Pacific Journal of Mathematics

Volume 282 No. 1 May 2016

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.

The subscription price for 2016 is US $440/year for the electronic version, and $600/year for print and electronic.
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. 282, No. 1, 2016
dx.doi.org/10.2140/pjm.2016.282.1

ON THE HALF-SPACE THEOREM FOR MINIMAL SURFACES IN HEISENBERG SPACE

TRISTAN ALEX

We propose a simple proof of the vertical half-space theorem for Heisenberg space.

1. Introduction

A half-space theorem states that the only properly immersed minimal surface which is contained in a half-space is a parallel translate of the boundary of the half-space, namely a plane. Hoffman and Meeks [1990] first proved it for $\mathbb{R}^3$. It fails in $\mathbb{R}^n$ or $\mathbb{H}^n$ for $n \geq 4$.

In recent years, there has been increased interest in homogeneous 3-manifolds (see [Abresch and Rosenberg 2004; Hauswirth et al. 2008]). The original proof of Hoffman and Meeks also works in Heisenberg space Nil_3 with respect to *umbrellas*, which are the exponential image of a horizontal tangent plane [Abresch and Rosenberg 2005]. Daniel and Hauswirth [2009] extended the theorem to vertical half-spaces of Heisenberg space, where vertical planes are defined as the inverse image of a straight line in the base of the Riemannian fibration $\mathrm{Nil}_3 \to \mathbb{R}^2$.

Vertical half-space theorem in Heisenberg space [Daniel and Hauswirth 2009]. *Let S be a properly immersed minimal surface in Heisenberg space. If S lies to one side of a vertical plane P, then S is a plane parallel to P.*

Essential for the proof of half-space theorems is the existence of a family of catenoids or generalized catenoids. Their existence is simple to establish in spaces where they can be represented as ODE solutions. For instance, horizontal umbrellas in Heisenberg space are invariant under rotations around the vertical axis, so they lead to an ODE. However, the lack of rotations about horizontal axes means that the existence of analogues of a horizontal catenoid amounts to establishing true PDE solutions. Daniel and Hauswirth use a Weierstraß-type representation to reduce this problem to a system of ODEs. Only after solving a period problem do they obtain the desired family of surfaces.

MSC2010: primary 53A10, 53C42; secondary 53A35.
Keywords: minimal immersions, half-space theorem, Heisenberg space.

In the present paper we introduce a simpler approach: we take a coordinate model of Heisenberg space and consider coordinate surfaces of revolution. Provided we can choose a family of surfaces whose mean curvature normal points into the half-space, the original maximum principle argument of Hoffman and Meeks will prove the theorem. Our approach is based on an idea by Bergner [2010], who generalized the classical half-space theorem to surfaces with negative Gaussian curvature such that the principal curvatures satisfy an inequality, and Sá Earp and Toubiana [1995], who consider special Weingarten surfaces with mean curvature satisfying an inequality.

It is an open problem to prove a vertical half-space theorem for $\mathrm{PSL}_2(\mathbb{R})$, where it would apply to surfaces whose mean curvature is the so-called *magic number* $H_0 = 1/2$, namely the limiting value of the mean curvature of large spheres. Here, it would state that surfaces with mean curvature $H_0 = 1/2$ lying on the mean convex side of a horocylinder can only be horocylinders, that is, the inverse image of a horocycle of the fibration $\mathrm{PSL}_2(\mathbb{R}) \to \mathbb{H}^2$. Our strategy could also work there. However, so far we have not been successful in establishing the desired family of generalized catenoids with $H \le H_0$.

2. The Euclidean half-space theorem

Euclidean half-space theorem [Hoffman and Meeks 1990]. *A properly immersed minimal surface S in $\mathbb{R}^3$ lying in a half-space H is a plane parallel to $P = \partial H$.*

Proof. By the standard maximum principle we can assume $\mathrm{dist}(S, P) = 0$ but $S \cap P = \varnothing$.

Let $\mathscr{C}_r \subset \mathbb{R}^3 \setminus H$ be a half catenoid with necksize r and $\partial\mathscr{C}_r \subset P$. By the properness of S, we can translate S by $\varepsilon > 0$ towards $\mathscr{C}_1$ such that S intersects P but stays disjoint to $\partial\mathscr{C}_r$ for all $r \in (0, 1]$.

As r tends to 0, the family of catenoids $\mathscr{C}_r$ converges to P minus a point. We claim that the set I of parameters for which $\mathscr{C}_r$ does not intersect S is open. Consider a catenoid $\mathscr{C}_{r_0}$ that does not intersect S. For each $r \in (0, 1)$ there exists a compact set K such that the distance between $\mathscr{C}_r$ and P is larger than 2ε in the complement of K. We may choose K in a way that this property holds for all r in a small neighborhood of r_0. This implies that the distance between S and all these $\mathscr{C}_r$ is larger than ε in the complement of K (see Figure 1).

However, within the compact set K, the distance between S and C_{r_0} is positive, so for all r in a (possibly smaller) neighborhood of r_0, this distance is still positive.

We conclude that in a small neighborhood of r_0,

$$\mathrm{dist}(\mathscr{C}_r, S) \ge \min\{\mathrm{dist}(\mathscr{C}_r \cap K, S \cap K), \mathrm{dist}(\mathscr{C}_r \cap K^c, S \cap K^c)\} > 0,$$

thereby proving our claim.

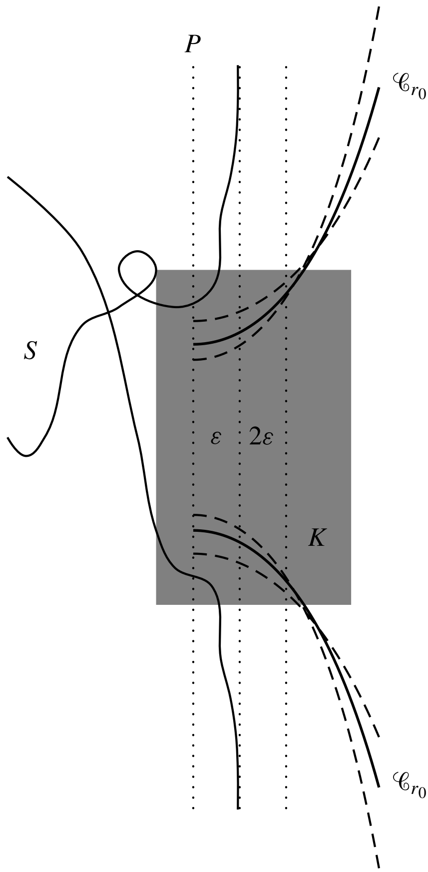


Figure 1. Proof of the Euclidean half-space theorem.

Therefore, the set of parameters for which $\mathscr{C}_r$ and S do intersect is closed, so there is a first catenoid $\mathscr{C}_{r_1}$ touching S at a point p. Since the boundaries of all $\mathscr{C}_r$ with $r \in (0, 1]$ are disjoint from S, the touching point p is an interior point, contradicting the maximum principle. $\square$

3. Coordinate surfaces of revolution

We take the following coordinates:

$$\mathrm{Nil}_3 := (\mathbb{R}^3, \mathrm{d}s^2), \qquad \mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + (2\tau x\,\mathrm{d}y - \mathrm{d}z)^2 \quad \text{with} \geq 0.$$

An orthonormal frame of the tangent space is given by

$$E_1 = \partial_x, \qquad E_2 = \partial_y + 2\tau x\,\partial_z, \qquad E_3 = \partial_z,$$

and the Riemannian connection in these coordinates is determined by

$$\begin{aligned} &\nabla_{E_1} E_2 = -\nabla_{E_2} E_1 = \tau E_3, && \nabla_{E_1} E_3 = \nabla_{E_3} E_1 = -\tau E_2, \\ &\nabla_{E_2} E_3 = \nabla_{E_3} E_2 = \tau E_1, && \nabla_{E_i} E_j = 0 \text{ in all other cases.} \end{aligned} \tag{1}$$

The Heisenberg space is a Riemannian fibration $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ with vanishing base curvature. The bundle curvature of Nil_3 is given by $\frac{1}{2} g(\nabla_{E_1} E_2 - \nabla_{E_2} E_1, E_3) = \tau$ and for $\tau = 0$ we recover $\mathbb{R}^3$.

Let us consider a curve $c(t) = (0, t, r(t))$ in Heisenberg space with a positive function r and $t \geq 0$. By rotating around the y-axis, we get an immersion

$$f : [t_0, \infty) \times [0, 2\pi) \to \mathrm{Nil}_3, \quad (t, \varphi) \mapsto \begin{pmatrix} -r(t) \sin \varphi \\ t \\ r(t) \cos \varphi \end{pmatrix}.$$

In order to apply the proof of Hoffman and Meeks, we will construct Euclidean surfaces of revolution around the y-axis. With the Heisenberg space metric, these rotations are not isometric, because the 4-dimensional isometry group of Nil_3 contains only translations and rotations around the vertical axis. Therefore, the mean curvature of such a surface will depend on the angle of rotation φ. We will need to find a surface with mean curvature vector pointing to the half-space to arrive at the desired contradiction with the maximum principle.

The tangent space of $M := f([t_0, \infty) \times [0, 2\pi))$ is spanned by

$$\begin{aligned} v_1 &= -r'(t) \sin \varphi E_1 + E_2 + (2\tau r(t) \sin \varphi + r'(t) \cos \varphi) E_3, \\ v_2 &= -r(t) \cos \varphi E_1 - r(t) \sin \varphi E_3, \end{aligned}$$

so the inner normal of M is

$$N = \frac{1}{W} \big(\sin \varphi E_1 + (r'(t) + 2\tau r(t) \sin \varphi \cos \varphi) E_2 - \cos \varphi E_3 \big),$$

where $W = \sqrt{1 + (2\tau r(t) \sin \varphi \cos \varphi + r'(t))^2}$.

We will now compute the first and second fundamental forms of M. We easily get

$$\begin{aligned} G_{ij} &= \mathrm{d}s^2(v_i, v_j) \\ &= \begin{pmatrix} r'(t)^2 \sin^2 \varphi + (2\tau r(t) \sin \varphi + r'(t) \cos \varphi)^2 + 1 & -2\tau r(t)^2 \sin^2 \varphi \\ -2\tau r(t)^2 \sin^2 \varphi & r(t)^2 \end{pmatrix} \end{aligned}$$

with determinant $\det G = r(t)^2 W^2$.

The most tedious part of the calculation is the second fundamental form. We have to compute

$$B_{ij} = \mathrm{d}s^2(\nabla_{v_i} v_j, N).$$

To start, (1) gives

$$\begin{aligned}
\nabla_{v_1} E_1 &= (-2\tau^2 r(t)\sin\varphi - \tau r'(t)\cos\varphi)E_2 - \tau E_3,\\
\nabla_{v_1} E_2 &= (2\tau^2 r(t)\sin\varphi + \tau r'(t)\cos\varphi)E_1 - \tau r'(t)\sin\varphi E_3,\\
\nabla_{v_1} E_3 &= \tau E_1 + \tau r'(t)\sin\varphi E_2.
\end{aligned}$$

We calculate

$$\begin{aligned}
\nabla_{v_1} v_1 &= -r''(t)\sin\varphi E_1 + (2\tau r'(t)\sin\varphi + r''(t)\cos\varphi)E_3\\
&\qquad - r'(t)\sin\varphi \nabla_{v_1} E_1 + \nabla_{v_1} E_2 + (2\tau r(t)\sin\varphi + r'(t)\cos\varphi)\nabla_{v_1} E_3\\
&= (-r''(t)\sin\varphi + 4\tau^2 r(t)\sin\varphi + 2\tau r'(t)\cos\varphi)E_1\\
&\qquad + (4\tau^2 r(t) r'(t)\sin^2\varphi + 2\tau r'(t)^2 \sin\varphi\cos\varphi)E_2\\
&\qquad + (2\tau r'(t)\sin\varphi + r''(t)\cos\varphi)E_3,
\end{aligned}$$

and obtain the first entry of B as

$$\begin{aligned}
B_{11} = \frac{1}{W}\big(&-r''(t) + 4\tau^2 r(t) r'(t)^2 \sin^2\varphi + 8\tau^3 r(t)^2 r'(t)\sin^3\varphi\cos\varphi\\
&+ 4\tau^2 r(t)\sin^2\varphi + 2\tau r'(t)^3\sin\varphi\cos\varphi + 4\tau^2 r(t) r'(t)^2\sin^2\varphi\cos^2\varphi\big).
\end{aligned}$$

The other three entries arise similarly from

$$\begin{aligned}
\nabla_{v_2} v_1 = \nabla_{v_1} v_2 &= -(\tau r(t)\sin\varphi + r'(t)\cos\varphi)E_1\\
&\qquad + \big(\tau r(t)(2\tau r(t)\sin\varphi\cos\varphi + r'(t)\cos(2\varphi))\big)E_2\\
&\qquad + (\tau r(t)\cos\varphi - r'(t)\sin\varphi)E_3,\\
\nabla_{v_2} v_2 &= r(t)\sin\varphi E_1 - 2\tau r(t)^2\sin\varphi\cos\varphi E_2 - r(t)\cos\varphi E_3.
\end{aligned}$$

They are

$$\begin{aligned}
B_{12} = B_{21} &= \frac{1}{W}\big(\tau r(t)\big(4\tau r(t) r'(t)\sin\varphi\cos^3\varphi\\
&\qquad + \tau^2 r(t)^2\sin^2(2\varphi) + r'(t)^2\cos(2\varphi) - 1\big)\big),\\
B_{22} &= -\frac{1}{W}\big(r(t)\big(\tau r(t)\sin(2\varphi)(\tau r(t)\sin(2\varphi) + r'(t)) - 1\big)\big).
\end{aligned}$$

We obtain the mean curvature H for our coordinate surface of revolution:

Lemma 1. *The mean curvature $H = H(t,\varphi)$ of f is given by*

$$\begin{aligned}
H &:= \tfrac{1}{2}\operatorname{tr}(G^{-1}B)\\
&= \frac{G_{22}B_{11} - G_{12}B_{21} - G_{21}B_{12} + G_{11}B_{22}}{2r(t)^2 W^2}\\
&= \frac{1 + r'(t)^2 - r(t)r''(t) + 4\tau^2 r(t)^2\sin^4\varphi + 2\tau r(t) r'(t)\sin\varphi\cos\varphi}{2r(t)W^3}.
\end{aligned}$$

4. Half-space theorem in Heisenberg space

As expected, for $\tau = 0$, Lemma 1 recovers the mean curvature for surfaces of revolution in Euclidean space. For $\tau \neq 0$, the two additional terms depending on φ in the nominator of H arise because the horizontal rotation is not an isometry of Heisenberg space. Our goal is to exhibit a family of surfaces of revolution satisfying $H \leq 0$ with respect to the normal N.

Consider the surface of revolution f_c generated by the curve by

$$r_c(t) := \exp\left(\tfrac{1}{c}\exp(ct)\right) \tag{2}$$

with $c > c_0 := 4\tau^2 + 2\tau + 1$. We claim that this surface satisfies $H \leq 0$ for $t > 0$. Indeed, the following estimate for the denominator of H holds:

$$\begin{aligned}
2r(t)W^3H &\leq 1 + r_c'(t)^2 - r_c(t)r_c''(t) + 4\tau^2 r_c(t)^2 + 2\tau r_c(t)r_c'(t)\\
&= 1 + r_c(t)^2(\exp(ct)(2\tau - c) + 4\tau^2)\\
&\leq 1 + r_c(t)^2(4\tau^2 + 2\tau - c) \leq 1 + 4\tau^2 + 2\tau - c \leq 0.
\end{aligned}$$

Since we consider a surface of revolution with an embedded meridian, the embeddedness of $M_c := f_c([t_0, \infty) \times [0, 2\pi))$ is obvious. Also, the boundary

$$\partial M_c = \left\{\exp\left(\tfrac{1}{c}\right) \cdot (\sin\varphi, 0, \cos\varphi) \colon \varphi \in [0, 2\pi)\right\}$$

is explicitly known.

It is also important to note that for each c and any given $\varepsilon > 0$, there exists a compact set such that the distance between M_c and the plane $\{y = 0\}$ is larger than ε in the complement of this compact set.

Let us summarize the result:

Lemma 2. *The coordinate surface of revolution whose meridian is defined by (2) satisfies, for $c > c_0$,*

1. $H \leq 0$ *with respect to the normal N,*
2. *for $c \to \infty$, the surface M_c converges uniformly to a subset of $\{y = 0\}$ on compact sets,*
3. *M_c is properly embedded, and*
4. $\partial M_c = \left\{\exp\left(\tfrac{1}{c}\right) \cdot (\sin\varphi, 0, \cos\varphi) : \varphi \in [0, 2\pi)\right\}$ *for all c.*

Using the surfaces M_c, our proof of the Euclidean half-space theorem literally applies to Heisenberg space.

Acknowledgement

I would like to thank my advisor Karsten Große-Brauckmann for his help.

References

[Abresch and Rosenberg 2004] U. Abresch and H. Rosenberg, "A Hopf differential for constant mean curvature surfaces in $\mathbb{S}^2 \times \mathbb{R}$ and $\mathbb{H}^2 \times \mathbb{R}$", *Acta Math.* **193**:2 (2004), 141–174. MR 2006h:53003 Zbl 1078.53053

[Abresch and Rosenberg 2005] U. Abresch and H. Rosenberg, "Generalized Hopf differentials", *Mat. Contemp.* **28** (2005), 1–28. MR 2006h:53004 Zbl 1118.53036

[Bergner 2010] M. Bergner, "A halfspace theorem for proper, negatively curved immersions", *Ann. Global Anal. Geom.* **38**:2 (2010), 191–199. MR 2011j:53009 Zbl 1217.53010

[Daniel and Hauswirth 2009] B. Daniel and L. Hauswirth, "Half-space theorem, embedded minimal annuli and minimal graphs in the Heisenberg group", *Proc. Lond. Math. Soc.* (3) **98**:2 (2009), 445–470. MR 2009m:53018 Zbl 1163.53036

[Hauswirth et al. 2008] L. Hauswirth, H. Rosenberg, and J. Spruck, "On complete mean curvature $1/2$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$", *Comm. Anal. Geom.* **16**:5 (2008), 989–1005. MR 2010d:53009 Zbl 1166.53041

[Hoffman and Meeks 1990] D. Hoffman and W. H. Meeks, III, "The strong halfspace theorem for minimal surfaces", *Invent. Math.* **101**:2 (1990), 373–377. MR 92e:53010 Zbl 0722.53054

[Sá Earp and Toubiana 1995] R. Sá Earp and E. Toubiana, "Sur les surfaces de Weingarten spéciales de type minimal", *Bol. Soc. Brasil. Mat.* (*N.S.*) **26**:2 (1995), 129–148. MR 96j:53006 Zbl 0864.53004

Received September 16, 2015. Revised October 30, 2015.

TRISTAN ALEX
FACHBEREICH MATHEMATIK
TECHNISCHE UNIVERSITÄT DARMSTADT
AG GEOMETRIE UND APPROXIMATION
SCHLOSSGARTENSTRASSE 7
D-64289 DARMSTADT
GERMANY
alex@mathematik.tu-darmstadt.de

PACIFIC JOURNAL OF MATHEMATICS
Vol. 282, No. 1, 2016
dx.doi.org/10.2140/pjm.2016.282.9

EXTENDING SMOOTH CYCLIC GROUP ACTIONS ON THE POINCARÉ HOMOLOGY SPHERE

NIMA ANVARI

Let X_0 denote a compact, simply connected, smooth 4-manifold with boundary the Poincaré homology 3-sphere $\Sigma(2, 3, 5)$ and with even negative definite E_8 intersection form. We obtain constraints on the rotation data if a free $\mathbb{Z}/p$-action on $\Sigma(2, 3, 5)$ extends to a smooth, homologically trivial action on X_0 with isolated fixed points, for any odd prime $p \geq 7$. The approach is to study the equivariant Yang–Mills instanton-one moduli space for cylindrical-end 4-manifolds. As an application we show that a smooth, homologically trivial $\mathbb{Z}/7$-action on $\#^8 S^2 \times S^2$ with isolated fixed points does not equivariantly split along a free action on $\Sigma(2, 3, 5)$.

1. Introduction

It is well known that the Poincaré homology 3-sphere $\Sigma(2, 3, 5)$ can be realized as the boundary of a smooth, compact, simply connected 4-manifold X_0 obtained by plumbing disk bundles over 2-spheres along the E_8 graph. Let π denote a cyclic group of prime order $p \geq 7$; then the Poincaré homology sphere admits a free π-action contained in the circle action that gives it the structure of a Seifert fibered manifold. In this paper, we obtain constraints for this action to extend smoothly and homologically trivially to X_0 with isolated fixed points. An action is *homologically trivial* if it induces the identity on integral homology $H_2(X_0, \mathbb{Z})$ and in this case the action necessarily has fixed sets in the interior of X_0.

When studying symmetries of 4-manifolds, typically there are known linear actions and one would like to understand how closely a general smooth group action resembles the linear models. In our case, the linear actions are obtained by plumbing equivariantly; in fact, the circle action on $\Sigma(2, 3, 5)$ extends to X_0. These actions, however, always contain a fixed 2-sphere, namely the central node in the E_8 graph. We ask if there can be any smooth extension with only isolated fixed points. We have local tangential representations at each fixed point described by rotation numbers, a pair (a, b) of nonzero integers modulo p, well-defined up to order and simultaneous change of sign. For a fixed generator t of π, the local

MSC2010: primary 57S17, 58D19; secondary 70S15.
Keywords: Yang Mills moduli spaces, gauge theory, group actions, Poincaré homology sphere.

representation is given by $t \cdot (z_1, z_2) = (t^a z_1, t^b z_2)$. The constraints are in the form of congruence relations satisfied by the rotation numbers.

Theorem A. *Let X_0 denote a compact, simply connected smooth 4-manifold with boundary $\Sigma(2,3,5)$ and negative definite intersection form E_8. For any prime $p > 5$, if a free $\mathbb{Z}/p$-action on $\Sigma(2,3,5)$ extends to a smooth, homologically trivial action on X_0, then the rotation data of the isolated fixed points are (a, b) such that $a + b \equiv \pm 1 \pmod p$ or $a + b \equiv \pm 7 \pmod p$.*

Remark 1.1. Note the action is automatically homologically trivial for $p > 7$, since the π-action on X_0 gives rise to an integral representation on $H_2(X_0; \mathbb{Z})$ and a decomposition (see [Curtis and Reiner 1962, p. 508; Edmonds 1989, p. 111])

$$H_2(X_0; \mathbb{Z}) = \mathbb{Z}[\pi]^r \oplus \mathbb{Z}^t \oplus \mathbb{Z}[\zeta_p]^c$$

as $\mathbb{Z}[\pi]$-modules with multiplicities $r, t, c \geq 0$ and $b_2(X_0) = rp + t + (p-1)c$. When $p > b_2(X_0) + 1$ we must have $r, c = 0$ and $t = b_2(X_0)$. When $p = 7$ the action need not be homologically trivial (see [Quebbemann 1981, Example 3.10, p. 168]), as the splitting $H_2(X_0; \mathbb{Z}) = \mathbb{Z}[\pi] \oplus \mathbb{Z}$ may occur (c must be even by [Edmonds 1989, Proposition 2.4(i)]; see also the algebraic result in [Hambleton and Riehm 1978, Proposition 10(c)]). For homologically trivial actions, the fixed set X_0^π consists of isolated points and 2-spheres, and the Lefschetz fixed point formula gives the Euler characteristic $\chi(X_0^\pi) = 9$.

The necessary conditions for a smooth extension from Theorem A can be checked against the Atiyah–Patodi–Singer G-signature formula for manifolds with boundary; see [Atiyah et al. 1975b]. This leads to the following rigidity result.

Theorem B. *Let X_0 denote a compact, simply connected, smooth 4-manifold with boundary $\Sigma(2,3,5)$ and negative definite E_8 intersection form. A free $\mathbb{Z}/7$-action on $\Sigma(2,3,5)$ does not extend to a smooth, homologically trivial action on X_0 with fixed set consisting of only isolated fixed points.*

As a consequence, we have the following corollary regarding equivariant embedding of the Poincaré homology sphere in $\#^8 S^2 \times S^2$.

Corollary C. *The 4-manifold $X = \#^8 S^2 \times S^2$, with a smooth, homologically trivial $\mathbb{Z}/7$-action with only isolated fixed points, does not contain an equivariant embedding of $\Sigma(2,3,5)$ with a free action.*

Remark 1.2. If $\Sigma(2,3,5)$ embeds smoothly in $\#^8 S^2 \times S^2$, it separates X into two smooth, spin 4-manifolds with boundary, each with even intersection form. By van der Blij's lemma and the nontriviality of the Rokhlin invariant of $\Sigma(2,3,5)$, each side must have signature divisible by 8 and since $b_2(X) = 16$, each side must have definite intersection form. The additivity of the signature shows that they must have opposite sign. So any embedding of $\Sigma(2,3,5)$ in $\#^8 S^2 \times S^2$ decomposes X as $X_0 \cup_{\Sigma(2,3,5)} X_1$ with intersection form $Q_X = E_8 \oplus -E_8$.

In the next section we summarize the results of equivariant Yang–Mills moduli spaces that are needed in the proof of Theorem A. In the equivariant setting, a crucial technical component is provided by equivariant transversality results, as developed by Hambleton and Lee [1992] and based on Bierstone's theory of equivariant general position [1977]. This provides a suitable perturbation theory that gives the moduli spaces the structure of a Whitney stratified space.

2. The equivariant moduli space

Let X_0 denote a smooth, compact, simply connected 4-manifold with negative definite intersection form E_8 whose boundary is the Poincaré homology sphere $\partial X_0 = \Sigma(2,3,5)$. Suppose we have a cyclic group $\pi = \mathbb{Z}/p$ of odd prime order acting smoothly on X_0 which is both homologically trivial and free on the boundary. Denote by (X, g) the cylindrical-end Riemannian manifold $X = X_0 \cup \mathrm{End}(X)$ where $\mathrm{End}(X)$ is orientation preserving isometric to $\Sigma(2,3,5) \times [0,\infty)$, with g a π-invariant metric which restricts to a product metric on end. Yang–Mills moduli spaces for cylindrical-end 4-manifolds have been studied extensively; see [Taubes 1987; Morgan et al. 1994; Donaldson 2002]. We briefly sketch the main ideas here and refer the reader to the sources for details.

Consider a principal $\mathrm{SU}(2)$ bundle P over X. By fixing a trivialization we obtain bundle maps which cover the π-action on X. Let $\mathcal{G}(\pi) = \{\hat{t} : P \to P \mid t \in \pi\}$; then there exists an exact sequence

$$1 \to \mathcal{G} \to \mathcal{G}(\pi) \to \pi \to 1 \tag{2-1}$$

where $\mathcal{G}$ is the gauge group of P. The natural action of $\mathcal{G}(\pi)$ on the space of connections $\mathcal{A}(P)$ is given by pullback; it is well-defined modulo gauge; and thus the space $\mathcal{B}(P) = \mathcal{A}/\mathcal{G}$ of connections up to gauge transformations inherits an action of $\mathcal{G}(\pi)/\mathcal{G} = \pi$.

Recall that the Yang–Mills energy functional acts on the space of connections by

$$\mathcal{YM}(A) = -\frac{1}{8\pi^2}\int_X \mathrm{Tr}(F_A \wedge *F_A) = \frac{1}{8\pi^2}\int_X |F_A|^2 = \|F_A\|^2_{L^2}, \tag{2-2}$$

where F_A is the $\mathrm{ad}(P)$-valued curvature 2-form of the connection, and $*$ is the Hodge star operator associated to the Riemannian metric. The Hodge star-operator on X extends to an involution on bundle valued 2-forms giving rise to a splitting $\Omega^2(\mathrm{ad}\, P) = \Omega^2_+(\mathrm{ad}\, P) \oplus \Omega^2_-(\mathrm{ad}\, P)$ into self-dual and anti-self-dual (ASD) 2-forms. The L^2-finite moduli spaces are anti-self-dual connections modulo gauge with finite

Yang–Mills action:

$$\mathcal{M}(X, g) = \{[A] \in \mathcal{B}(P) \mid F_A^+ = 0, \|F_A\|^2_{L^2} < \infty\}. \tag{2-3}$$

This space is π-invariant since π acts by isometries. It is a fundamental result that g-ASD connections with finite Yang–Mills energy are asymptotic to flat connections down the cylindrical-end; see [Donaldson 2002, p. 77]. Since flat connections are π-invariant under the pullback action, this defines a π-equivariant boundary map $\partial_\infty : \mathcal{M}(X, g) \to \mathcal{R}(\Sigma(2, 3, 5))$ where $\mathcal{R}(\Sigma)$ denotes the representation variety of flat SU(2)-connections modulo gauge. This gives a π-invariant partition of the moduli space according to its limiting flat connection:

$$\mathcal{M}(X, g) = \bigsqcup_{\alpha \in \mathcal{R}(\Sigma)} \mathcal{M}(X, \alpha). \tag{2-4}$$

The energy of a g-ASD connection A is congruent modulo $\mathbb{Z}$ to the Chern–Simons invariant of the limiting flat connection α and so the energy takes on a discrete set of values determined by the Chern–Simons invariant and we get a further π-invariant decomposition according to energy value $\mathcal{M}(X, \alpha) = \bigsqcup_{\ell \geq 0} \mathcal{M}_\ell(X, \alpha)$ with $\ell \equiv \mathrm{CS}(\alpha) \mod \mathbb{Z}$. The index of the δ-decay complex [Morgan et al. 1994]

$$0 \to \Omega^0_{3,\delta}(X, \operatorname{ad} P) \xrightarrow{d_A} \Omega^1_{2,\delta}(X, \operatorname{ad} P) \xrightarrow{d_A^+} \Omega^2_{1,\delta,+}(X, \operatorname{ad} P) \to 0 \tag{2-5}$$

gives the formal dimension of the moduli space

$$\dim \mathcal{M}_\ell(X, \alpha) = 8\ell - \tfrac{3}{2}(\chi(X) + \mathrm{Sign}(X)) - \tfrac{1}{2}(h^1_\alpha + h^0_\alpha) + \tfrac{1}{2}\rho_\alpha(\Sigma), \tag{2-6}$$

where $h^i_\alpha = \dim_{\mathbb{R}} H^i(\Sigma, \operatorname{ad} \alpha)$ for $i = 0, 1$ and $\rho_\alpha(\Sigma)$ is the Atiyah–Patodi–Singer rho invariant [Atiyah et al. 1975a]. The corresponding dimension formula for a Floer-type moduli space on the cylinder is given by

$$\dim \mathcal{M}_\ell(\Sigma \times \mathbb{R}, \alpha, \beta) = 8\ell - \tfrac{1}{2}(h_\alpha + h_\beta) + \tfrac{1}{2}(\rho_\beta(\Sigma) - \rho_\alpha(\Sigma)) \tag{2-7}$$

with $h_\alpha = h^1_\alpha + h^0_\alpha$, similarly for h_β, and $\ell \equiv \mathrm{CS}(\beta) - \mathrm{CS}(\alpha) \mod \mathbb{Z}$.

The moduli space of interest in this paper are ASD connections on X asymptotic to the trivial product connection and with unit L^2-norm. Since the intersection form of X is negative definite, the formal dimension is $\dim \mathcal{M}_1(X, \theta) = 5$. This is the π-equivariant instanton-one moduli space we study to extract information about the original π-action on X.

2.1. ***Uhlenbeck–Taubes compactification.*** Let us recall the compactness theorem of Uhlenbeck [Lawson 1985; Freed and Uhlenbeck 1991; Donaldson and Kronheimer 1990] for the instanton moduli spaces. Intuitively, if we are given an infinite sequence of uniformly bounded g-ASD connections without a convergent subsequence, then there exists a gauge equivalent subsequence which has a weak limit, where the limiting ASD connection has a curvature density that accumulates in integral amounts of the total energy around a finite number of points in X. For a moduli space with one unit of total energy, there can be at most one point where curvature becomes highly concentrated. Uhlenbeck compactness continues to hold in the cylindrical-end setting. After passing to a subsequence we can find a gauge equivalent sequence that converges on compact subsets, but since our manifold is noncompact there is the possibility that curvature escapes down the cylindrical-end. This corresponds to broken trajectories of flat connections for the Chern–Simons flow and leads to convergence without loss of energy; see [Morgan et al. 1994, 6.3.3; Donaldson 2002, 5.1]. Weak limits are defined as a tuple of gauge equivalence class of L^2-finite ASD connections $[A] := ([A_0], [A_1], \dots, [\theta])$ where $[A_0] \in \mathcal{M}_{\ell_0}(X, \alpha_0)$ and $[A_i] \in \mathcal{M}_{\ell_i}(\Sigma \times \mathbb{R}, \alpha_{i-1}, \alpha_i)$, α_i are flat connections on Σ and have compatible boundary values $\partial_\infty(A_i) = \partial_\infty(A_{i+1})$. The "ends" of the moduli space $\mathcal{M}_1(X, \theta)$ are parametrized by products of the form

$$\mathcal{M}_{\ell_0}(X, \alpha_0) \times \mathcal{M}_{\ell_1}(\Sigma \times \mathbb{R}, \alpha_0, \alpha_1) \times \cdots \times \mathcal{M}_{\ell_k}(\Sigma \times \mathbb{R}, \alpha_{k-1}, \theta). \tag{2-8}$$

with $\sum_i \ell_i = 1$.

We also have the analogue of the Taubes construction; see [Freed and Uhlenbeck 1991] for details and also [Buchdahl et al. 1990] for the equivariant case. Since X is negative definite, this provides an equivariant collar neighborhood in the moduli space and a partial compactification

$$\overline{\mathcal{M}}_1(X, \theta) = \mathcal{M}_1(X, \theta) \cup X \times (0, \lambda_0) \tag{2-9}$$

consisting of g-ASD connections with highly concentrated curvature. In particular, the equivariant moduli space $\mathcal{M}_1(X, \theta)$ is nonempty when the fixed set X^π is nonempty. For connections $[A] \in X \times (0, \lambda_0)$ Taubes also shows that $H^2_A = 0$ [Lawson 1985, Theorem 3.38, p. 81]; thus a neighborhood of the collar is a smooth 5-manifold and these connections are irreducible. The fixed set X^π give rise to a family of ASD connections which correspond to equivariant lifts of the π-action on X to a $\tilde{\pi} = \mathbb{Z}/2p$-action on the principal SU(2)-bundle; see [Braam and Matić 1993]. We study the π-equivariant compactification of the fixed set $\mathcal{M}_1(X, \theta)^\pi$ which originates in the Taubes collar to obtain information about the fixed set X^π.

2.2. ***Equivariant general position.*** In the nonequivariant setting, the argument in Freed and Uhlenbeck [1991] can be adapted to show that for a Baire set of metrics g which restrict to a product metric on the End(X), the moduli space $\mathcal{M}_1(X, \theta)$ is a smooth 5-dimensional manifold. In the equivariant setting we have a theorem of Cho [1991] on the existence of a Baire set of π-invariant metrics on X such that all the components of the fixed set $\mathcal{M}_1(X, \theta)^\pi$ are either empty or smooth manifolds. This π-invariant version of Freed and Uhlenbeck is also valid on cylindrical end

4-manifolds; see [Buchdahl et al. 1990]. Even though $\mathcal{M}_1(X,\theta)^\pi$ is smooth if nonempty, it may not have smooth π-invariant neighborhoods and in general the surrounding moduli space may be highly singular. Another approach would be to perturb the anti-self-duality equations at the chart level by passing to the Kuranishi model

$$\phi : H^1_A \to H^2_A \tag{2-10}$$

as in Donaldson [1983]. In the equivariant case, H^1_A and H^2_A are finite dimensional real π-representation spaces and the obstruction to the existence of an equivariant perturbation is

$$[H^1_A] - [H^2_A] \in R^+(\pi) \tag{2-11}$$

being an actual representation. Hambleton and Lee in [1992] applied the theory of equivariant general position of Bierstone [1977] to equivariant moduli spaces. For our setting, we use Wilson loop perturbations in free π-orbits of embedded circles in X. The nonequivariant case is described in [Donaldson 1987, pp. 400–401]. The perturbed section $F^+_A + \hat{\sigma}_+(A)$ is now $\mathcal{G}(\pi)$-equivariant and the perturbed moduli space inherits a π-action as before.

Since Bierstone general position is an open-dense condition, a generic equivariant perturbation of the ASD equations give the moduli spaces the structure of a Whitney stratified space, with open manifold strata and equivariant cone bundle neighborhoods; see [Bierstone 1977; Hambleton and Lee 1992] for details.

3. Proof of Theorem A

We begin with a lemma to determine the equivariant bundle structures. These are described by weights $\pm\lambda$ of the isotropy representation $\tilde{\pi} \to \mathrm{SU}(2)$ over a fixed point. Each of the fixed points $p_i \in X^\pi$ lies at the Taubes collar $X \times (0, \lambda_0)$ of the moduli space and is one end of a π-fixed arc γ_i. We would like to show that none of these arcs connect with each other in the irreducible component of the moduli space $\mathcal{M}_1(X,\theta)$.

Lemma 3.1 [Hambleton and Lee 1995, Lemma 17]. *If a fixed point has rotation numbers (a, b) then the equivariant lift it generates in the moduli space has an isotropy representation over the fiber of this point given by $\mathbb{Z}/2p$-weights $\pm(b - a)$ and over the other fixed points $\pm(a + b)$. Moreover, the γ_i represent distinct equivariant bundle structures and are therefore disjoint in $\mathcal{M}^*_1(X,\theta)$.*

Proof. Since the Euler characteristic $\chi(\mathrm{Fix}(X_0, \pi)) = 9$, we may suppose there are at least three fixed points of the π-action p_i, say with rotation numbers (a_1, b_1), (a_2, b_2), (a_3, b_3). Suppose that γ connects p_1 and p_2; the normal bundle information is propagated along this oriented arc and gives a canceling pair of rotation numbers $(a_2, b_2) = (a_1, -b_1)$. We will use the presence of the third distinct fixed point p_3 to get a contradiction. Because the point p_1 is fixed, there is a π-invariant ball $B(p_1)$ with a linear action and an equivariant degree one map $f_1 : X \to S^4$. We can pullback the equivariant bundle structure $Q \to S^4$ via f_1 and get an equivariant bundle $(X,\, f_1^* Q)$ with $\mathbb{Z}/2p$-weights $\pm(b_1 - a_1)$ over p_1 and $\pm(a_1 + b_1)$ over the other fixed points [Fintushel and Lawson 1986]. Similarly, we can do this with a map f_2 about the point p_2, this gives an equivariant bundle structure $(X,\, f_2^* Q')$. Since these bundle structures are equivalent, the isotropy at p_3 has to agree and a comparison shows that either $2a_1 \equiv 0 \pmod p$ or $2b_1 \equiv 0 \pmod p$, in either case we get a contradiction. □

The lemma shows that the fixed arcs γ_i generated by the fixed points in X must have an end that is not a component of the Taubes collar and according to the Uhlenbeck compactness results applicable here, these arcs must lead to energy or charge splitting down the cylindrical end. We first rule out the case of a trivial splitting:

Lemma 3.2. *The one-dimensional fixed set generated by the fixed points in the Taubes boundary $X \times (0, \lambda_0)$ cannot split energy in the equivariant compactification of $\mathcal{M}_1(X, \theta)$ by $\mathcal{M}_0(X, \theta) \times \mathcal{M}_1(\theta, \theta)$.*

Proof. The idea is that $\mathcal{M}_0(X, \theta)$ has zero energy, so it leaves behind a flat equivariant bundle which identifies the isotropy over the fibers of each fixed point. Suppose that γ is a one-parameter family of π-fixed ASD connections generated at the Taubes boundary from the fixed point with rotation numbers (a, b). Then the corresponding equivariant lift has isotropy over the fiber of this fixed point with weight $\pm(b - a)$ and $\pm(a + b)$ over the other fixed points. In such an energy splitting a flat equivariant bundle identifies the isotropy over all the points, so $a + b = \pm(b - a)$ and this forces either $2a \equiv 0 \pmod p$ or $2b \equiv 0 \pmod p$. Since p is odd and (a, b) are rotation numbers for a fixed point we get a contradiction. □

A nontrivial charge splitting will involve the flat connections of $\Sigma(2, 3, 5)$, which as representations α of the fundamental group into SU(2), are determined by rotation numbers (ℓ_1, ℓ_2, ℓ_3) and for $\Sigma(2, 3, 5)$ there are only two irreducible representations [Saveliev 2002]. We record in Table 1 the necessary values for index calculations.

The energy in $\mathcal{M}(\alpha_i, \theta)$ is given by $-\mathrm{CS}(\Sigma(2, 3, 5), \alpha_i) \bmod \mathbb{Z} \in (0, 1]$; see [Fintushel and Stern 1990, Saveliev 2002, p. 101]. In an energy splitting, the moduli space has an end given by a local diffeomorphism

$$\mathcal{M}_{\ell_0}(X, \alpha_0) \times_{\alpha_0} \mathcal{M}_{\ell_1}(\alpha_0, \alpha_1) \times_{\alpha_1} \cdots \times_{\alpha_{k-1}} \mathcal{M}_{\ell_k}(\alpha_{k-1}, \theta) \to \mathcal{M}_1(X, \theta),$$

where $\{\alpha_i\}_{i=1}^{k-1}$ are irreducible flat connections on $\Sigma(2, 3, 5)$; this then leads to a

α	(ℓ_1, ℓ_2, ℓ_3)	$\mu(\alpha)$	$\rho(\alpha)/2$	$-\mathrm{CS}(\alpha) \in (0, 1]$
1	$(1, 2, 2)$	5	$-97/30$	$49/120$
2	$(1, 2, 4)$	1	$-73/30$	$1/120$

Table 1. For each flat connection α of $\Sigma(2, 3, 5)$ are listed values for the Floer μ-index modulo 8, one-half the Atiyah–Patodi–Singer ρ-invariant and minus the Chern–Simons invariant of the given flat connection [Fintushel and Stern 1990]. The values for the ρ-invariant can be computed using a flat SO(3)-cobordism to a disjoint union of lens spaces; see [Saveliev 2002, p. 144].

dimension count

$$5 = \dim \mathcal{M}_{\ell_0}(X, \alpha_0) + \sum_{i=1}^{k} \dim \mathcal{M}_{\ell_i}(\alpha_{i-1}, \alpha_i)$$

with $\alpha_k = \theta$ and, as the convergence is without loss of energy, we have the condition

$$\sum_{i=0}^{k} \ell_i = 1.$$

The dimensions modulo 8 can be determined [Floer 1988] by the formulas

$$\begin{aligned} \dim \mathcal{M}(\alpha, \beta) &\equiv \mu(\alpha) - \mu(\beta) - \dim \mathrm{Stab}(\beta) \pmod 8, \\ \dim \mathcal{M}(X, \alpha) &\equiv -\mu(\alpha) - 3 \pmod 8, \end{aligned} \tag{3-1}$$

where μ is the Floer index and $\mu(\theta) = -3$. Imposing the energy condition allows one to determine the exact geometric dimensions. Since there are only 2 irreducible flat connections on $\Sigma(2, 3, 5)$ denoted by $\alpha_1 = (1, 2, 2)$ and $\alpha_2 = (1, 2, 4)$, we have only the possibilities listed in Table 2.

Let $\Sigma(b = 0; (a_1, b_1), (a_2, b_2), (a_3, b_3))$ be the Seifert invariants $\Sigma(a_1, a_2, a_3)$ and $\pi = \mathbb{Z}/p$ act as the standard action on $\Sigma(a_1, a_2, a_3)$. Then the orbit space $Q = \Sigma/\pi$ is a rational homology sphere with Seifert invariants $Q(b = 0; (a_i, \beta_i))$, where $\beta_i = pb_i$. We will need the formula for the Chern–Simons invariant of reducible flat connections on Q. Note that if we take the p-fold cover we get the trivial product connection on $\Sigma(a_1, a_2, a_3)$ and, as Chern–Simons invariants are multiplicative under finite covers, we expect an expression of the form

$$\mathrm{CS}(Q, \rho(k)) \equiv \frac{n}{p} \mod \mathbb{Z}$$

for some integer n, where $\rho : \pi_1(Q) \to U(1)$ is a reducible flat connection. The fundamental group of Q has presentation

$$\pi_1(Q) = \langle x_1, x_2, x_3, h \mid h \text{ central}, x_i^{a_i} h^{\beta_i} = 1, x_1 x_2 x_3 = 1 \rangle$$

	charge-splitting	dimension	energy
A	$\mathcal{M}(X,\alpha_1)\times\mathcal{M}(\alpha_1,\alpha_2)\times\mathcal{M}(\alpha_2,\theta)$	$0+4+1$	$71/120+2/5+1/120$
B	$\mathcal{M}(X,\alpha_1)\times\mathcal{M}(\alpha_1,\theta)$	$0+5$	$71/120+49/120$
C	$\mathcal{M}(X,\alpha_2)\times\mathcal{M}(\alpha_2,\theta)$	$4+1$	$119/120+1/120$
D	$\mathcal{M}(X,\theta)\times\mathcal{M}(\theta,\theta)$	$0+5$	$0+1$

Table 2. All possible energy splitting in the compactification of $\mathcal{M}_1(X,\theta)$. Note that the total energy in each case is 1.

with abelianization $\mathbb{Z}/p$ generated by the regular fiber h. A reducible representation sends $h\mapsto e^{2\pi ik/p}$ for some integer k and $x_i\mapsto 1$.

Theorem 3.3. *The Chern–Simons invariant satisfies* $\mathrm{CS}(Q,\rho(k))\equiv\dfrac{n_0k}{p}\pmod{\mathbb{Z}}$, *where n_0 is an integer such that $n_0a_1a_2a_3\equiv k\pmod{p}$.*

Proof. This congruence is obtained from Auckly's formula [1994] using a representation $\rho(n_0,n_1,n_2,n_3):\pi_1(Q)\to U(1)$, where n_0 satisfies $a_1a_2a_3\cdot n_0\equiv k\pmod{p}$ and $n_i=0$ for $i\neq 0$. The Seifert invariants satisfy

$$\sum_i\frac{\beta_i}{a_i}=\frac{p}{a_1a_2a_3}\tag{3-2}$$

and the formula for the Chern–Simons invariant of the corresponding flat connection is given in [Auckly 1994, p. 234][1] as

$$\mathrm{CS}(Q,\rho)\equiv\sum_{j=1}^{3}\frac{\rho_jn_j^2+n_j(n_0+c/2+\sum_{i=1}^{3}n_i/a_i)/(b+\sum_i\beta_i/a_i)}{a_j}+\frac{(n_0+c/2)(n_0+c/2+\sum_i n_i/a_i)}{b+\sum_i\beta_i/a_i}\pmod{\mathbb{Z}}$$

with $c=0$ and such that ρ_j satisfies $a_j\sigma_j-\beta_j\rho_j=1$ for some integers σ_j. This simplifies to $n_0k/p \bmod \mathbb{Z}$. □

The Chern–Simons invariants for the irreducible flat connections on Q can be computed by using an SO(3) flat cobordism to a disjoint union of lens spaces as with $\Sigma(a_1,a_2,a_3)$, but also again from [Auckly 1994, p. 232]. We now investigate whether any of the charge splittings given in Table 2 contain π-fixed ASD connections. It follows immediately from Lemma 3.2 that case D is ruled out. We now rule out the possibility of a 1-dimensional fixed set in the equivariant moduli space $(\mathcal{M}_1(X,\theta),\pi)$ splitting in case A of Table 2.

[1] Note that Auckly [1994] and Kirk–Klassen [1990] use opposite orientations on the Seifert fibered manifolds than the one used in this paper. As a result the Chern–Simons invariant differs by a sign.

Lemma 3.4. *The moduli space $\mathcal{M}_\ell(\alpha_1, \alpha_2)$ does not support π-fixed ASD connections with energy $\ell = 2/5$ for any odd prime $p \geq 7$.*

Proof. If there exists a π-fixed ASD connection with energy $\ell = 2/5$ in $\mathcal{M}^\pi_\ell(\alpha_1, \alpha_2)$ then it corresponds to an equivariant lift of the π-action to the principal bundle which leaves that connection invariant. Since a π-invariant connection descends to an SO(3) connection on the cylinder $Q \times \mathbb{R}$ where $Q = \Sigma(2, 3, 5)/\pi$ is a rational homology 3-sphere, the moduli space in the quotient must be nonempty. Let α'_1 and α'_2 denote the irreducible limiting flat connections on $Q \times \mathbb{R}$. The connection in the quotient has energy or Pontryagin charge $4\ell/p = 8/5p$, however, a nonempty moduli space must have energy that is congruent modulo $4\mathbb{Z}$ (see [Saveliev 2002, Remark 5.6, p. 102]) to the difference of the SO(3) Chern–Simons invariants $\mathrm{CS}(Q, \alpha'_2) - \mathrm{CS}(Q, \alpha'_1)$. It follows from Auckly's formula that this difference has the form $n/30$ for some integer n. But

$$\frac{n}{30} \not\equiv \frac{8}{5p} \mod 4\mathbb{Z}$$

since the former has denominator at most 30 and for the latter $p \geq 7$. It must be that the moduli space $\mathcal{M}^\pi_\ell(\alpha_1, \alpha_2)$ is empty. □

It remains to investigate the cases $\mathcal{M}_\ell(\alpha_i, \theta)$. The next proposition is more general and gives a necessary condition for $\Sigma(a_1, a_2, a_3) \times \mathbb{R}$ with an irreducible flat limit at $-\infty$ and the trivial connection θ at $+\infty$ to admit π-invariant ASD connections: the numerator in the energy must be a square integer. Since every reducible flat SU(2) connection on Q sends h to $\exp(2\pi ik/p)$ in $U(1) \subset \mathrm{SU}(2)$, the integer k is referred to as the holonomy number of the flat connection. The SO(3) holonomy number is obtained by applying the adjoint representation $\mathrm{ad} : \mathrm{SU}(2) \to \mathrm{SO}(3)$.

Proposition 3.5. *Suppose a principal* SU(2) *bundle over* $\Sigma(a_1, a_2, a_3) \times \mathbb{R}$ *admits* π*-invariant ASD connections with energy*

$$\ell \equiv \frac{e^2}{4a_1a_2a_3} \in (0, 1]$$

asymptotic to an irreducible flat connection α at $-\infty$ and the trivial connection at $+\infty$. Then this connection descends to an SO(3) *ASD connection on the quotient $Q \times \mathbb{R}$ with energy $4\ell/p$ which limits to an irreducible connection still denoted by α at $-\infty$ and a flat $U(1)$-reducible connection β at $+\infty$ which has* SO(3) *holonomy number $\pm e$* (mod p).

Proof. Since an invariant connection descends to an SO(3) ASD connection, the moduli space in the quotient is nonempty; this again gives the relation between the SO(3) Chern–Simons invariants

$$\mathrm{CS}(Q, \beta) - \mathrm{CS}(Q, \alpha) \equiv \frac{4\ell}{p} \equiv \frac{e^2}{pa_1a_2a_3} \mod 4\mathbb{Z} \tag{3-3}$$

But the Chern–Simons invariant of the reducible connection is given by

$$\mathrm{CS}(Q, \beta(k)) \equiv \frac{n_0 k}{p}$$

for some integer n_0 such that $n_0(a_1a_2a_3) \equiv k \pmod p$ and where k is the SO(3) holonomy number of the representation $\beta(k)$. On the other hand,

$$\mathrm{CS}(Q, \alpha) \equiv \frac{m}{a_1a_2a_3}$$

for some integer m. Taking the difference gives the relation

$$\frac{n_0k(a_1a_2a_3)-mp}{p(a_1a_2a_3)} \equiv \frac{e^2}{p(a_1a_2a_3)} \quad \mathrm{mod}\ 4\mathbb{Z}. \tag{3-4}$$

This implies that the numerators are congruent modulo $4p(a_1a_2a_3)\mathbb{Z}$ and

$$k^2 \equiv e^2 \pmod p. \tag{3-5}$$

Since $\mathbb{Z}/p$ has no zero divisors completes the proof. □

Proof of Theorem A. Suppose there exists a smooth extension to X_0 with isolated fixed points. If a fixed point of the π-action on X_0 has rotation numbers (a, b), where a, b are nonzero integers well-defined modulo p, then there is an equivariant lift corresponding to the 1-parameter family of π-fixed ASD connections in $\mathcal{M}_1^\pi(X, \theta)$ that it generates at the Taubes boundary. This is a $\tilde{\pi}$-action on the principal SU(2) bundle and has isotropy representation over the fixed point with weights $\pm(b-a)$ and action on

$$P|_{\mathrm{End}(X)} = \Sigma(2, 3, 5) \times [0, \infty) \times \mathrm{SU}(2)$$

is given by

$$\tilde{t} \cdot (x, s, U) = (tx, s, \phi(\tilde{t})U), \tag{3-6}$$

where $s \in \mathbb{R}$, $U \in \mathrm{SU}(2)$, and ϕ is the isotropy representation $\tilde{\pi} \to \mathrm{SU}(2)$ at ∞ with weights $\pm(a+b)$. We can mod out by the involution to get the π-equivariant adjoint SO(3)-bundle over $\Sigma(2, 3, 5) \times \mathbb{R}$ with action given by the adjoint representation sending t to $\mathrm{Diag}(1, t^{a+b})$ with $\mathbb{Z}/p = \langle t \rangle$. In the limit at $+\infty$ on $\Sigma(2, 3, 5) \times \mathbb{R}$ the trivial product connection descends to a flat reducible connection on Q whose SO(3) holonomy representation is isomorphic to the adjoint isotropy representation $\mathrm{ad}\,\phi$. Since this holonomy is either ± 1 and $\pm 7 \pmod p$ this completes the proof. □

The equivariant plumbing actions predict the existence of nonempty Floer type moduli spaces with fractional Yang–Mills energy; these dimensions can be computed by an index calculation using [Atiyah et al. 1975a]:

$$\dim \mathcal{M}_{4\ell/p}(Q \times \mathbb{R}, \alpha, \beta) = \tfrac{8\ell}{p} - \tfrac{1}{2}\big(h_\alpha + h_\beta\big) + \tfrac{1}{2}\big(\rho_\beta(Q) - \rho_\alpha(Q)\big). \tag{3-7}$$

Since α is irreducible and β is reducible, $h_\alpha = 0$ and $h_\beta = 1$. The rho invariants for reducible flat connections are determined using [Kwasik and Lawson 1993, p. 40]:

$$\rho_\beta(Q)(l) = -\frac{2}{p}\sum_{k=1}^{p-1}\sin^2\frac{\pi kl}{p} + \frac{2}{30p}\sum_{k=1}^{p-1}\csc^2\frac{\pi k}{p}\sin^2\frac{\pi kl}{p} + \sum_{i=1}^{3}\frac{2}{pa_i}\sum_{m_1=0}^{p-1}\sum_{m_2=1}^{a_{i-1}}\cot\frac{\pi m_2}{a_i}\cot\Big(\frac{\pi m_1}{p}-\frac{\pi m_2 b_i}{a_i}\Big)\sin^2\frac{\pi m_1 l}{p}, \tag{3-8}$$

where l is the rotation number for the holonomy representation of β in SO(3). For irreducible flat connections α, the rho invariants can be calculated by an SO(3)-flat cobordism to a union of lens spaces $L(a_i, pb_i)$ using the mapping cylinder for the Seifert fibration of Q [Yu 1991], as in the case of $\Sigma(a_1, a_2, a_3)$ [Saveliev 2002, p. 144]. In this way, the linear equivariant plumbing actions imply that the moduli space $\mathcal{M}_\ell(Q\times\mathbb{R}, \alpha_2, \beta)$ for $\ell = 1/120$ is nonempty with

$$\dim\mathcal{M}_\ell(Q\times\mathbb{R}, \alpha_2, \beta) = \tfrac{8}{p}\big(\tfrac{1}{120}\big) - \tfrac{1}{2} + \tfrac{1}{2}\big(\rho_\beta(Q)(1) - \rho_{\alpha_2}(Q)\big) = 1. \tag{3-9}$$

If we now imagine a nonlinear smooth π-extension to X_0, we do not know if $\mathcal{M}_\ell(Q\times\mathbb{R}, \alpha_1, \beta)$ for $\ell = 49/120$ is nonempty but we have the following formal dimension:

$$\dim\mathcal{M}_\ell(Q\times\mathbb{R}, \alpha_1, \beta) = \tfrac{8}{p}\big(\tfrac{49}{120}\big) - \tfrac{1}{2} + \tfrac{1}{2}\big(\rho_\beta(Q)(7) - \rho_{\alpha_1}(Q)\big) = 1. \tag{3-10}$$

We summarize this in the following theorem. The irreducible flat connections α_1 and α_2 on $\Sigma(2,3,5)$ descend to irreducible flat connections on the quotient $\Sigma(2,3,5)/\pi$, which we still denote by α_i.

Theorem 3.6. *Let Q denote the rational homology sphere quotient $\Sigma(2,3,5)/\pi$ and let $\ell = 49/120$. When the holonomy representation of the flat connection β is $\pm 7 \pmod p$, the formal dimension of the moduli space $\mathcal{M}_{4\ell/p}(Q\times\mathbb{R}, \alpha_1, \beta)$ of* SO(3)*-ASD connections on the cylinder $Q\times\mathbb{R}$ with energy $4\ell/p$ that limit to α_1 at $-\infty$ and to a reducible connection β at $+\infty$ is* 1. *Similarly, when $\ell = 1/120$ and the holonomy representation of the flat connection β is $\pm 1 \pmod p$, the formal dimension of $\mathcal{M}_{4\ell/p}(Q\times\mathbb{R}, \alpha_2, \beta)$ is* 1.

We have obtained congruence relations that give constraints on the rotation data for the fixed points of a smooth extension. The next natural step is to check these constraints against the G-signature formula and we do this in the next section.

4. G-signature for 4-manifolds with boundary

For smooth, closed, orientable 4-manifolds X, recall that the Hodge star operator induces an involution τ on the space of complexified sections of forms $\Omega^* =$

$\bigoplus_k C^\infty(\Lambda^k TX \otimes \mathbb{C})$, splitting it into ± 1 eigenspaces $\Omega^+ \oplus \Omega^-$. The signature operator $D^+ = d + d^*$ restricted to Ω^+ is an elliptic operator $D^+ : \Omega^+ \to \Omega^-$ whose index is the signature $\mathrm{Sign}(X) = b_2^+ - b_2^-$ of the nondegenerate quadratic form on $H^2(X; \mathbb{R})$. When a finite group G acts by orientation preserving isometries on X, the cotangent bundle, as an equivariant bundle over X, has an action that commutes with the Hodge star operator. So we obtain a G-invariant elliptic operator D^+ whose G-index is a complex virtual representation $\mathrm{Ind}_G(D^+) = H^+ - H^- \in R(G)$. The associated character, or Lefschetz number,

$$\mathrm{Sign}(X, g) = \mathrm{Tr}(g|_{H^2_+}) - \mathrm{Tr}(g|_{H^2_-})$$

is the g-signature. Note that when the action of G is homologically trivial, the g-signature coincides with the usual signature.

The g-signature can be computed from the fixed set by the Atiyah–Singer fixed point index theorem [1968]. Consider the case when G is a finite cyclic group of odd prime order p with generator $t = e^{2\pi i/p}$ and let $T_{p_i} X = \mathbb{C}^2(a_i, b_i)$ be the local tangential representation over the fixed points p_i for a homologically trivial action. Then

$$\mathrm{Sign}(X) = \sum_i \left(\frac{t^{a_i}+1}{t^{a_i}-1}\right)\left(\frac{t^{b_i}+1}{t^{b_i}-1}\right) - 4\sum_j \frac{\alpha_j t^{c_j}}{(t^{c_j}-1)^2} \tag{4-1}$$

where, for each j, α_j is the self-intersection of the fixed 2-sphere and c_j is the rotation number on its normal bundle.

Consider the situation where X_0 denotes a compact, simply connected, smooth 4-manifold with boundary $\partial X_0 = \Sigma$ an integral homology 3-sphere. If a free action on Σ by $\mathbb{Z}/p = \langle t \rangle$ extends to a locally linear, homologically trivial action on X_0 (not necessarily free), then the G-signature theorem for manifolds with boundary is given in Atiyah–Patodi–Singer [Atiyah et al. 1975b, p. 413]:

$$\mathrm{Sign}(X_0, t) = L(X_0, t) - \eta_t(0) \tag{4-2}$$

where $L(X_0, t)$ is the collection of terms occurring in the closed manifold case and $\eta_t(0)$ is the equivariant eta invariant of Σ or the G-signature defect. This invariant depends only on the 3-manifold Σ and not on how the action extends to X_0 nor on which 4-manifold it equivariantly bounds. To see this, suppose the action on Σ extends to another 4-manifold X_1. Then consider the G-signature theorem on $X_0 \cup_\Sigma -X_1$ to see that the signature defect terms are equal.

When the boundary ∂X_0 is a Seifert fibered homology sphere — thought of as a link of a complex surface singularity — it has a canonical negative definite resolution $\tilde{X}_0$ obtained by plumbing disk bundles over 2-spheres. Let $(\tilde{a}_i, \tilde{b}_i)$ denote the rotation numbers by equivariant plumbing [Fintushel 1977, p. 152] along the

E_8 graph. They are given by

$$\{(-4,5),(-3,4),(-2,3),(-2,3),(-1,2),(-1,2),(-1,2)\}.$$

The central node in the plumbing graph is a fixed 2-sphere with self-intersection number -2, and the rotation on the normal fiber is congruent to $1 \pmod{p}$. This gives the following formula for the equivariant eta invariant:

$$\eta_t(0) = \sum_{i=1}^{7} \left(\frac{t^{\tilde{a}_i}+1}{t^{\tilde{a}_i}-1} \right) \left(\frac{t^{\tilde{b}_i}+1}{t^{\tilde{b}_i}-1} \right) + \frac{8t}{(t-1)^2} + 8.$$

Proof of Theorem B. Suppose a free $\mathbb{Z}/7$-action on $\Sigma(2,3,5)$ extends to a smooth homologically trivial action on X_0 with fixed set consisting of only isolated fixed points with rotation data $\{(a_i,b_i)\}_{i=1}^{9}$. By Theorem A these rotation numbers must satisfy the congruence relations $a_i + b_i \equiv \pm 1$ or $0 \pmod{7}$. There are three types of rotation numbers that satisfy the first constraint: $(1,5)$, $(2,4)$, $(3,3)$ and three types that satisfy the second constraint:$(1,6)$, $(2,5)$, $(3,4)$. By the G-signature theorem, the rotation numbers must satisfy

$$-8 = \sum_{i=1}^{9} \left(\frac{t^{a_i}+1}{t^{a_i}-1} \right) \left(\frac{t^{b_i}+1}{t^{b_i}-1} \right) - \eta_t(0), \tag{4-3}$$

where $\eta_t(0)$ is determined from the equivariant plumbing action. One may check this formula directly for all possible rotation data of the types listed above. There will be 2002 independent G-signature checks since repetition of rotation numbers is allowed. This number may be significantly cut down in the following way. In (4-3) we can sum over nonidentity roots of unity to obtain

$$-8(p-1) = \sum_{i=1}^{9} \operatorname{def}(p; a_i, b_i) - \sum_{\substack{t^p=1 \\ t \neq 1}} \eta_{t^i}, \tag{4-4}$$

where

$$\operatorname{def}(p; a, b) = -\sum_{k=1}^{p-1} \cot \frac{\pi a k}{p} \cot \frac{\pi b k}{p}$$

are the G-signature defects. The second term above can easily be computed from the formula for η_t above and gives $\sum_{t^p=1, t\neq 1} \eta_{t^i} = 6$. The defect terms for the rotation data that satisfy the first constraint can also be computed for $p = 7$ and we have $\operatorname{def}(7;1,5) = 2 = -\operatorname{def}(7;2,4)$, $\operatorname{def}(7;3,3) = -10$. Similarly for the second constraint, all the defect terms sum to 10. If n_i are the number of rotation numbers of the types listed above respectively, we have the following system of linear Diophantine equations

$$-21 = n_1 - n_2 - 5n_3 + 5(n_4 + n_5 + n_6) \tag{4-5}$$

$$9 = n_1 + n_2 + n_3 + n_4 + n_5 + n_6. \tag{4-6}$$

The second equation holds since there must be 9 isolated fixed points for a homologically trivial extension. Note that not all rotation numbers can satisfy the second constraint $a_i + b_i \equiv 0 \pmod{p}$, since the left-hand side of (4-5) would not be divisible by 5. So there must be rotation numbers of the first type. There are 12 solutions $(n_1, n_2, n_3, n_4, n_5, n_6)$ to this system in total, which may be enumerated as follows:

$$(0, 1, 6, 0, 0, 2), (0, 1, 6, 0, 1, 1), (0, 1, 6, 0, 2, 0), (0, 1, 6, 1, 0, 1),$$
$$(0, 1, 6, 1, 1, 0), (0, 1, 6, 2, 0, 0), (0, 6, 3, 0, 0, 0), (1, 2, 5, 0, 0, 1),$$
$$(1, 2, 5, 0, 1, 0), (1, 2, 5, 1, 0, 0), (2, 3, 4, 0, 0, 0), (4, 0, 5, 0, 0, 0).$$

One can check that none of these candidates satisfies the G-signature formula in (4-3). This concludes the proof since we have shown that there are no solutions to the G-signature formula that satisfy the constraints for a smooth, homologically trivial extension. □

Proof of Corollary C. Suppose that (X, π) is a smooth, homologically trivial π-action on $X = \#^8 S^2 \times S^2$ with fixed set consisting of isolated fixed points. If $\Sigma(2, 3, 5)$ with a free π-action smoothly and equivariantly embeds in (X, π), we obtain a π-equivariant decomposition $X = X_0 \cup_{\Sigma(2,3,5)} X_1$ with intersection forms $\pm E_8$ on each side by Remark 1.2. If X_0 is not simply connected then, by van Kampen's theorem, there are no nontrivial representations $\pi_1(X_0) \to SU(2)$ whose restriction to $\Sigma(2, 3, 5)$ are trivial. In particular, no additional flat connections appear in the charge splitting case D and the corollary follows from Theorems A and B. □

Acknowledgements

This work was part of the author's PhD thesis under the supervision of Prof. Ian Hambleton at McMaster University. The author is grateful for his great patience, expertise and support while working on this project. Partial support was provided by Ontario Graduate Scholarship (OGS 2010–2012). The author would also like to thank Prof. Nikolai Saveliev for many helpful discussions and Prof. Ron Fintushel for pointing out the application of Theorem A to splitting $\#^8 S^2 \times S^2$; our original motivation was to consider equivariant embedding of $\Sigma(2, 7, 13)$ in homotopy $K3$ surfaces. A portion of this paper was written at the Max Planck Institute for Mathematics in Bonn, which the author wishes to thank for hospitality during his visit. Finally, the author is grateful to the referees for helpful comments and suggestions.

References

[Atiyah and Singer 1968] M. F. Atiyah and I. M. Singer, "The index of elliptic operators, III", *Ann. of Math.* (*2*) **87** (1968), 546–604. MR 38 #5245 Zbl 0164.24301

[Atiyah et al. 1975a] M. F. Atiyah, V. K. Patodi, and I. M. Singer, "Spectral asymmetry and Riemannian geometry. I", *Math. Proc. Cambridge Philos. Soc.* **77**:1 (1975), 43–69. MR 53 #1655a Zbl 0297.58008

[Atiyah et al. 1975b] M. F. Atiyah, V. K. Patodi, and I. M. Singer, "Spectral asymmetry and Riemannian geometry. II", *Math. Proc. Cambridge Philos. Soc.* **78**:3 (1975), 405–432. MR 53 #1655b Zbl 0314.58016

[Auckly 1994] D. R. Auckly, "Topological methods to compute Chern–Simons invariants", *Math. Proc. Cambridge Philos. Soc.* **115**:2 (1994), 229–251. MR 95f:57034 Zbl 0854.57013

[Bierstone 1977] E. Bierstone, "General position of equivariant maps", *Trans. Amer. Math. Soc.* **234**:2 (1977), 447–466. MR 57 #4221 Zbl 0318.57044

[Braam and Matić 1993] P. J. Braam and G. Matić, "The Smith conjecture in dimension four and equivariant gauge theory", *Forum Math.* **5**:3 (1993), 299–311. MR 94d:57063 Zbl 0779.57011

[Buchdahl et al. 1990] N. P. Buchdahl, S. Kwasik, and R. Schultz, "One fixed point actions on low-dimensional spheres", *Invent. Math.* **102**:3 (1990), 633–662. MR 92b:57047 Zbl 0735.57021

[Cho 1991] Y. S. Cho, "Finite group actions on the moduli space of self-dual connections. I", *Trans. Amer. Math. Soc.* **323**:1 (1991), 233–261. MR 91d:58030 Zbl 0724.57013

[Curtis and Reiner 1962] C. W. Curtis and I. Reiner, *Representation theory of finite groups and associative algebras*, Pure and Applied Mathematics **11**, Interscience, New York, 1962. Reprinted by American Mathematical Society, Providence, RI, 2006. MR 26 #2519 Zbl 0131.25601

[Donaldson 1983] S. K. Donaldson, "An application of gauge theory to four-dimensional topology", *J. Differential Geom.* **18**:2 (1983), 279–315. MR 85c:57015 Zbl 0507.57010

[Donaldson 1987] S. K. Donaldson, "The orientation of Yang–Mills moduli spaces and 4-manifold topology", *J. Differential Geom.* **26**:3 (1987), 397–428. MR 88j:57020 Zbl 0683.57005

[Donaldson 2002] S. K. Donaldson, *Floer homology groups in Yang–Mills theory*, Cambridge Tracts in Mathematics **147**, Cambridge University Press, 2002. MR 2002k:57078 Zbl 0998.53057

[Donaldson and Kronheimer 1990] S. K. Donaldson and P. B. Kronheimer, *The geometry of four-manifolds*, Clarendon/Oxford University Press, New York, 1990. MR 92a:57036 Zbl 0820.57002

[Edmonds 1989] A. L. Edmonds, "Aspects of group actions on four-manifolds", *Topology Appl.* **31**:2 (1989), 109–124. MR 90h:57050 Zbl 0691.57015

[Fintushel 1977] R. Fintushel, "Circle actions on simply connected 4-manifolds", *Trans. Amer. Math. Soc.* **230** (1977), 147–171. MR 56 #16659 Zbl 0362.57014

[Fintushel and Lawson 1986] R. Fintushel and T. Lawson, "Compactness of moduli spaces for orbifold instantons", *Topology Appl.* **23**:3 (1986), 305–312. MR 88e:57017 Zbl 0664.57006

[Fintushel and Stern 1990] R. Fintushel and R. J. Stern, "Instanton homology of Seifert fibred homology three spheres", *Proc. London Math. Soc.* (*3*) **61**:1 (1990), 109–137. MR 91k:57029 Zbl 0705.57009

[Floer 1988] A. Floer, "An instanton-invariant for 3-manifolds", *Comm. Math. Phys.* **118**:2 (1988), 215–240. MR 89k:57028 Zbl 0684.53027

[Freed and Uhlenbeck 1991] D. S. Freed and K. K. Uhlenbeck, *Instantons and four-manifolds*, 2nd ed., Mathematical Sciences Research Institute Publications **1**, Springer, New York, 1991. MR 91i:57019 Zbl 0559.57001

[Hambleton and Lee 1992] I. Hambleton and R. Lee, "Perturbation of equivariant moduli spaces", *Math. Ann.* **293**:1 (1992), 17–37. MR 93f:57023 Zbl 0734.57013

[Hambleton and Lee 1995] I. Hambleton and R. Lee, "Smooth group actions on definite 4-manifolds and moduli spaces", *Duke Math. J.* **78**:3 (1995), 715–732. MR 96m:57058 Zbl 0849.57033

[Hambleton and Riehm 1978] I. Hambleton and C. Riehm, "Splitting of Hermitian forms over group rings", *Invent. Math.* **45**:1 (1978), 19–33. MR 58 #2840 Zbl 0377.15010

[Kirk and Klassen 1990] P. A. Kirk and E. P. Klassen, "Chern–Simons invariants of 3-manifolds and representation spaces of knot groups", *Math. Ann.* **287**:2 (1990), 343–367. MR 91d:57008 Zbl 0681.57006

[Kwasik and Lawson 1993] S. Kwasik and T. Lawson, "Nonsmoothable Z_p actions on contractible 4-manifolds", *J. Reine Angew. Math.* **437** (1993), 29–54. MR 94d:57035 Zbl 0764.57024

[Lawson 1985] H. B. Lawson, Jr., *The theory of gauge fields in four dimensions*, CBMS Regional Conference Series in Mathematics **58**, American Mathematical Society, Providence, RI, 1985. MR 87d:58044 Zbl 0597.53001

[Morgan et al. 1994] J. W. Morgan, T. S. Mrowka, and D. Ruberman, *The L^2-moduli space and a vanishing theorem for Donaldson polynomial invariants*, Monographs in Geometry and Topology **2**, International Press, Cambridge, MA, 1994. MR 95h:57039 Zbl 0830.58005

[Quebbemann 1981] H.-G. Quebbemann, "Zur Klassifikation unimodularer Gitter mit Isometrie von Primzahlordnung", *J. Reine Angew. Math.* **326** (1981), 158–170. MR 82m:10039 Zbl 0452.10027

[Saveliev 2002] N. Saveliev, *Invariants for homology 3-spheres*, Encyclopaedia of Mathematical Sciences **140**, Springer, Berlin, 2002. MR 2004c:57026 Zbl 0998.57001

[Taubes 1987] C. H. Taubes, "Gauge theory on asymptotically periodic 4-manifolds", *J. Differential Geom.* **25**:3 (1987), 363–430. MR 88g:58176 Zbl 0615.57009

[Yu 1991] B. Z. Yu, "A note on an invariant of Fintushel and Stern", *Topology Appl.* **38**:2 (1991), 137–145. MR 92c:57020 Zbl 0783.57007

Received March 19, 2014. Revised November 5, 2015.

