Pacific Journal of Mathematics

Volume 231 No. 2 June 2007

PACIFIC JOURNAL OF MATHEMATICS

msp.org/pjm

Founded in 1951 by E. F. Beckenbach (1906–1982) and F. Wolf (1904–1989)

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

Regular subscription rate for 2007: $425.00 a year (10 issues). Special rate: $212.50 a year to individual members of supporting institutions.

Subscriptions, requests for back issues and changes of subscribers address should be sent to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94704-0163, U.S.A. The Pacific Journal of Mathematics is indexed by Mathematical Reviews, Zentralblatt MATH, PASCAL CNRS Index, Referativnyi Zhurnal, Current Mathematical Publications and Web of Knowledge (Science Citation Index).

The Pacific Journal of Mathematics (ISSN 0030-8730) at the University of California, c/o Department of Mathematics, 798 Evans Hall #3840, Berkeley, CA 94720-3840, is published twelve times a year. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices. POSTMASTER: send address changes to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94704-0163.

PJM peer review and production are managed by EditFLOW® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.257

THE EUCLIDEAN RANK OF HILBERT GEOMETRIES

OLIVER BLETZ-SIEBERT AND THOMAS FOERTSCH

We prove that the Euclidean rank of any 3-dimensional Hilbert geometry (D, h_D) is 1; that is, (D, h_D) does not admit an isometric embedding of the Euclidean plane. We show that for higher dimensions this remains true if the boundary ∂D of D is C^1.

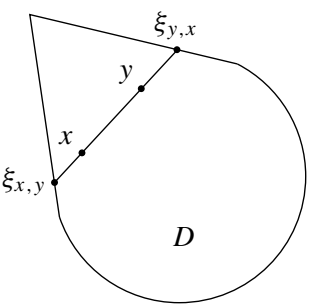
1. Introduction

In the late nineteenth century D. Hilbert informed F. Klein in a letter about his discovery of a method to construct metric spaces, generalizing Klein's model of the real hyperbolic space [Hilbert 1895]:

Let $\mathbb{E}^n$ denote the n-dimensional Euclidean space. The Euclidean distance of $x, y \in \mathbb{E}^n$ is written $|xy|$, the line segment between x and y is $[x, y]$, whereas $L(x, y) = \langle x, y\rangle$ denotes the whole affine line through x and y in $\mathbb{E}^n$.

Given an open bounded convex domain $D \subset \mathbb{E}^n$ with boundary $\partial D \subset \mathbb{E}^n$, the Hilbert metric $h_D : D \times D \to \mathbb{R}_0^+$ is defined via the cross-ratio: Given distinct points $x, y \in D$, take the points $\xi_{x,y}, \xi_{y,x}$ where $L(x, y)$ intersects ∂D and set

$$h_D(x, y) = \log \frac{|y\xi_{x,y}|\,|x\xi_{y,x}|}{|x\xi_{x,y}|\,|y\xi_{y,x}|}.$$



Formally, $\xi_{x,y} \in \langle x, y\rangle \cap \partial D$ is uniquely determined by the condition $|\xi_{x,y}x| < |\xi_{x,y}y|$. The cross ratio, of course, is invariant under projective transformations. For the basic properties of h_D see [Busemann 1955] and [de la Harpe 1993]; for example, the topology induced by h_D on D coincides with the subspace topology inherited from $\mathbb{E}^n$. We refer to the metric space (D, h_D) as a *Hilbert geometry*.

Hilbert proved that the straight line segments in (D, h_D) indeed are geodesics. In general, however, other geodesic segments also exist (see Lemma 2.1).

Both authors were supported by the Deutsche Forschungsgemeinschaft, grants BL 581/1-1 (Bletz-Siebert) and FO 353/1-1 (Foertsch).
MSC2000: 53C60, 51K99, 51F99.
Keywords: Hilbert geometries, Euclidean rank, asymptotic geometry.

It is natural to ask what metric spaces can be realized as Hilbert geometries. Hilbert himself pointed out that his construction, applied to an open n-dimensional Euclidean ball (or ellipsoid), yields Klein's model of n-dimensional real hyperbolic space.

Kelly and Strauss [1958] have proved that these are the only globally nonpositively Busemann curved Hilbert geometries, and therefore also the only Hilbert geometries which are CAT(0). (For these notions see [Bridson and Haefliger 1999; Jost 1997], for instance.)

More recently, Karlsson and Noskov [2002] showed that, nevertheless, there exist many Gromov-hyperbolic Hilbert geometries. (A geodesic metric space is called δ-hyperbolic, for a given $\delta \geq 0$, if each side of any geodesic triangle is contained in the union of the δ-neighborhoods of the other two sides; a δ-hyperbolic space is also called Gromov-hyperbolic.) Benoist [2003] then presented a necessary and sufficient condition for a Hilbert geometry to be Gromov-hyperbolic in terms of the boundary ∂D of D. Another characterization in terms of the spectrum of the Laplacian of (D, h_D) was recently obtained in [Colbois and Vernicos 2006].

A necessary — but by far not sufficient — condition for the Hilbert geometry on D to be Gromov-hyperbolic is that the boundary ∂D be C^1 and that $\overline{D}$ be strictly convex (see [Karlsson and Noskov 2002] or [Benoist 2003]). Thus there exist plenty of examples of Hilbert geometries that are *not* Gromov-hyperbolic.

Nonetheless, these examples still show certain "hyperbolic" features. This raises the question: *How close to being hyperbolic are Hilbert geometries in general?*

One particularly interesting class of non-Gromov-hyperbolic Hilbert geometries are those defined on the interior of simplices (the convex hull of $n+1$ points in $\mathbb{E}^n$ in general position). Although such geometries have been studied in detail for decades — see [Phadke 1974/75; Busemann and Phadke 1987], for example — it was not until much later that de la Harpe [1993] proved that, surprisingly, Hilbert geometries defined on the interior of simplices are isometric to normed vector spaces (see also [Nussbaum 1988]). In [Foertsch and Karlsson 2005] it was proved that these normed vector spaces are the only ones that can be realized as Hilbert geometries.

Whereas the papers mentioned above were concerned with the question of which metric spaces admit a realization as a Hilbert geometry, here we will be interested, more generally, in isometric embeddings into Hilbert geometries. This includes Hilbert geometries that are not uniquely geodesic, in particular those arising from nonstrictly convex domains — these cases are, in our view, even more interesting.

(Recall that a map $f : (X, d_X) \to (Y, d_Y)$ between metric spaces (X, d_X) and (Y, d_Y) is a *quasiisometric embedding* if there exist $\lambda \geq 1$ and $k \geq 0$ such that

$$\text{(1-1)} \quad \frac{1}{\lambda} d_X(x, x') - k \leq d_Y\big(f(x), f(x')\big) \leq \lambda\, d_X(x, x') + k \quad \text{for any } x, x' \in X.$$

If, moreover, $f(X)$ is a k-net in (Y, d_Y) (that is, Y is the k-neighborhood of $f(X)$), then f is called a quasiisometry. Recall further that if the inequalities in (1-1) hold with $\lambda = 1$, then f is called a rough-isometric embedding.)

This article is somehow motivated by the following result, which we quote from [Bridson and Haefliger 1999, Theorem III.H 1.9]: *There is no (quasi)isometric embedding of the Euclidean plane in a geodesic Gromov-hyperbolic metric space.*

Obviously, since every n-dimensional normed vector space is bilipschitz to $\mathbb{E}^n$, the Hilbert geometry defined on the interior of any simplex in $\mathbb{E}^n$ is quasiisometric to $\mathbb{E}^n$. However, as we will show, at least every 3-dimensional Hilbert geometry does not admit an isometric embedding of the Euclidean plane. To state the corresponding theorem, recall that the Euclidean rank, $\operatorname{rank}_E(X, d)$, of a metric space (X, d) is defined to be the supremum over all dimensions of Euclidean spaces that admit isometric embeddings into (X, d). Since affine line segments in a Hilbert geometry are geodesics, any Hilbert geometry has Euclidean rank at least 1.

Towards answering the question of how close to hyperbolic spaces a Hilbert geometry is, we will prove that Hilbert geometries have Euclidean rank 1 — at least in the cases mentioned in the abstract (though we believe this holds true for all cases):

Theorem 1.1. *Let (D, h_D) be a 3-dimensional Hilbert geometry. Then*

$$\operatorname{rank}_E(D, h_D) = 1.$$

We believe that Theorem 1.1 generalizes to arbitrary dimensions. For higher dimensions we can prove:

Theorem 1.2. *Let (D, h_D) be an n-dimensional Hilbert geometry. Then*

$$\operatorname{rank}_E(D, h_D) = 1$$

if ∂D does not contain a line segment, or if ∂D is C^1.

The main idea of the proof is to compare the Gromov products in $\mathbb{E}^2$ with those in the Hilbert geometry: in $\mathbb{E}^2$ the Gromov product of pairs of points on two geodesic rays — emanating from a fixed base point on different affine lines — is unbounded if the points tend (on the rays) to infinity. Now consider an isometric embedding of $\mathbb{E}^2$ into a Hilbert geometry and the geodesic rays which are the images of the two rays in $\mathbb{E}^2$. By a result of Karlsson and Noskov (Theorem 2.9), the endpoints of these geodesic rays span an affine line segment in the border of D (that is, this line segment does not intersect with D itself). Since this holds for all rays in $\mathbb{E}^2$ which are not collinear, this gives a lot of information about ∂D. The main difficulty is that the endpoints of geodesic rays do not depend continuously on the rays themselves (but compare Lemma 2.6). Therefore, there are many configurations to consider for

how the endpoints may be spread over ∂D. In dimension 3, we are able to exploit the facts mentioned to prove Theorem 1.1. In higher dimensions the geometry is much richer, and the variety of imaginable geometric configurations is much broader; this has kept us from solving the problem in higher dimensions. But we get Theorem 1.2 as a byproduct of our considerations.

We do not know whether there exist Hilbert geometries that admit isometric embeddings of open subsets of $\mathbb{E}^2$. To address this problem, the asymptotic methods used in this paper will not be of any help. However, an advantage of the asymptotic methods presented here is that they might even yield a stronger nonembedding result: To us it seems likely that Theorem 1.1 can be sharpened to prove that there does not exist even a rough-isometric embedding of the Euclidean plane into a Hilbert geometry. But it seems that new ideas are necessary to obtain such a result.

2. Preliminaries

A *geodesic segment* in a metric space X is an isometric embedding $\gamma : I \to X$ of an interval $I \subset \mathbb{E}^1$ into X. If $I = [0, \infty)$, we also call γ a *geodesic ray* (originating at $\gamma(0)$). The images of such maps γ can also be called geodesic segments and rays. A metric space X is *geodesic* if for every $x, y \in X$ there exists a geodesic segment connecting x to y, i.e., a geodesic $\gamma : [a, b] \to X$ such that $\gamma(a) = x$ and $\gamma(b) = y$. If for every $x, y \in X$ in a geodesic metric space X the image of a geodesic segment connecting x to y is unique, X is called *uniquely geodesic*.

Let D be an open bounded convex domain in $\mathbb{E}^n$. The closure $\overline{D}$ is also convex, and every affine line containing a point of D intersects ∂D in exactly two points. For distinct $a, b \in \partial D$, the open Euclidean segment from a to b, denoted $[a, b]^\circ := [a, b] - \{a, b\}$, lies either entirely in D or entirely in ∂D.

Convex hulls in Hilbert geometries. As mentioned in the introduction, Hilbert geometries are geodesic spaces, but are not, in general, uniquely geodesic. The following well-known lemma states this more precisely. In it, we use the notation $[x, z]_d$ for the *metric convex hull* of two points x and z in a metric space (X, d):

$$[x, z]_d = \big\{ y \in X \;\big|\; d(x, z) = d(x, y) + d(y, z) \big\}.$$

Lemma 2.1 [de la Harpe 1993, Proposition 2]. *Let (D, h_D) be a Hilbert geometry and let $x, z \in D$ be distinct. Then*

$$[x, z]_{h_D} = \big\{ y \in D \;\big|\; [\xi_{z,y}, \xi_{y,x}] \cup [\xi_{x,y}, \xi_{y,z}] \subset \partial D \big\};$$

in other words, there exists $y \in D \setminus \langle x, z \rangle$ such that $h_D(x, z) = h_D(x, y) + h_D(y, z)$ if and only if there exist open line segments $I, J \subset \partial D$ spanning an affine 2-plane in $\mathbb{E}^n$ and such that $\xi_{x,z} \in I$, $\xi_{z,x} \in J$.

In particular, if D is strictly convex (that is, if every affine line intersecting ∂D in at least two points also intersects D), the Hilbert geometry (D, h_D) is uniquely geodesic.

If Σ is an affine subspace of $\mathbb{E}^n$ intersecting ∂D but not D, then $F = \Sigma \cap \overline{D} = \Sigma \cap \partial D$ will be called a *boundary flat* of the Hilbert geometry (D, h_D); the border of F with respect to the subspace topology in Σ will be denoted by ∂F and its (relative) interior by $F^\circ = F \setminus \partial F$. A boundary flat is a convex set in $\mathbb{E}^n$.

We make some simple remarks on convexity, to be used several times later.

Remark 2.2. **(a)** ***Continuity of the boundary projection:*** Consider the map

$$(x, y) \mapsto (\xi_{x,y}, \xi_{y,x})$$

defined on $\{(x, y) \in D \times D \mid x \neq y\}$. Extend it to $\{(x, y) \in \overline{D} \times \overline{D} \mid x \neq y\}$ by setting $\xi_{x,y}$ and $\xi_{y,x}$ such that $[\xi_{x,y}, \xi_{y,x}] = \langle x, y\rangle \cap \overline{D}$ with $|x\xi_{x,y}| < |y\xi_{x,y}|$ and $|y\xi_{y,x}| < |x\xi_{y,x}|$. Then $(x, y) \mapsto (\xi_{x,y}, \xi_{y,x})$ is continuous on the set

$$\left\{(x, y) \in \overline{D} \times \overline{D} \;\middle|\; x \neq y \text{ and } [x, y] \not\subset \partial D\right\};$$

to be more precise, the maximal set on which the map is continuous consists exactly of the pairs of distinct points of $\overline{D}$ for which there is no line segment in the boundary of D containing both of them and even one of them in its relative interior.

(b) ***Edges converge to edges:*** If $\{p_n\}_{n\in\mathbb{N}}, \{q_n\}_{n\in\mathbb{N}}$ are sequences in ∂D converging to $p, q \in \partial D$ and such that $[p_n, q_n] \subset \partial D$ for all $n \in \mathbb{N}$, then $[p, q] \subset \partial D$.

(c) ***Triangles on the boundary:*** Assume that $a, b, c \in \partial D$ are not collinear, and denote by $\Delta(a, b, c)$ the closed affine triangle with vertices a, b and c, and by $\Delta^\circ(a, b, c) = \Delta(a, b, c) \setminus \big([a, b] \cup [a, c] \cup [b, c]\big)$ the relative interior of $\Delta(a, b, c)$. Then either $\Delta(a, b, c) \subset \partial D$ or $\Delta(a, b, c)^\circ \subset D$.

Geodesic rays in Hilbert geometries. Now we turn to geodesic rays in a Hilbert geometry (D, h_D). As one would expect, such rays do converge at infinity:

Theorem 2.3 [Foertsch and Karlsson 2005, Theorem 3]. *Let (D, h_D) be a Hilbert geometry.*

(i) *Every geodesic ray r in (D, h_D) converges to a point in ∂D, written $r(\infty)$.*

(ii) *Every complete geodesic γ in (D, h_D) has precisely two accumulation points $\gamma(\infty)$ and $\gamma(-\infty)$ in ∂D.*

Remark 2.4. Let r and $\tilde{r}$ be geodesic rays in a Hilbert geometry (D, h_D) such that $h_D(r(t), \tilde{r}(t))$ is a bounded function in t. If $r(\infty)$ lies in the (relative) interior of a boundary flat, the same is true of $\tilde{r}(\infty)$.

Lemma 2.5. *Let (D, h_D) be an n-dimensional Hilbert geometry and let*

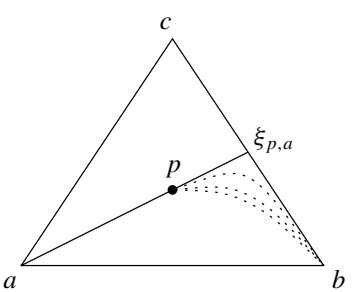
$$\gamma : (-\infty, \infty) \to D$$

be a geodesic in (D, h_D). Suppose that for $t_1 < t_2$ the straight line $L(\gamma(t_1), \gamma(t_2))$ intersects ∂D in the interior of two $(n-1)$-dimensional boundary flats. Then for all $t_0 < t_1$ there exists a neighborhood U of $\gamma(t_2)$ such that each point $p \in U$ lies on a geodesic through $\gamma(t_0)$ and $\gamma(t_1)$.

The lemma immediately implies that two points satisfying the condition on $\gamma(t_1)$ and $\gamma(t_2)$ cannot occur as image points of an isometric embedding $\varphi : \mathbb{E}^2 \to (D, h_D)$. For suppose $\varphi : \mathbb{E}^2 \to (D, h_D)$ were an isometric embedding and let γ denote the image of a straight line in $\mathbb{E}^2$ under φ, with $\varphi(x_1) = \gamma(t_1)$ and $\varphi(x_2) = \gamma(t_2)$. Then for any x_0 on the straight line through x_1 and x_2 in $\mathbb{E}^2$ satisfying $|x_0x_1| < |x_0x_2|$ we have $\varphi(x_0) = \gamma(t_0)$ for some $t_0 < t_1$. Now for any given neighborhood U of $\gamma(t_2)$ in (D, h_D), sufficiently small neighborhoods of x_2 in $\mathbb{E}^2$ are mapped via φ into U. This, however, is not possible for the neighborhood U as in Lemma 2.5, since $h_D(\gamma(t_0), \gamma(t_1)) + h_D(\gamma(t_1), u) = h_D(\gamma(t_0), u)$ for all $u \in U$, whereas in any arbitrarily small neighborhood V of x_2 in $\mathbb{E}^2$ there exists $v \in V$ with $|x_0x_1| + |x_1v| > |x_0v|$. This yields the desired contradiction.

Proof of Lemma 2.5. Fix $t_0 < t_1$. Then, by Lemma 2.1, $[\xi_{\gamma(t_1),\gamma(t_0)}, \xi_{\gamma(t_2),\gamma(t_1)}] \subset \partial D$ and $[\xi_{\gamma(t_0),\gamma(t_1)}, \xi_{\gamma(t_1),\gamma(t_2)}] \subset \partial D$. Since $\xi_{\gamma(t_1),\gamma(t_2)}$ and $\xi_{\gamma(t_2),\gamma(t_1)}$ lie in the interior of $(n-1)$-dimensional boundary flats F_1 and F_2, there exists a neighborhood U of $\gamma(t_2)$ in D such that for all $p \in U$ the projection $\xi_{\gamma(t_1),p}$ lies in the interior of F_1 and $\xi_{p,\gamma(t_1)}$ lies in the interior of F_2. The claim follows from Lemma 2.1. □

Consider sequences of geodesic rays r_i, $i \in \mathbb{N}$, in the Hilbert geometry (D, h_D) converging pointwise to some geodesic ray r in (D, h_D). In general their limit points $r_i(\infty)$ do not converge to $r(\infty)$! This can be observed in the Hilbert geometry $(\Delta^\circ, h_{\Delta^\circ})$ defined inside an open triangle $\Delta^\circ = \Delta^\circ(a, b, c)$ in $\mathbb{E}^2$ with vertices a, b and c. Fix $p \in \Delta^\circ$. Then $[p, \xi_{p,a}]$ is a geodesic ray converging to $\xi_{p,a} \in [b, c]^\circ$, yet $[p, \xi_{p,a}]$ is the pointwise limit of geodesic rays r_i in $(\Delta^\circ, h_{\Delta^\circ})$, which all converge to b.



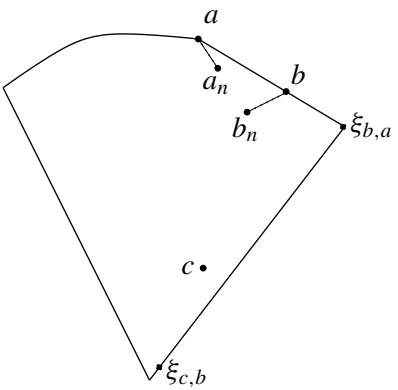
The Hilbert geometry of a 2-simplex is not uniquely geodesic.

The next lemma and its corollary investigate such behavior to some extent.

Lemma 2.6 (Jump Lemma). *Let (D, h_D) be a Hilbert geometry. Suppose $\{a_n\}_n$, $\{b_n\}_n$ and $\{c_n\}_n$ are sequences in D converging to distinct points $a, b \in \partial D$ and $c \in \overline{D}$, respectively, such that*

$$h_D(a_n, c_n) = h_D(a_n, b_n) + h_D(b_n, c_n) \quad \textit{for all } n \in \mathbb{N}.$$

(I) *If* $c \in D$, *then* $[a, b] \subsetneq [a, \xi_{b,a}] \subset \partial D$ *and* $[\xi_{b,a}, \xi_{c,b}] \subset \partial D$. *In particular,* $c \notin \langle a, b\rangle$ *and* $b \neq \xi_{b,a}$. (See figure.)



(II) *If* $c \in \partial D$, *one of the following* (*nondisjoint*) *cases occurs*:

(i) $[a, b] \cup [b, c] \subset \partial D$.

(ii) $[a, b] \subsetneq [a, \xi_{b,a}] \subset \partial D$ *and* $[c, \xi_{b,a}] \subset \partial D$, *for* $[b, c] \not\subset \partial D$.

(iii) $[c, b] \subsetneq [c, \xi_{b,c}] \subset \partial D$ *and* $[a, \xi_{b,c}] \subset \partial D$, *for* $[a, b] \not\subset \partial D$.

Proof. (I) Let $\widetilde{\xi}_{a,b}$ be an accumulation point of $\{\xi_{a_n,b_n}\}_{n\in\mathbb{N}}$; it might be different from $\xi_{a,b}$. Since $a_n \to a$ and $b_n \to b$ as $n \to \infty$, it follows that $\widetilde{\xi}_{a,b} \in \langle a, b\rangle \cap \partial D$ and $a \in [\widetilde{\xi}_{a,b}, b]$. Hence, with $[\xi_{a_n,b_n}, \xi_{b_n,c_n}] \subset \partial D$ and $b = \lim_{n\to\infty} \xi_{b_n,c_n}$, Remark 2.2 yields $[a, b] \subset \partial D$ and thus $c \notin \langle a, b\rangle$.

Now let $\widetilde{\xi}_{b,a}$ be an accumulation point of $\{\xi_{b_n,a_n}\}_{n\in\mathbb{N}}$. Then $\widetilde{\xi}_{b,a} \in \langle a, b\rangle \cap \partial D$ and $b \in [\widetilde{\xi}_{b,a}, a]$. Since $[\xi_{c,b}, \widetilde{\xi}_{b,a}] \subset \partial D$ (by Remark 2.2(b) and Lemma 2.1), it follows from Remark 2.2(c) (or directly from convexity) that $\widetilde{\xi}_{b,a} = \xi_{b,a}$, and therefore $[\xi_{b,a}, \xi_{c,b}] \subset \partial D$. Since $c \in [\xi_{c,b}, b] \cap D$, we deduce $b \neq \xi_{b,a}$ and thus $[a, b] \subsetneq [a, \xi_{b,a}] \subset \partial D$.

(II) Now suppose that $[b, c] \not\subset \partial D$. By Lemma 2.1 and Remark 2.2(a), $b = \xi_{b,c} = \lim_{n\to\infty} \xi_{b_n,c_n}$ and $c = \xi_{c,b} = \lim_{n\to\infty} \xi_{c_n,b_n}$. Again from the same lemma and remark, it follows that $[a, b] \subset \partial D$, since for every accumulation point $\widetilde{\xi}_{a,b}$ of $\{\xi_{a_n,b_n}\}_{n\in\mathbb{N}}$ we have $\widetilde{\xi}_{a,b} \in \langle a, b\rangle$.

Since $[b, c] \not\subset \partial D$, it follows just as in part (I) that $\widetilde{\xi}_{b,a} = \xi_{b,a}$, $[c, \xi_{b,a}] \subset \partial D$ and $b \neq \xi_{b,a}$, hence $[a, b] \subsetneq [a, \xi_{b,a}] \subset \partial D$. This proves (ii). Interchanging the roles of a and c, we get (iii). □

Corollary 2.7. *Let* r *and* r_i, $i \in \mathbb{N}$, *be geodesic rays in the Hilbert geometry* (D, h_D) *such that the* r_i *converge pointwise to* r, *and suppose that* $\lim_{i\to\infty} r_i(\infty)$ *exists. Then*

$$\Big[\lim_{i\to\infty} r_i(\infty), r(\infty)\Big] \subset \partial D.$$

If, moreover, $\lim_{i\to\infty} r_i(\infty) \neq r(\infty)$, *there exists* $\xi \in \partial D$ *such that*

$$r(\infty) \in \big[\xi, \lim_{i\to\infty} r_i(\infty)\big]^\circ.$$

Lemma 2.8. *Let* $\gamma : (-\infty, \infty) \to (D, h_D)$ *be a geodesic in the Hilbert geometry* (D, h_D) *such that*

$$[\gamma(-\infty), \gamma(\infty)] \subset \partial D.$$

Then

$$[\gamma(-\infty), \gamma(\infty)] = \langle \gamma(-\infty), \gamma(\infty)\rangle \cap \partial D.$$

Proof. From Lemma 2.1 we deduce

$$\big[\xi_{\gamma(-n),\gamma(0)}, \xi_{\gamma(0),\gamma(n)}\big] \cup \big[\xi_{\gamma(n),\gamma(0)}, \xi_{\gamma(0),\gamma(-n)}\big] \subset \partial D \quad \text{for all } n \in \mathbb{N}.$$

Hence, Remark 2.2(b) yields

$$\big[\xi_{\gamma(-\infty),\gamma(0)}, \xi_{\gamma(0),\gamma(\infty)}\big] \cup \big[\xi_{\gamma(\infty),\gamma(0)}, \xi_{\gamma(0),\gamma(-\infty)}\big] \subset \partial D,$$

from which the claim follows, since $\gamma(0) \in D$. □

Gromov products. Let (X, d) be a metric space. We use the standard notation

$$(x \cdot y)_o = \tfrac{1}{2}\big[d(x, o) + d(y, o) - d(x, y)\big],$$

where $x, y, o \in X$. The expression $(x \cdot y)_o$ is called the Gromov product of x and y with respect to the basepoint o. The following theorem will be used to study the boundary of images of isometric embeddings of the Euclidean plane $\mathbb{E}^2$ into Hilbert geometries:

Theorem 2.9 [Karlsson and Noskov 2002]. *Let* D *be a bounded convex domain. Let* $\{x_n\}_{n\in\mathbb{N}}, \{z_n\}_{n\in\mathbb{N}}$ *be two sequences of points in* D. *Assume that*

$$x_n \to \bar{x} \in \partial D, \quad z_n \to \bar{z} \in \partial D, \quad [\bar{x}, \bar{z}] \not\subset \partial D.$$

Then, for any fixed p_0, *there is a constant* $K = K(p_0, \bar{x}, \bar{z})$ *such that*

$$\limsup_{n\to\infty} (x_n \cdot z_n)_{p_0} \leq K.$$

For Gromov products in the Euclidean plane, we have:

Lemma 2.10 (Gromov products in $\mathbb{E}^2$). *Let* r_1, r_2 *be two geodesic rays in* $\mathbb{E}^2$, *parameterized by arc length, starting at* $o \in \mathbb{E}^2$ *and forming there an angle* $\alpha = \angle_o(r_1, r_2) \neq \pi$. *Then*

$$\lim_{n\to\infty} \big(r_1(n) \cdot r_2(n)\big)_o = \infty.$$

3. The proof of Theorem 1.2

Notation. Suppose that (D, h_D) is a Hilbert geometry admitting an isometric embedding $\varphi : \mathbb{E}^2 \to (D, h_D)$ of the Euclidean plane.

Fix an origin $o \in \mathbb{E}^2$. Consider the set $\mathcal{R}_o$ of geodesic rays in D that are images (under φ) of geodesic rays in $\mathbb{E}^2$ starting at o. We coordinatize $\mathcal{R}_o$ by the unit sphere $\mathbb{S}^1 \subset \mathbb{C} \cong \mathbb{E}^2$ around o: for $\alpha \in \mathbb{S}^1$ we denote by r_α the image in D of the geodesic ray starting at o and passing through α. We call the point $r_\alpha(\infty)$ in ∂D to which r_α converges the *endpoint* of r_α, and so get the *endpoint map*

$$\mathbb{S}^1 \to \partial D, \quad \alpha \mapsto r_\alpha(\infty),$$

whose image is the set $\mathscr{E}$ of endpoints. The endpoint map is not necessarily continuous — see remarks before the Jump Lemma 2.6 — but by that lemma or its corollary, the map *is* continuous at every $\alpha \in \mathbb{S}^1$ having the property that $r_\alpha(\infty)$ is not contained in the (relative) interior of a boundary flat. Indeed, if the endpoint map is not continuous at $\alpha \in \mathbb{S}^1$, then $r_\alpha(\infty)$ is contained in the (relative) interior of a line segment in ∂D. We may express this also as follows.

Proposition 3.1 (Endpoint Alternative). *If e is an endpoint,*

- *the point e is the limit of a sequence of distinct endpoints, or*
- *there is a line segment in ∂D containing e such that e or another endpoint is contained in the relative interior of the segment.*

(These alternatives are not mutually exclusive.)

Proof. Suppose a sequence $\{\alpha_n\}_{n\in\mathbb{N}}$ in $\mathbb{S}^1$ converges to $\alpha \in \mathbb{S}^1$ and that $\{r_{\alpha_n}(\infty)\}$ converges to some $e' := \lim_{n\to\infty} r_{\alpha_n}(\infty) \neq r_\alpha(\infty) =: e$. We show that

$$e \text{ lies in the relative interior of } \langle e, e' \rangle \cap \partial D. \tag{3-1}$$

Indeed, set $c_n := \varphi(o)$ for all $n \in \mathbb{N}$. The geodesic rays r_{α_n} converge pointwise to r_α with $r_\alpha(\infty) = e$, but $\lim_{n\to\infty} = e' \neq e$. By passing to a suitable subsequence of $\{\alpha_n\}_{n\in\mathbb{N}}$ (which, for simplicity, we denote again by $\{\alpha_n\}_{n\in\mathbb{N}}$), we can pick $b_n, a_n \in r_{\alpha_n}$ such that

$$h_D(a_n, c_n) = h_D(a_n, b_n) + h_D(b_n, c_n) \quad \text{for all } n \in \mathbb{N}$$

as well as $\lim_{n\to\infty} b_n = e$ and $\lim_{n\to\infty} a_n = e'$. Now apply part (I) of the Jump Lemma 2.6, with $b = e$ and $a = e'$, to obtain (3-1).

We will apply this fact in different cases. Suppose that $e = r_\alpha(\infty)$ is not the limit of a sequence of distinct endpoints. Then either

(a) $\delta \mapsto r_\delta(\infty)$ is locally constant at α, i.e., there exists $\varepsilon > 0$ such that $r_\beta(\infty) = e$ for all $\beta \in (\alpha - \varepsilon, \alpha + \varepsilon)$, or

(b) there exists a sequence $\{\alpha_n\}_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} \alpha_n = \alpha$ such that $\{r_{\alpha_n}(\infty)\}_{n\in\mathbb{N}}$ converges to some $e' := \lim_{n\to\infty} r_{\alpha_n}(\infty) \neq e$.

In case (b), we see immediately that the conditions leading to (3-1) are satisfied, and we are done.

Now suppose that (a) holds and let

$$\varepsilon_0 := \sup\big\{\varepsilon > 0 \mid r_\beta(\infty) = e \text{ for all } \beta \in (\alpha-\varepsilon, \alpha+\varepsilon)\big\}.$$

From Theorem 2.3(ii) it follows that $r_\beta(\infty)$ is not globally constant, which implies $\varepsilon_0 \le \pi$.

Let $\alpha' := \alpha \pm \varepsilon_0$ be such that $r_\beta(\infty)$ is not locally constant at α'. Then either

$$r_{\alpha'}(\infty) = e \quad \text{or} \quad r_{\alpha'}(\infty) \neq e.$$

In the first case, since ε_0 was chosen maximal, the choice of α' guarantees that there exists a sequence $\{\alpha'_n\}_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} \alpha'_n = \alpha$ such that $\{r_{\alpha'_n}(\infty)\}_{n\in\mathbb{N}}$ converges to some $e' := \lim_{n\to\infty} r_{\alpha'_n}(\infty) \neq e$. This follows since e is assumed not to be a limit of a sequence of distinct endpoints. Therefore, since $r_{\alpha'}(\infty) = e$, we are essentially back in alternative (b).

If instead $r_{\alpha'}(\infty) \neq e$, since $r_\beta(\infty) = e$ for all $\beta \in (\alpha-\varepsilon_0, \alpha+\varepsilon_0)$, there exists a sequence $\{\alpha'_n\}_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} \alpha'_n = \alpha'$ such that $r_{\alpha'_n}(\infty) = e \neq e' = r_{\alpha'}(\infty)$ for all $n \in \mathbb{N}$. Then again the conditions leading to (3-1) hold, with e and e' interchanged, showing that e' lies in the relative interior of $\langle e, e'\rangle \cap \partial D$. □

Corollary 3.2. *The set $\mathscr{E}$ is not contained in a single affine line. In particular, $\mathscr{E}$ contains more than two points.*

Proof. Suppose for a contradiction that there exists an affine line L such that $\mathscr{E} \subset L \cap \partial D$. Either $l := L \cap \partial D$ consists of exactly two points e and e', or l is a connected affine line segment.

In the first case, Theorem 2.3(ii) yields $\mathscr{E} = \{e, e'\}$. Thus the argument in the proof of the Endpoint Alternative implies that e or e' lie in the relative interior of $\langle e, e'\rangle \cap \partial D$; a contradiction.

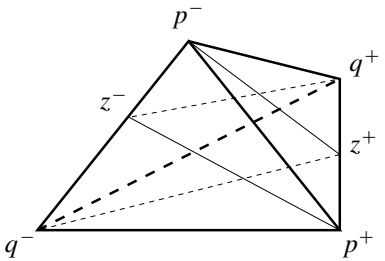
In the second case, when $l = L \cap \partial D$ is a connected affine line segment, let $e, e' \in \partial D$ be such that $l = [e, e']$. Lemma 2.8 implies that $r_\alpha(\infty) \in \{e, e'\}$ for all $\alpha \in \mathbb{S}^1$. Thus $\mathscr{E} \subset \{e, e'\}$, and Theorem 2.3(ii) yields $\mathscr{E} = \{e, e'\}$. Once again the argument in the proof of the Endpoint Alternative implies that e or e' lie in the relative interior of $\langle e, e'\rangle \cap \partial D$: a contradiction. □

Here is an immediate consequence of Theorem 2.9 and Lemma 2.10:

Corollary 3.3. *For $\alpha, \beta \in \mathbb{S}^1$ with $\alpha \neq -\beta$ we have $[r_\alpha(\infty), r_\beta(\infty)] \subset \partial D$.*

We will show that if there is an $\alpha \in \mathbb{S}^1$ with $[r_\alpha(\infty), r_{-\alpha}(\infty)] \not\subset \partial D$, then there is a tetrahedron (i.e. a 3-simplex) bordering D. Corollary 3.3 will be generalized in Lemma 4.2 for three-dimensional Hilbert geometries by showing that if D is the (interior of a) tetrahedron, then the case $[r_\alpha(\infty), r_{-\alpha}(\infty)] \not\subset \partial D$ does not occur.

Lemma 3.4. *Let (D, h_D) be an n-dimensional Hilbert geometry. Suppose there exists a geodesic $\gamma : (-\infty, \infty) \to D$ in $\mathbb{E}^2$ whose image satisfies $[\gamma(-\infty), \gamma(\infty)] \not\subset \partial D$. Then there exists a 3-dimensional affine space Σ in $\mathbb{E}^n$ such that $D \cap \Sigma^3$ is (bordered by) a 3-simplex.*



Proof. We can suppose that $\gamma|_{[0,\infty)} = r_1$, and we set

$$z^+ = r_1(\infty), \quad z^- = r_{-1}(\infty) = \gamma(-\infty).$$

(See the figure above, where the points p^+, p^-, q^+, and q^- are the vertices of a tetrahedron whose surface is contained in ∂D.)

Let $\{\alpha_n\}_n$ be a sequence in $\mathbb{S}^1 \setminus \{1, -1\}$ converging to 1 such that

$$\lim_{n\to\infty} r_{\alpha_n}(\infty) = q^+ \in \partial D \quad \text{and} \quad \lim_{n\to\infty} r_{-\alpha_n}(\infty) = q^- \in \partial D$$

exist. Thanks to Corollary 3.3, we have

$$[z^-, r_{\alpha_n}(\infty)],\ [z^+, r_{-\alpha_n}(\infty)],\ [z^-, r_{-\alpha_n}(\infty)],\ [z^+, r_{\alpha_n}(\infty)] \subset \partial D,$$

for all n, and since by Remark 2.2(b) edges converge to edges, we get

$$[z^-, q^+],\ [z^+, q^-],\ [z^-, q^-],\ [z^+, q^+] \subset \partial D.$$

Hence $q^+, q^- \in \partial D \setminus \{z^+, z^-\}$, since $[z^-, z^+] \not\subset \partial D$ by assumption. Furthermore, Remark 2.2(b) also implies $[q^+, q^-] \subset \partial D$.

Set $p^+ = \xi_{z^+,q^+}$ and $p^- = \xi_{z^-,q^-}$. From Corollary 2.7 it follows that $p^- \neq z^-$ and $p^+ \neq z^+$.

Now suppose $q^- = p^+$; since $[q^+, z^-] \subset \partial D$ as we saw above, Remark 2.2 yields $\Delta(q^+, q^-, p^-) \subset \partial D$, in particular $[z^+, z^-] \subset \partial D$; a contradiction. Thus we have $q^- \neq p^+$ and, similarly, $q^+ \neq p^-$.

From the Jump Lemma 2.6 we get $[z^-, p^+]$, $[z^+, p^-] \subset \partial D$, which implies that $q^+ \neq q^-$, for otherwise $[z^-, z^+] \subset \partial D$, which contradicts the assumption $[\gamma(-\infty), \gamma(\infty)] \not\subset \partial D$. Now Remark 2.2(c) finally implies

$$\Delta(q^+, q^-, p^-), \Delta(q^+, q^-, p^+), \Delta(p^+, p^-, q^+), \Delta(p^+, p^-, q^-) \subset \partial D;$$

that is, the surface of the tetrahedron spanned by q^+, q^-, p^+ and p^- is the part of ∂D contained in the affine space spanned by these four points. □

Proof of Theorem 1.2. Suppose there is an isometric embedding $\varphi : \mathbb{E}^2 \to (D, h_D)$; we will use the notation of page 265.

Lemma 2.8 and the Endpoint Alternative (Proposition 3.1) show there exists $\alpha \in \mathbb{S}^1 \setminus \{1, -1\}$ such that $r_\alpha(\infty) \neq r_1(\infty)$. Therefore, the existence of a line segment in ∂D follows from Corollary 3.3.

To prove that ∂D is not C^1, it obviously suffices to show that there exists an affine plane Σ in $\mathbb{E}^n$ intersecting D and such that there exist distinct points $a, b, c \in \Sigma \cap \partial D$ with $[a, b],\ [b, c] \subset \partial D$ and $a \notin \langle b, c\rangle$.

Suppose first that there exists a geodesic γ in D such that $[\gamma(-\infty), \gamma(\infty)] \subset \partial D$ and which is the image of a geodesic in $\mathbb{E}^2$ under φ. Then

$$\left[\xi_{\gamma(0),\gamma(-\infty)}, \xi_{\gamma(\infty),\gamma(0)}\right] \cup \left[\xi_{\gamma(0),\gamma(\infty)}, \xi_{\gamma(-\infty),\gamma(0)}\right] \subset \partial D.$$

Therefore, $a = \gamma(-\infty)$, $b = \gamma(\infty)$, $c = \xi_{\gamma(0),\gamma(-\infty)}$, and the affine plane Σ spanned by a, b and c are as claimed; note that $\gamma(0) \in \Sigma \cap D$.

Now suppose there is a geodesic $\gamma : (-\infty, \infty) \to D$ with $[\gamma(-\infty), \gamma(\infty)]^\circ \subset D$ that is the image of a geodesic in $\mathbb{E}^2$ under φ. With the notation as in the proof of Lemma 3.4, we set $a = z^-$, $b = q^+$ and $c = z^+$ and consider the affine plane Σ spanned by a, b and c. These points have the desired properties since $[z^+, z^-]^\circ \subset \Sigma \cap D$. □

4. The proof of Theorem 1.1

We continue to use the notation introduced in the last section, and we additionally assume $D \subset \mathbb{E}^3$.

To prove Lemma 4.2, a generalization of Corollary 3.3 in dimension 3 to the case $\alpha = -\beta$, we will use Lemma 4.1.

Lemma 4.1. *Let (D, h_D) be a 3-dimensional Hilbert geometry that admits an isometric embedding $\varphi : \mathbb{E}^2 \to (D, h_D)$. Suppose there exist 2-dimensional boundary faces $F_0, F_1, F_2, \subset \partial D$ such that $F_1 \cap F_0$ and $F_2 \cap F_0$ are one-dimensional boundary flats. Then the (relative) interior of $F_1 \cap F_0$ and $F_2 \cap F_0$ satisfy*

$$\mathscr{C} \cap (F_1 \cap F_0)^\circ = \varnothing \quad \text{or} \quad \mathscr{C} \cap (F_2 \cap F_0)^\circ = \varnothing.$$

Proof. Suppose $\mathscr{C} \cap (F_1 \cap F_0)^\circ \neq \varnothing \neq \mathscr{C} \cap (F_2 \cap F_0)^\circ$. We denote the affine plane containing F_0 by Σ_0. For $t \in (0, dist(\varphi(o), \Sigma_0))$ let Σ_t be the affine plane in Euclidean distance $dist(\Sigma_0, \Sigma_t) = t$ to Σ_0 which intersects D. For t sufficiently small, we can choose $e_t,\ f_t \in \Sigma_t \cap \varphi(\mathbb{E}^2)$ such that $\xi_{e_t, f_t} \in F_1$ and $\xi_{f_t, e_t} \in F_2$, which by Lemma 2.5 is not possible; this contradiction proves the claim. □

The generalization of Corollary 3.3 in dimension 3 mentioned above reads as follows:

Lemma 4.2. *Let (D, h_D) be a 3-dimensional Hilbert geometry that admits an isometric embedding $\varphi : \mathbb{E}^2 \to (D, h_D)$. Then*

$$\big[r_\alpha(\infty), r_\beta(\infty)\big] \subset \partial D \quad \text{for all } \alpha, \beta \in \mathbb{S}^1.$$

Slightly more generally: Any two points $e, f \in \overline{\mathscr{E}}$ span a line segment $[e, f] \subset \partial D$ in the boundary of D.

Proof. Due to Corollary 3.3, all we have to prove is that

$$\big[r_\alpha(\infty), r_{-\alpha}(\infty)\big] \subset \partial D \quad \text{for all } \alpha \in \mathbb{S}^1.$$

Suppose for a contradiction that there exists a geodesic through $o \in \mathbb{E}^2$, such that for its image γ under φ we get $[\gamma(-\infty), \gamma(\infty)] \not\subset \partial D$. Then, since D is three-dimensional, it follows from Lemma 3.4, that ∂D is a tetrahedron such that $z^- := \gamma(-\infty)$ and $z^+ := \gamma(\infty)$ are points on the relative interiors of two opposite edges.

Let $\alpha \in \mathbb{S}^1$ be such that $r_\alpha(\infty) = z^+$ and $r_{-\alpha}(\infty) = z^-$. We deduce from Corollary 3.3 that $[r_\beta(\infty), r_\alpha(\infty)] \cup [r_\beta(\infty), r_{-\alpha}(\infty)] \subset \partial D$ for any $\beta \in \mathbb{S}^1 \setminus \{-\alpha, \alpha\}$; this implies

$$\mathscr{E} \subset [q^+, q^-] \cup [q^+, p^-] \cup [p^+, q^-] \cup [p^+, p^-] \cup \{z^-, z^+\},$$

where we use the same notation as in the proof of Lemma 3.4. Thus the Jump Lemma 2.6 implies

$$\mathscr{E} \cap \big([q^+, q^-]^\circ \cup [q^+, p^-]^\circ \cup [p^+, q^-]^\circ \cup [p^+, p^-]^\circ\big) \neq \varnothing.$$

This, however, contradicts Lemma 4.1. To see this, suppose, for instance, that there exists $e \in [q^+, q^-]^\circ \cap \mathscr{E}$. The contradiction to Lemma 4.1 follows in this case by setting for instance $F_0 := \Delta(q^-, q^+, p^+)$, $F_1 := \Delta(p^+, p^-, q^+)$ and $F_2 := \Delta(q^+, q^-, p^-)$. The other cases follow in the same way. □

Being interested in the endpoint set $\mathscr{E}$, we have so far restricted our attention to those geodesic rays r_α in $\varphi(\mathbb{E}^2)$ the endpoints of which define $\mathscr{E}$, that is, the images of rays emanating from o. In the following we will also have to consider other geodesic rays in $\varphi(\mathbb{E}^2)$. The next lemma examines the relation of endpoints $r_\alpha(\infty)$ and those of geodesic rays r in $\varphi(\mathbb{E}^2)$ which are parallel to r_α.

Lemma 4.3. *Let r_α and r be geodesic rays in $\varphi(\mathbb{E}^2)$ which are in finite Hausdorff distance to each other. The following alternatives are possible.*

(1) *If $r_\alpha(\infty)$ is not contained in the interior of an affine line segment $[e, e'] \subset \partial D$, then $r_\alpha(\infty) = r(\infty)$.*

(2) *If $r_\alpha(\infty)$ is contained in the interior of an affine line segment $[e, e'] \subset \partial D$, then $r(\infty) \in [L(e, e') \cap \partial D]^\circ$.*

Proof. Suppose $r_\alpha(\infty) \neq r(\infty)$ and $\xi_{r_\alpha(\infty),r(\infty)} = r_\alpha(\infty)$. For all $\varepsilon > 0$ there exist Euclidean neighborhoods $U^\varepsilon_{r_\alpha}$ of $r_\alpha(\infty)$ and U^ε_r of $r(\infty)$ such that $|x\xi_{xy}| < \varepsilon$ for all $x \in U^\varepsilon_{r_\alpha} \cap D$ and $y \in U^\varepsilon_r \cap D$. The conclusion of the lemma follows from the definition of the Hilbert distance, since $|r(\infty)r_\alpha(\infty)| > 0$ and since the rays are supposed to be in finite Hausdorff distance. □

Lemma 4.4. *Let (D, h_D) be a 3-dimensional Hilbert geometry admitting an isometric embedding $\varphi : \mathbb{E}^2 \to (D, h_D)$ of the Euclidean plane. Suppose there are $\alpha, \beta \in \mathbb{S}^1$ such that the affine space $\Sigma = \langle r_\alpha(\infty), r_\beta(\infty), r_{-\beta}(\infty)\rangle$ induced by $r_\alpha(\infty), r_\beta(\infty)$ and $r_{-\beta}(\infty)$ in $\mathbb{E}^3$ intersects ∂D in a two-dimensional boundary flat $F = \Sigma \cap \partial D$, and assume that $r_\alpha(\infty)$ is not contained in an open line segment of ∂D. Then there exist $u, v \in \partial F$ with*

$$[u, r_\alpha(\infty)] \cup [v, r_\alpha(\infty)] \subset \partial F,$$

such that u, v and $r_\alpha(\infty)$ are affinely independent. Here ∂F denotes the boundary of F in Σ.

Proof. From Lemma 2.8 and our assumptions it follows that $r_\alpha(\infty)$, $r_{-\beta}(\infty)$, and $r_\beta(\infty)$ lie in ∂F.

Suppose that $r_{-\beta}(\infty)$ or $r_\beta(\infty)$ are contained in the interior of affine line segments which are contained in ∂D. Then those line segments are subsets of F, since otherwise, Remark 2.2(c) implies that ∂D is 3-dimensional. Since $r_{-\beta}(\infty)$ and $r_\beta(\infty)$ lie in ∂F, the same remark also yields that such a line segment in F containing $r_{-\beta}(\infty)$ or $r_\beta(\infty)$ in its relative interior has to be contained in ∂F.

(1) First assume that $r_{-\beta}(\infty)$ and $r_\beta(\infty)$ are not contained in the relative interior of line segments in ∂D.

Let γ denote the geodesic in $\varphi(\mathbb{E}^2)$ through $\varphi(o)$ determined by $\gamma(1) = r_\beta(1)$, i.e. $\gamma|_{[0,\infty]} = r_1$ and $\gamma(-t) = r_{-\beta}(t)$ for all $t > 0$. Let further $\gamma_n : (-\infty, \infty) \to D$ for $n \in \mathbb{N}$ be the geodesic in $\varphi(\mathbb{E}^2)$ determined by $|\gamma_n(t)\gamma(t)| = n$ for all $t \in (-\infty, \infty)$ and by $\operatorname{im}(r_\alpha) \cap \operatorname{im}(\gamma_n) \neq \varnothing$.

Then Lemma 4.3 implies $\gamma_n(-\infty) = r_{-\beta}(\infty)$ and $\gamma_n(\infty) = r_\beta(\infty)$ for all $n \in \mathbb{N}$. Set $b_n = \operatorname{im}(r_\alpha) \cap \operatorname{im}(\gamma_n)$; then $r_\alpha(\infty) = \lim_{n\to\infty} b_n$. Finally let Σ_n be the affine plane in $\mathbb{E}^3$ spanned by $r_{-\beta}(\infty)$, b_n and $r_\beta(\infty)$. Then $[\xi_{b_n r_\beta(\infty)}, r_{-\beta}(\infty)] \cup [\xi_{b_n r_{-\beta}(\infty)}, r_\beta(\infty)] \subset \Sigma_n \cap \partial D$. Since the b_n converge to $r_\alpha(\infty)$, it follows that

$$[r_\alpha(\infty), r_{-\beta}(\infty)] \cup [r_\alpha(\infty), r_\beta(\infty)] \subset \partial F,$$

and the claim follows with $u = r_{-\beta}(\infty)$ and $v = r_\beta(\infty)$.

(2) Now assume $r_{-\beta}(\infty)$ or $r_\beta(\infty)$ lie in the interior of an open line segment which is contained in ∂D. Then, as explained above, such a line segment has to be contained in ∂F.

To treat the remaining cases at once, we use the following notation. Let $e, e', f, f' \in \partial F$ be such that $r_{-\beta}(\infty) \in [e, e']^\circ \subset \partial F$ and $r_\beta(\infty) \in [f, f']^\circ \subset \partial F$. In case such points, say e and e', do not exist, we use the conventions $[e, e']^\circ = \{r_{-\beta}(\infty)\}$ and $[L(e, e') \cap \partial D]^\circ = \{r_{-\beta}(\infty)\} = L(e, e') \cap \partial D$.

From Lemma 4.3 it follows that for γ_n as defined in (1) we obtain $\gamma_n(-\infty) \in [L(e, e') \cap \partial D]^\circ$ as well as $\gamma_n(\infty) \in [L(f, f') \cap \partial D]^\circ$ for all $n \in \mathbb{N}$. By compactness of $L(e, e') \cap \partial D$ and $L(f, f') \cap \partial D$ there exists a subsequence of $\{\gamma_n\}_{n\in\mathbb{N}}$, which we again denote by $\{\gamma_n\}_{n\in\mathbb{N}}$, such that the $\gamma_n(-\infty)$ and the $\gamma_n(\infty)$ converge to some $u' \in L(e, e') \cap \partial D$ and some $v' \in L(f, f') \cap \partial D$.

Now let b_n be as in part (1) of the proof and let Σ_n denote the affine plane spanned by $\gamma_n(-\infty)$, b_n and $\gamma_n(\infty)$. Then

$$[\xi_{b_n,\gamma_n(-\infty)}\xi_{\gamma_n(\infty),b_n}] \quad \text{and} \quad [\xi_{b_n,\gamma_n(-\infty)}\xi_{\gamma_n(\infty),b_n}]$$

lie in $\Sigma_n \cap \partial D$.

If $\lim_{n\to\infty} \gamma_n(-\infty) = r_\alpha(\infty)$ or $\lim_{n\to\infty} \gamma_n(\infty) = r_\alpha(\infty)$, we deduce respectively that $[r_\alpha(\infty), r_{-\beta}(\infty)] \subset \partial F$ or $[r_\alpha(\infty), r_\beta(\infty)] \subset \partial F$, and we set $u = r_{-\beta}(\infty)$ or $v = r_\beta(\infty)$. Otherwise $u \neq r_\alpha(\infty) \neq v'$ and with $u := u'$ and $v := v'$ we conclude that

$$[r_\alpha(\infty), u] \cup [r_\alpha(\infty), v] \subset \partial F. \qquad \square$$

Theorem 1.1 follows as a corollary of the following two propositions, which restrict the possibilities for the configuration of the points in $\mathscr{E}$.

Proposition 4.5. *Let (D, h_D) be a 3-dimensional Hilbert geometry admitting an isometric embedding $\varphi : \mathbb{E}^2 \to (D, h_D)$ of the Euclidean plane. Then $\mathscr{E}$ is not contained in a single boundary flat.*

Proof. By Lemma 2.8 and the Endpoint Alternative (Proposition 3.1), the set $\mathscr{E}$ cannot be contained in a 1-dimensional boundary flat.

Now suppose that $\mathscr{E}$ is contained in a 2-dimensional boundary flat F. Let Σ be the affine plane in $\mathbb{E}^3$ containing F, and let ∂F denote the boundary of F in Σ. Then Lemma 2.8 already implies that $\mathscr{E} \subset \partial F$.

Claim. There exist two noncollinear line segments in ∂D each of which contains a point of $\mathscr{E}$ in its relative interior.

Proof. This follows from the Jump Lemma 2.6 and Lemma 4.4, but the argument requires distinguishing several cases. Fix $\beta \in \mathbb{S}^1$.

(1) If $r_{-\beta}(\infty)$ and $r_\beta(\infty)$ are both contained in (necessarily distinct) open line segments of ∂F, there is nothing to prove.

(2) Suppose next that $r_{-\beta}(\infty)$ is contained in an open line segment $[e, e']^\circ = [L(e, e') \cap \partial D]^\circ = [L(e, e') \cap \partial F]^\circ$, but $r_\beta(\infty)$ is not. Then, since e, e' and $r_\beta(\infty)$ are affinely independent, it follows from our analysis of the continuity properties

of $\delta \mapsto r_\delta(\infty)$ that there exists $\alpha \in \mathbb{S}^1$ with $r_\alpha(\infty) \neq r_\beta(\infty)$ and $r_\alpha(\infty) \notin [e, e']$. Without loss of generality we may assume that $r_\alpha(\infty)$ is not contained in an open line segment which is contained in ∂F, since otherwise $r_{-\beta}(\infty)$ and $r_\alpha(\infty)$ are points as desired. Lemma 4.4 then yields the existence of $u, v \in \partial F$ with $[u, r_\alpha(\infty)] \cup [v, r_\alpha(\infty)] \subset \partial F$. Again due to our analysis of the continuity properties of $\delta \mapsto r_\delta(\infty)$, there exists some $p \in ([u, r_\alpha(\infty)]^\circ \cup [v, r_\alpha(\infty)]^\circ) \cap \mathscr{C}$. Since $[u, r_\alpha(\infty)]$ and $[v, r_\alpha(\infty)]$ are not collinear with $[e, e']$ (by our choice of α), we see that p and $r_\alpha(\infty)$ form a tuple of points in $\mathscr{C}$ as desired.

(3) Now assume that neither $r_{-\beta}(\infty)$ nor $r_\beta(\infty)$ are contained in open line segments which are contained in ∂F. By Corollary 3.2 there exists $\alpha \in \mathbb{S}^1$ such that $r_\alpha(\infty)$, $r_{-\beta}(\infty)$ and $r_\beta(\infty)$ span the two-dimensional affine space Σ.

(3a) If $r_\alpha(\infty)$ and $r_{-\alpha}(\infty)$ are both contained in (necessarily distinct) open line segments which are contained in ∂F, again there is nothing to prove.

(3b) Suppose $r_\alpha(\infty)$ is contained in an open affine line segment which is contained in ∂F, but $r_{-\alpha}(\infty)$ is not. Then, by exactly the same arguments as above in (2) with $r_{-\beta}(\infty)$ and $r_\beta(\infty)$ replaced by $r_{-\alpha}(\infty)$ and $r_\alpha(\infty)$, the claim follows.

(3c) Finally suppose that $r_\alpha(\infty)$ is not contained in an open affine line segment contained in ∂F. Then, by Lemma 4.4, $[r_{-\beta}(\infty), r_\alpha(\infty)] \cup [r_\beta(\infty), r_\alpha(\infty)] \subset \partial F$. From our analysis of the continuity properties of $\delta \mapsto r_\delta(\infty)$, it follows that there exists

$$p \in ([r_{-\beta}(\infty), r_\alpha(\infty)]^\circ \cup [r_\beta(\infty), r_\alpha(\infty)]^\circ) \cap \mathscr{C}.$$

Assume $p \in [r_{-\beta}(\infty), r_\alpha(\infty)]^\circ$. If there also exists $q \in [r_\beta(\infty), r_\alpha(\infty)]^\circ \cap \mathscr{C}$, then p and q are points as desired. If not, then, due to the continuity properties of $\delta \mapsto r_\delta(\infty)$, there exists $q \in (\Sigma^- \cap \partial F \cap \mathscr{C}) \setminus \{r_{-\beta}(\infty), r_\beta(\infty)\}$, where Σ^- denotes the closure of the connected component of $\Sigma \setminus L(r_{-\beta}(\infty), r_\beta(\infty))$, which does not contain $r_\alpha(\infty)$.

If such a q is contained in an open line segment which is contained in ∂F, then p and q are points as desired. Otherwise Lemma 4.4 yields $[r_{-\beta}(\infty), q] \cup [r_\beta(\infty), q] \subset \partial F$ and there exists

$$q' \in ([r_{-\beta}(\infty), q]^\circ \cup [r_\beta(\infty), q]^\circ) \cap \mathscr{C}.$$

In this case p and q' are the points as desired, which completes the proof of the claim. □

To complete the proof of Proposition 4.5, we will apply Lemma 4.1 for a contradiction. Pick $\alpha, \beta \in \mathbb{S}^1$ so that $r_\alpha(\infty)$ and $r_\beta(\infty)$ are contained in the relative interior of noncollinear line segments l_α, l_β of ∂F. Then $[\xi_{r_\alpha(n), r_\alpha(0)}, \xi_{r_\alpha(0), r_{-\alpha}(n)}] \subset \partial D$ for all $n \in \mathbb{N}$, and therefore also $[\xi_{r_\alpha(\infty), r_\alpha(0)}, \xi_{r_\alpha(0), r_{-\alpha}(\infty)}] \subset \partial D$. Hence, by convexity, l_α and $[\xi_{r_\alpha(\infty), r_\alpha(0)}, \xi_{r_\alpha(0), r_{-\alpha}(\infty)}] = [r_\alpha(\infty), \xi_{r_\alpha(0), r_{-\alpha}(\infty)}]$ span a boundary

flat F_1 of ∂D which intersects F in l_α; see Remark 2.2(c). Similarly, l_β is part of the boundary of another flat F_2 contained in ∂D.

Setting $F_0 := F$, this contradicts Lemma 4.1. Hence, $\mathscr{C}$ is not contained in a 2-dimensional boundary flat. □

Proposition 4.6. *Let (D, h_D) be a 3-dimensional Hilbert geometry admitting an isometric embedding $\varphi : \mathbb{E}^2 \to (D, h_D)$ of the Euclidean plane. Then $\mathscr{C}$ is contained in a single 2-dimensional affine space $\Sigma \subset \mathbb{E}^3$.*

Proof. In order to reach a contradiction, suppose there exist endpoints $a, b, c, d \in \mathscr{C}$ not contained in a common affine plane. Then any three of them are not contained in a common affine line. The four points a, b, c, d span four triangles the edges of which lie in ∂D. In the following we distinguish several cases by the number of those triangles, which are contained in boundary flats, i.e. the interiors of which are not contained in D. Note that the border of such a triangle belongs to ∂D and that such a triangle is either contained in ∂D, or its intersection with ∂D is the boundary of the triangle; see Remark 2.2(c).

Case 0: *All four triangles are boundary flats.* In this case, D is the interior of the tetrahedron with vertices a, b, c, and d. The Endpoint Alternative yields that at least two of the edges carry a point of $\mathscr{C}$ in their relative interiors, while Lemma 4.2 implies that those edges belong to the boundary of the same boundary flat. This, however, is not possible, due to Lemma 4.1.

Case 1: *Exactly three triangles are boundary flats.* Assume that the triangles $\Delta(d, a, b)$, $\Delta(d, a, c)$ and $\Delta(d, b, c)$ are boundary flats, whereas $\Delta^\circ(a, b, c)$ is contained in D. Let $\mathscr{C}_{\text{hull}}(a, b, c, d)$ be the (Euclidean) convex hull of a, b, c and d. Then its interior $\mathscr{C}^\circ_{\text{hull}}(a, b, c, d)$ satisfies $\mathscr{C}^\circ_{\text{hull}}(a, b, c, d) \subset D$.

Now consider the points $\xi_{a,d}, \xi_{b,d}$ and $\xi_{c,d}$. At least two of them must correspond to the points a, b and c, since otherwise we get a contradiction due to Lemma 2.5. So assume that $\xi_{b,d} = b$ and $\xi_{c,d} = c$. If also $\xi_{a,d} = a$, there exists

$$e \in \mathscr{C} \cap \{[a, d]^\circ \cup [b, d]^\circ \cup [c, d]^\circ\},$$

as can be seen by combining the inclusion $\mathscr{C}^\circ_{\text{hull}}(a, b, c, d) \subset D$ with Lemma 4.2, since $\delta \mapsto r_\delta(\infty)$ has to leave d somehow. Suppose that $e \in [a, d]$, say; then we can just replace a by e and obtain $e \neq \xi_{e,d}$. Thus we can assume without loss of generality that $a \neq \xi_{a,d}$.

Again from Lemmas 4.2 and 2.5 we deduce that $\mathscr{C} \cap \langle b, c, d\rangle \cap \partial D = \{b, c, d\}$. Also, since $\mathscr{C}^\circ_{\text{hull}}(\xi_{a,d}, b, c, d) \subset D$, it follows that $\mathscr{C} \subset (\langle d, a, b\rangle \cup \langle d, a, c\rangle) \cap \partial D$.

Suppose next that there exists $e \in ([\xi_{a,d}, b]^\circ \cup [\xi_{a,d}, c]^\circ) \cap \mathscr{C}$. Then $\Delta(\xi_{a,d}, b, c) \subset \partial D$ and we reach a contradiction with Lemma 2.5, since $a \in [\xi_{a,d}, d]^\circ$.

Since $\delta \mapsto r_\delta(\infty)$ has to leave both b and c somehow, the Endpoint Alternative (Proposition 3.1) yields $\partial F^-(\xi_{a,d}, b, d) \cap \mathscr{C} \neq \varnothing$ as well as $\partial F^-(\xi_{a,d}, c, d) \cap \mathscr{C} \neq \varnothing$,

where, for instance, $F^-(\xi_{a,d}, b, d)$ denotes the intersection of the boundary flat $\langle \xi_{a,d}, b, d\rangle \cap \partial D$ with the open half-plane in $\langle \xi_{a,d}, b, d\rangle \setminus \langle \xi_{a,d}, b\rangle$ not containing d, and $\partial F^-(\xi_{a,d}, b, d)$ denotes its boundary in $\langle \xi_{a,d}, b, d\rangle \setminus \langle \xi_{a,d}, b\rangle$.

Set $\partial[F_b]^- := \partial F^-(\xi_{a,d}, b, d)$ and $\partial[F_c]^- := \partial F^-(\xi_{a,d}, c, d)$. We claim that

$$\bigcup_{q\in[\partial F_c]^+} [b, q] \subset \partial D \quad \text{and} \quad \bigcup_{q\in[\partial F_c]^-} [b, q] \subset \partial D. \tag{4-1}$$

Indeed, $\delta \mapsto r_\delta(\infty)$ has to move from c to b somehow. If $\delta \mapsto r_\delta(\infty)$ is continuous, the claim is clear from Lemma 4.2. The general case follows, since whenever $\delta \mapsto r_\delta(\infty)$ is not continuous, there exists $e \in [\partial F_b]^- \cap \mathscr{E}$ (or $e \in [\partial F_c]^- \cap \mathscr{E}$) in the relative interior of a line segment in $[\partial F_b]^-$ (or $[\partial F_b]^-$, respectively), and then Lemma 4.2 and Remark 2.2(c) lead to $[b, q] \subset \partial D$ (or $[b, q] \subset \partial D$) for all q contained in such a line segment.

There exist $e \in \mathscr{E} \cap [\partial F_b]^-$ and $e' \in \mathscr{E} \cap [\partial F_c]^-$. From Lemma 4.2 we find that $[e, e'] \subset \partial D$; hence, since $[\partial F_b]^- \cap [\xi_{ad}, b] = \varnothing$ and $[\partial F_b]^- \cap [\xi_{ad}, b] = \varnothing$, the segment $[e, e']$ is contained in the open half-space of $\mathbb{E}^3 \setminus \langle \xi_{a,d}, b, c\rangle$ that does not contain d.

Now take a sequence $\{q_n\}_{n\in\mathbb{N}}$ in $[\partial F_b]^-$ converging to ξ_{ad}. Then $[b, q_n]$ lies in ∂D for all $n \in \mathbb{N}$, and for n sufficiently large and $p \in [q_n, b]^\circ$ the straight line $\langle e, p\rangle$ intersects $\Delta^\circ(\xi_{a,d}, b, c)$. We reach a contradiction, since $e \in \partial D$ and $e \neq p \in \partial D$ imply that for the 3-dimensional $\mathscr{C}^\circ_{\text{hull}}(\xi_{a,d}, b, c, e)$ we obtain $\mathscr{C}^\circ_{\text{hull}}(\xi_{a,d}, b, c, e) \subset \partial D$ due to Remark 2.2(c).

Note that this argument also works if $d \in \partial D \setminus \mathscr{E}$ and there exists some $e \in \mathscr{E}$ in the relative interior of one of the edges of the closed triangles meeting in d.

Case 2: *Exactly two triangles are boundary flats.* We assume without loss of generality that $\Delta(a, b, d), \Delta(a, c, d) \subset \partial D$. We first fix some notation: $\Sigma(a, b, c)$ will denote the affine plane in $\mathbb{E}^3$ spanned by a, b and c, and $H^-(a, b, c; d)$, $H^+(a, b, c; d)$ will denote the open half-spaces of $\mathbb{E}^3 \setminus \Sigma(a, b, c)$ characterized by $d \notin H^-(a, b, c; d)$ and $d \in H^+(a, b, c; d)$. Also define the boundary flats

$$F_b = \Sigma(a, b, d) \cap \partial D, \quad F_c = \Sigma(a, c, d) \cap \partial D,$$

and let ∂F_b and ∂F_c be their boundaries in $\Sigma(a, b, d)$ and $\Sigma(a, c, d)$, respectively.

We now seek a contradiction with the existence of an isometric embedding of $\mathbb{E}^2$ into D.

(1) We first verify that $\mathscr{E} \subset \partial F_b \cup \partial F_c$.

Suppose there exists $p \in \mathscr{E} \cap (L(a, d) \cap \partial D)^\circ$. Since ∂D is 2-dimensional, Lemma 4.2 and Remark 2.2(c) yield $\mathscr{E} \cap \partial D \subset F_b \cup F_c$. Thus $\mathscr{E} \subset \partial F_b \cup \partial F_c$, since for any $p \in \mathscr{E} \cap F_b^\circ$ (or $p \in \mathscr{E} \cap F_c^\circ$), the condition $[c, p] \subset \partial D$ (or $[b, p] \subset \partial D$) contradicts the 2-dimensionality of ∂D, due to Remark 2.2(c).

Thus we may assume that $L(a,d)\cap\partial D=[a,d]$ and $[a,d]^\circ\cap\mathscr{C}=\varnothing$. From Remark 2.2(c) it follows that $C^\circ_{\text{hull}}(a,b,c,d)\cap\partial D$ is nonempty. Therefore, since $\Delta(a,b,d)\cup\Delta(a,c,d)\subset\partial D$, we conclude that

$$\partial D\cap\big(H^-(a,c,d;b)\cup H^-(a,b,d;c)\big)=\varnothing. \tag{4-2}$$

Moreover, we claim that

$$\mathscr{C}\cap H^+(a,c,d;b)\cap H^+(a,b,d;c)=\varnothing. \tag{4-3}$$

To verify this, we set $Q:=H^+(a,c,d;b)\cap H^+(a,b,d;c)$, $Q_0:=\Sigma(b,c,d)\cap Q$, $Q_+:=Q\cap H^+(b,c,d;a)$ and $Q_-:=Q\cap H^-(b,c,d;a)$. Then $Q=Q_0\cup Q_+\cup Q_-$. Now $Q_0\cap\mathscr{C}=\varnothing$, since otherwise $\Delta(b,c,d)\subset\partial D$. We also deduce $Q_+\cap\mathscr{C}=\varnothing$; indeed, if this intersection contained a point p, we would have $[p,d]\subset\partial D$ from Lemma 4.2, but since $p\notin\mathscr{C}_{\text{hull}}(a,b,c,d)$, this would imply that $[p,d]$ and $\Delta^\circ(a,b,c)$ are disjoint, contradicting $\Delta^\circ(a,b,c)\subset D$. Finally, $Q_-\cap\mathscr{C}=\varnothing$, since for $p\in Q_-\cap\mathscr{C}$, either $[p,a]\cap\Delta^\circ(b,c,d)\neq\varnothing$ and therefore $\Delta^\circ(b,c,d)\subset\partial D$, or $[p,d]\cap\Sigma(a,b,c)\cap Q_-\neq\varnothing$ and therefore $\Delta^\circ(a,b,c)\subset\partial D$. This proves our claim (4-3).

From (4-2) and (4-3) it follows $\mathscr{C}\subset F_b\cup F_c$, and just as above we deduce that $\mathscr{C}\subset\partial F_b\cup\partial F_c$.

(2) We will reach the desired contradiction by proving that $\Delta(\xi_{a,d},b,c)\subset\partial D$ or $\Delta(\xi_{d,a},b,c)\subset\partial D$; indeed, by our remark at the end of Case 1, either of these two inclusions precludes the existence of an isometric embedding $\varphi:\mathbb{E}^2\to D$.

Fix $p\in[\xi_{a,d},\xi_{d,a}]^\circ$. Define Σ^+, Σ^- as the components of $\Sigma(a,c,d)\setminus L(p,c)$ containing respectively $\xi_{a,d}$, $\xi_{d,a}$, and set $[\partial F_c]^+:=\big((\partial F_c\cap\Sigma^+)\setminus[\xi_{a,d},p]^\circ\big)\cup\{c\}$ and likewise for $[\partial F_c]^-$. Thus $[\partial F_c]^+$ is the piece of ∂F_c connecting $\xi_{a,d}$ to c in $\Sigma^+\cup\{c\}$. Define $[\partial F_b]^+$ and $[\partial F_b]^-$ similarly, replacing c by b. Then

$$\mathscr{C}\subset\partial F_b\cup\partial F_c=[\partial F_c]^+\cup[\partial F_c]^-\cup[\partial F_b]^+\cup[\partial F_b]^-\cup[\xi_{a,d},\xi_{d,a}].$$

An argument analogous to the one following (4-1) shows that

$$\bigcup_{q\in[\partial F_c]^+}[b,q]\subset\partial D\quad\text{or}\quad\bigcup_{q\in[\partial F_c]^-}[b,q]\subset\partial D.$$

Similarly,

$$\text{(a)}\ \bigcup_{q\in[\partial F_b]^+}[c,q]\subset\partial D\quad\text{or}\quad\text{(b)}\ \bigcup_{q\in[\partial F_b]^-}[c,q]\subset\partial D, \tag{4-4}$$

depending on whether $\delta\mapsto r_\delta(\infty)$ moves along $[\partial F_b]^+$ or $[\partial F_b]^-$. Without loss of generality we may assume that $\delta\mapsto r_\delta(\infty)$ moves along $[\partial F_c]^+$, that is to say, $\bigcup_{q\in[\partial F_c]^+}[b,q]\subset\partial D$. We have to consider cases (a) and (b) of (4-4).

(a) If there exists $q \in [\partial F_c]^+$ with $[\xi_{a,d}, q] \subset [\partial F_c]^+$, then, since $\delta \mapsto r_\delta(\infty)$ moves along $[\partial F_c]^+$ there exists $c' \in \mathscr{C} \cap [\xi_{a,d}, \xi_{q,\xi_{a,d}}]^\circ$. Replacing c by c', we obtain $\Delta(b, c', \xi_{a,d}) \subset \partial D$ by Remark 2.2(c), and we are done by our remark at the end of Case 1. Thus there exists a sequence $\{e_n\}_{n\in\mathbb{N}}$ of endpoints $e_n \in (\mathscr{C} \cap [\partial F_c]^+) \setminus \{\xi_{a,d}\}$ converging to $\xi_{a,d}$. Moreover, $[\partial F_c]^+ \cap [\xi_{a,d}, c]^\circ = \varnothing$. Since the $[c, e_n]^\circ$ converge to $[c, \xi_{a,d}]^\circ$ in $H^-(b, c, \xi_{a,d}; \xi_{d,a})$, it follows that, for n sufficiently large, there exist straight lines in $\mathbb{E}^3$ through distinct points $p \in [\partial F_c]^+ \subset \partial D$ and $q \in [c, e_n]^\circ \subset \partial D$, which intersect $\Delta^\circ(b, c, \xi_{a,d})$. Therefore, these intersection points do not belong to D and $\Delta^\circ(b, c, \xi_{a,d}) \subset \partial D$, due to Remark 2.2(c). Thus we are also done in this case.

(b) In this case there exists an endpoint $e \in \mathscr{C} \cap [\partial F_b]^- \setminus \{b, \xi_{a,d}\}$. For this endpoint we have $e \in \mathscr{C} \cap \Sigma(b, \xi_{a,d}, \xi_{d,a}) \cap H^-(b, c, \xi_{d,a}; \xi_{a,d})$, for otherwise $e \in [b, \xi_{a,d}]^\circ$ and therefore $\Delta^\circ(b, c, \xi_{a,d}) \subset \partial D$. Now just as for b, we also get for e the inclusion $\bigcup_{q\in[\partial F_c]^+} [e, q] \subset \partial D$. From the choice of e we further obtain $[e, \xi_{a,d}] \cap \Sigma(b, c, \xi_{d,a}) \in [b, \xi_{d,a}]^\circ$ as well as $[e, c] \cap \Sigma(b, c, \xi_{d,a}) = \{c\}$. Moving from $\xi_{a,d}$ to c along $[\partial F_c]^+$, the intersection of $[e, f]$ with $\Sigma(b, c, \xi_{d,a})$ moves continuously with f in $\Sigma(b, c, \xi_{d,a})$. But since $f \in [\partial F_c]^+$ implies $[e, f] \cap \Sigma(b, c, \xi_{d,a}) \notin \{b, \xi_{d,a}\}$, this point of intersection has to leave $[b, \xi_{a,d}]$ continuously in its relative interior. Since all these intersection points belong to ∂D, we deduce $\Delta^\circ(b, c, \xi_{d,a}) \subset \partial D$, which completes Case 2.

Case 3: *At most one of the triangles is a boundary flat.* We assume without loss of generality that none of the triangles with vertex d is a boundary flat. ($\Delta(a, b, c)$ might be a boundary flat or not.) Let $H = H^+(a, b, c; d)$ be the connected component of $\mathbb{E}^3 \setminus \Sigma(a, b, c)$ containing d.

(1) We first prove that $H \cap \mathscr{C} = \{d\}$.

In order to reach a contradiction, suppose that there exists $d' \in H \cap \mathscr{C}$ with $d' \neq d$. Since none of the three triangles with vertex d is a boundary flat, we deduce that $d' \notin \mathscr{C}_{\text{hull}}(a, b, c, d)$, the Euclidean convex hull of the points a, b, c, and d. In particular, d' is separated from one of the points a, b and c by the plane spanned by the two others and d.

Assume without loss of generality that d' is separated from a by $\Sigma(b, c, d)$. Then $[a, d']$ intersects $\Sigma(b, c, d)$ in one of the sides $[c, d]$ or $[d, b]$, since $\Delta(b, c, d)$ is not a boundary flat, and $[a, d'] \subset \partial D$, due to Lemma 4.2.

Without loss of generality, assume $[b, d] \cap [a, d'] \neq \varnothing$. Then $\Delta(a, b, d)$ is a boundary flat, by Remark 2.2(c); a contradiction.

(2) From $H \cap \mathscr{C} = \{d\}$ it follows by the Endpoint Alternative that the endpoint map jumps from d to some $f \in \mathbb{E}^3 \setminus H$. Let f' be the single point in $[d, f] \cap \langle a, b, c\rangle$. Due to Lemma 4.2 we have $[d, f] \subset \partial D$, and since none of the triangles with

vertex d is contained in a boundary flat, f' is separated in $\Sigma(a,b,c)$ from one of the points a, b and c by the line through the other two points.

Assume without loss of generality that f' and a are on different sides of $L(b,c)$ in $\Sigma(a,b,c)$. Since the endpoint map jumps from d to f, the claim in our proof of the Endpoint Alternative yields $f \in [d, \xi_{f,d}]^\circ$. Hence it follows from Remark 2.2 that $\Delta(a, \xi_{f,d}, d) \subset \partial D$.

Now $l = \Delta(a,d,\xi_{f,d}) \cap \Sigma(b,c,d) \subset \partial D$ is a line segment with vertex d. But, since $\Delta(b,c,d)$ is not a boundary flat, we have

$$\Sigma(b,c,d) \cap \partial D = [b,c] \cup [c,d] \cup [b,d],$$

and it follows that $[d,b] \subset l$ or $[c,d] \subset l$, contradicting the fact that neither $\Delta(a,b,d)$ nor $\Delta(a,c,d)$ are boundary flats. □

Proof of Theorem 1.1. From Proposition 4.5 and Proposition 4.6 it follows that all we have to show is that $\mathscr{C}$ is not contained in a two-dimensional plane which intersects D, i.e. that $\mathscr{C}$ does not consist of exactly three points a, b, and c with $\Delta^\circ(a,b,c) \subset D$, where $\Delta^\circ(a,b,c) = \Delta(a,b,c) \setminus ([a,b] \cup [a,c] \cup [b,c])$ is the relative interior of $\Delta(a,b,c)$. This, however, follows from our claim in the proof of the Endpoint Alternative, since none of the three points lies in the interior of an open line segment in ∂D contained in the line spanned by this and any of the other endpoints. □

Acknowledgment

We thank Anders Karlsson for useful discussions, and we are happy to express our deep gratitude to the University of Michigan for the hospitality and excellent working conditions.

References

[Benoist 2003] Y. Benoist, "Convexes hyperboliques et fonctions quasisymétriques", *Publ. Math. Inst. Hautes Études Sci.* 97 (2003), 181–237. MR 2005g:53066 Zbl 1049.53027

[Bridson and Haefliger 1999] M. R. Bridson and A. Haefliger, *Metric spaces of non-positive curvature*, Grundlehren der Math. Wissenschaften **319**, Springer, Berlin, 1999. MR 2000k:53038 Zbl 0988.53001

[Busemann 1955] H. Busemann, *The geometry of geodesics*, Pure and applied mathematics **6**, Academic Press, New York, 1955. MR 17,779a Zbl 0112.37002

[Busemann and Phadke 1987] H. Busemann and B. B. Phadke, *Spaces with distinguished geodesics*, Pure and applied mathematics **108**, Dekker, New York, 1987. MR 88g:53075 Zbl 0631.53001

[Colbois and Vernicos 2006] B. Colbois and C. Vernicos, "Bas du spectre et delta-hyperbolicité en géométrie de Hilbert plane", *Bull. Soc. Math. France* **134**:3 (2006), 357–381. MR MR2245997 Zbl 05144589

[Foertsch and Karlsson 2005] T. Foertsch and A. Karlsson, "Hilbert metrics and Minkowski norms", *J. Geom.* **83**:1-2 (2005), 22–31. MR 2007e:51021 Zbl 1084.52008

[de la Harpe 1993] P. de la Harpe, "On Hilbert's metric for simplices", pp. 97–119 in *Geometric group theory* (Sussex, 1991), vol. 1, edited by G. A. Niblo and M. A. Roller, London Math. Soc. Lecture Note Ser. **181**, Cambridge Univ. Press, Cambridge, 1993. MR 94i:52006 Zbl 0832.52002

[Hilbert 1895] D. Hilbert, "Ueber die gerade Linie als kürzeste Verbindung zweier Punkte", *Math. Ann.* **46** (1895), 91–96. Zbl 26.0540.02

[Jost 1997] J. Jost, *Nonpositive curvature: geometric and analytic aspects*, Lectures in Mathematics ETH Zürich, Birkhäuser, Basel, 1997. MR 98g:53070 Zbl 0896.53002

[Karlsson and Noskov 2002] A. Karlsson and G. A. Noskov, "The Hilbert metric and Gromov hyperbolicity", *Enseign. Math.* (2) **48**:1-2 (2002), 73–89. MR 2003f:53061 Zbl 1046.53026

[Kelly and Straus 1958] P. Kelly and E. Straus, "Curvature in Hilbert geometries", *Pacific J. Math.* **8** (1958), 119–125. MR 20 #2748 Zbl 0081.16401

[Nussbaum 1988] R. D. Nussbaum, *Hilbert's projective metric and iterated nonlinear maps*, Mem. Amer. Math. Soc. **391**, Amer. Math. Soc., Providence, RI, 1988. MR 89m:47046 Zbl 0666.47028

[Phadke 1974/75] B. B. Phadke, "A triangular world with hexagonal circles", *Geometriae Dedicata* **3** (1974/75), 511–520. MR 51 #4061 Zbl 0294.53040

Received March 16, 2006. Revised November 16, 2006.

OLIVER BLETZ-SIEBERT
MATHEMATISCHES INSTITUT
UNIVERSITÄT WUERZBURG
AM HUBLAND
97074 WUERZBURG
GERMANY

bletz-siebert@mathematik.uni-wuerzburg.de

THOMAS FOERTSCH
MATHEMATISCHES INSTITUT
RHEINISCHE FRIDRICH-WILHELMS-UNIVERSITÄT BONN
BERINGSTRASSE 1
53115 BONN
GERMANY

foertsch@math.uni-bonn.de
http://www.math.uni-bonn.de/people/foertsch/

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.279

A VOLUMISH THEOREM FOR THE JONES POLYNOMIAL OF ALTERNATING KNOTS

OLIVER T. DASBACH AND XIAO-SONG LIN

The Volume Conjecture claims that the hyperbolic volume of a knot is determined by the colored Jones polynomial.

Here we prove a "Volumish Theorem" for alternating knots in terms of the Jones polynomial, rather than the colored Jones polynomial: The ratio of the volume and certain sums of coefficients of the Jones polynomial is bounded from above and from below by constants.

Furthermore, we give experimental data on the relation of the growths of the hyperbolic volume and the coefficients of the Jones polynomial, both for alternating and nonalternating knots.

1. Introduction

Since the introduction of the Jones polynomial, there has been a strong desire to have a geometrical or topological interpretation for it rather than a combinatorial definition.

The first major success in this direction was arguably the proof of the Melvin–Morton Conjecture by Bar-Natan and Garoufalidis [1996] (see [Vaintrob 1997; Chmutov 1998; Lin and Wang 2001; Rozansky 1997] for different proofs): The Alexander polynomial is determined by the so-called colored Jones polynomial. For a knot K the colored Jones polynomial is given by the Jones polynomial and the Jones polynomials of cablings of K.

The next major conjecture that relates the Jones polynomial and its offsprings to classical topology and geometry was the Volume Conjecture of Kashaev, Murakami and Murakami (see [Murakami and Murakami 2001], for instance). This conjecture states that the colored Jones polynomial determines the Gromov norm of the knot complement. For hyperbolic knots the Gromov norm is proportional to the hyperbolic volume.

Dasbach was supported in part by NSF grants DMS-0306774 and DMS-0456275 (FRG)..
MSC2000: 57M25.
Keywords: Jones polynomial, hyperbolic volume, alternating knots, volume conjecture.

A proof of the Volume Conjecture for all knots would also imply that the colored Jones polynomial detects the unknot [Murakami and Murakami 2001]. This problem is still wide open; even for the Jones polynomial there is no counterexample known (see [Dasbach and Hougardy 1997], for example).

The purpose of this paper is to show a relation of the coefficients of the Jones polynomial and the hyperbolic volume of alternating knot complements. More specifically we prove:

Volumish Theorem. *For an alternating, prime, nontorus knot K let*

$$V_K(t) = a_n t^n + \cdots + a_m t^m$$

be the Jones polynomial of K. Then

$$v_8(\max(|a_{m-1}|, |a_{n+1}|) - 1) \leq \mathrm{Vol}(S^3 - K) \leq 10\, v_3(|a_{n+1}| + |a_{m-1}| - 1).$$

Here, $v_3 \approx 1.01494$ is the volume of an ideal regular hyperbolic tetrahedron and $v_8 \approx 3.66386$ is the volume of an ideal regular hyperbolic octahedron.

For the proof of this theorem we make use of a result from [Lackenby 2004] (with its proof), stating that the hyperbolic volume is linearly bounded from above and below by the twist number. Lackenby's upper bound was improved by Ian Agol and Dylan Thurston and his lower bound by Ian Agol, Peter Storm and Bill Thurston [Agol et al. 2005].

In an appendix we give some numerical data on the relation between other coefficients and the hyperbolic volume, for both alternating and nonalternating knots. These data gives some hope for a Volumish Theorem for nonalternating knots as well.

2. The Jones polynomial evaluation of the Tutte polynomial

Our goal is to relate the hyperbolic volume of alternating knot complements to the coefficients of the Jones polynomial. We make use of the computation of the Jones polynomial of alternating links via the Tutte polynomial.

Notation. Our objects are multigraphs, that is, graphs where parallel edges are allowed. Two edges are called parallel if they connect the same two vertices.

(a) A multigraph $G = (V, E)$ has a set V of vertices and a set E of edges.

(b) We denote by $\tilde{G} = (V, \tilde{E})$ a spanning subgraph of G where parallel edges are deleted. See Figure 1. The set of vertices V is the same.

(c) Each edge $e \in \tilde{E}$ in $\tilde{G}$ can be assigned a multiplicity $\mu(e)$, namely, the number of edges in G that are parallel to e. For example, the graph in Figure 1 has one edge with multiplicity 2, one with multiplicity 3, and one with multiplicity 4. All other edges have multiplicity 1.

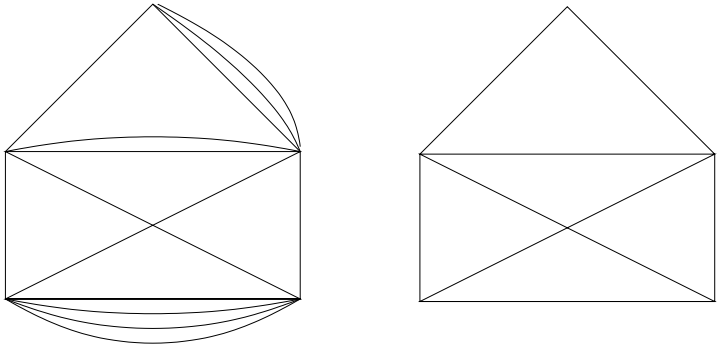


Figure 1. A multigraph G and its spanning subgraph $\tilde{G}$.

(d) We define $n(j)$ to be the number of edges $e \in \tilde{E}$ with $\mu(j) \geq j$. In particular $n(1) = |\tilde{E}|$. Thus the graph in Figure 1 has $n(2) = 3$, $n(3) = 2$, $n(4) = 1$.

(e) The number of components of a graph G is $k(G)$. If V is apparent from the context and $G = (V, E)$, we set $k(E) := k(G)$.

(f) The Tutte polynomial of a multigraph G (see [Bollobás 1998], for example) is

$$T_G(x, y) := \sum_{F \subseteq E} (x-1)^{k(F)-k(E)} (y-1)^{|F|-|V|+k(F)}.$$

The Tutte polynomial and the Jones polynomial for alternating links. Let K be an alternating link with an alternating plane projection $P(K)$. The region of the projection can be colored with two colors, say, purple and gold, such that two adjacent faces have different colors.

Two graphs are assigned to the projection, one corresponding to the purple regions and one to the golden regions. Every region gives rise to a vertex in the graph and two vertices are connected by an edge if the corresponding regions are adjacent to a common crossing. Such graphs are called checkerboard graphs.

Each edge comes with a sign as in Figure 2.

For an alternating link K all edges are either positive or negative. Thus we have a positive checkerboard graph and a negative checkerboard graph. These two graphs

Figure 2. A positive and a negative sign for the shaded region in the checkerboard graph.

are dual to each other. Let G be the positive checkerboard graph, a be the number of vertices in G and b be the number of vertices in the negative checkerboard graph.

The Jones Polynomial of an alternating link K with positive checkerboard graph G satisfies

$$V_K(t) = (-1)^w t^{(b-a+3w)/4} T_G(-t, -1/t);$$

see [Bollobás 1998], for instance. Here w is the writhe number, that is, the algebraic crossing number of the link projection.

Since we are interested in the absolute values of the Jones coefficients, all information relevant to us is contained in the evaluation $T_G(-t, -1/t)$ of the Tutte polynomial.

Reduction of multiple edges to simple edges. Our first step is to reduce the computation of the Tutte polynomial of a multigraph to the computation of a weighted Tutte polynomial of a spanning simple graph.

If $G = (V, E)$ is a connected graph without vertices of valence 1 (that is, without loops) and $\tilde{G} = (V, \tilde{E})$ is a spanning simple graph for it, we have

$$\begin{aligned} T_G\Big(-t, -\frac{1}{t}\Big) &= \sum_{F\subseteq E} (-t-1)^{k(F)-1}\Big(-\frac{1}{t}-1\Big)^{|F|-|V|+k(F)} \\ &= \sum_{\tilde{F}\subseteq \tilde{E}} (-t-1)^{k(\tilde{F})-1}\Big(-\frac{1}{t}-1\Big)^{-|V|+k(\tilde{F})} \\ &\quad \times\Bigg(\sum_{\substack{r(e_1)=1,\dots,r(e_j)=1 \\ e_1,\dots,e_j\in\tilde{F}}}^{\mu(e_1),\dots,\mu(e_j)} \binom{\mu(e_1)}{r(e_1)}\cdots\binom{\mu(e_j)}{r(e_j)}\Big(-\frac{1}{t}-1\Big)^{r(e_1)+\cdots+r(e_j)}\Bigg) \\ &= \sum_{\tilde{F}\subseteq\tilde{E}}\Bigg((-t-1)^{k(\tilde{F})-1}\Big(-\frac{1}{t}-1\Big)^{-|V|+k(\tilde{F})}\prod_{e\in\tilde{F}}\Big(\Big(-\frac{1}{t}\Big)^{\mu(e)}-1\Big)\Bigg). \end{aligned}$$

Setting $P(m) := \dfrac{(-1/t)^m - 1}{-1/t - 1} = 1 - t^{-1} + t^{-2} - \cdots \pm t^{-m+1}$, we have

$$(1)\quad T_G\Big(-t, -\frac{1}{t}\Big) = \sum_{\tilde{F}\subseteq\tilde{E}}\Bigg((-t-1)^{k(\tilde{F})-1}\Big(-\frac{1}{t}-1\Big)^{|\tilde{F}|-|V|+k(\tilde{F})}\prod_{e\in\tilde{F}} P(\mu(e))\Bigg).$$

Proposition 2.1 (Highest Tutte coefficients). *Let $G = (V, E)$ be a planar multigraph with spanning simple graph $\tilde{G} = (V, \tilde{E})$. Let the Tutte polynomial evaluate to*

$$T_G(-t, -1/t) = a_n t^n + a_{n+1} t^{n+1} + \cdots + a_{m-1} t^{m-1} + a_m t^m,$$

for suitable n and m.

Then the coefficients of the highest degree terms of $T_G(-t, -1/t)$ are:

(a) *The highest degree term t^m of $T_G(-t,-1/t)$ in t is $t^{|V|-1}$ with coefficient*

$$a_m = (-1)^{|V|-1}.$$

(b) *The second highest degree term is $t^{|V|-2}$, with coefficient*

$$a_{m-1} = (-1)^{|V|-1}\big(|V|-1-|\tilde{E}|\big).$$

Note that $|a_{m-1}| = |\tilde{E}|+1-|V|$.

(c) *The third highest degree term is $t^{|V|-3}$, with coefficient*

$$(-1)^{|V|}\left(-\binom{|V|-1}{2}+(|V|-2)|\tilde{E}|-n(2)-\binom{|\tilde{E}|}{2}+\text{tri}\right),$$

where tri *is the number of triangles in $\tilde{E}$. This term equals*

$$a_{m-2} = (-1)^{|V|}\left(-\binom{|a_{m-1}|+1}{2}-n(2)+\text{tri}\right).$$

Proof. It is easy to see that $|\tilde{F}|-|V|+k(F)\geq 0$ for all F. Therefore,

$$\left(-\frac{1}{t}-1\right)^{|\tilde{F}|-|V|+k(F)}\prod_{e\in\tilde{F}}P(\mu(e)) = \pm 1 + \text{ higher terms in } t^{-1}$$

This means that to determine the highest terms of $T_G(-t,-1/t)$ we have to analyze terms where $k(\tilde{F})$ is large.

Case $k(\tilde{F})=|V|$: this means that $|\tilde{F}|=0$. Thus the contribution in the sum in (1) is

$$(-t-1)^{|V|-1} = (-1)^{|V|-1}\left(t^{|V|-1}+(|V|-1)\,t^{|V|-2}+\binom{|V|-1}{2}t^{|V|-3}+\cdots+1\right).$$

Case $k(\tilde{F})=|V|-1$: this means that $|\tilde{F}|=1$. Thus the contribution is

$$(-t-1)^{|V|-2}\sum_{e\in\tilde{E}}P(\mu(e)).$$

Recalling that $n(j)$ is the number of edges in $\tilde{E}$ of multiplicity $\geq j$, we have

$$\sum_{e\in\tilde{E}}P(\mu(e)) = |\tilde{E}|-n(2)t^{-1}+n(3)t^{-2}-n(4)t^{-3}+\cdots.$$

Case $k(\tilde{F}) = |V| - 2$: this means that $|\tilde{F}|$ equals 2 or that $\tilde{F}$ is a triangle and $|\tilde{F}|$ equals 3. Thus the contribution is

$$\sum_{e,f\in\tilde{E}} (-t-1)^{|V|-3}\, P(\mu(e))\, P(\mu(f)) + \sum_{\substack{e,f,g\in\tilde{E}\\ (e,f,g)\text{ triangle}}} (-t-1)^{|V|-3}\Big(-\frac{1}{t}-1\Big) P(\mu(e))P(\mu(f))P(\mu(g)).$$

By combining these computations we get the result. □

3. An algebraic point of view

It is interesting to formulate the results of Proposition 2.1 in a purely algebraic way, as follows.

Let G be a multigraph and A its $N \times N$ adjacency matrix, so in particular $n(2)$ equals half the number of entries in A that exceed 1. Let $\tilde{A}$ be the matrix obtianed from A by replacing every nonzero entry A by 1. Thus, $\tilde{A}$ has only 1 and 0 as entries; further, the trace of $\tilde{A}^2$ is twice the number of edges of $\tilde{G}$ and the trace of $\tilde{A}^3$ is six times the number edges in $\tilde{G}$ (see [Biggs 1993], for example). Combining this with Proposition 2.1 immediately yields:

Corollary 3.1. *Let*

$$T_G(-t,-1/t) = a_n t^n + a_{n+1}t^{n+1} + \cdots + a_{m-1}t^{m-1} + a_m t^m$$

be the Jones evaluation of the Tutte Polynomial of a planar graph G.

$$\begin{aligned} |a_m| &= 1,\\ |a_{m-1}| &= \tfrac{1}{2}\operatorname{trace}\tilde{A}^2 - 1 - N,\\ |a_{m-2}| &= \binom{|a_{m-1}|+1}{2} + n(2) - \tfrac{1}{6}\operatorname{trace}\tilde{A}^3. \end{aligned}$$

4. The twist number and the volume of an hyperbolic alternating knot

(For information on hyperbolic structures on knot complements see [Callahan and Reid 1998], for instance.)

The figure-eight knot has minimal volume among all hyperbolic knot complements [Cao and Meyerhoff 2001]. For a hyperbolic knot K with crossing number $c > 4$, by a result of Colin Adams quoted in [Callahan and Reid 1998], the hyperbolic volume of the complement satisfies

$$\operatorname{Vol}(S^3 - K) \le (4c - 16)v_3,$$

where v_3 is the volume of a regular ideal hyperbolic tetrahedron.

Figure 3. A twist in a diagram of a knot .

For alternating knot complements a better general upper bound is known in terms of the twist number. As shown by Bill Menasco [1984], a nontorus alternating knot is hyperbolic.

The twist number of a diagram of an alternating knot is the minimal number of twists (see Figure 3) in it. Here, a twist can consist of a single crossing. The knot diagram shown on page 287 has twist number 8.

A twist corresponds to parallel edges in one of the checkerboard graphs. Let D be a diagram for an alternating knot K, and let $G = (V, E)$ and $G^* = (V^*, E^*)$ be the two checkerboard graphs, which are dual to one another, so $|E| = |E^*|$. We can now define the twist number by

$$\begin{aligned} T(K) &:= |E| - (|E| - |\tilde{E}|) - (|E^*| - |\tilde{E}^*|) \\ &= |E| - (|E| - |\tilde{E}|) - (|E| - |\tilde{E}^*|) = |\tilde{E}| + |\tilde{E}^*| - |E|. \end{aligned} \tag{2}$$

It is an easy exercise to see that

(a) $T(K)$ is indeed realized as the twist number of a diagram of K, and

(b) $T(K)$ is an invariant of all alternating projections of K. This follows from the Tait–Menasco–Thistlethwaite flyping theorem [Menasco and Thistlethwaite 1993]. Below we will give a different argument for it.

Theorem 4.1 (Lackenby [2004], Agol, D. Thurston).

$$v_3(T(K) - 2) \le \mathrm{Vol}(S^3 - K) < 10\, v_3(T(K) - 1),$$

where $\mathrm{Vol}(S^3 - K)$ *is the hyperbolic volume and* v_3 *is the volume of an ideal regular hyperbolic tetrahedron.*

Using work of Perelman the lower bound was improved by Agol, Storm and W. Thurston [Agol et al. 2005] to

$$\tfrac{1}{2} v_8 (T(K) - 2) \le \mathrm{Vol}(S^3 - K),$$

where $v_8 \approx 3.66386$ is the volume of an ideal regular hyperbolic octahedron.

5. Coefficients of the Jones polynomial

Let K be an alternating knot with reduced alternating diagram D having c crossings. From [Thistlethwaite 1987; Kauffman 1987; Murasugi 1987] we know that:

(a) the span of the Jones polynomial is c;

(b) the signs of the coefficients are alternating;

(c) the absolute values of the highest and lowest coefficients are 1.

Proposition 2.1 immediately leads to:

Theorem 5.1. *Let $V_K(t) = a_n t^n + a_{n+1}t^{n+1} + \cdots + a_m t^m$ be the Jones polynomial of an alternating knot K and let $G = (V, E)$ be a checkerboard graph of a reduced alternating projection of K. Then*:

(a) $|a_n| = |a_m| = 1$.

(b) $|a_{n+1}| + |a_{m-1}| = T(K)$.

(c) $|a_{n+2}| + |a_{m-2}| + |a_{m-1}||a_{n+1}| = \dfrac{T(K)+T(K)^2}{2} + n(2) + n^*(2) - \mathrm{tri} - \mathrm{tri}^*$,

where $n(2)$ is the number of edges in $\tilde{E}$ of multiplicity > 1 and $n^*(2)$ the corresponding number in the dual checkerboard graph.

The number tri *is the number of triangles in the graph $\tilde{G} = (V, \tilde{E})$ and* tri^* *corresponds to* tri *in the dual graph.*

(d) *In particular, the twist number is an invariant of reduced alternating projections of the knot.*

Proof. Let K be as in the statement, and let $G^* = (V^*, E^*)$ be the checkerboard graph dual to $G(V, E)$. We have $|E| = |E^*|$ and $|V| + |V^*| = |E| + 2$. Next recall from Equation (2) the definition of $T(K)$, which leads to

$$T(K) = (|\tilde{E}| - |V| + 1) + (|\tilde{E}^*| - |V^*| + 1).$$

The identities in the theorem then follow from Proposition 2.1. □

Volumish Theorem. *For an alternating, prime, nontorus knot K let*

$$V_K(t) = a_n t^n + \cdots + a_m t^m$$

be the Jones polynomial of K. Then

$$v_8(\max(|a_{m-1}|, |a_{n+1}|) - 1) \le \mathrm{Vol}(S^3 - K) \le 10\, v_3(|a_{n+1}| + |a_{m-1}| - 1).$$

Here, $v_3 \approx 1.01494$ is the volume of an ideal regular hyperbolic tetrahedron and $v_8 \approx 3.66386$ is the volume of an ideal regular hyperbolic octahedron.

Proof. The upper bound follows from Theorem 4.1 and 5.1. For the lower bound we need a closer look at [Lackenby 2004].

We can suppose that K admits a diagram such that both checkerboard graphs are imbedded so that every pair of edges connecting the same two vertices are adjacent to each other in the plane. This can be done through flypes.

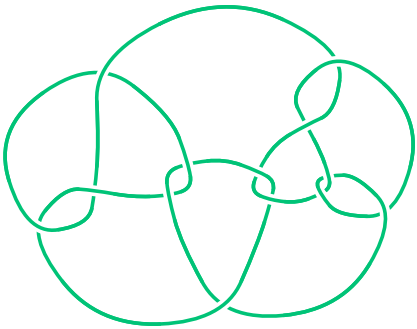


Figure 4. The alternating knot 13.123 in the Knotscape Census.

Suppose $G_p = (V_p, E_p)$ is the positive (colored in purple) and $G_g = (V_g, E_g)$ is the negative (colored in gold) checkerboard graph. Since $G_p^* = G_g$ we have $|E_p| = |E_g|$ and

$$|V_p| - |E_p| + |V_g| = 2 = |V_p| - |E_g| + |V_g|.$$

Let r_p and r_g be the number of vertices in G_p and G_g having valence at least 3. It is proved in [Lackenby 2004] (and the bound was improved in [Agol et al. 2005]) that

$$\mathrm{Vol}(S^3 - K) \geq v_8(\max(r_p, r_g) - 2).$$

If $\tilde{G}_p = (V_p, \tilde{E}_p)$ and $\tilde{G}_g = (V_g, \tilde{E}_g)$ are the reduced graphs of G_p and G_g than it is easy to see that

$$r_p = |V_p| - (|E_g| - |\tilde{E}_g|) = 2 - |V_g| + |\tilde{E}_g| = |a_{n+1}| + 1.$$

Similarly, $r_g = |a_{m-1}| + 1$ and the lower bound follows. □

Example. The checkerboard graph G of the knot in Figure 4 has $|V| = 8$ vertices, $|\tilde{E}| = 11$, $n(2) = 2$ and $\mathrm{tri} = 1$.

Its dual has $|V^*| = 7$ vertices, $|\tilde{E}^*| = 10$, $n^*(2) = 3$ and $\mathrm{tri}^* = 2$. Therefore, with the preceding notation for the coefficients of the Jones polynomial,

$$\begin{aligned}
|a_n| &= 1,\\
|a_{n+1}| &= |\tilde{E}| + 1 - |V| = 4,\\
|a_{n+2}| &= \binom{|a_{n+1}|+1}{2} + n(2) - \mathrm{tri} = 10 + 2 - 1 = 11,\\
|a_m| &= 1,\\
|a_{m-1}| &= |\tilde{E}^*| + 1 - |V^*| = 4,\\
|a_{m-2}| &= \binom{|a_{m-1}|+1}{2} + n^*(2) - \mathrm{tri}^* = 10 + 3 - 2 = 11.
\end{aligned}$$

The complete Jones polynomial of the knot is, according to Knotscape,

$$V_{13.121}(t) = t^{-12} - 4t^{-11} + 11t^{-10} - 23t^{-9} + 35t^{-8} - 47t^{-7} + 53t^{-6} - 52t^{-5} + 47t^{-4} - 34t^{-3} + 22t^{-2} - 11t^{-1} + 4 - t,$$

and the hyperbolic volume is

$$\mathrm{Vol}(S^3 - K) \approx 21.1052106828.$$

Appendix

Higher twist numbers of prime alternating knots on* 14 *crossings. Here we give experimental data on the relationship between the twist number, as computed using the Jones polynomial, and the hyperbolic volume of knots. All data are taken from Knotscape, written by Jim Hoste, Morwen Thistlethwaite and Jeff Weeks [Hoste et al. 1998]. We confined ourselves to knots with crossing number 14. As before, let $V_K(t) = a_n t^n + a_{n+1}t^{n+1} + \cdots + a_m t^m$ be the Jones polynomial of an alternating prime knot K.

As shown, the twist number is $T(K) = |a_{n+1}| + |a_{m-1}|$. We call $T_i(K) = |a_{n+i}| + |a_{m-i}|$ the higher twist numbers. In particular, $T(L) = T_1(L)$.

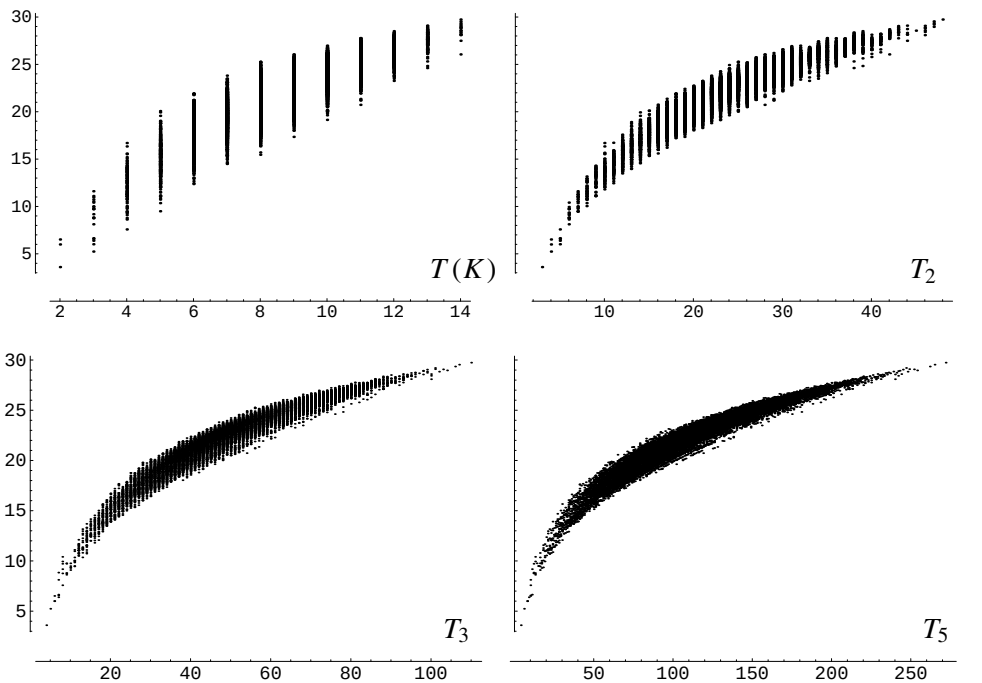


Figure 5. Twist numbers vs. volume correlation for 14-crossing alternating knows. Each dot stands for a knot in the census.

