in factorization homology and discrete Morse theory. We use features of both approaches to $S(\Gamma)$, so we must compare them. This comparison, which appears as [Proposition 5.11](#), is both a key technical step and a unification of perspectives.

For the case of configuration spaces of trees, [Theorem 1.2](#) was proven by direct computation of all Betti numbers by Maciążek and Sawicki [\[20, Equation \(23\)\]](#) and by Ramos [\[24, Theorem A\]](#), who also stated the upper-bound half of the general case as a conjecture.

Ramos’ methods involved a phenomenon similar to edge stabilization. In his work, the action of $\mathbb{Z}[E]$ is purely algebraic, occurring at the level of discrete Morse complexes, and limited to trees. We expect the two actions on homology to coincide.

1.5 Questions

Our work invites the following questions.

- (1) Is there an analogue of edge stabilization for *ordered* configuration spaces of graphs?

Lütgehetmann and Recio-Mitter [19, Proposition 5.6] recently constructed a stabilization map on ordered configuration spaces of graphs on an edge near an essential vertex using other information at the vertex. The nature of the optimal algebraic structure organizing these maps remains unclear.

- (2) Are there analogues of edge stabilization for higher-dimensional cells?

One conceptual explanation for edge stabilization is that the unordered configuration spaces of an edge are all contractible, which guarantees that a certain extension problem is unobstructed. In the context of both of these questions, the corresponding extension problems are obstructed. This fact explains why Lütgehetmann and Recio-Mitter's stabilization involves additional information beyond the choice of an edge. It also indicates that stabilization for higher-dimensional cells would likely have to involve a similar choice of further local information or, more radically, some kind of nonlocal invariants.

- (3) What is the significance of the invariant Δ_{Γ}^i ? For example, is its role here connected to its appearance in the “cut polynomials” of right-angled Artin groups studied by Papadima and Suciu [23, Section 4.2]?
- (4) Is there a simple characterization, extending Theorem 1.3 and Proposition 4.9, of the subsets $E_0 \subseteq E$ for which formality holds over $\mathbb{Z}[E_0]$?
- (5) Our proof of Theorem 1.2 shows that the W -tori of Definition 3.11 account for a positive fraction of the homology of configuration spaces of graphs; see Remark 3.14 for further discussion of this point. Can this estimate be improved and/or extended to other families of generators?

1.6 Linear outline

The paper following the introduction is divided into five sections. In Section 2, we introduce edge stabilization and the Świątkowski complex, and we connect the two via the statement of Theorem 2.10, which implies Theorem 1.1. Section 3 is concerned with the lower and upper bounds on growth necessary to establish Theorem 1.2, and Section 4 assembles the proof of Theorem 1.3. Finally, in Sections 5 and 6, we return

to complete the proof of [Theorem 2.10](#), in the process forging a connection between the cellular models of Abrams and Świątkowski.

1.7 Conventions

Bigradings of modules are by *degree* and *weight*, and all are nonnegative. The braiding isomorphism for a tensor product of modules has a sign which depends on degree and not on weight: if x and y have degree i and j , the braiding isomorphism takes $x \otimes y$ to $(-1)^{ij} y \otimes x$. We write $[m]$ for the degree shift functor by m and $\{n\}$ for the weight shift functor by n so that the degree i and weight j homogeneous component of $M[m]\{n\}$ is the degree $i - m$ and weight $j - n$ homogeneous component of M . In a differential graded context, differentials preserve weight.

The category $\mathcal{M}\text{od}$ has objects pairs (R, M) , where R is a weight-graded commutative unital ring and M a differential bigraded R -module. A morphism in $\mathcal{M}\text{od}$ between (R_1, M_1) and (R_2, M_2) is a pair (f, g) , where f is a weight-graded ring morphism from R_1 to R_2 and g is a differential bigraded R_1 -module morphism from M_1 to f^*M_2 (that is, M_2 with the action $rm := f(r)m$).

Acknowledgements

An and Drummond-Cole were supported by IBS-R003-D1. Knudsen was supported by NSF award 1606422. The authors would like to thank Daniel Lütgehetmann, Hyo Won Park, and Eric Ramos for illuminating conversations.

2 Edge stabilization

After establishing terminology and notation regarding graphs, we define the edge stabilization map $\sigma_e: B(\Gamma) \rightarrow B(\Gamma)$, where e is an edge of the graph Γ . At the level of homology, these maps give rise to an action by the polynomial ring generated by the edges of Γ . In [Theorem 2.10](#), we present a small chain complex computing this homology, together with its module structure.

2.1 A category of graphs

A *graph* Γ is a finite one-dimensional CW complex. Its 0-cells and open 1-cells are its *vertices* and *edges*, and the set of such is denoted by $V(\Gamma)$ and $E(\Gamma)$, or simply V and E , respectively. A *half-edge* is an end of an edge, and the set of such is denoted by $H(\Gamma)$ or simply H . The vertices of an edge $V(e)$ are the vertices

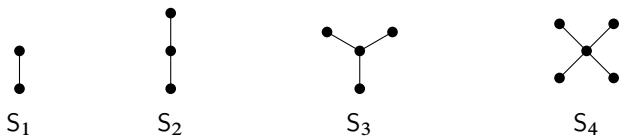


Figure 3: Star graphs.

contained in the closure of that edge in Γ . The edges of a vertex $E(v)$ are the edges incident to the vertex v . The half-edges of a vertex $H(v)$ (or of an edge $H(e)$) are the half-edges incident to v (contained in e). For h in H , we write $v(h)$ and $e(h)$ for the corresponding vertex and edge.

The *valence* of a vertex v is the cardinality of $H(v)$, denoted by $d(v)$. The vertex v is *isolated* if $d(v) = 0$ and *essential* if $d(v) \geq 3$. We shall sometimes write $V^{\geq 2}$ and V^{ess} for the set of vertices of valence at least 2 and the set of essential vertices, respectively. An edge with a 1-valent vertex is a *tail*. A *self-loop* at a vertex is an edge whose entire boundary is attached at that vertex.

Example 2.1 The cone on $\{1, \dots, n\}$ is a graph S_n with $n + 1$ vertices, with one of valence n and n of valence 1. These graphs are called *star graphs* and the cone point the *star vertex*. See Figure 3.

A *parametrization* of a graph Γ is a set of homeomorphisms $D_h: e(h) \rightarrow (0, 5)$ for $h \in H$ such that

- (1) D_h maps the h end of $e(h)$ to the 0 end of $(0, 5)$, and
- (2) if $h_1 \neq h_2 \in H(e)$, then $D_{h_2} = 5 - D_{h_1}$.

Making a choice of parametrization does not affect the homeomorphism type of configuration spaces. We will sometimes implicitly identify an edge of a parametrized graph with the interval $(0, 5)$. Up to homotopy, all constructions on parametrized graphs will be independent of the choice of parametrization.

Definition 2.2 Let $f: \Gamma_1 \rightarrow \Gamma_2$ be a continuous map between graphs. We say that f is a *graph morphism* if

- (1) the inverse image $f^{-1}(V(\Gamma_2))$ is contained in $V(\Gamma_1)$, and
- (2) the map f is injective.

We call a graph morphism a *smoothing* if it is a homeomorphism and a *graph embedding* if it preserves vertices. A graph morphism can be factored into a graph embedding



Figure 4: There is a graph morphism (in fact a smoothing) from left to right but not from right to left.

followed by a smoothing. At times, we may refer to the inverse of a smoothing as a *subdivision*, and we caution the reader that these are typically not graph morphisms.

The composite of graph morphisms is a graph morphism, and we obtain in this way a category $\mathcal{Gph}$. Although the objects of $\mathcal{Gph}$ are simply finite one-dimensional CW complexes, not all morphisms are cellular. A *subgraph* is the image of a graph embedding.

Since graph morphisms are injective, they induce maps at the level of configuration spaces. Thus, it is natural to view $H_*(B(-))$ as a functor from the category $\mathcal{Gph}$ to bigraded abelian groups, where the weight grading records the cardinality of a configuration. As we will see in the next section, there is more structure to be found.

We close this section with the following pair of definitions premised on our notion of a smoothing.

Definition 2.3 Let Γ be a graph.

- (1) We say that Γ is *smooth* if every smoothing with domain Γ is an isomorphism.
- (2) We say that Γ is *small* if, in any maximal smoothing of Γ , there is no vertex with three distinct edges. Otherwise, Γ is *large*.

Smoothness is almost, but not quite, equivalent to having no bivalent vertices; indeed, the cycle C (see Figure 2) is smooth.

2.2 Topological edge action

We now introduce the promised stabilization.

Definition 2.4 Let Γ be a parametrized graph and e an edge. *Edge stabilization* at e is the map $\sigma_e: B(\Gamma) \rightarrow B(\Gamma)$ that preserves partial configurations in the complement of $(2, 3) \subseteq e$ and replaces the partial configuration $\{x_1 \leq \dots \leq x_j\} \subseteq (2, 3)$ with

$$\left\{ \frac{1}{2}(2 + x_1), \frac{1}{2}(x_1 + x_2), \dots, \frac{1}{2}(x_{j-1} + x_j), \frac{1}{2}(x_j + 3) \right\}.$$

It is a direct verification that the map σ_e is continuous and independent of parametrization up to homotopy. See Figure 5 for a depiction of this map.

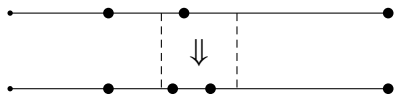


Figure 5: Edge stabilization redux.

Remark 2.5 Parametrizing edges by $(0, 5)$ and stabilizing in the subinterval $(2, 3)$ are choices designed to interface well with the arguments of Section 5. The more obvious choices of parametrizing edges by $(0, 1)$ and stabilizing in this entire interval (see Figure 1) lead to maps that differ only up to homotopy.

Remark 2.6 Stabilization by adding points to the tails of a graph has been considered [9, Section 5], and addition of points near the boundary of a manifold is a well-studied phenomenon; see [22, Section 4], for example. The existence of stabilization maps at internal edges is new, but see [24, Section 3.2] for a related algebraic stabilization mechanism in the context of trees.

Passing to homology, we obtain an action of the weight graded ring $\mathbb{Z}[E]$ on $H_*(B(\Gamma))$. This action is natural in the sense that a graph morphism from Γ_1 to Γ_2 induces a commutative diagram,

$$\begin{array}{ccc} \mathbb{Z}[E(\Gamma_1)] \otimes H_*(B(\Gamma_1)) & \longrightarrow & \mathbb{Z}[E(\Gamma_2)] \otimes H_*(B(\Gamma_2)) \\ \downarrow & & \downarrow \\ H_*(B(\Gamma_1)) & \longrightarrow & H_*(B(\Gamma_2)) \end{array}$$

In this way, the homology of configuration spaces of graphs lifts to a functor

$$H_*(B(-)): \mathcal{Gph} \rightarrow \mathcal{Mod}.$$

2.3 The Świątkowski complex

We now present a convenient chain model for $H_*(B(\Gamma))$, thought of as a functor valued in $\mathcal{Mod}$ via edge stabilization.

Definition 2.7 Let Γ be a graph and let R be a commutative ring. For $v \in V$, set $S(v) = \mathbb{Z}\langle \emptyset, v, h \in H(v) \rangle$. The *Świątkowski complex* of Γ (with coefficients in R) is the $R[E]$ -module

$$S(\Gamma; R) = R[E] \otimes_{\mathbb{Z}} \bigotimes_{v \in V} S(v),$$

endowed with the bigrading $|\emptyset| = (0, 0)$, $|v| = |e| = (0, 1)$, and $|h| = (1, 1)$, together with the differential ∂ determined by setting $\partial(h) = e(h) - v(h)$.

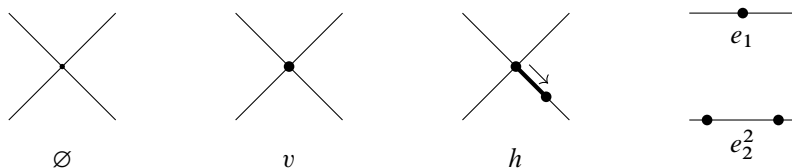


Figure 6: Local states near a vertex v and the state $e_1 e_2^2$.

Typically, the ring R will be $\mathbb{Z}$ or a field. When the coefficient ring is $\mathbb{Z}$, we omit it from the notation.

Remark 2.8 The following geometric heuristic is often helpful in dealing with the Świątkowski complex:

- (1) We think of a generator of $S(v)$ as prescribing the local “state” of a configuration near v . Specifically, the generator $\emptyset$ represents the absence of a particle at v , the generator v represents the presence of a stationary particle at v , and the generator $h \in H(v)$ represents an infinitesimal path of a particle exiting v in the direction of h . See Figure 6.
- (2) We think of a monomial in $\mathbb{Z}[E]$ as prescribing the state of a configuration on the edges of Γ ; for example, the generator $e_1 e_2^2$ represents the presence of a stationary particle on e_1 and two stationary particles on e_2 . See Figure 6.
- (3) A basis element of the Świątkowski complex is a prescription of a global state, which is to say that we choose a local state from each $S(v)$ and populate the edges with a monomial in E .
- (4) The first grading is the natural homological grading, since $h \in H$ is a path and all other generators involve stationary particles, and the second grading is given by counting the number of particles.
- (5) The differential takes a path to the difference of its endpoints.

Notation 2.9 (see Figure 7) We systematically omit all factors of $\emptyset$ and all tensor symbols from the notation, and we regard half-edge generators at different vertices as permutable up to sign. Thus, if Γ is the interval with vertices 0 and 1, half-edges h_0 and h_1 , and edge e , then we write

$$\begin{aligned} e^3 \otimes \emptyset \otimes \emptyset &= e^3, \\ e^3 \otimes \emptyset \otimes h_2 &= e^3 h_2, \\ e^3 \otimes h_1 \otimes h_2 &= e^3 h_1 h_2 = -e^3 h_2 h_1. \end{aligned}$$

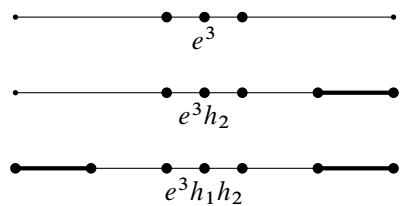


Figure 7: Three global states on the interval.

A graph morphism $f: \Gamma_1 \rightarrow \Gamma_2$ determines a map $S(f; R): S(\Gamma_1; R) \rightarrow S(\Gamma_2; R)$ as follows [2, Section 4.2]: an edge is sent to its image under f ; for a vertex v such that $f(v)$ is also a vertex, there is an evident map $S(v) \rightarrow S(f(v))$; and for a vertex such that $f(v)$ is an edge, we instead use the map $S(v) \rightarrow \mathbb{Z}[f(v)]$ sending $\emptyset$ to 1, v to e , and half-edges to 0. This map respects the bigrading, differential, and module structures, so we obtain a functor $S(-; R): \mathcal{Gph} \rightarrow \mathcal{Mod}$. Since ∂ is $R[E]$ –linear by definition, this module structure descends to homology.

Theorem 2.10 *There is a natural isomorphism*

$$H_*(B(\Gamma); R) \cong H_*(S(\Gamma; R))$$

of functors from $\mathcal{Gph}$ to $\mathcal{Mod}$.

At the level of bigraded abelian groups, this natural isomorphism was established in [2, Theorem 4.5]; see Theorem 5.4 below. To conclude the full statement over $\mathbb{Z}$, we will check that this isomorphism is compatible with the respective $\mathbb{Z}[E]$ –actions, a task that we take up in Section 5. The general case follows by the universal coefficients theorem. Since $S(\Gamma)$ is finitely generated over $\mathbb{Z}[E]$ by definition, and since $\mathbb{Z}[E]$ is Noetherian, Theorem 1.1 follows immediately from Theorem 2.10.

We close this section with an introduction of a smaller variant of the Świątkowski complex, which is often more convenient.

Definition 2.11 Let Γ be a graph. For each $v \in V$, let $\tilde{S}(v) \subseteq S(v)$ be the subspace spanned by $\emptyset$ and the differences $h_{ij} := h_i - h_j$ of half-edges. The *reduced Świątkowski complex* with coefficients in R is

$$\tilde{S}(\Gamma; R) := R[E] \otimes_{\mathbb{Z}} \bigotimes_{v \in V} \tilde{S}(v),$$

considered as a subcomplex and submodule of $S(\Gamma; R)$. To be explicit, the differential is determined by $\partial(h_{ij}) = e(h_i) - e(h_j)$.

The inclusion $\tilde{S}(\Gamma; R) \subseteq S(\Gamma; R)$ is an $R[E]$ -linear quasi-isomorphism. The $R = \mathbb{Z}$ case is [2, Proposition 4.9], and the general case follows by the universal coefficients theorem. Note that $\tilde{S}(-; R)$ is functorial for graph morphisms.

2.4 Vertex explosion and star classes

In this section, we review some tools from [2, Section 5] afforded by the Świątkowski complex, which will play an important role in the proofs of Theorems 1.2 and 1.3. The first of these is an exact sequence which is useful in reducing computations of $H_*(B(\Gamma))$ to computations for simpler graphs.

Definition 2.12 For $v \in V$, we write Γ_v for the *vertex explosion* of Γ at v , which is the graph obtained by

- (1) replacing the vertex v with $\{v\} \times H(v)$, and
- (2) modifying the attaching maps for half-edges at v by attaching h to (v, h) .

There is a graph morphism from Γ_v to Γ which takes each edge to itself, takes the vertex (v, h) to $e(h)$, and takes every other vertex to itself. Defining this morphism requires choices of precisely where in $e(h)$ to send (v, h) , but the isotopy class of this graph morphism is unique. See Figure 8.

Proposition 2.13 [2, Corollary 5.16] *Fix a half-edge $h_0 \in H(v)$. There is a short exact sequence of differential bigraded $R[E]$ -modules*

$$0 \rightarrow \tilde{S}(\Gamma_v; R) \rightarrow \tilde{S}(\Gamma; R) \xrightarrow{\psi} \bigoplus_{h \in H(v) \setminus \{h_0\}} \tilde{S}(\Gamma_v; R)[1]\{1\} \rightarrow 0,$$

where ψ sends an element of $\tilde{S}(\Gamma; R)$, written uniquely as

$$\beta + \sum_{h \in H(v) \setminus \{h_0\}} (h - h_0) \alpha_h$$

with β involving no half-edges incident on v , to $(\alpha_h)_{h \in H(v) \setminus \{h_0\}}$.

The result in [2] is only taken with coefficients in $\mathbb{Z}$. Since $\tilde{S}(\Gamma_v)$ is degreewise flat, tensoring with R preserves exactness.

In the homology long exact sequence corresponding to this short exact sequence of chain complexes, the connecting homomorphism δ from $\bigoplus H_*(\tilde{S}(\Gamma_v; R))\{1\}$ to $H_*(\tilde{S}(\Gamma_v; R))$ is given by the formula

$$\delta \beta_h = (e(h) - e(h_0)) \beta_h.$$

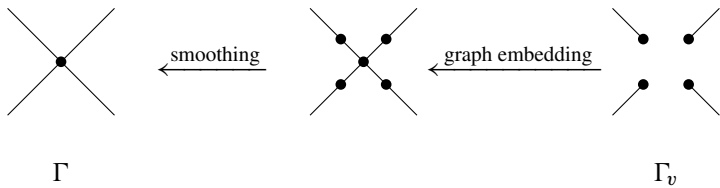


Figure 8: A local picture of vertex explosion along with an intermediate graph which admits a graph morphism from Γ_v and a smoothing to Γ .

Remark 2.14 This exact sequence is a degenerate example of a spectral sequence interpolating between the homology groups of the configuration spaces of Γ and those of the graph obtained by exploding a specified collection of vertices. This type of spectral sequence will play an important role in [Section 3](#).

The second tool is a type of atomic homology class. Recall that the configuration space of two points in the star S_3 is homotopy equivalent to a circle.

Definition 2.15 [\[2, Section 5.1\]](#) Let $S_3 \rightarrow \Gamma$ be a graph morphism. A *star class* is a class in $H_1(B_2(\Gamma); R)$ which is the image under the induced morphism of the generator α in $H_1(B_2(\Gamma); R) \simeq R$ represented by the chain $a \in S(S_3; R)$ given by

$$a := e_1(h_2 - h_3) + e_2(h_3 - h_1) + e_3(h_1 - h_2),$$

where h_i is the half-edge of e_i adjacent to the star vertex of S_3 .

We refer to the chain a , and to its image, as the *standard representative* of α . See [Figure 9](#). Given m graph morphisms $S_3 \rightarrow \Gamma$ with pairwise disjoint images, we obtain a degree m homology class, called the *external product* of the star classes [\[2, Definition 5.10\]](#), whose standard representative is the tensor product of the standard representatives of the factors.

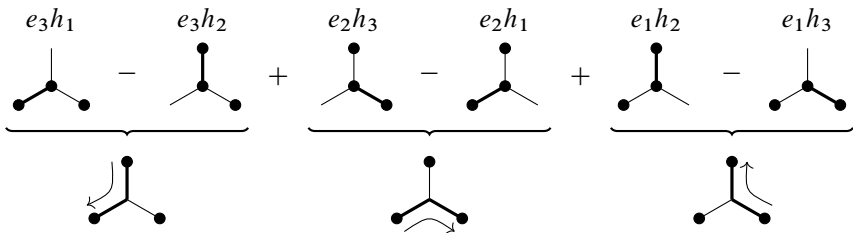


Figure 9: The standard representative of a star class.

3 Growth of Betti numbers

Theorem 1.1 implies that the $\mathbb{F}$ -Betti numbers of $B_k(\Gamma)$ are eventually given by a polynomial in k for any field $\mathbb{F}$ (see eg [7, Theorem 1.11]). In this section, we determine the exact degree of this polynomial.

3.1 Connectivity and growth

We write $D_\Gamma^i(\mathbb{F})$ for the degree of eventual polynomial growth of the dimension over $\mathbb{F}$ of $H_i(B_k(\Gamma); \mathbb{F})$ in the weight k . By convention, $D_\Gamma^i(\mathbb{F}) = -\infty$ if $H_i(B_k(\Gamma); \mathbb{F}) = 0$ for $k \gg 0$.

This degree of growth is controlled by a certain elementary connectivity invariant. In order to define this invariant, we introduce the following equivalence relation.

Definition 3.1 (see Figure 10) Let Γ be a smooth graph. A subset $W \subseteq V^{\geq 2}$ determines an equivalence relation $\sim_W$ on the edges of Γ , where $e \sim_W e'$ if and only if $[e] = [e']$ in $\pi_0(\Gamma \setminus W)$. Writing $E_W = E / \sim_W$ for the set of equivalence classes, we define

$$\Delta_\Gamma^W = |E_W| \quad \text{and} \quad \Delta_\Gamma^i = \max \Delta_\Gamma^W,$$

where the maximum is taken over subsets W of cardinality i of $V^{\geq 2}$. We use the convention that $\Delta_\Gamma^i = -\infty$ if $i > |V^{\geq 2}|$. If Γ is not smooth, define Δ_Γ^i as $\Delta_{\Gamma'}^i$, where Γ' is a smooth graph homeomorphic to Γ .

Remark 3.2 Although we work with a different set of conventions, this invariant was essentially defined by Ramos [24, page 2306], who conjectured Proposition 3.9 below.

Example 3.3 Here are some basic examples of the behavior of the invariant Δ_Γ^i :

- (1) If Γ is an isolated vertex, then $\Delta_\Gamma^0 = 0$.
- (2) If Γ is an isolated edge, then $\Delta_\Gamma^0 = 1$.
- (3) If Γ is a cycle, then $\Delta_\Gamma^0 = \Delta_\Gamma^1 = 1$.

In these examples, $\Delta_\Gamma^i = -\infty$ for all other values of i .

- (4) For $\Gamma = \Gamma_1 \sqcup \Gamma_2$ a disjoint union of (not necessarily connected) graphs,

$$\Delta_\Gamma^i = \max_{a+b=i} (\Delta_{\Gamma_1}^a + \Delta_{\Gamma_2}^b).$$

We recall Theorem 1.2:

Degree Theorem If Γ has at least one edge, then $D_\Gamma^i(\mathbb{F}) = \Delta_\Gamma^i - 1$.

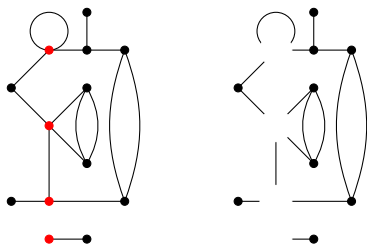


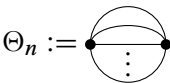
Figure 10: A graph with a choice of W in red, and the seven equivalence classes of edges under $\sim_W$.

Remark 3.4 A statement valid for an arbitrary graph is available for the sum

$$\sum_{\ell=0}^k \dim_{\mathbb{F}} H_i(B_{\ell}(\Gamma); \mathbb{F}),$$

which exhibits eventual polynomial growth of degree Δ_{Γ}^i .

Example 3.5 For $n \geq 3$, let Θ_n denote the Theta- n graph, which is the double cone on n points, as shown:



It is known that

$$\dim_{\mathbb{F}} H_i(B_k(\Theta_n); \mathbb{F}) = \begin{cases} 1 & \text{if } i = 0, \\ \binom{n}{2} & \text{if } i = 1, \\ \binom{n}{2} - 1 + \sum_{j=0}^2 (1-n)^j \binom{2}{j} \binom{k-j+n-1}{n-1} & \text{if } i = 2. \end{cases}$$

Calculating the zeroth Betti number is trivial, the first is given in [16, Lemma 3.14], and the second follows from these and knowledge of the Euler characteristic for $B_k(\Theta_n)$, which can be extracted from [10, Theorem 2] and is given explicitly in [2, Corollary 5.11]. These Betti numbers exhibit polynomial growth of degree 0 for $i \leq 1$ and $n - 1$ for $i = 2$, and these numbers are one less than the respective numbers of components obtained by removing i vertices. Thus, the theorem holds in this example.

Example 3.6 For the complete graph K_4 , the explicit formula for the i^{th} Betti number $\dim_{\mathbb{F}} H_i(B_k(K_4); \mathbb{F})$ in terms of k can be computed directly from [2, Section 5.6]. Table 1 shows the growth $D_{K_4}^i(\mathbb{F})$ of the Betti number and the invariant $\Delta_{K_4}^i$, verifying Theorem 1.2 for K_4 .

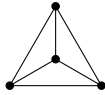
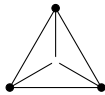
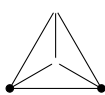
i	$\dim_{\mathbb{F}} H_i(B_k(K_4); \mathbb{F})$	$D_{K_4}^i(\mathbb{F})$	$\Delta_{K_4}^i$	$(K_4)_W$
0	1	0	1	
1	4 (for $k > 1$)	0	1	
2	$6k - 15$ (for $k > 2$)	1	2	
3	$4 \binom{k-3}{3}$	3	4	
4	$\binom{k-3}{5}$	5	6	
≥ 5	0	$-\infty$	$-\infty$	

Table 1: For the complete graph K_4 , a table containing the Betti number $\dim_{\mathbb{F}} H_i(B_k(K_4); \mathbb{F})$, the growth rate $D_{K_4}^i(\mathbb{F})$, the invariant $\Delta_{K_4}^i$, and the corresponding complement $(K_4)_W$.

Definition 3.7 We say that a graph Γ is *normal* if it is connected and smooth and has an essential vertex.

The bulk of the theorem is contained in the following two results, whose proofs will occupy the rest of the section. The second is essentially Conjecture 4.3 of [24].

Proposition 3.8 If Γ is normal, then $D_{\Gamma}^i(\mathbb{F}) \geq \Delta_{\Gamma}^i - 1$.

Proposition 3.9 If Γ is normal, then $D_{\Gamma}^i(\mathbb{F}) \leq \Delta_{\Gamma}^i - 1$.

Assuming these results for the moment, we complete the proof.

Proof of Theorem 1.2 We may assume that Γ is smooth. Since the presence or absence of finitely many isolated vertices in Γ does not change the eventual growth rate, we may further assume that each connected component of Γ contains an edge. Thus, each connected component of Γ is either normal, an interval, or a cycle. The

theorem is known for each of these components; indeed, Propositions 3.8 and 3.9 supply the normal case, and the other cases are classical. This observation forms the base case for an induction on $|\pi_0(\Gamma)|$.

For the inductive step, we make use of “big theta” notation [14, page 19] defined as

$$f(k) = \Theta(g(k)) \iff \exists c, C > 0 \quad \forall k \gg 0 \quad c \cdot g(k) \leq f(k) \leq C \cdot g(k).$$

Consider a decomposition $\Gamma = \Gamma_1 \sqcup \Gamma_2$ of a smooth graph into two nonempty graphs, and assume that the degree theorem holds for each of Γ_1 and Γ_2 . By the Künneth theorem, we have

$$\begin{aligned} \dim_{\mathbb{F}}(H_i(B_k(\Gamma); \mathbb{F})) &= \sum_{\ell=0}^k \sum_{a+b=i} \dim_{\mathbb{F}}(H_a(B_{\ell}(\Gamma_1); \mathbb{F})) \dim_{\mathbb{F}}(H_b(B_{k-\ell}(\Gamma_2); \mathbb{F})) \\ &= \sum_{a+b=i} \Theta \left(\sum_{\ell=0}^k \ell^{\Delta_{\Gamma_1}^a - 1} (k - \ell)^{\Delta_{\Gamma_2}^b - 1} \right) \\ &= \sum_{a+b=i} \Theta \left(k^{\Delta_{\Gamma_1}^a + \Delta_{\Gamma_2}^b - 1} \sum_{\ell=0}^k \frac{1}{k} \left(\frac{\ell}{k} \right)^{\Delta_{\Gamma_1}^a - 1} \left(1 - \frac{\ell}{k} \right)^{\Delta_{\Gamma_2}^b - 1} \right) \\ &= \sum_{a+b=i} \Theta(k^{\Delta_{\Gamma_1}^a + \Delta_{\Gamma_2}^b - 1}) \int_0^1 x^{\Delta_{\Gamma_1}^a - 1} (1 - x)^{\Delta_{\Gamma_2}^b - 1} dx, \end{aligned}$$

where the second equality uses our assumption on Γ_1 and Γ_2 . By our assumption on the connected components of Γ , both $\Delta_{\Gamma_1}^a$ and $\Delta_{\Gamma_2}^b$ are nonzero for every a and b , so the integral shown is a nonnegative rational number independent of k . The largest exponent of k present in this sum is

$$\max_{a+b=i} (\Delta_{\Gamma_1}^a + \Delta_{\Gamma_2}^b - 1) = \Delta_{\Gamma}^i - 1,$$

as desired; see Example 3.3. □

3.2 Lower bound

To prove Proposition 3.8, we will define a special kind of homology class, called a W -torus, depending on a set of vertices W . For a suitable choice of W , we will show that a W -torus exists and generates an $\mathbb{F}[E]$ -submodule with polynomial growth of the expected degree.

Throughout this section, we assume that Γ is normal. Given a set W of vertices, we write Γ_W for the graph obtained from Γ by successive explosion of each vertex of W .

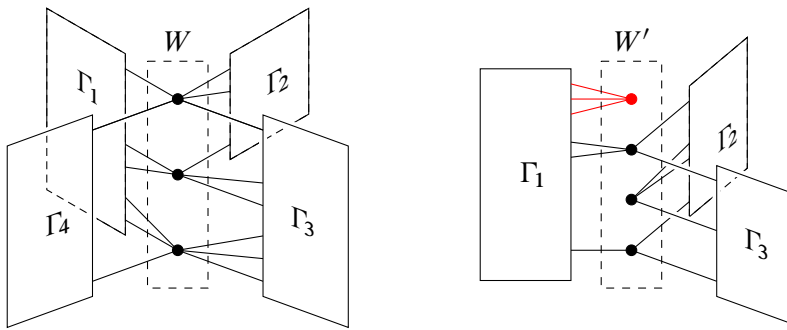


Figure 11: A subset W that is well-separating and a subset W' that is not well-separating (assuming $\Gamma_1 \setminus W'$ is connected).

Definition 3.10 A subset $W \subseteq V^{\text{ess}}$ is *well-separating* if the open star of each $v \in W$ intersects more than one connected component of Γ_W .

The empty set is vacuously well-separating. We fix a field $\mathbb{F}$ and a subset $W \subseteq V^{\text{ess}}$ for the remainder of this section, and we write $F_\ell H_*(B(\Gamma); \mathbb{F})$ for the filtration on homology induced by filtering $\tilde{S}(\Gamma; \mathbb{F})$ by the number of half-edge generators at vertices of W .

Definition 3.11 A class $\alpha \in H_{|W|}(B(\Gamma); \mathbb{F})$ is a *W -torus* if α is the external product of classes $\{\alpha_v\}_{v \in W}$, where α_v is a star class at v . We call α_v a *star factor* of α . We further say that α is *rigid* if W is well-separating and α lies in $F_{|W|} H_{|W|}(B(\Gamma); \mathbb{F})$ but not in $F_{|W|-1} H_{|W|}(B(\Gamma); \mathbb{F})$.

Example 3.12 There is a unique $\emptyset$ -torus, namely the class of the empty configuration in $H_0(B_0(\Gamma); \mathbb{F})$, and this class is rigid since $F_{-1} H_0(B(\Gamma); \mathbb{F}) = 0$.

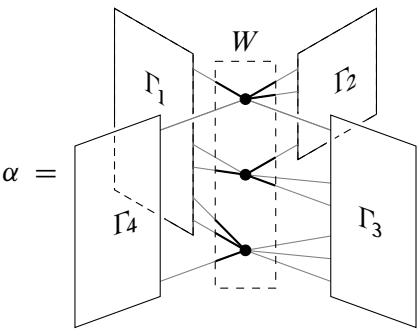


Figure 12: A W -torus which [Lemma 3.15](#) shows to be rigid.



Figure 13: The theta graph Θ_3 and a graph Θ_3^+ with a well-separating vertex set $W = \{v'\}$ and a nonrigid W -torus (indicated in bold).

Example 3.13 Consider the theta graph Θ_3 , which is the double cone on three points with vertices the cone points v and w as depicted in Figure 13. According to the Θ -relation [2, Lemma 5.8(4) and Definition 5.9(4)], any star class α_v at v is equal to some star class α_w at w . We have defined rigidity only for well-separating W because if W is not well-separating in Γ , pulling the Θ -relation back along a smoothing $\Theta'_3 \rightarrow \Theta_3$ and then pushing forward along a graph embedding $\Theta'_3 \rightarrow \Gamma$ shows that no W -torus could ever have a rigidity property.

We can use the same strategy to give examples of nonrigid W -tori in cases where W is well-separating. To wit, consider the graph Θ_3^+ of Figure 13. There is an evident graph morphism ι from Θ_3 to Θ_3^+ . The vertex set $W = \{v'\}$ is well-separating in Θ_3^+ but the star class $\iota(\alpha_v)$ is not rigid because it is equal to the star class $\iota(\alpha_w)$ by pushing forward the Θ -relation.

Remark 3.14 (universal generators) The definition and use of W -tori can be viewed as an outgrowth of the study of an important question in configuration spaces of graphs.

This question is to determine a *universal presentation* for the homology of configuration spaces of graphs (or of specific classes of graphs). By a universal presentation we mean

- (1) a set of homology classes (*universal generators*) which generate all homology groups of configuration spaces of graphs under pushforward along graph morphisms, edge stabilization, and external products;
- (2) a set of *universal relations* which generate all relations among universal generators under the same three operations.

For example, the empty configuration is a universal generator for degree 0 homology while star classes and *loop classes* (see [2, Definition 5.3]) are universal generators for degree 1 homology. The proof of Theorem 1.2 for the special case of trees in [20, Theorem V.3] amounts to showing that star classes are in fact universal generators for all homology groups for trees (it is known that this is not true for general graphs). In later work, the same authors show that the same statement holds for wheel graphs [21, Section 5.4].

One small step in the study of universal presentations is to ask quantitatively how much of the homology for a given graph or class of graphs is generated by such tori in general. Our proof of [Proposition 3.8](#) implies that for any graph and in any homological degree the amount of homology generated by such tori makes up a positive proportion of all homology as the weight increases. But this is a very crude estimate and it would be useful to have further information about this in order to know where to look for new universal generators and relations.

Set $R_W = \mathbb{F}[E_W]$ and write π_W for the projection from $\mathbb{F}[E]$ to R_W . If α is a W -torus, then the surjection $\mathbb{F}[E]\langle\alpha\rangle \rightarrow \mathbb{F}[E] \cdot \alpha$ factors through π_W , since any two edges in a connected component of Γ_W may be connected by a path of edges disjoint from the standard representative of α .

Lemma 3.15 *Let W be well-separating and let α be a W -torus. The following are equivalent:*

- (1) *The W -torus α is rigid.*
- (2) *The action of $\mathbb{F}[E]$ induces an isomorphism*

$$R_W\langle\alpha\rangle \xrightarrow{\cong} \mathbb{F}[E] \cdot \alpha.$$

- (3) *The standard representative of every star factor α_v involves edges in at least two distinct connected components of Γ_W .*

The third condition is a little technical but is also easier than the other two to check. In some sense it says that to guarantee rigidity, it is enough to avoid the class of “obvious” pitfalls illustrated in [Example 3.13](#). In particular, it has the following corollary.

Corollary 3.16 *For any well-separating subset W , there is a rigid W -torus.*

Proof Choose a star class α_v at each $v \in W$. Since W is well-separating, the star class α_v may be chosen to satisfy the star factor condition (3). $\square$

Remark 3.17 For an arbitrary set of vertices W , one could instead filter according to a maximal well-separating subset $W_0 \subseteq W$. The resulting generalized rigid W -tori do not exist for arbitrary W , $\mathbb{F}$, and Γ ; however, if $\mathbb{F} = \mathbb{Z}/2\mathbb{Z}$ or if Γ is planar, then such classes exist for any W . To construct one, use [Corollary 3.16](#) to generate a rigid W_0 -torus and form the external product with a star class at each vertex in $W \setminus W_0$. Either assumption implies that this external product is nonzero, and the proof of rigidity proceeds along similar lines.

In the situation of interest, we can guarantee that W is well-separating.