NIMA ANVARI
MATHEMATICS
UNIVERSITY OF MIAMI
1365 MEMORIAL DRIVE
CORAL GABLES, FL 33146
UNITED STATES

anvarin@math.miami.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 282, No. 1, 2016
dx.doi.org/10.2140/pjm.2016.282.27

A SHORT PROOF OF THE EXISTENCE OF SUPERCUSPIDAL REPRESENTATIONS FOR ALL REDUCTIVE p-ADIC GROUPS

RAPHAËL BEUZART-PLESSIS

Let G be a reductive p-adic group. We give a short proof of the fact that G always admits supercuspidal complex representations. This result has already been established by A. Kret using the Deligne–Lusztig theory of representations of finite groups of Lie type. Our argument is of a different nature and is self-contained. It is based on the Harish-Chandra theory of cusp forms and it ultimately relies on the existence of elliptic maximal tori in G.

Let p be a prime number and let F be a p-adic field (i.e., a finite extension of $\mathbb{Q}_p$). We denote by $\mathbb{O}$ the ring of integers of F and we fix a uniformizer $\varpi \in \mathbb{O}$. We also denote by $\mathrm{val} : F^\times \to \mathbb{Z}$ the normalized valuation (i.e., $\mathrm{val}(\varpi) = 1$). Let G be a connected reductive group defined over F. We will denote by $\mathfrak{g}$ the Lie algebra of G. A sentence like "Let $P = MN$ be a parabolic subgroup of G" will mean as usual that P is a parabolic subgroup of G defined over F, that N is its unipotent radical and that M is a Levi component of P also defined over F. More generally, all subgroups of G that we consider will be defined over F. We will also need to fix Haar measures on the various groups that we consider. The precise normalization of these Haar measures won't be important (unless we specify that they need to satisfy an explicit compatibility condition) and we will only make use of Haar measures on unimodular groups (e.g., F points of reductive or unipotent groups), so that the distinction between left and right Haar measures is irrelevant here and will be dropped from the notations.

Remark 1. We exclude fields of positive characteristic because we will use in a crucial way the exponential map. If $G = \mathrm{GL}_n$, then we can instead use the map $X \mapsto \mathrm{Id} + X$ and work over any nonarchimedean local field. For classical groups,

---

This work was supported by the Gould Fund and by the National Science Foundation under agreement No. DMS-1128155. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

MSC2010: primary 22E50; secondary 11F85.

Keywords: p-adic groups, supercuspidal representations, cusp forms.

we could probably also replace the exponential map by some Cayley map and considerably weaken the characteristic assumption.

Recall that a smooth representation of $G(F)$ is a pair (π, V_π) where V_π is a complex vector space (usually infinite dimensional) and $\pi : G(F) \to \mathrm{GL}(V_\pi)$ is a morphism such that for all vectors $v \in V_\pi$ the stabilizer $\mathrm{Stab}_{G(F)}(v)$ of v in $G(F)$ is an open subgroup. Let (π, V_π) be a smooth representation of $G(F)$ and let $P = MN$ be a parabolic subgroup of G. The Jacquet module of (π, V_π) with respect to P is the space of coinvariants

$$V_{\pi,N} = V_\pi / V_\pi(N),$$

where $V_\pi(N)$ is the subspace of V_π generated by the elements $v - \pi(n)v$ for all $v \in V_\pi$ and all $n \in N(F)$. It is the biggest quotient of V_π on which $N(F)$ acts trivially. There is a natural linear action π_N of $M(F)$ on $V_{\pi,N}$ and $(\pi_N, V_{\pi,N})$ is a smooth representation of $M(F)$. The functor $V_\pi \mapsto V_{\pi,N}$ is an exact functor from the category of smooth representations of $G(F)$ to the category of smooth representations of $M(F)$. Indeed, this follows rather easily from the following fact (see [Renard 2010, Proposition III.2.9]):

(1) Let $(N(F)_k)_{k\geqslant 0}$ be an increasing sequence of compact-open subgroups of $N(F)$ such that $N(F) = \bigcup_{k\geqslant 0} N(F)_k$ (such a sequence always exists). Then a vector $v \in V_\pi$ belongs to $V_\pi(N)$ if and only if there exists $k \geqslant 0$ such that

$$\int_{N(F)_k} \pi(n)v\, dn = 0.$$

Let (π, V_π) be an irreducible smooth representation of $G(F)$ (irreducible means that V_π is nontrivial and that it has no nonzero proper $G(F)$-invariant subspace). The representation (π, V_π) is said to be supercuspidal if for all proper parabolic subgroups $P = MN$ of G, the Jacquet module $V_{\pi,N}$ is zero. An equivalent condition is that the coefficients of (π, V_π) are compactly supported modulo the center (see [Renard 2010, Theorem VI.2.1]).

The purpose of this short note is to prove the following result.

Theorem 2. *$G(F)$ admits irreducible supercuspidal representations.*

Remark 3. This theorem has already been proved by A. Kret [2012]. The proof of Kret has the advantage of working in any characteristic (cf. Remark 1) and of being explicit (i.e., it exhibits a way to construct such supercuspidal representations by compact induction). Kret's proof eventually relies on the Deligne–Lusztig theory of representations of finite groups of Lie type and so can hardly be considered elementary. Although less complete and explicit than the results of [Kret 2012], the proof presented here has the advantage of being short and (almost) self-contained, using only elementary harmonic analysis arguments.

We will deduce Theorem 2 from the existence of nonzero compactly supported cusp forms, in the sense of Harish-Chandra, for the group $G(F)$. Before stating this existence result, we need to introduce some more definitions and notation. We will denote, as usual, by $C_c^\infty(G(F))$ the space of complex-valued functions on $G(F)$ that are smooth, i.e., locally constant, and compactly supported. We say that a function $f \in C_c^\infty(G(F))$ is a cusp form if for all proper parabolic subgroups $P = MN$ of G, we have

$$\int_{N(F)} f(xn)\, dn = 0, \quad \text{for all } x \in G(F)$$

(these functions are called supercusp forms in [Harish-Chandra 1970]). We shall denote by $C_{c,\mathrm{cusp}}^\infty(G(F)) \subseteq C_c^\infty(G(F))$ the subspace of cusp forms. As we said, Theorem 2 will follow from the following existence theorem.

Theorem 4. *We have $C_{c,\mathrm{cusp}}^\infty(G(F)) \neq 0$.*

Proof that Theorem 4 implies Theorem 2. Let us denote by ρ the action of $G(F)$ on $C_c^\infty(G(F))$ by right translation. Then $\big(\rho, C_c^\infty(G(F))\big)$ is a smooth representation of $G(F)$. Moreover, it is easy to see that the subspace $C_{c,\mathrm{cusp}}^\infty(G(F)) \subseteq C_c^\infty(G(F))$ is $G(F)$-invariant. We claim the following:

(2) For all proper parabolic subgroups $P = MN$ of G, we have

$$C_{c,\mathrm{cusp}}^\infty(G(F))_N = 0.$$

Let $P = MN$ be a proper parabolic subgroup of G and fix an increasing sequence $(N(F)_k)_{k\geqslant 0}$ of compact-open subgroups of $N(F)$ such that $N(F) = \bigcup_{k\geqslant 0} N(F)_k$. Let $f \in C_{c,\mathrm{cusp}}^\infty(G(F))$. By (1), it suffices to show the existence of an integer $k \geqslant 0$ such that

$$\int_{N(F)_k} \rho(n) f\, dn = 0,$$

or, what amounts to the same,

$$\int_{N(F)_k} f(xn)\, dn = 0, \quad \text{for all } x \in G(F). \tag{3}$$

Since $\mathrm{Supp}(f)$ (the support of the function f) is compact, there exists $k \geqslant 0$ such that

$$\mathrm{Supp}(f) \cap \big[\mathrm{Supp}(f)\big(N(F) \setminus N(F)_k\big)\big] = \varnothing. \tag{4}$$

We now show that (3) is satisfied for such k. Let $x \in G(F)$. If $x \notin \mathrm{Supp}(f)N(F)_k$, the term inside the integral (3) vanishes identically and there is nothing to prove. Assume from now on that $x \in \mathrm{Supp}(f)N(F)_k$. Up to translating x by an element of $N(F)_k$, we may as well assume that $x \in \mathrm{Supp}(f)$. Then, by (4) we have $xn \notin \mathrm{Supp}(f)$ for all $n \in N(F) \setminus N(F)_k$. It follows that

$$\int_{N(F)_k} f(xn)\,dn = \int_{N(F)} f(xn)\,dn.$$

But by definition of $C^\infty_{c,\mathrm{cusp}}(G(F))$, this last integral vanishes. This proves (3) and ends the proof of (2).

We now show how to deduce from (2) that Theorem 4 implies Theorem 2. Assume that Theorem 4 is satisfied. Then, we can find $f \in C^\infty_{c,\mathrm{cusp}}(G(F))$ which is nonzero. Denote by V_f the $G(F)$-invariant subspace of $C^\infty_{c,\mathrm{cusp}}(G(F))$ generated by f and let $V \subseteq V_f$ be a maximal $G(F)$-invariant subspace among those not containing f (which exists by Zorn's lemma). Then V_f/V is a smooth irreducible representation of $G(F)$. We claim that it is supercuspidal. Indeed, let $P = MN$ be a proper parabolic subgroup of G. By (2) and since the Jacquet module's functor is exact, we have $(V_f/V)_N = 0$. Thus, V_f/V is indeed a supercuspidal representation of $G(F)$ and this proves Theorem 2. □

We are now left with proving Theorem 4. The strategy is to prove first an analog result on the Lie algebra and then lift it to the group by means of the exponential map. Let $C^\infty_c(\mathfrak{g}(F))$ be the space of complex-valued smooth and compactly supported functions on $\mathfrak{g}(F)$. We say that a function $\varphi \in C^\infty_c(\mathfrak{g}(F))$ is a cusp form if for all proper parabolic subgroups $P = MN$ of G, we have

$$\int_{\mathfrak{n}(F)} \varphi(X+N)\,dn = 0, \quad \text{for all } X \in \mathfrak{g}(F),$$

where $\mathfrak{n}(F)$ denotes the F-points of the Lie algebra of N. We will denote by $C^\infty_{c,\mathrm{cusp}}(\mathfrak{g}(F)) \subseteq C^\infty_c(\mathfrak{g}(F))$ the subspace of cusp forms. The analog of Theorem 4 for the Lie algebra is the following lemma.

Lemma 5. *We have $C^\infty_{c,\mathrm{cusp}}(\mathfrak{g}(F)) \neq 0$.*

Before proving this lemma, we first explain how it implies Theorem 4.

Proof that Lemma 5 implies Theorem 4. Assume that Lemma 5 holds. Then, we can find a nonzero function $\varphi \in C^\infty_{c,\mathrm{cusp}}(\mathfrak{g}(F))$. The idea is to lift φ to a function on $G(F)$ using the exponential map. Of course, the exponential map is not necessarily defined on the support of φ. Hence, we need first to scale the function φ so that its support becomes small. Let us fix an element $\lambda \in F^\times$ which we will eventually assume to be sufficiently small. We define the function φ_λ by

$$\varphi_\lambda(X) = \varphi(\lambda^{-1}X), \quad X \in \mathfrak{g}(F).$$

We easily check that φ_λ is still a cusp form. Recall that there exists an open neighborhood $\omega \subseteq \mathfrak{g}(F)$ of 0 on which the exponential map exp is defined and such that it realizes an F-analytic isomorphism

$$\exp : \omega \simeq \Omega,$$

where $\Omega = \exp(\omega)$. Since $\mathrm{Supp}(\varphi_\lambda) = \lambda\,\mathrm{Supp}(\varphi)$, for λ sufficiently small we have

$$\mathrm{Supp}(\varphi_\lambda) \subseteq \omega.$$

We henceforth assume that λ is that sufficiently small. This allows us to define a function f_λ on $G(F)$ by setting

$$f_\lambda(g) = \begin{cases} \varphi_\lambda(X) & \text{if } g = \exp(X) \text{ for some } X \in \omega, \\ 0 & \text{otherwise,} \end{cases}$$

for all $g \in G(F)$. Note that we have $f_\lambda \in C_c^\infty(G(F))$, and obviously the function f_λ is nonzero. Hence, we will be done if we can prove the following:

(5) If λ is sufficiently small, the function f_λ is a cusp form.

Let us denote by $\log : \Omega \to \omega$ the inverse of the exponential map. Then, by the Campbell–Hausdorff formula, it is easy to see that we can find an $\mathcal{O}$-lattice L in the F-vector space $\mathfrak{g}(F)$ which is contained in ω and satisfies the following condition:

$$\log(e^X e^Y) \in X + Y + \varpi^{\mathrm{val}_L(X) + \mathrm{val}_L(Y)} L \tag{6}$$

for all $X, Y \in L$, where we have set $\mathrm{val}_L(X) = \sup\{k \in \mathbb{Z} : X \in \varpi^k L\}$ for all $X \in \mathfrak{g}(F)$. For all integers $n \geqslant 0$, set $K_n = \exp(\varpi^n L)$. It is easy to infer from (6) that K_n is an open-compact subgroup of $G(F)$ for all $n \geqslant 0$. Since φ is smooth and compactly supported, there exists $n_0 \geqslant 0$ such that translation by $\varpi^{n_0} L$ leaves φ invariant. Also, since φ is compactly supported, there exists $n_1 \geqslant 0$ such that $\mathrm{Supp}(\varphi) \subseteq \varpi^{-n_1} L$. We will show that (5) holds provided $\mathrm{val}(\lambda) \geqslant 2n_1 + n_0$. Assume this is so and set $n = \mathrm{val}(\lambda) - n_1$. Then we have

$$\mathrm{Supp}(\varphi_\lambda) = \lambda\,\mathrm{Supp}(\varphi) \subseteq \lambda \varpi^{-n_1} L = \varpi^n L. \tag{7}$$

Hence, it follows that

$$\mathrm{Supp}(f_\lambda) \subseteq K_n. \tag{8}$$

Let $P = MN$ be a proper parabolic subgroup of G and let $x \in G(F)$. Consider the integral

$$\int_{N(F)} f_\lambda(xn)\, dn. \tag{9}$$

If $xN(F) \cap K_n = \varnothing$, then by (8) the term inside the integral above vanishes identically and it follows that the integral is equal to zero. Assume from now on that $xK_n \cap N(F) \neq \varnothing$. Up to translating x by an element of $N(F)$, we may assume that $x \in K_n$. Then we may write $x = e^X$ for some $X \in \varpi^n L$. Using (8) again, and since K_n is a subgroup of $G(F)$, we see that the integral (9) is supported on $K_n \cap N(F)$. Thus, we have equalities

$$\int_{N(F)} f_\lambda(xn)\,dn = \int_{K_n\cap N(F)} f_\lambda(e^X n)\,dn. \tag{10}$$

Set $L_N = L \cap \mathfrak{n}(F)$. Then if we normalize measures correctly, the exponential map induces a measure preserving isomorphism $\varpi^n L_N \simeq K_n \cap N(F)$, so that we have

$$\int_{K_n\cap N(F)} f_\lambda(e^X n)\,dn = \int_{\varpi^n L_N} f_\lambda(e^X e^N)\,dn = \int_{\varpi^n L} \varphi_\lambda(\log(e^X e^N))\,dn. \tag{11}$$

By (6), for all $N \in \varpi^n L_N$ we have

$$\log(e^X e^N) \in X + N + \varpi^{2n} L. \tag{12}$$

Moreover, as φ is invariant by translation by $\varpi^{n_0} L$, the function φ_λ is invariant by translation by $\lambda\varpi^{n_0} L = \varpi^{n+n_1+n_0} L$ (recalling that $n = \mathrm{val}(\lambda) - n_1$). Since $\mathrm{val}(\lambda) \geqslant 2n_1 + n_0$, we also have $n \geqslant n_1 + n_0$. Hence, the function φ_λ is invariant by translation by $\varpi^{2n} L$ and so using (12), we deduce

$$\varphi_\lambda(\log(e^X e^N)) = \varphi_\lambda(X+N)$$

for all $N \in \varpi^n L_N$. From (10) and (11), it follows that

$$\int_{N(F)} f_\lambda(xn)\,dn = \int_{\varpi^n L} \varphi_\lambda(X+N)\,dn. \tag{13}$$

By (7) and since $X \in \varpi^n L$, the function $N \mapsto \varphi_\lambda(X+N)$ for $N \in \mathfrak{n}(F)$ is supported on $\varpi^n L_N$. Consequently, we have

$$\int_{\varpi^n L} \varphi_\lambda(X+N)\,dn = \int_{\mathfrak{n}(F)} \varphi_\lambda(X+N)\,dn.$$

As φ_λ is a cusp form, this last integral vanishes. Hence, using (13) we see that the integral (9) is also zero. Since it is true for all $x \in G(F)$ and all proper parabolic subgroups $P = MN$ of G, this shows that f_λ is a cusp form. Hence, (5) is indeed satisfied as soon as $\mathrm{val}(\lambda) \geqslant 2n_1 + n_0$, and this ends the proof that Lemma 5 implies Theorem 4. □

It only remains to establish Lemma 5. Recall that a maximal torus T in G is said to be elliptic if $A_T = A_G$, where A_T and A_G denote the maximal split subtori in T and the center of G, respectively. The proof of Lemma 5 will ultimately rely on the following existence result (see [Platonov and Rapinchuk 1994, Theorem 6.21]):

Theorem 6. *G admits an elliptic maximal torus.*

Proof of Lemma 5. Let us fix a symmetric nondegenerate bilinear form B on $\mathfrak{g}(F)$ which is $G(F)$-invariant. Such a bilinear form is easy to construct. On $\mathfrak{g}_{\mathrm{der}}(F)$, the derived subalgebra of $\mathfrak{g}(F)$, we have the Killing form B_{Kil} which is symmetric, $G(F)$-invariant and nondegenerate. Hence, we may take $B = B_{\mathfrak{z}} \oplus B_{\mathrm{Kil}}$ where $B_{\mathfrak{z}}$ is any nondegenerate symmetric bilinear form on $\mathfrak{z}_G(F)$, the center of $\mathfrak{g}(F)$. Let us also fix a nontrivial continuous additive character $\psi : F \to \mathbb{C}^\times$. Using those, we can define a Fourier transform on $C_c^\infty(\mathfrak{g}(F))$ by

$$\hat{\varphi}(Y) = \int_{\mathfrak{g}(F)} \varphi(X)\psi(B(X, Y))\, dX, \quad \varphi \in C_c^\infty(\mathfrak{g}(F)), Y \in \mathfrak{g}(F).$$

Of course, this Fourier transform also depends on the choice of a Haar measure on $\mathfrak{g}(F)$.

More generally, if V is a subspace of $\mathfrak{g}(F)$ and $V^\perp$ denotes the orthogonal of V with respect to B, we can also define a Fourier transform $C_c^\infty(V) \to C_c^\infty(\mathfrak{g}(F)/V^\perp)$, $\varphi \mapsto \hat{\varphi}$, by setting

$$\hat{\varphi}(Y) = \int_V \varphi(Y)\psi(B(X, Y))\, dY, \quad X \in \mathfrak{g}(F)/V^\perp,$$

where again we need to choose a Haar measure on V. It is easy to check that for compatible choices of Haar measures, the following diagram commutes:

$$\begin{array}{ccc} C_c^\infty(\mathfrak{g}(F)) & \xrightarrow{FT} & C_c^\infty(\mathfrak{g}(F)) \\ \big\downarrow \mathrm{res}_V & & \big\downarrow \int_{V^\perp} \\ C_c^\infty(V) & \xrightarrow{FT} & C_c^\infty(\mathfrak{g}(F)/V^\perp) \end{array}$$

whereby the horizontal arrows are Fourier transforms, the left vertical arrow is given by restriction to V and the right vertical arrow is given by integration over the cosets of $V^\perp$. For $P = MN$ a parabolic subgroup of G, we have $\mathfrak{p}(F)^\perp = \mathfrak{n}(F)$ (where $\mathfrak{p}$ stands for the Lie algebra of P). The commutativity of the above diagram in this particular case gives us (for some compatible choices of Haar measures) the formula

$$\int_{\mathfrak{n}(F)} \hat{\varphi}(X + N)\, dn = \int_{\mathfrak{p}(F)} \varphi(Y)\psi(B(X, Y))\, dY \tag{14}$$

for all $\varphi \in C_c^\infty(\mathfrak{g}(F))$ and all $X \in \mathfrak{g}(F)$.

Let T_{ell} be an elliptic maximal torus of G, whose existence is guaranteed by Theorem 6. Let $\mathfrak{t}_{\mathrm{ell}}$ be its Lie algebra and set $\mathfrak{t}_{\mathrm{ell,reg}} = \mathfrak{t}_{\mathrm{ell}} \cap \mathfrak{g}_{\mathrm{reg}}$ for the subset of G-regular elements in $\mathfrak{t}_{\mathrm{ell}}$. Denote by $\mathfrak{t}_{\mathrm{ell,reg}}(F)^G$ the subset of elements in $\mathfrak{g}_{\mathrm{reg}}(F)$ that are $G(F)$-conjugated to an element of $\mathfrak{t}_{\mathrm{ell,reg}}(F)$. Then, $\mathfrak{t}_{\mathrm{ell,reg}}(F)^G$ is an open subset of $\mathfrak{g}(F)$ (since the map $T_{\mathrm{ell}}(F) \backslash G(F) \times \mathfrak{t}_{\mathrm{ell,reg}}(F) \to \mathfrak{g}(F)$,

$(g, X) \mapsto g^{-1}Xg$, is everywhere submersive). In particular, we can certainly find a nonzero function $\varphi \in C_c^\infty(\mathfrak{g}(F))$ whose support is contained in $\mathfrak{t}_{\text{ell,reg}}(F)^G$. Let us fix such a function φ. We claim the following:

(15) The function $\hat{\varphi}$ is a cusp form.

Indeed, let $P = MN$ be a proper parabolic subgroup of G and let $X \in \mathfrak{g}(F)$. We need to see that the integral

$$\int_{\mathfrak{n}(F)} \hat{\varphi}(X+N)\,dn$$

is zero. By (14), this integral is equal to

$$\int_{\mathfrak{p}(F)} \varphi(Y)\psi(B(X,Y))\,dY.$$

Hence, we only need to show that $\mathrm{Supp}(\varphi) \cap \mathfrak{p}(F) = \varnothing$. By definition of φ, it even suffices to see that $\mathfrak{t}_{\text{ell,reg}}(F)^G \cap \mathfrak{p}(F) = \varnothing$. But this follows immediately from the fact that, P being proper, it does not contain any elliptic maximal torus of G. □

References

[Harish-Chandra 1970] Harish-Chandra, *Harmonic analysis on reductive p-adic groups*, Lecture Notes in Mathematics **162**, Springer, Berlin, 1970. MR 54 #2889 Zbl 0202.41101

[Kret 2012] A. Kret, "Existence of cuspidal representations of p-adic reductive groups", preprint, 2012. arXiv 1205.2771

[Platonov and Rapinchuk 1994] V. Platonov and A. Rapinchuk, *Algebraic groups and number theory*, Pure and Applied Mathematics **139**, Academic Press, Boston, 1994. MR 95b:11039 Zbl 0841.20046

[Renard 2010] D. Renard, *Représentations des groupes réductifs p-adiques*, Cours Spécialisés **17**, Société Mathématique de France, Paris, 2010. MR 2011d:22019 Zbl 1186.22020

Received June 9, 2015. Revised August 17, 2015.

RAPHAËL BEUZART-PLESSIS
DEPARTMENT OF MATHEMATICS
NATIONAL UNIVERSITY OF SINGAPORE
BLOCK 17, 10 LOWER KENT RIDGE ROAD
SINGAPORE 119076
SINGAPORE

rbeuzart@gmail.com

PACIFIC JOURNAL OF MATHEMATICS
Vol. 282, No. 1, 2016
dx.doi.org/10.2140/pjm.2016.282.35

QUANTUM GROUPS AND GENERALIZED CIRCULAR ELEMENTS

MICHAEL BRANNAN AND KAY KIRKPATRICK

We show that with respect to the Haar state, the joint distributions of the generators of Van Daele and Wang's free orthogonal quantum groups are modeled by free families of generalized circular elements and semicircular elements in the large (quantum) dimension limit. We also show that this class of quantum groups acts naturally as distributional symmetries of almost-periodic free Araki–Woods factors.

1. Introduction

There are intriguing connections between the representation theory of certain classes of compact matrix groups and independent Gaussian structures in probability theory. For instance, if one considers the N^2 matrix elements $\{u_{ij}\}_{1\le i,j\le N}$ of the fundamental representation of the $N\times N$ orthogonal group $O_N = O_N(\mathbb{R})$ on the Hilbert space $\mathbb{C}^N$, then it is well known that the joint moments of these variables with respect to the Haar probability measure are approximated by an independent and identically distributed, mean zero, variance $1/N$ family of real Gaussian random variables in the large N limit; see, for example, [Diaconis and Freedman 1987]. Intimately related to this asymptotic Gaussianity result is the celebrated theorem of Freedman [1962; Diaconis and Freedman 1987], which says that an infinite sequence $\boldsymbol{x} = (x_n)_{n\in\mathbb{N}}$ of real-valued random variables is a conditionally independent centered Gaussian family with common variance if and only if the sequence is *rotatable*: i.e., for each $N\in\mathbb{N}$, the joint distribution of the N-dimensional truncation $\boldsymbol{x}_N = (x_n)_{1\le n\le N}$ of $\boldsymbol{x}$ is invariant under rotations by O_N.

When one replaces the family of orthogonal groups $\{O_N\}_{N\in\mathbb{N}}$ by the unitary groups $\{U_N\}_{N\in\mathbb{N}}$, analogous results are known to hold where one replaces real Gaussian random variables by their complex-valued counterparts. The key ingredient for the above results is a certain asymptotic orthonormality property for canonical generators (weighted Brauer diagrams, in fact) of the spaces of intertwiners between the tensor powers of the fundamental representations of these groups in the large rank limit. This asymptotic feature of the representation theory is most concisely

MSC2010: primary 20G42, 46L54; secondary 46L65.
Keywords: quantum groups, free probability, free Araki–Woods factor, free quasifree state.

expressed via the so-called Weingarten calculus developed in [Collins 2003; Collins and Śniady 2006], with origins in the pioneering work of Weingarten [1978] on the asymptotics of unitary matrix integrals. For a broad treatment of probabilistic symmetries, we refer to the text [Kallenberg 2005].

Within the framework of operator algebras and noncommutative geometry, *compact quantum groups* provide a vast and rich generalization of the theory of compact groups. The operator algebraic theory of compact quantum groups was pioneered by Woronowicz — see [1987; 1998], for instance — and has led recently to many interesting examples and developments in the theory of operator algebras.

One can ask noncommutative probabilistic questions about compact quantum groups, because every compact quantum group $\mathbb{G}$ admits a natural analogue of the Haar probability measure (the Haar state). The last decade or so has seen a flurry of activity in this direction, particularly for free quantum groups and Voiculescu's free probability theory. For instance, the free orthogonal quantum groups O_N^+ and free unitary quantum groups U_N^+ discovered by Wang [1993] turn out to have interesting noncommutative probabilistic structures that share deep parallels with the aforementioned classical results for O_N and U_N. Most notable for our purposes are the works of Banica–Collins [2007] and Curran [2010]. Banica and Collins show that the rescaled matrix elements $\{\sqrt{N}u_{ij}\}_{1\le i,j\le N}$ of the fundamental representation of O_N^+ (respectively U_N^+) converge in joint distribution to a freely independent family of standard — mean zero, variance one — semicircular (respectively circular) elements in a free group factor. Curran provides a free probability analogue of Freedman's rotatability theorem: An infinite sequence $\boldsymbol{x} = (x_n)_{n\in\mathbb{N}}$ of self-adjoint noncommutative random variables in a W*-probability space (M, φ) is *quantum rotatable* if and only if there exists a W*-subalgebra $B \subseteq M$ and a φ-preserving conditional expectation $E : M \to B$ such that $\boldsymbol{x}$ is an identically distributed family of mean zero semicircular elements that is free with amalgamation over B. A similar result for U_N^+ is obtained in [loc. cit.].

In this paper, we consider similar noncommutative probabilistic questions for a broad class of compact quantum groups introduced by Van Daele and Wang [1996] that generalizes the construction of U_N^+ and O_N^+ (which they called *universal quantum groups*). To define such a universal quantum group, one uses an invertible matrix $F \in \mathrm{GL}_N(\mathbb{C})$ to deform the defining algebraic relations for the quantum groups O_N^+ and U_N^+, yielding a pair of new compact matrix quantum groups called O_F^+ and U_F^+. When $F = 1$ (the $N \times N$ identity matrix), we recover O_N^+ and U_N^+ as special cases (see Section 3 for precise definitions). We follow recent literature conventions and refer to O_F^+ as *free orthogonal quantum groups* and U_F^+ as *free unitary quantum groups* (both with parameter matrix F).

The quantum groups $\mathbb{G} = O_F^+, U_F^+$ (at least when F is not a multiple of a unitary matrix) are especially interesting because their Haar states are nontracial and because the corresponding quantum group von Neumann algebras $L^\infty(\mathbb{G})$ are known to be type III factors in many cases; see [De Commer et al. 2014; Vaes and Vergnioux 2007].

The main result of this paper is that in this far more general (possibly nontracial) setting, asymptotic freeness still emerges in the large rank limit. In Theorems 5.1, 5.2, and 5.4, we show that the joint distribution of the (suitably rescaled) matrix elements of the fundamental representation of the quantum group O_F^+ can be approximated by a freely independent family of noncommutative random variables consisting of semicircular elements and Shlyakhtenko's *generalized circular elements* [1997], which is built in a natural way from the initial data $F \in \mathrm{GL}_N(\mathbb{C})$. Generalized circular elements are nontracial deformations of Voiculescu's circular elements, and they arise as canonical generators of free Araki–Woods factors [loc. cit.]. Since free Araki–Woods factors are the natural nontracial (or type III) deformations of the free group factors, our asymptotic freeness results for O_F^+ provide a satisfactory generalization of the tracial asymptotic freeness results of [Banica and Collins 2007]. We also remark in Section 5C how similar asymptotic freeness results can be obtained for the free unitary quantum groups U_F^+.

Our proofs of asymptotic freeness results in O_F^+ and U_F^+ follow the general outline of the earlier work [Banica and Collins 2007; Collins and Śniady 2006], and we use a modified version of the Weingarten calculus for our situation. In our case, the formulas become a bit more unwieldy, a consequence of the extra parameters that arise from the nontrivial matrix $F \in \mathrm{GL}_N(\mathbb{C})$. On the other hand, there is one significant and interesting difference between our (nontracial) setting and the earlier asymptotic (free) independence results on groups and quantum groups where the Haar state is tracial. In the case of O_F^+ and U_F^+, $F \in \mathrm{GL}_N(\mathbb{C})$, we find that the error in approximation of joint moments by free variables is of order

$$O((\mathrm{Tr}(F^*F))^{-1}),$$

a bound that is in many cases much smaller than the traditional bound given by $O(1/N)$ in the classical case. This fact allows us to observe, for example, asymptotic freeness results in a fixed dimension N, by considering families of quantum groups $\{O_F^+\}_{F\in\Lambda\subset \mathrm{GL}_N(\mathbb{C})}$ where the *quantum dimension* $\mathrm{Tr}(F^*F)$ tends to infinity; see Theorem 5.4.

Based on our nontracial asymptotic freeness results described above, together with Curran's work on quantum rotatability [2010], it now becomes natural to ask whether the free quantum groups O_F^+ act nontrivially on free Araki–Woods factors in a free-quasifree state-preserving way. In Section 6, we answer this question in the affirmative, and as a result we observe that almost-periodic free Araki–Woods factors admit a wealth of quantum symmetries. A future goal of the authors is to find a suitable "type III" version of Freedman's theorem adapted to O_F^+, U_F^+, and free Araki–Woods factors. After a first version of this paper appeared, it was pointed out to the authors that the main result of Section 6, namely Theorem 6.5, can also be obtained as a special case of a very general result of S. Vaes [2005, Proposition 3.1].

We finish this section with an outline of the paper's organization. Section 2 discusses some preliminaries on quantum groups and free probability that are required. Section 3 defines the quantum groups O_F^+ and U_F^+ and gives Weingarten-type formulas for joint moments of the generators of O_F^+ with respect to the Haar state. Section 4 considers the large quantum dimension asymptotics of these Weingarten formulas. Section 5 gives various asymptotic freeness results for O_F^+, and includes a remark on how to extend our results on O_F^+ to some of their unitary counterparts U_F^+. Finally Section 6 considers O_F^+ as quantum symmetries of almost-periodic free Araki–Woods factors. This is achieved by associating to each O_F^+ a canonical free family of generalized circular elements whose joint distribution is invariant under quantum rotations by O_F^+.

2. Preliminaries

In this section we briefly review some concepts from free probability theory and compact quantum group theory. For more details, we refer the reader to [Nica and Speicher 2006] for free probability and to [Timmermann 2008; Woronowicz 1998] for quantum groups.

2A. ***Noncommutative probability spaces and free independence.*** A *noncommutative probability space (NCPS)* is a pair (A, φ), where A is a unital C*-algebra, and $\varphi : A \to \mathbb{C}$ is a state (i.e., a linear functional such that $\varphi(1_A) = 1$ and $\varphi(a^*a) \geq 0$ for all $a \in A$). Elements $a \in A$ are called *random variables*. Given a family of random variables $X = \{x_r\}_{r\in\Lambda} \subset (A, \varphi)$, the *joint distribution* of X is the collection of all *joint $*$-moments*

$$\bigl\{\varphi\bigl(P((x_r)_{r\in\Lambda})\bigr) : P \in \mathbb{C}\langle t_r, t_r^* : r \in \Lambda\rangle\bigr\},$$

where $\mathbb{C}\langle t_r, t_r^* : r \in \Lambda\rangle$ is the unital $*$-algebra of noncommutative polynomials in the variables $\{t_r\}_{r\in\Lambda}$, equipped with antilinear involution $t_r \mapsto t_r^*$. Given another family of random variables $Y = \{y_r\}_{r\in\Lambda}$ in a NCPS (B, ψ), we say that X and Y are *identically distributed* if

$$\varphi\bigl(P((x_r)_{r\in\Lambda})\bigr) = \psi\bigl(P((y_r)_{r\in\Lambda})\bigr) \quad \text{for all } P \in \mathbb{C}\langle t_r, t_r^* : r \in \Lambda\rangle.$$

Let (A, φ) be a NCPS. A family of $*$-subalgebras $\{A_r\}_{r\in\Lambda}$ of A is said to be *freely independent* (or simply *free*) if the following condition holds: for any choice of indices $r(1) \neq r(2), r(2) \neq r(3), \ldots, r(k-1) \neq r(k) \in \Lambda$ and any choice of centered random variables $x_{r(j)} \in A_{r(j)}$ (i.e., $\varphi(x_{r(j)}) = 0$), we have the equality

$$\varphi(x_{r(1)}x_{r(2)} \dots x_{r(k)}) = 0.$$

A family of random variables $X = \{x_r\}_{r\in\Lambda} \subset (A, \varphi)$ is said to be *free* if the family of unital $*$-subalgebras

$$\{A_r\}_{r\in\Lambda}, \quad A_r := \mathrm{alg}\left(1, x_r, x_r^*\right),$$

is free in the above sense. Let $S_\alpha = \{x_r^{(\alpha)}\}_{r\in\Lambda} \subset (A_\alpha, \varphi_\alpha)$ be a net of families of random variables and $S = \{x_r\}_{r\in\Lambda} \in (A, \varphi)$ be another family of random variables. We say that S_α *converges to* S *in distribution* (and write $S_\alpha \to S$) if, for any noncommutative polynomial $P \in \mathbb{C}\langle X_r : r \in \Lambda\rangle$,

$$\lim_\alpha \varphi_\alpha(P(S_\alpha)) = \varphi(P(S)).$$

2B. ***Fock spaces, semicircular elements, and generalized circular elements.*** Let H be a complex Hilbert space. The *full Fock space* is the Hilbert space

$$\mathcal{F}(H) = \bigoplus_{n=0}^{\infty} H^{\otimes n},$$

where we put $H^{\otimes 0} := \mathbb{C}\Omega$, where Ω is a fixed unit vector, called the *vacuum vector*. The *vacuum expectation* is the state $\varphi_\Omega : \mathcal{B}(\mathcal{F}(H)) \to \mathbb{C}$ given by $\varphi_\Omega(x) = \langle \Omega \mid x\Omega\rangle$, $x \in \mathcal{B}(\mathcal{F}(H))$.

For each $\xi \in H$, we define the *left creation operator* $\ell(\xi) \in \mathcal{B}(\mathcal{F}(H))$ by

$$\begin{cases} \ell(\xi)\Omega = \xi, \\ \ell(\xi)\eta = \xi \otimes \eta, \quad \eta \in H^{\otimes n},\ n \geq 1. \end{cases}$$

Note that $\|\ell(\xi)\|_{\mathcal{B}(\mathcal{F}(H))} = \|\xi\|_H$. Given a NCPS (A, φ), a *(standard) semicircular element* is a self-adjoint random variable $x \in A$ with the same distribution as $s(\xi) := \ell(\xi) + \ell(\xi)^* \in (\mathcal{B}(\mathcal{F}(H)), \varphi_\Omega)$, where $\xi \in H$ is a unit vector. Given $\alpha, \beta \in \mathbb{R}^+$, a random variable $x \in (A, \varphi)$ is called an (α, β)*-generalized circular element* if it has the same distribution as the element $\alpha\ell(\xi) + \beta\ell(\eta)^* \in (\mathcal{B}(\mathcal{F}(H)), \varphi_\Omega)$, where ξ, η are orthonormal vectors in H. One can readily verify that for an (α, β)-generalized circular element x,

$$\varphi(x^*x) = \alpha^2 \quad \text{and} \quad \varphi(xx^*) = \beta^2,$$

and this information completely determines the $*$-moments of x with respect to φ. We will call the numbers α^2 and β^2 the left and right variances of x, respectively.

Next, we want to state a well known theorem which gives a combinatorial characterization of the joint distribution of a free semicircular or generalized circular family. To do this, we first need some notation concerning noncrossing partitions that will be used below and throughout the remainder of the paper.

Notation 2.1. Let $k \in \mathbb{N}$ and denote by $[k]$ the ordered set $\{1, \dots, k\}$.

(1) The lattice of partitions of $[k]$ will be denoted by $\mathcal{P}(k)$, and the lattice of noncrossing partitions will be denoted by $\mathcal{NC}(k)$. The standard partial order on both lattices will be denoted by $\leq$.

(2) If $\pi \in \mathcal{P}(k)$ partitions $[k]$ into r disjoint, nonempty subsets $\mathcal{V}_1, \dots, \mathcal{V}_r$ (called *blocks*), we write $|\pi| = r$ and say that π *has r blocks*.

(3) Given a function $i : [k] \to \Lambda$, we denote by $\ker i$ the element of $\mathcal{P}(k)$ whose blocks are the equivalence classes of the relation

$$s \sim_{\ker i} t \iff i(s) = i(t).$$

Note that if $\pi \in \mathcal{P}(k)$, then $\pi \leq \ker i$ is equivalent to the condition that whenever s and t are in the same block of π, $i(s)$ must equal $i(t)$ (i.e., the function $i : [k] \to \Lambda$ is constant on the blocks of π).

(4) Elements of $\mathcal{P}(k)$ which partition $[k]$ into subsets with exactly two elements are called *pairings* and the set of pairings of $[k]$ is denoted by $\mathcal{P}_2(k)$. We also write $\mathcal{NC}_2(k) = \mathcal{P}_2(k) \cap \mathcal{NC}(k)$. If k is odd, we of course have $\mathcal{P}_2(k) = \mathcal{NC}_2(k) = \varnothing$.

(5) Given $\pi \in \mathcal{P}_2(k)$ and $s, t \in [k]$, we will always write $(s, t) \in \pi$ if $\{s, t\}$ is a block of π and $s < t$.

(6) Let $\epsilon : [k] \to \{1, *\}$ be a function. We let $\mathcal{NC}_2^\epsilon(k) \subset \mathcal{NC}_2(k)$ be the subset of all noncrossing pairings such that for all $(s, t) \in \pi$,

$$\epsilon(s) \neq \epsilon(t).$$

Theorem 2.2 [Nica and Speicher 2006, Chapters 7 and 15]. *Let $X = (x_r)_{r\in\Lambda}$ be a family of random variables in an NCPS (A, φ).*

(1) *If $x_r = x_r^*$ for each $r \in \Lambda$, then X is a free family of standard semicircular variables if and only if for any $k \in \mathbb{N}$ and $r : [k] \to \Lambda$,*

$$\begin{aligned}\varphi(x_{r(1)} \cdots x_{r(k)}) &= \sum_{\substack{\pi \in \mathcal{NC}_2(k)\\ \ker r \geq \pi}} \prod_{(s,t)\in\pi} \varphi(x_{r(s)} x_{r(t)}) \\ &= |\{\pi \in \mathcal{NC}_2(k) : \ker r \geq \pi\}|.\end{aligned}$$

(2) *Let $(\alpha_r, \beta_r)_{r\in\Lambda} \subset \mathbb{R}^+ \times \mathbb{R}^+$. Then X is a free family of (α_r, β_r)-generalized circular elements if and only if for any $k \in \mathbb{N}$, $r : [k] \to \Lambda$, and $\epsilon : [k] \to \{1, *\}$,*

$$\varphi\big(x_{r(1)}^{\epsilon(1)} \cdots x_{r(k)}^{\epsilon(k)}\big) = \sum_{\substack{\pi \in \mathcal{NC}_2^\epsilon(k)\\ \ker r \geq \pi}} \prod_{(s,t)\in\pi} \varphi\big(x_{r(s)}^{\epsilon(s)} x_{r(t)}^{\epsilon(t)}\big)$$

where, in the above equation,

$$\varphi(x_{r(s)}^{\epsilon(s)} x_{r(t)}^{\epsilon(t)}) = \begin{cases} \alpha_{r(s)}^2, & (\epsilon(s), \epsilon(t)) = (*, 1) \\ \beta_{r(s)}^2, & (\epsilon(s), \epsilon(t)) = (1, *) \end{cases}$$

2C. *Free Araki–Woods factors and generalized circular elements.* Let $H_{\mathbb{R}}$ be a real separable Hilbert space and let (U_t) be an orthogonal representation of $\mathbb{R}$ on $H_{\mathbb{R}}$. Let $H = H_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}$ be the complexified Hilbert space. If A is the infinitesimal generator of (the extension of) U_t on H (i.e., $U_t = A^{it}$), then it follows that the map $j : H_{\mathbb{R}} \hookrightarrow H$ defined by $j(\xi) = (2/(A^{-1}+1))^{1/2}\xi$ is an isometric embedding of $H_{\mathbb{R}}$ into H [Shlyakhtenko 1997]. Let $K_{\mathbb{R}} = j(H_{\mathbb{R}})$; then $K_{\mathbb{R}} \cap i K_{\mathbb{R}} = \{0\}$ and $K_{\mathbb{R}} + i K_{\mathbb{R}}$ is dense in H. The *free Araki–Woods factor* is the von Neumann algebra

$$\Gamma(H_{\mathbb{R}}, U_t)'' = \mathrm{W}^*(\ell(\xi) + \ell(\xi)^* : \xi \in K_{\mathbb{R}}) \subseteq \mathcal{B}(\mathcal{F}(H)).$$

The restriction of the vacuum expectation $\varphi_\Omega = \langle \Omega \mid \cdot\, \Omega \rangle$ on $\mathcal{B}(\mathcal{F}(H))$ to $\Gamma(H_{\mathbb{R}}, U_t)''$ is always a faithful normal state, and turns $(\Gamma(H_{\mathbb{R}}, U_t)'', \varphi_\Omega)$ into a noncommutative probability space.

We recall from [loc. cit.] that U_t is the trivial representation if and only if $\Gamma(H_{\mathbb{R}}, U_t)'' \cong L(\mathbb{F}_{\dim H_{\mathbb{R}}})$, the von Neumann algebra generated by the left regular representation of the free group on $\dim H_{\mathbb{R}}$ generators. Otherwise, $\Gamma(H_{\mathbb{R}}, U_t)''$ is a type III factor.

Free Araki–Woods factors arise naturally when one considers free families of generalized circular elements that we introduced earlier. More precisely, we have the following theorem, which follows easily from the results in [op. cit., Section 6].

Theorem 2.3 [Shlyakhtenko 1997]. *Let $X = (x_r)_{r \in \Lambda}$ be a free family of (α_r, β_r)-generalized circular elements in a noncommutative probability space (A, φ) and let $0 < \lambda_r = \min\{\alpha\beta^{-1}, \beta\alpha^{-1}\} \le 1$. Then there is a state-preserving $*$-isomorphism $(\mathrm{W}^*(X), \varphi) \cong (\Gamma(H_{\mathbb{R}}, U_t)'', \varphi_\Omega)$, where U_t is the almost-periodic orthogonal representation acting on the Hilbert space $H_{\mathbb{R}} = \bigoplus_{r \in \Lambda} \mathbb{R}^2$ given by*

$$U_t = \bigoplus_{r \in \Lambda} R_{\lambda_r}(t), \quad \textit{where } R_{\lambda_r}(t) = \begin{pmatrix} \cos(t \log \lambda_r) & -\sin(t \log \lambda_r) \\ \sin(t \log \lambda_r) & \cos(t \log \lambda_r) \end{pmatrix}.$$

Moreover, every free Araki–Woods factor $\Gamma(H_{\mathbb{R}}, U_t)''$ arising from an almost-periodic representation U_t arises in this fashion.

2D. *Compact quantum groups.* A compact quantum group $\mathbb{G}$ is a pair $(C(\mathbb{G}), \Delta)$ where $C(\mathbb{G})$ is a unital C*-algebra and $\Delta : C(\mathbb{G}) \to C(\mathbb{G}) \otimes C(\mathbb{G})$ is a unital

$*$-homomorphism satisfying

$$\begin{cases} (\iota\otimes\Delta)\Delta=(\Delta\otimes\iota)\Delta & \text{(coassociativity)},\\ [\Delta(C(\mathbb{G}))(1\otimes C(\mathbb{G}))]=[\Delta(C(\mathbb{G}))(C(\mathbb{G})\otimes 1)] & \\ \qquad\qquad =C(\mathbb{G})\otimes C(\mathbb{G}) & \text{(nondegeneracy)},\end{cases}$$

where $[S]$ denotes the norm-closed linear span of a subset $S\subset C(\mathbb{G})\otimes C(\mathbb{G})$. Here and in the rest of the paper, the symbol $\otimes$ will denote the minimal tensor product of C*-algebras, $\overline{\otimes}$ will denote the spatial tensor product of von Neumann algebras, and $\odot$ will denote the algebraic tensor product of complex associative algebras. The homomorphism Δ is called a *coproduct*.

For any compact quantum group $\mathbb{G}=(C(\mathbb{G}),\Delta)$, there exists a unique *Haar state* $h_{\mathbb{G}}:C(\mathbb{G})\to\mathbb{C}$ which satisfies the following left and right Δ-invariance property, for all $a\in C(\mathbb{G})$:

$$(h_{\mathbb{G}}\otimes\iota)\Delta(a)=(\iota\otimes h_{\mathbb{G}})\Delta(a)=h_{\mathbb{G}}(a)1. \tag{2-1}$$

Note that in general $h=h_{\mathbb{G}}$ is not faithful on $C(\mathbb{G})$. In any case, we can construct a GNS representation $\pi_h:C(\mathbb{G})\to\mathcal{B}(L^2(\mathbb{G}))$, where $L^2(\mathbb{G})$ is the Hilbert space obtained by separation and completion of $C(\mathbb{G})$ with respect to the sesquilinear form $\langle a\mid b\rangle=h(a^*b)$, and π_h is the natural extension to $L^2(\mathbb{G})$ of the left multiplication action of $C(\mathbb{G})$ on itself. The *von Neumann algebra of* $\mathbb{G}$ is given by

$$L^\infty(\mathbb{G})=\pi_h(C(\mathbb{G}))''\subseteq\mathcal{B}(L^2(\mathbb{G})).$$

We note that Δ_r extends to an injective normal $*$-homomorphism $\Delta_r:L^\infty(\mathbb{G})\to L^\infty(\mathbb{G})\overline{\otimes}L^\infty(\mathbb{G})$, and the Haar state on $C(\mathbb{G})$ lifts to a faithful normal Δ_r-invariant state on $L^\infty(\mathbb{G})$.

Let H be a finite dimensional Hilbert space and $U\in\mathcal{B}(H)\otimes C(\mathbb{G})$ be invertible (unitary). Then U is called a *(unitary) representation* of $\mathbb{G}$ if, following the leg numbering convention,

$$(\iota\otimes\Delta)U=U_{12}U_{13}. \tag{2-2}$$

If we fix an orthonormal basis of H, we can identify U with an invertible matrix $U=[u_{ij}]\in M_N(C(\mathbb{G}))$ and (2-2) means exactly that

$$\Delta(u_{ij})=\sum_{k=1}^N u_{ik}\otimes u_{kj}\quad(1\le i,j\le N).$$

Of course the unit $1\in C(\mathbb{G})$ is always a representation of $\mathbb{G}$, called the *trivial representation*.

Let $U\in\mathcal{B}(H_1)\otimes C(\mathbb{G})$ and $V\in\mathcal{B}(H_2)\otimes C(\mathbb{G})$ be two representations of $\mathbb{G}$. An *intertwiner* between U and V is a bounded linear map $T:H_1\to H_2$ such that

$(T \otimes \iota)U = V(T \otimes \iota)$. The Banach space of all such intertwiners is denoted by $\mathrm{Hom}_{\mathbb{G}}(U, V)$. When $U = 1$ is the trivial representation, we write $\mathrm{Hom}_{\mathbb{G}}(U, V) = \mathrm{Fix}(V) \subset H_2$, and call $\mathrm{Fix}(V)$ the *space of fixed vectors for* V. If there exists an invertible (unitary) intertwiner between U and V, they are said to be *(unitarily) equivalent*. A representation is said to be irreducible if its only self-intertwiners are the scalar multiples of the identity map. It is known that each irreducible representation of $\mathbb{G}$ is finite dimensional and every finite dimensional representation is equivalent to a unitary representation. In addition, every unitary representation is unitarily equivalent to a direct sum of irreducible representations.

A compact quantum group $\mathbb{G}$ is called a *compact matrix quantum group* if there exists a finite dimensional unitary representation $U = [u_{ij}] \in M_N(C(\mathbb{G}))$ whose matrix elements generate $C(\mathbb{G})$ as a C*-algebra. Such a representation U is called a *fundamental representation* of $\mathbb{G}$. In this case, we note that the Haar state h is faithful when restricted to the dense unital $*$-subalgebra $\mathrm{Pol}(\mathbb{G}) \subseteq C(\mathbb{G})$ generated by $\{u_{ij}\}_{1\le i,j\le N}$.

3. The free quantum groups O_F^+ and U_F^+

In this section we recall the definition of the free orthogonal and unitary quantum groups O_F^+ and U_F^+, introduced by Van Daele and Wang in [1996].

Notation 3.1. Given a complex $*$-algebra A and a matrix $X = [x_{ij}] \in M_N(A)$, we denote by $\overline{X}$ the matrix $[x_{ij}^*] \in M_N(A)$.

Definition 3.2 [Van Daele and Wang 1996]. Let $N \ge 2$ be an integer and let $F \in \mathrm{GL}_N(\mathbb{C})$.

(1) The *free unitary quantum group* U_F^+ (with parameter matrix F) is the compact quantum group given by the universal C*-algebra

$$C(U_F^+) = \mathrm{C}^*(\{v_{ij}\}_{1\le i,j\le N} : V = [v_{ij}] \text{ is unitary and } F\overline{V}F^{-1} \text{ is unitary}), \tag{3-1}$$

together with coproduct $\Delta : C(U_F^+) \to C(U_F^+) \otimes C(U_F^+)$ given by

$$\Delta(v_{ij}) = \sum_{k=1}^{N} v_{ik} \otimes v_{kj} \quad (1 \le i, j \le N).$$

(2) Let $c = \pm 1$ and assume that $F\bar{F} = c1$. The *free orthogonal quantum group* O_F^+ (with parameter matrix F) is the compact quantum group given by the universal C*-algebra

$$C(O_F^+) = \mathrm{C}^*(\{u_{ij}\}_{1\le i,j\le N} : U = [u_{ij}] \text{ is unitary and } U = F\overline{U}F^{-1}), \tag{3-2}$$

together with coproduct $\Delta : C(O_F^+) \to C(O_F^+) \otimes C(O_F^+)$ given by

$$\Delta(u_{ij}) = \sum_{k=1}^{N} u_{ik} \otimes u_{kj} \quad (1 \le i, j \le N).$$

Remark 3.3. The coproduct Δ is defined so that the matrices of generators $V = [v_{ij}]$ and $U = [u_{ij}]$ are always fundamental representations of the compact matrix quantum group U_F^+ and O_F^+.

Remark 3.4. Note that the above definition for O_F^+ makes sense for any $F \in \mathrm{GL}_N(\mathbb{C})$. The additional condition $F\bar{F} = \pm 1$ is equivalent to the requirement that U is always an irreducible representation of O_F^+. Indeed, Banica [1996] showed that U is irreducible if and only if $F\bar{F} = \pm\lambda 1$ ($\lambda > 0$); moreover we clearly have $O_F^+ = O_{\lambda^{-1/2}F}^+$.

We remark that for our asymptotic freeness results, our assumption that $F\bar{F} = \pm I$ is not a major restriction. Indeed, by a result of Wang [2002, Section 6], O_F^+ for generic $F \in \mathrm{GL}_N(\mathbb{C})$ can be decomposed into a free product of finitely many quantum groups $O_{F_i}^+$ and $U_{P_k}^+$ with F_i, P_k invertible matrices and $F_i\bar{F}_i = \pm 1$.

For the remainder of the paper, we will deal mostly with the free orthogonal quantum groups O_F^+. Later on in Section 5C we indicate how to extend some of our orthogonal results to the unitary case.

3A. ***Canonical F-matrices for O_F^+.*** Let $c \in \{\pm 1\}$ and let $F \in \mathrm{GL}_N(\mathbb{C})$ be such that $F\bar{F} = c1$. In [Bichon et al. 2006], it is shown that if $c = 1$, then there is an integer $0 \le k \le N/2$, a nondecreasing sequence $\rho = (\rho_i)_{i=1}^k \in (0, 1)^k$, and a unitary $w \in U_N$ such that

$$F_\rho^{(+1)} := w^t F w = \begin{pmatrix} 0 & D_k(\rho) & 0 \\ D_k(\rho)^{-1} & 0 & 0 \\ 0 & 0 & 1_{N-2k} \end{pmatrix}, \tag{3-3}$$

where $D_k(\rho)$ denotes the $k \times k$ diagonal matrix with diagonal entries given by the k-tuple ρ.

On the other hand if $c = -1$, then by [loc. cit.] $N = 2k$ must be even and there must exist a nondecreasing sequence $\rho = (\rho_i)_{i=1}^k \in (0, 1]^k$ and a unitary $w \in U_N$ such that

$$F_\rho^{(-1)} := w^t F w = \begin{pmatrix} 0 & D_k(\rho) \\ -D_k(\rho)^{-1} & 0 \end{pmatrix}. \tag{3-4}$$

Remark 3.5. Note that the Kac type quantum groups O_N^+ correspond to the case $F = 1_N$, which is exactly the canonical deformation matrix $F_\rho^{(+1)}$ with $k = 0$.

According to [loc. cit.], given two matrices $F_i \in \mathrm{GL}_{N_i}(\mathbb{C})$ such that $F_i\bar{F}_i = c_i 1$, the two free orthogonal quantum groups $O_{F_1}^+$ and $O_{F_2}^+$ are isomorphic if and only if $N_1 = N_2$, $c_1 = c_2$, and $F_2 = vF_1v^t$ for some unitary matrix $v \in U_{N_1}$. The corresponding equivalence relation on such matrices has fundamental domain given by all matrices of the form $F_\rho^{(\pm 1)}$. As a consequence, we call such matrices $F_\rho^{(\pm 1)}$ *canonical* F-matrices. The canonical F-matrices yield the most natural coordinate system in which to represent the isomorphism equivalence class of any given O_F^+.

3B. *Integration over O_F^+.* In this section, we consider the problem of evaluating arbitrary monomials in the generators $\{u_{ij}\}_{1\le i,j\le N}$ of $C(O_F^+)$ with respect to the Haar state $h_{O_F^+}$.

Notation 3.6. Fix an orthonormal basis $\{e_i\}_{i=1}^N$ for $\mathbb{C}^N$ and $F \in \mathrm{GL}_N(\mathbb{C})$. Define

$$\xi = \sum_{i=1}^N e_i \otimes e_i \quad \text{and} \quad \xi^F = (\mathrm{id} \otimes F)\xi = \sum_{i=1}^N e_i \otimes Fe_i.$$

For each $l \in \mathbb{N}$, $\pi \in \mathcal{NC}_2(2l)$, and $i : [2l] \to [N]$ define,

$$\delta_\pi^F(i) = \prod_{(s,t)\in\pi} F_{i(t)i(s)},$$

and put

$$\xi_\pi^F = \sum_{i:[2l]\to[N]} \delta_\pi^F(i)\, e_{i(1)} \otimes e_{i(2)} \otimes \cdots \otimes e_{i(2l)} \in (\mathbb{C}^N)^{\otimes 2l}.$$

For the purposes of integrating monomials over O_F^+ with respect to the Haar state, we are interested in the *l-th tensor power* of the fundamental representation $U = [u_{ij}]$ of O_F^+,

$$U^{\oplus l} := [u_{i(1)j(1)} \cdots u_{i(l)j(l)}] \in \mathcal{B}((\mathbb{C}^N)^{\otimes l}) \otimes C(O_F^+).$$

$U^{\oplus l}$ is evidently a representation of the quantum group O_F^+, and the following theorem of Banica describes the space of fixed vectors of these higher tensor powers of U.

Theorem 3.7 [Banica 1996]. *Let $N \ge 2$, $c \in \{\pm 1\}$, and $F \in \mathrm{GL}_N(\mathbb{C})$ be such that $F\bar{F} = c1$. Then for each $l \in \mathbb{N}$,*

$$\mathrm{Fix}(U^{\oplus 2l+1}) = \{0\},$$

and

$$\mathrm{Fix}(U^{\oplus 2l}) = \mathrm{span}\{\xi_\pi^F : \pi \in \mathcal{NC}_2(2l)\}.$$

Moreover, $\{\xi_\pi^F\}_\pi$ is a linear basis for $\mathrm{Fix}(U^{\oplus 2l})$.

With the preceding theorem in hand, we now use the Weingarten calculus to describe the Haar state on O_F^+ in terms of the Gram matrices associated to the bases $\{\xi_\pi^F\}_{\pi \in \mathcal{NC}_2(2l)}$ of $\mathrm{Fix}(U^{\oplus 2l})$. For each $l \in \mathbb{N}$, define an $|\mathcal{NC}_2(2l)| \times |\mathcal{NC}_2(2l)|$ matrix $G_{2l,F} = [G_{2l,F}(\pi,\sigma)]_{\pi,\sigma \in \mathcal{NC}_2(2l)}$ by

$$G_{2l,F}(\pi,\sigma) = \langle \xi_\pi^F \mid \xi_\sigma^F \rangle \quad (\pi,\sigma \in \mathcal{NC}_2(2l)).$$

Theorem 3.8. *Let $N \geq 2$, $c \in \{\pm 1\}$, and $F \in \mathrm{GL}_N(\mathbb{C})$ be such that $F\bar{F} = c1$. Set $N_F := \mathrm{Tr}(F^*F)$. Then for any $l \geq 1$, $G_{2l,F}$ is an invertible matrix and*

$$G_{2l,F}(\pi,\sigma) = c^{l+|\pi\vee\sigma|} N_F^{|\pi\vee\sigma|} \quad (\pi,\sigma \in \mathcal{NC}_2(2l)),$$

where $\pi \vee \sigma$ denotes the join of π and σ in the lattice $\mathcal{P}(2l)$.

Proof. The first assertion follows from Theorem 3.7. For the second assertion, fix $\pi, \sigma \in \mathcal{NC}_2(2l)$ and let $\pi \vee \sigma = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m\}$, where each $\mathcal{V}_a$ is a block of $\pi \vee \sigma$. Then we have

$$G_{2l,F}(\pi,\sigma) = \sum_{i:[2l]\to[N]} \overline{\delta_\pi^F(i)}\,\delta_\sigma^F(i) = \prod_{a=1}^m \Bigg(\sum_{i:\mathcal{V}_a\to[N]} \overline{\delta_{\pi|_{\mathcal{V}_a}}^F(i)}\,\delta_{\sigma|_{\mathcal{V}_a}}^F(i) \Bigg).$$

From the above equation we see that $G_{2l,F}(\pi,\sigma)$ is a multiplicative function of the blocks of $\pi \vee \sigma$, and therefore it suffices to prove the theorem when $\pi \vee \sigma = 1_{2l}$.

To prove this special case of the theorem, we proceed by induction on l: If $2l = 2$, then $G_{2l,F} = \langle \xi^F \mid \xi^F \rangle = \mathrm{Tr}(F^*F) = N_F = c^2 N_F$. Now assume $l \geq 2$ and that the desired result is true for all $\pi', \sigma' \in \mathcal{NC}_2(2l-2)$ with $\pi' \vee \sigma' = 1_{2l-2}$. Fix $\pi, \sigma \in \mathcal{NC}_2(2l)$ such that $\pi \vee \sigma = 1_{2l}$. Since π is noncrossing, we can fix an interval $\{r, r+1\}$ in π and let $\{a, r\}$, $\{b, r+1\}$ be the corresponding (unordered) pairs of σ that connect to r and $r+1$. (Note that σ does not pair r and $r+1$ because $|\pi \vee \sigma| = 1$ and $l \geq 2$.) Now let $\pi' \in \mathcal{NC}_2(2l-2)$ be the pairing obtained by deleting the block $\{r, r+1\}$ from π and let $\sigma' \in \mathcal{NC}_2(2l-2)$ be the pairing obtained by deleting the points $r, r+1$ from σ and pairing a and b. Note that by construction, $\pi' \vee \sigma' = 1_{2l-2}$.

Using the readily verified identities

$$c\xi^F = (\iota\otimes(\xi^F)^*\otimes\iota)(\xi^F\otimes\xi^F) = ((\xi^F)^*\otimes\iota)(\iota\otimes\xi^F\otimes\iota)\xi^F = (\iota\otimes(\xi^F)^*)(\iota\otimes\xi^F\otimes\iota)\xi^F,$$

it easily follows that $G_{2l,F}(\pi,\sigma) = \langle \xi_\pi^F | \xi_\sigma^F \rangle = c\langle \xi_{\pi'}^F \mid \xi_{\sigma'}^F \rangle$. We then have from our induction assumption that

$$G_{2l,F}(\pi,\sigma) = c(c^{l-1+|\pi'\vee\sigma'|} N_F) = c^{l+1} N_F. \qquad \square$$

For each $l \in \mathbb{N}$, denote by $W_{2l,F}$ the matrix inverse of $G_{2l,F}$. In the following theorem we give a Weingarten-type moment formula for the Haar state on O_F^+. Compare with [Banica and Collins 2007; Banica et al. 2009], where a version of the following result is obtained in the case where $F \in \mathrm{GL}_N(\mathbb{R})$ and $c = 1$.

Theorem 3.9. *For each pair of multi-indices $i, j : [l] \to [N]$,*

$$h_{O_F^+}(u_{i(1)j(1)}u_{i(2)j(2)}\cdots u_{i(l)j(l)}) = \begin{cases} 0 & \text{if } l \text{ is odd,} \\ \displaystyle\sum_{\pi,\sigma\in\mathcal{NC}_2(l)} W_{l,F}(\pi,\sigma)\overline{\delta_\pi^F(j)}\delta_\sigma^F(i) & \text{otherwise.} \end{cases}$$

Proof. We use the fact that if $V \in \mathcal{B}(H)\otimes C(\mathbb{G})$ is a unitary representation of a compact quantum group $\mathbb{G}$ with Haar state h, then $P_V = (\mathrm{id}\otimes h)(V)$ is the orthogonal projection onto $\mathrm{Fix}(V)$. Using this fact, the quantity we are interested in is the (i, j)-th matrix element of the projection $P_{U^{\oplus l}}$. Since $P_{U^{\oplus l}} = 0$ when l is odd (by Theorem 3.7), the first equality is immediate.

For the second equality, assume $l \geq 2$ is even. Let $\{\xi_\pi^F\}_{\pi\in\mathcal{NC}_2(l)}$ be the basis for $\mathrm{Fix}(U^{\oplus l})$ from Theorem 3.7 and define a new set $\{\tilde{\xi}_\pi^F\}_{\pi\in\mathcal{NC}_2(l)} \subset \mathrm{Fix}(U^{\oplus l})$ by

$$\tilde{\xi}_\pi^F = \sum_{\sigma\in\mathcal{NC}_2(l)} \mathrm{W}_{l,F}^{1/2}(\pi,\sigma)\xi_\sigma^F,$$

where $\mathrm{W}_{l,F}^{1/2}$ is the matrix square root of $W_{l,F}$. Then $\{\tilde{\xi}_\pi^F\}_{\pi\in\mathcal{NC}_2(l)}$ is an orthonormal basis for $\mathrm{Fix}(U^{\oplus l})$ by Theorem 3.8 and

$$P_{U^{\oplus l}} = \sum_{\pi\in\mathcal{NC}_2(l)} |\tilde{\xi}_\pi^F\rangle\langle\tilde{\xi}_\pi^F|.$$

Therefore,

$$\begin{aligned} h_{O_F^+}(u_{i(1)j(1)}u_{i(2)j(2)}\cdots u_{i(l)j(l)}) \\ &= \langle e_i \mid P_{U^{\oplus l}} e_j\rangle \\ &= \sum_{\rho\in\mathcal{NC}_2(l)} \langle\tilde{\xi}_\rho^F \mid e_j\rangle\langle e_i \mid \tilde{\xi}_\rho^F\rangle \\ &= \sum_{\pi,\sigma,\rho\in\mathcal{NC}_2(l)} \mathrm{W}_{l,F}^{1/2}(\rho,\pi)\langle\xi_\pi^F \mid e_j\rangle \mathrm{W}_{l,F}^{1/2}(\rho,\sigma)\langle e_i \mid \xi_\sigma^F\rangle \\ &= \sum_{\pi,\sigma\in\mathcal{NC}_2(l)} W_{l,F}(\pi,\sigma)\langle\xi_\pi^F \mid e_j\rangle\langle e_i \mid \xi_\sigma^F\rangle \\ &= \sum_{\pi,\sigma\in\mathcal{NC}_2(l)} W_{l,F}(\pi,\sigma)\overline{\delta_\pi^F(j)}\delta_\sigma^F(i). \end{aligned}$$

□

3C. ***Integrating $*$-monomials over O_F^+.*** Note that the defining relations for the generators $\{u_{ij}\}_{1\le i,j\le N}$ of $(C(O_F^+), h_{O_F^+})$ imply that the family $\{u_{ij}\}_{1\le i,j\le N}$ is self-adjoint. Moreover, taking the deformation matrix F to be in canonical form, as defined in Section 3A, we can write

$$F = F_\rho^{(c)} = \begin{pmatrix} 0 & D_k(\rho) & 0 \\ cD_k(\rho)^{-1} & 0 & 0 \\ 0 & 0 & 1_{N-2k} \end{pmatrix}.$$

where $D_k(\rho)$ denotes the $k \times k$ diagonal matrix with diagonal entries given by a k-tuple $\rho = (\rho_i)_{i=1}^k \subset (0, 1]^k$, and $N = 2k$ if $c = -1$. Then $F^{-1} = cF$, and we easily compute from (3-2) that

$$u_{ij}^* = (cFUF)_{ij} = c \sum_{1 \le r,s \le N} F_{ir} F_{sj} u_{rs} = \begin{cases} cF_{i,i+k} F_{j+k,j} u_{i+k,j+k} & 1 \le i, j \le 2k, \\ F_{i,i+k} u_{i+k,j} & 1 \le i \le 2k,\ j > 2k, \\ F_{j+k,j} u_{i,j+k} & 1 \le j \le 2k,\ i > 2k, \\ u_{ij} & i, j > 2k \end{cases} \tag{3-5}$$

where in the above equations we perform the additions $i \mapsto i+k,\ j \mapsto j+k \mod 2k$. Using these equations, it is easy to see that the fundamental representation $U = [u_{ij}]_{1 \le i,j \le N}$ of O_F^+ admits the following canonical block-matrix decomposition.

$$U = \begin{pmatrix} [u_{a,b}]_{1 \le a,b \le k} & [u_{a,b+k}]_{1 \le a,b \le k} & [u_{a,t}]_{\substack{1 \le a \le k \\ t \ge 2k+1}} \\ [c\rho_a^{-1}\rho_b^{-1} u_{a,b+k}^*]_{1 \le a,b \le k} & [\rho_a^{-1}\rho_b u_{a,b}^*]_{1 \le a,b \le k} & [\rho_a^{-1} u_{a,t}^*]_{\substack{1 \le a \le k \\ t \ge 2k+1}} \\ [u_{s,b}]_{\substack{1 \le b \le k \\ s \ge 2k+1}} & [\rho_b u_{s,b}^*]_{\substack{1 \le b \le k \\ s \ge 2k+1}} & [u_{s,t}]_{s,t \ge 2k+1} \end{pmatrix}. \tag{3-6}$$

Remark 3.10. From Equation (3-6), we see that the C*-algebra $C(O_F^+)$ is generated by the subset

$$\Big(\{u_{ij}\}_{\substack{1 \le i \le k \\ 1 \le j \le N}} \cup \{u_{ij}\}_{\substack{1 \le j \le k \\ 2k+1 \le i \le N}} \cup \{u_{ij}\}_{2k+1 \le i,j \le N}\Big) \subset \{u_{ij}\}_{1 \le i,j \le N}.$$

3C1. *General $*$-moments over O_F^+.* Let $\epsilon \in \{1, *\}$ and $i, j \in [N]$. Using (3-6) (or equivalently (3-5)), we can find unique numbers $i_\epsilon, j_\epsilon \in [N]$ and $t_F(i, j, \epsilon) \in \mathbb{R}$ such that

$$u_{ij}^\epsilon = t_F(i, j, \epsilon) u_{i_\epsilon j_\epsilon}. \tag{3-7}$$

In particular, arbitrary $*$-moments in the generators $\{u_{ij}\}_{1 \le i,j \le N}$ can be computed using the formula of Theorem 3.9.