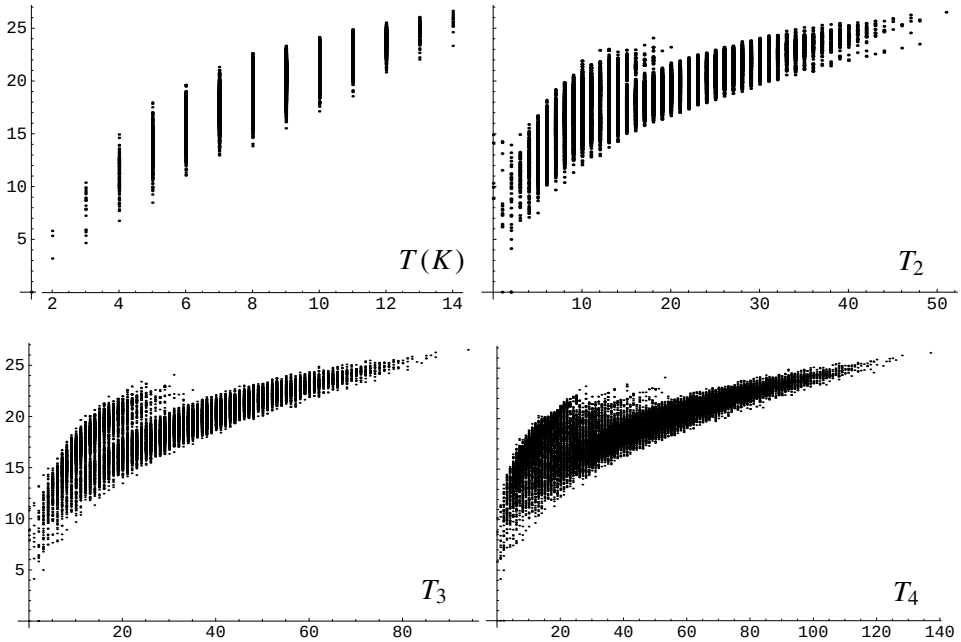


Figure 6. Twist numbers vs. volume correlation for 14-crossing nonalternating knots. Nonhyperbolic knots are assigned zero volume.

Higher twist numbers of prime nonalternating knots on* 14 *crossings. For nonalternating knots we keep the notation, although there is no direct geometrical justification known:

Again, let $V_L(t) = a_n t^n + a_{n+1}t^{n+1} + \cdots + a_m t^m$ be the Jones polynomial of a nonalternating knot L.

Define the twist number as $T(L) = |a_{n+1}| + |a_{m-1}|$. As in the alternating case, we call $T_i(L) = |a_{n+i}| + |a_{m-i}|$ the higher twist numbers. In particular, $T(L) = T_1(L)$.

The pictures give, for nonalternating knots with crossing number 14, the relation between the twist number (or one of the higher twist numbers T_2, T_3, T_4) and the volume.

Acknowledgment

We thank Ian Agol, Joan Birman, Charlie Frohman, Vaughan Jones, Lou Kauffman, James Oxley, and Neal Stoltzfus for helpful conversations at various occasions.

We also thank Alexander Stoimenow for pointing out some typos and Stavros Garoufalidis for historical remarks.

References

[Agol et al. 2005] I. Agol, P. Storm, and W. P. Thurston, "Lower bounds on volumes of hyperbolic Haken 3-manifolds", preprint, 2005, Available at arXiv:math.DG/0506338. With an appendix by N. Dunfield.

[Bar-Natan and Garoufalidis 1996] D. Bar-Natan and S. Garoufalidis, "On the Melvin–Morton–Rozansky conjecture", *Invent. Math.* **125**:1 (1996), 103–133. MR 97i:57004 Zbl 0855.57004

[Biggs 1993] N. Biggs, *Algebraic graph theory*, 2nd ed., Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1993. MR 95h:05105 Zbl 0797.05032

[Bollobás 1998] B. Bollobás, *Modern graph theory*, Graduate Texts in Mathematics **184**, Springer, New York, 1998. MR 99h:05001 Zbl 0902.05016

[Callahan and Reid 1998] P. J. Callahan and A. W. Reid, "Hyperbolic structures on knot complements", *Chaos Solitons Fractals* **9**:4-5 (1998), 705–738. MR 99e:57022 Zbl 0935.57017

[Cao and Meyerhoff 2001] C. Cao and G. R. Meyerhoff, "The orientable cusped hyperbolic 3-manifolds of minimum volume", *Invent. Math.* **146**:3 (2001), 451–478. MR 2002i:57016 Zbl 1028.57010

[Chmutov 1998] S. Chmutov, "A proof of the Melvin–Morton conjecture and Feynman diagrams", *J. Knot Theory Ramifications* **7**:1 (1998), 23–40. MR 99c:57012 Zbl 0892.57002

[Dasbach and Hougardy 1997] O. T. Dasbach and S. Hougardy, "Does the Jones polynomial detect unknottedness?", *Experiment. Math.* **6**:1 (1997), 51–56. MR 98c:57004 Zbl 0883.57006

[Hoste et al. 1998] J. Hoste, M. Thistlethwaite, and J. Weeks, "The first 1,701,936 knots", *Math. Intelligencer* **20**:4 (1998), 33–48. MR 99i:57015 Zbl 0916.57008

[Kauffman 1987] L. H. Kauffman, "State models and the Jones polynomial", *Topology* **26**:3 (1987), 395–407. MR 88f:57006 Zbl 0622.57004

[Lackenby 2004] M. Lackenby, "The volume of hyperbolic alternating link complements", *Proc. London Math. Soc.* (3) **88**:1 (2004), 204–224. With an appendix by Ian Agol and Dylan Thurston. MR 2004i:57008 Zbl 1041.57002

[Lin and Wang 2001] X.-S. Lin and Z. Wang, "Random walk on knot diagrams, colored Jones polynomial and Ihara-Selberg zeta function", pp. 107–121 in *Knots, braids, and mapping class groups: papers dedicated to Joan S. Birman* (New York, 1998), edited by J. Gilman et al., AMS/IP Stud. Adv. Math. **24**, Amer. Math. Soc., Providence, RI, 2001. MR 2003f:57026 Zbl 0992.57001

[Menasco 1984] W. Menasco, "Closed incompressible surfaces in alternating knot and link complements", *Topology* **23**:1 (1984), 37–44. MR 86b:57004 Zbl 0525.57003

[Menasco and Thistlethwaite 1993] W. Menasco and M. Thistlethwaite, "The classification of alternating links", *Ann. of Math.* (2) **138**:1 (1993), 113–171. MR 95g:57015 Zbl 0809.57002

[Murakami and Murakami 2001] H. Murakami and J. Murakami, "The colored Jones polynomials and the simplicial volume of a knot", *Acta Math.* **186**:1 (2001), 85–104. MR 2002b:57005 Zbl 0983.57009

[Murasugi 1987] K. Murasugi, "Jones polynomials and classical conjectures in knot theory", *Topology* **26**:2 (1987), 187–194. MR 88m:57010 Zbl 0628.57004

[Rozansky 1997] L. Rozansky, "Higher order terms in the Melvin–Morton expansion of the colored Jones polynomial", *Comm. Math. Phys.* **183**:2 (1997), 291–306. MR 98k:57016 Zbl 0882.57004

[Thistlethwaite 1987] M. B. Thistlethwaite, "A spanning tree expansion of the Jones polynomial", *Topology* **26**:3 (1987), 297–309. MR 88h:57007 Zbl 0622.57003

[Vaintrob 1997] A. Vaintrob, "Melvin–Morton conjecture and primitive Feynman diagrams", *Internat. J. Math.* **8**:4 (1997), 537–553. MR 99b:57027 Zbl 0890.57003

Received January 18, 2006.

OLIVER T. DASBACH
LOUISIANA STATE UNIVERSITY
DEPARTMENT OF MATHEMATICS
BATON ROUGE, LA 70803
UNITED STATES

kasten@math.lsu.edu
http://www.math.lsu.edu/~kasten

XIAO-SONG LIN
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
RIVERSIDE, CA 92521-0135
UNITED STATES
Deceased January 14, 2007

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.293

ON THE LOCAL NIRENBERG PROBLEM FOR THE Q-CURVATURES

PHILIPPE DELANOË AND FRÉDÉRIC ROBERT

The local image of each conformal Q-curvature operator on the sphere admits no scalar constraint, although identities of Kazdan–Warner type hold for its graph.

1. Introduction

Let (m, n) be positive integers such that $n > 1$, and $n \geq 2m$ in case n is even. We work on the standard n-sphere $(\mathbb{S}^n, g_0)$, with pointwise conformal metric $g_u = e^{2u} g_0$. (All objects will be taken to be smooth.)

We are interested in the structure near $u = 0$ of the image of the conformal $2m$-th order Q-curvature increment operator $u \mapsto \boldsymbol{Q}_{m,n}[u] = Q_{m,n}(g_u) - Q_{m,n}(g_0)$ (see Section 2), thus considering a local Nirenberg-type problem (Nirenberg's problem was for $m = 1$; see, for example, [Moser 1973; Kazdan and Warner 1974; 1975; Aubin 1982, p. 122]). At the infinitesimal level, the situation looks as follows (dropping henceforth the subscript (m, n)):

Lemma 1.1. *Let $L = d\boldsymbol{Q}[0]$ stand for the linearization at $u = 0$ of the conformal Q-curvature increment operator and Λ_1, for the $(n+1)$-space of first spherical harmonics on $(\mathbb{S}^n, g_0)$. Then L is self-adjoint and* $\operatorname{Ker} L = \Lambda_1$.

Further, the graph $\Gamma(\boldsymbol{Q}) := \{(u, \boldsymbol{Q}[u]), u \in C^\infty(\mathbb{S}^n)\}$ of $\boldsymbol{Q}$ in $C^\infty(\mathbb{S}^n) \times C^\infty(\mathbb{S}^n)$ admits scalar constraints which are the analogue for $\boldsymbol{Q}$ of the so-called Kazdan–Warner identities for the conformal scalar curvature (i.e., the case $m = 1$); see [Kazdan and Warner 1974; 1975; Bourguignon and Ezin 1987]. Here, a scalar constraint means a real-valued submersion defined near $\Gamma(\boldsymbol{Q})$ in $C^\infty(\mathbb{S}^n) \times C^\infty(\mathbb{S}^n)$ and vanishing on $\Gamma(\boldsymbol{Q})$. Specifically:

Theorem 1.2. *For each $(u, q) \in C^\infty(\mathbb{S}^n) \times C^\infty(\mathbb{S}^n)$ and each conformal Killing vector field X on $(\mathbb{S}^n, g_0)$, the condition $(u, q) \in \Gamma(\boldsymbol{Q})$ implies the vanishing of the*

Delanoë is supported by the CNRS.
MSC2000: primary 58J05; secondary 53C99.
Keywords: conformal Q-curvature, Nirenberg problem, Fredholm theory, Paneitz–Branson operators, local image, Kazdan–Warner identities.

integral $\int_{\mathbb{S}^n}(X\cdot q)\,d\mu_u$, *where* $d\mu_u = e^{nu}d\mu_0$ *stands for the Lebesgue measure of the metric* g_u. *In particular, there is no solution* $u \in C^\infty(\mathbb{S}^n)$ *to the equation*

$$Q(g_u) = z + \text{constant} \quad \text{with } z \in \Lambda_1.$$

Due to the naturality of Q (Remark 3.1) and the self-adjointness of $d\boldsymbol{Q}[u]$ in $L^2(M_n, d\mu_u)$ (Remarks 3.2 and 3.3), this theorem holds as a particular case of the more general Theorem 2.1 below.

Can one do better than Theorem 1.2 and drop the u variable occurring in the constraints and find constraints bearing on the sole image of the operator $\boldsymbol{Q}$? Since L is self-adjoint in $L^2(\mathbb{S}^n, g_0)$ (see [Graham and Zworski 2003]), Lemma 1.1 shows that the map $u \mapsto \boldsymbol{Q}[u]$ misses infinitesimally at $u = 0$ a vector space of dimension $n+1$. How does this translate at the local level? Calling a real-valued map K a *scalar constraint for the local image of* $\boldsymbol{Q}$ *near* 0 if K is a submersion defined near 0 in $C^\infty(\mathbb{S}^n)$ such that $K \circ \boldsymbol{Q} = 0$ near 0 in $C^\infty(\mathbb{S}^n)$, a spherical symmetry argument as in [Delanoë 2003, Corollary 5] shows that if the local image of $\boldsymbol{Q}$ admits a scalar constraint near 0, it must admit $n+1$ independent such constraints, the maximal number to be expected. In this context, our main result is quite in contrast with Theorem 1.2:

Theorem 1.3. *The local image of* $\boldsymbol{Q}$ *near*, 0 *admits no scalar constraint.*

The picture about the local image of the Q-curvature increment operator on $(\mathbb{S}^n, g_0)$ is completed with a remark:

Remark 1.4. The local Nirenberg problem for $\boldsymbol{Q}$ near 0 is governed by the nonlinear Fredholm formula (9) below. Thus, as in [Delanoë 2003, Corollary 5], a local result of Moser type [1973] holds: If $f \in C^\infty(\mathbb{S}^n)$ is close enough to zero and invariant under a nontrivial group of isometries of $(\mathbb{S}^n, g_0)$ acting without fixed points,[1] then $\mathfrak{D}(f) = 0$ in (9), so f lies in the local image of $\boldsymbol{Q}$.

The outline of the paper is as follows. In Section 2 we present an independent account on general Kazdan–Warner type identities, implying Theorem 1.2. Then we focus on Theorem 1.3: we recall basic facts for the Q-curvature operators on spheres in Section 3 and sketch the proof of Theorem 1.3 in Section 4, relying on [Delanoë 2003] and reducing it to Lemma 1.1 and another key lemma. In the last two sections we carry out the proofs of these lemmas, deferring to an Appendix some eigenvalues calculations.

2. General identities of Kazdan–Warner type

The following statement is essentially due to Jean–Pierre Bourguignon [1986]:

[1] This condition is more general than a free action.

Theorem 2.1. *Let M_n be a compact n-manifold and $g \mapsto D(g) \in C^\infty(M)$ be a scalar differential operator defined on the open cone of Riemannian metrics on M_n, and natural in the sense of* [Stredder 1975] (see (5) below). *Given a conformal class $\boldsymbol{c}$ and a Riemannian metric $g_0 \in \boldsymbol{c}$, sticking to the notation $g_u = e^{2u} g_0$ for $u \in C^\infty(M)$, consider the operator $u \mapsto \boldsymbol{D}[u] := D(g_u)$ and its linearization $L_u = d\boldsymbol{D}[u]$ at u. Assume that, for each $u \in C^\infty(M)$, the linear differential operator L_u is formally self-adjoint in $L^2(M, d\mu_u)$, where $d\mu_u = e^{nu} d\mu_0$ stands for the Lebesgue measure of g_u. Then, for any conformal Killing vector field X on $(M_n, \boldsymbol{c})$ and any $u \in C^\infty(M)$, we have*

$$\int_M X \cdot \boldsymbol{D}[u]\, d\mu_u = 0.$$

In particular, if $(M_n, \boldsymbol{c})$ is equal to $\mathbb{S}^n$ equipped with its standard conformal class, there is no solution $u \in C^\infty(\mathbb{S}^n)$ to the equation

$$\boldsymbol{D}[u] = z + \text{constant} \quad \textit{with } z \in \Lambda_1$$

(*a first spherical harmonic*).

Proof. We rely on Bourguignon's functional integral invariants approach and follow the proof of [Bourguignon 1986, Proposition 3] (using freely notations from p. 101 of the same paper), presenting its functional geometric framework with some care. We consider the affine Fréchet manifold Γ whose generic point is the volume form (possibly of odd type in case M is not orientable [de Rham 1960]) of a Riemannian metric $g \in \boldsymbol{c}$; we denote by ω_g the volume form of a metric g (recall the tensor ω_g is natural [Stredder 1975, Definition 2.1]). The metric $g_0 \in \boldsymbol{c}$ yields a global chart of Γ defined by

$$\omega_g \in \Gamma \quad \mapsto \quad u := \frac{1}{n} \log \frac{d\omega_g}{d\omega_{g_0}} \in C^\infty(M_n)$$

(viewing volume-forms like measures and using the Radon–Nikodym derivative) — in other words, such that $\omega_g = e^{nu} \omega_{g_0}$; changes of such charts are indeed affine (and pure translations). It will be easier, though, to avoid the use of charts on Γ, except for proving that a 1-form is closed (see below). The tangent bundle to Γ is trivial, equal to $T\Gamma = \Gamma \times \Omega^n(M_n)$ (setting $\Omega^k(A)$ for the k-forms on a manifold A), and there is a canonical Riemannian metric on Γ of Fischer type [Friedrich 1991], given at $\omega_g \in \Gamma$ by

$$\langle v, w \rangle := \int_M \frac{dv}{d\omega_g} \frac{dw}{d\omega_g} \omega_g \quad \text{for } (v, w) \in T_{\omega_g}\Gamma.$$

From Riesz's theorem, a tangent covector $a \in T^*_{\omega_g}\Gamma$ may thus be identified with a tangent vector $a^\sharp \in \Omega^n(M_n)$ or else with the function $da^\sharp/d\omega_g =: \rho_g(a) \in C^\infty(M_n)$

such that

$$a(\varpi) = \int_M \rho_g(a)\varpi \quad \text{for } \varpi \in T_{\omega_g}\Gamma. \tag{1}$$

We also consider the Lie group G of conformal maps on $(M_n, \boldsymbol{c})$, acting on the manifold Γ by

$$(\varphi, \omega_g) \in G \times \Gamma \to \varphi^*\omega_g \in \Gamma$$

(indeed, we have $\varphi^*\omega_g = \omega_{\varphi^* g}$ by naturality and $\varphi \in G \Rightarrow \varphi^* g \in \boldsymbol{c}$). For each conformal Killing field X on $(M_n, \boldsymbol{c})$, the flow of X as a map $t \in \mathbb{R} \to \varphi_t \in G$ yields a vector field $\bar{X}$ on Γ defined by

$$\omega_g \mapsto \bar{X}(\omega_g) := \frac{d}{dt}\left(\varphi_t^*\omega_g\right)_{t=0} \equiv L_X\omega_g$$

(L_X standing here for the Lie derivative on M_n). In this context, regardless of any Banach completion, one may define the (global) flow $t \in \mathbb{R} \to \bar{\varphi}_t \in \mathrm{Diff}(\Gamma)$ of $\bar{X}$ on the Fréchet manifold Γ by setting

$$\bar{\varphi}_t(\omega_g) := \varphi_t^*\omega_g \quad \text{for } \omega_g \in \Gamma;$$

indeed, the latter satisfies

$$\frac{d}{dt}\left(\varphi_t^*\omega_g\right) = \varphi_t^*(L_X\omega_g) \equiv L_X(\varphi_t^*\omega_g) = \bar{X}\left[\bar{\varphi}_t(\omega_g)\right]$$

(see [Kobayashi and Nomizu 1963, p. 33], for example). With the flow $(\bar{\varphi}_t)_{t\in\mathbb{R}}$ at hand, we can define the Lie derivative $L_{\bar{X}}$ of forms on Γ as usual, by setting $L_{\bar{X}}a := (d/dt)\,(\bar{\varphi}_t{}^*a)_{t=0}$. Finally, one can check Cartan's formula for $\bar{X}$, namely

$$L_{\bar{X}} = i_{\bar{X}}d + d i_{\bar{X}}, \tag{2}$$

where $i_{\bar{X}}$ denotes the interior product with $\bar{X}$, by verifying it for a generic function f on Γ and for its exterior derivative df (with d defined as in [Lang 1962]).

Following [Bourguignon 1986], and using our global chart $\omega_g \mapsto u$, we apply (2) to the 1-form σ on Γ defined at ω_g by the function $\rho_g(\sigma) := \boldsymbol{D}[u]$; see (1). Arguing as on p. 102 of the same reference, one readily verifies in the chart u (and using constant local vector fields on Γ) that the 1-form σ is closed due to the self-adjointness of the linearized operator L_u in $L^2(M_n, d\mu_u)$; furthermore (dropping the chart u), one derives at once the G-invariance of σ from the naturality of $g \mapsto D(g)$. We thus have $d\sigma = 0$ and $L_{\bar{X}}\sigma = 0$, hence $d(i_{\bar{X}}\sigma) = 0$ by (2). So the function $i_{\bar{X}}\sigma$ is constant on Γ; in other words, $\int_M \boldsymbol{D}[u]\, L_X\omega_u$ is independent of u, or else, integrating by parts, so is $\int_M X \cdot \boldsymbol{D}[u]\, d\mu_u$ (where $X\cdot$ stands for X acting as a derivation on real-valued functions on M_n).

To complete the proof of the first part of Theorem 2.1, we show that the integrand $X\cdot D(g_0)$ of the latter expression at $u = 0$ vanishes for a suitable choice of the metric

g_0 in the conformal class $\boldsymbol{c}$. We recall the Ferrand–Obata theorem [Lelong-Ferrand 1969; Obata 1971/72], according to which either the conformal group G is compact or $(M_n, \boldsymbol{c})$ is equal to $\mathbb{S}^n$ equipped with its standard conformal class. In the former case, averaging on G, we may pick $g_0 \in \boldsymbol{c}$ invariant under the action of G: with this choice, $D(g_0)$ is also G-invariant by naturality, hence $X \cdot D(g_0) \equiv 0$ as needed. In the latter case, as observed in the proof of Proposition 4.2 below, $D(g_0)$ is constant on $\mathbb{S}^n$, and again the desired result follows.

Finally, the last assertion of the theorem (consistently with Proposition 4.2 below and the Fredholm theorem if L_0 is elliptic) follows from the first one, by taking for the vector field X the gradient of z with respect to the standard metric of $\mathbb{S}^n$, which is known to be conformal Killing. □

3. Back to Q-curvatures on spheres: basic facts recalled

The special case $n = 2m$. Here we will consider the Q-curvature increment operator given by $\boldsymbol{Q}[u] = Q(g_u) - Q_0$, with

$$Q(g_u) = e^{-2mu}(Q_0 + P_0[u]) \tag{3}$$

where $Q_0 = Q(g_0)$ is equal to $Q_0 = (2m-1)!$ on $(\mathbb{S}^n, g_0)$, and where

$$P_0 = \prod_{k=1}^{m} \big(\Delta_0 + (m-k)(m+k-1)\big) \tag{4}$$

(see [Branson 1987; Beckner 1993]), Δ_0 denoting the positive laplacian relative to g_0. We call P_0 the Paneitz–Branson operator of the metric g_0.

Remark 3.1. Following [Branson 1995], one can define a Paneitz–Branson operator P_0 for *any* metric g_0 (given by a formula more general than (4) of course), and a Q-curvature $Q(g_0)$ transforming like (3) under the conformal change of metrics $g_u = e^{2u} g_0$. Importantly then, the map $g \mapsto Q(g) \in C^\infty(\mathbb{S}^n)$ is natural, meaning (see [Stredder 1975, Definition 2.1], for instance) that any diffeomorphism ψ satisfies

$$\psi^* Q(g) = Q(\psi^* g). \tag{5}$$

Remark 3.2. From (3) and the formal self-adjointness of P_0 in $L^2(\mathbb{S}^n, d\mu_0)$ [Graham and Zworski 2003, p. 91], one readily verifies that, for each $u \in C^\infty(\mathbb{S}^n)$, the linear differential operator $d\boldsymbol{Q}[u]$ is formally self-adjoint in $L^2(\mathbb{S}^n, d\mu_u)$.

The case $n \neq 2m$. The expression of the Paneitz–Branson operator on $(\mathbb{S}^n, g_0)$ becomes

$$P_0 = \prod_{k=1}^{m} \big(\Delta_0 + \big(\tfrac{1}{2}n - k\big)\big(\tfrac{1}{2}n + k - 1\big)\big) \tag{6}$$

(see [Guillarmou and Naud 2006, Proposition 2.2]), while that for the metric $g_u = e^{2u} g_0$ is given by

$$P_u(\cdot) = e^{-(\frac{1}{2}n+m)u} P_0\big(e^{(\frac{1}{2}n-m)u}\,\cdot\big), \tag{7}$$

with the Q-curvature of g_u given accordingly by $\big(\frac{1}{2}n-m\big)\,Q(g_u) = P_u(1)$. The analogue of Remark 3.1 still holds (now see [Graham et al. 1992; Graham and Zworski 2003]). We will consider the (renormalized) Q-curvature increment operator $\boldsymbol{Q}[u] = \big(\frac{1}{2}n-m\big)\,(Q(g_u) - Q_0)$, now with

$$\big(\tfrac{1}{2}n-m\big)\,Q_0 = \big(\tfrac{1}{2}n-m\big)\,Q(g_0) = P_0(1) = \prod_{k=0}^{2m-1}\big(k+\tfrac{1}{2}n-m\big). \tag{8}$$

Remark 3.3. Finally, we note again that the linearized operator $d\,\boldsymbol{Q}[u]$ is formally self-adjoint in $L^2(\mathbb{S}^n, d\mu_u)$. Indeed, a straightforward calculation yields

$$d\,\boldsymbol{Q}[u](v) = \big(\tfrac{1}{2}n-m\big)\,P_u(v) - \big(\tfrac{1}{2}n+m\big)\,P_u(1)\,v,$$

and the Paneitz–Branson operator P_u is known to be self-adjoint in $L^2(\mathbb{S}^n, d\mu_u)$ [Graham and Zworski 2003, p. 91].

For later use, and in all the cases for (m, n), we will set p_0 for the degree m polynomial such that $P_0 = p_0(\Delta_0)$.

4. Proof of Theorem 1.3

The case $m = 1$ was settled in [Delanoë 2003] with a proof robust enough to be followed again. For completeness, let us recall how it goes (see [Delanoë 2003] for details).

If $\mathcal{P}_1$ stands for the orthogonal projection of $L^2(\mathbb{S}^n, g_0)$ onto Λ_1, Lemma 1.1 and the self-adjointness of L imply [Delanoë 2003, Theorem 7] that the modified operator

$$u \mapsto \boldsymbol{Q}[u] + \mathcal{P}_1 u$$

is a local diffeomorphism of a neighborhood of 0 in $C^\infty(\mathbb{S}^n)$ onto another one: set $\mathcal{S}$ for its inverse and $\mathcal{D} = \mathcal{P}_1 \circ \mathcal{S}$ (defect map). Then $u = \mathcal{S} f$ satisfies the local nonlinear Fredholm-like equation

$$\boldsymbol{Q}[u] = f - \mathcal{D}(f). \tag{9}$$

By [Delanoë 2003, Theorem 2], if a local constraint exists for $\boldsymbol{Q}$ at 0, then $\mathcal{D} \circ \boldsymbol{Q} = 0$ (recalling the symmetry fact above). Fixing $z \in \Lambda_1$, we will prove Theorem 1.3 by showing that $\mathcal{D} \circ \boldsymbol{Q}[tz] \neq 0$ for small $t \in \mathbb{R}$; here is how. On the one hand, setting

$$u_t = \mathcal{S} \circ \boldsymbol{Q}[tz] := t u_1 + t^2 u_2 + t^3 u_3 + O(t^4),$$

Lemma 1.1 yields $u_1 = 0$; also, as and easily verified general fact, we have

$$\boldsymbol{Q}[u_t] + \mathcal{P}_1 u_t = t^2(L + \mathcal{P}_1)u_2 + t^3(L + \mathcal{P}_1)u_3 + O(t^4). \tag{10}$$

On the other hand, consider the expansion of $\boldsymbol{Q}[tz]$:

$$\boldsymbol{Q}[tz] = t^2 c_2[z] + t^3 c_3[z] + O(t^4), \tag{11}$$

and focus on its third order coefficient $c_3[z]$, for which we will prove:

Lemma 4.1. *Let (m, n) be positive integers such that $n > 1$ and $n \geq 2m$ in case n is even. Then*

$$\int_{\mathbb{S}^n} z\, c_3[z]\, d\mu_0 \neq 0.$$

Granted this lemma, we are done: indeed, the equality

$$\boldsymbol{Q}[u_t] + \mathcal{P}_1 u_t = \boldsymbol{Q}[tz],$$

combined with (10)–(11), yields

$$(L + \mathcal{P}_1)u_3 = c_3[z],$$

which, integrated against z, implies that

$$\int_{\mathbb{S}^n} z \mathcal{P}_1 u_3\, d\mu_0 \neq 0$$

(recalling that L is self-adjoint and $z \in \operatorname{Ker} L$ by Lemma 1.1). Therefore $\mathcal{P}_1 u_3 \neq 0$, hence also $\mathcal{D} \circ \boldsymbol{Q}[tz] \neq 0$.

Thus we have reduced the proof of Theorem 1.3 to that of Lemmas 1.1 and 4.1, which we now present.

Proof of Lemma 1.1. (1) *Proof of the inclusion $\Lambda_1 \subset \operatorname{Ker} L$.* We need neither ellipticity nor conformal covariance for this inclusion to hold; the naturality property (5) suffices. We state a general result that implies at once what we need:

Proposition 4.2. *Let $g \mapsto D(g)$ be any scalar natural differential operator on $\mathbb{S}^n$, defined on the open cone of Riemannian metrics, valued in $C^\infty(\mathbb{S}^n)$. For each $u \in C^\infty(\mathbb{S}^n)$, set $\boldsymbol{D}[u] = D(g_u) - D(g_0)$ and $L = d\boldsymbol{D}[0]$, where $g_u = e^{2u} g_0$. Then $\Lambda_1 \subset \operatorname{Ker} L$.*

Proof. Let us first observe that $D(g_0)$ must be constant. Indeed, for each isometry ψ of $(\mathbb{S}^n, g_0)$, the naturality of D implies $\psi^* D(g_0) \equiv D(g_0)$; so the result follows because the group of such isometries acts transitively on $\mathbb{S}^n$. Morally, since g_0 has constant curvature, this result is also expectable from the theory of Riemannian invariants (see [Stredder 1975] and references therein), here though, without any regularity (or polynomiality) assumption.

Given an arbitrary nonzero $z \in \Lambda_1$, let $S = S(z) \in \mathbb{S}^n$ stand for its corresponding south pole (where $z(S) = -M$ is minimum) and, for each small real t, let ψ_t denote the conformal diffeomorphism of $\mathbb{S}^n$ fixing S and composed elsewhere of: Ster_S, the stereographic projection with pole S, the dilation $X \in \mathbb{R}^n \mapsto e^{Mt} X \in \mathbb{R}^n$, and the inverse of Ster_S. As t varies, the family ψ_t satisfies

$$\psi_0 = I, \qquad \frac{d}{dt}(\psi_t)_{t=0} = -\nabla_0 z,$$

where ∇_0 denotes the gradient relative to g_0. If we set $e^{2u_t} g_0 = \psi_t^* g_0$, we get

$$\frac{d}{dt}(u_t)_{t=0} \equiv z.$$

Recalling that $D(g_0)$ is constant, the naturality of D implies

$$\boldsymbol{D}[u_t] = \psi_t^* D(g_0) - D(g_0) = 0;$$

in particular, differentiating this equation at $t = 0$ yields $Lz = 0$ hence we conclude that $\Lambda_1 \subset \mathrm{Ker}\, L$.

(2) *Proof of the reverse inclusion* $\ker L \subset \Lambda_1$. For a contradiction, assume the existence of a nonzero $v \in \Lambda_1^\perp \cap \mathrm{Ker}\, L$. If $\mathcal{B}$ is an orthonormal basis of eigenfunctions of Δ_0 in $L^2(\mathbb{S}^n, d\mu_0)$, there exists an integer $i \neq 1$ and a function $\varphi_i \in \Lambda_i \cap \mathcal{B}$ (where Λ_i henceforth denotes the space of i-th spherical harmonics) such that

$$\int_{\mathbb{S}^n} \varphi_i v \, d\mu_0 \neq 0$$

(actually $i \neq 0$, due to $\int_{\mathbb{S}^n} v \, d\mu_0 = 0$, obtained just by averaging $Lv = 0$ on $\mathbb{S}^n$). By the self-adjointness of L, we may write

$$0 = \int_{\mathbb{S}^n} \varphi_i L v \, d\mu_0 = \int_{\mathbb{S}^n} v L \varphi_i \, d\mu_0,$$

then infer (see below) that

$$0 = \big(p_0(\lambda_i) - p_0(\lambda_1)\big) \int_{\mathbb{S}^n} \varphi_i v \, d\mu_0,$$

and finally get the desired contradiction, because $p_0(\lambda_i) \neq p_0(\lambda_1)$ for $i \neq 1$ (see the Appendix). Here, we used the following auxiliary facts, obtained by differentiating (3) or (7) at $u = 0$ in the direction of $w \in C^\infty(\mathbb{S}^n)$:

$$\begin{aligned} n = 2m &\;\Rightarrow\; Lw = P_0(w) - n!\, w, \\ n \neq 2m &\;\Rightarrow\; Lw = \left(\tfrac{1}{2}n - m\right) P_0(w) - \left(\tfrac{1}{2}n + m\right) p_0(\lambda_0) w. \end{aligned}$$

From $\Lambda_1 \subset \operatorname{Ker} L$, we get, taking $w = z \in \Lambda_1$:

$$\begin{aligned} &n = 2m \ \Rightarrow\ p_0(\lambda_1) - n! = 0, \\ &n \neq 2m \ \Rightarrow\ \left(\tfrac{1}{2}n - m\right) p_0(\lambda_1) - \left(\tfrac{1}{2}n + m\right) p_0(\lambda_0) = 0. \end{aligned} \tag{12}$$

Moreover, taking $w = \varphi_i \in \Lambda_i$, we then have

$$\begin{aligned} &n = 2m \ \Rightarrow\ L\varphi_i = \big(p_0(\lambda_i) - p_0(\lambda_1)\big)\varphi_i, \\ &n \neq 2m \ \Rightarrow\ L\varphi_i = \left(\tfrac{1}{2}n - m\right)\big(p_0(\lambda_i) - p_0(\lambda_1)\big)\varphi_i. \end{aligned}$$ □

Proof of Lemma 4.1. (1) *The case $m = 2n$.* For fixed $z \in \Lambda_1$ and for $t \in \mathbb{R}$ close to 0, we compute the third order expansion of $\boldsymbol{Q}[tz]$. By Lemma 1.1 it vanishes up to first order. Noting the identity $\boldsymbol{Q}[v]/Q_0 \equiv e^{-nv}(1 + nv) - 1$, valid for all $v \in \Lambda_1$, we find at once

$$\frac{\boldsymbol{Q}[tz]}{Q_0} = -2m^2t^2z^2 + \tfrac{8}{3}m^3t^3z^3 + O(t^4);$$

in particular (with the notation of Section 1), we have $c_3[z] = \frac{8}{3}m^3Q_0z^3$, and Lemma 4.1 holds trivially.

(2) *The case $m \neq 2n$.* In this case, calculations are drastically simplified by picking the nonlinear argument of P_0 in $P_u(1)$, namely $w := \exp\big((\frac{1}{2}n - m)u\big)$ (see (7)), as new parameter for the local image of the conformal curvature-increment operator. Since w is close to 1, we further set $w = 1 + v$, so the conformal factor becomes

$$e^{2u} = (1 + v)^{4/(n-2m)}$$

and the renormalized Q-curvature increment operator accordingly becomes

$$\boldsymbol{Q}[u] \equiv \tilde{Q}[v] := (1 + v)^{1-2^\star} P_0(1 + v) - \left(\tfrac{1}{2}n - m\right) Q_0 \tag{13}$$

where $2^\star$ stands for $2n/(n - 2m)$ in our context (admittedly a loose notation, customary for critical Sobolev exponents). Of course, Lemma 1.1 still holds for the operator $\tilde{Q}$ (with $\tilde{L} := d\tilde{Q}[0] \equiv (2^\star/n)L$) and proving Theorem 1.3 for $\tilde{Q}$ is equivalent to proving it for $\boldsymbol{Q}$. Altogether, we may thus focus on the proof of Lemma 4.1 for $\tilde{Q}$ instead of $\boldsymbol{Q}$. (The reader can instead prove Lemma 4.1 directly for $\boldsymbol{Q}$, but it takes a few pages.)

Picking z and t as above, plugging $v = tz$ in (13), and using the equality

$$P_0(z) = p_0(\lambda_1)z \equiv (2^\star - 1)\left(\tfrac{1}{2}n - m\right) Q_0z,$$

obtained from (12), we readily calculate the expansion

$$\frac{1}{\left(\frac{1}{2}n - m\right) Q_0}\, \tilde{Q}[tz] = -\tfrac{1}{2}(2^\star - 2)(2^\star - 1)\, t^2z^2 + \tfrac{1}{3}(2^\star - 2)(2^\star - 1)2^\star\, t^3z^3 + O(t^4),$$

thus finding for its third order coefficient

$$\frac{1}{\left(\frac{1}{2}n-m\right)Q_0}\,\tilde{c}_3[z]=\tfrac{1}{3}(2^\star-2)(2^\star-1)2^\star\,z^3.$$

So Lemma 4.1 obviously holds, and with it Theorem 1.3. □

Appendix: Eigenvalue calculations

As well known (see [Berger et al. 1971], for instance), for each $i\in\mathbb{N}$, the i-th eigenvalue of Δ_0 on $\mathbb{S}^n$ equals $\lambda_i=i(i+n-1)$. Recalling (6), we must calculate

$$p_0(\lambda_i)=\prod_{k=1}^{m}\left(\lambda_i+\left(\tfrac{1}{2}n-k\right)\left(\tfrac{1}{2}n+k-1\right)\right).$$

Setting provisionally

$$r=\tfrac{1}{2}(n-1),\quad s_k=k-\tfrac{1}{2},$$

so that $\frac{1}{2}n-k=r-s_k$, $\frac{1}{2}n+k-1=r+s_k$ and $\lambda_i=i^2+2ir$, we can rewrite

$$\begin{aligned}p_0(\lambda_i)&=\prod_{k=1}^{m}\left((i+r)^2-s_k^2\right)\\&=\prod_{k=1}^{m}\left(\tfrac{1}{2}+i+r-k\right)\left(\tfrac{1}{2}+i+r+k-1\right)\equiv\prod_{k=0}^{2m-1}\left(\tfrac{1}{2}+i+r-m+k\right),\end{aligned}$$

getting (back to m, n and k only)

$$p_0(\lambda_i)=\prod_{k=0}^{2m-1}\left(i+\tfrac{1}{2}n-m+k\right).$$

In particular,

$$P_0(1)\equiv p_0(\lambda_0)=\left(\tfrac{1}{2}n-m\right)\prod_{k=1}^{2m-1}\left(\tfrac{1}{2}n-m+k\right)$$

as asserted in (8) (and consistently there with the value of Q_0 in case $n=2m$). An easy induction argument yields

$$p_0(\lambda_{i+1})=\frac{\left(\frac{1}{2}n+m+i\right)}{\left(\frac{1}{2}n-m+i\right)}\,p_0(\lambda_i)\quad\text{for all } i\in\mathbb{N}$$

(consistently when $i=0$ with (12)), which implies that $|p_0(\lambda_{i+1})|>|p_0(\lambda_i)|$ for all $i\in\mathbb{N}$, hence in particular $p_0(\lambda_i)\neq p_0(\lambda_1)$ for $i>1$, as required in the proof of Lemma 1.1. This also implies the final formula

$$p_0(\lambda_i)=\frac{\left(\frac{1}{2}n+m\right)\ldots\left(\frac{1}{2}n+m+i-1\right)}{\left(\frac{1}{2}n-m\right)\ldots\left(\frac{1}{2}n-m+i-1\right)}\,p_0(\lambda_0)\quad\text{for all } i\geq 1.$$

Acknowledgement

It is a pleasure to thank Colin Guillarmou for providing us with the explicit formula of the conformally covariant powers of the laplacian on the sphere [Guillarmou and Naud 2006, Proposition 2.2].

References

[Aubin 1982] T. Aubin, *Nonlinear analysis on manifolds: Monge–Ampère equations*, Grundlehren der Math. Wissenschaften **252**, Springer, New York, 1982. MR 85j:58002 Zbl 0512.53044

[Beckner 1993] W. Beckner, "Sharp Sobolev inequalities on the sphere and the Moser–Trudinger inequality", *Ann. of Math.* (2) **138**:1 (1993), 213–242. MR 94m:58232 Zbl 0826.58042

[Berger et al. 1971] M. Berger, P. Gauduchon, and E. Mazet, *Le spectre d'une variété riemannienne*, Lecture Notes in Math. **194**, Springer, Berlin, 1971. MR 43 #8025 Zbl 0223.53034

[Bourguignon 1986] J.-P. Bourguignon, "Invariants intégraux fonctionnels pour des équations aux dérivées partielles d'origine géométrique", pp. 100–108 in *Differential geometry* (Peñíscola 1985), edited by A. M. Naveira et al., Lecture Notes in Math. **1209**, Springer, Berlin, 1986. MR 88b:58145 Zbl 0623.53003

[Bourguignon and Ezin 1987] J.-P. Bourguignon and J.-P. Ezin, "Scalar curvature functions in a conformal class of metrics and conformal transformations", *Trans. Amer. Math. Soc.* **301**:2 (1987), 723–736. MR 88e:53054 Zbl 0622.53023

[Branson 1987] T. P. Branson, "Group representations arising from Lorentz conformal geometry", *J. Funct. Anal.* **74**:2 (1987), 199–291. MR 90b:22016 Zbl 0643.58036

[Branson 1995] T. P. Branson, "Sharp inequalities, the functional determinant, and the complementary series", *Trans. Amer. Math. Soc.* **347**:10 (1995), 3671–3742. MR 96e:58162 Zbl 0848.58047

[Delanoë 2003] P. Delanoë, "Local solvability of elliptic, and curvature, equations on compact manifolds", *J. Reine Angew. Math.* **558** (2003), 23–45. MR 2004e:53054 Zbl 1040.53047

[Friedrich 1991] T. Friedrich, "Die Fisher-Information und symplektische Strukturen", *Math. Nachr.* **153** (1991), 273–296. MR 93e:58008 Zbl 0792.62003

[Graham and Zworski 2003] C. R. Graham and M. Zworski, "Scattering matrix in conformal geometry", *Invent. Math.* **152**:1 (2003), 89–118. MR 2004c:58064 Zbl 1030.58022

[Graham et al. 1992] C. R. Graham, R. Jenne, L. J. Mason, and G. A. J. Sparling, "Conformally invariant powers of the Laplacian. I. Existence", *J. London Math. Soc.* (2) **46**:3 (1992), 557–565. MR 94c:58226 Zbl 0726.53010

[Guillarmou and Naud 2006] C. Guillarmou and F. Naud, "Wave 0-trace and length spectrum on convex co-compact hyperbolic manifolds", *Comm. Anal. Geom.* **14**:5 (2006), 945–967. MR 2287151 Zbl 05141318

[Kazdan and Warner 1974] J. L. Kazdan and F. W. Warner, "Curvature functions for compact 2-manifolds", *Ann. of Math.* (2) **99** (1974), 14–47. MR 49 #7949 Zbl 0273.53034

[Kazdan and Warner 1975] J. L. Kazdan and F. W. Warner, "Scalar curvature and conformal deformation of Riemannian structure", *J. Differential Geometry* **10** (1975), 113–134. MR 51 #1661 Zbl 0296.53037

[Kobayashi and Nomizu 1963] S. Kobayashi and K. Nomizu, *Foundations of differential geometry*, vol. I, Wiley-Interscience, New York and London, 1963. MR 27 #2945 Zbl 0119.37502

[Lang 1962] S. Lang, *Introduction to differentiable manifolds*, Wiley, New York, 1962. MR 27 #5192 Zbl 0103.15101

[Lelong-Ferrand 1969] J. Lelong-Ferrand, "Transformations conformes et quasiconformes des variétés riemanniennes; application à la démonstration d'une conjecture de A. Lichnerowicz", *C. R. Acad. Sci. Paris Sér. A-B* **269** (1969), A583–A586. MR 40 #7989 Zbl 0201.09701

[Moser 1973] J. Moser, "On a nonlinear problem in differential geometry", pp. 273–280 in *Dynamical systems* (Salvador, 1971), edited by M. Peixoto, Academic Press, New York, 1973. MR 49 #4018 Zbl 0275.53027

[Obata 1971/72] M. Obata, "The conjectures on conformal transformations of Riemannian manifolds", *J. Differential Geometry* **6** (1971/72), 247–258. MR 46 #2601 Zbl 0236.53042

[de Rham 1960] G. de Rham, *Variétés différentiables: formes, courants, formes harmoniques*, 2nd ed., ActualitÃ©s scientifiques et industrielles **1222**, Hermann, Paris, 1960. Translated as *Differentiable manifolds: forms, currents, harmonic forms*, Grundlehren der math. Wissenschaften **266**, Springer, New York, 1984. Zbl 0089.08105

[Stredder 1975] P. Stredder, "Natural differential operators on Riemannian manifolds and representations of the orthogonal and special orthogonal groups", *J. Differential Geometry* **10**:4 (1975), 647–660. MR 54 #3772 Zbl 0318.53046

Received May 4, 2006.

PHILIPPE DELANOË
UNIVERSITÉ DE NICE–SOPHIA ANTIPOLIS
LABORATOIRE J. A. DIEUDONNÉ
PARC VALROSE
F-06108 NICE CEDEX
FRANCE

delphi@math.unice.fr
http://www-math.unice.fr/∼delphi

FRÉDÉRIC ROBERT
UNIVERSITÉ DE NICE–SOPHIA ANTIPOLIS
LABORATOIRE J. A. DIEUDONNÉ
PARC VALROSE
F-06108 NICE CEDEX
FRANCE

frobert@math.unice.fr
http://www-math.unice.fr/∼frobert

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.305

KNOT COLOURING POLYNOMIALS

MICHAEL EISERMANN

We introduce a natural extension of the colouring numbers of knots, called colouring polynomials, and study their relationship to Yang–Baxter invariants and quandle 2-cocycle invariants.

For a knot K in the 3-sphere, let π_K be the fundamental group of the knot complement $\mathbb{S}^3 \setminus K$, and let $m_K, l_K \in \pi_K$ be a meridian-longitude pair. Given a finite group G and an element $x \in G$ we consider the set of representations $\rho : \pi_K \to G$ with $\rho(m_K) = x$ and define the colouring polynomial as $\sum_\rho \rho(l_K)$. The resulting invariant maps knots to the group ring $\mathbb{Z}G$. It is multiplicative with respect to connected sum and equivariant with respect to symmetry operations of knots. Examples are given to show that colouring polynomials distinguish knots for which other invariants fail, in particular they can distinguish knots from their mutants, obverses, inverses, or reverses.

We prove that every quandle 2-cocycle state-sum invariant of knots is a specialization of some knot colouring polynomial. This provides a complete topological interpretation of these invariants in terms of the knot group and its peripheral system. Furthermore, we show that the colouring polynomial can be presented as a Yang–Baxter invariant, i.e. as the trace of some linear braid group representation. This entails that Yang–Baxter invariants *can* detect noninversible and nonreversible knots.

1. Introduction and statement of results

To each knot K in the 3-sphere $\mathbb{S}^3$ we can associate its knot group, that is, the fundamental group of the knot complement, denoted by

$$\pi_K := \pi_1(\mathbb{S}^3 \setminus K).$$

This group is already a very strong invariant: it classifies unoriented prime knots [Whitten 1987; Gordon and Luecke 1989]. In order to capture the complete information, we consider a meridian-longitude pair $m_K, l_K \in \pi_K$: the group system

MSC2000: primary 57M25; secondary 57M27.
Keywords: fundamental group of a knot, peripheral system, knot group homomorphism, quandle 2-cocycle state-sum invariant, Yang–Baxter invariant.

(π_K, m_K, l_K) classifies oriented knots in the 3-sphere [Waldhausen 1968]. In particular, the group system allows us to tackle the problem of detecting asymmetries of a given knot (see Section 2C). Using this ansatz, M. Dehn [1914] proved that the two trefoil knots are chiral, and, half a century later, H. F. Trotter [1963] proved that pretzel knots are nonreversible. We will recover these results using knot colouring polynomials (see Section 2D).

Given a knot K, represented say by a planar diagram, we can easily read off the Wirtinger presentation of π_K in terms of generators and relations (Section 3A). In general, however, such presentations are very difficult to analyze. As R. H. Crowell and R. H. Fox [1963, §VI.5] put it:

> "What is needed are some standard procedures for deriving from a group presentation some easily calculable algebraic quantities which are the same for isomorphic groups and hence are so-called group invariants."

The classical approach is, of course, to consider abelian invariants, most notably the Alexander polynomial. In order to effectively extract nonabelian information, we consider the set of knot group homomorphisms $\mathrm{Hom}(\pi_K; G)$ to some finite group G. The aim of this article is to organize this information and to generalize colouring numbers to colouring polynomials. In doing so, we will highlight the close relationship to Yang–Baxter invariants and their deformations on the one hand, and to quandle cohomology and associated state-sum invariants on the other hand.

From colouring numbers to colouring polynomials. A first and rather crude invariant is given by the total number of G-representations, denoted by

$$F_G(K) := |\mathrm{Hom}(\pi_K; G)|.$$

This defines a map $F_G : \mathcal{K} \to \mathbb{Z}$ on the set $\mathcal{K}$ of isotopy classes of knots in $\mathbb{S}^3$. This invariant can be refined by further specifying the image of the meridian m_K, that is, we choose an element $x \in G$ and consider only those homomorphisms $\rho : \pi_K \to G$ satisfying $\rho(m_K) = x$. Their total number defines the knot invariant

$$F_G^x(K) := |\mathrm{Hom}(\pi_K, m_K; G, x)|.$$

Example 1.1. Let G be the dihedral group of order $2p$, where $p \geq 3$ is odd, and let $x \in G$ be a reflection. Then F_G^x is the number of p-colourings as introduced by Fox [1962; 1970], here divided by p for normalization such that $F_G^x(\bigcirc) = 1$.

We will call F_G^x the *colouring number* associated with (G, x), in the dihedral case just as well as in the general case of an arbitrary group. Obviously F_G can be recovered from F_G^x by summation over all $x \in G$. In order to exploit the information of meridian *and* longitude, we introduce knot colouring polynomials as follows:

Definition 1.2. Suppose that G is a finite group and x is one of its elements. The *colouring polynomial* $P_G^x : \mathcal{K} \to \mathbb{Z}G$ is defined as

$$P_G^x(K) := \sum_\rho \rho(l_K),$$

where the sum is taken over all homomorphisms $\rho : \pi_K \to G$ with $\rho(m_K) = x$.

By definition P_G^x takes its values in the semiring $\mathbb{N}G$, but we prefer the more familiar group ring $\mathbb{Z}G \supset \mathbb{N}G$. We recover the colouring number $F_G^x = \varepsilon P_G^x$ by composing with the augmentation map $\varepsilon : \mathbb{Z}G \to \mathbb{Z}$. As it turns out, colouring polynomials allow us in a simple and direct manner to distinguish knots from their mirror images, as well as from their reverse or inverse knots. We will highlight some examples below.

Elementary properties. The invariant P_G^x behaves very much like classical knot polynomials. Most notably, it nicely reflects the natural operations on knots: P_G^x is multiplicative under connected sum and equivariant under symmetry operations (Section 2C).

Strictly speaking, $P_G^x(K)$ is, of course, not a polynomial but an element in the group ring $\mathbb{Z}G$. Since l_K lies in the commutator subgroup π_K' and commutes with m_K, possible longitude images lie in the subgroup $\Lambda = C(x) \cap G'$. Very often this subgroup will be cyclic, $\Lambda = \langle t \mid t^n = 1\rangle$ say, in which case P_G^x takes values in the truncated polynomial ring $\mathbb{Z}\Lambda = \mathbb{Z}[t]/(t^n)$. Here is a first and very simple example:

Example 1.3. We choose the alternating group $G = \mathrm{A}_5$ with basepoint $x = (12345)$. Here the longitude subgroup $\Lambda = \langle x\rangle$ is cyclic of order 5. The colouring polynomials of the left- and right-handed trefoil knots are $1 + 5x$ and $1 + 5x^{-1}$ respectively, hence the trefoil knots are chiral. (There are five nontrivial colourings, one of which is shown in Section 3, Figure 5, and the other four are obtained by conjugating with x. This list is easily seen to be complete.)

Starting from scratch, i.e. from knot diagrams and Reidemeister moves, one usually appreciates Fox' notion of 3-colourability [1970] as the simplest proof of knottedness. In this vein, the preceding example is arguably one of the most elementary proofs of chirality, only rivalled by the Kauffman bracket leading to the Jones polynomial [Kauffman 1987].

Section 2D displays some further examples to show that colouring polynomials distinguish knots for which other invariants fail:

- They distinguish the Kinoshita–Terasaka knot from the Conway knot and show that none of them is inversible nor reversible nor obversible.
- They detect asymmetries of pretzel knots; thus, for example, they distinguish $B(3, 5, 7)$ from its inverse, reverse and obverse knot.

- They distinguish the (inversible) knot 8_{17} from its reverse.

We also mention two natural questions that will not be pursued here:

Question 1.4. Can knot colouring polynomials detect other geometric properties of knots? Applications to periodic knots and ribbon knots would be most interesting.

Question 1.5. Do colouring polynomials distinguish all knots? Since the knot group system (π_K, m_K, l_K) characterizes the knot K [Waldhausen 1968, Cor. 6.5], and knot groups are residually finite [Thurston 1982, Thm. 3.3], this question is not completely hopeless.

Colouring polynomials are Yang–Baxter invariants. Moving from empirical evidence to a more theoretical level, this article compares knot colouring polynomials with two other classes of knot invariants: Yang–Baxter invariants, derived from traces of Yang–Baxter representations of the braid group (Section 4), and quandle colouring state-sum invariants derived from quandle cohomology (Section 3). The result can be summarized as follows:

$$\left\{\begin{matrix}\text{Yang–Baxter}\\ \text{invariants}\end{matrix}\right\} \supset \left\{\begin{matrix}\text{colouring}\\ \text{polynomials}\end{matrix}\right\} \supset \left\{\begin{matrix}\text{quandle 2-cocycle}\\ \text{state-sum invariants}\end{matrix}\right\} \supset \left\{\begin{matrix}\text{col. polynomials}\\ \text{with } \Lambda \text{ abelian}\end{matrix}\right\}$$

P. J. Freyd and D. N. Yetter [1989, Prop. 4.2.5] have shown that every colouring number $F_G^x : \mathcal{H} \to \mathbb{Z}$ can be obtained from a certain Yang–Baxter operator c over $\mathbb{Z}$. We generalize this result to colouring polynomials:

Theorem 1.6 (Section 4C). *Suppose that G is a group with basepoint x such that the subgroup $\Lambda = C(x) \cap G'$ is abelian. Then the colouring polynomial $P_G^x : \mathcal{H} \to \mathbb{Z}\Lambda$ is a Yang–Baxter invariant of closed knots*: *there exists a Yang–Baxter operator $\tilde{c}$ over the ring $\mathbb{Z}\Lambda$, such that the associated knot invariant coincides with* (*a constant multiple*) *of P_G^x.*

In the general case, where Λ is not necessarily abelian, Section 4B gives an analogous presentation of P_G^x as a Yang–Baxter invariant of long knots (also called 1-tangles).

Corollary 1.7. *Since Λ is abelian in all our examples of Section 2D, it follows in particular that Yang–Baxter invariants* can *detect noninversible and nonreversible knots.*

Remark 1.8. It follows from our construction that $\tilde{c}$ is a deformation of c over the ring $\mathbb{Z}\Lambda$. Conversely, the deformation ansatz leads to quandle cohomology (see Section 4D). Elaborating this approach, M. Graña [2002] showed that quandle 2-cocycle state-sum invariants are Yang–Baxter invariants. The general theory of Yang–Baxter deformations of c_Q over the power series ring $\mathbb{Q}[\![h]\!]$ has been developed in [Eisermann 2005].

Remark 1.9. The celebrated Jones polynomial and, more generally, all quantum invariants of knots, can be obtained from Yang–Baxter operators that are formal power series deformations of the trivial operator. This implies that the coefficients in this expansion are of finite type [Bar-Natan 1995, §2.1]. Part of their success lies in the fact that these invariants distinguish many knots, and in particular they easily distinguish mirror images. It is still unknown, however, whether finite type invariants can detect noninversible or nonreversible knots.

For colouring polynomials the construction is similar in that P_G^x arises from a deformation of a certain operator c. There are, however, two crucial differences:

- The initial operator c models conjugation (and is not the trivial operator),
- Its deformation $\tilde{c}$ is defined over $\mathbb{Z}\Lambda$ (and not over a power series ring).

As a consequence, the colouring polynomial P_G^x is not of finite type, nor are its coefficients, nor any other real-valued invariant computed from it [Eisermann 2000b].

Quandle invariants are specialized colouring polynomials. A quandle, as introduced by D. Joyce [1982], is a set Q with a binary operation whose axioms model conjugation in a group, or equivalently, the Reidemeister moves of knot diagrams. Quandles have been intensively studied by different authors and under various names; we review the relevant definitions in Section 3. The Lifting Lemma proved in Section 3B tells us how to pass from quandle to group colourings and back without any loss of information. On the level of knot invariants this implies the following result:

Theorem 1.10 (Section 3B). *Every quandle colouring number F_Q^q is the specialization of some knot colouring polynomial P_G^x.*

Quandle cohomology was initially studied in order to construct invariants in low-dimensional topology: in [Carter et al. 1999; 2003b] it was shown how a 2-cocycle $\lambda \in Z^2(Q, \Lambda)$ gives rise to a state-sum invariant of knots, $S_Q^\lambda : \mathcal{K} \to \mathbb{Z}\Lambda$, which refines the quandle colouring number F_Q. We prove the following result:

Theorem 1.11 (Section 3E). *Every quandle 2-cocycle state-sum invariant of knots is the specialization of some knot colouring polynomial. More precisely, suppose that Q is a connected quandle, Λ is an abelian group, and $\lambda \in Z^2(Q, \Lambda)$ is a 2-cocycle with associated invariant $S_Q^\lambda : \mathcal{K} \to \mathbb{Z}\Lambda$. Then there exists a group G with basepoint x and a $\mathbb{Z}$-linear map $\varphi : \mathbb{Z}G \to \mathbb{Z}\Lambda$ such that $S_Q^\lambda = \varphi P_G^x \cdot |Q|$.*

This result provides a complete topological interpretation of quandle 2-cocycle state-sum invariants in terms of the knot group and its peripheral system. Conversely, we prove that state-sum invariants contain those colouring polynomials P_G^x for which the longitude group $\Lambda = C(x) \cap G'$ is abelian:

Theorem 1.12 (Section 3D). *Suppose that G is a colouring group with basepoint x such that the subgroup $\Lambda = C(x) \cap G'$ is abelian. Then the colouring polynomial P_G^x can be presented as a quandle 2-cocycle state-sum invariant. More precisely, the quandle $Q = x^G$ admits a 2-cocycle $\lambda \in Z^2(Q, \Lambda)$ such that $S_Q^\lambda = P_G^x \cdot |Q|$.*

How this article is organized. Section 2 recalls the necessary facts about the knot group and its peripheral system. It then discusses connected sum and symmetry operations with respect to knot colouring polynomials and displays some applications. The main purpose is to give some evidence as to the scope and the usefulness of these invariants.

In Section 3 we examine quandle colourings and explain how to replace quandle colourings by group colourings without any loss of information. The correspondence between quandle extensions and quandle cohomology is then used to show how quandle 2-cocycle state-sum invariants can be seen as specializations of colouring polynomials.

Section 4 relates colouring polynomials with Yang–Baxter invariants. After recalling the framework of linear braid group representations, we show how colouring polynomials can be seen as Yang–Baxter deformations of colouring numbers.

2. Knot groups and colouring polynomials

This section collects basic facts about the knot group and its peripheral system (Section 2A) and their homomorphic images (Section 2B). We explain how the connected sum and symmetry operations affect the knot group system and how this translates to colouring polynomials (Section 2C). We then display some examples showing that colouring polynomials are a useful tool in distinguishing knots where other invariants fail (Section 2D).

2A. ***Peripheral system.*** We use fairly standard notation, which we recall from [Eisermann 2003] for convenience. A *knot* is a smooth embedding $k : \mathbb{S}^1 \hookrightarrow \mathbb{S}^3$, considered up to isotopy. This is equivalent to considering the oriented image $K = k(\mathbb{S}^1)$ in $\mathbb{S}^3$, again up to isotopy. A *framing* of k is an embedding $f : \mathbb{S}^1 \times \mathbb{D}^2 \hookrightarrow \mathbb{S}^3$ such that $f|_{\mathbb{S}^1 \times 0} = k$. As basepoint of the space $\mathbb{S}^3 \setminus K$ we choose $p = f(1, 1)$. In the fundamental group $\pi_K := \pi_1(\mathbb{S}^3 \setminus K, p)$ we define the *meridian* $m_K = [f|_{1 \times \mathbb{S}^1}]$ and the *longitude* $l_K = [f|_{\mathbb{S}^1 \times 1}]$. Up to isotopy the framing is characterized by the linking numbers $\mathrm{lk}(K, m_K) \in \{\pm 1\}$ and $\mathrm{lk}(K, l_K) \in \mathbb{Z}$, and all combinations are realized. We will exclusively work with the *standard framing*, characterized by the linking numbers $\mathrm{lk}(K, m_K) = +1$ and $\mathrm{lk}(K, l_K) = 0$.

Up to isomorphism, the triple (π_K, m_K, l_K) is a knot invariant, and even a complete invariant: two knots K and K' are isotopic if and only if there is a group isomorphism $\phi : \pi_K \to \pi_{K'}$ with $\phi(m_K) = m_{K'}$ and $\phi(l_K) = l_{K'}$. This is a special

case of Waldhausen's theorem on sufficiently large 3-manifolds; see [Waldhausen 1968, Cor. 6.5] and [Burde and Zieschang 1985, §3C].

Besides closed knots $k : \mathbb{S}^1 \hookrightarrow \mathbb{S}^3$ it will be useful to consider long knots (also called 1-tangles), i.e. smooth embeddings $\ell : \mathbb{R} \hookrightarrow \mathbb{R}^3$ such that $\ell(t) = (t, 0, 0)$ for all parameters t outside of some compact interval. See [Eisermann 2003] for a detailed discussion with respect to knot groups and quandles.

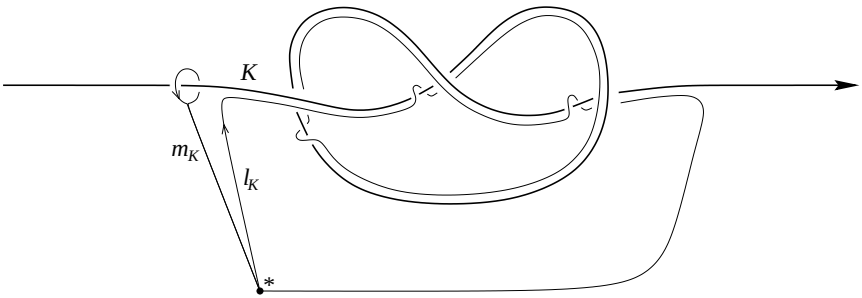


Figure 1. Meridian and longitude of a long knot.

2B. ***Colouring groups.*** Since knot groups are residually finite [Thurston 1982, Thm. 3.3], there are plenty of finite knot group representations. But which groups do actually occur as homomorphic images of knot groups? This question was raised by L. P. Neuwirth [1965], and first solved by F. González-Acuña:

Theorem 2.1 [González-Acuña 1975; Johnson 1980]. *A pointed group (G, x) is the homomorphic image of some knot group (π_K, m_K) if and only if G is finitely generated and $G = \langle x^G \rangle$.* □

The condition is necessary, because every knot group π_K is finitely generated by conjugates of the meridian m_K. (See the Wirtinger presentation, recalled in Section 3A.) For a proof of sufficiency we refer to the article of D. Johnson [1980], who has found an elegant and ingeniously simple way to construct a knot K together with an epimorphism $(\pi_K, m_K) \to (G, x)$. Here we restrict attention to *finite* groups:

Definition 2.2. Let G be a finite group and $x \in G$. The pair (G, x) is called a *colouring group* if the conjugacy class x^G generates the whole group G. For example, every finite simple group G is a colouring group with respect to any of its nontrivial elements $x \neq 1$.

Remark 2.3. Given a finite group G_0 and $x \in G_0$, every homomorphism

$$(\pi_K, m_K) \to (G_0, x)$$

maps to the subgroup $G_1 := \langle x^{G_0} \rangle$. If G_1 is strictly smaller than G_0, then we can replace G_0 by G_1. Continuing like this, we obtain a descending chain $G_0 \supset$

$G_1 \supset G_2 \supset \cdots$, recursively defined by $G_{i+1} = \langle x^{G_i} \rangle$. Since G_0 is finite, this chain must stabilize, and we end up with a colouring group $G_n = \langle x^{G_n} \rangle$. Hence, we can assume without loss of generality that (G, x) is a colouring group.

Given (G, x) let Λ^* be the set of longitude images $\rho(l_K)$, where ρ ranges over all knot group homomorphisms $\rho : (\pi_K, m_K) \to (G, x)$ and all knots K. Then Λ^* is a subgroup of G [Johnson and Livingston 1989]. Since meridian $m_K \in \pi_K$ and longitude $l_K \in \pi_K'$ commute, Λ^* is contained in the subgroup $\Lambda = C(x) \cap G'$, which will play an important rôle in subsequent arguments.

D. Johnson and C. Livingston [1989] have worked out a complete characterization of the subgroup Λ^* in terms of homological obstructions. As an application, consider a colouring group (G, x) that is perfect, i.e. $G' = G$, and has cyclic centralizer, say $C(x) = \langle x \rangle$. The main result of [Johnson and Livingston 1989] affirms that $\Lambda^* = \Lambda = C(x)$. All of our examples in Section 2D are of this type.

2C. ***Knot and group symmetries.*** The knot group π_K is obviously independent of orientations. In order to define the longitude, however, we have to specify the orientation of K, and the definition of the meridian additionally depends on the orientation of $\mathbb{S}^3$. Changing these orientations defines the following symmetry operations:

Definition 2.4. Let $K \subset \mathbb{S}^3$ be an oriented knot. The same knot with the opposite orientation of $\mathbb{S}^3$ is the *mirror image* or the *obverse* of K, denoted $K^\times$. (We can represent this as $K^\times = \sigma K$, where $\sigma : \mathbb{S}^3 \to \mathbb{S}^3$ is a reflection.) Reversing the orientation of the knot K yields the *reverse* knot $K^!$. Inverting both orientations yields the *inverse* knot K^*.

Please note that different authors use different terminology, in particular reversion and inversion are occasionally interchanged. Here we adopt the notation of J. H. Conway [1970].

Proposition 2.5. *Let K be an oriented knot with group system*

$$\check{\pi}(K) = (\pi_K, m_K, l_K).$$

Obversion, reversion and inversion affect the group system as follows:

$$\begin{aligned} \check{\pi}(K^\times) &= (\pi_K, m_K^{-1}, l_K), \\ \check{\pi}(K^!) &= (\pi_K, m_K^{-1}, l_K^{-1}), \\ \check{\pi}(K^*) &= (\pi_K, m_K, l_K^{-1}). \end{aligned}$$

The fundamental group of the connected sum $K \sharp L$ is the amalgamated product $\pi_K * \pi_L$ modulo $m_K = m_L$. Its meridian is m_K and its longitude is $l_K l_L$. □

Corollary 2.6. *Every colouring polynomial $P_G^x : \mathcal{K} \to \mathbb{Z}G$ is multiplicative, that is, we have $P_G^x(K \sharp L) = P_G^x(K) \cdot P_G^x(L)$ for any two knots K and L.* □

In order to formulate the effect of inversion, let $^* : \mathbb{Z}G \to \mathbb{Z}G$ be the linear extension of the inversion map $G \to G$, $g \mapsto g^{-1}$.

Corollary 2.7. *Every colouring polynomial $P_G^x : \mathcal{K} \to \mathbb{Z}G$ is equivariant under inversion, i.e. $P_G^x(K^*) = P_G^x(K)^*$ for every knot K. In particular, the colouring number $F_G^x(K)$ is invariant under inversion of K.* □

Obversion and reversion of knots can similarly be translated into symmetries of colouring polynomials, but to do so we need a specific automorphism of G:

Definition 2.8. An automorphism $^\times : G \to G$ with $x^\times = x^{-1}$ is called an *obversion* of (G, x). An antiautomorphism $^! : G \to G$ with $x^! = x$ is called a *reversion* of (G, x).

Obviously a group (G, x) possesses a reversion if and only if it possesses an obversion. They are in general not unique, because they can be composed with any automorphism $\alpha \in \mathrm{Aut}(G, x)$, for example conjugation by an element in $C(x)$.

Remark 2.9. The braid group B_n, recalled in Section 4A below, has a unique antiautomorphism $^! : \mathrm{B}_n \to \mathrm{B}_n$ fixing the standard generators $\sigma_1, \ldots, \sigma_{n-1}$. Analogously there exists a unique automorphism $^\times : \mathrm{B}_n \to \mathrm{B}_n$ mapping each standard generator σ_i to its inverse σ_i^{-1}. The exponent sum $\mathrm{B}_n \to \mathbb{Z}$ shows that this cannot be an inner automorphism.

These symmetry operations on braids correspond to the above symmetry operations on knots: if a knot K is represented as the closure of the braid β (see Section 4A), then the inverse braid β^{-1} represents the inverse knot K^*, the reverse braid $\beta^!$ represents the reverse knot $K^!$, and the obverse braid $\beta^\times$ represents the obverse knot $K^\times$.

Given an obversion and a reversion of (G, x), their linear extensions to the group ring $\mathbb{Z}G$ will also be denoted by $^\times : \mathbb{Z}G \to \mathbb{Z}G$ and $^! : \mathbb{Z}G \to \mathbb{Z}G$, respectively. We can now formulate the equivariance of the corresponding colouring polynomials:

Corollary 2.10. *Suppose that (G, x) possesses an obversion $^\times$ and a reversion $^!$. Then the colouring polynomial P_G^x is equivariant with respect to obversion and reversion, that is, we have $P_G^x(K^\times) = P_G^x(K)^\times$ and $P_G^x(K^!) = P_G^x(K)^!$ for every knot K. In this case the colouring numbers of K, K^*, $K^\times$, $K^!$ are the same.* □

Example 2.11. Every element x in the symmetric group S_n is conjugated to its inverse x^{-1}, because both have the same cycle structure. Any such conjugation defines an obversion $(\mathrm{S}_n, x) \to (\mathrm{S}_n, x^{-1})$. This argument also applies to alternating groups: given $x \in \mathrm{A}_n$ we know that x is conjugated to x^{-1} in S_n. Since A_n is normal in S_n, this conjugation restricts to an obversion $(\mathrm{A}_n, x) \to (\mathrm{A}_n, x^{-1})$. This need not be an inner automorphism.

On the other hand, some groups do not permit any obversion at all:

Example 2.12. Let $\mathbb{F}$ be a finite field and let $G = \mathbb{F} \rtimes \mathbb{F}^\times$ be its affine group. We have $\operatorname{Aut}(G) = \operatorname{Inn}(G) \rtimes \operatorname{Gal}(\mathbb{F})$, where $\operatorname{Gal}(\mathbb{F})$ is the Galois group of $\mathbb{F}$ over its prime field $\mathbb{F}_p$. If $\mathbb{F} = \mathbb{F}_p$, then every automorphism of G is inner and thus induces the identity on the abelian quotient $\mathbb{F}^\times$. If $p \geq 5$, we can choose an element $x = (a, b) \in G$ whose projection to $\mathbb{F}^\times$ satisfies $b \neq b^{-1}$. Hence there is no automorphism of G that maps x to x^{-1}. Indeed, searching all groups of small order with GAP [2006], we find that the smallest group having this property is $\mathbb{F}_5 \rtimes \mathbb{F}_5^\times$ of order 20.

For the sake of completeness we expound the following elementary result:

Proposition 2.13. *The affine group $G = \mathbb{F} \rtimes \mathbb{F}^\times$ satisfies $\operatorname{Aut}(G) = \operatorname{Inn}(G) \rtimes \operatorname{Gal}(\mathbb{F})$.*

Proof. The product in G is given by $(a, b)(c, d) = (a + bc, bd)$, and so $\operatorname{Gal}(\mathbb{F})$ can be seen as a subgroup of $\operatorname{Aut}(G)$, where $\phi \in \operatorname{Gal}(\mathbb{F})$ acts as $(a, b) \mapsto (\phi(a), \phi(b))$. Since $\operatorname{Inn}(G)$ is a normal subgroup of $\operatorname{Aut}(G)$ with $\operatorname{Inn}(G) \cap \operatorname{Gal}(\mathbb{F}) = \{\mathrm{id}_G\}$, we see that $\operatorname{Aut}(G)$ contains the semidirect product $\operatorname{Inn}(G) \rtimes \operatorname{Gal}(\mathbb{F})$.

It remains to show that every $\alpha \in \operatorname{Aut}(G)$ belongs to $\operatorname{Inn}(G) \rtimes \operatorname{Gal}(\mathbb{F})$. This is trivially true for $\mathbb{F} = \mathbb{F}_2$, so we will assume that $\mathbb{F}$ has more than two elements. It is then easily verified that $G' = \mathbb{F} \times \{1\}$. Let ζ be a generator of the multiplicative group $\mathbb{F}^\times$. We have $\alpha(1, 1) = (u, 1)$ with $u \in \mathbb{F}^\times$, and $\alpha(0, \zeta) = (v, \xi)$ with $v \in \mathbb{F}$, $\xi \in \mathbb{F}^\times$, $\xi \neq 1$. Conjugating by $w = (v(1 - \xi)^{-1}, u)$, we obtain $(u, 1)^w = (1, 1)$ and $(v, \xi)^w = (0, \xi)$. In the sequel we can thus assume $u = 1$ and $v = 0$. This implies $\alpha(0, b) = (0, \phi(b))$ with $\phi : \mathbb{F}^\times \to \mathbb{F}^\times$, $\zeta^n \mapsto \xi^n$ for all $n \in \mathbb{Z}$. Extending this by $\phi(0) = 0$ we obtain a bijection $\phi : \mathbb{F} \to \mathbb{F}$ satisfying $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in \mathbb{F}$. Moreover, we find $\alpha(a, 1) = (\phi(a), 1)$: this is clear for $a = 0$, and for $a \neq 0$ we have $(a, 1) = (0, a)(1, 1)$ and thus $\alpha(a, 1) = (0, \phi(a))(1, 1) = (\phi(a), 1)$. This proves that $\phi(a + b) = \phi(a) + \phi(b)$ for all $a, b \in \mathbb{F}$, whence $\phi \in \operatorname{Gal}(\mathbb{F})$. We conclude that $\alpha(a, b) = (\phi(a), \phi(b))$, as claimed. □

2D. ***Examples and applications.*** The preceding discussion indicates that symmetries of the group (G, x) affect the colouring polynomial $P_G^x(K)$ just as well as symmetries of the knot K. We point out several examples:

Example 2.14. Let p be a prime and let $G = \mathrm{PSL}_2 \mathbb{F}_p$ be equipped with basepoint $z = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ of order p. Inversion, obversion, and reversion are realized by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^* = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}, \qquad \begin{bmatrix} a & b \\ c & d \end{bmatrix}^\times = \begin{bmatrix} a & -b \\ -c & d \end{bmatrix}, \qquad \begin{bmatrix} a & b \\ c & d \end{bmatrix}^! = \begin{bmatrix} d & b \\ c & a \end{bmatrix}.$$

We have $C(z) = \langle z \rangle$. For $p = 2$ and $p = 3$ one finds that the longitude group $\Lambda = C(z) \cap G'$ is trivial. For $p \geq 5$ the group G is perfect (even simple), hence $\Lambda = \langle z \rangle$. We conclude that the colouring polynomial P_G^z is insensitive to reversion: we have $P_G^z(K) \in \mathbb{Z}\langle z \rangle$ and reversion fixes z and therefore all elements in $\mathbb{Z}\langle z \rangle$.

Example 2.15. Consider an alternating group $G = \mathrm{A}_n$ with $n \geq 3$, and a cycle $x = (123\ldots l)$ of maximal length, that is, $l = n$ for n odd and $l = n-1$ for n even. As we have pointed out above, a suitable conjugation in S_n produces an obversion $(G, x) \to (G, x^{-1})$. We have $C(x) = \langle x \rangle$. For $n = 3$ and $n = 4$ one finds that the longitude group $\Lambda = C(x) \cap G'$ is trivial. For $n \geq 5$ the group G is perfect (even simple), hence the longitude group is $\Lambda = \langle x \rangle$. Again we conclude that the colouring polynomial P_G^x is insensitive to reversion.

We observe that for $l = 3, 7, 11, \ldots$ an obversion of (G, x) cannot be realized by an inner automorphism: consider for example $G = \mathrm{A}_{11}$ and $x = (\mathtt{abcdefghijk})$: in S_{11} the centralizer is $C(x) = \langle x \rangle$ and consequently every permutation $\sigma \in \mathrm{S}_{11}$ with $x^\sigma = x^{-1}$ is of the form $\sigma = x^k(\mathtt{ak})(\mathtt{bj})(\mathtt{ci})(\mathtt{dh})(\mathtt{eg})$ and thus odd. The same argument shows that for $n = 5, 9, 13, \ldots$ an obversion of (G, x) can be realized by an inner automorphism.

Example 2.16. As a more exotic example, let us finally consider the Mathieu group M_{11}, i.e. the unique simple group of order $7920 = 2^4 \cdot 3^2 \cdot 5 \cdot 11$, and the smallest of the sporadic simple groups [Conway et al. 1985]. It can be presented as a subgroup of A_{11}, for example as

$$G = \langle x, y \rangle \quad \text{with} \quad x = (\mathtt{abcdefghijk}),\ y = (\mathtt{abcejikdghf}).$$

This presentation has been obtained from GAP and can easily be verified with any group-theory software by checking that G is simple of order 7920. The Mathieu group M_{11} is particularly interesting for us, because it does *not* allow an obversion. To see this it suffices to know that its group of outer automorphisms is trivial, in other words, every automorphism of M_{11} is realized by conjugation. In M_{11} the element x is not conjugated to its inverse — this is not even possible in A_{11} according to the preceding example. Hence there is no automorphism of M_{11} that maps x to x^{-1}.

Applied to colouring polynomials, this means that there is a priori no restriction on the invariants of a knot and its mirror image. As a concrete example we consider the Kinoshita–Terasaka knot K and the Conway knot C displayed in Figure 2.

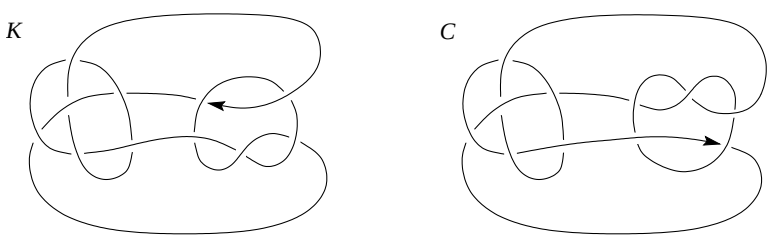


Figure 2. The Kinoshita–Terasaka knot and the Conway knot.

Both knots have trivial Alexander polynomial. They differ only by rotation of a 2-tangle, in other words they are mutants in the sense of Conway [1970]. Therefore neither the Jones, HOMFLYPT nor Kauffman polynomial can distinguish between K and C, see [Lickorish 1997]. With the help of a suitable colouring polynomial the distinction is straightforward:

Example 2.17. R. Riley [1971] has studied knot group homomorphisms to the simple group $G = \mathrm{PSL}_2 \mathbb{F}_7$ of order 168. Let z be an element of order 7, say $z = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$. Then the associated colouring polynomials are

$$\begin{aligned} P_G^z(K) &= P_G^z(C) = 1 + 7z^5 + 7z^6, \\ P_G^z(K^*) &= P_G^z(C^*) = 1 + 7z + 7z^2. \end{aligned}$$

This shows that both knots are chiral. By a more detailed analysis of their coverings, Riley could even show that K and C are distinct.

Example 2.18. To distinguish K and C we give a simple and direct argument using colouring polynomials. For every element $x \in \mathrm{PSL}_2 \mathbb{F}_7$ of order 3, say $x = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 1 \end{smallmatrix}\right]$, the associated colouring polynomial distinguishes K and C:

$$\begin{aligned} P_G^x(K) &= 1 + 6x, & P_G^x(C) &= 1 + 12x, \\ P_G^x(K^*) &= 1 + 6x^2, & P_G^x(C^*) &= 1 + 12x^2. \end{aligned}$$

Both invariants, P_G^z and P_G^x, show chirality but are insensitive to reversion.

These and the following colouring polynomials were calculated with the help of an early prototype of the computer program *KnotGRep*, an ongoing programming project to efficiently construct the set of knot group homomorphisms to a finite group. Even though general-purpose software may be less comfortable, our results can also be obtained from the Wirtinger presentation (Section 3A) using GAP or similar group-theoretic software.

Example 2.19. The alternating group $G = \mathrm{A}_7$ with basepoint $x = (1234567)$ yields

$$\begin{aligned} P_G^x(K) &= 1 + 7x^2 + 28x^5 + 28x^6, & P_G^x(C) &= 1 + 7x^2 + 7x^3 + 21x^5 + 14x^6, \\ P_G^x(K^*) &= 1 + 28x + 28x^2 + 7x^5, & P_G^x(C^*) &= 1 + 14x + 21x^2 + 7x^4 + 7x^5. \end{aligned}$$

Again this invariant distinguishes K et C and shows their chirality, but is insensitive to reversion, as explained in Example 2.15 above.

Example 2.20. More precise information can be obtained using the Mathieu group M_{11}, presented as the permutation group (G, x) in Example 2.16 above. For the

Kinoshita–Terasaka knot K and the Conway knot C one finds

$$\begin{aligned} P_G^x(K) &= 1+11x^3+11x^7, & P_G^x(C) &= 1+11x^3+11x^7,\\ P_G^x(K^*) &= 1+11x^4+11x^8, & P_G^x(C^*) &= 1+11x^4+11x^8,\\ P_G^x(K^\times) &= 1+11x^4+22x^8, & P_G^x(C^\times) &= 1+11x^4+11x^6+11x^8,\\ P_G^x(K^!) &= 1+22x^3+11x^7, & P_G^x(C^!) &= 1+11x^3+11x^5+11x^7. \end{aligned}$$

Consequently all eight knots are distinct; K and C are neither inversible nor obversible nor reversible. (This example was inspired by G. Kuperberg [1996], who used the colouring number F_G^x to distinguish the knot C from its reverse $C^!$.)

Usually it is very difficult to detect nonreversibility of knots. Most invariants fail to do so, including the usual knot polynomials. In view of the simplicity of our approach, the success of knot colouring polynomials is remarkable. We give two further examples:

Example 2.21. The family of pretzel knots $B(p_1, p_2, p_3)$, parametrized by odd integers p_1, p_2, p_3, is depicted in Figure 3, left. According to the classification of pretzel knots (see [Burde and Zieschang 1985], §12), the pretzel knot $B = B(3, 5, 7)$ is neither reversible nor obversible nor inversible. For the Mathieu group $G = M_{11}$ with basepoint x as in Example 2.20 we obtain

$$\begin{aligned} P_G^x(B) &= 1+11x, & P_G^x(B^\times) &= 1+11x^7,\\ P_G^x(B^*) &= 1+11x^{10}, & P_G^x(B^!) &= 1+11x^4. \end{aligned}$$

Again the colouring polynomial shows that the knot B possesses none of the three symmetries. Historically, pretzel knots were the first examples of nonreversible knots. Their nonreversibility was first proven by H. F. Trotter [1963] by representing the knot group system on a suitable triangle group acting on the hyperbolic plane.

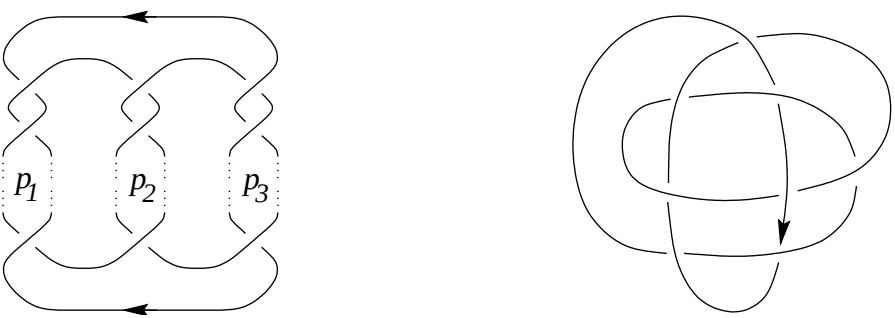


Figure 3. Left: the pretzel knot $B(p_1, p_2, p_3)$. Right: the knot 8_{17}.

Example 2.22. Figure 3, right, shows the knot 8_{17}, which is the smallest nonreversible knot. It is a 3-bridge knot but not a pretzel knot, and there is no general classification theorem available. To analyze this example we choose once more the Mathieu group M_{11} with basepoint x as above. The knot 8_{17} then has colouring polynomial $1 + 11x^5 + 11x^6$ whereas the reverse knot has trivial colouring polynomial 1. (Here even the colouring number F_G^x suffices to prove that this knot is nonreversible.) We remark that 8_{17} is inversible and that this symmetry is reflected in the symmetry of its colouring polynomials.

The colouring polynomial $P_G^x(K)$ is, by definition, an element in the group ring $\mathbb{Z}G$, and it actually lies in the much smaller ring $\mathbb{Z}\Lambda$. The following symmetry consideration further narrows down the possible values. It is included here to explain one of the observations that come to light in the previous examples, but it will not be used in the sequel.

Proposition 2.23. *Let (G, x) be a colouring group. If conjugation by x has order p^k for some prime p, then the colouring polynomial satisfies $P_G^x(K) \equiv 1 \pmod{p}$.*

Proof. The cyclic subgroup $\langle x \rangle$ acts on the set $\mathrm{Hom}(\pi_K, m_K ; G, x)$ by conjugation. The only fixed point is the trivial representation $(\pi(K), m_K) \to (\mathbb{Z}, 1) \to (G, x)$. This can be most easily seen by interpreting group homomorphisms $\rho : (\pi_K, m_K) \to (G, x)$ as colourings $f : (D, 0) \to (G, x)$ of a knot diagram D, see Section 3A below. If $f^x = f$ then all colours of f commute with x: following the diagram from the first to the last arc we see by induction that all colours are in fact equal to x. Since there is only one component, we conclude that f is the trivial colouring, corresponding to the trivial representation.

Every nontrivial representation ρ appears in an orbit of length p^ℓ, for some integer $\ell \geq 1$. Since $\rho(l_K)$ commutes with x, all representations in such an orbit have the same longitude image in G. The sum $P_G^x(K)$ thus begins with 1 for the trivial representation, and all other summands can be grouped to multiples of p. □

3. Quandle invariants are specialized colouring polynomials

The Wirtinger presentation allows us to interpret knot group homomorphisms as colourings of knot diagrams. Since such colourings involve only conjugation, they are most naturally treated in the category of quandles, as introduced by Joyce [1982]. We recall the basic definitions concerning quandles and quandle colourings in Section 3A, and explain in Section 3B how to pass from quandles to groups and back without any loss of information.

Quandle cohomology was studied in [Carter et al. 1999; 2003b], where it was shown how a 2-cocycle gives rise to a state-sum invariant of knots in $\mathbb{S}^3$. We recall this construction in Section 3D and show that every colouring polynomial P_G^x can

be presented as a quandle 2-cocycle state-sum invariant, provided that the subgroup $\Lambda = C(x) \cap G'$ is abelian (Theorem 3.24).

In order to prove the converse, we employ the cohomological classification of central quandle extensions established in [Eisermann 2003; Carter et al. 2003a], recalled in Section 3C below. This allows us to prove in Section 3E that every quandle 2-cocycle state-sum invariant is the specialization of a suitable knot colouring polynomial (Theorem 3.25).

3A. ***Wirtinger presentation, quandles, and colourings.*** Our exposition follows [Eisermann 2003], to which we refer for further details. We consider a long knot diagram as in Figure 1 and number the arcs consecutively from 0 to n. At the end of arc number $i-1$, we undercross arc number $\kappa i = \kappa(i)$ and continue on arc number i. We denote by $\varepsilon i = \varepsilon(i)$ the sign of this crossing, as depicted in Figure 4. The maps $\kappa : \{1, \dots, n\} \to \{0, \dots, n\}$ and $\varepsilon : \{1, \dots, n\} \to \{\pm 1\}$ are the *Wirtinger code* of the diagram.

Theorem 3.1. *Suppose that a knot L is represented by a long knot diagram with Wirtinger code (κ, ε) as above. Then the knot group allows the presentation*

$$\pi_L = \langle x_0, x_1, \dots, x_n \mid r_1, \dots, r_n \rangle \quad \textit{with relation } r_i \textit{ being} \quad x_i = x_{\kappa i}^{-\varepsilon i}\, x_{i-1}\, x_{\kappa i}^{\varepsilon i}.$$

As peripheral system we can choose $m_L = x_0$ and $l_L = \prod_{i=1}^{i=n} x_{i-1}^{-\varepsilon i}\, x_{\kappa i}^{\varepsilon i}$. □

For a proof see [Crowell and Fox 1963, §VI.3] or [Burde and Zieschang 1985, §3B]. The Wirtinger presentation works just as well for a closed knot diagram. Since arcs 0 and n are then identified, this amounts to adding the (redundant) relation $x_0 = x_n$ to the above presentation. The group is, of course, the same.

The Wirtinger presentation allows us to interpret knot group homomorphisms $\pi_L \to G$ as colourings. More precisely, a *G-colouring* of the diagram D is a map $f : \{0, \dots, n\} \to G$ such that $f(i) = f(\kappa i)^{-\varepsilon i} f(i-1) f(\kappa i)^{\varepsilon i}$. In other words, at each coloured crossing as in Figure 4 the colours a and c are conjugated via $a^b = c$. Such a colouring is denoted by $f : D \to G$. We denote by $\mathrm{Col}(D; G)$ the set of colourings of D with colours in G. For a long knot diagram D, we denote by $\mathrm{Col}(D, 0; G, x)$ the subset of colourings that colour arc number 0 with colour x. The Wirtinger presentation establishes natural bijections $\mathrm{Hom}(\pi_K; G) \cong \mathrm{Col}(D; G)$ and $\mathrm{Hom}(\pi_K, m_K; G, x) \cong \mathrm{Col}(D, 0; G, x)$.

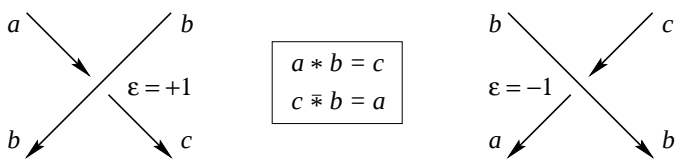


Figure 4. Wirtinger rules for colouring a knot diagram.

Example 3.2. Figure 5 shows a colouring of the left-handed trefoil knot (represented as a long knot) with elements in the alternating group A_5. Note that all definitions readily extend to closed knot diagrams.

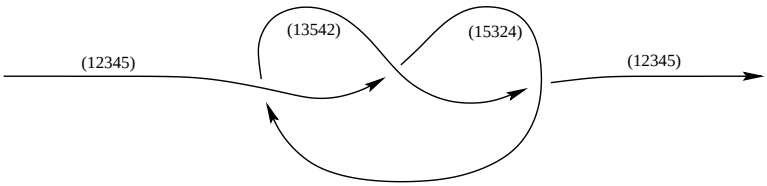


Figure 5. A_5-colouring of the left-handed trefoil knot.

The Wirtinger presentation of π_K involves only conjugation but not the group multiplication itself. The underlying algebraic structure can be described as follows:

Definition 3.3. A *quandle* is a set Q with two binary operations $*, \overline{*} : Q \times Q \to Q$ satisfying the following axioms for all $a, b, c \in Q$:

(Q1) $a * a = a$ (idempotency),

(Q2) $(a * b) \overline{*}\, b = (a \overline{*}\, b) * b = a$ (right invertibility),

(Q3) $(a * b) * c = (a * c) * (b * c)$ (self-distributivity).

As already mentioned, the notion (and name) was introduced in [Joyce 1982]. The same notion was studied by S. V. Matveev [1982] under the name "distributive groupoid", and by Kauffman [2001] who called it "crystal". Quandle axioms (Q2) and (Q3) are equivalent to saying that for every $b \in Q$ the right translation $\varrho_b : a \mapsto a * b$ is an automorphism of Q. Such structures were called "automorphic sets" by E. Brieskorn [1988]. The somewhat shorter term *rack* was preferred by R. Fenn and C. P. Rourke [1992]. The notion has been generalized to "crossed G-sets" by Freyd and Yetter [1989].

Definition 3.4. As before, let D be a long knot diagram, its arcs being numbered by $0, \ldots, n$. A *Q-colouring*, denoted $f : D \to Q$, is a map $f : \{0, \ldots, n\} \to Q$ such that at each crossing as in Figure 4 the three colours a, b, c satisfy the relation $a * b = c$. We denote by $\mathrm{Col}(D; Q)$ the set of Q-colourings, and by $\mathrm{Col}(D, 0; Q, q)$ the subset of colourings satisfying $f(0) = q$.

Proposition 3.5 [Joyce 1982]. *The quandle axioms ensure that each Reidemeister move $D \leftrightharpoons D'$ induces bijections $\mathrm{Col}(D; Q) \leftrightharpoons \mathrm{Col}(D'; Q)$ and $\mathrm{Col}(D, 0; Q, q) \leftrightharpoons \mathrm{Col}(D', 0; Q, q)$. In particular, if Q is finite, then the colouring numbers $F_Q(D) = |\mathrm{Col}(D; Q)|$ and $F_Q^q(D) = |\mathrm{Col}(D, 0; Q, q)|$ are knot invariants.* □

3B. ***From quandle colourings to group colourings and back.*** In many respects quandles are close to groups. For colourings we will now explain how to pass from quandles to groups and back without any loss of information.

Definition 3.6. A *quandle homomorphism* is a map $\phi : Q \to Q'$ that satisfies $\phi(a * b) = \phi(a) * \phi(b)$, and hence $\phi(a \,\overline{*}\, b) = \phi(a) \,\overline{*}\, \phi(b)$, for all $a, b \in Q$.

Definition 3.7. The automorphism group $\mathrm{Aut}(Q)$ consists of all bijective homomorphisms $\phi : Q \to Q$. We adopt the convention that automorphisms of Q act on the right, written a^ϕ, which means that their composition $\phi\psi$ is defined by $a^{(\phi\psi)} = (a^\phi)^\psi$ for all $a \in Q$.

Definition 3.8. The group $\mathrm{Inn}(Q) = \langle \varrho_b \mid b \in Q \rangle$ of *inner automorphisms* is the subgroup of $\mathrm{Aut}(Q)$ generated by all right translations $\varrho_b : a \mapsto a * b$. The quandle Q is called *connected* if the action of $\mathrm{Inn}(Q)$ on Q is transitive.

In view of the map $\varrho : Q \to \mathrm{Inn}(Q)$, $b \mapsto \varrho_b$, we also write $a^b = a * b$ for the operation in a quandle. Conversely, it will sometimes be convenient to write $a * b = b^{-1}ab$ for the conjugation in a group. In neither case will there be any danger of confusion.

Definition 3.9. A *representation* of a quandle Q on a group G is a map $\phi : Q \to G$ such that $\phi(a * b) = \phi(a) * \phi(b)$ for all $a, b \in Q$. In other words, the following diagram commutes:

$$\begin{array}{ccc} Q \times Q & \xrightarrow{\phi \times \phi} & G \times G \\ {\scriptstyle *}\downarrow & & \downarrow{\scriptstyle \mathrm{conj}} \\ Q & \xrightarrow{\phi} & G \end{array}$$

For example, the natural map $\varrho : Q \to \mathrm{Aut}(Q)$ satisfies $\varrho(a * b) = \varrho(a) * \varrho(b)$. We call ϱ the *inner* representation of Q. Moreover it satisfies $\varrho(a^g) = \varrho(a)^g$ for all $a \in Q$ and $g \in \mathrm{Aut}(Q)$. This is the prototype of an augmentation:

Definition 3.10. Let $\phi : Q \to G$ be a representation and let $\alpha : Q \times G \to Q$, $(a, g) \mapsto a^g$, be a group action. We call the pair (ϕ, α) an *augmentation* if $a * b = a^{\phi(b)}$ and $\phi(a^g) = \phi(a)^g$ for all $a, b \in Q$ and $g \in G$. In other words, the following diagram commutes:

$$\begin{array}{ccccc} Q \times Q & \xrightarrow{\mathrm{id} \times \phi} & Q \times G & \xrightarrow{\phi \times \mathrm{id}} & G \times G \\ {\scriptstyle *}\downarrow & & \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \mathrm{conj}} \\ Q & \xrightarrow{\mathrm{id}} & Q & \xrightarrow{\phi} & G \end{array}$$

Remark 3.11. We will usually reinterpret the group action α as a group homomorphism $\bar{\alpha} : G \to \mathrm{Aut}(Q)$, and denote the augmentation by $Q \xrightarrow{\phi} G \xrightarrow{\bar{\alpha}} \mathrm{Aut}(Q)$. If G is generated by the image $\phi(Q)$, then ϕ is equivariant and the action of G on Q is uniquely determined by the representation ϕ. In this case we simply say that $\phi : Q \to G$ is an augmentation. For example, every quandle Q comes equipped with the inner augmentation $\varrho : Q \to \mathrm{Inn}(Q)$.

Suppose that Q is a quandle and $\phi : Q \to G$ is a representation on some group G. Obviously every quandle colouring $\tilde{f} : D \to Q$ maps to a group colouring $f = \phi\tilde{f} : D \to G$. If ϕ is an augmentation, then this process can be reversed, and we can replace quandle colourings by group colourings without any loss of information:

Lemma 3.12. *Let $(Q, q) \xrightarrow{\phi} (G, g) \xrightarrow{\bar{\alpha}} \mathrm{Aut}(Q)$ be an augmentation of the quandle Q with basepoint $q \in Q$ on the group G with basepoint $x = \phi(q) \in G$. If D is a long knot diagram, then every group colouring $f : (D, 0) \to (G, x)$ can be lifted to a unique quandle colouring $\tilde{f} : (D, 0) \to (Q, q)$ such that $f = \phi\tilde{f}$. In other words, ϕ induces a bijection*

$$\phi_* : \mathrm{Col}(D, 0; Q, q) \xrightarrow{\sim} \mathrm{Col}(D, 0; G, x), \quad \tilde{f} \mapsto f = \phi\tilde{f}.$$

The lifted colouring $\tilde{f}$ begins with $\tilde{f}(0) = q$ and ends with $\tilde{f}(n) = q^{\rho(l_K)}$, where $\rho : (\pi_K, m_K) \to (G, x)$ is the knot group representation associated with f.

Proof. Every representation $\phi : (Q, q) \to (G, x)$ induces a map ϕ_* sending each quandle colouring $\tilde{f} : (D, 0) \to (Q, q)$ to the associated group colouring $\phi\tilde{f} : (D, 0) \to (G, x)$. In general ϕ_* is neither injective nor surjective, lest ϕ is an augmentation. To define the inverse map $\psi_* : \mathrm{Col}(D, 0; G, x) \to \mathrm{Col}(D, 0; Q, q)$, we use the action $\alpha : Q \times G \to Q$, which we temporarily denote by $(a, g) \mapsto a \bullet g$ for better readability.

The crucial ingredient in the proof is the commutativity of the bottom diagram on page 321. We first show how the condition $a * b = a \bullet \phi(b)$ ensures injectivity of ϕ_*. Let D be a long knot diagram with Wirtinger code (κ, ε). Assume that $\tilde{f}, \hat{f} : (D, 0) \to (Q, q)$ are colourings with $\phi\tilde{f} = \phi\hat{f}$. By hypothesis we have $\tilde{f}(0) = \hat{f}(0) = q$. By induction suppose that $\tilde{f}(i-1) = \hat{f}(i-1)$ for some $i \geq 1$. In the case of a positive crossing ($\varepsilon i = +1$) we then obtain

$$\begin{aligned}\tilde{f}(i) &= \tilde{f}(i-1) * \tilde{f}(\kappa i) = \tilde{f}(i-1) \bullet \phi\tilde{f}(\kappa i) \\ &= \hat{f}(i-1) \bullet \phi\hat{f}(\kappa i) = \hat{f}(i-1) * \hat{f}(\kappa i) = \hat{f}(i).\end{aligned}$$

The case of a negative crossing ($\varepsilon i = -1$) is analogous. We conclude that $\tilde{f} = \hat{f}$.

We now show how the equivariance condition $\phi(a \bullet g) = \phi(a) * g$ of the bottom diagram on page 321 ensures surjectivity. For every colouring $f : (D, 0) \to (G, x)$,

denoted by $i \mapsto x_i$, the colours $x_0, \ldots, x_n$ satisfy $x_i = x_{i-1} * x_{\kappa i}^{\varepsilon i}$. We define partial longitudes $\ell_0, \ldots, \ell_n$ by setting $\ell_i := \prod_{j=1}^{i} x_{j-1}^{-\varepsilon j} x_{\kappa j}^{\varepsilon j}$. In particular we have $x_0 = x_n = x$ and $x_i = x_0 * \ell_i$ for all $i = 0, \ldots, n$. By definition, $\ell_n = \rho(l_K)$ is the (total) longitude of the colouring f. We define $\tilde{f} : (D, 0) \to (Q, q)$ by assigning the colour $q_i = q \bullet \ell_i$ to arc number $i = 0, \ldots, n$. By hypothesis, $\phi : Q \to G$ is an equivariant map, whence

$$\phi(q_i) = \phi(q \bullet \ell_i) = \phi(q) * \ell_i = x * \ell_i = x_i. \tag{1}$$

At each positive crossing we find the following identity, using axiom (Q1):

$$q_{i-1} * q_{\kappa i} = (q_{i-1} \mathbin{\overline{*}} q_{i-1}) * q_{\kappa i} = (((q \bullet \ell_{i-1}) \bullet x_{i-1}^{-1}) \bullet x_{\kappa i} = q \bullet \ell_i = q_i. \tag{2}$$

Analogously at each negative crossing:

$$q_{i-1} \mathbin{\overline{*}} q_{\kappa i} = (q_{i-1} * q_{i-1}) \mathbin{\overline{*}} q_{\kappa i} = (((q \bullet \ell_{i-1}) \bullet x_{i-1}) \bullet x_{\kappa i}^{-1}) = q \bullet \ell_i = q_i. \tag{3}$$

We can thus define $\psi_* : \mathrm{Col}(D, 0; G, x) \to \mathrm{Col}(D, 0; Q, q)$ by $f \mapsto \tilde{f}$. Equation (1) shows that $\phi_* \psi_* = \mathrm{id}$, while (2) and (3) imply that $\psi_* \phi_* = \mathrm{id}$. □

Remark 3.13. Obviously, the condition $a * b = a^{\phi(b)}$ cannot be dropped because it connects the quandle operation $*$ with the group action α. Likewise, the equivariance condition $\phi(a^g) = \phi(a)^g$ cannot be dropped: as an extreme counterexample, consider a trivial quandle $Q = \{q\}$ and an arbitrary group (G, x). We have a unique representation $\phi : (Q, q) \to (G, x)$ and a unique group action $\alpha : Q \times G \to Q$. The map ϕ is equivariant if and only if $x \in Z(G)$. In general ϕ_* cannot be a bijection, because the only (Q, q)-colouring is the trivial one, while there may be nontrivial (G, x)-colourings.

The Lifting Lemma has the following analogue for closed knots:

Lemma 3.14. *Let $\phi : (Q, q) \to (G, x)$ be an augmentation of the quandle Q on the group G. If D is a closed knot diagram, then ϕ induces a bijection between* $\mathrm{Col}(D, 0; Q, q)$ *and those homomorphisms $\rho : (\pi_K, m_K) \to (G, x)$ satisfying $q^{\rho(l_K)} = q$.* □

As an immediate consequence we obtain:

Theorem 3.15. *Every quandle colouring number F_Q^q is the specialization of some knot colouring polynomial P_G^x.*

Proof. We consider an augmentation $\phi : (Q, q) \to (G, x)$ with $G = \langle \phi(Q) \rangle$, for example the inner augmentation on $\phi : Q \to G = \mathrm{Inn}(Q)$ with basepoint $x = \phi(q)$.

For long knots, Lemma 3.12 implies $F_Q^q = F_G^x$. Hence $F_Q^q = \varepsilon P_G^x$, where $\varepsilon : \mathbb{Z}G \to \mathbb{Z}$ is the augmentation map of the group ring, with $\varepsilon(g) = 1$ for all $g \in G$.

For closed knots we define the linear map $\varepsilon : \mathbb{Z}G \to \mathbb{Z}$ by setting $\varepsilon(g) = 1$ if $q^g = q$, and $\varepsilon(g) = 0$ if $q^g \neq q$. Then Lemma 3.14 implies that $F_Q^q = \varepsilon P_G^x$. □

This argument will be generalized in Section 3E, where we show that every quandle 2-cocycle state-sum invariant is the specialization of some colouring polynomial.

3C. ***Quandle coverings, extensions, and cohomology.*** We recall how quandle colourings can be used to encode longitudinal information (see [Eisermann 2003] for details). We consider a long knot diagram with meridians $x_0, \dots, x_n$ and partial longitudes $l_0, \dots, l_n$ as defined in the above proof of the Lifting Lemma. In particular we have $x_0 = x_n = m_K$ and $x_i = x_0 * l_i$ with $l_0 = 1$ and $l_n = l_K$. If we colour each arc not only with its meridian x_i but with the pair (x_i, l_i), then at each crossing we find that

$$x_i = x_{i-1} * x_{\kappa i}^{\varepsilon i} \quad \text{and} \quad l_i = l_{i-1} x_{i-1}^{-\varepsilon i} x_{\kappa i}^{\varepsilon i}.$$

This crossing relation can be encoded in a quandle as follows.