Lemma 3.18 *If $W \subseteq V^{\text{ess}}$ with $|W| = i$ is such that $\Delta_\Gamma^W = \Delta_\Gamma^i$, then either W is well-separating or $i = \Delta_\Gamma^i = 1$.*

Assuming these lemmas, we can establish the lower bound.

Proof of Proposition 3.8 Let $W \subseteq V^{\text{ess}}$ with $|W| = i$ be such that $\Delta_\Gamma^W = \Delta_\Gamma^i$. Except in the special case $i = \Delta_\Gamma^i = 1$, Lemma 3.18 guarantees that W is well-separating, so Corollary 3.16 supplies a rigid W -torus α . The dimension of $H_i(B_k(\Gamma); \mathbb{F})$ is no less than that of the submodule $\mathbb{F}[E] \cdot \alpha$, which is isomorphic to $R_W \langle \alpha \rangle$ by Lemma 3.15. Suppose α is in $H_i(B_{k_0}(\Gamma); \mathbb{F})$. Then the dimension over $\mathbb{F}$ of the space of monomials in R_W of weight $k - k_0$ is exactly $\binom{\Delta_\Gamma^W + (k - k_0) - 1}{\Delta_\Gamma^W - 1}$ and hence $\mathbb{F}[E] \cdot \alpha$ has polynomial growth of degree $\Delta_\Gamma^W - 1$.

For the special case $i = \Delta_\Gamma^i = 1$, we must verify that $H_1(B_k(\Gamma); \mathbb{F})$ is eventually nonzero, which is well known; see [16, Theorem 3.16], for example. $\square$

It remains to prove Lemmas 3.15 and 3.18. In the proof, we make use of the spectral sequence arising from the filtration $F_\ell H_*(B(\Gamma); \mathbb{F})$ introduced above, which we denote by $E_{p,q}^r$. This spectral sequence is a spectral sequence of $\mathbb{F}[E]$ -modules.

Proof of Lemma 3.15 The unique $\emptyset$ -torus is rigid, and clearly $\mathbb{F}[E] \cdot [\emptyset] \cong R_\emptyset$. In the remainder of the proof, we take $W \neq \emptyset$.

We prove first that (1) implies (2). Suppose that α is rigid, and write

$$[\alpha] = \alpha + F_{|W|-1} H_{|W|}(B(\Gamma); \mathbb{F}) \in E_{|W|,0}^\infty$$

for the resulting element in the associated graded. Since α is rigid, $[\alpha]$ cannot be zero. We wish to show that the left-hand map in the composite

$$R_W \langle \alpha \rangle \rightarrow \mathbb{F}[E] \cdot \alpha \rightarrow \mathbb{F}[E] \cdot [\alpha]$$

is an isomorphism. Since this map is surjective, it suffices to show that the composite is injective. We will show that $E_{|W|,0}^\infty$ is a submodule of a free R_W -module $R_W \langle X \rangle$ for some set X . Writing $\{[\alpha]_x\}_{x \in X}$ for the coordinates of $[\alpha]$ in this basis, the kernel of the map in question is $\bigcap_{x \in X} \text{Ann}_{R_W}([\alpha]_x)$. Since $[\alpha] \neq 0$ and R_W has no zero divisors, one of the terms of this intersection is zero, implying the claim.

Our spectral sequence is concentrated in the first quadrant, and therefore we have the containment $E_{|W|,0}^\infty \subseteq E_{|W|,0}^2$, and since $E_{|W|+1,0}^1 = 0$, we have the further

containment $E^2_{|W|,0} \subseteq E^1_{|W|,0}$. Both are containments of $\mathbb{F}[E]$ -submodules. Since Γ has no isolated vertices, there is a $\mathbb{F}[E]$ -linear isomorphism

$$E^1_{|W|,0} \cong H_0(B(\Gamma_W); \mathbb{F}) \otimes \mathbb{F}\langle X \rangle \cong R_W \otimes \mathbb{F}\langle X \rangle,$$

where X is the set of generators of $\tilde{S}(\Gamma)$ of the form $\bigotimes_{v \in W} h^v$ with h^v a difference of half-edge generators at v . This module is the desired R_W -free module.

We prove that (2) implies (3) by proving the contrapositive. Let α_v be a star factor represented by a graph morphism $\iota: S_3 \rightarrow \Gamma$ whose image intersects only one connected component of Γ_W , and extend ι to a graph morphism $\hat{\iota}: \Theta'_3 \rightarrow \Gamma$ with the same property, where Θ'_3 is a subdivision of Θ_3 (in other words, there is a smoothing $f: \Theta'_3 \rightarrow \Theta_3$). By pulling back via f the Θ -relation on Θ_3 as seen in Example 3.13, we have the Θ -relation on Θ'_3 as well and so $\alpha_v = \alpha_w$ for some star class α_w at w if we write w for the image of the other essential vertex of Θ'_3 under $\hat{\iota}$. Replacing α_v by α_w in the external product defining α , we conclude that α has a representative a that does not involve v . Since W is well-separating, there are edges e_i and e_j at v lying in distinct connected components of Γ_W , and appending the generator h_{ij} to a produces an element whose boundary is $(e_i - e_j)\alpha$. Thus, the map $R_W\langle\alpha\rangle \rightarrow \mathbb{F}[E] \cdot \alpha$ has nonzero kernel.

Finally, we prove that (3) implies (1). Throughout, products and tensor products are indexed by W . At each $v \in W$, choose a basis for the half-edge generators at v such that α_v is represented by $h^v_{12}(e^v_1 - e^v_3) - h^v_{13}(e^v_1 - e^v_2)$. By our assumption (3), we may assume that e^v_1 and e^v_3 intersect distinct connected components of Γ_W . We claim that $\bigotimes h^v_{12}$ appears with a nonzero coefficient in $\mathbb{F}[E]$ in every chain representative of α , implying in particular that $\alpha \notin F_{|W|-1}H_{|W|}(B(\Gamma); \mathbb{F})$, as desired.

We begin by examining the coefficient $\prod(e^v_1 - e^v_3)$ of $\bigotimes h^v_{12}$ in the standard representative a of α . We apply π_W to this coefficient to obtain an element of R_W . By construction, $\pi_W(e^v_1 - e^v_3) \neq 0$ for each v . Choose an ordering of $\pi_0(\Gamma_W)$ and write $\pi_W(e^v_1 - e^v_3) = \pm(\Delta^v_i - \Delta^v_j)$ with $\Delta^v_i < \Delta^v_j$ in the ordering. The leading term $\prod \Delta^v_i$ appears with coefficient ± 1 in the polynomial $\pi_W(\prod(e^v_1 - e^v_3))$, so $\prod(e^v_1 - e^v_3) \neq 0$, establishing the claim for the standard representative a .

Suppose now that $a - a' = \partial c$. In our preferred basis for the tensor product, c is necessarily a sum of elementary tensors of $|W| + 1$ half-edge generators, some of which are of the form

(1)
$$p(E)h_{ij} \otimes \bigotimes h^v_{12}$$

for some half-edge generator h_{ij} at a vertex away from W . Terms not of this form contain strictly fewer than $|W|$ tensor factors of the form h_{12}^v , and the boundary of such a term is a sum of terms containing strictly fewer than $|W|$ tensor factors of the form h_{12}^v . On the other hand, the boundary of (1) is a sum of terms containing strictly fewer than $|W|$ tensor factors of the form h_{12}^v , together with the term

$$p(E)(e_i - e_j) \otimes \bigotimes h_{12}^v.$$

By our assumption on h_{ij} , the edges e_i and e_j lie in the same connected component of Γ_W , so $\pi_W(p(E)(e_i - e_j)) = 0$. Thus, the coefficient of $\bigotimes h_{12}^v$ in a' coincides with the coefficient in a modulo $\ker \pi_W$, so this coefficient is nonzero by the previous paragraph. $\square$

Remark 3.19 The same techniques serve to establish a version of Lemma 3.15 for toric classes with some factors given by loop classes.

Proof of Lemma 3.18 If $|W| = 0$, or if $|W| = 1$ and $\Delta_\Gamma^W > 1$, then the claim is obvious. If $|W| > 1$ and W is not well-separating, then there is a vertex $v \in W$ whose open star intersects only one connected component of Γ_W . We claim that there is a second vertex $v' \in V^{\text{ess}} \setminus W$ lying in this component of Γ_W and sharing an edge e with some vertex of $W \setminus \{v\}$.

Given such a vertex, we replace v with v' in W to obtain a set W' with $\Delta_\Gamma^{W'} \geq \Delta_\Gamma^W + 1$. Indeed, the relations $\sim_W$ and $\sim_{W \setminus \{v\}}$ are identical, but the equivalence class of e under the relation $\sim_W$ is split into at least two distinct equivalence classes under $\sim_{W'}$. Thus, Δ_Γ^W does not achieve the maximum value Δ_Γ^i .

Assume for contradiction that such a v' does not exist, and suppose there is a reduced edge path in Γ beginning at v , terminating at some other vertex of W , and passing through no third vertex of W . Such a path is necessarily contained in the connected component of Γ_W containing the open star of v . Since this path is reduced and Γ has no bivalent vertices, the path passes only through essential vertices. We obtain a contradiction of the nonexistence of v' unless all such paths have length one; in other words, every edge at v terminates either at a vertex of W or a vertex of valence 1. Therefore, either the valence of v is 1, or the open star of v intersects more than one component of Γ_W . Either is a contradiction, so we conclude that no edge path exists between v and any other vertex of W . Since $|W| > 1$, it follows that Γ is disconnected, a contradiction. $\square$

3.3 Upper bound

Let Γ be a graph and $\mathbb{F}$ be a field. Given a subset $W \subseteq V^{\text{ess}}$ and a vertex v , we write $E_W(v) \subseteq E_W$ for the set of equivalence classes of edges adjacent to $v \in V$. We set

$$\tilde{S}_W(\Gamma; \mathbb{F}) := \tilde{S}(\Gamma_W; \mathbb{F}) \otimes_{\mathbb{F}[E]} \mathbb{F}[E_W].$$

The action of $\mathbb{F}[E]$ on $\mathbb{F}[E_W]$ is weight-respecting, so the complex $\tilde{S}_W$ is weight-graded. Since $\tilde{S}_\emptyset$ is simply the reduced Świątkowski complex over $\mathbb{F}$, [Proposition 3.9](#) is a special case of the following result.

Proposition 3.20 *Let Γ be smooth and contain no isolated vertices. For any $W \subseteq V^{\text{ess}}$, the dimension of $H_i(\tilde{S}_W(\Gamma; \mathbb{F}))$ is eventually polynomial in the weight of degree at most $\Delta_\Gamma^i - 1$.*

Remark 3.21 Geometrically, the weight k subcomplex of $\tilde{S}_W(\Gamma)$ corresponds to the space of configurations of k points in Γ which are permitted to collide at vertices in W ; see [\[4, Section 2\]](#) and [\[25, Section 1.1\]](#).

Note in particular that the behavior of $\tilde{S}_W(\Gamma)$ is different from that of $\tilde{S}(\Gamma_W)$. In both cases, there is no local homological information near W , but particles pass freely through vertices in W in the former while in the latter they avoid W . This is easiest to see in the star graph S_n , where the former models a contractible space and the latter a space with n contractible components.

The strategy of the proof of [Proposition 3.20](#) will be to show that the desired growth rate is already achieved on the E^2 page of a certain spectral sequence converging to the homology of $\tilde{S}_W(\Gamma; \mathbb{F})$. In order to introduce this spectral sequence, we require the following notation. Given a vertex $v \in V \setminus W$, we write S_v for a star graph of maximal valence equipped with a graph morphism $\iota: S_v \rightarrow \Gamma$ sending the star vertex to v . Setting $R_{W,v} = \mathbb{F}[E_W(v)]$, we write $C_{W,v}$ for the chain complex

$$C_{W,v} := \tilde{S}(S_v; \mathbb{F}) \otimes_{\mathbb{F}[E(S_v)]} R_{W,v}$$

of $R_{W,v}$ -modules. The chain complex $C_{W,v}$ is finitely generated over $R_{W,v}$ because $\tilde{S}(S_v; \mathbb{F})$ is finitely generated over $\mathbb{F}[E(S_v)]$.

Lemma 3.22 *Let v be an essential vertex of $V \setminus W$ in Γ . There is a convergent, homological, weight-graded spectral sequence of $\mathbb{F}[E_W]$ -modules,*

$$E_{p,q}^2 \cong H_p(\tilde{S}_W(\Gamma_v; \mathbb{F}) \otimes_{R_{W,v}} H_q(C_{W,v})) \Rightarrow H_{p+q}(\tilde{S}_W(\Gamma; \mathbb{F})).$$

$$\begin{array}{ccccccc} E_{0,1}^0 & \xleftarrow{d^1} & \cdots & \xleftarrow{d^1} & E_{p-1,1}^0 & \xleftarrow{d^1} & E_{p,1}^0 & \xleftarrow{d^1} & \cdots & \xleftarrow{d^1} & E_{|V|,1}^0 \\ \downarrow d^0 & & & & \downarrow d^0 & & \downarrow d^0 & & & & \downarrow d^0 \\ E_{0,0}^0 & \xleftarrow{d^1} & \cdots & \xleftarrow{d^1} & E_{p-1,0}^0 & \xleftarrow{d^1} & E_{p,0}^0 & \xleftarrow{d^1} & \cdots & \xleftarrow{d^1} & E_{|V|,0}^0 \end{array}$$

Figure 14: The potentially nonzero entries of the E^0 page of the spectral sequence for the bicomplex $(\tilde{S}_W(\Gamma; \mathbb{F}), \partial_v, \partial_-)$.

Proof We write the differential of $\tilde{S}_W(\Gamma; \mathbb{F})$ as $\partial = \partial_v + \partial_-$, where ∂_v is the sum of the terms of the differential involving half-edges at v . These two operators square to zero individually and commute, giving $\tilde{S}_W(\Gamma; \mathbb{F})$ the structure of a bicomplex of $\mathbb{F}[E_W]$ -modules. The desired spectral sequence is the spectral sequence of this bicomplex with zeroth differential ∂_v and first differential ∂_- ; see Figure 14. Since this bicomplex is concentrated in finitely many bidegrees, the spectral sequence collapses and in particular converges.

Using the decomposition $\tilde{S}(\Gamma) \cong \tilde{S}(\Gamma_v) \otimes_{\mathbb{Z}[E(v)]} \tilde{S}(S_v)$, we obtain the isomorphism

$$E^0 \cong (\tilde{S}_W(\Gamma_v; \mathbb{F}) \otimes_{R_{W,v}} C_{W,v}, 1 \otimes \partial)$$

of trigraded $\mathbb{F}[E_W]$ -modules, where the homological bidegree on the right-hand side is the natural bigrading of the tensor product. Since $\tilde{S}_W(\Gamma_v; \mathbb{F})$ is $\mathbb{F}[E_W]$ -free, and hence $R_{W,v}$ -free, the Künneth isomorphism gives

$$E^1 \cong (\tilde{S}_W(\Gamma_v; \mathbb{F}) \otimes_{R_{W,v}} H_*(C_{W,v}), \partial \otimes 1),$$

completing the proof. □

The key to the desired growth estimate is the following technical lemma.

Lemma 3.23 *Let (M, ∂_M) be a differential bigraded $R_{W,v}$ -module and N a bigraded $R_{W,v}$ -module with N finitely generated and concentrated in strictly positive homological degrees. If $\dim_{\mathbb{F}} H_j(M)$ is eventually polynomial in weight of degree at most d for every $0 \leq j < i$, then $\dim_{\mathbb{F}} H_i(M \otimes_{R_{W,v}} N, \partial_M \otimes 1)$ is as well.*

We assume this result for the moment.

Proof of Proposition 3.20 Let Γ be smooth with no isolated vertices. We proceed by induction on $|V^{\text{ess}} \setminus W|$.

For the base case $W = V^{\text{ess}}$, the complex $\tilde{S}_W(\Gamma; \mathbb{F})$ is isomorphic to $\mathbb{F}[E_W]$ concentrated in degree 0. In weight k , the dimension of this vector space is the number of

ways of putting k indistinguishable balls in E_W distinct bins, which is polynomial of degree $E_W - 1 = \Delta_\Gamma^{|W|} - 1$, as desired.

For the inductive step, suppose the statement has been shown true for $|V^{\text{ess}} \setminus W| < r$ and let $|V^{\text{ess}} \setminus W| = r > 0$. Fixing a vertex v in $V^{\text{ess}} \setminus W$, it will suffice to show that the $E_{i,0}^2$ and $E_{i-1,1}^2$ entries of the spectral sequence of Lemma 3.22 each have eventual polynomial growth of degree at most $\Delta_\Gamma^i - 1$ (since $C_{W,v}$ is concentrated in degrees 0 and 1, these are the only nonzero entries in the appropriate degree on the E^2 page).

We note first that $H_0(C_{W,v})$ is one-dimensional in each weight, since there is a degree 1 chain interpolating between any two edges of S_v . Thus,

$$\begin{aligned} E_{i,0}^2 &\cong H_i(\tilde{S}_W(\Gamma_v; \mathbb{F}) \otimes_{R_{W,v}} \mathbb{F}) \\ &\cong H_i(\tilde{S}(\Gamma_{W \sqcup \{v\}}; \mathbb{F}) \otimes_{\mathbb{F}[E]} \mathbb{F}[E_W] \otimes_{R_{W,v}} \mathbb{F}) \\ &\cong H_i(\tilde{S}_{W \sqcup \{v\}}(\Gamma; \mathbb{F})), \end{aligned}$$

and the latter has polynomial growth of degree at most $\Delta_\Gamma^i - 1$ by induction (we have used the isomorphism $\mathbb{F}[E_{W \sqcup \{v\}}] \cong \mathbb{F}[E_W] \otimes_{R_{W,v}} \mathbb{F}$).

Next, we have $E_{i-1,1}^2 \cong H_i(M \otimes_{R_{W,v}} N)$, where $M = \tilde{S}_W(\Gamma_v; \mathbb{F})$ and $N = H_1(C_{W,v})$. As previously noted, $C_{W,v}$ is finitely generated over the Noetherian ring $R_{W,v}$, so N is finitely generated, as well as concentrated in strictly positive degrees by definition. Moreover, M is degreewise free and thus degreewise flat. Since Γ was smooth and contained no isolated vertices, the same is true of Γ_v . Then the inductive hypothesis guarantees that, for $j < i$, the homology group $H_j(M)$ is eventually polynomial in the weight of degree at most

$$\Delta_{\Gamma_v}^j - 1 \leq \Delta_\Gamma^{j+1} - 1 \leq \Delta_\Gamma^i - 1,$$

so Lemma 3.23 implies that $\dim_{\mathbb{F}} E_{i-1,1}^2$ has the same property. □

We conclude with the proof of the technical lemma.

Proof of Lemma 3.23 Up to associated graded, the homology group $H_i(M \otimes_{R_{W,v}} N)$ is the sum of the i^{th} antidiagonal on the E^∞ page of a first-quadrant (hence convergent) Künneth spectral sequence. Thus, it suffices to bound the growth of the i^{th} antidiagonal of

$$E^2 \cong \text{Tor}_{R_{W,v}}(H_*(M), N).$$

To calculate these Tor groups, we resolve N . Since $R_{W,v}$ is Noetherian of finite global dimension, and since N is finitely generated, there is a finite-length resolution P of N by finitely generated projective $R_{W,v}$ -modules. By the Quillen–Suslin theorem, P is in fact a free resolution, generated by a trigraded $\mathbb{F}$ -vector space $G = \bigoplus G_{a,b}\{k\}$ in which every summand is finite-dimensional and almost all vanish. The trigraded complex $H_*(M) \otimes_{R_{W,v}} P$ computes the desired (weight-graded) Tor groups. Since N is concentrated in strictly positive homological degrees, P is as well, so the i^{th} antidiagonal of this complex is isomorphic to

$$\bigoplus_{j=0}^{i-1} \bigoplus_{a+b=i-j} \bigoplus_k H_j(M) \otimes_{\mathbb{F}} G_{a,b}\{k\},$$

which is a finite sum of vector spaces having eventual polynomial growth of degree at most d . □

4 Edge formality

We now undertake the in-depth study of one aspect of the chain level $R[E]$ -module structure induced by edge stabilization, namely the question of its formality.

4.1 Edge formality

Recall that a differential bigraded A -module is said to be *formal* if it is connected to its homology by a finite zigzag of (bigraded) quasi-isomorphisms:

$$(M, d) \xleftarrow{\simeq} (M_1, d_1) \xrightarrow{\simeq} (M_2, d_2) \xleftarrow{\simeq} \cdots \xrightarrow{\simeq} (H(M, d), 0).$$

Definition 4.1 We say that a graph Γ is *edge formal* over the commutative ring R if the singular chain complex of $B(\Gamma)$ with coefficients in R is formal as a differential bigraded $R[E]$ -module.

The goal of this section is to prove [Theorem 1.3](#), which gives a complete characterization of edge formal graphs independent of coefficient ring. We first recall the statement of this theorem; see [Definition 2.3](#) for a reminder on terminology.

Formality Theorem *A graph is edge formal if and only if it is small.*

In proving this theorem, we may replace the large and unwieldy complex of singular chains with the smaller complex $S(\Gamma; R)$; indeed, as we will see below in [Corollary 5.12](#), this $R[E]$ -module is quasi-isomorphic to the $R[E]$ -module of singular

chains with coefficients in R . Moreover, we may work interchangeably with the complexes $S(\Gamma; R)$ and $\tilde{S}(\Gamma; R)$. Since the underlying bigraded $R[E]$ -module of $S(\Gamma; R)$ (resp. $\tilde{S}(\Gamma; R)$) is free, we may use the following criterion, whose proof is a standard argument in homotopical algebra.

Lemma 4.2 *Let A be a weight-graded commutative ring and M a differential bigraded A -module which is projective in each homological degree. Then M is formal if and only if there is a map of differential bigraded A -modules $M \rightarrow H_*(M)$ inducing the identity on homology.*

Thus, [Theorem 1.3](#) amounts to characterizing the (non)existence of such maps.

Remark 4.3 We caution the reader that the notion of an edge formal graph has no direct connection with that of a formal topological space in the sense of rational homotopy theory. For partial results about this other kind of formality, see [\[15, Theorems 1.2 and 1.3\]](#), where a necessary and sufficient criterion is given for $B_4(\Gamma)$ to be formal as a space.

4.2 Small graphs are formal

We begin with a classification of smooth small graphs; see [Figure 2](#).

Lemma 4.4 *If Γ is a small graph that is both connected and smooth, then Γ is isomorphic to an isolated vertex, an interval I , a cycle C , a lollipop L , a figure-eight 8 , or a handcuff H .*

Proof Any vertex of a small graph must be either isolated, 1-valent, 2-valent (with either two distinct edges or a self-loop), 3-valent, with one self-loop, or 4-valent, with 2 self-loops. A connected graph with a vertex which is isolated, 2-valent with a self-loop, or 4-valent with 2 self-loops is necessarily an isolated vertex, the cycle graph C , or the figure-eight graph 8 , respectively. A graph containing a 2-valent vertex with two distinct edges cannot be smooth. Then in any remaining case, every vertex must be 1-valent or 3-valent with one self-loop. A connected graph all of whose vertices are of these two kinds must have precisely two vertices; then there are three cases, namely the interval I , the lollipop graph L , and the handcuff graph H . $\square$

This gives us most of what we need, since smoothings reflect formality.

Lemma 4.5 *Let $f: \Gamma \rightarrow \Gamma'$ be a smoothing. If Γ' is edge formal over R , then so is Γ .*

Proof The hypothesis guarantees the existence of the middle arrow in the diagram

$$S(\Gamma; R) \xrightarrow{S(f; R)} S(\Gamma'; R) \dashrightarrow H_*(S(\Gamma'; R)) \xleftarrow{H_*(S(f; R))} H_*(S(\Gamma; R))$$

of quasi-isomorphisms of $R[E(\Gamma)]$ -modules, where $R[E(\Gamma)]$ acts on the middle two entries by restriction. $\square$

We can also eliminate self-loops by making choices.

Lemma 4.6 *Let Γ be a graph with a self-loop e at the vertex v . Let Γ_- be the graph obtained by replacing e with an edge which is not a loop to a new vertex. Then the reduced Świątkowski complex $\tilde{S}(\Gamma; R)$ (unnaturally) decomposes as a direct sum of differential graded $R[E]$ -modules as*

$$\tilde{S}(\Gamma_v; R)[1]\{1\} \oplus \tilde{S}(\Gamma_-; R) \cong \tilde{S}(\Gamma; R).$$

In particular, Γ is edge formal over R if and only if both Γ_v and Γ_- are so.

Proof Let h_1 and h_2 be the two half-edges of e . Write all half-edge generators at v as differences $h_{1i} = h_1 - h_i$. We obtain a direct sum decomposition of $\tilde{S}(\Gamma; R)$ by grouping generators (using this choice of presentation at v) according to whether or not they involve h_{12} . Since h_{12} is closed, this decomposition respects the differential, and it respects the $R[E]$ -action by inspection.

Now, there is a graph morphism ι from Γ_- to Γ taking the half-edge of e to h_1 , and we obtain a map $(\alpha, \beta) \mapsto h_{12}\alpha + \iota_*(\beta)$. It is an immediate verification that this map realizes the direct sum decomposition. $\square$

Removing a vertex or turning a self-loop to an edge as in this lemma preserves smallness.

Proof of Theorem 1.3, “if” direction By Lemmas 4.5 and 4.6, we may assume that Γ is a smooth graph without self-loops. By Lemma 4.4, Γ is a disjoint union of isolated vertices and intervals. But then $\tilde{S}(\Gamma; R)$ has no differential, and there is nothing to prove. $\square$

4.3 Paradoxically decomposable cycles

We will derive the “only if” direction of Theorem 1.3 from the existence of certain cycles that we dub *paradoxically decomposable*.

Definition 4.7 Let A be a commutative ring, $I \subseteq A$ an ideal, M a differential graded A -module, and $b \in M$ a cycle.

- (1) We say that b is *decomposable* with respect to I if $b \in IM$.
- (2) We say that b is *paradoxically decomposable* with respect to I if
 - (a) b is decomposable with respect to I in M , and
 - (b) $[b]$ is *not* decomposable with respect to I in $H_*(M)$.

Lemma 4.8 Let M be a differential graded A -module that is projective in each degree, $I \subseteq A$ an ideal, and $b \in M$ a cycle. If b is paradoxically decomposable with respect to I , then M is not formal.

Proof Assume that M is formal. Then, by our hypothesis on M and Lemma 4.2, there is a map $f: M \rightarrow H_*(M)$ of A -modules inducing the identity on homology. Since b is decomposable with respect to I , we may write $b = \sum r_i b_i$ with $r_i \in I$, whence $f(b) = \sum r_i f(b_i)$; thus, $f(b)$ is decomposable with respect to I . But b is a cycle and f induces the identity on homology, so $f(b) = [b]$, a contradiction. $\square$

We briefly detour from our main discussion to present the following result, which is of independent interest, particularly from the point of view of [2, Theorem 3.20].

Proposition 4.9 Let Γ be a graph with tails e_1, e_2 , and e_3 in the same connected component, and let R be $\mathbb{Z}$ or $\mathbb{Z}/2\mathbb{Z}$. Then $S(\Gamma; R)$ is not formal over $R[e_1, e_2, e_3]$.

Proof We will show that the $S(\Gamma; R)$ has a paradoxically decomposable cycle with respect to the ideal I generated by e_1, e_2 , and e_3 .

Because e_1, e_2 , and e_3 lie in the same connected component, there is a subgraph $\Gamma_0 \subseteq \Gamma$ containing these three edges and admitting a smoothing $\Gamma_0 \rightarrow S_3$. We will show that the corresponding star class $\alpha \in H_*(B(\Gamma_0); R) \cong H_*(B(S_3); R)$ is not decomposable with respect to I , which will imply the claim, since a decomposable representative for this class is given by $\sum p_{ij} e_k$, where p_{ij} is the unique element in $S(\Gamma_0; R)$ with $\partial p_{ij} = e_i - e_j$.

Recall the canonical homomorphism $\sigma: H_1(B_2(\Gamma); R) \rightarrow \mathbb{Z}/2\mathbb{Z}$ that records the permutation of the endpoints of a braid. Since α is a star class, $\sigma(\alpha) = 1$. On the other hand, $\sigma(e\gamma) = 0$ for any $e \in \{e_1, e_2, e_3\}$ and $\gamma \in H_1(B_1(\Gamma); R)$ because e is a tail and thus is not part of the simple closed curve representing γ . Then the assumption that α is decomposable with respect to I implies the contradiction $\sigma(\alpha) = 0$. $\square$

We expect this statement is true with arbitrary coefficients.

Next, we will use paradoxically decomposable cycles to establish nonformality in two basic cases, from which we will deduce the statement for a general large graph. For the rest of this section, we shall only consider decomposability with respect to the ideal generated by E .

We write N for the number of essential vertices of Γ . As is well known, the homology of $B_k(\Gamma)$ vanishes above degree N ; indeed, the Świątkowski complex itself vanishes in these degrees (see [11, Theorem 3.3] for an earlier argument). Moreover, this top homology is nonzero, since there can be no boundaries in degree N , and since the external product of standard star class representatives at the essential vertices of Γ is a nonzero cycle in the Świątkowski complex.

Lemma 4.10 *Let Γ be a graph with no self-loops all of whose vertices are of valence 1 or 3. If Γ has a vertex of valence 3, then Γ is not edge formal.*

Proof By [2, Proposition 5.25], $H_N(B_k(\Gamma); R)$ is equal to 0 for $k < 2N$ and is one-dimensional for $k = 2N$, spanned by a W -torus, where W is the set of trivalent vertices. This class is not decomposable, since there is nothing in lower weight and degree N , but its standard representative is decomposable. The claim now follows from Lemma 4.8. $\square$

For the second result, we require the following preliminary notion:

Definition 4.11 We say that a vertex v is *simple* if there are no self-loops at v and no vertex w with multiple edges between v and w .

Lemma 4.12 *If Γ has a simple essential vertex, then Γ is not edge formal.*

Proof Let v be a simple essential vertex of Γ , and assume without loss of generality that Γ has no vertices of valence 2. Then $H_N(B(\Gamma); R) \neq 0$, so $H_{N-1}(B(\Gamma_v); R) \neq 0$ by Proposition 2.13. Choose a nonzero class β_v in the group $H_{N-1}(B_k(\Gamma_v); R)$ with k chosen minimally so that this group is nonzero, and write β for the external product of β_v and a star class at v . Since β is represented by a nonzero cycle b in top degree, we conclude that $\beta \neq 0$. There are no boundaries in top degree, so the cycle b is necessarily unique, and it is decomposable, since the standard representative of the star class is so. Thus, it suffices to show that β is not decomposable.

Applying Proposition 2.13 at v , we obtain the following piece of the exact sequence,

$$0 \rightarrow H_N(B_{k+1}(\Gamma); R) \rightarrow \bigoplus^{\ell-1} H_{N-1}(B_k(\Gamma_v); R) \xrightarrow{\delta} H_{N-1}(B_{k+1}(\Gamma_v); R),$$

where $\ell := d(v)$. Since β is a nonzero element of $H_N(B_{k+2}(\Gamma); R)$, the assumption of decomposability is the assumption that β is a nontrivial $R[E]$ -linear combination of elements from $H_N(B_{k+1}(\Gamma); R)$. Therefore it will suffice to show that $H_N(B_{k+1}(\Gamma); R)$ vanishes, ie that the connecting homomorphism δ is injective. Since Γ_v has $N-1$ essential vertices, this degree is the top homological degree for $B_k(\Gamma_v)$, and thus homology and cycles coincide. It follows that the kernel of δ consists of tuples of cycles $(b_2, \dots, b_\ell)$ in the weight N summand of $S_{N-1}(\Gamma_v; R)$ such that

$$\sum_{j=2}^{\ell} b_j(e_j - e_1) = 0$$

where e_j is the j^{th} edge incident on v . The e_j are distinct by the simplicity of v . We may assume without loss of generality that $b_2 \neq 0$.

Let $b_{j,1}$ be the chain obtained from b_j by replacing e_ℓ with e_1 . Since the differential is $R[E]$ -linear, each $b_{j,1}$ is again a cycle. We claim further that $b_{j,1}$ is nonzero if b_j was. To see this, we note that the difference $b_j - b_{j,1}$ is divisible by $(e_\ell - e_1)$, so that, if $b_{j,1} = 0$, we may conclude that b_j is the product of $(e_\ell - e_1)$ and a cycle of top degree in weight $k-1$, which is necessarily not a boundary. Since k was chosen to be minimal with respect to the existence of such a cycle, this is a contradiction.

Thus, the sum $\sum_{j=2}^{\ell-1} b_{j,1}(e_j - e_1)$ vanishes modulo $(e_\ell - e_1)$ and so vanishes, since none of the terms contain e_ℓ by construction. Applying this procedure repeatedly, we obtain a nonzero cycle $b_{2,\ell-2}$ with $b_{2,\ell-2}(e_2 - e_1) = 0$, a contradiction. $\square$

4.4 Large graphs are not formal

In order to reduce the case of a general large graph to the cases already considered, we will make use of the following device:

Definition 4.13 Let M_1 be a differential graded A_1 -module and M_2 be a differential graded A_2 -module. An (A_1, A_2) -retraction of M_2 onto M_1 consists of

- a retraction of rings $A_1 \xrightarrow{\iota} A_2 \xrightarrow{\pi} A_1$, and
- a retraction of $\mathbb{Z}$ -modules $M_1 \xrightarrow{i} M_2 \xrightarrow{p} M_1$



Figure 15: The theta graph and the A graph.

where i is A_1 –linear with respect to the A_1 –module structure on M_2 induced by ι and p is A_2 –linear with respect to the A_2 –module structure on M_1 induced by π .

The relation of this notion to questions of formality is the following:

Lemma 4.14 *Let M_1 and M_2 be differential graded A_1 – and A_2 –modules, respectively, and suppose that there exists an (A_1, A_2) –retraction of M_2 onto M_1 . If M_2 is projective in each degree and formal over A_2 , then M_1 is formal over A_1 .*

Proof Our hypotheses on M_2 and Lemma 4.2 guarantee the existence of a map $M_2 \rightarrow H_*(M_2)$ of differential graded A_2 –modules inducing the identity on homology. It follows that the map in homology induced by the composite

$$M_1 \rightarrow M_2 \rightarrow H_*(M_2) \rightarrow H_*(M_1)$$

is $H_*(p) \circ \text{id}_{H_*(M_2)} \circ H_*(i) = \text{id}_{H_*(M_1)}$, so it suffices to check that this map is A_1 –linear, which follows from a diagram chase. □

Definition 4.15 Let Γ and Γ' be graphs. An *algebraic retraction* (over R) from Γ' to Γ is a $(R[E(\Gamma)], R[E(\Gamma')])$ –retraction of $\tilde{S}(\Gamma'; R)$ onto $\tilde{S}(\Gamma; R)$. In the presence of an algebraic retraction, we say that Γ is an *algebraic retract* of Γ' (over R).

The composition of algebraic retractions is an algebraic retraction. An algebraic retraction need not arise from a topological retraction between configuration spaces.

The “only if” direction of Theorem 1.3 is an immediate consequence of the following result in combination with Lemmas 4.10, 4.12, and 4.14 (see Figure 15 for an explanation of unfamiliar terminology).

Lemma 4.16 *If Γ is a smooth large graph, then Γ has one of the following graphs as an algebraic retract:*

- (1) a graph with a simple essential vertex,
- (2) the theta graph Θ_3 , or
- (3) the A graph A .

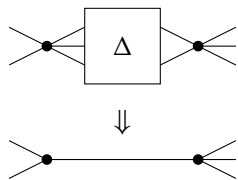


Figure 16: Depiction of a surgery.

The proof of this result relies on the fact that algebraic retractions arise geometrically as the result of *surgeries* on graphs (see [Figure 16](#)).

Lemma 4.17 *Let Γ and Γ' be graphs. Suppose Γ is obtained from Γ' by replacing a connected subgraph Δ , attached to the rest of Γ only at the vertices $v_1 \neq v_2$, with a single edge e_0 between v_1 and v_2 . That is,*

$$\Gamma' = (\Gamma \setminus e_0) \amalg_{v_1 \cup v_2} \Delta.$$

Then Γ is an algebraic retract of Γ' over any R .

We call this operation a *surgery*. Assuming this result momentarily, we complete the proof of [Theorem 1.3](#).

Proof of Lemma 4.16 A connected component is an algebraic retract and largeness is a local property so it suffices to consider the connected case. Suppose Γ has a vertex of valence strictly greater than three with a self-loop. The direct sum decomposition of [Lemma 4.6](#) gives an algebraic retraction from Γ to the graph Γ_- obtained by replacing a self-loop with a tail. The graph Γ_- is smooth, connected, and large if Γ was. Suppose Γ has a vertex of valence three with a self-loop. There is a surgery replacing the closed star of the vertex with a tail which preserves smoothness and largeness. Then iteratively we may assume Γ contains no self-loops.

Suppose that Γ has a pair of vertices sharing three edges. We obtain an algebraic retraction onto Θ_3 using [Lemma 4.17](#) with Δ the complement in Γ of two of these (open) edges. See [Figure 17](#), left. Suppose that Γ has a pair of vertices v and w sharing two edges e_1 and e_2 . Since Γ is smooth, each of v and w has a third edge. If there is a path of edges disjoint from e_1 and e_2 connecting v and w , then surgery produces an algebraic retraction onto Θ_3 . See [Figure 17](#), center. If not, two surgeries produce an algebraic retraction onto A . See [Figure 17](#), right.

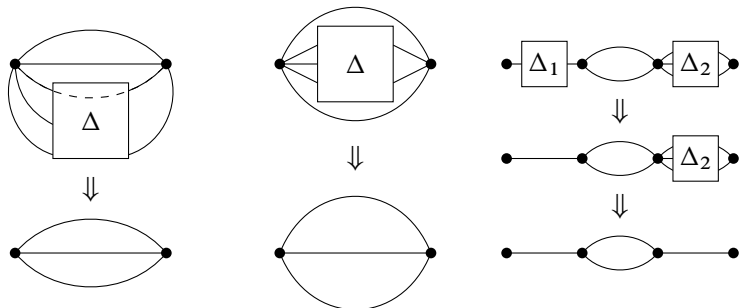


Figure 17: Surgeries for the proof of Lemma 4.16.