Proposition 3.11. *Let $l \in \mathbb{N}$, $i, j : [l] \to [N]$, and $\epsilon : [l] \to \{1, *\}$ be given. If l is odd, then*

$$h_{O_F^+}\big(u_{i(1)j(1)}^{\epsilon(1)} u_{i(2)j(2)}^{\epsilon(2)} \cdots u_{i(l)j(l)}^{\epsilon(l)}\big) = 0.$$

If l is even, then

$$h_{O_F^+}\big(u^{\epsilon(1)}_{i(1)j(1)}u^{\epsilon(2)}_{i(2)j(2)}\cdots u^{\epsilon(l)}_{i(l)j(l)}\big)$$
$$= \prod_{r=1}^{l} t_F(i(r), j(r), \epsilon(r))\, h_{O_F^+}\Bigg(\prod_{r=1}^{l} u_{i(r)_{\epsilon(r)}, j(r)_{\epsilon(r)}}\Bigg)$$
$$= \prod_{r=1}^{l} t_F(i(r), j(r), \epsilon(r)) \sum_{\pi,\sigma\in\mathcal{NC}_2(l)} W_{l,F}(\pi,\sigma)\,\overline{\delta^F_\pi(j_\epsilon)}\,\delta^F_\sigma(i_\epsilon),$$

where $i_\epsilon = (i_{\epsilon(1)}(1), i_{\epsilon(2)}(2), \dots, i_{\epsilon(l)}(l))$ *and* $j_\epsilon = (j_{\epsilon(1)}(1), j_{\epsilon(2)}(2), \dots, j_{\epsilon(l)}(l))$.

The proof of this result is immediate.

3C2. *Variances of the generators of $C(O_F^+)$.* The simplest (nonzero) $*$-moments are the left and right covariances of the generators $\{u_{ij}\}_{1\le i,j\le N}\subset C(O_F^+)$, i.e., the quantities

$$\langle u_{ij} \mid u_{kl}\rangle_L = h_{O_F^+}(u_{ij}^* u_{kl}) \quad \text{and} \quad \langle u_{kl} \mid u_{ij}\rangle_R = h_{O_F^+}(u_{ij} u_{kl}^*).$$

The left and right covariances can be computed using Proposition 3.11. Alternatively, we can compute these quantities using the Schur orthogonality relations

$$h_{O_F^+}(u_{ij}^* u_{kl}) = \frac{\delta_{jl}(Q^{-1})_{ki}}{N_F}, \quad h_{O_F^+}(u_{ij}u_{kl}^*) = \frac{\delta_{ik}Q_{lj}}{N_F} \quad (1\le i,j\le N), \tag{3-8}$$

where

$$Q = F^t\bar{F}, \quad Q^{-1} = FF^*, \quad \text{and } \operatorname{Tr}(Q) = \operatorname{Tr}(Q^{-1}) = N_F.$$

See, for example, [Woronowicz 1998] and the paragraphs following Theorem 7.2 of [Vaes and Vergnioux 2007]. In particular, when $F = F^{(c)}_\rho$ is a canonical F-matrix as above, then the structure of Q is simple:

$$Q = \begin{pmatrix} D_k(\rho)^{-2} & 0 & 0 \\ 0 & D_k(\rho)^2 & 0 \\ 0 & 0 & 1_{N-2k} \end{pmatrix}.$$

In particular, it follows from (3-8) that $\{u_{ij}\}_{1\le i,j\le N}$ is an orthogonal system with respect to the inner products $\langle\cdot\mid\cdot\rangle_L$ and $\langle\cdot\mid\cdot\rangle_R$ induced by $h_{O_F^+}$, and a simple calculation shows that the $N\times N$ matrix of left and right variances

$$\Phi = \big[(\langle u_{ij}\mid u_{ij}\rangle_L, \langle u_{ij}\mid u_{ij}\rangle_R)\big]_{1\le i,j\le N}$$

has the following block-matrix decomposition (compare with the decomposition of the fundamental representation of O_F^+ given by (3-6)).

$$\Phi = \frac{1}{N_F}\big[(Q_{ii}^{-1}, Q_{jj})\big]_{1\le i,j\le N}$$
$$= \frac{1}{N_F}\begin{pmatrix} [(\rho_a^2, \rho_b^{-2})]_{1\le a,b\le k} & [(\rho_a^2, \rho_b^2)]_{1\le a,b\le k} & [(\rho_a^2, 1)]_{k\times(N-2k)} \\ [(\rho_a^{-2}, \rho_b^{-2})]_{1\le a,b\le k} & [(\rho_a^{-2}, \rho_b^2)]_{1\le a,b\le k} & [(\rho_a^{-2}, 1)]_{k\times(N-2k)} \\ [(1, \rho_b^{-2})]_{(N-2k)\times k} & [(1, \rho_b^2)]_{(N-2k)\times k} & [(1,1)]_{k\times k} \end{pmatrix} \tag{3-9}$$

4. Large (quantum) dimension asymptotics

Using our Weingarten formulas (Theorem 3.9 and Proposition 3.11), we can study the large (quantum) dimension asymptotics of $*$-moments over O_F^+. Let $F \in \mathrm{GL}_N(\mathbb{C})$ be such that $F\bar{F} = c1$, and let $N_F = \mathrm{Tr}(F^*F)$. We will call the number N_F the *quantum dimension* of the fundamental representation $U = [u_{ij}]$ of O_F^+. Note that we always have $N_F \ge N$. The following proposition shows that the Weingarten matrices $W_{l,F}$ associated to O_F^+ are asymptotically diagonal as the quantum dimension N_F tends to infinity. This result should be compared with [Banica and Collins 2007, Theorem 6.1].

Theorem 4.1. *For each $l \in 2\mathbb{N}$, as $N_F \to \infty$,*

$$N_F^{l/2} W_{l,F}(\pi,\sigma) = \delta_{\pi,\sigma} + O(N_F^{-1}) \quad (\pi,\sigma \in \mathcal{NC}_2(l)).$$

Proof. According to Theorem 3.8, $W_{l,F} = G_{l,F}^{-1}$ and $G_{l,F}(\pi,\sigma) = c^{l/2+|\pi\vee\sigma|} N_F^{|\pi\vee\sigma|}$. Observe that $|\pi\vee\sigma| = l/2$ if and only if $\pi = \sigma$, and $|\pi\vee\sigma| \le l/2 - 1$ otherwise. Therefore, with respect to the operator norm, we have the asymptotic formula

$$G_{l,F} = N_F^{l/2} I + O(N_F^{l/2-1}) = N_F^{l/2}(I + O(N_F^{-1})).$$

Write $N_F^{-l/2} G_{l,F} = I + A_F$, where $\|A_F\| \le C_l N_F^{-1}$ for some $C_l \ge 0$. Then for sufficiently large N_F, we have the absolutely convergent power series expansion

$$N_F^{l/2} W_{l,F} = (I + A_F)^{-1} = \sum_{r=0}^{\infty} (-1)^r A_F^r = I + O(N_F^{-1}) \quad (N_F \to \infty).$$

The result now follows. □

The following proposition is a consequence of Theorem 4.1 and gives an asymptotic factorization of the normalized joint moments over O_F^+.

Proposition 4.2. *Fix $l \in 2\mathbb{N}$ and $i, j : [l] \to [N]$. Then there is a constant $D_l > 0$ (depending only on l) such that*

$$N_F^{l/2}\left|h_{O_F^+}(u_{i(1)j(1)}u_{i(2)j(2)}\cdots u_{i(l)j(l)}) - \sum_{\pi\in\mathcal{NC}_2(l)}\prod_{(s,t)\in\pi} h_{O_F^+}(u_{i(s)j(s)}u_{i(t)j(t)})\right| \\ \leq \frac{D_l}{N_F}\cdot\max_{\pi,\sigma\in\mathcal{NC}_2(l)}|\delta_\pi^F(j)\delta_\sigma^F(i)|.$$

Proof. Using Theorem 4.1, one can find a constant $D_l > 0$ such that

$$\sum_{\pi,\sigma\in\mathcal{NC}_2(l)}|N_F^{l/2}W_{l,F}(\pi,\sigma) - \delta_{\pi,\sigma}| \leq \frac{D_l}{N_F}.$$

Since it also follows from Theorem 3.9 that

$$\begin{aligned}\sum_{\pi\in\mathcal{NC}_2(l)}\prod_{(s,t)\in\pi} h_{O_F^+}(u_{i(s)j(s)}u_{i(t)j(t)}) &= \sum_{\pi\in\mathcal{NC}_2(l)}\prod_{(s,t)\in\pi} N_F^{-1}F_{i(t)i(s)}\overline{F_{j(t)j(s)}} \\ &= N_F^{-l/2}\sum_{\pi\in\mathcal{NC}_2(l)}\delta_\pi^F(i)\overline{\delta_\pi^F(j)},\end{aligned}$$

we obtain

$$\begin{aligned}N_F^{l/2}&\left|h_{O_F^+}(u_{i(1)j(1)}u_{i(2)j(2)}\cdots u_{i(l)j(l)}) - \sum_{\pi\in\mathcal{NC}_2(l)}\prod_{(s,t)\in\pi} h_{O_F^+}(u_{i(s)j(s)}u_{i(t)j(t)})\right| \\ &= \left|\sum_{\pi,\sigma\in\mathcal{NC}_2(l)} N_F^{l/2}(W_{l,F}(\pi,\sigma)-\delta_{\pi,\sigma})\overline{\delta_\pi^F(j)}\delta_\sigma^F(i)\right| \\ &\leq \frac{D_l}{N_F}\cdot\max_{\pi,\sigma\in\mathcal{NC}_2(l)}\left|\delta_\pi^F(j)\delta_\sigma^F(i)\right|.\end{aligned}$$ ☐

Using Propositions 3.11 and 4.2, we obtain a similar asymptotic factorization result for $*$-moments.

Corollary 4.3. *Fix $l \in 2\mathbb{N}$, $\epsilon : [l] \to \{1, *\}$, and $i, j : [l] \to [N]$. Then there is a constant $D_l > 0$ (depending only on l) such that*

$$N_F^{l/2}\left|h_{O_F^+}\big(u_{i(1)j(1)}^{\epsilon(1)}\cdots u_{i(l)j(l)}^{\epsilon(l)}\big) - \sum_{\pi\in\mathcal{NC}_2(l)}\prod_{(s,t)\in\pi} h_{O_F^+}\big(u_{i(s)j(s)}^{\epsilon(s)}u_{i(t)j(t)}^{\epsilon(t)}\big)\right| \\ \leq \frac{D_l}{N_F}\cdot\max_{\pi,\sigma\in\mathcal{NC}_2(l)}\left|\delta_\pi^F(j_\epsilon)\delta_\sigma^F(i_\epsilon)\right|\prod_{r=1}^{l}|t_F(i(r),j(r),\epsilon(r))|. \tag{4-1}$$

5. Asymptotic freeness in O_F^+

We now arrive at the main asymptotic freeness results of this paper.

Let us fix a canonical matrix

$$F = F_\rho^{(c)} = \begin{pmatrix} 0 & D_k(\rho) & 0 \\ cD_k(\rho)^{-1} & 0 & 0 \\ 0 & 0 & 1_{N-2k} \end{pmatrix} \in \mathrm{GL}_N(\mathbb{C}),$$

as in Section 3C. Recall from Remark 3.10 that the subset of (rescaled) matrix elements

$$\mathcal{S}_F = \{\sqrt{N_F}\,u_{ij}\}_{\substack{1\le i\le k\\ 1\le j\le N}} \cup \{\sqrt{N_F}\,u_{ij}\}_{\substack{1\le j\le k\\ 2k+1\le i\le N}} \cup \{\sqrt{N_F}\,u_{ij}\}_{2k+1\le i,j\le N}$$

generates the C*-algebra $C(O_F^+)$. In this section we show that the set $\mathcal{S}_F$ is asymptotically free in the following sense.

Theorem 5.1. *Let*

$$\mathcal{S} = \{y_{ij}\}_{\substack{1\le i\le k\\ 1\le j\le N}} \cup \{y_{ij}\}_{\substack{1\le j\le k\\ 2k+1\le i\le N}} \cup \{y_{ij}\}_{2k+1\le i,j\le N}$$

be a family of noncommutative random variables in an NCPS (A,φ) with the following properties.

(1) *$\mathcal{S}$ is freely independent.*

(2) *For each i, j, the elements $y_{ij} \in \mathcal{S}$ and $\sqrt{N_F}\,u_{ij} \in \mathcal{S}_F$ have the same left and right variances, given by the matrix entries of $N_F\Phi$ in (3-9).*

(3) *If either $i \le k$ or $j \le k$, then each y_{ij} is a generalized circular element.*

(4) *If $2k+1 \le i, j \le N$, then y_{ij} is a standard semicircular element.*

Then for each $l \in 2\mathbb{N}$, there is a constant $D_l > 0$ such that

$$\begin{aligned}(5\text{-}1)\quad &\Big|h_{O_F^+}\big(\sqrt{N_F}\,u_{i(1)j(1)}^{\epsilon(1)}\cdots\sqrt{N_F}\,u_{i(l)j(l)}^{\epsilon(l)}\big) - \varphi\big(y_{i(1)j(1)}^{\epsilon(1)}\cdots y_{i(l)j(l)}^{\epsilon(l)}\big)\Big| \\ &\qquad\le \frac{D_l}{N_F}\cdot \max_{\pi,\sigma\in\mathcal{NC}_2(l)}\big|\delta_\pi^F(j_\epsilon)\,\delta_\sigma^F(i_\epsilon)\big| \prod_{r=1}^{l}|t_F(i(r), j(r), \epsilon(r))|,\end{aligned}$$

for each $\epsilon : [l] \to \{1,*\}$ and $i, j : [l] \to [N]$.

Proof. Since $\mathcal{S}$ is a free family consisting of generalized circular elements and standard semicircular elements satisfying conditions (2)–(4) above, Theorem 2.2 gives

$$\begin{aligned}\varphi\big(y_{i(1)j(1)}^{\epsilon(1)}\cdots y_{i(l)j(l)}^{\epsilon(l)}\big) &= \sum_{\pi\in\mathcal{NC}_2(l)}\prod_{(s,t)\in\pi}\varphi\big(y_{i(s)j(s)}^{\epsilon(s)}y_{i(t)j(t)}^{\epsilon(t)}\big) \\ &= \sum_{\pi\in\mathcal{NC}_2(l)}\prod_{(s,t)\in\pi}N_F h_{O_F^+}\big(u_{i(s)j(s)}^{\epsilon(s)}u_{i(t)j(t)}^{\epsilon(t)}\big).\end{aligned}$$

The theorem now follows from Corollary 4.3. ∎

5A. *Asymptotic freeness in the large dimension limit.* Using Theorem 5.1, we see that the quantum groups O_F^+ provide asymptotic models for almost-periodic free Araki–Woods factors. That is, canonical generators of almost-periodic free Araki–Woods factors can be approximated in joint distribution by normalized coordinates over a suitable sequence of O_F^+ quantum groups.

To see this, let $\Gamma(H_{\mathbb{R}}, U_t)''$ be an almost-periodic free Araki–Woods factor. Then we can write $\Gamma(H_{\mathbb{R}}, U_t)'' = (z_i : i \in \mathbb{N})''$, where $(z_i)_{i\in\mathbb{N}}$ is a free family of $(1, \lambda_i)$-generalized circular elements z_i (with $1 < \lambda_i < \infty$). To approximate the joint $*$-distribution of $(z_i)_{i\in\mathbb{N}}$, define, for each $k \in \mathbb{N}$,

$$F(k) = \begin{pmatrix} 0 & D_{k+1}(1, \sqrt{\lambda_1}, \dots, \sqrt{\lambda_k})^{-1} \\ -D_{k+1}(1, \sqrt{\lambda_1}, \dots, \sqrt{\lambda_k}) & 0 \end{pmatrix} \in \mathrm{GL}_{2k+2}(\mathbb{C}).$$

Theorem 5.2. *The family of noncommutative random variables*

$$(z_{i,k})_{i=1}^k = \big(\sqrt{N_{F(k)}}\, u_{1,i+1}\big)_{i=1}^k \subset \big(C(O^+_{F(k)}), h_{O^+_{F(k)}}\big)$$

converges in joint distribution as $k \to \infty$ to $(z_i)_{i\in\mathbb{N}} \subset (\Gamma(H_{\mathbb{R}}, U_t)'', \varphi_\Omega)$.

Proof. By construction, $(z_{i,k})_{i=1}^k$ and $(z_i)_{i=1}^k$ have the same left and right variances. By Theorem 5.1, we then have for any $l \in 2\mathbb{N}$, $\epsilon : [l] \to \{1, *\}$, $i : [l] \to \mathbb{N}$, and $k = k(i)$ sufficiently large,

$$\begin{aligned} &\big|h_{O^+_{F(k)}}\big(z_{i(1),k}^{\epsilon(1)} \cdots z_{i(l),k}^{\epsilon(l)}\big) - \varphi_\Omega\big(z_{i(1)}^{\epsilon(1)} \cdots z_{i(l)}^{\epsilon(l)}\big)\big| \\ &\qquad \le \frac{D_l}{N_{F(k)}} \cdot \max_{\pi,\sigma\in\mathcal{NC}_2(l)} \big|\delta_\pi^{F(k)}(1_\epsilon)\,\delta_\sigma^{F(k)}((i+1)_\epsilon)\big| \prod_{r=1}^{l} |t_{F(k)}(1, i(r)+1, \epsilon(r))|. \end{aligned}$$

Since both quantities

$$\max_{\pi,\sigma\in\mathcal{NC}_2(l)} \big|\delta_\pi^{F(k)}(1_\epsilon)\,\delta_\sigma^{F(k)}((i+1)_\epsilon)\big| \quad \text{and} \quad \prod_{r=1}^{l} |t_{F(k)}(1, i(r)+1, \epsilon(r))|$$

are constant once l, i, and ϵ are fixed, and $N_{F(k)} = \mathrm{Tr}(F(k)^*F(k)) \ge \mathrm{Tr}(1) = 2k+2$, we conclude that

$$\big|h_{O^+_{F(k)}}\big(z_{i(1),k}^{\epsilon(1)} \cdots z_{i(l),k}^{\epsilon(l)}\big) - \varphi_\Omega\big(z_{i(1)}^{\epsilon(1)} \dots z_{i(l)}^{\epsilon(l)}\big)\big| \le \frac{\text{constant}}{2k+2} \to 0. \qquad \square$$

Remark 5.3. Using the same reasoning as in the proof of Theorem 5.2, it is easy to see that for the above sequence of quantum groups $(O^+_{F(k)})_{k\in\mathbb{N}}$, the entire family of normalized generators

$$\mathcal{S}_{F(k)} = \{\sqrt{N_{F(k)}}\, u_{ij}\}_{1\le i,j\le k} \cup \{\sqrt{N_{F(k)}}\, u_{i,j+k}\}_{1\le i,j\le k}$$

of $C(O^+_{F(k)})$ converges in distribution to a free family of generalized circular elements $\{y_{ij}\}_{1\le i,j<\infty} \cup \{w_{ij}\}_{1\le i,j<\infty}$ in an NCPS (A, φ) with the following left and right variances (determined by Theorem 5.1):

$$\varphi(y_{ij}^* y_{ij}) = \rho_i^2, \quad \varphi(y_{ij} y_{ij}^*) = \rho_j^{-2}, \quad \varphi(w_{ij}^* w_{ij}) = \rho_i^2, \quad \varphi(w_{ij} w_{ij}^*) = \rho_j^2,$$

where $\rho_1 = 1$ and $\rho_i = \lambda_{i-1}^{-1/2}$ for $i \geq 2$. Note also that there is a state-preserving $*$-isomorphism $\mathrm{W}^*(\{y_{ij}\}_{1\leq i,j<\infty} \cup \{w_{ij}\}_{1\leq i,j<\infty})$ and $\mathrm{W}^*((z_i)_{i\in\mathbb{N}}) = \Gamma(H_{\mathbb{R}}, U_t)''$, the almost-periodic free Araki–Woods factor we started with. (This isomorphism follows from [Shlyakhtenko 1997, Theorem 6.4]).

5B. ***Asymptotic freeness in finite dimensions.*** In Theorem 5.2, we saw that normalized generators of a suitable family of the O_F^+ converge in distribution to free random variables as the size N of the matrices $F \in \mathrm{GL}_N(\mathbb{C})$ go to infinity. On the other hand, the general estimate of Theorem 5.1 shows that in the nontracial setting, the rate of approximation to freeness is governed by the growth of the quantum dimension $N_F = \mathrm{Tr}(F^*F)$, and not the physical dimension N. This phenomenon allows one to consider scenarios where $N_F \to \infty$, while the dimension N of $F \in \mathrm{GL}_N(\mathbb{C})$ is fixed. This is illustrated by the next theorem.

Theorem 5.4. *Fix $k \in \mathbb{N}$ and a sequence $\rho = (\rho_1, \ldots, \rho_k) \in (0, 1)^k$ and let*

$$F(\gamma) = \begin{pmatrix} 0 & D_{k+1}(\rho) & 0 & 0 \\ D_{k+1}(\rho)^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma \\ 0 & 0 & \gamma^{-1} & 0 \end{pmatrix} \in \mathrm{GL}_{2k+2}(\mathbb{C}) \quad (0 < \gamma < 1).$$

Then the subset of generators

$$\tilde{\mathcal{S}}_{F(\gamma)} = \{\sqrt{N_{F(\gamma)}}\, u_{ij}\}_{\substack{1\leq i\leq k \\ 1\leq j\leq 2k}} \subset \big(C(O_{F(\gamma)}^+), h_{O_{F(\gamma)}^+}\big)$$

converges in distribution to a free family of generalized circular elements

$$\tilde{\mathcal{S}} = \{y_{ij}\}_{\substack{1\leq i\leq k \\ 1\leq j\leq 2k}}$$

in a NCPS (A, φ) with left and right variances given by

$$\varphi(y_{ij}^* y_{ij}) = \rho_i^2 \quad \textit{and} \quad \varphi(y_{ij} y_{ij}^*) = \begin{cases} \rho_j^{-2}, & 1 \leq j \leq k, \\ \rho_j^2, & k+1 \leq j \leq 2k. \end{cases}$$

Remark 5.5. Note that Theorem 5.4, makes a statement about the limiting distribution of a subset $\tilde{\mathcal{S}}_{F(\gamma)}$ of generators of $C(O_{F(\gamma)}^+)$. We cannot make a statement about the asymptotic freeness of the entire family of generators $\mathcal{S}_{F(\gamma)}$ in this setting since some of these variables do not have limiting distributions. For example,

$$h_{O_{F(\gamma)}^+}\big(\sqrt{N_{F(\gamma)}}\, u_{k+1,k+1}^{\epsilon(1)} \sqrt{N_{F(\gamma)}}\, u_{k+1,k+1}^{\epsilon(2)}\big) = \begin{cases} \gamma^2, & \epsilon(1) = *,\ \epsilon(2) = 1, \\ \gamma^{-2}, & \epsilon(1) = 1,\ \epsilon(2) = *, \end{cases}$$

which implies that $\sqrt{N_{F(\gamma)}}\, u_{k+1,k+1}$ does not have a limiting distribution as $\gamma \to 0$.

Proof of Theorem 5.4. By construction, the families $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{S}}_{F(\gamma)}$ have the same left and right variances (which are independent of γ). Therefore, we may apply Theorem 5.1 to obtain, for any $l \in 2\mathbb{N}$, $\epsilon : [l] \to \{1, *\}$, $i : [l] \to [k]$, $j : [l] \to [2k]$,

$$\begin{aligned}\Big|h_{O^+_{F(\gamma)}}\big(\sqrt{N_{F(\gamma)}}\,u^{\epsilon(1)}_{i(1)j(1)} \cdots \sqrt{N_{F(\gamma)}}\,u^{\epsilon(l)}_{i(l)j(l)}\big) - \varphi\big(y^{\epsilon(1)}_{i(1)j(1)} \cdots y^{\epsilon(l)}_{i(l)j(l)}\big)\Big| \\ \leq \frac{D_l}{N_{F(\gamma)}} \cdot \max_{\pi,\sigma \in \mathcal{NC}_2(l)} \big|\delta^{F(\gamma)}_{\pi}(j_\epsilon)\,\delta^{F(\gamma)}_{\sigma}(i_\epsilon)\big| \prod_{r=1}^{l} |t_{F(\gamma)}(i(r), j(r), \epsilon(r))|.\end{aligned}$$

Since the numerator of the above expression is fixed with respect to γ and $N_{F(\gamma)} = \gamma^2 + \gamma^{-2} + \sum_{i=1}^{k}(\rho_i^2 + \rho_i^{-2}) \to \infty$ as $\gamma \to 0$, the theorem follows. □

5C. ***A remark on the free unitary case.*** Let $c = \pm 1$ and $F \in \mathrm{GL}_N(\mathbb{C})$ be a canonical F-matrix. In this section we briefly comment on the free unitary quantum groups U_F^+.

In this case, it is known from the fundamental work of Banica [1997] that there is an injective $*$-homomorphism $L^\infty(U_F^+) \hookrightarrow L^\infty(\mathbb{T}) * L^\infty(O_F^+)$, the unital free product of $C(\mathbb{T})$ and $C(O_F^+)$, given by $v_{ij} \mapsto w u_{ij}$. Here, $w \in C(\mathbb{T})$ is canonical unitary coordinate function on the unit circle $\mathbb{T}$. Moreover, it is known that under the above embedding, $h_{U_F^+} = (\tau * h_{O_F^+})|_{C(U_F^+)}$, where τ denotes integration with respect to the Haar probability measure on $\mathbb{T}$.

In other words, the variables $\{v_{ij}\}_{1\le i,j\le N} \subset (C(U_F^+), h_{U_F^+})$ and $\{w u_{ij}\}_{1\le i,j\le N} \subset (C(\mathbb{T}) * C(O_F^+), \tau * h_{O_F^+})$ are identically distributed. Using this observation, together with some basic facts about free independence and the results we have already obtained on O_F^+, we arrive at the following unitary version of Theorem 5.1, whose proof we leave as an exercise to the reader. (Note that the extra freeness inside $C(U_F^+)$ given by the above free product model yields a slightly cleaner statement than Theorem 5.1.)

Theorem 5.6. *Fix a canonical deformation matrix*

$$F = F_\rho^{(c)} = \begin{pmatrix} 0 & D_k(\rho) & 0 \\ cD_k(\rho)^{-1} & 0 & 0 \\ 0 & 0 & 1_{N-2k} \end{pmatrix} \in \mathrm{GL}_N(\mathbb{C}),$$

and let $\mathcal{S} = \{y_{ij}\}_{1\le i,j\le N}$ be a family of noncommutative random variables in an NCPS (A, φ) with the following properties.

1. *$\mathcal{S}$ is freely independent.*
2. *For each i, j, the elements $y_{ij} \in \mathcal{S}$ and $\sqrt{N_F}\, v_{ij} \in C(U_F^+)$ have the same left and right variances, given by the matrix entries of $N_F \Phi$ in (3-9).*
3. *Each y_{ij} is a generalized circular element.*

Then for each $l \in 2\mathbb{N}$, there is a constant $D_l > 0$ such that

$$\left|h_{U_F^+}\left(\sqrt{N_F}\, v^{\epsilon(1)}_{i(1)j(1)} \cdots \sqrt{N_F}\, v^{\epsilon(l)}_{i(l)j(l)}\right) - \varphi\left(y^{\epsilon(1)}_{i(1)j(1)} \cdots y^{\epsilon(l)}_{i(l)j(l)}\right)\right| \\ \leq \frac{D_l}{N_F} \cdot \max_{\pi,\sigma \in \mathcal{NC}_2(l)} \left|\delta^F_\pi(j_\epsilon)\,\delta^F_\sigma(i_\epsilon)\right| \prod_{r=1}^{l} |t_F(i(r), j(r), \epsilon(r))|,$$

for each $\epsilon : [l] \to \{1, *\}$ and $i, j : [l] \to [N]$.

Remark 5.7. The most general version of Theorem 5.6 for U_F^+ with $F \in \mathrm{GL}_N(\mathbb{C})$ arbitrary is not known to the authors and requires further investigation.

6. O_F^+ and quantum symmetries of free Araki–Woods factors

In this final section we return to the free orthogonal quantum groups O_F^+ and investigate to what extent they can be regarded as quantum symmetries of free Araki–Woods factors. Inspired by the fact that the free group factors $L(\mathbb{F}_N)$ admit O_N^+ as natural quantum symmetries [Curran 2010], we are interested in finding canonical families of (nontracial) noncommutative random variables $(x_1, \dots, x_N)$ belonging to an NCPS (A, φ) whose joint distribution is O_F^+-invariant in the following sense.

Definition 6.1. Let (A, φ) be an NCPS and consider $\boldsymbol{x} = (x_1, \dots, x_N) \subset A$. Fix $F \in \mathrm{GL}_N(\mathbb{C})$ and let $U = [u_{ij}] \in M_N(C(O_F^+))$ be the fundamental representation of O_F^+. We say that $\boldsymbol{x}$ *has an O_F^+-invariant joint distribution* (or, that $\boldsymbol{x}$ is *O_F^+-rotatable*) if for any $l \in \mathbb{N}$, $i : [l] \to [N]$, and any $\epsilon : [l] \to \{1, *\}$,

$$\sum_{j:[l]\to N} u^{\epsilon(1)}_{i(1)j(1)} u^{\epsilon(2)}_{i(2)j(2)} \cdots u^{\epsilon(l)}_{i(l)j(l)}\, \varphi\left(x^{\epsilon(1)}_{j(1)} x^{\epsilon(2)}_{j(2)} \cdots x^{\epsilon(l)}_{j(l)}\right) \\ = \varphi\left(x^{\epsilon(1)}_{i(1)} x^{\epsilon(2)}_{i(2)} \cdots x^{\epsilon(l)}_{i(l)}\right) 1. \tag{6-1}$$

We note that the existence of an N-tuple $\boldsymbol{x}$ with the above O_F^+-invariance property is connected to the existence of an *action* of the quantum group O_F^+ on the von Neumann algebra generated by $\boldsymbol{x}$.

Definition 6.2. Let $\mathbb{G}$ be a compact quantum group with von Neumann algebra $L^\infty(\mathbb{G}) = \pi_h(C(\mathbb{G}))''$ and (extended) coproduct $\Delta_r : L^\infty(\mathbb{G}) \to L^\infty(\mathbb{G}) \mathbin{\overline{\otimes}} L^\infty(\mathbb{G})$. Let M a von Neumann algebra.

(1) A *left action* (or simply an *action*) of $\mathbb{G}$ on M is a normal, injective, and unital $*$-homomorphism $\alpha : M \to L^\infty(\mathbb{G}) \mathbin{\overline{\otimes}} M$ such that $(\iota \otimes \alpha) \circ \alpha = (\Delta \otimes \iota) \circ \alpha$. We denote the action of $\mathbb{G}$ on M by the notation $\mathbb{G} \curvearrowright^\alpha M$.

(2) If φ is a faithful normal state on M, we say that an action $\mathbb{G} \curvearrowright^\alpha M$ is *φ-preserving* if $(\iota \otimes \varphi) \circ \alpha = \varphi(\,\cdot\,) 1_{L^\infty(\mathbb{G})}$. Such an action has notation $\mathbb{G} \curvearrowright^\alpha (M, \varphi)$.

Remark 6.3. If $\boldsymbol{x} = (x_1, \dots, x_N) \subset (A, \varphi)$ is an N-tuple with an O_F^+-invariant joint distribution, then it is easy to see that

$$\alpha(x_i) = \sum_{j=1}^{N} \pi_h(u_{ij}) \otimes x_j \quad (1 \le i \le n)$$

defines a φ-preserving action $O_F^+ \curvearrowright^\alpha (\mathrm{W}^*(x_1, \dots, x_N), \varphi)$, where $\pi_h : C(O_F^+) \to L^\infty(O_F^+)$ is the GNS representation associated to the Haar state.

We now show that such O_F^+-invariant noncommutative random variables exist for any F. To do this, we first need a lemma.

Lemma 6.4. *Let $Q = F^t\bar{F}$, where $F \in \mathrm{GL}_N(\mathbb{C})$ is a canonical F matrix (so that, in particular, Q is diagonal), and let $U = [u_{ij}] \in M_N(C(O_F^+))$ be the fundamental representation of O_F^+. Then*

$$\sum_{r=1}^{N} u_{ir} u_{jr}^* = \delta_{ij} 1 \quad \textit{and} \quad \sum_{r=1}^{N} u_{ir}^* u_{jr} (Q^{-1})_{rr} = \delta_{ij} (Q^{-1})_{ii} 1 \quad (1 \le i, j \le N).$$

Proof. The first equality is a direct consequence of the fact that the fundamental representation U is unitary. To prove the second inequality, we use [Timmermann 2008, Proposition 3.2.17] which shows that $\overline{U}\,\overline{Q}^{-1} U^t = \overline{Q}^{-1}$. Since Q is diagonal and positive definite, the result follows. □

In the case where $F \in \mathrm{GL}_N(\mathbb{C})$ is canonical, the way to find O_F^+-rotatable generators of a free Araki–Woods factor is to fix a column of the fundamental representation of O_F^+ and take an N-tuple of freely independent generalized circular elements with the same left and right variances as this column (up to a common nonzero scaling factor). Again, we note that after completion of a first draft of this paper, it was pointed out to the authors that the following theorem can also be obtained as a consequence of [Vaes 2005, Proposition 3.1].

Theorem 6.5. *Let $F \in \mathrm{GL}_N(\mathbb{C})$ be a canonical F matrix and let $Q = F^t\bar{F}$ with diagonal entries $(Q_{ii})_{i=1}^N$. Let $\boldsymbol{x} = (x_1, \dots, x_N) \subset (A, \varphi)$ be a $*$-free family of generalized circular elements with left and right covariances given by*

$$\varphi(x_i^* x_i) = Q_{ii}^{-1}, \quad \varphi(x_i x_i^*) = 1, \quad (1 \le i \le N).$$

Then $\boldsymbol{x}$ has an O_F^+-invariant joint $*$-distribution. In other words, there is a φ-preserving action $O_F^+ \curvearrowright^\alpha (\mathrm{W}^*(x_1, \dots, x_N), \varphi)$ given by

$$\alpha(x_i) = \sum_{j=1}^{N} \pi_h(u_{ij}) \otimes x_j,$$

where $U = [u_{ij}]$ is the fundamental representation of O_F^+.

Proof. We must verify (6-1) for the N-tuple $\boldsymbol{x}$, for each choice of $l \in \mathbb{N}$, $i : [l] \to [N]$, and $\epsilon : [l] \to \{1, *\}$. To start, observe that when l is odd or $|\epsilon^{-1}(1)| \neq |\epsilon^{-1}(*)|$, then both sides of (6-1) are always zero. Therefore, we assume $l \in 2\mathbb{N}$ and that $|\epsilon^{-1}(1)| = |\epsilon^{-1}(*)| = l/2$.

We begin by considering the case $l = 2$, and fix $1 \le i(1), i(2) \le N$, $\epsilon(1) \neq \epsilon(2) \in \{1, *\}$. Then we have

$$\sum_{j(1),j(2)=1}^{N} u_{i(1)j(1)}^{\epsilon(1)} u_{i(2)j(2)}^{\epsilon(2)} \varphi\big(x_{j(1)}^{\epsilon(1)} x_{j(2)}^{\epsilon(2)}\big) = \sum_{j(1)=1}^{N} u_{i(1)j(1)}^{\epsilon(1)} u_{i(2)j(1)}^{\epsilon(2)} \varphi\big(x_{j(1)}^{\epsilon(1)} x_{j(1)}^{\epsilon(2)}\big).$$

If $\epsilon(1) = 1$ and $\epsilon(2) = *$, then $\varphi(x_{j(1)} x_{j(1)}^*) = 1$ and the above quantity equals

$$\sum_{j(1)=1}^{N} u_{i(1)j(1)} u_{i(2)j(1)}^* = \delta_{i(1),i(2)} 1 = \varphi(x_{i(1)} x_{i(2)}^*) 1$$

by Lemma 6.4. If $\epsilon(1) = *$ and $\epsilon(2) = 1$, then $\varphi(x_{j(1)}^* x_{j(1)}) = Q_{j(1)j(1)}^{-1}$ and the above quantity equals

$$\sum_{j(1)=1}^{N} u_{i(1)j(1)}^* u_{i(2)j(1)} Q_{j(1)j(1)}^{-1} = \delta_{i(1),i(2)} \varphi(x_{i(1)}^* x_{i(1)}) 1 = \varphi(x_{i(1)}^* x_{i(2)}) 1,$$

again by Lemma 6.4. In each case, we obtain

$$\sum_{j(1),j(2)=1}^{N} u_{i(1)j(1)}^{\epsilon(1)} u_{i(2)j(2)}^{\epsilon(2)} \varphi\big(x_{j(1)}^{\epsilon(1)} x_{j(2)}^{\epsilon(2)}\big) = \varphi\big(x_{i(1)}^{\epsilon(1)} x_{i(2)}^{\epsilon(2)}\big) 1. \tag{6-2}$$

Now let $2 < l \in 2\mathbb{N}$ and fix $\epsilon : [l] \to \{1, *\}$ and $i : [l] \to [N]$. Then from Theorem 2.2,

$$\begin{aligned}
&\sum_{j:[l]\to N} u_{i(1)j(1)}^{\epsilon(1)} u_{i(2)j(2)}^{\epsilon(2)} \cdots u_{i(l)j(l)}^{\epsilon(l)} \varphi\big(x_{j(1)}^{\epsilon(1)} x_{j(2)}^{\epsilon(2)} \cdots x_{j(l)}^{\epsilon(l)}\big) \\
&\qquad = \sum_{j:[l]\to N} u_{i(1)j(1)}^{\epsilon(1)} u_{i(2)j(2)}^{\epsilon(2)} \cdots u_{i(l)j(l)}^{\epsilon(l)} \Bigg(\sum_{\pi \in \mathcal{NC}_2^{\epsilon}(l)} \prod_{(s,t)\in\pi} \varphi\big(x_{j(s)}^{\epsilon(s)} x_{j(t)}^{\epsilon(t)}\big) \Bigg) \\
&\qquad = \sum_{\pi \in \mathcal{NC}_2^{\epsilon}(l)} \Bigg(\sum_{j:[l]\to N} u_{i(1)j(1)}^{\epsilon(1)} u_{i(2)j(2)}^{\epsilon(2)} \cdots u_{i(l)j(l)}^{\epsilon(l)} \prod_{(s,t)\in\pi} \varphi\big(x_{j(s)}^{\epsilon(s)} x_{j(t)}^{\epsilon(t)}\big) \Bigg).
\end{aligned}$$

Fix $\pi \in \mathcal{NC}_2^{\epsilon}(l)$ and consider the internal sum above. Since π is noncrossing, it contains a neighboring pair $(r, r+1)$. Applying (6-2) to the partial sum over

$1 \le j(r), j(r+1) \le N$, we obtain

$$\sum_{j:[l]\to[N]} u^{\epsilon(1)}_{i(1)i(1)} u^{\epsilon(2)}_{i(2)i(2)} \cdots u^{\epsilon(l)}_{i(l)i(l)} \prod_{(s,t)\in\pi} \varphi\big(x^{\epsilon(s)}_{j(s)} x^{\epsilon(t)}_{j(t)}\big)$$
$$= \varphi(x^{\epsilon(r)}_{i(r)} x^{\epsilon(r+1)}_{i(r+1)})$$
$$\times \Bigg(\sum_{j:[l]\setminus\{r,r+1\}\to[N]} u^{\epsilon(1)}_{i(1)i(1)} \cdots u^{\epsilon(r-1)}_{i(r-1)i(r-1)} u^{\epsilon(r+2)}_{i(r+2)i(r+2)} \cdots u^{\epsilon(l)}_{i(l)i(l)} \prod_{(s,t)\in\pi\setminus(r,r+1)} \varphi\big(x^{\epsilon(s)}_{j(s)} x^{\epsilon(t)}_{j(t)}\big)\Bigg).$$

Repeatedly applying the same principle to this new internal sum of lower order (note that $\pi \setminus (r, r+1)$ is again noncrossing), after a total of $l/2-1$ steps, we get

$$\sum_{j:[l]\to[N]} u^{\epsilon(1)}_{i(1)i(1)} u^{\epsilon(2)}_{i(2)i(2)} \cdots u^{\epsilon(l)}_{i(l)i(l)} \prod_{(s,t)\in\pi} \varphi\big(x^{\epsilon(s)}_{j(s)} x^{\epsilon(t)}_{j(t)}\big) = \prod_{(s,t)\in\pi} \varphi\big(x^{\epsilon(s)}_{i(s)} x^{\epsilon(t)}_{i(t)}\big)1. \tag{6-3}$$

Therefore,

$$\sum_{j:[l]\to N} u^{\epsilon(1)}_{i(1)j(1)} u^{\epsilon(2)}_{i(2)j(2)} \cdots u^{\epsilon(l)}_{i(l)j(l)} \varphi\big(x^{\epsilon(1)}_{j(1)} x^{\epsilon(2)}_{j(2)} \cdots x^{\epsilon(l)}_{j(l)}\big)$$
$$= \sum_{\pi\in\mathcal{NC}^{\epsilon}_2(l)} \prod_{(s,t)\in\pi} \varphi\big(x^{\epsilon(s)}_{i(s)} x^{\epsilon(t)}_{i(t)}\big)1 = \varphi\big(x^{\epsilon(1)}_{i(1)} x^{\epsilon(2)}_{i(2)} \cdots x^{\epsilon(l)}_{i(l)}\big)1. \qquad \square$$

Remark 6.6. It is clear that $(\mathrm{W}^*(x_1, \dots, x_N), \varphi) \cong (\Gamma(\mathcal{H}_{\mathbb{R}}, U_t)'', \varphi_\Omega)$ is a free Araki–Woods factor associated to a finite dimensional orthogonal representation $(U_t)_{r\in\mathbb{R}}$ (compare with Theorem 2.3). Moreover, it is interesting to note that the type classification for the von Neumann algebras $L^\infty(O^+_F)$ and $\Gamma(H_{\mathbb{R}}, U_t)''$ is the same. More precisely, if $\Gamma < \mathbb{R}^*_+$ is the subgroup generated by the eigenvalues of $Q \otimes Q^{-1}$, then both of these algebras are type II_1 when $Q = 1$, type III_λ if $\Gamma = \lambda^{\mathbb{Z}}$, and type III_1 otherwise. Compare [Shlyakhtenko 1997, Theorem 6.1; Vaes and Vergnioux 2007, Theorem 7.1].

Remark 6.7. Theorem 6.5 only considers the case of a canonical matrix F. For generic $F \in \mathrm{GL}_N(\mathbb{C})$ such that $F\bar{F} = c1$, recall from Section 3A that there is a canonical F-matrix $F^{(c)}_\rho \in \mathrm{GL}_N(\mathbb{C})$ and $v \in \mathcal{U}_N$ such that $F^{(c)}_\rho = vFv^t$ and $O^+_F \cong O^+_{F^{(c)}_\rho}$. Then $O^+_F \curvearrowright^{\alpha_F} (\Gamma(\mathcal{H}_{\mathbb{R}}, U_t)'', \varphi_\Omega)$, where $\Gamma(\mathcal{H}_{\mathbb{R}}, U_t)''$ is the free Araki–Woods factor on which $O^+_{F^{(c)}_\rho}$ acts in the sense of the above theorem. Indeed let $\boldsymbol{x} = (x_1, \dots, x_N)$ be the generalized circular system constructed in Theorem 6.5, let $\alpha_{F^{(c)}_\rho}$ be the corresponding action, and let $\boldsymbol{y} = v\boldsymbol{x}$. Then $\mathrm{W}^*(\boldsymbol{x}) = \mathrm{W}^*(\boldsymbol{y})$ and one readily checks from the defining relations that

$$W = vUv^*,$$

where $W = [w_{ij}]$ and $U = [u_{ij}]$ are the fundamental representations of O^+_F and $O^+_{F^{(c)}_\rho}$, respectively. A simple calculation then shows that condition (6-1) holds with the w_{ij} replacing the u_{ij} and the y_i replacing the x_i.

Acknowledgments

Brannan was partially supported by an NSERC Postdoctoral Fellowship. Kirkpatrick was partially supported by NSF Grant DMS-1106770 and NSF CAREER Award DMS-1254791. Both Brannan and Kirkpatrick wish to thank Todd Kemp for encouraging them to collaborate.

References

[Banica 1996] T. Banica, "Théorie des représentations du groupe quantique compact libre O(n)", *C. R. Acad. Sci. Paris Sér. I Math.* **322**:3 (1996), 241–244. MR 97a:46108 Zbl 0862.17010

[Banica 1997] T. Banica, "Le groupe quantique compact libre U(n)", *Comm. Math. Phys.* **190**:1 (1997), 143–172. MR 99k:46095 Zbl 0906.17009

[Banica and Collins 2007] T. Banica and B. Collins, "Integration over compact quantum groups", *Publ. Res. Inst. Math. Sci.* **43**:2 (2007), 277–302. MR 2008m:46137 Zbl 1129.46058

[Banica et al. 2009] T. Banica, B. Collins, and P. Zinn-Justin, "Spectral analysis of the free orthogonal matrix", *Int. Math. Res. Not.* **2009**:17 (2009), 3286–3309. MR 2011d:46133 Zbl 1179.46056

[Bichon et al. 2006] J. Bichon, A. De Rijdt, and S. Vaes, "Ergodic coactions with large multiplicity and monoidal equivalence of quantum groups", *Comm. Math. Phys.* **262**:3 (2006), 703–728. MR 2007a:46072 Zbl 1122.46046

[Collins 2003] B. Collins, "Moments and cumulants of polynomial random variables on unitary groups, the Itzykson–Zuber integral, and free probability", *Int. Math. Res. Not.* **2003**:17 (2003), 953–982. MR 2003m:28015 Zbl 1049.60091

[Collins and Śniady 2006] B. Collins and P. Śniady, "Integration with respect to the Haar measure on unitary, orthogonal and symplectic group", *Comm. Math. Phys.* **264**:3 (2006), 773–795. MR 2007c:60009 Zbl 1108.60004

[Curran 2010] S. Curran, "Quantum rotatability", *Trans. Amer. Math. Soc.* **362**:9 (2010), 4831–4851. MR 2011e:46105 Zbl 1203.46043

[De Commer et al. 2014] K. De Commer, A. Freslon, and M. Yamashita, "CCAP for universal discrete quantum groups", *Comm. Math. Phys.* **331**:2 (2014), 677–701. MR 3238527 Zbl 06346288

[Diaconis and Freedman 1987] P. Diaconis and D. Freedman, "A dozen de Finetti-style results in search of a theory", *Ann. Inst. H. Poincaré Probab. Statist.* **23**:2, suppl. (1987), 397–423. MR 88f:60072 Zbl 0619.60039

[Freedman 1962] D. A. Freedman, "Invariants under mixing which generalize de Finetti's theorem", *Ann. Math. Statist* **33** (1962), 916–923. MR 27 #6292 Zbl 0201.49501

[Kallenberg 2005] O. Kallenberg, *Probabilistic symmetries and invariance principles*, Springer, New York, 2005. MR 2006i:60002 Zbl 1084.60003

[Nica and Speicher 2006] A. Nica and R. Speicher, *Lectures on the combinatorics of free probability*, London Mathematical Society Lecture Note Series **335**, Cambridge Univ. Press, 2006. MR 2008k:46198 Zbl 1133.60003

[Shlyakhtenko 1997] D. Shlyakhtenko, "Free quasi-free states", *Pacific J. Math.* **177**:2 (1997), 329–368. MR 98b:46086 Zbl 0882.46026

[Timmermann 2008] T. Timmermann, *An invitation to quantum groups and duality: From Hopf algebras to multiplicative unitaries and beyond*, European Mathematical Society (EMS), Zürich, 2008. MR 2009f:46079 Zbl 1162.46001

[Vaes 2005] S. Vaes, "Strictly outer actions of groups and quantum groups", *J. Reine Angew. Math.* **578** (2005), 147–184. MR 2005k:46167 Zbl 1073.46047

[Vaes and Vergnioux 2007] S. Vaes and R. Vergnioux, "The boundary of universal discrete quantum groups, exactness, and factoriality", *Duke Math. J.* **140**:1 (2007), 35–84. MR 2010a:46166 Zbl 1129.46062

[Van Daele and Wang 1996] A. Van Daele and S. Wang, "Universal quantum groups", *Internat. J. Math.* **7**:2 (1996), 255–263. MR 97d:46090 Zbl 0870.17011

[Wang 1993] S. Wang, *General constructions of compact quantum groups*, Ph.D. thesis, University of California, Berkeley, 1993, Available at http://search.proquest.com/docview/304078126.

[Wang 2002] S. Wang, "Structure and isomorphism classification of compact quantum groups $A_u(Q)$ and $B_u(Q)$", *J. Operator Theory* **48**:3, suppl. (2002), 573–583. MR 2004b:46083 Zbl 1029.46089

[Weingarten 1978] D. Weingarten, "Asymptotic behavior of group integrals in the limit of infinite rank", *J. Mathematical Phys.* **19**:5 (1978), 999–1001. MR 57 #11421 Zbl 0388.28013

[Woronowicz 1987] S. L. Woronowicz, "Compact matrix pseudogroups", *Comm. Math. Phys.* **111**:4 (1987), 613–665. MR 88m:46079 Zbl 0627.58034

[Woronowicz 1998] S. L. Woronowicz, "Compact quantum groups", pp. 845–884 in *Symétries quantiques* (Les Houches, 1995), edited by A. Connes et al., North-Holland, Amsterdam, 1998. MR 99m:46164 Zbl 0997.46045

Received June 19, 2015.

Michael Brannan
Department of Mathematics
Texas A&M University
College Station, TX 77843
United States

mbrannan@math.tamu.edu

Kay Kirkpatrick
University of Illinois at Urbana-Champaign
Urbana, IL 61821
United States

kkirkpat@illinois.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 282, No. 1, 2016
dx.doi.org/10.2140/pjm.2016.282.63

VOLUMES OF MONTESINOS LINKS

KATHLEEN FINLINSON AND JESSICA S. PURCELL

We show that the volume of any Montesinos link can be bounded above and below in terms of the combinatorics of its diagram. This was known for Montesinos links with at most two tangles and those with at least five tangles. We complete the result for the remaining cases.

1. Introduction

W. Thurston [1982] proved that if K is a nontorus, nonsatellite knot, then its complement $S^3 \setminus K$ admits a complete hyperbolic metric. This metric is unique up to isometry by the Mostow–Prasad rigidity theorem [Mostow 1968; Prasad 1973]. Therefore, the hyperbolic volume of a knot complement is a knot invariant. This paper studies the hyperbolic volume of Montesinos links.

Montesinos links are built out of rational tangles, whose definition and properties we review below. It is known that any Montesinos link made up of just one or two rational tangles is a 2-bridge link, and therefore it admits an alternating diagram. Volumes of alternating links can be bounded below due to work of Lackenby [2004]. On the other hand, Futer, Kalfagianni, and Purcell found lower volume bounds for Montesinos links with at least three positive tangles [Futer et al. 2013, Theorems 8.6 and 9.1]; by taking the mirror image, this also gives lower bounds on Montesinos links with at least three negative tangles. Together, this gives lower volume bounds on all Montesinos links with five or more tangles. However, until now, volume bounds for Montesinos links with three or four tangles were unknown.

In this paper, we finish the case of Montesinos links with three or four tangles. We give a lower bound on the volume of any such link in terms of properties of a diagram, which can easily be read off the diagram of the Montesinos link. Specifically, we show volume is bounded in terms of the Euler characteristic of a graph obtained from the diagram. This graph is the reduced A- or B-state graph $\mathbb{G}'_\sigma$ (see Definition 3.2). Our main result is the following.

We acknowledge support by the National Science Foundation under grant number DMS-125687. We also thank David Futer and Efstratia Kalfagianni for helpful conversations.
MSC2010: 57M25, 57M27, 57M50.
Keywords: Montesinos link, volume, hyperbolic geometry, essential product disk.

Theorem 1.1. *Let K be a hyperbolic Montesinos link with a reduced, admissible diagram with at least three tangles. Then*

$$\mathrm{vol}(S^3 \setminus K) \geq v_8(\chi_-(\mathbb{G}'_\sigma) - 1).$$

Here $v_8 \approx 3.6638$ is the hyperbolic volume of a regular ideal octahedron, and $\mathbb{G}'_\sigma$ is the reduced state graph of $D(K)$ corresponding to either the all-A or the all-B state, depending on whether the diagram of K admits two or more positive tangles, or two or more negative tangles, respectively.

Every Montesinos link admits a reduced, admissible diagram (see Definitions 2.5 and 2.6, respectively). The notation $\chi_-(\,\cdot\,)$ denotes the negative Euler characteristic, defined to be

$$\chi_-(Y) = \sum \max\{-\chi(Y_i), 0\},$$

where the sum is over the components $Y_1, \ldots, Y_n$ of Y.

While Theorem 1.1 gives explicit diagrammatical bounds on volume, in many cases, we may estimate $\chi_-(\mathbb{G}'_\sigma)$ in terms of the *twist number* $t(K)$ of the diagram, which is even easier to read off of the diagram. The following theorem generalizes [Futer et al. 2013, Theorem 9.12].

Theorem 1.2. *Let K be a Montesinos link that admits a reduced, admissible diagram with at least two positive tangles and at least two negative tangles, and suppose further that K is not the $(2, -2, 2, -2)$ pretzel link. Then K is hyperbolic, and*

$$\tfrac{1}{4} v_8(t(K) - \#K - 8) \leq \mathrm{vol}(S^3 \setminus K) \leq 2 v_8 t(K),$$

where $v_8 \approx 3.6638$ is the hyperbolic volume of a regular ideal octahedron, $t(K)$ is the twist number of the diagram, and $\#K$ is the number of link components of K.

Outline of proof of Theorem 1.1. The proof applies results from [Futer et al. 2013], but we will restate them for self-containedness. In that paper, using the guts machinery of Agol, Storm, and Thurston [Agol et al. 2007], it is shown that volumes of many links, including hyperbolic Montesinos links with at least two positive or two negative tangles, can be bounded below by identifying *complex essential product disks* (EPDs) in the link complement (see Definition 3.18). In particular, Theorem 9.3 of [Futer et al. 2013] states that for diagrams of links satisfying particular hypotheses, which include the Montesinos links of this paper, we have the estimate

$$\mathrm{vol}(S^3 \setminus K) \geq v_8(\chi_-(\mathbb{G}'_A) - \|E_c\|), \tag{1}$$

where $\|E_c\|$ is the number of complex essential product disks.

In this paper, we show that for a Montesinos link with three or four tangles, the existence of a complex EPD leads to restrictions on the diagram. These restrictions, in turn, imply that at most one complex EPD may exist. This implies Theorem 1.1.

Organization. In Section 2, we review the definitions of rational tangles and Montesinos links. We will need to work with particular diagrams of these links, and we prove such diagrams exist and are prime. In Section 3, we recall the definition of A-adequacy and techniques from [Futer et al. 2013] that can be applied to A-adequate links to give a polyhedron whose combinatorial description is determined by the diagram. We review these results and apply them to the Montesinos links of interest. Section 4 contains the main technical results in the paper. Given a polyhedron for a Montesinos link, we search for complex EPDs that lie in the polyhedron. These are found by analyzing the combinatorics of the diagram and working through several cases. Finally, in Section 5, we put the results together to give the proofs of Theorems 1.1 and 1.2.

2. Tangles and Montesinos links

In this section, we recall the definitions of rational tangles and Montesinos links, and various properties of their diagrams that we will use in the sequel. Throughout, if K is a link in S^3, then $D(K) = D$ is the corresponding link diagram in the plane of projection, and we will assume that D is connected.

Rational tangles. A *rational tangle* is obtained by drawing two arcs of rational slope on the surface of a pillowcase, and then pushing the interiors into the 3-ball bounded by the pillowcase. Rational tangles have been studied in many contexts; see, for example, [Murasugi 1996]. We record here some well-known facts.

A rational number can be described by a continued fraction:

$$\frac{p}{q} = [a_n, a_{n-1}, \dots, a_1] = a_n + \cfrac{1}{a_{n-1} + \cfrac{1}{\ddots + \cfrac{1}{a_1}}}.$$

A continued fraction $[a_n, \dots, a_1]$ defines a rational tangle as follows. Label the four points on the pillowcase NW, NE, SW, and SE. If n is even, connect these points by attaching two arcs c_1 and c_2 connecting NE to SE and NW to SW as in Figure 1(a). Perform a homeomorphism of B^3 that rotates the points NW and NE $|a_1|$ times, twisting the two arcs to create a vertical band of crossings. The crossings will be positive or negative depending on the direction of twist, which is determined by the sign of a_1. In Figure 1(b), three positive crossings have been added. After twisting, relabel the points NW, NE, SW, and SE in their original orientation. Now perform a homeomorphism of B^3 to rotate NE and SE $|a_2|$ times,

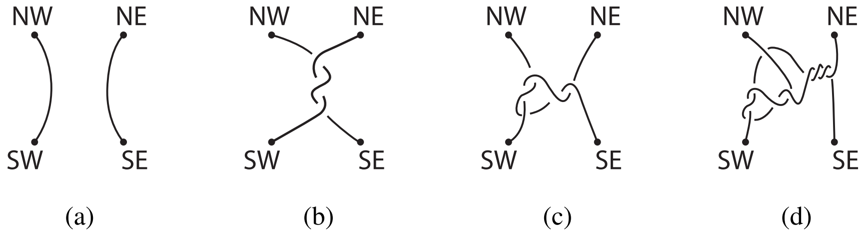


Figure 1. Building a rational tangle from the continued fraction $[4, -1, -2, 3]$.

adding positive or negative crossings in a horizontal band with sign corresponding to a_2. Repeating this process for each a_i, we obtain a rational tangle.

If n is odd, start by using two arcs to connect NW to NE and SW to SE. In this case we add a horizontal band of crossings first, and then continue as before, alternating between horizontal and vertical bands for each a_i.

Any rational tangle may be built by this process. As a convention, we require that a_n always corresponds to a horizontal band of crossings. Thus if we build a rational tangle ending with a vertical band, as in Figure 1(b), we insert a 0 into the corresponding continued fraction, representing a horizontal band of 0 crossings. For example, the continued fraction corresponding to the tangle in Figure 1(b) is $[0, 3]$. This convention ensures that any continued fraction completely specifies a single rational tangle. The tangle shown in Figure 1(a) has continued fraction expansion $\infty = [0, 0] = 0 + \frac{1}{0}$.

Proposition 2.1 [Conway 1970]. *Equivalence classes of rational tangles are in one-to-one correspondence with the set $\mathbb{Q} \cup \infty$. In particular, tangles $T(a_n, \ldots, a_1)$ and $T(b_m, \ldots, b_1)$ are equivalent if and only if the continued fractions $[a_n, \ldots, a_1]$ and $[b_m, \ldots, b_1]$ are equal.*

Using Proposition 2.1, we can put all our tangles into nice form. In particular, if a rational tangle corresponds to a positive rational number, then we can ensure its continued fraction expansion consists only of nonnegative integers. Similarly, if the tangle corresponds to a negative rational number, we can ensure the continued fraction expansion consists of nonpositive numbers. Thus in this paper, positive tangles have only positive crossings, and negative tangles have only negative crossings. This proves that we may divide all nontrivial rational tangles into two groups: *positive tangles* and *negative tangles*. In either case, the tangle has an alternating diagram.

In addition, we may require for a continued fraction with n integers that $a_i \neq 0$ for all $i < n$.

In the description of building rational tangles, we added vertical bands of crossings by rotating the points NW and NE, inserting the vertical band on the north of the tangle. Notice that we could have rotated SW and SE instead, adding a vertical

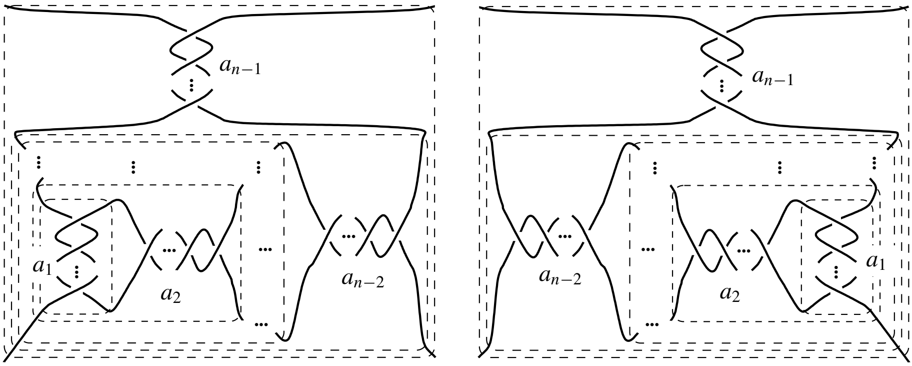


Figure 2. Shown are general forms of admissible tangles, positive on left, negative on right, for n even and $a_n = 0$. (For n odd, the band of a_1 crossings will be horizontal, a_2 vertical, etc. The band of a_{n-1} crossings will be vertical in all cases.)

band of crossings on the south of the tangle. These two methods are equivalent by a sequence of flypes. Likewise, we may add each horizontal band of crossings either on the west side of the tangle (by rotating the points NW and SW), or on the east side of the tangle (by rotating the points NE and SE). The following definition ensures a consistent choice.

Definition 2.2. (a) If T is a positive tangle, then an alternating diagram for T is *admissible* if all the vertical bands of crossings were added by rotating the points NW and NE, and all the horizontal bands of crossings were added by rotating the points NE and SE. See Figure 2, left.

(b) If T is a negative tangle, then an alternating diagram for T is *admissible* if all the vertical bands of crossings were added by rotating the points NW and NE, and all the horizontal bands of crossings were added by rotating the points NW and SW. See Figure 2, right.

By a sequence of flypes, any nontrivial tangle has an admissible diagram.

Montesinos links. Recall that the *numerator closure* $\mathrm{num}(T)$ of a tangle T is formed by connecting NW to NE and SW to SE by simple arcs with no crossings. The *denominator closure* $\mathrm{denom}(T)$ is formed by connecting NW to SW and NE to SE by simple arcs with no crossings.

Given two rational tangles T_1 and T_2 with slopes q_1 and q_2, we form their *sum* by connecting the NE and SE corners of T_1 to the NW and SW corners of T_2, respectively, with two disjoint arcs. If q_1 or q_2 is an integer, then the sum $T_1 + T_2$ is also a rational tangle; this is called a *trivial sum*.

The *cyclic sum* of $T_1, \ldots, T_r$ is the numerator closure of the sum $T_1 + \cdots + T_r$.

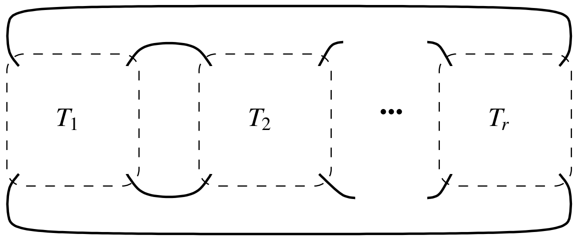


Figure 3. A Montesinos link with r tangles.

Definition 2.3. A *Montesinos link* is the cyclic sum of a finite ordered list of rational tangles $T_1, \ldots, T_r$ with corresponding slopes in $\mathbb{Q}$. See Figure 3.

A Montesinos link is determined by the integer r and an r-tuple of slopes $q_1, \ldots, q_r$ with $q_i \in \mathbb{Q}$. Note that if q_i is an integer, then it consists of a single band of horizontal crossings, which can be subsumed into an adjacent rational tangle in a sum. Thus we assume that $q_i \notin \mathbb{Z}$ to avoid trivial sums.

Theorem 2.4 [Bonahon and Siebenmann 2010, Theorem 12.8]. *Let K be a Montesinos link obtained as the cyclic sum of $r \geq 3$ rational tangles whose slopes are $q_1, \ldots, q_r \in \mathbb{Q} \setminus \mathbb{Z}$. Then K is determined up to isomorphism by the rational number $\sum_{i=1}^{r} q_i$ and the vector $((q_1 \bmod 1), \ldots, (q_r \bmod 1))$, up to dihedral permutation.*

Note that this theorem gives isomorphism up to dihedral permutation; however, we will only use isomorphism up to cyclic permutation. By Theorem 2.4, given K as the cyclic sum of $T_1, \ldots, T_r$, we can "combine" the integer parts of $q_1, \ldots, q_r$. The following definition makes use of this fact.

Definition 2.5. A diagram $D(K)$ is called a *reduced Montesinos diagram* if it is the cyclic sum of the diagrams T_i, and for each i, the diagram of T_i has either all positive or all negative crossings, and either

(1) all the slopes q_i of tangles T_i have the same sign, or

(2) $0 < |q_i| < 1$ for all i.

It is not hard to see that every Montesinos link with $r \geq 3$ has a reduced diagram. For example, if $q_i < 0$ while $q_j > 1$, one may add 1 to q_i and subtract 1 from q_j. By Theorem 2.4, this does not change the link type. One may continue in this manner until condition (1) of Definition 2.5 is satisfied.

We make one more definition.

Definition 2.6. A diagram $D(K)$ of the cyclic sum of $T_1, \ldots, T_r$ is an *admissible Montesinos diagram* if the diagram of T_i is an admissible tangle diagram for each i.

Since every tangle has an admissible diagram, every Montesinos link with $r \geq 3$ has a reduced, admissible diagram.

We will need to know that an admissible diagram of a Montesinos link is prime. Recall that a diagram is *prime* if for any simple closed curve γ meeting the diagram graph transversely in exactly two edges, the curve γ bounds a region of the projection plane with no crossings.

Proposition 2.7. *A reduced, admissible diagram of a Montesinos link with at least two (nontrivial) tangles is prime.*

We will prove Proposition 2.7 using two lemmas.

Lemma 2.8. *If T is a reduced, admissible diagram of a rational tangle with at least two crossings, then either T is a single vertical band of crossings and* $\mathrm{denom}(T)$ *is prime, or* $\mathrm{num}(T)$ *is prime.*

Proof. If T is a rational tangle with only a single vertical band of crossings, then the denominator closure of T is a $(2, q)$-torus link, with $q > 1$ by assumption on the number of crossings. Otherwise, the numerator closure of T will be an alternating diagram of a 2-bridge link. By [Menasco 1984], in either case the diagram will be prime. □

Lemma 2.9. *Suppose T_1 is a diagram of a connected nontrivial tangle such that either* $\mathrm{num}(T_1)$ *or* $\mathrm{denom}(T_1)$ *is prime, and suppose that T_2 is a reduced, admissible diagram of a rational tangle with at least one crossing. Then* $\mathrm{num}(T_1 + T_2)$ *is prime.*

Proof. Let D be the diagram of $\mathrm{num}(T_1 + T_2)$ and suppose that γ is a simple closed curve meeting D exactly twice.

Case 1. The curve γ meets D outside of both tangles. Then since the diagram has no crossings outside the two tangles, either γ bounds a portion of the diagram with no crossings on one side, or γ encloses T_1 on one side, T_2 on the other. But in the latter case, γ would have to meet D four times, contradicting the fact that it meets D exactly twice.

Case 2. The curve γ meets D twice in the tangle T_1. Then we may isotope γ to be contained entirely in T_1, that is, within the Conway sphere enclosing T_1. Then γ can be drawn into the numerator and denominator closures of T_1. One of these is prime; without loss of generality say $\mathrm{num}(T_1)$ is prime (otherwise rotate the diagram for the following argument). The curve γ must contain no crossings on one side. If it contains no crossings in its interior, then there are no crossings in the interior of γ in D, and we are done. So suppose γ contains no crossings on its exterior in $\mathrm{num}(T_1)$. Then the tangle must contain all its crossings on the interior of γ. Moreover, exactly two strands of the tangle run to the exterior of γ, and four strands must connect from the NW, NE, SE, and SW corners. See Figure 4. This is impossible for a connected tangle.

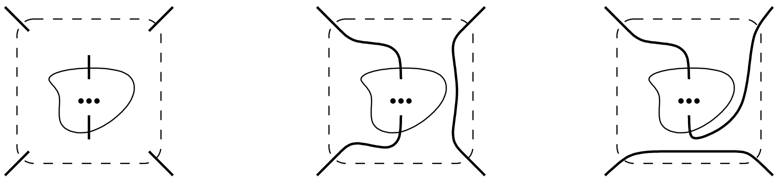


Figure 4. If γ has no crossings to the exterior of $\mathrm{num}(T_1)$, then T_1 cannot have a connected diagram.

Case 3. The curve γ meets D twice in the tangle T_2. If T_2 has prime numerator or denominator closure, then the same argument as in Case 2 applies, to guarantee that γ bounds no crossings on one side. By Lemma 2.8, the only remaining case is that T_2 consists of a single crossing. But then if γ meets D exactly twice in a tangle consisting of a single crossing, it cannot encircle that crossing, but must bound a region of the diagram with no crossings to the interior.

Case 4. The curve γ intersects T_1 exactly once. The tangle T_1 is bounded by a square meeting the diagram in four points, NW, NE, SE, and SW.

If γ exits T_1 by running through adjacent sides of the square, then we may form a new simple closed curve γ' meeting the diagram exactly twice by taking γ inside the square, and taking portions of the two sides of the square meeting in one of the corners (NW, NE, SE, or SW). This new curve γ' can be drawn into the diagrams of $\mathrm{num}(T_1)$ and $\mathrm{denom}(T_1)$. Since once of these is prime, without loss of generality $\mathrm{num}(T_1)$, γ' bounds no crossings on one side in that diagram. If the interior, then to the interior γ' bounds a single strand of the diagram, and we may slide γ' and γ along this strand to remove the intersection of γ with T_1. If γ' bounds no crossings to the exterior, then since there are three knot strands to the exterior, emanating from three of NW, NE, SE, SW, the tangle diagram is not connected. This is a contradiction.

If γ exits T_1 by running through the north and south sides of the square, then consider the portion of γ inside T_1, and form $\mathrm{denom}(T_1)$. We may connect north to south in $\mathrm{denom}(T_1)$ by an arc that does not meet the diagram of $\mathrm{denom}(T_1)$. Connecting this to γ, we obtain a closed curve meeting the diagram of $\mathrm{denom}(T_1)$ exactly once. This is impossible. Note this argument did not need $\mathrm{denom}(T_1)$ to be prime. Symmetrically, if γ exits T_1 by running through the east and west, then we may form a closed curve meeting $\mathrm{num}(T_1)$ exactly once, which is impossible. So we may assume Case 4 does not happen.

Case 5. The curve γ intersects T_2 exactly once. Again if T_2 contains more than one crossing, Lemma 2.8 and the argument of Case 4 will imply we can isotope γ outside of T_2. If T_2 contains exactly one crossing, and γ meets T_2 exactly once, then it must meet T_2 in one of the strands running from NE, NW, SW, SE to the center, and we may isotope it from that point of intersection to the corner without meeting any crossings. Thus we may assume, after isotopy, that Case 5 does not happen.

Thus in all cases, γ bounds no crossings on one side. □

Proof of Proposition 2.7. The proof is by induction on the number of tangles in the Montesinos link. If there are two tangles, then either the result holds by Lemma 2.9, or both tangles consist of a single crossing. In that case, their sum is a horizontal band of two crossings; hence the Montesinos link is a standard diagram of a $(2, 2)$-torus link, which is prime.

Now suppose that any reduced, admissible diagram of a Montesinos link with k tangles is prime, and consider a Montesinos link with $k+1$ tangles. The first k tangles to the right have a sum satisfying the hypotheses on T_1 in Lemma 2.9, and the $(k+1)$-st tangle satisfies the hypothesis on T_2. So by that lemma, the diagram of the Montesinos link is prime. □

Montesinos links of interest. Futer, Kalfagianni, and Purcell found a volume estimate for Montesinos links with at least three positive or three negative tangles [Futer et al. 2013]. A Montesinos link with only one or two tangles has an alternating diagram; its volume is bounded by [Lackenby 2004]. Thus the only types of Montesinos links whose volumes cannot be estimated by previous results are

(a) Montesinos links with two positive and one negative tangles,

(b) Montesinos links with one positive and two negative tangles,

(c) Montesinos links with two positive and two negative tangles.

Notice that the mirror image of a type (b) link is a type (a) link, and taking the mirror will not change the volume of the link complement. Thus we will ignore type (b) links in favor of type (a) links in our analysis. Notice also that there is only one "arrangement" of a type (a) link, up to cyclic permutation. However, there are two arrangements of type (c) links.