Lemma 3.16 [Eisermann 2003]. *Let G be a group that is generated by a conjugacy class $Q = x^G$. Then Q is a connected quandle with respect to conjugation $a * b = b^{-1}ab$ and its inverse $a \mathbin{\overline{*}} b = bab^{-1}$. Let G' be the commutator subgroup and define*

$$\tilde{Q} = \tilde{Q}(G, x) := \{\, (a, g) \in G \times G' \mid a = x^g \,\}.$$

The set $\tilde{Q}$ becomes a connected quandle when equipped with the operations

$$(a, g) * (b, h) = (a * b, ga^{-1}b) \quad \text{and} \quad (a, g) \mathbin{\overline{*}} (b, h) = (a \mathbin{\overline{*}} b, gab^{-1}).$$

The projection $p : \tilde{Q} \to Q$ given by $p(a, g) = a$ is a surjective quandle homomorphism. It becomes an equivariant map when we let G' act on Q by conjugation and on $\tilde{Q}$ by $(a, g)^b = (a^b, gb)$. In both cases G' acts transitively and as a group of inner automorphisms. □

The construction of the quandle $\tilde{Q}(G, x)$ has been tailor-made to capture longitude information. Considered purely algebraically, it is a covering in the following sense:

Definition 3.17. A surjective quandle homomorphism $p : \tilde{Q} \to Q$ is called a *covering* if $p(\tilde{x}) = p(\tilde{y})$ implies $\tilde{a} * \tilde{x} = \tilde{a} * \tilde{y}$ for all $\tilde{a}, \tilde{x}, \tilde{y} \in \tilde{Q}$. In other words, the inner representation $\tilde{Q} \to \mathrm{Inn}(\tilde{Q})$ factors through p. This property allows us to define an action of Q on $\tilde{Q}$ by setting $\tilde{a} * x := \tilde{a} * \tilde{x}$ with $\tilde{x} \in p^{-1}(x)$.

In the construction of Lemma 3.16, the projection $p : \tilde{Q} \to Q$ is a covering map. Moreover, covering transformations are given by the left action of $\Lambda = C(x) \cap G'$ defined by $\lambda \cdot (a, g) = (a, \lambda g)$. This action satisfies the following axioms:

(E1) $(\lambda\tilde{x}) * \tilde{y} = \lambda(\tilde{x} * \tilde{y})$ and $\tilde{x} * (\lambda\tilde{y}) = \tilde{x} * \tilde{y}$ for all $\tilde{x}, \tilde{y} \in \tilde{Q}$ and $\lambda \in \Lambda$.

(E2) Λ acts freely and transitively on each fibre $p^{-1}(x)$.

Axiom (E1) is equivalent to saying that Λ acts by automorphisms and the left action of Λ commutes with the right action of $\mathrm{Inn}(\tilde{Q})$. We denote such an action by $\Lambda \curvearrowright \tilde{Q}$. In this situation the quotient $Q := \Lambda\backslash\tilde{Q}$ carries a unique quandle structure that turns the projection $p : \tilde{Q} \to Q$ into a quandle covering.

Definition 3.18. An *extension* $E : \Lambda \curvearrowright \tilde{Q} \to Q$ consists of a surjective quandle homomorphism $\tilde{Q} \to Q$ and a group action $\Lambda \curvearrowright \tilde{Q}$ satisfying axioms (E1) and (E2). We call E a *central extension* if Λ is abelian.

Quandle extensions are an analogue of group extensions, and central quandle extensions come as close as possible to imitating central group extensions. Analogous to the case of groups, central quandle extensions are classified by the second cohomology group $H^2(Q, \Lambda)$. More precisely:

Theorem 3.19 [Eisermann 2003]. *Let Q be a quandle, let Λ be an abelian group, and let $\mathscr{E}(Q, \Lambda)$ be the set of equivalence classes of central extensions of Q by Λ. Given a central extension $E : \Lambda \curvearrowright \tilde{Q} \to Q$, each section $s : Q \to \tilde{Q}$ defines a 2-cocycle $\lambda : Q \times Q \to \Lambda$. If s' is another section, then the associated 2-cocycle λ' differs from λ by a 2-coboundary. The map $E \mapsto [\lambda]$ so constructed induces a natural bijection $\mathscr{E}(Q, \Lambda) \cong H^2(Q, \Lambda)$.* □

The relevant portion of the cochain complex $C^1 \xrightarrow{\delta^1} C^2 \xrightarrow{\delta^2} C^3$ is formed by n-cochains $\lambda : Q^n \to \Lambda$ satisfying $\lambda(a_1, \ldots, a_n) = 0$ whenever $a_i = a_{i+1}$ for some index i, and the first two coboundary operators $\delta^1(\mu)(a, b) = \mu(a) - \mu(a^b)$ and $\delta^2(\lambda)(a, b, c) = \lambda(a, c) - \lambda(a, b) + \lambda(a^c, b^c) - \lambda(a^b, c)$. For details, see [Carter et al. 1999; 2003b; Eisermann 2003].

3D. ***From colouring polynomials to state-sum invariants.*** Let D be a knot diagram and let f be a colouring of D with colours in Q. Suppose that Λ is an abelian group, written multiplicatively, and that $\lambda : Q^2 \to \Lambda$ is a 2-cocycle. For each coloured crossing p as in Figure 4, we define its *weight* by $\langle\lambda|p\rangle := \lambda(a, b)^{\varepsilon}$. The total weight of the colouring f is the product $\langle\lambda|f\rangle := \prod_p \langle\lambda|p\rangle$ over all crossings p. The *state-sum* of the diagram D is defined to be $S^{\lambda}_{Q}(D) := \sum_f \langle\lambda|f\rangle$, where the sum in $\mathbb{Z}\Lambda$ is taken over all colourings $f : D \to Q$. We recall some results:

Lemma 3.20 [Carter et al. 1999; 2003b]. *The state-sum S^{λ}_{Q} is invariant under Reidemeister moves and thus defines a knot invariant $S^{\lambda}_{Q} : \mathcal{K} \to \mathbb{Z}\Lambda$.* □

Lemma 3.21 [Carter et al. 2003b, Prop. 4.5]. *If the colouring $f : D \to Q$ is closed, that is $f(0) = f(n)$, then the weight $\langle\lambda|f\rangle$ is invariant under addition of coboundaries. As a consequence, the state sum S^{λ}_{Q} of a closed knot depends only on the cohomology class $[\lambda]$.* □

Lemma 3.22 [Eisermann 2005, Lem. 32]. *The diagonal action of* $\mathrm{Inn}(Q)$ *on* Q^n *induces the trivial action on* $H^*(Q, \Lambda)$. *As a consequence, for each closed colouring* $f : D \to Q$ *and every inner automorphism* $g \in \mathrm{Inn}(Q)$ *we have* $\langle \lambda | f^g \rangle = \langle {}^g\lambda | f \rangle = \langle \lambda | f \rangle$. □

This last result is well-known in group cohomology [Brown 1982, Prop. II.6.2]. It seems to be folklore in quandle cohomology, but I could not find a written account of it. The necessary argument is provided by [Eisermann 2005, Lem. 32] in the more general setting of Yang–Baxter cohomology, which immediately translates to Lemma 3.22.

Lemma 3.23 [Eisermann 2003, Lem. 50]. *Let* $p : (\tilde{Q}, \tilde{q}) \to (Q, q)$ *be a central quandle extension. Given a long knot diagram* D, *every colouring* $f : (D, 0) \to (Q, q)$ *uniquely lifts to a colouring* $\tilde{f} : (D, 0) \to (\tilde{Q}, \tilde{q})$ *such that* $f = p\tilde{f}$. *If* f *is closed then* $\tilde{f}(n) = \langle \lambda | f \rangle \cdot \tilde{q}$, *where* $[\lambda] \in H^2(Q, \Lambda)$ *is the cohomology class associated with the extension* p. □

These preliminaries being in place, we can now prove that every colouring polynomial P_G^x can be presented as a 2-cocycle state-sum invariant, provided that the subgroup $\Lambda = C(x) \cap G'$ is abelian.

Theorem 3.24. *Suppose that* G *is a colouring group with basepoint* x *such that the subgroup* $\Lambda = C(x) \cap G'$ *is abelian. Then the colouring polynomial* P_G^x *can be presented as a quandle 2-cocycle state-sum invariant. More precisely, the quandle* $Q = x^G$ *admits a 2-cocycle* $\lambda \in Z^2(Q, \Lambda)$ *such that* $S_Q^\lambda = P_G^x \cdot |Q|$.

Proof. Let $Q = x^G$ be the conjugacy class of x in the group G, and let $\tilde{Q} = \tilde{Q}(G, x)$ be the covering quandle constructed in Lemma 3.16. Since Λ is abelian, we obtain a central extension $\Lambda \curvearrowright \tilde{Q} \to Q$. Let $[\lambda] \in H^2(Q, \Lambda)$ be the associated cohomology class. As basepoints we choose $q = x$ in Q and $\tilde{q} = (x, 1)$ in $\tilde{Q}$.

Let D be a long diagram of some knot K, let $f : (D, 0) \to (Q, q)$ be a colouring, let $\rho : (\pi_K, m_K) \to (G, x)$ be the corresponding knot group homomorphism, and let $\tilde{f} : (D, 0) \to (\tilde{Q}, \tilde{q})$ be the lifting of f. On the one hand we have $\tilde{f}(n) = (x, \langle \lambda | f \rangle)$ from Lemma 3.23. On the other hand we have $\tilde{f}(n) = (x, \rho(l_K))$ from the Wirtinger presentation. Thus $\rho(l_K) = \langle \lambda | f \rangle$, and summing over all colourings $f : (D, 0) \to (Q, q)$ yields $P_G^x(K)$.

To obtain the state-sum S_Q^λ we have to sum over all colourings $f : D \to Q$. We have $\mathrm{Col}(D, Q) = \bigcup_{q' \in Q} \mathrm{Col}(D, 0; Q, q')$. Since Q is connected, for each $q' \in Q$ there exists $g \in G$ such that $q^g = q'$. Hence $f \mapsto f^g$ establishes a bijection between $\mathrm{Col}(D, 0; Q, q)$ and $\mathrm{Col}(D, 0; Q, q')$. By Lemma 3.22 we have $\langle \lambda | f \rangle = \langle \lambda | f^g \rangle$. Thus the state-sum over all colourings $f : (D, 0) \to (Q, q')$ again yields P_G^x. We conclude that $S_Q^\lambda(K) = P_G^x(K) \cdot |Q|$. □

3E. ***From state-sum invariants to colouring polynomials.*** Theorem 3.24 has the following converse, which allows us to express quandle 2-cocycle state-sum invariants by knot colouring polynomials.

Theorem 3.25. *Every quandle 2-cocycle state-sum invariant of knots is the specialization of some knot colouring polynomial. More precisely, suppose that Q is a connected quandle, Λ is an abelian group, and $\lambda \in Z^2(Q, \Lambda)$ is a 2-cocycle with associated invariant $S^\lambda_Q : \mathcal{K} \to \mathbb{Z}\Lambda$. Then there exists a group G with basepoint x and a linear map $\varphi : \mathbb{Z}G \to \mathbb{Z}\Lambda$ such that the colouring polynomial $P^x_G : \mathcal{K} \to \mathbb{Z}G$ satisfies $S^\lambda_Q = \varphi P^x_G \cdot |Q|$.*

Proof. We first construct a suitable group (G, x) together with a linear map $\varphi : \mathbb{Z}G \to \mathbb{Z}\Lambda$. Let $\Lambda \curvearrowright \tilde{Q} \xrightarrow{p} Q$ be the central extension associated with the 2-cocycle λ, as explained in Theorem 3.19. We put $G := \mathrm{Inn}(\tilde{Q})$. The inner representation $\tilde{\varrho} : \tilde{Q} \to G$ defines an augmented quandle in the sense of Section 3B. We choose a basepoint $\tilde{q} \in \tilde{Q}$ and set $x := \tilde{\varrho}(\tilde{q})$.

We choose $q = p(\tilde{q})$ as basepoint of Q. Let $s : Q \to \tilde{Q}$ be a section that realizes the 2-cocycle λ. Since p is a covering, we obtain a representation $\varrho : Q \to G$ by $\varrho = \tilde{\varrho} \circ s$. Conversely, we can define an action of G on Q by setting $a^g = p(s(a)^g)$. This turns the representation $\varrho : Q \to G$ into an augmentation and $p : \tilde{Q} \to Q$ into an equivariant map. Our notation being in place, we can now define the linear map

$$\varphi : \mathbb{Z}G \to \mathbb{Z}\Lambda \quad \text{by setting} \quad \varphi(g) = \begin{cases} 0 & \text{if } q^g \neq q, \\ \ell & \text{if } q^g = q \text{ and } \ell \in \Lambda \text{ such that } \tilde{q}^g = \ell \cdot \tilde{q}. \end{cases}$$

It remains to prove that $S^\lambda_Q = \varphi P^x_G \cdot |Q|$. Let K be a knot represented by a long knot diagram D. The Lifting Lemma 3.14 grants us a bijection between closed colourings $f : (D, 0) \to (Q, q)$ and those homomorphisms $\rho : (\pi_K, m_K) \to (G, x)$ that satisfy $q^{\rho(l_K)} = q$. Regarding the covering $\tilde{Q}$, we claim that $\tilde{q}^{\rho(l_K)} = \langle \lambda | f \rangle \cdot \tilde{q}$. To see this, let $\tilde{f} : (D, 0) \to (\tilde{Q}, \tilde{q})$ be the lifting of f. On the one hand we can apply the Lifting Lemma 3.14 to the augmentation $\tilde{Q} \to G$, which yields $\tilde{f}(n) = \tilde{q}^{\rho(l_K)}$. On the other hand we can apply Lemma 3.23, which yields $\tilde{f}(n) = \langle \lambda | f \rangle \cdot \tilde{q}$.

The map φ thus specializes the knot colouring polynomial $P^x_G(K)$ to the state-sum $\sum_f \langle \lambda | f \rangle$, at least if we restrict the summation to colourings $f : (D, 0) \to (Q, q)$. Since Q is connected, any other basepoint q' yields the same state-sum by Lemma 3.22. Summing over all $q' \in Q$, we thus obtain $S^\lambda_Q = \varphi P^x_G \cdot |Q|$, as claimed. □

4. Colouring polynomials are Yang–Baxter invariants

Freyd and Yetter [1989] have shown that the colouring number $F^x_G : \mathcal{K} \to \mathbb{Z}$ is a Yang–Baxter invariant. This means that F^x_G can be obtained as the trace of a linear braid group representation arising from a suitable Yang–Baxter operator c.

In this section we will show that the colouring polynomial $P_G^x : \mathcal{H} \to \mathbb{Z}\Lambda$ is also a Yang–Baxter invariant, obtained from a certain Yang–Baxter operator $\tilde{c}$ defined below. It will follow from our construction that $\tilde{c}$ is a deformation of c over $\mathbb{Z}\Lambda$.

4A. ***Braid group representations and Yang–Baxter invariants.*** The existence of Yang–Baxter invariants rests on two classical theorems: Artin's presentation of the braid groups and the Alexander–Markov theorem, which we will now recall. Our exposition closely follows [Eisermann 2005] and is included here for convenience.

Theorem 4.1 [Artin 1947]. *The braid group on n strands can be presented as*

$$\mathrm{B}_n = \left\langle \sigma_1, \dots, \sigma_{n-1} \;\middle|\; \begin{array}{ll} \sigma_i\sigma_j = \sigma_j\sigma_i & \text{for } |i-j| \geq 2 \\ \sigma_i\sigma_j\sigma_i = \sigma_j\sigma_i\sigma_j & \text{for } |i-j| = 1 \end{array} \right\rangle,$$

where the braid σ_i performs a positive half-twist of the strands i and $i+1$.

Definition 4.2. Let $\mathbb{K}$ be a commutative ring and V a $\mathbb{K}$-module. A *Yang–Baxter operator* (or *R-matrix*) is an automorphism $c : V \otimes V \to V \otimes V$ that satisfies the *Yang–Baxter equation*, also called *braid relation*:

$$(c \otimes \mathrm{id}_V)(\mathrm{id}_V \otimes c)(c \otimes \mathrm{id}_V) = (\mathrm{id}_V \otimes c)(c \otimes \mathrm{id}_V)(\mathrm{id}_V \otimes c) \qquad \text{in } \mathrm{Aut}_{\mathbb{K}}(V^{\otimes 3}).$$

Here and in the sequel tensor products are taken over $\mathbb{K}$ if no other ring is indicated.

Corollary 4.3. *Given a Yang–Baxter operator c and some integer $n \geq 2$, we can define automorphisms $c_i : V^{\otimes n} \to V^{\otimes n}$ by setting*

$$c_i = \mathrm{id}_V^{\otimes(i-1)} \otimes c \otimes \mathrm{id}_V^{\otimes(n-i-1)} \quad \text{for } i = 1, \dots, n-1.$$

The Artin presentation implies that there exists, for each n, a unique braid group representation $\rho_c^n : \mathrm{B}_n \to \mathrm{Aut}_{\mathbb{K}}(V^{\otimes n})$ defined by $\rho_c^n(\sigma_i) = c_i$. □

We orient braids from right to left as in Figure 6. Braid groups will act on the left, so that composition of braids corresponds to the usual composition of maps. The passage from braids to links is granted by the closure map $[\,] : \bigcup_n \mathrm{B}_n \to \mathscr{L}$ defined as follows: for each braid β we define its *closure* $[\beta]$ to be the link in $\mathbb{S}^3$ obtained by identifying opposite endpoints, as indicated in Figure 6.

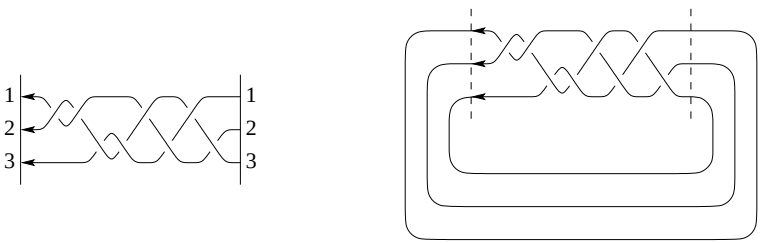


Figure 6. A braid β and its closure $[\beta]$.

Theorem 4.4 (Alexander and Markov; see [Birman 1974]). *Every link can be represented as the closure of some braid. Two braids represent the same link if and only if one can be transformed into the other by a finite sequence of the following* Markov moves:

(M1) *Pass from* $\beta \in \mathrm{B}_n$ *to* $\beta\sigma_n^{\pm 1} \in \mathrm{B}_{n+1}$, *or vice versa.* (stabilization)

(M2) *Pass from* $\beta \in \mathrm{B}_n$ *to* $\alpha^{-1}\beta\alpha$ *with* $\alpha \in \mathrm{B}_n$. (conjugation)

Constructing a link invariant $F : \mathscr{L} \to \mathbb{K}$ is thus equivalent to constructing a map $F : \bigcup_n \mathrm{B}_n \to \mathbb{K}$ that is invariant under Markov moves. The most natural approach is to consider traces of linear braid group representations: invariance under conjugation is automatic, so we only have to require invariance under stabilization:

Definition 4.5. Suppose that V is a free $\mathbb{K}$-module with finite basis. Let $c : V \otimes V \to V \otimes V$ be a Yang–Baxter operator. An automorphism $m : V \to V$ is called *Markov operator* for c if it satisfies

(m1) $\mathrm{tr}_2(\,(m \otimes m) \circ c^{\pm 1}\,) = m$, (the trace condition)

(m2) $c \circ (m \otimes m) = (m \otimes m) \circ c$. (commutativity)

Here the partial trace $\mathrm{tr}_2 : \mathrm{End}(V \otimes V) \to \mathrm{End}(V)$ is defined as follows. Let $(v_1, \ldots, v_n)$ be a basis of V over $\mathbb{K}$. Every $f \in \mathrm{End}(V \otimes V)$ uniquely corresponds to a matrix $f_{ij}^{k\ell}$ such that $f(v_i \otimes v_j) = \sum_{k,\ell} f_{ij}^{k\ell}\, v_k \otimes v_\ell$. We can then define $g = \mathrm{tr}_2(f) \in \mathrm{End}(V)$, $g(v_i) = \sum_k g_i^k\, v_k$, by the matrix $g_i^k = \sum_j f_{ij}^{kj}$. See [Kassel 1995, §II.3].

Corollary 4.6. *Given a Yang–Baxter operator* c *with Markov operator* m, *we define a family of maps* $F_n : \mathrm{B}_n \to \mathbb{K}$ *by* $F_n(\beta) = \mathrm{tr}(m^{\otimes n} \circ \rho_c^n(\beta))$. *Then the induced map* $F : \bigcup_n \mathrm{B}_n \to \mathbb{K}$ *is invariant under both Markov moves and thus defines a link invariant* $F : \mathscr{L} \to \mathbb{K}$. □

The proof of this corollary is straight-forward: the trace condition (m1) implies invariance under stabilization (M1), and commutativity (m2) implies invariance under conjugation (M2). Much more intricate is the question how to actually *find* such a Yang–Baxter–Markov operator (c, m). Attempts to construct solutions in a systematic way have led to the theory of quantum groups [Drinfel'd 1987]. For details we refer to the concise introduction [Kassel et al. 1997] or the textbook [Kassel 1995].

Remark 4.7. For some Yang–Baxter operators c there does not exist any Markov operator m at all. If it exists, m is in general not the identity, as in the case of the Jones polynomial or other quantum invariants. The Yang–Baxter operators derived from knot diagram colourings below are very special in that they allow the Markov operator $m = \mathrm{id}$, which is equivalent to saying that $\mathrm{tr}_2(c^{\pm 1}) = \mathrm{id}$.

4B. ***Colouring polynomials of long knots.*** Before we consider colouring polynomials, let us first recall how colouring numbers can be obtained from a suitable Yang–Baxter operator.

Theorem 4.8 [Freyd and Yetter 1989, Prop. 4.2.5 and the remark following its proof]. *Let Q be a quandle and let $\mathbb{K}Q$ be the free $\mathbb{K}$-module with basis Q. The quandle structure of Q can be linearly extended to a Yang–Baxter operator*

$$c_Q : \mathbb{K}Q \otimes \mathbb{K}Q \to \mathbb{K}Q \otimes \mathbb{K}Q \quad \text{with} \quad a \otimes b \mapsto b \otimes (a * b) \quad \text{for all} \quad a, b \in Q.$$

Axiom $(Q2)$ *ensures that c_Q is an automorphism, while Axiom* $(Q3)$ *implies the Yang–Baxter equation. If Q is finite, then* $(Q1)$ *ensures that* $\operatorname{tr}_2(c_Q^{\pm 1}) = \mathrm{id}$. *In this case the corresponding Yang–Baxter invariant $F_Q = \operatorname{tr} \circ \rho_Q$ coincides with the number of Q-colourings* (*defined in Section 3A*) *followed by the ring homomorphism $\mathbb{Z} \to \mathbb{K}$.* □

As an example consider a finite group G with basepoint x. The Yang–Baxter operator constructed from the quandle $Q = x^G$ then leads to the colouring number $F_Q = F_G^x \cdot |Q|$.

We will now move from colouring numbers to colouring polynomials: consider the quandle extension $\Lambda \curvearrowright \tilde{Q} \to Q$ as defined in Section 3C, where the quandle $Q = Q(G, x)$ is covered by $\tilde{Q} = \tilde{Q}(G, x)$, and the deck transformation group is $\Lambda = C(x) \cap G'$. As before, we linearly extend the quandle structure of $\tilde{Q}$ to a Yang–Baxter operator $c_{\tilde{Q}}$, and denote the associated linear braid group representation by $\rho_{\tilde{Q}}$. We will, however, not take the total trace as before, but rather use the partial trace $\operatorname{tr}' : \operatorname{End}_{\mathbb{K}}(\mathbb{K}\tilde{Q}^{\otimes n}) \to \operatorname{End}_{\mathbb{K}}(\mathbb{K}\tilde{Q})$, contracting the tensor factors $2, \ldots, n$.

Theorem 4.9. *Let (G, x) be a finite group such that the conjugacy class $Q = x^G$ generates G. Let $\tilde{Q} = \tilde{Q}(G, x)$ be the covering quandle and let $\rho_{\tilde{Q}}$ be the associated braid group representation. Suppose that the knot K is represented by a braid β. Then the partial trace $\operatorname{tr}'(\rho_{\tilde{Q}}(\beta)) : \mathbb{K}\tilde{Q} \to \mathbb{K}\tilde{Q}$ is given by multiplication with $P_G^x(K)$.*

Note that the free left action of Λ on $\tilde{Q}$ turns $\mathbb{K}\tilde{Q}$ into a free left module over $\mathbb{K}\Lambda$. In particular, multiplication by $P_G^x(K)$ is a $\mathbb{K}$-linear endomorphism. If $\mathbb{K}$ is of characteristic 0, then the endomorphism $\operatorname{tr}'(\rho_{\tilde{Q}}(\beta))$ uniquely determines $P_G^x(K)$.

Proof. We use the obvious bases $\tilde{Q}$ for $\mathbb{K}\tilde{Q}$ and $\tilde{Q}^n$ for $\mathbb{K}\tilde{Q}^{\otimes n}$. Each endomorphism $f : \mathbb{K}\tilde{Q}^{\otimes n} \to \mathbb{K}\tilde{Q}^{\otimes n}$ is then represented by a matrix $M_{q_1 q_2 \ldots q_n}^{p_1 p_2 \ldots p_n}$, indexed by elements p_i and q_j in the basis $\tilde{Q}$. The partial trace $\operatorname{tr}'(f) : \mathbb{K}\tilde{Q} \to \mathbb{K}\tilde{Q}$ is given by the matrix $T_{q_1}^{p_1} = \sum M_{q_1 p_2 \ldots p_n}^{p_1 p_2 \ldots p_n}$, where the sum is taken over all repeated indices $p_2, \ldots, p_n$.

By construction, each elementary braid σ_i acts as a permutation on the basis $\tilde{Q}^n$, thus each braid $\beta \in \mathrm{B}_n$ is represented by a permutation matrix with respect to

this basis. We interpret this action as colouring the braid β with elements of $\tilde{Q}$: we colour the right ends of the braid with $v = p_1 \otimes \cdots \otimes p_n$. Moving from right to left, at each crossing the new arc is coloured according to the Wirtinger rule as depicted in Figure 4. We thus arrive at the left ends of the braid being coloured with $\rho(\beta)v = q_1 \otimes \cdots \otimes q_n$. We conclude that colourings of the braid β that satisfy the trace conditions $p_2 = q_2, \ldots, p_n = q_n$ are in natural bijection with colourings of the corresponding long knot K.

We now turn to the remaining indices p_1 and q_1. Let us first consider the special case $p_1 = (x, 1)$ and $q_1 = (y, \lambda)$. From the preceding argument we see that $T_{q_1}^{p_1}$ equals the number of $\tilde{Q}$-colourings of the long knot K that start with $(x, 1)$ and end with (y, λ). According to Lemma 3.12, such colourings exist only for $y = x$ and $\lambda \in \Lambda$, hence we have $q_1 = \lambda \cdot p_1$. We conclude that $T_{q_1}^{p_1}$ equals the number of representations $(\pi_K, m_K, l_K) \to (G, x, \lambda)$. In total we get

$$\mathrm{tr}'(\rho(\beta))\,(p_1) = P_G^x(K) \cdot p_1.$$

The preceding construction is equivariant under the right-action of the group G' on the covering quandle $\tilde{Q}$. According to Lemma 3.16 this action is transitive: for every $p \in \tilde{Q}$ there exists $g \in G'$ and $p = p_1^g$, so we conclude that

$$\mathrm{tr}'(\rho(\beta))\,(p) = P_G^x(K) \cdot p.$$

This means that the endomorphism $\mathrm{tr}'(\rho(\beta)) : \mathbb{K}\tilde{Q} \to \mathbb{K}\tilde{Q}$ is given by multiplication with $P_G^x(K)$. □

Remark 4.10. The partial trace $\mathrm{tr}' : \mathrm{End}_{\mathbb{K}}(\mathbb{K}\tilde{Q}^{\otimes n}) \to \mathrm{End}_{\mathbb{K}}(\mathbb{K}\tilde{Q})$ corresponds to closing the strands $2, \ldots, n$ of the braid β, but leaving the first strand open: the object thus represented is a long knot. The natural setting for such constructions is the category of tangles and its linear representations [Kassel 1995]. The previous theorem then says that the long knot K is represented by the endomorphism $\mathbb{K}\tilde{Q} \to \mathbb{K}\tilde{Q}$ that is given by multiplication with $P_G^x(K)$.

If we used the complete trace $\mathrm{tr} : \mathrm{End}_{\mathbb{K}}(\mathbb{K}\tilde{Q}^{\otimes n}) \to \mathbb{K}$ instead, then we would obtain a different invariant $F_{\tilde{Q}} = \mathrm{tr} \circ \rho_{\tilde{Q}}$. By the preceding arguments, $F_{\tilde{Q}}(K)$ equals $|\tilde{Q}|$ times the number of representations $(\pi_K, m_K, l_K) \to (G, x, 1)$, which corresponds to the coefficient of the unit element in the colouring polynomial $P_G^x(K)$.

4C. ***Colouring polynomials of closed knots.*** We now show how the colouring polynomial P_G^x of closed knots can be obtained as the trace of a suitable Yang–Baxter representation. To this end we will modify the construction of the preceding paragraph in order to replace the partial trace tr' by the complete trace tr.

We proceed as follows: the quandle $Q = x^G$ admits an extension $\Lambda \curvearrowright \tilde{Q} \to Q$ as defined in Section 3C. The quandle structure of $\tilde{Q}$ linearly extends to a Yang–Baxter operator $c_{\tilde{Q}}$ on $\mathbb{K}\tilde{Q}$. The free Λ-action on $\tilde{Q}$ turns $\mathbb{K}\tilde{Q}$ into a free module over $\mathbb{A} = \mathbb{K}\Lambda$. If Λ is abelian, we can pass to an $\mathbb{A}$-linear operator

$$\tilde{c}_Q : \mathbb{K}\tilde{Q} \otimes_{\mathbb{A}} \mathbb{K}\tilde{Q} \to \mathbb{K}\tilde{Q} \otimes_{\mathbb{A}} \mathbb{K}\tilde{Q} \quad \text{with} \quad \tilde{a} \otimes \tilde{b} \mapsto \tilde{b} \otimes (\tilde{a} * \tilde{b}) \quad \text{for all} \quad \tilde{a}, \tilde{b} \in \tilde{Q}.$$

The difference between $c_{\tilde{Q}}$ and $\tilde{c}_Q$ is that the tensor product is now taken over $\mathbb{A}$, which means that everything is bilinear with respect to multiplication by $\lambda \in \Lambda$. In the following theorem and its proof all tensor products are to be taken over the ring $\mathbb{A}$, but for notational simplicity we will write $\otimes$ for $\otimes_{\mathbb{A}}$.

Theorem 4.11. *If (G, x) is a colouring group such that $\Lambda = C(x) \cap G'$ is abelian, then the colouring polynomial $P_G^x : \mathcal{K} \to \mathbb{Z}\Lambda$ is a Yang–Baxter invariant. More precisely, the preceding construction yields a Yang–Baxter–Markov operator $(\tilde{c}_Q, \mathrm{id})$ over the ring $\mathbb{A} = \mathbb{K}\Lambda$, and the associated knot invariant satisfies $\tilde{F}_Q = \varphi P_G^x \cdot |Q|$ where $\varphi : \mathbb{Z}\Lambda \to \mathbb{K}\Lambda$ is the natural ring homomorphism defined by $\varphi(\lambda) = \lambda$ for all $\lambda \in \Lambda$.*

If $\mathbb{K}$ is of characteristic 0, then $\tilde{F}_Q$ is equivalent to the knot colouring polynomial P_G^x. If $\mathbb{K}$ is of finite characteristic, then we may lose some information and $\tilde{F}_Q$ is usually weaker than P_G^x. In the worst case $|Q|$ vanishes in $\mathbb{K}$ and $\tilde{F}_Q$ becomes trivial.

Proof. It is a routine calculation to prove that $\tilde{c}_Q$ is a Yang–Baxter operator over $\mathbb{A}$: as before, axiom (Q2) implies that $\tilde{c}_Q$ is an automorphism, while axiom (Q3) ensures that $\tilde{c}_Q$ satisfies the Yang–Baxter equation. Axiom (Q1) implies the trace condition $\mathrm{tr}_2(\tilde{c}_Q^{\pm 1}) = \mathrm{id}$, hence $(\tilde{c}_Q, \mathrm{id})$ is a Yang–Baxter–Markov operator. We thus obtain a linear braid group representation $\tilde{\rho}_Q^n : \mathrm{B}_n \to \mathrm{Aut}_{\mathbb{A}}(\mathbb{K}\tilde{Q}^{\otimes n})$, whose character $\tilde{F}_Q = \mathrm{tr} \circ \tilde{\rho}_Q$ is Markov invariant and induces a link invariant $\tilde{F}_Q : \mathcal{L} \to \mathbb{A}$. Restricted to knots we claim that $\tilde{F}_Q = P_G^x \cdot |Q|$. The proof of the theorem parallels the proof of Theorem 4.9, but requires some extra care.

To represent $\tilde{c}_Q$ by a matrix, we have to choose a basis of $\mathbb{K}\tilde{Q}$ over $\mathbb{A}$. Let $s : Q \to \tilde{Q}$ be a section to the central extension $\Lambda \curvearrowright \tilde{Q} \to Q$. Then $B = s(Q)$ is a basis of $\mathbb{K}\tilde{Q}$ as an $\mathbb{A}$-module. For the basepoint x we can assume $s(x) = (x, 1)$, but otherwise there are no canonical choices. In general, s will not (and cannot) be a homomorphism of quandles, but we have $s(a) * s(b) = \lambda(a, b) \cdot s(a * b)$ with a certain 2-cocycle $\lambda : Q \times Q \to \Lambda$ that measures the deviation of s from being a homomorphism. Just as c_Q is represented by a permutation matrix, we see that $\tilde{c}_Q$ is represented by the same matrix except that the 1's are replaced with the elements $\lambda(a, b) \in \Lambda$. This is usually called a *monomial matrix* or *generalized permutation matrix*.

Since $\mathbb{K}\tilde{Q}$ is a free $\mathbb{A}$-module with finite basis $B = s(Q)$, the tensor product $\mathbb{K}\tilde{Q}^{\otimes n}$ is also free and has finite basis B^n. The trace $\mathrm{tr} \circ \tilde{\rho}(\beta)$ is calculated as the sum $\sum_{v \in B^n} \langle \tilde{\rho}(\beta)v | v \rangle$. Note that $\tilde{\rho}(\beta)$ is again a monomial matrix in the sense that each row and each column has exactly one nonzero entry. Hence a vector $v \in B^n$ contributes to the trace sum if and only if $\tilde{\rho}(\beta)v = \lambda(v)v$ with some $\lambda(v) \in \Lambda$. It remains to characterize eigenvectors and identify their eigenvalues.

Given a braid $\beta \in \mathrm{B}_n$ we can interpret the action of $\tilde{\rho}(\beta)$ as colouring the braid β: we colour the right ends of the braid with a basis vector $v \in B^n$,

$$v = (a_1, g_1) \otimes (a_2, g_2) \otimes \ldots \otimes (a_n, g_n).$$

Moving from right to left, at each crossing the new arc is coloured according to the Wirtinger rule as depicted in Figure 4. We thus arrive at the left ends of the braid, being coloured with

$$\tilde{\rho}(\beta)v = (b_1, h_1) \otimes (b_2, h_2) \otimes \cdots \otimes (b_n, h_n).$$

Since the tensor product is defined over $\mathbb{A}$, we have $\tilde{\rho}(\beta)v = \lambda(v)v$ if and only if $a_1 = b_1, a_2 = b_2, \ldots, a_n = b_n$. Hence each eigenvector $v \in B^n$ naturally corresponds to a Q-colouring of the closed braid $K = [\beta]$.

In order to identify the eigenvalue $\lambda(v)$, we will further assume that $(a_1, g_1) = (x, 1)$, where x is the basepoint of G. Such an eigenvector will be called *normalized*. Using the tensor product-structure over $\mathbb{A} = \mathbb{K}\Lambda$, we obtain

$$\tilde{\rho}(\beta)v = (x, \lambda) \otimes (a_2, g_2) \otimes \cdots \otimes (a_n, g_n) = \lambda(v)v$$

as in the proof of Theorem 4.9. We conclude that each normalized eigenvector $v \in B^n$ with $\tilde{\rho}(\beta)v = \lambda(v)v$ corresponds to a $\tilde{Q}$-colouring of the long knot, where the first arc is coloured by $(x, 1)$ and the last arc is coloured by (x, λ). This means that the eigenvalue $\lambda(v)$ is the associated colouring longitude.

We finally show that $\tilde{F}_Q = P_G^x \cdot |Q|$ by calculating the trace $\sum_{v \in B^n} \langle \tilde{\rho}(\beta)v | v \rangle$. Normalized eigenvectors $v \in \{(x, 1)\} \times B^{n-1}$ with $\tilde{\rho}(\beta)v = \lambda(v)v$ correspond to colourings $\rho : (\pi_K, m_k) \to (G, x)$ with $\rho(l_K) = \lambda(v)$. Summing over these vectors only, we thus obtain the colouring polynomial $P_G^x(K)$. To calculate the total sum we use again the fact that the right-action of G' on $\tilde{Q}$ is transitive. Hence for every $q \in Q$ there exists $g \in G'$ such that $s(q)^g = (x, 1)$. The action of g induces a bijection between the set of basis vectors $\{s(q)\} \times B^{n-1}$ and $\{(x, 1)\} \times B^{n-1}$. Since the preceding trace calculation is G'-invariant, each vector $v \in \{s(q)\} \times B^{n-1}$ contributes $P_G^x(K)$ to the trace. In total we obtain $\tilde{F}_Q = P_G^x \cdot |Q|$, as claimed. □

4D. ***Concluding remarks.*** It follows from our construction that $\tilde{c}_Q$ is a deformation of the Yang–Baxter operator c_Q. More precisely we have $\tilde{c}_Q(a \otimes b) = \lambda(a, b) \cdot c_Q(a, b)$ for all $a, b \in Q$ with a suitable map $\lambda : Q \times Q \to \Lambda$. Our

construction via quandle coverings and central extensions provides a geometric interpretation in terms of meridian-longitude information. This interpretation carries through all steps of our construction, which finally allows us to interpret the resulting Yang–Baxter invariant as a colouring polynomial.

Conversely, it is natural to consider the ansatz $\tilde{c}_Q(a \otimes b) = \lambda(a, b) \cdot c_Q(a, b)$ and to ask which λ turn $\tilde{c}_Q$ into a Yang–Baxter operator. This idea can, though in a restricted form, already be found in [Freyd and Yetter 1989, Thm. 4.2.6]. A direct calculation shows that $\tilde{c}_Q$ is a Yang–Baxter operator if and only if λ is a 2-cocycle in the sense of quandle cohomology. Moreover, two such deformations will be equivalent if the cocycles differ by a coboundary. This observation has been worked out by M. Graña [2002], who independently proved that quandle 2-cocycle state-sum invariants are Yang–Baxter invariants.

Acknowledgements

The author thanks the anonymous referee for a careful reading and numerous helpful comments. The results of Section 2 were part of the author's Ph.D. thesis [Eisermann 2000a], which was financially supported by the Deutsche Forschungsgemeinschaft through the Graduiertenkolleg Mathematik at the University of Bonn. Sections 3 and 4 were elaborated while the author held a postdoc position at the École Normale Supérieure de Lyon, whose hospitality is gratefully acknowledged.

References

[Artin 1947] E. Artin, "Theory of braids", *Ann. of Math.* (2) **48** (1947), 101–126. MR 8,367a Zbl 0030.17703

[Bar-Natan 1995] D. Bar-Natan, "On the Vassiliev knot invariants", *Topology* **34**:2 (1995), 423–472. MR 97d:57004 Zbl 0898.57001

[Birman 1974] J. S. Birman, *Braids, links, and mapping class groups*, Annals of Mathematics Studies **82**, Princeton University Press, Princeton, N.J., 1974. MR 51 #11477 Zbl 0305.57013

[Brieskorn 1988] E. Brieskorn, "Automorphic sets and braids and singularities", pp. 45–115 in *Braids* (Santa Cruz, CA, 1986), edited by J. S. Birman and A. Libgober, Contemp. Math. **78**, Amer. Math. Soc., Providence, RI, 1988. MR 90a:32024 Zbl 0716.20017

[Brown 1982] K. S. Brown, *Cohomology of groups*, Graduate Texts in Mathematics **87**, Springer, New York, 1982. Corrected reprint, 1994. MR 96a:20072 Zbl 0584.20036

[Burde and Zieschang 1985] G. Burde and H. Zieschang, *Knots*, Studies in Mathematics **5**, de Gruyter, Berlin, 1985. MR 87b:57004 Zbl 0568.57001

[Carter et al. 1999] J. S. Carter, D. Jelsovsky, S. Kamada, L. Langford, and M. Saito, "State-sum invariants of knotted curves and surfaces from quandle cohomology", *Electron. Res. Announc. Amer. Math. Soc.* **5** (1999), 146–156. MR 2002c:57014 Zbl 0995.57004

[Carter et al. 2003a] J. S. Carter, M. Elhamdadi, M. A. Nikiforou, and M. Saito, "Extensions of quandles and cocycle knot invariants", *J. Knot Theory Ramifications* **12**:6 (2003), 725–738. MR 2004g:57020 Zbl 1049.57008

[Carter et al. 2003b] J. S. Carter, D. Jelsovsky, S. Kamada, L. Langford, and M. Saito, "Quandle cohomology and state-sum invariants of knotted curves and surfaces", *Trans. Amer. Math. Soc.* **355**:10 (2003), 3947–3989. MR 2005b:57048 Zbl 1028.57003

[Conway 1970] J. H. Conway, "An enumeration of knots and links, and some of their algebraic properties", pp. 329–358 in *Computational Problems in Abstract Algebra (Proc. Conf., Oxford, 1967)*, Pergamon, Oxford, 1970. MR 41 #2661 Zbl 0202.54703

[Conway et al. 1985] J. H. Conway, R. T. Curtis, S. P. Norton, R. A. Parker, and R. A. Wilson, *Atlas of finite groups: Maximal subgroups and ordinary characters for simple groups*, Oxford University Press, Eynsham, 1985. MR 88g:20025 Zbl 0568.20001

[Crowell and Fox 1963] R. H. Crowell and R. H. Fox, *Introduction to knot theory*, Ginn and Co., Boston, 1963. MR 26 #4348 Zbl 0126.39105

[Dehn 1914] M. Dehn, "Die beiden Kleeblattschlingen", *Math. Ann.* **75**:3 (1914), 402–413. English translation in [Dehn 1987]. MR 1511799

[Dehn 1987] M. Dehn, *Papers on group theory and topology*, Springer, New York, 1987. Translated and edited by John Stillwell. MR 88d:01041

[Drinfel'd 1987] V. G. Drinfel'd, "Quantum groups", pp. 798–820 in *Proceedings of the International Congress of Mathematicians* (Berkeley, 1986), vol. 1, Amer. Math. Soc., Providence, RI, 1987. MR 89f:17017

[Eisermann 2000a] M. Eisermann, *Knotengruppen-Darstellungen und Invarianten von endlichem Typ*, Bonner Mathematische Schriften **327**, Universität Bonn Mathematisches Institut, Bonn, 2000. (Author's dissertation). MR 2003g:57014 Zbl 0948.57008

[Eisermann 2000b] M. Eisermann, "The number of knot group representations is not a Vassiliev invariant", *Proc. Amer. Math. Soc.* **128**:5 (2000), 1555–1561. MR 2000j:57009 Zbl 0939.57007

[Eisermann 2003] M. Eisermann, "Homological characterization of the unknot", *J. Pure Appl. Algebra* **177**:2 (2003), 131–157. MR 2003j:57009 Zbl 1013.57002

[Eisermann 2005] M. Eisermann, "Yang-Baxter deformations of quandles and racks", *Algebr. Geom. Topol.* **5** (2005), 537–562. MR 2006b:17023 Zbl 02221853

[Fenn and Rourke 1992] R. Fenn and C. Rourke, "Racks and links in codimension two", *J. Knot Theory Ramifications* **1**:4 (1992), 343–406. MR 94e:57006 Zbl 0787.57003

[Fox 1962] R. H. Fox, "A quick trip through knot theory", pp. 120–167 in *Topology of 3-manifolds and related topics*, Prentice-Hall, Englewood Cliffs, NJ, 1962. MR 25 #3522

[Fox 1970] R. H. Fox, "Metacyclic invariants of knots and links", *Canad. J. Math.* **22** (1970), 193–201. MR 41 #6197 Zbl 0195.54002

[Freyd and Yetter 1989] P. J. Freyd and D. N. Yetter, "Braided compact closed categories with applications to low-dimensional topology", *Adv. Math.* **77**:2 (1989), 156–182. MR 91c:57019 Zbl 0679.57003

[GAP 2006] *GAP: Groups, Algorithms, and Programming, version 4.4.9*, The GAP Group, 2006, Available at http://www.gap-system.org.

[González-Acuña 1975] F. González-Acuña, "Homomorphs of knot groups", *Ann. of Math.* (2) **102**:2 (1975), 373–377. MR 52 #576 Zbl 0323.57010

[Gordon and Luecke 1989] C. M. Gordon and J. Luecke, "Knots are determined by their complements", *J. Amer. Math. Soc.* **2**:2 (1989), 371–415. MR 90a:57006a Zbl 0672.57009

[Graña 2002] M. Graña, "Quandle knot invariants are quantum knot invariants", *J. Knot Theory Ramifications* **11**:5 (2002), 673–681. MR 2003e:57019 Zbl 1027.57014

[Johnson 1980] D. Johnson, "Homomorphs of knot groups", *Proc. Amer. Math. Soc.* **78**:1 (1980), 135–138. MR 80j:57004 Zbl 0435.57003

[Johnson and Livingston 1989] D. Johnson and C. Livingston, "Peripherally specified homomorphs of knot groups", *Trans. Amer. Math. Soc.* **311**:1 (1989), 135–146. MR 89f:57006 Zbl 0662.57003

[Joyce 1982] D. Joyce, "A classifying invariant of knots, the knot quandle", *J. Pure Appl. Algebra* **23**:1 (1982), 37–65. MR 83m:57007 Zbl 0474.57003

[Kassel 1995] C. Kassel, *Quantum groups*, Graduate Texts in Mathematics **155**, Springer, New York, 1995. MR 96e:17041 Zbl 0808.17003

[Kassel et al. 1997] C. Kassel, M. Rosso, and V. Turaev, *Quantum groups and knot invariants*, Panoramas et Synthèses **5**, Société Math. de France, Paris, 1997. MR 99b:57011 Zbl 0878.17013

[Kauffman 1987] L. H. Kauffman, "State models and the Jones polynomial", *Topology* **26**:3 (1987), 395–407. MR 88f:57006 Zbl 0622.57004

[Kauffman 2001] L. H. Kauffman, *Knots and physics*, Third ed., Series on Knots and Everything **1**, World Scientific, River Edge, NJ, 2001. MR 2002h:57012 Zbl 1057.57001

[Kuperberg 1996] G. Kuperberg, "Detecting knot invertibility", *J. Knot Theory Ramifications* **5**:2 (1996), 173–181. MR 97h:57018 Zbl 0858.57011

[Lickorish 1997] W. B. R. Lickorish, *An introduction to knot theory*, Graduate Texts in Mathematics **175**, Springer, New York, 1997. MR 98f:57015 Zbl 0886.57001

[Matveev 1982] S. V. Matveev, "Distributive groupoids in knot theory", *Mat. Sb.* (*N.S.*) **119** (1982), 78–88. In Russian; translated in *Math. USSR Sb.* **47** (1984), 73–83. MR 84e:57008 Zbl 0523.57006

[Neuwirth 1965] L. P. Neuwirth, *Knot groups*, Annals of Mathematics Studies **56**, Princeton University Press, Princeton, N.J., 1965. MR 31 #734 Zbl 0184.48903

[Riley 1971] R. Riley, "Homomorphisms of knot groups on finite groups", *Math. Comp.* **25** (1971), 603–619. Addendum in **25** (1971), no. 115, microfiche suppl. A-B. MR 45 #4399 Zbl 0224.55003

[Thurston 1982] W. P. Thurston, "Three-dimensional manifolds, Kleinian groups and hyperbolic geometry", *Bull. Amer. Math. Soc.* (*N.S.*) **6**:3 (1982), 357–381. MR 83h:57019 Zbl 0496.57005

[Trotter 1963] H. F. Trotter, "Non-invertible knots exist", *Topology* **2** (1963), 275–280. MR 28 #1618 Zbl 0136.21203

[Waldhausen 1968] F. Waldhausen, "On irreducible 3-manifolds which are sufficiently large", *Ann. of Math.* (2) **87** (1968), 56–88. MR 36 #7146 Zbl 0157.30603

[Whitten 1987] W. Whitten, "Knot complements and groups", *Topology* **26**:1 (1987), 41–44. MR 88f:57014 Zbl 0607.57004

Received July 24, 2004. Revised July 25, 2007.

MICHAEL EISERMANN
INSTITUT FOURIER
UNIVERSITÉ GRENOBLE I
100 RUE DES MATHS, BP 74
38402 ST MARTIN D'HÈRES
FRANCE

Michael.Eisermann@ujf-grenoble.fr
http://www-fourier.ujf-grenoble.fr/~eiserm

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.337

SOME NEW SIMPLE MODULAR LIE SUPERALGEBRAS

ALBERTO ELDUQUE

Two new simple modular Lie superalgebras will be obtained in characteristics 3 and 5, which share the property that their even parts are orthogonal Lie algebras and the odd parts their spin modules. The characteristic 5 case will be shown to be related, by means of a construction of Tits, to the exceptional ten-dimensional Jordan superalgebra of Kac.

1. Introduction

There are well-known constructions of the exceptional simple Lie algebras of type E_8 and F_4 which go back to Witt [1941], as $\mathbb{Z}_2$-graded algebras $\mathfrak{g} = \mathfrak{g}_{\bar{0}} \oplus \mathfrak{g}_{\bar{1}}$ with even part the orthogonal Lie algebras $\mathfrak{so}_{16}$ and $\mathfrak{so}_9$ respectively, and odd part given by their spin representations (see [Adams 1996]).

Brown [1982] found a new simple finite-dimensional Lie algebra over fields of characteristic 3 which presents the same pattern, but with $\mathfrak{g}_{\bar{0}} = \mathfrak{so}_7$.

Among the simple Lie superalgebras in Kac's classification [1977b], only the orthosymplectic Lie superalgebra $\mathfrak{osp}(1,4)$ presents the same pattern, since $\mathfrak{g}_{\bar{0}} = \mathfrak{sp}_4$ in this case, and $\mathfrak{g}_{\bar{1}}$ is its natural four-dimensional module. But $\mathfrak{sp}_4$ is isomorphic to $\mathfrak{so}_5$, so $\mathfrak{g}_{\bar{1}}$ is its spin module.

In [Elduque 2006], we found another instance of this phenomenon. There exists a simple Lie superalgebra over fields of characteristic 3 with even part isomorphic to $\mathfrak{so}_{12}$ and odd part its spin module.

This paper is devoted to settling the question of which other simple $\mathbb{Z}_2$-graded Lie algebras or Lie superalgebras display this pattern: the even part being an orthogonal Lie algebra and the odd part its spin module.

It turns out that, besides the previously mentioned examples and the example of $\mathfrak{so}_9$, which is the direct sum of $\mathfrak{so}_8$ and its natural module, but where, because of triality, this natural module can be replaced by the spin module, there appear exactly two other possibilities for Lie superalgebras, one in characteristic 3 with

Supported by the Spanish Ministerio de Educación y Ciencia and FEDER (MTM 2004-081159-C04-02) and by the Diputación General de Aragón (Grupo de Investigación de Álgebra).
MSC2000: primary 17B50; secondary 17B60.
Keywords: Lie superalgebra, Kac Jordan superalgebra, spin module.

even part isomorphic to $\mathfrak{so}_{13}$, and the other in characteristic 5, with even part isomorphic to $\mathfrak{so}_{11}$. These simple Lie superalgebras seem to appear here for the first time.

The characteristic-5 case will be shown to be strongly related to the ten-dimensional simple exceptional Kac Jordan superalgebra, by means of a construction due to Tits. As was proved by McCrimmon [2005], and indirectly hinted in [Elduque and Okubo 2000], the Grassmann envelope of this Jordan superalgebra satisfies the Cayley–Hamilton equation of degree 3 and hence, as shown in [Benkart and Zelmanov 1996] and [Benkart and Elduque 2003], this Jordan superalgebra $\mathcal{J}$ can be plugged into the second component of the Tits construction [1966], the first component being a Cayley algebra. The even part of the resulting Lie superalgebra is then isomorphic to $\mathfrak{so}_{11}$ and the odd part turns out to be its spin module.

The characteristic-3 case is related to the six-dimensional composition superalgebra $B(4, 2)$ (see [Elduque and Okubo 2002; Shestakov 1997]) and, therefore, to the exceptional Jordan superalgebra of 3×3 hermitian matrices $H_3(B(4, 2))$. This will be further discussed in [Cunha and Elduque 2006], where an extended Freudenthal magic square in characteristic 3 is considered.

Convention. *Throughout the paper, k will always denote an algebraically closed field of characteristic $\neq 2$.*

Overview. Section 2 reviews the basic properties of the orthogonal Lie algebras, associated Clifford algebras and spin modules in a way suitable to our purposes. In Section 3 we determine the simple $\mathbb{Z}_2$-graded Lie algebras and the simple Lie superalgebras whose even part is an orthogonal Lie algebra of type B and its odd part its spin module. The two new simple Lie superalgebras mentioned above appear here. Section 4 is devoted to type D, and here the objects that appear are either classical or a Lie superalgebra in characteristic 3 with even part $\mathfrak{so}_{12}$, which appeared for the first time in [Elduque 2006] related to a Freudenthal triple system, which in turn is constructed in terms of the Jordan algebra of the hermitian 3×3 matrices over a quaternion algebra. Finally, Section 5 is devoted to study the relationship of the exceptional Lie superalgebra that has appeared in characteristic 5, with even part isomorphic to $\mathfrak{so}_{11}$, to the Lie superalgebra obtained by means of Tits construction in terms of the Cayley algebra and of the exceptional ten-dimensional Jordan superalgebra of Kac.

2. Spin modules

Let V be a vector space of dimension $l \geq 1$ over the field k, let V^* be its dual vector space, and consider the $(2l+1)$-dimensional vector space $W = ku \oplus V \oplus V^*$, with

the regular quadratic form q given by

$$q(\alpha u + v + f) = -\alpha^2 + f(v), \tag{2.1}$$

for any $\alpha \in k$, $v \in V$ and $f \in V^*$.

Let $\mathfrak{Cl}(V \oplus V^*, q)$ be the Clifford algebra of the restriction of q to $V \oplus V^*$, and let $\mathfrak{Cl}_{\bar{0}}(W, q)$ be the even Clifford algebra of q. As a general rule, the multiplication in Clifford algebras will be denoted by a dot: $x \cdot y$. The linear map

$$V \oplus V^* \to \mathfrak{Cl}_{\bar{0}}(W, q) : \ x \mapsto u \cdot x = \tfrac{1}{2}(u \cdot x - x \cdot u) = \tfrac{1}{2}[u, x]^{\cdot}$$

extends to an algebra isomorphism

$$\Psi : \mathfrak{Cl}(V \oplus V^*, q) \to \mathfrak{Cl}_{\bar{0}}(W, q). \tag{2.2}$$

Let τ be the involution of $\mathfrak{Cl}(W, q)$ such that $\tau(w) = w$ for any $w \in W$, and let $\tau_{\bar{0}}$ be its restriction to $\mathfrak{Cl}_{\bar{0}}(W, q)$. Let τ' be the involution of $\mathfrak{Cl}(V \oplus V^*, q)$ such that $\tau'(x) = -x$ for any $x \in V \oplus V^*$. Then, for any $x \in V \oplus V^*$,

$$\tau_{\bar{0}}\big(\Psi(x)\big) = \tau_{\bar{0}}(u \cdot x) = \tau(x) \cdot \tau(u) = x \cdot u = -u \cdot x = u \cdot \tau'(x) = \Psi\big(\tau'(x)\big),$$

so Ψ in (2.2) is actually an isomorphism of algebras with involution:

$$\Psi : \big(\mathfrak{Cl}(V \oplus V^*, q), \tau'\big) \to \big(\mathfrak{Cl}_{\bar{0}}(W, q), \tau_{\bar{0}}\big). \tag{2.3}$$

Now consider the exterior algebra $\bigwedge V$. Multiplication here will be denoted by juxtaposition. This conveys a natural grading over $\mathbb{Z}_2$: $\bigwedge V = \bigwedge_{\bar{0}} V \oplus \bigwedge_{\bar{1}} V$. In other words, like Clifford algebras, $\bigwedge V$ is an associative superalgebra. For any $f \in V^*$, let $df : \bigwedge V \to \bigwedge V$ be the unique odd superderivation such that $(df)(v) = f(v)$ for any $v \in V \subseteq \bigwedge V$ (see, for instance, [Knus et al. 1998, §8]). Note that $(df)^2 = 0$.

Also, for any $v \in V$, the left multiplication by v gives an odd linear map $l_v : \bigwedge V \to \bigwedge V : x \mapsto vx$. Again $l_v^2 = 0$, and for any $v \in V$ and $f \in V^*$,

$$(l_v + df)^2 = l_v df + df l_v = l_{(df)(v)} = f(v)\,\mathrm{id} = q(v + f)\,\mathrm{id}. \tag{2.4}$$

The linear map $V \oplus V^* \to \mathrm{End}_k(\bigwedge V)$ defined by $v + f \mapsto l_v + df$ then induces an isomorphism

$$\Lambda : \mathfrak{Cl}(V \oplus V^*, q) \to \mathrm{End}_k(\bigwedge V). \tag{2.5}$$

Let $\bar{\ } : \bigwedge V \to \bigwedge V$ be the involution such that $\bar{v} = -v$ for any $v \in V$. Fix a basis $\{v_1, \ldots, v_l\}$ of V, and let $\{f_1, \ldots, f_l\}$ be its dual basis ($f_i(v_j) = \delta_{ij}$ for any $i, j = 1, \ldots, l$). Let $\Phi : \bigwedge V \to k$ be the linear function such that

$$\begin{aligned} &\Phi(v_1 \cdots v_l) = 1, \\ &\Phi(v_{i_1} \cdots v_{i_r}) = 0 \quad \text{for any } r < l \text{ and } 1 \le i_1 < \cdots < i_r \le l, \end{aligned} \tag{2.6}$$

that is, Φ is a determinant, and consider the bilinear form

$$\begin{aligned} b : \textstyle\bigwedge V \times \bigwedge V &\to k \\ (s, t) &\mapsto \Phi(\bar{s}t). \end{aligned} \tag{2.7}$$

Since

$$\begin{aligned} \Phi(\overline{v_1 \cdots v_l}) &= (-1)^l \Phi(v_l \cdots v_1) = (-1)^l (-1)^{\binom{l}{2}} \Phi(v_1 \cdots v_l) \\ &= (-1)^{\binom{l+1}{2}} \Phi(v_1 \cdots v_l) = (-1)^{\binom{l+1}{2}}, \end{aligned}$$

it follows that, for any $s, t \in \bigwedge V$,

$$b(t, s) = \Phi(\bar{t}s) = \Phi(\overline{\bar{s}t}) = (-1)^{\binom{l+1}{2}} \Phi(\bar{s}t) = (-1)^{\binom{l+1}{2}} b(s, t).$$

Hence,

$$\begin{aligned} &b \text{ is symmetric if and only if } l \equiv 0 \text{ or } 3 \pmod 4, \\ &b \text{ is skew-symmetric if and only if } l \equiv 1 \text{ or } 2 \pmod 4. \end{aligned} \tag{2.8}$$

Let τ_b be the adjoint involution of $\bigwedge V$ relative to b. Then, for any $v \in V$ and $s, t \in \bigwedge V$,

$$b\big(l_v(s), t\big) = \Phi(\overline{vs}t) = \Phi(\bar{s}\bar{v}t) = -\Phi(\bar{s}vt) = -b\big(s, l_v t\big),$$

so $\tau_b(l_v) = -l_v$. Also, if $f \in V^*$ and $v \in V$,

$$(df)(\bar{v}) = -(df)(v) = -f(v) = -\overline{f(v)} = (-1)^{|v|}\overline{(df)(v)},$$

where $\bigwedge V = \oplus_{i=0}^{l} \bigwedge^i V$ is the natural $\mathbb{Z}$-grading of $\bigwedge V$ and $|s| = i$ for $s \in \bigwedge^i V$. Also, assuming $(df)(\bar{s}) = (-1)^{|s|}\overline{(df)(s)}$ and $(df)(\bar{t}) = (-1)^{|t|}\overline{(df)(t)}$ for homogeneous $s, t \in \bigwedge V$,

$$\begin{aligned} (df)(\overline{st}) &= (df)(\bar{t}\bar{s}) = (df)(\bar{t})\bar{s} + (-1)^{|t|}\bar{t}(df)(\bar{s}) \\ &= (-1)^{|t|}\overline{(df)(t)}\bar{s} + (-1)^{|s|+|t|}\bar{t}\overline{(df)(s)} \\ &= (-1)^{|s|+|t|}\overline{(df)(s)t + (-1)^{|s|}s(df)(t)} \\ &= (-1)^{|st|}\overline{(df)(st)}. \end{aligned}$$

Hence $(df)(\bar{s}) = (-1)^{|s|}\overline{(df)(s)}$ for any homogeneous $s \in \bigwedge V$. Thus, for any $f \in V^*$ and $s, t \in \bigwedge V$,

$$\begin{aligned} b\big((df)(s), t\big) &= \Phi\big(\overline{(df)(s)}t\big) = (-1)^{|s|}\Phi\big((df)(\bar{s})t\big) \\ &= -\Phi\big(\bar{s}(df)(t)\big) \qquad \text{since } \Phi\big((df)(\textstyle\bigwedge V)\big) = 0 \\ &= -b\big(s, (df)(t)\big) \end{aligned}$$

and, therefore, $\tau_b(df) = -df$. As a consequence, the isomorphism Λ in (2.5) is actually an isomorphism of algebras with involution:

$$\Lambda : \big(\mathfrak{Cl}(V \oplus V^*, q), \tau'\big) \to \big(\mathrm{End}_k(\textstyle\bigwedge V), \tau_b\big). \tag{2.9}$$

The orthogonal Lie algebra $\mathfrak{so}_{2l+1} = \mathfrak{so}(W, q)$ is spanned by the linear maps

$$\sigma_{w_1,w_2} = q(w_1, \cdot)w_2 - q(w_2, \cdot)w_1 \tag{2.10}$$

where $q(w_1, w_2) = q(w_1+w_2) - q(w_1) - q(w_2)$ is the associated symmetric bilinear form.

But for any $w_1, w_2, w_3 \in W$, inside $\mathfrak{Cl}(W, q)$ one has

$$\begin{aligned}
[[w_1, w_2]^{\cdot}, w_3]^{\cdot} &= (w_1 \cdot w_2 - w_2 \cdot w_1) \cdot w_3 - w_3 \cdot (w_1 \cdot w_2 - w_2 \cdot w_1) \\
&= q(w_2, w_3)w_1 - w_1 \cdot w_3 \cdot w_2 - q(w_1, w_3)w_2 + w_2 \cdot w_3 \cdot w_1 \\
&\quad - q(w_1, w_3)w_2 + w_1 \cdot w_3 \cdot w_2 + q(w_2, w_3)w_1 - w_2 \cdot w_3 \cdot w_1 \\
&= -2\sigma_{w_1,w_2}(w_3).
\end{aligned}$$

Therefore, $\mathfrak{so}_{2l+1}$ embeds in $\mathfrak{Cl}_{\bar{0}}(W, q)$ by means of $\sigma_{w_1,w_2} \mapsto -\frac{1}{2}[w_1, w_2]^{\cdot}$, so $\mathfrak{so}_{2l+1}$ can be identified with the subspace $[W, W]^{\cdot}$ in $\mathfrak{Cl}_{\bar{0}}(W, q)$.

Under this identification, the action of $\mathfrak{so}_{2l+1} = \mathfrak{so}(W, q)$ on its natural module W corresponds to the adjoint action of $[W, W]^{\cdot}$ on W inside $\mathfrak{Cl}(W, q)$. Note that for any $x, y \in V \oplus V^*$,

$$\begin{aligned}
\Psi\big([x, y]^{\cdot}\big) = [u \cdot x, u \cdot y]^{\cdot} &= u \cdot x \cdot u \cdot y - u \cdot y \cdot u \cdot x \\
&= -u \cdot u \cdot (x \cdot y - y \cdot x) \\
&= x \cdot y - y \cdot x = [x, y]^{\cdot},
\end{aligned}$$

so Ψ acts "identically" on $\mathfrak{so}_{2l} = [V \oplus V^*, V \oplus V^*]^{\cdot} \subseteq \mathfrak{Cl}(V \oplus V^*, q)$.

The subspace $\mathfrak{h} = \mathrm{span}\,\{[v_i, f_i]^{\cdot} : i = 1, \ldots, l\}$ is a Cartan subalgebra of $\mathfrak{so}_{2l+1} \simeq [W, W]^{\cdot}$. Besides,

$$\begin{aligned}
[[v_i, f_i]^{\cdot}, u]^{\cdot} &= 0, \\
[[v_i, f_i]^{\cdot}, v_j]^{\cdot} &= 2\delta_{ij}v_j, \\
[[v_i, f_i]^{\cdot}, f_j]^{\cdot} &= -2\delta_{ij}f_j.
\end{aligned}$$

Hence, if $\epsilon_i : \mathfrak{h} \to k$ denotes the linear map with $\epsilon_i\big([v_j, f_j]^{\cdot}\big) = 2\delta_{ij}$, the weights of the natural module W relative to $\mathfrak{h}$ are 0 and $\pm\epsilon_i$, $i = 1, \ldots, l$, all of them of multiplicity 1; while there appears a root space decomposition

$$\mathfrak{so}_{2l+1} \simeq [W, W]^{\cdot} = \mathfrak{h} \oplus \big(\oplus_{\alpha\in\Delta}\mathfrak{g}_\alpha\big),$$

where

$$\Delta = \{\pm(\epsilon_i + \epsilon_j) : 1 \le i < j \le l\} \cup \{\pm\epsilon_i : 1 \le i \le l\} \cup \{\pm(\epsilon_i - \epsilon_j) : 1 \le i < j \le l\}.$$

Here $\mathfrak{g}_{\epsilon_i+\epsilon_j} = k[v_i, v_j]^{\cdot}$, $\mathfrak{g}_{-(\epsilon_i+\epsilon_j)} = k[f_i, f_j]^{\cdot}$, $\mathfrak{g}_{\epsilon_i-\epsilon_j} = k[v_i, f_j]^{\cdot}$, $\mathfrak{g}_{\epsilon_i} = k[u, v_i]^{\cdot}$, and $\mathfrak{g}_{-\epsilon_i} = k[u, f_i]^{\cdot}$, for any $i \neq j$. This root space decomposition induces a triangular decomposition

$$\mathfrak{so}_{2l+1} = \mathfrak{g}_- \oplus \mathfrak{h} \oplus \mathfrak{g}_+,$$

where $\mathfrak{g}_\pm = \oplus_{\alpha\in\Delta^\pm}\mathfrak{g}_\alpha$, with $\Delta^+ = \{\epsilon_i+\epsilon_j : 1 \le i < j \le l\} \cup \{\epsilon_i : 1 \le i \le l\} \cup \{\epsilon_i - \epsilon_j : 1 \le i < j \le l\}$, and $\Delta^- = -\Delta^+$.

The spin representation of $\mathfrak{so}_{2l+1}$ is given by the composition

$$\mathfrak{so}_{2l+1} \hookrightarrow \mathfrak{Cl}_{\bar{0}}(W, q) \xrightarrow{\Psi^{-1}} \mathfrak{Cl}(V \oplus V^*, q) \xrightarrow{\Lambda} \mathrm{End}_k(\textstyle\bigwedge V).$$

Denote this composition by

$$\rho = \Lambda \circ \Psi^{-1}|_{\mathfrak{so}_{2l+1}}, \tag{2.11}$$

and denote by $S = \bigwedge V$ the spin module. Note that for any $1 \le i \le l$ and any $1 \le i_1 < \cdots < i_r \le l$

$$\rho\big([v_i, f_i]^{\cdot}\big)(v_{i_1}\cdots v_{ir}) = \begin{cases} v_{i_1}\cdots v_{i_r} & \text{if } i = i_j \text{ for some } j, \\ -v_{i_1}\cdots v_{i_r} & \text{otherwise.} \end{cases}$$

Thus, $v_{i_1}\cdots v_{i_r}$ is a weight vector relative to $\mathfrak{h}$, with weight

$$\frac{1}{2}\Bigg(\sum_{i\in\{i_1,\ldots,i_r\}} \epsilon_i - \sum_{j\notin\{i_1,\ldots,i_r\}} \epsilon_j\Bigg),$$

and hence all the weights of the spin module have multiplicity 1.

Proposition 2.12. *Up to scalars, there is a unique $\mathfrak{so}_{2l+1}$-invariant bilinear map $S \times S \to \mathfrak{so}_{2l+1}$, $(s, t) \mapsto [s, t]$. This map is given by the formula*

$$\tfrac{1}{2}\,\mathrm{tr}\big(\sigma[s, t]\big) = b\big(\rho(\sigma)(s), t\big), \tag{2.13}$$

for any $\sigma \in \mathfrak{so}_{2l+1}$ and $s, t \in S$, where tr *denotes the trace of the natural representation of $\mathfrak{so}_{2l+1}$.*

Moreover, this bilinear map $[\,\cdot\,,\cdot\,]$ is symmetric if and only if l is congruent to 1 *or* 2 *modulo* 4*. Otherwise, it is skew-symmetric.*

Proof. First note that the trace form tr is $\mathfrak{so}_{2l+1}$-invariant, and so is b because Ψ and Λ in (2.3) and (2.9) are isomorphisms of algebras with involutions. Since both tr and b are nondegenerate, $[\,\cdot\,,\cdot\,]$ is well defined and $\mathfrak{so}_{2l+1}$-invariant. Now, the space of $\mathfrak{so}_{2l+1}$-invariant bilinear maps $S \times S \to \mathfrak{so}_{2l+1}$ is isomorphic to $\mathrm{Hom}_{\mathfrak{so}_{2l+1}}(S \otimes S, \mathfrak{so}_{2l+1})$ (all the tensor products are considered over the ground field k), or to the space of those tensors in $S\otimes S\otimes(\mathfrak{so}_{2l+1})^* \simeq S\otimes S\otimes\mathfrak{so}_{2l+1} \simeq \mathfrak{so}_{2l+1} \otimes S\otimes S^*$ (tr and b are nondegenerate) annihilated by $\mathfrak{so}_{2l+1}$, and hence to $\mathrm{Hom}_{\mathfrak{so}_{2l+1}}(\mathfrak{so}_{2l+1} \otimes S, S)$.

But $\mathfrak{so}_{2l+1} \otimes S$ is generated, as a module for $\mathfrak{so}_{2l+1}$, by the tensor product of any nonzero element (like $[v_1, v_2]^{\cdot}$) in the root space $(\mathfrak{so}_{2l+1})_{\epsilon_1+\epsilon_2}$ ($(\mathfrak{so}_3)_{\epsilon_1}$ if $l=1$), and any nonzero element (like 1) in the weight space $S_{-\frac{1}{2}(\epsilon_1+\cdots+\epsilon_l)}$. (Note that $\epsilon_1+\epsilon_2$ is the longest root in the lexicographic order given by $\epsilon_1 > \cdots > \epsilon_l > 0$, while $-\frac{1}{2}(\epsilon_1+\cdots+\epsilon_l)$ is the lowest weight in S.) The image of this basic tensor under any homomorphism of $\mathfrak{so}_{2l+1}$-modules lies in the weight space of weight $\frac{1}{2}(\epsilon_1+\epsilon_2-\epsilon_3-\cdots-\epsilon_l)$, which is one-dimensional. Hence, $\dim_k \operatorname{Hom}_{\mathfrak{so}_{2l+1}}(\mathfrak{so}_{2l+1} \otimes S, S) = 1$, as required.

The last part of the Proposition follows from (2.8). □

For future use, note that for any $w_1, w_2, w_3, w_4 \in W$,

$$\operatorname{tr}\big(\sigma_{w_1,w_2}\sigma_{w_3,w_4}\big) = 2\big(q(w_1,w_4)q(w_2,w_3) - q(w_1,w_3)q(w_2,w_4)\big),$$

and hence, under the identification $\mathfrak{so}_{2l+1} \simeq [W,W]^{\cdot}$ $(\sigma_{w_1,w_2} \mapsto -\frac{1}{2}[w_1,w_2]^{\cdot})$,

$$\frac{1}{2}\operatorname{tr}\big([w_1,w_2]^{\cdot}[w_3,w_4]^{\cdot}\big) = 4\big(q(w_1,w_4)q(w_2,w_3) - q(w_1,w_3)q(w_2,w_4)\big). \tag{2.14}$$

To deal with the Lie algebras $\mathfrak{so}_{2l}$ (type D), $l \geq 2$, consider the involution of $\mathfrak{Cl}(V \oplus V^*, q)$, which will be denoted by τ too, which is the identity on $V \oplus V^*$. Also consider the involution $\hat{}: \bigwedge V \to \bigwedge V$ such that $\hat{v} = v$ for any $v \in V \subseteq \bigwedge V$, and the nondegenerate bilinear form

$$\begin{aligned} \hat{b} : \textstyle\bigwedge V \times \bigwedge V &\to k \\ (s,t) &\mapsto \Phi(\hat{s}t), \end{aligned} \tag{2.15}$$

where Φ is as in (2.6). Here, with the same arguments as for (2.8),

$$\begin{aligned} &\hat{b} \text{ is symmetric if and only if } l \equiv 0 \text{ or } 1 \pmod 4, \\ &\hat{b} \text{ is skew-symmetric if and only if } l \equiv 2 \text{ or } 3 \pmod 4. \end{aligned} \tag{2.16}$$

Moreover, if l is even, then $\hat{b}\big(\bigwedge_{\bar{0}} V, \bigwedge_{\bar{1}} V\big) = 0$, so the restrictions of $\hat{b}$ to $S^+ = \bigwedge_{\bar{0}} V$ and $S^- = \bigwedge_{\bar{1}} V$ are nondegenerate. However, if l is odd, then both S^+ and S^- are isotropic subspaces relative to $\hat{b}$.

The nondegenerate bilinear form $\hat{b}$ induces the adjoint involution $\tau_{\hat{b}}$ on $\bigwedge V$ and, as before, the isomorphism Λ in (2.5) becomes an isomorphism of algebras with involution:

$$\Lambda : \big(\mathfrak{Cl}(V \oplus V^*, q), \tau\big) \to \big(\operatorname{End}_k(\textstyle\bigwedge V), \tau_{\hat{b}}\big). \tag{2.17}$$

Under this isomorphism, the even Clifford algebra $\mathfrak{Cl}_{\bar{0}}(V \oplus V^*, q)$ maps onto $\operatorname{End}_k(\bigwedge_{\bar{0}} V) \oplus \operatorname{End}_k(\bigwedge_{\bar{1}} V)$.

Also, as before, $\mathfrak{so}_{2l} = \mathfrak{so}(V \oplus V^*, q)$ can be identified with the subspace $[V \oplus V^*, V \oplus V^*]^{\cdot}$ of $\mathfrak{Cl}_{\bar{0}}(V \oplus V^*, q)$, $\mathfrak{h} = \operatorname{span}\{[v_i, f_i]^{\cdot} : i = 1, \ldots, l\}$ is a Cartan

subalgebra, the roots are $\{\pm\epsilon_i \pm \epsilon_j : 1 \le i < j \le l\}$, the set of weights of the natural module $V \oplus V^*$ are $\{\pm\epsilon_i : 1 \le i \le l\}$, all the weights appear with multiplicity one, and the composition

$$\mathfrak{so}_{2l} \hookrightarrow \mathfrak{Cl}_{\bar{0}}(V \oplus V^*, q) \xrightarrow{\Lambda} \mathrm{End}_k(\textstyle\bigwedge_{\bar{0}} V) \oplus \mathrm{End}_k(\bigwedge_{\bar{1}} V)$$

gives two representations

$$\rho^+ : \mathfrak{so}_{2l} \to \mathrm{End}_k(\textstyle\bigwedge_{\bar{0}} V) \quad \text{and} \quad \rho^- : \mathfrak{so}_{2l} \to \mathrm{End}_k(\bigwedge_{\bar{1}} V), \tag{2.18}$$

called the half-spin representations. The weights in $S^+ = \bigwedge_{\bar{0}} V$ (respectively $S^- = \bigwedge_{\bar{1}} V$) are the weights $\frac{1}{2}(\pm\epsilon_1 \pm \cdots \pm \epsilon_l)$, with an even (respectively odd) number of $+$ signs.

Proposition 2.19. (i) *If l is odd, $l \ge 3$, there is no nonzero $\mathfrak{so}_{2l}$-invariant bilinear map $S^+ \times S^+ \to \mathfrak{so}_{2l}$.*

(ii) *If l is even, there is a unique, up to scalars, such bilinear map, which is given by the formula*

$$\tfrac{1}{2}\,\mathrm{tr}\big(\sigma[s,t]\big) = \hat{b}\big(\rho^+(\sigma)(s), t\big), \tag{2.20}$$

for any $\sigma \in \mathfrak{so}_{2l}$ and $s, t \in S^+$. Moreover, this bilinear map $[\,\cdot\,,\cdot\,]$ is symmetric if and only if l is congruent to 2 or 3 modulo 4, and it is skew-symmetric otherwise.

Proof. If l is odd ($l \ge 3$), then $S^+ \otimes S^+$ is generated, as a module for $\mathfrak{so}_{2l}$, by $v_1 \cdots v_{l-1} \otimes v_{l-1} v_l$ (the tensor product or a nonzero highest weight vector and a nonzero lowest weight vector), and its image under any nonzero $\mathfrak{so}_{2l}$-invariant linear map $S^+ \otimes S^+ \to \mathfrak{so}_{2l}$ lies in the root space of root $\frac{1}{2}(\epsilon_1 + \cdots + \epsilon_{l-1} - \epsilon_l) + \frac{1}{2}(-\epsilon_1 - \cdots - \epsilon_{l-2} + \epsilon_{l-1} + \epsilon_l) = \epsilon_{l-1}$. But ϵ_{l-1} is not a root, so its image must vanish.

For l even $\hat{b}$ is nondegenerate on S^+ and, as in Proposition 2.12, it is enough to compute $\dim_k \mathrm{Hom}_{\mathfrak{so}_{2l}}(\mathfrak{so}_{2l} \otimes S^+, S^+)$, which is proven to be 1 with the same arguments given there. □

Remark 2.21. In $\mathfrak{Cl}_{\bar{1}}(V \oplus V^*, q)$ there are invertible elements a such that $a^{\cdot 2} \in k1$ and $a \cdot (V \oplus V^*) \cdot a^{-1} \subseteq V \oplus V^*$. For instance, one can take the element $a = [v_1, f_1]^{\cdot} \cdots [v_{l-1}, f_{l-1}]^{\cdot} \cdot (v_l + f_l)$, which satisfies $a^{\cdot 2} = (-1)^{l-1}$. (Note that $[v_i, f_i]^{\cdot 2} = v_i \cdot f_i \cdot v_i \cdot f_i + f_i \cdot v_i \cdot f_i \cdot v_i = v_i \cdot (1 - v_i \cdot f_i) \cdot f_i + f_i \cdot (1 - f_i \cdot v_i) \cdot v_i = v_i \cdot f_i + f_i \cdot v_i = f_i(v_i) = 1$ and $(v_i + f_i) \cdot (v_i + f_i) = v_i \cdot f_i + f_i \cdot v_i = 1$.) Consider the linear isomorphism $\phi_a : S^- = \bigwedge_{\bar{1}} V \to S^+ = \bigwedge_{\bar{0}} V$, $s \mapsto \rho(a)(s)$, and the order two automorphism $\mathrm{Ad}_a : \mathfrak{Cl}(V \oplus V^*, q) \to \mathfrak{Cl}(V \oplus V^*, q)$, $x \mapsto a \cdot x \cdot a^{-1}$. Ad_a preserves $V \oplus V^*$, and hence also $\mathfrak{so}_{2l} \simeq [V \oplus V^*, V \oplus V^*]^{\cdot}$. Then, for any $\mathfrak{so}_{2l}$-invariant bilinear map $[\,\cdot\,,\cdot\,] : S^+ \times S^+ \to \mathfrak{so}_{2l}$, one gets a $\mathfrak{so}_{2l}$-invariant bilinear

map $[\cdot,\cdot]^- : S^- \times S^- \to \mathfrak{so}_{2l}$ at once by

$$[s,t]^- = \mathrm{Ad}_a\big([\phi_a(s),\phi_a(t)]\big).$$

Therefore, it is enough to deal with the half spin representation S^+.

3. Type *B*

Let $\mathfrak{g} = \mathfrak{g}_{\bar{0}} \oplus \mathfrak{g}_{\bar{1}}$ be either a simple $\mathbb{Z}_2$-graded Lie algebra or a simple Lie superalgebra with $\mathfrak{g}_{\bar{0}} = \mathfrak{so}_{2l+1}$ and $\mathfrak{g}_{\bar{1}} = S$ (its spin module).

Because of Proposition 2.12, the product of two odd elements can be assumed to be given by the bilinear map $[s,t]$ in (2.13). Therefore, the possibilities for such a $\mathfrak{g}$ are given precisely by the values of l such that the product $[s,t]$ satisfies the Jacobi identity

$$J(s_1,s_2,s_3) = \rho([s_1,s_2])(s_3) + \rho([s_2,s_3])(s_1) + \rho([s_3,s_1])(s_2) = 0,$$

for any $s_1, s_2, s_3 \in S$. As in Section 2, ρ denotes the spin representation of $\mathfrak{so}_{2l+1}$.

But $S \otimes S \otimes S$ is generated, as a module for $\mathfrak{so}_{2l+1}$, by the elements $1 \otimes v_1 \cdots v_l \otimes v_{i_1} \cdots v_{i_r}$, where $0 \le r \le l$ and $1 \le i_1 < \cdots < i_r \le l$. As in Section 2, $\{v_1,\dots,v_l\}$ denotes a fixed basis of V and $\{f_1,\dots,f_l\}$ the corresponding dual basis in V^*. The trilinear map $S \times S \times S \to S$, $(s_1,s_2,s_3) \mapsto J(s_1,s_2,s_3)$ is $\mathfrak{so}_{2l+1}$-invariant, so it is enough to check for which values of l the Jacobian

$$J(1, v_1 \cdots v_l, v_{i_1} \cdots v_{i_r})$$

is 0 for any $0 \le r \le l$ and $1 \le i_1 < \cdots < i_r \le l$. By symmetry, it is enough to check the Jacobians

$$J(1, , v_1 \cdots v_l, v_1 \cdots v_r)$$

for $0 \le r \le l$.

Theorem 3.1. *Let $l \in \mathbb{N}$ and let $\mathfrak{g} = \mathfrak{g}_{\bar{0}} \oplus \mathfrak{g}_{\bar{1}}$ be the $\mathbb{Z}_2$-graded algebra with $\mathfrak{g}_{\bar{0}} = \mathfrak{so}_{2l+1}$, $\mathfrak{g}_{\bar{1}} = S$ (its spin module), and multiplication given by the Lie bracket of elements in $\mathfrak{so}_{2l+1}$, and by*

$$[\sigma,s] = -[s,\sigma] = \rho(\sigma)(s), \quad \rho \text{ as in (2.11)},$$
$$[s,t] \quad \text{given by (2.13)}.$$

for any $\sigma \in \mathfrak{g}_{\bar{0}}$ and $s,t \in \mathfrak{g}_{\bar{1}}$. Then:

(i) *$\mathfrak{g}$ is a Lie algebra if and only if either*

- *$l = 3$ and the characteristic of k is 3, and then $\mathfrak{g}$ is isomorphic to the 29-dimensional simple Lie algebra discovered by Brown [1982], or*
- *$l = 4$, and then $\mathfrak{g}$ is isomorphic to the simple Lie algebra of type F_4.*

(ii) *$\mathfrak{g}$ is a Lie superalgebra if and only if either*

- $l = 1$, *and then* $\mathfrak{g}$ *is isomorphic to the orthosymplectic Lie superalgebra* $\mathfrak{osp}(1, 2)$, *or*
- $l = 2$, *and then* $\mathfrak{g}$ *is isomorphic to the orthosymplectic Lie superalgebra* $\mathfrak{osp}(1, 4)$, *or*
- $l = 5$ *and the characteristic of* k *is* 5, *or*
- $l = 6$ *and the characteristic of* k *is* 3.

Proof. With the same notations as in Section 2, note that $v_1 \cdots v_r$ is a weight vector relative to $\mathfrak{h}$, of weight $\frac{1}{2}(\epsilon_1 + \cdots + \epsilon_r - \epsilon_{r+1} - \cdots - \epsilon_l)$ for any $0 \leq r \leq l$. Hence $[1, v_1 \cdots v_r] \in (\mathfrak{so}_{2l+1})_{-(\epsilon_{r+1}+\cdots+\epsilon_l)}$. In the same vein, $[v_1 \cdots v_l, v_1 \cdots v_r] \in (\mathfrak{so}_{2l+1})_{\epsilon_1+\cdots\epsilon_r}$. In particular,

$$[1, v_1 \cdots v_r] = 0 \text{ if } 0 \leq r \leq l-3; \quad [v_1 \cdots v_l, v_1 \cdots v_r] = 0 \text{ if } 3 \leq r \leq l. \tag{3.2}$$

Also, $[1, v_1 \cdots v_l] \in \mathfrak{h}$, so that $[1, v_1 \cdots v_l] = \sum_{i=1}^{l} \alpha_i [v_i, f_i]^{\cdot}$ for some $\alpha_1, \ldots, \alpha_l \in k$. By (2.13)

$$\frac{1}{2} \sum_{i=1}^{l} \alpha_i \operatorname{tr}\big(\sigma [v_i, f_i]^{\cdot}\big) = b\big(\rho(\sigma)(1), v_1 \cdots v_l\big). \tag{3.3}$$

Let $\sigma = [v_j, f_j]^{\cdot}$, then by (2.14)

$$\frac{1}{2} \sum_{i=1}^{l} \alpha_i \operatorname{tr}\big([v_j, f_j]^{\cdot} [v_i, f_i]^{\cdot}\big) = 4 f_i(v_j) f_j(v_i) = 4\delta_{ij},$$

while $\rho\big([v_j, f_j]^{\cdot}\big)(1) = [l_{v_j}, df_j](1) = -(df_j)(v_j) = -1$. Thus, (3.3) gives $4\alpha_j = -1$ for any $j = 1, \ldots, l$, so

$$[1, v_1 \cdots v_l] = -\frac{1}{4} \sum_{i=1}^{l} [v_i, f_i]^{\cdot}. \tag{3.4}$$

In the same vein, $[1, v_1 \cdots v_{l-1}] \in (\mathfrak{so}_{2l+1})_{-\epsilon_l}$, so $[1, v_1 \cdots v_{l-1}] = \alpha [u, f_l]^{\cdot}$ for some $\alpha \in k$, and by (2.13)

$$\tfrac{1}{2}\alpha \operatorname{tr}\big([u, v_l]^{\cdot} [u, f_l]^{\cdot}\big) = b\big(\rho([u, v_l]^{\cdot}(1), v_1 \cdots v_{l-1}\big). \tag{3.5}$$

The left-hand side is $4\alpha\big(-q(u, u)q(v_l, f_l)\big) = 8\alpha$, while on the right-hand side $\rho\big([u, v_l]^{\cdot}\big) = 2\rho\big(\Psi(v_l)\big) = 2\Lambda(v_l)$, so this side becomes

$$\begin{aligned} 2b(v_l, v_1 \cdots v_{l-1}) &= 2\Phi(\bar{v}_l v_1 \cdots v_{l-1}) = -2\Phi(v_l v_1 \cdots v_{l-1}) \\ &= 2(-1)^l \Phi(v_1 \cdots v_l) = 2(-1)^l. \end{aligned}$$

Therefore $\alpha = \frac{1}{4}(-1)^l$ and

$$[1, v_1 \cdots v_{l-1}] = \tfrac{1}{4}(-1)^l [u, f_l]^{\cdot}. \tag{3.6}$$

Similar arguments, which are left to the reader, give

$$[1, v_1 \cdots v_{l-2}] = \tfrac{1}{2}[f_{l-1}, f_l]^{\cdot}, \text{ if } l \geq 2, \tag{3.7}$$

$$[v_1 \cdots v_l, v_1] = -\tfrac{1}{4}(-1)^{\binom{l+1}{2}}[u, v_1]^{\cdot}, \tag{3.8}$$

$$[v_1 \cdots v_l, v_1 v_2] = -\tfrac{1}{2}(-1)^{\binom{l+1}{2}}[v_1, v_2]^{\cdot}. \tag{3.9}$$

Now, if $l \geq 7$ and $3 \leq r \leq l-3$,

$$\begin{aligned} J(1, v_1 \cdots v_l, v_1 \cdots v_r) &= [[1, v_1 \cdots v_l], v_1 \cdots v_r] && \text{(by (3.2))} \\ &= -\tfrac{1}{4}\sum_{i=1}^{l}[[v_i, f_i]^{\cdot}, v_1 \cdots v_r] && \text{(by (3.4))} \\ &= -\tfrac{1}{4}\sum_{i=1}^{l}\rho\big([v_i, f_i]^{\cdot}\big)(v_1 \cdots v_r) \\ &= -\tfrac{1}{4}(r-(l-r))v_1 \cdots v_r = \tfrac{1}{4}(l-2r)v_1 \cdots v_r. \end{aligned}$$

With $r = \frac{1}{2}(l-2)$ if l is even or $\frac{1}{2}(l-1)$ if l is odd, $l-2r$ ($= 1$ or 2) $\neq 0$, so the Jacobi identity is not satisfied.

Assume now that $l = 6$, so $[s, t]$ is symmetric in $s, t \in S$ by Proposition 2.12. Then

$$\begin{aligned} J(1, v_1 \cdots v_6, 1) &= 2[[1, v_1 \cdots v_6], 1] \quad \text{(because } [1, 1] = 0\text{; see (3.2))} \\ &= -\tfrac{1}{2}\sum_{i=1}^{6}[[v_i, f_i]^{\cdot}, 1] && \text{(by (3.4))} \\ &= -\tfrac{1}{2}\sum_{i=1}^{6}(-1) = 3, \end{aligned}$$

so the characteristic of k must be 3. Assuming this is so, it is easily checked that $J(1, v_1 \cdots v_6, v_1 \cdots v_r) = 0$ for any $0 \leq r \leq 6$.

For $l = 5$, the product $[s, t]$ is also symmetric (Proposition 2.12) and, as before,

$$J(1, v_1 \cdots v_5, 1) = 2[[1, v_1 \cdots v_5], 1] = -\tfrac{1}{2}\sum_{i=1}^{5}(-1) = \tfrac{5}{2},$$

so the characteristic of k must be 5, and then $J(1, v_1 \cdots v_5, v_1 \cdots v_r) = 0$ for any $0 \leq r \leq 5$.

For $l = 4$, the product $[s, t]$ is skew-symmetric, by Proposition 2.12. Therefore $J(1, v_1 \cdots v_4, 1) = J(1, v_1 \cdots v_4, v_1 \cdots v_4) = 0$ by skew-symmetry. The other instances of $J(1, v_1 \cdots v_4, v_1 \cdots v_r)$, $1 \leq r \leq 3$, are also checked to be trivial.

With $l = 3$, the product $[s, t]$ is skew-symmetric too. Hence, by (3.4), (3.6), and (3.7),

$$\begin{aligned} J(1, v_1v_2v_3, v_1) &= [[1, v_1v_2v_3], v_1] + [[v_1v_2v_3, v_1], 1] + [[v_1, 1], v_1v_2v_3] \\ &= -\tfrac{1}{4}\sum_{i=1}^{3}[[v_i, f_i]^{\cdot}, v_1] - \tfrac{1}{4}[[u, v_1]^{\cdot}, 1] - \tfrac{1}{2}[[f_2, f_3]^{\cdot}, v_1v_2v_3] \\ &= -\tfrac{1}{4}(1-1-1)v_1 - \tfrac{1}{4}(2v_1) - \tfrac{1}{2}(-1-1)v_1 = \tfrac{3}{4}v_1, \end{aligned}$$

and hence the characteristic must be 3. The other instance of the Jacobi identity to be checked: $J(1, v_1v_2v_3, v_1v_2) = 0$, also holds easily.

For $l = 1$ or $l = 2$, the Jacobi identity is satisfied too.

The assertions about which Lie algebras or superalgebras appear follows at once, since all the algebras and superalgebras mentioned in the statement of the Theorem satisfy the hypotheses. (For $\mathfrak{osp}(1, 4)$, the even part is isomorphic to the symplectic Lie algebra $\mathfrak{sp}_4$, and the odd part is its natural 4-dimensional module. However $\mathfrak{sp}_4$ is isomorphic to $\mathfrak{so}_5$, and viewed like this, the 4-dimensional module is the spin module. The same happens for $\mathfrak{osp}(1, 2)$.) □

Remark 3.10. Up to our knowledge, the modular Lie superalgebras that occur for $l = 5$ and $l = 6$ have not appeared previously in the literature. Note that the simplicity of $\mathfrak{so}_{2l+1}$ and the irreducibility of its spin module imply that these superalgebras are simple.

4. Type *D*

In this section the situation in which $\mathfrak{g}_{\bar{0}} = \mathfrak{so}_{2l}$ ($l \geq 2$), and $\mathfrak{g}_{\bar{1}} = S^+$ (half-spin module) will be considered. First note that it does not matter which half-spin representation is used (Remark 2.21). By Proposition 2.19, it is enough to deal with even values of l.

Theorem 4.1. *Let l be an even positive integer, and let $\mathfrak{g} = \mathfrak{g}_{\bar{0}} \oplus \mathfrak{g}_{\bar{1}}$ be the $\mathbb{Z}_2$-graded algebra with $\mathfrak{g}_{\bar{0}} = \mathfrak{so}_{2l}$, $\mathfrak{g}_{\bar{1}} = S^+$, and multiplication given by the Lie bracket of elements in $\mathfrak{so}_{2l}$, and by*

$$\begin{aligned} &[\sigma, s] = -[s, \sigma] = \rho^+(\sigma)(s), \quad \rho^+ \text{ as in (2.18)}, \\ &[s, t] \quad \text{given by (2.20)}. \end{aligned}$$

for any $\sigma \in \mathfrak{g}_{\bar{0}}$ and $s, t \in \mathfrak{g}_{\bar{1}}$. Then:

(i) $\mathfrak{g}$ *is a Lie algebra if and only if either*

- $l = 8$ *and then $\mathfrak{g}$ is isomorphic to the simple Lie algebra of type E_8, or*
- $l = 4$, *and then $\mathfrak{g}$ is isomorphic to the simple Lie algebra $\mathfrak{so}_9$ (of type B_4).*

(ii) $\mathfrak{g}$ *is a Lie superalgebra if and only if either*

- *$l = 6$, and the characteristic of k is 3, and then $\mathfrak{g}$ is isomorphic to the Lie superalgebra in [Elduque 2006, Theorem 3.2(v)], or*
- *$l = 2$, and then $\mathfrak{g}$ is isomorphic to the direct sum $\mathfrak{osp}(1,2) \oplus \mathfrak{sl}_2$, of the orthosymplectic Lie superalgebra $\mathfrak{osp}(1,2)$ and the three-dimensional simple Lie algebra.*

Proof. The restriction to $S^+ = \bigwedge_{\bar{0}} V$ of the bilinear form $\hat{b}$ in (2.15) coincides with the restriction of the bilinear form b in (2.7). Hence, as in the proof of Theorem 3.1, the equations (3.2), (3.4), (3.7), and (3.9) are all valid here.

If $l \geq 10$ and $4 \leq r \leq l-4$, r even,

$$\begin{aligned} J(1, v_1 \cdots v_l, v_1 \cdots v_r) &= [[1, v_1 \cdots v_l], v_1 \cdots v_r] = -\tfrac{1}{4} \sum_{i=1}^{l} [[v_i, f_i]^{\cdot}, v_1 \cdots v_r] \\ &= -\tfrac{1}{4}(r - (l-r)) v_1 \cdots v_r = \tfrac{1}{4}(l-2r) v_1 \cdots v_r, \end{aligned}$$

so, with $r = \frac{1}{2}(l-2)$ if l is congruent to 2 modulo 4, or $r = \frac{1}{4}(l-4)$ otherwise, $l - 2r$ equals 2 or 4, and the Jacobi identity is not satisfied.

For $l = 8$, $[s,t]$ is skew-symmetric and it is enough to check that the Jacobian $J(1, v_1 \cdots v_8, v_1 \cdots v_r)$ is 0 for $r = 2$, 4 or 6, which is straightforward.

For $l = 6$, $[s,t]$ is symmetric and

$$J(1, v_1 \cdots v_6, 1) = 2[[1, v_1 \cdots v_6], 1] = -\tfrac{1}{2} \sum_{i=1}^{6} [[v_i, f_i]^{\cdot}, 1] = -\tfrac{1}{2} \sum_{i=1}^{6} (-1) = 3,$$

so the characteristic of k must be 3 and then all the other instances of the Jacobi identity hold.

For $l = 4$, $[s,t]$ is skew-symmetric, and thus it is enough to deal with

$$\begin{aligned} &J(1, v_1 v_2 v_3 v_4, v_1 v_2) \\ &\quad = [[1, v_1 v_2 v_3 v_4], v_1 v_2] + [[v_1 v_2 v_3 v_4, v_1 v_2], 1] + [[v_1 v_2, 1], v_1 v_2 v_3 v_4] \\ &\quad = -\tfrac{1}{4} \sum_{i=1}^{4} [[v_i, f_i]^{\cdot}, v_1 v_2] - \tfrac{1}{2}[[v_1, v_2]^{\cdot}, 1] - \tfrac{1}{2}[[f_1, f_2]^{\cdot}, v_1 v_2 v_3 v_4] \\ &\quad = -\tfrac{1}{4}(1 + 1 - 1 - 1) v_1 v_2 - v_1 v_2 + v_1 v_2 = 0. \end{aligned}$$

It is well-known that $\mathfrak{g} = \mathfrak{so}_9$ is $\mathbb{Z}_2$-graded with $\mathfrak{g}_{\bar{0}} = \mathfrak{so}_8$ and $\mathfrak{g}_{\bar{1}}$ the natural module for $\mathfrak{so}_8$. But here, the triality automorphism permutes the natural and the two half-spin modules, so one can replace the natural module by any of its half-spin modules. Therefore, the Lie algebra that appears is isomorphic to $\mathfrak{so}_9$.

Finally, for $l = 2$, $[s,t]$ is symmetric and the Jacobi identity is easily checked to hold. Since $\mathfrak{so}_4$ is isomorphic to $\mathfrak{sl}_2 \oplus \mathfrak{sl}_2$ and the two half-spin representations are the two natural (two-dimensional) modules for each of the two copies

of $\mathfrak{sl}_2$. It follows then that $\mathfrak{g} = \mathfrak{g}_1 \oplus \mathfrak{g}_2$, where $(\mathfrak{g}_1)_{\bar{0}} \simeq \mathfrak{sl}_2$ and $(\mathfrak{g}_1)_{\bar{1}}$ the natural module for $\mathfrak{sl}_2$ (and hence $\mathfrak{g}_1 \simeq \mathfrak{osp}(1,2)$), while $\mathfrak{g}_2 = (\mathfrak{g}_2)_{\bar{0}} = \mathfrak{sl}_2$. Alternatively, the subspaces $\operatorname{span}\{[v_1, v_2]^{\cdot}, [f_1, f_2]^{\cdot}, [v_1, f_1]^{\cdot} + [v_2, f_2]^{\cdot}, 1, v_1 v_2\}$ and $\operatorname{span}\{[v_1, f_2]^{\cdot}, [v_2, f_1]^{\cdot} [v_1, f_1]^{\cdot} - [v_2, f_2]^{\cdot}\}$ are ideals of $\mathfrak{g}$, the first one being isomorphic to $\mathfrak{osp}(1,2)$, and the second one to $\mathfrak{sl}_2$. □

5. The Kac Jordan superalgebra and the Tits construction

The aim of this section is to show that the Lie superalgebra in Theorem 3.1 for $l = 5$ (and characteristic 5) is related to a well-known construction by Tits, applied to the Cayley algebra and the ten-dimensional Kac Jordan superalgebra over k.

This last superalgebra is easily described in terms of the smaller Kaplansky superalgebra. The tiny Kaplansky superalgebra is the three-dimensional Jordan superalgebra $K = K_{\bar{0}} \oplus K_{\bar{1}}$, with $K_{\bar{0}} = ke$ and $K_{\bar{1}} = U$, a two-dimensional vector space endowed with a nonzero alternating bilinear form $(\cdot\,|\,\cdot)$, and multiplication given by

$$e^2 = e, \quad ex = xe = \tfrac{1}{2}x, \quad xy = (x|y)e,$$

for any $x, y \in U$. The bilinear form $(\cdot\,|\,\cdot)$ can be extended to a supersymmetric bilinear form by means of $(e|e) = \frac{1}{2}$ and $(K_{\bar{0}}|K_{\bar{1}}) = 0$.

For any homogeneous $u, v \in K$, we know from [Benkart and Elduque 2002, (1.61)] that

$$[L_u, L_v] = L_u L_v - (-1)^{\bar{u}\bar{v}} L_v L_u = \tfrac{1}{2}\big(u(v|\cdot) - (-1)^{\bar{u}\bar{v}} v(u|\cdot)\big), \tag{5.1}$$

where L_x denotes the left multiplication by x, $\bar{x}$ being the degree of the homogeneous element x. Moreover, the Lie superalgebra of derivations of K is (see [Benkart and Elduque 2002])

$$\mathfrak{Der}\, K = [L_K, L_K] = \mathfrak{osp}(K) \ \big(\simeq \mathfrak{osp}(1,2)\big).$$

The Kac Jordan superalgebra is

$$\mathcal{J} = k1 \oplus (K \otimes K),$$

with unit element 1 and product determined by

$$(a \otimes b)(c \otimes d) = (-1)^{\bar{b}\bar{c}}\big(ac \otimes bd - \tfrac{3}{4}(a|c)(b|d)1\big), \tag{5.2}$$

for homogeneous elements $a, b, c, d \in K$ (see [Benkart and Elduque 2002, (2.1)]). By Proposition 2.7 and Theorem 2.8 of the same reference, the superspace spanned by the associators $(x, y, z) = (xy)z - x(yz) = (-1)^{\bar{y}\bar{z}}[L_x, L_z](y)$ is $(\mathcal{J}, \mathcal{J}, \mathcal{J}) = K \otimes K$, and the Lie superalgebra of derivations of $\mathcal{J}$ is $\mathfrak{Der}\,\mathcal{J} = [L_{\mathcal{J}}, L_{\mathcal{J}}]$, which acts trivially on 1 and leaves invariant $(\mathcal{J}, \mathcal{J}, \mathcal{J}) = K \otimes K$. Considered then as

subspaces of $\mathrm{End}_k(K \otimes K)$

$$\mathfrak{der}\, J = (\mathfrak{der}\, K \otimes \mathrm{id}) \oplus (\mathrm{id} \otimes \mathfrak{der}\, K) \quad \big(\simeq \mathfrak{osp}(1,2) \oplus \mathfrak{osp}(1,2)\big). \tag{5.3}$$

More precisely (see [Benkart and Elduque 2002, (2.4)]), for any homogeneous $a, b, c, d \in K$,

$$[L_{a\otimes b}, L_{c\otimes d}] = (-1)^{\bar{b}\bar{c}}\big([L_a, L_c] \otimes (b|d)\,\mathrm{id} + (a|c)\,\mathrm{id} \otimes [L_b, L_d]\big), \tag{5.4}$$

as endomorphisms of $K \otimes K$. (Note that, with the usual conventions for superalgebras, $\mathrm{id} \otimes \varphi$ acts on $a \otimes b$ as $(-1)^{\bar{\varphi}\bar{a}} a \otimes \varphi(b)$ for homogeneous φ and a, b.) In particular,

$$(\mathfrak{der}\, \mathcal{J})_{\bar{0}} = \big((\mathfrak{der}\, K)_{\bar{0}} \otimes \mathrm{id}\big) \oplus \big(\mathrm{id} \otimes (\mathfrak{der}\, K)_{\bar{0}}\big),$$

and $(\mathfrak{der}\, K)_{\bar{0}}$ is isomorphic to $\mathfrak{sp}(U) = \mathfrak{sp}_2 = \mathfrak{sl}_2$ (acting trivially on the idempotent e). The restriction of $(\mathfrak{der}\, \mathcal{J})_{\bar{0}}$ to the subspace $K_{\bar{1}} \otimes K_{\bar{1}} = U \otimes U$ of $K \otimes K$ then gives an isomorphism

$$(\mathfrak{der}\, \mathcal{J})_{\bar{0}} \cong \mathfrak{so}(U \otimes U) \quad \big(= (\mathfrak{sp}(U) \otimes \mathrm{id}) \oplus (\mathrm{id} \otimes \mathfrak{sp}(U))\big),$$

where $U \otimes U$ is endowed with the symmetric bilinear form given by

$$(u_1 \otimes u_2 | v_1 \otimes v_2) = (u_1|v_1)(u_2|v_2),$$

for any $u_1, u_2, v_1, v_2 \in U$.

Assume now that the characteristic of k is $\neq 2, 3$. Let (C, n) be a unital composition algebra over k with norm n. That is, n is a regular quadratic form satisfying $n(ab) = n(a)n(b)$ for any $a, b \in C$. (See [Schafer 1966, Chapter III] for basic facts about these algebras.)

Since the field k is assumed to be algebraically closed, C is isomorphic to either k, $k \times k$, $\mathrm{Mat}_2(k)$ or the Cayley algebra $\mathfrak{C}$ over k.

The map $D_{a,b} : C \to C$ given by

$$D_{a,b}(c) = [[a, b], c] - 3(a, b, c) \tag{5.5}$$

is the inner derivation determined by $a, b \in C$, and the Lie algebra $\mathfrak{der}\, C$ is spanned by these derivations. The subspace $C^0 = \{a \in C : n(1, a) = 0\}$ orthogonal to 1 is invariant under $\mathfrak{der}\, C$.

For later use, let us state some properties of the inner derivations of Cayley algebras. For any a, let $\mathrm{ad}_a = L_a - R_a$ (L_a and R_a denote, respectively, the left and right multiplication by the element a), and consider $\mathrm{ad}_C = \{\mathrm{ad}_a : a \in C\} = \{\mathrm{ad}_a : a \in C^0\}$.