Then if there is no algebraic retraction from Γ onto Θ_3 or A , the removal of self-loops constitutes an algebraic retraction onto a connected smooth large graph with a simple essential vertex. $\square$

To prepare for the proof of the surgery lemma, we observe that the Świątkowski complex enjoys a certain noncanonical functoriality for subdivisions. More precisely, the map induced by a smoothing $f: \Gamma_1 \rightarrow \Gamma_2$ admits a noncanonical left inverse, which we regard as corresponding to the subdivision f^{-1} . It suffices to construct this morphism in the case of a subdivision of a single edge e with half-edges h_1 and h_2 into two edges e_1 and e_2 meeting at the common vertex v with corresponding half-edges h_{v1} and h_{v2} . We define the map in this case by the assignments $e \mapsto e_1$, $h_1 \mapsto h_1$, and $h_2 \mapsto h_{v1} - h_{v2} + h_2$. See Figure 18. This map is a map of differential bigraded $R[E_2]$ -modules, where the action of $R[E_2]$ on the target is via the map $R[E_2] \rightarrow R[E_1]$ just constructed and hence also noncanonical. The same remarks hold for reduced complexes.

Proof of Lemma 4.17 There is a (noncanonical) graph embedding of a subdivision of Γ into Γ' , where only the edge e_0 is subdivided. Thus, for any surgery, there is a noncanonical map $\tilde{S}(\Gamma; R) \rightarrow \tilde{S}(\Gamma'; R)$. This noncanonical map admits a left inverse r , defined as follows. Write Γ' as $(\Gamma \setminus e_0) \amalg_{v_1 \cup v_2} \Delta$. Then

$$r(h) = \begin{cases} h_i & \text{if } h \in H(\Delta) \text{ and } v(h) = v_i, \\ 0 & \text{if } h \in H(\Delta) \text{ and } v(h) \neq v_i, \\ h & \text{if } h \notin H(\Delta), \end{cases} \quad \text{and} \quad r(e) = \begin{cases} e_0 & \text{if } e \in E(\Delta), \\ e & \text{if } e \notin E(\Delta). \end{cases}$$

The map on edges endows $\tilde{S}(\Gamma; R)$ with an $R[E(\Gamma')]$ -module structure, and linearity is a direct verification. By inspection, these maps constitute a retraction both at the level of modules and at the level of rings. $\square$

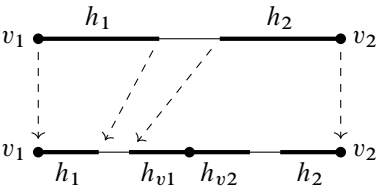


Figure 18: A choice of left inverse for the map induced by a smoothing.

5 Two isomorphisms

We return to the proof of Theorem 2.10. Along the way, we will encounter two very different routes to the Świątkowski complex. In Section 5.1, we review the method of [2], which has the advantage that it naturally outputs a functor on the category $\mathcal{Gph}$. On the other hand, there is the cubical deformation retract introduced by Świątkowski [27], which offers a more direct and geometric comparison to $B(\Gamma)$ at the cost of nonfunctorial choices. Our argument will combine the advantages of these two approaches.

In this section and the next, we write $C^{\text{sing}}(X)$ for the singular chain complex of the topological space X . If X is a CW complex, we denote the cellular chain complex of X by $C(X)$. We will make use of the existence of a zigzag

$$C^{\text{sing}} \stackrel{\sim}{\leftarrow} \bullet \stackrel{\sim}{\rightarrow} C$$

of quasi-isomorphisms connecting these complexes, which is natural for cellular maps. The specifics of the intermediate object will play no role here, but, for one option, see [2, Construction 2.22].

5.1 Functorial model

We now recall some of the work of [2]. The starting point is a cell complex introduced by Abrams [1, Definition 2.1].

Definition 5.1 Let X be a cell complex. The k^{th} unordered *configuration complex* of X is the subspace

$$B_k^\square(X) = \left(\bigcup_{\bar{c}_i \cap \bar{c}_j = \emptyset} c_1 \times \cdots \times c_k \right) / \Sigma_k \subseteq B_k(X),$$

where the union is taken over the set of k -tuples of disjoint open cells of X .

The configuration complexes of X are approximations to the configuration spaces of X by cell complexes, and the accuracy of this approximation improves with subdivision. Following [8, Section 3], we say that a graph Γ is *sufficiently subdivided for k* if, first, every path in Γ between essential vertices passes through at least $k - 1$ edges, and, second, every loop passes through at least $k + 1$ edges.

Theorem 5.2 [1, Theorem 2.1] *Let $\Gamma \rightarrow \Gamma'$ be a subdivision with Γ' sufficiently subdivided for k . The inclusion $B_k^\square(\Gamma') \rightarrow B_k(\Gamma)$ is a deformation retract.*

In order to work functorially and with all k at once, it is more convenient to consider all subdivisions simultaneously. Denoting by $\mathcal{P}$ the set of subdivisions of Γ , viewed as a category under refinement, Theorem 5.2 implies that the natural map

$$\operatorname{colim}_{\Gamma' \in \mathcal{P}} B_k^\square(\Gamma') \rightarrow B_k(\Gamma)$$

is a weak homotopy equivalence [2, Theorem 2.8].

Including arbitrary subdivisions has the further benefit of allowing one to work locally on Γ . In order to leverage this flexibility, we return to the heuristic interpretation of Remark 2.8, in which generators of $S(\Gamma)$ represent “states” in $B(\Gamma)$ obtained by prescribing “local states” near vertices and along edges, which are compatible on half-edges.

More precisely, define $S^\wedge(S_n)$ to be the subcomplex of $S(S_n)$ spanned over $\mathbb{Z}[E]$ by the symbols $\emptyset$, the star vertex and half-edges at the star vertex. For e an edge, viewed as a 1-cell, define $S^\wedge(e)$ as the subcomplex of $S(e)$ spanned by basis elements with no vertices or half-edges (then $S^\wedge(e)$ is canonically isomorphic to $\mathbb{Z}[e]$). For these atomic graphs, one can write down a map $C(B^\square(\Gamma')) \rightarrow S^\wedge(\Gamma)$ and check by hand that it is a quasi-isomorphism for sufficiently fine subdivisions; see [2, Sections 4.2–4.3].

Now, for Ξ a disjoint union of star graphs and edges, define $S^\wedge(\Xi)$ to be the tensor product over the connected components of Ξ of the corresponding subcomplexes defined above. Then there is an isomorphism

$$S(\Gamma) \cong S^\wedge\left(\coprod_{v \in V} S_{d(v)}\right) \otimes_{S^\wedge((\coprod_{e \in E} 1) \times_E H)} S^\wedge\left(\coprod_{e \in E} 1\right).$$

We now introduce a device that aids in piecing together these local identifications. Recall that each edge of Γ is identified with $(0, 5)$ via its parametrization. We define a

map $\pi: \Gamma \rightarrow [0, 1]$ by setting

$$\pi(t) = \begin{cases} t-1 & \text{if } t \in [1, 2] \subseteq e, \\ 1 & \text{if } t \in [2, 3] \subseteq e, \\ 4-t & \text{if } t \in [3, 4] \subseteq e, \\ 0 & \text{otherwise,} \end{cases}$$

and by sending every vertex of Γ to 0.

Definition 5.3 A *gap* in Γ is a subspace of the form $A = \pi^{-1}(A_0)$, where $A_0 \subseteq [0, 1]$ is a nonempty open subset such that

- (1) the complement $[0, 1] \setminus A_0$ is a (possibly empty) finite union of closed intervals of positive length, and
- (2) if $i \in \{0, 1\}$ lies in the closure of A_0 , then $i \in A_0$.

See [Figure 19](#). We write $\mathcal{G}$ for the poset of gaps.

Since the complement of a gap is a disjoint union of stars and intervals, we have a map $C(B^\square(\Gamma' \setminus A)) \rightarrow S^\wedge(\Gamma' \setminus A)$ whenever A is a gap and Γ' is a subdivision in which A is a union of cells. In this way, we obtain the following zigzag of quasi-isomorphisms:

$$\begin{array}{ccc} \operatorname{colim}_{\Gamma' \in \mathcal{P}} C^{\operatorname{sing}}(B^\square(\Gamma')) & \xrightarrow{\sim} & C^{\operatorname{sing}}(B(\Gamma)) \\ \uparrow \wr & & \\ \bullet & & \\ \downarrow \wr & & \\ \operatorname{colim}_{\Gamma' \in \mathcal{P}} C(B^\square(\Gamma')) & & \\ \uparrow \wr & & \\ \operatorname{hocolim}_{A \in \mathcal{G}^{\operatorname{op}}} \operatorname{colim}_{\Gamma' \in \mathcal{P}_A} C(B^\square(\Gamma' \setminus A)) & & \\ \downarrow \wr & & \\ \operatorname{hocolim}_{A \in \mathcal{G}^{\operatorname{op}}} S^\wedge(\Gamma \setminus A) & \xrightarrow{\sim} & S(\Gamma) \end{array}$$

Here, $\mathcal{P}_A \subseteq \mathcal{P}$ contains only those subdivisions of Γ in which A is a union of cells. For details on why these maps are quasi-isomorphisms and why the resulting isomorphism on homology is functorial, see [\[2, Section 4\]](#).

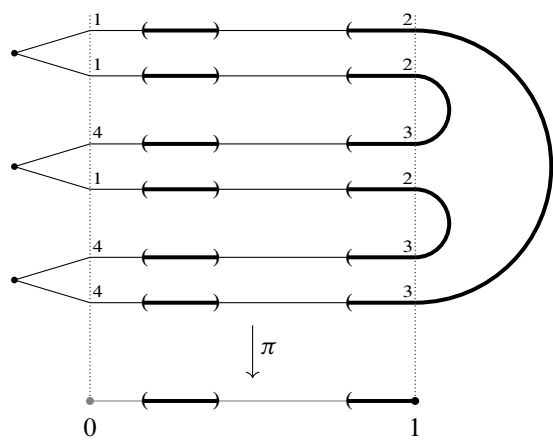


Figure 19: A gap in the complete graph K_3 .

Theorem 5.4 [2, Theorem 4.5] *There is a natural isomorphism*

$$H_*(B(\Gamma)) \cong H_*(S(\Gamma))$$

of functors from $\mathfrak{Gph}$ to bigraded abelian groups.

5.2 Cubical model

We recall the cubical model of $B(\Gamma)$ introduced in [27] and corrected in [4, Section 2.1]. Given a graph Γ , we write $A(\Gamma)$ for the set of labelings

$$\lambda\colon E \sqcup V \rightarrow \mathbb{Z}_{\geq 0} \sqcup V \sqcup H \sqcup \{\emptyset\}$$

such that $\lambda(e) \in \mathbb{Z}_{\geq 0}$ and $\lambda(v) \in \{\emptyset, v\} \sqcup H(v)$.

Construction 5.5 (Świątkowski) Define a space $\mathrm{UK}(\Gamma)$ as the quotient

$$\mathrm{UK}(\Gamma) = \coprod_{\lambda \in A(\Gamma)} \{\lambda\} \times [0, 1]^{\lambda^{-1}(H)} / \sim,$$

where the equivalence relation is determined as follows. Fix a half-edge h with edge e and vertex v . Suppose λ , λ_0 , and λ_1 are labelings which agree except on v and e , where

$$\begin{aligned} \lambda(v) &= h, & \lambda_0(v) &= v, & \lambda_1(v) &= \emptyset, \\ \lambda(e) &= n, & \lambda_0(e) &= n, & \lambda_1(e) &= n + 1. \end{aligned}$$

Then we glue according to the following identifications for $\epsilon \in \{0, 1\}$:

$$\{\lambda_\epsilon\} \times [0, 1]^{\lambda_\epsilon^{-1}(H)} \cong \{\lambda\} \times [0, 1]^{\lambda^{-1}(H \setminus \{h\})} \times \{\epsilon\} \quad \text{as a subset of } \{\lambda\} \times [0, 1]^{\lambda^{-1}(H)}.$$

To specify a point in $\text{UK}(\Gamma)$, it is enough to give a labeling $\lambda \in A(\Gamma)$ together with numbers $t(h) \in [0, 1]$ for each $h \in \lambda^{-1}(H)$.

We endow the set $A(\Gamma)$ with a bigrading by declaring that

$$|\lambda| = \left(|\lambda^{-1}(H)|, \sum_E \lambda(e) + |\lambda^{-1}(V)| + |\lambda^{-1}(H)| \right)$$

and note that $\text{UK}(\Gamma)$ splits as a disjoint union of cell complexes $\text{UK}_k(\Gamma)$ whose i -cells are those in bigrading (i, k) .

Observation 5.6 By inspection, the map

$$C(\text{UK}(\Gamma)) \rightarrow S(\Gamma), \quad \lambda \mapsto \left(\prod_E e^{\lambda(e)} \right) \otimes \bigotimes_V \lambda(v),$$

is a bigraded chain isomorphism.

Remark 5.7 To make [Observation 5.6](#) precise, we should take care with orientations. One way to specify an orientation on a cube which is a product of intervals indexed by a set J is to give an order on J up to even permutation. Similarly, one way to specify the correct sign on a tensor product of odd degree vector spaces indexed by a set J is to give an order on J up to even permutation. Then, with appropriate conventions as to which is the “positive” and which is the “negative” end of an interval, the map of the observation intertwines these conventions.

Remark 5.8 In previous work along these lines [\[4; 17; 27\]](#), neither 2-valent nor 1-valent vertices were considered, but there is no obstruction to this mild generalization.

We now show that the complex $\text{UK}(\Gamma)$ is homotopy equivalent to the configuration space of interest. Apart from slight modifications for simplicity and compatibility with our setup, this argument is essentially that of [\[4, Section 2.1\]](#).¹ In particular, our homotopy equivalence will be homotopic to the one considered in that work.

In order to compare $B(\Gamma)$ and $\text{UK}(\Gamma)$, we first deform $B(\Gamma)$ onto the subspace $\tilde{B}(\Gamma)$ of configurations x with the property that, for each vertex v , at most one coordinate

¹As observed by Lütgehetmann, the map defined by Świątkowski is not quite a retraction [\[17, page 24\]](#). Unfortunately, the replacement retraction constructed by Lütgehetmann is not continuous, as one can see by comparing the formulas for $t_x(s)$ appearing on page 22 for $k = 2$ and $k = 3$ in a situation in which the first of three particles in an edge approaches the initial endpoint of that edge. Chettih and Lütgehetmann [\[4\]](#) give a continuous retraction.

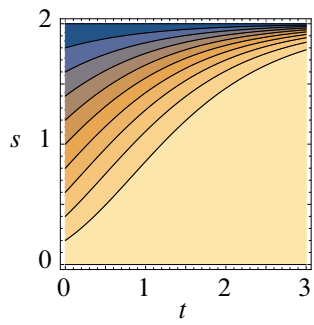


Figure 20: The homotopy of [Construction 5.10](#).

lies in the open star $\text{st}_1(v)$ of radius 1 at v . This deformation is achieved by radial expansion in each star simultaneously. Having taken this intermediate step, we define our comparison map ρ as the composite

$$\rho\colon B(\Gamma) \rightarrow \widetilde{B}(\Gamma) \rightarrow \text{UK}(\Gamma),$$

where the second map records the presence or absence of particles at vertices, the coordinate of any particle in $\text{st}_1(v) \setminus \{v\} \cong H(v) \times (0, 1)$, and the number of particles lying in the subinterval $[1, 4]$ of each edge. The details of the deformation are given below in [Construction 5.10](#); for now, we state the following result concerning ρ .

Proposition 5.9 *The map ρ is a homotopy equivalence.*

Construction 5.10 We define the deformation of $B(\Gamma)$ onto $\widetilde{B}(\Gamma)$ as follows. For each vertex v , we have $\overline{\text{st}_2(v)} \setminus \{v\} \cong H(v) \times (0, 2]$. On each of these intervals, for a fixed configuration x , we use the homotopy $(s, t) \mapsto 2(\frac{1}{2}s)^{e^{-t}}$ (see [Figure 20](#)) on all points of $x \cap (\{h\} \times (0, 2])$ simultaneously. Here, $s \in (0, 2]$ and $t \in [0, t(x, v)]$, where $t(x, v)$ is the least t such that the resulting configuration has at most one particle in $\text{st}_1(v)$.² Since $t(x, v)$ is continuous in x , and since $t(x, v) = 0$ for $x \in \widetilde{B}(\Gamma)$, this prescription defines a deformation retraction.

Proof of Proposition 5.9 Chettih and Lütgehetmann [[4](#), Section 2.1] define a cube complex $\widetilde{\text{UK}}_k(\Gamma)$ which is a deformation retract of the *ordered* configuration space of k points in Γ . Their cube complex is a Σ_k -cover of the weight k subcomplex of $\text{UK}(\Gamma)$. Their deformation retraction is equivariant and so passes to a deformation

²Explicitly, $t(x, v)$ is 0 if there are fewer than two points in $\text{st}_2(v)$ and otherwise $\log(1 - \log_2(s_v))$, where s_v is the distance from v to the second closest point in $\text{st}_2(v)$ in x .

retraction of $B(\Gamma)$ onto $\text{UK}(\Gamma)$. The composite $B(\Gamma) \xrightarrow{\rho} \text{UK}(\Gamma) \rightarrow B(\Gamma)$ with the inclusion of this quotient deformation retract differs from the identity only in the positions of particles in individual open edges. Therefore, the two maps are homotopic by edgewise straight line homotopies. It follows that ρ is a one-sided homotopy inverse to a homotopy equivalence and hence itself a homotopy equivalence. $\square$

5.3 Comparison of models

We have two isomorphisms $H_*(B(\Gamma)) \cong H_*(S(\Gamma))$. The isomorphism of [Theorem 5.4](#) is an isomorphism of functors on the category $\mathfrak{Sph}$. This naturality is a powerful tool in applications [\[2, Section 5\]](#). On the other hand, the isomorphism obtained by combining [Proposition 5.9](#) and [Observation 5.6](#) is more geometric in nature. Fortunately, we need not choose between these virtues.

Proposition 5.11 *The diagram of isomorphisms*

$$\begin{array}{ccc} H_*(B(\Gamma)) & \xrightarrow[\simeq]{\text{Proposition 5.9}} & H_*(\text{UK}(\Gamma)) \\ & \searrow \text{Theorem 5.4} \quad \swarrow \text{Observation 5.6} & \\ & H_*(S(\Gamma)) & \end{array}$$

commutes.

The proof of this result will occupy [Section 6](#). For now, we will use it to deduce the desired conclusion regarding edge stabilization.

For $\lambda \in A(\Gamma)$, write $e\lambda$ for the labeling that differs from λ only in that $e\lambda(e) = \lambda(e) + 1$ (in particular, $\lambda^{-1}(H) = (e\lambda)^{-1}(H)$). There is a version of edge stabilization at the level of $\text{UK}(\Gamma)$ which sends λ to $e\lambda$ and fixes all t_h coordinates. The induced $\mathbb{Z}[E]$ -action on $C(\text{UK}(\Gamma))$ coincides, through the isomorphism of [Observation 5.6](#), with the canonical action on $S(\Gamma)$.

Proof of Theorem 2.10 It suffices by [Proposition 5.11](#) to show that the map

$$\rho_* \colon H_*(B(\Gamma)) \rightarrow H_*(\tilde{B}(\Gamma)) \rightarrow H_*(\text{UK}(\Gamma))$$

is $\mathbb{Z}[E]$ -linear. By inspection, we have the commuting diagram

$$\begin{array}{ccc} \tilde{B}(\Gamma) & \longrightarrow & \text{UK}(\Gamma) \\ \sigma_e \downarrow & & \downarrow \\ \tilde{B}(\Gamma) & \longrightarrow & \text{UK}(\Gamma) \end{array}$$

so the second map in this composite is $\mathbb{Z}[E]$ -linear. In order to show that the first map is also $\mathbb{Z}[E]$ -linear, we note that the inverse, which is induced by the inclusion $\tilde{B}(\Gamma) \subseteq B(\Gamma)$, is $\mathbb{Z}[E]$ -linear. $\square$

Corollary 5.12 *The differential bigraded $\mathbb{Z}[E]$ -modules $C^{\text{sing}}(B(\Gamma))$ and $S(\Gamma)$ are quasi-isomorphic.*

Proof We have the zigzag of $\mathbb{Z}[E]$ -linear quasi-isomorphisms

$$C^{\text{sing}}(B(\Gamma)) \leftarrow C^{\text{sing}}(\tilde{B}(\Gamma)) \rightarrow C^{\text{sing}}(\text{UK}(\Gamma)) \leftarrow \bullet \rightarrow C(\text{UK}(\Gamma)) \cong S(\Gamma). \quad \square$$

Unlike the homology isomorphism, this quasi-isomorphism is not natural, since it relies on a choice of parametrization.

6 Long ends and the proof of Proposition 5.11

In this section, we compare the two isomorphisms $H_*(B(\Gamma)) \cong H_*(S(\Gamma))$. The key observation is that, for a certain class of subdivision $\Gamma \rightarrow \Gamma'$, the natural map

$$B_k^\square(\Gamma') \subseteq B_k(\Gamma') \cong B_k(\Gamma) \rightarrow \text{UK}_k(\Gamma)$$

from Abrams' model to Świątkowski's model is cellular. With this observation in hand, Proposition 5.11 follows after checking that this collection of special subdivisions is large enough to support the argument of [2].

6.1 Long ends

In this section, we fix a parametrization of a graph Γ , identifying each edge with $(0, 5)$. Given a subdivision $\Gamma \rightarrow \Gamma'$, we do *not* independently parametrize Γ' . Rather, we use the homeomorphism underlying the subdivision to identify Γ' with Γ . This allows us to specify the data of the subdivision Γ' up to isomorphism by naming a finite set of points in $(0, 5)$ for each edge of Γ .

Definition 6.1 We say that a subdivision $\Gamma \rightarrow \Gamma'$ has *long ends* if, for every edge of Γ , the induced subdivision $\{0, t_1, \dots, t_r, 5\}$ of $[0, 5]$ has $t_1 = 1$ and $t_r = 4$.

If Γ' has long ends, we identify a half-edge h of Γ with the corresponding 1-cell of length 1 in Γ' . The following observation is the heart of our comparison of the two models in question.

Lemma 6.2 *If $\Gamma \rightarrow \Gamma'$ has long ends, then the composite*

$$B_k^\square(\Gamma') \subseteq B_k(\Gamma) \xrightarrow{\rho_k} \text{UK}_k(\Gamma)$$

is cellular.

Proof A cell of $B_k^\square(\Gamma')$ is specified by a function μ from the set of cells of Γ' to $\{0, 1\}$ with the following properties:

- (1) $\sum_c \mu(c) = k$.
- (2) If $c_i \cap c_j \neq \emptyset$, then $\mu(c_i) + \mu(c_j) \leq 1$.

The degree of the cell μ is the sum of μ over the 1-cells of Γ' .

By inspection, the composite in question maps the cell μ onto the cell λ_μ with

$$\lambda_\mu(v) = \begin{cases} v & \text{if } \mu(v) = 1, \\ h & \text{if } \mu(h) = 1, \\ \emptyset & \text{otherwise,} \end{cases}$$

and $\lambda_\mu(e)$ defined by summing μ over all cells of the induced subdivision on $[1, 4]$. Since the degree of λ_μ is at most the degree of μ , this claim implies the lemma. $\square$

Write $\mathcal{L}$ for the category of subdivisions of Γ with long ends. Although this category is filtered, it is not convergent in the sense of [2, Definition 2.6]. Nevertheless, we have the following.

Lemma 6.3 *For every $k \geq 0$, the natural map*

$$\text{colim}_{\mathcal{L}} C(B_k^\square(\Gamma')) \rightarrow \text{colim}_{\mathcal{P}} C(B_k^\square(\Gamma'))$$

is a quasi-isomorphism.

Proof Let $\mathcal{P}_k \subseteq \mathcal{P}$ denote the subcategory of subdivisions of Γ that are sufficiently subdivided for k , and set $\mathcal{L}_k = \mathcal{L} \cap \mathcal{P}_k$. Since $\mathcal{L}_k$ is final in $\mathcal{L}$ and $\mathcal{P}_k$ in $\mathcal{P}$, the vertical arrows in the commuting diagram

$$\begin{array}{ccc} \text{colim}_{\mathcal{L}_k} C(B_k^\square(\Gamma')) & \longrightarrow & \text{colim}_{\mathcal{P}_k} C(B_k^\square(\Gamma')) \\ \downarrow & & \downarrow \\ \text{colim}_{\mathcal{L}} C(B_k^\square(\Gamma')) & \longrightarrow & \text{colim}_{\mathcal{P}} C(B_k^\square(\Gamma')) \end{array}$$

are isomorphisms, so it suffices to show that the top arrow is a quasi-isomorphism. By sufficient subdivision and [Theorem 5.2](#), every arrow in $\mathcal{P}_k$ induces a quasi-isomorphism on $C(B_k^\square(-))$, so the claim follows from the observation that the inclusion of $\mathcal{L}_k$ in $\mathcal{P}_k$ induces a weak homotopy equivalence on nerves, since both nerves are contractible. Indeed, both $\mathcal{L}$ and $\mathcal{P}$ are filtered, hence contractible, and final functors induce weak equivalences on nerves. $\square$

Lemma 6.4 *Let $A = \pi^{-1}(A_0)$ be a gap. If $A_0 \cap \{0, 1\} = \emptyset$, then the natural map*

$$\operatorname{colim}_{\Gamma' \in \mathcal{L} \cap \mathcal{P}_A} C(B^\square(\Gamma' \setminus A)) \rightarrow \operatorname{colim}_{\Gamma' \in \mathcal{P}_A} C(B^\square(\Gamma' \setminus A))$$

is a quasi-isomorphism.

Proof The category of subdivisions of $\Gamma' \setminus A$ that are sufficiently subdivided for a fixed k is final in the category of subdivisions restricted from $\mathcal{P}$. Since $A_0 \cap \{0, 1\} = \emptyset$, the same is true for $\mathcal{L}$, so the claim follows in the manner of [Lemma 6.3](#). $\square$

Lemma 6.5 *Suppose that $\Gamma \rightarrow \Gamma'$ has long ends, and let A be a gap in Γ that is a union of cells of Γ' . The following diagram of chain maps commutes:*

$$\begin{array}{ccccc} C(B^\square(\Gamma' \setminus A)) & \longrightarrow & C(B^\square(\Gamma')) & \xrightarrow{C(\rho)} & C(\operatorname{UK}(\Gamma)) \\ \downarrow & & & & \parallel \wr \\ S^\wedge(\Gamma \setminus A) & \longrightarrow & & & S(\Gamma) \end{array}$$

Proof The claim is immediate from the explicit description of the value of ρ on cells given in [Lemma 6.2](#) and the description of the left-hand vertical map given in [\[2, Definitions 4.13 and 4.15\]](#). $\square$

6.2 Proof of [Proposition 5.11](#)

We wish to compare two isomorphisms identifying $H_*(B(\Gamma))$ with $H_*(S(\Gamma))$. The first is induced on homology by the zigzag of quasi-isomorphisms in the right-hand portion of the diagram of [Figure 21](#), in which each of the square subdiagrams commutes. The first step in the proof is to verify that *all* of the maps in the diagram are quasi-isomorphisms, so that we may replace this zigzag with the outer zigzag in the diagram.

$$\begin{array}{ccc}
 \operatorname{colim}_{\Gamma' \in \mathcal{L}} C^{\operatorname{sing}}(B^{\square}(\Gamma')) & \longrightarrow & \operatorname{colim}_{\Gamma' \in \mathcal{P}} C^{\operatorname{sing}}(B^{\square}(\Gamma')) \xrightarrow{\sim} C^{\operatorname{sing}}(B(\Gamma)) \\
 (1) \uparrow & & \uparrow \wr \\
 \bullet & \longrightarrow & \bullet \\
 (2) \downarrow & & \downarrow \wr \\
 \operatorname{colim}_{\Gamma' \in \mathcal{L}} C(B^{\square}(\Gamma')) & \xrightarrow{(3)} & \operatorname{colim}_{\Gamma' \in \mathcal{P}} C(B^{\square}(\Gamma')) \\
 \uparrow & & \uparrow \wr \\
 \operatorname{hocolim}_{A \in \mathcal{G}^{\operatorname{op}}} \operatorname{colim}_{\Gamma' \in \mathcal{L} \cap \mathcal{P}_A} C(B^{\square}(\Gamma' \setminus A)) & \xrightarrow{(4)} & \operatorname{hocolim}_{A \in \mathcal{G}^{\operatorname{op}}} \operatorname{colim}_{\Gamma' \in \mathcal{P}_A} C(B^{\square}(\Gamma' \setminus A)) \\
 & & \downarrow \wr \\
 & & \operatorname{hocolim}_{A \in \mathcal{G}^{\operatorname{op}}} S^{\wedge}(\Gamma \setminus A) \xrightarrow{\sim} S(\Gamma)
 \end{array}$$

Figure 21: A commutative diagram of quasi-isomorphisms for the proof of Proposition 5.11.

- (1)–(2) These maps are induced by natural quasi-isomorphisms after taking the colimit over $\mathcal{L}$. Since $\mathcal{L}$ is filtered, the claim follows.
- (3) This quasi-isomorphism is supplied by Lemma 6.3.
- (4) This quasi-isomorphism follows from Lemma 6.4 and the observation that the poset of gaps satisfying the hypotheses of that lemma is homotopy initial in $\mathcal{G}$ (hence homotopy final in $\mathcal{G}^{\operatorname{op}}$).

The remaining arrows are quasi-isomorphisms by two-out-of-three.

The second isomorphism $H_*(B(\Gamma)) \rightarrow H_*(S(\Gamma))$ is induced by the map

$$\rho: B(\Gamma) \rightarrow \operatorname{UK}(\Gamma)$$

of topological spaces, together with the identification of $S(\Gamma)$ with cellular chains on $\operatorname{UK}(\Gamma)$. In order to compare this isomorphism with the previous, we note that Lemma 6.2 supplies the dashed fillers in the commuting diagram

$$\begin{array}{ccccc}
 C^{\operatorname{sing}}(B^{\square}(\Gamma')) & \longrightarrow & C^{\operatorname{sing}}(B(\Gamma)) & \xrightarrow{\sim} & C^{\operatorname{sing}}(\operatorname{UK}(\Gamma)) \\
 \wr \uparrow & & & & \uparrow \wr \\
 \bullet & \dashrightarrow & \bullet & & \\
 \wr \downarrow & & & & \downarrow \wr \\
 C(B^{\square}(\Gamma')) & \dashrightarrow & C(\operatorname{UK}(\Gamma)) & &
 \end{array}$$

whenever $\Gamma \rightarrow \Gamma'$ has long ends. Passing to the colimit over $\mathcal{L}$, we obtain the diagram

$$\begin{array}{ccc}
 \operatorname{colim}_{\Gamma' \in \mathcal{L}} C^{\operatorname{sing}}(B^{\square}(\Gamma')) & \xrightarrow{\sim} & C^{\operatorname{sing}}(B(\Gamma)) \xrightarrow{\sim} C^{\operatorname{sing}}(\operatorname{UK}(\Gamma)) \\
 \uparrow \wr & & \uparrow \wr \\
 \bullet & \xrightarrow{\quad\quad\quad} & \bullet \\
 \downarrow \wr & & \downarrow \wr \\
 \operatorname{colim}_{\Gamma' \in \mathcal{L}} C(B^{\square}(\Gamma')) & \xrightarrow{\quad\quad\quad} & C(\operatorname{UK}(\Gamma)) \\
 \uparrow \wr & & \parallel \wr \\
 \operatorname{hocolim}_{A \in \mathcal{G}^{\operatorname{op}}} \operatorname{colim}_{\Gamma' \in \mathcal{L} \cap \mathcal{P}_A} C(B^{\square}(\Gamma' \setminus A)) & & \\
 \downarrow \wr & & \\
 \operatorname{hocolim}_{A \in \mathcal{G}^{\operatorname{op}}} S^{\wedge}(\Gamma \setminus A) & \xrightarrow{\sim} & S(\Gamma)
 \end{array}$$

The upper portion of the diagram commutes by what has already been said, and the bottom portion of the diagram commutes by [Lemma 6.5](#) and the universal properties of the colimit and the homotopy colimit. We established in the first half of the proof that the first of the isomorphisms in question is induced on homology by the counterclockwise zigzag from $C^{\operatorname{sing}}(B(\Gamma))$ to $S(\Gamma)$. Since the clockwise zigzag induces the second of the isomorphisms, the proof is complete.

References

- [1] **AD Abrams**, *Configuration spaces and braid groups of graphs*, PhD thesis, University of California, Berkeley (2000) [MR](#) Available at <https://search.proquest.com/docview/304583880>
- [2] **BH An, G C Drummond-Cole, B Knudsen**, *Subdivisional spaces and graph braid groups*, *Doc. Math.* 24 (2019) 1513–1583 [MR](#)
- [3] **VI Arnold**, *The cohomology ring of the colored braid group*, *Mat. Zametki* 5 (1969) 227–231 [MR](#) In Russian; translated in *Math. Notes* 5 (1969) 138–140
- [4] **S Chetthi, D Lütgehetmann**, *The homology of configuration spaces of trees with loops*, *Algebr. Geom. Topol.* 18 (2018) 2443–2469 [MR](#)
- [5] **T Church, JS Ellenberg, B Farb**, *FI-modules and stability for representations of symmetric groups*, *Duke Math. J.* 164 (2015) 1833–1910 [MR](#)
- [6] **FR Cohen, TJ Lada, JP May**, *The homology of iterated loop spaces*, *Lecture Notes in Mathematics* 533, Springer (1976) [MR](#)

- [7] **D Eisenbud**, *Commutative algebra: with a view toward algebraic geometry*, Graduate Texts in Mathematics 150, Springer (1995) [MR](#)
- [8] **D Farley, L Sabalka**, *Discrete Morse theory and graph braid groups*, *Algebr. Geom. Topol.* 5 (2005) 1075–1109 [MR](#)
- [9] **D Farley, L Sabalka**, *On the cohomology rings of tree braid groups*, *J. Pure Appl. Algebra* 212 (2008) 53–71 [MR](#)
- [10] **Ś R Gal**, *Euler characteristic of the configuration space of a complex*, *Colloq. Math.* 89 (2001) 61–67 [MR](#)
- [11] **R Ghrist**, *Configuration spaces and braid groups on graphs in robotics*, from “Knots, braids, and mapping class groups: papers dedicated to Joan S Birman” (J Gilman, W W Menasco, X-S Lin, editors), *AMS/IP Stud. Adv. Math.* 24, Amer. Math. Soc., Providence, RI (2001) 29–40 [MR](#)
- [12] **J M Harrison, J P Keating, J M Robbins, A Sawicki**, *n -particle quantum statistics on graphs*, *Comm. Math. Phys.* 330 (2014) 1293–1326 [MR](#)
- [13] **J H Kim, K H Ko, H W Park**, *Graph braid groups and right-angled Artin groups*, *Trans. Amer. Math. Soc.* 364 (2012) 309–360 [MR](#)
- [14] **D E Knuth**, *Big omicron and big omega and big theta*, *SIGACT News* 8 (1976) 18–24
- [15] **K H Ko, J H La, H W Park**, *Graph 4-braid groups and Massey products*, *Topology Appl.* 197 (2016) 133–153 [MR](#)
- [16] **K H Ko, H W Park**, *Characteristics of graph braid groups*, *Discrete Comput. Geom.* 48 (2012) 915–963 [MR](#)
- [17] **D Lütgehetmann**, *Configuration spaces of graphs*, master’s thesis, Freie Universität Berlin (2014) Available at <https://tinyurl.com/luetgehetmann-master>
- [18] **D Lütgehetmann**, *Representation stability for configuration spaces of graphs*, preprint (2017) [arXiv](#)
- [19] **D Lütgehetmann, D Recio-Mitter**, *Topological complexity of configuration spaces of fully articulated graphs and banana graphs*, preprint (2018) [arXiv](#)
- [20] **T Maciążek, A Sawicki**, *Homology groups for particles on one-connected graphs*, *J. Math. Phys.* 58 (2017) art. id. 062103, 24 pages [MR](#)
- [21] **T Maciążek, A Sawicki**, *Non-abelian quantum statistics on graphs*, *Comm. Math. Phys.* 371 (2019) 921–973 [MR](#)
- [22] **D McDuff**, *Configuration spaces of positive and negative particles*, *Topology* 14 (1975) 91–107 [MR](#)
- [23] **S Papadima, A I Suciu**, *Algebraic invariants for right-angled Artin groups*, *Math. Ann.* 334 (2006) 533–555 [MR](#)
- [24] **E Ramos**, *Stability phenomena in the homology of tree braid groups*, *Algebr. Geom. Topol.* 18 (2018) 2305–2337 [MR](#)

- [25] **E Ramos**, *Configuration spaces of graphs with certain permitted collisions*, Discrete Comput. Geom. 62 (2019) 912–944 [MR](#)
- [26] **L Sabalka**, *On rigidity and the isomorphism problem for tree braid groups*, Groups Geom. Dyn. 3 (2009) 469–523 [MR](#)
- [27] **J Świątkowski**, *Estimates for homological dimension of configuration spaces of graphs*, Colloq. Math. 89 (2001) 69–79 [MR](#)

Department of Mathematics Education, Teachers College, Kyungpook National University
Daegu, South Korea

Center for Geometry and Physics, Institute for Basic Science
Pohang, South Korea

Department of Mathematics, Harvard University
Cambridge, MA, United States

anbyhee@knu.ac.kr, gabriel.c.drummond.cole@gmail.com,
ben.e.knudsen@gmail.com

<https://drummondcole.com/gabriel/academic/>

Proposed: Benson Farb

Seconded: Haynes R Miller, Mark Behrens

Received: 11 October 2018

Revised: 4 May 2019

Min-max minimal disks with free boundary in Riemannian manifolds

LONGZHI LIN

AO SUN

XIN ZHOU

We establish a min-max theory for constructing minimal disks with free boundary in any closed Riemannian manifold. The main result is an effective version of the partial Morse theory for minimal disks with free boundary established by Fraser. Our theory also includes as a special case the min-max theory for the Plateau problem of minimal disks, which can be used to generalize the famous work by Morse–Tompkins and Shiffman on minimal surfaces in $\mathbb{R}^n$ to the Riemannian setting.