Definition 2.10. A $++-$ *link* is a Montesinos link which is the cyclic sum of T_a, T_b, T_c, where T_a and T_b are positive tangles and T_c is a negative tangle. A $+-+-$ *link* is a Montesinos link which is the numerator closure of $T_a+T_b+T_c+T_d$, where T_a and T_c are positive tangles and T_b and T_d are negative tangles. A $++--$ *link* is a Montesinos link which is the numerator closure of $T_a+T_b+T_c+T_d$, where T_a and T_b are positive tangles and T_c and T_d are negative tangles.

Our goal is to find volume bounds for these types of Montesinos links. We take Definition 2.10 as the definition not only of $++-$ and $+-+-$ links, but also of the tangles T_a, T_b, T_c, and T_d. Notice that the definitions of T_a, T_b, and T_c change depending on whether we are talking about $++-$, $+-+-$, or $++--$ links.

Remark 2.11. In fact, we can actually consider only $++--$ links or only $+-+-$ links, as follows. Since we are assuming the link is reduced, each of our tangles has slope with absolute value at most 1, as in Definition 2.5(2). Thus in a $++--$ link, we may subtract 1 from the second slope and add 1 to the third. By Theorem 2.4, the result will be equivalent to a $+-+-$ link. So we only consider $+-+-$ links.

Notice that reduced diagrams for $++-$ and $+-+-$ links do not satisfy part (1) of Definition 2.5; therefore they must satisfy (2). This means that there are no integer parts of slopes of each tangle. In other words, if T_a has slope $[a_n, \dots, a_1]$, then we may assume that $a_n = 0$. Recall from page 66 that a_n corresponds to a horizontal band of crossings; so a_{n-1} corresponds to a vertical band of crossings, as in Figure 2. Recall also that $a_i \neq 0$ for $i \neq n$. Thus the tangles T_a, T_b, T_c, and T_d have the form of Figure 2, except possibly n even replaced with n odd, meaning the vertical band of a_1 crossings will be horizontal.

3. Estimating the guts

Our volume bounds use estimates developed by Futer, Kalfagianni, and Purcell to bound volumes of semiadequate links [Futer et al. 2013]. We will see that the $++-$ and $+-+-$ Montesinos links of interest are semiadequate, and so they fit into this machinery. In this section, we recall the definition of semiadequate links and review the relevant features of [Futer et al. 2013]. All the necessary details from that paper are contained here. However, one may consult that paper or the survey article [Futer et al. 2014] for additional information and for the proofs of the results that we cite.

Semiadequate links. Given a link diagram D and a crossing x of D, we define a new diagram by replacing the crossing x with a crossing-free *resolution*. There are two ways to resolve a crossing, shown in Figure 5: the A-resolution or the B-resolution.

Definition 3.1. A *state* σ is a choice of A- or B-resolution at each crossing of a diagram D. Applying a state σ to a diagram D yields a collection of crossing-free simple closed curves called *state circles*. If we attach an edge to the state circles at each removed crossing, i.e., attach the dashed edge shown in Figure 5, we obtain a trivalent graph H_σ. The edges coming from crossings, dashed in the figure, are called *segments*.

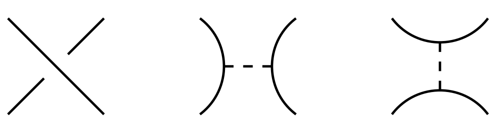
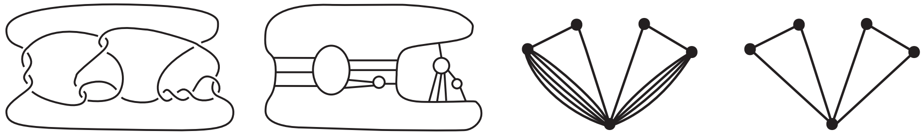


Figure 5. Left to right: a crossing, its A-resolution, and its B-resolution.



(a) (b) (c) (d)

Figure 6. (a) A $++-$ Montesinos link, (b) its graph H_A, (c) the state graph $\mathbb{G}_A$, (d) the reduced state graph $\mathbb{G}'_A$.

We are concerned mainly with the *all-A state*, which chooses the A-resolution at each crossing. We will occasionally mention the all-B resolution as well. The graph H_A is obtained by applying the all-A state to D and including segments. For an example, see Figure 6.

Definition 3.2. From H_A we create the *A-state graph* $\mathbb{G}_A$ by shrinking each state circle to a single vertex. We obtain the *reduced A-state graph* $\mathbb{G}'_A$ by removing multiple edges between pairs of vertices in $\mathbb{G}_A$. An example is shown Figure 6.

The following two lemmas concern H_A for admissible diagrams of Montesinos links, and they follow immediately from the structure of admissible tangles, as in Figure 2. Both lemmas are illustrated in Figure 7.

Lemma 3.3. *Let T be a positive admissible tangle with corresponding continued fraction $[0, a_{n-1}, \dots, a_2, a_1]$. Then for any Montesinos knot containing T, the graph H_A will have the following properties in a neighborhood of T.*

(1) *A portion of a state circle, call it S_0, runs from NW to SW, and a portion of another, call it S_{n-1}, runs from NE to SE.*

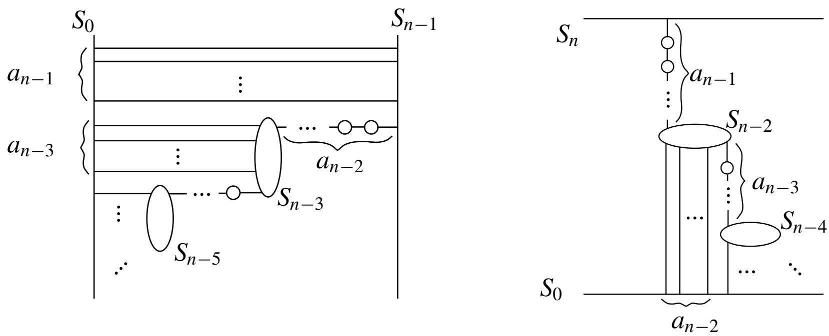
(2) *There are $a_{n-1} > 0$ horizontal segments running from S_0 to S_{n-1} at the north of the graph.*

(3) *For each a_i with $i \equiv n \pmod 2$, there exists a horizontal string of a_i state circles alternating with a_i segments, with the segment on the far east having one endpoint on S_{i+1}, south of any other segments, and with the final state circle on the far west denoted by S_{i-1}.*

(4) *For each a_i with $i \equiv (n-1) \pmod 2$, there are a_i horizontal segments connecting S_0 and S_i.* □

Lemma 3.4. *Let T be a negative admissible tangle with corresponding continued fraction $[0, a_{n-1}, \dots, a_2, a_1]$. Then for any Montesinos knot containing T, the graph H_A will have the following properties in a neighborhood of T.*

(1) *A portion of a state circle, call it S_n, runs from NW to NE, and a portion of another, call it S_0, runs from SW to SE.*



(a) positive (b) negative

Figure 7. General form of H_A in a neighborhood of an admissible positive tangle and an admissible negative tangle.

(2) *For each a_i with $i \equiv (n-1) \pmod 2$, there exists a vertical string of a_i state circles alternating with a_i segments, with the segment at the far north having an endpoint on S_{i+1}, to the east of all other segments on S_{i+1}. The final state circle at the far south is denoted by S_{i-1}.*

(3) *For each a_i with $i \equiv n \pmod 2$, there exist a_i vertical segments connecting S_0 and S_i.* □

The following definition is due to Lickorish and Thistlethwaite.

Definition 3.5. A link diagram $D(K)$ is called *A-adequate* if $\mathbb{G}_A$ has no 1-edge loops, and *B-adequate* if $\mathbb{G}_B$ has no 1-edge loops.

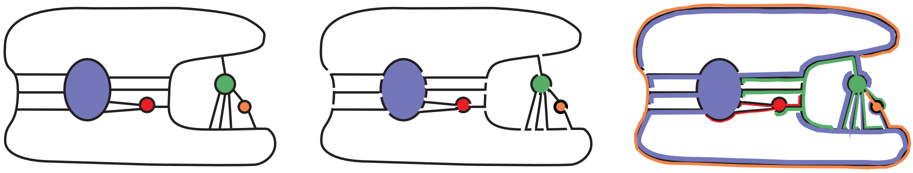
Theorem 3.6 [Lickorish and Thistlethwaite 1988]. *Let $D(K)$ be a reduced Montesinos diagram with $r > 0$ positive tangles and $s > 0$ negative tangles. Then $D(K)$ is A-adequate if and only if $r \geq 2$ and B-adequate if and only if $s \geq 2$. Since $r + s \geq 3$ in a reduced diagram, D must be either A- or B-adequate.*

Also note that if $r = 0$ or $s = 0$ then $D(K)$ is alternating, in which case it is both A- and B-adequate. We will see shortly why adequacy is a desirable property.

From H_A we may obtain a surface as follows. The state circles of H_A bound disjoint disks in the 3-ball below the projection plane. To these disks, attach a half-twisted band corresponding to each crossing in the original diagram. This forms a connected surface called the *A-state surface*, or simply S_A.

Theorem 3.7 [Ozawa 2011]. *Let D be a (connected) diagram of a link K. Then the surface S_A is essential in $S^3 \setminus K$ if and only if D is A-adequate.*

Define the manifold with boundary $M_A = S^3 \backslash\backslash S_A$, where $S^3 \backslash\backslash S_A$ is defined to be S^3 with a regular neighborhood of S_A removed.



(a) (b) (c)

Figure 8. Building the upper polyhedron: (a) shade innermost disks, (b) remove portions of state circle, (c) construct tentacles.

Definition 3.8. Let $M = S^3 \setminus N(K)$, and define the *parabolic locus* of M_A to be $P = \partial M_A \cap \partial M$. The parabolic locus consists of annuli. These annuli consist of the remnants of the knot diagram.

In [Futer et al. 2013], it was shown that M_A can be cut into ideal polyhedra. We will not describe the details of this cutting here, because we will not need those details; we will concern ourselves only with the results. The cutting produces finitely many polyhedra that lie below the projection plane, and a single polyhedron above, which we call the *upper polyhedron*. In this paper, we only need to study the upper polyhedron. It has a nice combinatorial description coming from the graph H_A, which we now recall.

To visualize the upper polyhedron, start with the state graph H_A. Recall that H_A lies in the projection plane and is composed of state circles and segments. We call a given state circle S *innermost* if S bounds a region in the projection plane which does not contain any segments of H_A. We shade each innermost disk, giving each a different color. These will correspond to the distinct *shaded faces* of the upper polyhedron. See Figure 8(a) for an example.

The faces extend from the innermost state circles as follows. Given a segment s of H_A, rotate H_A so that s is vertical. There are two distinct ways to perform this rotation; the procedure that follows is independent of that choice. Once s is vertical, erase a small part of the graph immediately northeast of s and a small part immediately southwest of s. Repeat this rotation and erasing for each segment in the graph. See, for example, Figure 8(b).

Finally, draw the "tentacles": Choose a segment s that meets one of the innermost state circles. The innermost state circle bounds a shaded face. Rotate H_A so that s is vertical with the shaded face on the top. The small hole to the northeast of s acts as a gate, allowing the shaded face to run through the hole, forming a *tentacle*. The tentacle runs in a thin band along H_A, adjacent to a segment and a state circle, running south and then east. It terminates when it runs into a segment. However, the tentacle may run *past* other segments on the opposite side of the state circle without terminating. When this occurs, the tentacle spawns a new tentacle, running through the hole in H_A adjacent to that segment. Each new tentacle also terminates when it hits a segment, and also spawns other tentacles when it runs past a segment. Continue until each tentacle has terminated. Now one shaded face is complete. Repeat this process for each innermost disk. Figure 8(c) shows a completed example of the upper polyhedron.

Definition 3.9. For a given shaded face, or a given tentacle of a shaded face, we will say the face or tentacle *originated* in the innermost state circle of the same color. The place where a tentacle terminates by running into a segment is called the *tail* of the tentacle. The place where a tentacle runs adjacent to a segment is called the *head* of the tentacle.

The upper polyhedron has faces that include the shaded faces, as well as white faces corresponding to unshaded regions of the diagram. Edges run from head to tail of tentacles, and separate white and shaded faces. Vertices are the remnants of H_A. In [Futer et al. 2013], it was proved that this process produces an ideal polyhedron, with ideal vertices on the parabolic locus.

Lemma 3.10 (Futer, Kalfagianni, and Purcell [Futer et al. 2013, Theorem 3.13]). *Let $D(K)$ be an A-adequate link diagram. Then the polyhedron P_A as described above is a checkerboard colored ideal polyhedron with* 4*-valent vertices.*

Here "checkerboard" means that white faces never share an edge, and neither do colored faces. After erasing the small holes as in Figure 8(b), each connected component of the remnants of H_A is one ideal vertex of P_A. Each such component consists of segments and portions of state circle, with segments meeting two distinct shaded faces on opposite sides, and portions of state circle meeting white faces at the endpoints of the ideal vertices. This is why the vertices are 4-valent.

The careful reader may notice we have not discussed the *nonprime arcs* in the polyhedral decomposition in [Futer et al. 2013]. This is because in Montesinos links, nonprime arcs only occur between adjacent negative tangles; see [Futer et al. 2013, Lemma 8.7]. We focus on $++-$ and $+-+-$ links, which have no adjacent negative tangles. Thus these diagrams have no nonprime arcs.

We are now ready to discuss the upper polyhedron of the polyhedral decomposition of $++-$ and $+-+-$ Montesinos links.

Lemma 3.11. *Let K be a $++-$ Montesinos link. Then K has a reduced, admissible diagram D with upper polyhedron P_A having the following properties.*

(1) *There is an innermost disk, denoted G, between the two positive tangles T_a and T_b.*

(2) *There are $a_{n-1} > 0$ segments running across T_a from west to east, with east endpoints on G. Similarly, there are $b_{n-1} > 0$ segments running across T_b from west to east, with their west endpoints on G.*

(3) *If T_a contains no state circles in the interior, then all white faces in T_a are bigons. Otherwise, there are a_{n-1} bigons at the north of T_a, and a nonbigon just below. The same holds for T_b.*

(4) *There is exactly one segment at the north of T_c. If T_c has only one state circle, then all white faces below that state circle in T_c are bigons.*

Proof. Immediate from Lemmas 3.3 and 3.4 and the definition of P_A. □

Lemma 3.12. *Let K be a $+-+-$ Montesinos link. Then K has a reduced, admissible diagram D with upper polyhedron P_A having the following properties.*

(1) *The tangle T_b is contained in a state circle, denoted G, lying between the two positive tangles.*

(2) *There are $a_{n-1} > 0$ segments running across T_a from west to east, with east endpoints on G. Similarly, there are c_{n-1} segments running across T_c from west to east, with west endpoints on G.*

(3) *If T_a has no state circles in the interior, then all white faces are bigons. Otherwise, there are a_{n-1} bigons at the north of T_a, and a nonbigon just below. Similarly for T_c.*

(4) *There is exactly one segment at the north of T_b. If T_b has only one state circle, then all white faces below that state circle in T_b are bigons. Similarly for T_d.*

Proof. Again immediate from Lemmas 3.3 and 3.4. □

Remark 3.13. In many of the figures below, for simplicity, we omit the step above of erasing a small portion of the graph near each segment, and just draw the tentacles without accurately portraying the ideal vertices of the diagram.

Essential product disks. By [Futer et al. 2013], the problem of bounding volumes of A-adequate links can be reduced to the problem of finding essential product disks in the upper polyhedron.

Definition 3.14. An *essential product disk* (*EPD*) *in the upper polyhedron* is a properly embedded essential disk in P_A whose boundary consists of two arcs in two shaded faces, and two points where the boundary meets the parabolic locus.

We consider the boundary of an EPD, which we will draw into the diagram of P_A. Figure 9 depicts a portion of the diagram of P_A with the boundary of an EPD. Given an EPD E in P_A, we generally pull ∂E into a *normal square*. A normal square is a disk D in normal form, with ∂D intersecting P_A in exactly four arcs. We will only need the special kinds of normal squares described in the next lemma.

Lemma 3.15 [Futer et al. 2013, Lemma 6.1]. *Let D be a diagram of a Montesinos link with an EPD, E, embedded in the upper polyhedron. Then ∂E can be pulled off the parabolic locus to give a normal square in P_A such that*

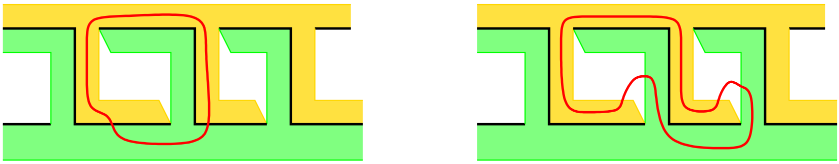


Figure 9. A portion of the upper polyhedron containing an EPD, showing the boundary before (left) and after (right) pulling into a normal square.

(1) *two opposite sides of the square run through shaded faces, which we color green and gold, with endpoints of each side on distinct edges of* P_A;

(2) *the other two edges of the square run through white faces, and cut off a single vertex of the white face (these are the white sides of the normal square)*;

(3) *the single vertex of the white face that is cut off by a white side of* ∂E *forms a triangle, so that when moving clockwise the edges of the triangle are colored gold, white, and green.*

Given an EPD E in P_A, two edges of the square run through shaded faces; we may choose how we color these faces green and gold. Then by property (3) of Lemma 3.15, we may force the triangles formed by the white faces to be oriented accordingly. Notice that the orientation given in property (3) will force the white edges to occur at tails of gold tentacles and heads of green. Figure 9, right, depicts a normal square in P_A oriented as in the above lemma.

Complex essential product disks. Recall that we are attempting to bound volumes of Montesinos links. A bound on volume is given by Equation (1), originally from [Futer et al. 2013, Theorem 9.3], which applies to reduced, admissible diagrams of hyperbolic Montesinos links. The term $\|E_c\|$ appears in this equation; we define it in this section. We also determine further information about EPDs in P_A and their relationships to each other.

Definition 3.16. Let S be a surface in P_A. A *parabolic compression disk* for S is an embedded disk E in P_A such that

(i) $E \cap S$ is a single arc in ∂E;

(ii) the rest of ∂E is an arc in ∂P_A that has endpoints disjoint from the parabolic locus P and that intersects P in exactly one transverse arc;

(iii) $E \cap S$ is not parallel in S to an arc in ∂S that contains at most one component of $S \cap P$.

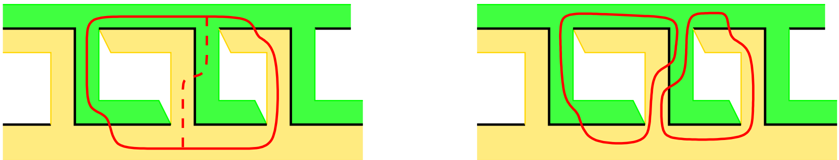


Figure 10. A parabolic compression: arc of parabolic compression disk (left) and two new EPDs (right).

Figure 10, left, shows a portion of a diagram of P_A. The red line represents the boundary of an EPD D. There is a parabolic compression disk E for D. The dashed red line shows the arc as in part (ii) of the above definition.

Definition 3.17. If D is an EPD in P_A with a parabolic compression disk E for D, then surgering D along E will produce a pair of new EPDs, D' and D''. We say that D and $D' \cup D''$ are *equivalent under parabolic compression*.

Figure 10, right, shows the two new EPDs D' and D'' obtained from the disk D in the left part of the figure. These new disks are equivalent to D under parabolic compression.

Definition 3.18. An essential product disk D in P_A is called

(1) *simple* if D is the boundary of a regular neighborhood of a white bigon face of P_A,

(2) *semisimple* if D is equivalent under parabolic compression to a union of simple disks (but D is not simple),

(3) *complex* if D is neither simple nor semisimple.

We see that the EPDs in the right half of Figure 10 are simple and the EPD in the left half is semisimple. The following lemma identifies semisimple EPDs more generally, using a diagrammatic condition.

Lemma 3.19. *Let D be an essential product disk in P_A that is not simple. Then D is semisimple if its boundary bounds a region in the projection plane that contains only bigon white faces of P_A.*

Proof. Suppose D is an EPD in P_A whose boundary contains only white bigon faces on one side, say the inside. The boundary of D must run through two shaded faces with two colors, say green and gold, and it switches between colors by running over an ideal vertex v. Because there are only bigon white faces to the inside, the white face adjacent to v must be a bigon. The bigon face must have edges also meeting green and gold shaded faces. Thus there is an arc with endpoints on ∂D running through the ideal vertex of the bigon opposite v. See Figure 11.

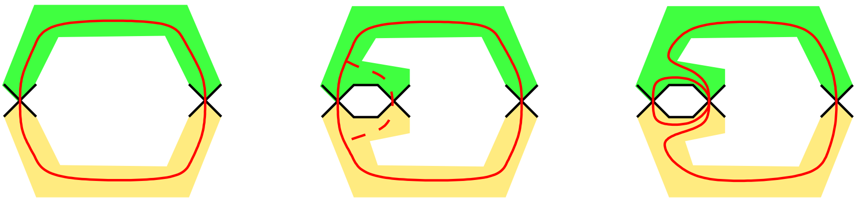


Figure 11. An EPD bounding bigons to the interior, left, must have ideal vertex adjacent to a bigon, middle, which allows us to parabolically compress, right.

This arc defines a parabolic compression disk for D. Thus D is equivalent under parabolic compression to two EPDs, one encircling the bigon, and one, D', whose boundary agrees with D except excludes this bigon. Now D' is an EPD bounding one fewer bigon in its interior. We may repeat the above process, compressing off another bigon. There are finitely many bigons in this region, so we continue until there is only one left. Then D is equivalent under parabolic compression to the union of simple EPDs. □

Lemma 3.20 [Futer et al. 2013, Lemma 5.8]. *There exists a set $E_s \cup E_c$ of essential product disks in P_A such that*

(1) *E_s is the set of all simple disks in P_A;*

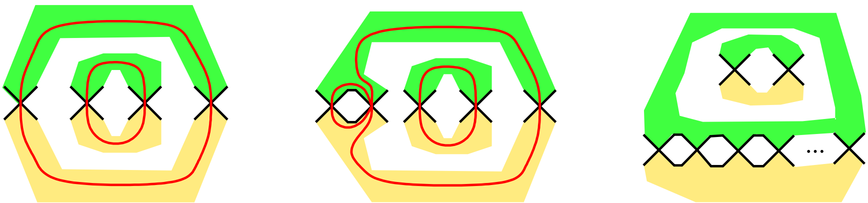
(2) *E_c consists of complex disks, and further, E_c is minimal in the sense that no disk in E_c is equivalent under parabolic compression to a subcollection of $E_s \cup E_c$;*

(3) *E_c is also maximal in the sense that if any complex disks are added to E_c, then E_c is no longer minimal.*

We take Lemma 3.20 as a definition of the minimal set E_c. The number of EPDs in E_c is denoted $\|E_c\|$. This is the term we wish to bound in Equation (1). It is not easy to tell directly from Definition 3.16, Definition 3.17, and Lemma 3.20 whether a set of complex disks E_c is minimal. The following lemma provides an easier way to characterize minimality.

Lemma 3.21. *Let E_s be the set of all simple EPDs in P_A, and let E_1 and E_2 be EPDs in P_A. Then E_2 is equivalent under parabolic compression to a subcollection of $E_s \cup E_1$ if ∂E_1 and ∂E_2 differ only by white bigons, i.e., one of the regions S bounded by ∂E_1 and one of the regions T bounded by ∂E_2 are such that $(S \cup T) \setminus (S \cap T)$ contains no nonbigon white faces.*

Proof. Consider first the case that ∂E_1 and ∂E_2 do not intersect. If S and T are disjoint, then both ∂E_1 and ∂E_2 bound disjoint regions containing only white bigon



(a) (b) (c)

Figure 12. Region between disjoint ∂E_1 and ∂E_2 consists of bigons.

faces, so both E_1 and E_2 are semisimple. Thus E_2 is equivalent under parabolic compression to EPDs in E_s alone; hence it is equivalent to a subcollection of $E_s \cup E_1$.

If ∂E_1 and ∂E_2 do not intersect, but S and T are not disjoint, then $(S \cup T) \setminus (S \cap T)$ is an annular region of the projection plane between ∂E_1 and ∂E_2, and no nonbigon white faces lie in this region. This is shown in Figure 12(a).

Focus on the left-most vertex, on ∂E_1. By Lemma 3.10, P_A is checkerboard colored; hence the face to the right of this vertex is a white face. Therefore, either this face lies inside ∂E_1 and ∂E_2, and the two EPDs both meet this ideal vertex, or the face to the right must be a bigon. If the face is a bigon, we find a parabolic compression disk whose parabolic compression arc is the line running from ∂E_1 through the opposite vertex of the bigon. Therefore, ∂E_1 is parabolically compressible to a union of two EPDs, one of which is simple, as in Figure 12(b). Replace E_1 with the other EPD. Then we are in the same situation as before, only with one fewer bigon between boundaries of the EPDs. Repeat, compressing off bigons until there are no more. We obtain a finite chain of bigons, which must connect with one of the other three vertices in the figure. In Figure 12(c), the chain connects to the other vertex on ∂E_1. However, this creates a contradiction, since a curve following the arc of the boundary of E_1 through the green, then running across the tops of the bigons in the green, gives a simple closed curve that does not bound a disk in the green face. This contradicts the fact that each shaded face of P_A is simply connected. Therefore the chain must connect with one of the vertices on ∂E_2. (Note this implies also that the faces that ∂E_1 and ∂E_2 meet must both be the same color, gold and green, as we have illustrated in the figures.) The same process may be repeated for the other two vertices. Then the faces between ∂E_1 and ∂E_2 consist of strings of adjacent bigons; hence E_1 is equivalent under parabolic compression to a subset of $E_s \cup E_2$.

Now suppose ∂E_1 and ∂E_2 intersect. In this case the region $(S \cup T) \setminus (S \cap T)$ consists of one or more connected components. Take one of these connected components, say H. Then ∂H consists of an arc of ∂E_1 and an arc of ∂E_2, with two intersections of ∂E_1 and ∂E_2. Suppose these intersections both take place in

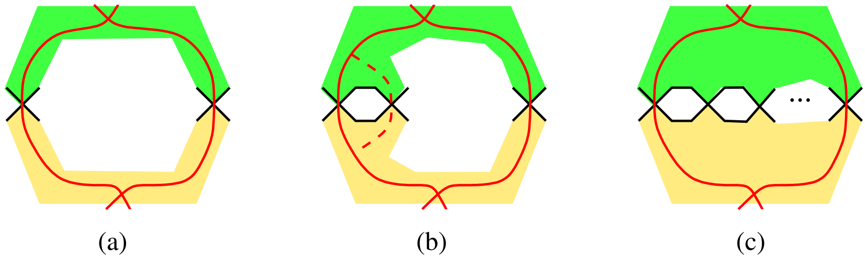


Figure 13. When ∂E_1 and ∂E_2 are not disjoint.

the green face. Since the green face is simply connected, we can push E_1 and E_2 to remove both of these intersections.

Thus we assume that the intersections of ∂E_1 and ∂E_2 occur in two different faces, green and gold as in Figure 13(a). Here the black lines represent edges of P_A. As before, by Lemma 3.10, the boundary of an EPD can only move from one colored face to another at an ideal vertex of P_A. Suppose H is the region in the middle of this figure, so there are no nonbigon white faces in this region. The face just to the right of the vertex at the left of the figure is a white face, which must be a bigon, as in Figure 13(b). Also shown is a compression arc corresponding to a parabolic compression disk for E_1. Therefore E_1 is equivalent under parabolic compression to two new EPDs, one of which is simple. Then we can repeat this process, compressing off simple EPDs in a chain until the chain connects to the other vertex of P_A (Figure 13(c)). Thus in this region, E_1 is equivalent under parabolic compression to a subset of $E_s \cup E_2$. Repeat the above arguments for each connected component of $\mathcal{H}$. We either push the intersections off, or compress bigons off of E_1 until E_1 is equivalent to a subset of $E_s \cup E_2$. □

Using Lemma 3.21 we may more easily find a minimal set E_c and therefore calculate $\|E_c\|$. For certain Montesinos links, $\|E_c\|$ is already known.

Theorem 3.22 (Futer, Kalfagianni, and Purcell [Futer et al. 2013, Proposition 8.16]). *Let $D(K)$ be a reduced, admissible, nonalternating Montesinos diagram with at least three positive tangles. Then $\|E_c\| = 0$.*

Finding complex EPDs. We now want to bound $\|E_c\|$ for classes of Montesinos links that do not fit under the umbrella of Theorem 3.22. We will do so directly by finding a minimal spanning set of complex EPDs in P_A. The following results show how we might begin to look for complex EPDs.

Lemma 3.23. *Let $D(K)$ be a reduced, admissible, A-adequate, nonalternating Montesinos diagram. Let E be a complex EPD in P_A. Then either*

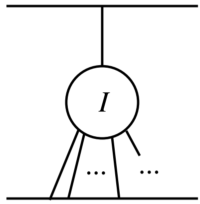


Figure 14. A negative tangle with slope $-1 \leq q \leq -\frac{1}{2}$.

(1) *∂E runs through a negative tangle N_i of slope $-1 \leq q \leq -\frac{1}{2}$, along segments of H_A that connect a single state circle, I, to the north and south sides of the tangle, as in Figure 14, or*

(2) *there are exactly two positive tangles P_1 and P_2, and ∂E runs along segments of H_A that run through P_1 and P_2 from east to west.*

Proof. By [Futer et al. 2013, Corollary 6.6], ∂E runs over tentacles adjacent to segments of a 2-edge loop in $\mathbb{G}'_A$. Further, Lemma 8.14 of that paper gives three possible forms of 2-edge loops in G_A. The first form of loop from that lemma corresponds to crossings in a single twist region, in which the all-A resolution is the short resolution. By [Futer et al. 2013, Lemma 5.17], we may remove all bigons in the short resolution of a twist region without changing $\|E_c\|$. Therefore we may ignore such loops. For a 2-edge loop of the second form, [Futer et al. 2013, Lemma 8.15] gives that ∂E must run through I. (Notice that the first paragraph in the proof of that lemma, proving this fact, does not use the hypothesis that $D(K)$ has at least three positive tangles.) Finally, a 2-edge loop of the third form runs over exactly two positive tangles. □

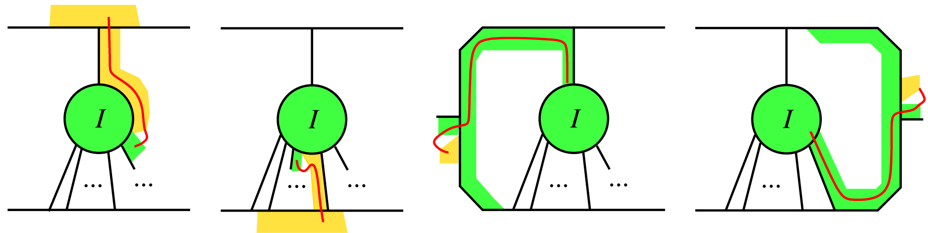
Notice Lemma 3.23 also applies to an EPD that has been pulled into a normal square. This is because we can pull an EPD E into a normal square without changing the segments that ∂E runs along.

Definition 3.24. A *type* (1) *EPD* is an EPD in P_A described by Lemma 3.23(1); similarly for a type (2) EPD.

By Lemma 3.23, every complex EPD in a Montesinos link must be either type (1) or type (2). This is not true for EPDs in general. Note also that an EPD may be both types (1) and (2).

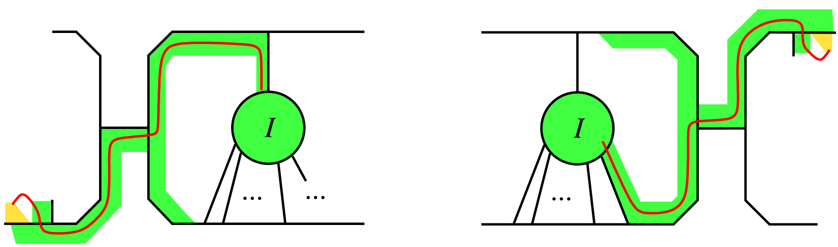
The following lemma gives further information about type (1) EPDs and follows from the proof of [Futer et al. 2013, Proposition 8.16].

Lemma 3.25. *Let K be a $++-$ or a $+-+-$ Montesinos link, and let E be a type* (1) *complex EPD in P_A as in Lemma 3.23, with the innermost disk I colored green, and E pulled into a normal square as in Lemma 3.15. Then the arcs of ∂E running over a white face may occur only in the following places*:



(a) North (b) South (c) North (d) South

Figure 15. Type (1N) and (1S) and type (2N) and (2S) white edges.



(a) North (b) South

Figure 16. Type (3N) and (3S) white edges.

(1) *On the innermost disk I, either at the tail of a tentacle running to I from the north or from the south, as in Figures 15*(*a*) *and 15*(*b*).

(2) *At the head of a tentacle running out of the negative block containing I, if ∂E runs from I to an adjacent positive tangle, as in Figures 15*(*c*) *and 15*(*d*).

(3) *On the next adjacent negative block, if ∂E runs downstream from I to an adjacent positive tangle, then across a segment spanning the positive tangle east to west, and then downstream across the outer state circle of the next negative block. See Figure 16. This situation cannot occur for $++-$ links.*

We will call these white arcs of types (1N), (1S), (2N), (2S), (3N), and (3S). Lemma 3.25 requires that I be colored green. However, recall that given an EPD, we may choose how to label faces green and gold. Hence classifying all type (1) EPDs will involve considering only the combinations resulting from the lemma.

Proof. See [Futer et al. 2013, Proposition 8.16]. In that proof, there are five types of arcs in white faces. However, types (4) and (5) require nonprime arcs, and we do not have any nonprime arcs in our diagrams. Notice that for a type (3) arc, we must have a string of tangles $-+-$ moving either east or west, where the first negative tangle is the one containing I. This string of tangles cannot occur in a Montesinos link with two positive tangles and one negative tangle. Thus $++-$ links may only have types (1) and (2) white edges. □

We will often obtain information about a complex EPD E by the following argument. Suppose E is pulled into a normal square with ∂E running through a given white face, cutting off a triangle. For any given white side, the two shaded faces met by ∂E can originate only at a few locations, which can be determined by considering the forms of the polyhedra as in Lemmas 3.11 and 3.12, and the A-resolutions of the tangles. By considering locations where the faces originate, we can identify ∂E. The following lemma is also useful.

Lemma 3.26 (Two-face argument). *Let E be a normal square such that ∂E meets shaded faces F_1 and F_2 at one arc in a white face W_1, and meets shaded faces F_1 and F_3 in another arc in a white face W_2. Then $F_2 = F_3$. Moreover, up to isotopy there is a unique embedded arc from W_1 to W_2 running through $F_2 = F_3$, and ∂E must follow this arc.*

Proof. The fact that $F_2 = F_3$ is obvious, since a normal square only runs through two shaded faces. The fact that there is a unique arc from W_1 to W_2 in F_2 follows from the fact that shaded faces are simply connected. □

4. Bounds on essential product disks

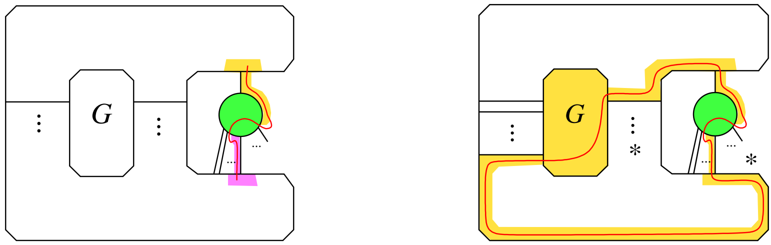
The main results of this section are Propositions 4.3 and 4.6, which bound $\|E_c\|$ for $++-$ and $+-+-$ Montesinos links, respectively. We prove these by looking first for complex EPDs that are of type (1), then for those that are not of type (1), which must necessarily be type (2).

The $++-$ case.

Lemma 4.1. *Let K be a $++-$ link with reduced, admissible diagram $D(K)$. Then there are only two type* (1) *complex EPDs in P_A, one with white faces of type* (1*N*) *and* (1*S*) *as shown in Figure 17*(*b*), *and the other with white faces of type* (2*N*) *and* (1*S*) *as shown in Figure 19*(*c*).

Proof. Let E be a type (1) complex EPD in P_A. Pull E into a normal square with two white edges as in Lemma 3.15. Lemma 3.25 tells us all the possible locations for white edges of ∂E. These possibilities are labeled (1N), (1S), (2N), and (2S). We must have one white edge north and one south. Therefore we must check only the following four combinations.

White edges (1N), (1S). By the notation (1N), (1S), we mean that our two white edges are described by type (1N) and type (1S) in Lemma 3.25. In this case, our EPD must have a portion as in Figure 17(a). By the two-face argument (Lemma 3.26), the faces to the north and south, colored gold and magenta, must agree. Therefore,



(a) (b)

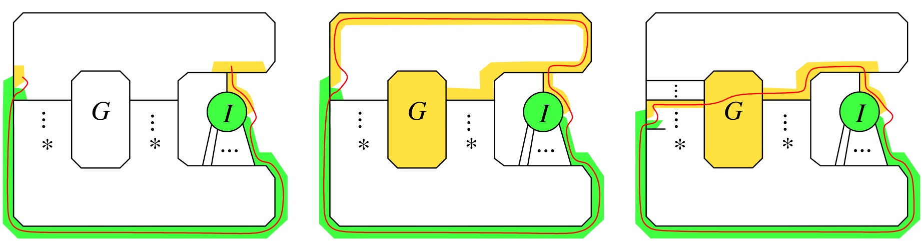
Figure 17. An EPD with white edges of type (1N) and (1S).

they must originate in the same innermost disk. We now consider where these faces might originate.

The face from the south, labeled magenta, must run around the bottom of the diagram on the inside, and therefore originate in the positive tangle T_a, or in the state circle G between T_a and T_b if T_a contains no state circles. The face from the north, labeled gold, runs from T_b, and therefore originates in G. Thus both originate in G and T_a contains no state circles. This is shown in Figure 17(b). Here ∂E may bound white faces that are not bigons on both sides, where indicated by an asterisk in the figure. Thus this may be a complex EPD.

White edges (1N), (2S). The boundary of an EPD with white edges of type (1N) and (2S) will be as in Figure 18(a), with the EPD following a green tentacle wrapped around the south of the diagram on the outside. A white edge of type (2S) requires that T_c has only one state circle, so that there are no nonbigons in T_c.

Now, the gold tentacle meeting I from the north originates in G. The gold tentacle meeting the green on the far west of the diagram will originate in T_a or in G; since the gold only has one innermost disk from which it originates, this must be G. From G, the gold face reaches the type (1N) white edge by running across T_b; but it may reach the (2S) white edge by running across either T_a or T_b. Therefore there are two possible EPDs, shown in Figures 18(b) and 18(c). In each



(a) (b) (c)

Figure 18. Semisimple EPDs with white edges of type (1N) and (2S).

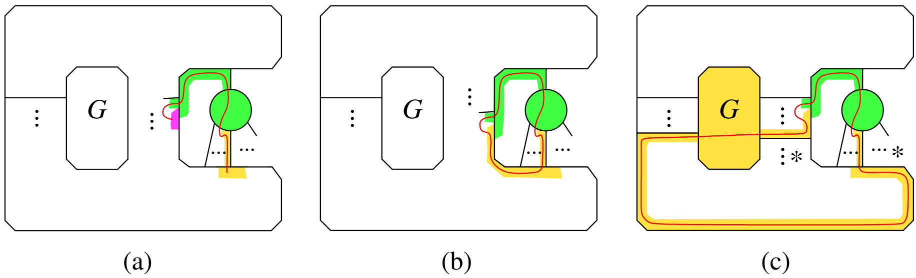


Figure 19. EPDs with white edges of type (2N) and (1S).

of these figures the only places where nonbigons may occur in the diagram are indicated by an asterisk. Note each of the EPDs bounds only bigons on one side, so they are both semisimple.

White edges (2N), (1S). A normal square with white edges (2N) and (1S) will have a diagram as in Figure 19(a). The gold face to the south wraps around the bottom of the diagram and originates in T_a or in G. The magenta face to the west, inside T_b, must either originate in T_b, or in G, or it is the same tentacle as the gold to the south.

Suppose first that the magenta face to the west is the same tentacle as the gold to the south of T_c. Then as shown in Figure 19(b), the EPD follows this tentacle from gold to the south to the tip of the face to the west, and the EPD is semisimple, encircling only bigons of T_c.

So suppose these are not the same tentacle. Since they lie in the same face, by Lemma 3.26, both must originate in G, as in Figure 19(c). Thus T_a has only bigon faces, but T_b and T_c might have nonbigon faces. The location of nonbigon faces is indicated by asterisks. Note that this may give a complex EPD.

White edges (2N), (2S). The (2N) and (2S) white edges give a diagram as in Figure 20(a). Note the gold and magenta faces must actually be the same face. We determine where gold and magenta tentacles may originate. This case is somewhat more complicated than the previous three. The magenta may originate in T_b or in G, or the magenta tentacle may agree with the tentacle running around the south of the diagram, and therefore originate at the furthest southwest state circle of T_a, or in G if T_a has no state circles. The gold may originate in G and run east to wrap around the north of the diagram, may originate in G and run west across a single segment in T_a, or may originate in T_a. Putting this information together, noting that gold and magenta originate in the same place (Lemma 3.26), we find that the face either originates in G, or in the far southwest state circle in T_a.

If the face originates in G there are four different possible diagrams, depending on whether gold or magenta run east or west, shown in Figures 20(b)–(e). If the

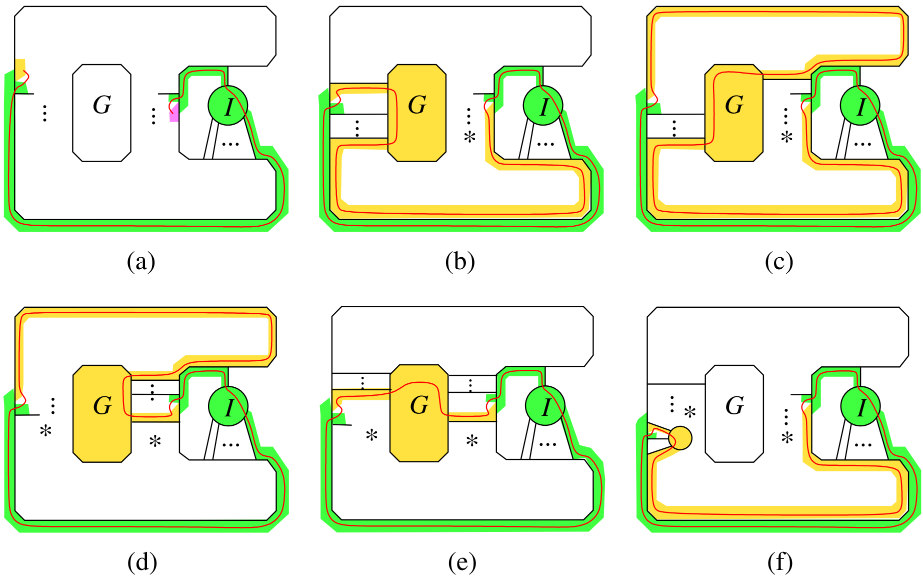


Figure 20. EPDs with white edges of type (2N), (2S).

face originates in T_a the diagram is as in Figure 20(f). Note that all of the EPDs pictured bound only bigons on one side, and therefore are semisimple.

This concludes the search for type (1) complex EPDs in P_A. The complex EPDs are depicted in Figures 17(b) and 19(c). □

Lemma 4.2. *Let K be a $++-$ link with reduced, admissible diagram $D(K)$. Then the only possible complex EPDs in P_A which are not of type* (1) *are those shown in Figures* 23(*c*) *and* 23(*d*).

Proof. Let E be a complex EPD in P_A that is not type (1); then E must be type (2). Therefore, ∂E runs along segments across both T_a and T_b. Each such segment has an endpoint on the innermost disk G. Since we may choose how to label the colors an EPD runs through, we choose to color the face bounded by G gold. By Lemma 3.15, we may pull E into a normal square with white sides at tails of the gold tentacles, and heads of the green.

For convenience, call the segment on the west α and the one on the east β. A priori, ∂E may run along either the north or south sides of these segments. However, by labeling G gold, we may restrict to the case that ∂E runs along the south of α and the north of β. This is because white sides occur at tails of gold and heads of green. There are no heads of any tentacles adjacent G, except those which run from innermost disk G itself, which are gold, so there can be no white sides adjacent G. Thus ∂E must run away from G through gold tentacles, which run south of α and north of β.

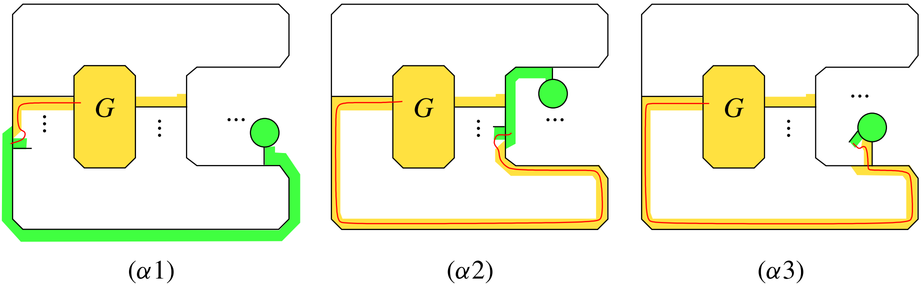


Figure 21. ∂E runs through a gold face across the south of α; shown are possible locations of white sides.

Tentacles running along the south side of α and the north side of β both originate in G. To find ∂E, we must look for a white side downstream from both α and β, and these must occur at a tail of a gold tentacle.

We have three possibilities for the location of the white side downstream from α:

(α1) The tentacle on the south side of α terminates in T_a. In this case the white side lies in T_a, and meets the head of a green tentacle coming from a tentacle originating in the far southeast of T_c. See Figure 21(α1).

(α2) The tentacle runs out of T_a, across the entire south of the diagram, and ∂E follows it to where it terminates in T_b. The white side lies in T_b, and meets a green tentacle originating in the far northwest of T_c. See Figure 21(α2).

(α3) The tentacle runs across the south of the diagram, meeting the head of a new gold tentacle in T_c, and ∂E runs into T_c. Such a tentacle must terminate on a state circle in T_c, and so the white side is in T_c, meeting a green tentacle in T_c originating at the state circle on which the gold tentacle terminates. See Figure 21(α3).

Similarly, there are three possibilities for the white side downstream from β:

(β1) The tentacle across the north of β terminates in T_b, and the white side lies in T_b. It meets a green tentacle originating at the state circle in the far northwest of T_c. See Figure 22(β1).

(β2) The tentacle across the north of β runs all the way around the north of the diagram, terminating in T_a, and ∂E follows this tentacle to T_a. The white side lies in T_a, and meets a green face with head on the tentacle running across the bottom of the diagram, originating in the far southeast state circle of T_c. See Figure 22(β2).

(β3) The tentacle across the north of β runs around the north of the diagram, but ∂E follows a new tentacle into T_c. This tentacle terminates on a state circle in T_c,

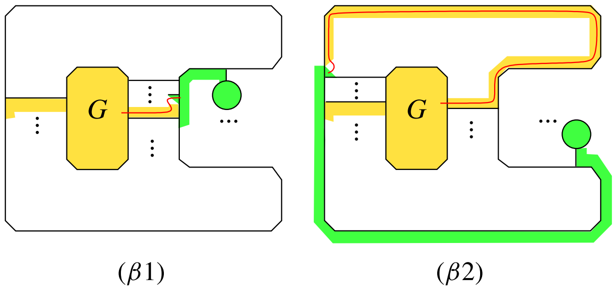


Figure 22. ∂E runs through a gold face across the north of β; shown are possible locations of white edges and origins of the green tentacles they meet.

and the white side must be in T_c, connecting to a green face that originates in this state circle. See Figure 22(β3).

The above give nine possible combinations, which we reduce. Note that in order for (α1) to combine with (β1) or (β3), there can only be one state circle in T_c and ∂E must run over it. This means E is a type (1) EPD, contradicting assumption. An identical argument rules out (α2) combined with (β2), and (α3) combined with either (β1) or (β3).

We work through the remaining combinations. First, (α1) and (β2), shown in Figure 23(a), gives rise to a semisimple EPD. Similarly, (α3) and (β2), shown in Figure 23(b), gives a semisimple EPD. The combination (α2) combined with (β1) may give a complex EPD, as does (α2) combined with (β3). These are shown in Figures 23(c) and 23(d). $\square$

Proposition 4.3. *Let K be a $++-$ link with reduced, admissible diagram $D(K)$. Then $\|E_c\| \leq 1$, where $\|E_c\|$ is the number of complex EPDs required to span P_A.*

Proof. To obtain the desired result, we need to show that if E and E' are two complex EPDs in P_A, then E' is equivalent under parabolic compression to a subset of $E \cup E_s$.

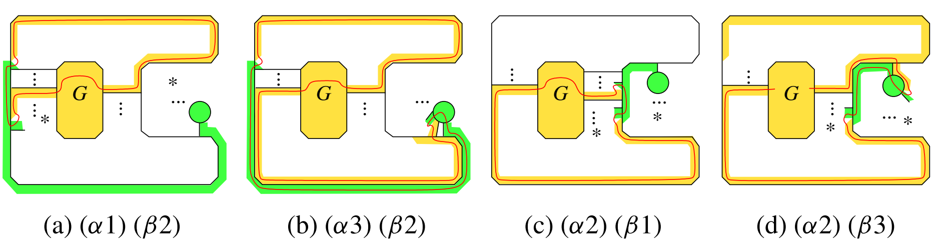


Figure 23. Possibilities for ∂E running over the south of α and the north of β.

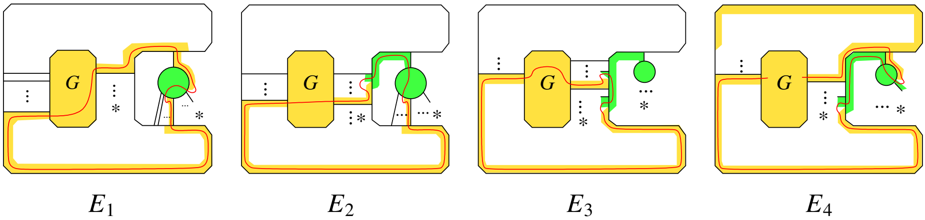


Figure 24. All complex EPDs in P_A for a $++-$ link.

In Lemmas 4.1 and 4.2, we found all possible complex EPDs in P_A. They are shown in Figures 17(b), 19(c), 23(c), and 23(d), which we reproduce for convenience in Figure 24. Denote the four complex EPDs by E_1, E_2, E_3, and E_4, as illustrated. We compare these pairwise.

First consider E_1 and E_2. Note both EPDs run over the southernmost segment of T_a. In T_b, E_1 runs over the northernmost segment, and E_2 runs over a segment besides the northernmost one that still spans T_b. Note they both run from the segment spanning T_b to the north of T_c, and so there are only bigons in T_b between them. In T_c, then run from north to south, and either one may run down any segments connecting the state circle I to the state circle on the south side of T_c. From there, both run along the inside south face, back to the southernmost segment of T_a. Thus there are only bigons in T_c and in T_b between the two EPDs. Therefore by Lemma 3.21, E_1 and E_2 are equivalent under parabolic compression.

The analysis of other pairs is similar. For E_1 and E_3, the two EPDs differ by only bigons in T_b, and bigons in T_c, including the bigon on the far west side of T_c. Similarly for E_1 and E_4. The pair E_2 and E_3 differ only by bigons in T_b and in T_c, as do E_2 and E_4. Finally, E_3 and E_4 differ only by bigons in T_b. So in all cases, Lemma 3.21 implies that the EPDs are equivalent under parabolic compression. □

The $+-+-$ case.

Lemma 4.4. *Let K be a $+-+-$ link with reduced, admissible diagram $D(K)$. Then a type* (1) *complex EPD in P_A is equivalent to one with white edges of type* (1*N*) *and* (1*S*), *shown in Figure 25*(*b*), *or of type* (2*N*) *and* (1*S*), *shown in Figure 28*(*c*), *or of type* (3*N*) *and* (2*S*), *shown in Figure 32*(*b*).

Proof. For E a type (1) complex EPD in P_A, we pull E into a normal square with two white edges (Lemma 3.15). Since E is type (1), ∂E runs through a negative tangle from north to south. We may assume, after possibly taking a cyclic permutation, that ∂E runs through T_d. Lemma 3.25 tells us all the possible locations for white edges of ∂E. These possibilities are labeled (1N), (1S), (2N), (2S), (3N), and (3S). Since ∂E runs over the segment to the north and a segment to the south of the state circle I in T_d, there are nine combinations to check.

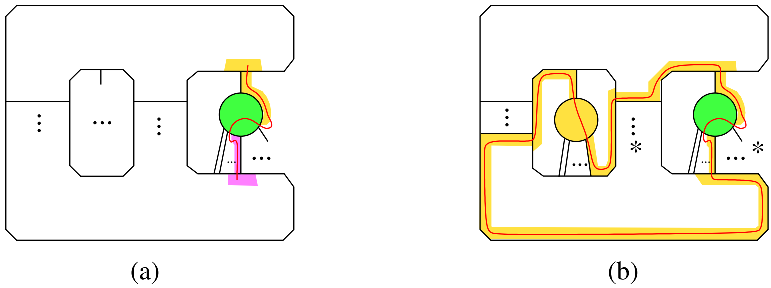


Figure 25. For $+-+-$ link, edges of type (1N) and (1S); possible complex EPD shown in (b).

White edges (1N), (1S). By the notation (1N), (1S), we mean that our two white edges are described by type (1N) and type (1S) in Lemma 3.25. Our EPD must have white edges as in Figure 25(a). By Lemma 3.26, the gold and magenta faces shown must agree. The magenta tentacle has its head on the tentacle that wraps all the way around the bottom of the diagram. This tentacle either originates in the far southwest state circle of T_a, or if T_a has no state circles, from the far northwest state circle of T_b. The gold tentacle has its head on the tentacle that wraps around the inside of the top of the diagram; it must run over a segment connecting east to west of T_c, and originates in the far southwest state circle of T_b. For these originating state circles to be the same, there must be only one state circle in T_b, and gold and magenta must originate there. Thus there are no state circles in T_a. This results in the diagram pictured in Figure 25(b). This diagram may contain nonbigon white faces where indicated by an asterisk, so this EPD may be complex.

White edges (1N), (2S). In this case, ∂E must run over faces of P_A that look like Figure 26(a). Since a white edge is of type (2S), we only have one state circle in T_d. The gold tentacle in this figure has its head on the tentacle running across the inside-top of the diagram; this originates in the furthest southeast state circle of T_b. By Lemma 3.26, the magenta tentacle must originate from the same state

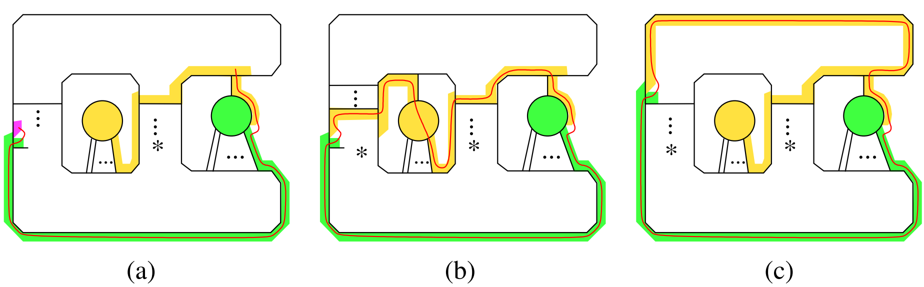


Figure 26. White edges of type (1N) and (2S).

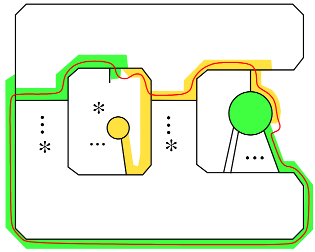
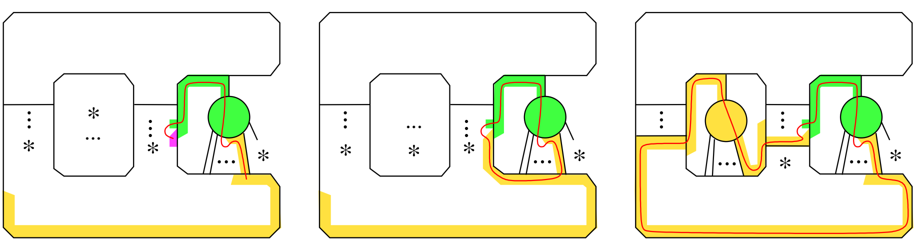


Figure 27. White edges of type (1N) and (3S).

circle. However, the magenta tentacle either originates in T_a, or runs through T_a and originates in the northwest of T_b, or runs across the inside-top of the diagram and originates in the southeast of T_b. Originating in T_a is impossible, but the other two options are both valid, provided in the case the face originates in the northwest of T_b there is only one state circle in T_b. These two valid options result in Figures 26(b) and 26(c), respectively. Note the EPD in Figure 26(c) bounds a single bigon on the outside, and so is simple. In Figure 26(b), nonbigons can only occur where indicated by an asterisk. This EPD is semisimple.

White edges (1N), (3S). White edges of type (1N) and (3S) produce a diagram as in Figure 27. Because a white edge is of type (3S), T_d must have only one state circle. In this case, both gold tentacles originate in the southeast corner of the tentacle T_b, as shown, and only one diagram is possible. White faces that are not bigons may occur only where indicated by an asterisk. Notice that the outside of ∂E in this case bounds only bigons, so the EPD is semisimple.

White edges (2N), (1S). A normal square with white edges of type (2N) and (1S) will have a diagram as in Figure 28(a). The gold tentacle has its head on the tentacle wrapping all along the inside-bottom of the diagram. Thus the gold face originates in the far southwest state circle in T_a, or if T_a has no state circles, in the far northwest state circle of T_b. The magenta tentacle either originates in T_c, or



(a) (b) (c)

Figure 28. White edges of type (2N) and (1S).

runs from east to west in T_c and originates in the southeast corner of T_b, or it might be the same tentacle that runs across the inside-bottom of the diagram.

If the magenta tentacle is the tentacle that runs across the inside-bottom of the diagram, then we can close up ∂E in such a way that it encircles only bigons in T_d, as in Figure 28(b). Because the gold face is simply connected, this is the only way to connect ∂E up to homotopy, and so E is semisimple in this case.

Since gold and magenta must originate in the same state circle, the only remaining possibility is that both originate in T_b, so T_a contains no state circles, and in T_b, the northwest and southeast state circles agree, so T_b has only one state circle. The result is shown in Figure 28(c). Nonbigons may be present where indicated by an asterisk; the EPD may be complex.

White edges (2N), (2S). A normal square with white edges of type (2N) and (2S) has a diagram as in Figure 29(a). Notice that type (2S) forces T_d to have only one state circle. The magenta tentacle from (2S) either

- originates in T_a, or
- runs across T_a and originates in the northwest of T_b, or
- comes from the tentacle running across the inside north of the diagram, and therefore originates in the southeast of T_b.

The gold tentacle from (2N) either

- originates in T_c, or

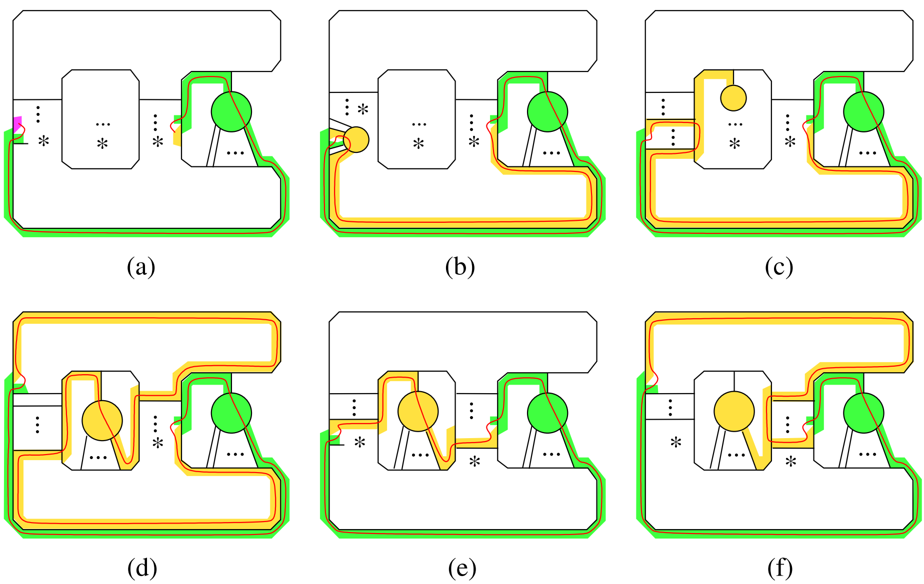
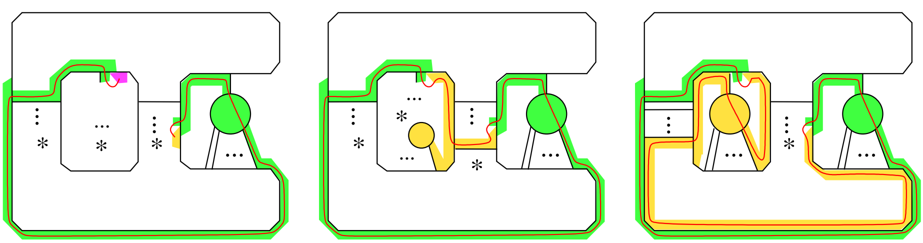


Figure 29. White edges of type (2N) and (2S).

- runs across T_c and originates in the southeast of T_b, or
- comes from the tentacle running across the inside south of the diagram, and therefore originates in the south of T_a, or
- if T_a has no state circles, comes from the tentacle running across the inside south of the diagram and originates in the northwest of T_b.

Since magenta and gold originate in the same place, (2N) cannot originate in T_c, but there are five possibilities remaining, as follows. First, gold originates in T_a; this is shown in Figure 29(b). Note it is semisimple. Second, gold originates in T_b, the tentacle from (2N) runs around the inside south of the diagram, and the tentacle from (2S) originates in the northwest of T_b; this is shown in Figure 29(c), and again it is semisimple. Third, gold originates in T_b, the tentacle from (2N) runs around the inside south of the diagram, and the tentacle from (2S) runs around the inside north of the diagram, as in Figure 29(d). Note this is also semisimple. Fourth, gold originates in T_b, the tentacle from (2N) runs west to east across T_c, and the tentacle from (2S) runs east to west across T_a, as in Figure 29(e), which is semisimple. Finally fifth, gold originates in T_b, the tentacle from (2N) runs west to east across T_c, and the tentacle from (2S) runs across the inside north of the diagram, as in Figure 29(f). Again this is semisimple, bounding only bigons to the outside.

White edges (2N), (3S). White edges (2N) and (3S) give the diagram in Figure 30(a). Since we have a (3S) white edge, there can be only one state circle in T_d. The magenta tentacle originates in the southeast state circle in T_b. Therefore, the gold tentacle must also originate in T_b. This can happen one of two ways: either the gold tentacle runs west to east across T_c, originating in the southeast of T_b, or the gold tentacle runs all across the inside south of the diagram, across a tentacle of T_a (which must have no state circles), and originates in the northeast of T_b (which must have exactly one state circle). These two options are shown in Figures 30(b) and 30(c). Note both bound only bigons to one side; hence both are semisimple.



(a) (b) (c)

Figure 30. White edges of type (2N) and (3S).

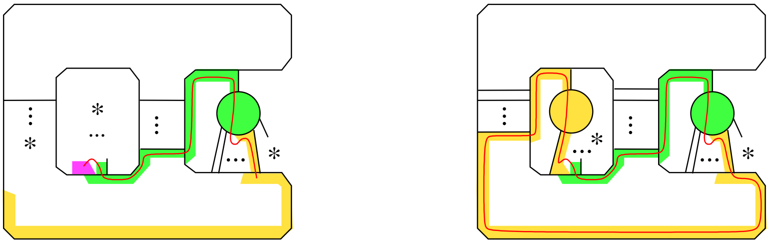


Figure 31. White edges of type (3N) and (1S).

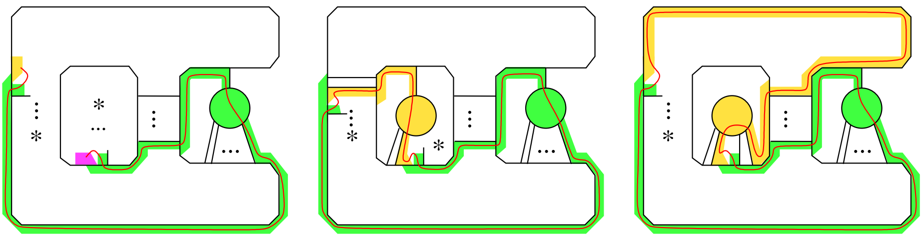


Figure 32. White edges of type (3N) and (2S).

White edges (3N), (1S). The diagram appears as in Figure 31(a). Note that a type (3N) white edge requires that T_c has no state circles. The magenta tentacle originates in T_b; hence the gold also must originate in T_b. This means T_a has no state circles, and the gold tentacle runs across T_a, originating in the northwest of T_b. Then the magenta originates in the same circle, and the diagram is as in Figure 31(b). Note this gives a semisimple EPD.

White edges (3N), (2S). In this case, the diagram is as shown in Figure 32(a). Note that (3N) implies that T_c has no state circles, and (2S) implies that T_d has only one state circle. The magenta tentacle originates in T_b. The gold face must also originate in T_b. This is possible if either the gold tentacle runs across T_a and originates in the northwest of T_b, or the gold tentacle runs across the inside north of the diagram and originates in the southeast of T_b. The two options are shown in Figures 32(b) and 32(c). In the first case, we obtain an EPD that may be complex. In the second case, the EPD is semisimple.

White edges (3N), (3S). The case of type (3N) and (3S) is shown in Figure 33. The gold faces meeting the green tentacles must both originate in T_b, at the southeast corner. Hence the EPD is as shown; it is semisimple.

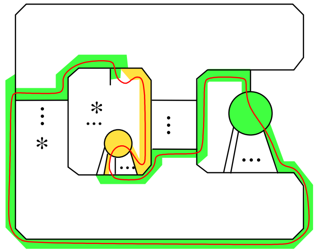


Figure 33. White edges of type (3N) and (3S).

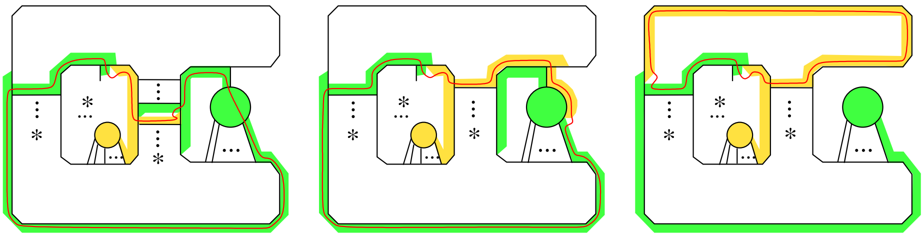
This concludes the search for type (1) complex EPDs in P_A. The complex EPDs found are in Figures 25(b), 28(c), and 32(b). □

Lemma 4.5. *Let K be a $+-+-$ link with reduced, admissible diagram $D(K)$. Then there are no complex EPDs in P_A which are not type* (1).

Proof. Let E be a complex EPD in P_A that is not type (1); then E must be type (2). That is, ∂E runs along segments across both T_a and T_c. For convenience, let's call these two segments α and β. Now, ∂E may run along either the north or south sides of these segments. Therefore we have four possibilities, listed below. By Lemma 3.15, we may pull E into a normal square with white edges at tails of the gold tentacles.

Case αN, βN. We interpret the notation αN, βN to mean that ∂E runs along the north side of α and the north side of β. The tentacle running along the north side of α originates in the southeast state circle in T_d. The tentacle running along the north side of β originates in the southeast state circle in T_b. These two cannot agree; hence one is green and one is gold. Say the tentacle running along α is green and the one along β is gold.

After running west to east along the north of α, ∂E must meet a white face at the head of a green tentacle. The only option is that there is a white face at the north of T_b. Similarly, after running west to east along the north of β, ∂E must meet a white face at the tail of gold. There are three options. First, if there is only one state circle in T_d, then the tentacle along the west side of T_d will be green, and ∂E could meet the head of a green tentacle on the east side of T_c. This is shown in Figure 34(a); note the result is type (1), contrary to assumption. Second, again if there is only one state circle in T_d, then ∂E could run across the north of T_d, into a tentacle in T_d, to the head of a green tentacle in T_d. This is shown in Figure 34(b). Note this is again type (1), contradicting assumption. Finally, ∂E could follow a gold tentacle all across the inside north of the diagram, meeting the head of a green tentacle in T_a. This is shown in Figure 34(c); note the result is a simple EPD.



(a) (b) (c)

Figure 34. ∂E runs along the north of α, north of β.

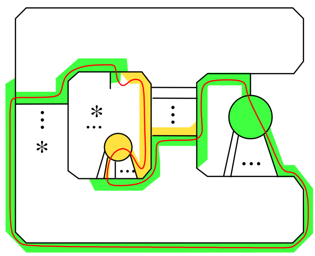


Figure 35. ∂E runs along the north of α, south of β.

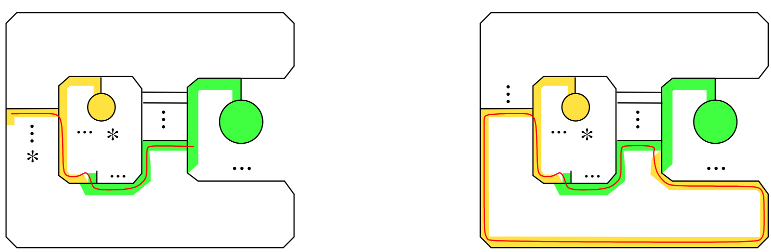
Case αN, βS. Again, the tentacle running over α to the north originates in the southeast state circle of T_d; color this face green. The tentacle running over β to the south originates in the northwest state circle of T_d. Suppose first that these state circles of T_d are not the same; so we color the face running over β to the south gold. Now, as in the previous case ∂E must run east out of α to a white edge, which must be at a head of green. There is only one possible head of green, at the north of T_b, as before. However, at this white edge ∂E runs into the face originating in the southeast of T_b; we color this face magenta. Then ∂E runs through the green, magenta, and gold faces, which are all distinct. This contradicts Lemma 3.26.

Therefore, it must be that the northwest and southeast state circles of T_d are the same, and this face is green. Note that ∂E runs from β to a white face, at the head of a green tentacle. The only way to meet the head of a green tentacle running west across β is if T_c has no state circles, β is the far south segment in T_c, and ∂E runs around the outside south of T_b to meet a white face just inside T_c. Then ∂E must be as shown in Figure 35. This is a type (1) EPD, contrary to assumption.

Case αS, βN. The tentacle running over the south of α originates in the northwest state circle in T_b. The tentacle running over the north of β originates in the southeast of T_b. If these two state circles are the same, then between α and β, ∂E must run over T_b from north to south, and E is of type (1); contradiction.

So assume the northwest and southeast state circles of T_b are distinct. Color the northwest state circle gold, and the southeast one green. After ∂E runs east to west along the gold face to the south of α, it must meet a white face at the tail of a gold tentacle. There are three ways this can happen, but we claim that all lead to a face originating in T_d and hence give rise to a distinct color. First, the gold tentacle to the south of α can terminate in T_a. In this case ∂E meets a white face in T_a and jumps to the tentacle running around the outside south of the diagram. This face originates in the southeast of T_d. So suppose, second, the gold tentacle south of α runs all along the inside south of the diagram, and ∂E follows it to where it terminates, in T_c. Then it jumps to a green tentacle originating in the northwest of T_d. So finally, suppose the gold tentacle south of α runs along the inside south of the diagram, but ∂E follows a new gold tentacle from the south of T_d into T_d. Then this tentacle has its tail on a state circle in T_d, which must be where the green face originates. All three options require the green face to originate in T_d and in T_b, which is impossible.