Lemma 5.6. *Let $\mathfrak{C}$ be the Cayley algebra over k* ($\mathrm{char}\, k \neq 2, 3$). *Then*:

(i) $\mathscr{C}^0$ *is invariant under* $\mathfrak{der}\, C$ *and* $\mathrm{ad}_{\mathscr{C}}$, *both of which annihilate* $k1$. *Moreover, as subspaces of* $\mathrm{End}_k(\mathscr{C}^0)$, $\mathfrak{so}(\mathscr{C}^0, n) = \mathfrak{der}\,\mathscr{C} \oplus \mathrm{ad}_{\mathscr{C}}$.

(ii) $[\mathrm{ad}_a, \mathrm{ad}_b] = 2D_{a,b} - \mathrm{ad}_{[a,b]}$ *for any* $a, b \in \mathscr{C}$.

(iii) $D_{a,b} + \frac{1}{2}\mathrm{ad}_{[a,b]} = 3\big(n(a,\cdot)b - n(b,\cdot)a\big)$ *for any* $a, b \in \mathscr{C}^0$.

Proof. First, in [Schafer 1966, Chapter III, §8] it is proved that $\mathfrak{der}\, C$ leaves invariant $\mathscr{C}^0$ and, as a subspace of $\mathrm{End}_k(\mathscr{C}^0)$, it is contained in the orthogonal Lie algebra $\mathfrak{so}(\mathscr{C}^0, n)$. The same happens for $\mathrm{ad}_{\mathscr{C}} = \mathrm{ad}_{\mathscr{C}^0}$, and $\mathfrak{der}\,\mathscr{C} \cap \mathrm{ad}_{\mathscr{C}} = 0$. Hence, by dimension count, $\mathfrak{so}(\mathscr{C}^0, n) = \mathfrak{der}\,\mathscr{C} \oplus \mathrm{ad}_{\mathscr{C}}$, which gives (i).

Now, $\mathscr{C}$ is an alternative algebra. That is, the associator $(a, b, c) = (ab)c - a(bc)$ is alternating on its arguments. Hence, for any $a, b, c \in \mathscr{C}$,

$$\big(L_{ab} - L_aL_b\big)(c) = (a, b, c) = -(a, c, b) = [L_a, R_b](c).$$

Interchange a and b and subtract to get $L_{[a,b]} - [L_a, L_b] = 2[L_a, R_b]$. Similarly,

$$\begin{aligned} R_{[a,b]} + [R_a, R_b] &= -2[L_a, R_b], \\ \mathrm{ad}_{[a,b]} &= [L_a, L_b] + [R_a, R_b] + 4[L_a, R_b]. \end{aligned}$$

Hence

$$\begin{aligned} D_{a,b} &= \mathrm{ad}_{[a,b]} - 3(a, b, \cdot) = \mathrm{ad}_{[a,b]} + 3(a, \cdot, b) \\ &= \mathrm{ad}_{[a,b]} - 3[L_a, R_b] = [L_a, L_b] + [R_a, R_b] + [L_a, R_b] \end{aligned} \tag{5.7}$$

and $[\mathrm{ad}_a, \mathrm{ad}_b] = [L_a - R_a, L_b - R_b] = [L_a, L_b] + [R_a, R_b] - 2[L_a, R_b] = D_{a,b} - 3[L_a, R_b] = 2D_{a,b} - \mathrm{ad}_{[a,b]}$, which yields (ii).

Now, for any $a \in \mathscr{C}$, we have (see [Schafer 1966, Chapter III, §4])

$$a^2 - n(1, a)a + n(a)1 = 0,$$

so for any $a \in \mathscr{C}^0$, $a^2 = -n(a)1$ and hence

$$ab + ba = -n(a, b)1, \tag{5.8}$$

for any $a, b \in \mathscr{C}^0$. Therefore, for any $a, b, c \in \mathscr{C}^0$,

$$\begin{aligned} &2\big(n(a, c)b - n(b, c)a\big) \\ &\quad = -(ac + ca)b - b(ac + ca) + (bc + cb)a - a(bc + cb) \\ &\quad = -(a, c, b) + (b, c, a) - (ca)b + (cb)a - b(ac) + a(bc) \\ &\quad = \big(2[L_a, R_b] + [R_a, R_b] + [L_a, L_b]\big)(c) \\ &\quad = \big(D_{a,b} + [L_a, R_b]\big)(c) = \big(\tfrac{2}{3}D_{a,b} + \tfrac{1}{3}\mathrm{ad}_{[a,b]}\big)(c), \end{aligned}$$

because of (5.7), which gives (iii). □

Let $J = J_{\bar{0}} \oplus J_{\bar{1}}$ be now a unital Jordan superalgebra with a normalized trace $t : J \to k$. That is, t is a linear map such that $t(1) = 1$, and $t(J_{\bar{1}}) = 0 = t\big((J, J, J)\big)$ (see [Benkart and Elduque 2003, §1]). Then $J = k1 \oplus J^0$, where $J^0 = \{x \in J : t(x) = 0\}$, which contains $J_{\bar{1}}$. For $x, y \in J^0$, $xy = t(xy)1 + x * y$, where $x * y = xy - t(xy)1$ is a supercommutative multiplication on J^0. Since $(J, J, J) = [L_J, L_J](J)$ is contained in J^0, the subspace J^0 is invariant under $\mathfrak{inder}\, J = [L_J, L_J]$ (the Lie superalgebra of inner derivations).

Given a unital composition algebra C and a unital Jordan superalgebra with a normalized trace J, consider the superspace

$$\mathcal{T}(C, J) = \mathfrak{der}\, C \oplus (C^0 \otimes J^0) \oplus \mathfrak{inder}\, J,$$

with the superanticommutative product $[\,\cdot\,,\,\cdot\,]$ specified by the following conditions (see [Benkart and Elduque 2003]):

- $\mathfrak{der}\, C$ is a Lie subalgebra and $\mathfrak{inder}\, J$ a Lie subsuperalgebra of $\mathcal{T}(C, J)$,
- $[\mathfrak{der}\, C, \mathfrak{inder}\, J] = 0$,
- $[D, a \otimes x] = D(a) \otimes x$, $[d, a \otimes x] = a \otimes d(x)$,
- $[a \otimes x, b \otimes y] = t(xy)D_{a,b} + [a, b] \otimes x * y - 2n(a, b)[L_x, L_y]$,

for all $D \in \mathfrak{der}\, C$, $d \in \mathfrak{inder}\, J$, $a, b \in C^0$ and $x, y \in J^0$.

If the Grassmann envelope $G(J)$ satisfies the Cayley–Hamilton equation

$$\mathrm{ch}_3(x) = 0$$

for 3×3-matrices, where

$$\mathrm{ch}_3(x) = x^3 - 3t(x)x^2 + \big(\tfrac{9}{2}t(x)^2 - \tfrac{3}{2}t(x^2)\big)x - \big(t(x^3) - \tfrac{9}{2}t(x^2)t(x) + \tfrac{9}{2}t(x)^3\big)1,$$

then $\mathcal{T}(C, J)$ is a Lie superalgebra; see [Benkart and Elduque 2003, Sections 3 and 4].

This construction, for algebras, was considered by Tits [1966], who used it to give a unified construction of the exceptional simple Lie algebras. In the above terms, it was considered in [Benkart and Zelmanov 1996; Benkart and Elduque 2003].

The Kac Jordan superalgebra $\mathcal{J}$ is endowed with a unique normalized trace, given necessarily by $t(1) = 1$ and $t(K \otimes K) = 0$. Note that if $f = f^2$ is an idempotent linearly independent to 1 in a unital Jordan superalgebra with a normalized trace t, and if the Grassmann envelope satisfies the Cayley–Hamilton equation $\mathrm{ch}_3(x) = 0$, in particular $\mathrm{ch}_3(f) = 0$ so, by linear independence, $t(f) - \frac{9}{2}t(f)^2 + \frac{9}{2}t(f)^3 = 0$, and $t(f)$ is either 0, $\frac{1}{3}$ or $\frac{2}{3}$. But, if $t(f)$ were 0, $0 = \mathrm{ch}_3(f)$ would be equal to f^3, hence to f, a contradiction. Hence $t(f) = \frac{1}{3}$ or $\frac{2}{3}$. In the Kac

superalgebra $\mathcal{J}$, the element $f = -\frac{1}{2} + 2e \otimes e$ is an idempotent with $t(f) = -\frac{1}{2}$. Hence the Grassmann envelope of $\mathcal{J}$ cannot satisfy the Cayley–Hamilton equation of degree 3 unless, $-\frac{1}{2} = \frac{1}{3}$ or $-\frac{1}{2} = \frac{2}{3}$, that is, unless the characteristic of k be 5 or 7. Actually, McCrimmon [2005] has shown that the Grassmann envelope $G(\mathcal{J})$ satisfies this Cayley–Hamilton equation if and only if the characteristic is 5. In retrospect, this explains the appearance of the nine-dimensional pseudocomposition superalgebras over fields of characteristic 5 (and only over these fields) in [Elduque and Okubo 2000, Example 9, Theorem 14 and concluding notes].

Assume from now on that the characteristic of the ground field k is 5.

Then, if C is a unital composition algebra, then $\mathcal{T}(C, \mathcal{J})$ is always a Lie superalgebra. Obviously $\mathcal{T}(k, \mathcal{J}) = \mathfrak{inder}\, \mathcal{J} = \mathfrak{der}\, \mathcal{J}$, which is isomorphic to

$$\mathfrak{osp}(1,2) \oplus \mathfrak{osp}(1,2)$$

(see (5.3)), and $\mathcal{T}(k \times k, \mathcal{J})$ is naturally isomorphic to $L_{\mathcal{J}^0} \oplus \mathfrak{der}\, \mathcal{J}$ which, in turn, is isomorphic to $\mathfrak{osp}(K \oplus K) \cong \mathfrak{osp}(2,4)$ (see [Benkart and Elduque 2002, Theorem 2.13]). Also, it is well-known that $\mathcal{T}(\mathrm{Mat}_2(k), \mathcal{J})$ is isomorphic to the Tits–Kantor–Koecher Lie superalgebra of $\mathcal{J}$, which is isomorphic to the exceptional Lie superalgebra of type $F(4)$. (This was used by Kac [1977a] in his classification of the complex finite-dimensional simple Jordan superalgebras.)

Our final result shows that the Lie superalgebra $\mathcal{T}(\mathcal{C}, \mathcal{J})$ is, up to isomorphism, the simple Lie superalgebra in Theorem 3.1 for $l = 5$.

Theorem 5.9. *Let $\mathcal{C}$ be the Cayley algebra and let $\mathcal{J}$ be the Kac Jordan superalgebra over an algebraically closed field k of characteristic* 5. *Then*:

(i) $\mathcal{T}(\mathcal{C}, \mathcal{J})_{\bar{0}}$ *is isomorphic to the orthogonal Lie algebra* $\mathfrak{so}_{11}$.

(ii) $\mathcal{T}(\mathcal{C}, \mathcal{J})_{\bar{1}}$ *is isomorphic to the spin module for* $\mathcal{T}(\mathcal{C}, \mathcal{J})_{\bar{0}}$.

Proof. For (i) consider the vector space

$$M = \mathcal{C}^0 \oplus (U \otimes U)$$

(recall that $U = K_{\bar{1}}$), endowed with the symmetric bilinear form Q such that

$$\begin{aligned} &Q(\mathcal{C}^0, U \otimes U) = 0, \\ &Q(x) = -n(x), \\ &Q(u_1 \otimes u_2, v_1 \otimes v_2) = -(u_1|v_1)(u_2|v_2), \end{aligned}$$

for $x \in \mathcal{C}^0$ and $u_1, u_2, v_1, v_2 \in U$. It will be shown that $\mathcal{T}(\mathcal{C}, \mathcal{J})_{\bar{0}}$ is isomorphic to the orthogonal Lie algebra $\mathfrak{so}(M, Q)$.

This last orthogonal Lie algebra is spanned by the maps

$$\sigma^Q_{x,y} = Q(x, \cdot\,)y - Q(y, \cdot\,)x$$

for $x, y \in M$, and for any $\sigma \in \mathfrak{so}(M, Q)$,

$$[\sigma, \sigma^Q_{x,y}] = \sigma^Q_{\sigma(x),y} + \sigma^Q_{x,\sigma(y)}.$$

Moreover, since $\mathscr{C}^0$ and $U \otimes U$ are orthogonal relative to Q,

$$\mathfrak{so}(M, Q) = \Big(\mathfrak{so}\big(\mathscr{C}^0, Q|_{\mathscr{C}^0}\big) \oplus \mathfrak{so}\big(U \otimes U, Q|_{U\otimes U}\big)\Big) \oplus \sigma^Q_{\mathscr{C}^0, U\otimes U} \tag{5.10}$$

(which gives a $\mathbb{Z}_2$-grading of $\mathfrak{so}(M, Q)$), where $\mathfrak{so}(\mathscr{C}^0, Q|_{\mathscr{C}^0})$ and $\mathfrak{so}(U\otimes U, Q|_{U\otimes U})$ are embedded in $\mathfrak{so}(M, Q)$ in a natural way, and $\mathfrak{so}(M, Q)$ is generated, as a Lie algebra, by $\sigma^Q_{\mathscr{C}^0, U\otimes U}$. Besides, for any $a, b \in \mathscr{C}^0$, and $u_1, u_2, v_1, v_2 \in U$,

$$[\sigma^Q_{a,u_1\otimes u_2}, \sigma^Q_{b,v_1\otimes v_2}] = Q(u_1 \otimes u_2, v_1 \otimes v_2)\sigma^Q_{a,b} + Q(a, b)\sigma^Q_{u_1\otimes u_2, v_1\otimes v_2}.$$

Also, by Lemma 5.6(iii), for any $a, b \in \mathscr{C}^0$,

$$\sigma^Q_{a,b} = -2D_{a,b} - \mathrm{ad}_{[a,b]}. \tag{5.11}$$

Now, the multiplication in $\mathscr{T}(\mathscr{C}, \mathscr{J})$ gives, for any $a, b \in \mathscr{C}^0$, $u, u_1, u_2, v, v_1, v_2 \in U$, $D \in \mathfrak{der}\,\mathscr{C}$ and $d \in (\mathfrak{der}\,\mathscr{J})_{\bar{0}}$:

$$[D, a \otimes (u \otimes v)] = D(a) \otimes (u \otimes v), \tag{5.12a}$$

$$[a \otimes (e \otimes e), b \otimes (u \otimes v)] = \tfrac{1}{4}[a, b] \otimes (u \otimes v) = -\mathrm{ad}_a(b) \otimes (u \otimes v), \tag{5.12b}$$

$$[d, a \otimes (u \otimes v)] = a \otimes d(u \otimes v), \tag{5.12c}$$

$$[D, d] = 0 = [d, a \otimes (e \otimes e)], \tag{5.12d}$$

$$[L_{u_1\otimes u_2}, L_{v_1\otimes v_2}]|_{U\otimes U} = \tfrac{1}{2}\sigma^Q_{u_1\otimes u_2, v_1\otimes v_2}|_{U\otimes U}, \tag{5.12e}$$

since

$$\begin{aligned}(u_1 \otimes u_2)\big((v_1 \otimes v_2)(w_1 \otimes w_2)\big) &= Q(v_1 \otimes v_2, w_1 \otimes w_2)(u_1 \otimes u_2)(e \otimes e - \tfrac{3}{4}1)\\ &= -\tfrac{1}{2}Q(v_1 \otimes v_2, w_1 \otimes w_2)(u_1 \otimes u_2),\end{aligned}$$

for any $u_1, u_2, v_1, v_2, w_1, w_2 \in U$.

Moreover, for any $a, b \in \mathscr{C}^0$ and $u_1, u_2, v_1, v_2 \in U$,

$$\begin{aligned}&[a\otimes(u_1 \otimes u_2), b \otimes (v_1 \otimes v_2)]\\ &\quad= t\big(u_1 \otimes u_2)(v_1 \otimes v_2)\big)D_{a,b} + [a, b] \otimes \big((u_1 \otimes u_2) * (v_1 \otimes v_2)\big)\\ &\qquad\qquad - 2n(a, b)[L_{u_1\otimes u_2}, L_{v_1\otimes v_2}]\\ &\quad= -2Q(u_1 \otimes u_2, v_1 \otimes v_2)D_{a,b} + [a, b] \otimes \big(Q(u_1 \otimes u_2, v_1 \otimes v_2)(e \otimes e)\big)\\ &\qquad\qquad - n(a, b)\big(2[L_{u_1\otimes u_2}, L_{v_1\otimes v_2}]\big)\\ &\quad= Q(u_1 \otimes u_2, v_1 \otimes v_2)\big(-2D_{a,b} + [a, b] \otimes (e \otimes e)\\ &\qquad\qquad + Q(a, b)\big(2[L_{u_1\otimes u_2}, L_{v_1\otimes v_2}]\big).\end{aligned}$$

Now, the equations in Lemma 5.6 and equations (5.11) and (5.12) prove that the linear map

$$\Phi_{\bar{0}} : \mathcal{T}(\mathscr{C}, \mathscr{J})_{\bar{0}} = \mathfrak{der}\, C \oplus \big(\mathscr{C}^0 \otimes (k(e \otimes e) \oplus U \otimes U)\big) \oplus (\mathfrak{der}\, \mathscr{J})_{\bar{0}} \to \mathfrak{so}(M, Q),$$

such that

- $\Phi_{\bar{0}}(D) = D$ for any $D \in \mathfrak{der}\, \mathscr{C}$ $\big(\subseteq \mathfrak{so}(\mathscr{C}^0, n) = \mathfrak{so}(\mathscr{C}^0, -n) \subseteq \mathfrak{so}(M, Q)\big)$, for any $D \in \mathfrak{der}\, \mathscr{C}$,
- $\Phi_{\bar{0}}(d) = d|_{U \otimes U}$ $\big(\in \mathfrak{so}(U \otimes U, Q) \subseteq \mathfrak{so}(M, Q)\big)$, for any $d \in (\mathfrak{der}\, \mathscr{J})_{\bar{0}}$,
- $\Phi_{\bar{0}}\big(a \otimes (e \otimes e)\big) = -\operatorname{ad}_a$ $\big(\in \mathfrak{so}(\mathscr{C}^0, -n) \subseteq \mathfrak{so}(M, Q)\big)$, for any $a \in \mathscr{C}^0$,
- $\Phi_{\bar{0}}\big(a \otimes (u \otimes v)\big) = \sigma^Q_{a, u \otimes v}$, for any $a \in \mathscr{C}^0$ and $u, v \in U$,

is an isomorphism of Lie algebras. This proves the first part of the Theorem.

Now, let us consider the linear map

$$\begin{aligned} \Psi : M &\longrightarrow \operatorname{End}_k\big(\mathscr{C} \otimes (U \oplus U)\big) \\ a \in \mathscr{C}^0 &\mapsto \quad L_a \otimes \begin{pmatrix} -\mathrm{id} & 0 \\ 0 & \mathrm{id} \end{pmatrix}, \\ u_1 \otimes u_2 &\mapsto \mathrm{id} \otimes \begin{pmatrix} 0 & (u_2|\cdot)u_1 \\ (u_1|\cdot)u_2 & 0 \end{pmatrix}. \end{aligned}$$

(The elements in $U \oplus U$ are written as $\binom{u_1}{u_2}$, and then $\operatorname{End}_k(U \oplus U)$ is identified with $\operatorname{Mat}_2\big(\operatorname{End}_k(U)\big)$.) For any $a \in \mathscr{C}^0$ and $u_1, u_2, v_1, v_2 \in U$,

$$\Psi(a)\Psi(u_1 \otimes u_2) + \Psi(u_1 \otimes u_2)\Psi(a) = 0,$$

$$\Psi(a)^2 = -n(a)\,\mathrm{id} = Q(a)\,\mathrm{id} \quad (\text{as } a(ab) = a^2 b = -n(a)b \text{ since } \mathscr{C} \text{ is alternative}),$$

$$\begin{aligned} &\Psi(u_1 \otimes u_2)\Psi(v_1 \otimes v_2) + \Psi(v_1 \otimes v_2)\Psi(u_1 \otimes u_2) \\ &\quad = \mathrm{id} \otimes \begin{pmatrix} (u_2|v_2)(v_1|\cdot)u_1 + (v_2|u_2)(u_1|\cdot)v_1 & 0 \\ 0 & (u_1|v_1)(v_2|\cdot)u_2 + (v_1|u_1)(u_2|\cdot)v_2 \end{pmatrix} \\ &\quad = \mathrm{id} \otimes \begin{pmatrix} -(u_1|v_1)(u_2|v_2)\,\mathrm{id} & 0 \\ 0 & -(u_1|v_1)(u_2|v_2)\,\mathrm{id} \end{pmatrix} \\ &\quad = Q(u_1 \otimes u_2, v_1 \otimes v_2)\,\mathrm{id}, \end{aligned}$$

since $(u_2|v_2)\big((v_1|w_1)u_1 + (w_1|u_1)v_1\big) = -(u_2|v_2)(u_1|v_1)w_1$, because $(u_1|v_1)w_1 + (v_1|w_1)u_1 + (w_1|u_1)v_1 = 0$, as this is an alternating trilinear map on a two-dimensional vector space.

Therefore, Ψ induces an algebra homomorphism

$$\mathfrak{Cl}(M, Q) \to \mathrm{End}_k\big(\mathscr{C} \otimes (U \oplus U)\big),$$

whose restriction $\Psi : \mathfrak{Cl}_{\bar{0}}(M, Q) \to \mathrm{End}_k\big(\mathscr{C} \otimes (U \oplus U)\big)$ is an isomorphism, by dimension count. Therefore, $\mathscr{C} \otimes (U \oplus U)$ is the spin module for $\mathfrak{so}(M, Q)$. Recall that $\mathfrak{so}(M, Q)$ embeds in $\mathfrak{Cl}_{\bar{0}}(M.Q)$ by means of $\sigma^Q_{x,y} \mapsto -\frac{1}{2}[x, y]^{\cdot}$. Since $\mathfrak{so}(M, Q)$ is generated by the elements $\sigma^Q_{a,u_1\otimes u_2}$ ($a \in \mathscr{C}^0$, $u_1, u_2 \in U$), the spin representation is determined by

$$\begin{aligned}\rho\big(\sigma^Q_{a,u_1\otimes u_2}\big) &= -\tfrac{1}{2}\Psi\big([a, u_1 \otimes u_2]^{\cdot}\big) = -\tfrac{1}{2}[\Psi(a), \Psi(u_1 \otimes u_2)] \\ &= -\Psi(a)\Psi(u_1 \otimes u_2) = L_a \otimes \begin{pmatrix} 0 & (u_2|\cdot)u_1 \\ -(u_1|\cdot)u_2 & 0 \end{pmatrix}.\end{aligned}$$

Now identify $\mathcal{T}(\mathscr{C}, \mathscr{J})_{\bar{0}}$ with $\mathfrak{so}(M, Q)$ through $\Phi_{\bar{0}}$, and identify $\mathcal{T}(\mathscr{C}, \mathscr{J})_{\bar{1}} = \mathscr{C}^0 \otimes \big((U \otimes e) \oplus (e \otimes U)\big) \oplus (\mathfrak{der}\,\mathscr{J})_{\bar{1}}$ with $\mathscr{C} \otimes (U \oplus U)$ by means of

$$\begin{aligned}\Phi_{\bar{1}} : \mathcal{T}(\mathscr{C}, \mathscr{J})_{\bar{1}} &\longrightarrow \mathscr{C} \otimes (U \oplus U) \\ a \otimes (u_1 \otimes e + e \otimes u_2) &\mapsto \quad a \otimes \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \\ [L_e, L_{u_1}] \otimes \mathrm{id} + \mathrm{id} \otimes [L_e, L_{u_2}] &\mapsto \ -\tfrac{1}{2}\Big(1 \otimes \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}\Big),\end{aligned}$$

for $a \in \mathscr{C}^0$ and $u_1, u_2 \in U$.

In $\mathcal{T}(\mathscr{C}, \mathscr{J})$, for any $a, b \in \mathscr{C}^0$, $u_1, u_2, v_1, v_2 \in U$, using (5.4) we get

$$\begin{aligned}&[a \otimes (u_1 \otimes u_2), b \otimes (v_1 \otimes e + e \otimes v_2)] \\ &\quad = [a, b] \otimes \tfrac{1}{2}\big(u_1 \otimes (u_2|v_2)e - (u_1|v_1)e \otimes u_2\big) \\ &\qquad - 2n(a, b)[L_{u_1\otimes u_2}, L_{v_1\otimes e + e\otimes v_2}] \\ &\quad = \tfrac{1}{2}[a, b] \otimes \big((u_2|v_2)u_1 \otimes e - e \otimes (u_1|v_1)u_2\big) \\ &\qquad - 2n(a, b)\big(-\tfrac{1}{2}(u_1|v_1)\,\mathrm{id} \otimes [L_{u_2}, L_e] + \tfrac{1}{2}[L_{u_1}, L_e] \otimes (u_2|v_2)\,\mathrm{id}\big) \\ &\quad = \tfrac{1}{2}[a, b] \otimes \big((u_2|v_2)u_1 \otimes e - e \otimes (u_1|v_1)u_2\big) \\ &\qquad - \tfrac{1}{2}n(a, b)\big(-2\big([L_e, L_{(u_2|v_2)u_1}] \otimes \mathrm{id} - \mathrm{id} \otimes [L_e, L_{(u_1|v_1)v_2}]\big)\big).\end{aligned}$$

That is,

$$\Big[a \otimes (u_1 \otimes u_2), \Phi_{\bar{1}}^{-1}\Big(b \otimes \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\Big)\Big] = \Phi_{\bar{1}}^{-1}\Big(ab \otimes \begin{pmatrix} 0 & (u_2|\cdot)u_1 \\ -(u_1|\cdot)u_2 & 0 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\Big),$$

or

$$\Big[\Phi_{\bar{0}}^{-1}\big(\sigma^Q_{a,u_1\otimes u_2}\big), \Phi_{\bar{1}}^{-1}\Big(b \otimes \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\Big)\Big] = \Phi_{\bar{1}}^{-1}\Big(\rho\big(\sigma^Q_{a,u_1\otimes u_2}\big)\Big(b \otimes \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\Big)\Big),$$

because $ab = -\frac{1}{2}n(a,b) + \frac{1}{2}[a,b]$ for any $a, b \in \mathscr{C}^0$ by (5.8). But also,

$$\begin{aligned}\big[a \otimes (u_1 \otimes u_2), [L_e, L_{v_1}] \otimes \mathrm{id} + \mathrm{id} \otimes [L_e, L_{v_2}]\big] \\ = a \otimes \big(-[L_e, L_{v_1}](u_1) \otimes u_2 + u_1 \otimes [L_e, L_{v_2}](u_2)\big) \\ = a \otimes \big(\tfrac{1}{2}(u_1|v_1)e \otimes u_2 - \tfrac{1}{2}u_1 \otimes (u_2|v_2)e\big),\end{aligned}$$

or

$$\left[a \otimes (u_1 \otimes u_2), \Phi_{\bar{1}}^{-1}\left(1 \otimes \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right)\right] = \Phi_{\bar{1}}^{-1}\left(a \otimes \begin{pmatrix} 0 & (u_2|\cdot)u_1 \\ -(u_1|\cdot)u_2 & 0 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right),$$

that is,

$$\left[\Phi_{\bar{0}}^{-1}\big(\sigma^{Q}_{a,u_1\otimes u_2}\big), \Phi_{\bar{1}}^{-1}\left(1 \otimes \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right)\right] = \Phi_{\bar{1}}^{-1}\left(\rho\big(\sigma^{Q}_{a,u_1\otimes u_2}\big)\left(1 \otimes \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right)\right),$$

and this shows that, if $\mathscr{T}(\mathscr{C}, \mathscr{J})_{\bar{0}}$ is identified with $\mathfrak{so}(M, Q)$ by means of $\Phi_{\bar{0}}$ and $\mathscr{T}(\mathscr{C}, \mathscr{J})_{\bar{1}}$ with $\mathscr{C} \otimes (U \oplus U)$ by means of $\Phi_{\bar{1}}$, the action of $\mathscr{T}(\mathscr{C}, \mathscr{J})_{\bar{0}}$ on $\mathscr{T}(\mathscr{C}, \mathscr{J})_{\bar{1}}$ is given, precisely, by the spin representation. □

The Lie superalgebra in Theorem 3.1 for $l = 6$ and characteristic 3, appears in the extended Freudenthal magic square in this characteristic [Cunha and Elduque 2006], as the Lie superalgebra $\mathfrak{g}\big(B(4,2), \overline{B(4,2)}\big)$, associated to two copies of the unique six-dimensional symmetric composition superalgebra. This is related to the six-dimensional simple alternative superalgebra $B(4,2)$ [Shestakov 1997], and hence to the exceptional Jordan superalgebra of 3×3 hermitian matrices over $B(4,2)$, which is exclusive of characteristic 3.

References

[Adams 1996] J. F. Adams, *Lectures on exceptional Lie groups*, Univ. of Chicago Press, Chicago, IL, 1996. MR 98b:22001 Zbl 0866.22008

[Benkart and Elduque 2002] G. Benkart and A. Elduque, "A new construction of the Kac Jordan superalgebra", *Proc. Amer. Math. Soc.* **130**:11 (2002), 3209–3217. MR 2003d:17024 Zbl 1083.17014

[Benkart and Elduque 2003] G. Benkart and A. Elduque, "The Tits construction and the exceptional simple classical Lie superalgebras", *Q. J. Math.* **54**:2 (2003), 123–137. MR 2004c:17016 Zbl 1045.17002

[Benkart and Zelmanov 1996] G. Benkart and E. Zelmanov, "Lie algebras graded by finite root systems and intersection matrix algebras", *Invent. Math.* **126**:1 (1996), 1–45. MR 97k:17044 Zbl 0871.17024

[Brown 1982] G. Brown, "Properties of a 29-dimensional simple Lie algebra of characteristic three", *Math. Ann.* **261**:4 (1982), 487–492. MR 84f:17008 Zbl 0485.17005

[Cunha and Elduque 2006] I. Cunha and A. Elduque, "An extended Freudenthal magic square in characteristic 3", preprint, 2006. math.RA/0605379

[Elduque 2006] A. Elduque, "New simple Lie superalgebras in characteristic 3", *J. Algebra* **296**:1 (2006), 196–233. MR 2006i:17028 Zbl 05019856

[Elduque and Okubo 2000] A. Elduque and S. Okubo, "Pseudo-composition superalgebras", *J. Algebra* **227**:1 (2000), 1–25. MR 2001c:17005 Zbl 0973.17005

[Elduque and Okubo 2002] A. Elduque and S. Okubo, "Composition superalgebras", *Comm. Algebra* **30**:11 (2002), 5447–5471. MR 2003i:17002 Zbl 1015.17002

[Kac 1977a] V. G. Kac, "Classification of simple Z-graded Lie superalgebras and simple Jordan superalgebras", *Comm. Algebra* **5**:13 (1977), 1375–1400. MR 58 #16806 Zbl 0367.17007

[Kac 1977b] V. G. Kac, "Lie superalgebras", *Advances in Math.* **26**:1 (1977), 8–96. MR 58 #5803 Zbl 0366.17012

[Knus et al. 1998] M.-A. Knus, A. Merkurjev, M. Rost, and J.-P. Tignol, *The book of involutions*, Colloquium Publications **44**, Amer. Math. Soc., Providence, RI, 1998. MR 2000a:16031 Zbl 0955.16001

[McCrimmon 2005] K. McCrimmon, "The Grassmann envelope of the Kac superalgebra sK_{10}", preprint, 2005, Available at http://www.uibk.ac.at/mathematik/loos/jordan/archive/gren/index.html.

[Schafer 1966] R. D. Schafer, *An introduction to nonassociative algebras*, Pure and Applied Mathematics **22**, Academic Press, New York, 1966. Reprinted Dover, New York, 1995. MR 96j:17001 Zbl 0145.25601

[Shestakov 1997] I. P. Shestakov, "Prime alternative superalgebras of arbitrary characteristic", *Algebra i Logika* **36**:6 (1997), 675–716. MR 99k:17006 Zbl 0904.17025

[Tits 1966] J. Tits, "Algèbres alternatives, algèbres de Jordan et algèbres de Lie exceptionnelles, I: Construction", *Indag. Math.* **28** (1966), 223–237. MR 36 #2658 Zbl 0139.03204

[Witt 1941] E. Witt, "Spiegelungsgruppen und Aufzählung halbeinfacher Liescher Ringe", *Abh. Math. Sem. Hansischen Univ.* **14** (1941), 289–322. MR 3,100f JFM 67.0077.03

Received December 30, 2005.

ALBERTO ELDUQUE
DEPARTAMENTO DE MATEMÁTICAS
UNIVERSIDAD DE ZARAGOZA
50009 ZARAGOZA
SPAIN

elduque@unizar.es
http://www.unizar.es/matematicas/algebra/elduque

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.361

SUBFACTORS FROM BRAIDED C^* TENSOR CATEGORIES

JULIANA ERLIJMAN AND HANS WENZL

We extend subfactor constructions originally defined for unitary braid representations to the setting of braided C^*-tensor categories. The categorical approach is then used to compute the principal graph of these subfactors. We also determine the dual principal graph for several important cases. Here invertibility of the so-called S-matrix of a subcategory and certain related group actions play an important role.

It was noted by Vaughan Jones that his examples of subfactors gave rise to unitary braid representations. By this we mean representations of the infinite braid group $\mathcal{B}_\infty$ defined by infinitely many generators $\sigma_1, \sigma_2, \ldots$ which satisfy the familiar braid relations. Subsequently, unitary braid representations were used by A. Ocneanu and by H. Wenzl to construct new examples of subfactors; here the subfactor is given by the subgroup $\mathcal{B}_{2,\infty}$ generated by $\sigma_2, \sigma_3, \ldots$. This construction was denoted as the one-sided subfactor construction by J. Erlijman, as opposed to her multisided subfactors. Here, for a given integer $s > 1$, the s-sided subfactor is obtained as a suitable inductive limit of the embeddings of the quotients of $\mathcal{B}_n^s = \mathcal{B}_n \times \cdots \times \mathcal{B}_n$ (s times) into $\mathcal{B}_{ns}$ for $n \to \infty$. She also computed the indices of these subfactors and their first relative commutants.

The main motivation for this paper was to calculate the higher relative commutants of Erlijman's subfactors. To do this it is convenient to generalize the above mentioned constructions to the setting of a braided C^*-tensor category $\mathcal{C}$ with only finitely many simple objects up to isomorphism. By definition of such a category, we obtain a unitary representation of $\mathcal{B}_n$ in $\mathrm{End}(X^{\otimes n})$ for any object X in $\mathcal{C}$. The constructions in our paper in the category setting follow closely the above-mentioned braid constructions. They reduce to them in case that $\mathrm{End}(X^{\otimes n})$ is generated by the quotients of $\mathcal{B}_n$ for all $n \in \mathbb{N}$, where X is a generating object of $\mathcal{C}$. However, the categorical setting makes it easier to calculate the higher relative commutants, and also contains new nontrivial examples.

The main results of our paper are as follows. We show that the first principal graph is given by the fusion graph of $(\mathcal{C}')^s$, where $\mathcal{C}'$ is a subcategory of $\mathcal{C}$ depending on the tensor powers of X in which the trivial object appears. The fusion graph

MSC2000: primary 46L37; secondary 18D10.
Keywords: subfactor, braided C^* tensor category.

describes the decomposition of the tensor product of s simple objects of $\mathcal{C}'$ into irreducible ones; see Theorem 4.6 for details. The situation is more complicated for the dual (or second) principal graph. If a certain matrix depending on the braiding structure, called the S-matrix for the category $\mathcal{C}'$, is invertible, the dual principal graph coincides with the principal graph.

We do not have a general complete result in the case of a noninvertible S-matrix. It is known that in this case there is a canonical subcategory $\mathcal{T}$ of $\mathcal{C}'$ which is equivalent to the representation category of a finite group G. If G is abelian, we obtain an action of G on the set of irreducible objects of $\mathcal{C}$, which is given by a labeling set Λ. The dual principal graph can now be fairly precisely characterized in terms of the orbits of the action of a group G_1^s on Λ^s; see Theorem 5.9 for details and, for an example, Proposition 6.1.

The basic idea of our paper is that we explicitly construct a number of $\mathcal{A}$–$\mathcal{B}$ bimodules, with $\{\mathcal{A}, \mathcal{B}\} \subset \{\mathcal{N}, \mathcal{M}\}$ and with $\mathcal{N} \subset \mathcal{M}$ being our s-sided inclusion. We show that these examples of bimodules are closed under induction and restriction. One deduces from this that the induction-restriction graph for these bimodules must coincide with the principal or dual principal graph under some mild additional assumptions.

Our findings are related to a number of results by different authors. If $s = 2$, our subfactors correspond to the subfactors obtained from the asymptotic inclusion of certain one-sided subfactors. In this case, the orbifold phenomenon for the dual principal graph has first been observed by Ocneanu for the example of the Jones subfactors. Further results have been obtained in [Evans and Kawahigashi 1998] and [Izumi 2000]. In particular, some of our proofs have been inspired by these results. More recently, after hearing a talk on this paper, M. Asaeda [2006] obtained an analogue of the s-sided construction under more general conditions.

More or less the same combinatorics as in our paper also appears in the work of Feng Xu [2000] on subfactors of type III_1 factors related to disconnected intervals. In spite of the similarity of principal graphs and indices, his construction of these subfactors is completely different from ours and relies on Wassermann's loop group construction, which has not appeared yet in print for all Lie types.

Here is a more detailed description of the contents of this paper. In the first chapter we review some basic results on bimodules in the type II_1 setting. The second chapter contains definitions concerning braided C^* tensor categories. In the third chapter we present the generalization of previous subfactor constructions to the setting of braided C^* tensor categories, as well as additional technical results. This is used in the following section to construct certain bimodules and compute the principal graph of these subfactors. In the last section we prove the already mentioned results about the dual principal graph. We then discuss examples of our construction including the case of the Jones subfactors.

1. Bimodules

1A. *Definitions.*

Definition 1.1. Let $\mathcal{A}$ and $\mathcal{B}$ be type II_1 factors, and let H be a Hilbert space.

(i) H is a *left $\mathcal{A}$-module* if there exists an action of $\mathcal{A}$ on H determined by a normal unital morphism $\lambda : \mathcal{A} \to B(H)$, where $B(H)$ is the von Neumann algebra of all bounded linear operators on H.

(ii) A *right $\mathcal{B}$-module* H is a left $\mathcal{B}^{\mathrm{opp}}$-module (here, $\mathcal{B}^{\mathrm{opp}}$ denotes the opposite algebra of $\mathcal{B}$).

(iii) H is an *$\mathcal{A}$–$\mathcal{B}$ bimodule* if it is a left $\mathcal{A}$-module, a right $\mathcal{B}$-module, and if the left and right actions intertwine. That is, if $\lambda : \mathcal{A} \to B(H)$ is the left action, and if $\rho : \mathcal{B}^{\mathrm{opp}} \to B(H)$ is the right action, then we must have that $\lambda(a)\rho(b) = \rho(b)\lambda(a)$ for all $a \in \mathcal{A}$, $b \in \mathcal{B}$.

(iv) If H and K are $\mathcal{A}$–$\mathcal{B}$ bimodules, we define the space of *intertwiners*, denoted by $\mathrm{Hom}_{\mathcal{A},\mathcal{B}}(H, K)$, to be the set of linear bounded operators $T : H \to K$ such that they intertwine the actions, in the sense that $T\lambda_H(a) = \lambda_K(a)T$ for all $a \in \mathcal{A}$ and $T\rho_H(b) = \rho_K(b)T$ for all $b \in \mathcal{B}$.

(v) Two $\mathcal{A}$–$\mathcal{B}$ bimodules H and K are *equivalent* or *isomorphic* if there exists a unitary operator in $\mathrm{Hom}_{\mathcal{A},\mathcal{B}}(H, K)$.

Definition 1.2. Let H be an $\mathcal{A}$–$\mathcal{B}$ bimodule with left action λ and right action ρ. The *inclusion generated by H* is the inclusion of factors given by

$$\lambda(\mathcal{A}) \subset \rho(\mathcal{B})'.$$

The *dual inclusion generated by H* is the inclusion of factors given by

$$\rho(\mathcal{B}) \subset \lambda(\mathcal{A})'.$$

Remark 1.3. Similarly, if we have an inclusion of type II_1-factors $\mathcal{N} \subset \mathcal{M}$, we can make $L^2(\mathcal{M}, \mathrm{tr})$ into an $\mathcal{M}$–$\mathcal{M}$, $\mathcal{M}$–$\mathcal{N}$, $\mathcal{N}$–$\mathcal{M}$ or $\mathcal{N}$–$\mathcal{N}$-bimodule via usual left and right multiplication. If $\mathcal{N} \subset \mathcal{M}$ is a reducible inclusion, i.e., the relative commutant $\mathcal{N}' \cap \mathcal{M}$ is larger than $\mathbb{C}1$, then we obtain further examples by reducing by projections in the relative commutant. For example, if $p \in \mathcal{N}' \cap \mathcal{M}$, we obtain the $\mathcal{N}$–$\mathcal{M}$ bimodule $L^2(p\mathcal{M}, \mathrm{tr})$.

If $\phi_i : \mathcal{M} \to \mathcal{M}$ are endomorphisms for $i = 1, 2$, we can also define an $\mathcal{M}$–$\mathcal{M}$-bimodule structure on $L^2(\mathcal{M}, \mathrm{tr})$ by perturbing the right and left actions by these endomorphisms, that is, by defining the action by $m_1.\xi.m_2 = \phi_1(m_1)\xi\phi_2(m_2)$.

All the examples of bimodules encountered in this paper are of one of these types or tensor products or direct summands of them.

Definition 1.4. Let $\mathcal{A}$ and $\mathcal{B}_i$ be type II_1 factors for $i = 1, 2$. Let H_i be $\mathcal{A}$–$\mathcal{B}_i$ bimodules with left actions λ_i and right actions ρ_i, respectively, for $i = 1, 2$, and assume that $\dim_{\mathcal{A}}(H_2) \le \dim_{\mathcal{A}}(H_1) < \infty$. Then we say that H_2 is *(left)-weakly reduced* or a *(left)-weak reduction* of H_2 if there exists a nonzero projection $p \in \mathcal{B}_1$ and an isomorphism $\Psi : \mathcal{B}_2 \cong p\mathcal{B}_1 p$ such that $H_1 p := \rho_1(b)H_1$ and H_2 are isomorphic $\mathcal{A} - \mathcal{B}_2$-bimodules; here the $\mathcal{A} - \mathcal{B}_2$-bimodule structure on $\mathcal{H}_1 p$ is defined by $a.\xi.b = \lambda_1(a)\xi\rho_1(\Psi(b))$ for $a \in A$, $b \in B_2$ and $\xi \in H_1 p$.

Remark 1.5. (1) Since right multiplication by p commutes with the left action of $\mathcal{A}$ and also with the commutant of the right action of $\mathcal{B}_1$, we obtain isomorphic inclusions $\lambda_1(\mathcal{A}) \subset \rho_1(\mathcal{B}_1)'$ and $\lambda_1(\mathcal{A})p \subset \rho_1(\mathcal{B}_1)'p$. It follows from this and the fact that isomorphic bimodules define isomorphic inclusions that a left weak reduction of a bimodule yields an isomorphic inclusion.

(2) If we perturb the right-action on an $\mathcal{A}$–$\mathcal{B}$ bimodule H by an outer automorphism α of $\mathcal{B}$, the resulting bimodule H_α is not isomorphic to H. However, it is a left weak reduction of H.

(3) One can similarly define a notion of (right)-weak reduction. We shall mostly be concerned with (left)-weak reduction, and will usually just call it weak reduction. Also, we shall often suppress the notation λ and ρ if it is clear from which side the algebras act.

1B. ***Tensor products.*** Tensor products of bimodules have been defined by Connes and Sauvageot. A good review with results for our paper can be found in [Bisch 1997].

Proposition 1.6. *Let H_i be $\mathcal{A}$–$\mathcal{B}_i$ bimodules for $i = 1, 2$, and let $\mathcal{D}$ be a type II_1 factor. If H_2 is weakly reduced from H_1, then also $L \otimes_{\mathcal{A}} H_2$ is weakly reduced from $L \otimes_{\mathcal{A}} H_1$, for any $\mathcal{D}$–$\mathcal{A}$ bimodule L.*

Proof. By definition, since H_2 is weakly reduced from H_2, there must exist a projection $p \in \mathcal{B}_1$ such that $H_1 p$ and H_2 are isomorphic as $\mathcal{A}$–$\mathcal{B}_2$ bimodules, assuming $\dim_{\mathcal{A}}(H_1) \ge \dim_{\mathcal{A}}(H_2)$. This isomorphism extends in an obvious way to a spatial isomorphism between $L \otimes_{\mathcal{A}} H_1 p = (L \otimes_{\mathcal{A}} H_1)(1 \otimes p)$ and $L \otimes_{\mathcal{A}} H_2$. □

1C. ***Higher relative commutants.*** Let $\mathcal{N} \subset \mathcal{M}$ be type II_1 factors with normalized trace tr. There exists a canonical extension $\mathcal{M}_1 \supset \mathcal{M}$, called Jones' basic construction for $\mathcal{N} \subset \mathcal{M}$, which is the von Neumann algebra generated by $\mathcal{M}$ acting via left multiplication on $L^2(\mathcal{M}, \mathrm{tr})$ and by the orthogonal projection $e_{\mathcal{N}}$ onto the subspace $L^2(\mathcal{N}, \mathrm{tr}) \subset L^2(\mathcal{M}, \mathrm{tr})$. It is well-known that the Jones index $[\mathcal{M} : \mathcal{N}]$ is finite if and only if $\mathcal{M}_1$ is again a type II_1 factor; it is given by $[\mathcal{M} : \mathcal{N}] = 1/\mathrm{tr}(e_{\mathcal{N}})$, with tr denoting the unique normalized trace on $\mathcal{M}_1$. In this case, we can apply the basic construction again for $\mathcal{M} \subset \mathcal{M}_1$ to obtain an extension $\mathcal{M}_2 \supset \mathcal{M}_1$. Iterating this

construction, we obtain a sequence of II_1 factors $\mathcal{N} \subset \mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots$. We obtain important invariants of the original inclusion $\mathcal{N} \subset \mathcal{M}$ via the so-called higher relative commutants $\mathcal{N}' \cap \mathcal{M}_k$ and $\mathcal{M}' \cap \mathcal{M}_k$. These are finite-dimensional C*-algebras. If there exists a uniform bound for the dimensions of the centers of the relative commutants, the subfactor $\mathcal{N} \subset \mathcal{M}$ is called a *finite depth subfactor*. In this case, the inclusion diagram for $\mathcal{N}' \cap \mathcal{M}_{2k} \subset \mathcal{N}' \cap \mathcal{M}_{2k+1}$ does not depend on k for k sufficiently large; the corresponding graph is called the *principal graph* of $\mathcal{N} \subset \mathcal{M}$. Similarly, one defines the *dual principal graph* from the inclusion of $\mathcal{M}' \cap \mathcal{M}_{2k} \subset \mathcal{M}' \cap \mathcal{M}_{2k+1}$ for k sufficiently large. These graphs are important invariants for the inclusion $\mathcal{N} \subset \mathcal{M}$.

The following results are presented in [Bisch 1997] in great detail and with precise references to the original sources.

Proposition 1.7. *Let $\mathcal{N} \subset \mathcal{M}$ be a finite depth subfactor with finite index. Then*

(a) *The inclusions $\mathcal{N} \subset \mathcal{M}_{2k+1}$, $\mathcal{N} \subset \mathcal{M}_{2k}$, $\mathcal{M} \subset \mathcal{M}_{2k+1}$, $\mathcal{M} \subset \mathcal{M}_{2k}$ are given by the bimodule $\mathcal{M}^{\otimes k} = \mathcal{M} \otimes_{\mathcal{N}} \mathcal{M} \otimes_{\mathcal{N}} \cdots \otimes_{\mathcal{N}} \mathcal{M}$ (k times), viewed, respectively, as an $\mathcal{N}$–$\mathcal{N}$, $\mathcal{N}$–$\mathcal{M}$, $\mathcal{M}$–$\mathcal{N}$ and $\mathcal{M}$–$\mathcal{M}$ bimodule.*

(b) *The embedding of $\mathcal{N}' \cap \mathcal{M}_k \subset \mathcal{N}' \cap \mathcal{M}_{k+1}$ coincides with the embedding of the algebras $\mathrm{End}_{\mathcal{M}-\mathcal{N}}(\mathcal{M}^{\otimes k}) \subset \mathrm{End}_{\mathcal{N}-\mathcal{N}}(\mathcal{M}^{\otimes k})$ for k even. If k is odd, the embedding of $\mathcal{N}' \cap \mathcal{M}_k \subset \mathcal{N}' \cap \mathcal{M}_{k+1}$ coincides with the embedding of $\mathrm{End}_{\mathcal{N}-\mathcal{N}}(\mathcal{M}^{\otimes k}) \subset \mathrm{End}_{\mathcal{M}-\mathcal{N}}(\mathcal{M}^{\otimes k+1})$, given by $x \in \mathrm{End}_{\mathcal{N}-\mathcal{N}}(\mathcal{M}^{\otimes k}) \to 1_{\mathcal{M}} \otimes x$.*

(c) *Analogous statements hold for the embedding of $\mathcal{M}' \cap \mathcal{M}_k \subset \mathcal{M}' \cap \mathcal{M}_{k+1}$; we only need to replace $\mathrm{Hom}_{\mathcal{X}-\mathcal{N}}$ by $\mathrm{Hom}_{\mathcal{X}-\mathcal{M}}$ in all the statements in (b), with $\mathcal{X} \in \{\mathcal{M}, \mathcal{N}\}$.*

Proof. Statement (a) is shown in [Bisch 1997], Proposition 3.2. Statement (b) can be found in [Bisch 1997], Corollaries 4.2 and 4.4 (with tensoring from the right instead of tensoring from the left, as we have chosen here). Statement (c) follows from (b) and (a). □

Let $\mathcal{N}, \mathcal{M}, \mathcal{B}$ be type II_1 factors with $\mathcal{N} \subset \mathcal{M}$ a subfactor of finite index. Let $\{H_\lambda\}_\lambda$ and $\{K_\nu\}_\nu$ be a collection of mutually nonisomorphic irreducible $\mathcal{N}$–$\mathcal{B}$ and $\mathcal{M}$–$\mathcal{B}$ bimodules, respectively. Observe that $\mathcal{M} \otimes_{\mathcal{N}} H_\lambda$ is an $\mathcal{M}$–$\mathcal{B}$ bimodule for any $\mathcal{N}$–$\mathcal{B}$ bimodule H_λ. Similarly, we can view any $\mathcal{M}$–$\mathcal{B}$ bimodule K_ν as an $\mathcal{N}$–$\mathcal{B}$ bimodule by restricting the left action to $\mathcal{N}$. We say that the system of bimodules $(\{H_\lambda\}_\lambda, \{K_\nu\}_\nu)$ is *closed under induction and restriction* if

– for each $\mathcal{N}$–$\mathcal{B}$ bimodule H_λ the induced $\mathcal{M}$–$\mathcal{B}$ bimodule $\mathcal{M} \otimes_{\mathcal{N}} H_\lambda$ is isomorphic to a direct sum of irreducible $\mathcal{M}$–$\mathcal{B}$ bimodules each of which is isomorphic to an element in $\{K_\nu\}_\nu$,

– for each $\mathcal{M}$–$\mathcal{B}$ bimodule K_ν the $\mathcal{N}$–$\mathcal{B}$ bimodule $\mathcal{M} \otimes_{\mathcal{M}} K_\nu$ obtained from K_ν by restricting the left action to $\mathcal{N}$ is isomorphic to a direct sum of irreducible $\mathcal{N}$–$\mathcal{B}$ bimodules each of which is isomorphic to an element in $\{H_\lambda\}_\lambda$.

The *induction-restriction graph* for our system of bimodules is the bipartite graph whose (say) odd vertices are labeled by the elements in $\{H_\lambda\}_\lambda$ and whose even vertices are labeled by the elements in $\{K_\nu\}_\nu$. A vertex labeled by H_λ is connected with a vertex labeled by K_ν by L_λ^ν edges, where L_λ^ν is the multiplicity of H_λ in K_ν, viewed as an $\mathcal{N}$–$\mathcal{B}$ bimodule. By Frobenius reciprocity (see [Bisch 1997, Theorem 1.18], for example), this number coincides with the multiplicity of K_ν in $\mathcal{M} \otimes_{\mathcal{N}} H_\lambda$.

Proposition 1.8. *Let* $(\{H_\lambda\}_\lambda, \{K_\nu\}_\nu)$ *be a system of* $\mathcal{N}$–$\mathcal{B}$*- and* $\mathcal{M}$–$\mathcal{B}$*-bimodules which is closed under induction and restriction.*

(a) *If* $\{H_\lambda\}_\lambda$ *contains a bimodule* H_0 *which is weakly reduced from the trivial* $\mathcal{N}$–$\mathcal{N}$*-bimodule* $\mathcal{N}$*, then the principal graph for* $\mathcal{N} \subset \mathcal{M}$ *is given by the connected component of the induction- restriction graph for* $(\{H_\lambda\}_\lambda, \{K_\nu\}_\nu)$ *which contains* H_0*.*

(b) *If* $\{K_\nu\}_\nu$ *contains a bimodule* K_0 *which is weakly reduced from the trivial* $\mathcal{M}$–$\mathcal{M}$*-bimodule* $\mathcal{M}$*, then the dual principal graph for* $\mathcal{N} \subset \mathcal{M}$ *is given by the connected component of the induction- restriction graph for* $(\{H_\lambda\}_\lambda, \{K_\nu\}_\nu)$ *which contains* K_0*.*

(c) *If* $\phi : \mathcal{B} \to \mathcal{B}$ *is an endomorphism of the* II_1 *factor* $\mathcal{B}$ *and* $p \in \phi(\mathcal{B})' \cap \mathcal{B}$ *such that* $[p\mathcal{B}p : p\phi(\mathcal{B})] = 1$*, then* $L^2(\mathcal{B}, \mathrm{tr})p$*, with action* $b_1.\xi p.b_2 = b_1 \xi p \phi(b_2)$ *for* $b_1, b_2 \in \mathcal{B}$*,* $\xi \in L^2(\mathcal{B}, \mathrm{tr})$ *is weakly reduced from the trivial* $\mathcal{B}$–$\mathcal{B}$ *bimodule* $L^2(\mathcal{B}, \mathrm{tr})$*.*

Proof. Part (a) follows from Proposition 1.6 and Proposition 1.7(b). Similarly, part (b) follows from Proposition 1.6 and Proposition 1.7(c). Part (c) follows almost immediately from Definition 1.4, using the fact that $p\mathcal{B}p = \phi(\mathcal{B})p$. □

Remark 1.9. In the setting of Proposition 1.8(a), there may be more than one bimodule H_λ which is weakly reduced from the trivial $\mathcal{N}$–$\mathcal{N}$-bimodule $\mathcal{N}$. Which of those will correspond to the trivial $\mathcal{N}$–$\mathcal{N}$-bimodule $\mathcal{N}$ will depend on the choice of the automorphism between $p\mathcal{N}p$ and $\mathcal{B}$. The resulting graph will be independent of this choice. A similar phenomenon may also occur in part (b).

Let H be an $\mathcal{A}$–$\mathcal{B}$ bimodule. We define $\mathrm{ind}(H)$ to be equal to the index $[\rho(\mathcal{B})' : \lambda(\mathcal{A})] = [\lambda(\mathcal{A})' : \rho(\mathcal{B})]$. In the following lemma, $(H_\lambda)_\lambda$ and $(K_\nu)_\nu$ are bimodules as in the last proposition, where we now assume for simplicity that they only denote the bimodules which label the vertices of a given principal graph. Moreover, we also assume the subfactor to be of finite depth, meaning that both sets only contain finitely many bimodules.

Lemma 1.10. *With notations as above, we have*:

(a) $\sum_\nu \mathrm{ind}(K_\nu) = \sum_\lambda \mathrm{ind}(H_\lambda)$.

(b) *Assume that the $\mathcal{A}$–$\mathcal{B}$-bimodule H decomposes as $H = \bigoplus m_i H_i$, with H_i irreducible $\mathcal{A}$–$\mathcal{B}$-bimodules, and let $l = \dim(\mathrm{End}_{\mathcal{A},\mathcal{B}}(H)) = \sum_i m_i^2$. Then we have $\sum_i \mathrm{ind}(H_i) \geq \mathrm{ind}(H)/l$, with equality only if $\dim_{\mathcal{A}}(H_i) = m_i \dim_{\mathcal{A}}(H)/l$ for all i.*

Proof. It is well-known that the inclusion of higher relative commutants $\mathcal{M}' \cap \mathcal{M}_k \subset \mathcal{N}' \cap \mathcal{M}_k$ defines periodic commuting squares which generate in the limit a subfactor of index $[\mathcal{M} : \mathcal{N}]$. Hence we can use the results of [Wenzl 1988], Theorem 1.5(iii). It follows that the index is equal to the quotient of the l^2-norms of the weight vectors of $\mathcal{M}' \cap \mathcal{M}_k$ and $\mathcal{N}' \cap \mathcal{M}_k$ for k sufficiently large. Let p_λ and p_μ be minimal idempotents in $\mathcal{M}' \cap \mathcal{M}_k$ and $\mathcal{N}' \cap \mathcal{M}_k$ respectively. Then we have $\mathrm{ind}(p_\nu \mathcal{M}_k) = \mathrm{tr}(p_\nu)^2 [\mathcal{M} : \mathcal{N}]^k$ and $\mathrm{ind}(p_\lambda \mathcal{M}_k) = \mathrm{tr}(p_\lambda)^2 [\mathcal{M} : \mathcal{N}]^{k+1}$. Solving for $\mathrm{tr}(p_\lambda)^2$ and $\mathrm{tr}(p_\nu)^2$, we obtain

$$[\mathcal{M} : \mathcal{N}] = \frac{\sum_\nu \mathrm{ind}(p_\nu \mathcal{M}_k)/[\mathcal{M} : \mathcal{N}]^k}{\sum_\lambda \mathrm{ind}(p_\lambda \mathcal{M}_k)/[\mathcal{M} : \mathcal{N}]^{k+1}}.$$

The claimed formula follows from this in the case that our system of bimodules labels the vertices of the principal graph. One obtains the claim for the dual principal graph by the same proof applied to the inclusion $\mathcal{M} \subset \mathcal{M}_1$.

Part (b) is proved using Lagrange multipliers as follows: Let $x_i = \dim_{\mathcal{A}}(H_i)$ and let $d = \dim_{\mathcal{A}} H$. Then the minimum of the function $f(x_1, \ldots, x_r) = \sum x_i^2$ subject to the condition $\sum m_i x_i = d$ is obtained for $2x_i = \lambda m_i$, and we deduce from the constraint that $d = \frac{\lambda}{2} \sum m_i^2 = l\lambda/2$. Hence $x_i = m_i d/l$ and

$$\sum_i (\dim_{\mathcal{A}} H_i)^2 \geq \frac{d^2}{l^2} \sum_i m_i^2 = d^2/l. \qquad (*)$$

Now observe that if p_i is the projection onto the submodule $H_i \subset H$, we have $\mathrm{tr}(p_i) = \dim_{\mathcal{A}}(H_i)/\dim_{\mathcal{A}}(H)$ and $\mathrm{ind}(H_i) = \mathrm{tr}(p_i)^2 \mathrm{ind}(H)$ (again see [Wenzl 1988], Theorem 1.5(iii)). The claim follows from this after multiplying $(*)$ by $\mathrm{ind}(H)/d^2$. □

2. Categories

In this section we deal with categories which can be considered as generalizations of the representation categories of finite groups. This allows us to deal simultaneously with categories of bimodules of von Neumann factors, fusion categories (which can be constructed using quantum groups or loop groups) and categories obtained from unitary braid representations. For more details, we refer to [Mac Lane 1998], [Freyd 1964] for general categorical notions, and to [Kassel 1995], [Turaev 1994] for tensor categories; our treatment of traces also uses results from [Longo and Roberts 1997].

2A. ***General definitions.*** We recall some basic definitions and set up notations.

In the following, $\mathcal{C}$ will always denote a strict monoidal complex tensor category with unit $\mathbb{1}$. This means that $\mathcal{C}$ is a category with a functor $\otimes : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$ called the *tensor product* which satisfies certain associativity conditions such as the *Pentagon Axiom*. There are similar axioms involving the morphisms $l_X : \mathbb{1} \otimes X \to X$ and $r_X : X \otimes \mathbb{1} \to X$ called the left and right *unit constraints*. Moreover, $\mathcal{C}$ being a complex category just means that the homomorphisms $\mathrm{Hom}(X, Y)$ form a complex vector space for any objects X and Y in $\mathcal{C}$.

The complex tensor category $\mathcal{C}$ is called a $*$ tensor category if there exists a contragredient complex conjugate functor $* : \mathcal{C} \to \mathcal{C}$ which is compatible with $\otimes$. This means in detail that:

- if $f \in \mathrm{Hom}(X, Y)$, then $f^* \in \mathrm{Hom}(Y, X)$,
- $(\alpha f)^* = \bar{\alpha} f^*$ for all $\alpha \in \mathbb{C}$ and $f \in \mathrm{Hom}(X, Y)$,
- $(fg)^* = g^* f^*$ for $f \in \mathrm{Hom}(X, Y)$ and $g \in \mathrm{Hom}(U, X)$,
- $(f \otimes g)^* = f^* \otimes g^*$ for $f \in \mathrm{Hom}(X, Y)$ and $g \in \mathrm{Hom}(U, V)$,
- $1_X^* = 1_X$ for the identity morphism 1_X for any object X in $\mathcal{C}$.

2B. ***Duality and Frobenius reciprocity.*** An object X in a strict monoidal category $\mathcal{C}$ is called *left rigid* if there exists an object $\bar{X} \in \mathcal{C}$ and a pair of morphisms $i_X : \mathbb{1} \to X \otimes \bar{X}$ and $d_X : \bar{X} \otimes X \to \mathbb{1}$ such that the maps $(1_X \otimes d_X)(i_X \otimes 1_X) : X \to X$ and $(d_X \otimes 1_{\bar{X}})(1_X \otimes i_{\bar{X}}) : \bar{X} \to \bar{X}$ are 1_X and $1_{\bar{X}}$. An object X is called right rigid if we can find an object $\bar{X}'$ and morphisms $i'_X : \mathbb{1} \to \bar{X}' \otimes X$ and $d'_X : X \otimes \bar{X}' \to \mathbb{1}$ satisfying analogous identities. It is easy to check that in a $*$ category any left rigid object is also right rigid, with $\bar{X}' = \bar{X}$, $i'_X = d_X^*$ and $d'_X = i_X^*$. Hence we will in the following only talk about rigid objects. A category $\mathcal{C}$ is called rigid if every object of $\mathcal{C}$ is rigid.

With this notion of duality, we also have the usual Frobenius reciprocity isomorphism between $\mathrm{Hom}(V, W \otimes \bar{X})$ and $\mathrm{Hom}(V \otimes X, W)$ for any objects V, W in $\mathcal{C}$. One checks easily that these isomorphisms are given by the maps

$$a \mapsto (1_W \otimes d_X) \circ (a \otimes 1_X) \quad \text{and} \quad b \mapsto (b \otimes 1_Y) \circ (1_V \otimes i_X)$$

for $a \in \mathrm{Hom}(V, W \otimes X)$ and $b \in \mathrm{Hom}(V \otimes Y, W)$. In particular, one obtains as a special case that $\dim \mathrm{Hom}(\mathbb{1}, X \otimes \bar{X}) = \dim \mathrm{End}(X) = 1$ if X is a simple object. Hence the morphisms i_X and d_X are unique up to scalar multiples for X simple. We shall say that the rigidity morphisms i_X and d_X are *normalized* if $i_X^* i_X = d_X d_X^* *$.

2C. ***Dimension, trace and conditional expectation.*** In the following we always assume the rigidity morphisms i_X and d_X to be normalized for any object X. If X is simple, this can always be assumed after some rescaling in view of the discussion

in the last section. For normalized rigidity morphisms, we can now define the dimension of a simple object X to be equal to the scalar

$$\dim(X) = i_X^* i_X = d_X d_X^*.$$

Of course, we would like the dimension to be additive with respect to a decomposition $X = \bigoplus W_i$, with the W_i being simple objects. To do so, we define morphisms $\phi_i : W_i \to X$ such that $\phi_i^* \phi_j = \delta_{ij} 1_{W_i}$ and $\sum_i \phi_i \phi_i^* = 1_X$, and we define

$$i_X = \sum (\phi_i \otimes \bar{\phi}_i) i_{W_i}, \quad d_X = \sum d_{W_i} (\bar{\phi}_i^* \otimes \phi_i^*), \tag{2-1}$$

where the $\bar{\phi}_i$ are the analogous morphisms for the decomposition of the dual $\bar{X} = \sum \bigoplus_i \bar{W}_i$. Then it is easy to check that these morphisms satisfy the rigidity axiom, and they are normalized if the ϕ_i are so. Moreover, one also checks that these morphisms yield the desired additivity property of the dimension function.

Additionally, the dimension function should be multiplicative with respect to the tensor product. If $X \otimes Y$ is a tensor product of simple objects X and Y, we obtain normalized rigidity morphisms

$$i_{X\otimes Y} = (1_X \otimes i_Y \otimes 1_{\bar{X}}) i_X, \quad d_{X\otimes Y} = d_Y (1_{\bar{X}} \otimes d_X \otimes 1_X).$$

It can be shown that these rigidity morphisms define the same dimension as the one we obtain from the decomposition $X \otimes Y \cong \bigoplus_i W_i$, with W_i simple and with rigidity morphisms as defined in the last paragraph. It will be convenient to represent the rigidity morphisms i_X and d_X, by the following pictures:

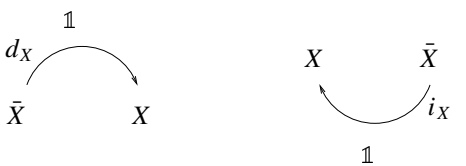


Figure 1. Rigidity morphisms.

In a $*$ tensor category we define the *categorical trace* of an endomorphism $f \in \mathrm{End}(X)$ by

$$\mathrm{Tr}_X(f) = i_X^* \circ (f \otimes 1_{\bar{X}}) \circ i_X \in \mathrm{End}(\mathbb{1}). \tag{2-2}$$

If $Z = \bigoplus m_i X_i$, where X_i is a simple object, and m_i is the multiplicity of X_i in Z, we can write an element $f \in \mathrm{End}(Z)$ in the form $f = \bigoplus f_i$, where $f_i \in \mathrm{End}(m_i X_i)$ can be viewed as an $m_i \times m_i$ matrix. Defining rigidity morphisms i_Z, d_Z with respect to this decomposition, and using (2-1), one checks easily that

$$\mathrm{Tr}_Z(f) = \sum \dim(X_i) \mathrm{Tr}(f_i),$$

where $\mathrm{Tr}(f_i)$ is the usual trace of a matrix. This shows that we obtain a well-defined trace for $\mathrm{End}(Z)$ for any object Z, and that $\mathrm{Tr}_Z(fg) = \mathrm{Tr}_Z(gf)$ for any $f, g \in \mathrm{End}(Z)$. Moreover, using this formula, one shows as well that we can define the trace also by

$$\mathrm{Tr}_X(f) = i_{\bar{X}}^* \circ (1_{\bar{X}} \otimes f) \circ i_{\bar{X}} \in \mathrm{End}(\mathbb{1}).$$

This shows that $*$-categories satisfy the axioms of a spherical category (see [Barrett and Westbury 1999]).

The *normalized trace* tr_X on $\mathrm{End}(X)$ is defined by $\mathrm{tr}_X(f) = \mathrm{Tr}_X(f)/(\dim X)$. In the following we will often just write Tr, tr for the trace or normalized trace when it is clear for which object it is defined.

Conditional expectations can also be very naturally defined using our categorical definitions. Let X be an object. Let $A = \mathrm{End}(X) \cong A \otimes 1_V \subset B = \mathrm{End}(X \otimes V)$. We define the map E_A from B onto A by

$$E_A(b) = \frac{1}{\dim V}(1_X \otimes i_V^*)(b \otimes 1_{\bar{V}})(1_X \otimes i_V);$$

in the tangle picture, $E_A(b)$ is obtained from b by closing up the tangle with color V and renormalizing by $1/\dim V$.

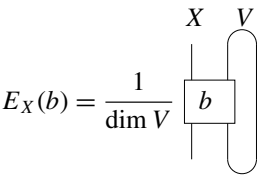


Figure 2. Conditional expectation.

It is known and easy to check that this definition of conditional expectation coincides with the usual definition of conditional expectation in operator algebras (see [Orellana and Wenzl 2002, Proposition 1.4], for instance). Actually, one can show more: Let X_1, X_2, X_3 be objects in our $*$ tensor category $\mathcal{C}$. Define the algebras $A = \mathrm{End}(X_2)$, $B = \mathrm{End}(X_1 \otimes X_2)$, $C = \mathrm{End}(X_2 \otimes X_3)$ and $D = \mathrm{End}(X_1 \otimes X_2 \otimes X_3)$. We can consider all these algebras as subalgebras of D, say by identifying A with $1_{X_1} \otimes \mathrm{End}(X_2) \otimes 1_{X_3}$. The next proposition now follows immediately from the graphical description of the conditional expectations.

Proposition 2.1. *The algebras A, B, C, D form a commuting square*; *that is*, $E_B E_C = E_A = E_C E_B$.

2D. ***Braided tensor categories.*** A strict monoidal category $\mathcal{C}$ is called *braided* if, for any objects X, Y in $\mathcal{C}$, there exists a natural isomorphism $c_{X,Y} : X \otimes Y \to Y \otimes X$ called the *braiding* such that:

$$c_{X,Y\otimes Z} = (1_Y \otimes c_{X,Z})(c_{X,Y} \otimes 1_Z)$$

and

$$c_{X\otimes Y,Z} = (c_{X,Z}\otimes 1_Y)(1_X\otimes c_{Y,Z}).$$

Naturality means that for any morphisms $f : X \to X'$ and $g : Y \to Y'$

$$(g\otimes f)\circ c_{X,Y} = c_{X',Y'}\circ(f\otimes g).$$

Finally, we also require that $c_{\mathbb{1},X} = 1_X = c_{X,\mathbb{1}}$ under the isomorphisms $\mathbb{1}\otimes X \cong X \cong X\otimes\mathbb{1}$.

2E. ***C* tensor categories.*** We call a complex $*$ tensor category a C^* tensor category if

(a) for any objects X, Y in $\mathcal{C}$ the space $\mathrm{Hom}(X, Y)$ is a Hilbert space with inner product $(a, b) = \mathrm{Tr}(b^*a)$ for $a, b \in \mathrm{Hom}(X, Y)$,

(b) for any objects X, Y in $\mathcal{C}$ the algebra $\mathrm{End}(Y)$ is a C^*-algebra acting on the Hilbert space $\mathrm{Hom}(X, Y)$.

Observe that these definitions imply that the dimensions of all objects are positive. A *braided C^* tensor category* is a C^* tensor category with a braiding for which all its braiding morphisms are unitary operators. For examples of C^*-tensor categories, see Section 6A.

3. The multisided construction

3A. ***Categorical setting.*** We shall use the following conventions: Let $\mathcal{C}$ be a finite braided C^* tensor category, where finite means that we only have finitely many equivalence classes of simple objects in $\mathcal{C}$. Let $\{X_\lambda, \lambda \in \Lambda\}$ be a set of representative nonequivalent simple objects, indexed by some labeling set Λ. We define d_λ to be the dimension of X_λ. We shall also assume that the category $\mathcal{C}$ is generated by an object X, so any simple object appears in some tensor power of X. We define algebras $A_n = \mathrm{End}(X^{\otimes n}) = \mathrm{End}_{\mathcal{C}}(X^{\otimes n})$. By the definition of A_n, the simple components of A_n are labeled by the equivalence classes of simple objects which appear in the n-th tensor power of X, i.e., by a certain subset Λ_n of Λ. We define the embeddings $\iota_r : a \in A_n \to a\otimes 1_r \in A_{n+r}$, where we will often omit the subscript r. It follows from the definitions that the vertices of the inclusion diagram for $\iota : A_n \to A_{n+1}$ are labeled by the elements of Λ_n and Λ_{n+1} respectively; the vertex labeled by $\lambda \in \Lambda_n$ is connected with the one labeled by μ by L^μ_λ edges, where L^μ_λ is the multiplicity of the object X_μ in $X_\lambda \otimes X$. We have the commuting diagram of embeddings

$$\begin{array}{ccc} 1_m\otimes A_n & \subset & A_{n+m} \\ {\scriptstyle 1\otimes\iota}\downarrow & & \downarrow{\scriptstyle \iota} \\ 1_m\otimes A_{n+1} & \subset & A_{n+m+1} \end{array} \tag{3-1}$$

We will also assume that the Bratteli diagram for the algebras (A_n) is *strongly connected*. This means that for any X_λ, there exists an r such that $X_\lambda \otimes X^{\otimes r}$ contains all irreducible representations which appear in $X^{\otimes|\lambda|+r}$, where $|\lambda|$ is the smallest integer such that $X_\lambda \in X^{\otimes|\lambda|}$. Equivalently, it means that for any projection $p \in A_n$ there exists an r such that the central support of p in A_{n+r} is 1. We define $k = k(X) = gcd\{n, \mathbb{1} \subset X^{\otimes n}\}$. Let $\mathcal{C}'$ be the subcategory of $\mathcal{C}$ generated by the simple objects in $X^{\otimes mk}$, $m \in \mathbb{N}$.

Lemma 3.1. *Let $\mathcal{C}$ be a finite C^*-tensor category, not necessarily braided. Then we have*

(a) $\Lambda_n = \Lambda_{n+k}$ *for n sufficiently large and* $\Lambda_n \cap \Lambda_m = \varnothing$ *if* $|n-m| < k$; *in particular* $\Lambda' := \Lambda_{nk}$ *for n sufficiently large labels the simple objects of $\mathcal{C}'$.*

(b) *The weight vector for the trace on the algebra A_n is* $\vec{v}_n = (d_\lambda/(\dim X)^n)_{\lambda\in\Lambda_n}$.

(c) *The inductive limit of* $(1_m \otimes A_n \subset A_{n+m})$, *for* $n \to \infty$, *defines an inclusion $B \subset A$ of hyperfinite II_1 factors with index* $(\dim X)^{2m}$.

(d) $\sum_{\lambda\in\Lambda_n} d_\lambda^2 = \frac{1}{k}\sum_{\lambda\in\Lambda} d_\lambda^2$ *for n sufficiently large.*

Proof. If the trivial object $\mathbb{1}$ appears in the r-th tensor power of X and $X_\lambda \subset X^{\otimes n}$, then we have

$$X_\lambda \cong X_\lambda \otimes \mathbb{1} \subset X_\lambda \otimes X^{\otimes r} \subset X^{\otimes n+r}.$$

Hence $\Lambda_n \subset \Lambda_{n+r}$ for all $n \in \mathbb{N}$. As Λ is finite, these inclusions become equalities for n sufficiently large. Applying this to any r such that $\mathbb{1} \subset X^{\otimes r}$, we can similarly prove $\Lambda_n = \Lambda_{n+k}$ for k the gcd of all such r and n sufficiently large. Finally, if $0 < m-n = k' < k$ and $\lambda \in \Lambda_n \cap \Lambda_m$, then we also have $\nu \in \Lambda_{n+r} \cap \Lambda_{m+r} = \Lambda_{n+k'+r}$ for any $X_\nu \subset X_\lambda \otimes X^{\otimes r}$ and $r \in \mathbb{N}$. As the Bratteli diagram for (A_n) is strongly connected, we obtain $\Lambda_{n+r} = \Lambda_{n+r+k'}$ for r sufficiently large. Using the convention $X_0 = \mathbb{1}$, we can find r such that $0 \in \Lambda_{n+r} = \Lambda_{n+r+k'}$, contradicting the definition of k. This shows (a).

Statement (b) follows from the fact that the value of the normalized trace of a projection p_λ corresponding to a simple object $X_\lambda \subset X^{\otimes n}$ is given by $\mathrm{tr}(p_\lambda) = d_\lambda/(\dim X)^n$.

For statement (c) observe that Diagram (3-1) defines a commuting square by Proposition 2.1. Moreover, the sequence of algebras as in the statement has a k-periodic pattern: By part (a), we have the same labeling sets for the algebras in Diagram (3-1) if we substitute n by $n+k$ everywhere, for n sufficiently large. Moreover, also the inclusion pattern remains the same by the discussion before Diagram (3-1). It follows from [Wenzl 1988], Theorem 1.5(iii), that the index $[A : B]$ is given by the ratio $\|\vec{v}_n\|^2/\|\vec{v}_{n+1}\|^2$, for n large enough. As this holds for

any sufficiently large n, we have

$$[A : B]^k = \prod_{i=1}^{k} \frac{\|\vec{v}_{n-1+i}\|^2}{\|\vec{v}_{n+i}\|^2} = \frac{\|\vec{v}_n\|^2}{\|\vec{v}_{n+k}\|^2}.$$

The claim now follows from the fact that $\vec{v}_n = (\dim X)^k \vec{v}_{n+k}$, by (a) and (b). Finally observe that $(\dim X)^2 \|\vec{v}_{n+1}\|^2 = \|\vec{v}_n\|^2$ implies $\sum_{\lambda\in\Lambda_n}(d_\lambda)^2 = \sum_{\mu\in\Lambda_{n+1}}(d_\mu)^2$ for all n sufficiently large. As $\Lambda_n \cap \Lambda_m = \varnothing$ whenever $|n-m| < k$, we obtain Statement (d). □

3B. ***Multisided construction.*** The subfactors constructed in the last section will sometimes be denoted as one-sided subfactors. We will now generalize the construction in [Erlijman 2001] to the setting of braided C^*-tensor categories, which we call multisided subfactors in analogy to the notation in [Erlijman 2001]. We will fix a positive integer s. For the s-sided construction, we will have to define an embedding of algebras $A_n^{\otimes s} \subset A_{ns}$ such that we will obtain a subfactor if we consider the inductive limit over n.

We shall need special braids $\gamma_n \in \mathcal{B}_{sn}$, which can be defined inductively by $\gamma_1 = 1_s$ and by Figure 3.

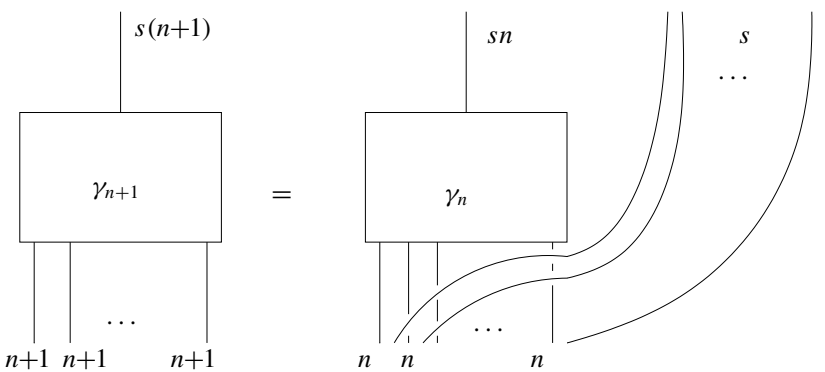


Figure 3. Inductive property of intertwining braids.

Alternatively, the braid γ_n can be described as follows: arrange the points labeled by the numbers 1 up to ns in a rectangular pattern with height n and width s. Now we can numerate the points either by first going down the columns, or by first going to the right in each row. This defines a permutation π mapping the i-th point in the column-first count to the i-th point in the row-first count. The braid γ_n is now defined by this permutation where the i-th lower point is connected with the $\pi(i)$-th upper point and where we assume all crossings to be positive (i.e., the strand going from southwest to northeast crosses over the one going from southeast to northwest). A picture for this braid can be found in [Erlijman 2003, p. 83].

Let $c = c_{X,X}$ be the braiding morphism for X. By definition, we obtain a unitary representation ρ of the braid group $\mathcal{B}_n$ into A_n by mapping the generator σ_i to $c_i = 1_{i-1} \otimes c \otimes 1_{n-1-i}$. We define the unitary $u_n = u_n^{(s)} = \rho(\gamma_n)$, with γ_n defined as in Figure 3. Finally, the embedding from $A_n^{\otimes s}$ into A_{ns} is given by first identifying $A_n^{\otimes s}$ with $\mathrm{End}(X^{\otimes n})^{\otimes s} \subset \mathrm{End}(X^{\otimes ns}) = A_{ns}$ and by then conjugating this with u_n, i.e., by

$$(a_1 \otimes \cdots \otimes a_s) \overset{\hat{u}_n}{\longmapsto} u_n(a_1 \otimes \cdots \otimes a_s)u_n^*;$$

throughout this paper, $\hat{u}$ will denote the inner automorphism given by conjugation via the unitary u unless stated otherwise. We now obtain the following diagram of maps, where the vertical arrows are labeled by $\iota^{\otimes s} = \iota_1^{\otimes s}$ and $\iota = \iota_s$ respectively:

(3-2)
$$\begin{array}{ccc} A_n^{\otimes s} & \xrightarrow{\hat{u}_n} & A_{ns} \\ \downarrow & & \downarrow \\ A_{n+1}^{\otimes s} & \xrightarrow[\hat{u}_{n+1}]{} & A_{(n+1)s} \end{array}$$

Then we have the following lemma which has essentially already been proved in [Erlijman 2001], Section 3.2; the case proved there would correspond to the special case in which A_n is generated by the image of $\mathcal{B}_n$.

Lemma 3.2. *The diagram (3-2) above commutes and also forms a commuting square. Moreover, the inclusion pattern is k-periodic.*

Proof. We check first that Diagram (3-2) is a commuting diagram: This is most easily seen by the following pictures (these proofs by pictures contain all the necessary details and translate faithfully to the algebraic proofs by simply rewriting the definitions already included in this article). We take $s = 3$ for simplicity. For $b \in \mathcal{A}_n^{\otimes s}$, we have:

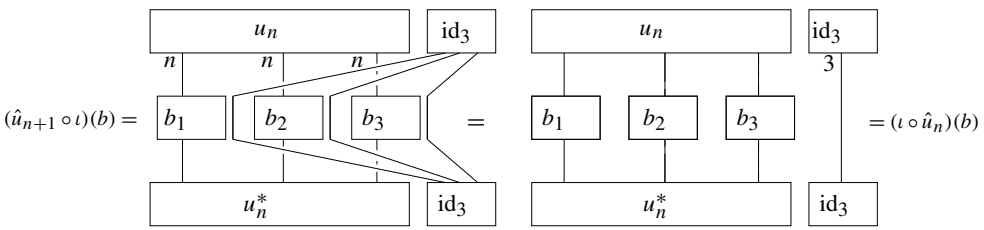


Figure 4. Diagram (3-2) is a commuting diagram.

Now we check that Diagram (3-2) is a commuting square, i.e., that $(E_{A_{sn}} \circ \hat{u}_{n+1})(b) = (\hat{u}_n \circ E_{A_n^{\otimes s}})(b)$ for $b \in A_{n+1}^{\otimes s}$. We use the categorical definition for a conditional expectation as described in Section 2C, Figure 2. For $b = b_1 \otimes \cdots \otimes b_s \in A_{n+1}^{\otimes s}$, we have

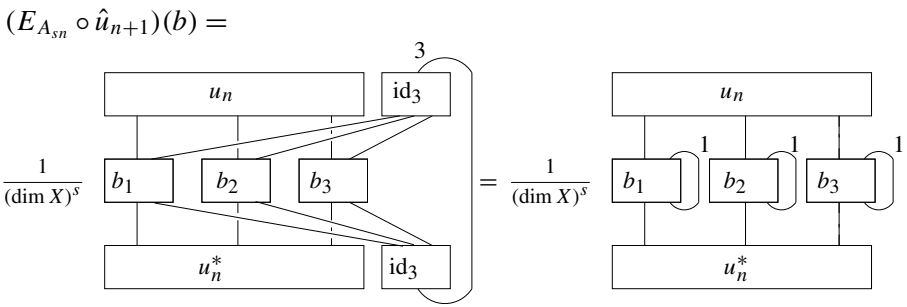


Figure 5. Diagram (3-2) is a commuting square.

which in turn equals $(\hat{u}_n \circ E_{A_n^{\otimes s}})(b)$. To show that the inclusion diagrams are k-periodic for large n, observe that Lemma 3.1(a) implies that we have a one-to-one correspondence between the labeling sets of simple components of $A_n^{\otimes s}$ and $A_{n+k}^{\otimes s}$ as well as between the components of A_{ns} and $A_{(n+1)s}$. This identification of edges is compatible with the number of edges between them, which again is just given by tensor product multiplicities. □

Theorem 3.3. *Fix $s \in \mathbb{N}$, $s > 1$. There is an embedding of the factor $\mathcal{N} := \liminf A_n^{\otimes s}$ (inductive limit) in $\mathcal{M} := \liminf A_{ns}$ given by $\hat{u} := \liminf \hat{u}_n$, with u_n as above. The index of the resulting inclusion is*

$$\left(\sum_{\lambda \in \Lambda'} d_\lambda^2 \right)^{s-1},$$

where Λ' is an indexing set for the simple objects of the subcategory $\mathcal{C}'$ as defined at the beginning of this subsection and $d_\lambda = \dim(X_\lambda)$.

Proof. This was done in [Erlijman 2001] in the case that the A_n's are generated by braid elements only. By Lemma 3.2, Diagram (3-2) is a periodic commuting square for large n. Thus, by [Wenzl 1988], Theorem 1.5(iii), $\hat{u} : \mathcal{N} \hookrightarrow \mathcal{M}$ is an inclusion of hyperfinite II_1 factors with index given by $\|\vec{t}_n\|^2/\|\vec{v}_n\|^2$ for n sufficiently large, where $\vec{t}_n$ and $\vec{v}_n$ are the trace vectors for the trace in $\mathcal{M}$ restricted to the finite-dimensional approximants $A_n^{\otimes s}$ and A_{ns}, respectively. For this observe that if $k|n$ the dimension vectors for $A_n^{\otimes s}$ and A_{ns} are given by $\vec{t}_{ns} = (d_{\vec{\lambda}}/(\dim X)^{ns})_{\vec{\lambda}}$ and $\vec{v}_{ns} = (d_\nu/(\dim X)^{ns})_\nu$, with $\vec{\lambda} \in (\Lambda')^s$ and $\nu \in \Lambda'$; here $d_{\vec{\lambda}} = \prod_{i=1}^s d_{\lambda_i}$. Hence we obtain

$$[\mathcal{M} : \mathcal{N}] = \frac{\|\vec{t}_n\|^2}{\|\vec{v}_n\|^2} = \frac{\sum_{\vec{\lambda} \in (\Lambda')^s} d_{\vec{\lambda}}^2}{\sum_{\nu \in \Lambda'} d_\nu^2} = \left(\sum_{\lambda \in \Lambda'} d_\lambda^2 \right)^{s-1}. \qquad \square$$

3C. ***More embeddings.*** We shall need a variation of the embeddings in the last section for the construction of certain bimodules.

Lemma 3.4. *Let $\vec{m} = (m_1, \dots, m_s)$, where $m_i \in \mathbb{Z}_{\geq 0}$, and $m_1 \geq m_2 \geq \cdots \geq m_s$, and let $|\vec{m}| = \sum_i |m_i|$. Then there exist unitaries $u_{\vec{m},n} = u_{\vec{m},n}(s) \in A_{|\vec{m}|+sn}$ such that we obtain k periodic commuting squares*

$$\begin{array}{ccc} A_{n+m_1} \otimes \cdots \otimes A_{n+m_s} & \xrightarrow{\hat{u}_{\vec{m},n}} & A_{|\vec{m}|+ns} \\ \downarrow & & \downarrow \\ A_{n+1+m_1} \otimes \cdots \otimes A_{n+1+m_s} & \xrightarrow[\hat{u}_{\vec{m},n+1}]{} & A_{|\vec{m}|+(n+1)s} \end{array} \tag{3-3}$$

which produce an inclusion $\hat{u}_{\vec{m}} : \mathcal{N} \to \mathcal{M}$ isomorphic to the map $\hat{u} : \mathcal{N} \to \mathcal{M}$ of Theorem 3.3.

Proof. We shall give diagrammatic representations of the unitaries $u_{\vec{m},n} = u_{\vec{m},n}(s) \in A_{|\vec{m}|+sn}$ as follows. Let $t_{\vec{m}} = t_{\vec{m}}(s)$ be the unitary in $A_{|\vec{m}|}$ given by

$t_{\vec{m}} =$

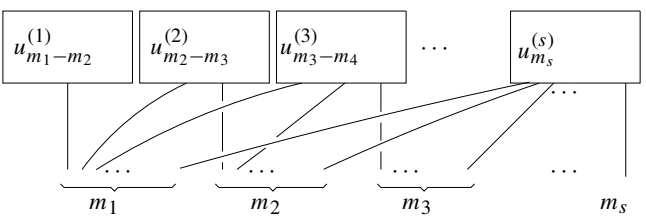


Figure 6

where the unitary $u_r^{(s)}$ is given by Figure 3 for $s > 1$ (with $n+1$ replaced by r) and is equal to id_r for $s = 1$, with any positive integer r.

The unitary $u_{\vec{m},n}$ is then defined from $t_{\vec{m}}$ and $u_n^{(s)}$ as in Figure 7.

$u^{(s)}_{\vec{m},n} =$

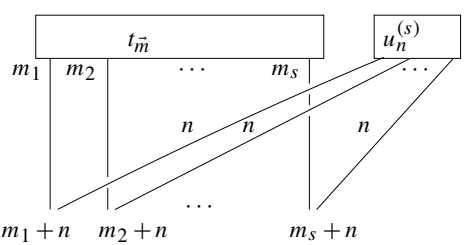


Figure 7

We proceed as in Lemma 3.2 to show that Diagram (3-3) is a commuting square. First we check that our diagram is a commuting diagram; we shall denote the vertical arrows by $\iota^{\otimes s}$ and ι respectively. Assume $s = 3$ again for simplicity. For $b \in A_{n+m_1} \otimes \cdots \otimes A_{n+m_s}$, we have $(\hat{u}_{\vec{m},n+1} \circ \iota^{\otimes s})(b) = = (\iota \circ \hat{u}_{\vec{m},n})(b)$. The commuting square property as well as k periodicity is shown in the same way as in Lemma 3.2.

It remains to show that the subfactor constructed in this lemma is conjugate to the one in Theorem 3.3. We define an automorphism Φ of the factor $\mathcal{M} =$

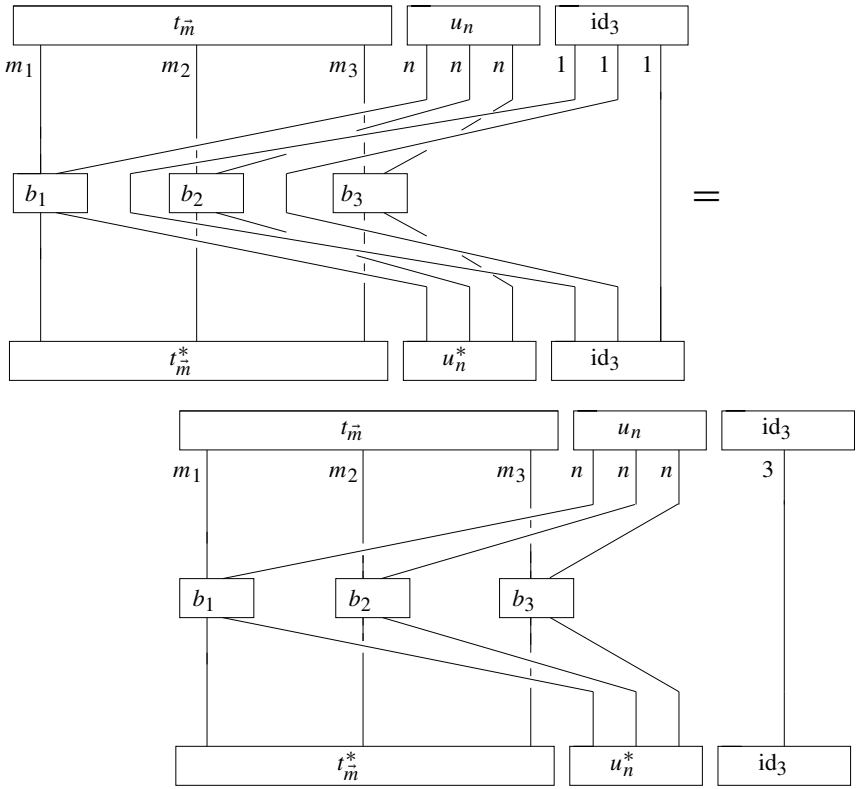


Figure 8. Diagram (3-3) is a commuting diagram.

$\liminf A_{sn+|\vec{m}|} = \liminf A_{s(n+m_1)}$ that will carry the subfactor defined here,

$$\hat{u}_{\vec{m}}(\mathcal{N}) = \liminf u_{\vec{m},n}(A_{n+m_1} \otimes \cdots \otimes A_{n+m_s})u^*_{\vec{m},n},$$

to the subfactor $\hat{u}(\mathcal{N}) = \liminf u_n A_n^{\otimes s} u_n^*$ from Theorem 3.3. Define Φ_n at the finite-dimensional level by

$$\begin{array}{ccc} \mathcal{M} & & \mathcal{M} \\ \cup & \Phi_n & \cup \\ A_{sn+|\vec{m}|} & \longrightarrow & A_{s(n+m_1)} \\ \cup & & \cup \\ a & \longmapsto & u_{n+m_1} b_n \iota(u^*_{\vec{m},n} a u_{\vec{m},n}) b_n^* u^*_{n+m_1}, \end{array}$$

where $b_n \in A_{s(n+m_1)}$ is a unitary described in Figure 9 on the next page, and where $\iota : A_{sn+|\vec{m}|} \to A_{s(n+m_1)}$ is the usual inclusion (recall $m_1 \geq m_i$). Observe that

$$b_n(A_{n+m_1} \otimes \cdots \otimes A_{n+m_s} \otimes 1_{sm_1-|\vec{m}|})b_n^*$$

equals the image of the natural inclusion map $A_{n+m_1} \otimes \cdots \otimes A_{n+m_s} \to A_{n+m_1}^{\otimes s}$.

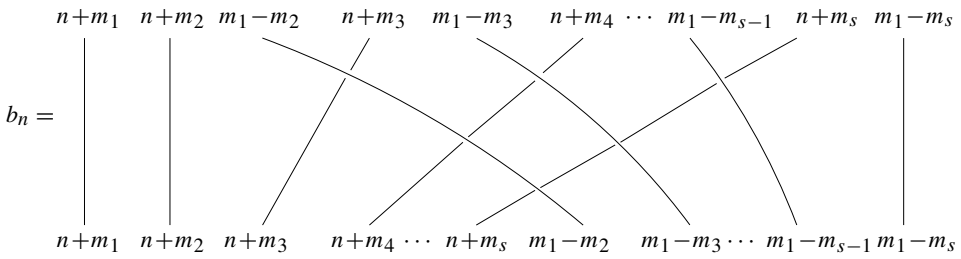


Figure 9. Pictorial description of $b_n \in A_{s(n+m_1)}$.

It is easy to check that the maps Φ_n are compatible with respect to n, and so we can define $\Phi : \mathcal{M} = \liminf A_{sn+|\vec{m}|} \to \mathcal{M} = \liminf A_{s(n+m_1)}$ by $\Phi := \liminf \Phi_n$. We observe that

$$u_{\vec{m},n}(a_1 \otimes \cdots \otimes a_s)u^*_{\vec{m},n} \overset{\Phi}{\mapsto} u_{n+m_1}(a_1 \otimes \cdots \otimes a_s)u^*_{n+m_1}$$

for $a_i \in A_{n+m_i}$, so Φ carries $\hat{u}_{\vec{m}}(\mathcal{N})$ to $\hat{u}(\mathcal{N})$. To check that Φ is an automorphism, one first observes that $\Phi = \liminf \Phi_n = \liminf \left(\hat{u}_{n+m_1} \circ \hat{b}_n \circ \iota \circ \hat{u}^*_{\vec{m},n}\right)$ (the "hat" morphisms denote the adjoint morphisms $\hat{w}(x) := wxw^*$). Then one checks that Φ has left inverse given by $\Phi_l^{-1} := \liminf \left(\hat{u}_{\vec{m},n+m_1} \circ \iota \circ \hat{b}^*_n \circ \hat{u}^*_{n+m_1}\right)$, where $\iota : A_{s(n+m_1)} \to A_{s(n+m_1)+m_1}$ also denotes the canonical inclusion, and a right inverse given by $\Phi_r^{-1} := \liminf \left(\hat{u}_{\vec{m},n} \circ \iota \circ b^*_{n-m_1} \circ \hat{u}^*_n\right)$, where $\iota : A_{sn} \to A_{sn+|\vec{m}|}$ again denotes the canonical inclusion, also observing that in the inductive limit the canonical inclusions turn out to be the identity map. □

3D. ***Endomorphisms.*** We now want to construct bimodules with respect to the just constructed factors $\mathcal{N}$ and $\mathcal{M}$ in the proof of the last theorem. This will be done according to the recipe described in Remark 1.3. To do so, we need to define the endomorphisms mentioned in the braid setting before, in the categorical setting.

Lemma 3.5. *Fix $m_i \in \mathbb{Z}_{\geq 0}$, $i = 1, 2, \ldots, s$, with $m_1 \geq m_2 \geq \cdots \geq m_s$.*

(a) *For $n \in \mathbb{N}$, the maps*

$$\begin{aligned} A_n^{\otimes s} \to \quad & A_{m_1+n} \otimes \cdots \otimes A_{m_s+n} \\ a_1 \otimes \cdots \otimes a_s \mapsto \quad & (1_{m_1} \otimes a_1) \otimes \cdots \otimes (1_{m_s} \otimes a_s) \end{aligned}$$

extend to an endomorphism $\mathrm{Shift}^{\mathcal{N}}_{\vec{m}} : \mathcal{N} \to \mathcal{N}$, *where* $\vec{m} := (m_1, \ldots, m_s)$.

(b) *Let $\hat{u}$ denote the embedding of $\mathcal{N} \hookrightarrow \mathcal{M}$. The endomorphism* $\mathrm{Shift}^{\mathcal{N}}_{\vec{m}}$ *extends to an endomorphism of $\mathcal{M}$, denoted by* $\mathrm{Shift}^{\mathcal{M}}_{\vec{m}}$. *In other words, we have a*

commuting diagram

$$\begin{array}{ccc} \mathcal{N} & \xrightarrow{\hat{u}} & \mathcal{M} \\ {\scriptstyle \mathrm{Shift}_{\vec{m}}^{\mathcal{N}}}\downarrow & & \downarrow{\scriptstyle \mathrm{Shift}_{\vec{m}}^{\mathcal{M}}} \\ \mathcal{N} & \xrightarrow[\hat{u}_{\vec{m}}]{} & \mathcal{M} \end{array}$$

(c) $(\mathrm{Shift}_{\vec{m}}^{\mathcal{M}} \circ \hat{u})$ *only depends on the norm* $|\vec{m}|$ *of* $\vec{m}$, *and it is of the form*

$$A_n^{\otimes s} \to A_{|\vec{m}|+sn}$$
$$(a_1 \otimes \cdots \otimes a_s) \mapsto 1_{|\vec{m}|} \otimes u_n(a_1 \otimes \cdots \otimes a_s)u_n^*.$$

Proof. (a) If $s = 1$, $\mathcal{N} = \mathcal{R} = \liminf A_n$, and we obtain the familiar one-sided shift Shift_m. For $s > 1$, $\mathcal{N} = \mathcal{R} \otimes \cdots \otimes \mathcal{R}$ (s factors) and $\mathrm{Shift}_{\vec{m}} = \mathrm{Shift}_{m_1} \otimes \cdots \otimes \mathrm{Shift}_{m_s}$. The following formalization of these facts will also be useful for the proofs of (b) and (c). Let $v_{\vec{m},n} \in A_{|\vec{m}|+sn}$ be the unitary image under ρ of the braid described by:

$v_{\vec{m},n} =$

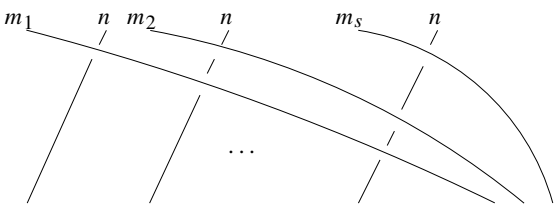


Figure 10

It is easy to see pictorially that for any element $a_1 \otimes \cdots \otimes a_s \in A_n^{\otimes s}$, the maps defined in the statement of (a) are given by

$$(a_1 \otimes \cdots \otimes a_n) \;\mapsto\; v_{\vec{m},n}(a_1 \otimes \cdots \otimes a_n \otimes \mathrm{id}_{|\vec{m}|})v_{\vec{m},n}^* \;\in A_{n+m_1} \otimes \cdots \otimes A_{n+m_s}.$$

That these maps extend to the von Neumann algebra inductive limit $\mathcal{N} = \liminf A_n^{\otimes s}$ follows from the fact that the following are commuting diagrams with respect to the canonical inclusions:

(3-4)
$$\begin{array}{ccccc} A_n^{\otimes s} & \hookrightarrow & A_n^{\otimes s} \otimes A_{|\vec{m}|} & \xrightarrow{\hat{v}_{\vec{m},n}} & A_{n+m_1} \otimes \cdots \otimes A_{n+m_s} \\ \downarrow & & \downarrow & & \downarrow \\ A_{n+1}^{\otimes s} & \hookrightarrow & A_{n+1}^{\otimes s} \otimes A_{|\vec{m}|} & \xrightarrow{\hat{v}_{\vec{m},n+1}} & A_{n+1+m_1} \otimes \cdots \otimes A_{n+1+m_s} \end{array}$$

and from the fact that the maps are norm and trace preserving. We denote the resulting endomorphism by $\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}$.

(b) We shall extend the map $\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}$ to $\mathcal{M}$ after embedding $\mathcal{N}$ in $\mathcal{M}$ via $\hat{u}$ (given by the inductive limit of conjugation of unitaries u_n or $u_{\vec{m},n}$ as in Figures 7 and 3). At

the finite-dimensional level we define $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}} : \liminf A_{sn} \to \liminf A_{|\vec{m}|+sn}$ by

$$\widehat{\omega_n} : \ A_{sn} \ \hookrightarrow \ A_{|\vec{m}|+sn} \ \to \ A_{|\vec{m}|+sn}, \tag{3-5}$$

where the first arrow stands for the standard inclusion $a \in A_{sn} \mapsto a \otimes 1 \in A_{|\vec{m}|+sn}$, and where the second arrow stands for conjugation by the unitary $\omega_n = \omega_n(s, \vec{m}) \in A_{|\vec{m}|+sn}$ defined by

$$\omega_n := u_{\vec{m},n} v_{\vec{m},n} (u_n^* \otimes \mathrm{id}_{|\vec{m}|}); \tag{3-6}$$

here $u_{\vec{m},n}$ and $v_{\vec{m},n}$ are given by Figures 6, 7, and 10. We give a diagrammatic representation for $s = 3$ in Figure 11, with $b \in A_{sn}$:

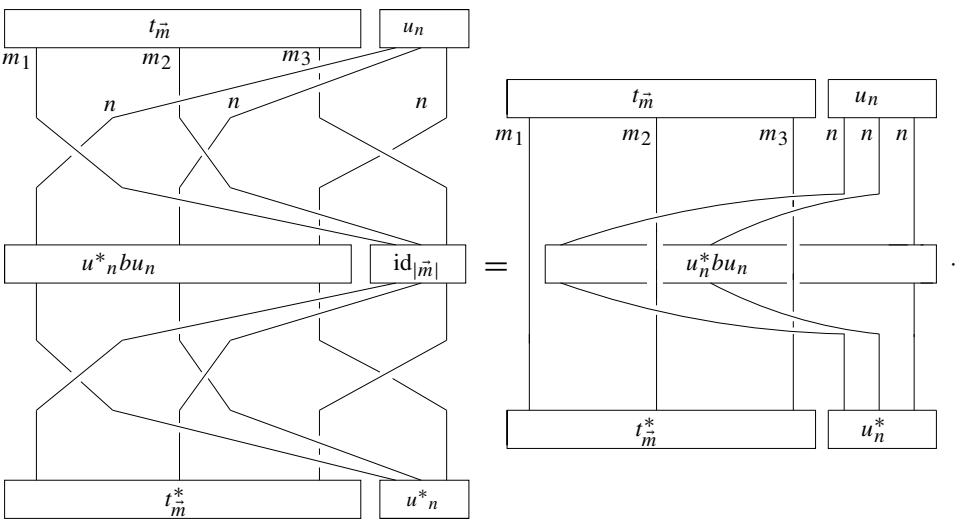


Figure 11. Pictorial representation of $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(b) \in A_{|\vec{m}|+sn}$, for $b \in A_{sn}$ (s=3).

We want to show that these maps extend to a well-defined map $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}$ on the inductive limit $\liminf A_{sn}$, i.e., we have to show that $\widehat{\omega}_{n+1}(\iota(b)) = \iota\big(\widehat{\omega}_n(b)\big)$, where we use the notation ι for the standard inclusions of $A_{sn} \to A_{s(n+1)}$ as well as for $A_{|\vec{m}|+sn} \to A_{|\vec{m}|+s(n+1)}$. To show this, we need the inductive property of the unitaries $u_n^{(s)}$ mentioned already at the braid level, seen in Figure 3, to write $u_{\vec{m},n+1}$ in terms of $u_{\vec{m},n}$ and of id_s. We then have for $b \in A_{sn}$ that $\widehat{\omega}_{n+1}(\iota(b)) = \iota\big(\widehat{\omega}_n(b)\big)$, as shown in Figure 12. Hence $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}} = \liminf \widehat{\omega}_n$ is well defined.