More precisely, we generalize, to the free boundary setting, the min-max construction of minimal surfaces using harmonic replacement introduced by Colding–Minicozzi. As a key ingredient to this construction, we show an energy convexity for weakly harmonic maps with mixed Dirichlet and free boundaries from the half unit 2–disk in $\mathbb{R}^2$ into any closed Riemannian manifold, which in particular yields the uniqueness of such weakly harmonic maps. This is a free boundary analogue of the energy convexity and uniqueness for weakly harmonic maps with Dirichlet boundary on the unit 2–disk proved by Colding and Minicozzi.

35R35, 49J35, 49Q05, 53C43

0. Introduction	472
1. Notation	479
2. Energy convexity for weakly harmonic maps with free boundary	480
3. Free boundary harmonic replacement: existence and regularity	487
4. Overview of the variational approach	491
5. Conformal parametrization	494
6. Construction of the tightening process	496
7. Compactness of maximal slices	511
8. Modifications for the proof of Theorem 0.3 and discussions	524
Appendix	526
References	528

0 Introduction

Given a closed Riemannian manifold $\mathcal{N}^n$ (that is isometrically embedded in $\mathbb{R}^N$) and an embedded submanifold Γ of codimension $l \geq 1$, a map from the unit 2-disk $D \subset \mathbb{R}^2$ into $\mathcal{N}$ with boundary lying on Γ is said to be a *minimal disk with free boundary* if it minimizes area up to the first order among all such maps. One physical model of such surfaces is the soap film whose boundary is constrained (but allowed to move freely) on the boundary of some smooth domain in $\mathbb{R}^3$. When a soap film achieves the equilibrium state, it will minimize the area up to first order. Geometrically, the stationary soap film would have vanishing mean curvature in the interior and meet the boundary of the given domain orthogonally. Therefore, the orthogonality condition at the boundary is called the *free boundary* condition.

After earlier works of Gergonne in 1816 and H Schwarz in 1890, Courant first studied systematically the free boundary problems for minimal surfaces in a series of seminal papers; see Courant [8, Chapter VI] and also Courant and Davids [9]). In particular, Courant and Davids proved that given an embedded closed surface S in $\mathbb{R}^3$ other than the sphere, there exists a minimal disk Σ with free boundary on S under certain linking conditions; see [8, pages 213–218]. Since then there have been immense research activities on this topic. To remove the topological assumption on the constraint surface S , Smyth [51] showed that if S is the boundary of a tetrahedron (which is *nonsmooth*), then there must exist exactly three minimal disks embedded inside the tetrahedron satisfying the free boundary condition. When S is a smooth topological two-sphere, Struwe [52] used the mountain pass lemma to establish the existence of at least one *unstable* minimal disk with free boundary lying on S . In higher dimensions and codimensions (for $(\mathcal{N}, \Gamma)$), Ye [57] obtained the existence of an area-minimizing disk with free boundary when the kernel of $\pi_1(\Gamma) \rightarrow \pi_1(\mathcal{N})$ is nontrivial. About 20 years ago, Fraser [15] developed a partial Morse theory for finding minimal disks with free boundary in any codimension using the perturbed energy approach of Sacks–Uhlenbeck [45], and proved the existence of solutions with bounded Morse index when the relative homotopy group $\pi_k(\mathcal{N}, \Gamma)$ is nontrivial for some $k \in \mathbb{N}$. Recently, Fraser–Schoen [16; 17] established deep relations between the minimal surfaces with free boundary in round balls and the extremal eigenvalue problems.

In this paper, we develop a direct variational theory for constructing minimal disks with free boundary for any pair $(\mathcal{N}, \Gamma)$, using the min-max method. In particular, given a k -parameter family of mappings from the unit disk D into $\mathcal{N}$ such that ∂D is mapped

into Γ , one can associate to it a min-max value analogous to the classical Morse theory. We prove that if the min-max value is nontrivial, then there exists a minimal disk together with (possibly empty) several minimal spheres (possibly with a puncture), which is usually called a *bubble tree* and whose areas sum up to the min-max value (the so-called *energy identity*). Moreover, we prove that every approximate sequence of maps converges to a bubble tree such that the energy identity holds true. By reflecting on this strong convergence property, our results can be viewed as an effective version of Fraser [15]; see more discussion in Remark 0.2(2).

Our theory can be also used to generalize the famous work of Morse–Tompkins [34] and Shiffman [49] on minimal surfaces in $\mathbb{R}^n$ to the Riemannian setting (Theorem 0.3). In particular, if we take Γ to be a Jordan curve in $\mathcal{N}$ and assume that Γ bounds two different strictly minimizing minimal disks $\bar{v}_i: D \rightarrow \mathcal{N}$, $i = 1, 2$, such that $\bar{v}_i|_{\partial D}: \partial D \rightarrow \Gamma$ is a monotone parametrization, then our theory produces finitely many harmonic disks u_k (and minimal spheres) such that when restricted to ∂D , only one will have degree 1 and all others will have degree 0. If additionally one has that the $u_k|_{\partial D}: \partial D \rightarrow \Gamma$ are monotone parametrizations, then in the special case when $\mathcal{N}$ has nonpositive curvature (so that there exist no punctured minimal spheres), our theory produces a third *nonminimizing* minimal disk, and therefore it provides a direct generalization of the work of Morse–Tompkins [34] and Shiffman [49]; see Section 8 for more details.

Another novelty of this paper is reflected by our constructive method. The *Schwarz alternating method* introduced by H Schwarz goes back to the late 1860s, and later it was generalized to an iterative method for finding the solution of an elliptic PDE on a domain which is the union of two overlapping subdomains. In [6], Colding and Minicozzi adapted this method and used the *harmonic replacement* to construct min-max minimal surfaces. During this repeated replacement procedure, at each step one replaces a map u by a map $\tilde{u}$ that coincides with u outside a disk and inside the disk is equal to an energy-minimizing map with the same boundary values as u . A key ingredient to this construction is a version of energy convexity for weakly harmonic maps with Dirichlet boundary and small energy on the unit 2-disk D , which also yields the (quantitative) uniqueness for such weakly harmonic maps; see also Lamm and the first author [27]. In this paper, we generalize this min-max construction of Colding–Minicozzi using harmonic replacement to construct minimal disks with free boundary in any closed Riemannian manifold. To this end, we will show an energy convexity for weakly harmonic maps with mixed Dirichlet and free boundaries from

the half unit 2-disk into any closed Riemannian manifold (Theorem 2.2), which is the free boundary analogue of Colding–Minicozzi’s energy convexity and uniqueness for weakly harmonic maps. We shall remark that a priori it is not at all clear if such energy convexity should hold, due to the complication of the free boundary component of the map. The key to this is an ϵ -regularity (gradient estimate) for weakly harmonic maps with mixed Dirichlet and free boundaries on the half unit 2-disk (Theorem 2.5).

Now we proceed to present the precise mathematical statements of our main results. To make the presentation simpler, we will focus on 1-parameter min-max constructions, though our results extend in a straightforward manner to k parameters. We will use $[0, 1]$ as the parameter space. Consider the total variational space

$$\Omega = \left\{ \sigma: D \times [0, 1] \rightarrow \mathcal{N} \mid \begin{array}{l} \sigma: [0, 1] \rightarrow C^0(\bar{D}, \mathcal{N}) \cap W^{1,2}(D, \mathcal{N}) \text{ is continuous,} \\ \sigma(\cdot, t)(\partial D) \subset \Gamma \text{ for all } t \in [0, 1], \\ \sigma(\cdot, 0) \text{ and } \sigma(\cdot, 1) \text{ are constant maps.} \end{array} \right\}.$$

Each $\beta \in \Omega$ will be called a *sweepout*. Given a map $\beta \in \Omega$, we define Ω_β to be the homotopy class of β in Ω .

Here and in the following, we write $E(\cdot)$ and $\text{Area}(\cdot)$ for the Dirichlet energy and area functionals on $C^0(\bar{D}, \mathcal{N}) \cap W^{1,2}(D, \mathcal{N})$. Associated to each homotopy class Ω_β , there is a min-max value, also called the *width* of Ω_β ,

$$(1) \quad W = W(\Omega_\beta) := \inf_{\gamma \in \Omega_\beta} \max_{s \in [0, 1]} \text{Area}(\gamma(\cdot, s)).$$

As the first main result, we establish a direct variational construction for minimal surfaces associated with this critical value W .

Theorem 0.1 *Given $\beta \in \Omega$ with $W = W(\Omega_\beta) > 0$, there is a sequence of sweepouts $\gamma^j \in \Omega_\beta$ with $\max_{s \in [0, 1]} E(\gamma^j(\cdot, s)) \rightarrow W$ such that for any given $\epsilon > 0$, there exist $\bar{j}$ and $\delta > 0$ such that if $j > \bar{j}$ and*

$$\text{Area}(\gamma^j(\cdot, s)) > W - \delta,$$

then there exist finitely many harmonic maps $u_k: D \rightarrow \mathcal{N}$ which have free boundary $u_k(\partial D) \subset \Gamma$, and finitely many (possibly empty) harmonic maps $\hat{u}_l: S^2 \rightarrow \mathcal{N}$, such that

$$d_V \left(\gamma^j(\cdot, s), \bigcup_k \{u_k\} \cup \bigcup_l \{\hat{u}_l\} \right) < \epsilon.$$

Here we have identified each map u_i with the varifold associated to the map, and d_V denotes the varifold distance. Moreover we have the energy identity

$$\sum_k \text{Area}(u_k) + \sum_l \text{Area}(\hat{u}_l) = W.$$

Remark 0.2 (i) By the work of Sacks–Uhlenbeck [45, Corollary 1.7] and Fraser [15] (see also [14, Lemma 1.1]), we know that these harmonic maps are conformal and hence parametrize minimal disks with free boundary on Γ or minimal spheres.

(ii) Our result is an effective version of Fraser’s result [15] in the sense that we obtain a strong convergence property. In particular, a sequence of maps $\{\gamma^{j_k}(\cdot, s_k)\}$ is usually called a min-max sequence if

$$\lim_{k \rightarrow \infty} \text{Area}(\gamma^{j_k}(\cdot, s_k)) = W.$$

We prove that every min-max sequence will subconverge to a set of minimal disks with free boundary and possibly some minimal spheres in the varifold sense. The essential ingredient is to prove that all the min-max sequences converge to a bubble tree limit and the energy identity holds. Note that the energy identity has caught a lot of attention from mathematicians in conformally invariant variational problems; see eg Chen and Tian [4], Ding and Tian [12], Ding, Li and Liu [11], Jost [25], Parker and Wolfson [36; 37] and Qing and Tian [40], and for quantification results for harmonic maps with free boundary, see eg Jost, Liu and Zhu [26] and Laurain and Petrides [30]. A similar strong convergence property was first proven by Colding and Minicozzi [6] for the min-max construction of minimal spheres, and it played an essential role in their proof of the finite-time extinction for certain 3-dimensional Ricci flow. A similar property was also obtained for the min-max construction of closed minimal surfaces of higher genus by the last author [58; 60], and for min-max construction of closed minimal surfaces via viscosity method by Rivière [43]. To the authors’ knowledge, our work is the first occasion to obtain such a strong property in the context of free boundary problems.

(iii) As a special case of our result, one can take $(\mathcal{N}, \Gamma)$ to be a compact manifold with convex boundary $(M, \partial M)$. Using the convex boundary as barriers, our theory applies in this case and the resulting minimal disks with free boundary on ∂M and minimal spheres will all lie inside M .

Our main result has an almost direct corollary for a min-max construction of minimal disks with fixed boundary. In particular, we now assume Γ to be a Jordan curve

in $\mathcal{N}$. Suppose $\bar{v}_0: D \rightarrow \mathcal{N}$ and $\bar{v}_1: D \rightarrow \mathcal{N}$ are two area-minimizing minimal disks (conformal and harmonic maps), where $\bar{v}_i|_{\partial D}: \partial D \rightarrow \Gamma$ is a monotone parametrization for $i = 1, 2$. The total variational space for the fixed boundary problem will be

$$\Omega_f = \left\{ \sigma: D \times [0, 1] \rightarrow \mathcal{N} \left| \begin{array}{l} \sigma: [0, 1] \rightarrow C^0(\bar{D}, \mathcal{N}) \cap W^{1,2}(D, \mathcal{N}) \text{ is continuous,} \\ \sigma(\cdot, t)(\partial D) \subset \Gamma, \\ \sigma(\cdot, 0) = \bar{v}_0 \text{ and } \sigma(\cdot, 1) = \bar{v}_1. \end{array} \right. \right\}.$$

Given a map $\beta \in \Omega_f$, we define Ω_β to be the homotopy class of β in Ω_f . The width W associated with Ω_β can be defined in the same way, namely

$$(2) \quad W = W(\Omega_\beta) := \inf_{\gamma \in \Omega_\beta} \max_{s \in [0, 1]} \text{Area}(\gamma(\cdot, s)).$$

Then $W \geq \max(\text{Area}(\bar{v}_0), \text{Area}(\bar{v}_1)) > 0$. The next result is a slight variant of [Theorem 0.1](#).

Theorem 0.3 *Given $\beta \in \Omega_f$ with $W = W(\Omega_\beta) > \max(\text{Area}(\bar{v}_0), \text{Area}(\bar{v}_1))$, there is a sequence of sweepouts $\gamma^j \in \Omega_\beta$ with $\max_{s \in [0, 1]} E(\gamma^j(\cdot, s)) \rightarrow W$ such that for any given $\epsilon > 0$, there exist $\bar{j}$ and $\delta > 0$ such that if $j > \bar{j}$ and*

$$\text{Area}(\gamma^j(\cdot, s)) > W - \delta,$$

then there exist finitely many harmonic disks $u_k: D \rightarrow \mathcal{N}$ with $u_k(\partial D) \subset \Gamma$ and finitely many (possibly empty) harmonic spheres $\hat{u}_l: S^2 \rightarrow \mathcal{N}$ with

$$d_V\left(\gamma^j(\cdot, s), \bigcup_k \{u_k\} \cup \bigcup_l \{\hat{u}_l\}\right) < \epsilon.$$

Moreover, when considered as restricted maps from ∂D to Γ , only one map among the u_k has degree 1, whereas all others have degree 0. We also have the energy identity $\sum_k \text{Area}(u_k) + \sum_l \text{Area}(\hat{u}_l) = W$.

Remark We postpone the discussions of this result until [Section 8](#).

We also want to mention the min-max theory for constructing minimal submanifolds with free boundary using geometric measure theory. In the 1960s, Almgren [\[2\]](#) initiated a program to develop a Morse theory for minimal submanifolds (with or without free boundary), and he obtained in [\[3\]](#) the existence of an integral varifold which is stationary with free boundary in the sense of first variation in any dimension and codimension; see more details in [\[31\]](#). Later along this direction, higher regularity was established for hypersurfaces. In particular, Grüter and Jost [\[19\]](#) proved the existence

of an unstable embedded minimal disk inside any compact *convex* domain in $\mathbb{R}^3$. Later, Jost [24, Theorem 4.1] generalized their work to any compact three-manifold which is diffeomorphic to a ball with mean convex boundary. Higher-dimensional results were developed very recently by De Lellis and Ramic [10] (for both free and fixed boundary problems in convex manifolds; see also Montezuma [33]), and by Li and Zhou [31] (for the free boundary problem in any compact manifolds with boundary). We refer to Marques and Neves [32] for other recent developments of the min-max theory using geometric measure theory.

Sketch of main ideas

Here we provide a brief summary of our main new ideas. Though the main scheme follows the approach laid out by Colding–Minicozzi [6] for minimal spheres (see also Zhou [58; 60] for minimal surfaces with higher genus), the presence of free boundary in our setting brings in several main new obstacles.

For the analytic aspect, there are two main ingredients that we have to establish for weakly harmonic maps with (partial) free boundary. The first ingredient is a version of energy convexity which says that the energy functional is strictly convex near a weakly harmonic map with mixed Dirichlet and free boundaries on the half 2–disk. Unlike the proof in [6] where Colding–Minicozzi used the moving frame method developed by Hélein [20] in order to get a Hardy-type estimate for weakly harmonic maps with Dirichlet boundary and small energy on the 2–disk D , we first use the Uhlenbeck–Rivière decomposition method developed by Rivière [41] to get a refined ϵ –regularity (gradient estimate) for weakly harmonic maps with mixed Dirichlet and free boundaries on the half 2–disk (Theorem 2.5), and then appeal to the first-order Hardy inequality (Lemma 2.4); cf Lamm and Lin [27], where the energy density $|\nabla u|^2$ is estimated in the local Hardy space $h^1(D)$. The key geometric observation in the proof of the energy convexity lies in two orthogonality conditions: one observed by Colding–Minicozzi [6], and the other by Zhou [59]. Very recently, this idea also permitted the first author and Lin [29] to obtain an energy convexity and uniqueness for *weakly intrinsic biharmonic maps* defined on the unit 4–ball with small bi-energy, which in particular yields a version of uniqueness for weakly harmonic maps in dimension 4. The other ingredient is the uniform continuity estimate up to the free boundary for weakly harmonic maps. Whereas the continuity up to the Dirichlet boundary was proven by Qing [39], we prove the uniform continuity for weakly harmonic maps with mixed Dirichlet and free boundaries by a careful covering argument using our previous gradient estimate.

We think that the results and techniques in both of the ingredients are of independent interest for the free boundary problem of harmonic maps.

The variational construction consists mainly of two parts. The first part is a two-stage tightening process, following closely the approach of Colding–Minicozzi [6] (this can be made a 2^k –stage process when the parameter space is k –dimensional). The main challenges here include the $W^{1,2} \cap C^0$ –continuity for (partial) free boundary harmonic replacements, and an interpolation construction to prove energy improvement inequalities. In fact, the key idea in both places is to construct comparison maps and then use the energy minimality conditions. One natural candidate of comparison maps is the linear interpolation in $\mathbb{R}^N$ of two maps $u_i: D \rightarrow \mathcal{N}$, $i = 1, 2$. Though such a map will go outside of $\mathcal{N}$, Colding–Minicozzi in [6] used the nearest-point projection to pull it back to $\mathcal{N}$. However the projection does not necessarily map the image of ∂D into Γ . Therefore we have to develop new methods to overcome this issue. To prove the $W^{1,2}$ –continuity, we find a way to reduce the problem to an interpolation between two curves; and in the second place, we choose to do interpolation in Fermi coordinates. Based on the two new results, the tightening process can then be carried through in the free boundary setting analogously to [6].

The second part is a bubbling convergence procedure for almost harmonic maps with free boundary. The bubbling convergence for almost harmonic maps on spheres was developed by Colding–Minicozzi [6, Appendix B], and for free boundary α –harmonic maps it has been investigated systematically by Fraser [15]. Our result can be viewed as a combination of the two results. Among several things, the most novel observation in this part is the asymptotic analysis for harmonic maps with free boundary defined on a long half-cylinder (which is conformally equivalent to a thin half-annulus). In particular, we prove that the angular energy is much smaller than the total energy of this map. Since boundary terms will appear in the integration by parts argument, we have to use our gradient estimate together with a delicate doubling argument to take care of these boundary terms, so as to carry out Colding–Minicozzi’s method (which works on cylinders).

Layout of the paper

The paper is organized as follows. We fix some notation in Section 1. In Section 2, we prove the energy convexity for weakly harmonic maps defined on the half-disk with mixed Dirichlet and free boundaries (Theorem 2.2). In Section 3, we prove that a

weakly harmonic map with mixed Dirichlet and free boundaries on the half-disk with continuous partial Dirichlet boundary is also continuous on the whole free boundary including the corner points; and we also present the proof on the existence of a weakly harmonic map with free boundary and with prescribed partial Dirichlet boundary, which we call the *partial free boundary harmonic replacement*. In [Section 4](#), we outline the main ingredients needed to establish the min-max theory in the free boundary setting. In [Section 5](#), we prove that the min-max values for the area and energy functionals are the same by using conformal reparametrizations. In [Section 6](#), we carry out the construction of the tightening process; in particular, we show how to use a two-stage harmonic replacement procedure to make the sweepout as tight as possible. In [Section 7](#), we prove that any min-max sequence of maps will converge to a bubble tree consisting of harmonic disks and harmonic spheres; here we also show that the bubbling convergence satisfies the energy identity, or equivalently, the total energy of the bubble tree is the same as the min-max value. In [Section 8](#), we point out necessary changes needed to adapt our theory to the min-max Plateau problem ([Theorem 0.3](#)).

Acknowledgements The authors would like to thank Professor William Minicozzi for encouragement and helpful comments, and also Professors Rugang Ye and Tony Tromba for helpful discussions. Zhou is partially supported by NSF grant DMS-1811293.

Added in proof We were informed that Laurain and Petrides have recently proved a similar result, and [Theorem 2.2](#) was announced in a seminar proceedings paper [\[28\]](#), but without a proof.

1 Notation

We first fix some notation.

- $\mathbb{R}^2$ denotes the Euclidean two-plane, where (x_1, x_2) (or (x, y)) and (r, θ) are the Cartesian and polar coordinates, respectively.
- $\mathbb{H}^2 = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 > 0\}$ denotes the upper half-plane.
- $D_r = D_r(0) = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 < r^2\}$ denotes the disk of radius r centered at the origin.
- $D_s^+ := D_s \cap \mathbb{H}^2 = \{(r, \theta) : 0 \leq r < s, 0 < \theta < \pi\}$ denotes the upper half-disk with radius s .
- For simplicity, we sometimes write $D^+ = D_1^+$.

Then we write $\partial D_s^+ = \partial_s^C \cup \partial_s^A$, which is the union of the chord (diameter) and the arc of the upper semicircle, ie

$$\partial_s^C = \{(r, \theta) : 0 \leq r \leq s, \theta = 0 \text{ or } \pi\} \quad \text{and} \quad \partial_s^A = \{(r, \theta) : r \equiv s, 0 \leq \theta \leq \pi\}.$$

Similarly, we will write

$$\partial^C = \partial_1^C \quad \text{and} \quad \partial^A = \partial_1^A.$$

We write $u: (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ if u is a map from D to $\mathcal{N}$ and $u(\partial D) \subset \Gamma$. Similarly, we write $u: (D_s^+, \partial_s^C) \rightarrow (\mathcal{N}, \Gamma)$ if u is a map from D_s^+ to $\mathcal{N}$ and $u(\partial_s^C) \subset \Gamma$.

Given $u: D \rightarrow \mathcal{N}$ or $u: D_s^+ \rightarrow \mathcal{N}$, the Dirichlet energy is defined as

$$E(u) := \frac{1}{2} \int_{D \text{ or } D_s^+} |\nabla u|^2 dx dy.$$

The Euler–Lagrange equation for the Dirichlet energy is the so-called *harmonic map equation*, which is a quasilinear system defined by

$$(3) \quad -\Delta u = A(u)(\nabla u, \nabla u),$$

where $A(u)$ is the second fundamental form of the embedding $\mathcal{N} \hookrightarrow \mathbb{R}^N$. See eg [20].

Definition 1.1 A $W^{1,2}$ -map $u: (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ is called a *weakly harmonic map with free boundary* if u satisfies the harmonic map equation weakly in D and

$$\frac{\partial u}{\partial r} \perp \Gamma \quad \text{along } \partial D.$$

Similarly a $W^{1,2}$ -map $u: (D_s^+, \partial_s^C) \rightarrow (\mathcal{N}, \Gamma)$ is called a *weakly harmonic map with partial free boundary* if u satisfies the harmonic map equation weakly in D_s^+ and

$$\frac{\partial u}{\partial y} \perp \Gamma \quad \text{along } \partial_s^C.$$

2 Energy convexity for weakly harmonic maps with partial free boundary

In this section, we present the first main result, that is, the energy convexity and uniqueness for weakly harmonic maps with mixed Dirichlet and free boundaries. This is not only one of the key ingredients of our min-max existence theory for minimal disks with free boundary, but it also has its own interest from the point of view of PDEs and calculus of variations.

We first recall Colding–Minicozzi’s energy convexity for weakly harmonic maps defined on the 2–disk D with Dirichlet boundary and small energy; see also [27].

Theorem 2.1 [6, Theorem 3.1] *There exists a constant $\varepsilon_0 > 0$, depending only on $\mathcal{N}$, such that if $u, v \in W^{1,2}(D, \mathcal{N})$, $u = v$ on ∂D , and u is weakly harmonic with $E(u) \leq \varepsilon_0$, then*

$$(4) \quad \frac{1}{2} \int_D |\nabla v - \nabla u|^2 dx \leq \int_D |\nabla v|^2 dx - \int_D |\nabla u|^2 dx.$$

In this section, we prove a free boundary analogue of Theorem 2.1. We will abuse notation and still denote the energy threshold as ε_0 . More precisely, we prove:

Theorem 2.2 (energy convexity for weakly harmonic maps with mixed Dirichlet and free boundaries) *There exists a constant $\varepsilon_0 > 0$, depending only on $\mathcal{N}$ and Γ , such that if $u, v \in W^{1,2}(D^+, \mathcal{N})$ with $u|_{\partial \mathcal{A}} = v|_{\partial \mathcal{A}}$, $u|_{\partial \mathcal{C}} \subset \Gamma$, $v|_{\partial \mathcal{C}} \subset \Gamma$, $E(u) \leq \varepsilon_0$, and u is a weakly harmonic map with partial free boundary, then we have the energy convexity*

$$(5) \quad \frac{1}{2} \int_{D^+} |\nabla v - \nabla u|^2 dx \leq \int_{D^+} |\nabla v|^2 dx - \int_{D^+} |\nabla u|^2 dx.$$

An immediate corollary of Theorem 2.2 is the uniqueness of weakly harmonic maps with mixed Dirichlet and free boundaries on D^+ .

Corollary 2.3 *There exists an $\varepsilon_0 > 0$, depending only on $\mathcal{N}$ and Γ , such that for any two weakly harmonic maps $u, v \in W^{1,2}(D^+, \mathcal{N})$ with Dirichlet boundary $u|_{\partial \mathcal{A}} = v|_{\partial \mathcal{A}}$ and free boundaries $u|_{\partial \mathcal{C}} \subset \Gamma$ and $v|_{\partial \mathcal{C}} \subset \Gamma$, if their energies satisfy $E(u) \leq \varepsilon_0$ and $E(v) \leq \varepsilon_0$, then we have $u \equiv v$ in D^+ .*

In order to prove Theorem 2.2, we will use the following first-order Hardy inequality and appeal to the refined ϵ –regularity for weakly harmonic maps with mixed Dirichlet and free boundaries (Theorem 2.5).

Lemma 2.4 (Hardy inequality) *Let u, v be in $W^{1,2}(D^+, \mathbb{R}^N)$ with $u|_{\partial \mathcal{A}} = v|_{\partial \mathcal{A}}$. Then we have*

$$(6) \quad \int_{D^+} |v - u|^2 \cdot \frac{1}{(1 - |x|)^2} dx \leq 4 \int_{D^+} |\nabla(v - u)|^2 dx,$$

where $x = (x_1, x_2) \in D^+$.

Proof First we extend u and v across ∂^C by reflection, setting $\tilde{u}(x_1, x_2) = u(-x_1, x_2)$ and $\tilde{v}(x_1, x_2) = v(-x_1, x_2)$ for $x_1 \leq 0$, to get $\tilde{v} - \tilde{u} \in W_0^{1,2}(D)$. Approximate $\tilde{v} - \tilde{u}$ in $W^{1,2}$ by a sequence of smooth functions with compact support $\tilde{w}_i \in C_c^\infty(D)$, and let w_i be the restriction of $\tilde{w}_i$ to D^+ . Now for each w_i we have

$$\begin{aligned}
 (7) \quad \int_{D^+} \frac{|w_i|^2}{(1-|x|)^2} dx &= \int_0^\pi \int_0^1 \frac{|w_i|^2 r}{(1-r)^2} dr d\theta \\
 &= \int_0^\pi \left[\left| \frac{|w_i|^2 r}{1-r} \right|_0^1 - \int_0^1 \left(\frac{|w_i|^2}{1-r} + \frac{2w_i \cdot (w_i)_r r}{1-r} \right) dr \right] d\theta \\
 &\leq 2 \left(\int_0^\pi \int_0^1 \frac{|w_i|^2 r}{(1-r)^2} dr d\theta \right)^{\frac{1}{2}} \left(\int_0^\pi \int_0^1 |(w_i)_r|^2 r dr d\theta \right)^{\frac{1}{2}} \\
 &\leq 2 \left(\int_{D^+} \frac{|w_i|^2}{(1-|x|)^2} dx \right)^{\frac{1}{2}} \left(\int_{D^+} |\nabla w_i|^2 dx \right)^{\frac{1}{2}},
 \end{aligned}$$

which yields

$$(8) \quad \int_{D^+} \frac{|w_i|^2}{(1-|x|)^2} dx \leq 4 \int_{D^+} |\nabla w_i|^2 dx.$$

Now (6) follows from (8) and Fatou's lemma. $\square$

The next result is a refined ϵ -regularity for weakly harmonic maps on D^+ with Dirichlet and free boundaries; cf [46; 30; 26] and Fraser [15]. This ϵ -regularity is crucial to the proof of Theorem 2.2. We shall remark that such ϵ -regularity is well-known for weakly harmonic maps defined on the 2-disk D with Dirichlet boundary, see eg Qing [39, Lemma 4].

Theorem 2.5 *There exists a constant $\varepsilon_0 > 0$, depending only on $\mathcal{N}$ and Γ , such that if $u \in W^{1,2}(D^+, \mathcal{N})$ is a weakly harmonic map with mixed Dirichlet boundary on ∂^A and free boundary $u|_{\partial^C} \subset \Gamma$ and $E(u) \leq \varepsilon_0$, then for any $x \in D^+ \cup (\partial^C)^\circ$ we have*

$$(9) \quad |\nabla u|(x) \leq \frac{C\sqrt{\varepsilon_0}}{1-|x|}$$

for some constant $C > 0$ that only depends on $\mathcal{N}$ and Γ .

Proof By the reflection across the free boundary Γ constructed by Scheven in [46], $u \in W^{1,2}(D^+)$ can be extended to $\tilde{u} \in W^{1,2}(D)$ so that $\tilde{u}$ weakly solves in D the system of equations

$$(10) \quad \operatorname{div}(Q\nabla\tilde{u}) = \omega \cdot Q\nabla\tilde{u}$$

for some $\omega = (\omega_j^i)_{1 \leq i, j \leq n} \in L^2(D, \mathfrak{so}(n) \otimes \wedge^1 \mathbb{R}^2)$ (ie $\omega_i^j = -\omega_j^i$) and $Q \in W^{1,2} \cap L^\infty(D, \mathrm{GL}(n, \mathbb{R}))$ such that

$$(11) \quad |\omega| \leq C |\nabla \tilde{u}| \quad \text{a.e. in } D \quad \text{and} \quad \|Q\|_{L^\infty(D)} + \|Q^{-1}\|_{L^\infty(D)} \leq C,$$

where $C > 0$ is a constant depending only on $\mathcal{N}$ and Γ ; see Jost, Liu and Zhu [26, Proposition 3.3]. Moreover, by the assumption $E(u) \leq \varepsilon_0$ and the reflection construction, we have

$$(12) \quad \|\nabla \tilde{u}\|_{L^2(D)} \leq C \sqrt{\varepsilon_0} \quad \text{and} \quad \|\omega\|_{L^2(D)} \leq C \sqrt{\varepsilon_0}.$$

Then, using the Uhlenbeck–Rivière decomposition method developed by Rivière [41] and Rivière–Struwe [44] in the study of regularity of elliptic PDEs with antisymmetric structure, we can obtain a constant $\alpha > 0$ such that (cf [42, page 50])

$$(13) \quad \sup_{0 < \rho < \frac{1}{4}, p \in D_{1/2}} \rho^{-\alpha} \int_{D_\rho(p)} |\Delta \tilde{u}| \, dx \leq C \sqrt{\varepsilon_0}.$$

Proof of (13) To see this, we first note that since $\tilde{u} - \bar{\tilde{u}}$ also satisfies (10) with the same Q and ω , where $\bar{\tilde{u}} = \frac{1}{\pi} \int_D \tilde{u} \, dx$ is the average of $\tilde{u}$ on D , without loss of generality we may assume that $\bar{\tilde{u}} = 0$. Then by the work of Sharp (see eg [47, Corollary 1.4, Proposition 3.1]) on the higher integrability for solutions to a system of PDEs similar to (10), we know that there exists a constant $C > 0$ such that

$$(14) \quad \|\nabla^2 \tilde{u}\|_{L^1(D_{1/2})} \leq C \|\tilde{u}\|_{L^1(D)} \leq C \|\nabla \tilde{u}\|_{L^1(D)} \leq C \sqrt{\varepsilon_0}.$$

See [48, Theorem 1.2] and [26, Theorem 2.4]. Here we have used the Poincaré–Wirtinger inequality in the second inequality. We note that this higher integrability (14) essentially follows from the stability of the local Hardy space $h^1(D)$ (see eg Lamm and Lin [27, Section A.2]) under multiplication by Hölder continuous functions, coupled with the Hölder continuity of $\tilde{u}$ in D proved by Scheven [46, Theorem 4.1].

Then by the continuous embedding of $W^{1,1}(D)$ into $L^{2,1}(D)$, where $L^{2,1}$ is the Lorentz space (see eg [22; 38; 55; 20]), we have

$$(15) \quad \|\nabla \tilde{u}\|_{L^{2,1}(D_{1/2})} \leq C \|\nabla \tilde{u}\|_{W^{1,1}(D_{1/2})} \leq C \|\nabla \tilde{u}\|_{L^1(D)} \leq C \sqrt{\varepsilon_0}.$$

Now for any $p \in D_{\frac{1}{2}}$ and $0 < r < \frac{1}{2}$, we use the Hodge decomposition (see eg [23, Corollary 10.5.1]) to find $A \in W^{1,2}(D_r(p), \mathbb{R}^n)$ and $B \in W_0^{1,2}(D_r(p), \mathbb{R}^n)$ such that

$$(16) \quad Q \nabla \tilde{u} = \nabla A + \nabla^\perp B \quad \text{in } D_r(p),$$

where $\nabla^\perp := (-\partial_{x_2}, \partial_{x_1})$ and we have

$$\|\nabla A\|_{L^2(D_r(p))} + \|\nabla B\|_{L^2(D_r(p))} \leq \tilde{C} \|\nabla \tilde{u}\|_{L^2(D_r(p))}.$$

Then we have, taking divergence on both sides of the first equation in (16),

$$(17) \quad \Delta A = \operatorname{div}(Q \nabla \tilde{u}) = \omega \cdot Q \nabla \tilde{u} \quad \text{in } D_r(p),$$

and, taking curl on both sides of the first equation in (16),

$$(18) \quad \begin{cases} \Delta B = \nabla^\perp Q \cdot \nabla \tilde{u} & \text{in } D_r(p), \\ B = 0 & \text{on } \partial D_r(p). \end{cases}$$

Now let $A = A_1 + A_2$ on $D_r(p)$ so that $\Delta A_2 = 0$ and

$$(19) \quad \begin{cases} \Delta A_1 = \operatorname{div}(Q \nabla \tilde{u}) & \text{in } D_r(p), \\ A_1 = 0 & \text{on } \partial D_r(p). \end{cases}$$

Then using [20, Theorem 3.3.3] (which implies that the standard L^p -theory extends to Lorentz spaces), we furthermore get

$$(20) \quad \|\nabla A_1\|_{L^{2,1}(D_r(p))} \leq C \|Q \nabla \tilde{u}\|_{L^{2,1}(D_r(p))} \leq C \|\nabla \tilde{u}\|_{L^{2,1}(D_r(p))}.$$

Hence by [20, Theorem 3.3.4] we conclude that

$$(21) \quad \begin{aligned} \|A_1\|_{L^\infty(D_r(p))} &\leq C(\|A_1\|_{L^{2,1}(D_r(p))} + \|\nabla A_1\|_{L^{2,1}(D_r(p))}) \\ &\leq C \|\nabla A_1\|_{L^{2,1}(D_r(p))} \\ &\leq C \sqrt{\varepsilon_0}, \end{aligned}$$

where we again used [20, Theorem 3.3.3] (which ensures that the Poincaré inequality extends to Lorentz spaces) and the fact that $A_1 = 0$ on $\partial D_r(p)$ in the second estimate. Thus, an integration by parts yields

$$(22) \quad \begin{aligned} \|\nabla A_1\|_{L^2(D_r(p))}^2 &= - \int_{D_r(p)} A_1 \Delta A_1 \, dx \\ &\leq \|A_1\|_{L^\infty(D_r(p))} \int_{D_r(p)} |\omega \cdot Q \nabla \tilde{u}| \, dx \\ &\leq C \sqrt{\varepsilon_0} \|\nabla \tilde{u}\|_{L^2(D_r(p))}^2, \end{aligned}$$

where we have used (12) and the fact that $Q \in W^{1,2} \cap L^\infty(D, \operatorname{GL}(n, \mathbb{R}))$. Note that again by integration by parts (using $A_1 = 0$ on $\partial D_r(p)$ and $\Delta A_2 = 0$) we have

$$(23) \quad \|\nabla A_2\|_{L^2(D_r(p))}^2 = \|\nabla A\|_{L^2(D_r(p))}^2 - \|\nabla A_1\|_{L^2(D_r(p))}^2 \leq \tilde{C} \|\nabla \tilde{u}\|_{L^2(D_r(p))}^2.$$

Now since A_2 is harmonic on $D_r(p)$ we know that for every $q \in D_r(p)$ the function

$$\rho \rightarrow \frac{1}{\rho^2} \int_{D_\rho(q)} |\nabla A_2|^2 dx$$

is increasing; see eg [42, Lemma IV.1]. Now let $\bar{C} > 0$ be such that $\|Q^{-1}\|_{L^\infty(D)} \leq \bar{C}$ and

$$\delta = \min \left\{ \frac{1}{4\sqrt{\tilde{C}\bar{C}}}, \frac{1}{2} \right\},$$

where $\tilde{C}$ is from (23). Then we have

$$(24) \quad \int_{D_{\delta r}(p)} |\nabla A_2|^2 dx \leq \frac{1}{16\tilde{C}\bar{C}} \int_{D_r(p)} |\nabla A_2|^2 dx \leq \frac{1}{16\bar{C}} \|\nabla \tilde{u}\|_{L^2(D_r(p))}^2.$$

Now using (18) and by the results of Coifman–Lions–Meyer–Semmes [5] or Wente’s lemma [56], we know

$$(25) \quad \|B\|_{L^\infty(D_r(p))} + \|\nabla B\|_{L^{2,1}(D_r(p))} \leq C \|\nabla \tilde{u}\|_{L^2(D_r(p))}^2.$$

Therefore, combining (16), (22), (24) and (25) we have

$$(26) \quad \begin{aligned} & \|\nabla \tilde{u}\|_{L^2(D_{\delta r}(p))}^2 \\ & \leq \|Q^{-1}\|_{L^\infty(D)} (2\|\nabla A_1\|_{L^2(D_r(p))}^2 + 2\|\nabla A_2\|_{L^2(D_{\delta r}(p))}^2 + \|\nabla B\|_{L^2(D_r(p))}^2) \\ & \leq C\sqrt{\varepsilon_0} \|\nabla \tilde{u}\|_{L^2(D_r(p))}^2 + \frac{1}{8} \|\nabla \tilde{u}\|_{L^2(D_r(p))}^2 + C \|\nabla \tilde{u}\|_{L^2(D_r(p))}^4. \end{aligned}$$

Choosing ε_0 sufficiently small we arrive at

$$(27) \quad \|\nabla \tilde{u}\|_{L^2(D_{\delta r}(p))}^2 \leq \frac{1}{2} \|\nabla \tilde{u}\|_{L^2(D_r(p))}^2$$

for any $p \in D_{\frac{1}{2}}$ and $0 < r < \frac{1}{2}$. Iterating this inequality gives the existence of a constant $\alpha > 0$ such that for all $p \in D_{\frac{1}{2}}$ and all $\rho < \frac{1}{4}$, one has

$$(28) \quad \rho^{-2\alpha} \int_{D_\rho(p)} |\nabla \tilde{u}|^2 dx \leq C \int_D |\nabla \tilde{u}|^2 dx \leq C\varepsilon_0.$$

Now by (10) we have

$$\Delta \tilde{u} = Q^{-1}(-\nabla Q \nabla \tilde{u} + \omega \cdot Q \nabla \tilde{u}).$$

Then using (11), for all $p \in D_{\frac{1}{2}}$ and all $\rho < \frac{1}{4}$ we have

$$\begin{aligned}
 (29) \quad & \rho^{-\alpha} \int_{D_\rho(p)} |\Delta \tilde{u}| \, dx \\
 & \leq C \rho^{-\alpha} \int_{D_\rho(p)} |\nabla Q| |\nabla \tilde{u}| + |\nabla \tilde{u}|^2 \, dx \\
 & \leq C \rho^{-\alpha} \left[\left(\int_{D_\rho(p)} |\nabla Q|^2 \, dx \right)^{\frac{1}{2}} \left(\int_{D_\rho(p)} |\nabla \tilde{u}|^2 \, dx \right)^{\frac{1}{2}} + \int_{D_\rho(p)} |\nabla \tilde{u}|^2 \, dx \right] \\
 & \leq C \sqrt{\varepsilon_0}.
 \end{aligned}$$

Here we have used $\|\nabla Q\|_{L^2(D)} \leq C$ for some $C > 0$ depending only on $\mathcal{N}$ and Γ . This proves (13). $\square$

Now using (13), a classical estimate on Riesz potentials then gives, for all $p \in D_{\frac{1}{4}}$,

$$(30) \quad |\nabla \tilde{u}|(p) \leq C \frac{1}{|x - p|} * \chi_{D_{1/2}} |\Delta \tilde{u}| + C \|\nabla \tilde{u}\|_{L^2(D)},$$

where $\chi_{D_{1/2}}$ is the characteristic function of the ball $D_{\frac{1}{2}}$. Together with injections proved by Adams in [1] (see also [18, Exercise 6.1.6]), the latter shows that

$$\|\nabla \tilde{u}\|_{L^r(D_{1/4})} \leq C \sqrt{\varepsilon_0}$$

for some $r > 2$. Reinjecting this into the equation (10) and bootstrapping the estimates, we get

$$\|\nabla \tilde{u}\|_{L^\infty(D_{1/8})} \leq C \sqrt{\varepsilon_0}.$$

In particular, $|\nabla \tilde{u}|(0) \leq C \sqrt{\varepsilon_0}$. Then by a scaling argument we get the desired estimate (9). $\square$

Now we are ready to prove Theorem 2.2.