Case αS, βS. As in the previous case, the tentacle on the south of α originates in the north of T_b. The tentacle on the south of β originates in the north of T_d. Color the tentacle south of α gold, and the one south of β green. Note that ∂E, after running east to west along β, must run to a white edge at the head of a green tentacle. The only way this is possible is if T_c contains no state circles, and β is the segment at the far south of T_c. The white face must occur at the south of T_b, jumping from the head of a green tentacle to the tail of a gold tentacle. This gold tentacle originates in a state circle in T_b. This state circle must agree with the origin of the gold tentacle at the south of α. In order to form an EPD that is not of type (1), ∂E must run over the far west side of T_b, as in Figure 36(a).



(a) (b)

Figure 36. ∂E runs along the south of α, south of β.

Now after ∂E runs east to west along the south of α, it must run to the tail of a gold tentacle and the head of a green. Just as in the previous case, there are three possibilities: the gold tentacle terminates in T_a, the gold tentacle runs all along the inside south of the diagram and terminates in T_c and ∂E follows it to T_c, or the gold tentacle runs along the inside south of the diagram, but ∂E follows a new gold tentacle into T_d. The first and last cases can only happen if there is just one state circle in T_d and ∂E runs from the south to the north of T_d, contradicting the fact that E is not type (1). The middle case is shown in Figure 36(b). Note it results in an EPD bounding a bigon, so it is not complex. $\square$

Proposition 4.6. *Let K be a $+-+-$ link with reduced, admissible diagram $D(K)$. Then $\|E_c\| \le 1$, where $\|E_c\|$ is the number of complex EPDs required to span P_A.*

Proof. Again to obtain the desired result, we need to show that if E and E' are two complex EPDs in P_A, then E' is equivalent under parabolic compression to a subset of $E \cup E_s$.

In Lemmas 4.4 and 4.5, we found all possible complex EPDs in P_A. These are all type (1), and appear in Figures 25(b), 28(c), and 32(b), which we reproduce in Figure 37 for convenience, calling the EPDs E_1, E_2, and E_3. We compare the complex EPDs pairwise.

Consider first the pair E_1 and E_2. Both run over the same segments of T_a and T_b. In T_c, E_1 runs over the northernmost segment while E_2 runs over a different, parallel segment. However, both then run to the northernmost segment of T_d, and then to the south of T_d, where they meet again. Thus these two EPDs differ only by a collection of bigons in T_c and possibly a collection of bigons in T_d. By Lemma 3.21, the two EPDs are equivalent under parabolic compression.

If E_1 and E_3 both appear in P_A, then note that the asterisks in both figures will be replaced by bigons, and again the two EPDs differ only by bigons. Finally, E_2 and E_3 also differ only by bigons (in fact, E_2 is a cyclic permutation of E_3). Lemma 3.21 implies the EPDs are equivalent under parabolic compression. $\square$

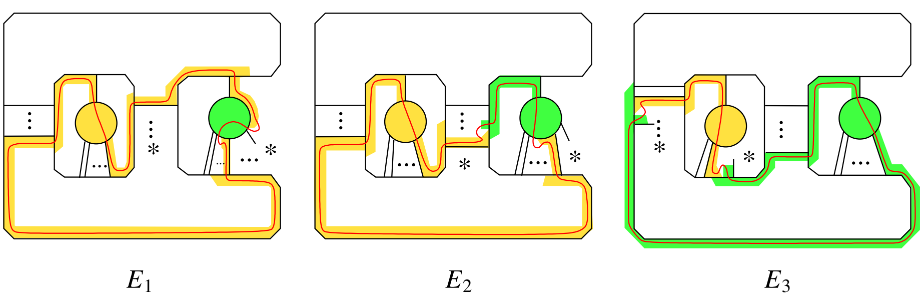


Figure 37. Complex EPDs in P_A for a $+-+-$ link.

5. Proofs of main results

To complete the proofs of the main theorems, we will use the following result.

Theorem 5.1 [Futer et al. 2013, Theorem 5.14]. *Let $D = D(K)$ be a prime A-adequate diagram of a hyperbolic link K. Then*

$$\mathrm{vol}(S^3 \setminus K) \geq v_8(\chi_-(\mathbb{G}'_A) - \|E_c\|),$$

where $\chi_-(\cdot)$ is the negative Euler characteristic, $\mathbb{G}'_A$ is the reduced A-state graph, and $\|E_c\|$ is the number of essential product disks required to span the upper polyhedron of the knot complement.

Proof of Theorem 1.1. Let K be a hyperbolic Montesinos link with reduced, admissible diagram. We will use Theorem 5.1 to bound $\mathrm{vol}(S^3 \setminus K)$. The hypotheses of the theorem are that K is prime and that D is A-adequate. The diagram D is prime by Proposition 2.7. By Theorem 3.6, the diagram is either A- or B-adequate. If not A-adequate, then it must be B-adequate, and so the mirror image is A-adequate. Taking the mirror image does not change the volume of the knot complement, so we may assume that the diagram is A-adequate.

Now we may use Theorem 5.1. Notice that we obtain the desired volume bound if we show $\|E_c\| \leq 1$. We divide into several simple cases.

Case 1. Suppose K has either all positive or all negative tangles. Then the diagram is alternating, so $\|E_c\| = 0$ (see [Lackenby 2004]).

Case 2. Suppose K has three tangles, with slopes not all the same sign. If K is a $++-$ link, then by Proposition 4.3, $\|E_c\| \leq 1$, so the desired volume bound holds. If K is a $+--$ tangle, then apply that proposition to the mirror image. These are the only links in this case, up to cyclic permutation.

Case 3. Suppose K has four tangles, with slopes not all the same sign. Up to cyclic permutation, there are four ways the positive and negative tangles can be arranged:

(1) $+-+-$: By Proposition 4.6, $\|E_c\| \leq 1$. Thus the desired volume bound holds.

(2) $++--$: As explained in Remark 2.11, this link is equivalent to a link of type $+-+-$, and so $\|E_c\| \leq 1$.

(3) $+++-$ or $+---$: By Theorem 3.22, $\|E_c\| = 0$ for such a link or its mirror image.

Case 4. Suppose K has five or more tangles. Then K must have at least three positive or at least three negative tangles. Thus Theorem 3.22 applies to K or its mirror image. □

We now turn our attention to the proof of Theorem 1.2. This theorem generalizes [Futer et al. 2013, Theorem 9.12], which requires at least three positive and at least three negative tangles. Much of the proof of that result goes through in the setting of fewer positive and negative tangles.

Lemma 5.2. *Let K be a Montesinos link that admits a reduced, admissible diagram D with at least two positive and at least two negative tangles. Let $\mathbb{G}'_A$ and $\mathbb{G}'_B$ be the reduced all-A and all-B graphs associated to D. Then*

$$-\chi(\mathbb{G}'_A) - \chi(\mathbb{G}'_B) \geq t(K) - Q_{1/2}(K) - 2,$$

where $Q_{1/2}(K)$ is the number of rational tangles whose slope has absolute value $|q| \in \left[\frac{1}{2}, 1\right)$.

Proof. We follow the proof of [Futer et al. 2013, Lemma 9.10]. Since D has at least two positive and at least two negative tangles, D is both A- and B-adequate. Let v_A be the number of vertices in $\mathbb{G}_A$, e_A the number of edges in $\mathbb{G}_A$, and e'_A the number of edges in $\mathbb{G}'_A$; and similarly for v_B, e_B, and e'_B. Then we have $-\chi(\mathbb{G}_A) = v_A - e_A$ and $-\chi(\mathbb{G}'_A) = v_A - e'_A$, and likewise for $-\chi(\mathbb{G}_B)$ and $-\chi(\mathbb{G}'_A)$.

Now construct the Turaev surface T for D, as in [Dasbach et al. 2008, Section 4]. The diagram D will be alternating on T, and the graphs $\mathbb{G}_A$ and $\mathbb{G}_B$ embed in T as graphs of the alternating projection and are dual to one another, and so the number of regions in the complement of $\mathbb{G}_A$ on T is equal to v_B. Because K is a cyclic sum of alternating tangles, T is a torus, just as in [Futer et al. 2013, Lemma 9.10]. Thus we have

$$v_A - e_A + v_B = \chi(T) = 0.$$

Now consider the number of edges of $\mathbb{G}_A$ that are discarded when we pass to $\mathbb{G}'_A$. By [Futer et al. 2013, Lemma 8.14], edges may be lost in three ways:

(1) If r is a twist region with $c(r) > 1$ crossings such that the A-resolution of r gives $c(r)$ parallel segments, then $c(r) - 1$ of these edges will be discarded when we pass to $\mathbb{G}'_A$.

(2) If N_i is a negative tangle with slope $q \in \left(-1, -\frac{1}{2}\right]$, then one edge of $\mathbb{G}_A$ will be lost from the 2-edge loop spanning N_i from north to south.

(3) If there are exactly two positive tangles P_1 and P_2, then one edge of $\mathbb{G}_A$ will be lost from the 2-edge loop that runs across P_1 and P_2 from east to west.

The same holds for $\mathbb{G}_B$, with A and B switched and positive and negative tangles switched. Combining this information, we obtain

$$\begin{aligned}(e_A - e'_A) + (e_B - e'_B) &\leq \sum \{c(r) - 1\} + \#\left\{i : |q_i| \in \left[\tfrac{1}{2}, 1\right)\right\} + 2 \\ &= c(D) - t(D) + Q_{1/2}(D) + 2,\end{aligned}$$

where the sum is over all twist regions r.

Since the edges of $\mathbb{G}_B$ are in one-to-one correspondence with the crossings of D, we have

$$\begin{aligned}-\chi(\mathbb{G}'_A) - \chi(G'_B) &= e'_A + e'_B - v_A - v_B \\ &= (e'_A + e'_B - e_A - e_B) + e_B + (e_A - v_A - v_B) \\ &\geq (-c(D) + t(D) - Q_{1/2}(D) - 2) + c(D) + 0 \\ &= t(D) - Q_{1/2}(D) - 2. \qquad \square\end{aligned}$$

Lemma 5.3. *Let D be a reduced, admissible Montesinos diagram with at least two positive tangles and at least two negative tangles. Then*

$$Q_{1/2} \leq \tfrac{1}{2}(t(D) + \#K),$$

where $\#K$ denotes the number of link components of K.

Proof. Again we follow the proof of [Futer et al. 2013, Lemma 9.11]. The number $Q_{1/2}(K)$ is equal to the number of tangles with slope q satisfying $|q| \in \left(\frac{1}{2}, 1\right)$ plus the number of tangles with slope q satisfying $|q| = \frac{1}{2}$.

A tangle R_i of slope $|q_i| \in \left(\frac{1}{2}, 1\right)$ has at least two twist regions, $t(R_i) \geq 2$. Thus $\frac{1}{2}t(R_i) \geq 1$ for such a tangle.

A tangle of slope $|q| = \frac{1}{2}$ has only one twist region. It can be replaced by a tangle of slope ∞ without changing the number of link components of the diagram, but such a replacement gives a diagram with a "break" in it. Thus if n is the number of tangles of slope $|q| = \frac{1}{2}$, then $n \leq \#K$. Hence we have

$$\begin{aligned}Q_{1/2}(D) &= \sum_{\{R_i \text{ tangle}:|q_i|\in(1/2,1)\}} 1 + \sum_{\{R_j \text{ tangle}:|q_j|=1/2\}} 1 \\ &\leq \sum_{\{R_i:|q_i|\in(1/2,1)\}} \tfrac{1}{2}t(R_i) + \sum_{\{R_j:|q_j|=1/2\}} \tfrac{1}{2}(t(R_j)+1) \\ &\leq \tfrac{1}{2}(t(D) + \#K) \qquad \square\end{aligned}$$

Proof of Theorem 1.2. If K admits a reduced, admissible diagram with at least two positive tangles, then by work of Bonahon and Siebenmann [2010], the complement of K must be hyperbolic, unless K is the $(2, -2, 2, -2)$ pretzel link (see also [Futer and Guéritaud 2009, Section 3.3]). We exclude this pretzel link.

For the lower bound in the theorem, by Lemmas 5.2 and 5.3, we have

$$-\chi(\mathbb{G}'_A) - \chi(\mathbb{G}'_B) \geq \tfrac{1}{2}(t(D) - \#K - 4).$$

By Theorem 3.7, S_A and S_B are both essential in $S^3 \setminus A$. Then a theorem of Agol, Storm, and Thurston [Agol et al. 2007, Theorem 9.1] applied to S_A and S_B

implies that

$$\operatorname{vol}(S^3 \setminus K) \geq \tfrac{1}{2} v_8(\chi_- \operatorname{guts}(S^3 \backslash\backslash S_A) + \chi_- \operatorname{guts}(S^3 \backslash\backslash S_B)).$$

By [Futer et al. 2013, Theorem 5.14], $\operatorname{guts}(S^3 \backslash\backslash S_A) = \chi_-(\mathbb{G}'_A) - \|E_c\|$; similarly for $\operatorname{guts}(S^3 \backslash\backslash S_B)$. By Theorem 5.1, along with Propositions 4.3 and 4.6, we may assume that $\|E_c\| \leq 1$ (possibly after a mutation of the diagram in the $++--$ case), for both the A and B cases. Thus

$$\begin{aligned}\operatorname{vol}(S^3 \setminus K) &\geq \tfrac{1}{2} v_8(\chi_-(\mathbb{G}'_A) + \chi_-(\mathbb{G}'_B)) - \tfrac{1}{2} v_8(\|E_c\|_A + \|E_c\|_B) \\ &\geq \tfrac{1}{4} v_8(t(D) - \#K - 4) - \tfrac{1}{2} v_8(2) \\ &\geq \tfrac{1}{4} v_8(t(D) - \#K - 8).\end{aligned}$$

As for the upper bound on volume, this goes straight through as in the proof of [Futer et al. 2013, Theorem 9.12] without change. That is, augment the Montesinos link by drilling out a link component B encircling two strands at the east of the tangle. The result is hyperbolic, and a belted sum of tangles as in [Adams 1985]. The estimates on volume follow. □

References

[Adams 1985] C. C. Adams, "Thrice-punctured spheres in hyperbolic 3-manifolds", *Trans. Amer. Math. Soc.* **287**:2 (1985), 645–656. MR 86k:57008 Zbl 0527.57002

[Agol et al. 2007] I. Agol, P. A. Storm, and W. P. Thurston, "Lower bounds on volumes of hyperbolic Haken 3-manifolds", *J. Amer. Math. Soc.* **20**:4 (2007), 1053–1077. MR 2008i:53086 Zbl 1155.58016

[Bonahon and Siebenmann 2010] F. Bonahon and L. Siebenmann, "New geometric splittings of classical knots, and the classification and symmetries of arborescent knots", preprint, 2010, available at http://www-bcf.usc.edu/~fbonahon/Research/Preprints/BonSieb.pdf. To appear in *Geometry and Topology Monographs*.

[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* (Oxford, 1967), edited by J. Leech, Pergamon, Oxford, 1970. MR 41 #2661 Zbl 0202.54703

[Dasbach et al. 2008] O. T. Dasbach, D. Futer, E. Kalfagianni, X.-S. Lin, and N. W. Stoltzfus, "The Jones polynomial and graphs on surfaces", *J. Combin. Theory Ser. B* **98**:2 (2008), 384–399. MR 2009d:57020 Zbl 1135.05015

[Futer and Guéritaud 2009] D. Futer and F. Guéritaud, "Angled decompositions of arborescent link complements", *Proc. London Math. Soc.* (3) **98**:2 (2009), 325–364. MR 2009m:57008 Zbl 1163.57003

[Futer et al. 2013] D. Futer, E. Kalfagianni, and J. S. Purcell, *Guts of surfaces and the colored Jones polynomial*, Lecture Notes in Mathematics **2069**, Springer, Heidelberg, 2013. MR 3024600 Zbl 1270.57002

[Futer et al. 2014] D. Futer, E. Kalfagianni, and J. S. Purcell, "Jones polynomials, volume and essential knot surfaces: a survey", pp. 51–77 in *Knots in Poland, III, part 1*, edited by J. H. Przytycki and P. Traczyk, Banach Center Publ. **100**, Polish Acad. Sci. Inst. Math., Warsaw, 2014. MR 3220475 Zbl 1285.57002

[Lackenby 2004] M. Lackenby, "The volume of hyperbolic alternating link complements", *Proc. London Math. Soc.* (3) **88**:1 (2004), 204–224. MR 2004i:57008 Zbl 1041.57002

[Lickorish and Thistlethwaite 1988] W. B. R. Lickorish and M. B. Thistlethwaite, "Some links with nontrivial polynomials and their crossing-numbers", *Comment. Math. Helv.* **63**:4 (1988), 527–539. MR 90a:57010 Zbl 0686.57002

[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

[Mostow 1968] G. D. Mostow, "Quasi-conformal mappings in n-space and the rigidity of hyperbolic space forms", *Inst. Hautes Études Sci. Publ. Math.* **34** (1968), 53–104. MR 38 #4679 Zbl 0189.09402

[Murasugi 1996] K. Murasugi, *Knot theory and its applications*, Birkhäuser, Boston, 1996. MR 97g:57011 Zbl 0864.57001

[Ozawa 2011] M. Ozawa, "Essential state surfaces for knots and links", *J. Aust. Math. Soc.* **91**:3 (2011), 391–404. MR 2900614 Zbl 1241.57012

[Prasad 1973] G. Prasad, "Strong rigidity of $\mathbb{Q}$-rank 1 lattices", *Invent. Math.* **21** (1973), 255–286. MR 52 #5875 Zbl 0264.22009

[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

Received March 9, 2015. Revised September 23, 2015.

KATHLEEN FINLINSON
DEPARTMENT OF APPLIED MATHEMATICS
UNIVERSITY OF COLORADO BOULDER
526 UCB
BOULDER, CO 80309-0526
UNITED STATES
kathleen.finlinson@colorado.edu

JESSICA S. PURCELL
DEPARTMENT OF MATHEMATICS
BRIGHAM YOUNG UNIVERSITY
292 TMCB
PROVO, UT 84602-6539
UNITED STATES
and
SCHOOL OF MATHEMATICAL SCIENCES
MONASH UNIVERSITY
9 RAINFOREST WALK, ROOM 401
MONASH UNIVERSITY, VIC 3800
AUSTRALIA
jessica.purcell@monash.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 282, No. 1, 2016
dx.doi.org/10.2140/pjm.2016.282.107

MINIMAL SURFACES WITH TWO ENDS WHICH HAVE THE LEAST TOTAL ABSOLUTE CURVATURE

SHOICHI FUJIMORI AND TOSHIHIRO SHODA

Dedicated to Professor Reiko Miyaoka on the occasion of her retirement

We consider complete noncatenoidal minimal surfaces of finite total curvature with two ends. A family of such minimal surfaces with least total absolute curvature is given. We also obtain a uniqueness theorem for this family from its symmetries.

1. Introduction

For a complete minimal surface in Euclidean space, an inequality stronger than the classical inequality of Cohn-Vossen holds, giving a lower bound for the total absolute curvature. It is then natural to ask whether there is a minimal surface which attains this minimum value for the total absolute curvature. We consider this problem and contribute to the theory of existence of minimal surfaces in Euclidean space. Our work connects with the Björling problem for minimal surfaces in Euclidean space.

Let $f : M \to \mathbb{R}^3$ be a minimal immersion of a 2-manifold M into Euclidean 3-space $\mathbb{R}^3$. We usually call f a *minimal surface* in $\mathbb{R}^3$. Choosing isothermal coordinates makes M a Riemann surface, and then f is called a *conformal minimal immersion*. The following representation formula is one of the basic tools in the theory of minimal surfaces.

Theorem 1.1 (Weierstrass representation [Osserman 1964]). *Let (g, η) be a pair of a meromorphic function g and a holomorphic differential η on a Riemann surface M such that*

$$(1\text{-}1) \qquad (1+|g|^2)^2\eta\bar{\eta}$$

Fujimori partially supported by JSPS Grant-in-Aid for Young Scientists (B) 25800047.
Shoda partially supported by JSPS Grant-in-Aid for Young Scientists (B) 24740047.
MSC2010: primary 53A10; secondary 49Q05, 53C42.
Keywords: minimal surface, finite total curvature, two ends.

gives a Riemannian metric on M. We set

$$\Phi := \begin{pmatrix} (1-g^2)\eta \\ i(1+g^2)\eta \\ 2g\eta \end{pmatrix}, \tag{1-2}$$

where $i=\sqrt{-1}$. Assume that

$$\operatorname{Re}\oint_\ell \Phi = \mathbf{0} \quad \textit{holds for any } \ell \in \pi_1(M). \tag{P}$$

Then

$$f = \operatorname{Re}\int_{z_0}^{z} \Phi : M \to \mathbb{R}^3 \quad (z_0 \in M) \tag{1-3}$$

defines a conformal minimal immersion.

The pair (g, η) in Theorem 1.1 is called the *Weierstrass data* of f.

Remark 1.2. The *period condition* (P) of the minimal surface is equivalent to

$$\oint_\ell \eta = \overline{\oint_\ell g^2\eta} \tag{1-4}$$

and

$$\operatorname{Re}\oint_\ell g\eta = 0 \tag{1-5}$$

for any $\ell \in \pi_1(M)$.

Remark 1.3. The first fundamental form ds^2 and the second fundamental form $\mathbb{I}$ of the surface (1-3) are given by

$$ds^2 = (1+|g|^2)^2\eta\bar{\eta}, \qquad \mathbb{I} = -\eta dg - \overline{\eta dg}.$$

Moreover, $g : M \to \mathbb{C} \cup \{\infty\}$ coincides with the composition of the Gauss map $G : M \to S^2$ of the minimal surface and stereographic projection $\sigma : S^2 \to \mathbb{C}\cup\{\infty\}$, that is, $g = \sigma \circ G$. So we call g the Gauss map of the minimal surface.

Next, we assume that a minimal surface is complete and of finite total curvature. These two conditions give rise to restrictions on the topological and conformal types of minimal surfaces.

Theorem 1.4 [Huber 1957; Osserman 1964]. *Let $f : M \to \mathbb{R}^3$ be a conformal minimal immersion. Suppose that f is complete and of finite total curvature.*

(1) *M is conformally equivalent to a compact Riemann surface $\overline{M}_\gamma$ of genus γ punctured at a finite number of points $p_1, \dots, p_n$.*

(2) *The Gauss map g extends to a holomorphic mapping $\hat{g} : \overline{M}_\gamma \to \mathbb{C}\cup\{\infty\}$.*

The removed points $p_1, \dots, p_n$ correspond to ends of the minimal surface.

The asymptotic behavior around each end p_i can be described by the order of the poles of $\Phi = (\Phi_1, \Phi_2, \Phi_3)$ in Theorem 1.1 at p_i. Let

$$d_i = \max_{1\le j\le 3}\{\operatorname{ord}(\Phi_j, p_i)\} - 1, \tag{1-6}$$

where $\operatorname{ord}(\Phi_j, p_i)$ is the order of the pole of Φ_j at p_i ($1 \le i \le n$, $1 \le j \le 3$). Condition (P) yields $\operatorname{residue}(\Phi, p_i) \in \mathbb{R}^3$, and thus $d_i \ge 1$. The following theorem shows the geometric properties of d_i, which includes a stronger inequality than the Cohn-Vossen inequality.

Theorem 1.5 [Osserman 1964; Jorge and Meeks 1983; Schoen 1983]. *Let the immersion $f : M \to \mathbb{R}^3$ be a minimal surface as in Theorem 1.4.*

(a) *The immersion f is proper.*

(b) *If $S^2(r)$ is the sphere of radius r, then $\frac{1}{r}(f(M) \cap S^2(r))$ consists of n closed curves $\Gamma_1, \dots, \Gamma_n$ in $S^2(1)$ which converge C^1 to closed geodesics $\gamma_1, \dots, \gamma_n$ of $S^2(1)$, with multiplicities $d_1, \dots, d_n$, as $r \to \infty$. Moreover,*

$$\begin{aligned}\frac{1}{2\pi}\int_M K\,dA &= \chi(\overline{M}_\gamma) - \sum_{i=1}^{n}(d_i+1) \\ &\le \chi(M) - n = \chi(\overline{M}_\gamma) - 2n = 2(1-\gamma-n),\end{aligned} \tag{1-7}$$

and equality holds if and only if each end is embedded.

The equation in the first line of (1-7) is called the *Jorge–Meeks formula*.

Moreover, a relation between the total (absolute) curvature and the degree of g is as follows. Note that since g extends to a holomorphic map $\hat{g}$ from a compact Riemann surface $\overline{M}_\gamma$ to a compact Riemann surface $\mathbb{C} \cup \{\infty\}$, we can define the degree of g by $\deg(g) := \deg(\hat{g})$. Since the Gaussian curvature of a minimal surface $M \to \mathbb{R}^3$ is always nonpositive, its total absolute curvature $\tau(M) := \int_M |K|\,dA$ is given by

$$\tau(M) = \int_M (-K)\,dA.$$

Recall that the total absolute curvature of a minimal surface in $\mathbb{R}^3$ is just the area under the Gauss map $g : M \to \mathbb{C} \cup \{\infty\} \cong S^2$, that is,

$$\tau(M) = (\text{the area of } S^2)\deg(g) = 4\pi\deg(g) \in 4\pi\mathbb{Z}.$$

(See, for example, (3.11) in [Hoffman and Osserman 1980] for details.) Hence (1-7) is rewritten as

$$\deg(g) \ge \gamma + n - 1 \tag{1-8}$$

and we consider sharpness of the inequality (1-8).

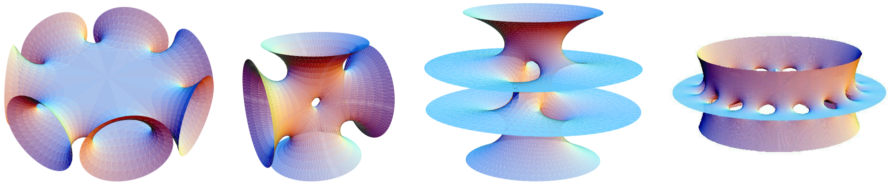


Figure 1. Minimal surfaces with $n \geq 3$ satisfying equality in (1-8). For details on these surfaces, see [Jorge and Meeks 1983; Berglund and Rossman 1995; Wohlgemuth 1997; Hoffman and Meeks 1990], for instance.

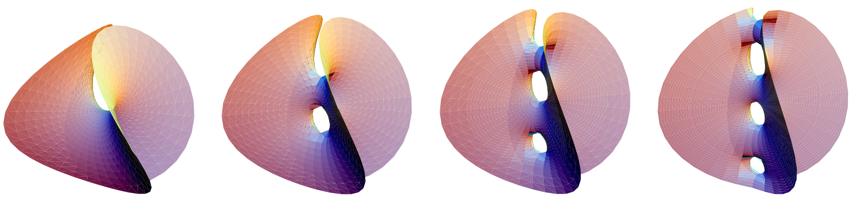


Figure 2. Minimal surfaces with $n = 1$ satisfying $\deg(g) = \gamma + 1$.

For $n \geq 3$, there exist many examples of minimal surfaces which satisfy equality in (1-8), $\deg(g) = \gamma + n - 1$. (See Figure 1.)

If $n = 1$, then a minimal surface satisfying $\deg(g) = \gamma$ must be a plane. (See [Hoffman and Karcher 1997, Remark 2.2], for instance.) Thus on a nonplanar minimal surface, $\deg(g) \geq \gamma + 1$. The existence of minimal surfaces with $\deg(g) = \gamma + 1$ was shown by C. C. Chen and F. Gackstatter [1982] (for $\gamma = 1, 2$), N. do Espírito Santo [1994] (for $\gamma = 3$), K. Sato [1996], and M. Weber and M. Wolf [1998]. (See Figure 2.)

Finally, we consider the case $n = 2$. In this case, the following uniqueness result is known.

Theorem 1.6 [Schoen 1983]. *Let $f : M \to \mathbb{R}^3$ be a complete conformal minimal surface of finite total curvature. If f has two ends and equality holds in* (1-8), *then f must be a catenoid.*

It follows that on a noncatenoidal minimal surface with two ends,

$$\deg(g) \geq \gamma + 2. \tag{1-9}$$

As a consequence, it is reasonable to consider:

Problem 1.7. For an arbitrary genus γ, does there exist a complete conformal minimal surface of finite total curvature with two ends satisfying equality in (1-9)?

In the case $\gamma = 0$, such minimal surfaces exist, and moreover, these minimal surfaces have been classified by F. J. López [1992]. (See Figure 3.) However, for the case $\gamma > 0$, no answer to Problem 1.7 is known. Our first main result is to give a partial answer to this problem.

Main Theorem 1. *If γ is equal to 1 or an even number, there exists a complete conformal minimal surface of finite total curvature with two ends which satisfies equality in* (1-9).

Note that if we do not assume the equality in (1-9), then there exists a complete conformal minimal surface of finite total curvature with two ends for an arbitrary genus $\gamma \geq 0$. (See [Fujimori and Shoda 2014], for instance.)

We prove Main Theorem 1 by explicit constructions in Section 2. We now discuss the asymptotic behavior for our minimal surfaces in terms of d_i. For a minimal surface as in Problem 1.7, we have $(d_1, d_2) = (1, 3)$ or $(2, 2)$. The case $(d_1, d_2) = (1, 3)$ corresponds to a minimal surface with an embedded end and an Enneper type end. Recall that an embedded end is asymptotic to a plane or a catenoid (see [Schoen 1983]). The minimal surface given in Section 2A has an embedded end which is asymptotic to a plane and $(d_1, d_2) = (1, 3)$ (see Corollary 2.3). The minimal surface introduced in Section 4A is another example with an embedded end which is asymptotic to a half catenoid and $(d_1, d_2) = (1, 3)$. The minimal surfaces with $(d_1, d_2) = (2, 2)$ are obtained in Section 2B (see Corollary 2.6).

The minimal surfaces given in Corollary 2.3 and Corollary 2.6 have symmetry groups with $4(\gamma + 1)$ elements. Next we consider the uniqueness theorem for the symmetries. Uniqueness is also one of the important problems for minimal surfaces, and there are many uniqueness theorems (see [Martín and Weber 2001; Hoffman and Meeks 1990]). Our other main theorem is as follows.

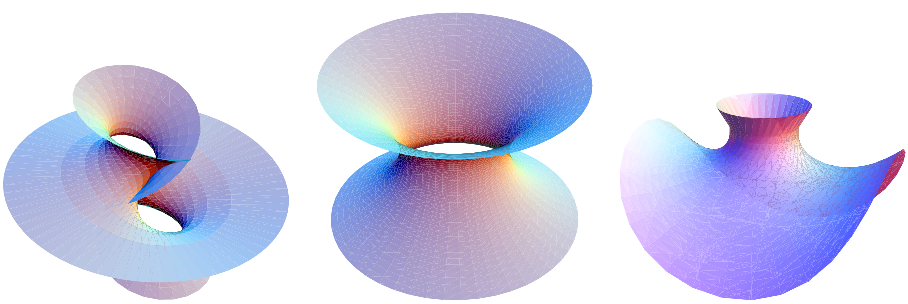


Figure 3. Examples for $\gamma = 0$. The surface in the middle is a double cover of a catenoid.

Main Theorem 2. *Let $f : M \to \mathbb{R}^3$ be a complete conformal minimal surface of finite total curvature with two ends and genus γ. Suppose that f satisfies equality in (1-9) and has $4(\gamma+1)$ symmetries. We assume either $\gamma = 1$ and $(d_1, d_2) = (1, 3)$, or γ is an even number and $(d_1, d_2) = (2, 2)$. Then f is one of the minimal surfaces given in Main Theorem 1.*

At the end of this section, we discuss our work from the point of view of the Björling problem for minimal surfaces. The classical Björling problem is to determine a piece of a minimal surface containing a given analytic strip. This was named after E. G. Björling in 1844. H. A. Schwarz gave an explicit solution to it. (See [Nitsche 1975], for instance.) Recently, Mira [2006] used the solution to the Björling problem to classify a certain class of minimal surfaces of genus 1. Also, Meeks and Weber [2007] produced an infinite sequence of complete minimal annuli by using the solution to the Björling problem and then gave a complete answer as to which curves appear as the singular set of a Colding–Minicozzi limit minimal lamination. Hence it is useful to study minimal surfaces from the point of view of the Björling problem. However, the existence of minimal surfaces of higher genus derived from the solution to the Björling problem seems to be unknown. In Section 2B, we show that our minimal surfaces, which have even numbers for the genus, are solutions to the Björling problem, and the generating curves are closed plane curves.

The paper is organized as follows: Section 2 contains constructions of concrete examples to prove Main Theorem 1, with the genus 1 case provided in Section 2A, and the even genus case in Section 2B. Section 2B also contains the result from the point of view of the Björling problem. We prove our uniqueness result in Section 3. In Section 4 we discuss remaining problems related to our work.

2. Construction of surfaces for Main Theorem 1

In this section we will construct the surfaces for proving Main Theorem 1. We will use the Weierstrass representation in Theorem 1.1, for which we need a Riemann surface M, a meromorphic function g, and a holomorphic differential η.

2A. ***The case $\gamma = 1$.*** Let $\overline{M}_\gamma$ be the Riemann surface

$$\overline{M}_\gamma = \left\{(z, w) \in (\mathbb{C} \cup \{\infty\})^2 \mid w^{\gamma+1} = z(z^2-1)^\gamma\right\}.$$

The surface we will consider is

$$M = \overline{M}_\gamma \setminus \{(0, 0), (\infty, \infty)\},$$

a Riemann surface of genus γ from which two points have been removed. We want to define a complete conformal minimal immersion of M into $\mathbb{R}^3$ by the Weierstrass

(z, w)	$(0, 0)$	$(1, 0)$	$(-1, 0)$	(∞, ∞)
g	0^1	0^γ	0^γ	$\infty^{2\gamma+1}$
η	$\infty^{\gamma+3}$			$0^{3\gamma+1}$

Table 1. Orders of zeros and poles of g and η.

representation in Theorem 1.1. To do this, set

$$g = cw, \qquad \eta = i\frac{dz}{z^2 w},$$

where $c \in \mathbb{R}_{>0}$ is a positive constant to be determined.

Let Φ be the $\mathbb{C}^3$-valued differential as in (1-2). We shall prove that (1-3) is a conformal minimal immersion of M.

We begin by showing by straightforward calculation how the following conformal diffeomorphisms κ_1 and κ_2 act on Φ.

Lemma 2.1 (symmetries of the surface). *Consider the conformal mappings*

$$\kappa_1(z, w) = (\bar{z}, \overline{w}), \qquad \kappa_2(z, w) = \bigl(-z, e^{\pi i/(\gamma+1)} w\bigr)$$

of M. Then

$$\kappa_1^*\Phi = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \overline{\Phi}, \qquad \kappa_2^*\Phi = \begin{pmatrix} -\cos\frac{\pi}{\gamma+1} & \sin\frac{\pi}{\gamma+1} & 0 \\ -\sin\frac{\pi}{\gamma+1} & -\cos\frac{\pi}{\gamma+1} & 0 \\ 0 & 0 & -1 \end{pmatrix} \Phi.$$

Since (1-1) gives a complete Riemannian metric on M (see Table 1), it suffices to show that f is well-defined on M for the right choice of c.

Theorem 2.2. *For any positive number γ, there exists a unique positive constant $c \in \mathbb{R}_{>0}$ for which the immersion f given in (1-3) is well-defined on M.*

Proof. To establish this theorem we must show (P) in Theorem 1.1. We will prove (1-4) and (1-5). But (1-5) follows from the exactness of $g\eta = ic\,dz/z^2 = d(-ic/z)$, and thus we will only have to show (1-4). We first check the residues of η and $g^2\eta$ at the ends $(0, 0)$, (∞, ∞). At the end $(0, 0)$, w is a local coordinate for the Riemann surface $\overline{M}_\gamma$, and then $z = z(w) = w^{\gamma+1}\{(-1)^\gamma + \mathcal{O}(w^{2\gamma+2})\}$. We have

$$\eta = \left(\frac{\alpha_1}{w^{\gamma+3}} + \mathcal{O}(w^{\gamma-1})\right) dw \quad \text{and} \quad g^2\eta = \left(\frac{\alpha_2}{w^{\gamma+1}} + \mathcal{O}(w^{\gamma+1})\right) dw,$$

where $\alpha_j \in \mathbb{C}$ (for $j = 1, 2$) are constants. These imply that both η and $g^2\eta$ have no residues at $(0, 0)$. Then the residue theorem yields that they have no residues at (∞, ∞) as well.

We next consider path-integrals along topological 1-cycles on $\overline{M}_\gamma$. We will give a convenient 1-cycle.

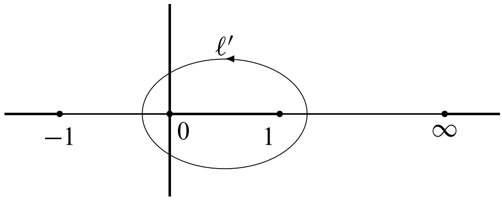


Figure 4. Projection to the z-plane of the loop $\ell' \in \pi_1(M)$.

Define a 1-cycle on $\overline{M}_\gamma$ by

$$\ell = \left\{(z, w) = \left(-t,\ \sqrt[\gamma+1]{-t(1-t^2)^\gamma}\, e^{\gamma\pi i/(\gamma+1)}\right) \mid -1 \le t \le 0\right\} \\ \cup \left\{(z, w) = \left(t,\ \sqrt[\gamma+1]{t(1-t^2)^\gamma}\, e^{-\gamma\pi i/(\gamma+1)}\right) \mid 0 \le t \le 1\right\}.$$

Recall that $(0, 0)$ corresponds to the end of f. Avoiding the end $(0, 0)$, we can deform ℓ to a 1-cycle ℓ' on M which is projected to a loop winding once around $[0, 1]$ in the z-plane. (See Figure 4.)

By the actions of the κ_j's, we can obtain all of the 1-cycles on M from ℓ'. If (P) holds for this ℓ', then

$$\operatorname{Re}\int_{\kappa_j\circ\ell'}\Phi = \operatorname{Re}\int_{\ell'}\kappa_j^*\Phi = K\operatorname{Re}\int_{\ell'}\Phi = \mathbf{0}$$

for some orthogonal matrix K, by Lemma 2.1. Hence all that remains to be done is to show that (1-4) holds for ℓ'.

We now calculate path-integrals of η and $g^2\eta$ along ℓ', and we want to reduce them to path-integrals along ℓ for simplicity. Note that both η and $g^2\eta$ have poles at $(0, 0)$. To avoid divergent integrals, here we add exact 1-forms which have principal parts of η and $g^2\eta$, respectively. It is straightforward to check

$$\frac{dz}{z^2w} - \frac{\gamma+1}{\gamma+2}\, d\left(\frac{z^2-1}{zw}\right) = \frac{\gamma}{\gamma+2}\frac{dz}{w}, \qquad \frac{w}{z^2}\, dz + \frac{\gamma+1}{\gamma}\, d\left(\frac{w}{z}\right) = \frac{2w}{z^2-1}\, dz.$$

So we have

$$\oint_{\ell'}\eta = \frac{i\gamma}{\gamma+2}\oint_{\ell'}\frac{dz}{w} = \frac{i\gamma}{\gamma+2}\oint_{\ell}\frac{dz}{w} = \frac{-2\gamma}{\gamma+2}\sin\frac{\gamma\pi}{\gamma+1}\int_0^1\frac{dt}{\sqrt[\gamma+1]{t(1-t^2)^\gamma}},$$

$$\oint_{\ell'} g^2\eta = 2ic^2\oint_{\ell'}\frac{w}{z^2-1}\, dz = 2ic^2\oint_{\ell}\frac{w}{z^2-1}\, dz = -4c^2\sin\frac{\gamma\pi}{\gamma+1}\int_0^1\sqrt[\gamma+1]{\frac{t}{1-t^2}}\, dt.$$

By setting

$$A_\gamma = \frac{\gamma}{\gamma+2}\int_0^1\frac{dt}{\sqrt[\gamma+1]{t(1-t^2)^\gamma}} \in \mathbb{R}_{>0}, \qquad B_\gamma = 2\int_0^1\sqrt[\gamma+1]{\frac{t}{1-t^2}}\, dt \in \mathbb{R}_{>0},$$

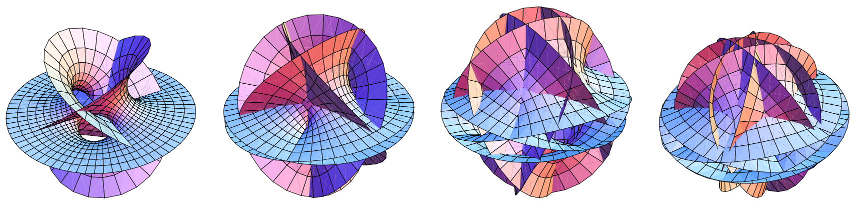


Figure 5. Minimal surfaces of genus γ with two ends which satisfy $\deg(g) = 2\gamma + 1$.

the equation (1-4) is reduced to $A_\gamma = c^2 B_\gamma$. Let us set

$$c = \sqrt{\frac{A_\gamma}{B_\gamma}} \in \mathbb{R}_{>0}.$$

This choice of c satisfies (1-4) and is the unique positive real number that does so. This completes the proof. (See Figure 5.) □

Since $\deg(g) = 2\gamma + 1$, $\deg(g) = \gamma + 2$ if and only if $\gamma = 1$. As a consequence, the next corollary follows:

Corollary 2.3. *There exists a complete conformal minimal surface of genus* 1 *with two ends which has least total absolute curvature.*

2B. ***The case γ is even.*** The following construction is similar to the construction in Section 2A. Crucial arguments are given after (2-3).

For an integer $k \geq 2$, let $\overline{M}_\gamma$ be the Riemann surface

$$\overline{M}_\gamma = \left\{ (z, w) \in (\mathbb{C} \cup \{\infty\})^2 \;\middle|\; w^{k+1} = z^2 \left(\frac{z-1}{z-a} \right)^k \right\},$$

where $a \in (1, \infty)$ is a constant to be determined. By the Riemann–Hurwitz formula, we see that

$$\gamma = \begin{cases} k & \text{if } k \text{ is even,} \\ k-1 & \text{if } k \text{ is odd.} \end{cases}$$

Note that the genus γ is always even. (See Figure 6.)

We set

$$\begin{aligned} M &= \overline{M}_\gamma \setminus \{(0, 0), (\infty, \infty)\}, \\ g &= cw \quad \text{for } c = a^{(k-2)/(2k+2)} \in \mathbb{R}_{>0}, \\ \eta &= \frac{dz}{zw}. \end{aligned}$$

Then (1-1) gives a complete Riemannian metric M. (See Table 2.)

Also, $\deg(g) = k + 2$ for all $k \geq 2$. Thus equality in (1-9) holds if and only if k is even. Hereafter we assume k is even.

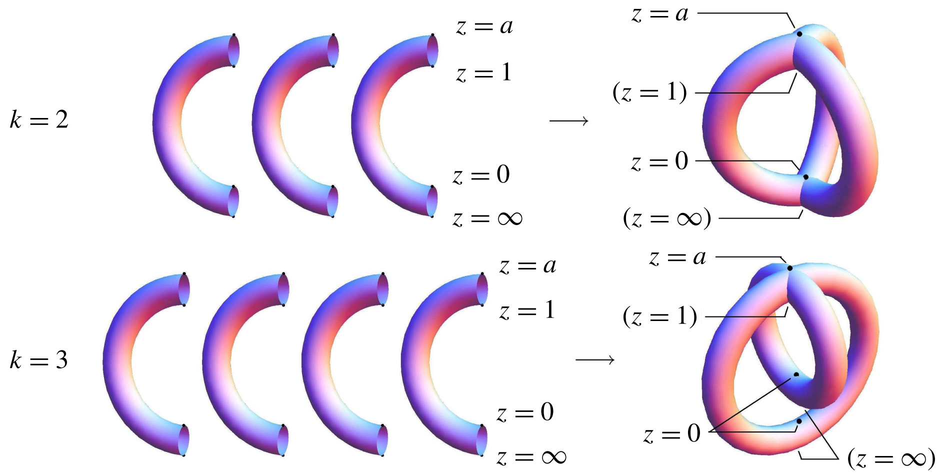


Figure 6. Riemann surfaces $\overline{M}_\gamma$ for $k = 2$ (top) and $k = 3$ (bottom). Both surfaces have genus 2. When k is odd, $\overline{M}_\gamma$ has self-intersections at $z = 0$ and $z = \infty$. In the sketch in the bottom row, we see two different $z = 0$ (and two different $z = \infty$, which are hidden from this viewpoint) but they are in fact the same points. The reason we place these points differently is to reveal their genus clearly.

(z, w)	$(0, 0)$	$(1, 0)$	(a, ∞)	(∞, ∞)
g	0^2	0^k	∞^k	∞^2
η	∞^4		0^{2k}	

(z, w)	$(0, 0)$	$(1, 0)$	(a, ∞)	(∞, ∞)
g	0^2	0^k	∞^k	∞^2
η	∞^3		0^{2k}	0^1

Table 2. Orders of zeros and poles of g and η when k is odd (top) and k is even (bottom).

Let Φ be the $\mathbb{C}^3$-valued differential as in (1-2). We shall prove that (1-3) is a conformal minimal immersion of M.

First, we observe the following symmetries κ_1, κ_2, and κ_3 of the surface.

Lemma 2.4 (Symmetries of the surface). *Consider the conformal mappings*

$$\kappa_1(z, w) = (\bar{z}, \overline{w}), \qquad \kappa_2(z, w) = \left(z, e^{2\pi i/(k+1)} w\right), \qquad \kappa_3(z, w) = \left(\frac{a}{z}, \frac{1}{c^2 w}\right)$$

of M. Then

$$\kappa_1^*\Phi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \overline{\Phi}, \qquad \kappa_2^*\Phi = \begin{pmatrix} \cos\frac{2\pi}{k+1} & -\sin\frac{2\pi}{k+1} & 0 \\ \sin\frac{2\pi}{k+1} & \cos\frac{2\pi}{k+1} & 0 \\ 0 & 0 & 1 \end{pmatrix} \Phi,$$

$$\kappa_3^*\Phi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \Phi.$$

As we have already seen the completeness of f, it suffices to show that f is well-defined on M for the right choice of $a \in (1, \infty)$.

Theorem 2.5. *For any positive even number k, there exists a unique constant $a \in (1, \infty)$ for which the immersion f given in (1-3) is well-defined on M.*

Proof. We will show (P) in Theorem 1.1. It is easy to verify that there are no residues at the ends $(0, 0)$ and (∞, ∞). So all that remains is to choose c so that (P) is satisfied. The equation (1-5) follows from the exactness of

$$g\eta = \frac{c}{z}\, dz = c \cdot d(\log z)$$

and $c \in \mathbb{R}$, and hence we will only have to show (1-4). To do this, we will give convenient 1-cycles.

Define a 1-cycle on $\overline{M}_\gamma$ as

$$\ell_1 = \left\{ (z, w) = \left(-t,\ \sqrt[k+1]{t^2\left(\frac{1+t}{a+t}\right)^k} \right) \,\middle|\, -1 \le t \le 0 \right\} \\ \cup \left\{ (z, w) = \left(t,\ \sqrt[k+1]{t^2\left(\frac{1-t}{a-t}\right)^k}\, e^{2\pi i/(k+1)} \right) \,\middle|\, 0 \le t \le 1 \right\}.$$

Recall that $(0, 0)$ corresponds to the end of f. Avoiding the end $(0, 0)$, we can deform ℓ_1 to a 1-cycle ℓ_1' on M which is projected to a loop winding once around $[0, 1]$ in the z-plane. We also define another 1-cycle on M as

$$\ell_2 = \left\{ (z, w) = \left(-t,\ \sqrt[k+1]{t^2\left(\frac{-t-1}{a+t}\right)^k}\, e^{k\pi i/(k+1)} \right) \,\middle|\, -a \le t \le -1 \right\} \\ \cup \left\{ (z, w) = \left(t,\ \sqrt[k+1]{t^2\left(\frac{t-1}{a-t}\right)^k}\, e^{-k\pi i/(k+1)} \right) \,\middle|\, 1 \le t \le a \right\}.$$

(See Figure 7.)

Again by the actions of the κ_j's, we can obtain all of the 1-cycles on M from ℓ_1' and ℓ_2. We now show that (1-4) holds for ℓ_1' and ℓ_2.

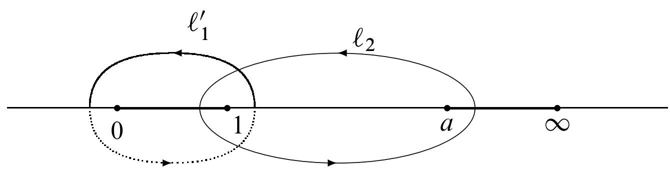


Figure 7. Projections to the z-plane of the loops ℓ_1' and $\ell_2 \in \pi_1(M)$.

First we calculate the path-integrals of η and $g^2\eta$ along ℓ_2. Then we have

$$\oint_{\ell_2} \eta = 2i \sin \frac{k\pi}{k+1} \int_1^a \sqrt[k+1]{\frac{(a-t)^k}{t^{k+3}(t-1)^k}}\, dt, \tag{2-1}$$

$$\begin{aligned}\oint_{\ell_2} g^2\eta &= -2ic^2 \sin \frac{k\pi}{k+1} \int_1^a \sqrt[k+1]{\frac{(t-1)^k}{t^{k-1}(a-t)^k}}\, dt \\ &= -2i \sin \frac{k\pi}{k+1} \int_1^a \sqrt[k+1]{\frac{(a-\tau)^k}{\tau^{k+3}(\tau-1)^k}}\, d\tau,\end{aligned} \tag{2-2}$$

where $\tau = a/t$. As a result, (1-4) holds for ℓ_2.

Next we calculate the path-integrals of η and $g^2\eta$ along ℓ_1', and we want to reduce them to the path-integrals along ℓ_1. Note that η has a pole at $(0, 0)$. To avoid a divergent integral, here we add an exact 1-form which has the principal part of η. It is straightforward to check

$$\eta - \frac{k+1}{2} d\Big(\frac{z-1}{w}\Big) = -\frac{k}{2} \frac{z-1}{w(z-a)}\, dz + \frac{1}{2} \frac{dz}{w}.$$

Thus we have

$$\begin{aligned}\oint_{\ell_1'} \eta &= \oint_{\ell_1'} \Big(-\frac{k}{2} \frac{z-1}{w(z-a)}\, dz + \frac{1}{2} \frac{dz}{w}\Big) \\ &= \oint_{\ell_1} \Big(-\frac{k}{2} \frac{z-1}{w(z-a)}\, dz + \frac{1}{2} \frac{dz}{w}\Big) = ie^{-\pi i/(k+1)} \sin \frac{\pi}{k+1}(kA_1 - A_2),\end{aligned}$$

where

$$A_1 = \int_0^1 \frac{(1-t)^{1/(k+1)}}{t^{2/(k+1)}(a-t)^{1/(k+1)}}\, dt,$$
$$A_2 = \int_0^1 \frac{(a-t)^{k/(k+1)}}{t^{2/(k+1)}(1-t)^{k/(k+1)}}\, dt.$$

Also, we have

$$\oint_{\ell_1'} g^2\eta = \oint_{\ell_1} g^2\eta = 2ie^{\pi i/(k+1)} \sin \frac{\pi}{k+1} a^{(k-2)/(k+1)} A_3,$$

where

$$A_3 = \int_0^1 \frac{(1-t)^{k/(k+1)}}{t^{(k-1)/(k+1)}(a-t)^{k/(k+1)}}\,dt.$$

Hence for the loop $\ell_1' \in \pi_1(M)$, (1-4) holds if and only if

$$kA_1 + 2a^{(k-2)/(k+1)}A_3 - A_2 = 0. \tag{2-3}$$

Now we evaluate the values A_1, A_2, and A_3. Since $1/a \le 1/(a-t) \le 1/(a-1)$, we see that

$$\frac{1}{a^{1/(k+1)}}B\Big(\frac{k-1}{k+1}, \frac{k+2}{k+1}\Big) \le A_1 \le \frac{1}{(a-1)^{1/(k+1)}}B\Big(\frac{k-1}{k+1}, \frac{k+2}{k+1}\Big),$$
$$\frac{1}{a^{k/(k+1)}}B\Big(\frac{2}{k+1}, \frac{2k+1}{k+1}\Big) \le A_3 \le \frac{1}{(a-1)^{k/(k+1)}}B\Big(\frac{2}{k+1}, \frac{2k+1}{k+1}\Big),$$

where $B(x, y)$ is the beta function defined by

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt \quad \text{for } \operatorname{Re} x > 0 \text{ and } \operatorname{Re} y > 0.$$

Also, since $a-1 \le a-t \le a$, we have

$$(a-1)^{k/(k+1)}B\Big(\frac{k-1}{k+1}, \frac{1}{k+1}\Big) \le A_2 \le a^{k/(k+1)}B\Big(\frac{k-1}{k+1}, \frac{1}{k+1}\Big).$$

It follows that for the case $a \to \infty$, we have $A_1 \to 0$, $a^{(k-2)/(k+1)}A_3 \to 0$, and $A_2 \to \infty$. As a result, the left-hand side of (2-3) is negative. On the other hand, for the case $a \to 1$, we have

$$\begin{aligned}
&kA_1 + 2a^{(k-2)/(k+1)}A_3 - A_2 \\
&\ge \frac{k}{a^{1/(k+1)}}B\Big(\frac{k-1}{k+1}, \frac{k+2}{k+1}\Big) + \frac{2}{a^{2/(k+1)}}B\Big(\frac{2}{k+1}, \frac{2k+1}{k+1}\Big) - a^{k/(k+1)}B\Big(\frac{k-1}{k+1}, \frac{1}{k+1}\Big) \\
&= \frac{1}{a^{1/(k+1)}}B\Big(\frac{k-1}{k+1}, \frac{1}{k+1}\Big) + \frac{2}{a^{2/(k+1)}}B\Big(\frac{2}{k+1}, \frac{2k+1}{k+1}\Big) - a^{k/(k+1)}B\Big(\frac{k-1}{k+1}, \frac{1}{k+1}\Big) \\
&\underset{a\to 1}{\longrightarrow} 2B\Big(\frac{2}{k+1}, \frac{2k+1}{k+1}\Big) > 0,
\end{aligned}$$

and here we use formula

$$B(x, y+1) = \frac{y}{x+y}B(x, y)$$

for the beta function. So, the left-hand side of (2-3) is positive.

Thus, the intermediate value theorem yields that there exists $a \in (1, \infty)$ which satisfies (2-3). Moreover, since each of A_1, $a^{(k-2)/(k+1)}A_3$, and $-A_2$ is a monotone decreasing function with respect to a, the left-hand side of (2-3) is a monotone decreasing function as well. This proves the uniqueness. (See Figure 8). $\square$

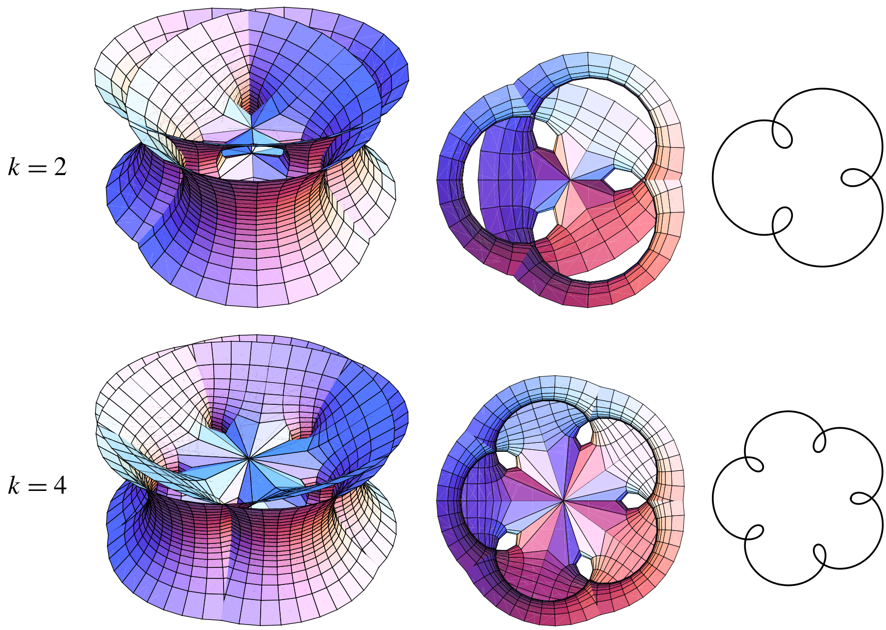


Figure 8. Minimal surfaces of genus k with two ends which satisfy $\deg(g)=k+2$. The middle columns show a half cut away from the surfaces by the xy-plane. The right columns show their intersection with the xy-plane.

Since $\deg(g) = \gamma + 2$, the next corollary follows.

Corollary 2.6. *For all even numbers γ, there exists a complete conformal minimal surface of genus γ with two ends which has least total absolute curvature.*

Combining Corollaries 2.3 and 2.6 proves Main Theorem 1.

Next we discuss the above minimal surfaces from the point of view of the Björling problem. As we mentioned in the introduction, there is a construction method for minimal surfaces from a given curve. We show that every minimal surface given in this subsection gives a solution to the Björling problem in the higher genus case and the generating curve is a closed planar geodesic.

Let l be a fixed point set of $\kappa_3 \circ \kappa_1$. Using (1-1), we see that $\kappa_3 \circ \kappa_1$ is an isometry, and thus l is a geodesic. An explicit description of l is given by

$$\frac{a}{\bar{z}} = z, \qquad \frac{1}{c^2\,\overline{w}} = w,$$

that is,

$$|z| = \sqrt{a}, \qquad |w| = \frac{1}{c}.$$

Hence we conclude that l is a closed geodesic. Moreover, by Lemma 2.4, l lies in the xy-plane, and therefore the assertion follows.

3. Uniqueness

In this section, we will prove Main Theorem 2 through four subsections.

3A. ***Symmetry.*** First, we refer to some basic results about symmetries of a minimal surface. (See p. 349 in [López and Martín 1999].)

Let $f : M \to \mathbb{R}^3$ be a conformal minimal immersion, with (g, η) its Weierstrass data. Suppose that $A : M \to M$ is a diffeomorphism. A is said to be a *symmetry* if there exists $O \in \mathcal{O}(3, \mathbb{R})$ and $v \in \mathbb{R}^3$ such that

$$(f \circ A)(p) = Of(p) + v.$$

Denote by $\mathrm{Sym}(M)$ the group of symmetries of M, and by $\mathrm{Iso}(M)$ the isometry group of M. Then, by definition, $\mathrm{Sym}(M)$ is a subgroup of $\mathrm{Iso}(M)$. Let $\mathrm{L}(M)$ be the group of holomorphic and antiholomorphic diffeomorphisms α of M satisfying

$$G \circ \alpha(p) = O \circ G(p),$$

where $G : M \to S^2$ is the Gauss map and $O \in \mathcal{O}(3, \mathbb{R})$ is a linear isometry of $\mathbb{R}^3$. We now assume that f is complete, and of finite total curvature. López and Martín pointed out that if one of the three differentials $(1 - g^2)\eta$, $i(1 + g^2)\eta$, and $2g\eta$ is not exact, then

$$\mathrm{L}(M) = \mathrm{Iso}(M) = \mathrm{Sym}(M).$$

Suppose that f has two ends. By Theorem 1.4, there exists a compact Riemann surface $\overline{M}_\gamma$ of genus γ and two points $p_1, p_2 \in \overline{M}_\gamma$ such that M is conformally equivalent to $\overline{M}_\gamma - \{p_1, p_2\}$. A symmetry of $f(M)$ extends to $\overline{M}_\gamma$ leaving the set $\{p_1, p_2\}$ invariant. By the Hurwitz theorem, the group $\mathrm{Sym}(M)$ is finite, and so up to a suitable choice of the origin, $\mathrm{Sym}(M)$ is a finite group Δ of orthogonal linear transformations of $\mathbb{R}^3$.

We assume that $\mathrm{Sym}(M)$ has $4(\gamma + 1)$ elements (for $\gamma \geq 1$) and also that $\mathrm{L}(M) = \mathrm{Iso}(M) = \mathrm{Sym}(M)$. If there is no symmetry in Δ such that either p_1 or p_2 is fixed, then Δ has at most 2 elements by a fundamental argument in linear algebra. Hence, we may assume without loss of generality that there exists a symmetry which fixes p_1. Up to rotations, we may assume $g(p_1) = 0$, and then Δ leaves the x_3-axis invariant.

We now focus on the following two cases: the case $\gamma = 1$ with $(d_1, d_2) = (1, 3)$, and the even genus case with $(d_1, d_2) = (2, 2)$ (for the definition of d_i, see (1-6)). For the former case, every symmetry in Δ leaves p_i invariant. So we see $g(p_2) = 0$ or ∞. For the latter case, we have $|\Delta| \geq 12$, and then there exist at least two symmetries which leave p_i invariant. Hence $g(p_2) = 0$ or ∞.

Let Δ_0 be the subgroup of holomorphic transformations in Δ, and denote by $\mathcal{R} \subset \Delta_0$ the cyclic subgroup of rotations around the x_3-axis. Clearly, we obtain that

$$[\Delta : \Delta_0] \leq 2, \qquad [\Delta_0 : \mathcal{R}] \leq 2. \tag{3-1}$$

So the subgroups $\Delta_0 \subset \Delta$ and $\mathcal{R} \subset \Delta_0$ are both normal.

Let R be the rotation around the x_3-axis with the smallest positive angle in Δ_0, that is, $\mathcal{R} = \langle R \rangle$. We first consider the quotient map $\pi_{\mathcal{R}} : \overline{M}_\gamma \to \overline{M}_\gamma / \mathcal{R}$. From (3-1), we see that

$$\deg(\pi_{\mathcal{R}}) = |\mathcal{R}| \geq \gamma + 1. \tag{3-2}$$

By the Riemann–Hurwitz formula, we have

$$\begin{aligned} |\mathcal{R}|(2 - 2\gamma(\overline{M}_\gamma/\mathcal{R})) &= 2 - 2\gamma + \sum_{p \in \overline{M}_\gamma} (\mu(p) - 1) \\ &= 2 - 2\gamma + 2(|\mathcal{R}| - 1) + \sum_{p \in M} (\mu(p) - 1), \end{aligned} \tag{3-3}$$

where $\gamma(\overline{M}_\gamma/\mathcal{R})$ is the genus of $\overline{M}_\gamma/\mathcal{R}$ and $\mu(p) - 1$ is the ramification index at p. Let $q'_1, \ldots, q'_t$ be ramified values of $\pi_{\mathcal{R}}$ except for the $\pi_{\mathcal{R}}(p_i)$'s, and $m_i - 1$ the ramification index at $p \in \pi_{\mathcal{R}}^{-1}(q'_i)$. Note that $2 \leq m_i \leq |\mathcal{R}|$ and the ramification index at p_i is $|\mathcal{R}| - 1$. Combining (3-2) and (3-3) yields

$$\begin{aligned} |\mathcal{R}|(2 - 2\gamma(\overline{M}_\gamma/\mathcal{R})) &= 2 - 2\gamma + 2(|\mathcal{R}| - 1) + \sum_{i=1}^{t} (m_i - 1) \frac{|\mathcal{R}|}{m_i} \\ &\geq 2 + \sum_{i=1}^{t} (m_i - 1) \frac{|\mathcal{R}|}{m_i} > 0. \end{aligned} \tag{3-4}$$

It follows that $\gamma(\overline{M}_\gamma/\mathcal{R}) = 0$. This, combined with (3-2) and (3-4), implies

$$2\gamma = \sum_{i=1}^{t} (m_i - 1) \frac{|\mathcal{R}|}{m_i} = |\mathcal{R}| \left(\sum_{i=1}^{t} \left(\frac{1}{2} - \frac{1}{m_i} \right) + \frac{t}{2} \right) \geq (\gamma + 1) \frac{t}{2}. \tag{3-5}$$

So $t \leq 4\gamma/(\gamma + 1) < 4$, and thus $t = 1, 2, 3$. We remark that we have $|\mathcal{R}| = \gamma + 1$, $2(\gamma + 1)$, and $4(\gamma + 1)$.

<u>The case $t = 1$.</u>

From the first equality of (3-5), we obtain $2\gamma = |\mathcal{R}|\left(1 - \frac{1}{m_1}\right)$. For the case $|\mathcal{R}| = \gamma + 1$, we get $\frac{1}{m_1} = -\frac{\gamma - 1}{\gamma + 1} \leq 0$, which is absurd. Next, if $|\mathcal{R}| = 2(\gamma + 1)$ then $m_1 = \gamma + 1$ holds. Finally, $|\mathcal{R}| = 4(\gamma + 1)$ gives

$$\frac{1}{2} > \frac{\gamma}{2(\gamma + 1)} = 1 - \frac{1}{m_1} \geq 1 - \frac{1}{2} = \frac{1}{2},$$

which leads to a contradiction.

The case $t = 2$.

We obtain $2\gamma = |\mathcal{R}|\big(2 - \frac{1}{m_1} - \frac{1}{m_2}\big)$ from the first equality of (3-5). Without loss of generality, we may assume $m_1 \leq m_2$. Then

$$2 - \frac{2}{m_1} \leq \frac{2\gamma}{|\mathcal{R}|} = 2 - \frac{1}{m_1} - \frac{1}{m_2} \leq 2 - \frac{2}{m_2}.$$

For the case $|\mathcal{R}| = \gamma + 1$, the inequality $\gamma + 1 \leq m_2 \leq |\mathcal{R}| = \gamma + 1$ holds, and thus $m_2 = \gamma + 1$ and $m_1 = \gamma + 1$. Next we consider the case $|\mathcal{R}| \geq 2(\gamma + 1)$. In this case, we have

$$2 - \frac{2}{m_1} \leq \frac{2\gamma}{|\mathcal{R}|} \leq \frac{\gamma}{\gamma+1} < 1,$$

and so $m_1 < 2$, which is absurd.

The case $t = 3$.

It follows from the first equality of (3-5) that $2\gamma = |\mathcal{R}|\big(3 - \frac{1}{m_1} - \frac{1}{m_2} - \frac{1}{m_3}\big)$. We may assume $m_1 \leq m_2 \leq m_3$. Then

$$3 - \frac{3}{m_1} \leq \frac{2\gamma}{|\mathcal{R}|} = 3 - \frac{1}{m_1} - \frac{1}{m_2} - \frac{1}{m_3} \leq 3 - \frac{3}{m_3}.$$

For the case $|\mathcal{R}| = \gamma + 1$, we have $3 - \frac{3}{m_1} \leq \frac{2\gamma}{|\mathcal{R}|} = \frac{2\gamma}{\gamma+1} < 2$. So $m_1 < 3$, and hence $m_1 = 2$. As a result,

$$\frac{5}{2} - \frac{2}{m_2} \leq \frac{2\gamma}{\gamma+1} = \frac{5}{2} - \frac{1}{m_2} - \frac{1}{m_3} \leq \frac{5}{2} - \frac{2}{m_3}.$$

The inequality $\frac{5}{2} - \frac{2}{m_2} \leq \frac{2\gamma}{\gamma+1}$ gives $m_2 \leq \frac{4(\gamma+1)}{\gamma+5} < 4$. Thus $m_2 = 2$ or 3. For the case $m_2 = 2$,

$$\frac{2\gamma}{\gamma+1} = 2 - \frac{1}{m_3}$$

holds, and so $m_3 = \frac{\gamma+1}{2}$. For the case $m_2 = 3$,

$$\frac{2\gamma}{\gamma+1} = \frac{13}{6} - \frac{1}{m_3}.$$

It follows that $m_3 = \frac{6(\gamma+1)}{\gamma+13} < 6$, and hence $m_3 = 3$, 4, or 5. So $\gamma = 11$, 23, or 59. For the case $|\mathcal{R}| \geq 2(\gamma + 1)$,

$$3 - \frac{3}{m_1} \leq \frac{2\gamma}{|\mathcal{R}|} \leq \frac{\gamma}{\gamma+1} < 1.$$

So $m_1 < \frac{3}{2}$, and this contradicts $m_1 \geq 2$.

As a consequence, we obtain Tables 3, 4, and 5.

Note that, for $t = 2$, $\pi_{\mathcal{R}}$ is a cyclic branched cover of S^2, of order $\gamma + 1$, whose branch points are the fixed points of $\mathcal{R}$, namely p_1, p_2, $\pi_{\mathcal{R}}^{-1}(q_1')$, $\pi_{\mathcal{R}}^{-1}(q_2')$.

If $t = 1$, then $\pi_{\mathcal{R}}^{-1}(q_1') = \{q_1, q_2\}$ for some two points $q_1, q_2 \in \overline{M}_\gamma$. If R leaves every q_i invariant, then m_1 must be $2(\gamma + 1)$. Thus $R(q_1) = q_2$ and $R(q_2) = q_1$.

$\lvert\mathcal{R}\rvert$	m_i
$\lvert\mathcal{R}\rvert=\gamma+1$	does not occur
$\lvert\mathcal{R}\rvert=2(\gamma+1)$	$m_1=\gamma+1$
$\lvert\mathcal{R}\rvert=4(\gamma+1)$	does not occur

Table 3. The case $t=1$.

$\lvert\mathcal{R}\rvert$	m_i
$\lvert\mathcal{R}\rvert=\gamma+1$	$m_1=m_2=\gamma+1$
$\lvert\mathcal{R}\rvert\geq 2(\gamma+1)$	does not occur

Table 4. The case $t=2$.

$\lvert\mathcal{R}\rvert$	m_i			
	m_1	m_2	m_3	γ
$\lvert\mathcal{R}\rvert=\gamma+1$	2	2	$(\gamma+1)/2$	odd (>1)
	2	3	3	11
	2	3	4	23
	2	3	5	59
$\lvert\mathcal{R}\rvert\geq 2(\gamma+1)$	does not occur			

Table 5. The case $t=3$.

So $R^2(q_i)=q_i$ and $f(q_1)=f(q_2)\in\{x_3\text{-axis}\}$. Now we consider the quotient map $\pi_{\langle R^2\rangle}:\overline{M}_\gamma\to\overline{M}_\gamma/\langle R^2\rangle$. From the Riemann–Hurwitz formula,

$$\begin{aligned}|\langle R^2\rangle|\bigl(2-2\gamma(\overline{M}/\langle R^2\rangle)\bigr)&=2-2\gamma+4(|\langle R^2\rangle|-1)\\&=2\gamma+2>0.\end{aligned}$$

Hence we obtain $\gamma(\overline{M}/\langle R^2\rangle)=0$. It follows that $\pi_{\langle R^2\rangle}$ is a cyclic branched cover of S^2, of order $\gamma+1$, whose branch points are p_1,p_2,q_1,q_2. This case corresponds to the case $t=2$, and thus we can determine the case $t=1$ after we consider the case $t=2$.

Next, we consider the quotient map $\pi_{\Delta_0}:\overline{M}_\gamma\to\overline{M}_\gamma/\Delta_0$ and repeat similar arguments as above. From the Riemann–Hurwitz formula, we obtain

$$2\gamma-2=|\Delta_0|\bigl(2\gamma(\overline{M}_\gamma/\Delta_0)-2\bigr)+\sum_{p\in\overline{M}_\gamma}(\mu(p)-1). \tag{3-6}$$

We now treat the two cases that there is a symmetry $\sigma\in\Delta_0$ satisfying $\sigma(p_1)=p_2$ or not. For our case, we may exclude the case $t=3$, and consider the case $t=2$. It follows from (3-1) that $|\Delta_0|=2(\gamma+1)$ and $\sigma\in\Delta_0\setminus\mathcal{R}$.