We still need to show that $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}$ extends $\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}$, i.e., that $(\mathrm{Shift}_{\vec{m}}^{\mathcal{M}} \circ \hat{u}) = (\hat{u}_{\vec{m}} \circ \mathrm{Shift}_{\vec{m}}^{\mathcal{N}})$. From the definition, for $a = a_1 \otimes \cdots \otimes a_s \in A_n^{\otimes s}$,

$$\begin{aligned}(\mathrm{Shift}_{\vec{m}}^{\mathcal{M}} \circ \liminf \hat{u}_n)(a) &= (\hat{\omega}_n \circ \iota \circ \hat{u}_n)(a) = (\hat{u}_{\vec{m},n} \circ \hat{v}_{\vec{m},n})(a \otimes \mathrm{id}_{|\vec{m}|}) \\ &= (\liminf \hat{u}_{\vec{m},n} \circ \mathrm{Shift}_{\vec{m}}^{\mathcal{N}})(a).\end{aligned}$$

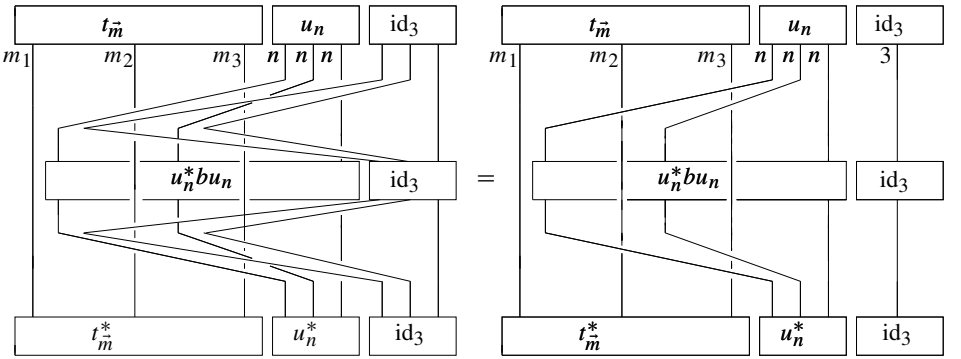
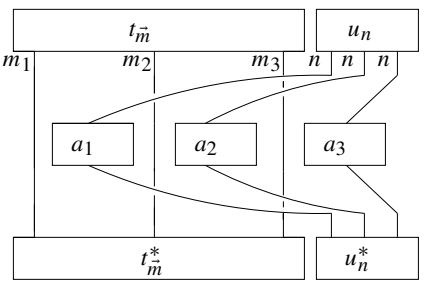


Figure 12. $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}$ is well-defined.

(c) This follows from the definition. Take $(a_1 \otimes \cdots \otimes a_s) \in A_n^{\otimes s}$. Using Figure 11, we obtain that $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(u_n(a_1 \otimes a_2 \otimes a_3)u_n^*)$ equals



and this in turn equals $1_{|\vec{m}|} \otimes u_n(a_1 \otimes a_2 \otimes a_3)u_n^*$. □

Proposition 3.6. *Let* $\mathrm{Shift}_{\vec{m}}$ *be as in Lemma 3.5.*

(a) $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(\mathcal{M}) \subset \mathcal{M}$ *is an inclusion of* II_1 *factors with index* $(\dim(X))^{2|\vec{m}|}$, *where* $|\vec{m}| = \sum m_i$ *and* $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(\mathcal{M})' \cap \mathcal{M}$ *has a subalgebra isomorphic to*

$$A_{m_1} \otimes \cdots \otimes A_{m_s}.$$

(b) $(\hat{u}_{\vec{m}} \circ \mathrm{Shift}_{\vec{m}}^{\mathcal{N}})(\mathcal{N}) \subset \mathcal{M}$ *is an inclusion of* II_1 *factors with index*

$$[\mathcal{M} : \mathcal{N}](\dim(X))^{2|\vec{m}|}$$

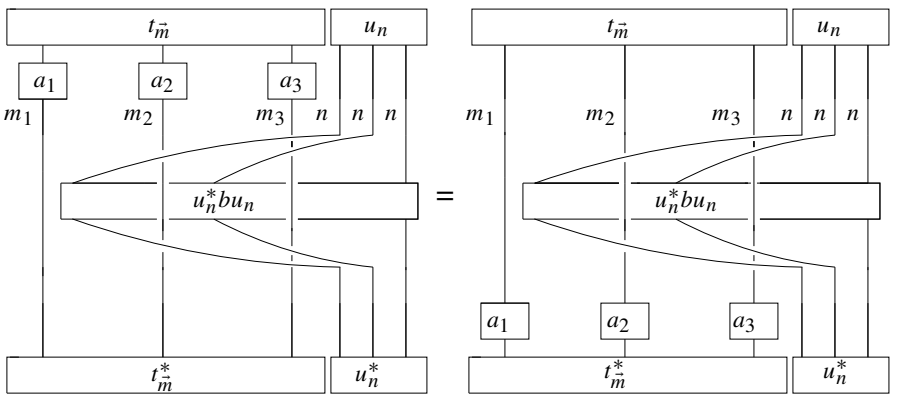
and relative commutant $(\hat{u}_{\vec{m}} \circ \mathrm{Shift}_{\vec{m}}^{\mathcal{N}})(\mathcal{N})' \cap \mathcal{M} \cong A_{|\vec{m}|}$.

(c) $\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}(\mathcal{N}) \subset \mathcal{N}$ *is an inclusion of* II_1 *factors with index* $(\dim(X))^{2|\vec{m}|}$ *and relative commutant* $\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}(\mathcal{N})' \cap \mathcal{N} \cong A_{m_1} \otimes \cdots \otimes A_{m_s}$.

Proof. For (a), we first show that the maps $\hat{\omega}_n$ in (3-5) define periodic commuting squares with respect to n (which generate $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(\mathcal{M}) \subset \mathcal{M}$ by definition). For this, one simply uses the fact that these maps are compositions involving the maps $\hat{v}_{\vec{m},n}$,

$\hat{u}_{\vec{m},n}$ and $\hat{u}_n$ (see (3-6)). The desired diagrams are then built from compositions of the periodic commuting squares in diagrams (3-4), (3-2) and (3-3); see Lemma 3.2 and Lemma 3.4. Hence the desired diagrams are commuting squares. Periodicity is shown as in Lemma 3.2, and we can use the formula for the index, as done there. It follows from parts (b) and (d) of Lemma 3.1 that the ratio of the square lengths of the weight vectors for A_{sn} and $A_{sn+|\vec{m}|}$ is equal to $(\dim X)^{2|\vec{m}|}$.

The statement on the relative commutant follows from the definition of $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}$. Let us represent $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(b)$, for $b \in A_{sn}$ ($s = 3$ to make things simpler) as it appears in Figure 11. Then for $a \in (t_{\vec{m}} \otimes 1_{sn})(A_{m_1} \otimes \cdots \otimes A_{m_s} \otimes 1_{sn})(t_{\vec{m}}^* \otimes 1_{sn}) \in A_{|\vec{m}|+sn}$ we have $a\,\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(b) = \mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(b)a$, as follows from this figure representing the two sides:



Hence, $(t_{\vec{m}} \otimes 1_{sn})(A_{m_1} \otimes \cdots \otimes A_{m_s} \otimes 1_{sn})(t_{\vec{m}}^* \otimes 1_{sn}) \cong A_{m_1} \otimes \cdots \otimes A_{m_s}$ commutes with $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(b)$ for $b \in A_{sn}$, for every n, so that $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(\mathcal{M})' \cap \mathcal{M}$ has a subalgebra isomorphic to $A_{m_1} \otimes \cdots \otimes A_{m_s}$. This proves the last statement of (a).

For (b), one observes that the generating square for $(\hat{u}_{\vec{m}} \circ \mathrm{Shift}_{\vec{m}}^{\mathcal{N}})(\mathcal{N}) \subset \mathcal{M}$ is obtained from the double-square given in (3-4) of proof of Lemma 3.5(a) (which defines the endomorphism $\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}$), composed with the square given in (3-3) (which defines the inclusion $\hat{u}_{\vec{m}} : \mathcal{N} \to \mathcal{M}$). These squares are commuting squares (the one in (3-4) because it involves maps that are trace preserving, and the one in (3-3) was shown in Lemma 3.4). So their composition, which generates $(\hat{u}_{\vec{m}} \circ \mathrm{Shift}_{\vec{m}}^{\mathcal{N}})(\mathcal{N}) \subset \mathcal{M}$, gives also a commuting square. The indices for parts (b) and (c) can now be computed as before, using Lemma 3.1. It only remains to show the statement about the relative commutant.

Lemma 3.5(c) implies that $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(u_n A_n^{\otimes s} u_n^*) = 1_{|\vec{m}|} \otimes u_n A_n^{\otimes s} u_n^*$ for every n. So $A_{|\vec{m}|} \otimes 1_{sn}$ commutes with $\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(u_n A_n^{\otimes s} u_n^*)$ for every n and $(\mathrm{Shift}_{\vec{m}}^{\mathcal{M}} \circ \hat{u})(\mathcal{N})' \cap \mathcal{M}$ $(= (\hat{u}_{\vec{m}} \circ \mathrm{Shift}_{\vec{m}}^{\mathcal{N}})(\mathcal{N})' \cap \mathcal{M})$ has a subalgebra isomorphic to $A_{|\vec{m}|}$. Conversely, for the other inclusion, we apply a dimension upper bound result for relative commutants of inclusions generated by periodic commuting squares (see [Wenzl 1988],

Theorem 1.6):

$$\dim\left((\mathrm{Shift}_{\vec{m}}^{\mathcal{M}}\circ\hat{u})(\mathcal{N})'\cap\mathcal{M}\right)\le\dim\left((1_{|\vec{m}|}\otimes u_nA_n^{\otimes s}u_n^*)'_p\cap(A_{|\vec{m}|+sn})_p\right)$$
$$\le\dim(A_{|\vec{m}|+sn})_p,$$

for any projection $p\in 1_{|\vec{m}|}\otimes u_nA_n^{\otimes s}u_n^*$, and n large. If n is divisible by k and sufficiently large, then $X^{\otimes n}$ contains a subobject isomorphic to $\mathbb{1}$; let $p_{\mathbb{1}}\in A_n$ be the projection onto it. If $p=1_{|\vec{m}|}\otimes u_n(p_{\mathbb{1}}^{\otimes s})u_n^*\in A_{|\vec{m}|+ns}$, then we have $pA_{|\vec{m}|+ns}p\cong A_{|\vec{m}|}$. This shows (b).

For (c), it is even easier than in (a) to show that the generating Diagram (3-4) for $\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}(\mathcal{N})\subset\mathcal{N}$ is a periodic commuting square; one can see that pictorially, as it was done in Lemmas 3.2 and 3.4, which is left to the reader. The statement about the relative commutant in (c) is proved in the same manner as in (b): By definition, $\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}(a_1\otimes\cdots\otimes a_s)=(1_{m_1}\otimes a_1)\otimes\cdots\otimes(1_{m_s}\otimes a_s)$. Thus, $(A_{m_1}\otimes 1_n)\otimes\cdots\otimes(A_{m_s}\otimes 1_n)$ commutes with $\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}(A_n^{\otimes s})$ for every n, and so $\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}(\mathcal{N})'\cap\mathcal{N}$ has a subalgebra isomorphic to $A_{m_1}\otimes\cdots\otimes A_{m_s}$. For the other inclusion we apply again the upper bound result for the dimension of the relative commutant:

$$\dim\left(\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}(\mathcal{N})'\cap\mathcal{N}\right)$$
$$\le\dim\left((1_{m_1}\otimes A_n)\otimes\cdots\otimes(1_{m_s}\otimes A_n)\right)'_p\cap(A_{n+m_1}\otimes\cdots\otimes A_{n+m_s})_p$$
$$\le\dim(A_{n+m_1}\otimes\cdots\otimes A_{n+m_s})_p,$$

for any projection $p\in(1_{m_1}\otimes A_n)\otimes\cdots\otimes(1_{m_s}\otimes A_n)$. One shows as in (b) that for $p=(1_{m_1}\otimes p_{\mathbb{1}})\otimes\cdots\otimes(1_{m_s}\otimes p_{\mathbb{1}})$ we have $(A_{n+m_1}\otimes\cdots\otimes A_{n+m_s})_p\cong A_{m_1}\otimes\cdots\otimes A_{m_s}$, from which one deduces (c). □

4. Bimodules and the principal graph

4A. ***Examples of bimodules.*** We are going to construct systems of bimodules in order to calculate the principal and the dual principal graph, as described in Proposition 1.8. This will be done using the endomorphisms Shift defined in the last section.

The $\mathcal{N}$–$\mathcal{N}$-bimodules. Let $\lambda_i\in\Lambda$ and let A_{m_i,λ_i} be the simple component of A_{m_i} corresponding to the simple object $X_{\lambda_i}\subset X^{\otimes m_i}$ with m_i being large multiples of k for $i=1,2,\ldots,s$. We first fix minimal projections $p_{\lambda_i}\in A_{m_i,\lambda_i}$. Define $p_{\vec{\lambda}}=p_{\lambda_1}\otimes\cdots\otimes p_{\lambda_s}$, where $\vec{\lambda}=(\lambda_1,\ldots,\lambda_s)$. The underlying Hilbert space will be given by

$$L^2(\mathcal{N},\mathrm{tr})p_{\vec{\lambda}}:=\{\zeta p_{\vec{\lambda}},\ \zeta\in L^2(\mathcal{N},\mathrm{tr})\}.$$

The $\mathcal{N}$–$\mathcal{N}$ bimodule structure is defined by

$$x.\xi.y=x\xi\mathrm{Shift}_{\vec{m}}^{\mathcal{N}}(y),\quad\text{for } x,\ y\in\mathcal{N},\ \xi\in L^2(\mathcal{N},\mathrm{tr})p_{\vec{\lambda}},$$

where we use the usual right and left multiplication in $\mathcal{N}$ on the right hand side. It follows from Proposition 3.6 that this indeed defines an $\mathcal{N}$–$\mathcal{N}$ bimodule structure on $L^2(\mathcal{N}, \mathrm{tr})p_{\vec{\lambda}}$.

Definition 4.1. The $\mathcal{N}$–$\mathcal{N}$ bimodules defined above will be denoted by $N_{\vec{\lambda},\vec{m}}$.

The $\mathcal{M}$–$\mathcal{N}$-bimodules. Again let $\vec{m} := (m_1, \ldots, m_s) \in \mathbb{N}^s$, with $m := m_1 + \cdots + m_s$. We fix a minimal projection $p_\mu \in (\mathrm{Shift}_{\vec{m}}^{\mathcal{M}} \circ \hat{u})(\mathcal{N})' \cap \mathcal{M} \cong A_m$ (see Proposition 3.6) belonging to the simple component of A_m labeled $\mu \in \Lambda$. The underlying Hilbert space for all these bimodules will be given by

$$L^2(\mathcal{M}, \mathrm{tr})p_\mu := \{\zeta p_\mu \ / \ \zeta \in L^2(\mathcal{M}, \mathrm{tr})\}.$$

The $\mathcal{M}$–$\mathcal{N}$ bimodule structure is defined by

$$x.\xi.y = x\xi(\mathrm{Shift}_{\vec{m}}^{\mathcal{M}} \circ \hat{u})(y), \quad \text{for } x \in \mathcal{M},\ y \in \mathcal{N},\ \xi \in L^2(\mathcal{M}, \mathrm{tr})p_\mu.$$

Definition 4.2. The $\mathcal{M}$–$\mathcal{N}$-bimodules defined above will be denoted by $H_{\mu,\vec{m}}$.

The $\mathcal{N}$–$\mathcal{M}$-bimodules. With notations as in the last definition, we define similarly $\mathcal{N}$–$\mathcal{M}$-bimodules based on Hilbert spaces $p_\mu L^2(\mathcal{M}, \mathrm{tr}) := \{p_\mu\zeta \ / \ \zeta \in L^2(\mathcal{M}, \mathrm{tr})\}$, and with the $\mathcal{N}$–$\mathcal{M}$ bimodule structure defined by

$$x.\xi.y = (\mathrm{Shift}_{\vec{m}}^{\mathcal{M}} \circ \hat{u})(x)\xi y, \quad \text{for } x \in \mathcal{N},\ y \in \mathcal{M},\ \xi \in p_\mu L^2(\mathcal{M}, \mathrm{tr}).$$

Definition 4.3. The $\mathcal{N}$–$\mathcal{M}$-bimodules defined above will be denoted by $K_{\mu,\vec{m}}$.

The $\mathcal{M}$–$\mathcal{M}$-bimodules. Similarly as for the $\mathcal{N}$–$\mathcal{N}$-bimodules, we fix minimal projections $p_{\lambda_i} \in A_{m_i,\lambda_i}$, with $\lambda_i \in \Lambda$, but now only requiring that $\sum m_i$ being divisible by k. The underlying Hilbert space for all these bimodules will be given by

$$p_{\vec{\lambda}} L^2(\mathcal{M}, \mathrm{tr}) := \{p_{\vec{\lambda}}\zeta / \ \zeta \in L^2(\mathcal{M}, \mathrm{tr})\}.$$

The $\mathcal{M}$–$\mathcal{M}$ bimodule structure is defined by

$$x.\xi.y = \mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(x)\xi y, \quad \text{for } x, y \in \mathcal{M},\ \xi \in p_{\vec{\lambda}} L^2(\mathcal{M}, \mathrm{tr}),$$

Definition 4.4. The $\mathcal{M}$–$\mathcal{M}$-bimodules defined above will be denoted by $M_{\vec{\lambda},\vec{m}}$.

Lemma 4.5. *Let the notation be as above.*

(a) *If we view both $N_{\vec{\lambda},\vec{m}}$ and $H_{\nu,\vec{m}}$ as left $\mathcal{N}$-modules, then*

$$\dim_{\mathcal{N}} N_{\vec{\lambda},\vec{m}} = d_{\vec{\lambda}}/(\dim X)^{|\vec{m}|} \quad \text{and} \dim_{\mathcal{N}} H_{\nu,\vec{m}} = d_\nu[\mathcal{M} : \mathcal{N}]/(\dim X)^{|\vec{m}|}.$$

Moreover, we have $\mathrm{ind}(N_{\vec{\lambda},\vec{m}}) = d_{\vec{\lambda}}^2 = \mathrm{ind}(M_{\vec{\lambda},\vec{m}})$, *where* $d_{\vec{\lambda}} = \prod d_{\lambda_i}$, *and* $\mathrm{ind}(H_{\nu,\vec{m}}) = d_\nu^2[\mathcal{M} : \mathcal{N}]$.

(b) *If $|\vec{m}| = |\vec{k}|$, then $H_{\mu,\vec{m}} \cong H_{\mu,\vec{k}}$ as $\mathcal{M}$–$\mathcal{N}$-bimodules, and $K_{\mu,\vec{m}} \cong K_{\mu,\vec{k}}$ as $\mathcal{N}$–$\mathcal{M}$-bimodules.*

(c) *If* $|\vec{m}| = |\vec{k}|$, *we have*

$$\mathrm{Hom}_{\mathcal{M}-\mathcal{M}}(M_{\vec{\lambda},\vec{m}}, M_{\vec{\mu},\vec{k}}) \subset \mathrm{Hom}_{\mathcal{N}-\mathcal{M}}(M_{\vec{\lambda},\vec{m}}, M_{\vec{\mu},\vec{k}}) \cong \mathrm{Hom}_{\mathcal{C}}(X_{\vec{\lambda}}, X_{\vec{\mu}}),$$

where $X_{\vec{\lambda}} = \otimes_{i=1}^{s} X_{\lambda_i}$ *and* $X_{\vec{\mu}} = \otimes_{i=1}^{s} X_{\mu_i}$.

Proof. It is well known (see [Jones 1983], for instance) that $\dim_{\mathcal{N}} L^2(\mathcal{N}, \mathrm{tr})p = \mathrm{tr}(p)$ and $\dim_{\mathcal{N}} L^2(\mathcal{M}, \mathrm{tr})q = \mathrm{tr}(q)[\mathcal{M} : \mathcal{N}]$ for any projections $p \in \mathcal{N}$, $q \in \mathcal{M}$. The dimension statements in (a) follow. For the index statements in (a), let ℓ and r denote left and right multiplication by $\mathcal{N}$ on $L^2(\mathcal{N}, \mathrm{tr})$ or suitable submodules of it. Observe that $\ell(\mathcal{N})'_{|L^2(\mathcal{N},\mathrm{tr})p}$ is equal to $r(p\mathcal{N}p)$ for any $p \in \mathrm{Shift}_{\vec{m}}(\mathcal{N})' \cap \mathcal{N}$. Recall that $\mathrm{Shift}_{\vec{m}}(\mathcal{N}) \subset \mathcal{N}$ has index $(\dim X)^{2|\vec{m}|}$. Moreover, $\mathrm{tr}(p_{\vec{\lambda}}) = d_{\vec{\lambda}}/(\dim X)^{|\vec{m}|}$ for a minimal idempotent $p_{\vec{\lambda}} \in \mathrm{Shift}_{\vec{m}}(\mathcal{N})' \cap \mathcal{N}$; see Proposition 3.6. Using the formula for local indices [Wenzl 1988, Theorem 1.5(iii)] and the index formula in Proposition 3.6(c), we obtain

$$\mathrm{ind}(N_{\vec{\lambda},\vec{m}}) = [p_{\vec{\lambda}}\mathcal{N}p_{\vec{\lambda}} : p_{\vec{\lambda}}\mathrm{Shift}_{\vec{m}}(\mathcal{N})] = \mathrm{tr}(p_{\vec{\lambda}})^2(\dim X)^{2|\vec{m}|} = (d_{\vec{\lambda}})^2.$$

The indices for $H_{\nu,\vec{m}}$ and $M_{\vec{\lambda},\vec{m}}$ are computed similarly. By Lemma 3.5, (c), we have $\mathrm{Shift}_{\vec{m}}^{\mathcal{N}} = \mathrm{Shift}_{\vec{k}}^{\mathcal{N}}$, from which (b) follows.

Let ${}_{\vec{m}}L^2(\mathcal{M}, \mathrm{tr})$ be the Hilbert space $L^2(\mathcal{M}, \mathrm{tr})$ with $\mathcal{M}$–$\mathcal{M}$ bimodule structure $x.\xi.y = \mathrm{Shift}_{\vec{m}}^{\mathcal{M}}(x)\xi y$ for $x, y \in \mathcal{M}$ and $\xi \in L^2(\mathcal{M}, \mathrm{tr})$. Define ${}_{\vec{k}}L^2(\mathcal{M}, \mathrm{tr})$ similarly. These bimodules are isomorphic as $\mathcal{N}$–$\mathcal{M}$ bimodules, again by Lemma 3.5(c). This, combined with Lemma 3.5(b), results in

$$\mathrm{Hom}_{\mathcal{M}-\mathcal{M}}({}_{\vec{m}}L^2(\mathcal{M}, \mathrm{tr}), {}_{\vec{k}}L^2(\mathcal{M}, \mathrm{tr})) \subset \mathrm{Hom}_{\mathcal{N}-\mathcal{M}}({}_{\vec{m}}L^2(\mathcal{M}, \mathrm{tr}), {}_{\vec{k}}L^2(\mathcal{M}, \mathrm{tr})) \cong$$

$$\cong \mathrm{End}_{\mathcal{N}-\mathcal{M}}({}_{\vec{m}}L^2(\mathcal{M}, \mathrm{tr})) \cong A_{|\vec{m}|} = \mathrm{End}_{\mathcal{C}}(X^{\otimes|\vec{m}|}), \qquad (*)$$

where the second isomorphism follows from Proposition 3.6(b), and (b). By construction, we have $M_{\vec{\lambda},\vec{m}} = p_{\vec{\lambda}}({}_{\vec{m}}L^2(\mathcal{M}, \mathrm{tr}))$ and $M_{\vec{\mu},\vec{k}} = p_{\vec{\mu}}({}_{\vec{k}}L^2(\mathcal{M}, \mathrm{tr}))$, where $p_{\vec{\lambda}} = p_{\lambda_1} \otimes \cdots \otimes p_{\lambda_s}$ and $p_{\vec{\mu}} = p_{\mu_1} \otimes \cdots \otimes p_{\mu_s}$. Hence we can interpret an element $f \in \mathrm{Hom}_{\mathcal{M}-\mathcal{M}}(M_{\vec{\lambda}}, M_{\vec{\mu}})$ as an element in $\mathrm{Hom}_{\mathcal{N}-\mathcal{M}}({}_{\vec{m}}L^2(\mathcal{M}, \mathrm{tr}), {}_{\vec{k}}L^2(\mathcal{M}, \mathrm{tr}))$ which satisfies $p_{\vec{\mu}} f p_{\vec{\lambda}} = f$. Using this together with $(*)$ proves claim (c). □

4B. ***Principal graph.*** Let $\vec{\lambda} = (\lambda_1, \ldots, \lambda_s) \in (\Lambda')^s$, and let $L_{\vec{\lambda}}^{\nu}$ be the multiplicity of the object X_ν in $\otimes X_{\lambda_i}$. Observe that $L_{\vec{\lambda}}^{\nu}$ is also equal to the rank of the projection $\bigotimes p_{\lambda_i}$ in the simple component of $A_{|\vec{\lambda}|}$ labeled by ν.

In the following we will fix a vector $\vec{m} = (m_i)$ where all its coordinates are divisible by k, and with m_i large enough that all simple objects of $\mathcal{C}'$ will appear in $X^{\otimes m_i}$ for $i = 1, \ldots, s$. We shall hence omit $\vec{m}$ in the indices of the bimodules and will just write $N_{\vec{\lambda}}$ and K_ν for $N_{\vec{\lambda},\vec{m}}$ and $K_{\nu,\vec{m}}$, respectively.

Theorem 4.6. *With the notation as above*:

(a) *The bimodules* $N_{\vec{\lambda}}$ *and* H_ν *defined above are irreducible.*

(b) *The principal graph for $\mathcal{N} \subset \mathcal{M}$ is the connected component of the fusion graph from $(\mathcal{C}')^s$ to $\mathcal{C}'$ which contains the trivial object of $\mathcal{C}$. Recall that the even vertices of the fusion graph are labeled by s-tuples of elements of Λ', the odd vertices are labeled by the elements of Λ', and the vertex labeled by $\vec{\lambda} = (\lambda_1, \ldots, \lambda_s)$ is connected with the vertex labeled by ν by $L_{\vec{\lambda}}^{\nu}$ edges.*

(c) *The subfactor $\mathcal{N} \subset \mathcal{M}$ has finite depth.*

Proof. Statement (a) follows from Proposition 3.6. For (b), it suffices to calculate the principal graph for the isomorphic inclusion $\mathcal{N} \subset \mathcal{M}$ given by $\hat{u}_{\vec{m}}$ (see Lemma 3.4). In this case the $\mathcal{M}$–$\mathcal{N}$ bimodule structure of $L^2(\mathcal{M}, \mathrm{tr})$ is given by $x.\xi.y = x\xi\hat{u}_{\vec{m}}(y)$. It follows from the definitions that $L^2(\mathcal{M}, \mathrm{tr}) \otimes_{\mathcal{N}} N_{\vec{\lambda}} \cong L^2(\mathcal{M}, \mathrm{tr}) p_{\vec{\lambda}} \cong \oplus L_{\vec{\lambda}}^{\nu} H_{\nu}$; the decomposition of $L^2(\mathcal{M}, \mathrm{tr}) p_{\vec{\lambda}}$ into irreducible $\mathcal{M}$–$\mathcal{N}$ bimodules follows from Proposition 3.6(b) and the remarks at the beginning of this subsection. Hence our system of bimodules $(N_{\vec{\lambda}})_{\vec{\lambda} \in (\Lambda')^s}$ and $(H_{\nu})_{\nu \in \Lambda'}$ is closed under induction. To prove closedness under restriction, observe that the multiplicity of the $\mathcal{N}$–$\mathcal{N}$ bimodule $N_{\vec{\lambda}}$ in the $\mathcal{M}$–$\mathcal{N}$-bimodule H_{ν}, viewed as an $\mathcal{N}$–$\mathcal{N}$-bimodule, is equal to $L_{\vec{\lambda}}^{\nu}$, by Frobenius reciprocity. To show that $H_{\nu} \cong \bigoplus_{\vec{\lambda}} L_{\vec{\lambda}}^{\nu} N_{\vec{\lambda}}$ as an $\mathcal{N}$–$\mathcal{N}$-bimodule, it suffices to prove that both sides have the same dimension, i.e., by Lemma 4.5(a), that

$$[\mathcal{M} : \mathcal{N}] d_{\nu} = \sum_{\vec{\lambda}} L_{\vec{\lambda}}^{\nu} d_{\vec{\lambda}}. \tag{4-1}$$

For this observe that the dimension vectors for $A_n^{\otimes s}$ and A_{ns}, with n a multiple of k, are given by $\vec{t}_{ns} = (d_{\vec{\lambda}}/(\dim X)^{ns})_{\vec{\lambda}}$ and $\vec{v}_{ns} = (d_{\nu}/(\dim X)^{ns})_{\nu}$, with $\vec{\lambda} \in (\Lambda')^s$ and $\nu \in \Lambda'$. Observe that the subfactor $\mathcal{N} \subset \mathcal{M}$ is generated by the periodic sequence $(A_n^{\otimes s} \subset A_{ns})$, with the inclusion matrix for $A_n^{\otimes s} \subset A_{ns}$ given by $G = (L_{\vec{\lambda}}^{\nu})$ with $\vec{\lambda}$ and ν as above, provided $k|n$. Hence it follows from [Wenzl 1988], Theorem 1.5(ii), that $G\vec{v}_{ns} = [\mathcal{M} : \mathcal{N}]\vec{t}_{ns}$. This implies (4-1). Finally, if we choose $\vec{\lambda} = (\mathbb{1}, \mathbb{1}, \ldots, \mathbb{1})$ (s times), where $\mathbb{1}$ stands for the trivial object of $\mathcal{C}$, $\mathrm{ind}(N_{\vec{\lambda}}) = 1$ and hence $N_{\vec{\lambda}}$ is weakly reduced from the trivial $\mathcal{N}$–$\mathcal{N}$ bimodule by Proposition 1.8(c). This shows (b), by Proposition 1.8(a). Statement (c) is a consequence of (b). □

Remark 4.7. The fusion graph from $(\mathcal{C}')^s$ to $\mathcal{C}'$ may not be connected. An easy example is obtained for $\mathcal{C}$ being the representation category of a finite abelian group G, where it decomposes into $|G|$ connected components.

5. Dual principal graph

5A. ***Ring lemma.*** The precise structure of $\mathrm{Shift}_{\vec{m}}(\mathcal{M})' \cap \mathcal{M}$ is still open after Proposition 3.6. To say more about this, we need the following lemma. Similar techniques have appeared before in topological quantum field theory, and within subfactors in work of Ocneanu and others; see [Evans and Kawahigashi 1998; Müger

2003], for example. Dual principal graphs in a similar setting (corresponding to the case $s = 2$) have also been calculated in [Izumi 2000] by somewhat different techniques.

Lemma 5.1. *If $a \in \mathrm{Shift}_{\vec{m}}(\mathcal{M})' \cap \mathcal{M}$, take $\tilde{a} := t_{\vec{m}}^* a t_{\vec{m}}$ with $t_{\vec{m}} \in A_{|\vec{m}|}$ as in Figure 6. Then, for $r = 2, \ldots, s$, we have*

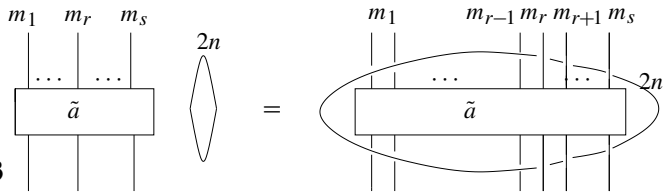


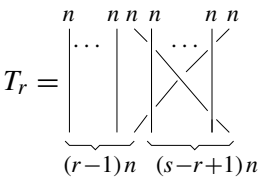
Figure 13

(*the picture is the translation of an algebraic expression of the form $\iota_X^r(\tilde{a} \otimes 1_{sn})\iota_X^{r*} = \iota_X^r x_r(\tilde{a} \otimes 1_{sn})x_r^*\iota_X^{r*}$ for certain morphisms x_r and ι_X^r defined below*).

Proof. By Proposition 3.6(b) and our definition of the inductive limit, we have $\mathrm{Shift}_{\vec{m}}(\mathcal{M})' \cap \mathcal{M} \cap A_{|\vec{m}|+ns} \subset A_{|\vec{m}|} \otimes 1_{ns}$. Take $t_{\vec{m}} \in A_{|\vec{m}|}$ as in Figure 6 in Lemma 3.4. If $a \in \mathrm{Shift}_{\vec{m}}(\mathcal{M})' \cap \mathcal{M}$ then set

$$\tilde{a} \otimes 1_{sn} := (t_{\vec{m}}^* \otimes u_n^*)a(t_{\vec{m}} \otimes u_n) = t_{\vec{m}}^* a t_{\vec{m}} \otimes 1_{sn} \in A_{|\vec{m}|} \otimes 1_{sn},$$

and note that $\tilde{a} \otimes 1_{sn} \in \big((t_{\vec{m}}^* \otimes u_n^*)\mathrm{Shift}_{\vec{m}}(\mathcal{M})(t_{\vec{m}} \otimes u_n)\big)' \cap \mathcal{M}$. In particular, take the element $x_r := (t_{\vec{m}}^* \otimes u_n^*)\mathrm{Shift}_{\vec{m}}(u_n T_r u_n^*)(t_{\vec{m}} \otimes u_n)$, for $r = 2, \ldots, s$, where $T_r \in A_{sn}$ is obtained from the braiding morphisms and can be represented by



We use Figure 11 in the proof of Lemma 3.5 to see that x_r is as in Figure 14.

Also note that x_r is a unitary, so that $(\tilde{a} \otimes 1_{sn})x_r = x_r(\tilde{a} \otimes 1_{sn})$ implies $(\tilde{a} \otimes 1_{sn}) = x_r(\tilde{a} \otimes 1_{sn})x_r^*$:

To obtain the relations in our statement in Figure 15, we proceed by closing strands in Figure 15 with cups and caps to form the loops (the caps and cups correspond to dual morphisms as described in Section 2B). This is done as follows: Let rh and lh be the left and right hand sides of Figure 15. Then we also obtain $rh \otimes 1_{(\bar{X})^{sn}} = lh \otimes 1_{(\bar{X})^{sn}}$. We now multiply both sides with $1_{X^{\otimes|\vec{m}|}} \otimes i_{X^{\otimes sn}}$ from the right (below) and by its conjugate from the left (above). The morphism $i_{X^{\otimes sn}}$ and its conjugate correspond to the pictures in Figure 16, which are obtained from the properties of the duality morphisms; see Section 2B. It is easy to check that we

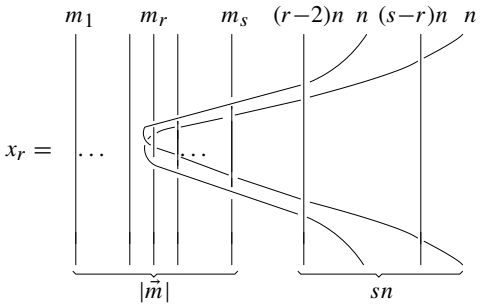


Figure 14. $x_r := (t^*_{\vec{m}} \otimes u^*_n)\mathrm{Shift}_{\vec{m}}(u_n T_r u^*_n)(t_{\vec{m}} \otimes u_n)$.

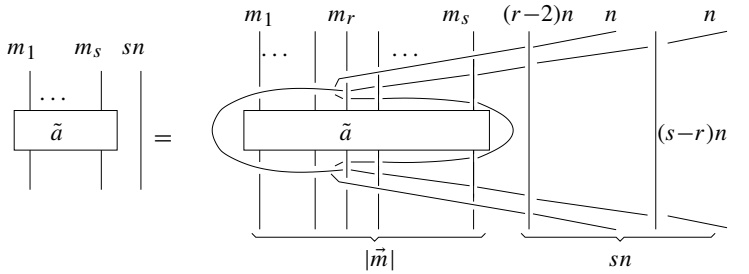


Figure 15. $(\tilde{a} \otimes 1_{sn}) = x_r(\tilde{a} \otimes 1_{sn})x^*_r$.

Figure 16. $\iota^*_{X^{\otimes(sn)}}$ and $\iota_{X^{\otimes(sn)}}$.

obtain $(s-2)n$ unlinked circles on the right hand side, which correspond to the scalar $(\dim X)^{(s-2)n}$. Canceling this with the same number of circles on the left hand side, we obtain the picture as claimed in the statement. □

Corollary 5.2. *The equality in Lemma 5.1 still holds if the rings on both sides are labeled by an irreducible object in $\mathcal{C}'$.*

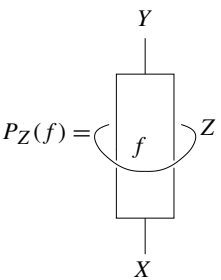
Proof. Assume that $k|n$. Then the proof of Lemma 5.1 works as well if we multiply T_r by $1_{(r-1)n} \otimes p_{\mathbb{1}} \otimes 1_{(s-r+1)n} \otimes p_\mu$ where $p_{\mathbb{1}}$ and p_μ are projections onto irreducible objects appearing in $X^{\otimes n}$ isomorphic to $\mathbb{1}$ and to X_μ, respectively. Going through the proof of Lemma 5.1, we obtain the statement of the corollary at the end. □

5B. ***Notations and preliminaries.*** For any braided semisimple tensor category $\mathcal{C}$ we can define a scalar $s_{\lambda\mu} = \mathrm{Tr}(c_{\mu,\lambda}c_{\lambda,\mu})$, where $c_{\lambda,\mu}$ is the braiding morphism for $X_\lambda \otimes X_\mu$. The S-matrix is then given by $(s_{\lambda\mu})$, where the rows and columns are labeled by the simple objects of $\mathcal{C}$.

Let now $\mathcal{D}$ be a full subcategory of $\mathcal{C}$. We define $\mathcal{T}_\mathcal{D}$ to be the set of simple objects X_λ in $\mathcal{D}$ for which $s_{\lambda\mu} = \dim(X_\lambda)\dim(X_\mu)$ for all simple objects X_μ in $\mathcal{C}'$. We will primarily be interested only in the cases $\mathcal{D} = \mathcal{C}$ and $\mathcal{D} = \mathcal{C}'$. We usually assume $\mathcal{D}$ to be fixed, in which case we may just write $\mathcal{T}$ for $\mathcal{T}_\mathcal{D}$.

Let $X = \bigoplus_\lambda m_\lambda X_\lambda$, $Y = \bigoplus n_\lambda X_\lambda$ be objects in $\mathcal{C}$, and let $f : X \to Y$ be a morphism. Then f can be written as $f = \bigoplus f_\lambda$, where $f_\lambda : m_\lambda X_\lambda \to n_\lambda X_\lambda$. For given $f : X \to Y$, we define the morphism $f_\mathcal{T} : X_\mathcal{T} \to Y_\mathcal{T}$, where $f_\mathcal{T} = \bigoplus_{X_\lambda \in \mathcal{T}} f_\lambda$, and $X_\mathcal{T}$, $Y_\mathcal{T}$ are defined accordingly. Also, we define $p_\mathcal{T}(X) \in \mathrm{End}(X)$ to be the projection from X onto $X_\mathcal{T}$.

For a fixed object Z in $\mathcal{C}$ and a morphism $f : X \to Y$ we define the morphism $P_Z(f) : X \to Y$ by



Of course this picture corresponds to an algebraic expression involving rigidity and braiding morphisms. One can also check that for $Z = Z_1 \otimes Z_2$, the operation P_Z is also given by a picture involving two parallel rings labeled by Z_1 and Z_2. If X_λ, X_μ are simple objects in $\mathcal{C}$, it follows from the definitions that $P_{X_\mu}(1_{X_\lambda}) = (s_{\lambda\mu}/d_\lambda)1_{X_\lambda}$. For a formal linear combination $\Omega = \sum_\mu \omega_\mu X_\mu$, with X_μ simple objects in $\mathcal{C}$, the morphism $P_\Omega(f)$ can also be expressed as the sum $\sum_\mu \omega_\mu P_{X_\mu}(f)$. The following lemma is well-known and follows from the definitions:

Lemma 5.3. *With notations above*,

$$P_{X_\mu}(f) = \sum_\lambda \frac{s_{\lambda\mu}}{d_\lambda} f_\lambda \quad \textit{and} \quad P_\Omega(f) = \sum_{\lambda,\mu} \omega_\mu \frac{s_{\lambda\mu}}{d_\lambda} f_\lambda.$$

We now state a straightforward generalization of the results in [Bruguières 2000, Lemma 1.3].

Proposition 5.4. *Fix the category $\mathcal{D}$ and let $\mathcal{T} = \mathcal{T}_\mathcal{D}$. There exists a linear combination $\Omega = \sum_{\mu\in\Lambda'} \omega_\mu X_\mu$ such that $P_\Omega(f) = f_\mathcal{T}$ for any morphism f in $\mathcal{D}$. Moreover, $\sum_\mu \omega_\mu d_\mu = 1$.*

Proof. We adapt the arguments in the proofs of [Bruguières 2000, Lemmas 1.2 and 1.3] to our setting. By Lemma 5.3, we have to find scalars ω_μ, $\mu \in \Lambda'$ such that $\sum_{\mu\in\Lambda'} \omega_\mu(s_{\lambda\mu}/d_\lambda)$ is equal to 1 or 0 depending on whether $X_\lambda \in \mathcal{T}$ or not. Observe that the second statement will also follow from this as $s_{\lambda\mu} = d_\lambda d_\mu$ for $X_\lambda \in \mathcal{T}$.

To do so, pick an object $X = \bigoplus_{\lambda\in\Lambda(\mathcal{D})} m_\lambda X_\lambda$ in $\mathcal{D}$ with $m_\lambda \neq 0$ for all $\lambda \in \Lambda(\mathcal{D})$. Let z_λ denote the corresponding minimal idempotent in the center of $\mathrm{End}(X)$. Then $P_{X_\mu}(z_\lambda) = \frac{s_{\lambda\mu}}{d_\lambda} z_\lambda$. It also follows immediately by drawing pictures that $P_{Z_1\otimes Z_2}(f) = P_{Z_1}(P_{Z_2}(f))$ for any $f \in \mathrm{End}(X)$ (see also the proof of [Bruguières 2000], Lemma 1.2). Hence we obtain a representation of the fusion algebra of $\mathcal{C}'$ on V, the $\mathbb{C}$-span of the idempotents z_λ, $\lambda \in \Lambda(\mathcal{D})$, with each P_{X_μ} acting via a diagonal matrix with respect to the basis of z_λ's. It follows from Lemma 5.3 that P_{X_μ} acts via the same scalar on the central idempotent z_λ as on $z_{\mathbb{1}}$, for all simple objects X_μ in $\mathcal{C}'$, if and only if $\lambda \in \mathcal{T}$. Hence the projection onto $\mathrm{span}\{z_\lambda, X_\lambda \in \mathcal{T}\}$ is in the image of the fusion algebra, which is spanned by the P_{X_μ}'s. So we can find scalars ω_μ such that this projection is written as $\sum_{\mu\in\Lambda'} \omega_\mu P_{X_\mu}$. The claim follows from this. □

5C.. Let $f : \bigotimes_{i=1}^s X_{\lambda_i} \to \bigotimes_{i=1}^s X_{\mu_i}$ be a morphism. We define, for $r = 1, \dots, s$, the morphism $\hat{f}_r : \bigotimes_{i=r+1}^s X_{\lambda_i} \otimes \overline{X}_{\mu_i} \to \bigotimes_{i=1}^r \overline{X}_{\lambda_i} \otimes X_{\mu_i}$ using rigidity and braiding morphisms for suitable objects as indicated in Figure 17; if $r = s$, the source of $\hat{f}_s$ is defined to be $\mathbb{1}$. For instance, we have

$$\hat{f}_1 = \alpha \circ (1_{\bar{\lambda}_1} \otimes f \otimes 1_{\bar{\mu}_2} \otimes \cdots \otimes 1_{\bar{\mu}_s}) \circ \beta,$$

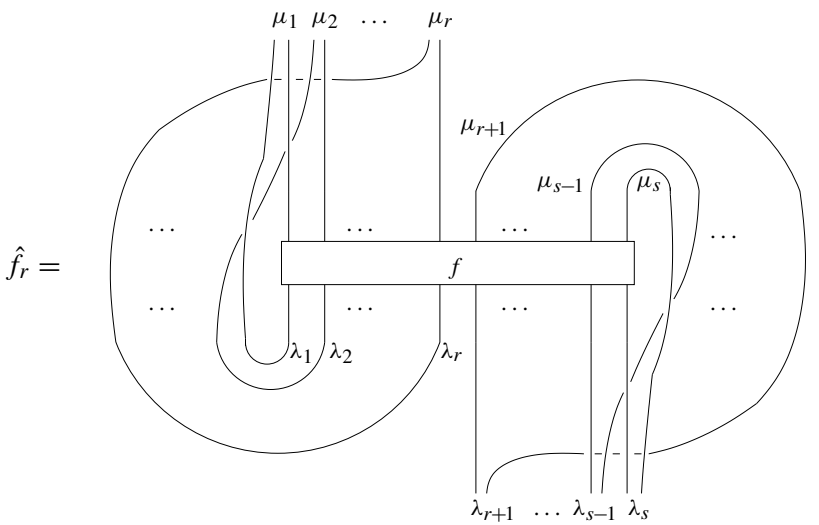


Figure 17. $\hat{f}_r : \bigotimes_{i=r+1}^s X_{\lambda_i} \otimes \overline{X}_{\mu_i} \to \bigotimes_{i=1}^r \overline{X}_{\lambda_i} \otimes X_{\mu_i}$.

for suitable morphisms α and β. We set $\hat{f} = \hat{f}_s$.

Corollary 5.5. *Let* $f \in \mathrm{Hom}_{\mathcal{M}-\mathcal{M}}(M_{\vec{\lambda}}, M_{\vec{\mu}})$ *(with the notation as explained at the beginning of Section 4B), viewed as an element in* $\mathrm{Hom}_{\mathcal{C}}(X_{\vec{\lambda}}, X_{\vec{\mu}})$ (see Lemma 4.5(c)), *and let* P_Ω *be as in Proposition 5.4. Then* $\hat{f}_r = P_\Omega(\hat{f}_r) = (\hat{f}_r)_{\mathcal{T}}$.

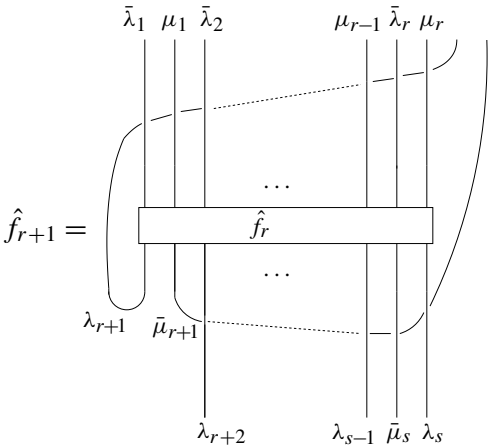
Proof. Fix r, and put a ring around f as it was done for $\tilde{a}$ in Lemma 5.1. By Corollary 5.2 the equality there also holds if we label the ring by $\Omega = \sum \omega_\mu X_\mu$, with the ω_μ as in Proposition 5.4. Observe that the ring on the left hand side becomes the scalar $\sum_\mu \omega_\mu d_\mu = 1$, by Proposition 5.4. Now multiply both sides with suitable morphisms which change f to $\hat{f}_r$, such that all strands ending up go under the ring, and all strands ending at the bottom go above the ring. Then the right-hand side is equal to $P_\Omega(\hat{f}_r)$, which is equal to the left-hand side, $\hat{f}_r$. But by Proposition 5.4 $P_\Omega(\hat{f}_r) = (\hat{f}_r)_{\mathcal{T}}$. □

Lemma 5.6. *If* $f \in \mathrm{Hom}(M_{\vec{\lambda}}, M_{\vec{\mu}})$ *then* $\hat{f} = (\otimes_{i=1}^{s} p_{\mathcal{T}}(\overline{X}_{\lambda_i} \otimes X_{\mu_i}))\hat{f}$.

Proof. We will prove by induction on r that $\hat{f}_r = (\otimes_{i=1}^{r} p_{\mathcal{T}}(\overline{X}_{\lambda_i} \otimes X_{\mu_i}))\hat{f}_r$. For $r = 1$, we have

$$\hat{f}_1 = P_\Omega(\hat{f}_1) = (\hat{f}_1)_{\mathcal{T}},$$

by Corollary 5.5. This proves the claim for $r = 1$, as the target of the morphism $\hat{f}_1$ is $\bar{X}_{\lambda_1} \otimes X_{\mu_1}$. For the induction step we use the inductive formula for $\hat{f}_{r+1}$, as given in the figure:



We obtain from this and the induction assumption that

$$\hat{f}_{r+1} = [(\otimes_{i=1}^{r} p_{\mathcal{T}}(\overline{X}_{\lambda_i} \otimes X_{\mu_i})) \otimes 1_{\overline{X}_{\lambda_{r+1}} \otimes X_{\mu_{r+1}}}]\hat{f}_{r+1}.$$

By Corollary 5.5 (as for the case $r = 1$) we also obtain

$$\hat{f}_{r+1} = P_\Omega(\hat{f}_{r+1}) = p_{\mathcal{T}}(\otimes_{i=1}^{r+1} \overline{X}_{\lambda_i} \otimes X_{\mu_i})\hat{f}_{r+1}.$$

If X_λ is an object in $\mathcal{T}$, then so is $\bar{X}_\lambda$. It follows that the tensor product of simple objects $X_\lambda \otimes X_\mu$ is in $\mathcal{T}$ for $X_\lambda \in \mathcal{T}$ only if also X_μ is in $\mathcal{T}$. One deduces from this and the last two formulas that

$$\begin{aligned}\hat{f}_{r+1} &= p_{\mathcal{T}}(\textstyle\bigotimes_{i=1}^{r+1} \bar{X}_{\lambda_i} \otimes X_{\mu_i})\big((\textstyle\bigotimes_{i=1}^{r} p_{\mathcal{T}}(\bar{X}_{\lambda_i} \otimes X_{\mu_i})) \otimes 1_{\bar{X}_{\lambda_{r+1}} \otimes X_{\mu_{r+1}}}\big)\hat{f}_{r+1} \\ &= \textstyle\bigotimes_{i=1}^{r+1} p_{\mathcal{T}}(\bar{X}_{\lambda_i} \otimes X_{\mu_i})\hat{f}_{r+1}.\end{aligned}$$

This proves the claim by induction on r. □

5D.. It can be shown under fairly weak conditions that the category $\mathcal{T}$ is equivalent to the representation category of a finite group G, see the papers [Bruguières 2000] and [Müger 2000]. In the following, we shall require in addition that $\mathcal{T}$ is equivalent to the representation category of a finite *abelian* group G, for any choice of $\mathcal{D}$. In this case, every simple object in the subcategory $\mathcal{T}$ is invertible. Moreover, we can and will label the simple objects of $\mathcal{T}$ by the elements of G in such a way that $X_g \otimes X_h \cong X_{gh}$ for any $g, h \in G$. Then we get a G-action on the index set Λ defined by $X_{g.\lambda} = X_g \otimes X_\lambda$. We shall also need the subgroup G_1^s of G^s consisting of all s-tuples $(g_1, g_2, \ldots, g_s)$ which satisfy $g_1 g_2 \cdots g_s = 1$. The just defined G-action extends to an action of G_1^s on Λ^s in the obvious way.

Proposition 5.7. *Under the above assumptions we have*

(a) $\mathrm{Hom}(M_{\vec{\lambda}}, M_{\vec{\mu}}) \neq 0$ *only if there exists a* $g \in G_1^s$ *such that* $\vec{\mu} = g.\vec{\lambda}$.

(b) $\dim \mathrm{End}(M_{\vec{\lambda}}) \leq |Stab_{G_1^s}\vec{\lambda}|$.

Proof. We use notations as in Lemma 5.6. By our assumptions, we have $p_{\mathcal{T}}(\bar{X}_{\lambda_i} \otimes X_{\mu_i}) = 0$ unless we can find an element $g_i \in G$ such that $X_{g_i} \subset \bar{X}_{\lambda_i} \otimes X_{\mu_i}$. This implies $g_i.\lambda_i = \mu_i$, and hence $\vec{\mu} = g.\vec{\lambda}$ for some $g \in G^s$. Moreover, we have a nonzero morphism from $\mathbb{1}$ to $\otimes X_{g_i}$ if and only if $\prod g_i = 1$. This shows that $g \in G_1^s$, by Lemma 5.6.

By the discussion in the previous paragraph, the dimension of

$$\mathrm{Hom}(\mathbb{1}, \otimes_i p_G(\bar{X}_{\lambda_i} \otimes X_{\lambda_i}))$$

is equal to the cardinality of all s-tuples $g = (g_i)$ of elements of G for which $g.\vec{\lambda} = \vec{\lambda}$ and whose product $\prod g_i$ is equal to 1. These are exactly the elements of $\mathrm{Stab}_{G_1^s}\vec{\lambda}$. The claim now follows from the fact that the map $f \mapsto \hat{f}$ is injective; indeed, it is easy to construct a left-inverse by multiplying $\hat{f}$ by a suitable combination of $\cap$'s and $\cup$'s to get back f. □

Theorem 5.8. *If the S-matrix for the category* $\mathcal{C}'$ *is invertible, the dual principal graph for the inclusion* $\mathcal{N} \subset \mathcal{M}$ *coincides with its principal graph. In particular, each* $\mathcal{M}$–$\mathcal{M}$ *bimodule* $M_{\vec{\lambda}}$, *with* $\vec{\lambda} = (\lambda_i)$ *such that each* λ_i *labels a simple object in* $\mathcal{C}'$ *is irreducible.*

Proof. We will use the results of Lemma 5.6 and of Proposition 5.7 for the category $\mathscr{C}'$ (recall that its simple objects appear in tensor powers of X whose exponents are divisible by k). If the S-matrix is invertible, the group G corresponding to the category $\mathscr{T}_{\mathscr{C}'}$ is the trivial group. Hence there are no nonzero morphisms between $M_{\vec{\lambda}}$ and $M_{\vec{\mu}}$ for $\vec{\lambda} \neq \vec{\mu}$, and each $\mathcal{M}$–$\mathcal{M}$-bimodule $M_{\vec{\lambda}}$ is irreducible by Proposition 5.7. It follows from the definitions (see before Theorem 4.6) that the multiplicity of a simple $\mathcal{N}$–$\mathcal{M}$ bimodule K_ν in the simple $\mathcal{M}$–$\mathcal{M}$ bimodule $M_{\vec{\lambda}}$ is equal to $L^\nu_{\vec{\lambda}}$.

Observe that $\operatorname{ind}(K_\nu) = d_\nu^2[\mathcal{M} : \mathcal{N}]$ and $\operatorname{ind}(M_{\vec{\lambda}}) = \prod_i d^2_{\lambda_i}$. It follows that

$$\sum_{\nu\in\Lambda'} d_\nu^2[\mathcal{M} : \mathcal{N}] = \left(\sum_{\nu\in\Lambda'} d_\nu^2\right)^s = \sum_{\vec{\lambda}\in(\Lambda')^s} \prod_i d^2_{\lambda_i}.$$

Hence $\sum_{\nu\in\Lambda'} \operatorname{ind}(K_\nu) = \sum_{\vec{\lambda}\in(\Lambda')^s} \operatorname{ind}(M_{\vec{\lambda}})$. Hence any simple $\mathcal{N}$–$\mathcal{M}$-bimodule in a higher relative commutant has as a weak reduction an element in $(K_\nu)_{\nu\in\Lambda'}$, by Theorem 4.6. As our original inclusion $\mathcal{N} \subset \mathcal{M}$ is of finite depth by Theorem 4.6(c), it follows from Lemma 1.10(a) that there can not be any additional $\mathcal{M}$–$\mathcal{M}$-bimodules in the higher relative commutants. □

5E. ***Noninvertible S-matrix.*** We shall make the following assumptions: We assume that the category $\mathscr{T}$ for our chosen category $\mathscr{D} = \mathscr{C}$ is equivalent to the representation category of a finite abelian group G, and, moreover, that $|G| = k$, with k as defined in Section 3A. This also implies that $|G_1^s| = k^{s-1}$. For $\lambda \in \Lambda$ we also define $|\lambda|$ to be the residue class mod k such that $|\lambda| \equiv n \bmod k$ whenever $X_\lambda \subset X^{\otimes n}$.

Theorem 5.9. *Let the conditions be as just stated.*

(a) $\operatorname{End}_{\mathcal{M}-\mathcal{M}}(M_{\vec{\lambda}})$ *has dimension* $|\operatorname{Stab}_{G_1^s}\vec{\lambda}|$ *for any* $\vec{\lambda} \in \Lambda_0^s := \{\vec{\lambda} \in \Lambda^s : k | \sum |\lambda_i|\}$.

(b) *The even vertices of the dual principal graph of the inclusion* $\mathcal{N} \subset \mathcal{M}$ *are labeled by the equivalence classes of irreducible components of the bimodules* $M_{\vec{\lambda}}$, *with* $\vec{\lambda} \in \Lambda_0^s$.

Proof. Let $M_{\vec{\lambda}} = \bigoplus_i m_i Q_{\vec{\lambda},i}$ be the decomposition of the $\mathcal{M}$–$\mathcal{M}$ bimodule $M_{\vec{\lambda}}$ into irreducible $\mathcal{M}$–$\mathcal{M}$-bimodules, the m_i being multiplicities. Then it follows from Lemma 1.10(b), and Proposition 5.7 that

$$\sum_i \operatorname{ind}(Q_{\vec{\lambda},i}) \geq \frac{\operatorname{ind}(M_{\vec{\lambda}})}{\dim(\operatorname{End}(M_{\vec{\lambda}}))} \geq \frac{\operatorname{ind}(M_{\vec{\lambda}})}{|\operatorname{Stab}_{G_1^s}\vec{\lambda}|}.$$

Now let $(Q_j)_j = \bigcup_{\vec{\lambda}}(Q_{\vec{\lambda},i})_i$ be the collection of nonisomorphic representatives of irreducible $\mathcal{M}$–$\mathcal{M}$ submodules of any module $M_{\vec{\lambda}}$ with $\vec{\lambda} \in \Lambda_0^s$. Then

$$\sum_j \operatorname{ind}(Q_j) \geq \sum_{G_1^s-\text{orbits}\in\Lambda_0^s} \frac{\operatorname{ind}(M_{\vec{\lambda}})}{|\operatorname{Stab}_{G_1^s}\vec{\lambda}|} = \frac{1}{k^{s-1}} \sum_{\vec{\lambda}\in\Lambda_0^s} \operatorname{ind}(M_{\vec{\lambda}}).$$

Using Lemma 4.5(a) and Lemma 3.1(d) one sees that this equals

$$= \frac{1}{k^{s-1}} \left(\frac{1}{k} \sum_{\vec{\lambda} \in \Lambda^s} d_{\vec{\lambda}}^2 \right) = \left(\sum_{\lambda \in \Lambda'} d_\lambda^2 \right)^s .$$

But the last sum is equal to $\sum_{\nu \in \Lambda'} \mathrm{ind}(K_\nu)$, as was already shown in the proof of Theorem 5.8. Hence the inequalities above must be equalities, and our set of bimodules $(Q_j)_j$ must already exhaust all possible $\mathcal{M}$–$\mathcal{M}$-bimodules in the higher relative commutant, by Lemma 1.10. □

Remark 5.10. If the stabilizer $\mathrm{Stab}_{G_1^s} \vec{\lambda}$ is trivial, which usually is the case for most labels, the bimodule $M_{\vec{\lambda}}$ is irreducible, and its decomposition into $\mathcal{N}$–$\mathcal{M}$-bimodules is again determined by the fusion coefficients $L_{\vec{\lambda}}^{\nu}$. Unfortunately, our theorem does not say anything about what $\mathrm{End}(M_{\vec{\lambda}})$ looks like if $|\mathrm{Stab}_{G_1^s} \vec{\lambda}| \geq 4$. For example, if the stabilizer has four elements, $\mathrm{End}(M_{\vec{\lambda}})$ could be isomorphic to $\mathbb{C}^4$ or to the 2×2 matrices. Neither does it say how the submodules of $\mathcal{M}_{\vec{\lambda}}$ decompose into irreducible $\mathcal{N}$–$\mathcal{M}$ modules in these cases.

6. Examples

6A. ***Examples of C*-tensor categories.*** (1) The easiest example for our set-up is the representation category $Rep(G)$ of finite-dimensional unitary representations of a finite group. In order to avoid degenerate trivial cases, we take for X in our construction an object such that some tensor power of it contains the whole group ring $\mathbb{C}G$ as a subobject. For example, for G a finite cyclic group, we could take the direct sum of the trivial and of a faithful one-dimensional representation. For these examples, the braiding structure is just given by the permutation of tensor factors, which commutes with the group action. This makes the S-matrix a rank 1 matrix, meaning it is noninvertible unless G is trivial. However, at least in principle, the dual principal graph can be computed from a general result about fixed point algebras of a group K and its subgroup H. In our setting, $K = G^s$ and $H \cong G$, which is embedded by $g \in G \mapsto (g, g, \ldots, g)$ (s times). See [Kosaki et al. 1997] for details.

In the special case when the subgroup K is normal, we obtain principal and dual principal graphs of the factor group H/K. This is the case in our setting if G is abelian.

(2) Let ρ be a II_1 factor representation of the infinite braid group $\mathcal{B}_\infty$ such that the Jones index for the inclusion of factors $\rho(\mathcal{B}_{2,\infty})'' \subset \rho(\mathcal{B}_\infty)''$ is finite. Let us define $A_n = \rho(\mathcal{B}_{n+1,\infty})' \cap \rho(B_\infty)''$ (recall that finite index implies that the relative commutant is finite-dimensional). We moreover assume that there exists, for some $k \in \mathbb{N}$, a projection $p \in A_k$ such that $p\rho(\mathcal{B}_\infty)''p = p\rho(B_{k+1,\infty})''$. It is possible to

define from this a C^*-tensor category, with the objects being the projections in A_n. Most of this has already been done in [Wenzl 1993], Section 2, without mentioning categories. We shall not do this here. We just remark that the constructions of this paper will work in this setting without explicitly exhibiting the category; this has already been done in [Erlijman 2001]. In particular, this can be applied to the Jones subfactors as well as to the Hecke algebra and BCD type subfactors.

(3) Let $U_q\mathfrak{g}$ be the Drinfeld–Jimbo deformation of the universal enveloping algebra $U\mathfrak{g}$ of a semisimple Lie algebra $\mathfrak{g}$. It is well-known that the category of its finite-dimensional representations has a braiding structure. It can not be unitarized except for $q = 1$. If q is a root of unity $\neq 1$, one can define a special class of representations called tilting modules which again forms a braided tensor category. It can be shown that the category of tilting modules has a semisimple quotient with only finitely many simple modules up to equivalence; this is often referred to as a fusion category (see [Andersen 1992],[Andersen and Paradowski 1995]). Moreover, for q being certain roots of unity (usually of the form $q = e^{\pm 2\pi i/l}$ for suitable integers l (see [Wenzl 1998] for precise values), this quotient can be unitarized. This yields a large and important class of C^* tensor categories. Using the one-sided subfactor construction, one obtains the Jones subfactors for X being the $U_q sl_2$-analog of the 2-dimensional representation of sl_2. Similarly, Hecke algebra subfactors and BCD type subfactors can be obtained from fusion categories of quantum groups of classical Lie types.

These C^*-fusion categories can also be obtained by a completely different construction using the category of positive energy representation of a loop group. The difficulty in this construction comes from the fact that one can not use the usual tensor product for representations; instead one has to define a new, so-called fusion tensor product (see [Wassermann 1998]).

(4) Let $N \subset M$ be an inclusion of II_1 factors with finite index and finite depth. Then the category of N–N bimodules obtained as direct sums of summands of the bimodules

$$M^{\otimes n} = M \otimes_N M \otimes_N \cdots \otimes_N M$$

(n times), $n \in \mathbb{N}$ defines a C^*-tensor category which may or may not be braided. One can similarly also define the C^*-tensor category of M-M bimodules generated by $M^{\otimes n}$.

If these categories are not braided, one can apply a general construction, called the categorical quantum double construction to construct from our category of bimodules a larger braided C^* tensor category. It was shown that this category is equivalent to the category of $\mathcal{M}$–$\mathcal{M}$-bimodules for the asymptotic inclusion $\mathcal{N} \subset \mathcal{M}$ derived from $N \subset M$; see [Müger 2003]. If the original category already was

braided, the asymptotic inclusion coincides with the 2-sided inclusion constructed in this paper.

(5) Our constructions of bimodules in this paper are based on certain endomorphisms of II_1 factors. The approach to categories via endomorphisms has been used for a long time for type III factors in the framework of algebraic quantum field theory (see [Longo and Roberts 1997; Fredenhagen et al. 1989; Xu 2000], for example). Here subtleties involving coupling constants do not matter, and objects are given directly by morphisms.

6B. ***Examples for our construction.*** (1) We first list examples of C^*-tensor categories with invertible S-matrix.

(a) The S-matrix for the full fusion tensor categories as constructed in [Andersen 1992],[Andersen and Paradowski 1995] is invertible under the conditions for unitarizability, as stated in [Wenzl 1998]. Hence if we can find an object X such that *all* irreducible representation of the fusion category appear in some tensor power of X, we have $\mathscr{C}' = \mathscr{C}$ and the dual principal graph is equal to the principal graph. Such representations can be found in all cases, but usually can not be chosen to be irreducible. For instance, for Lie type A (the case of Jones subfactors and Hecke algebra subfactors), one can choose $X = \mathbb{1} \oplus V$, where V is the analog of the vector representation.

(b) Similarly, the S-matrix for the quantum double of a C^* tensor category is always invertible (see [Müger 2003], for instance). Hence, as soon as we have found an object X for which all irreducible representations of the double category appear in some tensor power of X, the dual principal graph of our s-sided inclusion with respect to X is equal to the principal graph.

(2) It turns out that our construction not only depends on the category $\mathscr{C}$, but also on the choice of the object X. Even though in the case of the fusion tensor categories the S-matrix for $\mathscr{C}$ is invertible, the S matrix for the category $\mathscr{C}'$ may not be invertible. For instance, for type A if one takes $X = V$, the S-matrix for $\mathscr{C}'$ is invertible only if the degree of the root of unity is coprime to k. If this is not the case, however, our results for noninvertible S-matrices apply. This will be shown in more detail in the following subsection at an example.

6C. ***Subfactors related to Jones subfactors.*** We illustrate our examples in some detail for the fusion category $\mathscr{C}$ of $U_q sl_2$, with $q = e^{2\pi i/l}$. There also exist other, more elementary methods to construct these categories using the Temperley–Lieb algebras; see [Turaev 1994], for example. As mentioned before, this is also one of the cases where the subfactor constructions can be done on the level of braid representations, as it was carried out in the original paper [Erlijman 2001].

We give a brief description of this category. Up to isomorphism, we have exactly $l-1$ simple objects in $\mathcal{C}$, which are denoted by $[i]$, $1 \le i \le l-1$. The decomposition of tensor products is given by

$$[i] \otimes [j] = [|i-j|+1] \oplus [|i-j|+3] \oplus \cdots \oplus [m], \tag{6-1}$$

where m is the minimum of $i+j-1$ and $2l-1-i-j$. One sees easily that $[1]$ corresponds to the trivial object. It follows from the tensor product rules by induction on n that all simple objects in $[2]^{\otimes n}$ are labeled by even numbers if n is odd, and by odd numbers if n is even. Hence $k=2$ and the simple objects of $\mathcal{C}'$ are labeled by odd numbers. This explicitly describes the principal graph for $\mathcal{N} \subset \mathcal{M}$, constructed with $X=[2]$, by Theorem 4.6.

Observe that $[i] \otimes [l-1] = [l-i]$ for all $1 \le i < l$. Hence the objects $[1]$ and $[l-1]$ together with the operation $\otimes$ form a group G which is isomorphic to $\mathbb{Z}/2\mathbb{Z}$. Moreover, the S matrix is well-known to be of the form $S = (\sin(ij\pi/l))$, up to a scalar.

It is very easy to check that if l is even, then $\sin(i(l-1)\pi/l) = \sin(i\pi/l)$ for any odd $i = 1, 3, \ldots, l-1$. Hence the category $\mathcal{T}$ contains at least the objects $[1]$ and $[l-1]$. It contains no more simple objects as obviously $\sin(i\pi/l) = \sin(ij\pi/l)$ for $1 < j < l$ only if $j = l-1$. So the conditions at the beginning of Section 5B are satisfied with $|G| = 2 = k$. We have shown most of the following

Proposition 6.1. *Let $\mathcal{N} \subset \mathcal{M}$ be the subfactor constructed from the s-sided inclusion from the Jones subfactor at an l-th root of unity, with l even. Then we have*

(a) *The even vertices of the principal graph are labeled by all s-tuples of odd positive numbers less than l and the odd vertices are labeled by all odd positive numbers less than l. The number of edges between two vertices can be computed from the tensor product rule stated in (6-1).*

(b) *Each s-tuple of positive integers less than l whose sum is even and which contains the number $l/2$ at most once labels an even vertex of the dual principal graph; the number of edges emanating from such a vertex can be computed as in (a). The $\mathcal{M}$–$\mathcal{M}$ bimodules $M_{\vec{\lambda}}$ labeled by an s-tuple $\vec{\lambda}$ containing the number $l/2$ exactly $r > 1$ times satisfies* $\dim(\mathrm{End}(M_{\vec{\lambda}})) = 2^{r-1}$.

Proof. Part (a) follows from Theorem 4.6 and our explicit description of the simple objects of $\mathcal{C}'$. For part (b), we have already checked the conditions stated at the beginning of Section 5B. It remains to calculate $\mathrm{Stab}_{G_1^s} \vec{\lambda}$ for any $\vec{\lambda} \in \Lambda^s$. Recall that the action of the nontrivial element of G on our labeling set is given by $i \mapsto l-i$. Obviously, the only fixed point is $l/2$ for l odd. It is now not hard to show that $\vec{\lambda} \in \Lambda^s$ has a nontrivial stabilizer in G_1^s if and only if $r \ge 2$ of its components are equal to $l/2$, and that in this case the stabilizer has exactly 2^{r-1} elements. Statement (b) now follows from Theorem 5.9. □

Remark 6.2. If $s = 3$, part (b) of the last proposition completely determines the number of edges in the dual principal graph except for the decomposition of the bimodule $M_{\vec{\lambda}}$ with $\vec{\lambda} = (l/2, l/2, l/2)$, which could decompose into the direct sum of four nonisomorphic irreducible $\mathcal{M}$–$\mathcal{M}$ bimodules or into the direct sum of two isomorphic irreducible $\mathcal{M}$–$\mathcal{M}$ bimodules.

Acknowledgement

We thank the referee for a careful reading of the manuscript and for several useful suggestions, which lead to an improved presentation.

References

[Andersen 1992] H. H. Andersen, "Tensor products of quantized tilting modules", *Comm. Math. Phys.* **149**:1 (1992), 149–159. MR 94b:17015 Zbl 0760.17004

[Andersen and Paradowski 1995] H. H. Andersen and J. Paradowski, "Fusion categories arising from semisimple Lie algebras", *Comm. Math. Phys.* **169**:3 (1995), 563–588. MR 96e:17026 Zbl 0827.17010

[Asaeda 2006] M. Asaeda, "Quantum multiple construction of subfactors", preprint, UC Riverside, 2006. math.OA/0602265

[Barrett and Westbury 1999] J. W. Barrett and B. W. Westbury, "Spherical categories", *Adv. Math.* **143**:2 (1999), 357–375. MR 2000c:18007 Zbl 0930.18004

[Bisch 1997] D. Bisch, "Bimodules, higher relative commutants and the fusion algebra associated to a subfactor", pp. 13–63 in *Operator algebras and their applications* (Waterloo, ON, 1994/1995), edited by P. A. Fillmore and J. A. Mingo, Fields Inst. Commun. **13**, Amer. Math. Soc., Providence, RI, 1997. MR 97i:46109 Zbl 0894.46046

[Bruguières 2000] A. Bruguières, "Catégories prémodulaires, modularisations et invariants des variétés de dimension 3", *Math. Ann.* **316**:2 (2000), 215–236. MR 2001d:18009 Zbl 0943.18004

[Erlijman 2001] J. Erlijman, "Multi-sided braid type subfactors", *Canad. J. Math.* **53**:3 (2001), 546–564. MR 2002f:46119 Zbl 1031.46072

[Erlijman 2003] J. Erlijman, "Multi-sided braid type subfactors, II", *Canad. Math. Bull.* **46**:1 (2003), 80–94. MR 2004a:46060 Zbl 1032.46074

[Evans and Kawahigashi 1998] D. E. Evans and Y. Kawahigashi, "Orbifold subfactors from Hecke algebras, II: Quantum doubles and braiding", *Comm. Math. Phys.* **196**:2 (1998), 331–361. MR 2000i:46058 Zbl 0910.46043

[Fredenhagen et al. 1989] K. Fredenhagen, K.-H. Rehren, and B. Schroer, "Superselection sectors with braid group statistics and exchange algebras, I: General theory", *Comm. Math. Phys.* **125**:2 (1989), 201–226. MR 91c:81047 Zbl 0682.46051

[Freyd 1964] P. Freyd, *Abelian categories: An introduction to the theory of functors*, Harper & Row, New York, 1964. MR 29 #3517 Zbl 0121.02103

[Izumi 2000] M. Izumi, "The structure of sectors associated with Longo-Rehren inclusions, I: General theory", *Comm. Math. Phys.* **213**:1 (2000), 127–179. MR 2002k:46160 Zbl 1032.46529

[Jones 1983] V. F. R. Jones, "Index for subfactors", *Invent. Math.* **72**:1 (1983), 1–25. MR 84d:46097 Zbl 0508.46040

[Kassel 1995] C. Kassel, *Quantum groups*, Graduate Texts in Mathematics **155**, Springer, New York, 1995. MR 96e:17041 Zbl 0808.17003

[Kosaki et al. 1997] H. Kosaki, A. Munemasa, and S. Yamagami, "On fusion algebras associated to finite group actions", *Pacific J. Math.* **177**:2 (1997), 269–290. MR 98i:46064 Zbl 0882.46030

[Longo and Roberts 1997] R. Longo and J. E. Roberts, "A theory of dimension", *K-Theory* **11**:2 (1997), 103–159. MR 98i:46065 Zbl 0874.18005

[Mac Lane 1998] S. Mac Lane, *Categories for the working mathematician*, Second ed., Graduate Texts in Mathematics **5**, Springer, New York, 1998. MR 2001j:18001 Zbl 0906.18001

[Müger 2000] M. Müger, "Galois theory for braided tensor categories and the modular closure", *Adv. Math.* **150**:2 (2000), 151–201. MR 2001a:18008 Zbl 0945.18006

[Müger 2003] M. Müger, "From subfactors to categories and topology, II: The quantum double of tensor categories and subfactors", *J. Pure Appl. Algebra* **180**:1-2 (2003), 159–219. MR 2004f:18014 Zbl 1033.18003

[Orellana and Wenzl 2002] R. C. Orellana and H. G. Wenzl, "q-centralizer algebras for spin groups", *J. Algebra* **253**:2 (2002), 237–275. MR 2003i:17022 Zbl 1085.17012

[Turaev 1994] V. G. Turaev, *Quantum invariants of knots and 3-manifolds*, de Gruyter Studies in Mathematics **18**, de Gruyter, Berlin, 1994. MR 95k:57014 Zbl 0812.57003

[Wassermann 1998] A. Wassermann, "Operator algebras and conformal field theory, III: Fusion of positive energy representations of LSU(N) using bounded operators", *Invent. Math.* **133**:3 (1998), 467–538. MR 99j:81101 Zbl 0944.46059

[Wenzl 1988] H. Wenzl, "Hecke algebras of type A_n and subfactors", *Invent. Math.* **92**:2 (1988), 349–383. MR 90b:46118 Zbl 0663.46055

[Wenzl 1993] H. Wenzl, "Braids and invariants of 3-manifolds", *Invent. Math.* **114**:2 (1993), 235–275. MR 94i:57021 Zbl 0804.57007

[Wenzl 1998] H. Wenzl, "C^* tensor categories from quantum groups", *J. Amer. Math. Soc.* **11**:2 (1998), 261–282. MR 98k:46123 Zbl 0888.46043

[Xu 2000] F. Xu, "Jones–Wassermann subfactors for disconnected intervals", *Commun. Contemp. Math.* **2**:3 (2000), 307–347. MR 2001f:46094 Zbl 0966.46043

Received November 23, 2005. Revised February 28, 2007.

JULIANA ERLIJMAN
DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF REGINA
REGINA, SASKATCHEWAN S4S 0A2
CANADA
erlijman@math.uregina.ca
http://www.math.uregina.ca/~erlijman/

HANS WENZL
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
9500 GILMAN DRIVE, DEPT 0112
LA JOLLA, CA 92093-0112
UNITED STATES
hwenzl@ucsd.edu
http://www.math.ucsd.edu/~wenzl/

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.401

AN ELEMENTARY, EXPLICIT, PROOF OF THE EXISTENCE OF QUOT SCHEMES OF POINTS

TROND STØLEN GUSTAVSEN, DAN LAKSOV AND ROY MIKAEL SKJELNES

We give an easy and elementary construction of quotient schemes of modules of relatively finite rank under very general conditions. The construction provides a natural, explicit description of an affine covering of the quotient schemes, that is useful in many situation.

Introduction

Quotient schemes were introduced by A. Grothendieck [1966], and belong among the fundamental tools in algebraic geometry. They generalize simultaneously the Hilbert schemes and the Grassmann schemes. In most cases where these schemes appear it suffices to know that they exist. However, there are cases when it is crucial to have an explicit description of the schemes. We give here, for any morphism of schemes $\mathrm{Proj}(\mathscr{R}) \to S$, where S is arbitrary and $\mathscr{R}$ is a quasicoherent graded $\mathcal{O}_S$-algebra, an elementary construction of quotient schemes parametrizing equivalence classes of surjections from a quasicoherent $\mathcal{O}_{\mathrm{Proj}(\mathscr{R})}$-module to coherent modules that are of relatively finite rank over S. The construction provides a natural and explicit description of an affine covering of the quotient schemes. In a previous article [Gustavsen et al. 2007] we indicated the usefulness of such a description in the case of Hilbert schemes of points, and further evidence of this is given by M. Huibregtse [2002; 2006]. Our proof of the existence of the quotient schemes is a simplification and clarification of the constructions of these works.

The main new idea is the description of a local version of the quot functor. More precisely, let A be a ring, B an A-algebra and E and F modules over A with E free of finite rank, and fix an A-module homomorphism $s : E \to B \otimes_A F$. We parametrize B-module structures on E, together with B-module homomorphisms $u : B \otimes_A F \to E$ such that $us = \mathrm{id}_E$.

When S is locally noetherian and $\mathscr{R}$ is locally finitely generated by elements of degree one, our existence result for the quotient schemes of modules of relatively finite rank follows from the more general results of Grothendieck [1966]. Apparently the first detailed existence proof of Grothendieck's result was provided by

MSC2000: 14D20, 14A15.
Keywords: Hilbert schemes, quotient schemes.

A. Altman and S. Kleiman [1980], valid under quite general conditions, where S is not assumed to be locally noetherian. A much used method for obtaining these results was given by D. Mumford [1966] (see also [Sernesi 2006]). In contrast to these approaches our method relies on simple algebraic constructions and avoids embeddings into high dimensional grassmannians via Castelnuovo–Mumford regularity. As a consequence our local description of the quotient schemes is, in most cases appearing in applications, in terms of explicit natural equations in the affine space of commuting matrices of smallest possible size, that is, the size is equal to the finite rank of the modules.

1. The local quot functor

1.1. Let A be a commutative ring with unit, and let $A \to B$ be an A-algebra. Moreover, let E and F be A-modules where E is free of finite rank, and let

$$s : E \to B \otimes_A F$$

be an A-module homomorphism. We want to describe all B-module structures on E, together with B-module homomorphisms $u : B \otimes_A F \to E$ such that $us = \mathrm{id}_E$. Recall that two surjections are considered *equivalent* if their kernels coincide, and that a B-module structure on an A-module E corresponds to an A-algebra homomorphism $B \to \mathrm{End}_A(E)$.

More precisely, we want to describe, for every A-algebra $A \to A'$, the set consisting of an $A' \otimes_A B$-module structure on $A' \otimes_A E$ together with an $A' \otimes_A B$-module homomorphism $A' \otimes_A B \otimes_A F \xrightarrow{u} A' \otimes_A E$ such that the composite A'-module homomorphism

$$A' \otimes_A E \xrightarrow{\mathrm{id}_{A'} \otimes_A s} A' \otimes_A B \otimes_A F \xrightarrow{u} A' \otimes_A E$$

is the identity. This clearly defines a functor from A-algebras to sets.

The main objective of Sections 1, 2 and 3 is to show that this functor is representable by an A-algebra Q^s, and to give a simple explicit description of this algebra.

We first find a slightly different description of this functor.

Notation 1.2. We shall need *evaluation* and *trace* maps. We recall that when G and K are A-modules with K free of finite rank, we obtain for every submodule D of $\mathrm{Hom}_A(G, K)$ the *evaluation* homomorphisms

$$\mathrm{ev} : D \otimes_A G \to K \quad \text{or} \quad \mathrm{ev} : G \otimes_A D \to K$$

that maps $x \otimes_A \varphi$, or $\varphi \otimes_A x$, to $\varphi(x)$. Moreover, we have the *trace* homomorphism

$$\mathrm{tr} : A \to K \otimes_A K^{\check{}}$$

obtained from the dual homomorphism of $\mathrm{ev} : K^{\check{}} \otimes_A K \to A$.

Proposition 1.3. *There is a natural bijection between*

(1) *the set of B-module structures on E provided with a surjective homomorphism $u : B \otimes_A F \to E$ of B-modules, and*

(2) *the set of A-algebra homomorphisms $\varphi : B \to \mathrm{End}_A(E)$ provided with an A-module homomorphism $v : F \to E$ such that the composite map*

$$B \otimes_A F \xrightarrow{\varphi \otimes_A v} \mathrm{End}_A(E) \otimes_A E \xrightarrow{\mathrm{ev}} E$$

is surjective.

The bijection maps a pair (v, φ) in (2) *to* $\mathrm{ev}(\varphi \otimes_A v)$ *in* (1), *where E has the B-module structure given by φ.*

Proof. We first note that B-module structures on E correspond to A-algebra homomorphisms $\varphi : B \to \mathrm{End}(E)$.

Given a B-module structure on E, a surjection $u : B \otimes_A F \to E$ of B-modules determines a B-module structure on E uniquely. Moreover an A-module homomorphism $u : B \otimes_A F \to E$ determines an A-module homomorphism $v : F \to E$ by *restriction of scalars*, and conversely, u is determined by v by *extension* of scalars.

Finally, an A-module homomorphism $v : F \to E$ and a B-module structure $\varphi : B \to \mathrm{End}_A(E)$ on E makes the composite map

$$B \otimes_A F \xrightarrow{\varphi \otimes_A v} \mathrm{End}_A(E) \otimes_A E \xrightarrow{\mathrm{ev}} E$$

into a B-module homomorphism, since ev is an $\mathrm{End}_A(E)$-module homomorphism and $\varphi : B \to \mathrm{End}_A(E)$ is a ring homomorphism. □

Corollary 1.4. *The bijection of the proposition induces a bijection between*

(1) *B-module structures on E provided with a homomorphism of B-modules $u : B \otimes_A F \to E$ such that*

$$E \xrightarrow{s} B \otimes_A F \xrightarrow{u} E$$

is the identity, and

(2) *A-algebra homomorphisms $\varphi : B \to \mathrm{End}_A(E)$ provided with an A-module homomorphism $v : F \to E$ such that*

$$E \xrightarrow{s} B \otimes_A F \xrightarrow{\varphi \otimes_A v} \mathrm{End}_A(E) \otimes_A E \xrightarrow{\mathrm{ev}} E$$

is the identity.

Proof. This follows from the proposition since the surjectivity of u and $\mathrm{ev}(\varphi \otimes_A v)$ is automatic. □

2. The local Hilbert scheme

The material of this section will basically give a construction of Hilbert schemes of points, as in [Gustavsen et al. 2007], but the presentation here is different from that of that paper. We use that the A-algebra $\mathrm{Sym}_A(G \otimes_A \mathrm{End}_A(E)^\vee)$ parametrizes module homomorphisms $u : G \to \mathrm{End}_A(E)$. Then we explicitly construct a residue algebra H of $\mathrm{Sym}_A(G \otimes_A \mathrm{End}_A(E)^\vee)$ that parametrizes those u such that the elements of the image commute. Then H also parametrizes A-algebra homomorphisms $\mathrm{Sym}_A(G) \to \mathrm{End}_A(E)$.

Notation 2.1. Let G be an A-module and K a free A-module of finite rank. We denote by

$$u_K : \mathrm{Sym}_A(G \otimes_A K^\vee) \otimes_A G \to \mathrm{Sym}_A(G \otimes_A K^\vee) \otimes_A K$$

the $\mathrm{Sym}_A(G \otimes_A K^\vee)$-module homomorphism defined by $u_K(1 \otimes_A x) = x \otimes_A \mathrm{tr}(1_A)$ for all $x \in G$, where $G \otimes_A K^\vee$ is considered as a submodule of $\mathrm{Sym}_A(G \otimes_A K^\vee)$.

Lemma 2.2. *For every A-algebra homomorphism $A \to A'$ there is a natural bijection between*

(1) *A'-module homomorphisms $A' \otimes_A G \to A' \otimes_A K$, and*

(2) *A-algebra homomorphisms $\mathrm{Sym}_A(G \otimes_A K^\vee) \to A'$.*

The bijection maps $\varphi : \mathrm{Sym}_A(G \otimes_A K^\vee) \to A'$ to the homomorphism u defined by $u(1_{A'} \otimes_A x) = (\varphi \otimes_A \mathrm{id}_K) u_K(1 \otimes_A x)$ for all $x \in G$.

Proof. The set (2) is mapped to the set (1) via three isomorphisms:

(1) $\mathrm{Hom}_{A\text{-alg}}(\mathrm{Sym}_A(G \otimes_A K^\vee), A') \to \mathrm{Hom}_A(G \otimes_A K^\vee, A')$ that follows from the definition of the symmetric algebra.

(2) $\mathrm{Hom}_A(G \otimes_A K^\vee, A') \to \mathrm{Hom}_A(G, A' \otimes_A K)$, that is a canonical standard isomorphism, when K is free, that maps $u : G \otimes_A K^\vee \to A'$ to the composite homomorphism $G \xrightarrow{\mathrm{id}_G \otimes_A \mathrm{tr}} G \otimes_A K^\vee \otimes_A K \xrightarrow{u \otimes_A \mathrm{id}_K} A' \otimes_A K$.

(3) $\mathrm{Hom}_A(G, A' \otimes_A K) \to \mathrm{Hom}_{A'}(A' \otimes_A G, A' \otimes_A K)$, that is the standard isomorphism obtained by *extension of scalars*. □

Notation 2.3. Let $u : G \to K$ be a homomorphism of A-modules with K free of finite rank. We denote by $\mathfrak{I}_Z(u)$ the image of the composite homomorphism

$$G \otimes_A K^\vee \xrightarrow{u \otimes_A \mathrm{id}\, K^\vee} K \otimes_A K^\vee \xrightarrow{\mathrm{ev}} A.$$

Lemma 2.4. *For every A-algebra $A \to A'$ the A'-module homomorphism*

$$A' \otimes_A G \xrightarrow{\mathrm{id}_{A'} \otimes_A u} A' \otimes_A K$$

is zero if and only if the homomorphism $A \to A'$ factors via the residue homomorphism $A \to A/\mathfrak{I}_Z(u)$.

Proof. This is an immediate consequence of the definition of $\mathfrak{I}_Z(u)$. □

Notation 2.5. Let

$$v : \mathrm{Sym}_A(G \otimes_A \mathrm{End}_A(E)^\vee) \otimes_A G \otimes_A G \to \mathrm{Sym}_A(G \otimes_A \mathrm{End}_A(E)^\vee) \otimes_A \mathrm{End}_A(E)$$

be the $\mathrm{Sym}_A(G \otimes_A \mathrm{End}_A(E)^\vee)$-module homomorphism defined by

$$v(1 \otimes_A x \otimes_A y) = u_{\mathrm{End}_A(E)}(1 \otimes_A x) u_{\mathrm{End}_A(E)}(1 \otimes_A y) - u_{\mathrm{End}_A(E)}(1 \otimes_A y) u_{\mathrm{End}_A(E)}(1 \otimes_A x),$$

where $u_{\mathrm{End}_A(E)}$ is defined in paragraph 2.1. We denote by H the residue algebra of $\mathrm{Sym}_A(G \otimes_A \mathrm{End}_A(E)^\vee)$ modulo the ideal $\mathfrak{I}_Z(v)$ and let

$$\rho_H : \mathrm{Sym}_A(G \otimes_A \mathrm{End}_A(E)^\vee) \to H$$

be the residue homomorphism. From the $\mathrm{Sym}_A(G \otimes_A \mathrm{End}_A(E)^\vee)$-module homomorphism $u_{\mathrm{End}_A(E)}$ and the A-algebra homomorphism ρ_H we obtain an H-module homomorphism $w : H \otimes_A G \to H \otimes_A \mathrm{End}_A(E)$ defined by $w(1 \otimes_A x) = (\rho_H \otimes_A 1_{\mathrm{End}_A(E)}) u_{\mathrm{End}_A(E)}(1 \otimes_A x)$ for all $x \in G$. It follows from the definition of $\mathfrak{I}_Z(w)$ and H that the elements of the image of w commute. Consequently w gives a unique H-algebra homomorphism

$$\mu_H : H \otimes_A \mathrm{Sym}_A(G) \to H \otimes_A \mathrm{End}_A(E)$$

such that $\mu_H(1_H \otimes_A x) = w(1_H \otimes_A x)$ for all $x \in G$.

Lemma 2.6. *Let $A \to A'$ be an A-algebra. We have a bijection between*

(1) *A'-algebra homomorphisms $A' \otimes_A \mathrm{Sym}_A(G) \to A' \otimes_A \mathrm{End}_A(E)$, and*

(2) *A-algebra homomorphisms $H \to A'$.*

The bijection maps $\varphi : H \to A'$ to the homomorphism u defined by $u(1_{A'} \otimes_A x) = (\varphi \otimes_A \mathrm{id}_{\mathrm{End}_A(E)}) \mu_H(1_H \otimes_A x)$ for all $x \in G$.

Proof. It follows from Lemma 2.2 that there is a bijection between A'-module homomorphisms $u : A' \otimes_A G \to A' \otimes_A \mathrm{End}_A(E)$ and A-algebra homomorphisms $\varphi : \mathrm{Sym}_A(G \otimes \mathrm{End}_A(E)^\vee) \to A'$. By the definition of $\mathfrak{I}_Z(w)$ and Lemma 2.4 the homomorphism φ factors via a homomorphism $\chi : H \to A'$ if and only if the elements $u(1_{A'} \otimes_A x)$ commute for all $x \in X$. However, the maps $u : A' \otimes_A G \to A' \otimes_A \mathrm{End}_A(E)$ whose images consist of commuting elements correspond to A'-algebra homomorphisms $\psi : A' \otimes_A \mathrm{Sym}_A(G) \to A' \otimes_A \mathrm{End}_A(E)$ for which

$$\psi(1_{A'} \otimes_A x) = u(1_{A'} \otimes_A x)$$

for all $x \in G$. □

Notation 2.7. Let G be an A-module and $\mathfrak{I}$ an ideal in $\mathrm{Sym}_A(G)$. Denote by $\iota : \mathfrak{I} \to \mathrm{Sym}_A(G)$ the inclusion map, let $B = \mathrm{Sym}_A(G)/\mathfrak{I}$, and let $\rho_B : \mathrm{Sym}_A(G) \to B$ be the residue homomorphism. We denote by v the composite H-module homomorphism

$$H \otimes_A \mathfrak{I} \xrightarrow{\mathrm{id}_H \otimes_A \iota} H \otimes_A \mathrm{Sym}_A(G) \xrightarrow{\mu_H} H \otimes_A \mathrm{End}(E). \tag{2-1}$$

Moreover we let H_B be the residue algebra of H modulo the ideal $\mathfrak{I}_Z(v)$ and we denote the residue homomorphism by

$$\rho_{H_B} : H \to H_B.$$

We tensor the modules of (2-1) by H_B over H and obtain a homomorphism of H_B-modules

$$H_B \otimes_A \mathfrak{I} \xrightarrow{\mathrm{id}_{H_B} \otimes_A \iota} H_B \otimes_A \mathrm{Sym}_A(G) \xrightarrow{\mathrm{id}_{H_B} \otimes_H \mu_H} H_B \otimes_A \mathrm{End}_A(E)$$

such that the composite homomorphism is zero by the definition of $\mathfrak{I}_Z(v)$. Consequently we obtain a homomorphism of H_B-algebras

$$\mu_{H_B} : H_B \otimes_A B \to H_B \otimes_A \mathrm{End}_A(E)$$

such that $\mu_{H_B}(1 \otimes_A \rho_B(f)) = (\rho_{H_B} \otimes_A \mathrm{id}_{\mathrm{End}_A(E)})\mu_H(1 \otimes_A f)$ for all $f \in \mathrm{Sym}_A(G)$.

The algebra H_B parametrizes all B-module structures on E, and μ_{H_B} is the *universal homomorphism*.

Proposition 2.8. *Let $A \to A'$ be an A-algebra. We have a bijection between*

(1) *A'-algebra homomorphisms $A' \otimes_A B \to A' \otimes_A \mathrm{End}_A(E)$, and*

(2) *A-algebra homomorphisms $H_B \to A'$.*

The bijection maps $\varphi : H_B \to A'$ to the homomorphism u defined by $u(1_{A'} \otimes_A f) = (\varphi \otimes_A \mathrm{id}_{\mathrm{End}_A(E)})\mu_{H_B}(1_{H_B} \otimes_A f)$ for all $f \in B$.

Proof. It follows from Lemma 2.4 that an A-algebra homomorphism $H \to A'$ factors via $\rho_{H_B} : H \to H_B$ if and only if the composite homomorphism

$$A' \otimes_A \mathfrak{I} \xrightarrow{\mathrm{id}_{A'} \otimes_A \iota} A' \otimes_A \mathrm{Sym}_A(G) \xrightarrow{\mathrm{id}_{A'} \otimes_H \mu_H} A' \otimes_A \mathrm{End}_A(E)$$

is zero. This last condition holds if and only if $\mathrm{id}_{A'} \otimes_H \mu_H$ factors via an A'-algebra homomorphism $A' \otimes_A B \to A' \otimes_A \mathrm{End}_A(E)$. □

3. The local quot scheme

It follows from Lemma 2.2 that the A-algebra $\mathrm{Sym}_A(F\otimes_A E^\vee)$ parametrizes homomorphisms $F\to E$. We use this, together with the properties of the A-algebra H_B, to construct a residue algebra Q^s of $\mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A H_B$ that parametrizes the *local quot functor* of Section 1.1.

Notation 3.1. We defined in 2.1 an A-module homomorphism

$$u_E : \mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A F \to \mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A E$$

such that $u_E(1\otimes_A y) = y\otimes_A \mathrm{tr}(1)$ for all $y\in F$.

Let $A\to B$ be an A-algebra and fix a presentation $0\to\mathfrak{I}\to\mathrm{Sym}_A(G)\to B\to 0$ of B as in 2.7. Moreover, fix an A-module homomorphism

$$s : E\to B\otimes_A F$$

and let

$$v : \mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A H_B\otimes_A E \to \mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A H_B\otimes_A E$$

be the composition of the $\mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A H_B$-module homomorphisms

$$\begin{aligned}
\mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A H_B\otimes_A E &\xrightarrow{1\otimes_A 1\otimes_A s} \mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A H_B\otimes_A B\otimes_A F\\
&\xrightarrow{\sim} \mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A F\otimes_A H_B\otimes_A B\\
&\xrightarrow{u_E\otimes_A \mu_{H_B}} \mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A E\otimes_A H_B\otimes_A \mathrm{End}_A(E)\\
&\xrightarrow{\sim} \mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A H_B\otimes_A \mathrm{End}_A(E)\otimes_A E\\
&\xrightarrow{\mathrm{id}\otimes_A \mathrm{id}\otimes_A \mathrm{ev}} \mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A H_B\otimes_A E,
\end{aligned}$$

where the isomorphisms without names are the appropriate permutations of the factors in the tensor products. Let Q^s be the residue algebra of $\mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A H_B$ modulo the ideal $\mathfrak{I}_Z(v-\mathrm{id})$ and let

$$\rho_{Q^s} : \mathrm{Sym}_A(F\otimes_A E^\vee)\otimes_A H_B\to Q^s$$

be the residue homomorphism.

Denote by $\rho_1 : \mathrm{Sym}_A(F\otimes_A E^\vee)\to Q^s$ and $\rho_2 : H_B\to Q^s$ the A-algebra homomorphisms that determine ρ_{Q^s}, that is, $\rho_1(f)=\rho_{Q^s}(f\otimes_A 1)$ and $\rho_2(g)=\rho_{Q^s}(1\otimes g)$ for all $f\in\mathrm{Sym}_A(F\otimes_A E^\vee)$ and $g\in H_B$. We obtain a universal Q^s-algebra homomorphism

$$\mu_{Q^s} : Q^s\otimes_A B\to Q^s\otimes_A \mathrm{End}_A(E)$$

defined by $\mu_{Q^s}(1_{Q^s}\otimes_A f)=(\rho_2\otimes_A \mathrm{id}_{\mathrm{End}_A(E)})\mu_{H_B}(1\otimes_A f)$ for all $f\in B$ and a Q^s-module homomorphism $u'_{Q^s} : Q^s\otimes_A F\to Q^s\otimes_A E$ defined by $u'_{Q^s}(1_{Q^s}\otimes_A y)=$

$(\rho_1 \otimes_A \mathrm{id}_E) u_E (1 \otimes_A y)$ for all $y \in F$. When we give $Q^s \otimes_A E$ the $Q^s \otimes_A B$-module structure given by μ_{Q^s} we denote by

$$u_{Q^s} : Q^s \otimes_A B \otimes_A F \to Q^s \otimes_A E$$

the $Q^s \otimes_A B$-module homomorphism obtained from u'_{Q^s} by extension of scalars. Identifying $Q^s \otimes_A B \otimes_A F$ with $(Q^s \otimes_A B) \otimes_{Q^s} (Q^s \otimes_A F)$ and $Q^s \otimes_A \mathrm{End}_A(E) \otimes_A E$ with $(Q^s \otimes_A \mathrm{End}_A(E)) \otimes_{Q^s} (Q^s \otimes_A E)$ we obtain a composite homomorphism

$$Q^s \otimes_A E \xrightarrow{\mathrm{id}_{Q^s} \otimes_A s} Q^s \otimes_A B \otimes_A F \xrightarrow{\mu_{Q^s} \otimes_A u'_{Q^s}} Q^s \otimes_A \mathrm{End}_A(E) \otimes_A E \xrightarrow{\mathrm{id}_{Q^s} \otimes_A \mathrm{ev}} Q^s \otimes_A E$$

that is the identity by the definition of $\mathfrak{I}_Z(v - \mathrm{id})$ and Q^s.

Theorem 3.2. *The A-algebra Q^s represents the local quot functor defined in Section 1.1. The universal homomorphisms are $\mu_{Q^s} : Q^s \otimes_A B \to Q^s \otimes_A \mathrm{End}_A(E)$ and $u_{Q^s} : Q^s \otimes_A B \otimes_A F \to Q^s \otimes_A E$.*

More precisely, let $A \to A'$ be an A-algebra. We have bijections between the following three sets:

(1) *$A' \otimes_A B$-module structures on $A' \otimes_A E$ and $A' \otimes_A B$-linear homomorphisms $u : A' \otimes_A B \otimes_A F \to A' \otimes_A E$ for the $A' \otimes_A B$-module structure such that*

$$A' \otimes_A E \xrightarrow{1 \otimes_A s} A' \otimes_A B \otimes_A F \xrightarrow{u} A' \otimes_A E$$

is the identity.

(2) *A'-module homomorphisms $v : A' \otimes_A F \to A' \otimes_A E$ and A'-algebra homomorphisms $\varphi : A' \otimes_A B \to A' \otimes_A \mathrm{End}(E)$ such that the composite homomorphism*

(3-1) $$A' \otimes_A E \xrightarrow{\mathrm{id}_{A'} \otimes s} A' \otimes_A B \otimes_A F \xrightarrow{\varphi \otimes_A v} A' \otimes_A \mathrm{End}_A(E) \otimes_A E \xrightarrow{\mathrm{id}_{A'} \otimes_A \mathrm{ev}} A' \otimes_A E$$

is the identity.

(3) *A-algebra homomorphisms $Q^s \to A'$.*

The bijection from the set (2) *to the set* (1) *is described in Proposition 1.3 and the bijection from the set* (3) *to the set* (2) *is defined as follows*:

Let $\varphi : Q^s \to A'$ be an A-algebra homomorphism. The homomorphism $\varphi \rho_{Q^s} : \mathrm{Sym}_A(F \otimes_A E^\vee) \otimes_A H_B \to A'$ is determined by A-algebra homomorphisms $\rho_1 : \mathrm{Sym}_A(F \otimes_A E^\vee) \to A'$ and $\rho_2 : H_B \to A'$. The homomorphism ρ_2 defines an A'-algebra homomorphism $\psi : A' \otimes_A B \to A' \otimes_A \mathrm{End}_A(E)$ by Proposition 2.8, and ρ_1 defines, by Lemma 2.2 with $E = K$ and $G = F$, an A'-module homomorphism $v : A' \otimes_A F \to A' \otimes_A E$. We extend v to an $A' \otimes_A B$-module homomorphism $A' \otimes_A B \otimes_A F \to A' \otimes_A E$, when $A' \otimes_A E$ has the $A' \otimes_A B$-module structure given by ψ.

Proof. The bijection between (1) and (2) is given in Corollary 1.4 when we use the canonical isomorphism of A'-module $A' \otimes_A \mathrm{End}_A(E) \to \mathrm{End}_{A'}(A' \otimes_A E)$.

From the description of the map from the set in (3) to the set in (2) given in the Theorem it follows that the map is a bijection, since an A-algebra homomorphism $\mathrm{Sym}_A(F \otimes_A E^\vee) \otimes_A H_B \to A'$ factors via ρ_{Q^s} if and only if the composite homomorphism (3-1) is the identity.

That the map from the set (3) to the set (1) is functorial follows from the explicit description of the maps in the Theorem. □

4. The quot functor

Definition 4.1. Let $f : X \to S$ be a morphism of schemes. For every morphism $g : S' \to S$ we write

$$\begin{array}{ccc} X' = X \times_S S' & \xrightarrow{g'} & X \\ \downarrow{\scriptstyle f'} & & \downarrow{\scriptstyle f} \\ S' & \xrightarrow{g} & S. \end{array}$$

Let $\mathcal{M}$ be a quasicoherent $\mathcal{O}_X$-module that is flat over S and such that $\mathrm{Supp}\,\mathcal{M}$ is a scheme that is finite over S. Here $\mathrm{Supp}\,\mathcal{M}$ is the subscheme of X defined by the *annihilator* of $\mathcal{M}$. We say that $\mathcal{M}$ is of *relative rank* n over S if $f_*\mathcal{M}$ is a locally free $\mathcal{O}_S$-module of rank n. The latter condition is equivalent to the condition that $\dim_{\kappa(s)}(f_*\mathcal{M} \otimes_{\mathcal{O}_X} \kappa(s)) = n$ for all points $s \in S$ (see [Laksov et al. 2000]).

Let $\mathcal{F}$ be a quasicoherent $\mathcal{O}_X$-module. We denote by $\mathrm{Quot}^n_{\mathcal{F}/X/S}$ the functor from S-schemes to sets that to a morphism $g : S' \to S$ associates the set $\mathrm{Quot}^n_{\mathcal{F}/X/S}(S')$ of classes of surjections $g'^*\mathcal{F} \to \mathcal{M}$ of $\mathcal{O}_{X'}$-modules, where $\mathcal{M}$ is a coherent $\mathcal{O}_{X'}$-module which is flat over S' with support that is a finite scheme over S' with *relative rank* n (see [Grothendieck 1966] or [Deligne 1973]).

Let $\mathcal{E}$ be a locally free $\mathcal{O}_S$-module of rank n and fix an $\mathcal{O}_S$-module homomorphism $s : \mathcal{E} \to f_*\mathcal{F}$. We denote by $\mathrm{Quot}^s_{\mathcal{F}/X/S}$ the subfunctor of $\mathrm{Quot}^n_{\mathcal{F}/X/S}$ that to an S-scheme S' associates the set $\mathrm{Quot}^s_{\mathcal{F}/X/S}(S')$ of all equivalence classes of surjections $g'^*\mathcal{F} \to \mathcal{M}$ such that the composite homomorphism

$$g^*\mathcal{E} \xrightarrow{g^*s} g^*f_*\mathcal{F} \to f'_*g'^*\mathcal{F} \to f'_*\mathcal{M}$$

is surjective, that is, an isomorphism.

When convenient we write $\mathrm{Quot}^n_{\mathcal{F}}$ and $\mathrm{Quot}^s_{\mathcal{F}}$ for the functors $\mathrm{Quot}^n_{\mathcal{F}/X/S}$, respectively $\mathrm{Quot}^s_{\mathcal{F}/X/S}$, and indicate in the text that the functors are taken relative to the homomorphism $f : X \to S$.

Lemma 4.2. *Let $A \to B$ be an A-algebra and let M be a B-module that is free of rank n as an A-module. Then $B/\mathrm{Ann}_B(M)$ is integral over A.*

Proof. Since M is finitely generated as an A-module it is finitely generated as a B-module, say by $x_1, \dots, x_n$. We obtain a B-module homomorphism $B \to \prod_{i=1}^n M$ that maps b to $(bx_1, \dots, bx_n)$. This gives an injection $B/\mathrm{Ann}_B(M) \to \prod_{i=1}^n M$ of B-modules. In particular, $\prod_{i=1}^n M$ is a faithful $B/\mathrm{Ann}_B(M)$-module. Moreover, the composite homomorphism $A \to B/\mathrm{Ann}_B(M) \to \prod_{i=1}^n M$ is injective since M is free as an A-module. For every element $f \in B/\mathrm{Ann}_B(M)$, the module $A[f]$ is contained in the finitely generated A-module $\prod_{i=1}^n M$ that is faithful over $A[f]$. Hence f is integral over A; see [Lang 1993, Chapter VII, §1 INT3]. □

Proposition 4.3. *Let $f : X \to S$ be a morphism of affine schemes. Assume that $\mathscr{E}$ is a free $\mathbb{O}_S$-module and that $\mathscr{F} = f^*\mathscr{F}_0$ where $\mathscr{F}_0$ is an $\mathbb{O}_S$-module. Then the functor $\mathsf{Quot}^s_{\mathscr{F}/X/S}$ is representable by Q^s for $S = \mathrm{Spec}(A)$, $X = \mathrm{Spec}(B)$, $E = \Gamma(S, \mathscr{E})$, $F = \Gamma(S, \mathscr{F}_0)$ and $s : E \to B \otimes_A F$ corresponds to the homomorphism s.*

Proof. In the correspondence between affine schemes over S and A-algebras we see that, in order to represent $\mathsf{Quot}^s_{\mathscr{F}/X/S}$, we must represent the functor that to A' associates the $A' \otimes_A B$-module homomorphisms $u : A' \otimes_A B \otimes_A F \to M$ where M is a free A'-module of rank n such that

$$A' \otimes_A E \xrightarrow{\mathrm{id}_{A'} \otimes_A s} A' \otimes_A B \otimes_A F \xrightarrow{u} M$$

is surjective. This functor is representable by Theorem 3.2, taken into account that the condition $\mathrm{Supp}\, M = A' \otimes_A B/\mathrm{Ann}_{A' \otimes_A B}(M)$ is finite over A' is automatically fulfilled by Lemma 4.2. □

Like several of the reductions of this section, the next result is well known.