Proof of Theorem 2.2 In order to prove the energy convexity (5), it suffices to show

$$(31) \quad \Psi \geq -\frac{1}{2} \int_{D^+} |\nabla(v - u)|^2 \, dx,$$

where (using the boundary conditions and the harmonic map equation)

$$\begin{aligned}
 (32) \quad \Psi &:= \int_{D^+} |\nabla v|^2 \, dx - \int_{D^+} |\nabla u|^2 \, dx - \int_{D^+} |\nabla(v - u)|^2 \, dx \\
 &= 2 \int_{D^+} \langle \nabla(v - u), \nabla u \rangle \, dx \\
 &= 2 \int_{D^+} \langle v - u, A(u)(\nabla u, \nabla u) \rangle \, dx + 2 \int_{\partial C} \left\langle v - u, \frac{\partial u}{\partial \nu} \right\rangle ds.
 \end{aligned}$$

Here $\nu = (0, -1)$ is the outward unit normal to ∂^C . Now we note that for any $p, q \in \mathcal{N}$ (resp. Γ), there exists a constant $C > 0$ depending only on $\mathcal{N}$ (resp. Γ) such that

$$|(p - q)^\perp| \leq C|p - q|^2,$$

where the superscript $\perp$ denotes the normal component of the vector $p - q$ at q ; see eg [7, Lemma A.1]. Therefore, using

$$-\Delta u = A(u)(\nabla u, \nabla u)^\perp T_u \mathcal{N},$$

the Cauchy–Schwarz inequality together with (32) yields

$$(33) \quad \Psi \geq -C_1 \int_{D^+} |v - u|^2 |\nabla u|^2 dx - C_2 \int_{\partial^C} |v - u|^2 \left| \frac{\partial u}{\partial \nu} \right| ds,$$

where $C_1 > 0$ depends only on $\mathcal{N}$ and $C_2 > 0$ depends only on Γ . Now by Lemma 2.4, Theorem 2.5 and using the facts that $v = u$ on ∂^A and $x_2 = 0$ on $\partial^C = [-1, 1] \times \{0\}$, we have

$$\begin{aligned} \Psi &\geq -C_3 \varepsilon_0 \int_{D^+} \frac{|v - u|^2}{(1 - |x|)^2} dx - C_4 \sqrt{\varepsilon_0} \int_{\partial^C} \frac{|v - u|^2}{1 - |x_1|} ds \\ &\geq -4C_3 \varepsilon_0 \int_{D^+} |\nabla(v - u)|^2 dx + C_4 \sqrt{\varepsilon_0} \int_{-1}^1 \int_0^{\sqrt{1-x_1^2}} \partial_{x_2} \left(\frac{|v - u|^2}{1 - |x_1|} \right) dx_2 dx_1 \\ &= -4C_3 \varepsilon_0 \int_{D^+} |\nabla(v - u)|^2 dx + C_4 \sqrt{\varepsilon_0} \int_{D^+} \frac{\langle v - u, (v - u)_{x_2} \rangle}{1 - |x_1|} dx \\ &\geq -4C_3 \varepsilon_0 \int_{D^+} |\nabla(v - u)|^2 dx - \frac{1}{4} \int_{D^+} |\nabla(v - u)|^2 dx - C_5 \sqrt{\varepsilon_0} \int_{D^+} \frac{|v - u|^2}{(1 - |x|)^2} dx \\ &\geq -C_6 \sqrt{\varepsilon_0} \int_{D^+} |\nabla(v - u)|^2 dx - \frac{1}{4} \int_{D^+} |\nabla(v - u)|^2 dx. \end{aligned}$$

In the second-to-last inequality we have used Young's inequality. Choosing ε_0 sufficiently small that $C_6 \sqrt{\varepsilon_0} \leq \frac{1}{4}$, we get the desired estimate (31). $\square$

3 Existence and regularity of partial free boundary harmonic replacement

In this section, we discuss the existence and regularity of the (partial) free boundary version of the harmonic replacement. The fixed boundary harmonic replacement was discussed in Colding–Minicozzi [6].

3.1 Continuity of weakly harmonic maps with partial free boundary

In this section, we prove the C^0 regularity up to the boundary for weakly harmonic maps with mixed Dirichlet and free boundaries and small energy. For weakly harmonic maps with fixed continuous Dirichlet boundary, this C^0 boundary regularity was proved by Qing [39].

Let $u: (D^+, \partial^C) \rightarrow (\mathcal{N}, \Gamma)$ be a weakly harmonic map with mixed Dirichlet and free boundaries and small energy $E(u) \leq \varepsilon_0$ (where ε_0 is from Theorem 2.5) such that the Dirichlet boundary $u: \partial^A \rightarrow \mathcal{N}$ is continuous. Then by Helein's interior regularity [20], u is smooth in D^+ ; by Qing's (Dirichlet) boundary regularity [39], u is continuous up to the interior of the Dirichlet boundary on ∂^A ; moreover, by the (free) boundary regularity result of Scheven [46], u is also smooth up to the interior of the free boundary on ∂^C . So the only thing left to be verified is that u is continuous up to the two corner points $(1, 0)$ and $(1, \pi)$ (in polar coordinates on $\overline{D^+}$), ie the endpoints of the Dirichlet boundary on ∂^A or the free boundary on ∂^C .

It suffices to prove the continuity around $p = (1, \pi)$.¹ We will first prove that there exists a sequence of points $\{x_i\}$ on ∂^C , converging to p , such that the $u(x_i)$ are all close to $u(p)$; next we show that all the intermediate points in $u([x_i, x_{i+1}])$ are also close to $u(p)$. For the convenience of later proof, we will parallel shift D^+ in $\mathbb{R}^2$ so that $p = (0, 0)$. Now $D^+ = \{(x, y) \in \mathbb{R}^2 : y > 0, (x-1)^2 + y^2 < 1\}$ and $\partial^A = \{(x, y) \in \mathbb{R}^2 : y > 0, (x-1)^2 + y^2 = 1\}$, and we will use polar coordinates on $\mathbb{R}^2$.

We first need a variant of the Courant–Lebesgue lemma (cf [39, Lemma 2]) at the corner point p .

Lemma 3.1 *Given $0 < l < \frac{1}{2}$, there exists $l' \in (l, 2l)$ such that*

$$(34) \quad \int_0^{\alpha(l')} \left| \frac{\partial u}{\partial \theta} \right|^2 (l', \theta) d\theta \leq \frac{1}{\log 2} E(u|_{(D_{2l}(0) \setminus D_l(0)) \cap D^+}),$$

where $\alpha(l')$ is the angle such that $(l', \alpha(l')) \in \partial^A$. Consequently, we have

$$(35) \quad |u(l', \theta_1) - u(l', \theta_2)| \leq \sqrt{\frac{N}{\log 2}} E(u|_{(D_{2l}(0) \setminus D_l(0)) \cap D^+})^{\frac{1}{2}} |\theta_1 - \theta_2|^{\frac{1}{2}},$$

where N is the dimension of the ambient space of the embedding of $\mathcal{N}$ into $\mathbb{R}^N$.

¹Given Scheven's reflection construction and equation (10), one may also use the Uhlenbeck–Rivière decomposition method as in Müller–Schikorra [35] to prove boundary regularity.

Proof The proof is virtually the same as the proof of [39, Lemma 2]:

$$E(u|_{(D_{2l}(0) \setminus D_l(0)) \cap D^+}) \geq \int_l^{2l} \int_0^{\alpha(r)} \left(\left| \frac{\partial u}{\partial r} \right|^2 + \frac{1}{r^2} \left| \frac{\partial u}{\partial \theta} \right|^2 \right) r \, d\theta \, dr,$$

where $\alpha(r)$ is the angle such that

$$(r, \alpha(r)) \in \partial^A = \{(x, y) \in \mathbb{R}^2 : y > 0, (x-1)^2 + y^2 = 1\}.$$

Therefore there exists $l' \in (l, 2l)$ such that

$$\int_0^{\alpha(l')} \left| \frac{\partial u}{\partial \theta} \right|^2 (l', \theta) \, d\theta \leq \frac{1}{\log 2} E(u|_{(D_{2l}(0) \setminus D_l(0)) \cap D^+}).$$

Then we have

$$\begin{aligned} |u(l', \theta_1) - u(l', \theta_2)| &\leq \left| \int_{\theta_1}^{\theta_2} \frac{\partial u}{\partial \theta} (l', \theta) \, d\theta \right| \\ &\leq \sqrt{\frac{N}{\log 2}} E(u|_{(D_{2l}(0) \setminus D_l(0)) \cap D^+})^{\frac{1}{2}} |\theta_1 - \theta_2|^{\frac{1}{2}}. \quad \square \end{aligned}$$

Now we are ready to prove the continuity up to the corners on the boundary of D^+ .

Theorem 3.2 *There exists an $\varepsilon_0 > 0$ such that if $u: (D^+, \partial^C) \rightarrow (\mathcal{N}, \Gamma)$ is a weakly harmonic map with mixed Dirichlet and free boundaries with $E(u) \leq \varepsilon_0$, and u is continuous on ∂^A , then u is continuous on $\overline{D^+}$.*

Proof Since u is continuous on ∂^A , by the discussion prior to Lemma 3.1 it suffices to show that given $\varepsilon > 0$, there exists $\delta > 0$ such that for $x \in \partial^C$ and $|x - p| \leq \delta$ we have $|u(x) - u(p)| \leq \varepsilon$. By the continuity of u on ∂^A , we can find $\delta_1 > 0$ such that if $p_l = (l, \alpha(l)) \in \partial^A$ and $|p_l - p| \leq \delta_1$, then $|u(p_l) - u(p)| \leq \frac{\varepsilon}{10}$.

Now we identify the unit interval $[0, 1]$ with half of $\partial^C = [-1, 1] \times \{0\} \subset \mathbb{R}^2$ and abuse notation so that $x \in [0, 1]$ represents a point on ∂^C . Consider the decomposition

$$[0, 1] = \bigcup_{k=0}^{\infty} [2^{-k-1}, 2^{-k}] = \bigcup_{k=0}^{\infty} I_k.$$

Note that we have

$$\sum_{k=0}^{\infty} E(u|_{(D_{2^{-k-1}} \setminus D_{2^{-k}}) \cap D^+}) \leq \varepsilon_0.$$

Then there is some $K_1 > 0$ such that for all $k \geq K_1$, we have

$$E(u|_{(D_{2^{-k-1}} \setminus D_{2^{-k}}) \cap D^+}) \leq \frac{\varepsilon^2}{1000N^2}.$$

Now using Lemma 3.1, for any $k \geq K_1$ we can pick $x_k \in I_k$ such that

$$|u(x_k) - u(p)| \leq |u(x_k) - u(p_{|x_k|})| + |u(p_{|x_k|}) - u(p)| \leq \frac{\varepsilon}{8}.$$

Now let us consider a family of half-disks $D_k^+ \subset D^+$, where D_k^+ has corner points $(2^{-k-2}, 0)$ and $(2^{-k} + 2^{-k-2}, 0)$. Then the center of D_k^+ is $y_k := (2^{-k-1} + 2^{-k-2}, 0)$ and the radius of D_k^+ is $r_k := 2^{-k-1}$. Each D_k^+ can only overlap with at most four of the other D_k^+ , so we have

$$\sum_{k=0}^{\infty} E(u|_{D_k^+}) \leq 4\varepsilon_0.$$

Thus, for any $\varepsilon_0 > 0$, there exists a $K_2 > 0$ such that for any $k \geq K_2$ we have $E(u|_{D_k^+}) \leq \varepsilon_0$. Then by rescaling and Theorem 2.5, for any $y \in [2^{-k-1}, 2^{-k}]$ with $k \geq K_2$ we have

$$(36) \quad |\nabla u(y)| \leq C 2^{k+1} \sqrt{\varepsilon_0}.$$

Then integration along the interval between y and x_k gives

$$|u(y) - u(x_k)| \leq 2C \sqrt{\varepsilon_0}.$$

If we choose ε_0 small enough (depending only on ε), then for any $y \in [2^{-k-1}, 2^{-k}]$ with $k \geq K_2$ we have $|u(y) - u(x_k)| \leq \frac{\varepsilon}{4}$. So for all $y \in (0, 2^{-K})$, where $K = \max\{K_1, K_2\}$, we have

$$|u(y) - u(p)| \leq \varepsilon.$$

We conclude the continuity of u at p and hence finish the proof of the theorem. $\square$

As a corollary of Theorem 3.2, we have:

Corollary 3.3 *A weakly harmonic map u with mixed Dirichlet and free boundaries on D^+ is continuous on $\overline{D^+}$ provided that u is continuous on ∂^A .*

3.2 Existence of (partial) free boundary harmonic replacement

Theorem 3.4 *There exists $\varepsilon_0 > 0$ such that for any $v \in C^0(\overline{D^+}, \mathcal{N}) \cap W^{1,2}(D^+, \mathcal{N})$ with $v(\partial^C) \subset \Gamma$ and $E(v) \leq \varepsilon_0$, there exists a unique harmonic map*

$$u \in C^0(\overline{D^+}, \mathcal{N}) \cap W^{1,2}(D^+, \mathcal{N})$$

such that $E(u) \leq \varepsilon_0$, $u = v$ on ∂^A , $u(\partial^C) \subset \Gamma$ and u meets Γ orthogonally along ∂^C .

Remark 3.5 The map u in Theorem 3.4 is usually called the (partial) free boundary harmonic replacement of v .

Proof of Theorem 3.4 Combining Hélein's interior regularity for weakly harmonic maps on two-dimensional domains, Qing's (Dirichlet) boundary regularity [39], the free boundary regularity of Scheven [46] and Theorem 3.2, we know that the (partial) free boundary harmonic replacement u of v , if it exists, is smooth in D^+ and continuous in $\overline{D^+}$.

The (partial) free boundary harmonic replacement can be constructed as follows: suppose $v \in C^0(\overline{D^+}, \mathcal{N}) \cap W^{1,2}(D^+, \mathcal{N})$ is such that $v(\partial^C) \subset \Gamma$ and $E(v) \leq \varepsilon_0$. Let us define the space $\mathcal{F}$ to be the space of maps

$$\mathcal{F} := \{w \in W^{1,2}(D^+, \mathcal{N}) : w|_{\partial^A} = v|_{\partial^A}, w(\partial^C) \subset \Gamma\}.$$

Now we choose an energy-minimizing sequence $u^i \in \mathcal{F}$, ie

$$\lim_{i \rightarrow \infty} E(u^i) = \inf\{E(w) : w \in \mathcal{F}\}.$$

By the Rellich compactness theorem, we can find a subsequence of $\{u^i\}$ that weakly converges to a $W^{1,2}$ map u . By first variation of the energy functional at u we know that u is a weakly harmonic map with mixed Dirichlet and free boundaries on D^+ . Then indeed we know that $\{u^i\}$ converges strongly in $W^{1,2}$ to u by the energy convexity Theorem 2.2 and Poincaré inequality Lemma A.2. $\square$

4 Overview of the variational approach

In this section, we provide an overview of the proof of Theorem 0.1.

As in the proof of the classical Plateau problem, the area functional is too weak to control the maps, so we have to change gear to the energy functional. The next result says that one can take an approximating sequence of sweepouts in a given homotopy class so that their maximal energy converges to the maximal area. In particular, we have:

Theorem 4.1 *Given $\beta \in \Omega$ with $W = W(\Omega_\beta) > 0$, there exists a sequence of sweepouts $\gamma^j \in \Omega_\beta$ with*

$$\max_{s \in [0,1]} E(\gamma^j(\cdot, s)) \searrow W.$$

Remark The proof is given in Section 5. Note that we have

$$W \leq \max_{s \in [0,1]} \text{Area}(\gamma^j(\cdot, s)) \leq \max_{s \in [0,1]} E(\gamma^j(\cdot, s)) \searrow W.$$

The next ingredient is a tightening theorem. We first fix some notation concerning balls in D .

Definition 4.2 A *generalized ball* B in the unit disk D is either an interior ball of D or the intersection with D of a ball of $\mathbb{R}^2$ centered at some boundary point of ∂D . That is: $B \subset D$ or $B = B_r(x) \cap D$ where $x \in \partial D$ and $r < \frac{1}{2}$. We will call the first case a *classical ball* and the second case a *boundary ball*.

Given $\rho > 0$, when B is a classical ball we let $\rho B \subset D$ denote the ball with the same center as B and radius ρ times that of B ; when B is a boundary ball, we can define $\rho B \subset D$ as follows: there exists a unique fractional linear transformation Π_B from D to the upper half-plane $\mathbb{H}^2$ such that $\Pi_B(B) = D^+$, and ρB is defined as $\Pi_B^{-1}(D_\rho^+)$.

Remark In the following, we will frequently identify a boundary ball B with its image $\Pi_B(B)$ in $\mathbb{H}^2$.

The following result plays the role of deformation lemma in nonlinear analysis; see eg [54].

Theorem 4.3 Given $\beta \in \Omega$ with $W = W(\Omega_\beta) > 0$, there exist a constant $\epsilon_1 > 0$ and a continuous function $\Psi: [0, \infty) \rightarrow [0, \infty)$ with $\Psi(0) = 0$, both depending on $(\mathcal{N}, \Gamma)$, such that given any $\tilde{\gamma} \in \Omega$ with no nonconstant harmonic slices other than $\tilde{\gamma}(\cdot, 0)$ and $\tilde{\gamma}(\cdot, 1)$, there exists $\gamma \in \Omega_{\tilde{\gamma}}$ such that $E(\gamma(\cdot, t)) \leq E(\tilde{\gamma}(\cdot, t))$ for each t , and for each t with $E(\tilde{\gamma}(\cdot, t)) \geq \frac{1}{2}W$ we have the following property:

(*) If $\mathcal{B}$ is any finite collection of disjoint generalized closed balls in D with

$$\int_{\bigcup_B \mathcal{B}} |\nabla \gamma(\cdot, t)|^2 < \epsilon_1,$$

and if $v: \bigcup_B \frac{1}{8}B \rightarrow \mathcal{N}$ is the free boundary harmonic replacement of $\gamma(\cdot, t)$ on $\bigcup_B \frac{1}{8}B$, then

$$\int_{\bigcup_B \frac{1}{8}B} |\nabla \gamma(\cdot, t) - \nabla v|^2 \leq \Phi(E(\tilde{\gamma}(\cdot, t)) - E(\gamma(\cdot, t))).$$

Remark The proof is given in Section 6.

We also need the following compactness result. Let ϵ_{SU} and ϵ_F be the small thresholds (depending only on $\mathcal{N}$ and $(\mathcal{N}, \Gamma)$, respectively) from [39, Lemma 4] (cf [45, Main Estimate 3.2]) and Theorem 2.5 (cf [15, Proposition 1.7]), respectively, giving us interior estimates for harmonic maps or free boundary harmonic maps with energy less than ϵ_{SU} and ϵ_F , respectively.

Theorem 4.4 Let $\epsilon_1, E_0 > 0$ be such that $\epsilon_1 < \min\{\epsilon_{SU}, \epsilon_F\}$, and suppose that $u^j: (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ is a sequence of maps in $C^0(\bar{D}, \mathcal{N}) \cap W^{1,2}(D, \mathcal{N})$ with $E_0 \geq E(u^j)$ which satisfies

$$(37) \quad \text{Area}(u^j) > E(u^j) - \frac{1}{j}$$

as well as the following condition:

(†) For any finite collection $\mathcal{B}$ of disjoint generalized closed balls in D with

$$\int_{\bigcup_{\mathcal{B}} B} |\nabla u^j|^2 < \epsilon_1,$$

if $v: \bigcup_{\mathcal{B}} \frac{1}{8}B \rightarrow \mathcal{N}$ is the harmonic replacement of u^j on $\bigcup_{\mathcal{B}} \frac{1}{8}B$, then

$$\int_{\bigcup_{\mathcal{B}} \frac{1}{8}B} |\nabla u^j - \nabla v|^2 \leq \frac{1}{j}.$$

Then a subsequence of the varifolds for the maps u^j converges to a collection of harmonic maps $v_0, \dots, v_m: (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ with free boundary, and harmonic spheres $\tilde{v}_1, \dots, \tilde{v}_l: S^2 \rightarrow \mathcal{N}$. Moreover, the energy identity holds:

$$\sum_{i=0}^m E(v_i) + \sum_{k=1}^l E(\tilde{v}_k) = \lim_{j \rightarrow \infty} E(u^j).$$

Remark The proof is given in [Section 7](#). Note that $E(v_i) = \text{Area}(v_i)$ and $E(\tilde{v}_j) = \text{Area}(\tilde{v}_j)$.

Now we prove the main [Theorem 0.1](#) using [Theorems 4.1, 4.3](#) and [4.4](#).

Proof of Theorem 0.1 Choose a sequence $\tilde{\gamma}^j \in \Omega_\beta$ as in [Theorem 4.1](#) and assume that

$$\max_{t \in [0,1]} E(\tilde{\gamma}^j(\cdot, t)) < W + \frac{1}{j}.$$

We can slightly change the parametrization so that $\tilde{\gamma}^j$ maps a small open subset of D to a point, so that each slice cannot be harmonic unless it is a constant map; we refer to [\[6, footnote 8\]](#) for details. Applying [Theorem 4.3](#) to $\tilde{\gamma}^j$ gives a sequence $\gamma^j \in \Omega_\beta$. We will show that $\{\gamma^j\}$ has the desired properties.

Let us argue by contradiction. Let $\mathcal{G}^W$ be the collection of harmonic maps from $S^2 \rightarrow \mathcal{N}$ and free boundary harmonic maps $(D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ such that the sum of the energies is exactly W . Suppose $\{\gamma^j\}$ does not have the desired property, which

means that there exists some $\varepsilon > 0$ such that there exist $j_k \rightarrow \infty$ and $s_k \in [0, 1]$ with $d_V(\gamma^{j_k}(\cdot, s_k), \mathcal{G}^W) \geq \varepsilon$ and $\text{Area}(\gamma^{j_k}(\cdot, s_k)) > W - \frac{1}{k}$. Then by $E(u) \geq \text{Area}(u)$ we get $\lim_{k \rightarrow \infty} E(\gamma^{j_k}(\cdot, s_k)) = W$, and

$$E(\tilde{\gamma}^{j_k}(\cdot, s_k)) - E(\gamma^{j_k}(\cdot, s_k)) \leq E(\tilde{\gamma}^{j_k}(\cdot, s_k)) - \text{Area}(\gamma^{j_k}(\cdot, s_k)) \leq \frac{1}{k} + \frac{1}{j_k} \rightarrow 0.$$

Since the tightening process decreases the energy, we get

$$E(\gamma^{j_k}(\cdot, s_k)) - \text{Area}(\gamma^{j_k}(\cdot, s_k)) \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

By [Theorem 4.3](#) we have that if $\mathcal{B}$ is any finite collection of disjoint generalized closed balls in D with

$$\int_{\bigcup_{\mathcal{B}} B} |\nabla \gamma^{j_k}(\cdot, s_k)|^2 < \varepsilon_1,$$

and if $v: \bigcup_{\mathcal{B}} \frac{1}{8} B \rightarrow \mathcal{N}$ is the harmonic replacement of $\gamma^{j_k}(\cdot, s_k)$ on $\bigcup_{\mathcal{B}} \frac{1}{8} B$, then

$$\int_{\bigcup_{\mathcal{B}} \frac{1}{8} B} |\nabla \gamma^{j_k}(\cdot, s_k) - \nabla v|^2 \leq \Phi\left(\frac{1}{k} + \frac{1}{j_k}\right) \rightarrow 0.$$

Therefore applying [Theorem 4.4](#) gives a subsequence of the $\gamma^{j_k}(\cdot, s_k)$ that varifold-converges to a collection of desired harmonic disks with free boundary and harmonic spheres. The energy identity implies that the sum of the energies of the limit is exactly W . However this contradicts $d_V(\gamma^{j_k}(\cdot, s_k), \mathcal{G}^W) \geq \varepsilon$. This finishes the proof of [Theorem 0.1](#). $\square$

5 Conformal parametrization

We now prove [Theorem 4.1](#). The main idea of the proof follows [\[6, Appendix D\]](#); see also [\[58, Section 3\]](#) and [\[60, Section 3\]](#). For a given sweep-out, we will find conformal reparametrizations of the regularization of this sweep-out, so that the energy of each slice of this family can not be much larger than its area.

Proof of [Theorem 4.1](#) First we claim that for a given sweep-out $\tilde{\gamma}(\cdot, t) \in \Omega_{\tilde{\gamma}}$, we can find a regularized sweep-out $\gamma(\cdot, t) \in \Omega_{\tilde{\gamma}}$ which lies in $C^0([0, 1], C^2(D, \mathcal{N}))$ as a map of t and is such that $\gamma(\cdot, t)$ is close to $\tilde{\gamma}(\cdot, t)$ uniformly in the $W^{1,2} \cap C^0$ -norm for all $t \in [0, 1]$. This follows from a standard argument using mollification just like [\[6, Lemma D.1\]](#). Here we only point out the necessary modifications of [\[6, Lemma D.1\]](#) to make sure that the images of ∂D under each slice $\gamma(\cdot, t)$ lie in the constraint submanifold Γ . In particular, near the boundary ∂D we can first enlarge the domain D

to $D_{1+\alpha}$ (for small $\alpha > 0$) by reflecting $\tilde{\gamma}(\cdot, t)$ across Γ in the Fermi coordinates around Γ (cf [15]) and then do classical mollification under the Fermi coordinates. The mollified maps when restricted to D will map ∂D to Γ by our construction. By the classical mollification result [13, page 250, Theorem 1] and Lemma A.1, we can make sure that the obtained maps are close to $\tilde{\gamma}(\cdot, t)$ uniformly in $W^{1,2} \cap C^0$ near ∂D . In the interior of D , we can just mollify $\tilde{\gamma}(\cdot, t)$ in $\mathbb{R}^N$. To combine them, we can choose a partition of unity to glue these two mollifications together, and by the same argument as [13, page 252, Theorem 3] the obtained maps are also close to $\tilde{\gamma}(\cdot, t)$ uniformly in $W^{1,2} \cap C^0(D)$. Up to this step, under the obtained maps the boundary ∂D goes into Γ , but the interior of D may get out of $\mathcal{N}$. Finally, one can follow [6, Lemma D.1] to project these maps to $\mathcal{N}$ using the nearest-point projection to get the desired $\gamma(\cdot, t)$. By choosing the mollification parameter small enough, we can make sure $\max_{t \in [0,1]} \|\gamma(\cdot, t) - \tilde{\gamma}(\cdot, t)\|$ is as small as we want. Note that an explicit homotopy between γ and $\tilde{\gamma}$ is given by letting the mollification parameter go to 0. So we finish the sketch of the proof of the claim.

Then $\gamma(\cdot, t)$ induces a continuous one-parameter family of C^1 metrics

$$g(t) = \gamma(t)^*(\text{metric on } \mathcal{N})$$

on D . This family of metrics may be degenerate, so we define the perturbed metrics as $\tilde{g}(t) = g(t) + \varepsilon g_0$, where g_0 is the standard flat metric on D . Then by the same methods as in [6, Lemma D.6], [58, Proposition 3.1] and [60, Proposition 3.4], we can construct a family of conformal reparametrizations $h_t: D_{g_0} \rightarrow D_{\tilde{g}(t)}$ (which fix three given points on ∂D) that varies continuously in $C^0 \cap W^{1,2}(D, D)$. Then with the conformality we can control the energy:

$$\begin{aligned} (38) \quad E(\gamma(\cdot, t) \circ h_t) &= E(h_t: D_{g_0} \rightarrow D_{g(t)}) \\ &\leq E(h_t: D_{g_0} \rightarrow D_{\tilde{g}(t)}) = \text{Area}(D_{\tilde{g}(t)}) \\ &= \int_D (\det(g_0^{-1}g(t)) + \varepsilon \text{Tr}(g_0^{-1}g(t)) + \varepsilon^2)^{\frac{1}{2}} d\text{Vol}_{g_0} \\ &\leq \text{Area}(D_{g(t)}) + \pi(\varepsilon^2 + 2\varepsilon \sup_t |g_0^{-1}g(t)|)^{\frac{1}{2}}. \end{aligned}$$

Choose $\varepsilon > 0$ so that $\pi(\varepsilon^2 + 2\varepsilon \sup_t |g_0^{-1}g(t)|)^{\frac{1}{2}} < \frac{\delta}{2}$, and we get

$$E(\gamma(\cdot, t) \circ h_t) - \text{Area}(D_{g_t}) \leq \frac{\delta}{2}.$$

If $\tilde{\gamma}^j$ is a sequence of sweepouts in Ω_β , then $\gamma^j \circ h_t$ constructed as above is a sequence as desired. $\square$

6 Construction of the tightening process

This section is devoted to the proof of [Theorem 4.3](#).

6.1 Continuity of harmonic replacement

In this part we want to prove that the free boundary harmonic replacement process is actually continuous as a map from $C^0(\overline{D^+}) \cap W^{1,2}(D^+)$ to itself, if we restrict to maps with small energy. This generalizes the continuity of the harmonic replacement process on classical interior balls by Colding–Minicozzi [\[6\]](#).

Theorem 6.1 *Let ε_0 be as in [Theorem 2.2](#) and set*

$$\mathcal{M} = \{u \in C^0(\overline{D^+}, \mathcal{N}) \cap W^{1,2}(D^+, \mathcal{N}) : E(u) \leq \varepsilon_0, u(\partial^C) \subset \Gamma\}.$$

Given $u \in \mathcal{M}$, let $w \in \mathcal{M}$ be the unique free boundary harmonic replacement of u (produced by [Theorem 3.4](#)). Then $u \rightarrow w$ is continuous as a map from $C^0(\overline{D^+}) \cap W^{1,2}(D^+)$ to itself.

6.1.1 $W^{1,2}$ –continuity First we prove that this map is $W^{1,2}$ continuous. In [\[6\]](#), Colding–Minicozzi’s idea was to construct a comparison map by interpolating in $\mathbb{R}^N$ between two maps of the same boundary value in $W^{1,2}(D, \mathcal{N})$, and then projecting the interpolation back to $\mathcal{N}$. In our free boundary setting, however, such an interpolation–projection trick may not leave the image of ∂^C lying inside Γ . Here we prove the $W^{1,2}$ –continuity by contradiction. The main idea is that if the $W^{1,2}$ –continuity fails, then we can find a sequence u_i converging to u_∞ , but the replacements w_i of the u_i have energy strictly greater than the energy of w_∞ . In this scenario, we are able to construct some comparison maps v_i sharing the same fixed boundary value with w_i , but having energy strictly smaller than that of w_i for i large. This contradicts the uniqueness of free boundary harmonic maps with small energy; see [Corollary 2.3](#).

The key ingredient in our proof is the construction of certain new comparison maps. We first collect a few preliminary results for the construction.

There are five components in the comparison map. One of the components consists of “small” free boundary harmonic maps. The following lemma shows that the energy of these maps is actually small.

Lemma 6.2 *There exists a $\delta_0 > 0$ such that if a Lipschitz map $f: [0, \pi] \rightarrow \mathcal{N}$ satisfies $f(0) \in \Gamma$, $f(\pi) \in \Gamma$ and $\int_0^\pi |f'|^2 \leq \delta \leq \delta_0$, then there exists a map $v: D^+ \rightarrow \mathcal{N}$ such that $v(1, \theta) = f(\theta)$, $v(\partial^C) \subset \Gamma$ and $E(v) \leq C\delta$ for some universal constant $C > 0$ depending only on $\mathcal{N}$, Γ and δ_0 .*

Proof By the gradient bound of f and the fact that $f(0), f(\pi) \in \Gamma$, we can assume that the image of f lies in the Fermi coordinates $\{y^1, \dots, y^n\}$ around Γ , where [Lemma A.1](#) in the [appendix](#) applies. We can further assume that locally Γ is identified as a subset of $\{y^{k+1} = \dots = y^n = 0\}$ and $f(0) = 0$. Define

$$v(r, \theta) = r \cdot f(\theta),$$

where $r \cdot f(\theta)$ is the scalar multiplication under Fermi coordinates; hence v is a map from $D^+ \rightarrow \mathcal{N}$, and $v(r, 0)$ and $v(r, \pi)$ both lie in Γ for any $r \in [0, 1]$, thus $v(\partial^C) \subset \Gamma$.

It is only left to check that the energy of v is small. By [Lemma A.1](#), we only need to check that the energy of v is small as a map to the Euclidean space $\mathbb{R}^N$. In particular, [Lemma A.1](#) implies that $\int_0^\pi |f'|^2 \leq (1 + \alpha)\delta$ under the Euclidean metric for some small $\alpha > 0$. Also, as $f(0) = 0$, we get $|f(\theta)| \leq \sqrt{\pi(1 + \alpha)\delta}$ for $\theta \in [0, \pi]$. Now we have the energy estimates for v as a map into standard Euclidean space $\mathbb{R}^N$

$$\begin{aligned} (39) \quad E(v) &= \int_0^1 \int_0^\pi \left(\left| \frac{\partial v}{\partial r} \right|^2 + \frac{1}{r^2} \left| \frac{\partial v}{\partial \theta} \right|^2 \right) r \, d\theta \, dr \\ &\leq \int_0^1 \int_0^\pi (|f|^2 + |f'|^2) r \, d\theta \, dr \leq C\delta \end{aligned}$$

for some universal constant $C > 0$. This completes the proof. $\square$

Another component in our comparison map is a modified interpolation band. First we recall a lemma in [\[6\]](#) to construct the interpolation of band between two circles.