Case 1. *The point p_1 can be transformed to p_2.*

If there exists such σ, then the ramification index at p_i must be $|\Delta_0|/2-1$. So (3-6) can be reduced to

$$2\gamma-2=|\Delta_0|\bigl(2\gamma(\overline{M}/\Delta_0)-2\bigr)+2(|\Delta_0|/2-1)+\sum_{p\in M}(\mu(p)-1).$$

Hence,

$$\begin{aligned}2\gamma&=|\Delta_0|\bigl(2\gamma(\overline{M}/\Delta_0)-1\bigr)+\sum_{p\in M}(\mu(p)-1)\\&=2(\gamma+1)\bigl(2\gamma(\overline{M}/\Delta_0)-1\bigr)+\sum_{p\in M}(\mu(p)-1)\\&\geq 2(\gamma+1)\bigl(2\gamma(\overline{M}/\Delta_0)-1\bigr).\end{aligned}\tag{3-7}$$

So the case $\gamma(\overline{M}/\Delta_0)>0$ leads to a contradiction, and thus $\gamma(\overline{M}/\Delta_0)=0$ holds. As a consequence, (3-7) can be reduced to

$$4\gamma+2=\sum_{p\in M}(\mu(p)-1).\tag{3-8}$$

Let $r'_1,\dots,r'_s$ be ramified values of π_{Δ_0} except for the $\pi_{\Delta_0}(p_i)$'s, and m_i-1 the ramification index at $p\in\pi_{\Delta_0}^{-1}(r'_i)$. Note $2\leq m_i\leq|\Delta_0|$, and (3-8) can be rewritten as

$$4\gamma+2=\sum_{i=1}^{s}(m_i-1)\frac{|\Delta_0|}{m_i}=2(\gamma+1)\sum_{i=1}^{s}\Bigl(1-\frac{1}{m_i}\Bigr).\tag{3-9}$$

If $s=1$, then (3-9) yields $2\gamma<0$, which is absurd. Hence $s\geq 2$, and (3-9) becomes

$$2\gamma=2(\gamma+1)\Biggl(\sum_{i=1}^{s}\Bigl(1-\frac{1}{m_i}\Bigr)-1\Biggr)=2(\gamma+1)\Biggl(\frac{s-2}{2}+\sum_{i=1}^{s}\Bigl(\frac{1}{2}-\frac{1}{m_i}\Bigr)\Biggr).$$

So

$$1>\frac{2\gamma}{2(\gamma+1)}=\frac{s-2}{2}+\sum_{i=1}^{s}\Bigl(\frac{1}{2}-\frac{1}{m_i}\Bigr)\geq\frac{s-2}{2}.$$

As a result, $2\leq s<4$ follows, and thus $s=2$ or 3.

<u>The case $s=2$.</u>

Equation (3-9) implies

$$2\gamma=2(\gamma+1)\Bigl(1-\frac{1}{m_1}-\frac{1}{m_2}\Bigr),$$

that is,

$$\frac{1}{m_1}+\frac{1}{m_2}=\frac{1}{\gamma+1}$$

holds. The inequalities $2\leq m_i\leq|\Delta_0|=2(\gamma+1)$ yield $m_1=m_2=2(\gamma+1)$.

The case $s = 3$.

From (3-9),

$$2\gamma = 2(\gamma+1)\Big(2 - \frac{1}{m_1} - \frac{1}{m_2} - \frac{1}{m_3}\Big). \tag{3-10}$$

Without loss of generality, we may assume $m_1 \le m_2 \le m_3$. In this case,

$$2 - \frac{3}{m_1} \le 2 - \frac{1}{m_1} - \frac{1}{m_2} - \frac{1}{m_3} \le 2 - \frac{3}{m_3}.$$

By (3-10), we obtain

$$2(\gamma+1)\Big(2 - \frac{3}{m_1}\Big) \le 2(\gamma+1)\Big(2 - \frac{1}{m_1} - \frac{1}{m_2} - \frac{1}{m_3}\Big) = 2\gamma.$$

Thus we have

$$m_1 \le \frac{3(\gamma+1)}{\gamma+2} < 3.$$

Hence $m_1 = 2$. Moreover, let us consider the case $m_1 = 2$ and $m_2 \le m_3$. Then

$$2(\gamma+1)\Big(\frac{3}{2} - \frac{2}{m_2}\Big) \le 2(\gamma+1)\Big(2 - \frac{1}{2} - \frac{1}{m_2} - \frac{1}{m_3}\Big) = 2\gamma,$$

and so

$$m_2 \le \frac{4(\gamma+1)}{\gamma+3} < 4.$$

It follows that $m_2 = 2$ or 3. For the case $m_2 = 2$, we have

$$2\gamma = 2(\gamma+1)\Big(1 - \frac{1}{m_3}\Big),$$

and thus $m_3 = \gamma + 1$. If $m_2 = 3$, then

$$2\gamma = 2(\gamma+1)\Big(\frac{7}{6} - \frac{1}{m_3}\Big)$$

and so $m_3 = \frac{6(\gamma+1)}{\gamma+7} < 6$. As a consequence, $(m_3, \gamma) = (3, 5)$, $(4, 11)$, or $(5, 29)$.

Case 2. *The point p_1 cannot be transformed to p_2.*

If there does not exist σ satisfying $\sigma(p_1) = p_2$, then the ramification index at p_i must be $|\Delta_0| - 1$. It follows that (3-6) can be reduced to

$$2\gamma - 2 = |\Delta_0|\big(2\gamma(\overline{M}/\Delta_0) - 2\big) + 2(|\Delta_0| - 1) + \sum_{p\in M}(\mu(p) - 1),$$

and thus

$$\begin{aligned} 2\gamma &= 2|\Delta_0|\gamma(\overline{M}/\Delta_0) + \sum_{p\in M}(\mu(p) - 1) \\ &= 4(\gamma+1)\gamma(\overline{M}/\Delta_0) + \sum_{p\in M}(\mu(p) - 1). \end{aligned} \tag{3-11}$$

<table>
<tr><th>s</th><th colspan="4">m_i</th></tr>
<tr><td>$s=2$</td><td colspan="4">$m_1=m_2=2(\gamma+1)$</td></tr>
<tr><td rowspan="5">$s=3$</td><td>m_1</td><td>m_2</td><td>m_3</td><td>γ</td></tr>
<tr><td>2</td><td>2</td><td>$\gamma+1$</td><td>arbitrary</td></tr>
<tr><td>2</td><td>3</td><td>3</td><td>5</td></tr>
<tr><td>2</td><td>3</td><td>4</td><td>11</td></tr>
<tr><td>2</td><td>3</td><td>5</td><td>29</td></tr>
</table>

Table 6. The case p_1 can be transformed to p_2.

s	m_i
$s=1$	$m_1=\gamma+1$

Table 7. The case p_1 cannot be transformed to p_2.

Equation (3-11) yields $\gamma(\overline{M}/\Delta_0)=0$, and so

$$2\gamma=\sum_{p\in M}(\mu(p)-1). \tag{3-12}$$

Suppose that $r_1',\dots,r_s'$ are ramified values of π_{Δ_0} except for the $\pi_{\Delta_0}(p_i)$'s, and m_i-1 the ramification index at $p\in\pi_{\Delta_0}^{-1}(r_i')$ as above. Now (3-12) can be rewritten as

$$2\gamma=\sum_{i=1}^{s}(m_i-1)\frac{|\Delta_0|}{m_i}=2(\gamma+1)\sum_{i=1}^{s}\Bigl(1-\frac{1}{m_i}\Bigr).$$

As a result,

$$1>\frac{2\gamma}{2\gamma+2}=\sum_{i=1}^{s}\Bigl(1-\frac{1}{m_i}\Bigr)=\sum_{i=1}^{s}\Bigl(\frac{1}{2}+\Bigl(\frac{1}{2}-\frac{1}{m_i}\Bigr)\Bigr)\geq\frac{s}{2},$$

and hence $1\leq s<2$, that is, $s=1$. Thus, we have $m_1=\gamma+1$.

Therefore, we obtain Tables 6 and 7.

3B. ***Weierstrass data for the case $\gamma=1$ with $(d_1,d_2)=(1,3)$.*** By Tables 3–5, we first consider the case $t=2$. Then $|\mathcal{R}|=2$ and we find $|\Delta_0|=4$ by (3-1). Set q_1,q_2 as two branch points of $\pi_{\mathcal{R}}$ distinct from the p_i's and $p_i':=\pi_{\mathcal{R}}(p_i)$, $q_i':=\pi_{\mathcal{R}}(q_i)$. Since $\pi_{\mathcal{R}}$ is a cyclic branched double cover of S^2, $\overline{M}_1$ can be given by

$$v^2=(u-p_1')^{m_1h_1}(u-p_2')^{m_2h_2}(u-q_1')^{m_3h_3}(u-q_2')^{m_4},$$

where $h_i\in\{1,-1\}$ $(i=1,2,3)$, $(2,m_i)=1$, and $R(u,v)=(u,-v)$.

Since $(d_1,d_2)=(1,3)$, there does not exist $\sigma\in\Delta_0$ with $\sigma(p_1)=p_2$. By Table 7 on Δ_0, there is a transformation $\tau\in\Delta_0\setminus\mathcal{R}$ satisfying $\tau(q_1)=q_2$, and thus $m_3=m_4$.

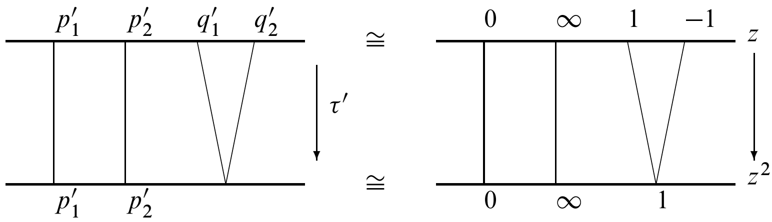


Figure 9. The Riemann surface $\overline{M}_1$.

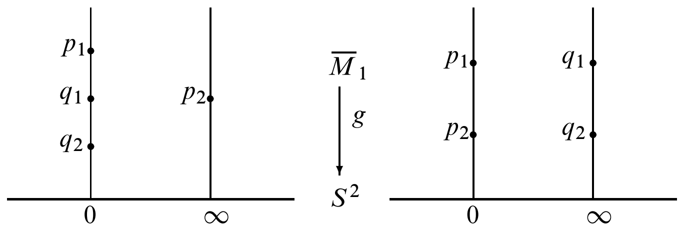


Figure 10. The possibilities of the Gauss map.

The transformation τ induces a degree 2 transformation $\tau' : \overline{M}_1/\mathcal{R} \to \overline{M}_1/\Delta_0$, that is, a transformation on S^2 such that $\tau'(q'_1) = \tau'(q'_2)$. Choosing suitable variables (z, w), we can represent $\tau(z, w) = (-z, *)$, $\tau'(z) = z^2$, $p'_1 = 0$, $p'_2 = \infty$, $q'_1 = 1$, $q'_2 = -1$, and moreover, $\overline{M}_1$ can be rewritten as $w^2 = z(z^2-1)$ (see Figure 9).

Now consider the Gauss map. Since p_1, p_2, q_1, q_2 are fixed points of $\mathcal{R}$ (that is, fixed points by rotations around the x_3-axis), we have $g(\{p_1, p_2, q_1, q_2\}) \subset \{0, \infty\}$. Note that q_1 can be transformed to q_2 by the biholomorphism τ. On the other hand, p_1 cannot be transformed to p_2. It follows that we essentially only need to consider the two cases depicted in Figure 10.

In the right-hand side case in Figure 10, the ramification index at q_i is $\frac{\gamma+2}{2} - 1 \notin \mathbb{Z}$ since the ramification index at q_1 of g must coincide with the ramification index at q_2 of g. Hence we only consider the left-hand side case in Figure 10. The ramification index at p_2 may be $1-1$, $2-1$, or $3-1$. If it is $2-1$, then $g^{-1}(\infty)$ consists of p_2 and a simple pole $q \in \overline{M}_1$. Then $R(q)$ must be a pole of g, but $R(q) \notin \{p_2, q\}$. This contradicts $R(q) \in g^{-1}(\infty)$. So the divisor of g is given by

$$(g) = \begin{cases} p_1 + q_1 + q_2 - 3p_2, \\ p_1 + q_1 + q_2 - p_2 - Q - R(Q) \end{cases}$$

for a point Q. Since $\tau(p_2) = p_2$, τ leaves {the poles of g} invariant. For the latter case, if we take $Q^* := \tau(Q)$, then Q^* must be a pole of g which is distinct from the $R^i(Q)$'s. It leads to a contradiction, and so we only consider the former case. In this case, the divisor of the meromorphic function $z(z^2-1)$ coincides with that of g^2, thus $g^2 = c'z(z^2-1)$ holds for some constant c'. Hence, $\overline{M}_1$ and g can be rewritten as

$$w^2 = z(z^2-1), \qquad g = cw$$

for some constant c, and $R(z, w) = (z, -w)$, $\tau(z, w) = (-z, iw)$. Then, the divisor of η is obtained by

$$(\eta) = \begin{cases} -2p_1 + 2p_2, \\ -4p_1 + 4p_2, \end{cases}$$

since $(d_1, d_2) = (1, 3)$. Thus, by a similar argument,

$$\eta^2 = \begin{cases} c'' \dfrac{(dz)^2}{z^3(z^2-1)} \\ c'' \dfrac{(dz)^2}{z^5(z^2-1)} \end{cases} = \begin{cases} c''\Big(\dfrac{dz}{zw}\Big)^2 \\ c''\Big(\dfrac{dz}{z^2w}\Big)^2 \end{cases}$$

hold for some constant c''. As a consequence, we obtain $\eta = c''' \frac{dz}{zw}$, $c''' \frac{dz}{z^2w}$ for some constant c'''. The latter case is given in Section 2A, and we shall prove that the former case does not occur in Section 3D. Note that the case $t = 1$ does not occur in this case.

3C. ***Weierstrass data for the even genus case with $(d_1, d_2) = (2, 2)$.*** We treat the case $t = 2$ ($|\mathcal{R}| = \gamma + 1$, $|\Delta_0| = 2(\gamma + 1)$). By Table 4, $\pi_{\mathcal{R}} : \overline{M}_\gamma \to \overline{M}_\gamma/\mathcal{R}$ is a cyclic branched cover of S^2. Thus $\overline{M}_\gamma$ can be represented by

$$v^{\gamma+1} = (u - p_1')^{m_1h_1}(u - p_2')^{m_2h_2}(u - q_1')^{m_3h_3}(u - q_2')^{m_4},$$

where $(m_i, \gamma + 1) = 1$, $h_i = \pm 1$, and $R(u, v) = (u, e^{2\pi i/(\gamma+1)}v)$.

Case 1. *The point p_1 cannot be transformed to p_2.*

We assume that there does not exist $\sigma \in \Delta_0$ such that $\sigma(p_1) = p_2$. From the table on Δ_0, there exists a transformation $\tau \in \Delta_0 \setminus \mathcal{R}$ satisfying $\tau(q_1) = q_2$, and thus $m_3 = m_4$. This τ induces a degree 2 transformation $\tau' : \overline{M}_1/\mathcal{R} \to \overline{M}_1/\Delta_0$, that is, a transformation on S^2 such that $\tau'(q_1') = \tau'(q_2')$. Choosing suitable variables (z, w), we can represent $\tau(z, w) = (-z, *)$, $\tau'(z) = z^2$, $p_1' = 0$, $p_2' = \infty$, $q_1' = 1$, $q_2' = -1$, and moreover, $\overline{M}_\gamma$ can be rewritten as $w^{\gamma+1} = z^{m_1h_1}(z^2-1)^{m_3}$ and $R(z, w) = (z, e^{2\pi i/(\gamma+1)}w)$ (see Figure 11).

Now we consider the Gauss map. Since p_1, p_2, q_1, q_2 are fixed points of $\mathcal{R}$ (that is, fixed points by rotations around the x_3-axis), we find $g(\{p_1, p_2, q_1, q_2\}) \subset \{0, \infty\}$. Note that q_1 can be transformed to q_2 by the biholomorphism τ. On the other hand, p_1 cannot be transformed to p_2. It follows that we essentially only need to consider the two cases as in Figure 12.

The left-hand side case in Figure 12.

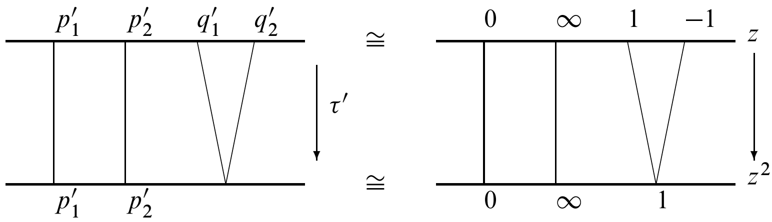


Figure 11. The Riemann surface $\overline{M}_\gamma$.

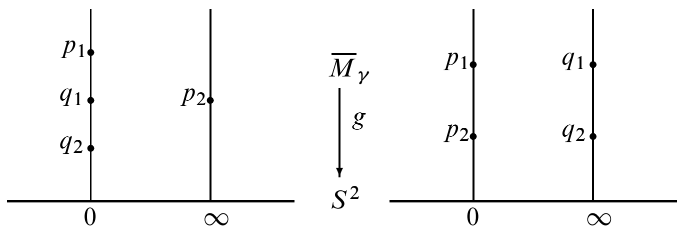


Figure 12. The possibilities of the Gauss map.

The divisor of g is given by

$$(g) = \begin{cases} (\gamma+2-2N)p_1 + Nq_1 + Nq_2 - (\gamma+2)p_2 \\ (\gamma+2-2N)p_1 + Nq_1 + Nq_2 - p_2 - Q - R(Q) - \cdots - R^\gamma(Q) \end{cases}$$

for a point Q. Note that $N > 0$ and $\gamma + 2 - 2N > 0$. Since $\tau(p_2) = p_2$, τ leaves {the poles of g} invariant. For the latter case, if we take $Q^* := \tau(Q)$, then Q^* must be a pole of g which is distinct from the $R^i(Q)$'s. This leads to a contradiction, and so we only consider the former case. Then the divisor of η is given by

$$(\eta) = -3p_1 + (2\gamma+1)p_2.$$

Hence the divisor of $g\eta$ is obtained by

$$(g\eta) = (\gamma - 2N - 1)p_1 + Nq_1 + Nq_2 + (\gamma-1)p_2.$$

If $\gamma - 2N - 1 \geq 0$, then $g\eta$ is holomorphic. Thus f is bounded and this leads to a contradiction. As a result, $\gamma - 2N - 1 < 0$ follows. The inequality $\gamma + 2 - 2N > 0$ yields $N < \frac{\gamma}{2} + 1$. Also, since γ is even, $N \leq \frac{\gamma}{2}$ holds. So we have $\gamma = 2N$.

It follows that the divisor of $g^{\gamma+1}$ coincides with that of $z^2(z^2-1)^{\gamma/2}$. Therefore, $\overline{M}_\gamma$ and g can be rewritten as

$$w^{\gamma+1} = z^2(z^2-1)^{\gamma/2}, \qquad g = cw$$

for some constant c, and $R(z, w) = (z, e^{2\pi i/(\gamma+1)} w)$, $\tau(z, w) = (-z, w)$. Furthermore, the divisor of $\eta^{\gamma+1}$ coincides with that of $(dz)^{\gamma+1}/(z^{\gamma-1} g^{2(\gamma+1)})$. Hence

$$\eta^{\gamma+1} = c'' \frac{(dz)^{\gamma+1}}{z^{\gamma-1} g^{2(\gamma+1)}} \left(= c'' z^2 \frac{(dz)^{\gamma+1}}{z^{\gamma+1} g^{2(\gamma+1)}} \right)$$

for some constant c''. By setting $z = u^{\gamma+1}$ and $w = u^2 v$, $\overline{M}_\gamma$ can be rewritten as

$$v^{\gamma+1} = (u^{2(\gamma+1)} - 1)^{\gamma/2},$$

and moreover,

$$g = cu^2 v, \qquad \eta = c''' \frac{u}{g^2} du$$

for some constant c'''. However, in this case, its genus is greater than γ, and such a case is excluded.

The right-hand side case in Figure 12.

The divisor of g is obtained by

$$(g) = (\gamma + 2 - N) p_1 + N p_2 - \frac{\gamma+2}{2} q_1 - \frac{\gamma+2}{2} q_2,$$

where $N > 0$ and $\gamma + 2 - N > 0$. Also, the divisor of η is given by

$$(\eta) = -3p_1 - 3p_2 + (\gamma + 2) q_1 + (\gamma + 2) q_2.$$

Thus the divisor of $g\eta$ is obtained by

$$(g\eta) = (\gamma - N - 1) p_1 + (N - 3) p_2 + \frac{\gamma+2}{2} q_1 + \frac{\gamma+2}{2} q_2.$$

If $\gamma - N - 1 \geq 0$ and $N - 3 \geq 0$, then $g\eta$ is holomorphic. Hence $\gamma - N - 1 < 0$ or $N - 3 < 0$. From the inequality $\gamma + 2 - N > 0$, the case $\gamma - N - 1 < 0$ corresponds to $\gamma = N, N - 1$. The case $N - 3 < 0$ implies $N = 1$ or 2. Essentially, we consider the cases $N = 1, 2$. The divisor of $g^{\gamma+1}$ coincides with

$$\frac{z^{\gamma+2-N}}{(z^2 - 1)^{(\gamma+2)/2}}.$$

As a consequence, $\overline{M}_\gamma$ and g can be rewritten as

$$w^{\gamma+1} = \frac{(z^2 - 1)^{(\gamma+2)/2}}{z^{\gamma+2-N}}, \qquad g = \frac{c}{w}$$

for some constant c. If $N = 1$, then $\gamma + 2 - N$ and $\gamma + 1$ are not coprime. So $N = 2$, and $R(z, w) = (z, e^{2\pi i/(\gamma+1)} w)$, $\tau(z, w) = (-z, w)$. Furthermore, the divisor $\eta^{\gamma+1}$ coincides with that of

$$\frac{(z^2 - 1)^2}{z^{\gamma+3}} (dz)^{\gamma+1}.$$

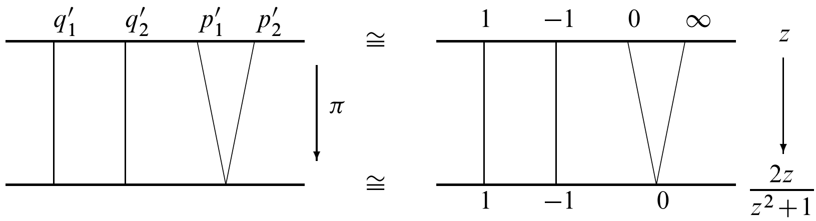


Figure 13. The Riemann surface $\overline{M}_\gamma$ (for the case $s = 2$).

As a consequence,

$$\eta^{\gamma+1} = c' \frac{(z^2-1)^2}{z^{\gamma+3}} (dz)^{\gamma+1} \left(= \frac{c'}{z^6} \left(\frac{z^3 w^4 \, dz}{(z^2-1)^2} \right)^{\gamma+1} \right)$$

for some constant c'. By similar arguments as above, we may exclude this case except the case $\gamma = 2$. For the case $\gamma = 2$, $\overline{M}_\gamma$ and g can be rewritten as

$$w^3 = \frac{(z^2-1)^2}{z^2}, \qquad g = \frac{c}{w},$$

and moreover, we find $\eta = c'' \frac{w}{z} \, dz$ for some constants c, c''. However, this surface has a transformation $\sigma \in \Delta_0$ defined by $\sigma(z, w) = \left(\frac{1}{z}, w\right)$, and we have $\sigma(p_1) = p_2$. This contradicts our assumption.

Case 2. *The point p_1 can be transformed to p_2.*

Suppose that there exists $\sigma \in \Delta_0$ such that $\sigma(p_1) = p_2$. By Table 6, we consider two cases, namely the case $s = 2$ and the case $s = 3$. Note that $\sigma(p_2) = p_1$ and $\sigma \in \Delta_0 \setminus \mathcal{R}$.

The case $s = 2$.

By Table 6, every q_i must be branch points of π_{Δ_0} with the ramified index $2(\gamma+1)-1$. Hence, $\sigma(q_i) = q_i$ for $i = 1, 2$, and moreover, σ induces a degree 2 transformation $\sigma' : \overline{M}_\gamma/\mathcal{R} \to \overline{M}_\gamma/\Delta_0$ which is a transformation on S^2 and the q_i''s are two fixed points of σ' (see Figure 13).

By suitable variables (z, w), we have $\sigma(z, w) = \left(\frac{1}{z}, *\right)$, $\sigma'(z) = \frac{2z}{z^2+1}$, $p_1' = 0$, $p_2' = \infty$, $q_1' = 1$, $q_2' = -1$. Also, $\overline{M}_\gamma$ is given by

$$w^2 = z^{m_1 h_1} (z-1)^{m_2 h_2} (z+1)^{m_3 h_3}.$$

We consider the Gauss map. Essentially, we treat the two cases shown in Figure 14.

For the case in the left-hand side of Figure 14, the divisor of g is given by

$$(g) = \begin{cases} (\gamma+2-2N)q_1 + Np_1 + Np_2 - (\gamma+2)q_2 \\ (\gamma+2-2N)q_1 + Np_1 + Np_2 - q_2 - Q - R(Q) - \cdots - R^\gamma(Q), \end{cases}$$

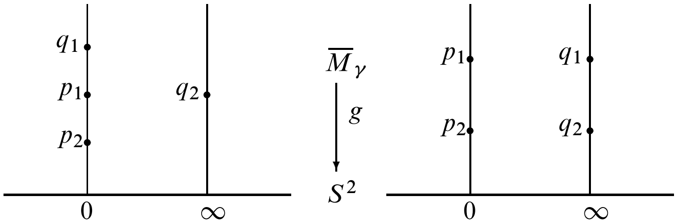


Figure 14. The possibilities of the Gauss map.

where $N > 0$ and $\gamma + 2 - 2N > 0$. Since $\sigma(q_2) = q_2$, σ leaves {the poles of g} invariant. For the latter case, if we take $Q^* := \sigma(Q)$, then Q^* must be a pole of g which is distinct from the Q_i's. This leads to a contradiction, and so we only consider the former case. In this case, we see

$$(\eta) = -3p_1 - 3p_2 + (2\gamma + 4)q_2.$$

Then

$$(g\eta) = (\gamma + 2 - 2N)q_1 + (N-3)p_1 + (N-3)p_2 + (\gamma+2)q_2.$$

It follows that $N - 3 < 0$. With $N > 0$, this yields $N = 1$ or 2.

Thus the divisor of $g^{\gamma+1}$ coincides with that of

$$\frac{z^N(z-1)^{\gamma+2-2N}}{(z+1)^{\gamma+2}}.$$

Therefore, $\overline{M}_\gamma$ and g can be rewritten as

$$w^{\gamma+1} = \frac{z^N(z-1)^{\gamma+2-2N}}{(z+1)^{\gamma+2}} \quad (\text{for } N = 1, 2), \qquad g = cw$$

for some constant c, and $R(z, w) = (z, e^{2\pi i/(\gamma+1)}w)$, $\sigma(z, w) = \left(\frac{1}{z}, w\right)$. Furthermore, by similar arguments, η can be obtained by

$$\eta^{\gamma+1} = c' \frac{(z+1)^{\gamma+4}}{z^{\gamma+3}(z-1)^\gamma}(dz)^{\gamma+1} = \begin{cases} c' \dfrac{z+1}{z-1}\left(\dfrac{(z-1)\,dz}{z(z+1)w^2}\right)^{\gamma+1} & \text{for } N = 1, \\ c'\left(\dfrac{z+1}{z-1}\right)^2\left(\dfrac{dz}{zw}\right)^{\gamma+1} & \text{for } N = 2, \end{cases}$$

for some constant c'. So we may exclude this case, like the previous case.

Next we consider the case in the right-hand side of Figure 14. Then the divisor of g is obtained by

$$(g) = \frac{\gamma+2}{2}p_1 + \frac{\gamma+2}{2}p_2 - Nq_1 - (\gamma + 2 - N)q_2,$$

where $N > 0$ and $\gamma + 2 - N > 0$. In this case, the divisor of η is given by

$$(\eta) = -3p_1 - 3p_2 + 2Nq_1 + (2\gamma + 4 - 2N)q_2.$$

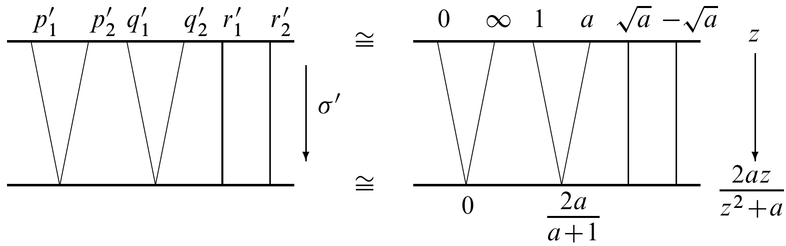


Figure 15. The Riemann surface $\overline{M}_\gamma$ (the case $s = 3$).

Thus

$$(g\eta) = \frac{\gamma-4}{2} p_1 + \frac{\gamma-4}{2} p_2 + N q_1 + (\gamma+2-N) q_2.$$

Hence $\gamma - 4 < 0$ and so $\gamma = 2$. Moreover, from the inequality $\gamma + 2 - N > 0$, we obtain $N = 1$, 2, or 3. If $N = 1$, then $\gamma + 2 - N$ and $\gamma + 1$ are not coprime. Also, if $N = 3$, then N and $\gamma + 1$ are not coprime. So we have $N = 2$. As a result, the divisor of $g^{\gamma+1} (= g^3)$ coincides with that of

$$\frac{z^{(\gamma+2)/2}}{(z-1)^N (z+1)^{\gamma+2-N}} \left(= \frac{z^2}{(z-1)^2 (z+1)^2} \right).$$

Therefore, $\overline{M}_2$ and g can be rewritten as

$$w^3 = \frac{(z-1)^2 (z+1)^2}{z^2}, \qquad g = \frac{c}{w}$$

for some constant c, and $R(z, w) = (z, e^{2\pi i/3} w)$, $\sigma(z, w) = \left(\frac{1}{z}, w\right)$. Also, we have $\eta = c' \frac{w}{z}\, dz$ for some constant c'. We shall prove that this case does not occur in Section 3D.

The case $s = 3$

From Table 6, we have that the p_i's and q_i's must be branch points of π_{Δ_0} with the ramified index $(\gamma + 1) - 1$. As a result, there exist two sets $\{r_1^{(1)}, \ldots, r_{\gamma+1}^{(1)}\}$, $\{r_1^{(2)}, \ldots, r_{\gamma+1}^{(2)}\}$ of branch points with the ramified index $2 - 1$ of π_{Δ_0} satisfying $\pi_{\Delta_0}(r_i^{(1)}) = \pi_{\Delta_0}(r_j^{(1)})$ and $\pi_{\Delta_0}(r_i^{(2)}) = \pi_{\Delta_0}(r_j^{(2)})$ for $1 \le i, j \le \gamma + 1$. Note that $r_i^{(1)}$ and $r_i^{(2)}$ are distinct from the p_i's and q_i's. Hence, $\sigma(q_1) = q_2$ and $\sigma(q_2) = q_1$, and moreover, σ induces a degree 2 transformation $\sigma' : \overline{M}_\gamma/\mathcal{R} \to \overline{M}_\gamma/\Delta_0$, that is, a transformation on S^2 such that $r_i' := \pi_{\mathcal{R}}(r_j^{(i)})$'s are two branch points of σ' (see Figure 15).

Choosing suitable variables (z, w), for $a \in \mathbb{C} \setminus \{0\}$, we have $\sigma(z, w) = \left(\frac{a}{z}, *\right)$, $\sigma'(z) = 2az/(z^2 + a)$, $p_1' = 0$, $p_2' = \infty$, $q_1' = 1$, $q_2' = a$, $r_1' = \sqrt{a}$, $r_2' = -\sqrt{a}$. Also, $\overline{M}_\gamma$ is given by $w^{\gamma+1} = z^{m_1 h_1} (z - 1)^{m_2 h_2} (z - a)^{m_2 h_3}$ (see Figure 15).

We now consider the Gauss map, essentially the two cases in Figure 16.

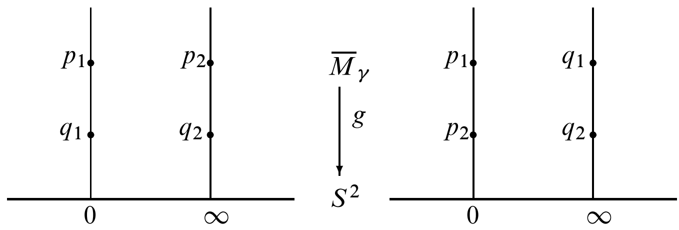


Figure 16. The possibilities of the Gauss map.

Then the divisor of g is obtained by

$$(g) = \begin{cases} (\gamma+2-N)p_1 + Nq_1 - (\gamma+2-N)p_2 - Nq_2 & \text{(the LHS case)}, \\ \dfrac{\gamma+2}{2}p_1 + \dfrac{\gamma+2}{2}p_2 - \dfrac{\gamma+2}{2}q_1 - \dfrac{\gamma+2}{2}q_2 & \text{(the RHS case)}, \end{cases}$$

where $N > 0$ and $\gamma+2-N > 0$. The divisor of η is given by

$$(\eta) = \begin{cases} -3p_1 + (2\gamma-2N+1)p_2 + 2Nq_2, \\ -3p_1 - 3p_2 + (\gamma+2)q_1 + (\gamma+2)q_2. \end{cases}$$

So the divisor of $g\eta$ is obtained by

$$(g\eta) = \begin{cases} (\gamma-N-1)p_1 + (\gamma-N-1)p_2 + Nq_1 + Nq_2, \\ \dfrac{\gamma-4}{2}p_1 + \dfrac{\gamma-4}{2}p_2 + \dfrac{\gamma+2}{2}q_1 + \dfrac{\gamma+2}{2}q_2. \end{cases}$$

For the former case, if $\gamma-N-1 \geq 0$, then $g\eta$ is holomorphic. Thus $\gamma-N-1 < 0$ holds. Also, the inequality $\gamma+2-N > 0$ yields $N = \gamma$ or $\gamma+1$. For the latter case, $\gamma-4 \geq 0$ cannot hold, and hence $\gamma = 2$. As a consequence,

$$g^{\gamma+1} = \begin{cases} c' z^{\gamma+2-N}\left(\dfrac{z-1}{z-a}\right)^N & \text{(the LHS case)}, \\ c'\left(\dfrac{z}{(z-1)(z-a)}\right)^{(\gamma+2)/2} & \text{(the RHS case)}, \end{cases}$$

for some constant c'. If $N = \gamma+1$, then $\gamma+1$ and N are not coprime. Thus $N = \gamma$ follows. Therefore, $\overline{M}_\gamma$ and g can be rewritten as

$$\begin{cases} w^{\gamma+1} = z^2\left(\dfrac{z-1}{z-a}\right)^\gamma, \quad g = cw & \text{(the LHS case)}, \\ w^3 = \left(\dfrac{(z-1)(z-a)}{z}\right)^2, \quad g = \dfrac{c}{w} & \text{(the RHS case)}, \end{cases}$$

and

$$R(z,w)=(z,e^{2\pi i/(\gamma+1)}w),$$

$$\sigma(z,w)=\begin{cases}\left(\dfrac{a}{z},\dfrac{a^{(\gamma+2)/(\gamma+1)}}{w}\right) & \text{(the LHS case)},\\ \left(\dfrac{a}{z},w\right) & \text{(the RHS case)}.\end{cases}$$

Also, for some constant c', we have

$$\eta=\begin{cases}c'\dfrac{dz}{zw} & \text{(the LHS case)},\\ c'\dfrac{w}{z}\,dz & \text{(the RHS case)}.\end{cases}$$

$\Delta\setminus\Delta_0\neq\varnothing$ implies that there exists a degree 2 antiholomorphic transformation. Hence $a\in\mathbb{R}$ and the antiholomorphic transformation can be represented by the map $(z,w)\mapsto(\bar z,\overline{w})$. The former case corresponds to our result in Section 2B, and the latter case is considered in Section 3D. Note that the case $t=1$ does not occur.

3D. ***Well-definedness.*** In this subsection, we consider the well-definedness for the following two cases:

$\overline{M}_\gamma$	g	η	symmetries
$w^2=z(z^2-1)$	cw	$c'\dfrac{dz}{zw}$	$R(z,w)=(-z,iw)$
$w^3=\dfrac{(z-1)^2(z-a)^2}{z^2}$	$\dfrac{c}{w}$	$c'\dfrac{w}{z}\,dz$	$R(z,w)=(z,e^{2\pi i/3}w)$, $\sigma(z,w)=\left(\dfrac{a}{z},w\right)$

Note that $c,c'\in\mathbb{C}\setminus\{0\}$, $a\in\mathbb{R}\setminus\{0,1\}$. M is given by

$$M=\begin{cases}\overline{M}_1\setminus\{(0,0),(\infty,\infty)\} & \text{(the former case)},\\ \overline{M}_2\setminus\{(0,\infty),(\infty,\infty)\} & \text{(the latter case)}.\end{cases}$$

The case $a=-1$ corresponds to the surface which we treat for the case $s=2$. Note that the Weierstrass data $(e^{i\theta}g,e^{-i\theta}\eta)$ produces the same minimal surface as (g,η) rotated by an angle θ around the x_3-axis. So after a suitable rotation of the surface, we may assume $c\in\mathbb{R}_+$. Also, multiplying a positive real number into η is just a homothety, so we may assume that $|c'|=1$.

Our claim is that all cases do not occur.

<u>The former case.</u>

First we consider $\Phi={}^t(\Phi_1,\Phi_2,\Phi_3)$ in Theorem 1.1:

$$\Phi_1=\left(\frac{1}{w}-c^2w\right)c'\frac{dz}{z},\qquad \Phi_2=i\left(\frac{1}{w}+c^2w\right)c'\frac{dz}{z},\qquad \Phi_3=2cc'\frac{dz}{z}.$$

For the residue of Φ_3 at $z=0$ to be real, we see that $c'=\pm 1$. We may choose $c'=1$. We shall use the notation in the proof of Theorem 2.2 for $\gamma=1$.

Straightforward calculation yields

$$\frac{dz}{zw} - d\Big(\frac{2w}{z}\Big) = -\frac{z}{w}\,dz.$$

Thus we have

$$\oint_{\ell'} \eta = -\oint_{\ell'} \frac{z}{w}\,dz = -\oint_{\ell} \frac{z}{w}\,dz = -2i\int_0^1 \sqrt{\frac{t}{1-t^2}}\,dt,$$

$$\oint_{\ell'} g^2\eta = c^2\oint_{\ell} \frac{w}{z}\,dz = -2ic^2\int_0^1 \sqrt{\frac{1-t^2}{t}}\,dt.$$

Equation (1-4) implies

$$-\int_0^1 \sqrt{\frac{t}{1-t^2}}\,dt = c^2\int_0^1 \sqrt{\frac{1-t^2}{t}}\,dt.$$

So we have $c^2<0$ and this contradicts $c>0$.

<u>The latter case.</u>

First we consider $\Phi = {}^t(\Phi_1, \Phi_2, \Phi_3)$ in Theorem 1.1:

$$\Phi_1 = \Big(w-\frac{c^2}{w}\Big)c'\frac{dz}{z}, \qquad \Phi_2 = i\Big(w+\frac{c^2}{w}\Big)c'\frac{dz}{z}, \qquad \Phi_3 = 2cc'\frac{dz}{z}.$$

For the residue of Φ_3 at $z=0$ to be real, we see that $c'=\pm 1$. We may choose $c'=1$.

We shall show that for any $c>0$ and $a\in\mathbb{R}\setminus\{0,1\}$, the period condition (P) cannot be satisfied.

A straightforward calculation yields

$$\eta + \frac{3}{2}\,dw = \Big(\frac{w}{z-1}+\frac{w}{z-a}\Big)\,dz. \tag{3-13}$$

Note that the right-hand side of this equation is a holomorphic differential on $\overline{M}_2\setminus\{(\infty,\infty)\}$.

We now consider the following three cases: $a>1$, $0<a<1$, $a<0$.

(i) The case $a>1$:

We set

$$\ell = \left\{(z,w) = \left(t, \sqrt[3]{\frac{(1-t)^2(a-t)^2}{t^2}}\right) \,\Big|\, 0\le t\le 1\right\} \cup \left\{(z,w) = \left(-t, e^{2\pi i/3}\sqrt[3]{\frac{(1+t)^2(a+t)^2}{t^2}}\right) \,\Big|\, -1\le t\le 0\right\}.$$

From (3-13), we have

$$\oint_\ell \eta = -\bigl(1-e^{2\pi i/3}\bigr)\int_0^1 \left(\sqrt[3]{\frac{(a-t)^2}{t^2(1-t)}}+\sqrt[3]{\frac{(1-t)^2}{t^2(a-t)}}\right)dt, \tag{3-14}$$

$$\int_\ell g^2\eta = c^2\bigl(1-e^{-2\pi i/3}\bigr)\int_0^1 \frac{dt}{\sqrt[3]{t(1-t)^2(a-t)^2}}. \tag{3-15}$$

Thus (1-4) is equivalent to

$$-\int_0^1 \left(\sqrt[3]{\frac{(a-t)^2}{t^2(1-t)}}+\sqrt[3]{\frac{(1-t)^2}{t^2(a-t)}}\right)dt = c^2\int_0^1 \frac{dt}{\sqrt[3]{t(1-t)^2(a-t)^2}}.$$

But this is impossible because the left-hand side is a negative real number and the right-hand side is a positive real number.

(ii) The case $0<a<1$:
We set

$$\ell=\left\{(z,w)=\left(at,\sqrt[3]{\frac{(at-1)^2(1-t)^2}{t^2}}\right)\,\middle|\,0\le t\le 1\right\} \\ \cup\left\{(z,w)=\left(-at,e^{2\pi i/3}\sqrt[3]{\frac{(at+1)^2(1+t)^2}{t^2}}\right)\,\middle|\,-1\le t\le 0\right\}.$$

From (3-13), we have

$$\oint_\ell \eta = -\bigl(1-e^{2\pi i/3}\bigr)\int_0^1 \left(a\sqrt[3]{\frac{(1-t)^2}{t^2(1-at)}}+\sqrt[3]{\frac{(1-at)^2}{t^2(1-t)}}\right)dt, \tag{3-16}$$

$$\int_\ell g^2\eta = c^2\bigl(1-e^{-2\pi i/3}\bigr)\int_0^1 \frac{dt}{\sqrt[3]{t(1-t)^2(1-at)^2}}. \tag{3-17}$$

Thus (1-4) is equivalent to

$$-\int_0^1 \left(a\sqrt[3]{\frac{(1-t)^2}{t^2(1-at)}}+\sqrt[3]{\frac{(1-at)^2}{t^2(1-t)}}\right)dt = c^2\int_0^1 \frac{dt}{\sqrt[3]{t(1-t)^2(1-at)^2}},$$

but again this is impossible by the same reason as in the case (i).

(iii) The case $a<0$:

We set

$$\ell = \left\{(z,w) = \left(at, \sqrt[3]{\frac{(1-at)^2(1-t)^2}{t^2}}\right) \,\middle|\, 0 \le t \le 1\right\} \cup \left\{(z,w) = \left(-at, e^{2\pi i/3}\sqrt[3]{\frac{(1+at)^2(1+t)^2}{t^2}}\right) \,\middle|\, -1 \le t \le 0\right\},$$

$$\ell' = \left\{(z,w) = \left(t, \sqrt[3]{\frac{(1-t)^2(t-a)^2}{t^2}}\right) \,\middle|\, 0 \le t \le 1\right\} \cup \left\{(z,w) = \left(-t, e^{2\pi i/3}\sqrt[3]{\frac{(1+t)^2(t+a)^2}{t^2}}\right) \,\middle|\, -1 \le t \le 0\right\}.$$

From (3-13), we have

$$\oint_\ell \eta = -\big(1-e^{2\pi i/3}\big)\int_0^1 \left(a\sqrt[3]{\frac{(1-t)^2}{t^2(1-at)}} + \sqrt[3]{\frac{(1-at)^2}{t^2(1-t)}}\right) dt, \tag{3-18}$$

$$\int_\ell g^2\eta = c^2\big(1-e^{-2\pi i/3}\big)\int_0^1 \frac{dt}{\sqrt[3]{t(1-t)^2(1-at)^2}}. \tag{3-19}$$

Thus (1-4) is equivalent to

$$-\int_0^1 \left(a\sqrt[3]{\frac{(1-t)^2}{t^2(1-at)}} + \sqrt[3]{\frac{(1-at)^2}{t^2(1-t)}}\right) dt = c^2\int_0^1 \frac{dt}{\sqrt[3]{t(1-t)^2(1-at)^2}}. \tag{3-20}$$

The right-hand side is clearly positive. So now we estimate the left-hand side:

$$\begin{aligned}
(\text{LHS}) &= -a\int_0^1 \sqrt[3]{\frac{(1-t)^2}{t^2(1-at)}}\,dt - \int_0^1 \sqrt[3]{\frac{(1-at)^2}{t^2(1-t)}}\,dt \\
&\le -a\int_0^1 \sqrt[3]{\frac{(1-t)^2}{t^2}}\,dt - \int_0^1 \sqrt[3]{\frac{1}{t^2(1-t)}}\,dt \\
&= -aB\big(\tfrac13,\tfrac53\big) - B\big(\tfrac13,\tfrac23\big) \\
&= -\big(\tfrac23 a + 1\big)B\big(\tfrac13,\tfrac23\big),
\end{aligned}$$

where $B(x,y)$ is the classical beta function mentioned in Section 2B. Hence, if $-\big(\frac23 a+1\big) \le 0$, (3-20) never holds. That is,

$$\text{if } -\tfrac32 \le a < 0, \text{ (3-20) never holds.} \tag{3-21}$$

On the other hand, we have

$$\oint_{\ell'} \eta = -\left(1-e^{2\pi i/3}\right) \int_0^1 \left(\sqrt[3]{\frac{(t-a)^2}{t^2(1-t)}} - \sqrt[3]{\frac{(1-t)^2}{t^2(t-a)}} \right) dt, \tag{3-22}$$

$$\int_{\ell'} g^2\eta = c^2\left(1-e^{-2\pi i/3}\right) \int_0^1 \frac{dt}{\sqrt[3]{t(1-t)^2(t-a)^2}}, \tag{3-23}$$

from (3-13). Thus (1-4) is equivalent to

$$-\int_0^1 \left(\sqrt[3]{\frac{(t-a)^2}{t^2(1-t)}} - \sqrt[3]{\frac{(1-t)^2}{t^2(t-a)}} \right) dt = c^2 \int_0^1 \frac{dt}{\sqrt[3]{t(1-t)^2(t-a)^2}}. \tag{3-24}$$

The right-hand side is again positive. So again we estimate the left-hand side:

$$\begin{aligned}
(\text{LHS}) &= -\int_0^1 \sqrt[3]{\frac{(t-a)^2}{t^2(1-t)}}\, dt + \int_0^1 \sqrt[3]{\frac{(1-t)^2}{t^2(t-a)}}\, dt \\
&\le -\int_0^1 \sqrt[3]{\frac{(-a)^2}{t^2(1-t)}}\, dt + \int_0^1 \sqrt[3]{\frac{(1-t)^2}{t^2(-a)}}\, dt \\
&= -(-a)^{2/3} B\left(\tfrac{1}{3}, \tfrac{2}{3}\right) + (-a)^{-1/3} B\left(\tfrac{1}{3}, \tfrac{5}{3}\right) \\
&= (-a)^{-1/3}\left(a+\tfrac{2}{3}\right) B\left(\tfrac{1}{3}, \tfrac{2}{3}\right).
\end{aligned}$$

Hence, if $a+\frac{2}{3} \le 0$, (3-24) never holds. That is,

$$\text{if } a \le -\tfrac{2}{3}, \text{ (3-24) never holds.} \tag{3-25}$$

Combining (3-21) and (3-25), the period condition cannot be solved.

Main Theorem 2 is an immediate consequence of the above arguments.

4. Remaining problems

In this section we introduce remaining problems related to this work.

4A. ***The case that γ is odd and greater than 1.*** For the case that the genus $\gamma > 1$ is odd, a complete minimal surface of finite total curvature $f : M = \overline{M}_\gamma \setminus \{p_1, p_2\} \to \mathbb{R}^3$ which satisfies equality in (1-9) is yet to be found. However, Matthias Weber [2015] has constructed the following examples numerically.

Example 4.1 (Weber). Let γ be a positive integer. Define

$$F_1(z; a_1, a_3, \ldots, a_{2\gamma-1}) = \prod_{i=1}^{\gamma} (z - a_{2i-1}),$$

$$F_2(z; a_2, a_4, \ldots, a_{2\gamma}) = \prod_{i=1}^{\gamma} (z - a_{2i}),$$

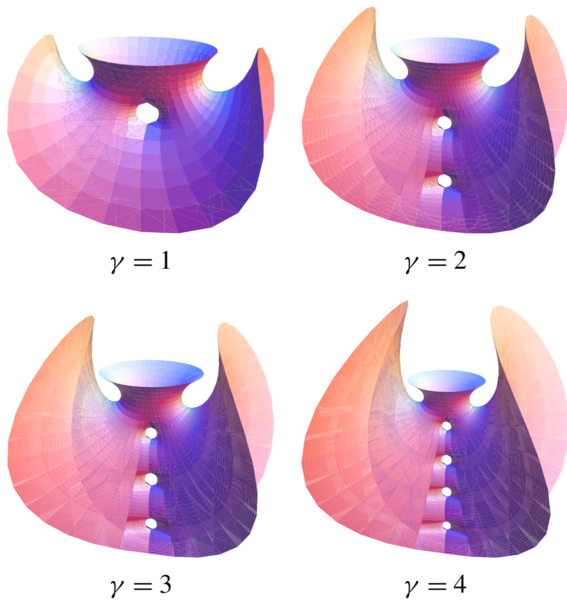


Figure 17. Minimal surfaces of genus γ with two ends which satisfy $\deg(g) = \gamma + 2$.

where $1 = a_1 < a_2 < \cdots < a_{2\gamma}$ are constants to be determined. Define a compact Riemann surface $\overline{M}_\gamma$ of genus γ by

$$\overline{M}_\gamma = \left\{ (z, w) \in (\mathbb{C} \cup \{\infty\})^2 \;\middle|\; w^2 = z \frac{F_1(z; a_1, a_3, \dots, a_{2\gamma-1})}{F_2(z; a_2, a_4, \dots, a_{2\gamma})} \right\}.$$

Let $M = \overline{M}_\gamma \setminus \{(0,0), (\infty, \infty)\}$. We set

$$g = c \frac{w}{z+1} \quad \text{for } c > 0, \qquad \eta = \frac{(z+1)^2}{zw} \, dz.$$

Then there exist constants $c, a_2, a_3, \dots, a_{2\gamma}$ such that (P) holds. (See Figure 17.)

For $\gamma = 1$, we can prove the existence of the surface rigorously. However, for other cases, since the surface does not have enough symmetry, the rigorous proof of the existence still remains an open problem.

4B. ***Existence of nonorientable minimal surfaces.*** Our work is devoted to minimal surfaces satisfying $\deg(g) = \gamma + 2$. On the other hand, it is important to consider the existence of nonorientable minimal surfaces with $\deg(g) = \gamma + 3$. Now, we review nonorientable minimal surfaces in $\mathbb{R}^3$.

Let $f' : M' \to \mathbb{R}^3$ be a minimal immersion of a nonorientable surface into $\mathbb{R}^3$. Then the oriented two sheeted covering space M of M' naturally inherits a Riemann surface structure and we have a canonical projection $\pi : M \to M'$. We can also define a map $I : M \to M$ such that $\pi \circ I = \pi$, which is an antiholomorphic involution on M without fixed points. Here M' can be identified with $M/\langle I \rangle$. In this way, if $f : M \to \mathbb{R}^3$ is a conformal minimal surface and there is an antiholomorphic involution $I : M \to M$ without fixed points so that $f \circ I = f$, then we can define a nonorientable minimal surface $f' : M' = M/\langle I \rangle \to \mathbb{R}^3$. Conversely, every nonorientable minimal surface is obtained in this procedure.

Suppose that $f' : M' = M/\langle I \rangle \to \mathbb{R}^3$ is complete and of finite total curvature. Then we can apply Theorems 1.4 and 1.5 to the conformal minimal immersion $f : M \to \mathbb{R}^3$. Furthermore, we have a stronger restriction on the topology of M' or M. In fact, Meeks [1981] showed that the Euler characteristic $\chi(\overline{M}_\gamma)$ and $2\deg(g)$ are congruent modulo 4, where g is the Gauss map of f. By these facts, we can observe that for every complete nonorientable minimal surface of finite total curvature, $\deg(g) \geq \gamma + 3$ holds.

For $\gamma = 0$ and $\gamma = 1$, Meeks' Möbius strip [1981] and López's Klein bottle [1993] satisfy $\deg(g) = \gamma + 3$. But, for $\gamma \geq 2$, no examples with $\deg(g) = \gamma + 3$ are known. So, it is interesting to give a minimal surface satisfying $\deg(g) = \gamma + 3$ with an antiholomorphic involution without fixed points. This problem appeared in [López and Martín 1999] and [Martín 2005].

Acknowledgements

The authors thank Professor Reiko Miyaoka for a fruitful discussion when the authors attended her informal lectures on minimal surfaces at Kyushu University. The authors also thank Professors Shin Kato, Francisco J. López, Wayne Rossman, Masaaki Umehara, Matthias Weber, and Kotaro Yamada for their valuable comments. They would further like to thank the referee for comments that significantly improved the results here.

References

[Berglund and Rossman 1995] J. Berglund and W. Rossman, "Minimal surfaces with catenoid ends", *Pacific J. Math.* **171**:2 (1995), 353–371. MR 97a:53007 Zbl 1147.53302

[Chen and Gackstatter 1982] C. C. Chen and F. Gackstatter, "Elliptische und hyperelliptische Funktionen und vollständige Minimalflächen vom Enneperschen Typ", *Math. Ann.* **259**:3 (1982), 359–369. MR 84d:53005 Zbl 0468.53008

[Espírito Santo 1994] N. do Espírito Santo, "Complete minimal surfaces in $\mathbb{R}^3$ with type Enneper end", *Ann. Inst. Fourier* (*Grenoble*) **44**:2 (1994), 525–557. MR 95h:53008 Zbl 0803.53006

[Fujimori and Shoda 2014] S. Fujimori and T. Shoda, "A family of complete minimal surfaces of finite total curvature with two ends", pp. 19–31 in *Differential geometry of submanifolds and its related topics*, edited by S. Maeda et al., World Scientific, Hackensack, NJ, 2014. MR 3203469 Zbl 1303.53014

[Hoffman and Karcher 1997] D. A. Hoffman and H. Karcher, "Complete embedded minimal surfaces of finite total curvature", pp. 5–93 in *Geometry, V*, edited by R. Osserman and R. V. Gamkrelidze, Encyclopaedia of Mathematical Sciences **90**, Springer, Berlin, 1997. MR 98m:53012 Zbl 0890.53001

[Hoffman and Meeks 1990] D. A. Hoffman and W. H. Meeks, III, "Embedded minimal surfaces of finite topology", *Ann. of Math.* (2) **131**:1 (1990), 1–34. MR 91i:53010 Zbl 0695.53004

[Hoffman and Osserman 1980] D. A. Hoffman and R. Osserman, *The geometry of the generalized Gauss map*, Memoirs of the American Mathematical Society **28**:236, American Mathematical Society, Providence, RI, 1980. MR 82b:53012 Zbl 0469.53004

[Huber 1957] A. Huber, "On subharmonic functions and differential geometry in the large", *Comment. Math. Helv.* **32** (1957), 13–72. MR 20 #970 Zbl 0080.15001

[Jorge and Meeks 1983] L. P. Jorge and W. H. Meeks, III, "The topology of complete minimal surfaces of finite total Gaussian curvature", *Topology* **22**:2 (1983), 203–221. MR 84d:53006 Zbl 0517.53008

[López 1992] F. J. López, "The classification of complete minimal surfaces with total curvature greater than -12π", *Trans. Amer. Math. Soc.* **334**:1 (1992), 49–74. MR 93a:53008 Zbl 0771.53005

[López 1993] F. J. López, "A complete minimal Klein bottle in $\mathbb{R}^3$", *Duke Math. J.* **71**:1 (1993), 23–30. MR 94e:53005 Zbl 0796.53006

[López and Martín 1999] F. J. López and F. Martín, "Complete minimal surfaces in $\mathbb{R}^3$", *Publ. Mat.* **43**:2 (1999), 341–449. MR 2002c:53010 Zbl 0951.53001

[Martín 2005] F. Martín, "Complete nonorientable minimal surfaces in $\mathbb{R}^3$", pp. 371–380 in *Global theory of minimal surfaces* (Berkeley, 2001), edited by D. A. Hoffman, Clay Mathematics Proceedings **2**, American Mathematical Society, Providence, RI, 2005. MR 2006e:53020 Zbl 1106.53006

[Martín and Weber 2001] F. Martín and M. Weber, "On properly embedded minimal surfaces with three ends", *Duke Math. J.* **107**:3 (2001), 533–559. MR 2002b:53004 Zbl 1044.53006

[Meeks 1981] W. H. Meeks, III, "The classification of complete minimal surfaces in $\mathbb{R}^3$ with total curvature greater than -8π", *Duke Math. J.* **48**:3 (1981), 523–535. MR 82k:53009 Zbl 0472.53010

[Meeks and Weber 2007] W. H. Meeks, III and M. Weber, "Bending the helicoid", *Math. Ann.* **339**:4 (2007), 783–798. MR 2008k:53020 Zbl 1156.53011

[Mira 2006] P. Mira, "Complete minimal Möbius strips in $\mathbb{R}^n$ and the Björling problem", *J. Geom. Phys.* **56**:9 (2006), 1506–1515. MR 2007d:53012 Zbl 1107.53007

[Nitsche 1975] J. C. C. Nitsche, *Vorlesungen über Minimalflächen*, Chapter 1–5, pp. 1–430, Grundlehren der mathematischen Wissenschaften **199**, Springer, Berlin, 1975. Translated and revised in *Lectures on minimal surfaces, 1: Introduction, fundamentals, geometry and basic boundary value problems*, Cambridge University Press, 1989. MR 56 #6533 Zbl 0319.53003

[Osserman 1964] R. Osserman, "Global properties of minimal surfaces in E^3 and E^n", *Ann. of Math.* (2) **80** (1964), 340–364. MR 31 #3946 Zbl 0134.38502

[Sato 1996] K. Sato, "Construction of higher genus minimal surfaces with one end and finite total curvature", *Tohoku Math. J.* (2) **48**:2 (1996), 229–246. MR 97d:53008 Zbl 1010.53501

[Schoen 1983] R. M. Schoen, "Uniqueness, symmetry, and embeddedness of minimal surfaces", *J. Differential Geom.* **18**:4 (1983), 791–809. MR 85f:53011 Zbl 0575.53037

[Weber 2015] M. Weber, Virtual minimal surface museum, 2015, http://www.indiana.edu/∼minimal.

[Weber and Wolf 1998] M. Weber and M. Wolf, "Minimal surfaces of least total curvature and moduli spaces of plane polygonal arcs", *Geom. Funct. Anal.* **8**:6 (1998), 1129–1170. MR 99m:53020 Zbl 0954.53007

[Wohlgemuth 1997] M. Wohlgemuth, "Minimal surfaces of higher genus with finite total curvature", *Arch. Rational Mech. Anal.* **137**:1 (1997), 1–25. MR 98j:53015 Zbl 0874.53007

Received April 1, 2015.

SHOICHI FUJIMORI
DEPARTMENT OF MATHEMATICS
OKAYAMA UNIVERSITY
OKAYAMA 700-8530
JAPAN
fujimori@math.okayama-u.ac.jp

TOSHIHIRO SHODA
FACULTY OF CULTURE AND EDUCATION
SAGA UNIVERSITY
1 HONJO-MACHI
SAGA-CITY 840-8502
JAPAN
tshoda@cc.saga-u.ac.jp

PACIFIC JOURNAL OF MATHEMATICS
Vol. 282, No. 1, 2016
dx.doi.org/10.2140/pjm.2016.282.145

MULTIPLICITÉ DU SPECTRE DE STEKLOV SUR LES SURFACES ET NOMBRE CHROMATIQUE

PIERRE JAMMES

On démontre plusieurs résultats sur la multiplicité des premières valeurs propres de Steklov sur les surfaces compactes à bord. On améliore certaines bornes sur la multiplicité, en particulier pour la première valeur propre, et on montre qu'elles sont optimales sur plusieurs surfaces de petit genre. Dans un article précédent, on a défini un nouvel invariant chromatique des surfaces à bord et on a conjecturé qu'il est relié à la multiplicité de la première valeur propre de Steklov. Dans le present article on étudie cet invariant et on démontre une des inégalités de la conjecture.

We prove several results about the multiplicity of the first Steklov eigenvalues on compact surfaces with boundary. We improve some bounds on the multiplicity, especially for the first eigenvalue, and we prove they are sharp on some surfaces of small genus. In a previous article, we defined a new chromatic invariant of surfaces with boundary and conjectured that this invariant is related to the bound on the first eigenvalue. In the present article, we study this invariant, and prove one of the inequalities of this conjecture.

1. Introduction

L'étude de la multiplicité des valeurs propres de Steklov a fait récemment l'objet des travaux [Fraser et Schoen 2015; Jammes 2014; Karpukhin et al. 2014]. Ces trois articles montrent que sur une surface compacte à bord donnée, la multiplicité de la k-ième valeur propre de Steklov est majorée en fonction de k et de la topologie (voir le paragraphe qui suit pour la définition du spectre de Steklov). Dans [Jammes 2014], j'ai montré que ce phénomène est spécifique à la dimension 2 et qu'en dimension plus grande, on peut prescrire arbitrairement le début du spectre de Steklov, avec multiplicité. J'ai aussi conjecturé qu'en dimension 2, la multiplicité maximale de la première valeur propre non nulle est déterminée par un invariant topologique de même nature que le nombre chromatique mais spécifique aux surfaces à bord (cf. la conjecture 1.6 ci dessous).

MSC2010 : 35P15, 57M15, 58J50.
Mots-clefs : Steklov eigenvalues, multiplicity, chromatic number.

L'objet de cet article est triple. D'abord, améliorer certaines bornes sur la multiplicité du spectre de Steklov en dimension 2, en particulier pour la première valeur propre et sur les surfaces de petit genre. Ensuite, construire des exemples de première valeur propre multiple, ce qui permet de montrer que certaines bornes sont optimales. Enfin, on étudie l'invariant chromatique des surfaces à bord définie dans [Jammes 2014], qu'on appellera *nombre chromatique relatif*, on justifiera en particulier cette dénomination et on calculera sa valeur sur les surfaces de petit genre.

La confrontation de ces différents résultats permet de consolider la conjecture faite dans [Jammes 2014] sur le lien entre la multiplicité de la première valeur propre et le nombre chromatique relatif en montrant qu'elle est vérifiée dans tous les cas où on connaît la borne optimale et qui sont résumés dans la table 1 ci-dessous.

1A. ***Définitions et notations.*** Avant d'énoncer les résultats précis, nous allons rappeler les notions en jeu et préciser quelques définitions.

Soit M une variété compacte à bord et $\gamma \in C^\infty(M)$, $\rho \in C^\infty(\partial M)$ deux fonctions densités strictement positives sur M et ∂M respectivement (on peut travailler avec des hypothèses de régularité plus faible sur les densités mais ça n'est pas crucial dans la suite). Le problème aux valeurs propres de Steklov consiste à résoudre l'équation, d'inconnues $\sigma \in \mathbb{R}$ et $f : \overline{M} \to \mathbb{R}$,

$$\begin{cases} \operatorname{div}(\gamma \nabla f) = 0 & \text{dans } M, \\ \gamma \dfrac{\partial f}{\partial \nu} = \sigma \rho f & \text{sur } \partial M, \end{cases} \tag{1.1}$$

où ν est un vecteur unitaire sortant normal au bord. On parle de problème de Steklov homogène quand $\gamma \equiv 1$ et $\rho \equiv 1$, et le cas $\gamma \not\equiv 1$ se rattache au problème de Calderón. L'ensemble des réels σ solutions du problème forme un spectre discret positif noté

$$0 = \sigma_0(M, g, \rho, \gamma) < \sigma_1(M, g, \rho, \gamma) \leq \sigma_2(M, g, \rho, \gamma) \leq \cdots \tag{1.2}$$

On omettra les références à ρ et γ quand ces densités sont uniformément égales à 1.

On s'intéressera à la multiplicité de ces valeurs propres, et si on se donne une surface compacte à bord Σ, on notera $m_k(\Sigma)$ la multiplicité maximale de $\sigma_k(\Sigma, g, \rho, \gamma)$ quand on fait varier g, ρ et γ.

On établira des liens entre ce spectre et les graphes plongés dans la surface. On introduit pour cela les deux définitions suivantes :

Définition 1.3. Soit Γ un graphe fini et Σ une surface compacte à bord. Un plongement de Γ dans Σ sera appelé *plongement propre* si ce plongement envoie tous les sommets de Γ sur $\partial\Sigma$.

Remarque 1.4. Les graphes qui se plongent proprement dans le disque sont connus sous le nom de graphes planaires extérieurs (*outerplanar graphs*).

Définition 1.5. Si Σ est une surface compacte à bord, on appelle *nombre chromatique relatif* de Σ, noté $\mathrm{Chr}_0(\Sigma)$, la borne supérieure des nombres chromatiques des graphes finis qui admettent un plongement propre dans Σ.

Avec ces notations, la conjecture énoncée dans [Jammes 2014] peut se reformuler ainsi (l'équivalence entre les deux énoncés est l'un des objets de la section 3, cf. remarque 3.4) :

Conjecture 1.6. Pour toute surface Σ compacte à bord, on a

$$m_1(\Sigma) = \mathrm{Chr}_0(\Sigma) - 1.$$

Cette conjecture adapte au spectre de Steklov une conjecture analogue énoncée par Y. Colin de Verdière dans [Colin de Verdière 1987] et reliant la multiplicité de la 2-ième valeur propre des opérateurs de Schrödinger et le nombre chromatique usuel.

Enfin, on utilisera les notations suivantes : $\mathbb{S}^2$, $\mathbb{D}^2$, $\mathbb{P}^2$, $\mathbb{M}^2$, $\mathbb{T}^2$ et $\mathbb{K}^2$ désigneront respectivement la sphère de dimension 2, le disque, le plan projectif, le ruban de Möbius, le tore et la bouteille de Klein. Si Σ est une surface close, on désignera par Σ_p la surface Σ privée de p disques disjoints, où p est un entier strictement positif. Par exemple, on a $\mathbb{M}^2 = \mathbb{P}^2_1$ et $\mathbb{D}^2 = \mathbb{S}^2_1$.

1B. *Résultats.* Un premier résultat de cet article est de démontrer qu'on a au moins une inégalité dans l'égalité de la conjecture 1.6 :

Théorème 1.7. *Pour toute surface Σ une surface compacte à bord*, *on a*

$$m_1(\Sigma) \geq \mathrm{Chr}_0(\Sigma) - 1.$$

On va s'attacher ensuite à améliorer les bornes connues sur la multiplicité des valeurs propres, en commençant par considérer une surface quelconque. Rappelons d'abord le résultat de [Karpukhin et al. 2014], qui est la meilleure majoration connue pour une surface Σ et une valeur propre σ_k quelconques (elle est démontré dans [Karpukhin et al. 2014] pour $\gamma \equiv 1$ mais sans hypothèse de régularité sur ρ) :

Théorème 1.8 [Karpukhin et al. 2014]. *Si Σ est une surface compacte à bord de caractéristique d'Euler χ et dont le bord possède l composantes connexes*, *alors*

$$m_k(\Sigma) \leq 2k - 2\chi - 2l + 5 \tag{1.9}$$

et

$$m_k(\Sigma) \leq k - 2\chi + 4 \tag{1.10}$$

pour tout $k \geq 1$, *cette dernière égalité étant stricte pour les surfaces non simplement connexes.*

Les majorations de multiplicité que nous allons montrer sont de plusieurs types. Dans les deux théorèmes qui suivent, nous donnons deux majorations de m_k sur n'importe quelle surface : le premier précise l'inégalité (1.10) en utilisant les outils développés dans [Fraser et Schoen 2015], le deuxième théorème est une application des techniques développées par B. Sévennec [1994; 2002] pour majorer la multiplicité de la première valeur propre du laplacien. Nous donnerons ensuite des majorations spécifiques aux premières valeurs propres des surfaces de petit genre.

Théorème 1.11. *Si Σ est une surface compacte à bord de caractéristique d'Euler χ et dont le bord possède l composantes connexes, alors pour tout $k \geq 1$:*

$$m_k(\Sigma) \leq k - 2\chi + 3. \tag{1.12}$$

Remarque 1.13. La majoration (1.12) est optimale pour $k = 1$ sur le disque et le ruban de Möbius, et dans les deux cas l'égalité est atteinte par un problème homogène, c'est-à-dire avec $\rho \equiv 1$ et $\gamma \equiv 1$ (pour la métrique canonique dans le cas du disque et pour une métrique S^1-invariante construite dans [Fraser et Schoen 2015] pour le ruban de Möbius).

Remarque 1.14. L'inégalité (1.12) étend les cas d'inégalité stricte de (1.10) à toutes les surfaces. Les techniques de [Fraser et Schoen 2015] qui autorisent cette démonstration unifiée permettent aussi les raffinements obtenus dans les théorèmes 1.20 et 1.21.

Contrairement au théorème 1.11 et aux travaux [Fraser et Schoen 2015; Karpukhin et al. 2014; Jammes 2014], le théorème qui suit n'utilise pas les résultats de Cheng [1976] sur la structure de l'ensemble nodal des fonctions propres mais les techniques topologiques de Sévennec, ce qui autorise des hypothèses de régularité beaucoup plus faibles (voir le théorème 5.11 et la remarque 5.13).

Théorème 1.15. *Soit Σ une surface compacte à bord de caractéristique d'Euler χ et dont le bord possède l composantes connexes. Si $\chi + l \leq -1$, alors*

$$m_1(\Sigma) \leq 5 - \chi - l. \tag{1.16}$$

Si $l = 1$ et $\chi \leq -2$, alors

$$m_1(\Sigma) \leq 3 - \chi. \tag{1.17}$$

Ces majorations restent valables si g, ρ et γ sont L^∞ sans régularité supplémentaire.

Remarque 1.18. Si on note $\bar{\chi}$ la caractéristique d'Euler de la surface close obtenue en collant un disque sur chaque bord de Σ, la majoration (1.16) s'écrit $m_1(\Sigma) \leq 5 - \bar{\chi}$. Elle est optimale pour $\bar{\chi} = -1$ et $l \geq 2$, et $\bar{\chi} = -2$ ou -3 et $l \geq 3$. L'inégalité (1.17)

$p=$	1	2	3	4	$p=$	1	2	3	4
$\mathbb{S}^2$	2	3	3	3	$\#3\mathbb{P}^2$	5	6	6	6
$\mathbb{P}^2$	4	4	5	5	$\mathbb{T}^2\#\mathbb{T}^2$	6	7	7	7
$\mathbb{K}^2$	?	5	5	5	$\#4\mathbb{P}^2$	6	?	7	7
$\mathbb{T}^2$	5	?	6	6	$\#5\mathbb{P}^2$	7	?	8	8

Table 1. Valeur de m_1 sur les surfaces de petit genre.

peut s'écrire $m_1(\Sigma) \leq 4 - \bar{\chi}$, elle est optimale pour $\bar{\chi} = 1, 2, 3$. Les valeurs propres multiples correspondantes sont données par le théorème 1.7 et le calcul des nombres chromatiques relatifs fait dans la section 3.

Dans les trois théorèmes qui suivent, on donne des bornes spécifiques aux surfaces de petit genre. Le premier adapte des résultats analogues obtenus par G. Besson [1980] et Colin de Verdière [1987] pour le laplacien.

Théorème 1.19. *$m_1(\mathbb{T}^2_p) = 6$ pour $p \geq 3$ et $m_1(\mathbb{K}^2_p) = 5$ pour $p \geq 2$.*

Avec des techniques similaires mais en utilisant des arguments spécifiques au spectre de Steklov, on améliore deux bornes :

Théorème 1.20. *$m_1(\mathbb{P}^2_2) = 4$ et $m_1(\mathbb{T}^2_1) = 5$.*

Enfin, on majore la multiplicité de la 2-ième valeur propre du disque au moyen d'une stratégie indépendante de toutes celles évoquées jusqu'ici.

Théorème 1.21. *$m_2(\mathbb{D}^2) = 2$.*

Remarque 1.22. On sait déjà que $m_1(\mathbb{D}^2) = 2$. Par ailleurs, M. Karpukhin, G. Kokarev et I. Polterovich [2014] montrent que $m_k(\mathbb{D}^2) = 2$ si k est assez grand. Ces résultats laissent penser qu'on a $m_k(\mathbb{D}^2) = 2$ quel que soit k. Cette multiplicité est atteinte par la métrique canonique.