Lemma 4.4. *Let $f : X \to S$ be a homomorphism of affine schemes. Assume that $\mathscr{E}$ is a free $\mathbb{O}_S$-module. Then $\mathsf{Quot}^s_{\mathscr{F}/X/S}$ is representable.*

More precisely, there is a free $\mathbb{O}_S$-module $\mathscr{F}_0$, a surjection $u : f^*\mathscr{F}_0 \to \mathscr{F}$ of $\mathbb{O}_X$-modules, and a homomorphism of $\mathbb{O}_S$-modules $s_0 : \mathscr{E} \to f_*f^*\mathscr{F}_0$ such that $s = (f_*u)s_0$. These homomorphisms make $\mathsf{Quot}^s_{\mathscr{F}/X/S}$ into a closed subfunctor of $\mathsf{Quot}^{s_0}_{f^*\mathscr{F}_0/X/S}$.

Proof. Let $F = \Gamma(X, \mathscr{F})$ and $E = \Gamma(S, \mathscr{E})$. Choose a surjection of A-modules $F_0 \to F$ with F_0 free. We then obtain, by extension of scalars, a surjection of B-modules $B \otimes_A F_0 \to F$, and consequently a surjection $u : f^*\mathscr{F}_0 \to \mathscr{F}$ of $\mathbb{O}_X$-modules with $\mathscr{F}_0 = \widetilde{F_0}$. We lift $s : E \to F$ to an A-module homomorphism $E \to B \otimes_A F_0$ via the surjection $B \otimes_A F_0 \to F$. The corresponding lifting $s_0 : \mathscr{E} \to f_*f^*\mathscr{F}_0$ has the property that $s = (f_*u)s_0$.

Clearly, given an arbitrary A-algebra $A \to A'$, we have a map $\mathsf{Quot}^n_{\mathscr{F}/X/S}(A') \to \mathsf{Quot}^n_{f^*\mathscr{F}_0/X/S}(A')$ that takes $\mathscr{F} \xrightarrow{v} \mathscr{M}$ to the composite homomorphism

$$f^*\mathscr{F}_0 \xrightarrow{u} \mathscr{F} \xrightarrow{v} \mathscr{M},$$

and this map defines a closed immersion of functors $\mathrm{Quot}^n_{\mathscr{F}/X/S} \to \mathrm{Quot}^n_{f^*\mathscr{F}_0/X/S}$. Moreover the closed immersion maps $\mathrm{Quot}^s_{\mathscr{F}/X/S}$ into $\mathrm{Quot}^{s_0}_{f^*\mathscr{F}_0/X/S}$ because

$$\mathscr{E} \xrightarrow{s} f_*\mathscr{F} \xrightarrow{v} \mathscr{M}$$

is surjective if and only if $\mathscr{E} \xrightarrow{s_0} f_*f^*\mathscr{F}_0 \xrightarrow{f_*u} f_*\mathscr{F} \xrightarrow{v} \mathscr{M}$ is surjective.

For the representability of $\mathrm{Quot}^{s_0}_{f^*\mathscr{F}_0/X/S}$, use Theorem 3.2 and Proposition 4.3. □

Proposition 4.5. *Let $f : X \to S$ be a homomorphism of affine schemes. Then $\mathrm{Quot}^n_{\mathscr{F}/X/S}$ is representable.*

More precisely, the functor $\mathrm{Quot}^n_{\mathscr{F}/X/S}$ is covered by open affine subfunctors $\mathrm{Quot}^s_{\mathscr{F}/X/S}$ for all $\mathcal{O}_S$-module homomorphisms $s : \mathscr{E} \to f_*\mathscr{F}$, where $\mathscr{E}$ is a free $\mathcal{O}_S$-module of rank n.

Proof. It is clear that the subfunctors $\mathrm{Quot}^s_{\mathscr{F}/X/S}$ of $\mathrm{Quot}^n_{\mathscr{F}/X/S}$ are open. We have to show that $\mathrm{Quot}^n_{\mathscr{F}/X/S}$ is covered by these functors. Let $S = \mathrm{Spec}(A)$, $X = \mathrm{Spec}(B)$, $F = \Gamma(\mathrm{Spec}(B), \mathscr{F})$ and $E = \Gamma(\mathrm{Spec}(A), \mathscr{E})$. For every A-algebra $A \to A'$ we let the A'-module homomorphism $u : A' \otimes_A F \to M$ correspond to an element $g'^*\mathscr{F} \to \mathscr{M}$ in $\mathrm{Quot}^n_{\mathscr{F}/X/S}(\mathrm{Spec}(A'))$. For every maximal ideal $\mathfrak{p}$ in A' we can, since M is a finitely generated A'-module, find an element $a' \in A' \setminus \mathfrak{p}$ and an A-module homomorphism $s_M : E \to F$ such that the composite $A'_{a'}$-module homomorphism $A'_{a'} \otimes_A E \xrightarrow{s_M} A'_{a'} \otimes_A F \to M_{a'}$ is surjective. This shows that the image $g'^*\mathscr{F}|f'^{-1}\,\mathrm{Spec}(A'_{a'}) \to \mathscr{M}|f'^{-1}\,\mathrm{Spec}(A'_{a'})$ of the element $g'^*\mathscr{F} \to \mathscr{M}$ by the map $\mathrm{Quot}^n_{\mathscr{F}/X/S}(\mathrm{Spec}(A')) \to \mathrm{Quot}^n_{\mathscr{F}/X/S}(\mathrm{Spec}(A'_{a'}))$ lies in $\mathrm{Quot}^s_{\mathscr{F}/X/S}(\mathrm{Spec}(A'_{a'}))$ so the set $\mathrm{Quot}^s_{\mathscr{F}/X/S}(\mathrm{Spec}(A'_{a'}))$ *covers* $g'^*\mathscr{F}|f'^{-1}\,\mathrm{Spec}(A'_{a'}) \to \mathscr{M}|f'^{-1}\,\mathrm{Spec}(A'_{a'})$ considered as an element in $\mathrm{Quot}^n_{\mathscr{F}/X/S}(\mathrm{Spec}(A'_{a'}))$.

The representability of $\mathrm{Quot}^s_{\mathscr{F}/X/S}$ follows from Lemma 4.4 and it follows that the Zariski sheaf $\mathrm{Quot}^n_{\mathscr{F}/X/S}$ is representable. □

Lemma 4.6. *Let R be a graded A-algebra. For every prime ideal $\mathfrak{p}$ of A write $\kappa(\mathfrak{p}) = A_\mathfrak{p}/\mathfrak{p}A_\mathfrak{p}$.*

Let Z be a closed subscheme of $\mathrm{Proj}(\kappa(\mathfrak{p}) \otimes_A R)$ that is finite over $\mathrm{Spec}(\kappa(\mathfrak{p}))$. Then there is an element $a \in A$ not in $\mathfrak{p}$ and an element $f \in R_a$ such that Z is contained in the open subscheme $\mathrm{Spec}(\kappa(\mathfrak{p}) \otimes_{A_a} (R_a)_{(f)}) = \mathrm{Spec}(\kappa(\mathfrak{p})) \times_{\mathrm{Spec}(A_a)} \mathrm{Spec}((R_a)_{(f)})$ of $\mathrm{Proj}(\kappa(\mathfrak{p}) \otimes_{A_a} R_a) = \mathrm{Spec}(\kappa(\mathfrak{p})) \times_{\mathrm{Spec}(A_a)} \mathrm{Proj}(R_a)$.

Proof. Since Z is finite over $\mathrm{Spec}(\kappa(\mathfrak{p}))$ the fiber of the induced morphism $Z \to \mathrm{Spec}(\kappa(\mathfrak{p}))$ consists of a finite number of points, corresponding to homogeneous prime ideals $\mathfrak{q}_1, \ldots, \mathfrak{q}_k$ in $\kappa(\mathfrak{p}) \otimes_A R$ that do not contain the irrelevant ideal. Their union consequently does not contain the irrelevant ideal. Hence we can find a homogeneous element $g \in \kappa(\mathfrak{p}) \otimes_A R$ of positive degree that is not contained in

any of the ideals $\mathfrak{q}_1, \ldots, \mathfrak{q}_k$. Thus Z is contained in the open subscheme

$$\operatorname{Spec}((\kappa(\mathfrak{p}) \otimes_A R)_{(g)})$$

of $\operatorname{Proj}(\kappa(\mathfrak{p}) \otimes_A R)$.

Clearly we can find an element $a \in A$ not in $\mathfrak{p}$ and an element $f \in R_a$ such that $1_{\kappa(\mathfrak{p})} \otimes_{A_a} f$ is the image of g by the natural isomorphism $\kappa(\mathfrak{p}) \otimes_A R \to \kappa(\mathfrak{p}) \otimes_{A_a} R_a$. However, then $\operatorname{Spec}((\kappa(\mathfrak{p}) \otimes_A R)_{(g)}) = \operatorname{Spec}(\kappa(\mathfrak{p}) \otimes_{A_a} (R_a)_{(f)})$, and we have proved the Lemma. □

Theorem 4.7. *Let $\mathcal{R}$ be a quasicoherent sheaf that is a graded $\mathcal{O}_S$-algebra, let $X = \operatorname{Proj}(\mathcal{R})$, and let $\mathcal{F}$ be a quasicoherent $\mathcal{O}_X$-module. Then $\operatorname{Quot}^n_{\mathcal{F}/X/S}$ is representable.*

More precisely, for every open affine subscheme $\operatorname{Spec}(A)$ of S we write $R = \Gamma(\operatorname{Spec}(A), \mathcal{R})$. Then $\operatorname{Quot}^n_{\mathcal{F}/X/S}$ is covered by open subfunctors naturally isomorphic to $\operatorname{Quot}^n_{\mathcal{F}|\operatorname{Spec}(R_{(r)})}$ relative to $\operatorname{Spec}(R_{(r)}) \to \operatorname{Spec}(A)$ for all $\operatorname{Spec}(A)$ in an open covering of S and all homogeneous elements r of positive degree in R.

Proof. For every affine open subset U of S we consider $\operatorname{Quot}^n_{\mathcal{F}|f^{-1}(U)}$ relative to $f^{-1}(U) \to U$ as a subfunctor of $\operatorname{Quot}_{\mathcal{F}/\operatorname{Proj}(\mathcal{R})/S}$ by letting

$$\operatorname{Quot}^n_{\mathcal{F}|f^{-1}(U)}(S') = \varnothing$$

when $g : S' \to S$ does not factor via U. It is clear that the subfunctors $\operatorname{Quot}^n_{\mathcal{F}|f^{-1}(U)}$ are open and that they cover $\operatorname{Quot}^n_{\mathcal{F}/\operatorname{Proj}(\mathcal{R})/S}$. Consequently we can assume that $S = \operatorname{Spec}(A)$ is affine.

For every $a \in A$ and every $r \in R_a$ we can consider the functor $\operatorname{Quot}_{a,r} = \operatorname{Quot}^n_{\mathcal{F}|\operatorname{Spec}((R_a)_{(r)})}$ relative to $\operatorname{Spec}(R_a)_{(r)} \to \operatorname{Spec}(A_a)$ as a subfunctor of the functor $\operatorname{Quot}_a = \operatorname{Quot}^n_{\mathcal{F}|\operatorname{Proj}(R_a)}$ relative to $\operatorname{Proj}(R_a) \to \operatorname{Spec}(A_a)$. This is because, if $g : S' \to \operatorname{Spec}(A_a)$ is a morphism and $g'^*\mathcal{F}|f^{-1}(\operatorname{Spec}(R_a)_{(r)}) \to \mathcal{M}$ represents an element in $\operatorname{Quot}_{a,r}(S')$, then $\operatorname{Supp}\mathcal{M} \subseteq f^{-1}(\operatorname{Spec}((R_a)_{(r)}))$ and $\operatorname{Supp}\mathcal{M}$ is finite over $\operatorname{Spec}(A_a)$, and $\operatorname{Spec}((R_a)_{(r)}) \to \operatorname{Spec}(A_a)$ is separated so $\operatorname{Supp}\mathcal{M}$ is closed in $\operatorname{Proj}(R_a) \times_{\operatorname{Spec}(A_a)} S'$. Thus we can extend $\mathcal{M}$ uniquely by zero to an $\mathcal{O}_{\operatorname{Proj}(R_a)\times_{\operatorname{Spec}(A_a)}S'}$-module $\mathcal{N}$ and the surjection $g'^*\mathcal{F}|f'^{-1}(\operatorname{Spec}((R_a)_{(r)})) \to \mathcal{M}$ can be extended to a surjection $g'^*\mathcal{F} \to \mathcal{N}$ representing an element in $\operatorname{Quot}_a(\operatorname{Spec}(A_a))$. It is clear that the subfunctors $\operatorname{Quot}_{a,r}$ are open. It remains to show that they cover $\operatorname{Quot}_{\mathcal{F}/\operatorname{Proj}(R)/\operatorname{Spec}(A)}$. For this it suffices to show that for every prime ideal $\mathfrak{p} \subset A$ and every surjection $g'^*\mathcal{F} \to \mathcal{M}$, with $g' : \operatorname{Spec}(\kappa(\mathfrak{p})) \times_S \operatorname{Proj}(R_a) \to S' = \operatorname{Spec}(\kappa(\mathfrak{p}))$, where $\mathcal{M}$ has finite support over $\operatorname{Spec}(\kappa(\mathfrak{p}))$ of relative rank n, there is an $a \in A \setminus \mathfrak{p}$ and a homogeneous element r in R_a such that the support of $\mathcal{M}$ is contained in the open subscheme $\operatorname{Spec}(\kappa(\mathfrak{p}) \otimes_A (R_a)_{(r)}) = \operatorname{Spec}(\kappa(\mathfrak{p})) \times_{\operatorname{Spec}(A_a)} \operatorname{Spec}((R_a)_{(r)})$ of $\operatorname{Proj}(\kappa(\mathfrak{p}) \otimes_{A_a} R_a) = \operatorname{Spec}(\kappa(\mathfrak{p})) \times_{\operatorname{Spec}(A_a)} \operatorname{Proj}(R_a)$. However this is the assertion of Lemma 4.6. □

5. Applications

5.1. ***Special cases.*** The two best known special cases of Theorem 4.7 are the Hilbert schemes and the Grassmann schemes. As in the Theorem we let $X = \mathrm{Proj}(\mathcal{R})$ with $\mathcal{R}$ quasicoherent on the S-scheme X, and $\mathcal{F}$ is a quasicoherent $\mathcal{O}_X$-module.

When $\mathcal{F} = \mathcal{O}_X$ we obtain the existence and construction of the Hilbert scheme $\mathrm{Hilb}^n_{X/S}$ of n points in X, as in [Gustavsen et al. 2007].

When $X = S$ the Theorem gives the existence and the standard description of the Grassmann scheme $\mathrm{Grass}^n_{\mathcal{F}/S}$ of n-quotients of the quasicoherent $\mathcal{O}_X$-module $\mathcal{F}$.

5.2. ***Example: Quot schemes over the affine line.*** Let $A[T]$ be the polynomial ring over A in the variable T. Moreover, let F be a free module of rank r, and let $S = \mathrm{Spec}(A)$, $X = \mathrm{Spec}(A[T])$ and $\mathcal{F} = A[T] \otimes_A F$. We shall show how our construction gives an open covering of $\mathrm{Quot}^n_{\mathcal{F}/X/S}$ by affine spaces of relative dimension rn over $\mathrm{Spec}(A)$, with an intersection that contains an affine space of relative dimension n. In particular, since $\mathrm{Quot}^n_{\mathcal{F}/X/S}$ can be covered by mutually intersecting open sets that are affine spaces over $\mathrm{Spec}(A)$ of relative dimension rn, it is irreducible of relative dimension rn if and only if $\mathrm{Spec}(A)$ is irreducible.

Let E be a vector space with basis $e_1, \ldots, e_n$, and let $f_1, \ldots, f_r$ be a basis for F. From Proposition 4.5 we have that $\mathrm{Quot}^n_{\mathcal{F}/X/S}$ can be covered by open affine subsets $\mathrm{Quot}^s_{\mathcal{F}/X/S}$, where the maps $s : E \to A[T] \otimes_A F$ are linear.

Let X_{ij} for $i, j = 1, \ldots, n$ and Y_{ij} for $i = 1, \ldots, n$, $j = 1, \ldots, r$ be independent variables over A, and let (X_{ij}), be the $n \times n$-matrix with coordinates X_{ij}, and (Y_{ij}) the n-vector with coordinates Y_{ij} for fixed j. We denote by C the ring of polynomials over A in all these variables, and define maps

$$\varphi : C \otimes_A A[T] \to C \otimes_A \mathrm{End}_A(E) \quad \text{and} \quad v : C \otimes_A F \to C \otimes_A E$$

of C-algebras, respectively of C-modules, by $\varphi(T) = (X_{ij})$ with respect to the basis $e_1, \ldots, e_n$, respectively by $v(1 \otimes f_j) = (Y_{ij})$, with respect to the bases $e_1, \ldots, e_n$ and $f_1, \ldots, f_r$.

In Sections 2 and 3 we proved that $\mathrm{Quot}^s_{\mathcal{F}/X/S}$ is given as the residue algebra of C modulo the ideal generated by the relations we obtain requiring that the composite homomorphism

(5-1)
$$C \otimes_A E \xrightarrow{\mathrm{id}_C \otimes_A s} C \otimes_A A[T] \otimes_A F \xrightarrow{\varphi \otimes_A v} C \otimes_A \mathrm{End}_A(E) \otimes_A E \xrightarrow{\mathrm{id}_C \otimes_A \mathrm{ev}} C \otimes_A E$$

be the identity. It follows from the Cayley–Hamilton Theorem used on $\varphi(T) = (X_{ij})$ that $\psi := (\mathrm{id}_C \otimes_A \mathrm{ev})(\varphi \otimes_A v)$ is surjective if and only if its restriction to the C-submodule of $C \otimes_A A[T] \otimes_A F$ generated by the elements $1 \otimes T^i \otimes f_j$, for

$i = 0, \dots, n-1$ and $j = 1, \dots, r$, is surjective. We will describe a collection of maps $s : E \to A[T] \otimes_A F$ so that $\mathrm{Quot}^s_{\mathscr{F}/X/S}$ cover $\mathrm{Quot}^n_{\mathscr{F}/X/S}$. To do this we may clearly assume that s maps the elements of $e_1, \dots, e_n$ to the elements of the form $T^i \otimes f_j$ for $i = 0, \dots, n-1$ and $j = 1, \dots, r$.

Let s be given, and let p_j be the number of elements of the form $T^i \otimes f_j$, for fixed j that are in the image by s of the elements $e_1, \dots, e_n$. In particular $p_1 + \cdots + p_r = n$. Since ψ is a $C[T]$-module homomorphism for the $C[T]$-module structure on $C \otimes_A E$ given by φ we conclude that, if the restriction of ψ to the C-module generated by $s(e_1), \dots, s(e_n)$ is surjective, then ψ restricted to the C-module generated by $1 \otimes f_j, \dots, T^{p_j - 1} \otimes f_j$ for $j = 1, \dots, r$ is surjective. Consequently, we can cover $\mathrm{Quot}^n_{\mathscr{F}/X/S}$ by the sets $\mathrm{Quot}^s_{\mathscr{F}/X/S}$ where s is given by

$$s(e_{q_{j-1}+i+1}) = T^i \otimes f_j \quad \text{for} \quad j = 1, \dots, r \quad \text{and} \quad i = 0, \dots, p_j - 1 \tag{5-2}$$

with $q_j = p_1 + \cdots + p_j$ and $q_0 = 0$.

The image of $1 \otimes e_{q_{j-1}+i+1}$ by (5-1) is $\varphi(T)^i v(1 \otimes f_j)$. For $i = 0$ we obtain, in particular, that the condition that (5-1) is the identity is $(Y_{ij}) = e_{q_{j-1}+1}$ for those $j = 1, \dots, r$ that satisfy $p_j \geq 1$. Hence the image of $1 \otimes e_{q_{j-1}+i+1}$ by (5-1) is $\varphi(T)^i e_{q_{j-1}+1}$ for these j. By induction on i we easily verify that the condition that (5-1) is the identity is that column number i in (X_{ij}) is equal to $e_{q_{j-1}+i+1}$ for $i = 1, \dots, p_j - 1$.

Consequently, the condition for Equation (5-1) to be the identity is that

$$v(1 \otimes f_j) = 1 \otimes e_{q_{j-1}+1}$$

for those $j = 1, \dots, r$ that satisfy $p_j \geq 1$, and that $\varphi(T)$ has column number i equal to e_{i+1} when i is different from $q_1, \dots, q_r$, and with no conditions on columns number $q_1, \dots, q_r$. Consequently $\mathrm{Quot}^s_{\mathscr{F}/X/S}$ is affine space of relative dimension nr when s is determined by Equation (5-2). Note that, independently of the choice of s satisfying the conditions (5-2), $\mathrm{Quot}^s_{\mathscr{F}/X/S}$ contains the affine space of relative dimension n consisting of the homomorphisms φ such that $\varphi(T)$ is the companion matrix of the polynomial $T^n - X_{n\,n} T^{n-1} - X_{n-1\,n} T^{n-2} - \cdots - X_{1\,n}$.

References

[Altman and Kleiman 1980] A. B. Altman and S. L. Kleiman, "Compactifying the Picard scheme", *Adv. in Math.* **35**:1 (1980), 50–112. MR 81f:14025a Zbl 0427.14015

[Deligne 1973] P. Deligne, "Cohomologie à supports propres", pp. 250–480 (exposé XVII) in *Théorie des topos et cohomologie étale des schémas* (SGA4), vol. 3, edited by M. Artin et al., Lecture Notes in Mathematics **305**, Springer, Berlin, 1973. MR 50 #7132 Zbl 0255.14011

[Grothendieck 1966] A. Grothendieck, "Techniques de construction et théorèmes d'existence en géométrie algébrique, IV: Les schémas de Hilbert", pp. 249–276 (exposé 221) in *Séminaire Bourbaki 1960/1961*, Benjamin, Paris, 1966. Also in vol. 6 of the reprint by Soc. Math. de France, Paris, 1995.

[Gustavsen et al. 2007] T. S. Gustavsen, D. Laksov, and R. M. Skjelnes, "An elementary, explicit, proof of the existence of Hilbert schemes of points", *J. Pure Appl. Algebra* **210**:3 (2007), 705–720.

[Huibregtse 2002] M. E. Huibregtse, "A description of certain affine open subschemes that form an open covering of $\mathrm{Hilb}^n_{\mathbf{A}^2_k}$", *Pacific J. Math.* **204** (2002), 97–143. MR 2003h:14008 Zbl 1068.14007

[Huibregtse 2006] M. E. Huibregtse, "An elementary construction of the multigraded Hilbert scheme of points", *Pacific J. Math.* **223**:2 (2006), 269–315. MR MR2221028 Zbl 05129581

[Laksov et al. 2000] D. Laksov, Y. Pitteloud, and R. M. Skjelnes, "Notes on flatness and the Quot functor on rings", *Comm. Algebra* **28**:12 (2000), 5613–5627. MR 2001i:13009 Zbl 0990.13005

[Lang 1993] S. Lang, *Algebra*, 3rd ed., Addison-Wesley, Reading, MA, 1993. MR 33 #5416 Zbl 0848.13001

[Mumford 1966] D. Mumford, *Lectures on curves on an algebraic surface*, Annals of Mathematics Studies **59**, Princeton Univ. Press, Princeton, NJ, 1966. MR 35 #187 Zbl 0187.42701

[Sernesi 2006] E. Sernesi, *Deformations of algebraic schemes*, Grundlehren der Math. Wiss **334**, Springer, Berlin, 2006. MR MR2247603 Zbl 1102.14001

Received January 28, 2007. Revised May 16, 2007.

TROND STØLEN GUSTAVSEN
DEPARTMENT OF ECONOMICS
BI NORWEGIAN SCHOOL OF MANAGEMENT
N-0442 OSLO
NORWAY

stolen@math.uio.no

DAN LAKSOV
DEPARTMENT OF MATHEMATICS
KTH
S-100 44 STOCKHOLM
SWEDEN

laksov@math.kth.se

ROY MIKAEL SKJELNES
DEPARTMENT OF MATHEMATICS
KTH
S-100 44 STOCKHOLM
SWEDEN

skjelnes@math.kth.se

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.417

SYMPLECTIC ENERGY AND LAGRANGIAN INTERSECTION UNDER LEGENDRIAN DEFORMATIONS

HAI-LONG HER

Let M be a compact symplectic manifold, and $L \subset M$ be a closed Lagrangian submanifold which can be lifted to a Legendrian submanifold in the contactization of M. For any Legendrian deformation of L satisfying some given conditions, we get a new Lagrangian submanifold L'. We prove that the number of intersection $L \cap L'$ can be estimated from below by the sum of $\mathbb{Z}_2$-Betti numbers of L, provided they intersect transversally.

1. Introduction

V. I. Arnold formulated in [1965; 1978, Appendix 9] his famous conjectures on the number of fixed points of Hamiltonian diffeomorphisms of any compact symplectic manifold and the number of intersection points of any Lagrangian submanifold with its Hamiltonian deformations in a symplectic manifold. If M is a symplectic manifold, $L \subset M$ is a Lagrangian submanifold, and ψ_M is a Hamiltonian diffeomorphism, his conjectures can be written in topological terms as

$\#\mathrm{Fix}(\psi_M) \geq$ sum of Betti numbers of M, with all fixed points nondegenerate;
$\#\mathrm{Fix}(\psi_M) \geq$ cuplength of M, fixed points possibly degenerate;

$\#(L \cap \psi_M(L)) \geq$ sum of Betti numbers of L, with intersection points transverse;
$\#(L \cap \psi_M(L)) \geq$ cuplength of L, with intersection points possibly nontransverse.

Much effort has gone into proving these two conjectures. The pioneering works are [Conley and Zehnder 1983; Gromov 1985; Floer 1988a; 1988b; 1989a; 1989b]. Floer originally developed the seminal method, motivated by the variational method used by Conley and Zehnder and the elliptic PDE techniques introduced by Gromov, which is now called Floer homology theory, and solved many special cases of Arnold's conjectures. Fukaya and Ono [1999] and Liu and Tian [1998] independently proved the first conjecture for general compact symplectic manifolds in the nondegenerate case. The conjecture for general symplectic manifolds in the degenerate case is still open.

MSC2000: 57R22, 53D40, 53D12, 53D10.
Keywords: Arnold conjecture, Floer homology, Lagrangian intersection, Symplectic energy.

For the second conjecture, Floer [1988a; 1989b] gave the proof under the additional assumption $\pi_2(M, L) = 0$. We write his result for the case that all intersections are transverse.

Theorem 1.1 (Floer). *Let L be a closed Lagrangian submanifold of a compact (or tame) symplectic manifold (M, ω) satisfying $\pi_2(M, L) = 0$, and ψ_M be a Hamiltonian diffeomorphism, then $\#(L \cap \psi_M(L)) \geq \dim H_*(L, \mathbb{Z}_2)$, if all intersections are transverse.*

In general, the condition $\pi_2(M, L) = 0$ can not be removed. For instance, let L be a circle in $\mathbb{R}^2$, then $\pi_2(\mathbb{R}^2, L) \neq 0$, however, there always exists a Hamiltonian diffeomorphism which can translate L arbitrarily far from its original position.

To prove his theorem, Floer introduced the so-called Floer homology group for Lagrangian pairs and showed that it is isomorphic to the homology of L under the condition above. The definition of Floer homology for Lagrangian pairs was generalized by Oh [1993a; 1993b; 1995] in the class of monotone Lagragian submanifolds with minimal Maslov number being at least 3. However, for general Lagrangian pairs, the Floer homology is hard to define due to the bubbling off phenomenon and some essentially topological obstructions [Fukaya et al. 2000], which is much different from the Hamiltonian fixed point case.

Therefore, if we want to throw away the additional assumption, we have to restrict the class of Hamiltonian diffeomorphisms. For the simplest case that ψ_M is C^0-small perturbation of the identity, the Lagrangian intersection problem is equivalent to the one for zero sections of cotangent bundles, which is proved by Hofer [1985] and Laudenbach and Sikorav [1985]. Yu. V. Chekanov [1996; 1998] also gave a version of Lagrangian intersection theorem which used the notion of symplectic energy introduced by Hofer [1990] (for $(\mathbb{R}^{2n}, \omega_0)$) and Lalonde and McDuff [1995] (for general symplectic manifolds). Following their notations, we denote by $\mathcal{H}(M)$ the space of compactly supported smooth functions on $[0, 1] \times M$. Any $H \in \mathcal{H}(M)$ defines a time dependent Hamiltonian flow ϕ_H^t on M. All such time-1 maps $\{\phi_H^1, H \in \mathcal{H}(M)\}$ form a group, denoted by $\mathrm{Ham}(M)$. Now we define a norm on $\mathcal{H}(M)$:

$$\|H\| = \int_0^1 \left(\max_{x \in M} H(t, x) - \min_{x \in M} H(t, x)\right) dt,$$

and we can define the energy of a $\psi \in \mathrm{Ham}(M)$ by

$$E(\psi) = \inf_H \left\{\|H\| \mid \psi = \phi_H^1, H \in \mathcal{H}(M)\right\}.$$

For a compact symplectic manifold (M, ω), there always exists an almost complex structure J compatible with ω, so (M, ω, J) is a compact almost complex manifold, and we denote by $\mathcal{J}$ the set of all such J. Let $\sigma_S(M, J)$ and $\sigma_D(M, L, J)$ denote the minimal area of a J-holomorphic sphere in M and of a J-holomorphic

disc in M with boundary in L, respectively. If there is no such J-holomorphic curve, these values will be infinity. Otherwise, minimums are obtained by the Gromov compactness theorem [1985], and they are always positive. We write $\sigma(M, L, J) = \min(\sigma_S(M, J), \sigma_D(M, L, J))$, and $\sigma(M, L) = \sup_{J\in\mathcal{J}} \sigma(M, L, J)$.

Theorem 1.2 [Chekanov 1998]. *If $E(\psi) < \sigma(M, L)$, then*

$$\#(L \cap \psi(L)) \geq \dim H_*(L, \mathbb{Z}_2),$$

provided all intersections are transverse.

Remark. For the nontransverse case, under similar assumptions, C.-G. Liu [2005] also found an estimate for Lagrangian intersections by cup-length of L.

In this paper, we give an analogous Lagrangian intersection theorem, but the Hamiltonian deformation ψ will be replaced by a "Legendrian deformation" $\tilde{\psi}$ (which will be explained in the sequel). In fact, K. Ono has shown such a result, still under the assumption $\pi_2(M, L) = 0$, as we now discuss.

Suppose the symplectic structure ω is in an integral cohomology class and there exists a principal circle bundle $\pi : N \to M$ with a connection so that the curvature form is ω. This means for a connection form α, one has $d\alpha = \pi^*\omega$. We see that the horizontal distribution $\xi = \operatorname{Ker}\alpha$ is a cooriented contact structure on N. We say L satisfies the Bohr–Sommerfeld condition if $\alpha|_L$ is flat, or, in other words, it can be lifted to a Legendrian submanifold Λ in N.

Theorem 1.3 [Ono 1996]. *Let a contact isotopy $\{\tilde{\psi}_t \mid 0 \leq t \leq 1\}$ be given on N. If L is a Lagrangian submanifold of M that can be lifted to a Legendrian submanifold Λ in N, and $\pi_2(M, L) = 0$, then*

$$\#(L \cap \pi \circ \tilde{\psi}_1(\Lambda)) \geq \dim H_*(L, \mathbb{Z}_2),$$

provided L and $\pi \circ \tilde{\psi}_1(\Lambda)$ intersect transversally.

Remark. Since a Hamiltonian isotopy of M can be lifted to a contact isotopy of N, Ono's theorem is a generalization of the theorem of Floer already stated.

Eliashberg, Hofer, and Salamon [Eliashberg et al. 1995] independently obtained a result similar to Ono's. They overcame some difficulties due to the noncompactness of the symplectization manifold, and their arguments involve some complicated conditions for avoiding bubbling off.

In the present paper, we discard Ono's assumption that $\pi_2(M, L) = 0$, while adding a restrictive condition on the class of Legendrian deformation $\tilde{\psi}$. Let $\tilde{L}$ be the image of Λ under the principal S^1-action on N. We denote by (SN, ω_ξ) the symplectization of the contact manifold (N, ξ) with cooriented contact structure ξ, where the symplectic structure ω_ξ is induced from the standard 1-form of cotangent bundle T^*N. Then $\tilde{L}$ is a compact Lagrangian submanifold in SN. There is a

natural projection $p : SN \to N$, and each section corresponds to a splitting $SN = N \times \mathbb{R}_+ = N \times (e^{-\infty}, +\infty]$. The contactomorphism $\tilde{\psi}$ can be lifted to a $\mathbb{R}_+$-equivariant Hamiltonian symplectomorphism Ψ on SN. We denote $\mathscr{L} = p^{-1}(\Lambda)$, which is also a Lagrangian submanifold in SN. Then we can see that there is a one-to-one correspondence between $\tilde{L} \cap \Psi(\mathscr{L})$ and $L \cap \pi \circ \tilde{\psi}_1(\Lambda)$. However, the symplectization SN is not compact. So, the ordinary method of Floer Lagrangian intersection needs to be modified.

Following [Ono 1996], we can replace the symplectization (SN, ω_ξ) manifold by another symplectic manifold (Q, Ω), which may be considered as a symplectic filling in the negative end, so Q coincides with SN in the part $N \times [e^{-C}, +\infty] \supset \tilde{L}$, where $C > 0$ is a sufficiently large number. We note that Q is a 2-plane bundle over M and is diffeomorphic to the associated complex line bundle $N \times_{S^1} \mathbb{C}$. We define the compatible almost complex structure by J' on Q in the following way. Since Q is the associated complex line bundle, the connection α on N gives the decomposition of $TQ = \mathrm{Ver}(Q) \oplus \mathrm{Hor}(Q)$. We have a ω-compatible almost complex structure J on M, and then we lift J to an almost complex structure on $\mathrm{Hor}(Q)$. Also we define the almost complex structure on each fiber by choosing the standard complex structure J_0 on complex plane $\mathbb{C}$. Then we let $J' = J \oplus J_0$, so J' is uniquely determined by the choice of J on M and a connection on N. Furthermore, Ono [1996, Section 6] showed that if we choose a generic J on M in the sense of the construction of Floer homology for (M, L), then J' is also a regular or generic almost complex structure on Q. If we write $\Pi : Q \to M$ for the natural projection, then it is a (J', J)-holomorphic map. Therefore, a map $u = \Pi \circ \tilde{u} : \Sigma \to M$ is J-holomorphic if and only if $\tilde{u} : \Sigma \to M$ is J'-holomorphic. We can see that, for $r > 1$, the image of the positive end $N \times \{r\} \subset SN$ in Q is J'-*convex*. Hence we can choose the Ω-compatible almost complex structure so that it coincides with J' outside of a compact set. For simplicity, we denote using J this almost complex structure on Q, if we can do so without danger of confusion.

Ono also proved that there is an a priori C^0-bound for connecting orbits in Q (note that all J-holomorphic curves we are concerned with are contained in a compact subset $K \subset Q$, while K depends on the choice of the contact isotopy $\{\psi_t\}$), and the bubbling off argument can go through as in the case of compact symplectic manifold. So the minimal area of J-holomorphic spheres and J-holomorphic discs bounding Lagrangian submanifolds $\tilde{L}$ and $\mathscr{L}$ can be achieved. We denote it by

$$\begin{aligned}\sigma(Q, \tilde{L}, \mathscr{L}, J) &\\ = \min\big(&\sigma_S(Q, J)|_K, \sigma_D(Q, \tilde{L}, J)|_K, \sigma_D(Q, \mathscr{L}, J)|_K, \sigma_D(Q, \mathscr{L}, \tilde{L}, J)|_K\big)\end{aligned}$$

and we set

$$\sigma(Q, \tilde{L}, \mathscr{L}) = \sup_{J_M \in \mathscr{J}} \sigma(Q, \tilde{L}, \mathscr{L}, J_M \oplus J_0).$$

We will show that we can find a compactly supported Hamiltonian diffeomorphism $\Psi' \in \mathrm{Ham}(Q)$ such that for a compact set K, the two images of Ψ and Ψ' coincide. For detailed explanation, we refer to [Ono 1996], or to Section 2. We denote a contactomorphism by ψ, so our main result is:

Theorem 1.4. *Let M be a compact symplectic manifold and N be the principal S^1-bundle $\pi : N \to M$ defined above. Given a contact isotopy $\{\psi_t \mid 0 \le t \le 1\}$ on N, suppose L is a closed Lagrangian submanifold of M which can be lifted to a Legendrian submanifold Λ in N, and $E(\Psi') < \sigma(Q, \tilde{L}, \mathcal{L})$, then*

$$\#(L \cap \pi \circ \psi_1(\Lambda)) \ge \dim H_*(L, \mathbb{Z}_2),$$

provided L and $\pi \circ \psi_1(\Lambda)$ intersect transversally.

2. Preliminaries on symplectic and contact geometry

We say that a given $(2n+1)$-dimensional manifold N is a contact manifold if there exists a contact structure ξ which is a completely nonintegrable tangent hyperplane distribution. It is obvious that ξ can locally be defined as the kernel of a 1-form α satisfying $\alpha \wedge (d\alpha)^n \neq 0$. If the contact structure is coorientable, then α can be globally defined. We only consider the cooriented contact structure in this paper. The contact manifold is denoted by (N, ξ), and α is called a contact form. A diffeomorphism ψ of N is called a contactomorphism if it preserves the cooriented contact structure ξ. We call $\{\psi_t, 0 \le t \le 1\}$ a contact isotopy if $\psi_0 = \mathrm{id}$ and every ψ_t is a contactomorphism, and $X_t = d\psi_t/dt$ is the contact vector field on N.

For any symplectic manifold (M, ω) there exists an almost complex structure J on M. We say the almost complex is *compatible* with the symplectic manifold if $\omega(J\,\cdot, J\,\cdot) = \omega(\cdot, \cdot)$ and $\omega(\cdot, J\,\cdot) > 0$, which can give the Riemannian metric on M.

Let N be an oriented codimension 1 submanifold in an almost complex manifold (Q, J), and ξ_x be the maximal J-invariant subspace of the tangent space T_xN, then ξ_x has codimension 1. N is said to be *J-convex* if for any defining 1-form α for ξ, we have $d\alpha(v, Jv) > 0$ for all nonzero $v \in \xi_x$. This implies ξ is a contact structure on N. It is a fact that if N is J-convex then no J-holomorphic curve in Q can touch (or be tangent to) N from the inside (from the negative side) [Gromov 1985; McDuff 1991].

Symplectization. We denote by $SN = S_\xi(N)$ the $\mathbb{R}_+$-subbundle of the cotangent bundle T^*N whose fibers at $q \in N$ are all nonzero linear forms in T_q^*N, which is compatible with the contact hyperplane $\xi_q \subset T_qN$. There is a canonical 1-form pdq on T^*N, and if we let $\alpha_\xi = pdq|_{SN}$, then $\omega_\xi = d\alpha_\xi$ is a symplectic structure on SN. Thus, we call $(SN,\ \omega_\xi)$ the symplectization of the contact manifold $(N,\ \xi)$.

We see that a contact form $\alpha : N \to SN$ is a section of this $\mathbb{R}_+$-bundle $p : SN \to N$. Hence we have a splitting $SN = N \times \mathbb{R}_+$.

An n-dimensional submanifold $\Lambda \subset (N, \xi)$ is called Legendrian if it is tangent to the distribution ξ, that is to say, Λ is Legendrian if and only if $\alpha|_\Lambda = 0$. The preimage $\mathcal{L} = p^{-1}(\Lambda)$ is an $\mathbb{R}_+$-invariant Lagrangian cone in (SN, ω_ξ). Conversely, any Lagrangian cone in the symplectization projects onto a Legendrian submanifold in (N, ξ).

SN carries a canonical conformal symplectic $\mathbb{R}_+$-action. Every contactomorphism φ uniquely lifts to a $\mathbb{R}_+$-equivariant symplectomorphism $\tilde{\varphi}$ of SN, which is also a Hamiltonian diffeomorphism of SN. Conversely, each $\mathbb{R}_+$-equivariant symplectomorphism of SN projects to a contactomorphism of (N, ξ). A function F on SN is called a contact Hamiltonian if it is homogeneous of degree 1. that is, if $F(cx) = cF(x)$ for all $c \in \mathbb{R}_+$, $x \in SN$.

The Hamiltonian flow generated by a contact Hamiltonian function is $\mathbb{R}_+$-equivariant; it defines a contact isotopy on (N, ξ). Therefore, any contact isotopy $\{\varphi_t\}$ is generated in this sense by a uniquely defined time-dependant contact Hamiltonian $F_t : SN \to \mathbb{R}$. There is a one-to-one correspondence between a contact vector field X_t and a function on N: $f_t = \alpha(X_t)$, also called a contact Hamiltonian function.

Contactization. If a symplectic manifold (M, ω) is exact ($\omega = d\alpha$), it can be contactized. The contactization $C(M, \omega)$ is the manifold $N = M \times S^1$ (or $M \times \mathbb{R}$) endowed with the contact form $dz - \alpha$. Here we denote by z the projection to the second factor and still denote by α its pull-back under the projection $N \to M$.

However, the contactization can sometimes be defined even when ω is not exact. Suppose that the form ω represents an integral cohomology class $[\omega] \in H^2(M)$. The contactization $C(M, \omega)$ of (M, ω) can be constructed as follows. Let

$$\pi : N \to M$$

be a principal S^1-bundle with the Euler class equal to $[\omega]$. This bundle admits a connection whose curvature form just is ω. This connection can be viewed as a S^1-invariant 1-form α on N. The nondegeneracy of ω implies that α is a contact form and, therefore, $\xi = \{\alpha = 0\}$ is a contact structure on N. The contact manifold (N, ξ) is, by definition, the contactization $C(M, \omega)$ of the symplectic manifold (M, ω). A change of the connection α leads to a contactomorphic manifold.

We note that a Hamiltonian vector field on (M, ω) can be lifted to a contact vector field on N. In fact, a Hamiltonian function H on M and its Hamiltonian vector field X_H satisfy $dH = \iota(X_H)\omega$. We know there exists a one-to-one correspondence between contact vector fields and functions on N, so we obtain a contact vector field $\tilde{X}_H$ on N by $\alpha(\tilde{X}_H) = \pi^* H$. Also, we have $\pi_* \tilde{X}_H = X_H$. Thus, any Hamiltonian isotopy on M is lifted to a contact isotopy on N.

If $L \subset M$ is a Lagrangian submanifold then the connection α over it is flat. The pull-back $\pi^{-1}(L) \subset N$ under the projection, which is also the image of the S^1-action of a Legendrian lift Λ, denoted by $\tilde{L}$, is a Lagrangian submanifold in SN and is foliated by Legendrian leaves obtained by integrating the flat connection over L. If the holonomy defined by the connection α is integrable over L then the Lagrangian submanifold $\tilde{L}$ is foliated by closed Legendrian submanifolds in N. In this case, the connection over L is trivial. If this condition is satisfied then L is called exact (Bohr–Sommerfeld condition), and the Lagrangian submanifold $\tilde{L}$ is foliated by closed Legendrian lifts of L.

A Legendrian submanifold $\Lambda \subset (N, \xi)$ has a neighborhood U contactomorphic to the 1-jet space $J^1(\Lambda)$. Then $\tilde{L} \cap U$ can be identified under the contactomorphism with the so-called "0-wall": $W = \Lambda \times \mathbb{R} \subset J^1(\Lambda)$, which is just the set of 1-jets of function with 0 differential.

***Modifying* (SN, ω_ξ).** Now, given a contact isotopy $\{\psi_t \mid 0 \le t \le 1\}$ of (N, ξ), it can be lifted to a Hamiltonian isotopy $\{\Psi_t \mid 0 \le t \le 1\}$ of SN. Then, from the definition and properties listed above, we have a one-to-one correspondence between $L \cap \pi \circ \psi_1(\Lambda)$ and $\tilde{L} \cap \Psi_1(p^{-1}(\Lambda))$. They also coincide with $\tilde{L} \cap \psi_1(\Lambda)$, and all intersections are transversal. Therefore, it is natural to define Floer homology for such a pair of Lagrangian submanifolds $\tilde{L}$ and $\mathscr{L} = p^{-1}(\Lambda)$. However, as we all know, symplectization SN is not compact, so the ordinary method can not be applied directly. We adopt Ono's argument [1996] to overcome this difficulty.

We see that N is compact, thus there exists large $C > 0$ such that the trace of N under the isotopy $\{\Psi_t | 0 \le t \le 1\}$ is contained in a compact set $N \times [e^{-C}, e^{C}]$, and $N \times [e^{-C}, e^{C}]$ is disjoint from $\Psi_t(SN \setminus [e^{-D}, e^{D}])$, $t \in [0, 1]$, for some number $D > C$. The domain we are concerned with is $N \times [e^{-D}, +\infty)$. The isotopy $\{\Psi_t\}$ is generated by a Hamiltonian $H : [0, 1] \times SN \to \mathbb{R}$. We can find another function H', so that H' equals H on $N \times [e^{-C}, e^{C}]$ and equals zero outside of $N \times [e^{-D}, e^{D}]$. Then we get a new Hamiltonian isotopy $\{\Psi'_t | 0 \le t \le 1\}$ with compact support.

Since the boundary of the bundle $N \times [e^{-D-\epsilon}, +\infty)$ is of contact type, by symplectic filling techniques the symplectization $(SN = N \times \mathbb{R}_+, \omega_\xi)$ can be replaced by a new symplectic manifold (Q, Ω) which is diffeomorphic to the associated complex line bundle $N \times_{S^1} \mathbb{C} \to M$. In fact, Ono showed there exists a symplectic embedding $\mathscr{F}$ from $N \times (e^{-D-\epsilon}, +\infty)$ into (Q, Ω) ($\mathscr{F}$ is a symplectomorphism between $N \times (e^{-D-\epsilon}, +\infty)$ and $N \times_{S^1} \mathbb{C} - \{0\text{-section}\}$, see [Ono 1996, Appendix] for details). Therefore, we study the Lagrangian intersection problem for $Q, \mathscr{F}\tilde{L}, \mathscr{F}(\mathscr{L} \cap N \times (e^{-D-\epsilon}, +\infty))$ under Hamiltonian isotopy Φ_t generated by a Hamiltonian defined on Q, which equals $H' \circ \mathscr{F}^{-1}$ on $N \times_{S^1} \mathbb{C} - \{0\text{-section}\}$, and equals zero on the 0-section. For simplicity, we still denote them by $\tilde{L}, \mathscr{L}, H$.

The positive end of Q is J-convex; that is, for a given $E > 1$, the product $N \times \{E\} \subset Q$ is a J-convex codimension-1 submanifold. So no J-holomorphic curves can touch it, and, especially, there exists a C^0 bound for every J-holomorphic disc $u : D^2 \to Q$ with boundary in Lagrangian submanifolds $\tilde{L}$ and $\Phi_t(\mathcal{L})$ (compare also [Ono 1996]). For the general case, we consider $u : \Pi = \mathbb{R} \times [0, 1] \to Q$ with $u(\tau, 0) \subset \mathcal{L}$ and $u(\tau, 1) \subset \tilde{L}$, $\tau \in \mathbb{R}$, which is regarded as the connecting orbit between $x_-(t) = \lim_{\tau \to -\infty} u(\tau, t)$ and $x_+(t) = \lim_{\tau \to +\infty} u(\tau, t)$, solving the perturbed Cauchy–Riemann equation

$$\frac{\partial u}{\partial \tau} = -J \frac{\partial u}{\partial t} + \nabla H(t, u(\tau, t)).$$

In this situation, Gromov [1985] showed how to define an almost complex structure $\tilde{J}_H$ on the product $\tilde{Q} = \Pi \times Q$, such that the $\tilde{J}_H$-holomorphic sections of $\tilde{Q}$ are precisely the graph $\tilde{u}$ of solutions of the equation above. We can see that $\tilde{Q}$ is $\tilde{J}_H$-convex, so there is a C^0-bound for $\tilde{J}_H$-holomorphic curves in $\tilde{Q}$. The same then thing happens to the connecting orbits in Q.

3. Variation and functional

From the discussion above, we know that we have got a symplectic manifold (Q, Ω), and two Lagrangian submanifolds $\tilde{L}$ and $\mathcal{L}$. Then we will establish a homology theory for the pair $(\tilde{L}, \mathcal{L})$ in Q, and study critical points of the symplectic action functional defined on (some covering of) the space of paths in Q, starting from $\mathcal{L}$ with ends on $\tilde{L}$.

Let $H \in \mathcal{H}(Q)$ satisfy $\|H\| < \sigma(Q, \tilde{L}, \mathcal{L}, J)$, and $\Psi^t_{(s)}$, $s \in [0, 1]$, be the time-t flow generated by Hamiltonian sH (note that $\Psi^1_{(s)}$ is the lift of the contactomorphism $\psi^1_{(s)}$). And set

$$\mathcal{L}_s = \Psi^1_{(s)}(\mathcal{L}), \quad \Lambda_s = \psi^1_{(s)}(\Lambda) \subset N.$$

We suppose that $\tilde{L}$ intersects $\mathcal{L}_1$ transversally.

Let Σ be the connected component of constant paths in the path space

$$\{\gamma \in C^\infty([0, 1], Q) | \gamma(0) \in \mathcal{L},\ \gamma(1) \in \tilde{L}\}.$$

We define the closed 1-form α on Σ by

$$\langle \alpha(\gamma), v \rangle = \int_0^1 \Omega(\dot{\gamma}(t), v(t))\, dt, \quad v(t) \in TQ|_{\gamma(t)},\ \text{for } t \in [0, 1].$$

We also write the function $\theta : \Sigma \to \mathbb{R}$ as

$$\theta(\gamma) = -\int_0^1 H(t, \gamma(t))\, dt.$$

Note that the zeroes of $\alpha_s = \alpha + s d\theta$ are just time-1 trajectories generated by the flow $\Psi^t_{(s)}$ that start from $\mathcal{L}$ and end on $\tilde{L}$. If γ is the zero of α_s, then the ends of all $\gamma(1)$ are just the intersection points of $\tilde{L}$ with $\mathcal{L}_s$, which are one-to-one correspondent to the zeroes of α_s. The purpose of this paper is to estimate from below the number of zeroes of α_1 .

Since H_t is compactly supported on Q, let $b_+ = \int_0^1 \max_{x\in Q} H(t,x)\,dt$, and $b_- = \int_0^1 \min_{x\in Q} H(t,x)\,dt$. Then $\|H\| = b_+ - b_-$, $-b_+ \le \theta(\gamma) \le -b_-$, for all $\gamma \in \Sigma$. We introduce the Riemannian structure on Σ through the metric

$$(v_1, v_2) = \int_0^1 \Omega(v_1(t), Jv_2(t))\,dt.$$

Since

$$\begin{aligned}(\mathrm{grad}_\alpha(\gamma), v) &= \langle \alpha(\gamma), v\rangle \\ &= \int_0^1 \Omega(\dot\gamma(t), v(t))\,dt = \int_0^1 \Omega(J\dot\gamma(t), Jv(t))\,dt = (J\dot\gamma, v),\end{aligned}$$

so the gradient of the closed 1-form α is given by $J\dot\gamma$, similarly, the gradient of the closed 1-form α_s is $\mathrm{grad}_{\alpha_s} = J\dot\gamma - s\nabla H$.

Now, we consider the minimal covering $\pi : \tilde\Sigma \to \Sigma$ such that the form $\pi^*\alpha$ is exact (meaning that there is a functional F on $\tilde\Sigma$ such that $\pi^*\alpha = dF$), and its structure group Γ is free abelian. Denote $F_s = F + s(\theta\circ\pi)$, so $dF_s = \pi^*\alpha_s$. The gradient ∇F_s of the functional F_s, with respect to the lift of the Riemannian structure on Σ, is a Γ-invariant vector field on $\tilde\Sigma$, and $\pi_*\nabla F_s = \mathrm{grad}_{\alpha_s}$. Then we consider the moduli space of thus gradient flows connecting a pair of critical points (x_-, x_+) of F_s

$$\begin{aligned}&M_s(x_-, x_+) \\ &\quad= \left\{ u : \mathbb{R} \to \tilde\Sigma \;\middle|\; \frac{du(\tau)}{d\tau} = -\nabla F_s(u(\tau)),\ u \text{ is not constant},\ \lim_{\tau\to\pm\infty} u(\tau) = x_\pm \right\}.\end{aligned}$$

Denote by $\mathcal{M}_s = \bigcup_{x_\pm} M_s(x_-, x_+)$ the collection, and the nonparameterized space by $\hat M_s(x_-, x_+) = M_s(x_-, x_+)/\mathbb{R}$, and the natural quotient map $q : M_s \to \hat M_s$. Choosing a regular Ω-compatible almost complex structure J on Q [Ono 1996], we can assume that there is a dense set $T \subset [0,1]$ such that for all $s \in T$, $M_s(x_-, x_+)$ are finite dimensional smooth manifolds, consequently, $\tilde L$ intersects $\mathcal{L}_s$ transversally.[1]

[1]Recall that the J used here is just the $J' = J \oplus J_0$ given in the introduction; by choosing a generic ω-compatible almost complex structure J on M we can obtain the regular or generic Ω-compatible structure J' on Q. The arguments in [Ono 1996] for J'-holomorphic maps can apply to our H-perturbed J'-holomorphic map by similar statements as those in [Floer et al. 1995]. We can overcome the similar problem which appears in the continuation argument of Section 6.

We define the length of a gradient trajectory $u \in M_s(x_-, x_+)$ by $l_s(u) = F_s(x_-) - F_s(x_+)$. If $\hat{u} \in \hat{M}_s$, then we define its length naturally by $l_s(\hat{u}) = l_s(u)$, where $\hat{u} = q \circ u$. Denote $\Pi = \mathbb{R} \times [0, 1]$, then the map $\bar{u} : \Pi \to Q$, defined by $\bar{u}(\tau, t) = \pi(u(\tau))(t)$, satisfies the following perturbed Cauchy–Riemann equation

$$\frac{\partial \bar{u}(\tau, t)}{\partial \tau} = -J(\bar{u}(\tau, t))\frac{\partial \bar{u}(\tau, t)}{\partial t} + s\nabla H(t, \bar{u}(\tau, t)),$$

with limits

$$\lim_{\tau \to \pm\infty} \bar{u}(\tau, t) = \pi(x_\pm) = \bar{x}_\pm(t).$$

It is easy to see that $l_0(u) = \int_{-\infty}^{+\infty} u^* dF = \int_\Pi \bar{u}^*\Omega$.

If $u \in M_0$, then $\bar{u}$ is a J-holomorphic map from Π to Q. From Oh's removing of boundary singularities theorem [1992], $\bar{u}$ can be extended to a J-holomorphic curve $\bar{u}' : (D^2, \partial^+ D^2, \partial^- D^2) \to (Q, \tilde{L}, \mathcal{L})$, where $D^2 = \bar{\Pi}$ is the two-point compactification of Π. Since $l_0(u) = \int_\Pi \bar{u}^*\Omega = \int_{D^2} (\bar{u}')^*\Omega$, we know that $l_0(u) \geq \sigma_D(Q, \tilde{L}, \mathcal{L}, J)$.

4. Defining and computing homology for C_ε^0

We denote by Y_s the set of critical points of F_s, and by C_s the vector space spanned by Y_s over $\mathbb{Z}_2$.

Since Y_s is Γ-invariant, C_s has a structure of free K-module with rank$= \#(\tilde{L} \cap \mathcal{L}_s)$, $s \in T$, where $K = \mathbb{Z}_2[\Gamma]$. Our aim in this section is to establish some homology for the complex C_ε, where ε is small enough. We write the following definition similar as the one given by Chekanov [1998].

Definition 4.1. Fix $\delta > 0$, satisfying $\Delta := \|H\| + \delta < \sigma(Q, \tilde{L}, \mathcal{L}, J)$. A gradient trajectory $u \in M_s$ is said to be short if $l_s(u) \leq \Delta$, and be very short if $l_s(u) \leq \delta$.

Now we denote the area by $A(u) = \int_\Pi \bar{u}^*\Omega$, and $h(u) = s\int_{-\infty}^{+\infty} u^* d(\theta \circ \pi)$, then still write $l(u) = l_s(u) = A(u) + h(u)$, we have

Lemma 4.2. *If u is very short, in the sense that $l(u) \leq \delta$, then the area $A(u) \leq \Delta$.*

Proof. Since $\theta = -\int_0^1 H(t, \gamma(t))\, dt \in [-b_+, -b_-]$, then

$$h(u) = s\int_{-\infty}^{+\infty} u^* d(\theta \circ \pi) = s\theta(\pi(u(\tau)))\Big|_{-\infty}^{+\infty} \geq s(b_- - b_+),$$

so

$$A(u) = l(u) - h(u) \leq \delta - (b_- - b_+) = \|H\| + \delta = \Delta. \qquad \square$$

The next lemma was essentially proved by Chekanov [1998, Lemma 6]; we adapt it here for our setting, with some modifications.

Lemma 4.3. . *For a small neighborhood U of $\tilde{L}$ in Q, there exists a $\varepsilon_0 > 0$, such that for any positive $\varepsilon < \varepsilon_0$, every short gradient trajectory $u \in M_\varepsilon$ is very short, and for every short u we have $\bar{u}(\Pi) \subset U$.*

Proof. We work it by contradiction. For the first claim, we suppose there is a sequence $u_n \in M_{s_n}$ and a positive number c with $\delta \le c \le \Delta$ so that when $s_n \to 0$ then $l_{s_n}(u_n) \to c$. By Gromov's compactness theorem, there are some subsequence of $\bar{u}_n = \pi(u_n)$ convergent to $\bar{u}_\infty$ which is a collection of J-holomorphic spheres and J-holomorphic discs bounding $\tilde{L}$ and/or $\mathscr{L}$. Then the total symplectic area of this limit collection is just $l_0(u_\infty) = c$ which by the assumptions of Theorem 1.4 is larger than $\sigma(Q, \tilde{L}, \mathscr{L})$, but $c \le \Delta < \sigma(Q, \tilde{L}, \mathscr{L})$, so the claim holds. For the second claim, the argument is similar. Note that if the image $\bar{u}_\infty(\Pi)$ of the limit collection is not contained in U, then at least one of the J-curve is not contained in U which is nonconstant and its area will be larger than $\sigma(Q, \tilde{L}, \mathscr{L}, J) > \Delta$, this contradicts the (very) shortness condition. □

Then, we denote by $M'_\varepsilon \subset M_\varepsilon$ the set of all short gradient trajectories (nonparameterized short gradient trajectories), and likewise for $\hat{M}'_\varepsilon \subset \hat{M}_\varepsilon$. And we can define the $\mathbb{Z}_2$-linear map $\partial : C_\varepsilon \to C_\varepsilon$ by

$$\partial(x) = \sum_{y \in Y_\varepsilon} \#\{\text{isolated points of } \hat{M}'_\varepsilon(x, y)\} y \quad \text{for all } x \in Y_\varepsilon.$$

Let $\varepsilon \in T$ be sufficiently small and satisfy the conditions of Lemma 4.3. Choose an element $x_0 \in Y_\varepsilon$, then we can define a subclass $Y^0_\varepsilon \subset Y_\varepsilon$ by

$$Y^0_\varepsilon = \{x \in Y_\varepsilon \mid |F_\varepsilon(x) - F_\varepsilon(x_0)| \le \delta\}.$$

Then we see that the projection π bijectively maps Y^0_ε onto the set of zeroes of the form α_ε, and we get the bijection $Y^0_\varepsilon \times \Gamma \to Y_\varepsilon$ given by $(y, a) \mapsto a(y)$, which induces the isomorphism $C^0_\varepsilon \otimes K \to C_\varepsilon$, where $C^0_\varepsilon \subset C_\varepsilon$ is spanned over $\mathbb{Z}_2$ by Y^0_ε. Now, for sufficiently small $\varepsilon \in T$, we can establish the homology for $(C_\varepsilon, \partial)$

Lemma 4.4. (1) *The map ∂ is K-linear, well defined, and $\partial(C^0_\varepsilon) \subset C^0_\varepsilon$.*

(2) *If $\varepsilon \in T$ is sufficiently small, then $\partial^2 = 0$.*

(3) *The homology $H(C^0_\varepsilon, \partial) \cong H_*(\Lambda, \mathbb{Z}_2)$.*

Proof. Step 1. Since the gradient flow is Γ-invariant, ∂ is naturally K-linear. We know that the bubbling off can not occur. Indeed, since ε is sufficiently small, then $u \in M_\varepsilon$ is very short, $l_\varepsilon(u) \le \delta$, by Lemma 4.2, the area $A(u) \le \Delta < \sigma(Q, \tilde{L}, \mathscr{L}, J)$, and from the assumption in our theorem, the area of any J-holomorphic sphere or J-holomorphic disc bounding $\tilde{L}$ and $\mathscr{L}$ is larger than $\sigma(Q, \tilde{L}, \mathscr{L}, J)$. Thus, $\hat{M}'_\varepsilon(x, y)$ is compact and the number of its isolated points is finite.

Step 2. Suppose $\varepsilon \in T$ satisfy the conditions in Lemma 4.3. If $\|H\| = 0$, $\Delta = \delta$, then $H \equiv const.$ and $\psi_H \equiv \mathrm{id}$, it is a trivial case. If $\|H\| > 0$, we can always choose a fixed $\delta < \frac{1}{2}\|H\| < \frac{1}{2}\Delta$. Consider a pair of isolated trajectories $u_1 \in \hat{M}'_\varepsilon(x, y)$, $u_2 \in \hat{M}'_\varepsilon(y, z)$. Then there exists a unique 1-dimensional connected component $\mathscr{C} \subset \hat{M}_\varepsilon(x, z)$ such that (u_1, u_2) is one of the two ends of compactification of $\mathscr{C}$ [Floer 1988a]. Since the length is additive under gluing, we have $l_\varepsilon(u) = l_\varepsilon(u_1) + l_\varepsilon(u_2) < 2\delta < \Delta$, for all $u \in \mathscr{C}$. By Lemma 4.3, u is also very short. From Lemma 4.2, we know the bubbling off doesn't occur, too. Then the other end of $\mathscr{C}$ can be compactified by a pair of isolated trajectories $u'_1 \in \hat{M}_\varepsilon(x, y)$, $u'_2 \in \hat{M}_\varepsilon(y, z)$. Also $l_\varepsilon(u'_1) + l_\varepsilon(u'_2) = l_\varepsilon(u_1) + l_\varepsilon(u_2) < \Delta$, thus u'_1, u'_2 are short trajectories. So we know that, for $x, z \in Y_\varepsilon$, the number of isolated trajectories $(u_1, u_2) \in \hat{M}'_\varepsilon(x, y) \times \hat{M}'_\varepsilon(y, z)$ is even, that is, $\partial^2(x) = 0$.

Step 3. From Lemma 4.3, we know that for any $u \in \hat{M}'_\varepsilon(x, y)$, $\bar{u}(\bar{\Pi}) \subset U$. That is to say, if ε is small enough, then $\Lambda_\varepsilon = \psi^1_{(\varepsilon)}(\Lambda)$ is always contained in a small neighborhood $U' = U \cap N$ of the Legendrian submanifold Λ in N. By Darboux's theorem, U' is contactomorphic to a neighborhood of the 0-section in the 1-jet space $J^1(\Lambda)$. This contactomorphism moves $\tilde{L} \cap U'$ onto the 0-wall W, which is defined as the space of 1-jets of functions with 0 differential. Thus a Legendrian submanifold Λ', which is C^1-close to Λ and transverse to W, corresponds to a Morse function $\beta : \Lambda \to \mathbb{R}$ so that the intersection points of $\tilde{L}$ and Λ' are in one-to-one correspondence with the critical points of the function β. We can explicitly choose a metric on Λ and a generic almost complex structure J (recall the footnote in Section 3) on the symplectization SN in such a way that the gradient trajectories in $\mathcal{M}_\varepsilon$ would be in one-to-one correspondence with the gradient trajectories of the function β connecting the corresponding critical points of this function. Thus we can identify our complex C^0_ε with the Morse chain complex for the function β (here we may also reduce the problem to Lagrangian intersections in M by applying continuation argument and projecting the manifold to M, the method of equating Floer and Morse complex is standard, we refer the reader to [Schwarz 1993]), so we have an isomorphism $H(C^0_\varepsilon, \partial) \cong H_*(\Lambda, \mathbb{Z}_2)$. □

5. Homology algebra

Under the condition $\pi_2(Q, \cdot) = 0$ or the monotonicity assumption [Floer 1988a; Oh 1993a; 1993b; 1995], the Floer homology $HF_*(C_s, \partial)$ of the complex C_s, $s \in T \subset [0, 1]$, can be defined. Then we can use the classical continuation method (see [Floer 1989b; McDuff 1990] or the papers by Oh just cited) to prove the isomorphism between $HF_*(C_\varepsilon, \partial)$ and $HF_*(C_1, \partial)$, that means to construct chain homotopy $\Phi'\Phi \sim \mathrm{id}_\varepsilon$ (and $\Phi\Phi' \sim \mathrm{id}_1$), where $\Phi : (C_\varepsilon, \partial) \to (C_1, \partial)$, $\Phi' : (C_1, \partial) \to (C_\varepsilon, \partial)$ are chain homomorphisms defined similarly as the definition of ∂, except

for considering the moduli space of continuation trajectories. That is to say, in order to prove $\Phi'\Phi \sim \mathrm{id}_\varepsilon$, we should show there exists a chain homomorphism $\boldsymbol{h} : (C_\varepsilon, \partial) \to (C_\varepsilon, \partial)$, so that

$$\Phi'\Phi - \mathrm{id} = \boldsymbol{h}\partial - \partial\boldsymbol{h}.$$

However, in general case, we can not define appropriately any homology for C_s unless s is small enough. Then we may only prove a weaker "homotopy", which was called λ-*homotopy* by Chekanov. In fact, for the aim of estimating from below the number of critical points of the functional F_s, this λ-homotopy is enough.

We shall use the following homology algebraic result, introduced and proved by Chekanov [1998].

Let Γ be a free abelian group equipped with a monomorphism $\lambda : \Gamma \to \mathbb{R}$, which we call a *weight function*. Set

$$\Gamma^+ = \{a \in \Gamma | \lambda(a) > 0\}, \qquad \Gamma^- = \{a \in \Gamma | \lambda(a) < 0\}.$$

Let k be a commutative ring. Consider the group ring $K = k[\Gamma]$. For a k-module M, we have the natural decomposition $M \otimes K = M^+ \oplus M^0 \oplus M^-$. where $M^+ = \Gamma^+(M)$, $M^0 = M$, $M^- = \Gamma^-(M)$. Consider the projections

$$p^+ : M \otimes K \to M^+ \oplus M^0, \qquad p^- : M \otimes K \to M^0 \oplus M^-.$$

Assume that (M, ∂) is a differential k-module, then ∂ naturally extends to a K-linear differential on $M \otimes K$.

Definition 5.1. We say two linear maps $\phi_0, \phi_1 : M \otimes K \to M \otimes K$ are λ-homotopic if there exists a K-linear map $\boldsymbol{h} : M \otimes K \to M \otimes K$ such that

$$p^+(\phi_0 - \phi_1 + \boldsymbol{h}\partial + \partial\boldsymbol{h})p^- = 0.$$

Lemma 5.2 [Chekanov 1998]. *Let λ be a weight function on a free abelian group Γ. Assume (M, ∂) to be a differential k-module and N to be a K-module, where $K = k[\Gamma]$. If the maps $\Phi^+ : M \otimes K \to N$ and $\Phi^- : N \to M \otimes K$ are K-linear and $\Phi^-\Phi^+$ is λ-homotopic to the identity, then* $\mathrm{rank}_K N \geq \mathrm{rank}_k H(M, \partial)$.

6. Proofs of the main results

Given a (s_-, s_+) continuation function $\rho : \mathbb{R} \to [0, 1]$ satisfying

$$\rho(\tau) = \begin{cases} s_- & \text{if } \tau < -r, \\ s_+ & \text{if } \tau > r, \end{cases}$$

where $r \in \mathbb{R}$, we can define the moduli space of continuation trajectories

$$M_\rho(x_-, x_+) = \left\{ u : \mathbb{R} \to \tilde{\Sigma} \;\middle|\; \frac{du(\tau)}{d\tau} = -\nabla F_{\rho(\tau)}(u(\tau)),\ \lim_{\tau \to \pm\infty} u(\tau) = x_\pm \right\},$$

where $x_\pm \in Y_{s_\pm}$. And we denote the collection by $\mathcal{M}_\rho = \bigcup_{x_-,x_+} M_\rho(x_-, x_+)$. The length of a continuation trajectory is defined by $l_\rho(u) = F_{s_-}(x_-) - F_{s_+}(x_+)$.

Choose a monotone $(\varepsilon, 1)$ continuation function ρ_+ and a monotone $(1, \varepsilon)$ continuation function ρ_-. For generic H, $M_{\rho_\pm}(x_-, x_+)$ are smooth manifolds. We will say a continuation trajectory $u \in M_{\rho_+}(x_-, x_+)$ (or $u \in M_{\rho_-}(x_-, x_+)$) is short if $l_{\rho_+}(u) \le \delta + (1-\varepsilon)b_+$ (resp. $l_{\rho_-}(u) \le \delta + (\varepsilon - 1)b_-$). The subspace of all short trajectories is denoted by $M'_{\rho_\pm}(x_-, x_+) \subset \mathcal{M}_{\rho_\pm}$.

In the best possible situation, that is, if no sequence of continuation trajectories reaches the negative end (the zero section of $Q \to M$),[2] we can simply construct the $\mathbb{Z}_2$-linear continuation map $\Phi^+ : C_\varepsilon \to C_1$, $\Phi^- : C_1 \to C_\varepsilon$ as

$$\Phi^+(x) = \sum_{y \in Y_1} \#\{\text{isolated points of } M'_{\rho_+}(x, y)\}y,$$

$$\Phi^-(y) = \sum_{z \in Y_\varepsilon} \#\{\text{isolated points of } M'_{\rho_-}(x, z)\}z,$$

where $x \in Y_\varepsilon$. The next lemma implies that this definition of $\Phi^\pm$ is sound.

Lemma 6.1. *If $u \in M_{\rho_+}(x_-, x_+)$, then $l_{\rho_+}(u) \ge (1-\varepsilon)b_-$. If $u \in M_{\rho_-}(x_-, x_+)$, then $l_{\rho_-}(u) \ge (\varepsilon - 1)b_+$. And the sum in the definition of $\Phi^\pm$ is finite.*

Proof. Recall $F_s = F + s\theta \circ \pi$, $s \in [0, 1]$. For a (s_-, s_+) continuation function $\rho : \mathbb{R} \to [0, 1]$, we have $F_{\rho(\tau)} = F + \rho(\tau)\theta \circ \pi$. So the length of a continuation trajectory is

$$\begin{aligned} l_\rho(u) &= F_{s_-}(x_-) - F_{s_+}(x_+) \\ &= -\int_{-\infty}^{+\infty} u^* dF_{\rho(\tau)} = -\int_{-\infty}^{+\infty} u^* dF - \int_{-\infty}^{+\infty} u^* d(\rho(\tau)\theta \circ \pi) \\ &= A(u) + h(u), \end{aligned}$$

where we set

$$\begin{aligned} A(u) &= -\int_{-\infty}^{+\infty} u^* dF = \int_\Pi \bar{u}^* \Omega = \int_{-\infty}^{+\infty} \left(\frac{du(\tau)}{d\tau}, \frac{du(\tau)}{d\tau} \right) d\tau \\ &= \int_{-\infty}^{+\infty} \|\nabla F_{\rho(\tau)}\|^2 d\tau \ge 0, \end{aligned}$$

$$h(u) = -\int_{-\infty}^{+\infty} u^* d(\rho(\tau)\theta \circ \pi) = -\int_{-\infty}^{+\infty} \frac{d\varrho(\tau)}{d\tau} \theta(\pi(u(\tau)))\, d\tau.$$

Recall that

$$\theta = -\int_0^1 H\, dt \in [-b_+, -b_-],$$

[2]The possibility that there is such a sequence was pointed out to the author by one of referees. In this case, we have to modify the continuation map.

if $u \in M_{\rho_+}(x_-, x_+)$,

$$l(u) \geq h(u) \geq (1-\varepsilon)b_-;$$

if $u \in M_{\rho_-}(x_-, x_+)$,

$$l(u) \geq h(u) \geq (\varepsilon-1)b_+.$$

Thus, For a short trajectory $u \in M'_{\rho_\pm}(x_-, x_+)$,

$$A(u) = l(u) - h(u) \leq \delta + (1-\varepsilon)(b_+ - b_-) = \|H\| + \delta = \Delta.$$

Since $A(u) = \int_\Pi \bar{u}^*\Omega \leq \Delta < \sigma(Q, \tilde{L}, J)$, by Gromov's arguments, no bubbling can occur, then spaces $M'_{\rho_\pm}(x_-, x_+)$ are compact, and the finiteness of the set of isolated points of $M'_{\rho_\pm}(x, y)$ is verified. □

However, in the general case, the ideal assumption is not always satisfied. If a sequence of continuation trajectories converges to a curve reaching the negative end of $\mathcal{L}$, the boundary of moduli space of continuation trajectories will contain such curve. So we need modify the definition of $\Phi^\pm$ to get a well-defined continuation map. In fact, Ono [1996, p. 218] had dealt with a similar problem by considering the algebraic intersection number of the continuation trajectories with the zero section O_M of $Q \to M$. Under Ono's assumption $\pi_2(M, L) = 0$, bubbling off of holomorphic discs contained in the zero section of Q with boundary on L never occurs. In our case, since we have the bound for energy, such kind of bubbling off of discs is also avoided. Thus, we can define homomorphisms $\Phi^+_k : C_\varepsilon \to C_1$, $\Phi^-_k : C_1 \to C_\varepsilon$ as

$$\Phi^+_k(x) = \sum_{y \in Y_1} \mathrm{Int}^+_{2,k}(x, y)y, \quad \Phi^-_k(y) = \sum_{z \in Y_\varepsilon} \mathrm{Int}^-_{2,k}(y, z)z,$$

where $\mathrm{Int}^+_{2,k}(x, y)$ ($\mathrm{Int}^-_{2,k}(y, z)$) is the mod-2 number of isolated points u in the continuational moduli spaces $M'_{\rho_+}(x, y)$ and $M'_{\rho_-}(y, z)$, respectively, having algebraic intersection number $u \cdot O_M = k/2$.

With the restriction of bound of energy, we can make similar discussions as in [Ono 1996] to get the finiteness of $\mathrm{Int}^\pm_{2,0}$. So $\Phi^\pm_0$ are just our favorite continuation maps. We will not list the detailed arguments here and refer the reader to [Ono 1996] for original discussion. In the following, we will still denote the continuation maps by $\Phi^\pm$ for simplicity.

Then we can use the homology algebraic result given in Section 5 to prove Theorem 1.4, provided there exists a λ-homotopy between $\Phi^-\Phi^+$ and the identity.

Proof of Theorem 1.4. Take $k = \mathbb{Z}_2$, $K = \mathbb{Z}_2[\Gamma]$, $M = C^0_\varepsilon$, $M \otimes K = C_\varepsilon$, $N = C_1$, and let Γ be the structure group of the covering. The weight function $\lambda : \Gamma \to \mathbb{R}$ can be defined as $\lambda(a) = F(a(x)) - F(x)$. We also have decompositions $Y_\varepsilon = Y^+_\varepsilon \cup Y^0_\varepsilon \cup Y^-_\varepsilon$, $C_\varepsilon = C^+_\varepsilon \oplus C^0_\varepsilon \oplus C^-_\varepsilon$, where $Y^\pm_\varepsilon = \Gamma^\pm(Y^0_\varepsilon)$.

Assume that we have got a λ-homotopy $\boldsymbol{h} : C_\varepsilon \to C_\varepsilon$. By Lemmas 5.2 and 6.1 we have

$$\begin{aligned}\#(L \cap \pi \circ \psi_1(\Lambda)) &= \#(\tilde{L} \cap \psi_1(\Lambda)) = \#(\tilde{L} \cap \Psi_1(\mathscr{L})) = \operatorname{rank}_K C_1 \\ &\geq \operatorname{rank}_k H(C_\varepsilon^0, \partial) = \dim H_*(\Lambda, \mathbb{Z}_2) = \dim H_*(L, \mathbb{Z}_2). \quad \square\end{aligned}$$

In the rest of this section, we show a sketchy proof of the existence of λ-homotopy.

Lemma 6.2. *There exists a λ-homotopy $\boldsymbol{h} : C_\varepsilon \to C_\varepsilon$ between $\Phi^-\Phi^+$ and the identity.*

Proof. We follow the arguments of Chekanov and state his main thought. For constructing the homomorphism $\boldsymbol{h}$, we use a family of $(\varepsilon, \varepsilon)$ continuation functions $\mu_c,\ c \in [0, +\infty)$ satisfying these conditions:

(1) $\mu_0(\tau) \equiv \varepsilon$.

(2) $du_c(\tau)/d\tau \geq 0$ if $\tau < 0$ and $du_c(\tau)/d\tau \leq 0$ if $\tau > 0$.

(3) $c \mapsto \mu_c(0)$ is a monotone map from $[0, +\infty)$ onto $[\varepsilon, 1]$.

(4) when c is large enough $\mu_c(\tau) = \begin{cases} \rho_+(\tau + c), & \text{if } \tau \leq 0; \\ \rho_-(\tau - c), & \text{if } \tau \geq 0. \end{cases}$

We denote the moduli space by

$$M_\mu(x_-, x_+) = \{(c, u) \mid u \in M_{\mu_c}(x_-, x_+)\},\ x_\pm \in Y_\varepsilon.$$

For generic H, $M_\mu(x_-, x_+)$ are smooth manifolds.

As for the $(\varepsilon, 1)$ or $(1, \varepsilon)$ continuation trajectories earlier in this section, under the assumption that any sequence of μ_c-continuation trajectories reaches the negative end of $\mathscr{L}$,[3] we can define the $\mathbb{Z}_2$-linear map for C_ε^0 as

$$\boldsymbol{h}(x) = \sum_{y \in Y_\varepsilon^0} \#\{\text{isolated points of } M'_\mu(x, y)\} y,\ x \in Y_\varepsilon^0,$$

where $M'_\mu(x, y)$ is the subset of the moduli space $M_\mu(x, y)$ which contains only short μ_c-continuation trajectories, a μ_c-continuation trajectory $u \in M_{\mu_c}(x_-, x_+)$ is called short if its length $l(u) = l_{\mu_c}(u) \leq \delta$. Moreover, For any $u \in M_{\mu_c}(x_-, x_+)$, we have $l_{\mu_c}(u) = A(u) + h(u) \geq h(u) \geq (\mu_c(0) - \varepsilon)b_- + (\varepsilon - \mu_c(0))b_+ \geq b_- - b_+ = -\|H\|$, and if $l(u) \leq \delta$, then $A(u) = l(u) - h(u) \leq \delta + \|H\| \leq \Delta$.

The map $\boldsymbol{h}$ can be extended naturally to a K-linear map on C_ε. Since for $u \in M'_\mu(x, y)$, $l_{\mu_c}(u) \leq \delta,\ A(u) \leq \Delta$, the bubbling off does not occur, $M'_\mu(x, y)$ is compact and the sum is finite, thus the map $\boldsymbol{h}$ is well defined.

[3] Otherwise, we will again adopt Ono's argument to take into consideration the algebraic intersection number. The way to modify the definition of the map $\boldsymbol{h}$ is similar to the previously mentioned way to modify $\Phi^\pm$.

To prove $\boldsymbol{h}$ is a λ-homotopy, we have to verify

$$p^+(x + \Phi^-\Phi^+ x + \boldsymbol{h}\partial x + \partial \boldsymbol{h} x) = 0,$$

for all $x \in Y_\varepsilon^0 \cup Y_\varepsilon^-$. This will follow from the standard gluing argument involving the ends of the 1-dimensional part $\aleph$ of $M_\mu(x, z)$, $z \in Y_\varepsilon^+ \cup Y_\varepsilon^0$. Since $x \in Y_\varepsilon^0 \cup Y_\varepsilon^-$, $z \in Y_\varepsilon^+ \cup Y_\varepsilon^0$, and ε is small enough, we know $l(u) \le \delta$ for $u \in \aleph$. Indeed, $l(u) = F_\varepsilon(x) - F_\varepsilon(z)$, and there exist x' and z' in Y_ε^0 and $a \in \Gamma^+ \cup \Gamma^0$, $b \in \Gamma^0 \cup \Gamma^-$ such that $z = a(z')$, $x = b(x')$, and $F_\varepsilon(x) - F_\varepsilon(x') = \lambda(b) \le 0$, $F_\varepsilon(z') - F_\varepsilon(z) = -\lambda(a) \le 0$, also we know that $F_\varepsilon(x') - F_\varepsilon(z') \le \delta$ since $x', z' \in Y_\varepsilon^0$, this implies $l(u) \le \delta$, so $A(u) \le \Delta$.[4] This gets rid of the bubbling off. Then the compactification of $\aleph$ shows that the left-hand side of the formula above equals

$$\sum_{z \in Y_\varepsilon^+ \cup Y_\varepsilon^0} \#\{S(x, z)\} z,$$

and the number $\#\{S(x, z)\}$ is even.[5] This ends the proof of the lemma and of Theorem 1.4. For more details, see [Chekanov 1998; Floer 1989b; McDuff 1990]. □

Acknowledgement

The author thanks Professor Yiming Long for his constant encouragement. He also thanks both referees for carefully checking the paper and pointing out statements that were unclear in the earlier version. He especially thanks the referee who raised an important point about possible bubbling-off of holomorphic discs from the continuation trajectories, which the author didn't take into consideration in the earlier version of the paper, and gave a suggestion for revising the paper.

The author also thanks Silvio Levy for a careful editing that made the English more fluent.

References

[Arnold 1965] V. Arnold, "Sur une propriété topologique des applications globalement canoniques de la mécanique classique", *C. R. Acad. Sci. Paris* **261** (1965), 3719–3722. MR 33 #1861

[Arnold 1978] V. I. Arnold, *Mathematical methods of classical mechanics*, Graduate Texts in Mathematics **60**, Springer, New York, 1978. MR 57 #14033b Zbl 0386.70001

[Chekanov 1996] Y. V. Chekanov, "Hofer's symplectic energy and Lagrangian intersections", pp. 296–306 in *Contact and symplectic geometry* (Cambridge, 1994), edited by C. B. Thomas, Publ. Newton Inst. **8**, Cambridge Univ. Press, Cambridge, 1996. MR 98b:58058 Zbl 0867.58026

[4] From this proof of the inequality $l(u) \le \delta$, the reader can see why we only check λ-homotopy.

[5] Beyond the standard gluing argument, one should verify that the other ends of the compactification, which are pairs of continuation trajectories, are still all short. This is not difficult to do. Also, for the modified continuation maps $\Phi_0^\pm$, we can verify that the continuation trajectories in the other ends have the same 0-algebraic intersection number with the zero section O_M, see [Ono 1996].

[Chekanov 1998] Y. V. Chekanov, "Lagrangian intersections, symplectic energy, and areas of holomorphic curves", *Duke Math. J.* **95**:1 (1998), 213–226. MR 99k:58034 Zbl 0977.53077

[Conley and Zehnder 1983] C. C. Conley and E. Zehnder, "The Birkhoff–Lewis fixed point theorem and a conjecture of V. I. Arnol'd", *Invent. Math.* **73** (1983), 33–49. MR 85e:58044 Zbl 0516.58017

[Eliashberg et al. 1995] Y. Eliashberg, H. Hofer, and D. Salamon, "Lagrangian intersections in contact geometry", *Geom. Funct. Anal.* **5**:2 (1995), 244–269. MR 96c:58034 Zbl 0844.58038

[Floer 1988a] A. Floer, "Morse theory for Lagrangian intersections", *J. Differential Geom.* **28**:3 (1988), 513–547. MR 90f:58058 Zbl 0674.57027

[Floer 1988b] A. Floer, "The unregularized gradient flow of the symplectic action", *Comm. Pure Appl. Math.* **41**:6 (1988), 775–813. MR 89g:58065 Zbl 0633.53058

[Floer 1989a] A. Floer, "Symplectic fixed points and holomorphic spheres", *Comm. Math. Phys.* **120**:4 (1989), 575–611. MR 90e:58047 Zbl 0755.58022

[Floer 1989b] A. Floer, "Witten's complex and infinite-dimensional Morse theory", *J. Differential Geom.* **30**:1 (1989), 207–221. MR 90d:58029 Zbl 0678.58012

[Floer et al. 1995] A. Floer, H. Hofer, and D. Salamon, "Transversality in elliptic Morse theory for the symplectic action", *Duke Math. J.* **80**:1 (1995), 251–292. MR 96h:58024 Zbl 0846.58025

[Fukaya and Ono 1999] K. Fukaya and K. Ono, "Arnold conjecture and Gromov–Witten invariant", *Topology* **38**:5 (1999), 933–1048. MR 2000j:53116 Zbl 0946.53047

[Fukaya et al. 2000] K. Fukaya, Y.-G. Oh, H. Ohta, and K. Ono, "Lagrangian intersection theory, anomaly and obstruction", Preprint, Kyoto Univ., 2000.

[Gromov 1985] M. Gromov, "Pseudoholomorphic curves in symplectic manifolds", *Invent. Math.* **82**:2 (1985), 307–347. MR 87j:53053 Zbl 0592.53025

[Hofer 1985] H. Hofer, "Lagrangian embeddings and critical point theory", *Ann. Inst. H. Poincaré Anal. Non Linéaire* **2**:6 (1985), 407–462. MR 87i:58059 Zbl 0591.58009

[Hofer 1990] H. Hofer, "On the topological properties of symplectic maps", *Proc. Roy. Soc. Edinburgh Sect. A* **115**:1-2 (1990), 25–38. MR 91h:58042 Zbl 0713.58004

[Lalonde and McDuff 1995] F. Lalonde and D. McDuff, "The geometry of symplectic energy", *Ann. of Math.* (2) **141**:2 (1995), 349–371. MR 96a:58089 Zbl 0829.53025

[Laudenbach and Sikorav 1985] F. Laudenbach and J.-C. Sikorav, "Persistance d'intersection avec la section nulle au cours d'une isotopie hamiltonienne dans un fibré cotangent", *Invent. Math.* **82**:2 (1985), 349–357. MR 87c:58042 Zbl 0592.58023

[Liu 2005] C.-G. Liu, "Cup-length estimate for Lagrangian intersections", *J. Differential Equations* **209**:1 (2005), 57–76. MR 2005g:53176 Zbl 02136639

[Liu and Tian 1998] G. Liu and G. Tian, "Floer homology and Arnold conjecture", *J. Differential Geom.* **49**:1 (1998), 1–74. MR 99m:58047 Zbl 0917.58009

[McDuff 1990] D. McDuff, "Elliptic methods in symplectic geometry", *Bull. Amer. Math. Soc.* (*N.S.*) **23**:2 (1990), 311–358. MR 91i:58046 Zbl 0723.53018

[McDuff 1991] D. McDuff, "Symplectic manifolds with contact type boundaries", *Invent. Math.* **103**:3 (1991), 651–671. MR 92e:53042 Zbl 0719.53015

[Oh 1992] Y.-G. Oh, "Removal of boundary singularities of pseudo-holomorphic curves with Lagrangian boundary conditions", *Comm. Pure Appl. Math.* **45**:1 (1992), 121–139. MR 92k:58065 Zbl 0743.58018

[Oh 1993a] Y.-G. Oh, "Floer cohomology of Lagrangian intersections and pseudo-holomorphic disks, I", *Comm. Pure Appl. Math.* **46**:7 (1993), 949–993. MR 95d:58029a Zbl 0795.58019

[Oh 1993b] Y.-G. Oh, "Floer cohomology of Lagrangian intersections and pseudo-holomorphic disks, II: ($\mathbf{C}P^n$, $\mathbf{R}P^n$)", *Comm. Pure Appl. Math.* **46**:7 (1993), 995–1012. MR 95d:58029b Zbl 0795.58020

[Oh 1995] Y.-G. Oh, "Floer cohomology of Lagrangian intersections and pseudo-holomorphic disks, III: Arnol'd–Givental conjecture", pp. 555–573 in *The Floer memorial volume*, edited by H. Hofer et al., Progr. Math. **133**, Birkhäuser, Basel, 1995. MR 96k:58037 Zbl 0842.58033

[Ono 1996] K. Ono, "Lagrangian intersection under Legendrian deformations", *Duke Math. J.* **85**:1 (1996), 209–225. MR 97g:58036 Zbl 0868.58035

[Schwarz 1993] M. Schwarz, *Morse homology*, Progress in Mathematics **111**, Birkhäuser, Basel, 1993. MR 95a:58022 Zbl 0806.57020

Received January 11, 2006. Revised April 10, 2007.

HAI-LONG HER
CHERN INSTITUTE OF MATHEMATICS
NANKAI UNIVERSITY
TIANJIN 300071
CHINA
Current address:
INSTITUTE OF MATHEMATICAL SCIENCE
NANJING UNIVERSITY
NANJING 210093
CHINA
hailongher@ims.nju.edu.cn

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.437

HARMONIC NETS IN METRIC SPACES

JÜRGEN JOST AND LEONARD TODJIHOUNDE

We investigate harmonic maps from weighted graphs into metric spaces that locally admit unique centers of gravity, like Alexandrov spaces with upper curvature bounds. We prove an existence result by constructing an iterative geometric process that converges to such maps, called harmonic nets for short.

1. Introduction

This paper deals with harmonic maps from weighted graphs into metric spaces. Such maps can be considered as a generalization of geodesic lines in Riemannian manifolds. A geodesic, considered as a map $\gamma : [0, 1] \to N$ from the unit interval to the Riemannian manifold N and parametrized proportionally to arclength, is characterized by the property that for all sufficiently close $0 \le a < b \le 1$, the point $\gamma((a+b)/2)$ is the unique midpoint of $\gamma(a)$ and $\gamma(b)$, that is,

$$\gamma\Big(\frac{a+b}{2}\Big) = \operatorname{argmin}_{q\in N}\big(d^2(\gamma(a), q) + d^2(\gamma(b), q)\big). \tag{1}$$

This leads us to represent a geodesic as a string of points in N, each of which is the midpoint of its two neighbors. At the same time, this allows us flexible refinements: we can insert additional points as midpoints of consecutive ones already present. For that, it is useful to also consider the following slight generalization of (1):

$$\gamma(ta + (1-t)b) = \operatorname{argmin}_{q\in N}\big(td^2(\gamma(a), q) + (1-t)d^2(\gamma(b), q)\big),$$

where $0 < t < 1$.

A midpoint is a center of gravity of two points. In a Riemannian manifold, such centers of gravity exist locally uniquely, that is, when the points whose center is to be constructed are sufficiently close. Globally, uniqueness need not be true. Therefore, we may need to localize in the image.

It is then clear how to conceptualize a harmonic map from a weighted graph into N. We simply require that the images of the nodes of the graph by appropriately weighted centers of gravity of their neighbors. Here, in order to localize in the image, we might need to refine the graph by subdividing edges.

MSC2000: 58E20, 53C43, 53C22.
Keywords: harmonic map, weighted graph, center of gravity, refinement.

Harmonic maps from graphs into compact Riemannian manifolds were studied in [Kotani and Sunada 2001]. Our approach, however, naturally leads to a generalization to any metric space locally admitting unique centers of gravity — and this includes the important class of Alexandrov spaces with upper curvature bounds (see [Berestovskij and Nikolaev 1993] as a systematic reference).

Thus, we show the existence of harmonic maps from weighted graphs into such metric spaces. Such maps are called *harmonic nets* in short. (In the case of a space of nonpositive curvature in the sense of Alexandrov or Busemann, this result is contained in a general existence result for harmonic maps; see [Jost 1994; 1997].) The proof is not difficult. It is based on the iterative replacement of image points by the centers of gravity of the images of their neighbors, following the strategy described in [Jost 1998], together with suitable adaptive refinements to keep the constructions local. Here it is important that the domain, that is, our graph, can be treated as a one-dimensional object. While two dimensions represent a borderline case, in higher dimensions, general constructions of harmonic maps are only possible when the target space possesses nonpositive curvature. The reason is that the energy functional we are employing is quadratic, and therefore the scaling behavior is different in dimensions 1, 2, and greater than 2. The essential features of our scheme are local uniqueness and the scaling property of our functions.

Our constructions possess certain similarities with some schemes employed in numerical analysis, like the standard difference scheme for the numerical solution of the Laplace equation or adaptive refinements in multigrid methods. A key conceptual feature of our approach is that we systematically exploit the local uniqueness of solutions and that we need to make explicit only the images of a discrete set of points, because then all other images are implicitly determined by that local uniqueness. Therefore, as in good numerical schemes, we never have to work out or store more information than needed.

2. Geometric concepts

Let (N, d) be a complete metric space. For conciseness, we usually write N in place of (N, d), the metric d being understood. We say that N *admits refinements* if any $p, q \in N$ have a *midpoint*, that is, if there exists $m \in N$ such that

$$d(m, p) = d(m, q) = \tfrac{1}{2} d(p, q).$$

Definition 2.1. Suppose that N admits refinements. We define the radius $r(N)$ of unique refinement as the largest $r \in [0, \infty]$ with the property that for any two $p, q \in N$ with $d(p, q) \leq 2r$, their refinement (midpoint) $m = m(p, q)$ is unique.

More generally, we say that $q \in N$ is a *center of gravity* of the finitely many points $p_1, \ldots, p_n \in N$ with positive weights $w_1, \ldots, w_n$ if

$$q = \operatorname{argmin}_p \sum_{j=1}^{n} w_j^2 d^2(p, p_j).$$

We define the convexity radius $c(N)$ as the largest $c \in [0, \infty]$ with the property that whenever $p, p_1, \ldots, p_n$ are points in N with $d(p, p_i) \le c$ for $i = 1, \ldots, n$, the center of gravity of $p_1, \ldots, p_n$ (with any positive weights) is unique.

Since a midpoint is a center of gravity, we obviously have $0 \le c(N) \le r(N)$.

Let Γ be a finite weighted graph with vertex set I and edge set E, where each $e \in E$ has a weight $w(e) > 0$. We say that two vertices i, j are neighbors if they are connected by an edge. Thus, for our purposes, a graph is a discrete set I together with a symmetric neighborhood (adjacency) relation $\sim$ and (symmetric) weights $w(i, j) = w(e)$ for neighboring vertices i, j connected by the edge e.

We define the *refinement* Γ_r of Γ as the graph with vertex set $I \cup E$ and adjacency relation defined as follows: $i \in I$ and $e \in E$ are neighbors if $i \in e$ in Γ, and there are no other pairs of neighbors in Γ_r. The weights are $w(i, e) = \sqrt{2} w(e)$, where $w(e)$ is the weight of the edge e in Γ. Further refinements of Γ_r are defined iteratively.

A map f from Γ to N assigns to every $i \in I$ some point $p = f(i)$ in N. We define the *energy* of such a map $f : \Gamma \to N$ as

$$E(f) = \sum_{i \in I} E_i(f), \quad \text{where} \quad E_i(f) = \sum_{j \in I;\ i \sim j} w^2(i, j) d^2(f(i), f(j)).$$

In particular, for $i \sim j$,

$$d^2(f(i), f(j)) \le \frac{1}{w^2(i, j)} E_i(f). \tag{2}$$

We say that the map f is *harmonic* if for all $i \in I$, $f(i)$ is a center of gravity of the points $f(j)$, $j \sim i$, with weights $w_j = w^2(i, j)$.

3. Characterization by angles in tangent cones

The above concepts of refinement and center of gravity find their natural place in the context of Alexandrov's metric spaces. For a systematic development of this theory that we shall use in this section, see [Berestovskij and Nikolaev 1993].

These spaces enjoy particular properties when their (Alexandrov) curvature is bounded from above. It is part of the definition of such a space of curvature bounded from above that any two sufficiently close points can be joined by a shortest geodesic which then is in fact unique and depends continuously on these endpoints. (We may also parametrize it by arclength — and call it an arclength geodesic — if convenient.) Some of the general notions in the theory, however, do not need the assumption of an upper curvature bound. That assumption then is rather employed to derive geometric properties of the objects defined in the theory.

An important concept here is the tangent cone of a metric space at a point. Let (N, d) be a metric space and $\gamma_1, \gamma_2 : [0, \varepsilon] \to (N, d)$ arclength geodesics emanating from a point $P \in N$. Consider points $Q \in \gamma_1$, $R \in \gamma_2$ different from P. An (upper) angle $\theta(\gamma_1, \gamma_2)$ between γ_1 and γ_2 is defined by

$$\cos\theta(\gamma_1, \gamma_2) := \overline{\lim}_{Q,R\to P} \frac{d^2(P,Q) + d^2(P,R) - d^2(Q,R)}{2d(P,R)d(P,Q)}.$$

In a metric space whose curvature is bounded from above by a constant $K \geq 0$, we have the following characterization of the angle between γ_1 and γ_2 (see [Bridson and Haefliger 1999, II.1-II.3]):

$$\cos\theta(\gamma_1, \gamma_2) = \lim_{s\to 0} \frac{d(P, \gamma_2(\varepsilon)) - d(\gamma_1(s), \gamma_2(\varepsilon))}{s},$$

provided in case $K > 0$ that ε is less than the diameter of the comparison model space of constant curvature K.

A geodesic curve γ starting at a point $P \in N$ has a direction if $\theta(\gamma, \gamma) = 0$ and two curves have the same direction if the angle between them is equal to zero. This is an equivalence relation on the space of curves starting from the same point $P \in N$ and the completion of the set of equivalence classes (endowed with the distance induced by the angle) is called the space of directions $\Omega_P(N)$ of N at the point P. The tangent cone $T_P N$ of (N, d) at a point $P \in N$ is the cone over the space of directions, that is, $\Omega_P(N) \times \mathbb{R}_+$ with points in $\Omega_P(N) \times \{0\}$ identified together.

We will denote a tangent element by $[\gamma, x]$, where $\gamma \in \Omega_P(N)$, $x \geq 0$ and elements $[\gamma, 0]$ are identified with the origin O_p of $T_P N$.

The distance d in N induces on $T_P N$ a distance function $\tilde{d}_P$ defined by

$$\tilde{d}_P^2([\gamma_1, x_1], [\gamma_2, x_2]) = \begin{cases} x_1^2 + x_2^2 - 2x_1x_2\cos\theta(\gamma_1, \gamma_2) & \text{if } \theta(\gamma_1, \gamma_2) < \pi, \\ x_1 + x_2 & \text{if } \theta(\gamma_1, \gamma_2) \geq \pi. \end{cases}$$

For those $[\gamma, x]$ admitting a unique geodesic from P with direction γ that can be extended up to distance x, we define the endpoint of that geodesic segment as the exponential image of $[\gamma, x]$. The inverse of this exponential map, the projection map from the subset of N where it is defined to the tangent cone $T_P N$, is denoted by π_P. For simply connected, complete, nonpositively curved metric spaces, π_P is defined everywhere, distance nonincreasing and distance preserving in the radial direction; see [Wang 2000].

The following important result was proved in [Nikolaev 1995]:

Lemma 3.1. *Let (N, d) a metric space of curvature at most K, where $K \geq 0$. The tangent cone at a point of N is a space of nonpositive curvature in the sense of Alexandrov.*

Let P, Q, R be points in N and define $Q_s \equiv (1-s)\,P + s\,Q$ as the point on a distance-realizing geodesic joining P and Q at distance $s.d(P,Q)$ from P. We have the Taylor expansions

$$d^2(Q_s, R) = d^2(P, R) - 2sd(P, R)\cos\theta_P(Q, R) + a(s),$$
$$\tilde{d}^2_P(\pi_P(Q_s), \pi_P(R)) = d^2(P, R) - 2sd(P, R)\cos\theta_P(Q, R) + b(s),$$

in which $\lim_{s\to 0} a(s)/s = 0$ and $\lim_{s\to 0} b(s)/s = 0$. Here $\theta_P(Q, R)$ denotes the angle subtended between Q and R at P.

Proposition 3.2 [Izeki and Nayatani 2005; Wang 2000]. *Let $f : \Gamma \to (N, d)$ be a harmonic map.*

(i) *For any $i \in I$, the point $\pi_{f(i)}(f(i))$ minimizes*

$$\sum_{j\sim i} w^2(i, j)\tilde{d}^2_{f(i)}(\cdot\,, \pi_{f(i)} f(j)) \quad \text{in } T_{f(i)}N.$$

(ii) *For any $i \in I$ and any $V \in T_{f(i)}N$ we have*

$$\sum_{j\sim i} w^2(i, j)\big\langle V, \pi_{f(i)} f(j)\big\rangle \le 0,$$

where $\langle\ ,\ \rangle$ denotes the inner product defined on $T_{f(i)}N$ by

$$\big\langle [\gamma_1, x_1], [\gamma_2, x_2]\big\rangle = x_1 x_2 \cos\theta(\gamma_1, \gamma_2).$$

(iii) *For any $i \in I$, the center of gravity in $T_{f(i)}$ of the points $(\pi_{f(i)} f(j))_{j\sim i}$ with weights*

$$\left(\frac{w^2(i, j)}{w(i)}\right)_{j\sim i}$$

coincides with the origin $O_i := \pi_{f(i)} f(i)$ of $T_{f(i)}N$, where $w(i) = \sum_{j\sim i} w^2(i, j)$.

Inequality (ii) will be interpreted as the critical condition for harmonic nets.

4. Refining maps

If N admits refinements, we can construct a refinement $f_r : \Gamma_r \to N$ of a map $f : \Gamma \to N$ by assigning to every edge e connecting i and j in Γ some midpoint of $f(i)$ and $f(j)$. We observe that for each $i \in \Gamma$, we have

$$E_i(f_r) = \tfrac{1}{2} E_i(f) \tag{3}$$

where on the left hand side i is considered as an element of Γ_r. Also, by symmetry,

$$\sum_{i\in I} E_i(f_r) = \sum_{e\in E} E_e(f_r) = \tfrac{1}{2} E(f_r), \tag{4}$$

where we consider the i's and e's as vertices of Γ_r. From (3) and (4) we obtain:

Lemma 4.1. *If $f_r : \Gamma_r \to N$ is a refinement of $f : \Gamma \to N$, then $E(f_r) = E(f)$. If f is harmonic, so is its refinement.*

The converse holds when distances between images are sufficiently small, that is, when midpoints between images of neighbors are unique.

From (2) and Lemma 4.1, we conclude that by performing sufficiently many successive refinements, we may assume that all distances between the images of any two neighboring vertices are smaller than some prescribed $\epsilon > 0$, for example smaller than $r(N)$ or $c(N)$ when that quantity is positive.

5. Homotopy classes

For the present purposes, we write $r(f)$ and $r(\Gamma)$ instead of f_r and Γ_r because we wish to consider the refinement as an operation that can be iterated. For example, $r^2(\Gamma) = (\Gamma_r)_r$ is obtained as the refinement of Γ_r. A refinable map $f : \Gamma \to N$ is then considered as a collection of iteratively refined maps $r^n(f) : r^n(\Gamma) \to N$ for $n \in \mathbb{N}$.

Now assume that the refinement radius $r(N)$ is positive. We say that two maps $f_1, f_2 : \Gamma \to N$ are *geodesically close* if for every $i \in \Gamma$ the distance $d(f_1(i), f_2(i))$ is at most $2r(N)$; that is, if the images of i under f_1 and f_2 have a unique midpoint. A refinement of the pair f_1, f_2 is then defined to be the triple $f_1, f_{1,2}, f_2$, where

$$f_{1,2}(i) \text{ is the midpoint of } f_1(i) \text{ and } f_2(i) \quad \text{for every } i \in \Gamma.$$

Two refinable maps $f, g : \Gamma \to N$ are *geometrically homotopic* if there exist refinable maps $f_0 = f, f_1, f_2, \ldots, f_A = g$ (where $A \in \mathbb{N}$) such that $r^n(f_{j-1})$ and $r^n(f_j)$ are geodesically close for any $n \in \mathbb{N}$ and any $1 \le j \le A$. This finite sequence can again be refined by putting in midpoint maps between consecutive sequence elements. Geometric homotopy is obviously an equivalence relation.

6. Construction of harmonic nets

We assume that N admits centers of gravity. By subdividing suitable edges of Γ as above, we may assume that Γ is bipartite, that is, its vertex set is a disjoint union $I = I_1 \cup I_2$ such that all the neighbors of any point in one of those subsets are contained in the other one. On the space $C = C(\Gamma, N)$ of maps $f : \Gamma \to N$, we define maps $\rho_\alpha : C \to C$, $\alpha = 1, 2$ with $\rho_\alpha(f)$ being the map obtained from $f : \Gamma \to N$ by replacing the image of every $f(i)$ for $i \in I_\alpha$ by a center of gravity of the $f(j)$ for $j \sim i$. As long as the centers of gravity are not unique, we need to make choices here, but in the situation where $c(N) > 0$, we can assume that Γ has been sufficiently refined (depending on an upper bound E for the energy of f) so that the images $f(j)$ of the neighbors of any $i \in \Gamma$ possess a unique center of gravity. (This follows from (2) and the fact that the edge weights get multiplied

by a factor of $\sqrt{2}$, that is, become larger, under each refinement.) In that case, the maps $\rho_\alpha(f)$ are unambiguously defined for all f with $E(f) \le E$.