Lemma 6.3 *There exists $\delta_0 > 0$ such that for $\delta < \delta_0$ the following statement holds. Let $f, g: [0, \pi] \rightarrow \mathcal{N}$ be two C^0 maps such that*

$$|f - g| \leq \delta, \quad \int |f'|^2 \leq \delta' \quad \text{and} \quad \int |g'|^2 \leq \delta'$$

for some $\delta' > 0$. Then there exists a $\rho \in (0, \frac{1}{2}]$ and a map $\tilde{v}: D^+ \setminus D_{1-\rho}^+ \rightarrow \mathcal{N}$ such that

$$\tilde{v}(1 - \rho, \theta) = f(\theta), \quad \tilde{v}(1, \theta) = g(\theta) \quad \text{and} \quad E(\tilde{v}) \leq C\delta^{\frac{1}{2}}\delta'^{\frac{1}{2}}$$

for some constant $C > 0$ depending only on $\mathcal{N}$.

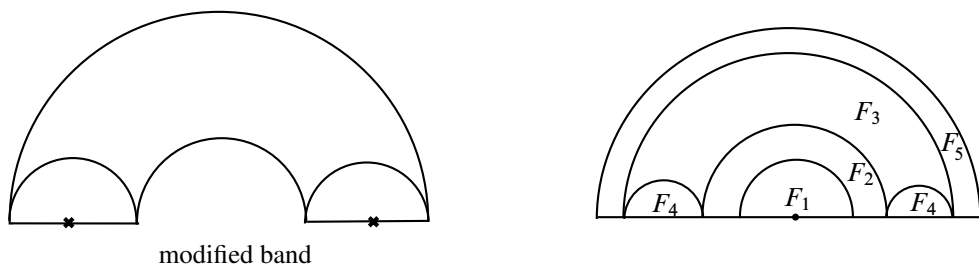


Figure 1

Proof We refer the proof to [6, Lemma 3.11], where the construction works for the whole circles, but it can be generalized to half-circles without any modification. Note that our assumption here is even stronger. $\square$

With this construction on hand, we can construct an interpolation between two arcs on a modified band. Let us first define a modified band. A *modified band* $MB_{a,b}$ is defined as

$$(40) \quad MB_{a,b} = \{(r, \theta) : r \in [a, b], \theta \in [0, \pi]\} \setminus \bigcup_{\xi=0, \pi} \{(r, \theta) : |(r, \theta) - (\frac{1}{2}(b+a), \xi)| < \frac{1}{2}(b-a)\}.$$

See Figure 1. In the following context we may use ϕ to denote the angle parameter of the arcs of the removed disks.

Lemma 6.4 *There exists $\delta_0 > 0$ such that for $\delta < \delta_0$ the following statement holds. Let $f, g: [0, \pi] \rightarrow \mathcal{N}$ be two C^0 maps such that $\|f - g\|_{L^\infty} \leq \delta$, $\int |f'|^2 \leq \delta'$ and $\int |g'|^2 \leq \delta'$ for some $\delta' > 0$. Then there exists $\rho \in (0, \frac{1}{2}]$ and a map $v: MB_{1-\rho, 1} \rightarrow \mathcal{N}$ such that $v(1-\rho, \theta) = f(\theta)$, $v(1, \theta) = g(\theta)$ and $E(v) \leq C\delta^{\frac{1}{2}}\delta'^{\frac{1}{2}}$ for some constant C depending only on $\mathcal{N}$ and ρ . Moreover, for the arc of the removed half-disks of the modified band (reparametrized by angle parameter ϕ), we have*

$$\int_0^\pi |v'(\phi)|^2 ds(\phi) \leq C\delta^2,$$

where $ds(\phi)$ is the intrinsic arc-length integral.

Proof We construct a change of variables. Construct a map $h: D^+ \setminus D_{1-\rho}^+ \rightarrow MB_{1-\rho, 1}$ as follows: given $s \in [1-\rho, 1]$, we consider the arc $MB_{1-\rho, 1} \cap \{r \equiv s\}$ and denote the length of this arc by l_s ; for any fixed $(r, \theta) \in D^+ \setminus D_{1-\rho}^+$, let h map it to (r, θ') ,

where the ratios satisfy

$$[|\theta - \frac{\pi}{2}| : \frac{\pi}{2}] = [|\theta' - \frac{\pi}{2}| : l_r].$$

Then the Jacobian and the gradient of this change of variables is bounded by some universal constant. Let $\tilde{v}$ be given by Lemma 6.3; then $v = \tilde{v} \circ (h^{-1})$ is the desired interpolation map. The only thing we need to check is the last estimate. In particular, the energy of v on the boundary of the removed half-disks is bounded by that of $\tilde{v}$ on $\theta = 0$ times a universal constant, and the energy of $\tilde{v}$ therein is bounded. $\square$

Now we start proving that the harmonic replacement $u \rightarrow w$ is continuous as a map from $C^0 \cap W^{1,2}$ to $W^{1,2}$ for u with small energy.

Proof of $W^{1,2}$ -continuity We prove by contradiction. Suppose the $W^{1,2}$ -continuity of the harmonic replacement process fails. Then we can find a sequence of maps $\{u_i: D^+ \rightarrow \mathcal{N}\}$ converging to u_∞ in $C^0(\overline{D^+}, \mathcal{N}) \cap W^{1,2}(D^+, \mathcal{N})$, but whose free boundary harmonic replacements $\{w_i\}$ do not converge to the corresponding replacement w_∞ (of u_∞) in $W^{1,2}$. Note that a subsequence of w_i (still denoted by w_i) must converge weakly to w_∞ , so by lower-semicontinuity of energy, $E(w_\infty) \leq \liminf E(w_i)$.

The fact that w_i does not converge to w_∞ in $W^{1,2}$ implies that there exists $\varepsilon > 0$ and a subsequence (still denoted by w_i) such that $\|w_i - w_\infty\|_{W^{1,2}}^2 \geq 2\varepsilon$ holds, ie

$$\int \langle w_i - w_\infty, w_i - w_\infty \rangle + \langle \nabla(w_i - w_\infty), \nabla(w_i - w_\infty) \rangle \geq 2\varepsilon.$$

The weak convergence implies that $\|w_i - w_\infty\|_{L^2} \rightarrow 0$ and $\int \langle \nabla w_i, \nabla w_\infty \rangle \rightarrow E(w_\infty)$. Thus we get $E(w_i) - E(w_\infty) \geq \varepsilon$ for i large.

We divide D^+ into five different pieces, $D^+ = \bigcup_{j=1}^5 F_j$, where

- (i) $F_1 = \{(r, \theta) : r \in [0, \eta\lambda(1 - \rho)]\}$,
- (ii) $F_2 = \{(r, \theta) : r \in [\eta\lambda(1 - \rho), \lambda(1 - \rho)]\}$,
- (iii) $F_3 = MB_{\lambda(1-\rho), \lambda}$,
- (iv) $F_4 = \{y \in D^+ : |y - (\lambda - \frac{1}{2}\lambda\rho, 0)| \leq \frac{1}{2}\lambda\rho\} \cup \{y \in D^+ : |y - (\lambda - \frac{1}{2}\lambda\rho, \pi)| \leq \frac{1}{2}\lambda\rho\}$,
- (v) $F_5 = \{(r, \theta) : r \in [\lambda, 1]\}$.

See Figure 1. Here $\eta, \rho, \lambda \in (0, 1)$ will be determined later. Note that the F_j have some common boundaries, and we will see in the construction below that the maps on the common boundaries share the same value. Let us choose a radius $r_0 > \frac{1}{2}$ such that $\int_{D^+ \setminus D_{1-r_0}^+} |\nabla u_\infty|^2 \leq \frac{\varepsilon}{16}$. Then given i large enough that $\|u_i - u_\infty\|_{W^{1,2}}^2 \leq \frac{\varepsilon}{16}$, we have $\int_{D^+ \setminus D_{1-r_0}^+} |\nabla u_i|^2 \leq \frac{\varepsilon}{8}$. Using the coarea formula, we can pick some radius $r_1 \in (r_0, 1)$ such that $\int_{r=r_1} |\partial_\theta u_\infty|^2 \leq C(r_0)\varepsilon$ and $\int_{r=r_1} |\partial_\theta u_i|^2 \leq C(r_0)\varepsilon$.

Construction of v_i

- On F_1 , we define v_i to be w_∞ after rescaling to a suitable scale to fit F_1 .
- On F_2 , we define v_i to be the part of u_∞ defined on $D^+ \setminus D_{1-r_1}^+$ after inversion (in polar coordinates, we change r to $2 - r_1 - r$) and rescaling. Note that the energy of v_i on F_2 is bounded by the energy of u_∞ on $D^+ \setminus D_{1-r_1}^+$ times a universal constant 4.
- On F_3 , we define v_i to be the interpolation between $u_\infty|_{\{r=r_1\}}$ and $u_i|_{\{r=r_1\}}$ from Lemma 6.4 on the modified band, then rescale to fit F_3 .
- On F_4 , we define v_i to be the map constructed in Lemma 6.2 (up to rescaling), which provides two maps with small energy.
- On F_5 , we define v_i to be the part of u_i defined on $D^+ \setminus D_{1-r_1}^+$ (up to rescaling).

Here ρ comes from Lemma 6.4, $\eta = 1 - r_1$ and $\lambda = 1 - r_1$. These parameters are chosen in order to guarantee that in the definition of the v_i , the rescalings are possible; ie the ratios between inner radius and outer radius of the bands or modified bands do not change.

Now we claim some properties of the v_i that we just constructed. First, on ∂^A the fixed boundary, $v_i = u_i$; on ∂^C the free boundary, v_i always has image in Γ . Second, v_i is continuous. So it is an eligible comparison map (with w_∞).

Finally we estimate the energy of v_i . We will repeatedly use (without mention) the fact that the energy is invariant under conformal reparametrization on the domain. Note that the energy of v_i on F_1 equals $E(w_\infty)$; on F_2 and F_5 the total energy of v_i is bounded by $\varepsilon/3$; on F_3 by Lemma 6.4, the energy of v_i is bounded by a constant times $\|u_\infty - u_i\|_{C^0}^{1/2}$, and on F_4 by Lemma 6.2 the energy of v_i is bounded by a constant times $\|u_\infty - u_i\|_{C^0}^{1/2}$. Combining all the pieces gives

$$E(v_i) \leq E(w_\infty) + \frac{2\varepsilon}{3} + C\|u_\infty - u_i\|_{C^0}^{1/2}.$$

Since u_i converges to u_∞ in C^0 , when i is large enough $C\|u_\infty - u_i\|_{C^0}^{1/2} \leq \frac{\varepsilon}{6}$. In conclusion, when i is large enough, $E(v_i) \leq E(w_\infty) + \frac{5\varepsilon}{6} \leq E(w_i) - \frac{\varepsilon}{6} < E(w_i)$.

Note that v_i shares the same fixed boundary data with w_i , and thus we have constructed a map v_i with same fixed boundary data as w_i along ∂^A and $v_i(\partial^C) \subset \Gamma$, but v_i has energy strictly less than that of w_i on D^+ ; this is a contradiction to the energy-minimizing property of the free boundary harmonic replacement ([Theorem 2.2](#)). Therefore the harmonic replacement is continuous in $W^{1,2}$. $\square$

6.1.2 C^0 -continuity

Proof of C^0 -continuity Now we are ready to prove the C^0 -continuity of the harmonic replacement process. We argue by contradiction. If the harmonic replacement process is not C^0 , we can find a sequence of maps $u_i: D^+ \rightarrow \mathcal{N}$ converging to u_∞ in $C^0(\overline{D^+}) \cap W^{1,2}(D^+)$ while for the corresponding free boundary harmonic replacements w_i (of u_i) we have $\|w_i - w_\infty\|_{C^0(\overline{D^+})} \geq \varepsilon > 0$.

From the proof of [Theorem 3.2](#), we know that the w_i are equicontinuous on $\overline{D^+}$. Thus by the Arzelà–Ascoli theorem, up to a subsequence w_i must converge to some w'_∞ in $C^0(\overline{D^+})$. Now using the $W^{1,2}$ -continuity of the harmonic replacement process that we have just proved, we have $\|w_i - w_\infty\|_{W^{1,2}(D^+)} \rightarrow 0$ as $i \rightarrow \infty$, which implies $w_\infty = w'_\infty$. This contradicts $\|w_i - w_\infty\|_{C^0(\overline{D^+})} \geq \varepsilon > 0$. $\square$

6.2 Uniform continuity of energy improvement

In this part, we prove two inequalities regarding energy improvements for two sets of free boundary harmonic replacements on generalized balls; see [Definition 4.2](#). These inequalities will play a key role in the proof of [Theorem 4.3](#). Similar results were proved by Colding–Minicozzi [[6](#), Section 3] for fixed boundary harmonic replacements. We suggest to first-time readers to come back to this section after reading [Section 6.3](#).

Let ε_0 be the minimum of the ε_0 in [Theorems 2.1](#) and [2.2](#). We adopt the following notation: given a map $u \in C^0(\overline{D}, \mathcal{N}) \cap W^{1,2}(D, \mathcal{N})$ with $u(\partial D) \subset \Gamma$ and a finite collection $\mathcal{B}$ of disjoint *generalized* closed balls in D such that the energy of u on $\bigcup_{B \in \mathcal{B}} B := \bigcup_{B \in \mathcal{B}} B$ is at most $\frac{1}{3}\varepsilon_0$, let $H(u, \mathcal{B}): D \rightarrow \mathcal{N}$ denote the map that coincides with u on $D \setminus \bigcup_{B \in \mathcal{B}} B$ and is equal to the harmonic replacements of u on $\bigcup_{B \in \mathcal{B}} B$; cf [Theorem 2.1](#) for the harmonic replacement with fixed boundary used by Colding–Minicozzi, and see [Remark 3.5](#) for the free boundary harmonic replacement, where the

existence is guaranteed by [Theorem 3.4](#). Note we also have $H(u, \mathcal{B})(\partial D) \subset \Gamma$, and without loss of generality we will call $H(u, \mathcal{B})$ the *free boundary harmonic replacement of u on $\mathcal{B}$* . Given two such disjoint collections $\mathcal{B}_1$ and $\mathcal{B}_2$, we use $H(u, \mathcal{B}_1, \mathcal{B}_2)$ to denote $H(H(u, \mathcal{B}_1), \mathcal{B}_2)$. Recall that for $\alpha \in (0, 1]$, we denote by $\alpha\mathcal{B}$ the collection of concentric balls with radii that are shrunk by the factor α in the sense of [Definition 4.2](#).

First we have the following interpolation formula for free boundary replacements; cf [\[6, Lemma 3.11\]](#). Let κ denote the radius for which the Fermi coordinates system exists in a tubular neighborhood of radius κ surrounding Γ where [Lemma A.1](#) in the [appendix](#) applies.

Lemma 6.5 *There exists $\tau > 0$ such that given two $C^0 \cap W^{1,2}$ maps $f, g: \partial_R^A \rightarrow \mathcal{N}$ with $f(0), f(\pi), g(0), g(\pi) \in \Gamma$, if f and g agree at one point on ∂_R^A and satisfy*

$$R \int_{\partial_R^A} |f' - g'|^2 \leq \tau^2 \quad \text{and} \quad |f(\theta) - f(0)| \leq \frac{\kappa}{3} \quad \text{for all } 0 \leq \theta \leq \pi,$$

then there exist $\rho \in (0, R/2]$ and a $C^0 \cap W^{1,2}$ -map $w: D_R^+ \setminus D_{R-\rho}^+ \rightarrow \mathcal{N}$ such that

$$w(R - \rho, \theta) = f(\theta), \quad w(R, \theta) = g(\theta),$$

the image of $w|_{\partial_R^C \setminus \partial_{R-\rho}^C}$ lies in Γ , and the following estimate holds:

$$\int_{D_R^+ \setminus D_{R-\rho}^+} |\nabla w|^2 \leq C \left(R \int_{\partial_R^A} |f'|^2 + |g'|^2 \right)^{\frac{1}{2}} \left(R \int_{\partial_R^A} |f' - g'|^2 \right)^{\frac{1}{2}},$$

where $C > 0$ is a universal constant depending only on κ , τ , Γ and $\mathcal{N}$.

Remark 6.6 This lemma generalizes an interpolation formula by Colding–Minicozzi for maps defined on circles [\[6, Lemma 3.11\]](#) to maps defined on half-circles, and the constructed interpolating map has the chord boundary $\partial_R^C \setminus \partial_{R-\rho}^C$ lying on Γ . In [\[6, Lemma 3.11\]](#), Colding–Minicozzi first took the linear interpolation of f, g in $\mathbb{R}^N$, and then projected it back to the ambient manifold $\mathcal{N}$. However this method doesn't work in the free boundary setting because the projection $\mathbb{R}^N \rightarrow \mathcal{N}$ may not map the boundary ∂_R^C to the constraint submanifold Γ .

Here we use the Fermi coordinate system to construct the desired interpolation between f and g . In the proof we have to work in two different coordinate systems, and [Lemma A.1](#) will be used to show the equivalence.

Proof of Lemma 6.5 There is one point $0 \leq \theta \leq \pi$ such that $f(\theta) = g(\theta)$. Choose $\tau = \kappa/3\pi$. Then by the assumptions and integrating, we get $|f(\theta) - g(\theta)| \leq \frac{\kappa}{2}$ for all $\theta \in [0, \pi]$, so that $|g(\theta) - g(0)| \leq \kappa$ for all $\theta \in [0, \pi]$.

As a result, we can assume the images of f and g both lie in a convex neighborhood U_κ of $f(0)$, and in U we can pick the Fermi coordinate systems $\{y^1, \dots, y^n\}$ as in Lemma A.1, where Γ is a subset of $\{y^{k+1} = \dots = y^n = 0\}$. In Lemma A.1, g^1 , g^2 and g^3 denote, respectively, the metric of $\mathcal{N}$, the flat metric in $\{y^1, \dots, y^n\}$ and the flat metric of $\mathbb{R}^N$. We will use $|\cdot|$ to denote the norm under the metric g^3 .

Since the statement is scaling-invariant, it suffices to prove the case $R = 1$. For $\rho \in (0, \frac{1}{2}]$ to be determined, define $w: D^+ \setminus D_{1-\rho}^+ \rightarrow \mathbb{R}^N$ by

$$w(r, \theta) = f(\theta) + \left(\frac{r + \rho - 1}{\rho} \right) (g(\theta) - f(\theta)).$$

On $\partial^C \setminus \partial_{1-\rho}^C$, since $f(0), f(\pi), g(0), g(\pi) \in \Gamma$, we know that in the Fermi coordinate chart, $f(0), f(\pi), g(0)$ and $g(\pi)$ lie in the plane $\{y^{k+1} = \dots = y^n = 0\}$. As a result, the interpolation function w also has the image $w|_{\partial^C \setminus \partial_{1-\rho}^C}$ lying in the same plane, and thus lying in Γ .

The energy density in this coordinate system is

$$e(w) = \sum_{i,j=1}^n g_{ij}^1 \left(\frac{\partial w^i}{\partial x} \frac{\partial w^j}{\partial x} + \frac{\partial w^i}{\partial y} \frac{\partial w^j}{\partial y} \right) \leq (1 + \alpha) \|\nabla w\|_{g^2}^2.$$

Now we proceed to prove the estimate of $\|\nabla w\|_{g^2}^2$. First we need a Wirtinger-type inequality. Suppose $h(\cdot)$ is a function on ∂^A and $h(s) = 0$ for some $s \in [0, \pi]$. Then for any $t \in [0, \pi]$, we have

$$|h(t)| = |h(t) - h(s)| \leq \int_s^t |h'(x)| dx \leq \int_{\partial^A} |h'|.$$

Integrating the square of both sides for $t \in [0, \pi]$, we get

$$\int_{\partial^A} |h(t)|^2 \leq \pi \left(\int_{\partial^A} |h'| \right)^2 \leq \pi^2 \int_{\partial^A} |h'|^2.$$

Then we get, setting $h = f - g$,

$$\int_{\partial^A} |f - g|^2 \leq \pi^2 \int_{\partial^A} |f' - g'|^2.$$

Thus

$$\begin{aligned}
 & E(D^+ \setminus D_{1-\rho}^+) \\
 & \leq \int_{D^+ \setminus D_{1-\rho}^+} (1+\alpha) \|\nabla w\|_{g^2}^2 \\
 & \leq (1+\alpha) \int_{1-\rho}^1 \left(\frac{1}{\rho^2} \int_0^\pi \|f-g\|_{g^2}^2(\theta) d\theta + \frac{1}{r^2} \int_0^\pi (\|f'\|_{g^2}^2 + \|g'\|_{g^2}^2)(\theta) d\theta \right) r dr \\
 & \leq (1+\alpha)^2 \int_{1-\rho}^1 \left(\frac{1}{\rho^2} \int_0^\pi |f-g|^2(\theta) d\theta + \frac{1}{r^2} \int_0^\pi (|f'|^2 + |g'|^2)(\theta) d\theta \right) r dr \\
 & \leq (1+\alpha)^2 \left(\frac{16}{\rho} \int_0^\pi |f'-g'|^2(\theta) d\theta + 2\rho \int_0^\pi (|f'|^2 + |g'|^2)(\theta) d\theta \right) \\
 & \leq 18(1+\alpha)^2 \left(\int_{\partial A} |f'|^2 + |g'|^2 \right)^{\frac{1}{2}} \left(\int_{\partial A} |f'-g'|^2 \right)^{\frac{1}{2}},
 \end{aligned}$$

once we pick $\rho = (\int_{\partial A} |f'-g'|^2)^{\frac{1}{2}} / (\int_{\partial A} |f'|^2 + |g'|^2)^{\frac{1}{2}}$. □

Now we are ready to prove the energy improvement inequalities for free boundary harmonic replacements. Similar results were first obtained by Colding–Minicozzi for fixed boundary harmonic replacements [6, Lemma 3.8]. Again we let ε_0 be the minimum of the ε_0 in Theorems 2.1 and 2.2.

Lemma 6.7 *There is a constant $k > 0$ such that if $u: D \rightarrow \mathcal{N}$ is in $C^0 \cap W^{1,2}$ and $\mathcal{B}_1$ and $\mathcal{B}_2$ are two finite collections of disjoint closed generalized balls in D such that the energy of u on each $\bigcup_{B_i} B$ is at most $\frac{1}{3}\varepsilon_0$, then*

$$(41) \quad E(u) - E(H(u, \mathcal{B}_1, \mathcal{B}_2)) \geq k \left(E(u) - E\left(H\left(u, \frac{1}{2}\mathcal{B}_2\right)\right) \right)^2.$$

Furthermore, for any $\mu \in [\frac{1}{8}, \frac{1}{2}]$ we have

$$\begin{aligned}
 (42) \quad & \frac{(E(u) - E(H(u, \mathcal{B}_1)))^{\frac{1}{2}}}{k} + E(u) - E(H(u, 2\mu\mathcal{B}_2)) \\
 & \geq E(H(u, \mathcal{B}_1)) - E(H(u, \mathcal{B}_1, \mu\mathcal{B}_2)).
 \end{aligned}$$

Proof The proof is analogous to the proof of [6, Lemma 3.8], and the main difficulty arises from those boundary balls. We will include the details for completeness, and focus on how we use the new interpolation result, ie Lemma 6.5.

Let $\mathcal{B}_1 = \{B_\alpha^1\}$ and $\mathcal{B}_2 = \{B_j^2\}$. We need to clarify the second replacement. Observe that the total energy of u on the union of the balls $\mathcal{B}_1 \cup \mathcal{B}_2$ is at most $\frac{2}{3}\varepsilon_0$, and the free

boundary harmonic replacement on $\mathcal{B}_1$ does not change the map outside these balls and is energy-nonincreasing, then it follows that the total energy of $H(u, \mathcal{B}_1)$ on $\mathcal{B}_2$ is at most $\frac{2}{3}\varepsilon_0$.

We will divide $\mathcal{B}_2$ into two disjoint subsets $\mathcal{B}_{2,+}$ and $\mathcal{B}_{2,-}$, set

$$\mathcal{B}_{2,+} = \{B_j^2 \in \mathcal{B}_2 : \frac{1}{2}B_j^2 \subset B_\alpha^1 \text{ for some } B_\alpha^1 \in \mathcal{B}_1\} \quad \text{and} \quad \mathcal{B}_{2,-} = \mathcal{B}_2 \setminus \mathcal{B}_{2,+}.$$

Since the balls in $\mathcal{B}_2$ are disjoint, we have

$$E(u) - E(H(u, \frac{1}{2}\mathcal{B}_2)) = (E(u) - E(H(u, \frac{1}{2}\mathcal{B}_{2,+}))) + (E(u) - E(H(u, \frac{1}{2}\mathcal{B}_{2,-}))).$$

Now we have two cases.

Case 1 Suppose

$$E(u) - E(H(u, \frac{1}{2}\mathcal{B}_{2,+})) \geq E(u) - E(H(u, \frac{1}{2}\mathcal{B}_{2,-}));$$

then

$$E(u) - E(H(u, \frac{1}{2}\mathcal{B}_{2,+})) \geq \frac{1}{2}(E(u) - E(H(u, \frac{1}{2}\mathcal{B}_2))).$$

Since the balls in $\frac{1}{2}\mathcal{B}_{2,+}$ are contained in balls in $\mathcal{B}_1$ and harmonic replacements minimize energy, we get

$$E(H(u, \mathcal{B}_1, \mathcal{B}_2)) \leq E(H(u, \mathcal{B}_1)) \leq E(H(u, \frac{1}{2}\mathcal{B}_{2,+})),$$

so that

$$\begin{aligned} (43) \quad E(u) - E(H(u, \mathcal{B}_1, \mathcal{B}_2)) &\geq E(u) - E(H(u, \frac{1}{2}\mathcal{B}_{2,+})) \\ &\geq \frac{1}{2}(E(u) - E(H(u, \frac{1}{2}\mathcal{B}_2))) \\ &\geq k(E(u) - E(H(u, \frac{1}{2}\mathcal{B}_2)))^2 \end{aligned}$$

if $k \leq 1/(2\varepsilon_0)$.

Case 2 Suppose

$$E(u) - E(H(u, \frac{1}{2}\mathcal{B}_{2,+})) \leq E(u) - E(H(u, \frac{1}{2}\mathcal{B}_{2,-}));$$

then

$$E(u) - E(H(u, \frac{1}{2}\mathcal{B}_{2,-})) \geq \frac{1}{2}(E(u) - E(H(u, \frac{1}{2}\mathcal{B}_2))).$$

Let $\tau > 0$ be given by [Lemma 6.5](#). We can assume that

$$(44) \quad 9 \int_D |\nabla u - \nabla H(u, \mathcal{B}_1)|^2 \leq \tau^2.$$

Otherwise Theorems 2.1 and 2.2 give (41) with $k = \tau^2/(2\varepsilon_0^2)$. In fact, if

$$9 \int_D |\nabla u - \nabla H(u, \mathcal{B}_1)|^2 > \tau^2,$$

then applying Theorems 2.1 and 2.2 to each classical ball or boundary ball in $\mathcal{B}_1$, we get

$$E(u) - E(H(u, \mathcal{B}_1, \mathcal{B}_2)) \geq E(u) - E(H(u, \mathcal{B}_1)) \geq \frac{1}{2} \int_D |\nabla u - \nabla H(u, \mathcal{B}_1)|^2 > \frac{\tau^2}{18}.$$

Noting that $E(u) - E(H(u, \frac{1}{2}\mathcal{B}_2)) \leq \frac{1}{3}\varepsilon_0$, we get the desired estimate.

Assuming (44), we want to show the following estimate for each $B_j^2 \in \mathcal{B}_{2,-}$:

$$\begin{aligned} (45) \quad & \int_{B_j^2} |\nabla H(u, \mathcal{B}_1)|^2 - \int_{B_j^2} |\nabla H(u, \mathcal{B}_1, B_j^2)|^2 \\ & \geq \int_{\frac{1}{2}B_j^2} |\nabla u|^2 - \int_{\frac{1}{2}B_j^2} |\nabla H(u, \frac{1}{2}B_j^2)|^2 \\ & \quad - C \left(\int_{B_j^2} |\nabla u|^2 + |\nabla H(u, \mathcal{B}_1)|^2 \right)^{\frac{1}{2}} \left(\int_{B_j^2} |\nabla u - \nabla H(u, \mathcal{B}_1)|^2 \right)^{\frac{1}{2}}. \end{aligned}$$

If (45) is true, summing it over balls in $\mathcal{B}_{2,-}$ and using the Cauchy–Schwartz inequality for discrete sums $|\sum a_j b_j| \leq (\sum a_j^2)^{\frac{1}{2}} (\sum b_j^2)^{\frac{1}{2}}$ and Theorems 2.1 and 2.2, we get

$$\begin{aligned} (46) \quad & E(H(u, \mathcal{B}_1)) - E(H(u, \mathcal{B}_1, \mathcal{B}_{2,-})) \\ & \geq E(u) - E(H(u, \frac{1}{2}\mathcal{B}_{2,-})) - C\varepsilon_0^{\frac{1}{2}} (E(u) - E(H(u, \mathcal{B}_1)))^{\frac{1}{2}} \\ & \geq E(u) - E(H(u, \frac{1}{2}\mathcal{B}_{2,-})) - C\varepsilon_0^{\frac{1}{2}} (E(u) - E(H(u, \mathcal{B}_1, \mathcal{B}_{2,-})))^{\frac{1}{2}}. \end{aligned}$$

Then noting that

$$\begin{aligned} E(H(u, \mathcal{B}_1)) - E(H(u, \mathcal{B}_1, \mathcal{B}_{2,-})) & \leq E(u) - E(H(u, \mathcal{B}_1, \mathcal{B}_{2,-})) \\ & \leq E(u) - E(H(u, \mathcal{B}_1, \mathcal{B}_2)) \leq \frac{2}{3}\varepsilon_0 < \varepsilon_0, \end{aligned}$$

we get the desired estimate (41).

So we only need to prove (45) to conclude Case 2. If B_j^2 is a classical ball, this is just the result in [6]. Now let us consider the case when B_j^2 is the boundary ball D_R^+ of radius R centered at 0 in the upper half-plane $\mathbb{H}^2$. Set $u_1 = H(u, \mathcal{B}_1)$. By the coarea

formula, there exists some $r \in [\frac{3}{4}R, R]$ with

$$\begin{aligned} \int_{\partial_r^A} |\nabla u_1 - \nabla u|^2 &\leq \frac{9}{R} \int_{\frac{3}{4}R}^R \left(\int_{\partial_s^A} |\nabla u_1 - \nabla u|^2 \right) ds \leq \frac{9}{r} \int_{D_R^+} |\nabla u_1 - \nabla u|^2, \\ \int_{\partial_r^A} (|\nabla u_1|^2 + |\nabla u|^2) &\leq \frac{9}{R} \int_{\frac{3}{4}R}^R \left(\int_{\partial_s^A} |\nabla u_1|^2 + |\nabla u|^2 \right) ds \\ &\leq \frac{9}{r} \int_{D_R^+} (|\nabla u_1|^2 + |\nabla u|^2) \leq \frac{24\varepsilon_0}{R}. \end{aligned}$$

The second estimate indicates that the length of the image of $u|_{\partial_r^A}$ is bounded by a universal constant times ε_0 . So if ε_0 is smaller than a constant multiple of κ (where κ is the radius for Fermi coordinates of Γ ; see [Lemma 6.5](#)), we will get $|u(x) - u(0)| \leq \frac{1}{2}\kappa$ for all $x \in \partial_r^A$. Also note that $B_R \in \mathcal{B}_{2,-}$, by definition $\frac{1}{2}B_R$ does not completely lie in $\mathcal{B}_1$, so ∂_r^A does not lie in $\mathcal{B}_1$ for all $r \in [\frac{1}{2}R, R]$. Hence u_1 and u must agree at one point on ∂_r^A . Then we can apply [Lemma 6.5](#) to get some $\rho \in (0, \frac{1}{2}r]$ and a map $w: D_r^+ \setminus D_{r-\rho}^+ \rightarrow \mathcal{N}$, with $w(r, \theta) = u_1(r, \theta)$ and $w(r - \rho, \theta) = u(r, \theta)$, such that

$$\int_{D_r^+ \setminus D_{r-\rho}^+} |\nabla w|^2 \leq C \left(\int_{D_r^+} |\nabla u|^2 + |\nabla u_1|^2 \right)^{\frac{1}{2}} \left(\int_{D_r^+} |\nabla u - \nabla u_1|^2 \right)^{\frac{1}{2}}.$$

The map $x \rightarrow H(u, D_r^+)(rx/(r - \rho))$ maps $D_{r-\rho}^+$ to $\mathcal{N}$ and agrees with w on $\partial_{r-\rho}^A$. So we get a map from D_R^+ to $\mathcal{N}$ which is equal to $H(u, B_1)$ on $D_R^+ \setminus D_r^+$, equal to w on $D_r^+ \setminus D_{r-\rho}^+$ and equal to $H(u, D_r^+)(r \cdot / (r - \rho))$ on $D_{r-\rho}^+$. This new map gives an upper bound for the energy of $H(u_1, D_R^+)$:

$$\int_{D_R^+} |\nabla H(u_1, D_R^+)|^2 \leq \int_{D_R^+ \setminus D_r^+} |\nabla u_1|^2 + \int_{D_r^+ \setminus D_{r-\rho}^+} |\nabla w|^2 + \int_{D_r^+} |\nabla H(u, D_r^+)|^2.$$

Using the previous estimate and the fact $||\nabla u_1|^2 - |\nabla u|^2| \leq (|\nabla u_1| + |\nabla u|)|\nabla(u - u_1)|$, we get

$$\begin{aligned} &\int_{D_R^+} |\nabla u_1|^2 - \int_{D_R^+} |\nabla H(u_1, D_R^+)|^2 \\ &\geq \int_{D_r^+} |\nabla u_1|^2 - \int_{D_r^+} |\nabla H(u, D_r^+)|^2 - \int_{D_r^+ \setminus D_{r-\rho}^+} |\nabla w|^2 \\ &\geq \int_{D_r^+} |\nabla u|^2 - \int_{D_r^+} |\nabla H(u, D_r^+)|^2 - C \left(\int_{D_R^+} |\nabla u|^2 + |\nabla u_1|^2 \right)^{\frac{1}{2}} \left(\int_{D_R^+} |\nabla(u - u_1)|^2 \right)^{\frac{1}{2}}. \end{aligned}$$

Since $r > \frac{1}{2}R$, we have

$$\begin{aligned} \int_{D_{R/2}^+} |\nabla u|^2 &= \int_{D_r^+} |\nabla u|^2 - \int_{D_r^+ \setminus D_{R/2}^+} |\nabla u|^2 \\ &= \int_{D_r^+} |\nabla u|^2 - \int_{D_r^+} |\nabla H(u, D_{R/2}^+)|^2 + \int_{D_{R/2}^+} |\nabla H(u, D_{R/2}^+)|^2 \\ &\leq \int_{D_r^+} |\nabla u|^2 - \int_{D_r^+} |\nabla H(u, D_r^+)|^2 + \int_{D_{R/2}^+} |\nabla H(u, D_{R/2}^+)|^2. \end{aligned}$$

Combining this with the previous estimate we get (45). Hence we can conclude the proof of (41).

Now one can prove (42) by the same argument as in [6], based on the proof of (41) above. This completes the proof. $\square$

6.3 Tightening process

Now we have developed the necessary tools to prove Theorem 4.3. In fact, the proof is virtually the same as the proof of [6, Theorem 2.1]. All the main ingredients in the free boundary setting needed in the proof have been obtained in previous sections. We give the proof here for completeness and the convenience of readers.

Again, let ε_0 be the minimum of the ε_0 in Theorems 2.1 and 2.2. Given a sweep-out $\sigma \in \Omega$ and $0 < \varepsilon \leq \varepsilon_0$, we define the maximal improvement for free boundary harmonic replacement on families of generalized closed balls with energy at most ε by

$$e_{\sigma, \varepsilon}(t) = \sup_{\mathcal{B}} \left\{ E(\sigma(\cdot, t)) - E\left(H\left(\sigma(\cdot, t), \frac{1}{2}\mathcal{B}\right)\right) \right\}.$$

Here the supremum is taken over all collections $\mathcal{B}$ of disjoint closed generalized balls in D where the total energy of $\sigma(\cdot, t)$ on $\mathcal{B}$ is at most ε . Note that $e_{\sigma, \varepsilon}(t)$ is positive if $\sigma(\cdot, t)$ is not harmonic.

We first show that the maximal improvement of a given slice (which is not harmonic) can control the maximal improvement of any nearby slices.

Lemma 6.8 *Given $t \in (0, 1)$, if $\sigma(\cdot, t)$ is not harmonic and $0 < \varepsilon \leq \varepsilon_0$, then there is an open interval I^t containing t such that $e_{\sigma, \varepsilon/2}(s) \leq 2e_{\sigma, \varepsilon}(t)$ for all s in the double interval $2I^t$.*

Proof By Theorem 6.1 and [6, Corollary 3.4], there exists $\delta_1 > 0$ depending on t such that if $\|\sigma(\cdot, t) - \sigma(\cdot, s)\|_{C^0 \cap W^{1,2}} < \delta_1$ and $\mathcal{B}$ is a finite collection of disjoint generalized closed balls in D where both $\sigma(\cdot, t)$ and $\sigma(\cdot, s)$ have energy at most ε_0 , then

$$(47) \quad \left| E(H(\sigma(\cdot, s), \tfrac{1}{2}\mathcal{B})) - E(H(\sigma(\cdot, t), \tfrac{1}{2}\mathcal{B})) \right| \leq \tfrac{1}{2}e_{\sigma, \varepsilon}(t).$$

Since $t \rightarrow \sigma(\cdot, t)$ is continuous as a map to $C^0 \cap W^{1,2}$, we can choose I^t so that for all $s \in 2I^t$ we have $\|\sigma(\cdot, t) - \sigma(\cdot, s)\|_{C^0 \cap W^{1,2}} < \delta_1$ and

$$(48) \quad \int_D |\nabla \sigma(\cdot, t)|^2 - |\nabla \sigma(\cdot, s)|^2 \leq \min\left\{\tfrac{\varepsilon}{2}, \tfrac{1}{2}e_{\sigma, \varepsilon}(t)\right\}.$$

Now suppose that $s \in 2I^t$ and the energy of $\sigma(\cdot, s)$ is at most $\frac{\varepsilon}{2}$ on a collection $\mathcal{B}$. It follows from (48) that the energy of $\sigma(\cdot, t)$ on this $\mathcal{B}$ is at most ε . Combining (47) with (48) we get

$$\left| E(\sigma(\cdot, s)) - E(H(\sigma(\cdot, s), \tfrac{1}{2}\mathcal{B})) - E(\sigma(\cdot, t)) + E(H(\sigma(\cdot, t), \tfrac{1}{2}\mathcal{B})) \right| \leq e_{\sigma, \varepsilon}(t).$$

Since this estimate applies to any such $\mathcal{B}$, and noting that $e_{\sigma, \varepsilon}$ is monotone nondecreasing in ε , we get that $e_{\sigma, \varepsilon/2}(s) \leq 2e_{\sigma, \varepsilon}(t)$. $\square$

Theorem 4.3 indicates that our tightening process should effectively decrease the energy of those nonharmonic slices with large energy, ie when $E \geq \frac{1}{2}W$. The next lemma shows that we can find a harmonic replacement to decrease the energy of those slices by a certain amount.