La table 1 rassemble les valeurs connues de $m_1(\Sigma_p)$, où Σ est une surface close (dans cette table, la notation $\#n\Sigma$ désigne la somme connexe de n copies de la surface Σ). Pour les surfaces Σ concernées, la valeur de m_1 ne dépend pas de p quand p≥ 4. Rappelons que la majoration de $\mathbb{S}^2_p$ pour $p \geq 2$ est montrée dans [Fraser et Schoen 2015; Jammes 2014; Karpukhin et al. 2014] et l'égalité découle du théorème 1.7 pour p quelconque et d'un exemple construit dans [Fraser et Schoen 2011] pour le cas $p = 2$.

Toutes les bornes contenues dans la table 1 sont bien conformes à la conjecture 1.6 (comparer avec les nombres chromatiques relatifs indiqués dans la table 2, section 3).

Les majorations de multiplicité pour la première valeur propre du laplacien induisent des critères de plongement de graphes dans les surfaces. Dans le cas du spectre de Steklov, on peut en tirer des critères de plongement propre dans les surfaces à bord. Cet aspect sera précisé dans la section 4 (corollaire 4.2). Comme par ailleurs on sait caractériser spectralement les graphes qui admettent un plongement non entrelacé dans $\mathbb{R}^3$, on obtient, par un cheminement inattendu, le critère de plongement qui suit. Rappelons qu'un plongement d'un graphe dans $\mathbb{R}^3$ est non entrelacé si tout ensemble de cycles disjoints forme un entrelacs trivial.

Corollaire 1.23. *Si un graphe admet un plongement propre dans $\mathbb{M}^2$ ou $\mathbb{P}^2_2$, alors il admet un plongement non entrelacé dans $\mathbb{R}^3$.*

Remarque 1.24. L'hypothèse que les sommets du graphe soient sur le bord est indispensable ; sans elle on peut trouver facilement des contre-exemples, à commencer par le graphe complet à 6 sommets.

Remarque 1.25. Si la conjecture 1.6 est vraie pour $\mathbb{K}^2_1$, alors le corollaire 1.23 s'applique aussi à cette surface.

On discutera dans la section 4 (remarque 4.8) l'existence de démonstrations plus élémentaires de ce corollaire.

Dans la section 2A, nous rappellerons quelques résultats techniques concernant le spectre de Steklov et les opérateurs sur les graphes. La section 3 sera consacrée à l'étude du nombre chromatique relatif : on justifiera en particulier cette dénomination et on calculera sa valeur sur les surfaces de petit genre. On verra dans la section 4 comment construire des valeurs propres multiples sur une surface à l'aide d'un graphe proprement plongé et on en déduira le théorème 1.7 et le corollaire 1.23. Enfin, on démontrera dans la section 5 les différentes bornes sur la multiplicité.

2. Rappels

2A. ***Le spectre de Steklov.*** On va rappeler ici quelques propriétés du spectre de Steklov que nous utiliserons. Il est l'ensemble des réels σ pour lesquels le problème

$$\begin{cases} \operatorname{div}(\gamma\nabla f)=0 & \text{dans } M, \\ \gamma\dfrac{\partial f}{\partial \nu}=\sigma\rho f & \text{sur } \partial M, \end{cases} \tag{2.1}$$

où ν est un vecteur unitaire sortant normal au bord, admet des solutions non triviales. Les variétés que nous considérerons auront un bord C^1 par morceaux, ce qui est suffisant pour que le problème soit bien défini.

Le spectre de Steklov $(\sigma_k(M,g,\gamma,\rho))_k$ est le spectre d'un opérateur Dirichlet–Neumann $H^1(\partial M)\to L^2(\partial M)$ défini par $\Lambda_{\rho,\gamma}u=\frac{\gamma}{\rho}\frac{\partial\mathcal{H}_\gamma u}{\partial\nu}$, où $\mathcal{H}_\gamma u$ est le prolongement harmonique de u pour la densité γ, c'est-à-dire que $\operatorname{div}(\gamma\nabla(\mathcal{H}_\gamma u))=0$. Il est auto-adjoint pour la norme de Hilbert $\|u\|^2=\int_{\partial M}u^2\rho\,\mathrm{d}v_g$ (voir [Bandle 1980; Sylvester et Uhlmann 1990; Uhlmann 2009] ; ces références traitent les cas $\rho\equiv 1$ ou $\gamma\equiv 1$ mais l'adaptation au cas général est aisée).

Pour montrer le théorème 1.7 on utilisera la caractérisation variationnelle suivante du spectre :

$$\sigma_k(M, g, \rho) = \inf_{V_{k+1} \in H^1(M)} \sup_{f \in V_{k+1} \setminus \{0\}} \frac{\int_M |\mathrm{d}f|^2 \gamma \,\mathrm{d}v_g}{\int_{\partial M} f^2 \rho \,\mathrm{d}v_g}, \tag{2.2}$$

où V_k parcours les sous-espaces de dimension k de l'espace de Sobolev $H^1(M)$.

On aura aussi recours à un problème de Steklov avec condition de Neumann sur une partie du bord. Si on partitionne ∂M en deux domaines (ou unions finies de domaines) $\partial_S M$ et $\partial_N M$, ce problème consiste à considérer la variante suivante du problème de Steklov :

$$\begin{cases} \operatorname{div}(\gamma \nabla f) = 0 & \text{dans } M, \\ \gamma \dfrac{\partial f}{\partial \nu} = \sigma \rho f & \text{sur } \partial M_S, \\ \dfrac{\partial f}{\partial \nu} = 0 & \text{sur } \partial M_N, \end{cases} \tag{2.3}$$

c'est-à-dire qu'on demande à la fonction harmonique f de vérifier la condition de Neumann sur ∂M_N. Cette condition revient à poser $\rho \equiv 0$ sur ∂M_N. Le spectre obtenu est celui d'un opérateur Dirichlet–Neumann défini sur $\partial_S M$. On notera $(\sigma_k(M, \partial M_S, g, \gamma, \rho))_k$ son spectre.

2B. ***Opérateurs sur les graphes.*** On rappelle ici la définition des laplaciens et des opérateurs de Schrödinger sur les graphes, en se référant par exemple à [Colin de Verdière 1998].

Soit Γ un graphe fini et S son ensemble de sommets. La géométrie du graphe est déterminée par la donnée, pour chaque arête a reliant deux sommets x et y, d'un réel l_a qui représente la longueur de cette arête. L'ensemble des l_a sera appelé une métrique sur le graphe Γ. Le laplacien agissant sur les fonctions $f : S \to S$ est alors défini par

$$\Delta f(x) = \sum_{a \sim x} \frac{f(x) - f(y_a)}{l_a}, \tag{2.4}$$

où la somme porte sur l'ensemble des arêtes d'extrémité x et où y_a désigne l'autre extrémité de a. Cet opérateur a un spectre positif, sa plus petite valeur propre est simple et vaut 0.

Si on se donne une fonction V sur l'ensemble des sommets, on peut définir un opérateur de Schrödinger H_V sur Γ par $H_V f(x) = \Delta f(x) + V(x) f(x)$. Une propriété de ces opérateurs nous sera utile : si la première valeur propre de H_V est nulle, alors son spectre est celui de la forme quadratique d'un laplacien relativement à une norme de Hilbert $|f| = \sum_{x \in S} \mu_x f^2(x)$. Les coefficients μ_x s'interprètent comme une mesure sur l'espace des sommets du graphe.

3. Nombre chromatique relatif d'une surface à bord

On va montrer dans cette section différents résultats concernant le nombre chromatique relatif des surfaces à bord, et en particulier calculer cet invariant sur les surfaces de petit genre.

Rappelons d'abord que le nombre chromatique d'un surface close Σ, c'est-à-dire la borne supérieure des nombres chromatiques des graphes qu'on peut plonger dans Σ, est donné par la formule

$$\operatorname{Chr}(\Sigma) = \left\lfloor \frac{7+\sqrt{49-24\chi(\Sigma)}}{2} \right\rfloor, \tag{3.1}$$

sauf pour la bouteille de Klein, pour laquelle $\operatorname{Chr}(\mathbb{K}^2) = 6$. L'étude de cet invariant, amorcée par P. J. Heawood [1890] et L. Heffter [1891] a été achevée par G. Ringel et J. Youngs [1968] en genre non nul, et K. Appel et W. Haken [1976] pour la sphère. On peut consulter [Ringel et Youngs 1968] pour un survol historique.

Notre premier résultat sur le nombre chromatique relatif est un encadrement analogue à la majoration du nombre chromatique obtenue par Heawood [1890]. Rappelons que si Σ est une surface close, on note Σ_p la surface obtenue en lui enlevant p disques disjoints.

Théorème 3.2. *Le nombre chromatique relatif* $\operatorname{Chr}_0(\Sigma_p)$ *possède les propriétés suivantes* :

(a) $\operatorname{Chr}_0(\Sigma_p)$ *est une fonction croissante de* p *et vérifie les inégalités*

$$\operatorname{Chr}(\Sigma)-1 \leq \operatorname{Chr}_0(\Sigma_p) \leq \inf\left(\operatorname{Chr}(\Sigma), \frac{5+\sqrt{25-24\chi(\Sigma)+24p}}{2}\right). \tag{3.3}$$

(b) $\operatorname{Chr}_0(\Sigma_p)$ *est le nombre de sommets du plus grand graphe complet proprement plongeable dans* Σ_p.

(c) $\operatorname{Chr}_0(\Sigma_1) = \operatorname{Chr}(\Sigma)-1$ *et* $\operatorname{Chr}_0(\Sigma_p) = \operatorname{Chr}(\Sigma)$ *si* $p \geq (\operatorname{Chr}(\Sigma)-1)/2$.

Remarque 3.4. Le fait que $\operatorname{Chr}_0(\Sigma_p)$ soit réalisé par un graphe complet établit l'équivalence entre la conjecture 1.6 et la conjecture énoncée dans [Jammes 2014].

Ce théorème permet de calculer la valeur exacte du nombre chromatique relatif sur un certain nombre de surfaces. On va compléter cette liste par des surfaces de petit genre. D'abord dans le cas où la caractéristique d'Euler de Σ vérifie $\chi(\Sigma) \geq -1$. On aura alors la valeur de $\operatorname{Chr}_0(\Sigma_p)$ pour tout p :

Proposition 3.5. $\operatorname{Chr}_0(\mathbb{P}^2_2) = 5$, $\operatorname{Chr}_0(\mathbb{K}^2_2) = 6$ *et* $\operatorname{Chr}_0(\#3\mathbb{P}^2_2) = 7$.

On va aussi calculer quelques nombres chromatiques relatifs supplémentaires dans le cas où $-2 \geq \chi(\Sigma) \geq -7$. La liste ne sera pas exhaustive mais elle contiendra tous les cas pour lesquels on sait démontrer la conjecture 1.6 :

$p =$	1	2	3	4	5
$\mathbb{S}^2$	3	4	4	4	4
$\mathbb{P}^2$	5	5	6	6	6
$\mathbb{K}^2$	5	6	6	6	6
$\mathbb{T}^2$	6	6	7	7	7
$\chi = -1$	6	7	7	7	7
$\#2\mathbb{T}^2$	7	8	8	8	8

$p =$	1	2	3	4	5
$\#4\mathbb{P}^2$	7	?	8	8	8
$\chi = -3$	8	8	9	9	9
$\chi = -4$	8	?	9	9	9
$\chi = -5$	9	9	9	10	10
$\#4\mathbb{T}^2$	9	9	10	10	10
$\chi = -7$	9	?	10	10	10

Table 2. Nombre chromatique relatif des surfaces de petit genre.

Proposition 3.6. *Soit* Σ *une surface close.*

(1) $\mathrm{Chr}_0(\#2\mathbb{T}^2_2) = \mathrm{Chr}_0(\#4\mathbb{P}^2_3) = 8$.

(2) *Si* $\chi(\Sigma) = -3$ *ou* -4, *alors* $\mathrm{Chr}_0(\Sigma_3) = 9$.

(3) *Si* $\chi(\Sigma) = -5$ *alors* $\mathrm{Chr}_0(\Sigma_4) = 10$.

(4) $\mathrm{Chr}_0(\#4\mathbb{T}^2_3) = 10$.

La table 2 rassemble en fonction de Σ et p les nombres chromatiques qu'on peut calculer à l'aide du théorème 3.2, des propositions 3.5 et 3.6 et en utilisant la monotonie du nombre chromatique relatif par rapport à p et par somme connexe. On se limite aux surfaces Σ telles que $\chi(\Sigma) \geq -7$.

On va montrer séparément le résultat le plus technique du théorème 3.2 :

Lemme 3.7.
$$\mathrm{Chr}_0(\Sigma_p) \leq \frac{5 + \sqrt{25 - 24\chi(\Sigma) + 24p}}{2}.$$

Démonstration. On pose $c_p = \lfloor (5 + \sqrt{25 - 24\chi(\Sigma) + 24p})/2 \rfloor$. Étant donné un graphe Γ plongé proprement dans Σ_p, on va montrer qu'on peut le colorier avec c_p couleurs. Quitte à ajouter (temporairement) des sommets et des arêtes, on peut supposer d'une part que toutes les faces de la décomposition de Σ_p induite par Γ sont simplement connexes, et d'autre part que $\partial\Sigma_p$ est recouvert par des arêtes. Il peut apparaître des arêtes multiples (plusieurs arêtes ayant les mêmes sommets) ou des boucles (arête reliant un sommet à lui-même) lors de cette étape mais ça n'est pas gênant pour la suite. On note s et a le nombre de sommets et d'arêtes de Γ, δ le degré minimal de ses sommets et f le nombre de faces de la décomposition de Σ_p induite par Γ.

On commence par majorer δ. Comme chaque face est bordée par au moins trois arêtes et que chaque arête (sauf celles qui sont sur le bord, qui sont en nombre égal au nombre de sommets) est adjacente à deux faces, on a $3f \leq 2a - s$. De plus, le degré minimal vérifie les inégalités $\delta s \leq 2a$ et $\delta + 1 \leq s$. Il découle alors de la formule d'Euler–Poincaré, en notant $\chi = \chi(\Sigma) - p$ la caractéristique d'Euler de

Σ_p et en supposant que $\delta \geq 4$, que

$$\begin{aligned} 6\chi &= 6f + 6s - 6a \leq 4s - 2a \\ &\leq 4s - \delta s = (4-\delta)s \leq (4-\delta)(\delta+1) \\ &\leq -\delta^2 + 3\delta + 4. \end{aligned} \tag{3.8}$$

Comme δ vérifie l'inéquation $\delta^2 - 3\delta - 4 + 6\chi \leq 0$, on en déduit que

$$\delta \leq \lfloor (3 + \sqrt{25 - 24\chi})/2 \rfloor = c_p - 1.$$

Si $\delta \leq 4$, cette inégalité est trivialement vérifiée dès que $\chi \leq 0$.

On supprime ici les boucles qui sont apparues en ajoutant des arêtes dans l'étape précédente ; la majoration de δ reste valide.

On construit ensuite un coloriage par une double récurrence sur p, et sur s à p fixé. Dans la récurrence sur s, on aura besoin que la majoration de δ s'applique bien à tous les graphes considérés.

Si $s \leq c_p$, le graphe Γ se colorie évidemment avec c_p couleurs. Dans le cas contraire, on note x un sommet de degré minimal et on considère le graphe Γ' obtenu en supprimant le sommet x et les arêtes qui lui sont adjacentes. La récurrence se décompose en deux cas, selon que la composante de bord où est situé x porte d'autres sommets de Γ où non. Soulignons que si $p = 1$, tous les sommets sont sur la même composante de bord donc seul le premier cas se présentera.

Dans le premier cas, on ajoute une arête qui joint les deux voisins de x sur la même composante de bord (s'il ne reste qu'un seul point sur cette composante, l'arête supplémentaire forme une boucle) et on appelle encore Γ' le graphe obtenu. La démonstration de la majoration $\delta \leq c_p - 1$ obtenue pour Γ s'applique alors aussi à Γ'. Comme Γ' a $s-1$ sommets, on peut lui appliquer l'hypothèse de récurrence et le colorier avec c_p couleurs. On applique le même coloriage à tous les sommets de Γ sauf x, et comme le degré δ de x est majoré par $c_p - 1$ on peut colorier x avec une couleur différente de ses voisins.

Dans le second cas, la composante de bord de x ne porte aucun sommet de Γ' et on se ramène à la surface Σ_{p-1} en collant un disque le long de cette composante de bord. Le graphe Γ' est proprement plongé dans Σ_{p-1}, donc admet un coloriage à c_{p-1} couleurs par hypothèse de récurrence. Comme la constante c_p croît avec p il admet un coloriage à c_p couleurs. On conclut comme dans le cas précédent. □

Démonstration du théorème 3.2. On va noter temporairement $\kappa(\Sigma_p)$ le nombre de sommets du plus grand graphe complet proprement plongeable dans Σ_p. Il est clair qu'un plongement dans Σ_p induit un plongement dans Σ_{p+1} et dans Σ, donc $\kappa(\Sigma_p)$ et $\mathrm{Chr}_0(\Sigma_p)$ sont des fonctions croissantes de p et

$$\kappa(\Sigma_p) \leq \mathrm{Chr}_0(\Sigma_p) \leq \mathrm{Chr}(\Sigma). \tag{3.9}$$

Supposons que $p=1$. Si on se donne un plongement propre du graphe complet K_n à n sommets dans Σ_1, on peut construire un plongement de K_{n+1} dans Σ en collant un disque sur le bord de Σ_1, en ajoutant un sommet sur ce disque et en le reliant par des arêtes aux sommets de K_n (c'est possible puisqu'ils sont tous sur le bord du disque). Réciproquement, étant donné un graphe complet K_{n+1} étant plongé dans Σ, on peut enlever un disque contenant un seul sommet et déplacer les autres sommets de manière à ce qu'ils se situent sur le bord du disque. On obtient un plongement de K_n dans Σ_1 dont les sommets sont sur $\partial\Sigma_1$. On en déduit que $\kappa(\Sigma_1)=\mathrm{Chr}(\Sigma)-1$ et donc que

$$\mathrm{Chr}(\Sigma)-1 \leq \kappa(\Sigma_p) \leq \mathrm{Chr}_0(\Sigma_p) \leq \mathrm{Chr}(\Sigma), \tag{3.10}$$

ce qui, avec le lemme 3.7, conclut la démonstration du point (a) du théorème.

On peut maintenant montrer le point (b), c'est-à-dire que $\kappa(\Sigma_p)=\mathrm{Chr}_0(\Sigma_p)$. Ces deux nombres ne peuvent prendre que les valeurs $\mathrm{Chr}(\Sigma)-1$ et $\mathrm{Chr}(\Sigma)$. Si $\mathrm{Chr}_0(\Sigma_p)=\mathrm{Chr}(\Sigma)-1$, alors on a immédiatement l'égalité dans les deux premières inégalités de (3.10). Supposons donc que $\mathrm{Chr}_0(\Sigma_p)=\mathrm{Chr}(\Sigma)$ et considérons un graphe Γ proprement plongé dans Σ_p, dont le nombre chromatique est $\mathrm{Chr}(\Sigma)$ et qui est critique, c'est-à-dire qu'on ne peut pas lui enlever d'arêtes sans diminuer son nombre chromatique. Cette propriété de criticité est intrinsèque au graphe Γ, donc il est aussi critique comme graphe plongé dans Σ. Or, on sait qu'un tel graphe est nécessairement un graphe complet (cf. [Bollobás 1998, Chapter V, p. 156–157]), donc on a égalité dans l'inégalité (3.9).

On a déjà montré que $\mathrm{Chr}_0(\Sigma_1)=\kappa(\Sigma_1)=\mathrm{Chr}(\Sigma)-1$. Il reste à montrer que $\mathrm{Chr}_0(\Sigma_p)=\mathrm{Chr}(\Sigma)$ si $p\geq(\mathrm{Chr}(\Sigma)-1)/2$ pour conclure le point (c) du théorème. On considère le graphe complet à $\mathrm{Chr}(\Sigma)$ sommets qu'on plonge dans Σ et on construit un plongement de ce graphe dans Σ_p en enlevant p disques (pour p suffisamment grand) à Σ de manière à ce que les disques ne rencontrent pas les arêtes et que les sommets soient situés sur le bord des disques. Pour déterminer une valeur de p adéquate, on remarque qu'en plaçant le premier disque de manière quelconque, on peut placer (au moins) trois sommets du graphe sur son bord. On peut ensuite regrouper les autres sommets par deux et placer un disque adjacent à ces deux sommets et qui longe l'arête qui les relie. En fonction de la parité du nombre de sommets, il peut en rester un qui nécessite un disque supplémentaire. Au total, on a utilisé $\lceil(\mathrm{Chr}(\Sigma)-1)/2\rceil$ disques. □

Démonstration de la proposition 3.5. Le fait que $\mathrm{Chr}_0(\mathbb{P}^2_2)=5$ découle des théorèmes 1.7 et 1.20.

Le calcul de $\mathrm{Chr}_0(\mathbb{K}^2_2)$ repose sur une amélioration de l'estimation faite par le théorème 3.2 dans le cas où $p\geq(\mathrm{Chr}(\Sigma)-1)/2$ en utilisant le plongement explicite d'un graphe complet dans la surface. Dans le cas de la bouteille de Klein,

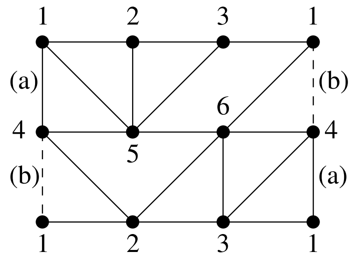


Figure 1. Plongement de K_6 dans $\mathbb{K}^2$.

le plongement du graphe complet à 6 sommets est représenté sur la figure 1 (les lettres indiquent comment sont identifiés les cotés droite et gauche).

En enlevant les faces (125) et (346), on obtient un plongement propre de K_6 dans $\mathbb{K}^2_2$.

Il reste à calculer $\mathrm{Chr}_0(\#3\mathbb{P}^2_2)$. On sait déjà que $\mathrm{Chr}_0(\#3\mathbb{P}^2_2) \leq 7$ (théorème 3.2), il suffit donc d'exhiber un plongement propre de K_7. On procède de la manière suivante : topologiquement, on enlève deux disques au ruban de Möbius $\mathbb{M}^2$ et on colle un pantalon (une sphère privée de trois disques) sur les deux composantes de bord ainsi crées. Pour construire le graphe, on part du graphe K_5 plongé dans le ruban de Möbius et on ajoute deux sommets situés sur le troisième bord du pantalon. Les détails de la construction sont représentés sur la figure 2.

En haut de la figure est représenté le graphe K_5 plongé dans $\mathbb{M}^2$. On enlève deux disques situés dans les faces (123) et (451). En bas est représenté le pantalon, les deux bords intérieurs étant recollés sur $\mathbb{M}^2$. Les sommets 6 et 7 sont bien reliés entre eux et aux cinq premiers sommets. □

Démonstration de la proposition 3.6. Pour calculer les nombres chromatiques traités ici, on procède comme on l'a fait pour $\mathrm{Chr}_0(\mathbb{K}^2_2)$: on considère le découpage de

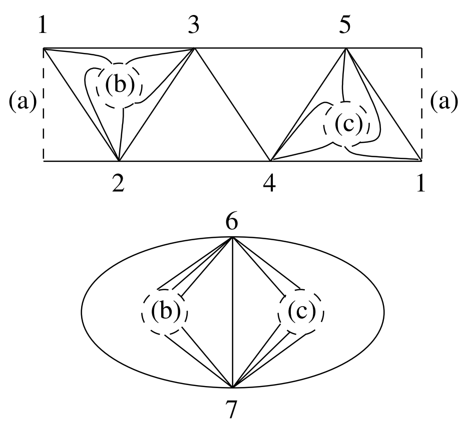


Figure 2. Plongement propre de K_7 dans $\#3\mathbb{P}^2_2$.

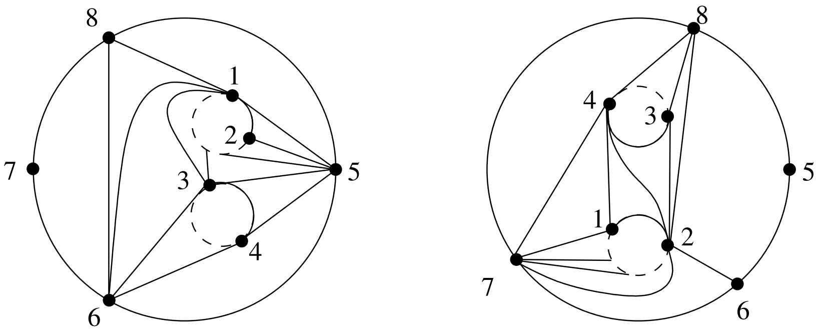


Figure 3. Plongement de K_8 dans $\mathbb{T}^2\#\mathbb{T}^2$.

Σ induit par le plongement d'un graphe complet maximal et on cherche comment enlever des disques de manière à placer tous les sommets sur le bord.

On commence par le cas $\chi = -2$. Des plongements de K_8 dans $\mathbb{T}^2\#\mathbb{T}^2$ et $\#4\mathbb{P}^2_2$ ont été décrits respectivement par L. Heffter [1891] et I. N. Kagno [1935]. La figure 3 représente un plongement de K_8 dans $\mathbb{T}^2\#\mathbb{T}^2$: la surface $\mathbb{T}^2\#\mathbb{T}^2$ est obtenue en recollant les deux pantalons représentés sur la figure le long de leurs bords, la position des sommets de K_8 indiquant l'orientation du recollement. On peut remarquer que l'arête (34) peut être placée arbitrairement d'un coté ou de l'autre du bord des pantalons portant les deux sommets. En fonction de ce choix, on obtient une surface orientable ou non.

On peut noter que la décomposition de la surface induite par le plongement comporte un quadrilatère (5628). Si on enlève cette face, on peut ensuite enlever deux autres disques, par exemple le long des arêtes (13) et (47), de manière à obtenir un plongement propre de K_8 dans $(\mathbb{T}^2\#\mathbb{T}^2)_3$ ou $(\#4\mathbb{P}^2)_3$ selon le choix d'orientation. Par conséquent, $\mathrm{Chr}_0((\mathbb{T}^2\#\mathbb{T}^2)_3) = \mathrm{Chr}_0(\#4\mathbb{P}^2_3) = 8$.

La figure 3 ne permet pas de montrer que $\mathrm{Chr}_0((\mathbb{T}^2\#\mathbb{T}^2)_2) = 8$ car les huit sommets ne sont pas sur le bord de deux quadrilatères. Cependant, le plongement de K_8 dans $(\mathbb{T}^2\#\mathbb{T}^2)_2$ décrit dans [Ringel 1974, p. 23, Table (2.9)] vérifie cette propriété, les deux quadrilatères étant (0246) et (1357).

Le plongement de K_9 dans $\#5\mathbb{P}^2$ a été construit par H.S.M. Coxeter [1943]. Ce plongement triangule la surface et on peut vérifier qu'on peut trouver trois faces dont les sommets sont tous distincts (par exemple les faces (129), (678) et (345) de la figure 11 de [Coxeter 1943]). En enlevant ces faces, on obtient un plongement propre de K_9 dans $\#5\mathbb{P}^2_3$.

Le plongement de K_9 dans $\#3\mathbb{T}^2$ exhibé dans [Heffter 1891] comporte une face hexagonale ayant 5 sommets distincts. En enlevant cette face et deux autres disques, on peut donc obtenir un plongement propre de K_9 dans $\#3\mathbb{T}^2_3$. Le cas de $\#6\mathbb{P}^2_3$ s'obtient par somme connexe de $\#5\mathbb{P}^2_3$ avec $\mathbb{P}^2$.

R. C. Bose [1939] a donné la construction d'un plongement de K_{10} dans $\#7\mathbb{P}^2_3$, elle est aussi étudiée (et illustrée par une figure) par Coxeter [1943]. Le graphe triangule la surface et on peut trouver deux triangles n'ayant pas de sommets communs (voir la figure 14 de [Coxeter 1943]). Les quatre autres sommets du graphe peuvent se placer sur le bord de deux disques ; en enlevant quatre disques, on peut donc bien placer tous les sommets sur le bord.

Le plongement de K_{10} dans $\#4\mathbb{T}^2$ exhibé dans [Heffter 1891] comporte des quadrilatères, chacun ayant ses quatre sommets distincts. Après avoir enlevé un de ces quadrilatères, il reste 6 sommets à placer sur un bord, ce qu'on peut faire en enlevant deux autres faces triangulaires (les trois faces sont $(2, 1, 10, 4)$, $(3, 6, 8)$, $(1, 4, 5)$ dans la construction de Heffter). □

4. Construction de valeurs propres multiples

Cette section est consacrée à la démonstration du théorème 1.7. Elle est très similaire à la démonstration du théorème 1.2 de [Jammes 2014]. On va donc rappeler les différentes étapes en indiquant les modifications à apporter. Le principe consiste à faire tendre le début du spectre de la surface vers celui d'un laplacien combinatoire d'un graphe complet, puis d'utiliser la propriété de stabilité du spectre de ce graphe (cf. [Colin de Verdière 1988] les rappels ci dessous) pour exhiber une valeur propre multiple sur la surface.

Par cette méthode on va en fait montrer un résultat analogue pour un graphe fini quelconque, cela nous sera utile en particulier pour démontrer le corollaire 1.23. L'énoncé général fait intervenir l'invariant de graphe forgé par Colin de Verdière [1990] (voir aussi [Colin de Verdière 1998]). Rappelons que cet invariant, qu'on notera $\mu(\Gamma)$, peut se définir comme étant la multiplicité maximale de la deuxième valeur propre des opérateurs de Schrödinger combinatoires sur le graphe Γ pour lesquels cette valeur propre multiple vérifie l'hypothèse de transversalité d'Arnol'd formalisée dans [Colin de Verdière 1988] (on dit alors que la multiplicité est stable). Si on note M la matrice de cet opérateur de Schrödinger, cette hypothèse de tranversalité consiste en ce que dans l'espace des matrices symétriques , le sous-espace des matrices d'opérateurs de Schrödinger et le sous-espace des matrices possédant une deuxième valeur propre de même multiplicité se coupent transversalement. En pratique, on manipulera des laplaciens à densité sur le graphe plutôt que des opérateurs de Schrödinger, les deux points de vue étant équivalents (cf. section 2B). On peut consulter [Colin de Verdière 1988 ; 1998] ou [Jammes 2009] pour une description plus détaillée de cette propriété de transversalité et de son utilisation.

Théorème 4.1. *Soit Γ un graphe admettant un plongement propre dans une surface compacte à bord Σ. Il existe des densités $\gamma \in C^\infty(\Sigma)$, $\rho \in C^\infty(\partial\Sigma)$ et une métrique g sur Σ telles que la multiplicité de $\sigma_1(\Sigma, g, \rho, \gamma)$ égale $\mu(\Gamma)$.*

Le théorème 1.7 s'en déduit en remarquant d'une part qu'un graphe complet à n sommets vérifie $\mu(K_n) = n-1$ [Colin de Verdière 1988, section 4, théorème 1], et d'autre part que si $n = \mathrm{Chr}_0(\Sigma)$, alors K_n admet un plongement propre dans Σ (théorème 3.2 de la section précédente).

On peut aussi en déduire le critère de plongement qui suit :

Corollaire 4.2. *Si Γ admet un plongement propre dans Σ, alors $\mu(\Gamma) \leq m_1(\Sigma)$.*

Le corollaire 1.23 est une conséquence de ce dernier ; il sera traité à la section 4A.

La première étape de la démonstration du théorème 4.1 consiste à construire une variété dont le début du spectre de Steklov tend vers celui d'un laplacien combinatoire pour une métrique donnée sur Γ, cette variété pouvant s'interpréter comme un voisinage tubulaire de Γ. En dimension 2, cette variété, qu'on notera Ω, est formée de n demi-disques (n étant le nombre de sommets de Γ), deux demi-disques étant reliés par un rectangle fin si les sommets correspondants sont reliés par une arête. Sur ce domaine, on considère le spectre de Steklov–Neumann (voir les rappels de la section 2A) : le bord de Steklov est formé des diamètres des demi-disques, et on pose la condition de Neumann sur le reste du bord.

Lemme 4.3 [Jammes 2014, théorème 4.3]. *Il existe une famille de métriques g_ε sur Ω telle que les n premières valeurs propre de Steklov de (Ω, g_ε) tendent vers le spectre de Δ_Γ, avec convergence des espaces propres.*

Remarque 4.4. Dans [Jammes 2014], ce théorème est démontré pour des densités uniformes sur $\partial\Omega_S$ et sur les sommets du graphes. Si on munit les sommets du graphe d'une autre mesure, on peut adapter la démonstration en munissant chaque composante de $\partial\Omega_S$ (les diamètres des demi-disques) de la densité correspondante. On peut donc faire tendre le spectre de Ω non seulement vers le spectre d'un laplacien sur Γ, mais aussi d'un opérateur de Schrödinger.

Le deuxième ingrédient de la démonstration est un résultat de convergence du spectre d'une surface vers celui d'un de ses domaines. Sur un domaine U d'une surface à bord Σ, on considérera le problème de Steklov–Neumann avec $\partial U_S = \partial U \cap \partial\Sigma$ comme bord de Steklov et $\partial U_N = \partial U \backslash \partial U_S$ comme bord de Neumann :

Lemme 4.5. *Soit (Σ, g) une surface riemannienne compacte à bord, ρ une densité sur $\partial\Sigma$ et U un domaine de Σ à bord C^1 par morceaux tel que $\partial U_S = \partial U \cap \partial\Sigma$ soit non vide. Il existe des familles de densités $\gamma_\varepsilon \in C^\infty(\Sigma)$ et de métriques g_ε telles que $\sigma_k(\Sigma, g_\varepsilon, \gamma_\varepsilon, \rho) \to \sigma_k(U, \partial U_S, g, \rho_{|\partial U_S})$ avec convergence des espaces propres quand ε tend vers* 0.

La démonstration est la même que celle du théorème 3.8 de [Jammes 2014], avec l'adaptation à la dimension 2 introduite par Colin de Verdière [1987]. On introduit une densité singulière $\bar{\gamma}_\eta$ qui vaut 1 sur U et η sur $\Sigma\backslash U$, une métrique $\bar{g}_\eta$ égale à

g sur U et à $\eta^2 g$ sur $\Sigma \backslash U$ et on travaille avec les familles de formes quadratiques et de normes de Hilbert induites par $\bar{\gamma}_\eta$ et $\bar{g}_\eta$:

$$Q_\eta(f) = \inf_{\tilde{f}_{|\partial M}=f} \left(\int_U |\mathrm{d}\tilde{f}|^2 \,\mathrm{d}v_g + \eta^3 \int_{M\backslash U} |\mathrm{d}\tilde{f}|^2 \,\mathrm{d}v_g \right), \tag{4.6}$$

et $|f|_\eta = \int_{\partial U_S} f^2 \,\mathrm{d}v_g + \eta^2 \int_{\partial M \backslash \partial U_S} f^2 \,\mathrm{d}v_g$. Le reste de la démonstration (convergence de spectre et lissage de la densité et de la métrique) est identique à [Jammes 2014].

Démonstration du théorème 4.1. Soit Γ admettant un plongement propre dans Σ. On peut trouver un voisinage de Γ dans Σ difféomorphe au domaine Ω décrit précédemment. Étant donnée une métrique et une mesure sur Γ, on peut trouver une famille de métriques g_ε sur Ω dont le début du spectre de Steklov–Neumann tend vers le spectre du laplacien combinatoire sur Γ (lemme 4.3). Pour toute métrique sur Ω on peut trouver une famille de densités sur Σ et $\partial\Sigma$ telle que le spectre de Steklov de Σ tende vers celui de Ω (lemme 4.5). Par conséquent, il existe des familles de métriques g_ε de densités ρ_ε (étendant la métrique g_ε et la densité ρ de Ω) et de densités γ_ε sur Σ telles que le début du spectre de $(\Sigma, g_\varepsilon, \rho_\varepsilon, \gamma_\varepsilon)$ tende vers le spectre du laplacien combinatoire sur Γ.

En munissant Γ de la métrique et de la mesure réalisant la multiplicité $\mu(\Gamma)$, on peut faire tendre le spectre de Σ vers un spectre limite (la métrique et la densité étant alors dégénérées) ayant la multiplicité souhaitée. Pour que $\sigma_1(\Sigma, g_\varepsilon, \rho_\varepsilon, \gamma_\varepsilon)$ soit de multiplicité $\mu(\Gamma)$ pour une métrique g_ε et des densités ρ_ε et γ_ε lisses, on utilise la propriété de stabilité de la multiplicité : quitte à déformer l'opérateur sur Γ, on peut trouver un $\varepsilon > 0$ tel que $\sigma_1(\Sigma, g_\varepsilon, \gamma_\varepsilon)$ soit de multiplicité $\mu(\Gamma)$. □

4A. ***Critère de plongement non entrelacé.*** Cette section est consacrée au corollaire 1.23, dont la démonstration repose sur les propriétés de l'invariant μ des graphes défini par Colin de Verdière et déjà utilisé dans le paragraphe précédent. On utilisera en particulier le fait que cet invariant permet de caractériser les graphes non entrelacés :

Théorème 4.7 [Robertson et al. 1995; Bacher et Colin de Verdière 1995; Lovász et Schrijver 1998]. *Un graphe Γ est non entrelacé si et seulement si $\mu(\Gamma) \leq 4$.*

Le fait $\mu(\Gamma) \leq 4$ implique le non entrelacement du graphe découle de la caractérisation par mineurs exclus des graphes non entrelacés due à N. Robertson, P. Seymour et R. Thomas [1995] et du calcul de μ sur les graphes de la famille de Petersen fait dans [Bacher et Colin de Verdière 1995]. La réciproque a été montrée par L. Lovász et A. Schrijver [1998].

Le corollaire 1.23 découle de la remarque suivante : sachant que $m_1(\mathbb{M}^2) = m_1(\mathbb{P}^2_2) = 4$, si Γ admet un plongement propre dans $\mathbb{M}^2$ ou $\mathbb{P}^2_2$, alors $\mu(\Gamma) \leq 4$ d'après le théorème 4.1. Le théorème 4.7 permet alors de conclure.

Remarque 4.8. Si $\Gamma \to \Sigma$ est un plongement propre d'un graphe dans une surface et $\Sigma \to \mathbb{R}^3$ un plongement de cette surface dans $\mathbb{R}^3$, on peut espérer par composition obtenir un plongement non entrelacé de Γ. Si le plongement usuel $\mathbb{M}^2 \to \mathbb{R}^3$ semble suffire pour tout plongement propre $\Gamma \to \mathbb{M}^2$, il n'en va pas de même pour $\mathbb{P}^2_2$: si on considère le plongement $\mathbb{P}^2_2 \to \mathbb{R}^3$ obtenu en enlevant un disque au plongement $\mathbb{M}^2 \to \mathbb{R}^3$, il existe des plongements propres $\Gamma \to \mathbb{M}^2$ tels que le plongement $\Gamma \to \mathbb{R}^3$ induit soit entrelacé (on utilise le fait que le bord et l'âme du ruban de Möbius forment un entrelacs non trivial). Mais cela n'exclut pas l'existence, pour un plongement $\Gamma \to \mathbb{M}^2$ fixé, d'un autre plongement $\mathbb{P}^2_2 \to \mathbb{R}^3$ tel que le plongement de Γ soit non entrelacé.

5. Bornes sur la multiplicité

5A. ***Propriétés de l'ensemble nodal d'une fonction propre.*** Ce paragraphe sera consacré aux propriétés générales de l'ensemble nodal des fonctions propres du problème de Steklov. Il s'agit généralement de rappels de résultats montrés dans [Fraser et Schoen 2015; Jammes 2014; Karpukhin et al. 2014], ou remontant à l'étude des fonctions propres du laplacien. Certaines sont reformulées ou précisées. Elles sont regroupées en deux théorèmes : le théorème 5.1 rassemble les propriétés locales de l'ensemble nodal et le théorème 5.5 les propriétés topologiques globales. Ces résultats ont généralement été démontrés dans le cas homogène mais leur démonstration reste valide dans le cas général. Si une fonction f s'annule en un point p, on appelle ordre d'annulation de f en p le plus petit entier k tel que $\nabla^k f \neq 0$.

Théorème 5.1 [Cheng 1976; Fraser et Schoen 2015]. *Soit p un point de l'ensemble nodal d'une fonction propre f du problème de Steklov sur Σ. On note k l'ordre d'annulation de f en p.*

Au voisinage de p l'ensemble nodal est la réunion de k arcs de courbes s'intersectant en p, de courbure géodésique nulle en p et formant un système équiangulaire. De plus :

(1) *Si p est un point intérieur à Σ, alors p est l'extrémité de $2k$ arcs nodaux.*

(2) *Si p est sur le bord de Σ, alors p est l'extrémité de k arcs nodaux rencontrant transversalement le bord.*

On peut en déduire, comme dans [Cheng 1976], que l'ensemble nodal est la réunion de sous-variétés de dimension 1 immergées dans Σ, chaque sous-variété étant soit un cercle ne rencontrant pas le bord, soit un intervalle dont les extrémités sont sur le bord. On appellera *lignes nodales* ces sous-variétés. On peut aussi interpréter l'ensemble nodal comme un graphe plongé dans Σ, les sommets étant les points critiques de la fonction propre. On appellera *arêtes nodales* les arêtes de ce graphe. Soulignons que les arêtes nodales sont plongées alors que les lignes nodales sont seulement immergées.

On aura aussi besoin du lemme qui suit et qui assure l'existence d'une fonction propre s'annulant à un ordre élevé quand la multiplicité est grande :

Lemme 5.2. *Soit k un entier positif, p un point de Σ et E un espace propre du problème de Steklov. Si l'une des deux conditions suivantes est vérifiée*

(1) *p est un point intérieur à Σ et E est de dimension au moins $2k$;*

(2) *p est sur le bord de Σ et E est de dimension au moins $k+1$;*

alors E contient une fonction qui s'annule à l'ordre k en p.

Démonstration. Le cas où p est intérieur à Σ est traité dans [Besson 1980]. On va donc supposer sur $p \in \partial\Sigma$ et que E est de dimension au moins $k+1$. Par une déformation conforme, on peut identifier localement la surface au demi-plan supérieur muni des coordonnées (x, y) et le point p au point $(0, 0)$.

Comme E est de dimension au moins $k+1$, on peut trouver une fonction non nulle de E telle que

$$f(p) = \frac{\partial f}{\partial x}(p) = \cdots = \frac{\partial^{k-1} f}{\partial x^{k-1}}(p) = 0. \tag{5.3}$$

On sait aussi que f vérifie l'équation aux valeurs propres $\partial f/\partial y = \sigma\rho f$ en tout point du bord. En dérivant cette relation par rapport à x et en l'évaluant au point p, on obtient que

$$\frac{\partial f}{\partial y}(p) = \frac{\partial^2 f}{\partial x \partial y}(p) = \cdots = \frac{\partial^k f}{\partial x^{k-1}\partial y}(p) = 0. \tag{5.4}$$

Ces relations suffisent pour conclure si $k = 1$ ou 2.

Enfin, la fonction f est harmonique relativement à la densité γ, c'est-à-dire qu'elle vérifie la relation $\operatorname{div}(\gamma \nabla f) = 0$. Par conséquent, $\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$ peut s'écrire comme une expression d'ordre 0 ou 1 en f, et donc, en utilisant les relations déjà obtenues, $\frac{\partial^2 f}{\partial y^2}(p) = 0$ si $k \geq 3$. En dérivant successivement l'équation $\operatorname{div}(\gamma \nabla f) = 0$, on obtient par récurrence que toutes les dérivées partielles de f d'ordre au moins 2 par rapport à y sont nulles en p jusqu'à l'ordre souhaité. □

Théorème 5.5. *La décomposition nodale de Σ vérifie les propriétés suivantes :*

(1) *Tous les domaines nodaux rencontrent le bord de la surface Σ.*

(2) *Les domaines nodaux et les composantes connexes du graphe nodal d'une fonction propre sont incompressibles dans la surface Σ.*

(3) *Si Γ est l'union d'une ou plusieurs composantes connexes du graphe nodal, alors on a la relation $\chi(\Gamma) \geq \chi(\Sigma)$.*

Démonstration. Les premières propriétés énoncées dans ce théorème sont déjà connues (voir par exemple [Jammes 2014]), mais nous allons rappeler leur preuve.

Si un domaine nodal ne rencontre pas le bord de Σ, alors la fonction propre est nulle sur toute la frontière du domaine, donc elle est nulle à l'intérieur du domaine puisqu'elle est harmonique et par conséquent elle est nulle partout selon la propriété d'unique prolongement.

Supposons qu'un domaine nodal D contienne une courbe contractile dans Σ mais pas dans D. Alors le disque bordé par cette courbe contient un domaine nodal distinct de D et qui ne rencontre pas le bord de Σ. Or on vient de voir que c'est impossible, donc toute courbe de D contractile dans Σ est contractile dans D. Le même argument montre que le graphe nodal est incompressible.

Enfin, si on note Γ la réunion d'une ou plusieurs composantes connexes du graphe nodal, et D_i les composantes connexes de $\Sigma\backslash\Gamma$, la formule d'Euler–Poincaré nous dit que

$$\chi(\Sigma) = \chi(\Gamma) + \sum_i \chi(D_i). \tag{5.6}$$

Pour un i donné, la caractéristique d'Euler de l'intérieur de D_i est au plus égale à 1. Mais comme D_i contient nécessairement un domaine nodal de la fonction propre, il rencontre le bord. Donc D_i est la réunion de son intérieur et d'un (ou plusieurs) intervalle ouvert du bord de Σ, donc $\chi(D_i) \leq 0$. Par conséquent, $\chi(\Gamma) \geq \chi(\Sigma)$. □

Remarque 5.7. L'argument d'incompressibilité de l'ensemble nodal permet en fait de montrer un résultat plus fort mais dont nous ne ferons pas usage : l'ensemble nodal ne contient pas de courbe fermée qui borde un domaine.

5B. ***Majoration de multiplicité.*** Le but de cette section est de démontrer les bornes sur la multiplicité données par les théorèmes 1.11 et 1.15.

Démonstration du théorème 1.11. La démonstration de l'inégalité (1.12) reprend la technique de N. Nadirashvili [1987] avec les améliorations que permet le problème de Steklov. On fixe un entier $k \geq 1$ et on note m la multiplicité de la k-ième valeur propre. On se donne aussi un point $p \in \partial\Sigma$. Selon le lemme 5.2, on peut trouver une fonction non nulle dans le k-ième espace propre qui s'annule à l'ordre $m-1$ en p. Le point p est donc l'extrémité de $m-1$ lignes nodales (théorème 5.1).

Parmi les lignes nodales d'extrémité p, il y en a au plus $1-\chi(\Sigma)$ dont les deux extrémités sont p. En effet, si on note q le nombre de lignes allant de p à p, la composante connexe de p dans le graphe nodal aura une caractéristique d'Euler au plus égal à $1-q$, et donc $\chi(\Sigma) \leq 1-q$ d'après le théorème 5.5. On déduit de cette remarque que p est l'extrémité d'au moins $m+\chi(\Sigma)-2$ lignes distinctes.

On peut alors majorer m en appliquant la formule d'Euler–Poincaré à la surface Σ. Contrairement à ce qui est fait dans la démonstration du théorème 5.5, on considérera la surface *ouverte*, c'est-à-dire en ignorant le bord et les sommets du graphe nodal qui s'y trouve. On notera $\dot{\Gamma}$ le graphe ainsi obtenu. On a alors

$$\chi(\Sigma) = \chi(\dot{\Gamma}) + \sum_i \chi(D_i). \tag{5.8}$$

Chaque domaine nodal vérifie $\chi(D_i) \leq 1$ donc $\sum_i \chi(D_i) \leq k+1$ selon le théorème de Courant. Comme $\dot{\Gamma}$ est la réunion d'au moins $m + \chi(\Sigma) - 2$ lignes nodales homéomorphes à des intervalles ouverts, sa caractéristique d'Euler vérifie $\chi(\dot{\Gamma}) \leq -m - \chi(\Sigma) + 2$. On obtient finalement que

$$m \leq k + 3 - 2\chi(\Sigma). \tag{5.9}$$ □

Avant d'entamer la démonstration du théorème 1.15, on va montrer une propriété des fonctions propres qui sera utile pour établir l'inégalité (1.17) :

Lemme 5.10. *Soit f une fonction propre de Steklov et $c > 0$ un réel positif. Chaque composante connexe de l'ensemble $f^{-1}(]-c, +\infty[)$ contient au moins un domaine nodal positif de f. En particulier, si f est associée à la première valeur propre non nulle, cet ensemble est connexe.*

Démonstration. Soit D une composante connexe de $f^{-1}(]-c, +\infty[)$ qui ne contient pas de domaine nodal positif. Comme f est harmonique, le maximum de f sur D est atteint en un point x situé sur le bord de la surface.

On a nécessairement $f(x) < 0$. En effet, si $f(x) > 0$ alors D contient un domaine nodal positif. Si $f(x) = 0$, alors x est l'extrémité d'une ligne nodale de f qui traverse le domaine, donc f change de signe dans D, en particulier elle prend des valeurs strictement positive dans D ce qui est impossible.

Au point x, la fonction f vérifie l'équation aux valeurs propres $\partial f/\partial\nu = \rho\sigma f$. Par conséquent, $(\partial f/\partial\nu)(x)$ est strictement négatif. Comme ν est un vecteur *sortant* normal au bord, on a une contradiction avec le fait que x soit le maximum de f sur D.

Pour finir, d'après ce qui précède le nombre de composantes connexes de $f^{-1}(]-c, +\infty[)$ est majoré par le nombre de domaines nodaux positifs de f. Or, si f est associée à la première valeur propre non nulle elle a exactement deux domaines nodaux, un positif et un négatif. Par conséquent, $f^{-1}(]-c, +\infty[)$ n'a qu'une composante connexe. □

Le théorème 1.15 repose sur un théorème de Sévennec :

Théorème 5.11 [Sévennec 2002, théorème 5]. *Si Σ est une surface close de caractéristique d'Euler strictement négative et E un espace de fonction continues sur Σ tel que pour toute fonction $f \in E\backslash\{0\}$, les ensembles $f^{-1}(]0, +\infty[)$ et $f^{-1}([0, +\infty[)$ sont connexes et non vides, alors $Dim(E) \leq 5 - \chi(\Sigma)$.*

Remarque 5.12. Le principe du maximum assure que pour une fonction propre de Steklov $f \neq 0$, l'ensemble $f^{-1}([0,+\infty[)$ n'a pas de composante connexe disjointe de $f^{-1}(]0,+\infty[)$.

Remarque 5.13. Pour appliquer le théorème 5.11 aux fonctions propres de σ_1, on utilise le théorème de Courant qui assure que ces fonctions propres ont exactement deux domaines nodaux. Or, G. Alessandrini [1998] a montré qu'il reste valable en dimension 2 si g et γ sont seulement L^∞. Comme on n'a pas besoin de la régularité des fonctions propres sur le bord de la surface, on peut aussi supposer que ρ est L^∞.

Démonstration du théorème 1.15. On commence par l'inégalité (1.16). En collant un disque sur chaque composante du bord de Σ, on obtient une surface close $\bar{\Sigma}$ de caractéristique d'Euler $\bar{\chi} = \chi + l$.

On prolonge chaque fonction propre f de la première valeur propre sur chacun de ces disques de la manière suivante : en notant p le centre du disque et x un point générique sur le bord, on fixe $f(p) = 0$ et on interpole linéairement f sur le segment $[p,x]$. La fonction f ainsi prolongée est continue, et elle a deux domaines nodaux, un positif et un négatif (en effet, comme elle est de signe constant sur $]p,x]$, les points intérieurs aux disques appartiennent aux mêmes domaines nodaux que les points de Σ). Compte tenu de la remarque 5.12, cette construction assure que les ensembles $f^{-1}([0,+\infty[)$ et $f^{-1}(]-\infty,0])$ sont eux aussi connexes. D'après le théorème de Sévennec, cet espace est de dimension au plus $5-\bar{\chi}$. On a donc $m_1 \leq 5-\bar{\chi} = 5-\chi-l$.

Pour montrer l'inégalité (1.17), on va encore appliquer le théorème 5 de Sévennec mais à un espace plus grand : on considère l'espace E engendré par les fonctions propres prolongées à $\bar{\Sigma}$ comme précédemment et par la fonction φ définie par $\varphi \equiv 1$ sur Σ, $\varphi(p) = -1$ et en interpolant la fonction de manière affine sur les rayons du disque (comme on suppose que $l = 1$, il n'y a qu'un disque et le point p est unique). L'espace E est constitué de fonction continues, on doit donc montrer que ces fonctions (à l'exception de la fonction nulle) ont exactement deux domaines nodaux, un positif et un négatif.

Une fonction de E est une combinaison linéaire $f_{a,b} = a \cdot f + b \cdot \varphi$, $a,b \in \mathbb{R}$, f étant une fonction propre prolongée à $\bar{\Sigma}$. Si a ou b est nul, il est clair que la fonction a un seul domaine nodal positif.

Sans nuire à la généralité, on peut supposer que $a = 1$ et que b est strictement positif. La fonction $f_{a,b}$ est égale à $f+b$ sur Σ et $f_{a,b}(p) = -b$. Selon le lemme 5.10 appliqué à la fonction f, l'ensemble $f_{a,b}^{-1}(]0,+\infty[) = f^{-1}(]-b,+\infty[)$ restreint à Σ est connexe. Comme $f_{a,b}$ est affine sur les rayons du disque et négative en son centre, les points du disque où $f_{a,b}$ est positive sont connectés au domaine nodal positif sur Σ. La fonction $f_{a,b}$ n'a donc globalement qu'un seul domaine nodal positif.

En restriction à Σ, la fonction $f_{a,b}$ peut avoir plusieurs domaines nodaux négatifs. Cependant, ils sont connectés à p par des rayons sur lesquels $f_{a,b}$ est négative. Il n'y a donc qu'un seul domaine nodal négatif sur $\overline{\Sigma}$. Comme E est de dimension m_1+1, l'application du théorème de Sévennec donne $m_1+1 \leq 5-\bar{\chi} = 4-\chi$, soit $m_1 \leq 3-\chi$. $\square$

5C. ***Première valeur propre en petit genre.*** On commencera dans ce paragraphe par démontrer le théorème 1.19, puis on montrera séparément les calculs des trois valeurs de m_k annoncées dans par le théorème 1.20, les deux premières dans ce paragraphe et la 3-ième dans le suivant.

Démonstration du théorème 1.19. Pour majorer la multiplicité il s'agit, comme dans [Fraser et Schoen 2015; Jammes 2014; Karpukhin et al. 2014], de se ramener à la démonstration utilisée dans le cas des opérateurs de Schrödinger sur les surfaces closes. La minoration découle du théorème 1.7.

Dans le cas de $\mathbb{T}^2_p$, on fixe un point x intérieur à la surface et on suppose que la multiplicité de σ_1 est au moins 7. Il existe alors un espace de dimension 2 de fonctions propres qui s'annulent à l'ordre 3 en x, et on note f_θ le cercle unité de cet espace. Si on ferme la surface en contractant chaque bord de $\mathbb{T}^2_p$ en un point, les lignes nodales de f_θ qui atteignent le bord se prolongent sur le quotient en des courbes C^1 par morceaux. On peut alors reprendre les arguments de [Besson 1980] pour montrer que sur le quotient, l'ensemble nodal est formé d'exactement trois lacets disjoints non homotopes et aboutir à une contradiction en faisant varier θ.

Sur $\mathbb{K}^2_p$, on se ramène de la même manière à la démonstration du théorème 4.1 de [Colin de Verdière 1987] en quotientant chaque bord sur un point. $\square$

Avant de traiter le cas de $\mathbb{P}^2_2$, on va montrer que $m_1(\mathbb{M}^2) = 4$ par une méthode légèrement différente de celle du théorème 1.11. Cette démonstration servira de base pour les deux propositions qui vont suivre.

On commence par montrer qu'il existe une fonction propre pour laquelle il existe au moins 6 extrémités de lignes nodales qui rejoignent le bord. Soit x un point du bord de $\mathbb{M}^2$. Si la multiplicité de σ_1 est supérieure ou égale à 5, il existe une fonction propre pour laquelle x est l'extrémité d'au moins quatre lignes nodales. S'il est l'extrémité d'au moins 5 lignes, alors il y a au moins six extrémités de lignes qui rejoignent le bord puisque ce nombre est toujours pair. S'il est l'extrémité d'exactement quatre lignes alors le signe de la fonction propre est constant le long du bord au voisinage de x, par conséquent la fonction propre change de signe ailleurs sur le bord (sinon la fonction est positive ou nulle partout sur le bord, donc partout sur $\mathbb{M}^2$) et il y a donc au moins six extrémités de lignes le long du bord. Si on définit une application $\mathbb{M}^2 \mapsto \mathbb{P}^2$ par contraction du bord de $\mathbb{M}^2$ sur un point (qu'on notera encore x), on peut considérer l'image de la décomposition nodale de $\mathbb{M}^2$ par cette application : elle décompose $\mathbb{P}^2$ en deux domaines, et x est l'extrémité

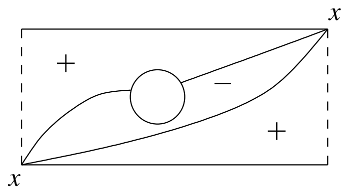


Figure 4. Décomposition nodale de $\mathbb{P}_2^2$.

de six arcs nodaux. Or Besson [1980] a montré qu'une telle décomposition de $\mathbb{P}^2$ est impossible.

Proposition 5.14. $$m_1(\mathbb{P}_2^2) = 4.$$

Démonstration. On représente $\mathbb{P}_2^2$ comme étant $\mathbb{M}^2$ privé d'un disque. On parlera de bord de $\mathbb{M}^2$ et de bord du disque pour désigner les deux bords de $\mathbb{P}_2^2$.

Le point de départ de la démonstration est le même que celle qui précède. On suppose que la multiplicité de $\sigma_1(\mathbb{P}_2^2)$ est 5. Étant donné un point x du bord de $\mathbb{M}^2$, il existe une fonction propre f telle que x soit localement l'extrémité de quatre lignes nodales. On a vu précédemment qu'il ne peut pas y avoir d'autres extrémités de lignes nodales sur le bord de $\mathbb{M}^2$. Le signe de f le long de ce bord est donc constant et on le supposera positif. Comme les deux domaines nodaux de f rencontrent le bord de $\mathbb{P}_2^2$, le domaine négatif rencontre nécessairement le bord du disque.

Si on considère un arc qui part orthogonalement de x (avec par conséquent deux lignes nodales de part et d'autre, selon le théorème 5.1) et qui rejoint un autre point du bord en restant dans le domaine nodal positif, il découpe nécessairement le ruban $\mathbb{M}^2$ en un rectangle comme sur la figure 4. La décomposition nodale de $\mathbb{P}_2^2$ est donc nécessairement topologiquement équivalente à celle de cette figure, à ceci près que le bord du disque peut être entièrement contenu dans le domaine nodal négatif.

Après avoir contracté le bord du disque sur un point qu'on notera p, on pourra donc toujours trouver dans $\mathbb{M}^2$ un lacet γ_x d'extrémité x, unique à homotopie près, contenu dans le domaine nodal négatif, passant par p et homotope à un générateur du groupe fondamental de $\mathbb{M}^2$.

Supposons maintenant qu'on déplace continûment le point x. Le lacet γ_x peut alors se déformer continûment avec la contrainte de toujours passer par p. Si on fait faire à x un tour complet du bord de $\mathbb{M}^2$ en partant d'un point x_0, on revient à la même décomposition de $\mathbb{P}_2^2$, donc le lacet γ_x devient homotope au lacet initial γ_{x_0}. Or, chacun des arcs de γ_{x_0} allant de x_0 à p serait homotope à sa concaténation avec un générateur du groupe fondamental du bord, ce qui est impossible. □

Proposition 5.15. $$m_1(\mathbb{T}_1^2) = 5.$$

Démonstration. On procède par l'absurde en supposant que $\sigma_1(\mathbb{T}_1^2)$ est de multiplicité 6 pour une métrique g et des densités ρ, γ données.

Soit x un point du bord de $\mathbb{T}_1^2$. Selon le théorème 5.1 et le lemme 5.2, il existe une fonction propre f qui s'annule à l'ordre 5 en x et telle que cinq arcs nodaux partent de x. Notons que cette fonction est unique : dans le cas contraire, on pourrait en choisir une s'annulant à l'ordre 6 et on aurait une contradiction comme dans le théorème 1.11. Comme les extrémités de lignes nodales rejoignant le bord sont nécessairement en nombre pair, il existe une 6-ième extrémité en un point x' du bord, x' étant distinct de x (sinon f s'annulerait à l'ordre 6 en x).

Comme dans la démonstration précédente, on va déplacer le point x le long du bord. Les arguments développés par Besson [1980] permettront d'aboutir à une contradiction. Si on contracte le bord sur un point — notons-le $\bar{x}$ — il y a six arcs nodaux partant de $\bar{x}$. On sait alors que l'ensemble nodal est la réunion de trois lacets non homotopes entre eux et qui ne s'intersectent qu'en $\bar{x}$ (cf. la démonstration du théorème 3.C.1 dans [Besson 1980]). Cette remarque permet alors de transposer le reste de la démonstration de [Besson 1980] : si x se déplace continûment en partant d'un point x_0, on peut construire une homotopie entre les lignes nodales de la fonction propre f_x correspondante et les lignes nodales de f_{x_0} (il est crucial ici que $x \neq x'$ quel que soit x). Après un tour complet de x le long du bord, les classes d'homotopies des lignes nodales sont donc les conjuguées des lignes de f_{x_0} par la classe d'homotopie du bord. Or elles doivent être identiques aux lignes nodales de f_{x_0}. Il y a contradiction car les classes d'homotopies des lignes nodales sont des classes d'homotopies non triviales dans le tore $\mathbb{T}^2$ obtenu par contraction du bord, elles ne commutent donc pas avec la classe du bord. □

5D. ***Deuxième valeur propre du disque.*** Pour finir, on montre la majoration de $m_2(\mathbb{D}^2)$ annoncée par le théorème 1.21. On sait déjà, d'après le théorème 1.11, que $m_2(\mathbb{D}^2) \leq 3$. On va supposer qu'il y a égalité pour aboutir à une contradiction.

On se donne une métrique g et une densité γ telle que la multiplicité de $\sigma_2(\mathbb{D}^2)$ soit 3. Selon le théorème de Courant, une fonction propre de $\sigma_2(\mathbb{D}^2)$ a deux ou trois domaines nodaux. Comme ces domaines sont simplement connexes, la décomposition nodale du disque est nécessairement topologiquement équivalente à l'une des trois indiquées sur la figure 5.

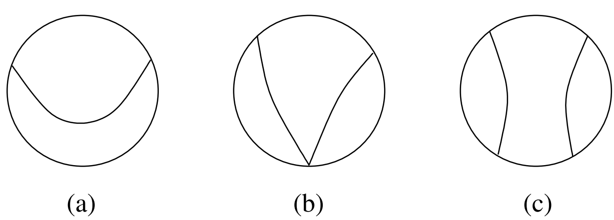


Figure 5. Domaines nodaux sur le disque.

On considère la sphère unité de l'espace propre de $\sigma_2(\mathbb{D}^2)$, qu'on notera S^2, et on note respectivement D_a (resp. D_b et D_c) l'ensemble des fonctions propres de S^2 dont la décomposition nodale est celle de la figure 5(a) (resp. 5(b) et 5(c)). La contradiction découlera de l'étude de la partition de S^2 ainsi formée.

Commençons par noter que, comme la multiplicité est égale à 3, le théorème 5.1 et le lemme 5.2 montrent qu'il existe nécessairement des fonctions propres réalisant la situation (b) de la figure 5 et donc D_b est non vide. Plus précisément, pour chaque point x de $\partial\mathbb{D}^2$, il existe une droite de fonction propre s'annulant à l'ordre 2 en x, et cette droite varie continûment avec x (elle est unique car dans le cas contraire, on pourrait trouver une fonction s'annulant à l'ordre 3 en x). Comme cette droite coupe S^2 en deux points, l'ensemble D_b est la réunion de deux cercles, l'un correspondant aux fonctions ayant un seul domaine nodal positif, l'autre à celles en ayant deux (en particulier, ces cercles sont disjoints).

Le complémentaire de D_b est donc formé de trois composantes connexes, deux qui sont antipodales et homéomorphes à des disques et une homéomorphe à un cylindre. Comme D_c a nécessairement deux composantes connexes antipodales, l'une formée des fonctions ayant un domaine nodal positif et l'autre des fonctions en ayant deux, D_c est la réunion des deux disques et D_a est le cylindre.

On choisit l'une des deux composantes de D_c, par exemple celle dont les fonctions ont deux domaines nodaux positifs, qu'on notera D_c^+, et on construit une application de D_c^+ dans le cercle de la manière suivante : on se donne une orientation sur $\partial\mathbb{D}^2$, et pour chaque fonction de D_c^+ on considère la paire de points de $\partial\mathbb{D}^2$ où la fonction s'annule en décroissant (par rapport à l'orientation du bord). Ces deux points sont nécessairement disjoints (sinon il y aurait au moins trois extrémités de lignes nodales se rejoignant en un même points du bord, ce qu'on a déjà exclu). On obtient ainsi une application continue partant de D_c^+ et dont l'image est $S^1 \times S^1$ privé de la diagonale et quotientée par $(x, y) \sim (y, x)$. Cette image est homotope à un cercle, ce qui permet de définir une application $D_c^+ \to S^1$. Or, le long du bord de D_c^+, c'est-à-dire d'une des composantes de D_b, cette application est homotopiquement non triviale, ce qui fournit la contradiction cherchée.

Remerciements

Je remercie Y. Colin de Verdière et I. Polterovich pour leurs commentaires sur la première version de cet article, ainsi qu'un rapporteur anonyme dont les remarques ont permis d'améliorer le texte.

Bibliographie

[Alessandrini 1998] G. Alessandrini, "On Courant's nodal domain theorem", *Forum Math.* **10**:5 (1998), 521–532. MR 99g:35091 Zbl 0909.35098

[Appel et Haken 1976] K. Appel et W. Haken, "Every planar map is four colorable", *Bull. Amer. Math. Soc.* **82**:5 (1976), 711–712. MR 54 #12561 Zbl 0331.05106

[Bacher et Colin de Verdière 1995] R. Bacher et Y. Colin de Verdière, "Multiplicités des valeurs propres et transformations étoile-triangle des graphes", *Bull. Soc. Math. France* **123**:4 (1995), 517–533. MR 96k:05129 Zbl 0845.05068

[Bandle 1980] C. Bandle, *Isoperimetric inequalities and applications*, Monographs and Studies in Mathematics **7**, Pitman, Boston, 1980. MR 81e:35095 Zbl 0436.35063

[Besson 1980] G. Besson, "Sur la multiplicité de la première valeur propre des surfaces Riemanniennes", *Ann. Inst. Fourier* (*Grenoble*) **30**:1 (1980), 109–128. MR 81h:58059 Zbl 0417.30033

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

[Bose 1939] R. C. Bose, "On the construction of balanced incomplete block designs", *Ann. Eugenics* **9** (1939), 353–399. MR 1,199b Zbl 0023.00102

[Cheng 1976] S. Y. Cheng, "Eigenfunctions and nodal sets", *Comment. Math. Helv.* **51**:1 (1976), 43–55. MR 53 #1661 Zbl 0334.35022

[Coxeter 1943] H. S. M. Coxeter, "The map-coloring of unorientable surfaces", *Duke Math. J.* **10** (1943), 293–304. MR 5,48f Zbl 0060.41602

[Fraser et Schoen 2011] A. Fraser et R. Schoen, "The first Steklov eigenvalue, conformal geometry, and minimal surfaces", *Adv. Math.* **226**:5 (2011), 4011–4030. MR 2012f:58054 Zbl 1215.53052

[Fraser et Schoen 2015] A. Fraser et R. Schoen, "Sharp eigenvalue bounds and minimal surfaces in the ball", *Invent. Math.* (online publication May 2015), 1–68.

[Heawood 1890] P. J. Heawood, "Map-colour theorem", *Quart. J. Pure Appl. Math.* **24**:96 (1890), 332–338. JFM 22.0562.02

[Heffter 1891] L. Heffter, "Ueber das Problem der Nachbargebiete", *Math. Ann.* **38**:4 (1891), 477–508. MR 1510685 JFM 23.0543.01

[Jammes 2009] P. Jammes, "Sur la multiplicité des valeurs propres d'une variété compacte", pp. 1–11 dans *Actes du Séminaire de Théorie Spectrale et Géométrie* (Grenoble, 2007–2008), Séminaire de Théorie Spectrale et Géométrie **26**, Université Grenoble I, Saint-Martin-d'Hères, 2009. MR 2011d:58075 Zbl 1235.58022

[Jammes 2014] P. Jammes, "Prescription du spectre de Steklov dans une classe conforme", *Anal. PDE* **7**:3 (2014), 529–549. MR 3227426 Zbl 1304.35452

[Kagno 1935] I. N. Kagno, "A note on the Heawood color formula", *J. Math. Phys.* **14** (1935), 228–231. JFM 61.1347.01

[Karpukhin et al. 2014] M. Karpukhin, G. Kokarev et I. Polterovich, "Multiplicity bounds for Steklov eigenvalues on Riemannian surfaces", *Ann. Inst. Fourier* (*Grenoble*) **64**:6 (2014), 2481–2502. MR 3331172 Zbl 1321.58027

[Lovász et Schrijver 1998] L. Lovász et A. Schrijver, "A Borsuk theorem for antipodal links and a spectral characterization of linklessly embeddable graphs", *Proc. Amer. Math. Soc.* **126**:5 (1998), 1275–1285. MR 98j:05059 Zbl 0886.05055

[Nadirashvili 1987] N. S. Nadirashvili, "Кратные собственные значения оператора Лапласа", *Mat. Sb.* (*N.S.*) **133**(175):2 (1987), 223–237. Translated as "Multiple eigenvalues of the Laplace operator" in *Math. USSR Sb.* **61** :1 (1988), 225–238. MR 89a:58113 Zbl 0672.35049

[Ringel 1974] G. Ringel, *Map color theorem*, Grundlehren der Mathematischen Wissenschaften **209**, Springer, New York, 1974. MR 50 #1955 Zbl 0287.05102

[Ringel et Youngs 1968] G. Ringel et J. W. T. Youngs, "Solution of the Heawood map-coloring problem", *Proc. Nat. Acad. Sci. USA* **60** (1968), 438–445. MR 37 #3959 Zbl 0155.51201

[Robertson et al. 1995] N. Robertson, P. Seymour et R. Thomas, "Sachs' linkless embedding conjecture", *J. Combin. Theory Ser. B* **64**:2 (1995), 185–227. MR 96m:05072 Zbl 0832.05032

[Sévennec 1994] B. Sévennec, "Multiplicité du spectre des surfaces: une approche topologique", pp. 29–36 dans *Actes du Séminaire de Théorie Spectrale et Géométrie* (Grenoble, 1993–1994), Séminaire de Théorie Spectrale et Géométrie **12**, Université Grenoble I, Saint-Martin-d'Hères, 1994. MR 1714546 Zbl 0909.58059

[Sévennec 2002] B. Sévennec, "Multiplicity of the second Schrödinger eigenvalue on closed surfaces", *Math. Ann.* **324**:1 (2002), 195–211. MR 2003h:58045 Zbl 1053.58014

[Sylvester et Uhlmann 1990] J. Sylvester et G. Uhlmann, "The Dirichlet to Neumann map and applications", pp. 101–139 dans *Inverse problems in partial differential equations* (Arcata, CA, 1989), édité par D. Colton et al., SIAM, Philadelphia, 1990. MR 91d:35063 Zbl 0713.35100

[Uhlmann 2009] G. Uhlmann, "Electrical impedance tomography and Calderón's problem", *Inv. Prob.* **25**:12 (2009), Article ID #123011. Zbl 1181.35339

[Colin de Verdière 1987] Y. Colin de Verdière, "Construction de Laplaciens dont une partie finie du spectre est donnée", *Ann. Sci. École Norm. Sup.* (4) **20**:4 (1987), 599–615. MR 90d:58156 Zbl 0636.58036

[Colin de Verdière 1988] Y. Colin de Verdière, "Sur une hypothèse de transversalité d'Arnol'd", *Comment. Math. Helv.* **63**:2 (1988), 184–193. MR 90c:58183 Zbl 0672.58046

[Colin de Verdière 1990] Y. Colin de Verdière, "Sur un nouvel invariant des graphes et un critère de planarité", *J. Combin. Theory Ser. B* **50**:1 (1990), 11–21. MR 91m:05068 Zbl 0742.05061

[Colin de Verdière 1998] Y. Colin de Verdière, *Spectres de graphes*, Cours Spécialisés **4**, Société Mathématique de France, Paris, 1998. MR 99k:05108 Zbl 0913.05071

Received November 18, 2014. Revised September 8, 2015.