The function ρ_α decreases (or rather, does not increase) the energy density $E_i(f)$ for points $i \in I_\alpha$, but not necessarily for those in the complement of I_α. Nevertheless, since by symmetry $\sum_{i\in I_1} E_i(f) = \sum_{i\in I_2} E_i(f) = \frac{1}{2}E(f)$ (see (4)), we have:

Lemma 6.1. *We have $E(\rho_\alpha(f)) \le E(f)$ for all f.*

Lemma 6.2. *We have $E(\rho_2(\rho_1(f))) = E(f)$ if and only if f is harmonic.*

Theorem 6.3. *Let (N, d) be a compact metric space that admits centers of gravity. Let Γ be a finite weighted graph. Then, for any map $f : \Gamma \to N$, the iterations $f_n := (\rho_2\rho_1)^n f$ contain a subsequence converging to a harmonic map.*

Proof. Since N is compact, we can find some sequence $\nu(n)$ of positive integers going to infinity for which $f_{\nu(n)}(i)$ converges to some point $f_0(i) \in N$ for every vertex i of the finite graph Γ. We have

$$f_{\nu(n+1)} = (\rho_2\rho_1)^{\mu(n)} f_{\nu(n)} \quad \text{for some } \mu(n) \in \mathbb{N}.$$

Since the metric d behaves continuously under convergence (since it defines the topology of N), we have $E(f_0) = \lim_{n\to\infty} E(f_{\nu(n)})$. At the same time,

$$\lim E(f_{\nu(n+1)}) = \lim E((\rho_2\rho_1)^{\mu(n)} f_{\nu(n)}) \le \lim E(f_{\nu(n)}) = E(f_0),$$

the inequality coming from Lemma 6.1 because $\mu(n) \ge 1$. Thus, equality has to hold throughout. Moreover, $\rho_2\rho_1 f_{\nu(n)}$ converges to $\rho_2\rho_1 f_0$, and so

$$\begin{aligned} E(\rho_2\rho_1 f_0) &= \lim E((\rho_2\rho_1)^{\mu(n)+1} f_{\nu(n)}) && \text{(as before)} \\ &= E(f_0) && \text{(from the preceding observation.)} \end{aligned}$$

Lemma 6.2 then implies that f_0 is harmonic. □

The assumption of the theorem that the space N admits centers of gravity is satisfied when N has an upper curvature bound. For $k \in \mathbb{R}$, we denote by D_k the diameter of the n-dimensional, complete, simply connected model space with constant sectional curvature k. We then have, from Alexandrov theory:

Lemma 6.4 [Berestovskij and Nikolaev 1993]. *Let X be an Alexandrov space with curvature bounded above by k. For every $x \in X$, there exists a positive number $R_x \in (0, \frac{1}{2}D_k]$ such that the closed metric ball of radius R_x centered at x is a convex subset in X.*

Remark. When N has nonpositive curvature in the sense of Alexandrov or Busemann, our theorem is contained in a general theorem from [Jost 1994], and when N is a compact Riemannian manifold, it follows from the fact that any homotopy class contains at least one harmonic map [Kotani and Sunada 2001].

Since distances of neighboring image points are controlled by the energy of a map — see (2) — we see that if the refinement radius $r(N)$ is positive, we may control the geometric homotopy class by assuming an energy bound and sufficiently refining the graph Γ.

References

[Berestovskij and Nikolaev 1993] V. N. Berestovskij and I. G. Nikolaev, "Multidimensional generalized Riemannian spaces", pp. 165–250 in *Geometry IV*, edited by Y. G.Reshetnyak, Encyclopaedia Math. Sci. **70**, Springer, Berlin, 1993. MR MR1263965

[Bridson and Haefliger 1999] M. R. Bridson and A. Haefliger, *Metric spaces of non-positive curvature*, Grundlehren der Math. Wissenschaften **319**, Springer, Berlin, 1999. MR 2000k:53038 Zbl 0988.53001

[Izeki and Nayatani 2005] H. Izeki and S. Nayatani, "Combinatorial harmonic maps and discrete-group actions on Hadamard spaces", *Geom. Dedicata* **114** (2005), 147–188. MR 2006k:58024 Zbl 02227558

[Jost 1994] J. Jost, "Equilibrium maps between metric spaces", *Calc. Var. Partial Differential Equations* **2**:2 (1994), 173–204. MR 98a:58049 Zbl 0798.58021

[Jost 1997] J. Jost, *Nonpositive curvature: geometric and analytic aspects*, Birkhäuser, Basel, 1997. MR 98g:53070 Zbl 0896.53002

[Jost 1998] J. Jost, *Riemannian geometry and geometric analysis*, 2nd ed., Springer, Berlin, 1998. Zbl 0997.53500

[Kotani and Sunada 2001] M. Kotani and T. Sunada, "Standard realizations of crystal lattices via harmonic maps", *Trans. Amer. Math. Soc.* **353**:1 (2001), 1–20. MR 2001j:58028 Zbl 0960.58009

[Nikolaev 1995] I. Nikolaev, "The tangent cone of an Aleksandrov space of curvature $\leq K$", *Manuscripta Math.* **86**:2 (1995), 137–147. MR 95m:53062 Zbl 0822.53043

[Wang 2000] M.-T. Wang, "Generalized harmonic maps and representations of discrete groups", *Comm. Anal. Geom.* **8**:3 (2000), 545–563. MR 2001m:58039 Zbl 0977.58018

Received July 7, 2006.

JÜRGEN JOST
MAX PLANCK INSTITUTE FOR MATHEMATICS IN THE SCIENCES
INSELSTRASSE 22-26
D-04103 LEIPZIG
GERMANY
jjost@mis.mpg.de

LEONARD TODJIHOUNDE
INSTITUT DE MATHÉMATIQUES ET DE SCIENCES PHYSIQUES
B.P. 613
PORTO-NOVO
BENIN
leonardt@imsp-uac.org

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.445

THE QUANTITATIVE HOPF THEOREM AND FILLING VOLUME ESTIMATES FROM BELOW

LUOFEI LIU

Let V^n be a compact, oriented Riemannian manifold and S^n the standard sphere. We study the problem of obtaining upper bounds for the dilatation invariants of maps $V^n \to S^n$ of nonzero degree. The dilatation upper bounds are then used to estimate Gromov's filling volume from below.

1. Introduction

Let V be a compact, connected, oriented n-manifold. The famous Hopf theorem says that there is an one-to-one correspondence between the degree and the homotopy class of continuous maps from V to S^n. If V is equipped with a Riemannian metric and S^n is equipped the canonical metric (of sectional curvature $+1$), we wish to measure the geometrical complexity of a map $V \to S^n$ of degree q. A natural measure of the geometrical complexity $f : X \to Y$ of a map between metric spaces is its dilatation, defined by

$$\operatorname{dil} f = \sup_{\substack{x,x' \in X \\ x \neq x'}} \frac{d_Y(f(x), f(x'))}{d_X(x, x')},$$

where d_X, d_Y are the respective metrics. If f is a Lipschitz map, $\operatorname{dil} f$ is the minimal Lipschitz constant C satisfying $d_Y(f(x), f(x')) \leq C d_X(x, x')$. If $\operatorname{Dil}_q V$ denotes the infimum of the dilatation of maps $V \to S^n$ with degree $\pm q$, an upper bound for $\operatorname{Dil}_q V$ can be regarded as a quantitative version of the Hopf theorem. For V a flat torus, M. Gromov proved such a bound in terms of the length $\operatorname{sys}_1 V$ of the systole (shortest noncontractible closed curve) in V.

Theorem 1.1 [Gromov 1999b, 2.12, p. 33]. *Let $\mathbb{T}^n$ be a flat torus. There exists a map $f : \mathbb{T}^n \to (S^n, \text{can})$ with nonzero degree and dilatation at most 1 if and only if $\operatorname{sys}_1 \mathbb{T}^n \geq 2\pi$.*

It has been Gromov's constant concern to study quantitative problems in algebraic topology and differential topology [1978; 1999a; 1999b, Chapters 2 and 7]).

MSC2000: primary 53C23, 53C20; secondary .
Keywords: dilatation, degree, filling volume.

In this direction, A. Nabutovsky and R. Rotman [2003] have studied a quantitative Hurewicz theorem. For estimates of dilatation invariants of maps between spheres, see [Peng and Tang 2002] and references therein.

Here is a slight generalization of a problem posed by Gromov after the theorem just quoted.

Problem [Gromov 1999b, 2.13, p. 35]. *Let V be a compact, connected, oriented n-manifold. What condition on the metric of V guarantees that there exist mappings $f : V \to S^n$ with degree q and dilatation* 1? *In other words, given a metric on V, for which values of q do there exist mappings $V \to S^n$ with degree q and dilatation less than some constant D?*

L. Guth [2005] has obtained important results concerning the dilatation of a nonzero degree map from a Riemannian surface to the standard 2-sphere. Following Guth, we denote by $\mathrm{HS}(V)$ the hypersphericity of a Riemannian n-manifold, defined as the maximal R such that there is a contracting map ($\mathrm{dil} \le 1$) of nonzero degree from V to the n-sphere of radius R. Replacing nonzero degree by degree q, one defines the degree-q hypersphericity $\mathrm{HS}_q(V)$ of a Riemannian manifold. It is clear that

$$\mathrm{HS}(V) \ge \mathrm{HS}_q(V) = (\mathrm{Dil}_q\ V)^{-1} \quad \text{for } q \ne 0.$$

In this paper we prove several results concerning dilatation. Our method is to construct directly some maps $V \to S^n$ of nonzero degree, and then estimate their dilatation using (geo)metric invariants of V.

Recall that the q-th packing radius, $\mathrm{pack}_q(X)$, of a compact metric space X in the sense of K. Grove and S. Markvorsen [1995] is the largest $r \ge 0$ for which X contains q disjoint open r-balls. It is clear that

$$\tfrac{1}{2}\operatorname{diam} X = \operatorname{pack}_2 X \ge \cdots \ge \operatorname{pack}_q X \ge \cdots \to 0.$$

Theorem A. (Proved in Section 2.) *Let V be a compact, oriented Riemannian n-manifold and let K be the supremum of the sectional curvature of V.*

(1) *In the case $K > 0$, if q is large enough to satisfy*

$$\operatorname{pack}_q V < \min\{\pi/\sqrt{K},\ \operatorname{Inj} V\},$$

there exists a map $f : V \to S^n$ of degree q such that

$$\operatorname{dil} f \le \frac{\pi\sqrt{K}}{\sin\bigl(\sqrt{K}\ \operatorname{pack}_q V\bigr)}.$$

Hence

$$\mathrm{HS}_q(V) \ge \sin\bigl(\sqrt{K}\ \operatorname{pack}_q V\bigr)/(\pi\sqrt{K}).$$

(2) *In the case* $K \leq 0$, *if* q *is large enough to satisfy*

$$\operatorname{pack}_q V \leq \operatorname{Inj} V,$$

there exists a map $f : V \to S^n$ *of degree* q *and*

$$\operatorname{dil} f \leq \frac{\pi}{\operatorname{pack}_q V}.$$

Hence $\operatorname{HS}_q(V) \geq (\operatorname{pack}_q V)/\pi$.

Here $\operatorname{Inj} V$, the injectivity radius of V, is the infimum of the injectivity radius at any point in V. For a flat torus $\mathbb{T}^n$ we have $\operatorname{Inj} \mathbb{T}^n = \frac{1}{2} \operatorname{sys}_1 \mathbb{T}^n$.

We denote by $\operatorname{Inj}_{\max} V = \max\{\operatorname{Inj}_x V : x \in V\}$ the largest injectivity radius of a compact Riemannian manifold V. The following corollary may be regarded as a slight generalization of the "if" part of Theorem 1.1.

Corollary 1.2. *Let* V *be a compact Riemannian* n*-manifold.*

(1) *If* $\sup \operatorname{sec}(V) \leq 1/\pi^2$ *and* $\operatorname{Inj}_{\max} V \geq \frac{1}{2}\pi^2$, *there is a map* $f : V \to S^n$ *with degree* 1 *and dilatation* 1.

(2) *If* $\sup \operatorname{sec}(V) \leq 0$ *and* $\operatorname{Inj}_{\max} V \geq \pi$, *there is a map* $f : V \to S^n$ *with degree* 1 *and dilatation* 1.

By deepening the results of D. Burago and S. Ivanov [1995], Gromov [1999b, p. 259] proved a sharp inequality relation between $\operatorname{Vol}(V)$ and $\operatorname{stsys}_1 V$ (the stable 1-systole of V — see page 452) under certain topological restrictions on V. In [Ivanov and Katz 2004; Katz and Lescop 2005], this inequality is stated as follows:

Theorem 1.3 (Burago–Ivanov–Gromov Inequality). *Assume that a compact, oriented Riemannian manifold* V *satisfies* $\dim(V) =$ *first Betti number* $b_1(V) =$ *real cuplength of* $V = n$. *Then*

$$\operatorname{Vol}(V) \geq (\gamma_n)^{-n/2} (\operatorname{stsys}_1 V)^n$$

where γ_n *denotes the classical Hermite constant.*

Theorem 1.3 and [Ivanov and Katz 2004, Lemma 7.3] lead us to:

Theorem B. *Assume that a compact, oriented Riemannian manifold* V *satisfies* $dim(V) = b_1(V) =$ *real cuplength of* $V = n$. *Then there exists a map* $f : V \to S^n$ *of nonzero degree such that*

$$\operatorname{dil} f \leq \frac{2\pi\sqrt{n}}{\operatorname{stsys}_1 V}.$$

Hence $\operatorname{HS}(V) \geq (\operatorname{stsys}_1 V)/(2\pi\sqrt{n})$.

The proof of Theorem B, which we give in Section 3, makes heavy use of the construction in [Ivanov and Katz 2004, Section 7], itself based on the techniques of [Burago and Ivanov 1995].

An immediate application of the quantitative Hopf theorem is that we can estimate volume from below by the formula

$$\mathrm{Vol}(V) \geq \frac{|\deg f|}{(\mathrm{dil}\, f)^n} \mathrm{Vol}(S^n, \mathrm{can}).$$

However, in general, using estimates for the Jacobian of a Lipschitz map, one can obtain better lower volume bounds [Ivanov and Katz 2004; Gromov 1999b, pp. 255, 257].

A second goal of this paper is to use quantitative Hopf theorem to estimate the filling volume of a Riemannian manifold. We believe the quantitative versions of the Hopf theorem to be of independent interest. For example, we have used them to estimate lower bounds for the filling radius of a Riemannian manifold in [Liu 2005].

Let V be a compact n-manifold equipped with a distance d (not necessarily Riemannian), and let $L^\infty(V)$ be the Banach space of bounded Borel functions f on V with the norm $\|f\| = \sup_x |f(x)|$. The map $i : V \to L^\infty(V)$ defined by $x \mapsto f_x(\cdot) = \mathrm{dist}(x, \cdot)$ is an isometric (distance-preserving) embedding. The filling volume $\mathrm{FillVol}(V)$ of (V, d) in the Gromov's sense is, roughly speaking, the infimum of the volumes of $(n+1)$-dimensional submanifolds in $L^\infty(V)$ whose boundary is $i(V)$. For details, see [Gromov 1983] or Section 4 below.

Our main tool to estimate filling volume from below is the following mapping property:

Theorem C. (Proved in Section 5.) *Let V be a compact, oriented n-manifold with a metric d and let $f : V \to S^n$ a map of nonzero degree. Then*

$$\mathrm{FillVol}(V) > \frac{|\deg f|}{(\mathrm{dil}\, f)^{n+1}} \left(\arccos \frac{n-1}{n+1}\right)^{n+1}.$$

In the definition of hypersphericity, it is not necessary that V be Riemannian, merely that V have a metric d. Hence an equivalent statement of Theorem C is that for any compact, oriented n-manifold V with a metric d,

$$\mathrm{FillVol}(V, d) > \left(\arccos \frac{n-1}{n+1}\right)^{n+1} |q|\ \mathrm{HS}(V, d)^{n+1}.$$

Define

$$c_n = \left(\frac{1}{2\pi} \arccos \frac{n-1}{n+1}\right)^{n+1}.$$

From Theorem C we derive corollaries of Theorems 1.1, A and B, respectively:

Corollary 1.4. *Let $\mathbb{T}^n$ be a flat torus. Then*

$$\mathrm{FillVol}(\mathbb{T}^n) > c_n\,(\mathrm{sys}_1\,\mathbb{T}^n)^{n+1}.$$

Corollary 1.5. *Let V be a compact, oriented Riemannian n-manifold.*

(1) *If $K = \sup \mathrm{sec}(V) > 0$ and $\mathrm{pack}_q V < \min\{\pi/\sqrt{K}, \mathrm{Inj}\, V\}$, then*

$$\mathrm{FillVol}(V) > q\left(\frac{\sin\left(\sqrt{K}\,\mathrm{pack}_q V\right)}{\pi\sqrt{K}}\,\arccos\frac{n-1}{n+1}\right)^{n+1}.$$

(2) *If $K = \sup \mathrm{sec}(V) \le 0$ and $\mathrm{pack}_q V \le \mathrm{Inj}\, V$, then*

$$\mathrm{FillVol}(V) > 2^{n+1} c_n\, q\,(\mathrm{pack}_q V)^{n+1}.$$

Corollary 1.6. *Let V be a compact Riemannian manifold of equal dimension, first Betti number and real cuplength. Then*

$$\mathrm{FillVol}(V) > n^{-(n+1)/2} c_n\,(\mathrm{stsys}_1 V)^{n+1}.$$

The following statements are alternative forms of Corollaries 1.4 and 1.6, relating the volumes of Riemannian manifolds with boundary to the boundary metric invariants. Their proofs appear in Section 4. (We omit a similar Corollary 1.5'.)

Corollary 1.4′. *Let W be a $(n+1)$-dimensional solid torus with boundary $\partial W = \mathbb{T}^n$. Given a flat metric g_0 on $\mathbb{T}^n$, g is any Riemannian metric on W which satisfies $d_g|_{\partial W} \ge d_{g_0}$, then*

$$\mathrm{Vol}(W, g) > c_n \cdot \mathrm{sys}_1(\mathbb{T}^n, g_0)^{n+1}.$$

Corollary 1.6′. *Assume that the topological restrictions of a closed n-manifold V are as Corollary 1.6, W is a $(n+1)$-manifold with boundary $\partial W = V$. Given a Riemannian metric g_0 on V, for any Riemannian metric g on W satisfying $d_g|_{\partial W} \ge d_{g_0}$, we have*

$$\mathrm{Vol}(W, g) \ge n^{-(n+1)/2} c_n \cdot \mathrm{stsys}_1(V, g_0)^{n+1}.$$

The constants in these corollaries are not sharp. We do not look for anything like the optimal constants in all estimate inequalities of this paper; but these inequalities ensure the upper bounds of dilatation invariants, or the lower bounds of filling volume, in terms of packing radius, stable 1-systole, etc.

2. Dilatation estimates with upper bounds of sectional curvature

The proof of Theorem A is a direct extension of the arguments used to prove Theorem 1.1 and [Liu 2005, Propositions 3.2 and 3.3]. For completeness, we will give a detailed proof of Theorem A in this section.

Let $f : X \to Y$ be a map between metric spaces. The local dilatation at $x \in X$ is the number

$$\operatorname{dil}_x f = \lim_{\varepsilon \to 0} (\operatorname{dil}(f|_{B(x,\varepsilon)})),$$

where $B(x, \varepsilon)$ is an open ball in X of radius ε centered at x. If (X, d) is a path metric space (for example, a Riemannian manifold), then $\operatorname{dil} f = \sup_{x \in X} \operatorname{dil}_x f$. If X and Y are Riemannian manifolds, and f is differentiable, then $\operatorname{dil}_x f = \|df_x\|$, where $df_x : T_x X \to T_{f(x)} Y$ is the differential of f at x.

Proof of Theorem A. (1) Assume $K = \sup \sec(V) > 0$. Let $B(x_1, r), \ldots, B(x_q, r)$ be disjoint r-balls of V centered at the points $x_1, \ldots, x_q$, where

$$r = \operatorname{pack}_q V < \min\{\pi/\sqrt{K}, \operatorname{Inj} V\}.$$

Denote by $B_T(x_i, r)$ the ball of radius r *in the tangent space* $T_{x_i} V$, centered at the origin. Since $r < \pi/\sqrt{K}$, the map $\exp_{x_i} |B_T(x_i, r)$ is nonsingular for each i; see the remark after 1.29 in [Cheeger and Ebin 1975]. Using the geometric Rauch theorem [Chavel 1993, Theorem 7.3 and subsequent Remark 7.1], we have

$$\frac{|d_v(\exp_{x_i})(X)|}{|X|} \geq \frac{S_K(|v|)}{|v|} = \frac{\sin(\sqrt{K}|v|)}{\sqrt{K}|v|}$$

for any $v \in B_T(x_i, r)$ and any $X \in T_{x_i} V$, where $d_v(\exp_{x_i}) : T_v(B_T(x_i, r)) \to T_y V$, with $y = \exp_{x_i}(v)$, denotes the differential of $\exp_{x_i}$ at v and the tangent space $T_v(B_T(x_i, r))$ is naturally identified with $T_{x_i} V$. Since $\exp_{x_i} : B_T(x_i, r) \to B(x_i, r)$ is a diffeomorphism, we have $d_y(\exp_{x_i}^{-1}) = (d_v(\exp_{x_i}))^{-1}$ for $v \in B_T(x_i, r)$ and $y = \exp_{x_i}(v) \in B(x_i, r)$. Let $d_v(\exp_{x_i})(X) = Y$. We have

$$\frac{|d_y(\exp_{x_i}^{-1})(Y)|}{|Y|} \leq \frac{\sqrt{K}|v|}{\sin(\sqrt{K}|v|)}, \quad \operatorname{dil}_y \exp_{x_i}^{-1} \leq \frac{\sqrt{K}|v|}{\sin(\sqrt{K}|v|)}.$$

Note that $\sqrt{K}|v|/\sin(\sqrt{K}|v|)$ is an increasing function of $|v| \in (0, r]$. Thus

$$\operatorname{dil} \exp_{x_i}^{-1} \leq \sup_{v \in B_T(x_i, r)} \frac{\sqrt{K}|v|}{\sin(\sqrt{K}|v|)} = \frac{\sqrt{K} \cdot r}{\sin(\sqrt{K} \cdot r)}.$$

Fix a point $p \in S^n$ and consider the composite $f_i : B(x_i, r) \to S^n$ defined by

$$B(x_i, r) \xrightarrow{\exp_{x_i}^{-1}} B_T(x_i, r) \xrightarrow{\varphi_1} B_T(x_i, \pi) \xrightarrow{\varphi_2} B_T(p, \pi) \xrightarrow{\varphi_3} S^n, \tag{1}$$

where φ_1 is the obvious diffeomorphism with dilatation π/r, φ_2 is the obvious isometry onto a ball $B_T(p, \pi)$ of radius π centered at the origin in the tangent space to S^n at p, and φ_3 has degree 1, dilatation 1, and maps $\partial B_{T, S^n}(p, \pi)$ to the antipode p' of p in S^n. The map f_i sends $\partial B_V(x_i, r)$ to p'.

It follows that $\deg f_i = 1$ and

$$\operatorname{dil} f_i \le \operatorname{dil}\varphi_3 \operatorname{dil}\varphi_2 \operatorname{dil}\varphi_1 \operatorname{dil}\exp_{x_i}^{-1} \le \frac{r\sqrt{K}}{\sin(r\sqrt{K})}.$$

Finally, define $f : V \to S^n$ by

$$f(x) = \begin{cases} f_i(x) & \text{for } x \in B(x_i, r), \\ p' & \text{elsewhere.} \end{cases}$$

It is clear that $\deg f = q$ and $\operatorname{dil} f \le r\sqrt{K}/\sin(r\sqrt{K})$.

(2) Now suppose $K = \sup \sec(V) \le 0$. Take q disjoint open balls $B(x_i, r)$ with $r = \operatorname{pack}_q V$. In this case, since V has no conjugate points, we require only $r = \operatorname{pack}_q V \le \operatorname{Inj} V$. In analogy with the above argument, we also have $\exp_{x_i}^{-1} : B(x_i, r) \to B_T(x_i, r)$, the composite map $f_i : B(x_i, r) \to S^n$, and $f : V \to S^n$. Since V is nonpositively curved, $\exp_{x_i}^{-1}$ is distance-decreasing. Then $\operatorname{dil}\exp_{x_i}^{-1} \le 1$ and $\operatorname{dil} f_i \le \pi/r$, hence $\operatorname{dil} f \le \pi/r$. This completes the proof of Theorem A. □

In proving Theorem A we have shown Corollary 1.2. Indeed, take a point $x_0 \in V$ such that $\operatorname{Inj}_{x_0} V = \operatorname{Inj}_{\max} V =: r$, we have a map $B_V(x_0, r) \to S^n$, whose extension $\varphi_{x_0} : V \to S^n$ is of degree one. We then estimate $\operatorname{dil}\varphi_{x_0}$. For details see [Liu 2005].

3. Abel–Jacobi map, stable 1-systoles and the proof of Theorem B

Let V be a compact, oriented n-manifold with first Betti number $b_1(V) = n$ and let $i_* : H_1(V, \mathbb{Z}) \to H_1(V, \mathbb{R})$ be the map induced by chain inclusion. We denote by $H_1(V, \mathbb{Z})_{\mathbb{R}}$ the image $\Im(i_*(H_1(V, \mathbb{Z})))$, which is a lattice in $H_1(V, \mathbb{R}) \simeq \mathbb{R}^n$. Let $J_1(V) = H_1(V, \mathbb{R})/H_1(V, \mathbb{Z})_{\mathbb{R}} \simeq \mathbb{T}^n$ be the Jacobi torus of V. Up to homotopy, we have the Abel–Jacobi map

$$\mathscr{A}_V : V \to J_1(V)$$

induced by the harmonic one-forms on V, originally introduced by A. Lichnerowicz [1969] (see also [Gromov 1999b, p. 249; Bangert and Katz 2004; Katz and Lescop 2005). The covering $\tilde{V} \to V$ is the pull-back along $\mathscr{A}_V$ of the universal covering $\mathbb{R}^n \to J_1(V)$, with the group of deck transformations $H_1(V, \mathbb{Z})_{\mathbb{R}} \simeq \mathbb{Z}^n$, as depicted in the commutative diagram

$$\begin{array}{ccc} \tilde{V} & \xrightarrow{\tilde{\mathscr{A}}_V} & H_1(V, \mathbb{R}) \simeq \mathbb{R}^n \\ \downarrow & & \downarrow \\ V & \xrightarrow[\mathscr{A}_V]{} & J_1(V) \simeq \mathbb{T}^n \end{array}$$

The key point in the following topological lemma is that $H^1(\mathscr{A}_V) : H^1(\mathbb{T}^n, \mathbb{R}) \to H^1(V, \mathbb{R})$ is an isomorphism and $H^*(\mathscr{A}_V)$ preserves cup products.

Lemma 3.1 [Gromov 1983, Section 7.5]. *Let V be a compact, oriented n-manifold with first Betti number $b_1(V) = n$. If the real cuplength is also n, then $\mathcal{A}_V$ has nonzero degree.* □

If V is equipped with a Riemannian metric g, there is a corresponding stable norm $\|\cdot\|_{\mathrm{st}}$ on $H_1(V, \mathbb{R})$ (see [Gromov 1999b, pp. 245–247; Bangert and Katz 2004], for example). We define the *stable 1-systole* of (V, g) as

$$\mathrm{stsys}_1(V) = \min\{\|h\|_{\mathrm{st}} : h \in H_1(V, \mathbb{Z})_{\mathbb{R}} \setminus \{0\}\}.$$

The construction of f in Theorem B is a combination $h \circ \psi$ where $\psi : V \to \mathbb{T}^n$ (a flat torus) comes from [Ivanov and Katz 2004, Section 7], based on the techniques of [Burago and Ivanov 1995].

From the stable norm $\|\cdot\|_{\mathrm{st}}$ on $H_1(V, \mathbb{R})$ one derives a Euclidean norm $|\cdot|_e$ such that $|\cdot|_e \geq \|\cdot\|_{\mathrm{st}}$ and

$$|\cdot|_e^2 = \sum_{i=1}^{N} a_i L_i^2 \quad \text{with} \quad N \leq \frac{n(n+1)}{2}, \quad a_i > 0 \text{ for } i = 1, \ldots, N, \quad \sum_{i=1}^{N} a_i = n,$$

where the $L_i : H_1(V, \mathbb{R}) \to \mathbb{R}$ $(i = 1, \ldots, N)$ are linear functions with $\|L_i\|_{\mathrm{st}} = |L_i|_e = 1$ (see [Burago and Ivanov 1995, Theorem 1.3]). The linear map $L : H_1(V, \mathbb{R}) \to \mathbb{R}^N$ defined by

$$L(x) = \left(\sqrt{a_1} L_1(x), \ldots, \sqrt{a_N} L_N(x)\right)$$

is an isometry from $(H_1(V, \mathbb{R}), |\cdot|_e)$ onto a linear subspace $L(H_1(V, \mathbb{R}))$ of $\mathbb{R}^N$, equipped with the restriction of the standard coordinate metric $|\cdot|_E$ of $\mathbb{R}^N$.

Since the stable norm $\|\cdot\|_{\mathrm{st}}$ satisfies

$$\|\gamma\|_{\mathrm{st}} \leq d_g(x, x + \gamma)$$

for all $\gamma \in H_1(V, \mathbb{Z})_{\mathbb{R}}$ and all $x \in \tilde{V}$, where $+\gamma$ denotes the action on $\tilde{V}$ of an element γ of the deck transformation group $H_1(V, \mathbb{Z})_{\mathbb{R}}$, we have:

Lemma 3.2 [Ivanov and Katz 2004, Lemma 7.3]. *For each linear function $L_i : H_1(V, \mathbb{R}) \to \mathbb{R}$ with $\|L_i\|_{\mathrm{st}} = 1$ there is a Lipschitz map $\varphi_i : \tilde{V} \to \mathbb{R}$ such that $\mathrm{dil}\, \varphi_i \leq 1$ and*

$$\varphi_i(x + \gamma) = \varphi_i(x) + L_i(\gamma) \qquad \text{for } x \in \tilde{V},\ \gamma \in H_1(V, \mathbb{Z})_{\mathbb{R}}.$$ □

Proposition 3.3 [Ivanov and Katz 2004, Section 7]. *Let (V, g) be a compact, oriented Riemannian manifold with $\dim(V) = b_1(V) =$ real cuplength$(V) = n$. There exists a Lipschitz map $\psi : (V, d_g) \to (J_1(V), |\cdot|_e)$ of nonzero degree such that $\mathrm{dil}\, \psi \leq \sqrt{n}$, where $|\cdot|_e$ on $J_1(V)$ is the flat Riemannian metric induced from $H_1(V, \mathbb{R}), |\cdot|_e)$.*

Proof. From Lemma 3.2, define $F : \tilde{V} \to \mathbb{R}^N$ by

$$F(x) = \big(\sqrt{a_1}\varphi_1(x), \sqrt{a_2}\varphi_2(x), \ldots, \sqrt{a_N}\varphi_N(x)\big).$$

It is easy to check that L and F are $H_1(V, \mathbb{Z})_{\mathbb{R}}$-equivariant with respect to the following action of $H_1(V, \mathbb{Z})_{\mathbb{R}}$ on $\mathbb{R}^N$

$$\mathbb{R}^N \times H_1(V, \mathbb{Z})_{\mathbb{R}} \to \mathbb{R}^N \quad (u, \gamma) \to u + L(\gamma)$$

where L is as the beginning of this section.

Let $P : \mathbb{R}^N \to L(H_1(V, \mathbb{R}))$ be the orthogonal projection. Since the composite map

$$L^{-1} \circ P \circ F : \tilde{V} \to \mathbb{R}^N \to L(H_1(V, \mathbb{R})) \to H_1(V, \mathbb{R})$$

is $H_1(V, \mathbb{Z})_{\mathbb{R}}$-equivariant, we can lower it to a map between base spaces:

$$\psi : V \to H_1(V, \mathbb{R})/H_1(V, \mathbb{Z})_{\mathbb{R}} = J_1(V).$$

From $\operatorname{dil} \varphi_i \le 1$ we have

$$\operatorname{dil} F \le \Big(\sum_{i=1}^{N} \operatorname{dil}(\sqrt{a_i}\varphi_i)^2 \Big)^{1/2} \le \Big(\sum_{i=1}^{N} a_i \Big)^{1/2} = \sqrt{n},$$

$$\operatorname{dil}(L^{-1} \circ P \circ F) \le \operatorname{dil} L^{-1} \operatorname{dil} P \operatorname{dil} F \le \sqrt{n},$$

with respect to d_g on $\tilde{V}$ and $|\cdot|_e$ on $H_1(V, \mathbb{R})$. This implies that $\operatorname{dil} \psi \le \sqrt{n}$.

Finally, there is a linear homotopy $\{G_t\}$ between $L^{-1} \circ P \circ F$ and $\tilde{\mathcal{A}}_V$, i.e., $G_t(x) = t(L^{-1} \circ P \circ F(x)) + (1 - t)(\tilde{\mathcal{A}}_V(x))$. It is easy to see that each G_t is $H_1(V, \mathbb{Z}_{\mathbb{R}})$-equivariant; hence it may lower to a homotopy between ψ and $\mathcal{A}_V$. Therefore $\deg \psi = \deg \mathcal{A}_V$. By Lemma 3.1, this completes the proof of the proposition. □

Remark 3.4. By estimating the Jacobian of ψ, Ivanov and Katz [2004] proved that ψ is volume-decreasing.

Proof of Theorem B. Let $(J_1(V), |\cdot|_e)$ be as above. For $x \in J_1(V)$, let $B(x, r)$ be the ball in $J_1(V)$ of radius $r = \operatorname{Inj}(J_1(V), |\cdot|_e)$ centered at x, and $B_T(x, r)$ the ball of radius r centered at the origin of the tangent space $T_x J_1(V)$. As in the proof of Theorem A, we fix $p \in S^n$ and define a map $h : B(x, r) \to S^n$ through the composition

$$B(x, r) \xrightarrow{\exp_x^{-1}} B_T(x, r) \xrightarrow{\varphi_1} B_T(x, \pi) \xrightarrow{\varphi_2} B_T(p, \pi) \xrightarrow{\varphi_3} S^n,$$

where the notation is just as in diagram (1) (page 450). This map h sends $\partial B(x, r)$ to p'. Since $(J_1(V), |\cdot|_e)$ is flat, we have $\operatorname{dil} \exp_x^{-1} \le 1$. Extend h to $J_1(V) \to S^n$

by setting $h(J_1(V)\backslash B(x,r)) = p'$. As in the proof of Theorem A, h has degree 1 and its dilatation satisfies $\operatorname{dil} h \le \operatorname{dil}\varphi_3 \operatorname{dil}\varphi_2 \operatorname{dil}\varphi_1 \operatorname{dil}\exp_x^{-1} \le \pi/r = \pi/\operatorname{Inj} J_1(V)$.

Next, a closed geodesic C in the flat torus $(J_1(V), |\cdot|_e)$ is the projection of a segment $[x, x+\gamma]$ joining x and $x+\gamma$ in the Euclidean space $(H_1(V,\mathbb{R}), |\cdot|_e)$, for $\gamma \in H_1(V,\mathbb{Z})_{\mathbb{R}}$. Moreover, C is not homologous to zero (hence not homotopic to zero since $\pi_1(J_1(V))$ is abelian) if and only if $\gamma \neq 0$. The length of C equals the displacement $d_{|\cdot|_e}(x, x+\gamma)$, which does not depend on x; see [Gromov 1996, 1.A]. Hence

$$\begin{aligned}\operatorname{sys}_1(J_1(V), |\cdot|_e) &= \inf\{|\gamma|_e : \gamma \in H_1(V,\mathbb{Z}) \setminus 0\}\\ &\ge \inf\{\|\gamma\|_{\mathrm{st}} : \gamma \in H_1(V,\mathbb{Z}) \setminus 0\} = \operatorname{stsys}_1 V.\end{aligned}$$

Thus $\operatorname{Inj}(J_1(V), |\cdot|_e) = \frac{1}{2}\operatorname{sys}_1(J_1(V), |\cdot|_e) \ge \frac{1}{2}\operatorname{stsys}_1 V$, and $\operatorname{dil} h \le 2\pi/(\operatorname{stsys}_1 V)$.

Composing h with the map ψ of Proposition 3.3 yields a Lipschitz map $f = h\circ\psi : (V, d_g) \to (S^n, \mathrm{can})$ satisfying $\deg f = \deg h \deg\psi = \deg \mathscr{A}_V \neq 0$ and

$$\operatorname{dil} f \le \operatorname{dil} h \operatorname{dil}\psi \le \frac{2\pi\sqrt{n}}{\operatorname{stsys}_1 V)}.$$

This completes the proof of the theorem. □

4. Filling volume and its mapping property

We recall from [Gromov 1983] the definition of filling volume. Let (X, d) be a metric space and $\sigma : \triangle^{n+1} \to X$ a singular simplex. Define

$$\operatorname{Vol}(\sigma) = \inf_g\{\operatorname{Vol}_g(\triangle^{n+1}\},$$

where the infimum is taken over all Riemannian metrics g on $\triangle^{n+1}$ such that $d_X(\sigma(x), \sigma(y)) \le d_g(x, y)$ for all $x, y \in \triangle^{n+1}$. For a singular chain $c = \sum_i r_i\sigma_i$, we can define

$$\operatorname{Vol}(c) = \sum_i |r_i| \operatorname{Vol}(\sigma_i),$$

where the coefficients r_i may be real numbers, integers, or integers mod 2. When (X, g) is a Riemannian manifold, $\operatorname{Vol}(c)$, with respect to d_g, is just the usual Riemannian volume of a singular chain.

Let z be an n-dimensional singular $\mathbb{G}$-cycle in X, and $\mathbb{G} = \mathbb{R}$, $\mathbb{Z}$ or $\mathbb{Z}_2$. We define

$$\operatorname{FillVol}(z \hookrightarrow X; \mathbb{G}) = \inf\{\operatorname{Vol}(c) : c \text{ are } (n+1)\text{-chains in } X, \partial c = z\},$$

where the coefficients of the chains c lie in $\mathbb{G}$.

Let V be a compact submanifold in X and $[V]$ the fundamental class of V (if V is not oriented, $[V]$ denotes the fundamental $\mathbb{Z}_2$-class in the group $H_{\dim(V)}(V, \mathbb{Z}_2) \simeq \mathbb{Z}_2$). We define the filling volume relative to the imbedding $i : V \hookrightarrow X$ by

$$\mathrm{FillVol}(V \hookrightarrow X; \mathbb{G}) = \inf\{\mathrm{FillVol}(z) : z \text{ represents } i_*[V]\}.$$

Let V be a compact manifold equipped with a metric d and the isometric imbedding $V \hookrightarrow L^\infty(V)$ is as Section 1. We define the absolute filling volume of V as

$$\mathrm{FillVol}(V; \mathbb{G}) =: \mathrm{FillVol}(V \hookrightarrow L^\infty(V); \mathbb{G}).$$

To simplify the notation, we will generally write $\mathrm{FillVol}(V)$ as $\mathrm{FillVol}(V, \mathbb{G})$, understanding the coefficient group $\mathbb{G}$ to be $\mathbb{Z}$ or $\mathbb{Z}_2$, depending on whether or not V is oriented.

The following proposition shows a mapping property of filling volume

Proposition 4.1. *Let $f : V \to W$ be a map between compact, connected, n-dimensional manifolds V, W with metrics d_V, d_W respectively.*

(i) *If V, W are oriented and $|\deg f| = 1$, then*

$$\mathrm{FillVol}(V; \mathbb{Z}) \geq \frac{1}{(\mathrm{dil}\, f)^{n+1}} \mathrm{FillVol}(W; \mathbb{Z}).$$

(ii) *If V, W are oriented and f is of nonzero degree, then*

$$\mathrm{FillVol}(V; \mathbb{R}) \geq \frac{|\deg f|}{(\mathrm{dil}\, f)^{n+1}} \mathrm{FillVol}(W; \mathbb{R}).$$

(iii) *If V, W are not oriented and $\deg f \neq 0 \, mod 2$, then*

$$\mathrm{FillVol}(V; \mathbb{Z}_2) \geq \frac{1}{(\mathrm{dil}\, f)^{n+1}} \mathrm{FillVol}(W; \mathbb{Z}_2).$$

Lipschitz Extension Lemma 4.2 [Gromov 1983, p. 8]. *Let V be a compact submanifold of a metric space X, and let W be a metric space. Every Lipschitz map $\varphi : V \to L^\infty(W)$ has a Lipschitz extension $\tilde{\varphi} : X \to L^\infty(W)$ with $\mathrm{dil}\, \tilde{\varphi} = \mathrm{dil}\, \varphi$.*

For a detailed proof, see [Liu 2005].

Proof of Proposition 4.1. We may suppose that f is a Lipschitz map (if not, the desired inequalities are obvious). By the Lipschitz Extension Lemma 4.2, $f : V \to W$ has a Lipschitz extension $\tilde{f} : L^\infty(V) \to L^\infty(W)$ with $\mathrm{dil}\, \tilde{f} = \mathrm{dil}\, f$. Let the isometric imbedding $i_V : V \hookrightarrow L^\infty(V)$ and $i_W : W \hookrightarrow L^\infty(W)$ be as above. Let $c = \sum r_i \sigma_i$ be any $(n+1)$-chain in $L^\infty(V)$ with $[\partial c] = i_*[V]$, where the $\sigma_i : \triangle^{n+1} \to L^\infty(V)$ are singular simplexes. Then $\tilde{f}(c) = \sum r_i \tilde{f}(\sigma_i)$ is an $(n+1)$-chain in $L^\infty(W)$. Let g be any Riemannian metric on $\triangle^{n+1}$ with $d_g(x, y) \geq d_{L^\infty(V)}(\sigma_i(x), \sigma_i(y))$ for any $x, y \in \triangle^{n+1}$. Set $\alpha := \mathrm{dil}\, f$ and consider the new Riemannian metric $\alpha^2 g$ on $\triangle^{n+1}$. For any $x, y \in \triangle^{n+1}$ we have

$$d_{L^\infty(W)}(\tilde{f}\sigma_i(x), \tilde{f}\sigma_i(y)) \leq \alpha d_{L^\infty(V)}(\sigma_i(x), \sigma_i(y)) \leq \alpha d_g(x, y) = d_{\alpha^2 g}(x, y).$$

By the definition of the volume of a singular simplex,

$$\mathrm{Vol}(\tilde{f}\sigma_i) \le \mathrm{Vol}(\triangle^{n+1}, \alpha^2 g) = \alpha^{n+1}\,\mathrm{Vol}(\triangle^{n+1}, g).$$

In view of the freedom of g, we have proved that $\mathrm{Vol}(\tilde{f}\sigma_i) \le \alpha^{n+1}\,\mathrm{Vol}(\sigma_i)$; hence $\mathrm{Vol}(\tilde{f}(c)) \le \alpha^{n+1}\,\mathrm{Vol}(c)$.

In cases (i) and (iii), $\partial\tilde{f}(c)$ represents $\pm(i_W)_*[W]$. Indeed,

$$[\partial\tilde{f}(c)] = \tilde{f}_*[\partial c] = \tilde{f}_* \circ (i_V)_*[V] = (i_W)_* \circ f_*[V] = \deg f\ (i_W)_*[W].$$

In case (ii), the real singular chain $|\deg f|^{-1}\tilde{f}(c)$ represents $\pm(i_W)_*[W]$ and

$$\mathrm{Vol}\Big(\frac{1}{|\deg f|}\tilde{f}(c)\Big) = \frac{1}{|\deg f|}\,\mathrm{Vol}(\tilde{f}(c)) \le \frac{1}{|\deg f|}\alpha^{n+1}\,\mathrm{Vol}(c).$$

This completes the proof of the proposition. □

Proposition 4.3. *Let V^n be a compact manifold with a metric d and let X be a metric space. For any isometric imbedding $j : V \hookrightarrow X$, we have*

$$\mathrm{FillVol}(V \hookrightarrow X) \ge \mathrm{FillVol}(V \hookrightarrow L^\infty(V)) = \mathrm{FillVol}(V).$$

Proof. Again using Lipschitz Extension Lemma 4.2, $i : V \hookrightarrow L^\infty(V)$ has a Lipschitz extension $\varphi : X \to L^\infty(V)$ with $\mathrm{dil}\,\varphi \le 1$. Let z be any n-cycle in X which represent $j_*[V]$ and c be any $(n+1) - chain$ in X with $\partial c = z$. Then $[\partial\varphi(c)] = \varphi_*[\partial c] = \varphi_*[z] = \varphi_* \circ j_*[V] = i_*[V]$, namely $\partial\varphi(c)$ represent $i_*[V]$. From the argument of Proposition 4.1, we know that $\mathrm{Vol}(\varphi(c)) \le \mathrm{Vol}(c)$. The arbitrariness of z and c implies that $\mathrm{FillVol}(V \hookrightarrow L^\infty(V)) \le \mathrm{FillVol}(V \hookrightarrow X)$. □

Let W be a compact manifold with boundary ∂W, with a Riemannian metric g. The *chordal metric* on ∂W is the (non-Riemannian) metric $d_g|_{\partial W}$ on ∂W defined by g-shortest paths in W. (See [Croke and Katz 2003] for details.)

Corollary 4.4 [Gromov 1983, p. 12]. $\mathrm{FillVol}(\partial W, d_g|_{\partial W}) \le \mathrm{Vol}(W, g)$.

Proof. For the isometric imbedding $(\partial W, d_g|_{\partial W}) \hookrightarrow (W, d_g)$, applying Proposition 4.3 we have

$$\mathrm{Vol}(W, g) = \mathrm{FillVol}(\partial W \hookrightarrow W) \ge \mathrm{FillVol}(\partial W, d_g|_{\partial W}).$$ □

Thanks to Proposition 4.1 and Corollary 4.4, we obtain Corollaries 1.4′ and 1.6′ from Corollaries 1.4 and 1.6.

5. Filling volume estimates by cube Besicovitch inequality

In this section we adopt the notation $I = [-1, 1]$. Then I^n is a topological cube in $\mathbb{R}^n$, with boundary $\partial I^n = \bigcup_{i=1}^n (F_i^{-1} \cup F_i^1)$, where F_i^{-1} is the set of points in I^n whose i-th coordinate equals -1, and likewise for F_i^1. The sets F_i^{-1} and F_i^1 are called $(n-1)$-faces of the cube; the notion of opposite faces is obvious.

Theorem 5.1 [Gromov 1983, pp. 85–86]. *Let V be a compact, oriented n-manifold with a metric d and let $f : V \to \partial I^{n+1} (\simeq S^n)$ be a continuous map with nonzero degree. Then*

$$\mathrm{FillVol}(V) > |\deg f| \prod_{i=1}^{n+1} d\big(f^{-1}(F_i^{-1}), f^{-1}(F_i^1)\big).$$

We regard this inequality still as a Besicovitch inequality, and we will use it in deducing Theorem C.

Proof of Theorem C. We establish a homeomorphism $\varphi : S^n \to \partial I^{n+1}$ by central projection:

$$\varphi(u) = \frac{u}{\|u\|_\infty} \quad \text{for } u \in S^n, \qquad \varphi^{-1}(x) = \frac{x}{|x|_E} \quad \text{for } x \in \partial I^{n+1},$$

where $\|\cdot\|_\infty$ is the l^∞-norm (whose unit sphere is ∂I^{n+1}) and $|\cdot|_E$ is the Euclidean norm. Next we determine $d_S\big(\varphi^{-1}(F_1^{-1}), \varphi^{-1}(F_1^1)\big)$, where d_S is the spherical distance on S^n. Take $x = (1, x_2, \ldots, x_{n+1}) \in F_1^1$, $y = (-1, y_2, \ldots, y_{n+1}) \in F_1^{-1}$, and assume without loss of generality that $|x|_E \geq |y|_E$. Then

$$\left|\frac{x}{|x|_E} - \frac{y}{|y|_E}\right|_E = \frac{1}{|x|_E}\left|x - \frac{|x|_E}{|y|_E} y\right|_E \geq \frac{2}{|x|_E} \geq \frac{2}{\max_{x \in \partial I^{n+1}} |x|_E} = \frac{2}{\sqrt{n+1}}.$$

For any

$$u = \frac{x}{|x|_E} \in \varphi^{-1}(F_1^1) \subset S^n \quad \text{and} \quad v = \frac{y}{|y|_E} \in \varphi^{-1}(F_1^{-1}) \subset S^n,$$

the cosine theorem implies

$$|u - v|_E^2 = 2 - 2\cos d_S(u, v).$$

Therefore

$$\begin{aligned} d_S(u, v) &= \arccos(1 - \tfrac{1}{2}|u - v|_E^2) \\ &= \arccos\left(1 - \frac{1}{2}\left|\frac{x}{|x|_E} - \frac{y}{|y|_E}\right|_E^2\right) \geq \arccos\Big(1 - \frac{2}{n+1}\Big) = \arccos\frac{n-1}{n+1}. \end{aligned}$$

Taking

$$x_0 = (1, 1, \ldots, 1) \in F_1^1, \qquad y_0 = (-1, 1, \ldots, 1) \in F_1^{-1},$$

$$u_0 = \frac{x_0}{|x_0|_E} = \frac{x_0}{\sqrt{n+1}} \in f^{-1}(F_1^1), \qquad v_0 = \frac{y_0}{|y_0|_E} = \frac{y_0}{\sqrt{n+1}} \in f^{-1}(F_1^{-1}),$$

we have

$$d_S(u_0, v_0) = \arccos\left(1 - \frac{1}{2}\left|\frac{x_0}{|x_0|_E} - \frac{y_0}{|y_0|_E}\right|_E^2\right) = \arccos\frac{n-1}{n+1}.$$

Hence $d_S\big(\varphi^{-1}(F_1^1), \varphi^{-1}(F_1^{-1})\big) = \arccos\dfrac{n-1}{n+1}$, and more generally

$$d_S\big(\varphi^{-1}(F_i^1), \varphi^{-1}(F_i^{-1})\big) = \arccos\frac{n-1}{n+1} \quad \text{for } i = 1, \dots, n+1.$$

Next, we estimate $d_S\big((\varphi\circ f)^{-1}(F_i^1), (\varphi\circ f)^{-1}(F_i^{-1})\big)$ for the composite map $\varphi\circ f : V \to \partial I^{n+1}$. For any $x \in (\varphi\circ f)^{-1}(F_i^1)$ and $y \in (\varphi\circ f)^{-1}(F_i^{-1})$, we have $f(x) \in \varphi^{-1}(F_i^1)$, $f(y) \in \varphi^{-1}(F_i^{-1})$. Therefore

$$\begin{aligned} d_V(x, y) &\geq \frac{1}{\operatorname{dil} f} d_S(f(x), f(y)) \\ &\geq \frac{1}{\operatorname{dil} f} d_S(\varphi^{-1}(F_i^1), \varphi^{-1}(F_i^{-1})) = \frac{1}{\operatorname{dil} f}\arccos\frac{n-1}{n+1}. \end{aligned}$$

Since $\deg(\varphi\circ f) = \deg f$, applying the Besicovitch inequality (Theorem 5.1) to $\varphi\circ f$, we get

$$\operatorname{FillVol}(V) \geq \frac{|\deg f|}{(\operatorname{dil} f)^{n+1}} \cdot \Big(\arccos\frac{n-1}{n+1}\Big)^{n+1}. \qquad \square$$

Remark 5.2. As is well-known, it is difficult to compute Gromov invariants. There is still not a single Riemannian manifold whose filling volume is known. Gromov's filling volume conjecture [1983, p. 13], which is still open in all dimensions, says that $\operatorname{FillVol}(S^n, \text{can}) = \frac{1}{2}\operatorname{Vol}(S^{n+1}, \text{can})$. The case $n = 1$ can be broken into separate problems depending on the genus of the filling [Gromov 1983, pp. 59–60; Bangert et al. 2005; Croke and Katz 2003; Katz and Lescop 2005]. Taking $W =$the canonical positive hemisphere (S_+^{n+1}, g), obviously, $\partial S_+^{n+1} = S^n$ and $d_g|_{S^n} =$the canonical distance of S^n. From Corollary 4.4, we have

$$\operatorname{FillVol}(S^n) \leq \operatorname{Vol}(S_+^{n+1}) = \tfrac{1}{2}\operatorname{Vol}(S^{n+1}).$$

On the other hand, as an intermediate result of the argument of Theorem C, we have

$$\operatorname{FillVol}(S^n) > \Big(\arccos\frac{n-1}{n+1}\Big)^{n+1}.$$

Unfortunately, this rough lower bound is far from the exact value conjectured by Gromov. It is also inferior to the easy estimate obtained as follows. Since the Euclidean unit ball B^{n+1} is flat and simply connected, we know from [Gromov 1983, 2.1] that

$$\operatorname{FillVol}(S^n, \text{Euclid}) = \operatorname{FillVol}(\partial B^{n+1}, \text{Euclid}) = \operatorname{Vol}(B^{n+1}) = \frac{\pi^{(n+1)/2}}{\Gamma\big(\frac{n+1}{2}+1\big)}.$$

The canonical Riemannian metric on S^n (of diameter π) dominates the Euclidean metric of diameter 2, so $\operatorname{FillVol}(S^n, \text{can}) \geq \operatorname{FillVol}(S^n, \text{Euclid})$ (Proposition 4.1 is a generalization of this fact).

Acknowledgements

This article is part of a Ph.D. thesis written under the direction of Fuquan Fang at the Nankai Institute of Mathematics. The author thanks Professor Fang for his guidance. Special thanks go to Mikhail Katz for constructive criticism of the first version of the manuscript, which prompted an overall improvement in the work.

References

[Bangert and Katz 2004] V. Bangert and M. Katz, "An optimal Loewner-type systolic inequality and harmonic one-forms of constant norm", *Comm. Anal. Geom.* **12** (2004), 703–732. MR 2005k:53035 Zbl 1068.53027

[Bangert et al. 2005] V. Bangert, C. Croke, S. Ivanov, and M. Katz, "Filling area conjecture and ovalless real hyperelliptic surfaces", *Geom. Funct. Anal.* **15**:3 (2005), 577–597. MR 2007b:53068 Zbl 1082.53033

[Burago and Ivanov 1995] D. Burago and S. Ivanov, "On asymptotic volume of tori", *Geom. Funct. Anal.* **5**:5 (1995), 800–808. MR 96h:53041 Zbl 0846.53043

[Chavel 1993] I. Chavel, *Riemannian geometry—a modern introduction*, Cambridge Tracts in Mathematics **108**, Cambridge University Press, Cambridge, 1993. MR 95j:53001 Zbl 0810.53001

[Cheeger and Ebin 1975] J. Cheeger and D. G. Ebin, *Comparison theorems in Riemannian geometry*, Mathematical Library **9**, North-Holland, Amsterdam, 1975. MR 56 #16538 Zbl 0309.53035

[Croke and Katz 2003] C. B. Croke and M. Katz, "Universal volume bounds in Riemannian manifolds", pp. 109–137 in *Surveys in differential geometry* (Boston, 2002), vol. 8, International Press, Somerville, MA, 2003. MR 2005d:53061 Zbl 1051.53026

[Gromov 1978] M. Gromov, "Homotopical effects of dilatation", *J. Differential Geom.* **13**:3 (1978), 303–310. MR 82d:58017 Zbl 0427.58010

[Gromov 1983] M. Gromov, "Filling Riemannian manifolds", *J. Differential Geom.* **18**:1 (1983), 1–147. MR 85h:53029 Zbl 0515.53037

[Gromov 1996] M. Gromov, "Systoles and intersystolic inequalities", pp. 291–362 in *Actes de la Table Ronde de Géométrie Différentielle* (Luminy, 1992), Sémin. Congr. **1**, Soc. Math. France, Paris, 1996. MR 99a:53051 Zbl 0877.53002

[Gromov 1999a] M. Gromov, "Quantitative homotopy theory", pp. 45–49 in *Prospects in mathematics* (Princeton, 1996), edited by H. Rossi, Amer. Math. Soc., Providence, 1999. MR 2000a:57056 Zbl 0927.57022

[Gromov 1999b] M. Gromov, *Metric structures for Riemannian and non-Riemannian spaces*, Prog. in Math. **152**, Birkhäuser, Boston, 1999. MR 2000d:53065 Zbl 0953.53002

[Grove and Markvorsen 1995] K. Grove and S. Markvorsen, "New extremal problems for the Riemannian recognition program via Alexandrov geometry", *J. Amer. Math. Soc.* **8**:1 (1995), 1–28. MR 95j:53066 Zbl 0829.53033

[Guth 2005] L. Guth, "Lipschitz maps from surfaces", *Geom. Funct. Anal.* **15**:5 (2005), 1052–1099. MR 2007e:53029

[Ivanov and Katz 2004] S. V. Ivanov and M. G. Katz, "Generalized degree and optimal Loewner-type inequalities", *Israel J. Math.* **141** (2004), 221–233. MR 2005k:53047 Zbl 1067.53031

[Katz and Lescop 2005] M. Katz and C. Lescop, "Filling area conjecture, optimal systolic inequalities, and the fiber class in abelian covers", pp. 181–200 in *Geometry, spectral theory, groups, and*

dynamics: in memory of Robert Brooks (Haifa, 2004), edited by M. Entov et al., Contemporary Math. **387**, Amer. Math. Soc., Providence, RI, 2005. MR 2006h:53030 Zbl 1088.53029

[Lichnerowicz 1969] A. Lichnerowicz, "Applications harmoniques dans un tore", *C. R. Acad. Sci. Paris Sér. A* **269** (1969), 912–916. MR 40 #6469 Zbl 0184.25002

[Liu 2005] L. Liu, "The mapping properties of filling radius and packing radius and their applications", *Differential Geom. Appl.* **22**:1 (2005), 69–79. MR 2005h:53062 Zbl 1073.53056

[Nabutovsky and Rotman 2003] A. Nabutovsky and R. Rotman, "Upper bounds on the length of a shortest closed geodesic and quantitative Hurewicz theorem", *J. Eur. Math. Soc.* **5**:3 (2003), 203–244. MR 2004f:53043 Zbl 1039.53042

[Peng and Tang 2002] C. Peng and Z. Tang, "Dilatation of maps between spheres", *Pacific J. Math.* **204**:1 (2002), 209–222. MR 2003c:53092 Zbl 1050.58012

Received December 20, 2005.

LUOFEI LIU
DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
JISHOU UNIVERSITY
JISHOU, HUNAN, 416000
CHINA

luofei_liu@yahoo.com.cn

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.461

ON THE VARIATION OF A SERIES ON TEICHMÜLLER SPACE

GREG MCSHANE

Using the Kerckhoff–Wolpert formula for the variation of the length of a geodesic along a Fenchel–Nielsen twist, we prove that a certain series defines a constant function.

1. Introduction: two questions

In [McShane 1998] we showed that

$$\sum \frac{2}{1+\exp l_\gamma} = 1,$$

where l_γ is the length of the closed geodesic freely homotopic to γ and the sum extends over all simple closed curves γ on a hyperbolic once punctured torus. The proof we gave is based on the geometry of (simple) geodesics on a complete hyperbolic surface with at least one puncture.

Here we present a proof of an analogous identity that emphasises the rôle of the different elements of the modular group and the infinitesimal geometry of the Teichmüller space, which appears via a variational formula due to Steve Kerckhoff [1983] and Scott Wolpert [1983b]. Since the difficultly is not greatly increased, we work with surfaces with geodesic boundary as in [Mirzakhani 2003].

We begin with two questions of Troels Jørgensen.

Question 1. *Can the identity above be proved using the* Markoff cubic

$$a^2+b^2+c^2-abc=0,\ \ a,b,c>2\,?$$

Bowditch [1996] answered this by giving a proof that uses a summation argument over the edges of the tree $\mathbf{T}$ of solutions to this equation.

Question 2. *Can the same identity be proved using the* Kerckhoff–Wolpert formula *for variation of length*?

MSC2000: 51M25, 53C22.
Keywords: Riemann surfaces, moduli, geodesic lengths, length spectrum, Fenchel–Nielsen twist.

We recall what the formula says. If μ_1, μ_2 are closed simple geodesics on a hyperbolic surface, then

$$dl_{\mu_1} t(\mu_2) = \sum_{z \in \mu_1 \cap \mu_2} \cos(\theta_z), \tag{1}$$

where $t(\mu_2)$ is the Fenchel–Nielsen vector field associated to μ_2 and the sum is over all the intersections between the geodesics. This formula was generalised by Goldman [1984; 1986] in terms of the natural Poisson bracket on the representation space of a surface group into a semisimple Lie group.

It is Jørgensen's second question that we address here. Our starting point is the observation that by clever "accounting", Bowditch avoids considering the divergent series

$$\sum_{\{a,b,c\} \in \mathbf{T}} \left(\frac{a}{bc} + \frac{b}{ca} + \frac{c}{ab} \right).$$

The series is divergent since $\mathbf{T}$ is infinite and, since a, b, c is a solution of the cubic, the value of each term is 1.

Here we consider another divergent series obtained by summing over the edges of $\mathbf{T}$. We show that, after a suitable regularization, this yields another series constant on Teichmüller space. The ingredients are

- a (formal) argument, using an involution of the surface, showing that the series defines a constant function;
- a geometric recipe, using Dehn twists, for finding an explicit expression for the terms in the series;
- a method for finding the value of the series.

This allows us to give a proof of:

Theorem 1. *For a one-holed torus M,*

$$\sum_{\gamma} \arctan \frac{\cosh(l_\delta/4)}{\sinh(l_\gamma/2)} = \frac{3\pi}{2},$$

where the sum extends over all simple closed geodesics γ on M and l_δ is the length of the boundary geodesic δ.

Relation with Mirzakhani's identities. Mirzakhani [2003] has obtained identities for hyperbolic surfaces with nonempty geodesic boundary by adapting the methods of [McShane 1998]. For the one-holed torus she obtains

$$\ell_\delta = \sum_{\gamma} 2 \log \frac{e^{\ell_\delta/2} + e^{\ell_\gamma}}{e^{-\ell_\delta/2} + e^{\ell_\gamma}},$$

where the sum is over all essential simple curves γ. When one quotients the (holed) torus by the elliptic involution J one obtains an orbifold M/J with a single geodesic boundary component of length $\ell_\delta/2$ and three cone singularities of angle π. Philosophically M/J surface is a "hyperbolic four-holed sphere with boundary geodesic of lengths $\ell_\delta/2, i\pi, i\pi, i\pi$", this assertion being interpreted as follows. The fundamental group of the four-holed sphere is generated by four simple loops $\alpha_1, \alpha_2, \alpha_3, \delta$ which meet only at the base point and each of which is homotopic to a different boundary geodesic. The loops $\alpha_1, \alpha_2, \alpha_3, \delta$ are said to be *peripheral*; after choosing orientations appropriately, we have $\pi_1 = \langle \alpha_1, \alpha_2, \alpha_3, \delta : \delta = \alpha_1\alpha_2\alpha_3 \rangle$.

- In the $\mathrm{SL}(2, \mathbb{R})$ character variety of π_1, there is a representation ρ such that the quotient is isometric to M/J.
- The monodromy around the loop δ is hyperbolic with translation length $\ell_\delta/2$.
- The monodromies of the remaining three peripheral curves $\alpha_1, \alpha_2, \alpha_3$ are elliptic of order 2.

For a hyperbolic four-holed sphere with three geodesic boundary components α_1, α_2, and α_3 such that $\ell_{\alpha_1} = \ell_{\alpha_2} = \ell_{\alpha_3} = L > 0$, Mirzakhani's identities give

$$L = \sum_{\gamma \in A_i} 2 \log \frac{e^{L/2} + e^{(\ell_\gamma + L)/2}}{e^{-L/2} + e^{(\ell_\gamma + L)/2}}, \tag{2}$$

where A_i the set of simple curves which bound a pair of pants with δ and α_i. Every nonperipheral simple curve belongs to one of the A_i so summing gives

$$3L = \sum_{\gamma} 2 \log \frac{e^{L/2} + e^{(\ell_\gamma + L)/2}}{e^{-L/2} + e^{(\ell_\gamma + L)/2}}, \tag{3}$$

where now the sum is over *all* simple essential γ. Both the right and left sides of (3) are restrictions to the $\mathrm{SL}(2, \mathbb{R})$ character variety of complex analytic functions defined on a neighborhood $\mathrm{SL}(2, \mathbb{C})$ character variety. By analytic continuation one expects this to be true at $L = \pi$. After some algebra and noting that the set of nonperipheral simple curves on the four-holed sphere and those on the one-holed torus are in one-to-one correspondence we obtain the identity in Theorem 1. We note that Ser Tan seems to have known of the existence of a version of (3) for some time and has given a different treatment based on the ideas of [Zhang et al. 2005].

Sketch of proof of Theorem 1. Let $\mathcal{P}$ be the left-hand side of the identity in the theorem. We show that the variation $d\mathcal{P}$ of the series vanishes identically on the Teichmüller space. To do this we find a series $d\mathcal{Q}$ (see Section 2) which converges absolutely to $d\mathcal{P}$ and a rearrangement showing that the former is identically 0; since Teichmüller space is connected the series $\mathcal{P}$ is constant. Our series $d\mathcal{Q}$ is, at least formally, the derivative of a divergent series $\mathcal{Q}$ which has a nice geometric

definition. To define $\mathcal{Q}$, we introduce the following notation which we shall use throughout the article: Let α, β be a pair of oriented essential simple closed curves meeting in a single point and let $\alpha \vee \beta$ denote the signed angle between the closed simple geodesics in the free homotopy classes determined by α and β (see Section 4 for discussion). The series $\mathcal{Q}$ is, roughly speaking, the sum over all such $\alpha \vee \beta$.

In dealing with $d\mathcal{Q}$ ($d\mathcal{P}$) we adopt the a point of view similar to that of [Kerckhoff 1983]; we work with the Fenchel–Nielsen geometry of the cotangent bundle (see also [Wolpert 1982; 1983a]). We evaluate the pairing of $d\mathcal{Q}$ with the Fenchel–Nielsen vector field $t(\mu)$ associated to a simple closed geodesic μ on $\Sigma_{g,n}$. By a result of Wolpert [1979; 1982] there are finitely many simple closed geodesics μ_i such that the associated Fenchel–Nielsen vector fields t_{μ_i} generate the tangent space at every point in the Teichmüller space of a surface of finite type. A 1-form vanishes if and only if its pairing with these fields vanishes. One concludes immediately, since the function $x \mapsto l_\delta(x)$, $\mathcal{T}_{1,1} \to \mathbb{R}^+$ has connected level sets, that the value of the series depends only on the length of l_δ. It is actually more convenient to think of the torus T as being an (embedded) convex subsurface of a closed surface $\Sigma_{g,n}$. The inclusion $i : T \to M$ induces a faithful representation of the mapping class group of T in the mapping class group of M. The advantage of this approach is that one can twist along closed geodesics in M that are not contained in $i(T)$ and as such vary the length of the boundary curve l_δ.

The main technical result, on which our proof hinges, is this:

Theorem 2. *Let $T \subset \Sigma_{g,n}$ be an embedded convex subsurface and let $\mathcal{MCG}^*(T)$ denote the image of the mapping class group of T in the mapping class group of $\Sigma_{g,n}$. Let $\mu \subset \Sigma_{g,n}$ be a simple closed geodesic with $t(\mu)$ the associated Fenchel–Nielsen vector field then the series*

$$d\mathcal{Q}.t(\mu) = \sum_{[g]\in\mathcal{MCG}^*(T)} d([g].(\alpha \vee \beta)).t(\mu)$$

converges absolutely and its sum vanishes.

We use the following corollary of his formula, also due to Wolpert [1983a], to prove absolute convergence (Section 7)

$$|dl_{\mu_1}.t(\mu_2)| \leq \sum_{x\in\mu_1\cap\mu_2} 1 = \operatorname{card}(\mu_1 \cap \mu_2) := i(\mu_1, \mu_2).$$

where $i(\mu_1, \mu_2)$ is the *geometric intersection number*.

Value of the series. The vanishing of $d\mathcal{P}$ is one of two parts of the proof of the Theorem 1 the other being the determination of the value of the series at some point. This value is shown to be $3\pi/2$ in [McShane 2004] for any point in the Teichmüller space of the punctured torus, that is when $\ell_\delta = 0$. It is worth note that

the usual heuristic of setting $\ell_\gamma = 0$, the result of which is $\pi/2$, gives an incorrect value.

Remark. The original motivation for this work was to extend the formula to spaces of representations of surface groups in Lie groups other than $\mathrm{SL}(2,\mathbb{R})$ for which Goldman has established analogues of the above variational formula in [Goldman 1984]. For example it is not hard to show that our method works in the case of $\mathrm{SL}(2,\mathbb{C})$, see [Series 2001] or [Goldman 2004] for a discussion of the formula in this case. Though our method should in principle give results in this direction in *certain special cases* (in Section 9 we explain why it appears to break down irretrievably in general), we note that identities have subsequently been established [Labourie and Mcshane 2006] using a different method. Bowditch's method has been extended to treat many other cases in [Zhang et al. 2004a; 2004b; 2005].

2. Another divergent series

Let M be a *one-holed torus*. The fundamental group of M is freely generated by two loops γ_1, γ_2 that meet in a single point and such that their commutator is a loop, δ, around the hole. Denote by $\mathcal{T}_1(l_\delta)$ the *Teichmüller space of* M. The mapping class group, $\mathcal{MCG}$, is defined to be the group of orientation preserving diffeomorphisms fixing the boundary pointwise up to isotopy $\pi_0(\mathrm{Diffeo}^+(M, \partial M))$.

Let M^* be a punctured torus. By the work of Nielsen and Mangel,

$$\pi_0(\mathrm{Diffeo}(M^*)) \cong \mathrm{Aut}(\pi_1)/\mathrm{Inn}(\pi_1) \cong \mathrm{GL}_2(\mathbb{Z}).$$

The mapping class group has index 2 in $\pi_0(\mathrm{Diffeo}(M^*))$ and so is isomorphic to $\mathrm{SL}_2(\mathbb{Z})$. Three (conjugacy classes of) elements of $\mathrm{SL}_2(\mathbb{Z})$ are of interest to us:

$$J = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Under the Nielsen–Mangel isomorphism, the elliptic involution goes to J and there is a (primitive) Dehn twist that goes to T. Let $\mathcal{MCG}^*$ denote the mapping class group modulo its center. From work of Ivanov, for instance, we know that the canonical action of $\mathcal{MCG}^*$ on Teichmüller space is effective and that, for a punctured torus, the center of the mapping class group is generated by the elliptic involution. The center of the mapping class group of the holed torus $\mathcal{MCG}$ is generated by a half Dehn twist round the boundary curve δ and the quotient is isomorphic to $\mathcal{MCG}^*$.

We write $[g] \in \mathcal{MCG}^*$ for the mapping class of an orientation preserving diffeomorphism g. Let $\alpha \in \pi_1$, $\alpha \neq 1$ be a closed loop and let $[\alpha]$ denote its free homotopy class, or equivalently its π_1 conjugacy class. If $g, g' \in [g]$, $[g] \in \mathcal{MCG}^*$ and $\alpha, \alpha' \in \pi_1$ are freely homotopic then $g(\alpha), g'(\alpha)$ are freely homotopic, that

is, $[g(\alpha)] = [g(\alpha')] = [g'(\alpha')]$; hence $\mathcal{MCG}^*$ admits an action on the set of free homotopy classes of closed curves,

$$([g], [\alpha]) \mapsto [g(\alpha)].$$

If M is negatively curved then for each nonperipheral loop $\alpha \neq 1$ there is a unique closed geodesic representing its free homotopy class, which we shall also denote by $[\alpha]$. By identifying a closed geodesic with its free homotopy class we may say that the mapping class group acts on the set of closed geodesics. The set of simple closed loops is invariant under diffeomorphism and so the set of simple closed geodesics geodesics, $\mathcal{G}_0$, is $\mathcal{MCG}^*$-invariant.

For $\alpha \neq 1$, let ℓ_α denote the length of the closed geodesic $[\alpha]$ with respect to the metric on M.

Let α, β be a pair of oriented simple closed loops on M meeting in exactly one point. The associated closed geodesics $[\alpha], [\beta]$ are simple and meet in a single point; let $\alpha \vee \beta$ denote the *signed angle* between them (see Section 4 for a precise definition). Now α, β freely generate $\pi_1(M)$; the only automorphism that simultaneously fixes them is the trivial automorphism. The elliptic involution reverses orientations for each of α, β and so the automorphism of $\pi_1(M)$ that it induces does not fix the corresponding pair of elements. Since $\mathcal{MCG}^*$ is a quotient of the group of automorphisms of $\pi_1(M)$, the stabiliser of $([\alpha], [\beta])$ in $\mathcal{MCG}^*$ is trivial. Moreover, since each $g \in \mathcal{MCG}^*$ is a homeomorphism, the pair of geodesics $[g(\alpha)], [g(\beta)]$ again meet in a single point, so

$$[g].(\alpha \vee \beta) := g(\alpha) \vee g(\beta)$$

is well defined and the sign is preserved since g preserves the orientation. After possibly exchanging α, β we can suppose that $[g].(\alpha \vee \beta) > 0$ for all $g \in \mathcal{MCG}^*$.

Consider the formal series

$$\mathfrak{Q} = \sum_{[g] \in \mathcal{MCG}^*} [g].(\alpha \vee \beta).$$

We determine its "sum" using a coset decomposition and hyperbolic geometry:

Step I. Let γ be a simple closed curve in a one-holed torus and let T_γ be the Dehn twist along γ. Observe that $\mathcal{MCG}^*$ acts transitively on $\mathcal{G}_0$ and the stabiliser of $\gamma \in \mathcal{G}_0$ is precisely $\langle T_\gamma \rangle$, so the map $\mathcal{MCG}^*/\langle T_\gamma \rangle \to \mathcal{G}_0$ defined by

$$[h] \mapsto [h(\gamma)]$$

is a bijection. We rewrite $\mathfrak{Q}$ as a sum running over the set of coset representatives $\mathcal{MCG}^*/\langle T_\gamma \rangle$ for γ satisfying

$$[\beta] = [T_\gamma(\alpha)] \tag{4}$$

(see Section 4). We obtain

$$\begin{aligned}\mathcal{Q} &= \sum_{[h]\in \mathcal{MCG}^*/\langle T_\gamma\rangle} \left(\sum_{n\in\mathbb{Z}} [h\circ T_\gamma^n].(\alpha \vee T_\gamma(\alpha)) \right) \\ &= \sum_{[h]\in \mathcal{MCG}^*/\langle T_\gamma\rangle} \left(\sum_{n\in\mathbb{Z}} [T_{h(\gamma)}^n].(h(\alpha) \vee T_{h(\gamma)}) \circ h(\alpha) \right) \\ &= \sum_{\gamma'\in\mathcal{G}_0} \left(\sum_{n\in\mathbb{Z}} T_{\gamma'}^n(\alpha') \vee T_{\gamma'}^{n+1}(\alpha') \right),\end{aligned}$$

where the outer sum is over all oriented simple closed geodesics $\mathcal{G}_0$ and α' is any simple closed geodesic that meets γ' exactly once. The passage from the first to the second line is justified by the equation

$$h\circ T_\gamma \circ h^{-1} = T_{h(\gamma)}.$$

Step II. We evaluate the inner sum over $\mathbb{Z}$ using the following result, the proof of which is postponed to the end of the next section. (For the notion of Weierstrass points, see page 470.)

Lemma 2.1. *For each γ', there exists a (Weierstrass) point $w \in M$, such that for all $i > 0$*

$$[T_{\gamma'}^{i-1}(\alpha')] \cap [T_{\gamma'}^{i}(\alpha')] = \{w\}.$$

Using the existence of such a w, by induction one obtains, for all $m < n$,

$$\sum_{m<i\le n} T_{\gamma'}^{i-1}(\alpha') \vee T_{\gamma'}^{i}(\alpha') = T_{\gamma'}^{m}(\alpha') \vee T_{\gamma'}^{n}(\alpha').$$

Moreover, there exist complete simple geodesics $\gamma^{-\infty}$, $\gamma^{+\infty}$ passing through w and spiraling to γ' such that $T_{\gamma'}^n(\alpha') \to \gamma^{\pm\infty}$ as $n \to \pm\infty$, where the convergence is uniform on compact sets of M. In particular,

$$\gamma^{-\infty} \vee \gamma^{+\infty} := \lim_{m,n\to\infty} T_{\gamma'}^n(\alpha') \vee T_{\gamma'}^n(\alpha').$$

is well defined. The inner sum telescopes over n and one obtains

$$\sum_{n\in\mathbb{Z}} T_{\gamma'}^n(\alpha') \vee T_{\gamma'}^{n+1}(\alpha') = \gamma^{-\infty} \vee \gamma^{+\infty}.$$

A calculation, which we carry out in Section 8, shows that

$$\gamma^{-\infty} \vee \gamma^{+\infty} = \pi - 2\arctan \frac{\cosh(l_\delta/4)}{\sinh(l_{\gamma'}/2)}. \tag{5}$$

Step III. We show that $\mathcal{Q}$ is constant using a different coset decomposition. There is an element $q \in \mathcal{MCG}^*$ of order 2, corresponding to the image of the matrix $Q \in \mathrm{SL}_2(\mathbb{Z})$ in $\mathrm{PSL}_2(\mathbb{Z})$, such that

$$q(\alpha) = \beta, \;\; q(\beta) = \alpha^{-1}.$$

For any $[g] \in \mathcal{MCG}^*$, the images $[g(\alpha)]$ and $[g(\alpha)^{-1}]$ determine the same undirected geodesic, so

$$g(\alpha) \vee g(\beta) + g(\beta) \vee g(\alpha^{-1}) = \pi.$$

Rewriting $\mathcal{Q}$ as a sum over cosets of $\mathcal{MCG}^*/\langle q \rangle$, one obtains

$$\mathcal{Q} = \sum_{g \in \mathcal{MCG}^*/\langle q \rangle} [g].(\alpha \vee \beta) + [g].(\beta \vee \alpha^{-1}) = \sum_{\mathcal{MCG}^*/\langle q \rangle} \pi \tag{6}$$

Although this last identity clearly implies that our series is divergent, it also suggests that the variation of $\mathcal{Q}$ vanishes when viewed as a 1-form on Teichmüller space. Under the hypothesis of absolute convergence, one expects that a suitable regularization defines a constant function. Here the regularization that works is

$$\mathcal{P} : x \mapsto \sum_{\gamma} 2 \arctan \frac{\cosh(l_\delta(x)/4)}{\sinh(l_\gamma(x)/2)}, \quad \mathcal{T}_1(l_\delta) \to \mathbb{R}.$$

From their expansions as infinite series, we have

$$d\mathcal{P} = d\mathcal{Q} = \sum_{[g]} d([g].(\alpha \vee \beta)).$$

Absolute convergence (Theorem 2) allows one to pair off terms as in (6) above:

$$d\left([g].(\alpha \vee \beta) + [g].(\beta \vee \alpha^{-1})\right).t(\mu) = 0,$$

so the sum for $d\mathcal{Q}.t(\mu)$ vanishes identically. Since Teichmüller space is connected, $\mathcal{P}$ is constant.

3. Markoff triples

For completeness we present a brief review of the theory of the representation variety of a free group on two generators, define a topological Markoff triple and describe the relationship with the elliptic involution. See [Goldman 2003] for background.

To each $x \in \mathcal{T}_1(l_\delta)$ one associates $\rho_x \in \mathrm{Hom}(\pi_1, \mathrm{SL}_2(\mathbb{R}))$, a discrete irreducible representation of π_1 such that the surface M with its metric is isometric to $\mathbb{H}^2/\rho_x$. Strictly speaking, a point x determines a representation into $\mathrm{PSL}_2(\mathbb{R})$, but since the surface is uniformized by a discrete torsion-free fuchsian group, we may lift into

$\mathrm{SL}_2(\mathbb{R})$. The advantage of working in the latter is that the trace provides a natural map.

For any $\rho_x \in [\rho_x]$ the length of the geodesic $[\gamma]$ on the surface determined by x satisfies

$$2\cosh\frac{l_\gamma(x)}{2} = |\operatorname{tr}\rho_x(\gamma)|,$$

for any $\gamma \in [\gamma]$; thus traces of matrices and length functions of geodesics are "interchangeable".

Fix a *topological Markoff triple*, that is, a triple of simple loops $\gamma_1, \gamma_2, \gamma_3$ such that

(1) the pairwise intersections $\{[\gamma_i] \cap [\gamma_j], i \neq j\}$ form a triple of distinct points,

(2) the loops γ_1, γ_2 freely generate $\pi_1(M, \gamma_1 \cap \gamma_2)$.

(3) $\gamma_3 = \gamma_1^{-1}\gamma_2$, and

(4) the commutator $\delta = \gamma_1\gamma_2\gamma_1^{-1}\gamma_2^{-1}$ represents a loop freely homotopic to the boundary curve of M, which we also denote by δ.

The *trace map* is defined to be

$$\rho \mapsto (\operatorname{tr}\rho(\gamma_1), \operatorname{tr}\rho(\gamma_2), \operatorname{tr}\rho(\gamma_3)),\ \ \mathrm{Hom}(\pi_1, \mathrm{SL}_2(\mathbb{R})) \to \mathbb{R}^3.$$

It is not difficult, by finding explicit matrices for $\rho(\gamma_1)$, $\rho(\gamma_2)$, to show this map is surjective. Observe that $\mathrm{SL}_2(\mathbb{R})$ acts by conjugation on

$$\mathrm{Hom}(\pi_1, \mathrm{SL}_2(\mathbb{R}))$$

and the trace map is clearly constant on the orbit $[\rho]$ of the representation ρ; we write $\mathrm{Hom}(\pi_1, \mathrm{SL}_2(\mathbb{R}))/\mathrm{SL}_2(\mathbb{R})$ for the space of orbits. Clearly the trace map induces a map, which we will still call the trace map, on $\mathrm{Hom}(\pi_1, \mathrm{SL}_2(\mathbb{R}))/\mathrm{SL}_2(\mathbb{R})$.

A diffeomorphism h of M acts on $\pi_1(M)$ on the left by the automorphism h_*, so h acts on the left on the representation variety by $(h, \rho) \mapsto \rho \circ h_*^{-1}$. If h is isotopic to the identity then $h_*^{-1} = i_\gamma$ for some $\gamma \in \pi_1$, i.e., h_*^{-1} is an inner automorphism, and so $[\rho \circ h_*^{-1}] = [\rho]$ for all ρ. From this we see that $[\rho \circ h_*^{-1}]$ depends only on the coset it determines modulo the group of isotopies, that is, only on its mapping class $[h]$. Thus one obtains a representation of $\mathcal{MCG}^*$ as a group acting on $\mathbb{R}^3$. A more explicit description of the action of $\mathcal{MCG}^*$ on $\mathbb{R}^3$ can be given by noting that $\mathcal{MCG}^*$ is isomorphic to $\mathrm{PSL}_2(\mathbb{Z})$ and that $\mathrm{PSL}_2(\mathbb{Z})$ splits as the free product $\mathbb{Z}/2\mathbb{Z} * \mathbb{Z}/3\mathbb{Z}$. The generator of $\mathbb{Z}/3\mathbb{Z}$ acts by cyclic permutation on the coordinates x_i whilst, using the trace relations in $\mathrm{SL}_2(\mathbb{R})$, the generator of $\mathbb{Z}/2\mathbb{Z}$ acts by a *quadratic reflection*

$$(x_1, x_2, x_3) \mapsto (x_2, x_1, x_1x_2 - x_3).$$

A calculation using the generators shows that the so-called Markoff cubic

$$\kappa(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 - x_1x_2x_3$$

is an invariant function for this action. Geometrically this is a consequence of the fact that every automorphism of the free group preserves the conjugacy class $[\delta] = [\gamma_1\gamma_2\gamma_1^{-1}\gamma_2^{-1}]$ and, by the trace relations, we have

$$\kappa(\operatorname{tr}\rho(\gamma_1), \operatorname{tr}\rho(\gamma_2), \operatorname{tr}\rho(\gamma_3)) = \operatorname{tr}\rho(\delta) + 2 = -2\cosh(l_\delta/2) + 2.$$

One identifies the component of the level set $\kappa = -2\cosh(l_\delta/2) + 2$ contained in $X = \{x_1, x_2, x_3 > 2\}$ with the Teichmüller space $\mathcal{T}_1(l_\delta)$ [Wolpert 1983a]. It follows that every function on $\mathcal{T}_1(l_\delta)$ can be expressed as a function of the lengths of the γ_i; we shall give an explicit expression for $\alpha \vee \beta$ in the next section.

Using the generators above, it is easy to check that the set $X = \{x_1, x_2, x_3 > 2\}$ is invariant under the $\mathrm{PSL}_2(\mathbb{Z})$-action and, moreover, the action is effective and properly discontinuous on X. The Bass–Serre tree of $\mathrm{PSL}_2(\mathbb{Z})$ is isomorphic to the infinite (regular) trivalent tree and so any orbit of the $\mathrm{PSL}_2(\mathbb{Z})$-action can be given the structure of a tree in the obvious way; this is the generalized (real) Markoff tree. The classical *Markoff tree* has vertices the integer solutions of

$$\kappa(x_1, x_2, x_3) = 0, x_1, x_2, x_3 > 2.$$

(The integer solutions to the Markoff equation above form a single $\mathrm{PSL}_2(\mathbb{Z})$ orbit).

Observations about Markoff triples. Given a configuration of loops $\gamma_i' \in [\gamma_i']$ satisfying the four conditions above, there is an obvious automorphism of π_1 that takes γ_i to γ_i', and this automorphism is f_* for some diffeomorphism f. So the Markoff tree enumerates the configurations that appear in the formal series $\mathfrak{D}$.

Also note that for simple loops $\gamma_1, \gamma_2, \gamma_3$, we have $\gamma_3 = \gamma_1^{-1}\gamma_2$ if and only if $T_{\gamma_3}(\gamma_1) = \gamma_2$, and this is none other than the condition (4).

Markoff triples and Weierstrass points. It is well known that M always admits an isometry J corresponding to the central element of $\mathcal{MCG}^*$ and that the quotient M/J is an orbifold: a disk with three cone points, one for each of the fixed points of J. Unoriented simple geodesics are invariant for this involution and it is easy to see that the three intersection points of a Markoff triple of geodesics $[\gamma_i] \cap [\gamma_j]$, $i \neq j$, coincide with the fixed points of J. The fixed points of J are *Weierstrass points*.

Proof of Lemma 2.1. Let w be the Weierstrass point not on γ_3. For each i, $T_{\gamma_3}^i(\gamma_1)$, $T_{\gamma_3}^{i+1}(\gamma_1)$ both intersect $T_{\gamma_3}(\gamma_3) = \gamma_3$ exactly once. So these three curves form a Markoff triple and $T_{\gamma_3}^i(\gamma_1) \cap T_{\gamma_3}^{i+1}(\gamma_1)$ is exactly the Weierstrass point not on γ_3. □

4. Signed angles

We now define the signed angle between two geodesics at an intersection (compare [Kerckhoff 1983; Series 2001; Jorgensen 2000] for a discussion of signed complex lengths in general.) Subsequently we find an explicit expression in terms of lengths of geodesics and study its behaviour on Teichmüller space.

Definition. Let $\alpha \neq \beta$ be a pair of oriented geodesics in $\mathbb{H}^2$ meeting in at a point x. There is a well defined *signed angle* between α, β at x; this is a function from ordered pairs of oriented geodesics to $]-\pi, \pi[$:

$$(\alpha, \beta) \mapsto \alpha \vee_x \beta.$$

One way to define $\alpha \vee \beta$ is to work in the disc model for $\mathbb{H}^2$. After conjugating we may assume

$$\alpha = (-1, 1), \quad \beta = (-e^{i\theta}, e^{i\theta})$$

for some $\theta \in]-\pi, \pi[$, so $\alpha \cap \beta = \{0\}$. Now $z \mapsto e^{i\theta} z$ is the unique rotation fixing 0 and taking 1 to $e^{i\theta}$ and hence α (oriented in the direction from -1 to 1) to β (oriented in the direction from $-e^{i\theta}$ to $e^{i\theta}$.) Set $\alpha \vee_x \beta = \theta$.

For a pair of geodesics $[\alpha] \neq [\beta]$ meeting at a point x in a surface M one defines the signed angle at x by lifting to $\mathbb{H}^2$. When $[\alpha] \neq [\beta]$ meet at a single point x in the surface we shall omit x and use simply $\alpha \vee \beta$ to denote this angle.

Computation of the signed angle. Let $[\alpha], [\beta]$ be a pair of simple closed geodesics meeting in a single point $x \in M$. For completeness, we show how to calculate the angle $\alpha \vee \beta$ in terms of l_α, l_β and the length of the boundary l_δ (compare [Wolpert 1983a; 1983b]). Let $\gamma = \alpha\beta^{-1} \in \pi_1(M, x)$; from the preceding section $[\alpha], [\beta], [\gamma]$ is a Markoff triple of geodesic and the pairwise intersections are the Weierstrass points. The quotient of $[\alpha] \cup [\beta] \cup [\gamma]$ is an embedded geodesic triangle on M/J with vertices at the three cone points. Its side lengths are $l_\alpha/2, l_\beta/2, l_\gamma/2$, and the (hyperbolic) cosine rule yields

$$\cosh(l_\gamma/2) = \cosh(l_\alpha/2)\cosh(l_\beta/2) - \sinh(l_\alpha/2)\sinh(l_\beta/2)\cos(\alpha \vee \beta), \tag{7}$$

so that

$$\alpha \vee \beta = \arccos \frac{\cosh(l_\alpha/2)\cosh(l_\beta/2) - \cosh(l_\gamma/2)}{\sinh(l_\alpha/2)\sinh(l_\beta/2)}. \tag{8}$$

One sees immediately that $\alpha \vee \beta$ is continuous on Teichmüller space.

Finally we derive another expression for $\alpha \vee \beta$, which will be useful in the proof of Theorem 2. Replacing in the Markoff cubic using (7) one obtains

$$\sinh^2(l_\alpha/2)\sinh^2(l_\beta/2)\sin^2(\alpha \vee \beta) = \cosh^2(l_\delta/4), \tag{9}$$

where δ is the boundary geodesic. If, $\alpha \vee \beta \in]0, \pi/2[$, then

$$\alpha \vee \beta = \arcsin \frac{\cosh(l_\delta/4)}{\sinh(l_\alpha/2)\sinh(l_\beta/2)}. \tag{10}$$

Remark. Another way of thinking of the relation (9) is as a hyperbolic version of the usual formula for the area of a euclidean torus:

$$2l_\alpha l_\beta \sin(\alpha \vee \beta) = \text{area of torus},$$

where α, β are closed Euclidean geodesics meeting in a single point with angle $\alpha \vee \beta$.

The reader is left to check that, unfortunately, the analogous series for the variation of the Euclidean angles does not converge absolutely, so we obtain no new identity on the moduli space of Euclidean structures.

5. Differentiability

We now study the regularity of $\alpha \vee \beta$ as we vary the surface over Teichmüller space. It is well known [Abikoff 1980] that for any closed geodesic γ the function

$$x \mapsto l_\gamma(x), \mathcal{T}_1(l_\delta) \to \mathbb{R}^+$$

is differentiable and even real analytic. It is not difficult to see from (8) that $\alpha \vee \beta$ is also real analytic.

From the expression (10) for the angle obtained above, we have

$$d(\alpha \vee \beta) = \cosh(l_\delta/4) \frac{\coth(l_\alpha/2)\, dl_\alpha + \coth(l_\beta/2)\, dl_\beta}{4\big(\sinh^2(l_\alpha/2)\sinh^2(l_\beta/2) - \cosh^2(l_\delta/4)\big)^{1/2}},$$

provided $|\alpha \vee \beta| \neq \pi/2$ (by equation (9)) — in other words, this equality holds on the complement of the subset where $|\alpha \vee \beta|$ attains its maximum. On this critical set the numerator $\coth(l_\alpha/2)\, dl_\alpha + \coth(l_\beta/2)\, dl_\beta$ vanishes and (it is left to the reader to check that) the right-hand side defines a form which extends by continuity to the whole of Teichmüller space.

6. The length spectrum of simple geodesics

We prove two lemmata used in the proof of Theorem 2 in the next section. For a discussion of length spectra in general see [Schmutz Schaller 1998].

Notation. Sections 4 and 6 deal with lengths of geodesics, and the mapping class group will not figure explicitly. To make this clear we set

$$B^+ := \mathcal{MCG}^*.(\alpha, \beta).$$

Let $M_{g,n}$ be a hyperbolic surface of genus g with n punctures and x the corresponding point in the moduli space. Let $\mathcal{G}_0$ be the set of all simple closed geodesics. Define the *simple length spectrum*, denoted by $\sigma_0(x) \subset \mathbb{R}^+$, to be the set $\{l_\gamma, \gamma \in \mathcal{G}_0\}$ counted *with* multiplicities. It is more useful to think in terms of the associated *counting function*

$$N(\mathcal{G}_0, t) := \operatorname{card}\{[\gamma] \in \mathcal{G}_0, l_\gamma(x) < t\}.$$

There are two important features of the simple length spectrum:

(1) The infimum over all lengths $l_\gamma(x)$ of all closed geodesics is strictly positive and attained for a simple closed geodesic, the *systole* $\operatorname{sys} x$. We shall also denote by $\operatorname{sys} x$ the length of this geodesic.

(2) $\sigma_0(M)$ is *discrete*, that is $N(\mathcal{G}_0, t)$ is finite for all $t \geq 0$, and moreover has *polynomial growth*, specifically

$$N(\mathcal{G}_0, t) \leq A t^{6g-6+2n}$$

for some $A = A(x) > 0$ [Rivin 2001; Rees 1981].

Lemma 6.1. *Let $x \in \mathcal{M}_1(l_\delta)$. For all $t > 0$ there exists $N = N(t, x) > 0$ such that the inequality*

$$\sinh(l_\alpha(x)/2) \sinh(l_\beta(x)/2) \geq t,$$

holds for all but N pairs $(\alpha, \beta) \in B^+$.

Proof. The quantity $\sinh(l_\alpha/2) \sinh(l_\beta/2)$ is at least

$$\tfrac{1}{2}\big(\sinh\big(\tfrac{1}{2}l_\alpha\big) \sinh\big(\tfrac{1}{2}\operatorname{sys} x\big) + \sinh\big(\tfrac{1}{2}\operatorname{sys} x\big) \sinh\big(\tfrac{1}{2}l_\beta\big)\big) \;\geq\; \tfrac{1}{2}\big(l_\alpha + l_\beta\big) \sinh\big(\tfrac{1}{2}\operatorname{sys} x\big).$$

The lemma follows by the discreteness of the length spectrum. □

The Collar Lemma. Useful information about the length spectrum can be obtained from the Collar Lemma [Buser 1992, Chapter 4]. Given a closed simple geodesic μ there is an embedded collar (regular tubular neighbourhood of μ) such that

$$(\text{width of collar round } \mu) \geq w(l_\mu),$$

for $w(s) := 2 \operatorname{arcsinh}(1/\sinh(s/2))$. One bounds from below the length of any closed geodesic γ such that $\gamma \cap \mu \neq \varnothing$ by the intersection number times the width the collar round μ, that is,

$$i(\gamma, \mu) \leq \frac{l_\gamma}{w(l_\mu)}, \tag{11}$$

where $i(\gamma, \mu) := \operatorname{card}(\gamma \cap \mu)$ is the *geometric intersection number*.

7. Bounding the variation of $\mathcal{D}$: proof of Theorem 2

Fix a metric on M, let $x \in \mathcal{M}_1(l_\delta)$ be the corresponding point in the moduli space, and fix a closed simple geodesic μ.

Claim. *There exists* $K = K(\operatorname{sys} x, l_\mu, l_\delta)$ *such that*

$$\sum_{B^+} |d(\alpha \vee \beta).t(\mu)| \leq K \left(\sum_{\gamma \in \mathcal{G}_0} l_\gamma e^{-l_\gamma/2} \right)^2.$$

Convergence follows since the simple length spectrum grows polynomially.

Proof of Claim. We start with the formula obtained for $\alpha \vee \beta$ in Section 3:

$$d(\alpha \vee \beta) = \cosh r \frac{\coth a \, da + \coth b \, db}{(\sinh^2 a \sinh^2 b - \cosh r)^{1/2}},$$

where we have set, to simplify notation,

$$a = l_\alpha/2, \quad b = l_\beta/2, \quad r = l_\delta/4.$$

For the geodesic μ Wolpert's corollary gives

$$\left| d(\alpha \vee \beta).t(\mu) \right| \leq \left| \frac{\cosh r \left(i(\alpha, \mu) \coth a + i(\beta, \mu) \coth b \right)}{\sinh a \sinh b \, (\sinh^2 a \sinh^2 b - \cosh^2 r)^{1/2}} \right|.$$

Firstly, note that $\coth a, \coth b \leq \coth(\frac{1}{2} \operatorname{sys} x)$ since $a, b \geq \operatorname{sys} x/2$. Secondly, replacing for $i(\alpha, \mu), i(\beta, \mu)$ using (11) above we obtain the following upper bound for the variation:

$$\frac{\cosh(r) \coth(\operatorname{sys} x/2)}{w(l_\mu)} \times \frac{l_\alpha + l_\beta}{(\sinh^2(a) \sinh^2(b) - \cosh^2(r))^{1/2}}.$$

Note that the leading factor does not depend on l_α, l_β.

Thirdly, by Lemma 6.1 for all but finitely many pairs (α, β) in B^+ one has

$$\sinh^2 a \sinh^2 b - \cosh^2 r \geq \tfrac{1}{2} \sinh^2 a \sinh^2 b \geq \tfrac{1}{8} \exp(a + b).$$

Finally, the sum over all the configurations satisfies

$$\sum_{B^+} (l_\alpha + l_\beta) e^{-\frac{1}{2}(l_\alpha + l_\beta)} \leq \frac{2}{\operatorname{sys} x} \left(\sum_\alpha l_\alpha e^{-l_\alpha/2} \right) 2,$$

since

$$\frac{2}{\operatorname{sys} x} l_\alpha l_\beta \geq l_\alpha + l_\beta.$$

The claim follows immediately, and with it Theorem 2. □

8. Calculation of the term in Theorem 1

As promised in Section 2 we now derive the formula (5). We pick up at the end of *step two* Section 2 and adopt the same notation. In particular, we have a geodesic γ', a Weierstrass point w not on γ' and a pair of simple geodesics $\gamma^{\pm}$ spiralling to γ'. We lift to $\mathbb{H}^2$ and use hyperbolic trigonometry to do the calculation (Figure 1).

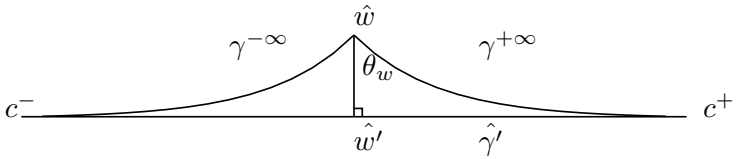


Figure 1. The triangle in $\mathbb{H}^2$ used to calculate $\gamma^{-\infty} \vee \gamma^{+\infty}$.

Let $\hat{\gamma}'$ be a lift of γ' to $\mathbb{H}^2$ with endpoints c^-, c^+ in $\partial\mathbb{H}^2$. For a point $\hat{w} \in \mathbb{H}^2$, let $\hat{w}' \in \hat{\gamma}$ denote the nearest point to $\hat{w}$ on $\hat{\gamma}'$; we write $|\hat{w}'\hat{w}|$ for the hyperbolic distance from $\hat{w}$ to $\hat{w}'$. The triangle $\hat{w}'\hat{w}c^+$ is a hyperbolic triangle with a right angle at $\hat{w}'$ and an angle $\theta_{\hat{w}}$ at $\hat{w}$. Trigonometry yields (see [Buser 1992, Theorem 2.2.2(iv)])

$$\cot(\theta_{\hat{w}}) = \sinh(|\hat{w}'\hat{w}|). \tag{12}$$

Note that the side opposite $\hat{w}$ has infinite length but the passage to the limit in the formula (iv) is valid.

We now choose $\hat{w}$ to be a lift of the Weierstrass point w such that $\hat{w}'\hat{w}$ projects to a simple arc on the surface. The geodesic $\hat{w}c^+$ projects to a simple geodesic on the surface that spirals to γ' and, without loss of generality, we may assume that this is $\gamma^{+\infty}$. Likewise the projection of $\hat{w}c^-$ is simple and spirals to γ' so this must be $\gamma^{-\infty}$. It follows from this that

$$|\gamma^{-\infty} \vee \gamma^{+\infty}| = 2\theta_{\hat{w}}.$$

It only remains to express the quantity $|\hat{w}'\hat{w}|$ in terms of the lengths ℓ_δ, ℓ_γ. Note first that this quantity is the distance from w to γ'. Now cut along γ' to obtain a pair of pants P. The Weierstrass point w is now the midpoint of the unique simple common perpendicular, labelled $w'w''$ in Figure 2, running between the two boundary components of the pants corresponding to γ'. By further cutting along common perpendiculars (see Figure 2) one obtains four congruent right angled pentagons; we shall work with the pentagon whose vertices are labelled $abcw'w$. There are three sides of this pentagon that concern us. The side labelled ww' has length $|\hat{w}'\hat{w}|$, the side adjacent, labelled $w'c$, has length $\ell_{\gamma'}/2$ and the side "opposite" this pair, labelled ab, has length $\ell_\delta/4$. The lengths of these three sides

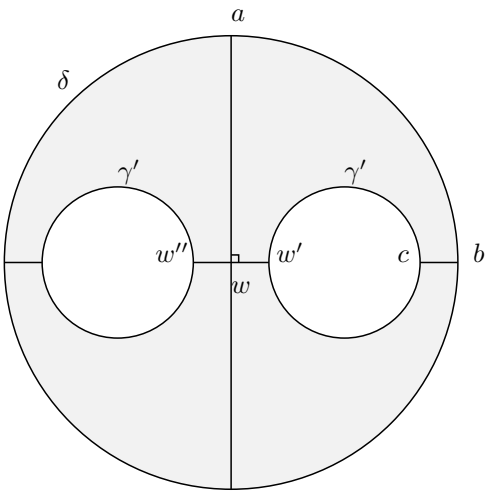


Figure 2. A torus cut along a geodesic γ' to obtain a pair of pants. The pants has been subdivided into four congruent pentagons.

are related by (see [Buser 1992])

$$\sinh(|\hat{w}\hat{w}'|)\sinh(\ell_{\gamma'}/2) = \cosh(\ell_\delta/4) \tag{13}$$

so that, replacing in (12), one has

$$\cot(\tfrac{1}{2}|\gamma^{-\infty} \vee \gamma^{+\infty}|) = \cot(\theta_{\hat{w}}) = \frac{\cosh(\ell_\delta/4)}{\sinh(\ell_{\gamma'}/2)}.$$

This is equivalent to (5).

It is amusing, as an afterthought, to note that the relation (13) can be deduced from (9) in Section 4. One reglues the pants P to obtain a holed torus such that the endpoints of the common perpendicular w', w'' are identified. In this way one obtains a holed torus with a simple closed geodesic α' which meets γ' perpendicularly in a single point. The formula (9) applies with $\alpha = \alpha'$, $\beta = \gamma'$ and, noting that $\gamma' \vee \alpha' = \pi/2$, yields (13).

9. When the method breaks down

The curious reader might wonder why we do not give a proof of the original identity using this method. We explain here how our method breaks down in general.

One can consider an analogous series which is roughly speaking the sum of all of Penner's h-lengths. Recall that the h-lengths are the lengths of the connected components (horocyclic arcs) obtained when one removes the intersection with the edges of an ideal triangulation from the horocycles; see [Penner 1987] for a details. In the case of the punctured torus there are three h-lengths satisfying a

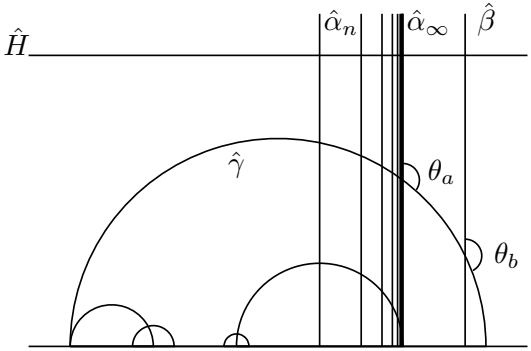


Figure 3. A choice of lifts of α_∞, β, γ to $\mathbb{H}^2$, and the angles θ_a, θ_b.

single relation and these can be identified with the quantities $a/bc, b/ac, c/ab$ arising in the discussion of Bowditch's method in the introduction.

Every punctured torus contains a cusp region X of area 2, the boundary is an embedded topological circle H. We choose once and for all an orientation for this circle. Let α be a simple directed bicuspidal geodesic; it meets H in exactly two points a^-, a^+ joining a^- to a^+. If α, β are a pair of disjoint simple directed bicuspidal geodesics then define the directed h-coordinate $\alpha \vee_h \beta$ to be the distance along the oriented curve H from a^- to b^-. Using the elliptic involution one sees that replacing a^- by a^+ and b^- by b^+ one gets the same number.

For any pair of disjoint simple directed bicuspidal geodesics one also has the identity

$$\beta \vee_h \alpha + \alpha \vee_h \beta = 2$$

since the length of H is 2. The mapping class group acts on pairs of disjoint simple directed bicuspidal geodesics and there are exactly two orbits. For any pair of disjoint simple directed bicuspidal geodesics α, β there is exactly one closed simple geodesic that meets each of α, β exactly once. After choosing orientations appropriately we may assume β is the image of α under the Dehn twist T_γ. One checks that the sum

$$\sum_{n\in\mathbb{Z}} T_\gamma^n(\alpha) \vee T_\gamma^{n+1}(\alpha)$$

converges to $1 - 1/(1 + e^{\ell_\gamma})$.

What goes wrong is the following.

Let β' denote the unique closed simple geodesic disjoint from β and $T_{\beta'}$ the corresponding Dehn twist. The geodesics β' and α meet (exactly once) and consider the associated sequence of directed geodesics $T_{\beta'}^n(\alpha), n > 0$. Let $a_n^- \in H$ denote the corresponding sequence of points realising the value of $T_{\beta'}^n(\alpha) \vee_h \beta$. There is a geodesic α_∞ asymptotic to β with a single endpoint up the cusp such that a_n^- tends to the intersection $a_\infty = \alpha_\infty \cap H$ as $n \to \infty$. Note that $a_\infty \neq \beta^-$.

Now choose another simple closed geodesic $\gamma \neq \beta$; note that these two geodesics must meet at least once. For large n value of $d(T^n_{\beta'}(\alpha) \vee_h \beta).\tau_\gamma$ is bounded below by $|\cos(\theta_a) - \cos(\theta_b)|$ where θ_a is the angle between a lift of α_∞ to $\mathbb{H}^2$ and the first lift of γ that it meets and θ_b the angle between a lift of β and the first lift of γ that it meets (see figure) . One sees in this way that the corresponding series for the variation does not converge since terms do not tend to 0.

Note that this reasoning fails in the case considered in the text. This is because if α', β' are simple closed geodesics which meet in a single point then $T^n_{\beta'}(\alpha') \vee \beta' \to 0$ as $n \to \infty$; see [McShane 2004] for details.

Acknowledgements

It is a pleasure to thank Troels Jørgensen for the original question and Pete Storm, Benson Farb, Frederic Paulin, Juan Souto for useful conversations and encouragement. Finally, as the referee points out, and we are happy to acknowledge, this article owes much to [Wolpert 1983a], which is (still) an excellent introduction to the moduli of the punctured torus.