Lemma 6.9 *If $W > 0$ and $\tilde{\gamma} \in \Omega$ has no harmonic slices other than $\tilde{\gamma}(\cdot, 0)$ and $\tilde{\gamma}(\cdot, 1)$, then we get an integer m depending on $\tilde{\gamma}$, along with m collections $\mathcal{B}_1, \dots, \mathcal{B}_m$ of generalized closed balls in D where the balls in each collection $\mathcal{B}_j$ are pairwise disjoint, and m continuous functions $r_1, \dots, r_m: [0, 1] \rightarrow [0, 1]$ such that, for each t :*

- (i) *At most two $r_j(t)$ are positive and $\sum_{\mathcal{B} \in \mathcal{B}_j} \int_{r_j(t)\mathcal{B}} |\nabla \tilde{\gamma}(\cdot, t)|^2 < \frac{1}{3}\varepsilon_0$ for each j .*
- (ii) *If $E(\tilde{\gamma}(\cdot, t)) \geq \frac{1}{2}W$, then there exists $j(t)$ such that the harmonic replacement on $(\frac{1}{2}r_{j(t)})\mathcal{B}_{j(t)}$ decreases energy by at least $\frac{1}{8}e_{\tilde{\gamma}, \varepsilon_0/8}(t)$.*

Proof The energy of the slices is continuous in t , so the set $I = \{t : E(\tilde{\gamma}(\cdot, t)) \geq \frac{1}{2}W\}$ is compact. For each $t \in I$, choose a finite collection $\mathcal{B}^t$ of disjoint closed balls in D with $\int_{\bigcup \mathcal{B}^t} |\nabla \tilde{\gamma}(\cdot, t)|^2 \leq \frac{1}{4}\varepsilon_0$ so that

$$(49) \quad E(\tilde{\gamma}(\cdot, t)) - E(H(\tilde{\gamma}(\cdot, t), \tfrac{1}{2}\mathcal{B}^t)) \geq \tfrac{1}{2}e_{\tilde{\gamma}, \varepsilon_0/4}(t) > 0.$$

Note that $0, 1 \notin I$. [Lemma 6.8](#) gives an open interval I^t containing t such that for all $s \in 2I^t$,

$$(50) \quad e_{\tilde{\gamma}, \varepsilon_0/8}(s) \leq 2e_{\tilde{\gamma}, \varepsilon_0/4}(t).$$

Using the continuity of $\tilde{\gamma}(\cdot, s)$ in $C^0 \cap W^{1,2}$, we can shrink I^t so that $\tilde{\gamma}(\cdot, s)$ has energy at most $\frac{1}{3}\varepsilon_0$ in B^t for $s \in 2I^t$ and in addition, using [Theorem 6.1](#) and [\[6, Corollary 3.4\]](#), we have

$$(51) \quad \left| E(\tilde{\gamma}(\cdot, s)) - E(H(\tilde{\gamma}(\cdot, s), \tfrac{1}{2}B^t)) - E(\tilde{\gamma}(\cdot, t)) + E(H(\tilde{\gamma}(\cdot, t), \tfrac{1}{2}B^t)) \right| \leq \tfrac{1}{4}e_{\tilde{\gamma}, \varepsilon_0/4}(t).$$

Since I is compact, we can cover I by finitely many of the I^t , say $I^{t_1}, \dots, I^{t_m}$. By discarding some of them, we can arrange that each t lies in at least one $\overline{I^{t_j}}$ and at most two consecutive $\overline{I^{t_j}}$. That is, we get a family of intervals I^{t_j} such that I^{t_j} only intersects $I^{t_{j-1}}$ and $I^{t_{j+1}}$, and $I^{t_{j-1}}$ and $I^{t_{j+1}}$ do not intersect each other. Now for each $j = 1, \dots, m$, we choose a continuous function $r_j(t): [0, 1] \rightarrow [0, 1]$ with the following properties: $r_j(t) = 1$ on $\overline{I^{t_j}}$ and $r_j(t) = 0$ for $t \notin 2I^{t_j} \cap (I^{t_{j-1}} \cup I^{t_j} \cup I^{t_{j+1}})$; $r_j(t) = 0$ on the intervals that do not intersect $\overline{I^{t_j}}$.

Property (1) follows directly from the construction of the $r_j(t)$, and property (2) follows from combining (49), (50) and (51). $\square$

Now using this lemma we can prove [Theorem 4.3](#).

Proof of Theorem 4.3 Let $B_1, \dots, B_m$ and $r_1, \dots, r_m$ be given by [Lemma 6.9](#). We will use an m -step replacement process to define γ .

We first set $\gamma^0 = \tilde{\gamma}$. Then for $j = 1, 2, \dots, m$ define $\gamma^j(\cdot, t) = H(\gamma^{j-1}(\cdot, t), r_j(t)B_j)$. Finally we set $\gamma = \gamma^m$.

We first claim that this is a well-defined process and γ is again in $\Omega\tilde{\gamma}$. In fact, property (1) in [Lemma 6.9](#) implies that each energy-minimizing map replaces a map with energy at most $\frac{2}{3}\varepsilon_0 < \varepsilon_0$. Moreover, [Theorem 6.1](#) and [\[6, Corollary 3.4\]](#) imply that the replacement depends continuously on the boundary values, which are themselves continuous in t . Finally it is clear that γ is homotopic to $\tilde{\gamma}$ since continuously shrinking the disjoint closed balls on which we make harmonic replacement gives an explicit homotopy.

Now we show that this γ satisfies the requirements of [Theorem 4.3](#). Suppose $t \in [0, 1]$ is chosen with $E(\tilde{\gamma}(\cdot, t)) \geq \frac{1}{2}W$, then property (2) of [Lemma 6.9](#) implies that the

harmonic replacement for $\tilde{\gamma}(\cdot, t)$ on $(\frac{1}{2}r_j(t))\mathcal{B}_{j(t)}$ decreases the energy by at least $\frac{1}{8}e_{\tilde{\gamma}, \varepsilon_0}(t)$. Thus using (41) of Lemma 6.7 we get

$$(52) \quad E(\tilde{\gamma}(\cdot, t)) - E(\gamma(\cdot, t)) \geq k\left(\frac{1}{8}e_{\tilde{\gamma}, \varepsilon_0/8}(t)\right)^2 \quad \text{for some } k > 0.$$

Suppose that $\mathcal{B}$ is a finite collection of disjoint generalized closed balls in D such that the energy of $\gamma(\cdot, t)$ on $\mathcal{B}$ is at most $\frac{1}{12}\varepsilon_0$. We can assume $\gamma^k(\cdot, t)$ has energy at most $\frac{1}{8}\varepsilon_0$ on $\mathcal{B}$ for every k , since otherwise using Theorems 2.1 and 2.2 we could choose Ψ to be a linear function $\Psi(x) = Cx$ for some $C > 0$ sufficiently large. Now we apply (42) in Lemma 6.7 twice with $\mu = \frac{1}{8}$ and then $\mu = \frac{1}{4}$ to get

$$(53) \quad \begin{aligned} E(\gamma(\cdot, t)) - E(H(\gamma(\cdot, t), \tfrac{1}{8}\mathcal{B})) \\ \leq E(\tilde{\gamma}(\cdot, t)) - E(H(\tilde{\gamma}(\cdot, t), \tfrac{1}{2}\mathcal{B})) + \tfrac{2}{k}(E(\tilde{\gamma}(\cdot, t)) - E(\gamma(\cdot, t)))^{\frac{1}{2}} \\ \leq e_{\tilde{\gamma}, \varepsilon_0/8}(t) + \tfrac{2}{k}(E(\tilde{\gamma}(\cdot, t)) - E(\gamma(\cdot, t)))^{\frac{1}{2}}. \end{aligned}$$

Combining (52) and (53) with Theorems 2.1 and 2.2, we complete the proof by choosing $\Psi(x) = Cx + Cx^{\frac{1}{2}}$ for some sufficiently large $C > 0$ and $\varepsilon_1 = \frac{1}{12}\varepsilon_0$. $\square$

7 Compactness of maximal slices

This section is devoted to the proof of Theorem 4.4. In particular, we will prove that any sequence of maps (as slices of approximating sweepouts) whose energy converges to the width will converge to a bubble tree of free boundary harmonic disks and harmonic spheres. Similar bubble tree convergence was first studied by Fraser [15] for α -harmonic disks with free boundary (where the notion of α -maps was introduced by Sacks–Uhlenbeck [45]). In our proof, we adopt the schemes in both [6, Appendix B] and [15]. We generalize the notion of almost harmonic maps with their asymptotic analysis and compactness in [6, Appendix B] to our free boundary setting.

7.1 Compactness of free boundary almost harmonic maps

We first introduce our notion of almost harmonic maps with free boundary. Note that generalized balls were defined in Definition 4.2.

Definition 7.1 We say a sequence of maps $u^j: (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ is ε_1 -almost harmonic in the free boundary sense if:

(B_0) For any generalized ball $B \subset D$ with $\int_B |\nabla u^j|^2 < \varepsilon_1$, there is a harmonic replacement $v: \frac{1}{8}B \rightarrow \mathcal{N}$ of u^j (with possible free boundary along $v(\partial \frac{1}{8}B \cap \partial D) \subset \Gamma$ if it is not empty) which satisfies the bound

$$\int_{\frac{1}{8}B} |\nabla u^j - \nabla v|^2 \leq \frac{1}{j}.$$

Next we have the following preliminary compactness result for a sequence of almost harmonic maps in the free boundary sense with finite energy. In particular, any such sequence converges to a harmonic map with free boundary in $W^{1,2}$ locally away from only finitely many points in D . Let ε_{SU} and ε_F be as in [Section 4](#).

Theorem 7.2 *Let $\varepsilon_1 > 0$ be such that $\varepsilon_1 < \min\{\varepsilon_{SU}, \varepsilon_F\}$. Let $u^j: (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ be a sequence of $W^{1,2}$ -maps satisfying property (B_0) in [Definition 7.1](#) and having $E(u^j) \leq E_0 < \infty$. Then there exist a finite collection of points $\{x_1, \dots, x_k\} \subset \bar{D}$, a subsequence of maps (still denoted by u^j), and a harmonic map $u: (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ with free boundary, such that $u^j \rightarrow u$ weakly in $W^{1,2}(D)$ and for any compact subset $K \subset \bar{D} \setminus \{x_1, \dots, x_k\}$, $u^j \rightarrow u$ strongly in $W^{1,2}(K)$. Furthermore, the measures $|\nabla u^j|^2 dx$ converge to a measure ν on $\bar{D}$ with $\nu(\{x_i\}) \geq \varepsilon_1$ for all $1 \leq i \leq k$ and $\nu(\bar{D}) \leq E_0$.*

Proof After passing to a subsequence we can assume that the u^j converge weakly in $W^{1,2}(D)$ to a $W^{1,2}$ -map $u: D \rightarrow \mathcal{N}$, and the measures $|\nabla u^j|^2 dx$ converge weakly to a limiting measure ν on $\bar{D}$ with $\nu(\bar{D}) \leq E_0$.

So there are at most E_0/ε_1 points $x_1, \dots, x_k \in \bar{D}$, with $\lim_{r \rightarrow 0} \nu(B_r(x_j)) \geq \varepsilon_1$.

Next we show that away from these points the convergence is strong in $W^{1,2}$ and u is a harmonic map with free boundary in Γ . Given any $x \in \bar{D} \setminus \{x_1, \dots, x_k\}$, by definition there exist a generalized ball B_x and an integer J_x such that $\int_{B_x} |\nabla u^j|^2 < \varepsilon_1$ for $j > J_x$. If x is an interior point we can choose B_x to be a classical ball which is contained in D , otherwise B_x will be some boundary ball. By condition (B_0) we get a harmonic replacement $v_x^j: \frac{1}{8}B_x \rightarrow \mathcal{N}$ such that

$$\int_{\frac{1}{8}B_x} |\nabla u^j - \nabla v_x^j| \leq \frac{1}{j}.$$

Note that the energy of v_x^j on the ball $\frac{1}{8}B_x(x)$ is less than ε_1 , so by the ε -regularity of Qing [[39](#), Lemma 4] and [Theorem 2.5](#), we get a uniform C^k -bound for v_x^j in $\frac{1}{9}B_x$ for $k \geq 1$. Hence a subsequence v_x^j converges strongly in $W^{1,2}(\frac{1}{9}B_x)$ to a harmonic

map $v_x: \frac{1}{9}B_x \rightarrow \mathcal{N}$ with possible free boundary along $v_x(\partial \frac{1}{9}B_x \cap \partial D) \subset \Gamma$. By the triangle inequality we get

$$\int_{\frac{1}{9}B_x} |\nabla u^j - \nabla v_x|^2 \leq 2 \int_{\frac{1}{9}B_x} |\nabla u^j - \nabla v_x^j|^2 + 2 \int_{\frac{1}{9}B_x} |\nabla v_x^j - \nabla v_x|^2 \rightarrow 0$$

as $j \rightarrow \infty$.

We can also derive the L^2 -convergence of u^j to v_x on $\frac{1}{9}B_x$ by the inequality

$$\int_{\frac{1}{9}B_x} |u^j - v_x|^2 \leq 2 \int_{\frac{1}{9}B_x} |u^j - v_x^j|^2 + 2 \int_{\frac{1}{9}B_x} |v_x^j - v_x|^2$$

and the Poincaré inequality (when B_x is a classical ball of D) or its variant [Lemma A.2](#) (when B_x is a boundary ball of D).

We have proved that the sequence u^j converges to v_x strongly in $W^{1,2}(\frac{1}{9}B_x)$, and hence $u = v_x$ in $\frac{1}{9}B_x$. Therefore we conclude that u is a free boundary harmonic map on $D \setminus \{x_1, \dots, x_k\}$. Furthermore, for K relatively compact in $\bar{D} \setminus \{x_1, \dots, x_k\}$, the $W^{1,2}(K)$ -convergence of u^j to u follows from a standard covering argument.

Finally, since u has finite energy, we can apply the removable singularity theorem by Sacks–Uhlenbeck [\[45, Theorem 3.6\]](#) or Fraser [\[15, Theorem 1.10\]](#) at each x_i for interior points or free boundary points respectively, so that u extends to a global harmonic map on the whole of D with free boundary along $u(\partial D) \subset \Gamma$. $\square$

7.2 Harmonic maps on half-cylinders

In this and the following subsections, we will generalize the analysis of harmonic maps defined on cylinders in [\[6, Appendix B\]](#) to harmonic maps with free boundary defined on half-cylinders. The analysis of harmonic maps and almost harmonic maps on cylinders in [\[6\]](#) is essential to the proof of the so-called *energy identity*. More precisely, in the blow-up process (see [Section 7.4](#)), the energy could be lost (so that the energy identity fails) only when some energy escapes from the “necks” (modeled by cylinders or half-cylinders), and this is the case we want to rule out in our scenario. In the free boundary setting, not only spherical bubbles but also disk bubbles may appear during the blow-up process. So we need to carefully analyze the maps defined on the “necks” between the spheres and disks, ie cylinders and half-cylinders.

Let us set up some notation. Let $\mathcal{C}_{a,b}$ denote the flat half-cylinder $[a, b] \times [0, \pi]$, where $[0, \pi]$ can be viewed as half of a circle S^1 . We will use (t, θ) as parameters on $\mathcal{C}_{a,b}$. Note that $\mathcal{C}_{a,b}$ is conformally equivalent to the half-annulus in plane: $[e^a, e^b] \times [0, \pi]$



Figure 2

under polar coordinates, see Figure 2. We will also abuse the notation $\partial^C \mathcal{C}$ to denote the part of the boundary $[a, b] \times \{0, \pi\}$, and later this will be the free boundary part of our maps and we will only focus on the boundary behavior of the maps along this part. Moreover, when we say a map $u: \mathcal{C} \rightarrow \mathcal{N}$ is a harmonic map with free boundary on Γ , we will always assume $u|_{\partial^C \mathcal{C}}$ is the free boundary with $u(\partial^C \mathcal{C}) \subset \Gamma$.

Theorem 7.3 *Given $\delta > 0$, there exist $\varepsilon_2 > 0$ and $l > 1$ depending on δ such that if u is a nonconstant C^3 -harmonic map from the flat half-cylinder $\mathcal{C}_{-3l, 3l}$ to $\mathcal{N}$ with free boundary along Γ , and the energy $E(u) \leq \varepsilon_2$, then*

$$(54) \quad \int_{\mathcal{C}_{-l, l}} |u_\theta|^2 < \delta \int_{\mathcal{C}_{-2l, 2l}} |\nabla u|^2.$$

Roughly speaking, this theorem implies that for nonconstant free boundary harmonic maps with small energy, the θ -energy (on a sub-half-cylinder) is far smaller than the total energy. Then by the Cauchy–Schwartz inequality this implies that the area of the image of u is strictly less than the energy of u .

In order to prove Theorem 7.3, we follow the idea of Colding–Minicozzi to prove a differential inequality for various energies of a free boundary harmonic map. In particular we need some free boundary versions of the lemmas in [6, Appendix B].

Lemma 7.4 *Given a C^3 -free boundary harmonic map u from $\mathcal{C}_{-3l, 3l}$ to $\mathcal{N} \subset \mathbb{R}^N$ with $E(u) \leq \varepsilon_2$ for some $\varepsilon_2 > 0$ sufficiently small, we have*

$$(55) \quad \partial_t^2 \int_t |u_\theta|^2 \geq \frac{1}{C} \int_t |u_\theta|^2 - C \int_t |\nabla u|^4 - C \sqrt{\varepsilon_2} \frac{1}{\lambda} \int_{t-\lambda}^{t+\lambda} \int_s |\nabla u|^2 ds$$

for all $t \in [-2l, 2l]$ and $\lambda \in (0, 1)$. Here $\int_t(\cdot)$ denotes the integral over the t -slice $\{(t, \theta) \in \mathcal{C}_{-3l, 3l} : \theta \in [0, \pi]\}$, and $C > 0$ is some universal constant depending only on $\mathcal{N}$ and Γ .

Proof We can assume $\varepsilon_2 \leq \varepsilon_0$, as in [Theorem 2.5](#). Differentiating $\int_t |u_\theta|^2$ twice gives

$$\begin{aligned}
 (56) \quad \frac{1}{2} \partial_t^2 \int_t |u_\theta|^2 &= \int_t |u_{t\theta}|^2 + \int_t \langle u_\theta, u_{tt\theta} \rangle \\
 &= \int_t |u_{t\theta}|^2 - \int_t \langle u_{\theta\theta}, u_{tt} \rangle + \langle u_\theta, u_{tt} \rangle|_0^\pi \\
 &= \int_t |u_{t\theta}|^2 - \int_t \langle u_{\theta\theta}, (\Delta u - u_{\theta\theta}) \rangle + \langle u_\theta, A^\Gamma(u_t, u_t) \rangle|_0^\pi \\
 &\geq \int_t |u_{t\theta}|^2 + \int_t |u_{\theta\theta}|^2 - \sup_{\mathcal{N}} |A^\mathcal{N}| \int_t |u_{\theta\theta}| |\nabla u|^2 \\
 &\quad - 2 \sup_{\Gamma} |A^\Gamma| C \sqrt{\varepsilon_2} \frac{1}{\lambda} \int_{t-\lambda}^{t+\lambda} \int_s |\nabla u|^2 ds.
 \end{aligned}$$

Here A^Γ and $A^\mathcal{N}$ are the second fundamental forms of $\Gamma \hookrightarrow \mathcal{N}$ and $\mathcal{N} \hookrightarrow \mathbb{R}^N$. In the second equality we used integration by parts; in the third inequality, since u is a free boundary map, on the boundary (where $\theta = 0$ or $\theta = \pi$) u_θ is a tangent vector of $\mathcal{N}$ and also perpendicular to Γ , and $u_{tt} = \nabla_{u_t}^\mathcal{N} u_t$ when projected to the tangent space $T_u \mathcal{N}$, where u_t is tangent to Γ ; in the last inequality we used $|\Delta u| \leq \sup_{\mathcal{N}} |A^\mathcal{N}| |\nabla u|^2$ by the harmonic map [\(3\)](#), and also the gradient estimate ([Theorem 2.5](#)) applied to half-disks of radius λ centered at $(t, 0)$ and (t, π) . By the Cauchy–Schwartz inequality we get

$$\frac{1}{2} \partial_t^2 \int_t |u_\theta|^2 \geq \int_t |u_{t\theta}|^2 + \frac{3}{4} \int_t |u_{\theta\theta}|^2 - C \int_t |\nabla u|^4 - C \sqrt{\varepsilon_2} \frac{1}{\lambda} \int_{t-\lambda}^{t+\lambda} \int_s |\nabla u|^2 ds$$

for some universal constant $C > 0$ depending only on $\mathcal{N}$ and Γ .

Next we claim: for u satisfying the assumptions of the theorem,

$$(57) \quad \int_t |u_\theta|^2 \leq C \int_t |u_{\theta\theta}|^2 \quad \text{for } t \in [-2l, 2l],$$

where $C > 0$ is a constant depending only on $\mathcal{N}$ and Γ . Note that once we have this inequality we get the desired inequality [\(55\)](#) in the lemma.

To prove [\(57\)](#), we reflect the map $u(t, \cdot): [0, \pi] \rightarrow \mathbb{R}^N$ across Γ to obtain a map defined on the circle $\hat{u}(t, \cdot): S^1 \rightarrow \mathbb{R}^N$. In particular, let P_Γ denote the nearest-point projection map from a tubular neighborhood of Γ in $\mathbb{R}^N$ to Γ . When the tubular neighborhood is chosen small enough, we can assume that

$$|DP_\Gamma| \leq 1 \quad \text{and} \quad |D^2 P_\Gamma| \leq C$$

for some universal constant $C > 0$. By the gradient estimate [Theorem 2.5](#), we can assume that the image of $u(\mathcal{C}_{-2l, 2l})$ lies in this tubular neighborhood when $\varepsilon_2 > 0$ is

chosen small enough. Now define the map $\hat{u}(t, \cdot): S^1 \rightarrow \mathbb{R}^N$ by

$$\hat{u}(t, \theta) = \begin{cases} u(t, \theta) & \text{when } \theta \in [0, \pi], \\ 2P_\Gamma(u(t, -\theta)) - u(t, -\theta) & \text{when } \theta \in [-\pi, 0]. \end{cases}$$

By the free boundary assumption, we know that $\hat{u}$ is $C^{1,1}$ on S^1 . As $\int_{S^1} \hat{u}_\theta = 0$, by the Wirtinger inequality we get $\int_{S^1} |\hat{u}_\theta|^2 \leq \int_{S^1} |u_{\theta\theta}|^2$, and hence we can deduce that

$$\begin{aligned} (58) \quad \int_0^\pi |u_\theta|^2 &\leq \int_0^\pi |u_{\theta\theta}|^2 + \int_0^\pi |(2P_\Gamma(u(t, \theta)) - u(t, \theta))_{\theta\theta}|^2 \\ &\leq \int_0^\pi |u_{\theta\theta}|^2 + \int_0^\pi |2DP_\Gamma(u_{\theta\theta}) - u_{\theta\theta} + 2D^2P_\Gamma(u_\theta, u_\theta)|^2 \\ &\leq C \int_0^\pi |u_{\theta\theta}|^2 + C \int_0^\pi |u_\theta|^4. \end{aligned}$$

Note that by the gradient estimates [Theorem 2.5](#), $C \int_0^\pi |u_\theta|^4 \leq C^2 \varepsilon_2 \int_0^\pi |u_\theta|^2$. So the desired estimates (57) follow by taking $\varepsilon_2 > 0$ small enough, so that $C^2 \varepsilon_2 < \frac{1}{2}$. $\square$

The next lemma is an ODE comparison lemma.

Lemma 7.5 Suppose f is a nonnegative C^2 function on $[-2l, 2l] \subset \mathbb{R}$ satisfying

$$(59) \quad f'' \geq \frac{1}{4C} f - a$$

for some constants $C, a > 0$. If $\max_{[-l, l]} f \geq 8Ca$, then

$$(60) \quad \int_{-2l}^{2l} f \geq 4\sqrt{2Ca} \sinh\left(\frac{l}{2\sqrt{2C}}\right).$$

Proof Let $\tilde{f}(t) = f(\sqrt{4C}t)$. Then we get a differential inequality for $\tilde{f}$,

$$\tilde{f}'' \geq \tilde{f} - 4Ca,$$

where $\tilde{f}$ is defined on $[-l/\sqrt{C}, l/\sqrt{C}]$, and

$$\max_{[-l/2\sqrt{C}, l/2\sqrt{C}]} \tilde{f} \geq 8Ca.$$

Then applying [\[6, Lemma B.4\]](#) gives

$$\int_{-l/\sqrt{C}}^{l/\sqrt{C}} \tilde{f} \geq 8\sqrt{2Ca} \sinh\left(\frac{l}{2\sqrt{2C}}\right),$$

which implies

$$\int_{-2l}^{2l} f \geq 4\sqrt{2Ca} \sinh\left(\frac{l}{2\sqrt{2C}}\right). \quad \square$$

Now we can prove the main theorem of this subsection.

Proof of Theorem 7.3 First we analyze $\int_t (|u_t|^2 - |u_\theta|^2)$. Differentiating it and applying integration by parts gives

$$(61) \quad \frac{1}{2} \partial_t \int_t (|u_t|^2 - |u_\theta|^2) = \int_t (\langle u_t, u_{tt} \rangle - \langle u_\theta, u_{\theta t} \rangle) = \int_t \langle u_t, u_{tt} + u_{\theta\theta} \rangle = 0.$$

Here again we use the fact that u_θ and u_t are perpendicular to each other on the free boundary, and $u_{tt} + u_{\theta\theta} = \Delta u$, which is normal to $\mathcal{N}$ and hence perpendicular to u_t . Thus $\int_t (|u_t|^2 - |u_\theta|^2)$ is a constant, and

$$(62) \quad \int_t (|u_t|^2 - |u_\theta|^2) = \frac{1}{4l} \int_{C_{-2l, 2l}} (|u_t|^2 - |u_\theta|^2) \leq \frac{1}{4l} \int_{C_{-2l, 2l}} |\nabla u|^2.$$

Moreover, we get

$$(63) \quad \int_t |\nabla u|^2 = 2 \int_t |u_\theta|^2 + \int_t (|u_t|^2 - |u_\theta|^2) \leq 2 \int_t |u_\theta|^2 + \frac{1}{4l} \int_{C_{-2l, 2l}} |\nabla u|^2.$$

Let us choose ε_2 smaller than the values of ε_0 in [45, Lemma 3.4] and Theorem 2.5. Then the interior gradient estimates for harmonic maps (see [45, Lemma 3.4]) and for free boundary harmonic maps (Theorem 2.5) imply that

$$(64) \quad \sup_{C_{-2l, 2l}} |\nabla u|^2 \leq C \varepsilon_2.$$

Let $f(t) = \int_t |u_\theta|^2$. Then by Lemma 7.4 we get

$$(65) \quad \begin{aligned} f''(t) &\geq \frac{1}{C} f(t) - C \varepsilon_2 \int_t |\nabla u|^2 - C \sqrt{\varepsilon_2} \frac{1}{\lambda} \int_{t-\lambda}^{t+\lambda} \int_s |\nabla u|^2 ds \\ &\geq \frac{1}{C} f(t) - 2C \varepsilon_2 f(t) - 2C \sqrt{\varepsilon_2} \frac{1}{\lambda} \int_{t-\lambda}^{t+\lambda} f(t) \\ &\quad - \frac{C \varepsilon_2}{4l} \int_{C_{-2l, 2l}} |\nabla u|^2 - \frac{2C \sqrt{\varepsilon_2}}{4l} \int_{C_{-2l, 2l}} |\nabla u|^2 \\ &\geq \frac{1}{2C} \left(f(t) - \frac{1}{4\lambda} \int_{t-\lambda}^{t+\lambda} f \right) - \frac{C \sqrt{\varepsilon_2}}{l} \int_{C_{-2l, 2l}} |\nabla u|^2, \end{aligned}$$

where $C > 0$ depends only on $\mathcal{N}$ and Γ , and we can further assume that ε_2 is small enough that $C^2 \varepsilon_2 \leq \frac{1}{4}$ and $2C^2 \varepsilon_2^{1/4} \leq 1$.

Letting $\lambda \rightarrow 0$, by continuity of f we get the differential inequality

$$(66) \quad f''(t) \geq \frac{1}{4C} f(t) - a,$$

where $a = (C\sqrt{\varepsilon_2}/l) \int_{C_{-2l,2l}} |\nabla u|^2$. Then we apply [Lemma 7.5](#) to get either

$$(67) \quad \max_{[-l,l]} f < 8 \frac{C^2 \sqrt{\varepsilon_2}}{l} \int_{C_{-2l,2l}} |\nabla u|^2$$

or

$$(68) \quad \int_{C_{-2l,2l}} |u_\theta|^2 = \int_{-2l}^{2l} f(t) dt \geq 4\sqrt{2}C \sqrt{\varepsilon_2} \frac{\sinh(l/2\sqrt{2}C)}{l} \int_{C_{-2l,2l}} |\nabla u|^2.$$

If we choose l large enough then the second inequality can not hold. Then we get

$$(69) \quad \int_{C_{-l,l}} |u_\theta|^2 \leq 2l \max_{[-l,l]} f < 8C^2 \sqrt{\varepsilon_2} \int_{C_{-2l,2l}} |\nabla u|^2.$$

Then the inequality [\(54\)](#) holds if $8C^2 \sqrt{\varepsilon_2} < \delta$. So we can choose ε_2 small and then choose l large to get the desired inequality. $\square$

7.3 Almost harmonic maps on half-cylinders

The main results in this part generalize the results in the previous subsection to almost harmonic maps on half-cylinders.

Let us first fix some notation. Given a half-cylinder C_{r_1, r_2} , we will view it as its conformally equivalent half-annulus $D_{e^{r_2}}^+ \setminus D_{e^{r_1}}^+ \subset D^+$. A generalized ball $B \subset D_{e^{r_2}}^+ \setminus D_{e^{r_1}}^+$ is either a ball in the interior of the annulus $D_{e^{r_2}}^+ \setminus D_{e^{r_1}}^+$, or is a half-ball centered along the chord boundary $\partial_{e^{r_2}}^C \setminus \partial_{e^{r_1}}^C$. When B is a half-ball, we write $\partial^C B = \partial B \cap \partial^C$. Note that this definition is the same as that in [Definition 4.2](#).

Definition 7.6 Given $\mu > 0$ and a half-cylinder C_{r_1, r_2} , we will say that a $W^{1,2}$ -map $u: C_{r_1, r_2} \rightarrow \mathcal{N}$ with $u(\partial^C C) \subset \Gamma$ is μ -almost harmonic with free boundary if for any finite collection of disjoint generalized balls $\mathcal{B}$ in the conformally equivalent half-annulus $D_{e^{r_2}}^+ \setminus D_{e^{r_1}}^+$, there is a free boundary harmonic replacement $v: \bigcup_{\mathcal{B}} \frac{1}{8} B \rightarrow \mathcal{N}$ with free boundary along $u(\bigcup_{\mathcal{B}} \frac{1}{8} \partial^C B) \subset \Gamma$ such that

$$\int_{\bigcup_{\mathcal{B}} \frac{1}{8} B} |\nabla u - \nabla v|^2 \leq \frac{\mu}{2} \int_{C_{r_1, r_2}} |\nabla u|^2.$$

The first lemma of this subsection shows that for almost harmonic maps with free boundary the estimate in the previous subsection still holds.

Lemma 7.7 Given $\delta > 0$ there exists $\mu > 0$ depending on δ , $\mathcal{N}$ and Γ such that if $u: \mathcal{C}_{-3l,3l} \rightarrow \mathcal{N}$ is a μ -almost harmonic map with free boundary and $E(u) \leq \varepsilon_2$, where l is from [Theorem 7.3](#), then

$$(70) \quad \int_{\mathcal{C}_{-l,l}} |u_\theta|^2 \leq \delta \int_{\mathcal{C}_{-3l,3l}} |\nabla u|^2.$$

Proof We will argue by contradiction. Suppose the lemma does not hold. Then there exists a sequence u^j of $\frac{1}{j}$ -almost harmonic maps from $\mathcal{C}_{-3l,3l}$ to $\mathcal{N}$ with free boundary $u_j(\partial^C \mathcal{C}_{-3l,3l}) \subset \Gamma$, energy $E(u^j) \leq \varepsilon_2$, and

$$(71) \quad \int_{\mathcal{C}_{-l,l}} |u_\theta^j|^2 > \delta \int_{\mathcal{C}_{-3l,3l}} |\nabla u^j|^2.$$

Now we have two cases depending on whether the energy of the limit is zero.

Case 1 Suppose $\limsup_{j \rightarrow \infty} E(u^j) > 0$. Then up to a subsequence, $\int_{\mathcal{C}_{-l,l}} |u_\theta^j|^2$ is uniformly bounded from below by [\(71\)](#). We will apply the compactness result ([Theorem 7.2](#)) to this sequence. In particular, we can use the same argument as [Theorem 7.2](#) to find a subsequence that converges weakly to a free boundary harmonic map $u: (\mathcal{C}_{-3l,3l}, \partial^C \mathcal{C}_{-3l,3l}) \rightarrow (\mathcal{N}, \Gamma)$, and strongly in $W^{1,2}$ on any compact subset of $\mathcal{C}_{-3l,3l}$. Note that since $E(u^j) \leq \varepsilon_2$, there will be no energy concentration points. The uniform lower bound of $\int_{\mathcal{C}_{-l,l}} |u_\theta^j|^2$ and the $W^{1,2}$ -convergence on $\mathcal{C}_{-l,l}$ imply that u can not be a constant map. Finally by the lower semicontinuity of energy along $W^{1,2}$ -weak convergence,

$$\int_{\mathcal{C}_{-l,l}} |u_\theta|^2 \geq \delta \int_{\mathcal{C}_{-3l,3l}} |\nabla u|^2,$$

which contradicts [Theorem 7.3](#).

Case 2 Suppose $\lim_{j \rightarrow \infty} E(u^j) = 0$. We will use a blow-up argument. Let

$$v^j = \frac{u^j - u^j(0)}{E(u^j)^{1/2}}.$$

This is a sequence of maps from $\mathcal{C}_{-3l,3l}$ to $\mathcal{N}_j = (\mathcal{N} - u^j(0))/(E(u^j)^{1/2})$. Here $0 = (0, 0)$ is a boundary point on $\mathcal{C}_{-3l,3l}$, so $u^j(0) \in \Gamma$; hence we can always see the free boundary $\Gamma_j = (\Gamma - u^j(0))/(E(u^j)^{1/2})$ in the blow-up process.

Note that $E(v^j) = 1$ and by [\(71\)](#), we have

$$\int_{\mathcal{C}_{-l,l}} |v_\theta^j|^2 > \delta > 0.$$

Furthermore, the sequence of the v^j are still $\frac{1}{j}$ -almost harmonic because this property is invariant under dilation. So we can argue as before to get a subsequence that converges in $W^{1,2}$ on compact subsets of $\mathcal{C}_{-3l,3l}$ to a free boundary harmonic map $v: \mathcal{C}_{-3l,3l} \rightarrow \mathbb{R}^n \subset \mathbb{R}^N$ with free boundary $\Gamma = \mathbb{R}^k \subset \mathbb{R}^N$. As before, we get

$$\int_{\mathcal{C}_{-l,l}} |v_\theta|^2 \geq \delta,$$

which again contradicts [Theorem 7.3](#). (Note that for free boundary harmonic maps into $(\mathbb{R}^n, \mathbb{R}^k)$ we do not need the assumption $E(v) \leq \varepsilon_2$; see also [\[6, Remark B.3\]](#).) $\square$

With this lemma we can prove that the θ -energy of a free boundary almost harmonic map on a long half-cylinder would be far less than the total energy.

Theorem 7.8 *Given $\delta > 0$ there exists $\nu > 0$ depending on δ , $\mathcal{N}$ and Γ such that if m is any positive integer and u is ν -almost harmonic from $\mathcal{C}_{-(m+3)l,3l}$ to $\mathcal{N}$ with free boundary along $u(\partial^C \mathcal{C}_{-(m+3)l,3l}) \subset \Gamma$, and $E(u) \leq \varepsilon_2$, then*

$$(72) \quad \int_{\mathcal{C}_{-ml,0}} |u_\theta|^2 \leq 7\delta \int_{\mathcal{C}_{-(m+3)l,3l}} |\nabla u|^2.$$

Proof The proof follows by covering $\mathcal{C}_{-(m+3)l,3l}$ by sub-half-cylinders of length $6l$, together with [Lemma 7.7](#). We refer to the proof of [\[6, Proposition B.19\]](#) for details. $\square$

The following simple lemma will be useful in the next subsection.