References

[Abikoff 1980] W. Abikoff, *The real analytic theory of Teichmüller space*, Lecture Notes in Mathematics **820**, Springer, Berlin, 1980. MR 82a:32028 Zbl 0452.32015

[Bowditch 1996] B. H. Bowditch, "A proof of McShane's identity via Markoff triples", *Bull. London Math. Soc.* **28**:1 (1996), 73–78. MR 96i:58137 Zbl 0854.57009

[Buser 1992] P. Buser, *Geometry and spectra of compact Riemann surfaces*, Progress in Mathematics **106**, Birkhäuser, Boston, 1992. MR 93g:58149 Zbl 0770.53001

[Goldman 1984] W. M. Goldman, "The symplectic nature of fundamental groups of surfaces", *Adv. in Math.* **54**:2 (1984), 200–225. MR 86i:32042 Zbl 0574.32032

[Goldman 1986] W. M. Goldman, "Invariant functions on Lie groups and Hamiltonian flows of surface group representations", *Invent. Math.* **85**:2 (1986), 263–302. MR 87j:32069 Zbl 0619.58021

[Goldman 2003] W. M. Goldman, "The modular group action on real SL(2)-characters of a one-holed torus", *Geom. Topol.* **7** (2003), 443–486. MR 2004k:57001 Zbl 1037.57001

[Goldman 2004] W. M. Goldman, "The complex-symplectic geometry of SL(2, $\mathbb{C}$)-characters over surfaces", pp. 375–407 in *Algebraic groups and arithmetic*, edited by S. G. Dani and G. Prasad, Tata Inst. Fund. Res., Mumbai, 2004. MR 2005i:53110 Zbl 1089.53060

[Jorgensen 2000] T. Jorgensen, "Composition and length of hyperbolic motions", pp. 211–220 in *In the tradition of Ahlfors and Bers* (Stony Brook, NY, 1998), edited by I. Kra and B. Maskit, Contemp. Math. **256**, Amer. Math. Soc., Providence, RI, 2000. MR 2001d:30080 Zbl 0973.53030

[Kerckhoff 1983] S. P. Kerckhoff, "The Nielsen realization problem", *Ann. of Math.* (2) **117**:2 (1983), 235–265. MR 85e:32029 Zbl 0528.57008

[Labourie and Mcshane 2006] F. Labourie and G. Mcshane, "Cross ratios and identities for higher Thurston theory", preprint, 2006. math/0611245

[McShane 1998] G. McShane, "Simple geodesics and a series constant over Teichmuller space", *Invent. Math.* **132**:3 (1998), 607–632. MR 99i:32028 Zbl 0916.30039

[McShane 2004] G. McShane, "Weierstrass points and simple geodesics", *Bull. London Math. Soc.* **36**:2 (2004), 181–187. MR 2004j:57023 Zbl 1052.57021

[Mirzakhani 2003] M. Mirzakhani, "Simple geodesics on hyperbolic surfaces and volumes of moduli space", preprint, 2003.

[Penner 1987] R. C. Penner, "The decorated Teichmüller space of punctured surfaces", *Comm. Math. Phys.* **113**:2 (1987), 299–339. MR 89h:32044 Zbl 0642.32012

[Rees 1981] M. Rees, "An alternative approach to the ergodic theory of measured foliations on surfaces", *Ergodic Theory Dynam. Syst.* **1**:4 (1981), 461–488 (1982). MR 84c:58065 Zbl 0539.58018

[Rivin 2001] I. Rivin, "Simple curves on surfaces", *Geometriae Dedicata* **87**:1-3 (2001), 345–360. MR 2003c:57018 Zbl 1002.53027

[Schmutz Schaller 1998] P. Schmutz Schaller, "Geometry of Riemann surfaces based on closed geodesics", *Bull. Amer. Math. Soc. (N.S.)* **35**:3 (1998), 193–214. MR 99b:11098 Zbl 1012.30029

[Series 2001] C. Series, "An extension of Wolpert's derivative formula", *Pacific J. Math.* **197**:1 (2001), 223–239. MR 2001m:30055 Zbl 1065.30044

[Wolpert 1979] S. Wolpert, "The length spectra as moduli for compact Riemann surfaces", *Ann. of Math.* (2) **109**:2 (1979), 323–351. MR 80j:58067 Zbl 0441.30055

[Wolpert 1982] S. Wolpert, "The Fenchel-Nielsen deformation", *Ann. of Math.* (2) **115**:3 (1982), 501–528. MR 83g:32024 Zbl 0496.30039

[Wolpert 1983a] S. Wolpert, "On the Kähler form of the moduli space of once punctured tori", *Comment. Math. Helv.* **58**:2 (1983), 246–256. MR 84j:32028 Zbl 0527.30031

[Wolpert 1983b] S. Wolpert, "On the symplectic geometry of deformations of a hyperbolic surface", *Ann. of Math.* (2) **117**:2 (1983), 207–234. MR 85e:32028 Zbl 0518.30040

[Zhang et al. 2004a] Y. Zhang, S. P. Tan, and Y. L. Wong, "Generalizations of McShane's identity to hyperbolic cone-surfaces", preprint, 2004. math.GT/0411184

[Zhang et al. 2004b] Y. Zhang, S. P. Tan, and Y. L. Wong, "Necessary and sufficient conditions for McShane's identity and variations", preprint, 2004. math.GT/0411184

[Zhang et al. 2005] Y. Zhang, S. P. Tan, and Y. L. Wong, "Generalized Markoff maps and McShane's identity", preprint, 2005. math.GT/0502464

Received February 23, 2006. Revised February 8, 2007.

GREG MCSHANE
LABORATOIRE ÉMILE PICARD
UNIVERSITÉ PAUL SABATIER
118 ROUTE DE NARBONNE
31062 TOULOUSE CEDEX 4
FRANCE

greg.mcshane@gmail.com
picard.ups-tlse.fr/∼mcshane

PACIFIC JOURNAL OF MATHEMATICS
Vol. 231, No. 2, 2007
dx.doi.org/10.2140/pjm.2007.231.481

ON THE GEOMETRIC AND THE ALGEBRAIC RANK OF GRAPH MANIFOLDS

JENNIFER SCHULTENS AND RICHARD WEIDMANN

For any $n \in \mathbb{N}$ we construct graph manifolds of genus $4n$ whose fundamental group is $3n$-generated.

1. introduction

A *Heegaard surface* of an orientable closed 3-manifold M is an embedded orientable surface S such that $\overline{M - S}$ consists of 2 handlebodies V_1 and V_2. This decomposition of M is called a *Heegaard splitting* and denoted by $M = V_1 \cup_S V_2$. We say that the splitting is of *genus* g if S is of genus g. It is not difficult to see that any orientable closed 3-manifold admits a Heegaard splitting. If M admits a Heegaard splitting of genus g but no Heegaard splitting of smaller genus then we say that M has *Heegaard genus* g and write $g(M) = g$.

Clearly any curve in a handlebody can be homotoped to its boundary. It follows that for any Heegaard splitting $M = V_1 \cup_S V_2$ every curve in M can be homotoped into V_1. Thus the map induced by the inclusion of V_1 into M maps a generating set of $\pi_1(V_1)$ to a generating set of $\pi_1(M)$. Since $\pi_1(V_1)$ is generated by g elements, $\pi_1(M)$ is also generated by g elements. Thus $g(M) \geq r(M)$, where $r(M)$ denotes the minimal number of generators of $\pi_1(M)$. Sometimes we will refer to $g(M)$ as the *geometric rank* and to $r(M)$ as the *algebraic rank* of M.

F. Waldhausen [1978] asked whether the converse inequality also holds: is $g(M) = r(M)$? A positive answer would have implied the Poincaré conjecture. First counterexamples however were found by M. Boileau and H. Zieschang [1984]. These examples were Seifert fibered manifolds with $g(M) = 3$ and $r(M) = 2$. The work of Y. Moriah and J. Schultens [1998] further shows that this class extends to higher-genus examples: Seifert manifolds with $g(M) = n + 1$ and $r(M) = n$. In [Weidmann 2003] a class of graph manifolds was found for which $g(M) = 3$ and $r(M) = 2$. The original Boileau–Zieschang examples can be interpreted as a special case of these graph manifolds.

Schultens was supported in part by NSF Grants #0203680 and #0603736. Both authors were supported by the Max-Planck-Institut für Mathematik, Bonn, Germany.
MSC2000: primary 54C40, 14E20; secondary 46E25, 20C20.
Keywords: 3-manifolds, rank, Heegaard genus, graph manifold.

We here show how the phenomenon observed in [Weidmann 2003] generalizes and how it can occur multiple times within a single graph manifold. This yields graph manifolds where the difference between the algebraic and the geometric rank is arbitrarily high.

We wish to thank the referee for an insightful reading of our manuscript and for numerous helpful comments.

2. Formulation of the main results

Let M be a closed graph manifold. We will always assume that M comes equipped with its characteristic tori $\mathcal{T} = \mathcal{T}_M$ and a fixed Seifert fibration on every component of $\overline{M - \mathcal{T}}$. Recall that the Seifert fibrations are unique up to isotopy except for components homeomorphic to Q, the Seifert space with base orbifold the disk with two cone points of order 2. The space Q can also be fibered as the orientable circle bundle over the Möbius band. We will refer to the components of $\overline{M - \mathcal{T}}$ as the *Seifert pieces* of M. The Seifert pieces of M are up to isotopy precisely the maximal Seifert submanifolds of M. We will mostly work with *totally orientable* graph manifolds, that is, orientable graph manifolds whose Seifert pieces have orientable base orbifold. This makes the Seifert fibrations unique up to isotopy on all Seifert pieces.

Let N be a Seifert piece of M. Denote the fiber of N by f. Let $T_1, \ldots, T_n$ be the boundary components of N and let $\gamma_i \subset T_i$ be the curve corresponding to the fiber of that Seifert piece L_i which is reached by travelling from N transversely through T_i. Note that we possibly have $N = L_i$. The maximality of the Seifert piece N guarantees that for all i the intersection number of f with γ_i does not vanish.

We then define $\hat{N}$ to be the manifold $N(\gamma_1, \ldots, \gamma_n)$ obtained from N by performing a Dehn filling with slope γ_i at each boundary component T_i. It is clear that the Seifert fibration of N can be extended to a Seifert fibration of $\hat{N}$ as f has nontrivial intersection number with all γ_i.

In the following we will denote the base orbifold of a Seifert piece N by $\mathbb{O}(N)$. We will denote an orbifold by its topological type with a list of the orders of cone points, where ∞ stands for a boundary component. We will denote the disc by D, the sphere by S^2, the annulus by A, the orientable surface of genus g, for $g > 0$, by F_g and the projective plane by P^2.

Theorem 1. *Let M be a closed graph manifold consisting of two Seifert pieces N_1 and N_2 glued along T, where $\mathbb{O}(N_1) = F_g(r, \infty)$, $\mathbb{O}(N_2) = D(p, q)$ with $(p, q) = 1$ and $\min(p, q) \leq 2g + 1$ such that the intersection number of the fibers of N_1 and N_2 equals 1.*

Then $\pi_1(M)$ is generated by $2g+1$ elements. Furthermore if M admits a Heegaard splitting of genus $2g+1$ then one of the following holds:

(1) *N_2 is the exterior of a s-bridge knot with $s \leq 2g+1$ and the fiber of N_1 is identified with the meridian of N_2, that is, $\hat{N}_2 = S^3$.*

(2) *$\hat{N}_1$ admits a horizontal Heegaard splitting of genus $2g$.*

(Conversely, M has a Heegaard splitting of genus $2g+1$ if (1) or (2) is satisfied. This splitting is vertical in N_1 and pseudohorizontal in N_2 in case (1) and pseudohorizontal in N_1 and vertical in N_2 in case (2). This follows from the proof of Theorem 1.)

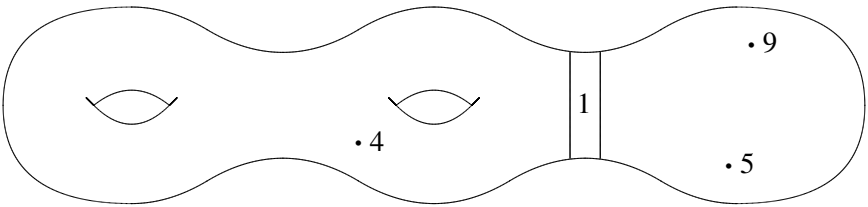


Figure 1. A graph manifold with 5-generated fundamental group.

We will further see that all manifolds of this type admit a Heegaard splitting of genus $2g+2$. Furthermore, most of these manifolds do not admit a Heegaard splitting of genus $2g+1$ as for any given pair of such manifolds N_1 and N_2 there are at most three gluing maps that yield a graph manifold of genus $2g+1$. It is also possible to show that $\pi_1(M)$ cannot be generated by less than $2g+1$ elements, but the argument is complicated.

A careful analysis of these examples shows that the phenomenon is of a local nature, it can therefore be reproduced multiple times within a graph manifold with a more complex underlying graph. Hence:

Theorem 2. *For any $n \in \mathbb{N}$ there exists a graph manifold M_n with $3n$-generated fundamental group that has Heegaard genus at least $4n$.*

This paper is organized as follows. In Section 3 we review the structure theorem for Heegaard splittings of totally orientable graph manifolds as proven in [Schultens 2004]. Then we study in more detail how Heegaard surfaces can intersect the Seifert pieces that are the building blocks of our examples. In Sections 5 and 6 we give the proofs of Theorems 1 and 2. We conclude by describing a class orientable Seifert manifolds with $2n$-generated fundamental group which we believe to be of Heegaard genus $3n$. However, these manifolds are not totally orientable.

3. Heegaard splittings of totally orientable graph manifolds

A graph manifold M is *totally orientable* if M is orientable and every Seifert piece N of M fibers over an orientable base space. In [Schultens 2004] it is shown that the Heegaard splittings of totally orientable graph manifolds have a structure that can be completely described. To do so, one considers a decomposition of M into edge manifolds and vertex manifolds. The edge manifolds are the submanifolds of the form $T \times I$, where T is one of the characteristic tori, $\mathcal{T}$, of M. The vertex manifolds are the components of the complement of the edge manifolds. Note that each vertex manifold is homeomorphic to a component of $M - \mathcal{T}$.

Heegaard splittings themselves are rather unwieldy. Instead we work with the surfaces arising in what is called a "strongly irreducible untelescoping" of a Heegaard splitting. We use the terms pseudohorizontal, horizontal, pseudovertical and vertical to describe the possible structure for the restriction of such a surface to the vertex manifolds. The restriction of such a surface to the edge manifolds takes three possible forms. It too plays a nontrivial role in the structure of the Heegaard splitting of a graph manifold.

A two-sided surface F in a 3-manifold M is said to be *weakly reducible* if there are disjoint essential curves a, b in F that bound disks D_a, D_b whose interior is disjoint from F and such that near their boundary D_a, D_b lie on opposite sides of F. A two-sided surface F in a 3-manifold M is said to be *strongly irreducible* if it is not weakly reducible.

Heegaard splittings correspond to handle decompositions. Given a 3-manifold M and a decomposition $M = V \cup_S W$ into two handlebodies, one handlebody, say V, provides the 0-handles and 1-handles and the other, W, provides the 2-handles and 3-handles. Without loss of generality, there is only one 0-handle and one 3-handle. Corresponding to $M = V \cup_S W$ we then have a handle decomposition in which all 1-handles are attached before any of the 2-handles. An *untelescoping* of a Heegaard splitting is a rearrangement of the order in which the 1-handles and 2-handles are attached. In the handle decomposition obtained we first attach the 0-handle, then some 1-handles, then some 2-handles, then some 1-handles, then some 2-handles, etc and finally, the 3-handle. We specify an untelescoping by a collection of surfaces $S_1, F_1, S_2, F_2, \ldots, F_{n-1}, S_n$. These surfaces are obtained as follows: S_1 is the boundary of the submanifold of M obtained by attaching the 0-handle and the first batch of 1-handles. F_1 is the boundary of the submanifold of M obtained by attaching the 0-handle, the first batch of 1-handles and the first batch of 2-handles. S_2 is the boundary of the submanifold of M obtained by attaching the 0-handle, the first batch of 1-handles, the first batch of 2-handles and the second batch of 1-handles. F_2 is the boundary of the submanifold of M obtained by attaching the 0-handle, the first batch of 1-handles, the first batch of 2-handles, the second

batch of 1-handles and the second batch of 2-handles, and so on. An untelescoping $S_1, F_1, S_2, F_2, \ldots, S_n$ is said to be strongly irreducible if each S_i is a strongly irreducible surface in M and each F_i is an incompressible surface in M. Note that a Heegaard splitting can be considered a trivial untelescoping S. If it is strongly irreducible, then it is its own strongly irreducible untelescoping.

For the discussion here it will be useful to note that each of the S_i and each of the F_i is separating, and that each pair S_i, F_i cobound a submanifold homeomorphic to $S_i \times I$ with 2-handles attached to $S_i \times \{1\}$. In particular, $\chi(S_i) < \chi(F_i)$. Similarly for F_i and S_{i+1}. The following theorem summarizes the discussion in [Scharlemann and Thompson 1994], [Scharlemann 2002] and [Scharlemann and Schultens 1999, Lemma 2].

Theorem 3. *Let M be a 3-manifold and $M = V \cup_S W$ a Heegaard splitting. Then $M = V \cup_S W$ has a strongly irreducible untelescoping $S_1, F_1, S_2, F_2, \ldots, S_n$. Furthermore,*

$$-\chi(S) = \sum_{i=1}^{n} \big(\chi(F_{i-1}) - \chi(S_i)\big).$$

A surface in a Seifert fibered space is *horizontal* if it is everywhere transverse to the fibration. It is *pseudohorizontal* if it is horizontal away from a fiber e and intersects a regular neighborhood $N(e)$ of e in an annulus that is a bicollar of e. (In [Moriah and Schultens 1998] the Heegaard splittings of a Seifert fibered space with pseudohorizontal splitting surface are called *horizontal Heegaard splittings*.)

Let F be a surface in a 3-manifold M and α an arc with interior in $M \setminus F$ and endpoints on F. Let $C(\alpha)$ be a collar of α in M. The boundary of $C(\alpha)$ consists of an annulus A together with two disks D_1, D_2, which we may assume to lie in F. We call the process of replacing F by $(F \setminus (D_1 \cup D_2)) \cup A$ *performing ambient 1-surgery on F along α*.

A surface S in a Seifert fibered space is *vertical* if it consists of regular fibers. It is *pseudovertical* if there is a vertical surface $\mathcal{V}$ and a collection of arcs Γ with interior disjoint from $\mathcal{V}$ that projects to an embedded collection of arcs such that S is obtained from $\mathcal{V}$ by ambient 1-surgery along Γ.

The definition of a standard Heegaard splitting for a graph manifold is rather lengthy. Let M be a graph manifold. A strongly irreducible untelescoping $S_1, F_1, S_2, F_2, \ldots, S_n$ of a Heegaard splitting $M = V \cup_S W$ is standard if it is as follows:

(1) Each F_i intersects each vertex manifold either in a horizontal or in a vertical surface (or $\varnothing$).

(2) Each F_i is either a torus entirely contained in an edge manifold or intersects an edge manifold in spanning annuli (or $\varnothing$);

(3) Each S_i intersects each vertex manifold in either a horizontal, pseudohorizontal, vertical or pseudovertical surface (or $\varnothing$).

(4) $\bigcup_i S_i$ intersects each edge manifold M_e = (torus) $\times$ [0, 1] in one of three possible ways: $\left(\bigcup_i S_i\right) \cap M_e$ consists of incompressible annuli; or $S_i \cap M_e$ can be obtained from a collection of incompressible annuli by ambient 1-surgery along an arc that is isotopic to an embedded arc in the boundary of the edge manifold; or there is a pair of simple closed curves $c, c' \subset$ (torus) such that $c \cap c'$ consists of a single point p and $S_i \cap M_e$ is the portion of the boundary of a collar of $c \times \{0\} \cup p \times [0, 1] \cup c' \times \{1\}$ that lies in the interior of M_e. Furthermore, each edge manifold must be met by at least one of the S_i.

Recall that for each i, F_i and S_i are separating. Thus if F_i or S_i intersects an edge manifold M_e in spanning annuli, then it must do so in an even number of spanning annuli. It is a nontrivial fact that if $S_1, F_1, S_2, F_2, \ldots, S_n$ meets M_e in spanning annuli, then between any two components of $F_i \cap M_e$ there must be at least two components of either $S_i \cap M_e$ or $S_{i+1} \cap M_e$. (This follows, for instance, by the argument used in the proof of Lemma 12.)

The Heegaard splitting $M = V \cup_S W$ is *standard* if every strongly irreducible untelescoping $S_1, F_1, S_2, F_2, \ldots, S_n$ of $M = V \cup_S W$, the union $\bigcup_i F_i \cup \bigcup_i S_i$ can be isotoped to be standard.

We recall the main theorem in [Schultens 2004], some of whose many consequences we will need.

Theorem 4. *Let $M = V \cup_S W$ be an irreducible Heegaard splitting of a totally orientable graph manifold. Then $M = V \cup_S W$ is standard.*

We assume that M is a totally orientable graph manifold, $M = V \cup_S W$ a Heegaard splitting and $S_1, F_1, \ldots, F_{n-1}, S_n$ a strongly irreducible untelescoping of $M = V \cup_S W$ that is standard. Then:

Fact 1. For N a vertex or edge manifold of M,

$$\sum_i \left(\chi(F_{i-1} \cap N) - \chi(S_i \cap N)\right) \geq 0.$$

Fact 2. Suppose e is an edge that abuts v. And suppose N_e, N_v, respectively, are the edge and vertex manifolds corresponding to e, v, respectively. Further suppose that $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_v$ is vertical and pseudovertical and a component $\tilde{S}$ of $\left(\bigcup_i S_i\right) \cap N_e$ is as in c). Then any annuli in $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_e$ that are parallel into ∂N_v can be isotoped to lie entirely in N_v. After this isotopy, $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_v$ is still vertical and pseudovertical.

Fact 3. Suppose e is an edge that abuts v. And suppose N_e, N_v, respectively, are the edge and vertex manifolds corresponding to e, v, respectively. Further suppose that $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_v$ is horizontal or pseudohorizontal. Then $\left(\bigcup_i F_i\right) \cap N_e$ does not contain a torus.

Fact 4. Suppose N_v is a vertex manifold and that a component $\tilde{S}$ of $\left(\bigcup_i S_i\right) \cap N_v$ is pseudohorizontal. Then $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_v = \tilde{S}$.

In the proof of the main theorem in [Schultens 2004] the strongly irreducible generalized Heegaard splitting $S_1, F_1, S_2, F_2, \ldots, S_n$ is isotoped to be standard. The first step in doing so involves isotoping $\bigcup_i F_i$ so that it intersects the boundaries of the edge manifolds in essential curves. After that, $\bigcup_i F_i$ remains fixed. We may assume without loss of generality that the number of components of this intersection is minimal. We will always assume that this is the case. As an immediate consequence we obtain our final fact:

Fact 5. Suppose N_e is an edge manifold. Then each annulus cut out by the intersection $\left(\bigcup_i F_i\right) \cap \partial N_e$ is essential in the compression body in which it is contained.

Lemma 5. *Let M be a graph manifold, N a Seifert piece and $S_1, F_1, S_2, F_2, \ldots, S_n$ a strongly irreducible untelescoping of a Heegaard splitting such that for some i $S_i \cap N$ is pseudohorizontal. Let f be the fiber contained in $S_i \cap N$. Then $(S_i \cap N) - \eta(f)$, for $\eta(f)$ a small regular neighborhood of f, is disconnected.*

Proof. Consider $(S_i \cap N) - \eta(f)$ in $N - \eta(f)$. Since S_i is separating in M, $(S_i \cap N) - \eta(f)$ is separating in $N - \eta(f)$. On the other hand, $(S_i \cap N) - \eta(f)$ is horizontal. A connected horizontal surface is not separating. It follows that $(S_i \cap N) - \eta(f)$ must have an even number of components, in this case at least two. Hence it is disconnected. □

Suppose that M is a closed totally orientable graph manifold and that $S_1, F_1, S_2, F_2, \ldots, S_n$ is a strongly irreducible untelescoping of a Heegaard splitting $M = V \cup_S W$. Suppose further that $S_1, F_1, S_2, F_2, \ldots, S_n$ has been isotoped to be standard. This implies in particular that for any vertex manifold N, $\left(\bigcup_i F_i \cup \bigcup_i S_i\right)$ meets ∂N in parallel simple closed curves. Thus to any vertex manifold N of M we associate the manifold N_S, which is the manifold obtained from N by performing a Dehn fillings at each components B of ∂N that meets $\left(\bigcup_i F_i \cup \bigcup_i S_i\right)$ along a slope represented by the curves $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap B$. Here N_S is not canonical. It depends on a specific (not necessarily unique) positioning of an (not necessarily unique) untelescoping. But we merely introduce this notation to discuss consequences of the existence of certain setups. N_S is a Seifert manifold if N contains a horizontal or pseudohorizontal component of $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N$, as $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap \partial N$ then consist of curves that have nontrivial intersection number with the fibre of N.

Lemma 6. *Suppose that for some i, $S_i \cap N$ is pseudohorizontal. Then the Seifert manifold N_S has a Heegaard surface S' such that $S' \cap N = S_i \cap N$. The corresponding Heegaard splitting is a horizontal Heegaard splitting of N_S. If $S_i \cap N$ is planar then S' is homeomorphic to S^2.*

Proof. Recall Fact 4 above: it tells us that if $S_i \cap N$ is pseudohorizontal, then $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N$ consists of a single component which we denote by $\tilde{S}$.

We may extend $\tilde{S}$ to a Heegaard surface of N_S by gluing meridional discs of the glued in solid tori to the boundary components of $\tilde{S}$. The corresponding Heegaard splitting for N_S is horizontal. If $\tilde{S}$ is planar then all boundary components get capped off which results in S^2. The assertion follows. □

4. Some lemmata

The lemmata in this section will enable us to compute the Heegaard genus of certain graph manifolds in the next section. We start by discussing the possible pseudohorizontal surfaces in the relevant Seifert manifolds. Some proofs rely on the theory of 2-dimensional orbifolds and their covering theory as discussed in [Scott 1983]. These lemmata will be used in our discussion of Heegaard splittings and their untelescopings. But many of these results are more general. We do not necessarily require S to be the splitting surface of a Heegaard splitting or to be a surface in an untelescoping. Lemma 14 concerns vertical and pseudovertical surfaces.

Lemma 7. *Let M be a graph manifold and N be a Seifert piece with $\mathbb{O}(N) = D(p, q)$ and $(p, q) = 1$. If $S \cap N$ is a planar surface that is pseudohorizontal, then*

(1) *N_S is homeomorphic to S^3 and*

(2) *$S \cap \partial N$ contains exactly $2p$ or $2q$ components.*

Note that N_S being homeomorphic to S^3 is equivalent to N being the exterior of an r-bridge knot with meridian μ parallel to $\partial N \cap S$, where $r = \min(p, q)$.

Proof. Possibly after exchanging p and q we can assume that S is horizontal in the space $\bar{N}$ obtained from N after removing a regular neighborhood of the exceptional fiber corresponding to the cone point of order q or by removing a neighborhood of a regular fiber. Clearly $\bar{N}$ is a Seifert space with $\mathbb{O}(\bar{N}) = A(p)$ or $\mathbb{O}(\bar{N}) = A(p, q)$. Let T_1 be the boundary component of $\bar{N}$ that bounds the drilled out solid torus and T_2 be the boundary of N. Let $\bar{S}$ be a component of $S \cap \bar{N}$. Clearly $\bar{S}$ is planar as it is a subsurface of a planar surface.

Since we assume that S is pseudohorizontal in N, it follows that $\bar{S} \cap T_1$ consists of a single loop α. Let γ be one component of $\bar{S} \cap T_2$ and let g be an element of $\pi_1(\bar{N})$ corresponding to γ. Recall that all other components of $\bar{S} \cap T_2$ are parallel to γ. Let n be the intersection number of γ with the fiber.

Since $\bar{S}$ is horizontal in $\bar{N}$, there exists a finite sheeted orbifold covering $\pi : \bar{S} \to O(\bar{N})$, in particular $\pi_*(\pi_1(\bar{S}))$ is of finite index in $\pi_1(O(\bar{N}))$. We distinguish the cases $O(\bar{N}) = A(p)$ and $O(\bar{N}) = A(p, q)$.

Case 1: $O(\bar{N}) = A(p)$. We have $\pi_1(A(p)) = \langle x, y | x^p \rangle$, where the generator y corresponds to the boundary curve corresponding to T_2. This implies in particular that $\pi_*(g)$ is conjugate to y^n.

Since $\bar{S}$ is planar this implies that $\pi_1(\bar{S})$ is generated by homotopy classes that correspond to the components of $\bar{S} \cap T_2$; that is, $\pi_*(\pi_1(\bar{S}))$ is generated by conjugates of the element y^n. Let $N(y^n)$ be the normal closure of y in $\pi_1(A(p))$. Clearly $\pi_1(A(p))/N(y^n) \cong \mathbb{Z}_n * \mathbb{Z}_p$ is infinite unless $n = 1$. Since $\pi_*(\pi_1(\bar{S})) \subset N(y^n)$, this implies that $n = 1$ as otherwise $\pi_*(\pi_1(\bar{S}))$ is contained in a subgroup of infinite index in $\pi_1(A(p))$ and is therefore of infinite index itself. Thus we can assume that $n = 1$ and that $\pi_*(\pi_1(\bar{S})) \subset N(y)$.

The orbifold covering space $\tilde{S}$ corresponding to $N(y)$ is an orbifold without cone points and is homeomorphic to the $(p+1)$-punctured sphere. Denote the corresponding covering map by $\tilde{\pi}$.

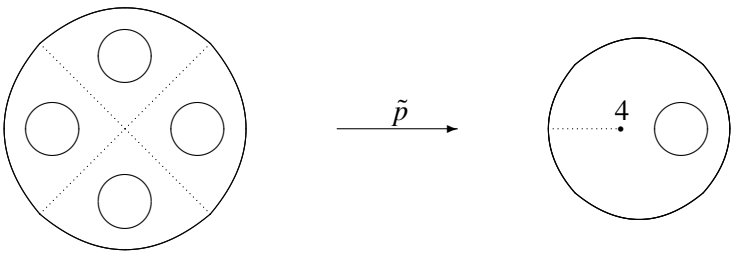


Figure 2. The 4-sheeted covering of $A(4)$ by a 5-punctured sphere.

Since $\pi_*(\pi_1(\bar{S})) \subset N(y)$, it follows that there exists a covering $\pi' : \bar{S} \to \tilde{S}$ such that $\pi = \tilde{\pi} \circ \pi'$.

Claim. *π' is a homeomorphism.*

As for both $\bar{S}$ and $\tilde{S}$, all but one boundary component map onto a curve corresponding to the element y it follows that π' is a homeomorphism when restricted to any of these boundary components. In particular π' extends to a covering $\pi'_\# : \bar{S}_\# \to \tilde{S}_\#$, where $\bar{S}_\#$ and $\tilde{S}_\#$ are the spaces obtained from $\bar{S}$ and $\tilde{S}$ by gluing discs to these boundary components. Since $\bar{S}_\#$ and $\tilde{S}_\#$ are discs, the map thus obtained is a homeomorphism. Thus the original π' was a homeomorphism, which proves the claim.

The second assertion is now immediate, because $S \cap \bar{N}$ is obtained from two copies of $\bar{S}$ by identifying two boundary components. All resulting boundary components lie in T_2. The first assertion follows from Lemma 6.

Case 2: $O(\bar{N}) = A(p, q)$. We have $\pi_1(A(p, q)) = \langle x, y, z \mid y^p, z^q \rangle$, where the generator x corresponds to the boundary curve corresponding to T_2. We see as

in the first case that $\pi_*(O(\bar{N}))$ lies in the kernel of the map $\phi : \pi_1(A(p,q)) \to \pi_1(A(p,q))/N(x^n)$. As $\pi_1(A(p,q))/N(y^n) \cong \mathbb{Z}_n * \mathbb{Z}_p * \mathbb{Z}_q$ is infinite for all $n \in \mathbb{N}$, this implies that $\pi_*(O(\bar{N}))$ is of infinite index in $\pi_1(A(p,q))$, which contradicts our assumption. □

Lemma 8. *Let M be a graph manifold and let N be a Seifert piece with $\mathcal{O}(N) = F_g(p,\infty)$ or $\mathcal{O}(N) = F_g(p,\infty,\infty)$. Suppose that $S \cap N$ is pseudohorizontal and $\chi(S \cap N) > -8g$ or $\chi(S \cap N) > -8g-4$, respectively.*

(1) *$S \cap T$ has two components for every component T of ∂N.*

(2) *$S \cap N$ extends to the splitting surface of a horizontal Heegaard surface of genus $2g$ of N_S.*

Proof. We only deal with the case $\mathcal{O}(N) = F_g(p,\infty)$ the other case is analogous.

Suppose that $S \cap N$ is pseudohorizontal with respect to the exceptional fiber or a regular fiber and let $\bar{N}$ be the space obtained by drilling out the neighborhood of this fiber. Let $\bar{S}$ be a component of $\bar{N} \cap S$. Recall that $S \cap N$ is obtained from two copies of $\bar{S}$ by identifying them along a boundary component. In particular we have that $\chi(S \cap N) = 2\chi(\bar{S})$.

Now $\bar{S}$ is a finite sheeted covering of $\mathcal{O}(\bar{N})$, where $\mathcal{O}(\bar{N}) = F_g(\infty,\infty)$ or $\mathcal{O}(\bar{N}) = F_g(p,\infty,\infty)$ depending on what kind of fiber was drilled out. Suppose that the covering is n-sheeted. In the case $\mathcal{O}(\bar{N}) = F_g(p,\infty,\infty)$, we must have $n \geq p$; otherwise the covering space must be a orbifold with singularities. Thus we have

$$\chi(S \cap N) = 2\chi(\bar{S}) = 2n\chi(\mathcal{O}(\bar{N})).$$

Since $\chi(\mathcal{O}(\bar{N})) = -2g$ or $\chi(\mathcal{O}(\bar{N})) = -2g-1+1/p$, it follows immediately from the hypothesis on the Euler characteristic that $n = 1$. Thus $\mathcal{O}(\bar{N}) = F_g(\infty,\infty)$: the exceptional fiber was drilled out. Assertion (1) is now immediate and (2) follows from the proof of Lemma 6. □

It will be important that many Seifert manifolds do not admit a pseudohorizontal surface of small genus indiscriminately of what graph manifold they belong to.

Lemma 9. *Let N be a Seifert manifold with $\mathcal{O}(N) = F_g(p,\infty)$ such that the exceptional fiber has invariant (α,β) with $1 \leq \beta < \alpha$.*

(1) *If $\alpha = 2$, there exist two slopes γ on ∂N such that $N(\gamma)$ admits a horizontal Heegaard splitting of genus $2g$.*

(2) *If $\alpha \neq 2$ and $\beta \in \{1, \alpha-1\}$, there exists one slope γ on ∂N such that $N(\gamma)$ admits a horizontal Heegaard splitting of genus $2g$.*

(3) *In all other cases $N(\gamma)$ has no Heegaard splitting of genus $2g$ if $\gamma \neq f$.*

Proof. If γ is the fiber then $N(\gamma)$ is not a Seifert manifold. In particular $N(\gamma)$ admits no horizontal Heegaard splitting as those are only defined for Seifert manifolds. If the intersection number m of γ with the fiber is greater than 1 then $M(\gamma)$ is a Seifert space with base orbifold $F_g(p,m)$ which has no Heegaard splitting of genus $2g$ by [Boileau and Zieschang 1984, Proposition 1.4(i)]. Suppose now that $m=1$. Let $e\in\mathbb{Z}$ be the Euler class of the Seifert space. By [Boileau and Zieschang 1984, Proposition 1.4(iii)] it follows that $N(\gamma)$ admits no Heegaard splitting of genus $2g$ unless $\beta-e\alpha=\pm1$. It is clear that there exists two values for e such that the equation holds if $\beta=1$ and $\alpha=2$, that there exists one solution if $\beta\in\{1,\alpha-1\}$ and none otherwise. The corresponding Heegaard splittings are constructed in [Boileau and Zieschang 1984, Section 1.10]. This proves the assertion. □

Lemma 10. *Let N be a Seifert manifold with $\mathbb{O}(N)=D(p,q)$ and $(p,q)=1$. Then N contains no compact planar horizontal surface.*

Proof. Suppose that S is a compact planar horizontal surface in N. Then there exists a finite sheeted orbifold covering $p:S\to D(p,q)$. Since all components of ∂S are parallel on ∂N, there exists a number $n\in\mathbb{N}$ such the restriction of p to any component of ∂S is a n-sheeted covering. This implies that we can extend p to a orbifold covering $p:S^2\to S^2(p,q,n)$ by gluing a disc to any component of ∂S and a disc with a cone point of order n to $D(p,q)$. If $n=1$ this yields a contradiction as $S^2(p,q,1)=S^2(p,q)$ is a bad orbifold which admits no covering by a manifold. If $n\neq1$, then $S^2(p,q,n)$ must be a spherical orbifold with universal cover the sphere. Moreover, N_S is a Seifert manifold with $\mathbb{O}(N_S)=S(p,q,n)$. As such it is irreducible. This yields a contradiction, as $S\subset N$ extends to a horizontal, hence incompressible, sphere in N_S. □

Lemma 11. *Let M be a graph manifold and let N be a Seifert piece with $\mathbb{O}(N)=F_g(p,\infty)$ or $\mathbb{O}(N)=F_g(p,\infty,\infty)$. If $S\cap N$ is horizontal, then $\chi(S\cap N)\le-4g+1$ or $\chi(S\cap N)\le-4g-p+1$, respectively.*

Proof. Suppose that S is a horizontal incompressible surface in N that covers regular points of $F_g(p,\infty)$ k times. Here $k\ge p\ge2$. By the Riemann–Hurwitz formula, $\chi(S)=k\big(-2g+\frac{1}{p}\big)\le p\big(-2g+\frac{1}{p}\big)=-2pg+1\le-4g+1$ or $\chi(S)=k\big(-2g-1+\frac{1}{p}\big)\le p\big(-2g-1+\frac{1}{p}\big)=-2pg-p+1\le-4g-p+1$, respectively. □

Lemma 12. *Let M be a graph manifold and let N be a Seifert piece with $\mathbb{O}(N)=F_g(p,\infty)$ or $\mathbb{O}(N)=F_g(p,\infty,\infty)$. Let $M=V\cup_S W$ be a Heegaard splitting and $S_1,F_1,\dots,F_{n-1},S_n$ an untelescoping. If $S_1,F_1,\dots,F_{n-1},S_n$ meets N in such a way that $F_i\cap N$ and $S_i\cap N$ are horizontal for each i, then*

$$\sum_i\big(\chi(F_{i-1}\cap N)-\chi(S_i\cap N)\big)\ge\begin{cases}8g-2 & \text{if } \mathbb{O}(N)=F_g(p,\infty),\\ 8g+2p-2 & \text{if } \mathbb{O}(N)=F_g(p,\infty,\infty).\end{cases}$$

Proof. The surfaces $S_1 \cap N$, $F_1 \cap N$, $\dots$, $F_{n-1} \cap N$, $S_n \cap N$ are disjoint and horizontal, hence they must be parallel. Let B be one of the components of ∂N. Consider the collection of torus knots $S_1 \cap B$, $F_1 \cap B$, $\dots$, $F_{n-1} \cap B$, $S_n \cap B$. Let γ be a torus knot on B that intersects each of the components in this collection of torus knots exactly once.

Now note that the untelescoping of the Heegaard splitting induces a Morse function of M and hence on B. If we assume that γ has as few critical points as possible, then near a maximum, γ meets two adjacent components of F_i or S_i for some i. If it meets two adjacent components of F_i, then the annulus cut out of B by these two components of F_i is inessential. But this contradicts Fact 5. Hence it meets two adjacent components of S_i. The same is true near a minimum of γ. Finally, if $\bigcup_i F_i$ meets N, then there are distinct adjacent components, as one such pair must lie above $\left(\bigcup_i F_i\right) \cap B$ and another below $\left(\bigcup_i F_i\right) \cap B$. Hence there are at least two more components of $\left(\bigcup_i S_i\right) \cap N$ than of $\left(\bigcup_i F_i\right) \cap N$. The lemma then follows from Lemma 11. □

Lemma 13. *Let N be a Seifert manifold with $\mathbb{O}(N) = D(p, q)$ with $(p, q) = 1$ and S be a properly embedded surface.*

(1) *If $S \cap N$ is horizontal, then there is an $l \geq 1$ such that $|S \cap N| = l$, $\chi(S \cap N) = lpq\left(-1 + \frac{1}{p} + \frac{1}{q}\right)$ and* genus$(S \cap N) \geq 1$.

(2) *If $S \cap N$ is pseudohorizontal, then $\chi(S \cap N) \leq -2\min(p, q) + 2$. Furthermore, either $S \cap N$ is as in Lemma 7, or* genus$(S \cap N) \geq 2$.

Proof. (1) Clearly $S \cap N$ is a finite sheeted cover of $D(p, q)$. The degree of this covering must be a positive multiple of pq, say lpq. It is clear that $S \cap N$ has l components. The second assertion follows from the Riemann–Hurwitz formula as $\chi(D(p, q)) = -1 + \frac{1}{q} + \frac{1}{q}$. The last assertion holds as by Lemma 10, S is nonplanar, so genus$(S \cap N) \geq 1$.

(2) Suppose first that $S \cap N$ is pseudohorizontal with respect to the fiber e. Let $N' = N - \eta(e)$ and S' be a component of $S \cap N'$. Recall that S' is horizontal by the definition of a pseudohorizonal surface.

If e is a regular fiber then S' must cover $A(p, q)$ at least pq times, that is, we have $\chi(S') \leq pq(-2 + \frac{1}{p} + \frac{1}{q}) = -2pq + p + q$ and therefore $\chi(S) = 2\chi(S') \leq -4pq + 2p + 2q \leq -2\min(p, q) + 2$. The remaining assertion follows from the proof of Lemma 7 which implies that S' cannot be planar.

Thus we can assume that e is an exceptional fiber. Suppose that e is the exceptional fiber of index q and let $N' = N - \eta(e)$. Suppose that H' is a horizontal incompressible surface in N' that covers regular points k times. Clearly $k \geq p$. Then $\chi(H') = k\left(-1 + \frac{1}{p}\right) \leq p\left(-1 + \frac{1}{p}\right) = -p + 1$. Thus if $S \cap N$ is pseudohorizontal

with respect to e, then

$$\chi(S \cap N) \leq 2\chi(H') \leq -2p + 2 \leq -2\min(p, q) + 2.$$

An analogous argument establishes this inequality in the case that e is the exceptional fiber of index p; the last comment follows immediately from Lemma 7. □

Lemma 14. *Let M be a graph manifold and let N be a vertex manifold. Let $M = V \cup_S W$ be a Heegaard splitting and $S_1, F_1, \ldots, F_{n-1}, S_n$ an untelescoping. Suppose that $F_i \cap N$ is vertical for each i and $S_i \cap N$ is vertical or pseudovertical for each i. Then*

$$\sum_i \big(\chi(F_{i-1} \cap N) - \chi(S_i \cap N)\big) \geq -2\chi(H) + 2s + \sum_i \big(|F_{i-1} \cap \partial N| - |S_i \cap \partial N|\big),$$

where H is the underlying surface of $\mathbb{O}(N)$ and s the number of exceptional fiberes.

Moreover, if If $\mathbb{O}(N) = F_g(p, \infty)$ and $\big(\bigcup_i S_i\big) \cap \partial N \neq \varnothing$, then

$$\sum_i \big(\chi(F_{i-1} \cap N) - \chi(S_i \cap N)\big) \geq 4g + 2 + \sum_i \big(|F_{i-1} \cap \partial N| - |S_i \cap \partial N|\big).$$

If $\mathbb{O}(N) = F_g(p, \infty, \infty)$, denote the components of ∂N by ∂N_1 and ∂N_2. If $\big(\bigcup_i S_i\big) \cap \partial N_j \neq \varnothing$ for $j = 1, 2$, then

$$\sum_i \big(\chi(F_{i-1} \cap N) - \chi(S_i \cap N)\big) \geq 4g + 4 + \sum_i \big(|F_{i-1} \cap \partial N| - |S_i \cap \partial N|\big).$$

Proof. We denote $\mathbb{O}(N)$ by F so long as we need not distinguish between the cases. Since $F_i \cap N$ is vertical, $F_i \cap N$ consists of saturated annuli and tori. Since $S_i \cap N$ is vertical or pseudovertical, $S_i \cap N$ is obtained from saturated annuli $A_1^i, \ldots, A_{n_i}^i$ and tori $T_1^i, \ldots, T_{k_i}^i$ (some of them parallel to components of $F_{i-1} \cap N$) by performing ambient 1-surgery along arcs $\beta_1^i, \ldots, \beta_{m_i}^i$ that project to disjoint imbedded arcs $b_1^i, \ldots, b_{m_i}^i$ disjoint from the projection of $A_1^i, \ldots, A_{n_i}^i$ and $T_1^i, \ldots, T_{k_i}^i$ except at their endpoints.

For the purposes of the computation in this lemma, we may amalgamate

$$\big(\textstyle\bigcup_i F_i \cup \bigcup_i S_i\big) \cap N.$$

Though it may not be possible to amalgamate $\bigcup_i F_i \cup \bigcup_i S_i$ without destroying its simultaneous structure on all vertex and edge manifolds, it is possible to perform an amalgamation without destroying the structure in a given vertex manifold. Said differently, a partial amalgamation in a given vertex manifold extends to a partial amalgamation in the graph manifold (though nothing can be said, for instance, about the structure of the resulting non strongly irreducible untelescoping of $M = V \cup_S W$ in edge manifolds adjacent to the given vertex manifold). Here the result of such an amalagamation with respect to N is a surface $\tilde{S}$ such that $\tilde{S} \cap N$

is pseudovertical. (For details on amalgamation involving vertical and pseudovertical surfaces see [Schultens 1993, Proposition 2.10], though note the difference in terminology.)

Since $\tilde{S} \cap N$ is pseudovertical, it is obtained from saturated annuli $A_1, \ldots, A_{\tilde{n}}$ and tori $T_1, \ldots, T_{\tilde{k}}$ by performing ambient 1-surgery along arcs $\beta_1, \ldots, \beta_{\tilde{m}}$ that project to disjoint imbedded arcs $b_1, \ldots, b_{\tilde{m}}$. These arcs are disjoint from the projections $a_1, \ldots, a_{\tilde{n}}$ of $A_1, \ldots, A_{\tilde{n}}$ and $t_1, \ldots, t_{\tilde{k}}$ of $T_1, \ldots, T_{\tilde{k}}$ except at their endpoints. Here each b_j corresponds either to b_l^i or to an arc dual to b_l^i for some l, i, and conversely. Furthermore,

$$-\chi(\tilde{S} \cap N) = 2\tilde{m} = 2\sum_i m_i = \sum_i \big(\chi(F_{i-1} \cap N) - \chi(S_i \cap N)\big)$$

and

$$|\tilde{S} \cap \partial N| = 2\tilde{n} = \sum_i \big(|S_i \cap \partial N| - |F_{i-1} \cap \partial N|\big).$$

Recall that $\tilde{S}$ cuts a submanifold of M that contains N into two compression bodies. Thus the (not necessarily connected) submanifolds into which $\tilde{S} \cap N$ cuts N can be analyzed from two perspectives: On the one hand, they result from cutting compression bodies along incompressible annuli. Recall that incompressible annuli are either essential or boundary parallel. Cutting a compression body along a boundary parallel annulus merely cuts off a solid torus. Cutting a compression body along an essential annulus yields either one or two compression bodies.

On the other hand, the submanifolds into which $\tilde{S} \cap N$ cuts N contain Seifert fibered submanifolds of N; specifically, the Seifert fibered submanifolds of N that project to the appropriate components of the complement of the graph $\Gamma = \big(\bigcup_j a_j\big) \cup \big(\bigcup_i t_i\big) \cup \big(\bigcup_l b_l\big) \cup \partial F$ in F. This is impossible unless the Seifert fibered spaces in question are fibered over a disk with at most one cone point (i.e., solid tori) or fibered over an annulus with no cone point. Each such solid torus or $(annulus) \times S^1$ must meet $\tilde{S}$. Furthermore, exactly one of the boundary components of any such $(annulus) \times S^1$ must lie in ∂N.

We denote the set of vertices of Γ by $V\Gamma$ and the set of edges by $E\Gamma$. We may assume that each vertex of Γ is either of valence two or of valence three. Each vertex on a circular component (corresponding either to a boundary component without attached b_i or to some t_i without attached b_i) is of valence two and each endpoint of an arc a_j and each endpoint of an arc b_l is a vertex of valence three. Then $\#V\Gamma = 2\tilde{n} + 2\tilde{m} + k$ and $\#E\Gamma = 3\tilde{n} + 3\tilde{m} + k$, where k is the number of circular components of Γ.

Denote the underlying surface of F by H. Now Γ induces a decomposition of H into 0-cells, 1-cells, 2-cells and annuli. Denote the union of the 2-cells and annuli by $D\Gamma$. Each such annulus must be cobounded by a component of ∂H. Let l be the number of annuli.

This implies that

$$\chi(H) = \#V\Gamma - (\#E\Gamma) + (\#D\Gamma - l).$$

Combining these insights we obtain

$$\begin{aligned}\sum_i \big(\chi(F_{i-1}\cap N) - \chi(S_i\cap N)\big) + \sum_i \big(|S_i\cap \partial N| - |F_{i-1}\cap \partial N|\big) &= 2\tilde{m} + 2\tilde{n}\\ &= -4\tilde{n} - 4\tilde{m} + 6\tilde{n} + 6\tilde{m} - 2(\#D\Gamma - l) + 2(\#D\Gamma - l)\\ &= -2\chi(H) + 2(\#D\Gamma - l).\end{aligned}$$

Thus $\sum_i \big(\chi(F_{i-1}\cap N) - \chi(S_i\cap N)\big)$ is at least

$$-2\chi(H) + 2(\#D\Gamma - l) + \sum_i \big(|F_{i-1}\cap \partial N| - |S_i\cap \partial N|\big),$$

that is,

$$\sum_i \big(\chi(F_{i-1}\cap N) - \chi(S_i\cap N)\big) \geq -2\chi(H) + 2s + \sum_i \big(|F_{i-1}\cap \partial N| - |S_i\cap \partial N|\big),$$

because every cone point must lie in a disk component. Now note that $\tilde{S}$ induces a bicoloring on the components of the complement of Γ in F according to which side of $\tilde{S}$ the Seifert fibered space that projects to that component lies. Thus $\#D\Gamma \geq 2$.

In the cases $F = F_g(p,\infty)$ or $F = F_g(p,\infty,\infty)$, $\#D\Gamma - l \geq 1$ because there must be a disk containing the cone point. Furthermore, if $l > 0$, then the result of cutting H along Γ yields annuli cobounded by boundary components of ∂H. This is impossible if $F = F_g(p,\infty)$ and $\big(\bigcup_i S_i\big)\cap \partial N \neq \varnothing$ or if $F = F_g(p,\infty,\infty)$ and $\big(\bigcup_i S_i\big)\cap \partial N_j \neq \varnothing$, for $j = 1, 2$, where N_1 and N_2 are the boundary components of N. Thus the additional formulas hold. □

Lemma 15. *Let M be a graph manifold and N a Seifert fibered submanifold with $\mathbb{O}(N) = D(p,q)$. Let $M = V\cup_S W$ be a Heegaard splitting and $S_1, F_1, \ldots, F_{n-1}, S_n$ an untelescoping. If $F_i\cap N$ is vertical for each i and $S_i\cap N$ is vertical or pseudovertical for each i, then*

$$\sum_i \big(\chi(F_{i-1}\cap N) - \chi(S_i\cap N)\big) \geq 2 + \sum_i \big(|F_{i-1}\cap \partial N| - |S_i\cap \partial N|\big).$$

Proof. This follows immediately from Lemma 14. □

5. The proof of Theorem 1

In order to give the proof of Theorem 1 we will first show that the fundamental groups can in fact be generated by $2g+1$ elements and then that only the manifolds listed admit a Heegaard splitting of genus $2g+1$.

Lemma 16. *The manifolds described in Theorem 1 have $2g+1$-generated fundamental groups.*

Proof. We first recall the presentations of the fundamental groups of N_1 and N_2: $\pi_1(N_1) = \langle a_1, b_1, \dots, a_g, b_g, s, t, f_1 \mid R \rangle$, with

$$R = \{[a_1, f_1], \dots, [a_g, f_1], [b_1, f_1], \dots, [b_g, f_1], [s, f_1], [t, f_1], \\ s^r = f_1^{\beta}, [a_1, b_1] \dots [a_g, b_g] s t = f_1^e\}$$

and $\pi_1(N_2) = \langle x, y, f_2 \mid [x, f_2], [y, f_2], x^p = f_2^{\beta_1}, y^q = f_2^{\beta_2} \rangle$.

The manifold M is obtained from N_1 and N_2 by identifying their boundaries, so it follows from van Kampen's theorem that

$$\pi_1(M) = \pi_1(N_1) *_C \pi_1(N_2) \text{ with } C \cong \mathbb{Z}^2.$$

Note that $f_1 = xyf_2^l$ for some $l \in \mathbb{Z}$ as we assume that the intersection number between f_1 and f_2 is 1. We will first establish a claim that is implicit in [Rost and Zieschang 1987].

Claim. *There exist $n = \min(p, q)$ conjugates of f_1 in $\pi_1(N_2)$ that generate a subgroup that maps surjectively onto the orbifold group $\pi_1(D(p, q))$.*

Proof. It suffices to show that n conjugates of xy generated the quotient group $\pi_1(D(p, q)) = \langle x, y | x^p, y^q \rangle$. We can assume that $n = q < p$. The assertion then follows as we can choose the conjugates to be xy, $yx = x^{-1}(xy)x, \dots, x^{-n+2}yx^{n-1} = x^{-(n-1)}(xy)x^{n-1}$ which implies that their product (in the same order) is $xy^n x^{n-1} = xx^{n-1} = x^n$. As n and p are coprime it follows that $\langle x^n \rangle = \langle x \rangle$ which implies that x lies in the subgroup generated by the n conjugates, it follows that also $y = x^{-1}xy$ lies in the subgroup, which proves the claim. □

In fact we need something stronger:

Claim. *We can choose elements $h_1, \dots, h_{n-1} \in \pi_1(N_2)$ such that*

$$U = \langle f_1, h_1 f_1 h_1^{-1}, \dots, h_{n-1} f_1 h_{n-1}^{-1} \rangle$$

maps surjectively onto the base group and that additionally $h_i \in U$ for $1 \le i \le n-1$.

Proof. Choose k_i such that $\langle f_1, k_1 f_1 k_1^{-1}, \dots, k_{n-1} f_1 k_{n-1}^{-1} \rangle$ maps surjectively. For any k_i choose $h_i \in U$ and $z_i \in \mathbb{Z}$ such that $k_i = h_i f_2^{z_i}$. Clearly such h_i and z_i exist as we assume that U maps surjectively one $\pi_1(D(p, q))$ and as the kernel is generated by f_2. Since f_1 and f_2 commute, we have $k_i f_1 k_i^{-1} = h_i f_2^{z_i} f_1 f_2^{-z_i} h_i^{-1} = h_i f_1 h_i^{-1}$. This implies that $U = \langle f_1, h_1 f_1 h_1^{-1}, \dots, h_{n-1} f_1 h_{n-1}^{-1} \rangle$, and the claim follows. □

Note that U is a subgroup of finite index in $\pi_1(N_2)$ and that we can choose the elements h_i such that $\pi_1(N_2) = U$ if and only if N_2 is the exterior of a torus knot with meridian f_1. It is however always true that $\pi_1(N_2) = \langle U, C \rangle$ as $f_2 \in C$.

Note further that the subgroup $\langle s, f_1\rangle$ of $\pi_1(N_1)$ is generated by a single element g_0 which corresponds to the core of the solid torus corresponding to the exceptional fiber of N_1. It follows that $g_0^k = f_1$ for some $k \in \mathbb{Z}$. In order to prove the lemma we describe elements $g_1, \dots, g_{2g} \in \pi_1(M)$ such that $\pi_1(M) = \langle g_0, \dots, g_{2g}\rangle$.

Recall that by assumption $n \le 2g+1$. Put $h_i = 1$ for $n \le i \le 2g$, and define

$$g_i := \begin{cases} h_i a_i & \text{for } 1 \le i \le g, \\ h_i b_{i-g} & \text{for } g+1 \le i \le 2g. \end{cases}$$

Claim. $U \subset \langle g_0, \dots, g_{2g}\rangle$.

Proof. It suffices to show that f_1 and the elements $h_i f_1 h_i^{-1}$ lie in $\langle g_0, \dots, g_{2g}\rangle$ for $1 \le i \le 2g$. Clearly $f_1 \in \langle g_0, \dots, g_{2g}\rangle$ as $f_1 = g_0^k$. Furthermore $h_i f_1 h_i^{-1} \in \langle g_0, \dots, g_{2g}\rangle$ for $1 \le i \le g$ as $g_i g_0^k g_i^{-1} = h_i a_i f_1 a_i^{-1} h_i^{-1} = h_i f_1 h_i^{-1}$. The same argument shows that $g_i g_0^k g_i^{-1} = h_i f_1 h_i^{-1}$ for $g+1 \le i \le 2g$, proving the claim. □

Since $h_i \in U$ for $1 \le i \le 2g$, this implies that $h_i \in \langle g_0, \dots, g_{2g}\rangle$ for $1 \le i \le 2g$ and therefore $h_i^{-1} g_i \in \langle g_0, \dots, g_{2g}\rangle$ for $1 \le i \le 2g$. Since $h_i^{-1} g_i = a_i$ for $1 \le i \le g$ and $h_i^{-1} g_i = b_{i-g}$ for $g+1 \le i \le 2g$, it follows that all a_i and b_i lie in $\langle g_0, \dots, g_{2g}\rangle$. Furthermore both f_1 and s are powers of g_0 and lie in $\langle g_0, \dots, g_{2g}\rangle$. The last generator t can be written as a product in the remaining generators by the last relation. Thus all generators of $\pi_1(N_1)$ lie in $\langle g_0, \dots, g_{2g}\rangle$ which shows that $\pi_1(N_1) \subset \langle g_0, \dots, g_{2g}\rangle$. Thus $C \subset \langle g_0, \dots, g_{2g}\rangle$ and therefore $\pi_1(N_2) = \langle U, C\rangle \subset \langle g_0, \dots, g_{2g}\rangle$. This shows that $\pi_1(M) = \langle g_0, \dots, g_{2g}\rangle$, proving Lemma 16. □

Lemma 17. *Let M be a manifold as described in Theorem 1 and let $M = V \cup_S W$ be a Heegaard splitting. Then one of the following holds*:

(1) *$S \cap N_1$ is vertical, $S \cap N_2$ is planar and pseudohorizontal with respect to the exceptional fiber e of index p and $q \le 2g+1$.*

(2) *$S \cap N_1$ is as in Lemma 8, $S \cap N_2$ consists of a single annulus and* genus $S = 2g+1$.

(3) genus $S \ge 2g+2$.

Proof. Let M be a manifold as described in Theorem 1 and let $M = V \cup_S W$ be a Heegaard splitting. Furthermore, let $S_1, F_1, \dots, F_{n-1}, S_n$ be a strongly irreducible untelescoping of $M = V \cup_S W$ that is standard.

Case 1: $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_1$ and $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_2$ are vertical or pseudovertical. If $\left(\bigcup_i F_i \cup \bigcup_i S_i\right)$ meets the edge manifold N_e between N_1 and N_2 in annuli including spanning annuli, then M must be a Seifert fibered space. But this contradicts our assumption that the fibers of N_1 and N_2 have intersection number 1.

If $\bigcup_i F_i$ meets the edge manifold N_e in a torus, then we may assume that $\bigcup_i S_i$ is disjoint from N_e. (Annuli that are boundary parallel in N_e can be isotoped into

the vertex manifolds.) Then Lemma 14 tells us that

$$\sum_i \big(\chi(F_{i-1}\cap N_1)-\chi(S_i\cap N_1)\big)\geq 4g+2+\sum_i\big(|F_{i-1}\cap\partial N_1|-|S_i\cap\partial N_1|\big)\geq 4g$$

and Lemma 15 tells us that

$$\sum_i \big(\chi(F_{i-1}\cap N_2)-\chi(S_i\cap N_2)\big)\geq 2+\sum_i\big(|F_{i-1}\cap\partial N_2|-|S_i\cap\partial N_2|\big)=2;$$

hence by Theorem 3, 2 genus $S-2=-\chi(S)\geq 4g+2$; thus genus $S\geq 2g+2$.

Otherwise $\big(\bigcup_i F_i\big)\cup\big(\bigcup_i S_i\big)$ meets the edge manifold between N_1 and N_2 in boundary parallel annuli and one component of Euler characteristic -2 contained in $\big(\bigcup_i S_i\big)\cap N_e$. Any boundary parallel annuli in $\big(\bigcup_i F_i\cup\bigcup_i S_i\big)\cap N_e$ can be isotoped into N_1 or N_2. It then follows from Lemmas 14 and 15 that

$$\begin{aligned}&\sum_i\big(\chi(F_{i-1})-\chi(S_i)\big)\\&\quad=\sum_i\big(\big(\chi(F_{i-1}\cap N_1)-\chi(S_i\cap N_1)\big)\\&\qquad+\sum_i\big(\chi(F_{i-1}\cap N_2)-\chi(S_i\cap N_2)\big)+\sum_i\big(\chi(F_{i-1}\cap N_e)-\chi(S_i\cap N_e)\big)\\&\quad\geq(4g+2-2)+(2-2)+2=4g+2.\end{aligned}$$

Hence, by Theorem 3, 2 genus $S-2=-\chi(S)\geq 4g+2$, whence genus $S\geq 2g+2$.

Case 2: $\big(\bigcup_i F_i\cup\bigcup_i S_i\big)\cap N_1$ *is horizontal.* Recall Fact 1 following Theorem 4. It tells us that $\sum_i\big(\chi(F_{i-1}\cap N)-\chi(S_i\cap N)\big)\geq 0$ for any vertex or edge manifold N. It follows that

$$\sum_i\big(\chi(F_{i-1})-\chi(S_i)\big)\geq\sum_i\big(\chi(F_{i-1}\cap N_1)-\chi(S_i\cap N_1)\big).$$

By Lemma 12, $\sum_i\big(\chi(F_{i-1}\cap N_1)-\chi(S_i\cap N_1)\big)\geq 8g-2$, so

$$\sum_i\big(\chi(F_{i-1})-\chi(S_i)\big)\geq 8g-2.$$

Hence, by Theorem 3, we have 2 genus $S-2=-\chi(S)\geq 8g-2$, that is,

$$\text{genus } S\geq 4g\geq 2g+2.$$

Case 3: A component of $\big(\bigcup_i S_i\big)\cap N_1$ *is pseudohorizontal.* Denote the pseudohorizontal component of $\big(\bigcup_i S_i\big)\cap N_1$ by $\tilde{S}$. By Lemma 8, either $\tilde{S}$ is as in that lemma and $\big(\bigcup_i S_i\big)\cap N_2$ consists of a single annulus, or genus $S\geq 2g+2$. This puts us in situation (2) or (3), respectively.

Case 4: $\big(\bigcup_i F_i\cup\bigcup_i S_i\big)\cap N_2$ *is horizontal.* If $\big(\bigcup_i F_i\cup\bigcup_i S_i\big)\cap N_1$ is horizontal, then the result follows by Case 2. If a component of $\big(\bigcup_i S_i\big)\cap N_1$ is

pseudohorizontal, then the result follows by Case 3. Thus we may assume that $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_1$ is vertical or pseudovertical.

We may assume that any boundary parallel annuli in the edge manifold N_e that are parallel into N_1 have been isotoped into N_1. (This does not change the Euler characteristics of the surfaces nor the fact that $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_1$ is vertical or pseudovertical.)

Fact 3 tells us that $\left(\bigcup_i F_i\right) \cap N_e$ does not contain a torus. Hence $\bigcup_i S_i \cap \partial N_1 \neq \varnothing$. It also follows from Fact 5 that

$$\left|\left(\bigcup_i F_i\right) \cap \partial N_1\right| = \left|\left(\bigcup_i F_i\right) \cap \partial N_2\right|.$$

Since there are no annuli in $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_e$ that are parallel into N_1, we obtain $\sum_i |S_i \cap \partial N_1| \leq \sum_i |S_i \cap \partial N_2|$, and hence

$$\sum_i \left(|F_{i-1} \cap \partial N_1| - |S_i \cap \partial N_1|\right) \geq \sum_i \left(|F_{i-1} \cap \partial N_2| - |S_i \cap \partial N_2|\right).$$

The components of $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_2$ are all parallel. If H is such a component, then

$$\chi(H) = 2 - 2\,\text{genus}\, H - |H \cap \partial N_2|.$$

Recall that by Lemma 10, genus $H \geq 1$. Thus

$$\begin{aligned}
&\sum_i \left(\chi(F_{i-1} \cap N_2) - \chi(S_i \cap N_2)\right) \\
&\quad = (2\,\text{genus}\, H - 2) \sum_i \left(|S_i \cap N_2| - |F_{i-1} \cap N_2|\right) - \sum_i \left(|F_{i-1} \cap \partial N_2| - |S_i \cap \partial N_2|\right) \\
&\quad \geq - \sum_i \left(|F_{i-1} \cap \partial N_2| - |S_i \cap \partial N_2|\right).
\end{aligned}$$

Now

$$\begin{aligned}
&\sum_i \left(\chi(F_{i-1}) - \chi(S_i)\right) \\
&\qquad = \sum_i \left(\chi(F_{i-1} \cap N_1) - \chi(S_i \cap N_1)\right) + \sum_i \left(\chi(F_{i-1} \cap N_e) - \chi(S_i \cap N_e)\right) \\
&\qquad\qquad + \sum_i \left(\chi(F_{i-1} \cap N_2) - \chi(S_i \cap N_2)\right).
\end{aligned}$$

Then Fact 1 tells us that $\sum_i \left(\chi(F_{i-1} \cap N_e) - \chi(S_i \cap N_e)\right) \geq 0$, so

$$\begin{aligned}
&\sum_i \left(\chi(F_{i-1}) - \chi(S_i)\right) \\
&\qquad\qquad \geq \sum_i \left(\chi(F_{i-1} \cap N_1) - \chi(S_i \cap N_1) + \chi(F_{i-1} \cap N_2) - \chi(S_i \cap N_2)\right).
\end{aligned}$$

Thus, by Lemma 14,

$$\begin{aligned}\sum_i \big(\chi(F_{i-1}) - \chi(S_i)\big) &\geq 4g+2+\sum_i \big(|F_{i-1}\cap \partial N_1| - |S_i \cap \partial N_1|\big) - \sum_i \big(|F_{i-1}\cap \partial N_2| - |S_i\cap \partial N_2|\big)\\ &\geq 4g+2.\end{aligned}$$

By Theorem 3, therefore, we conclude that $2\,\text{genus}\, S - 2 = -\chi(S) \geq 4g+2$, whence $\text{genus}\, S \geq 2g+2$.

Case 5: A component of $\big(\bigcup_i S_i\big)\cap N_2$ *is pseudohorizontal.* Here, too, if

$$\big(\textstyle\bigcup_i F_i \cup \bigcup_i S_i\big)\cap N_1$$

is horizontal, the result follows by Case 2. If a component of $\big(\bigcup_i S_i\big)\cap N_1$ is pseudohorizontal, it follows by Case 3. Thus we assume that $\big(\bigcup_i F_i \cup \bigcup_i S_i\big)\cap N_1$ is vertical or pseudovertical.

By the same reasoning as in Case 4, we can assume that

$$\sum_i \big(|F_{i-1}\cap \partial N_1| - |S_i\cap \partial N_1|\big) \geq \sum_i \big(|F_{i-1}\cap \partial N_2| - |S_i\cap \partial N_2|\big).$$

Denote the pseudohorizontal component of $\big(\bigcup_i S_i\big)\cap N_2$ by $\tilde{S}$ and note that here $\big(\big(\bigcup_i F_i \cup \bigcup_i S_i\big) - \tilde{S}\big)\cap N_2 = \varnothing$. Thus

$$\chi(\tilde{S}) = 2 - 2\,\text{genus}\,\tilde{S} - |\tilde{S}\cap \partial N_2|.$$

By Lemma 13, either $\tilde{S}$ is as in Lemma 7 or $\text{genus}\,\tilde{S}\geq 2$. In the former case, we have $|\partial \tilde{S}| = 2q$ and $\chi(\tilde{S}) = 2-2q$. If, moreover, $\text{genus}\, S \leq 2g+1$, then $-4g \leq \chi(S) = \sum_i \big(\chi(S_i) - \chi(F_{i-1})\big) \leq \chi(\tilde{S}) = 2-2q$. Thus $q \leq 2g+1$.

In the second case ($\text{genus}\,\tilde{S}\geq 2$), we have

$$\sum_i \big(\chi(F_{i-1}\cap N_2) - \chi(S_i\cap N_2)\big) = -\chi(\tilde{S}) = 2\,\text{genus}\,\tilde{S} - 2 + |\tilde{S}\cap \partial N_2| \geq |\tilde{S}\cap \partial N_2|.$$

Arguing as in Case 4, we obtain

$$\begin{aligned}\sum_i \big(\chi(F_{i-1}) - \chi(S_i)\big) &\geq \sum_i \big(\chi(F_{i-1}\cap N_1) - \chi(S_i\cap N_1) + \chi(F_{i-1}\cap N_2) - \chi(S_i\cap N_2)\big)\\ &\geq 4g+2+\sum_i \big(|F_{i-1}\cap \partial N_1| - |S_i\cap \partial N_1|\big) - \sum_i \big(|F_{i-1}\cap \partial N_2| - |S_i\cap \partial N_2|\big)\\ &\geq 4g+2.\end{aligned}$$

Again, by Theorem 3, we have $2\,\text{genus}\, S - 2 = -\chi(S) \geq 4g+2$, whence $\text{genus}\, S \geq 2g+2$. □

Proof of Theorem 1. Consider the options allowed by Lemma 17. If option (1) occurs, then Lemma 7 implies that N_2 is a q-bridge knot complement and the fiber of N_1 is identified with the meridian of N_2. This puts us in case (1) of Theorem 1. If option (2) occurs in Lemma 17, then $\hat{N}_1$ admits a horizontal Heegaard splitting of genus $2g$ by Lemma 8 and we are in case (2) of Theorem 1. If option (3) occurs there is nothing to show. □

6. The proof of Theorem 2

In this section we construct for any $n \in \mathbb{N}$ such that $n \geq 3$ a graph manifold M_n such that $\pi_1(M_n)$ is $3n$-generated but that the Heegaard genus of M_n is $4n$. We denote the graph underlying M_n by Γ_n. Γ_n is a tree on $2n+1$ vertices $z, c_1, \ldots, c_n, d_1, \ldots, d_n$ and $2n$ edges $e_1, \ldots, e_n, f_1, \ldots, f_n$ such that c_i and d_i are the endpoints of e_i and that d_i and z are the endpoints of f_i.

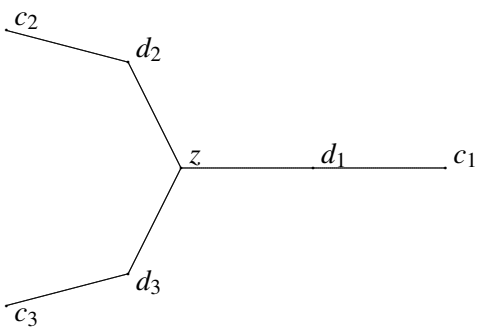


Figure 3. The tree Γ_3.

The closed graph manifold M_n is then constructed as follows, where we denote the Seifert piece corresponding to a vertex v by N_v.

(1) The intersection number between the fibers of the adjacent Seifert spaces is 1 at any torus of the JSJ decomposition.

(2) $\mathbb{O}(N_z)$ is a n-punctured sphere with one cone point of order $20n$ and $\hat{N}_z = S^3$.

(3) $\mathbb{O}(N_{d_i}) = T^2(\infty, \infty, 20n)$ and N_{d_i} admits no pseudohorizontal surface that has genus 2.

(4) $\mathbb{O}(N_{c_i})$ is of type $D(3, q)$ with $q \geq 20n$ and $(3, q) = 1$ but N_{c_i} is not homeomorphic to the exterior of a 2-bridge knot in S^3.

Remark 18. Note that (2) is equivalent to stating that N_z is the exterior of a Seifert fibered n component n-bridge link in S^3, in particular $\pi_1(N_z)$ is generated by the fibers of the N_{d_i}. The existence of the spaces N_{d_i} satisfying (3) is an immediate consequence of Lemma 9.

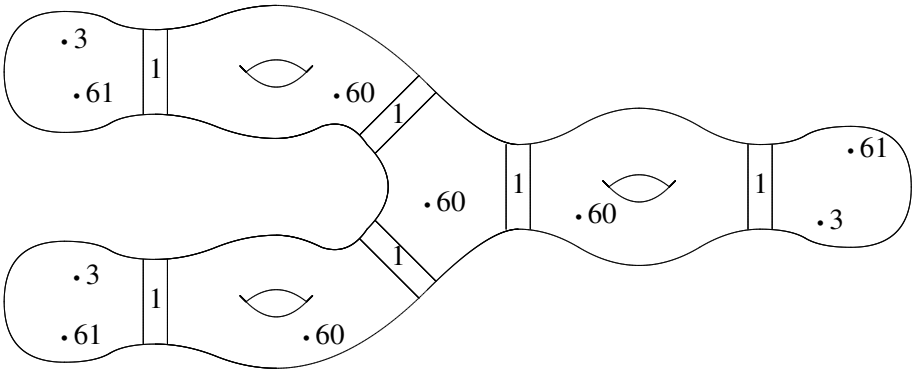


Figure 4. A graph manifold M with $g(M) = 12$ and $r(M) \leq 9$.

The first part of the proof of Theorem 2 is again a simple calculation:

Lemma 19. *$\pi_1(M_n)$ can be generated by $3n$ elements.*

Proof. The proof is almost identical to the proof of Lemma 16 and we frequently omit explicit calculations if they are identical. Recall that

$$\pi_1(N_{d_i}) = \langle a_i, b_i, s_i, t_{i1}, t_{i2}, f_i \mid R_i \rangle \text{ with}$$

$$R_i = \{[a_i, f_i], [b_i, f_i], [s_i, f_i], [t_{i1}, f_i], [t_{i2}, f_i], s_i^{5n} = f_i^{\beta_i}, [a_i b_i] t_{i1} t_{i2} s_i = f_i^{e_i}\}$$

where t_{i1} corresponds the the boundary component between N_{d_i} and N_z and t_{i2} corresponds to the boundary component between N_{d_i} and N_{c_i}.

Recall from the proof of Lemma 16 that there exist elements $h_{i1}, h_{i2} \in \pi_1(N_{c_i})$ such that $U_i = \langle f_i, h_{i1} f_i h_{i1}^{-1}, h_{i2} f_i h_{i2}^{-1} \rangle$ is a subgroup of finite index in $\pi_1(N_{c_i})$ that maps surjectively onto the fundamental group of $\mathbb{O}(N_{c_i})$ and that $h_{i1}, h_{i2} \in U_i$.

We will show that $\pi_1(M_n)$ is generated by the generators $g_1, \ldots, g_{3n}$ defined as follows:

(1) g_i is the generator of the cyclic group $\langle f_i, s_i \rangle$ for $1 \leq i \leq n$.

(2) $g_{n+i} = h_{i1} a_i$ for $1 \leq i \leq n$.

(3) $g_{2n+i} = h_{i2} b_i$ for $1 \leq i \leq n$.

Let $H = \langle g_1, \ldots, g_{3n} \rangle$. We show that $H = \pi_1(M_n)$.

Note first that $\pi_1(N_z) \subset H$ as $g_i \in H$ implies $f_i \in H$ for $1 \leq i \leq n$ and $\pi_1(N_z)$ is generated by the f_i. This implies that $t_{i1} \in H$ for $1 \leq i \leq n$.

The same calculation as in the proof of Lemma 16 further shows that $U_i \subset H$ for $1 \leq i \leq n$. It follows that $a_i, b_i \in H$ for $1 \leq i \leq n$. Thus $\pi_1(N_{d_i}) \subset H$ as $\pi_1(N_{d_i})$ is generated by $a_i, b_i, s_i, f_i, t_{i_1}$ and s_i and f_i are powers of g_i.

It follows further that $\pi_1(N_{c_i}) \subset H$ as $\pi_1(N_{c_i})$ is generated by U_i and C_i, where $C_i = \pi_1(N_{c_i}) \cap \pi_1(N_{d_i})$. □

To conclude the proof of Theorem 2 it clearly suffices to establish the following:

Proposition 20. *The Heegaard genus of M_n is at least $4n$.*

In proving this we will tacitly use that small genus Heegaard splittings have very special untelescopings — a fact that deserves its own result since it has independent interest:

Lemma 21. *Let $M_n = V \cup_S W$ be a Heegaard splitting of M_n. Then either $g(S) \geq 4n$ or there is a strongly irreducible untelescoping $S_1, F_1, \ldots, F_{k-1}, S_k$ of $M_n = V \cup_S W$ such that for any vertex manifold N no component of $S_i \cap N$ or $F_i \cap N$ is horizontal. In particular all F_i are vertical incompressible tori.*

Proof. Suppose that some component F of $S_i \cap N$ or $F_i \cap N$ is horizontal for some i and some vertex manifold N. Note first that no component of ∂F bounds a disk as any component is an essential curve in an incompressible torus. It follows that $\chi(F) \geq \chi(F_i)$ (or $\chi(F) \geq \chi(S_i)$), where F_i (or S_i) is the surface containing F.

Note first that $F \cap N$ is a covering of the base space $\mathbb{O}$ of N of degree at least $20n$. It is furthermore easy to see that we have $\chi(\mathbb{O}) \leq -\frac{1}{2}$ for any choice of N. If follows that $\chi(F \cap N) \leq -10n$ and therefore $\chi(F_i) \leq -10n$ (or $\chi(S_i) \leq -10n$). This however implies that the genus of F_i (or S_i) is greater than $5n$ which implies that the Heegaard surface S is of genus at least $5n$. This proves the assertion. □

Proof of Proposition 20. To see that M_n admits no Heegaard splitting of genus less than $4n$, proceed along the same lines as in the proof of Lemma 17. Let $M_n = V \cup_S W$ be a Heegaard splitting and let $S_1, F_1, \ldots, F_{k-1}, S_k$ be a strongly irreducible untelescoping of $M_n = V \cup_S W$. We consider the various possible cases for $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_{c_j}$ and $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_{d_j}$.

Case 1: Fix j and suppose that $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_{c_j}$ and $\left(\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_{d_j}\right)$ are vertical or pseudovertical. In this case it is impossible for $\left(\left(\bigcup_i F_i \cup \bigcup_i S_i\right)\right)$ to meet the edge manifold N_{e_j} between N_{c_j} and N_{d_j} in spanning annuli. Moreover, any boundary parallel annuli in N_{e_j} can be isotoped into N_{c_j} and N_{d_j} and any boundary parallel annuli in N_{g_j} that are parallel into N_{d_j} can be isotoped into N_{d_j}. (This does not change the fact that $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_{c_j}$ and $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_{d_j}$ are vertical or pseudovertical and serves to facilitate our counting argument.) In conjunction with Fact 5, this tells us that

$$\sum_i \left(-|S_i \cap \partial N_{d_j}| + |F_{i-1} \cap \partial N_{d_j}|\right) \geq \sum_i \left(-|S_i \cap \partial N_{c_j}| + |F_{i-1} \cap \partial N_{c_j}| - |S_i \cap \partial N_z^j| + |F_{i-1} \cap \partial N_z^j|\right),$$

where ∂N_z^j is the component of ∂N_z that meets the edge manifold N_{g_j} between N_z and N_{d_j}.

Now either $\bigcup_i F_i$ meets N_{e_j} in an essential torus, or $\bigcup_i S_i$ meets N_{e_j} in the only other possible configuration. In the first case, we obtain

$$\sum_i \big(-|S_i \cap \partial N_{c_j}| + |F_{i-1} \cap \partial N_{c_j}|\big) = 0$$

and

$$\sum_i \big(-|S_i \cap \partial N_{d_j}| + |F_{i-1} \cap \partial N_{d_j}|\big) \geq \sum_i \big(-|S_i \cap \partial N_z^j| + |F_{i-1} \cap \partial N_z^j|\big).$$

In the second case we obtain $\sum_i \big(-|S_i \cap \partial N_{c_j}| + |F_{i-1} \cap \partial N_{c_j}|\big) = -2$ and

$$\sum_i \big(-|S_i \cap \partial N_{d_j}| + |F_{i-1} \cap \partial N_{d_j}|\big) \geq -2 + \sum_i \big(-|S_i \cap \partial N_z^j| + |F_{i-1} \cap \partial N_z^j|\big).$$

We further distinguish the cases in which $\bigcup_i F_i$ meets or does not meet the edge manifold N_{f_j} in an essential torus.

Case 1.1: $\bigcup_i F_i$ meets N_{e_j} in an essential torus. By Lemmas 14 and 15, we have

$$\begin{aligned}
&\sum_i \big(\chi(F_{i-1} \cap N_{d_j}) - \chi(S_i \cap N_{d_j}) + \chi(F_{i-1} \cap N_{c_j}) - \chi(S_i \cap N_{c_j})\big) \\
&\quad \geq 4 + 2 + \sum_i \big(|F_{i-1} \cap \partial N_{d_j}| - |S_i \cap \partial N_{d_j}|\big) + 2 + \sum_i \big(|F_{i-1} \cap \partial N_{c_j}| - |S_i \cap \partial N_{c_j}|\big) \\
&\quad \geq 4 + 2 + \sum_i \big(|F_{i-1} \cap \partial N_z^j| - |S_i \cap \partial N_z^j|\big) + 2 \\
&\quad \geq 8 + \sum_i \big(|F_{i-1} \cap \partial N_z^j| - |S_i \cap \partial N_z^j|\big).
\end{aligned}$$

Case 1.2: $\bigcup_i F_i$ meets neither N_{f_j} nor N_{e_j} in an essential torus. By Lemmas 14 and 15, we have

$$\begin{aligned}
&\sum_i \big(\chi(F_{i-1} \cap N_{d_j}) - \chi(S_i \cap N_{d_j}) + \chi(F_{i-1} \cap N_{c_j}) - \chi(S_i \cap N_{c_j}) \\
&\qquad + \chi(F_{i-1} \cap N_{e_j}) - \chi(S_i \cap N_{e_j})\big) \\
&\geq 4 + 4 + \sum_i \big(|F_{i-1} \cap \partial N_{d_j}| - |S_i \cap \partial N_{d_j}|\big) + 2 + \sum_i \big(|F_{i-1} \cap \partial N_{c_j}| - |S_i \cap \partial N_{c_j}|\big) + 2 \\
&\geq 4 + 4 - 2 + \sum_i \big(|F_{i-1} \cap \partial N_z^j| - |S_i \cap \partial N_z^j|\big) + 2 - 2 + 2 \\
&\geq 8 + \sum_i \big(|F_{i-1} \cap \partial N_z^j| - |S_i \cap \partial N_z^j|\big).
\end{aligned}$$

Case 1.3: $\bigcup_i F_i$ meets N_{f_j} in an essential torus but does not meet N_{e_j} in an essential torus. Here Lemmas 14 and 15 yield only

$$\begin{aligned}
&\sum_i \big(\chi(F_{i-1} \cap N_{d_j}) - \chi(S_i \cap N_{d_j}) + \chi(F_{i-1} \cap N_{c_j}) - \chi(S_i \cap N_{c_j}) \\
&\qquad + \chi(F_{i-1} \cap N_{e_j}) - \chi(S_i \cap N_{e_j})\big) \\
&\geq 4 + 2 + \sum_i \big(|F_{i-1} \cap \partial N_{d_j}| - |S_i \cap \partial N_{d_j}|\big) + 2 + \sum_i \big(|F_{i-1} \cap \partial N_{c_j}| - |S_i \cap \partial N_{c_j}|\big) + 2 \\
&\geq 4 + 2 - 2 + 2 - 2 + 2 \geq 6.
\end{aligned}$$

In this case $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_z$ must be vertical or pseudovertical. (See Fact 3 above.)

Note that in all cases we have

$$\sum_i \left(\chi(F_{i-1} \cap N_{c_j}) - \chi(S_i \cap N_{c_j}) + \chi(F_{i-1} \cap N_{e_j}) - \chi(S_i \cap N_{e_j})\right) \geq 2.$$

Case 2: Fix j and suppose a component of $\bigcup_i S_i \cap N_{d_j}$ is pseudohorizontal. As we have seen, in this case

$$S' = \left(\bigcup_i S_i \cup \bigcup_i F_i\right) \cap N_{d_j}$$

is connected. In particular, $\bigcup_i F_i \cap N_{d_j} = \varnothing$. By construction, the genus of a pseudohorizontal surface is even. Recall our assumption that N_{d_j} admits no pseudohorizontal surface of genus 2. Thus the genus of S' is at least 4. Hence

$$\begin{aligned}\sum_i \left(\chi(F_{i-1} \cap N_{d_j}) - \chi(S_i \cap N_{d_j})\right) &= 0 - \chi(S' \cap N_{d_j}) \\ &\geq 6 + b = 6 - \sum_i \left(|F_{i-1} \cap \partial N_{d_j}| - |S_i \cap \partial N_{d_j}|\right),\end{aligned}$$

where b is the number of boundary components of S'. Since S' is a separating surface, it meets each boundary component of N_{d_j} at least twice. Consequently, $b \geq 4$. It thus follows from Fact 1 that

$$\begin{aligned}\sum_i \big(&\chi(F_{i-1} \cap N_{f_j}) - \chi(S_i \cap N_{f_j}) + \chi(F_{i-1} \cap N_{d_j}) - \chi(S_i \cap N_{d_j}) \\ &+ \chi(F_{i-1} \cap N_{e_j}) - \chi(S_i \cap N_{e_j}) + \chi(F_{i-1} \cap N_{c_j}) - \chi(S_i \cap N_{c_j})\big) \\ &\geq 6 + 4 = 10.\end{aligned}$$

Case 3: Fix j and suppose a component of $\left(\bigcup_i S_i\right) \cap N_{c_j}$ is pseudohorizontal. It will suffice to consider the case in which $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_{d_j}$ is vertical or pseudovertical. Denote the pseudohorizontal component of $\left(\bigcup_i S_i\right) \cap N_{c_j}$ by $\tilde{S}$ and note that here $\left(\left(\bigcup_i F_i \cup \bigcup_i S_i\right) - \tilde{S}\right) \cap N_{c_j} = \varnothing$. By assumption, N_{c_j} is not the exterior of a 2-bridge knot in $\mathbb{S}^3$, thus by Lemmas 7 and 13, genus $\tilde{S} \geq 2$. Hence,

$$\chi(F_{i-1} \cap N_{c_j}) - \chi(S_i \cap N_{c_j})) = -\chi(\tilde{S} \cap N_{c_j}) \geq -\chi(\tilde{S}) \geq 2 + c,$$

where c is the number of boundary components of $\tilde{S}$.

Recall that when $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_{d_j}$ is vertical or pseudovertical, we may isotope any annuli in $\left(\bigcup_i S_i\right) \cap N_{f_i}$ or $\left(\bigcup_i S_i\right) \cap N_{e_i}$ that are parallel into N_{d_j} into N_{d_j} without altering the fact that $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_{d_j}$ is vertical or pseudovertical. Therefore we may assume that there are no annuli in $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_{f_j}$ or $\left(\bigcup_i F_i \cup \bigcup_i S_i\right) \cap N_{e_j}$ that are parallel into N_{d_j}. Thus

$$\begin{aligned}\sum_i \big(|F_{i-1}\cap\partial N_{d_j}| - |S_i\cap\partial N_{d_j}|\big)&\\ \geq \sum_i \big(|F_{i-1}\cap\partial N_z^j| - |S_i\cap\partial N_z^j|\big) + \sum_i \big(|F_{i-1}\cap\partial N_{c_j}| - |S_i\cap\partial N_{c_j}|\big)&\\ = \sum_i \big(|F_{i-1}\cap\partial N_z^j| - |S_i\cap\partial N_z^j|\big) - c.&\end{aligned}$$

Hence, by Lemma 14,

$$\begin{aligned}\sum_i \big(\chi(F_{i-1}\cap N_{d_j}) - \chi(S_i\cap N_{d_j}) + \chi(F_{i-1}\cap N_{c_j}) - \chi(S_i\cap N_{c_j})\big)&\\ \geq 4+4+\sum_i \big(|F_{i-1}\cap\partial N_{d_j}| - |S_i\cap\partial N_{d_j}|\big) + 2 + c&\\ \geq 10 + \sum_i \big(|F_{i-1}\cap\partial N_z^j| - |S_i\cap\partial N_z^j|\big).&\end{aligned}$$

Putting these computations together we must consider the various options for $\big(\bigcup_i F_i \cup \bigcup_i S_i\big)\cap N_z$:

Case A: $\big(\bigcup_i F_i \cup \bigcup_i S_i\big)\cap N_z$ *is vertical and pseudovertical.* In this case the options for $\big(\bigcup_i F_i \cup \bigcup_i S_i\big)\cap N_{f_j}$ are severely limited. If $\big(\bigcup_i F_i \cup \bigcup_i S_i\big)\cap N_{d_j}$ is vertical and pseudovertical, then $\big(\bigcup_i F_i \cup \bigcup_i S_i\big)\cap N_{f_j}$ cannot consist of spanning annuli. So either $\bigcup_i F_i$ meets N_{f_j} in an essential torus, or $\bigcup_i S_i$ meets N_{f_j} in the only other possible configuration. If $\big(\bigcup_i F_i \cup \bigcup_i S_i\big)\cap N_{d_j}$ is pseudohorizontal, then N_{f_j} cannot meet a toral component of $\bigcup_i F_i$. So it must consist either of spanning annuli or the only other possible configuration.

Define

$$J_0 = \Big\{j : \sum_i \big(|F_{i-1}\cap\partial N_z^j| - |S_i\cap\partial N_z^j|\big) = 0\Big\}.$$

Then $\sum_i \big(\chi(F_{i-1}\cap N_{f_j}) - \chi(S_i\cap N_{f_j})\big) = 0$ for $j\in J_0$.

Denote by J_1 the set of j not in J_0 such that $\big(\bigcup_i F_i\big)\cup\big(\bigcup_i S_i\big)\cap N_{d_j}$ are vertical or pseudovertical. Then

$$\sum_i \big(|F_{i-1}\cap\partial N_z^j| - |S_i\cap\partial N_z^j|\big) = -2$$

and

$$\sum_i \big(\chi(F_{i-1}\cap N_{f_j}) - \chi(S_i\cap N_{f_j})\big) = 2 \quad \text{for } j\in J_1.$$

Denote by J_2 the set of j such that $\big(\bigcup_i F_i \cup \bigcup_i S_i\big)\cap N_{d_j}$ is pseudohorizontal; it is easy to see that J_0, J_1, J_2 are disjoint and their union equals J. We have

$$\sum_i \big(|F_{i-1}\cap\partial N_{d_j}| - |S_i\cap\partial N_{d_j}|\big) \geq \sum_i \big(|F_{i-1}\cap\partial N_z^j| - |S_i\cap\partial N_z^j|\big) \quad \text{for } j\in J_2.$$

Further, by Lemma 14,

$$\sum_i \big(\chi(F_{i-1}\cap N_z) - \chi(S_i\cap N_z)\big) \geq -2(2-n)+2+\sum_i \big(|F_{i-1}\cap\partial N_z| - |S_i\cap\partial N_z|\big).$$

Thus $-\chi(S)$ equals

$$\begin{aligned}
&\sum_i \big(\chi(F_{i-1}) - \chi(S_i)\big)\\
&\quad \geq \sum_i \big(\chi(F_{i-1}\cap N_z) - \chi(S_i\cap N_z)\\
&\qquad + \sum_j \Big(\chi(F_{i-1}\cap N_{c_j}) - \chi(S_i\cap N_{c_j}) + \chi(F_{i-1}\cap N_{d_j}) - \chi(S_i\cap N_{d_j})\\
&\qquad\qquad + \chi(F_{i-1}\cap N_{e_j}) - \chi(S_i\cap N_{e_j}) + \chi(F_{i-1}\cap N_{f_j}) - \chi(S_i\cap N_{f_j})\Big)\Big)\\
&\quad = \sum_i \big(\chi(F_{i-1}\cap N_z) - \chi(S_i\cap N_z)\big)\\
&\qquad + \sum_j \sum_i \big(\chi(F_{i-1}\cap N_{c_j}) - \chi(S_i\cap N_{c_j}) + \chi(F_{i-1}\cap N_{d_j}) - \chi(S_i\cap N_{d_j})\\
&\qquad\qquad + \chi(F_{i-1}\cap N_{e_j}) - \chi(S_i\cap N_{e_j}) + \chi(F_{i-1}\cap N_{f_j}) - \chi(S_i\cap N_{f_j})\big)\\
&\quad \geq -2(2-n-1) + \sum_i \big(|F_{i-1}\cap \partial N_z| - |S_i\cap \partial N_z|\big)\\
&\qquad + \sum_{j\in J_0}\sum_i \big(\chi(F_{i-1}\cap N_{c_j}) - \chi(S_i\cap N_{c_j}) + \chi(F_{i-1}\cap N_{d_j}) - \chi(S_i\cap N_{d_j})\\
&\qquad\qquad + \chi(F_{i-1}\cap N_{e_j}) - \chi(S_i\cap N_{e_j}) + \chi(F_{i-1}\cap N_{f_j}) - \chi(S_i\cap N_{f_j})\big)\\
&\qquad + \sum_{j\in J_1}\sum_i \big(\chi(F_{i-1}\cap N_{c_j}) - \chi(S_i\cap N_{c_j}) + \chi(F_{i-1}\cap N_{d_j}) - \chi(S_i\cap N_{d_j})\\
&\qquad\qquad + \chi(F_{i-1}\cap N_{e_j}) - \chi(S_i\cap N_{e_j}) + \chi(F_{i-1}\cap N_{f_j}) - \chi(S_i\cap N_{f_j})\big)\\
&\qquad + \sum_{j\in J_2}\sum_i \big(\chi(F_{i-1}\cap N_{c_j}) - \chi(S_i\cap N_{c_j}) + \chi(F_{i-1}\cap N_{d_j}) - \chi(S_i\cap N_{d_j})\\
&\qquad\qquad + \chi(F_{i-1}\cap N_{e_j}) - \chi(S_i\cap N_{e_j}) + \chi(F_{i-1}\cap N_{f_j}) - \chi(S_i\cap N_{f_j})\big)\\
&\quad \geq -2(1-n) + \sum_i \sum_j \big(|F_{i-1}\cap \partial N_z^j| - |S_i\cap \partial N_z^j|\big)\\
&\qquad + \sum_{j\in J_0} 6 + \sum_{j\in J_1}(8-2+2) + \sum_{j\in J_2}\Big(6 - \sum_i \big(|F_{i-1}\cap \partial N_z^j| - |S_i\cap \partial N_z^j|\big)\Big)\\
&\quad = -2(1-n) + \sum_{j\in J_0}\sum_i \big(|F_{i-1}\cap \partial N_z^j| - |S_i\cap \partial N_z^j|\big)\\
&\qquad + \sum_{j\in J_1}\sum_i \big(|F_{i-1}\cap \partial N_z^j| - |S_i\cap \partial N_z^j|\big) + \sum_{j\in J_2}\sum_i \big(|F_{i-1}\cap \partial N_z^j| - |S_i\cap \partial N_z^j|\big)\\
&\qquad + \sum_{j\in J_0} 6 + \sum_{j\in J_1}(8-2+2) + \sum_{j\in J_2}\Big(6 - \sum_i \big(|F_{i-1}\cap \partial N_z^j| - |S_i\cap \partial N_z^j|\big)\Big)\\
&\quad = -2 + 2n + 0 + \sum_{j\in J_1}(-2) + \sum_{j\in J_2}\sum_i \big(|F_{i-1}\cap \partial N_z^j| - |S_i\cap \partial N_z^j|\big)\\
&\qquad + \sum_{j\in J_0} 6 + \sum_{j\in J_1}(8-2+2) + \sum_{j\in J_2} 6 - \sum_{j\in J_2}\sum_i \big(|F_{i-1}\cap \partial N_z^j| - |S_i\cap \partial N_z^j|\big)\\
&\quad = -2 + 2n + \sum_{j\in J_0} 6 + \sum_{j\in J_1} 8 + \sum_{j\in J_2} 6 \geq -2 + 2n + 6n = 8n - 2.
\end{aligned}$$

This shows that genus $S \geq 4n$.

Case B: A component of $\big(\bigcup_i S_i\big)\cap N_z$ *is pseudohorizontal.* Denote this component by $\tilde{S}$ and note that $\big(\big(\bigcup_i F_i \cup \bigcup_i S_i\big)-\tilde{S}\big)\cap N_z = \varnothing$. Now $\chi(\tilde{S}) = 2-2\,\text{genus}\,\tilde{S} - |\partial\tilde{S}|$ and

$$\sum_i \big(|F_{i-1}\cap\partial N_z| - |S_i\cap\partial N_z|\big) = -|\partial\tilde{S}|.$$

Define J_0, J_1, J_2 as above and note that here $J_0 = \varnothing$. Then $-\chi(S)$ is given by

$$\begin{aligned}
&\sum_i \big(\chi(F_{i-1}) - \chi(S_i)\big)\\
&\quad\geq \sum_i \big(\chi(F_{i-1}\cap N_z) - \chi(S_i\cap N_z)\big)\\
&\qquad + \sum_i\sum_j \big(\chi(F_{i-1}\cap N_{c_j}) - \chi(S_i\cap N_{c_j}) + \chi(F_{i-1}\cap N_{d_j})\\
&\qquad\qquad - \chi(S_i\cap N_{d_j}) + \chi(F_{i-1}\cap N_{e_j}) - \chi(S_i\cap N_{e_j})\big)\\
&\quad= \sum_i \big(\chi(F_{i-1}\cap N_z) - \chi(S_i\cap N_z)\big)\\
&\qquad + \sum_i\sum_{j\in J_1} \big(\chi(F_{i-1}\cap N_{c_j}) - \chi(S_i\cap N_{c_j}) + \chi(F_{i-1}\cap N_{d_j})\\
&\qquad\qquad - \chi(S_i\cap N_{d_j}) + \chi(F_{i-1}\cap N_{e_j}) - \chi(S_i\cap N_{e_j})\big)\\
&\qquad + \sum_i\sum_{j\in J_2} \big(\chi(F_{i-1}\cap N_{c_j}) - \chi(S_i\cap N_{c_j}) + \chi(F_{i-1}\cap N_{d_j})\\
&\qquad\qquad - \chi(S_i\cap N_{d_j}) + \chi(F_{i-1}\cap N_{e_j}) - \chi(S_i\cap N_{e_j})\big)\\
&\quad= -2 + 2\,\text{genus}\,\tilde{S} + |\partial\tilde{S}| + \sum_{j\in J_1}\Big(8 + \sum_i \big(|F_{i-1}\cap\partial N_z^j| - |S_i\cap\partial N_z^j|\big)\Big) + \sum_{j\in J_2} 10\\
&\quad= -2 + 2\,\text{genus}\,\tilde{S} + \sum_j 8 \geq -2 + 8n.
\end{aligned}$$

Hence genus $S \geq 4n$. □

7. Some comments on nontotally orientable graph manifolds

In the proofs of Theorem 1 and Theorem 2 we make extensive use of the structure theorem for Heegaard splittings of totally orientable graph manifolds [Schultens 2004]. We believe however that similar statements are true for graph manifolds in general. This suggests a more straightforward generalization of the examples provided in [Weidmann 2003] which are not totally orientable.

Note that the verification that the manifolds constructed in [Weidmann 2003] are not of Heegaard genus 2 relies on the classification of 3-manifolds with nonempty characteristic submanifold that have a genus 2 Heegaard splitting as given by T. Kobayashi [1984].

Thus we conjecture that the manifolds M_n constructed below are of Heegaard genus $3n$, the same argument as above shows that they can be generated by $2n$ elements.

The graph underlying the manifold M_n is again Γ_n and the Seifert piece corresponding to the vertex v is again denoted by N_v. Moreover:

(1) The intersection number between the fibers of the adjacent Seifert spaces is 1 at any torus of the JSJ decomposition.

(2) $\mathbb{O}(N_z)$ is a n-punctured sphere with at most one cone point and $\hat{N}_z = S^3$.

(3) $\mathbb{O}(N_{d_i}) = P^2(\infty, \infty, 5n)$ and N_{d_i} admits no pseudohorizontal surface that has genus 2.

(4) $\mathbb{O}(N_{c_i})$ is of type $D(2, q)$ with odd q but N_{c_i} is not homeomorphic to the exterior of a 2-bridge knot in S^3.

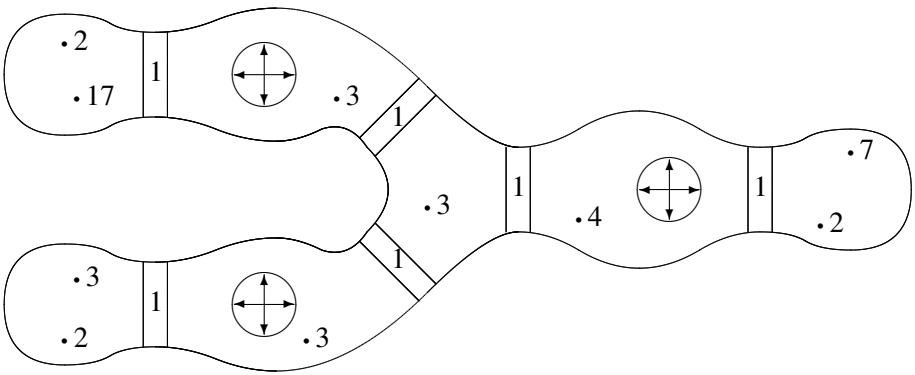


Figure 5. A graph manifold M with $g(M) = 9$ and $r(M) \leq 6$?

References

[Boileau and Zieschang 1984] M. Boileau and H. Zieschang, "Heegaard genus of closed orientable Seifert 3-manifolds", *Invent. Math.* **76**:3 (1984), 455–468. MR 86a:57008 Zbl 0538.57004

[Kobayashi 1984] T. Kobayashi, "Structures of the Haken manifolds with Heegaard splittings of genus two", *Osaka J. Math.* **21**:2 (1984), 437–455. MR 85k:57011 Zbl 0568.57007

[Moriah and Schultens 1998] Y. Moriah and J. Schultens, "Irreducible Heegaard splittings of Seifert fibered spaces are either vertical or horizontal", *Topology* **37**:5 (1998), 1089–1112. MR 99g:57021 Zbl 0926.57016

[Rost and Zieschang 1987] M. Rost and H. Zieschang, "Meridional generators and plat presentations of torus links", *J. London Math. Soc.* (2) **35**:3 (1987), 551–562. MR 88e:57010 Zbl 0587.57005

[Scharlemann 2002] M. Scharlemann, "Heegaard splittings of compact 3-manifolds", pp. 921–953 in *Handbook of geometric topology*, edited by R. Daverman and R. Sher, Elsevier, Amsterdam, 2002. MR 2002m:57027 Zbl 0985.57005

[Scharlemann and Schultens 1999] M. Scharlemann and J. Schultens, "The tunnel number of the sum of n knots is at least n", *Topology* **38**:2 (1999), 265–270. MR 2000b:57013 Zbl 0929.57003

[Scharlemann and Thompson 1994] M. Scharlemann and A. Thompson, "Thin position for 3-manifolds", pp. 231–238 in *Geometric topology* (Haifa, 1992), edited by C. Gordon et al., Contemp. Math. **164**, Amer. Math. Soc., Providence, RI, 1994. MR 95e:57032 Zbl 0818.57013

[Schultens 1993] J. Schultens, "The classification of Heegaard splittings for (compact orientable surface) $\times S^1$", *Proc. London Math. Soc.* (3) **67** (1993), 425–448. MR 94d:57043 Zbl 0789.57012

[Schultens 2004] J. Schultens, "Heegaard splittings of graph manifolds", *Geom. Topol.* **8** (2004), 831–876. MR 2005f:57031 Zbl 1055.57023

[Scott 1983] P. Scott, "The geometries of 3-manifolds", *Bull. London Math. Soc.* **15**:5 (1983), 401–487. MR 84m:57009 Zbl 0561.57001

[Waldhausen 1978] F. Waldhausen, "Some problems on 3-manifolds", pp. 313–322 in *Algebraic and geometric topology* (Stanford, 1976), vol. 2, edited by R. J. Milgram, Proc. Sympos. Pure Math. **32**, Amer. Math. Soc., Providence, R.I., 1978. MR 80g:57013 Zbl 0397.57007

[Weidmann 2003] R. Weidmann, "Some 3-manifolds with 2-generated fundamental group", *Arch. Math.* (*Basel*) **81**:5 (2003), 589–595. MR 2004j:57033 Zbl 1041.57008

Received October 11, 2005.

JENNIFER SCHULTENS
DEPARTMENT OF MATHEMATICS
1 SHIELDS AVE
UNIVERSITY OF CALIFORNIA
DAVIS, CA 95616
UNITED STATES

jcs@math.ucdavis.edu
http://www.math.ucdavis.edu/~jcs/

RICHARD WEIDMANN
DEPARTMENT OF MATHEMATICS
HERIOT-WATT UNIVERSITY
RICCARTON
EDINBURGH EH14 4AS
SCOTLAND, UK

R.Weidmann@ma.hw.ac.uk
http://www.ma.hw.ac.uk/~richardw/

CONTENTS

Volume 231, no. 1 and no. 2

Guidelines for Authors

Authors may submit articles at msp.org/pjm/about/journal/submissions.html and choose an editor at that time. Exceptionally, a paper may be submitted in hard copy to one of the editors; authors should keep a copy.

By submitting a manuscript you assert that it is original and is not under consideration for publication elsewhere. Instructions on manuscript preparation are provided below. For further information, visit the web address above or write to pacific@math.berkeley.edu or to Pacific Journal of Mathematics, University of California, Los Angeles, CA 90095–1555. Correspondence by email is requested for convenience and speed.

Manuscripts must be in English, French or German. A brief abstract of about 150 words or less in English must be included. The abstract should be self-contained and not make any reference to the bibliography. Also required are keywords and subject classification for the article, and, for each author, postal address, affiliation (if appropriate) and email address if available. A home-page URL is optional.

Authors are encouraged to use LaTeX, but papers in other varieties of TeX, and exceptionally in other formats, are acceptable. At submission time only a PDF file is required; follow the instructions at the web address above. Carefully preserve all relevant files, such as LaTeX sources and individual files for each figure; you will be asked to submit them upon acceptance of the paper.

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used but tags will be converted to the house format (see a current issue for examples).

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple or Mathematica. Many drawing tools such as Adobe Illustrator and Aldus FreeHand can produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, the authors should check with production by sending an email to pacific@math.berkeley.edu.

Each figure should be captioned and numbered, so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript will all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables, which should be used sparingly.

Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Page proofs will be made available to authors (or to the designated corresponding author) at a website in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

PACIFIC JOURNAL OF MATHEMATICS

Volume 231 No. 2 June 2007