Lemma 7.9 *Suppose $u: \mathcal{C}_{-(m+3)l,3l} \rightarrow \mathcal{N}$ is a map satisfying*

$$\int_{\mathcal{C}_{-ml,0}} |u_\theta|^2 \leq \frac{1}{9} E(u) \quad \text{and} \quad \int_{\mathcal{C}_{-(m+3)l,-ml} \cup \mathcal{C}_{0,3l}} |\nabla u|^2 \leq \frac{1}{9} E(u).$$

Then

$$\text{Area}(u) \leq \frac{8}{9} E(u).$$

Proof First note that

$$\text{Area}(u) = \int_{\mathcal{C}_{-(m+3)l,3l}} (|u_\theta|^2 |u_t|^2 - \langle u_\theta, u_t \rangle^2)^{1/2} \leq \int_{\mathcal{C}_{-(m+3)l,3l}} |u_\theta| |u_t|.$$

By $ab \leq \frac{4}{3}a^2 + \frac{3}{16}b^2$, we get

$$\begin{aligned}
 (73) \quad \int_{C_{-ml,0}} |u_\theta| |u_t| &\leq \frac{4}{3} \int_{C_{-ml,0}} |u_\theta|^2 + \frac{3}{16} \int_{C_{-ml,0}} |u_t|^2 \\
 &\leq \frac{1}{3} \int_{C_{-ml,0}} |u_\theta|^2 + \frac{3}{16} \int_{C_{-ml,0}} |u_t|^2 + \frac{1}{9} E(u) \\
 &\leq \frac{7}{9} E(u).
 \end{aligned}$$

Combining with the second assumption, we have

$$(74) \quad \text{Area}(u) \leq \frac{7}{9} E(u) + \frac{1}{9} E(u) = E(u) - \frac{1}{9} E(u) \leq \frac{8}{9} E(u). \quad \square$$

7.4 Proof of Theorem 4.4

Now we are ready to prove Theorem 4.4. We will follow the same scheme as in [6, Section B.6]. One key point in our setting is that we may get two different kinds of bubbles. We may get spherical bubbles as in [6, Section B.6], as well as free boundary disk bubbles as [15]. The new techniques developed in previous sections will be essentially used to study these bubbles.

A boundary ball B of D is always the intersection of a classical ball $B_r(x)$ of $\mathbb{R}^2$ with D , ie $B = B_r(x) \cap D$ for some $x \in \partial D$ and $r > 0$. In the following proof we say r is the radius of B and sometime abuse notation to write $B = B_r(x)$.

We use $\tilde{\Pi}$ to denote a fixed conformal map that maps the upper half-plane $\overline{\mathbb{H}^2}$ to the unit disk $\bar{D}$ which maps $(0, 1, \infty)$ to three given distinct points on ∂D . We denote by p^+ the image of ∞ and by D_1^+ the image of $D_1^+ \subset \overline{\mathbb{H}^2}$ under the map Π . For a given boundary ball $B_r(x)$, we define the *conformal dilation of $B_r(x)$* to be the map $\Psi_{r,x}: \bar{D} \rightarrow \bar{D}$ such that $\Psi_{r,x} = \tilde{\Pi} \circ \Phi_{r,x} \circ \tilde{\Pi}^{-1}$, where $\Phi_{r,x}$ is the composition of a dilation of $\overline{\mathbb{H}^2}$ by the factor $\frac{1}{r}$ and a translation of $\overline{\mathbb{H}^2}$ by $-\tilde{\Pi}^{-1}(x)$. (Note $\tilde{\Pi}^{-1}(x)$ is a boundary point of $\overline{\mathbb{H}^2}$.)

Proof of Theorem 4.4 We divide the whole proof into two parts. The first part is about the bubbling compactness, and the second part is about the energy identity. Note that bubbling convergence with energy identity implies varifold convergence by [6, Proposition A.3]. (Even though in [6, Proposition A.3] the domain of the maps is the sphere, the proof works in our case for a disk domain with no change.)

Bubbling convergence Let u^j be a sequence as in the theorem. Then property $(\dagger)$ implies property (B_0) in Definition 7.1. By the compactness result Theorem 7.2,

we can find a free boundary harmonic map $v_0: (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ (which may be trivial), and a finite collection of singular points $\mathcal{S}_0 \subset \bar{D}$, such that a subsequence (still denoted by u^j) converges to v_0 weakly in $W^{1,2}(D)$ and strongly in $W^{1,2}(K)$ for any compact subset $K \subset \bar{D} \setminus \mathcal{S}_0$. The measures $|\nabla u^j|^2 dx$ converge to a measure ν_0 with $\nu_0(\bar{D}) \leq E_0$, and at each singular point $x \in \mathcal{S}_0$, we have $\nu_0(\{x\}) \geq \varepsilon_1$.

Next we want to renormalize the maps near the singular points. Let us start with boundary points. Suppose $x \in \mathcal{S}_0$ lies on ∂D . Let $\varepsilon_3 > 0$ be smaller than $\frac{1}{2}\varepsilon_1$ and ε_2 . Fix a radius $\rho > 0$ so that x is the only singular point in the boundary ball $B_{2\rho}(x)$ and $\int_{B_\rho(x)} |\nabla v_0|^2 \leq \varepsilon_3$. For each j , we choose $r_j > 0$ to be the smallest radius for which

$$\inf_{y \in B_{\rho-r_j}(x) \cap \partial D} \int_{B_\rho(x) \setminus B_{r_j}(y)} |\nabla u^j|^2 = \varepsilon_3,$$

and choose a point $y_j \in \partial D$ so that $B_{r_j}(y_j) \subset B_\rho(x)$ with $\int_{B_\rho(x) \setminus B_{r_j}(y_j)} |\nabla u^j|^2 = \varepsilon_3$. Since u_j converges strongly to v_0 on any compact subset of $B_\rho(x) \setminus \{x\}$, by the energy bound we get $y_j \rightarrow x$ and $r_j \rightarrow 0$.

For each j , since the energy functional is invariant under conformal changes, the dilated sequence of maps $\tilde{u}_1^j = u^j \circ \Psi_{r_j, y_j}$ still satisfies the almost harmonic property (B_0) in Definition 7.1, and they all have the same energy as the u^j . Using the compactness result Theorem 7.2 again, we get a subsequence (still denoted by $\tilde{u}_1^j$), and a finite collection of singular points $\mathcal{S}_1 \subset \bar{D}$, and a free boundary harmonic map $v_1: D \rightarrow \mathcal{N}$, such that $\tilde{u}_1^j$ converges to v_1 weakly in $W^{1,2}(D)$ and strongly in $W^{1,2}(K)$ for any compact $K \subset \bar{D} \setminus \mathcal{S}_1$. Moreover, the measures $|\nabla \tilde{u}_1^j|^2 dx$ converge to a measure ν_1 on $\bar{D}$.

The choice of $B_{r_j}(y_j)$ guarantees that

$$\nu_1(\bar{D} \setminus \{p^+\}) \leq \nu_0(\{x\}) \quad \text{and} \quad \nu_1(D^-) \leq \nu_0(\{x\}) - \varepsilon_3.$$

Next we want to show the following claim:

Claim *The maximal energy concentration at any $y \in \mathcal{S}_1 \setminus \{p^+\}$ is at most $\nu_0(\{x\}) - \varepsilon_3$.*

Proof of the claim Note that any such point y satisfies $\nu_1(\{y\}) \geq \varepsilon_1 > \varepsilon_3$, hence it can only lie in D^- . Then the fact $\nu_1(D^-) \leq \nu_0(\{x\}) - \varepsilon_3$ implies $\nu_1(\{y\}) \leq \nu_0(\{x\}) - \varepsilon_3$.

Now we iterate this blowing-up construction at every boundary singular point in $\mathcal{S}_0$ and $\mathcal{S}_1$, and we will get subsequent singular sets $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, \dots$, and dilated sequences of maps $\{u^j\}, \{\tilde{u}_1^j\}, \{\tilde{u}_2^j\}, \dots$, one more singular set after each blowing-up process.

From the claim we know that this process must terminate after at most E_0/ε_3 steps, and we have in total m singular sets $\mathcal{S}_0, \mathcal{S}_1, \dots, \mathcal{S}_m$. Then for each sequence of dilated maps $\tilde{u}_\alpha^j$, $\alpha \in \{0, \dots, m\}$, there are no boundary singular points away from $\mathcal{S}_\alpha$. Lastly we can apply the blowing-up process in [6, Appendix B.6] to each $\tilde{u}_\alpha^j$ at those interior singular points, and finally get that the sequence u^j converges to a collection of free boundary harmonic disks $v_0, v_1, \dots, v_m: (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ and harmonic spheres $\tilde{v}_1, \dots, \tilde{v}_k: S^2 \rightarrow \mathcal{N}$. Note that these harmonic spheres arise as blow-up limits near interior singular points.

Energy identity In this part we will show the summation of the energy of all the v_i is equal to the limit of the energy of the u^j , ie no energy was lost in the bubbling convergence process.

In order to prove this, we need to reexamine what happens to the energy during the blowing-up process. The “no loss of energy” for blowing up of interior singular points has already been proved by Colding–Minicozzi in [6, Appendix B.6, Step 4], so we only need to analyze the case for blowing up at boundary singular points.

At each blowing-up step, the energy is taken away from a singular point x and then goes to one of two places:

- it can show up in the new limiting free boundary harmonic disk of a singular point in $\partial D \setminus \{p^+\}$, or
- it can disappear at p^+ .

In the first case, the energy is accounted for in the final summation and no energy is lost. So we only need to rule out energy loss in the second case. With out loss of generality, we can only prove the “no loss of energy” for the first blow-up process, ie for the convergence of $\{\tilde{u}_1^j\}$. Note that if there is energy loss, then $\nu_1(\bar{D} \setminus \{p^+\}) < \nu_0(\{x\})$.

We argue by contradiction. Suppose $\nu_1(\bar{D} \setminus \{p^+\}) \leq \nu_0(\{x\}) - \hat{\delta}$ for some $\hat{\delta} > 0$. Note that we must have $\hat{\delta} \leq \varepsilon_3$. Thus we can choose two sequences of radii $s_j > t_j$ so that each $A_j = B_{s_j}(y_j) \setminus B_{t_j}(y_j)$ is a half-cylinder with

$$(75) \quad s_j \rightarrow 0, \quad \frac{t_j}{r_j} \rightarrow \infty, \quad \int_{A_j} |\nabla u^j|^2 \geq \hat{\delta} > 0.$$

Actually we may choose s_j close to ρ , and t_j close to s_j/λ_j , and $s_j/t_j > \lambda_j$ for a sequence $\lambda_j \rightarrow \infty$. After a conformal change, the A_j are a sequence of half-cylinders whose length goes to ∞ . Moreover, there is quite a bit of freedom in choosing s_j

and t_j , ie we can change s_j and t_j a little bit and the above conditions are still satisfied. So we may also assume u^j has small energy near the two ends of the half-cylinder A_j as the second condition in [Lemma 7.9](#).

Theorem 7.3 (with $\delta = \frac{1}{63}$) together with $\hat{\delta} \leq \varepsilon_3 \leq \varepsilon_2$ implies the theta energy of u^j on A_j is small, so the first condition of [Lemma 7.9](#) is satisfied for j sufficiently large. Then by [Lemma 7.9](#), we get that the area of the image of u^j on A_j must be strictly less than the energy of u^j on A_j for j large, which is a contradiction to the area assumption (37). Thus we complete the proof. $\square$

8 Modifications for the proof of [Theorem 0.3](#) and discussions

In this part, we record necessary modifications to adapt the proof of [Theorem 0.1](#) to [Theorem 0.3](#). There are only two places where we have to do some modifications.

Modification for [Theorem 4.1](#) Again, the first step is to show that we can approximate a given $\gamma(\cdot, t) \in \Omega$ by some $\tilde{\gamma}(\cdot, t) \in \Omega$ which lies in $C^0([0, 1], C^2(\bar{D}, \mathcal{N}))$. So we need to do mollifications on γ . However, direct mollifications as we did in [Section 5](#) cannot work here, because the endpoint maps $\gamma(\cdot, 0) = \bar{v}_0$ and $\gamma(\cdot, 1) = \bar{v}_1$ may change after mollifications. In order to handle this issue, we first mollify the whole family $\gamma(\cdot, t)$ as in [Section 5](#) to get a continuous family of C^2 maps $\bar{\gamma}(\cdot, t): (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$; next, by reparametrizing $t \rightarrow s(t) = (1 - 2\mu)t + \mu$, we get a new family $\tilde{\gamma}(\cdot, s)$ which is defined for $s \in [\mu, 1 - \mu]$. Moreover, since $s(t) \rightarrow t$ as $\mu \rightarrow 0$ and $\bar{\gamma}(\cdot, t)$ is C^0 as a function of t to $C^2(\bar{D}, \mathcal{N})$, for any given $\varepsilon > 0$, we can choose μ small so that $\max_{s \in [\mu, 1 - \mu]} \|\tilde{\gamma}(\cdot, s) - \bar{\gamma}(\cdot, s)\|_{C^2} \leq \varepsilon$.

Note that $\gamma(\cdot, 0)$ and $\gamma(\cdot, 1)$ are both smooth, so by varying the mollification parameter we can connect $\gamma(\cdot, 0)$ and $\gamma(\cdot, 1)$ to $\tilde{\gamma}(\cdot, \mu)$ and $\tilde{\gamma}(\cdot, 1 - \mu)$ respectively, and hence get a continuous family of C^2 -maps $\tilde{\gamma}(\cdot, s): (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ for $s \in [0, 1]$. The slices of $\tilde{\gamma}(\cdot, s)$ for $s \in [0, \mu]$ are mollifications of $\gamma(\cdot, 0)$ and the slices of $\tilde{\gamma}(\cdot, s)$ for $s \in [1 - \mu, 1]$ are mollifications of $\gamma(\cdot, 1)$. Thus we get a regularization of $\gamma(t)$ which stays close to $\gamma(t)$ in $C^0(\bar{D}, \mathcal{N}) \cap W^{1,2}(D, \mathcal{N})$. The conformal reparametrization procedure works in the same way as [Section 5](#). Note that since $\tilde{\gamma}(\cdot, 0) = \bar{v}_0$ and $\tilde{\gamma}(\cdot, 1) = \bar{v}_1$ are both conformal harmonic maps, the pull-back metrics of g on $\mathcal{N}$ are already conformal to the standard metric g_0 on D , so the conformal reparametrization maps $h(t)$ satisfy $h(0) = h(1) = \text{id}$. Therefore after the conformal reparametrization procedure, the endpoint maps are still $\bar{v}_0$ and $\bar{v}_1$, so it is a legitimate sweepout homotopic to γ . We know that for this family $\tilde{\gamma}(\cdot, s)$, the area and the energy are close.

Modification for Theorem 4.3 In the tightening process, we only want to pull tight the slice with energy larger than the energy of $\bar{v}_0$ and $\bar{v}_1$. Therefore, we need to modify the theorem so that in the statement, $(*)$ holds for those t with $E(\tilde{\gamma}(\cdot, t)) \geq W/\lambda$, where $\lambda > 1$ is chosen with $W/\lambda > \max(\text{Area}(\bar{v}_0), \text{Area}(\bar{v}_1))$. All the proofs in Section 6 work in exactly the same way by changing $\frac{1}{2}W$ to W/λ .

Modification for Theorem 4.4 In the bubbling convergence procedure, we have one additional assumption on the sequence u^j : that is, when restricted to ∂D , $u^j: \partial D \rightarrow \Gamma$ is monotone. This is because the monotonicity is preserved when taking the limit.

Now we show that there is only one possible disk bubble. In order to show this, we only need to argue that each blow-up on the boundary can only generate a punctured sphere rather than a harmonic disk. Let's say u_k is generated from blowing up a boundary point on D . Then by the classical Courant–Lebesgue lemma, the length of $u_k|_{\partial D}$ is shorter than $\frac{1}{3}$ of the total length of Γ . Since the limiting series after conformal dilation are all monotone on the boundary, u_k should also be monotone on the boundary, which implies that u_k maps the whole boundary to a single point, ie u_k is a punctured sphere.

Some further discussions One main goal of the min-max construction for the fixed boundary problem is to produce a third nonminimizing minimal disk spanning Γ , which is a direct generalization of the work of Morse–Tompkins [34] and Shiffman [49] (see also Struwe [53]) to the Riemannian setting. There is one issue left open in our current result. In fact, it will be good if one can restrict the sweepouts to all those $\sigma \in \Omega_f$ where $\sigma(\cdot, t): \partial D \rightarrow \Gamma$ is a monotone parametrization for each $t \in [0, 1]$; if this could be done, our proof will show that one of the disk bubbles has monotone boundary parametrization, and all other disk bubbles must map the boundary ∂D to a point on Γ , so that they are punctured harmonic spheres. Thus if we assume additionally that the ambient manifold $\mathcal{N}$ has nonpositive curvature, then by the uniqueness of harmonic maps all of these (punctured) harmonic spheres must be constant, so the min-max solution we obtain in Theorem 0.3 gives an effective construction of a third nonminimizing (unstable) minimal disk; cf Hohrein's PhD thesis [21] for the general existence result of unstable minimal surfaces of higher genus in manifolds with nonpositive curvature. Unfortunately our current mollification process could possibly destroy the monotonicity property of the boundary parametrization in general. It will also be good to reduce the regularity of the boundary curve Γ to be just rectifiable, but our theory needs a nice Fermi neighborhood of Γ , which requires smoothness.

Appendix

For the definition of Fermi coordinate system near $\Gamma \subset \mathcal{N}$ we refer to [31, Appendix A]. We will say a quantity A is α -close to quantity B if we have $1 - \alpha \leq A/B \leq 1 + \alpha$.

Lemma A.1 *There is a constant $\kappa > 0$ depending only on $\mathcal{N}$, Γ and the isometric embedding $\mathcal{N} \hookrightarrow \mathbb{R}^N$, such that for any $x \in \Gamma$, the κ -neighborhood U_κ of x has a local Fermi coordinate system, satisfying the following condition: let g^1 be the metric of $\mathcal{N}$ under Fermi coordinates, g^2 be the standard Euclidean metric under the Fermi coordinate system, and g^3 be the standard Euclidean metric of $\mathbb{R}^N$. Then there is a constant $\alpha > 0$ depending on κ such that:*

- (i) *For any vector $V \in T_p \mathcal{N} \subset T_p \mathbb{R}^N$ where $p \in U_\kappa$, $\|V\|_{g^k}$ is α -close to $\|V\|_{g^l}$ for $k, l \in \{1, 2, 3\}$.*
- (ii) *For any pair of points $p_1, p_2 \in U_\kappa$, $\|p_1 - p_2\|_{g^2}$ is α -close to $\|p_1 - p_2\|_{g^3}$.*

Proof We only need to show that κ exists for any $x \in \Gamma$, then by a covering argument we can prove κ exists globally. Fix $x \in \Gamma$.

(i) By the definition of Fermi coordinates, we can choose the Fermi coordinates so that $g_{ij}^1(x) = g_{ij}^2(x) = \delta_{ij}$. Then if κ is small enough, locally $(1 - \alpha) \text{id} \leq (g^1)^{-1} g^2 \leq (1 + \alpha) \text{id}$, so we get (i) for $k, l \in \{1, 2\}$. Since $\mathcal{N} \hookrightarrow \mathbb{R}^N$ is an isometric embedding, the lengths of V measured by g^1 and g^3 are close when κ is small enough, hence (i) is true.

(ii) First we show $\|p_1 - p_2\|_{g^2}$ is α -close to the distance between p_1, p_2 in $\mathcal{N}$. Note that for any curve connecting p_1 with p_2 in U_κ , the length of the curve evaluated under g^1 and g^2 should be α -close, since the lengths of the tangent vectors of the curve are α -close. Since the length between p_1 and p_2 is the shortest distance among all curves, we see that $\|p_1 - p_2\|_{g^2}$ is α -close to the distance between p_1 and p_2 on $\mathcal{N}$ (note that we can choose κ small so that U_κ is convex).

Since at each point x of $\mathcal{N}$ scaling up makes $\mathcal{N}$ converge to the tangent space at x , the distance for points close to x on $\mathcal{N}$ is equivalent to the distance in $\mathbb{R}^N$. Then apply this argument to all Fermi neighborhoods with small κ we get that $\|p_1 - p_2\|_{g^3}$ is α -close to the distance between p_1, p_2 in $\mathcal{N}$. It is easy to see that $\|p_1 - p_2\|_{g^3} \leq \text{dist}_{\mathcal{N}}(p_1, p_2)$, thus $\text{dist}_{\mathcal{N}}(p_1, p_2)$ is α -close to $\|p_1 - p_2\|_{g^3}$ for some α . This proves the lemma. $\square$

We also need the following Poincaré-type inequality with partial zero boundary values:

Lemma A.2 Given a function $f: D_r^+ \rightarrow \mathbb{R}$ with $f|_{\partial_r^A} = 0$, there exists a constant $C = C(r)$ such that

$$\int_{D_r^+} f^2 \leq C \int_{D_r^+} |\nabla f|^2.$$

Proof We just extend f to the whole of D_r by letting $f(x, -y) = f(x, y)$ in Cartesian coordinates. Then f is a function on D_r which vanishes on the boundary, hence we can apply the classical Poincaré inequality. Note that this reflection doubles $\int f^2$ and $\int |\nabla f|^2$ at the same time, so we get the desired Poincaré inequality. $\square$

A.1 Bubble convergence implies varifold convergence

The main goal of this section is to prove that bubble convergence implies varifold convergence; cf [6, Proposition A.3]. Let us recall some notions from geometric measure theory; for more details see [6, Section 1.3], and for further details see [50]. Let $\pi: G_2\mathcal{N} \rightarrow \mathcal{N}$ be the Grassmannian bundle of 2-planes over $\mathcal{N}$, and let us consider the pairs (X, F) where X is a compact surface (not necessarily connected) and $F: X \rightarrow G_2\mathcal{N}$ is a measurable map such that $f := \pi \circ F$ is in $W^{1,2}(X, \mathcal{N})$. We also use J_f to denote the Jacobian of f . We say that a sequence $X_i = (X_i, F_i)$ with uniformly bounded areas *varifold converges* to (X, F) if for all $h \in C^0(G_2\mathcal{N})$ we have

$$\int_{X_i} h \circ F_i J_{f_i} \rightarrow \int_X h \circ F J_f.$$

This is a kind of weak notion of convergence of measures. There exists a distance function d_V , the varifold distance, which induces this topology.

Here is one important example: a varifold induced by a map. Let $u: D \rightarrow \mathcal{N}$ be a $W^{1,2}$ -map, then a pair (X, F) induced by u is constructed as follows: X is just D , and $F: X \rightarrow G_2\mathcal{N}$ is given by sending x to $du(T_x X)$. This is only defined on the measurable space where J_u is nonzero, but we can extend it arbitrarily to all of X since the corresponding Radon measure on $G_2\mathcal{N}$ (ie the *varifold induced by u*) given by $h \mapsto \int_X h \circ F J_u$ is independent of the extension.

Proposition A.3 If a sequence $v^j: D \rightarrow \mathcal{N}$ of $W^{1,2}$ -maps bubble converges to a finite collection of smooth maps $\{u_0, \dots, u_m\}$ such that either $u_i: (D, \partial D) \rightarrow (\mathcal{N}, \Gamma)$ is a harmonic disk with free boundary or $u_i: S^2 \rightarrow \mathcal{N}$ is a harmonic sphere, and the energy identity holds, then this sequence also varifold converges to the varifold induced by $\{u_0, \dots, u_m\}$.

We want to emphasize that the energy equality, which appears in the last part of [Theorem 4.4](#), plays a key role in the proof.

Proof (see also [\[6, proof of Proposition A.3\]](#)) For each v^j we let V^j denote the corresponding map to $G_2\mathcal{N}$. Similarly for each u_i let U_i denote the corresponding map to $G_2\mathcal{N}$. By the construction of bubble convergence (in the proof of [Theorem 4.4](#)), we can choose $m+1$ sequences of domains $\Omega_0^j, \dots, \Omega_m^j \subset D$ that are pairwise disjoint for each j , so that for each $i = 0, \dots, m$, applying $\Psi_{i,j}^{-1}$ (the inverse of the corresponding conformal dilation) to Ω_i^j gives a sequence of domains converging to either $D \setminus S_i$ (if it is a disk bubble) or $S^2 \setminus S_i$ (if it is a sphere bubble), and they account for all the energy by the energy identity, ie

$$\lim_{j \rightarrow \infty} \int_{D \setminus (\cup_i \Omega_i^j)} |\nabla v^j|^2 = 0.$$

In order to show varifold convergence, we only need to show for any $h \in C^0(G_2\mathcal{N})$ that

$$\int_D h \circ U_i J_{u_i} = \lim_{j \rightarrow \infty} \int_{\Omega_i^j} h \circ V^j J_{v^j} = \lim_{j \rightarrow \infty} \int_{\Psi_{i,j}^{-1}(\Omega_i^j)} h \circ V^j \circ \Psi_{i,j} J_{(v^j \circ \Psi_{i,j})},$$

where the last equality is the change of variables formula.

Given $\varepsilon > 0$ and i , let Ω_ε^i be the set where $J_{u_i} \geq \varepsilon$. Then we only need to show

$$\int_{\Omega_\varepsilon^i} h \circ U_i J_{u_i} = \lim_{j \rightarrow \infty} \int_{\Psi_{i,j}^{-1}(\Omega_i^j)} h \circ V^j \circ \Psi_{i,j} J_{(v^j \circ \Psi_{i,j})}.$$

Note that $J_{v^j \circ \Psi_{i,j}} \rightarrow J_{u_i}$ in L^1 since $v_i^j \rightarrow u_i$ in $W^{1,2}$, so the measure of

$$\{x \in \Omega_\varepsilon^i : J_{v^j \circ \Psi_{i,j}} < \frac{\varepsilon}{2}\}$$

goes to zero; the $W^{1,2}$ -convergence implies that for given $\delta > 0$, the measure of

$$\{x \in \Omega_\varepsilon^i : J_{v^j \circ \Psi_{i,j}} \geq \frac{\varepsilon}{2}, |V^j \circ \Psi_{i,j} - U_i| \geq \delta\}$$

goes to zero. Then by the dominated convergence theorem we get the desired identity. Thus we conclude the varifold convergence. $\square$

References

- [1] **DR Adams**, [A note on Riesz potentials](#), Duke Math. J. 42 (1975) 765–778 [MR](#)

- [2] **F J Almgren, Jr**, *The homotopy groups of the integral cycle groups*, Topology 1 (1962) 257–299 [MR](#)
- [3] **F J Almgren, Jr**, *The theory of varifolds*, mimeographed notes, Princeton University (1965)
- [4] **J Chen, G Tian**, *Compactification of moduli space of harmonic mappings*, Comment. Math. Helv. 74 (1999) 201–237 [MR](#)
- [5] **R Coifman, P-L Lions, Y Meyer, S Semmes**, *Compensated compactness and Hardy spaces*, J. Math. Pures Appl. 72 (1993) 247–286 [MR](#)
- [6] **T H Colding, W P Minicozzi, II**, *Width and finite extinction time of Ricci flow*, Geom. Topol. 12 (2008) 2537–2586 [MR](#)
- [7] **T H Colding, W P Minicozzi, II**, *Width and mean curvature flow*, Geom. Topol. 12 (2008) 2517–2535 [MR](#)
- [8] **R Courant**, *Dirichlet's principle, conformal mapping, and minimal surfaces*, Springer (1977) [MR](#)
- [9] **R Courant, N Davids**, *Minimal surfaces spanning closed manifolds*, Proc. Nat. Acad. Sci. U.S.A. 26 (1940) 194–199 [MR](#)
- [10] **C De Lellis, J Ramic**, *Min-max theory for minimal hypersurfaces with boundary*, Ann. Inst. Fourier (Grenoble) 68 (2018) 1909–1986 [MR](#)
- [11] **W Ding, J Li, Q Liu**, *Evolution of minimal torus in Riemannian manifolds*, Invent. Math. 165 (2006) 225–242 [MR](#)
- [12] **W Ding, G Tian**, *Energy identity for a class of approximate harmonic maps from surfaces*, Comm. Anal. Geom. 3 (1995) 543–554 [MR](#)
- [13] **L C Evans**, *Partial differential equations*, 2nd edition, Graduate Studies in Mathematics 19, Amer. Math. Soc., Providence, RI (2010) [MR](#)
- [14] **A M Fraser**, *On the free boundary variational problem for minimal disks*, PhD thesis, Stanford University (1998) [MR](#) Available at <https://search.proquest.com/docview/3044448599>
- [15] **A M Fraser**, *On the free boundary variational problem for minimal disks*, Comm. Pure Appl. Math. 53 (2000) 931–971 [MR](#)
- [16] **A Fraser, R Schoen**, *The first Steklov eigenvalue, conformal geometry, and minimal surfaces*, Adv. Math. 226 (2011) 4011–4030 [MR](#)
- [17] **A Fraser, R Schoen**, *Sharp eigenvalue bounds and minimal surfaces in the ball*, Invent. Math. 203 (2016) 823–890 [MR](#)
- [18] **L Grafakos**, *Modern Fourier analysis*, 2nd edition, Graduate Texts in Mathematics 250, Springer (2009) [MR](#)
- [19] **M Grüter, J Jost**, *On embedded minimal disks in convex bodies*, Ann. Inst. H. Poincaré Anal. Non Linéaire 3 (1986) 345–390 [MR](#)

- [20] **F Hélein**, *Harmonic maps, conservation laws and moving frames*, 2nd edition, Cambridge Tracts in Mathematics 150, Cambridge Univ. Press (2002) [MR](#)
- [21] **J Hohrein**, *Existence of unstable minimal surfaces of higher genus in manifolds of nonpositive curvature*, PhD thesis, Universität Heidelberg (1994)
- [22] **R A Hunt**, *On $L(p, q)$ spaces*, Enseign. Math. 12 (1966) 249–276 [MR](#)
- [23] **T Iwaniec, G Martin**, *Geometric function theory and non-linear analysis*, Oxford Univ. Press (2001) [MR](#)
- [24] **J Jost**, *Existence results for embedded minimal surfaces of controlled topological type, II*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. 13 (1986) 401–426 [MR](#)
- [25] **J Jost**, *Two-dimensional geometric variational problems*, Wiley, Chichester (1991) [MR](#)
- [26] **J Jost, L Liu, M Zhu**, *The qualitative behavior at the free boundary for approximate harmonic maps from surfaces*, Math. Ann. 374 (2019) 133–177 [MR](#)
- [27] **T Lamm, L Lin**, *Estimates for the energy density of critical points of a class of conformally invariant variational problems*, Adv. Calc. Var. 6 (2013) 391–413 [MR](#)
- [28] **P Laurain**, *Analyse des problèmes conformément invariants*, from “Séminaire Laurent Schwartz–Équations aux dérivées partielles et applications, 2016–2017” (F Golse, F Merle, editors), Ed. Éc. Polytech., Palaiseau (2017) Exposé XII [MR](#)
- [29] **P Laurain, L Lin**, *Energy convexity of intrinsic bi-harmonic maps and applications, I: Spherical target*, preprint (2018) [arXiv](#)
- [30] **P Laurain, R Petrides**, *Regularity and quantification for harmonic maps with free boundary*, Adv. Calc. Var. 10 (2017) 69–82 [MR](#)
- [31] **M Li, X Zhou**, *Min-max theory for free boundary minimal hypersurfaces, I: Regularity theory*, preprint (2016) [arXiv](#)
- [32] **F C Marques, A Neves**, *Min-max theory of minimal surfaces and applications*, from “Mathematical Congress of the Americas” (J A de la Peña, J A López-Mimbela, M Nakamura, J Petean, editors), Contemp. Math. 656, Amer. Math. Soc., Providence, RI (2016) 13–25 [MR](#)
- [33] **R Montezuma**, *A mountain pass theorem for minimal hypersurfaces with fixed boundary*, preprint (2018) [arXiv](#)
- [34] **M Morse, C Tompkins**, *The existence of minimal surfaces of general critical types*, Ann. of Math. 40 (1939) 443–472 [MR](#)
- [35] **F Müller, A Schikorra**, *Boundary regularity via Uhlenbeck–Rivière decomposition*, Analysis (Munich) 29 (2009) 199–220 [MR](#)
- [36] **T H Parker**, *Bubble tree convergence for harmonic maps*, J. Differential Geom. 44 (1996) 595–633 [MR](#)
- [37] **T H Parker, J G Wolfson**, *Pseudo-holomorphic maps and bubble trees*, J. Geom. Anal. 3 (1993) 63–98 [MR](#)

- [38] **S Poornima**, *An embedding theorem for the Sobolev space $W^{1,1}$* , Bull. Sci. Math. 107 (1983) 253–259 [MR](#)
- [39] **J Qing**, *Boundary regularity of weakly harmonic maps from surfaces*, J. Funct. Anal. 114 (1993) 458–466 [MR](#)
- [40] **J Qing, G Tian**, *Bubbling of the heat flows for harmonic maps from surfaces*, Comm. Pure Appl. Math. 50 (1997) 295–310 [MR](#)
- [41] **T Rivière**, *Conservation laws for conformally invariant variational problems*, Invent. Math. 168 (2007) 1–22 [MR](#)
- [42] **T Rivière**, *Conformally invariant variational problems*, preprint (2012) [arXiv](#)
- [43] **T Rivière**, *A viscosity method in the min-max theory of minimal surfaces*, Publ. Math. Inst. Hautes Études Sci. 126 (2017) 177–246 [MR](#)
- [44] **T Rivière, M Struwe**, *Partial regularity for harmonic maps and related problems*, Comm. Pure Appl. Math. 61 (2008) 451–463 [MR](#)
- [45] **J Sacks, K Uhlenbeck**, *The existence of minimal immersions of 2-spheres*, Ann. of Math. 113 (1981) 1–24 [MR](#)
- [46] **C Scheven**, *Partial regularity for stationary harmonic maps at a free boundary*, Math. Z. 253 (2006) 135–157 [MR](#)
- [47] **B Sharp**, *Higher integrability for solutions to a system of critical elliptic PDE*, Methods Appl. Anal. 21 (2014) 221–240 [MR](#)
- [48] **B Sharp, M Zhu**, *Regularity at the free boundary for Dirac-harmonic maps from surfaces*, Calc. Var. Partial Differential Equations 55 (2016) art. id. 27 [MR](#)
- [49] **M Shiffman**, *The Plateau problem for non-relative minima*, Ann. of Math. 40 (1939) 834–854 [MR](#)
- [50] **L Simon**, *Lectures on geometric measure theory*, Proceedings of the Centre for Mathematical Analysis, Australian National University 3, Aust. National Univ. Centre for Mathematical Analysis, Canberra (1983) [MR](#)
- [51] **B Smyth**, *Stationary minimal surfaces with boundary on a simplex*, Invent. Math. 76 (1984) 411–420 [MR](#)
- [52] **M Struwe**, *On a free boundary problem for minimal surfaces*, Invent. Math. 75 (1984) 547–560 [MR](#)
- [53] **M Struwe**, *Plateau's problem and the calculus of variations*, Mathematical Notes 35, Princeton Univ. Press (1988) [MR](#)
- [54] **M Struwe**, *Variational methods: applications to nonlinear partial differential equations and Hamiltonian systems*, 4th edition, Ergeb. Math. Grenzgeb. 34, Springer (2008) [MR](#)
- [55] **L Tartar**, *Imbedding theorems of Sobolev spaces into Lorentz spaces*, Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. 1 (1998) 479–500 [MR](#)

- [56] **H C Wente**, *An existence theorem for surfaces of constant mean curvature*, J. Math. Anal. Appl. 26 (1969) 318–344 [MR](#)
- [57] **R Ye**, *On the existence of area-minimizing surfaces with free boundary*, Math. Z. 206 (1991) 321–331 [MR](#)
- [58] **X Zhou**, *On the existence of min-max minimal torus*, J. Geom. Anal. 20 (2010) 1026–1055 [MR](#)
- [59] **X Zhou**, *On the free boundary min-max geodesics*, Int. Math. Res. Not. 2016 (2016) 1447–1466 [MR](#)
- [60] **X Zhou**, *On the existence of min-max minimal surface of genus $g \geq 2$* , Commun. Contemp. Math. 19 (2017) art. id. 1750041 [MR](#)

*Mathematics Department, University of California, Santa Cruz
Santa Cruz, CA, United States*

*Department of Mathematics, Massachusetts Institute of Technology
Cambridge, MA, United States*

*Department of Mathematics, University of California, Santa Barbara
Santa Barbara, CA, United States*

lzlin@ucsc.edu, aosun@mit.edu, zhou@math.ucsb.edu

Proposed: Tobias H Colding

Received: 28 November 2018

Seconded: Gang Tian, David Gabai

Accepted: 18 July 2019

Guidelines for Authors

Submitting a paper to Geometry & Topology

Papers must be submitted using the upload page at the [GT website](#). You will need to choose a suitable editor from the list of editors' interests and to supply MSC codes.

The normal language used by the journal is English. Articles written in other languages are acceptable, provided your chosen editor is comfortable with the language and you supply an additional English version of the abstract.

Preparing your article for Geometry & Topology

At the time of submission you need only supply a PDF file. Once accepted for publication, the paper must be supplied in $\LaTeX$, preferably using the journal's class file. More information on preparing articles in $\LaTeX$ for publication in GT is available on the [GT website](#).

arXiv papers

If your paper has previously been deposited on the arXiv, we will need its arXiv number at acceptance time. This allows us to deposit the DOI of the published version on the paper's arXiv page.

References

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited at least once in the text. Use of Bib $\TeX$ is preferred but not required. Any bibliographical citation style may be used, but will be converted to the house style (see a current issue for examples).

Figures

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Fuzzy or sloppily drawn figures will not be accepted. For labeling figure elements consider the [pinlabel](#) $\LaTeX$ package, but other methods are fine if the result is editable. If you're not sure whether your figures are acceptable, check with production by sending an email to graphics@msp.org.

Proofs

Page proofs will be made available to authors (or to the designated corresponding author) in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

GEOMETRY & TOPOLOGY

Volume 24

Issue 1 (pages 1–532)

2020

Coalgebraic formal curve spectra and spectral jet spaces	1
ERIC PETERSON	
Sutured manifolds and polynomial invariants from higher rank bundles	49
ALIAKBAR DAEMI and YI XIE	
Contact handles, duality, and sutured Floer homology	179
ANDRÁS JUHÁSZ and IAN ZEMKE	
Conical metrics on Riemann surfaces, I: The compactified configuration space and regularity	309
RAFE MAZZEO and XUWEN ZHU	
$GL_2\mathbb{R}$ –invariant measures in marked strata: generic marked points, Earle–Kra for strata and illumination	373
PAUL APISA	
Recognition of being fibered for compact 3–manifolds	409
ANDREI JAIKIN-ZAPIRAIN	
Edge stabilization in the homology of graph braid groups	421
BYUNG HEE AN, GABRIEL C DRUMMOND-COLE and BEN KNUDSEN	
Min-max minimal disks with free boundary in Riemannian manifolds	471
LONGZHI LIN, AO SUN and XIN ZHOU	