PIERRE JAMMES
CNRS, LJAD, UMR 7351
UNIV. NICE SOPHIA ANTIPOLIS
PARC VALROSE
06100 NICE
FRANCE

pjammes@unice.fr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 282, No. 1, 2016
dx.doi.org/10.2140/pjm.2016.282.173

E-POLYNOMIAL OF THE SL(3, $\mathbb{C}$)-CHARACTER VARIETY OF FREE GROUPS

SEAN LAWTON AND VICENTE MUÑOZ

We compute the E-polynomial of the character variety of representations of a rank r free group in SL(3, $\mathbb{C}$). Expanding upon techniques of Logares, Muñoz and Newstead (*Rev. Mat. Complut.* 26:2 (2013), 635–703), we stratify the space of representations and compute the E-polynomial of each geometrically described stratum using fibrations. Consequently, we also determine the E-polynomial of its smooth, singular, and abelian loci and the corresponding Euler characteristic in each case. Along the way, we give a new proof of results of Cavazos and Lawton (*Int. J. Math.* 25:6 (2014), 1450058).

1. Introduction

Let Γ be a finitely generated group, and let G be a complex reductive algebraic group. The space of G-representations is

$$\mathcal{R}(\Gamma, G) = \{\rho : \Gamma \to G \mid \rho \text{ is a group morphism}\}.$$

Writing a presentation $\Gamma = \langle x_1, \ldots, x_n | R_1, \ldots, R_s \rangle$, we have that $\rho \in \mathcal{R}(\Gamma, G)$ is determined by the images $A_i = \rho(x_i)$, $1 \le i \le n$. Hence we can write $\rho = (A_1, \ldots, A_n)$. These matrices are subject to the relations $R_j(A_1, \ldots, A_n) = \mathrm{Id}$, $1 \le j \le s$. Hence

$$\mathcal{R}(\Gamma, G) \cong \{(A_1, \ldots, A_n) \in G^n \mid R_1(A_1, \ldots, A_n) = \cdots = R_s(A_1, \ldots, A_n) = \mathrm{Id}\}$$

is an affine algebraic set, since G is algebraic.

There is an action of G by conjugation on $\mathcal{R}(\Gamma, G)$, which is equivalent to the action of $PG = G/Z(G)$, where $Z(G)$ is the center of G, since the center acts trivially. The *G-character variety of* Γ is the GIT quotient

$$\mathcal{M}(\Gamma, G) = \mathcal{R}(\Gamma, G) /\!\!/ G,$$

which is an affine algebraic set by construction. Note that if we write $X := \mathcal{R}(\Gamma, G) = \mathrm{Spec}(S)$, then $X /\!\!/ G = \mathrm{Spec}(S^G)$.

MSC2010: 14D20, 20C15, 14L30, 20E05.
Keywords: E-polynomial, free group, SL(3, $\mathbb{C}$), character variety.

Every element $g \in \Gamma$ determines a character $\chi_g : X \to \mathbb{C}$, $\chi_g(\rho) = \mathrm{tr}(\rho(g))$, with respect to an embedding $G \hookrightarrow \mathrm{GL}(n, \mathbb{C})$. These regular functions $\chi_g \in S$ are invariant by conjugation, and hence $\chi_g \in S^G$. Consider the algebra of characters

$$T = \mathbb{C}[\chi_g \mid g \in \Gamma] \subset S^G,$$

and let $\chi(\Gamma, G) = \mathrm{Spec}(T)$. There is a well-defined surjective map $\mathcal{M}(\Gamma, G) \to \chi(\Gamma, G)$, which is an isomorphism when $G = \mathrm{SL}(n, \mathbb{C})$ among other examples; see [Sikora 2013].

In this paper we are interested in the character variety for the free group on r elements $\Gamma = F_r$ and for the group $G = \mathrm{SL}(3, \mathbb{C})$. We compute the E-polynomial (also known as Hodge–Deligne polynomial) of $\mathcal{M}(F_r, \mathrm{SL}(3, \mathbb{C}))$. The E-polynomial of $\mathcal{M}(F_r, \mathrm{SL}(2, \mathbb{C}))$ has been computed in [Cavazos and Lawton 2014] by arithmetic methods (using the Weil conjectures). Recently, in [Mozgovoy and Reineke 2015], the E-polynomials of $\mathcal{M}(F_r, \mathrm{PGL}(n, \mathbb{C}))$ have also been computed by arithmetic methods, where the result is given in the form of a generating function.

Here we use a geometric technique, introduced in [Logares et al. 2013], to compute E-polynomials of character varieties. This consists of stratifying the space of representations geometrically, and computing the E-polynomials of each stratum using the behavior of E-polynomials with fibrations. This technique is used in [Logares et al. 2013] for the case of $\Gamma = \pi_1(X)$ for a surface X of genus $g = 1, 2$ and $G = \mathrm{SL}(2, \mathbb{C})$ (and also with one puncture, fixing the holonomy around the puncture). The case of $g = 3$ is worked out in [Martínez and Muñoz 2015a], the case of $g \geq 4$ in [Martínez and Muñoz 2015b], and the case of $g = 1$ with two punctures appears in [Logares and Muñoz 2014]. To implement this geometric technique for character varieties for $\mathrm{SL}(n, \mathbb{C})$, for $n \geq 3$, we need to introduce the *equivariant Hodge–Deligne polynomial* with respect to a finite group action on an affine variety. This will be useful for studying character varieties of surface groups in $\mathrm{SL}(n, \mathbb{C})$, $n \geq 3$.

We start by recovering the E-polynomials $e(\mathcal{M}(F_r, \mathrm{SL}(2, \mathbb{C})))$ of [Cavazos and Lawton 2014] and $e(\mathcal{M}(F_r, \mathrm{PGL}(2, \mathbb{C})))$ of [Mozgovoy and Reineke 2015], verifying that they are equal. Then we move to rank 3 to compute $e(\mathcal{M}(F_r, \mathrm{SL}(3, \mathbb{C})))$ and $e(\mathcal{M}(F_r, \mathrm{PGL}(3, \mathbb{C})))$. They turn out to be equal again. The latter one coincides, as expected, with the polynomial obtained in [Mozgovoy and Reineke 2015].

Unlike the methods used to obtain $e(\mathcal{M}(F_r, \mathrm{PGL}(3, \mathbb{C})))$ in [Mozgovoy and Reineke 2015], our method provides an explicit geometric description of, and the E-polynomial for, each stratum. By results in [Florentino and Lawton 2012] this additional information determines the E-polynomial of the smooth and singular loci of $\mathcal{M}(F_r, \mathrm{SL}(3, \mathbb{C}))$, and by [Florentino and Lawton 2014] also determines the E-polynomial of the abelian character variety $\mathcal{M}(\mathbb{Z}^r, \mathrm{SL}(3, \mathbb{C}))$.

Our main theorem is thus:

Theorem 1. *The E-polynomials $e(\mathcal{M}(F_r, \mathrm{SL}(3,\mathbb{C})))$ and $e(\mathcal{M}(F_r, \mathrm{PGL}(3,\mathbb{C})))$ are both equal to*

$$(q^8-q^6-q^5+q^3)^{r-1}+(q-1)^{2r-2}(q^{3r-3}-q^r)$$
$$+\tfrac{1}{6}(q-1)^{2r-2}q(q+1)+\tfrac{1}{2}(q^2-1)^{r-1}q(q-1)+\tfrac{1}{3}(q^2+q+1)^{r-1}q(q+1)$$
$$-(q-1)^{r-1}q^{r-1}(q^2-1)^{r-1}(2q^{2r-2}-q).$$

From the definition of the E-polynomial of a variety X, the classical Euler characteristic is given by $\chi(X)=e(X;1,1)$. Consequently, we deduce:

Corollary 2. *Let $r\geq 2$. Then $\mathcal{M}(F_r, \mathrm{SL}(3,\mathbb{C}))$, $\mathcal{M}(F_r, \mathrm{PGL}(3,\mathbb{C}))$, and* (by [Florentino and Lawton 2009]) *$\mathcal{M}(F_r, \mathrm{SU}(3))$, have Euler characteristic given by $2\cdot 3^{r-2}$. The Euler characteristic of $\mathcal{M}(\mathbb{Z}^r, \mathrm{SL}(3,\mathbb{C}))$, and* (by [Florentino and Lawton 2014]) *also $\mathcal{M}(\mathbb{Z}^r, \mathrm{SU}(3))$, is given 3^{r-2}.*

2. Hodge structures and E-polynomials

Our main goal is to compute the E-polynomial (Hodge–Deligne polynomial) of the $\mathrm{SL}(3,\mathbb{C})$-character variety of a free group. We will follow the methods in [Logares et al. 2013], so we collect some basic results from [loc. cit.] in this section.

We start by reviewing the definition of the Hodge–Deligne polynomial. A pure Hodge structure of weight k consists of a finite dimensional complex vector space H with a real structure, and a decomposition $H=\bigoplus_{k=p+q} H^{p,q}$ such that $H^{q,p}=\overline{H^{p,q}}$, the bar meaning complex conjugation on H. A Hodge structure of weight k gives rise to the so-called Hodge filtration, which is a descending filtration $F^p=\bigoplus_{s\geq p} H^{s,k-s}$. We define $\mathrm{Gr}^p_F(H):=F^p/F^{p+1}=H^{p,k-p}$.

A mixed Hodge structure consists of a finite dimensional complex vector space H with a real structure, an ascending (weight) filtration $\cdots\subset W_{k-1}\subset W_k\subset\cdots\subset H$ (defined over $\mathbb{R}$) and a descending (Hodge) filtration F such that F induces a pure Hodge structure of weight k on each $\mathrm{Gr}^W_k(H)=W_k/W_{k-1}$. We define $H^{p,q}:=\mathrm{Gr}^p_F\,\mathrm{Gr}^W_{p+q}(H)$ and write $h^{p,q}$ for the *Hodge number* $h^{p,q}:=\dim H^{p,q}$.

Let Z be any quasiprojective algebraic variety (possibly nonsmooth or noncompact). The cohomology groups $H^k(Z)$ and the cohomology groups with compact support $H^k_c(Z)$ are endowed with mixed Hodge structures [Deligne 1971; 1974]. We define the *Hodge numbers* of Z by

$$h^{k,p,q}_c(Z)=h^{p,q}(H^k_c(Z))=\dim \mathrm{Gr}^p_F\,\mathrm{Gr}^W_{p+q}\,H^k_c(Z).$$

The Hodge–Deligne polynomial, or E-polynomial, is defined as

$$e(Z)=e(Z)(u,v):=\sum_{p,q,k}(-1)^k h^{k,p,q}_c(Z)u^p v^q.$$

The key property of Hodge–Deligne polynomials that permits their calculation is that they are additive for stratifications of Z. If Z is a complex algebraic variety and $Z = \bigsqcup_{i=1}^{n} Z_i$, where all Z_i are locally closed in Z, then

$$e(Z) = \sum_{i=1}^{n} e(Z_i).$$

Also, by [Logares et al. 2013, Remark 2.5], if $G \to X \to B$ is a principal fiber bundle with G a connected algebraic group, then $e(X) = e(G)e(B)$. In general we shall use this as $e(X/G) = e(X)/e(G)$ when $B = X/G$. In particular, if Z is a G-space, and there is a subspace $B \subset Z$ such that $B \times G \to Z$ is surjective and it is an H-homogeneous space for a connected subgroup $H \subset G$, then

$$e(Z) = e(B)e(G)/e(H). \tag{1}$$

Definition 3. Let X be a complex quasiprojective variety on which a finite group F acts. Then F also acts on the cohomology $H_c^*(X)$ respecting the mixed Hodge structure. So $[H_c^*(X)] \in R(F)$, the representation ring of F. The *equivariant Hodge–Deligne polynomial* is defined as

$$e_F(X) = \sum_{p,q,k} (-1)^k [H_c^{k,p,q}(X)]\, u^p v^q \in R(F)[u, v].$$

Note that the map $\dim : R(F) \to \mathbb{Z}$ gives $\dim(e_F(X)) = e(X)$.

For instance, for an action of $\mathbb{Z}_2$, there are two irreducible representations T, N, where T is the trivial representation, and N is the nontrivial representation. Then $e_{\mathbb{Z}_2}(X) = aT + bN$. Clearly

$$e(X) = a + b, \quad e(X/\mathbb{Z}_2) = a.$$

In the notation of [Logares et al. 2013, Section 2], $a = e(X)^+$, $b = e(X)^-$. Note that if X, X' are spaces with $\mathbb{Z}_2$-actions, then writing

$$e_{\mathbb{Z}_2}(X) = aT + bN \quad \text{and} \quad e_{\mathbb{Z}_2}(X') = a'T + b'N,$$

we have $e_{\mathbb{Z}_2}(X \times X') = (aa' + bb')T + (ab' + ba')N$ and so

$$e((X \times X')/\mathbb{Z}_2) = aa' + bb' = e(X)^+ e(X')^+ + e(X)^- e(X')^-. \tag{2}$$

When $h_c^{k,p,q} = 0$ for $p \neq q$, the polynomial $e(Z)$ depends only on the product uv. This will happen in all the cases that we shall investigate here. In this situation, it is conventional to use the variable $q = uv$. If this happens, we say that the variety is *of balanced type*. For instance, $e(\mathbb{C}^n) = q^n$.

3. E-polynomial of the SL(2, $\mathbb{C}$)-character variety of free groups

Let F_r denote the free group on r generators. Then the space of representations of F_r in the group SL(2, $\mathbb{C}$) is

$$\mathcal{R}_{r,2} = \mathrm{Hom}(F_r, \mathrm{SL}(2, \mathbb{C})) = \{(A_1, \dots, A_r) \mid A_i \in \mathrm{SL}(2, \mathbb{C})\} = \mathrm{SL}(2, \mathbb{C})^r.$$

The group PGL(2, $\mathbb{C}$) acts on $\mathcal{R}_{r,2}$ by simultaneous conjugation of all matrices, and the character variety is defined as the GIT quotient

$$\mathcal{M}_{r,2} = \mathcal{R}_{r,2} /\!\!/ \mathrm{PGL}(2, \mathbb{C}).$$

We aim to compute the E-polynomial of $\mathcal{M}_{r,2}$ using the methods developed in [Logares et al. 2013] and to recover the results of [Cavazos and Lawton 2014]. We have the following sets:

- Reducible representations $\mathcal{R}^{\mathrm{red}}_{r,2} \subset \mathcal{R}_{r,2}$ and the corresponding set $\mathcal{M}^{\mathrm{red}}_{r,2} \subset \mathcal{M}_{r,2}$ of characters of reducible representations. A representation $\rho = (A_1, \dots, A_r)$ is reducible if and only if all A_i share at least one eigenvector.
- Irreducible representations $\mathcal{R}^{\mathrm{irr}}_{r,2} \subset \mathcal{R}_{r,2}$ and the corresponding set $\mathcal{M}^{\mathrm{irr}}_{r,2} \subset \mathcal{M}_{r,2}$ of characters of irreducible representations. This is the complement of $\mathcal{R}^{\mathrm{red}}_{r,2}$. It consists of the representations ρ such that PGL(2, $\mathbb{C}$) acts freely on ρ, and the orbit PGL(2, $\mathbb{C}$) $\cdot\, \rho$ is closed. Therefore $\mathcal{M}^{\mathrm{irr}}_{r,2} = \mathcal{R}^{\mathrm{irr}}_{r,2} / \mathrm{PGL}(2, \mathbb{C})$.

3.1. ***The reducible locus.*** Let us start by computing $e(\mathcal{M}^{\mathrm{red}}_{r,2})$. For a reducible representation, we have a basis of $\mathbb{C}^2$ in which

$$\rho = \left(\begin{pmatrix} \lambda_1 & * \\ 0 & \lambda_1^{-1} \end{pmatrix}, \begin{pmatrix} \lambda_2 & * \\ 0 & \lambda_2^{-1} \end{pmatrix}, \dots, \begin{pmatrix} \lambda_r & * \\ 0 & \lambda_r^{-1} \end{pmatrix} \right).$$

The associated point is determined by $(\lambda_1, \dots, \lambda_r) \in (\mathbb{C}^*)^r$, modulo $(\lambda_1, \dots, \lambda_r) \sim (\lambda_1^{-1}, \dots, \lambda_r^{-1})$. Note that the action of $\lambda \mapsto \lambda^{-1}$ on $X = \mathbb{C}^*$ has $e(X)^+ = q$ and $e(X)^- = -1$. Writing $X_i = \mathbb{C}^*$, $i = 1, \dots, r$, we have that

$$\begin{aligned} e(X_1 \times \cdots \times X_r)^+ &= \sum_{\epsilon \in A} \prod_{i=1}^{r} e(X_i)^{\epsilon_i} \\ &= q^r + \binom{r}{2} q^{r-2} + \binom{r}{4} q^{r-4} + \cdots + \binom{r}{2[r/2]} q^{r-2[r/2]} \\ &= \tfrac{1}{2}((q+1)^r + (q-1)^r), \end{aligned}$$

where $A = \{(\epsilon_1, \dots, \epsilon_r) \in (\pm 1)^r \mid \prod \epsilon_i = +1\}$. Also

$$\begin{aligned} e(X_1 \times \cdots \times X_r)^- &= e(X_1 \times \cdots \times X_r) - e(X_1 \times \cdots \times X_r)^+ \\ &= (q-1)^r - \tfrac{1}{2}((q+1)^r + (q-1)^r) \\ &= \tfrac{1}{2}((q-1)^r - (q+1)^r). \end{aligned}$$

Also note that $e(\mathcal{M}_{r,2}^{\mathrm{red}}) = e((X_1 \times \cdots \times X_r)/\mathbb{Z}_2) = e(X_1 \times \cdots \times X_r)^+$.

3.2. ***The reducible representations.*** Now we move to the computation of $e(\mathcal{R}_{r,2}^{\mathrm{red}})$. We stratify the space as $\mathcal{R}_{r,2}^{\mathrm{red}} = R_0 \cup R_1 \cup R_2 \cup R_3$, where:

- R_0 consists of $(A_1, \ldots, A_r) = (\pm\,\mathrm{Id}, \ldots, \pm\,\mathrm{Id})$. So $e(R_0) = 2^r$.
- R_1 consists of
$$\rho \sim \left(\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_1^{-1} \end{pmatrix}, \begin{pmatrix} \lambda_2 & 0 \\ 0 & \lambda_2^{-1} \end{pmatrix}, \ldots, \begin{pmatrix} \lambda_r & 0 \\ 0 & \lambda_r^{-1} \end{pmatrix} \right),$$
that is, abelian representations (all matrices are diagonalizable with respect to the same basis). Here $(\lambda_1, \ldots, \lambda_r) \neq (\pm 1, \ldots, \pm 1)$. Therefore this space is parametrized by
$$\big(\mathrm{PGL}(2,\mathbb{C})/D \times \big((\mathbb{C}^*)^r - \{(\pm 1, \ldots, \pm 1)\}\big)\big)/\mathbb{Z}_2,$$
where D is the space of diagonal matrices. We know that $e(\mathrm{PGL}(2,\mathbb{C})/D)^+ = q^2$, $e(\mathrm{PGL}(2,\mathbb{C})/D)^- = q$ by [Logares et al. 2013, Proposition 3.2]. For $B = (\mathbb{C}^*)^r - \{(\pm 1, \ldots, \pm 1)\}$, we have $e(B)^+ = \frac{1}{2}((q+1)^r + (q-1)^r) - 2^r$ and $e(B)^- = \frac{1}{2}((q-1)^r - (q+1)^r)$, by our computation above. Therefore
$$\begin{aligned} e(R_1) &= e(\mathrm{PGL}(2,\mathbb{C})/D)^+ e(B)^+ + e(\mathrm{PGL}(2,\mathbb{C})/D)^- e(B)^- \\ &= q^2 \tfrac{1}{2}((q+1)^r + (q-1)^r - 2^r) + q\tfrac{1}{2}((q-1)^r - (q+1)^r) \\ &= \tfrac{1}{2}(q^2-q)(q+1)^r + \tfrac{1}{2}(q^2+q)(q-1)^r - q^2 2^r\,. \end{aligned}$$
- R_2 consists of
$$\rho \sim \left(\begin{pmatrix} \pm 1 & a_1 \\ 0 & \pm 1 \end{pmatrix}, \begin{pmatrix} \pm 1 & a_2 \\ 0 & \pm 1 \end{pmatrix}, \ldots, \begin{pmatrix} \pm 1 & a_r \\ 0 & \pm 1 \end{pmatrix} \right),$$
where $(a_1, \ldots, a_r) \in \mathbb{C}^r - \{0\}$. Let B_2 be the space of representations as above with respect to the canonical basis. Therefore, there is a canonical surjective map $B_2 \times \mathrm{PGL}(2,\mathbb{C}) \twoheadrightarrow R_2$. The fibers of this map are given by $H_2 = \left\{ \left(\begin{smallmatrix} a & b \\ 0 & a^{-1} \end{smallmatrix}\right) \right\} \cong \mathbb{C}^* \times \mathbb{C}$. That is, $H_2 \to B_2 \times \mathrm{PGL}(2,\mathbb{C}) \to R_2$ is a fibration to which we apply Formula (1) to obtain
$$e(R_2) = \frac{e(B_2)e(\mathrm{PGL}(2,\mathbb{C}))}{e(H_2)} = \frac{2^r(q^r-1)(q^3-q)}{q(q-1)} = 2^r(q^r-1)(q+1).$$
- R_3 consists of
$$\rho \sim \left(\begin{pmatrix} \lambda_1 & b_1 \\ 0 & \lambda_1^{-1} \end{pmatrix}, \begin{pmatrix} \lambda_2 & b_2 \\ 0 & \lambda_2^{-1} \end{pmatrix}, \ldots, \begin{pmatrix} \lambda_r & b_r \\ 0 & \lambda_r^{-1} \end{pmatrix} \right),$$
where $\lambda_i \in \mathbb{C}^*$, $(\lambda_1, \ldots, \lambda_r) \neq (\pm 1, \ldots, \pm 1)$. Here, $(b_1, \ldots, b_r) \in \mathbb{C}^r$ and the upper diagonal matrices $\left(\begin{smallmatrix} 1 & y \\ 0 & 1 \end{smallmatrix}\right)$ transform
$$(b_1, \ldots, b_r) \mapsto (b_1 + y(\lambda_1 - \lambda_1^{-1}), \ldots, b_r + y(\lambda_r - \lambda_r^{-1})).$$

As $(\lambda_1, \dots, \lambda_r) \neq (\pm 1, \dots, \pm 1)$, this action is nontrivial. Note that $(b_1, \dots, b_r)$ does not live in the line spanned by $(\lambda_1 - \lambda_1^{-1}, \dots, \lambda_r - \lambda_r^{-1})$. There is a fibration $H_3 \to B_3 \times \mathrm{PGL}(2, \mathbb{C}) \to R_3$ where $H_3 = \left\{ \begin{pmatrix} a & b \\ 0 & a^{-1} \end{pmatrix} \right\} \cong \mathbb{C}^* \times \mathbb{C}$. Thus

$$\begin{aligned} e(R_3) &= (q^r - q)((q-1)^r - 2^r) e(\mathrm{PGL}(2, \mathbb{C}))/q(q-1) \\ &= \frac{q^{r-1} - 1}{q - 1}((q-1)^r - 2^r)(q^3 - q) \\ &= (q^{r-1} - 1)(q-1)^{r-1}(q^3 - q) - 2^r \frac{q^{r-1} - 1}{q - 1}(q^3 - q). \end{aligned}$$

Now we add all the subsets together:

$$\begin{aligned} e(\mathcal{R}_{r,2}^{\mathrm{red}}) &= e(R_0) + e(R_1) + e(R_2) + e(R_3) \\ &= \tfrac{1}{2}(q^2 - q)(q+1)^r + \tfrac{1}{2}(q^2 + q)(q-1)^r \\ &\quad + (q^{r-1} - 1)(q-1)^{r-1}(q^3 - q). \end{aligned}$$

3.3. ***The irreducible locus.*** Recall that $\mathcal{R}_{r,2}^{\mathrm{irr}} = \mathrm{SL}(2, \mathbb{C})^r - \mathcal{R}_{r,2}^{\mathrm{red}}$, so

$$\begin{aligned} e(\mathcal{R}_{r,2}^{\mathrm{irr}}) = (q^3 - q)^r - \tfrac{1}{2}(q^2 - q)(q+1)^r - \tfrac{1}{2}(q^2 + q)(q-1)^r \\ - (q^{r-1} - 1)(q-1)^{r-1}(q^3 - q), \end{aligned}$$

and

$$\begin{aligned} e(\mathcal{M}_{r,2}^{\mathrm{irr}}) &= \frac{e(\mathcal{R}_{r,2}^{\mathrm{irr}})}{q^3 - q} \\ &= (q^3 - q)^{r-1} - \tfrac{1}{2}(q+1)^{r-1} - \tfrac{1}{2}(q-1)^{r-1} - (q^{r-1} - 1)(q-1)^{r-1}. \end{aligned}$$

Finally,

$$\begin{aligned} e(\mathcal{M}_{r,2}) &= e(\mathcal{M}_{r,2}^{\mathrm{irr}}) + e(\mathcal{M}_{r,2}^{\mathrm{red}}) = e(\mathcal{M}_{r,2}^{\mathrm{irr}}) + \tfrac{1}{2}((q+1)^r + (q-1)^r) \\ &= (q^3 - q)^{r-1} + \tfrac{1}{2}q(q+1)^{r-1} + \tfrac{1}{2}q(q-1)^{r-1} - q^{r-1}(q-1)^{r-1}. \end{aligned}$$

This agrees with [Cavazos and Lawton 2014].

4. E-polynomial of the PGL(2, $\mathbb{C}$)-character variety of free groups

Let us compute the E-polynomial of $\mathcal{M}(F_r, \mathrm{PGL}(2, \mathbb{C}))$. The space of representations will be denoted

$$\overline{\mathcal{R}}_{r,2} = \mathrm{Hom}(F_r, \mathrm{PGL}(2, \mathbb{C})) = \{(A_1, \dots, A_r) \mid A_i \in \mathrm{PGL}(2, \mathbb{C})\} = \mathrm{PGL}(2, \mathbb{C})^r.$$

Note that $\mathrm{PGL}(2, \mathbb{C}) = \mathrm{SL}(2, \mathbb{C})/\{\pm \mathrm{Id}\}$, so $\overline{\mathcal{R}}_{r,2} = \mathcal{R}_{r,2}/\{(\pm \mathrm{Id}, \dots, \pm \mathrm{Id})\}$. The character variety is

$$\overline{\mathcal{M}}_{r,2} = \overline{\mathcal{R}}_{r,2} /\!\!/ \mathrm{PGL}(2, \mathbb{C}).$$

We denote by $\overline{\mathcal{R}}^{\mathrm{red}}_{r,2}$ and $\overline{\mathcal{R}}^{\mathrm{irr}}_{r,2}$ the subsets of reducible and irreducible representations, respectively, of $\overline{\mathcal{R}}_{r,2}$. We denote by $\overline{\mathcal{M}}^{\mathrm{red}}_{r,2}$ and $\overline{\mathcal{M}}^{\mathrm{irr}}_{r,2}$ the corresponding spaces in $\overline{\mathcal{M}}_{r,2}$.

The reducible locus. We first compute $e(\overline{\mathcal{M}}^{\mathrm{red}}_{r,2})$. A reducible representation in $\overline{\mathcal{M}}^{\mathrm{red}}_{r,2}$ is determined by the eigenvalues $(\lambda_1, \dots, \lambda_r) \in (\mathbb{C}^*)^r$, modulo $\lambda_i \sim -\lambda_i$, $1 \le i \le r$, and $(\lambda_1, \dots, \lambda_r) \sim (\lambda_1^{-1}, \dots, \lambda_r^{-1})$. So it is determined by $(\lambda_1^2, \dots, \lambda_r^2) \in (\mathbb{C}^*)^r$, modulo $(\lambda_1^2, \dots, \lambda_r^2) \sim (\lambda_1^{-2}, \dots, \lambda_r^{-2})$. This space is isomorphic to the one in Section 3.1, so $e(\overline{\mathcal{M}}^{\mathrm{red}}_{r,2}) = \frac{1}{2}((q+1)^r + (q-1)^r)$.

The reducible representations. Now we compute $e(\overline{\mathcal{R}}^{\mathrm{red}}_{r,2})$. We stratify it as

$$\overline{\mathcal{R}}^{\mathrm{red}}_{r,2} = \bar{R}_0 \cup \bar{R}_1 \cup \bar{R}_2 \cup \bar{R}_3,$$

where:

- $\bar{R}_0$ consists of one point $(A_1, \dots, A_r) = (\mathrm{Id}, \dots, \mathrm{Id})$. So $e(R_0) = 1$.
- $\bar{R}_1$ consists of

$$\rho \sim \left(\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_1^{-1} \end{pmatrix}, \begin{pmatrix} \lambda_2 & 0 \\ 0 & \lambda_2^{-1} \end{pmatrix}, \dots, \begin{pmatrix} \lambda_r & 0 \\ 0 & \lambda_r^{-1} \end{pmatrix} \right),$$

where the eigenvalues are determined by $(\lambda_1^2, \dots, \lambda_r^2) \neq (1, \dots, 1)$. This space is parametrized by $\big(\mathrm{PGL}(2,\mathbb{C})/D \times ((\mathbb{C}^*)^r - \{(1, \dots, 1)\})\big)/\mathbb{Z}_2$, where D is the space of diagonal matrices. Using that $e(\mathrm{PGL}(2,\mathbb{C})/D)^+ = q^2$, $e(\mathrm{PGL}(2,\mathbb{C})/D)^- = q$, and $e(B)^+ = \frac{1}{2}((q+1)^r + (q-1)^r) - 1$, $e(B)^- = \frac{1}{2}((q-1)^r - (q+1)^r)$, for $B = ((\mathbb{C}^*)^r - \{(1, \dots, 1)\})$, we have

$$\begin{aligned} e(\bar{R}_1) &= e(\mathrm{PGL}(2,\mathbb{C})/D)^+ e(B)^+ + e(\mathrm{PGL}(2,\mathbb{C})/D)^- e(B)^- \\ &= q^2 \tfrac{1}{2}((q+1)^r + (q-1)^r - 1) + q \tfrac{1}{2}((q-1)^r - (q+1)^r) \\ &= \tfrac{1}{2}(q^2 - q)(q+1)^r + \tfrac{1}{2}(q^2+q)(q-1)^r - q^2. \end{aligned}$$

- $\bar{R}_2$ consists of

$$\rho \sim \left(\begin{pmatrix} 1 & a_1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & a_2 \\ 0 & 1 \end{pmatrix}, \dots, \begin{pmatrix} 1 & a_r \\ 0 & 1 \end{pmatrix} \right),$$

where $(a_1, \dots, a_r) \in \mathbb{C}^r - \{0\}$. Then

$$e(\bar{R}_2) = e(R_2)/2^r = (q^r - 1)(q+1).$$

- $\bar{R}_3$ consists of

$$\rho \sim \left(\begin{pmatrix} \lambda_1 & b_1 \\ 0 & \lambda_1^{-1} \end{pmatrix}, \begin{pmatrix} \lambda_2 & b_2 \\ 0 & \lambda_2^{-1} \end{pmatrix}, \dots, \begin{pmatrix} \lambda_r & b_r \\ 0 & \lambda_r^{-1} \end{pmatrix} \right),$$

where $\lambda_i \in \mathbb{C}^*$, $(\lambda_1^2, \dots, \lambda_r^2) \neq (1, \dots, 1)$. Here

$$(b_1, \dots, b_r) \in \mathbb{C}^r - \langle(\lambda_1 - \lambda_1^{-1}, \dots, \lambda_r - \lambda_r^{-1})\rangle.$$

There is a fibration $H_3 \to B_3 \times \mathrm{PGL}(2, \mathbb{C}) \to \bar{R}_3$ where B_3 parametrizes $(\lambda_1^2, \dots, \lambda_r^2)$ and $(b_1, \dots, b_r)$, and $H_3 = \left\{\left(\begin{smallmatrix} a & b \\ 0 & a^{-1}\end{smallmatrix}\right)\right\} \cong \mathbb{C}^* \times \mathbb{C}$. Then

$$\begin{aligned} e(\bar{R}_3) &= (q^r - q)((q-1)^r - 1)e(\mathrm{PGL}(2, \mathbb{C}))/q(q-1) \\ &= \frac{q^{r-1} - 1}{q - 1}((q-1)^r - 1)(q^3 - q). \end{aligned}$$

Now we add all subsets together to obtain:

$$\begin{aligned} e(\bar{\mathcal{R}}_{r,2}^{\mathrm{red}}) &= e(\bar{R}_0) + e(\bar{R}_1) + e(\bar{R}_2) + e(\bar{R}_3) \\ &= \tfrac{1}{2}(q^2 - q)(q+1)^r + \tfrac{1}{2}(q^2 + q)(q-1)^r + (q^{r-1} - 1)(q-1)^{r-1}(q^3 - q) \\ &= e(\mathcal{R}_{r,2}^{\mathrm{red}}). \end{aligned}$$

The irreducible locus. Clearly, as $e(\mathrm{SL}(2, \mathbb{C})) = q^3 - q = e(\mathrm{PGL}(2, \mathbb{C}))$ and $e(\bar{\mathcal{R}}_{r,2}^{\mathrm{red}}) = e(\mathcal{R}_{r,2}^{\mathrm{red}})$, we have that $e(\bar{\mathcal{R}}_{r,2}^{\mathrm{irr}}) = e(\mathcal{R}_{r,2}^{\mathrm{irr}})$. Therefore $e(\overline{\mathcal{M}}_{r,2}^{\mathrm{irr}}) = e(\mathcal{M}_{r,2}^{\mathrm{irr}})$. Finally, since $e(\overline{\mathcal{M}}_{r,2}^{\mathrm{red}}) = e(\mathcal{M}_{r,2}^{\mathrm{red}})$, we have that

$$\begin{aligned} e(\overline{\mathcal{M}}_{r,2}) &= e(\mathcal{M}_{r,2}) \\ &= (q^3 - q)^{r-1} + \tfrac{1}{2}q(q+1)^{r-1} + \tfrac{1}{2}q(q-1)^{r-1} - q^{r-1}(q-1)^{r-1}. \end{aligned}$$

5. E-polynomial of the $\mathrm{SL}(3, \mathbb{C})$-character variety for F_1

Having given a new geometric derivation of the E-polynomial for $\mathcal{M}_{r,2}$ and $\overline{\mathcal{M}}_{r,2}$, in the next sections we work out the E-polynomial of $\mathcal{M}_{r,3}$ and $\overline{\mathcal{M}}_{r,3}$ in a similar fashion.

However, in this section we first address the $r = 1$ case. Although it is easy to see that $\mathcal{M}_{r,n} \cong \mathbb{C}^{n-1}$ via the coefficients of the characteristic polynomial, and hence $e(\mathcal{M}_{r,n}) = q^{n-1}$, this case will motivate the more complicated stratification, and the use of the *equivariant* E-polynomial, needed to compute the general E-polynomials for $\mathcal{M}_{r,3}$ and $\overline{\mathcal{M}}_{r,3}$ when $r \geq 2$.

We begin with the E-polynomials for $\mathrm{GL}(3, \mathbb{C})$, $\mathrm{SL}(3, \mathbb{C})$, and $\mathrm{PGL}(3, \mathbb{C})$. Like in the previous sections, we then stratify $\mathrm{Hom}(F_1, \mathrm{SL}(3, \mathbb{C}))$ by orbit type and compute the E-polynomial for each strata.

Lemma 4. $e(\mathrm{SL}(3, \mathbb{C})) = e(\mathrm{PGL}(3, \mathbb{C})) = (q^3 - 1)(q^3 - q)q^2 = q^8 - q^6 - q^5 + q^3$.

Proof. Consider $\mathbb{C}^n$, and let V_k be the Stiefel manifold of k linearly independent vectors in $\mathbb{C}^n$. Then, there is a (Zariski locally trivial) fibration $\mathbb{C}^n - \mathbb{C}^{k-1} \to V_k \to V_{k-1}$. Therefore $e(V_k) = \prod_{i=0}^{k-1}(q^n - q^i)$. So $e(\mathrm{GL}(n, \mathbb{C})) = e(V_n) = \prod_{i=0}^{n-1}(q^n - q^i)$.

Now there is a (Zariski locally trivial) fibration $\mathbb{C}^* \to \mathrm{GL}(n,\mathbb{C}) \to \mathrm{PGL}(n,\mathbb{C})$, hence $e(\mathrm{PGL}(n,\mathbb{C})) = e(\mathrm{GL}(n,\mathbb{C}))/(q-1) = q^{n-1}\prod_{i=0}^{n-2}(q^n - q^i)$.

For $\mathrm{SL}(n,\mathbb{C})$, the choice of $(v_1,\dots,v_{n-1}) \in V_{n-1}$ determines an affine hyperplane

$$\{v \in \mathbb{C}^n \mid \det(v_1,\dots,v_{n-1},v) = 1\}.$$

This gives a (Zariski locally trivial) affine bundle $\mathbb{C}^{n-1} \to \mathrm{SL}(n,\mathbb{C}) \to V_{n-1}$, and hence $e(\mathrm{SL}(n,\mathbb{C})) = q^{n-1}\prod_{i=0}^{n-2}(q^n - q^i)$. □

Now let us consider the representations of F_1 to $\mathrm{SL}(3,\mathbb{C})$. This is equivalent to studying the conjugation action of $\mathrm{PGL}(3,\mathbb{C})$ on $X := \mathrm{SL}(3,\mathbb{C})$. For this action, there are 6 strata types. In the following list, we write down all 6 strata, but include the computation of their E-polynomials for only the first 5. This is because the computation is apparent from the geometric description of each stratum alone in those cases.

- X_0 is formed by matrices of type

$$\begin{pmatrix} \xi & 0 & 0 \\ 0 & \xi & 0 \\ 0 & 0 & \xi \end{pmatrix}.$$

Here $\xi^3 = 1$, so X_0 consists of 3 points and $e(X_0) = 3$.

- X_1 is formed by matrices of type

$$\begin{pmatrix} \xi & 0 & 0 \\ 0 & \xi & 1 \\ 0 & 0 & \xi \end{pmatrix}.$$

Here $\xi^3 = 1$, so ξ admits 3 values. The stabilizer of this matrix is

$$U_1 = \left\{ \begin{pmatrix} \mu^{-2} & 0 & b \\ a & \mu & c \\ 0 & 0 & \mu \end{pmatrix} \right\} \cong \mathbb{C}^* \times \mathbb{C}^3.$$

So

$$\begin{aligned} e(X_1) &= 3e(\mathrm{PGL}(3,\mathbb{C})/U_1) \\ &= 3(q^3-1)(q^3-q)q^2/q^3(q-1) = 3q^4 + 3q^3 - 3q - 3. \end{aligned}$$

- X_2 is formed by matrices of type

$$\begin{pmatrix} \xi & 1 & 0 \\ 0 & \xi & 1 \\ 0 & 0 & \xi \end{pmatrix}.$$

Here $\xi^3 = 1$, so ξ admits 3 values. The stabilizer of this matrix is

$$U_2 = \left\{ \begin{pmatrix} 1 & b & c \\ 0 & 1 & b \\ 0 & 0 & 1 \end{pmatrix} \right\} \cong \mathbb{C}^2.$$

So

$$\begin{aligned} e(X_1) &= 3e(\mathrm{PGL}(3, \mathbb{C})/U_2) \\ &= 3(q^3 - 1)(q^3 - q)q^2/q^2 = 3q^6 - 3q^4 - 3q^3 + 3q. \end{aligned}$$

- X_3 is formed by matrices of type

$$\begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda^{-2} \end{pmatrix},$$

where $\lambda \in \mathbb{C}^* - \{\xi | \xi^3 = 1\}$. The stabilizer of this matrix is

$$U_3 = \left\{ \begin{pmatrix} A & 0 \\ 0 & (\det A)^{-1} \end{pmatrix} \middle| A \in \mathrm{GL}(2, \mathbb{C}) \right\} \cong \mathrm{GL}(2, \mathbb{C}).$$

So

$$\begin{aligned} e(X_3) &= (q - 4)e(\mathrm{PGL}(3, \mathbb{C})/U_3) \\ &= (q - 4)(q^3 - 1)(q^3 - q)q^2/(q^2 - 1)(q^2 - q) = q^5 - 3q^4 - 3q^3 - 4q^2. \end{aligned}$$

- X_4 is formed by matrices of type

$$\begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda^{-2} \end{pmatrix},$$

where $\lambda \in \mathbb{C}^* - \{\xi \mid \xi^3 = 1\}$. The stabilizer of this matrix is

$$U_4 = \left\{ \begin{pmatrix} \mu & b & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \mu^{-2} \end{pmatrix} \right\} \cong \mathbb{C}^* \times \mathbb{C}.$$

So

$$\begin{aligned} e(X_4) &= (q - 4)e(\mathrm{PGL}(3, \mathbb{C})/U_4) \\ &= (q - 4)(q^3 - 1)(q^3 - q)q^2/q(q - 1) \\ &= q^7 - 3q^6 - 4q^5 - q^4 + 3q^3 + 4q^2. \end{aligned}$$

- X_5 is formed by matrices of type

$$\begin{pmatrix} \lambda & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \gamma \end{pmatrix},$$

where $\lambda, \mu, \gamma \in \mathbb{C}^*$ are different and $\lambda\mu\gamma = 1$. The stabilizer is isomorphic to the diagonal matrices $D \cong \mathbb{C}^* \times \mathbb{C}^*$. The parameter space is

$$B = \{(\lambda, \mu) \in (\mathbb{C}^*)^2 \mid \lambda \neq \mu^{-2}, \mu \neq \lambda^{-2}, \mu \neq \lambda\}.$$

The map $\mathrm{PGL}(3, \mathbb{C})/D \times B \to X_5$ is a 6:1 cover. Moreover,

$$X_5 \cong (\mathrm{PGL}(3, \mathbb{C})/D \times B)/\Sigma_3,$$

where the symmetric group Σ_3 acts on $\mathrm{PGL}(3, \mathbb{C})$ by permuting the columns and acts on the triple $(\lambda, \mu, \gamma = \lambda^{-1}\mu^{-1})$ by permuting the entries.

We now compute $e(X_5)$ using the *equivariant* E-polynomial. Consider the finite group $F = \Sigma_3$. The representation ring $R(F)$ is generated by three irreducible representations:

- T is the (one-dimensional) trivial representation.
- S is the sign representation. This is one-dimensional and given by the sign map $\Sigma_3 \to \{\pm 1\} \subset \mathrm{GL}(1, \mathbb{C})$.
- V is the two-dimensional representation given as follows. Take $St = \mathbb{C}^3$ the standard 3-dimensional representation. This is generated by e_1, e_2, e_3 and Σ_3 acts by permuting the elements of the basis. Then $T = \langle e_1 + e_2 + e_3 \rangle$ and we can decompose $St = T \oplus V$.

The representation ring $R(\Sigma_3)$ has a multiplicative structure given by: $T \otimes T = T$, $T \otimes S = S$, $T \otimes V = V$, $S \otimes S = T$, $S \otimes V = V$, $V \otimes V = T \oplus S \oplus V$.

Lemma 5.

$$e_{\Sigma_3}(B) = (q^2 - q + 1)T + S - 2(q-2)V.$$

$$e_{\Sigma_3}(\mathrm{PGL}(3, \mathbb{C})/D) = q^6 T + q^3 S + (q^5 + q^4)V.$$

Proof. Write $e_{\Sigma_3}(X) = aT + bS + cV$, for a quasiprojective variety X with a Σ_3-action. Then $a = e(X/\Sigma_3)$. If we consider the cycle $(1, 2)$ and the subgroup $H = \langle (1, 2) \rangle$, there is a map $R(F) \to R(H)$ which sends $T \mapsto T$, $S \mapsto N$ and $V \mapsto T + N$. Then $e_H(X) = aT + bN + c(T + N) = (a + c)T + (b + c)N$. Therefore, $a + c = e(X/H)$. As $e(X) = a + b + 2c$, we can compute a, b, c by knowing these E-polynomials.

For $B = \{(\lambda, \mu) \in (\mathbb{C}^*)^2 \mid \lambda \neq \mu^{-2}, \mu \neq \lambda^{-2}, \mu \neq \lambda\}$, the three curves $\lambda = \mu^{-2}$, $\mu = \lambda^{-2}$, $\mu = \lambda$ intersect at the three points $\{(\xi, \xi) \mid \xi^3 = 1\}$. Hence $e(B) = (q-1)^2 - 3(q-4) - 3 = q^2 - 5q + 10$.

Now Σ_3 acts on (λ, μ, γ) and the quotient space is parametrized by $s = \lambda + \mu + \gamma$, $t = \lambda\mu + \lambda\gamma + \mu\gamma$ and $p = \lambda\mu\gamma = 1$, that is, by $(s, t) \in \mathbb{C}^2$. We have to remove the cases $s = \lambda + \lambda^{-2} + \lambda,\ t = \lambda^{-1} + \lambda^2 + \lambda^{-1}$. This defines a rational curve in $\mathbb{C}^2$. It has two points at infinity. The map $\lambda \mapsto (2\lambda + \lambda^{-2}, 2\lambda^{-1} + \lambda^2)$ is an embedding. Therefore $e(B/\Sigma_3) = q^2 - (q-1) = q^2 - q + 1$.

The action by H permutes (λ, μ), hence the quotient is parametrized by $s' = \lambda + \mu$, $p' = \lambda\mu \neq 0$. We have to remove the cases $s' = \lambda + \lambda^{-2}$, $p' = \lambda^{-1}$, that is, $s' = (p')^{-1} + (p')^2$; and $s' = 2\lambda$, $p' = \lambda^2$, i.e., $4p' = (s')^2$. They intersect at three points. Then $e(B/H) = q(q-1) - 2(q-1) + 3 = q^2 - 3q + 5$.

Thus

$$e_{\Sigma_3}(B) = (q^2 - q + 1)T + S - 2(q-2)V.$$

For $C = \mathrm{PGL}(3, \mathbb{C})/D$, the space C consists of points in $(\mathbb{P}^2)^3 - \Delta$, where Δ is the diagonal (triples of coplanar points). Certainly, a matrix in $\mathrm{GL}(3, \mathbb{C})$ can be written as (v_1, v_2, v_3), where v_1, v_2, v_3 are linearly independent vectors. Taking a quotient by the diagonal matrices corresponds to the vectors up to a scalar: $[v_1], [v_2], [v_3]$. Therefore, $e(C) = (q^3-1)(q^3-q)q^2/(q-1)^2 = q^6 + 2q^5 + 2q^4 + q^3$.

The group Σ_3 acts by permuting the vectors, so $C/\Sigma_3 = \mathrm{Sym}^3\,\mathbb{P}^2 - \overline{\Delta}$, where $\overline{\Delta}$ consists of linearly dependent triples $([v_1], [v_2], [v_3])$. If they are equal, the set has $e(\mathbb{P}^2) = q^2 + q + 1$. If they are collinear, there is a fibration with fiber $\mathrm{Sym}^3(\mathbb{P}^1) - \Delta$ and base $(\mathbb{P}^2)^\vee$. This has E-polynomial $(1 + q + q^2 + q^3 - 1 - q)(1 + q + q^2) = q^5 + 2q^4 + 2q^3 + q^2$. Also $e(\mathrm{Sym}^3\,\mathbb{P}^2) = q^6 + q^5 + 2q^4 + 2q^3 + 2q^2 + q + 1$. Therefore

$$e(C/\Sigma_3) = q^6 + q^5 + 2q^4 + 2q^3 + 2q^2 + q + 1 - (q^5 + 2q^4 + 2q^3 + q^2 + q^2 + q + 1) = q^6.$$

The group H acts by permuting the first two vectors, so $C/H = \mathrm{Sym}^2\,\mathbb{P}^2 \times \mathbb{P}^2 - \overline{\Delta}'$, where $\overline{\Delta}'$ consists of linearly dependent triples $([v_1], [v_2], [v_3])$. If $[v_1] = [v_2]$, we have the E-polynomial $(q^2 + q + 1)(q^2 + q + 1) = q^4 + 2q^3 + 3q^2 + 2q + 1$. If $[v_1] \neq [v_2]$, they lie in $\mathrm{Sym}^2\,\mathbb{P}^2 - \Delta$ and we have the E-polynomial

$$\big(q^4 + q^3 + 2q^2 + q + 1 - (q^2 + q + 1)\big)(q + 1) = q^5 + 2q^4 + 2q^3 + q^2.$$

Also $e(\mathrm{Sym}^2\,\mathbb{P}^2 \times \mathbb{P}^2) = (q^4 + q^3 + 2q^2 + q + 1)(q^2 + q + 1)$, so

$$\begin{aligned} e(C/H) = q^6 &+ 2q^5 + 4q^4 + 4q^3 + 4q^2 + 2q + 1 \\ &- (q^5 + 2q^4 + 2q^3 + q^2 + q^4 + 2q^3 + 3q^2 + 2q + 1) \\ = q^6 &+ q^5 + q^4. \end{aligned}$$

This produces the polynomial

$$e_{\Sigma_3}(C) = q^6 T + q^3 S + (q^5 + q^4)V. \qquad \square$$

Remark 6. If we consider $B' = \{(\lambda, \mu) \in (\mathbb{C}^*)^2\}$, then the proof of Lemma 5 says that $e_{\Sigma_3}(B') = q^2 T + S - qV$.

Now suppose $e_{\Sigma_3}(X) = aT + bS + cV$ and $e_{\Sigma_3}(X') = a'T + b'S + c'V$. Then

$$e_{\Sigma_3}(X \times X') = (aa' + bb' + cc')T + (ab' + ba' + cc')S + (ac' + ca' + bc' + cb' + cc')V,$$

and hence

$$e((X \times X')/\Sigma_3) = aa' + bb' + cc' . \tag{3}$$

We finally obtain the E-polynomial for the sixth strata X_5:

$$\begin{aligned} e(X_5) &= e((B \times C)/\Sigma_3) \\ &= (q^2 - q + 1)q^6 + q^3 - 2(q-2)(q^5 + q^4) \\ &= q^8 - q^7 - q^6 + 2q^5 + 4q^4 + q^3 . \end{aligned}$$

Now we add the strata together:

$$e(X_0)+e(X_1)+e(X_2)+e(X_3)+e(X_4)+e(X_5) = q^8 - q^6 - q^5 + q^3 = e(\mathrm{SL}(3, \mathbb{C})),$$

as expected.

Remark 7. All elements of $X = \mathrm{SL}(3, \mathbb{C})$ are reducible. The semisimple ones are given by diagonal matrices with entries λ, μ, γ with $\lambda\mu\gamma = 1$. So they are parametrized by $s = \lambda + \mu + \gamma$, $t = \lambda\mu + \lambda\gamma + \mu\gamma = \lambda^{-1} + \gamma^{-1} + \mu^{-1}$, for $(s, t) \in \mathbb{C}^2$. Hence $e(\mathcal{M}_{1,3}) = q^2$, as noted at the beginning of this section.

6. E-polynomials of character varieties for F_r, $r > 1$, and $\mathrm{SL}(3, \mathbb{C})$

In this section we prove (most of) our main theorem (Theorem 1) by computing the E-polynomial for $\mathcal{M}_{r,3}$; the rest of Theorem 1 is proved in Section 7. The computation is similar to the computation in Section 3 except the stratification is more complicated and the *equivariant* E-polynomial is needed, as was demonstrated in Section 5.

Indeed, we want to study the space of representations

$$\begin{aligned} \mathcal{R}_{r,3} &= \mathrm{Hom}(F_r, \mathrm{SL}(3, \mathbb{C})) = \{\rho : F_r \to \mathrm{SL}(3, \mathbb{C})\} \\ &= \{(A_1, \dots, A_r) \mid A_i \in \mathrm{SL}(3, \mathbb{C})\} = \mathrm{SL}(3, \mathbb{C})^r \end{aligned}$$

and the corresponding character variety

$$\mathcal{M}_{r,3} = \mathrm{Hom}(F_r, \mathrm{SL}(3, \mathbb{C})) /\!\!/ \mathrm{PGL}(3, \mathbb{C}).$$

Much of the algebraic structure of $\mathcal{M}_{r,3}$ has been worked out in [Lawton 2007; 2008; 2010].

Let us start by computing the E-polynomial of the space of reducible representations $\mathcal{R}^{\mathrm{red}}_{r,3} \subset \mathrm{Hom}(F_r, \mathrm{SL}(3, \mathbb{C}))$.

We now list the stratification and the computation of the E-polynomial for each stratum for $\mathcal{R}^{\mathrm{red}}_{r,3}$.

(i) $R_0 = R_{01} \cup R_{02}$. R_{01} is formed by representations $\rho = (A_1, \dots, A_r)$ which have a common eigenvector and such that the quotient representation is irreducible, that is,

$$A_i = \begin{pmatrix} \lambda_i^{-2} & b_i & c_i \\ 0 & \multicolumn{2}{c}{\multirow{2}{*}{$\lambda_i B_i$}} \\ 0 & & \end{pmatrix},$$

where $(B_1, \dots, B_r) \in \mathcal{R}_{2,r}^{\mathrm{irr}}$. Let B_{01} be the space of representations of such form with respect to the standard basis. The stabilizer of B_{01} (i.e., the set $H_{01} \subset \mathrm{PGL}(3, \mathbb{C})$ sending B_{01} to itself) is

$$H_{01} = \left\{ \begin{pmatrix} (\det B)^{-1} & a & b \\ 0 & \multicolumn{2}{c}{\multirow{2}{*}{B}} \\ 0 & & \end{pmatrix} \right\} \cong \mathrm{GL}(2, \mathbb{C}) \times \mathbb{C}^2.$$

This means that there is a fibration $H_{01} \to B_{01} \times \mathrm{PGL}(3, \mathbb{C}) \to R_{01}$. Hence

$$e(R_{01}) = (q-1)^r q^{2r} e(\mathcal{R}_{2,r}^{\mathrm{irr}}) \frac{e(\mathrm{PGL}(3, \mathbb{C}))}{q^2 e(\mathrm{GL}(2, \mathbb{C}))}.$$

R_{02} is formed by representations $\rho = (A_1, \dots, A_r)$ which have a common two-dimensional space and upon which it acts irreducibly, that is,

$$A_i = \begin{pmatrix} \multicolumn{2}{c}{\multirow{2}{*}{$\lambda_i B_i$}} & 0 \\ & & 0 \\ b_i & c_i & \lambda_i^{-2} \end{pmatrix},$$

where $(B_1, \dots, B_r) \in \mathcal{R}_{2,r}^{\mathrm{irr}}$. The stabilizer is now

$$H_{02} = \left\{ \begin{pmatrix} \multicolumn{2}{c}{\multirow{2}{*}{B}} & 0 \\ & & 0 \\ a & b & (\det B)^{-1} \end{pmatrix} \right\} \cong \mathrm{GL}(2, \mathbb{C}) \times \mathbb{C}^2.$$

Hence

$$e(R_{02}) = (q-1)^r q^{2r} e(\mathcal{R}_{2,r}^{\mathrm{irr}}) \frac{e(\mathrm{PGL}(3, \mathbb{C}))}{q^2 e(\mathrm{GL}(2, \mathbb{C}))}.$$

The intersection $R_{01} \cap R_{02}$ consists of those representations with $b_i = c_i = 0$, which have stabilizer $\mathrm{GL}(2, \mathbb{C})$, hence

$$e(R_{01} \cap R_{02}) = (q-1)^r e(\mathcal{R}_{2,r}^{\mathrm{irr}}) \frac{e(\mathrm{PGL}(3, \mathbb{C}))}{e(\mathrm{GL}(2, \mathbb{C}))}.$$

Finally $e(R_0) = e(R_{01}) + e(R_{02}) - e(R_{01} \cap R_{02}) = 2e(R_{01}) - e(R_{01} \cap R_{02})$. Note that the remaining representations have a full invariant flag.

(ii) R_1 is formed by representations $\rho = (A_1, \ldots, A_r)$ such that the eigenvalues of all A_i are equal (and hence cubic roots of unity). This consists of the following substrata:

- R_{11} consisting of matrices
$$A_i = \begin{pmatrix} \xi_i & 0 & 0 \\ 0 & \xi_i & 0 \\ 0 & 0 & \xi_i \end{pmatrix},$$
where $\xi_i^3 = 1$. So $e(R_{11}) = 3^r$.
- R_{12} formed by matrices of type
$$A_i = \begin{pmatrix} \xi_i & 0 & 0 \\ 0 & \xi_i & a_i \\ 0 & 0 & \xi_i \end{pmatrix},$$
with $\xi_i^3 = 1$ and $(a_1, \ldots, a_r) \neq 0$. Then the stabilizer is
$$H_{12} = \left\{ \begin{pmatrix} \mu^{-1}\gamma^{-1} & 0 & b \\ a & \mu & c \\ 0 & 0 & \gamma \end{pmatrix} \right\} \cong (\mathbb{C}^*)^2 \times \mathbb{C}^3.$$
So
$$e(R_{12}) = 3^r (q^r - 1) \frac{e(\mathrm{PGL}(3, \mathbb{C}))}{(q-1)^2 q^3}.$$
- R_{13} formed by matrices of type
$$A_i = \begin{pmatrix} \xi_i & 0 & a_i \\ 0 & \xi_i & b_i \\ 0 & 0 & \xi_i \end{pmatrix},$$
with $\xi_i^3 = 1$ and $(a_1, \ldots, a_r), (b_1, \ldots, b_r)$ linearly independent. Note that when they are linearly dependent, one may arrange a basis so that it belongs to the stratum R_{12}. Then the stabilizer is
$$H_{13} = \left\{ \begin{pmatrix} \multicolumn{2}{c}{\multirow{2}{*}{A}} & b \\ & & c \\ 0 & 0 & (\det A)^{-1} \end{pmatrix} \right\} \cong \mathrm{GL}(2, \mathbb{C}) \times \mathbb{C}^2.$$
Hence
$$e(R_{13}) = 3^r (q^r - 1)(q^r - q) \frac{e(\mathrm{PGL}(3, \mathbb{C}))}{(q^2 - 1)(q^2 - q) q^2}.$$

- R_{14} formed by matrices of type

$$A_i = \begin{pmatrix} \xi_i & a_i & b_i \\ 0 & \xi_i & 0 \\ 0 & 0 & \xi_i \end{pmatrix},$$

with $\xi_i^3 = 1$ and $(a_1, \dots, a_r)$, $(b_1, \dots, b_r)$ linearly independent. Note again that when they are linearly dependent, one may arrange a basis so that it belongs to the stratum R_{12}. Then the stabilizer is

$$H_{14} = \left\{ \begin{pmatrix} (\det A)^{-1} & b \;\; c \\ 0 & \\ 0 & A \end{pmatrix} \right\} \cong \mathrm{GL}(2, \mathbb{C}) \times \mathbb{C}^2.$$

Hence

$$e(R_{14}) = 3^r (q^r - 1)(q^r - q) \frac{e(\mathrm{PGL}(3, \mathbb{C}))}{(q^2 - 1)(q^2 - q)q^2}.$$

- R_{15} formed by matrices of type

$$A_i = \begin{pmatrix} \xi_i & a_i & b_i \\ 0 & \xi_i & c_i \\ 0 & 0 & \xi_i \end{pmatrix},$$

with $\xi_i^3 = 1$ and $(a_1, \dots, a_r)$, $(c_1, \dots, c_r)$ are both nonzero (if one of them is zero, then we are back in the case R_{13}). Then the stabilizer is

$$H_{15} = \left\{ \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} \right\}.$$

Hence

$$e(R_{14}) = 3^r (q^r - 1)^2 q^r \frac{e(\mathrm{PGL}(3, \mathbb{C}))}{(q-1)^2 q^3}.$$

All together, we have

$$e(R_1) = 3^r \big(1 + (1 + q + q^2)(q^{3r+1} + q^{3r} - 2q^{2r+1} + q - 1)\big).$$

(iii) R_2 is formed by matrices with eigenvalues $(\lambda_i, \lambda_i, \mu_i)$. Let $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_r)$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_r)$, with $\boldsymbol{\lambda} - \boldsymbol{\mu} \neq \mathbf{0}$. Note that $\mu_i = \lambda_i^{-2}$, so the parameter space has E-polynomial $(q-1)^r - 3^r$. We have the following substrata:

- R_{21} consists of representations of type

$$A_i = \begin{pmatrix} \lambda_i & 0 & 0 \\ 0 & \lambda_i & 0 \\ 0 & 0 & \mu_i \end{pmatrix}.$$

The stabilizer is $P(\mathrm{GL}(2,\mathbb{C})\times\mathbb{C}^*)\cong\mathrm{GL}(2,\mathbb{C})$, so

$$e(R_{21})=((q-1)^r-3^r)\frac{e(\mathrm{PGL}(3,\mathbb{C}))}{(q^2-1)(q^2-q)}\,.$$

- R_{22} consists of representations of type

$$A_i=\begin{pmatrix}\lambda_i & a_i & b_i\\ 0 & \lambda_i & c_i\\ 0 & 0 & \mu_i\end{pmatrix},$$

with $\boldsymbol{a}\neq\mathbf{0}$, $\boldsymbol{b}=(b_1,\dots,b_r)$, $\boldsymbol{c}=(c_1,\dots,c_r)\in\mathbb{C}^r$. The stabilizer is

$$H_{22}=\left\{\begin{pmatrix}a & b & c\\ 0 & d & e\\ 0 & 0 & f\end{pmatrix}\right\}.$$

Hence

$$e(R_{22})=((q-1)^r-3^r)(q^r-1)q^{2r}\frac{e(\mathrm{PGL}(3,\mathbb{C}))}{(q-1)^2q^3}\,.$$

- R_{23} consists of representations of type

$$A_i=\begin{pmatrix}\lambda_i & 0 & 0\\ 0 & \lambda_i & a_i\\ 0 & 0 & \mu_i\end{pmatrix},$$

with $\boldsymbol{a}\notin\langle\boldsymbol{\lambda}-\boldsymbol{\mu}\rangle$. If $\boldsymbol{a}=x(\boldsymbol{\lambda}-\boldsymbol{\mu})$, $x\in\mathbb{C}$, then we can arrange a basis so that this belongs to the stratum R_{21}. The stabilizer is

$$H_{23}=\left\{\begin{pmatrix}a & 0 & 0\\ b & c & d\\ 0 & 0 & e\end{pmatrix}\right\}.$$

So

$$e(R_{23})=((q-1)^r-3^r)(q^r-q)\frac{e(\mathrm{PGL}(3,\mathbb{C}))}{(q-1)^2q^2}\,.$$

- R_{24} consists of representations of type

$$A_i=\begin{pmatrix}\lambda_i & 0 & a_i\\ 0 & \lambda_i & b_i\\ 0 & 0 & \mu_i\end{pmatrix},$$

with $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{\lambda}-\boldsymbol{\mu}$ linearly independent (if they where linearly dependent, one can arrange a basis so that we go back to case R_{23}). The stabilizer is

$$H_{24}=\left\{\begin{pmatrix} A & & a \\ & & b \\ 0 & 0 & (\det A)^{-1}\end{pmatrix}\right\}.$$

Hence

$$e(R_{24})=((q-1)^r-3^r)(q^r-q)(q^r-q^2)\frac{e(\mathrm{PGL}(3,\mathbb{C}))}{(q^2-1)(q^2-q)q^2}\,.$$

- R_{25} consists of representations of type

$$A_i=\begin{pmatrix}\mu_i & a_i & b_i\\ 0 & \lambda_i & c_i\\ 0 & 0 & \lambda_i\end{pmatrix},$$

with $\boldsymbol{a}\notin\langle\boldsymbol{\lambda}-\boldsymbol{\mu}\rangle$, $\boldsymbol{c}\neq\boldsymbol{0}$. The stabilizer is

$$H_{25}=\left\{\begin{pmatrix}a & b & c\\ 0 & d & e\\ 0 & 0 & f\end{pmatrix}\right\}.$$

Hence

$$e(R_{25})=((q-1)^r-3^r)(q^r-1)(q^r-q)q^r\,\frac{e(\mathrm{PGL}(3,\mathbb{C}))}{(q-1)^2q^3}\,.$$

- R_{26} consists of representations of type

$$A_i=\begin{pmatrix}\mu_i & 0 & b_i\\ 0 & \lambda_i & c_i\\ 0 & 0 & \lambda_i\end{pmatrix},$$

with $\boldsymbol{b}\notin\langle\boldsymbol{\lambda}-\boldsymbol{\mu}\rangle$, $\boldsymbol{c}\neq\boldsymbol{0}$. (If $\boldsymbol{b}$ is a multiple of $\boldsymbol{\lambda}-\boldsymbol{\mu}$, then we can arrange with a suitable basis that $\boldsymbol{b}=0$, and this belongs to the substrata R_{22}). The stabilizer is

$$H_{26}=\left\{\begin{pmatrix}a & 0 & b\\ 0 & c & d\\ 0 & 0 & e\end{pmatrix}\right\}.$$

Hence

$$e(R_{26})=((q-1)^r-3^r)(q^r-1)(q^r-q)\frac{e(\mathrm{PGL}(3,\mathbb{C}))}{(q-1)^2q^2}\,.$$

- R_{27} consists of representations of type

$$A_i = \begin{pmatrix} \mu_i & a_i & 0 \\ 0 & \lambda_i & 0 \\ 0 & 0 & \lambda_i \end{pmatrix},$$

with $\boldsymbol{a} \notin \langle \boldsymbol{\lambda} - \boldsymbol{\mu} \rangle$. The stabilizer is

$$H_{27} = \left\{ \begin{pmatrix} a & b & 0 \\ 0 & c & 0 \\ 0 & d & e \end{pmatrix} \right\}.$$

Hence

$$e(R_{27}) = ((q-1)^r - 3^r)(q^r - q) \frac{e(\mathrm{PGL}(3, \mathbb{C}))}{(q-1)^2 q^2} \, .$$

- R_{28} consists of representations of type

$$A_i = \begin{pmatrix} \mu_i & a_i & b_i \\ 0 & \lambda_i & 0 \\ 0 & 0 & \lambda_i \end{pmatrix},$$

with $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\lambda} - \boldsymbol{\mu}$ linearly independent (otherwise we can reduce to the case R_{27}). The stabilizer is

$$H_{28} = \left\{ \begin{pmatrix} (\det A)^{-1} & b \ \ c \\ 0 & \\ 0 & \quad A \end{pmatrix} \right\}.$$

Hence

$$e(R_{28}) = ((q-1)^r - 3^r)(q^r - q)(q^r - q^2) \frac{e(\mathrm{PGL}(3, \mathbb{C}))}{(q^2-q)(q^2-1)q^2} \, .$$

- R_{29} consists of representations of type

$$A_i = \begin{pmatrix} \lambda_i & a_i & b_i \\ 0 & \mu_i & c_i \\ 0 & 0 & \lambda_i \end{pmatrix},$$

with $\boldsymbol{a}, \boldsymbol{c} \notin \langle \boldsymbol{\lambda} - \boldsymbol{\mu} \rangle$. The stabilizer is

$$H_{29} = \left\{ \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} \right\}.$$

Hence

$$e(R_{29}) = ((q-1)^r - 3^r)(q^r - q)^2 q^r \frac{e(\mathrm{PGL}(3, \mathbb{C}))}{(q-1)^2 q^3} \, .$$

All together, we have

$$e(R_2) = ((q-1)^r - 3^r)(q^2+q+1)(3q^{3r+1} + 3q^{3r} - 2q^{2r+2} - 4q^{2r+1} + q^3).$$

(iv) R_3 is formed by matrices with eigenvalues $(\lambda_i, \mu_i, \gamma_i)$ where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_r)$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_r)$, and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_r)$ are distinct. Note that $\lambda_i \mu_i \gamma_i = 1$ for all $1 \le i \le r$. The base B_r parametrizing $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\gamma})$ has E-polynomial $e(B_r) = (q-1)^{2r} - 3(q-1)^r + 2 \cdot 3^r$.

- R_{31} consists of representations of type

$$A_i = \begin{pmatrix} \lambda_i & 0 & 0 \\ 0 & \mu_i & 0 \\ 0 & 0 & \gamma_i \end{pmatrix}.$$

Then the stabilizer is $D \times \Sigma_3$, where D is the diagonal matrices. So we have to compute the E-polynomial of the quotient $R_{31} = (\mathrm{PGL}(3,\mathbb{C})/D \times B_r)/\Sigma_3$. We start by computing $e_{\Sigma_3}(B_r)$. Let $B_r' = \{(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\gamma}) \in (\mathbb{C}^*)^{3r} \mid \boldsymbol{\lambda}\boldsymbol{\mu}\boldsymbol{\gamma} = (1, \dots, 1)\}$. This is $B_r' = (B')^r$, in the notation of Remark 6. Then

$$e_{\Sigma_3}(B_r') = e_{\Sigma_3}(B')^r = (q^2 T + S - qV)^r.$$

Using the properties $T \otimes T = T$, $T \otimes S = S$, $T \otimes V = V$, $S \otimes S = T$, $S \otimes V = V$, $V \otimes V = T \oplus S \oplus V$, it is easy to see that $V^b = a_b V + a_{b-1}(T+S)$, where $a_b = a_{b-1} + 2a_{b-2}$, with $a_0 = 0$, $a_1 = 1$. This recurrence solves as $a_b = (2^b - (-1)^b)/3$. Therefore

$$\begin{aligned}
e_{\Sigma_3}(B_r') &= (q^2T + S - qV)^r \\
&= \sum \frac{r!}{(r-a-b)!a!b!} q^{2(r-a-b)} S^a (-q)^b V^b \\
&= \sum \frac{r!}{(r-a)!a!} q^{2(r-a)} S^a \\
&\quad + \sum_{b>0} \frac{r!}{(r-a-b)!a!b!} q^{2(r-a-b)} S^a (-q)^b V^b \\
&= \tfrac{1}{2}((q^2+1)^r + (q^2-1)^r)T + \tfrac{1}{2}((q^2+1)^r - (q^2-1)^r)S \\
&\quad + \sum_{b>0} \frac{r!}{(r-a-b)!a!b!} q^{2(r-a-b)} S^a (-q)^b \\
&\qquad \times \big(\tfrac{1}{3}(2^b - (-1)^b)V + \tfrac{1}{3}(2^{b-1} - (-1)^{b-1})(T+S)\big) \\
&= \tfrac{1}{2}\big((q^2+1)^r + (q^2-1)^r\big)T + \tfrac{1}{2}\big((q^2+1)^r - (q^2-1)^r\big)S \\
&\quad + \tfrac{1}{3}\big((q^2-2q+1)^r - (q^2+q+1)^r\big)V \\
&\quad + \tfrac{1}{3}\big(\tfrac{1}{2}(q^2-2q+1)^r + (q^2+q+1)^r - \tfrac{3}{2}(q^2+1)^r\big)(T+S)
\end{aligned} \tag{4}$$

$$= \left(\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{6}(q-1)^{2r} + \tfrac{1}{3}(q^2+q+1)^r\right)T$$
$$+ \left(-\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{6}(q-1)^{2r} + \tfrac{1}{3}(q^2+q+1)^r\right)S$$
$$+ \tfrac{1}{3}\left((q-1)^{2r} - (q^2+q+1)^r\right)V.$$

Now we have to look at the part that we removed:

$$C_r = \{(\boldsymbol{\lambda}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^{-2}) \mid \boldsymbol{\lambda} \in (\mathbb{C}^*)^r\} \cup \{(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{-2}, \boldsymbol{\lambda}) \mid \boldsymbol{\lambda} \in (\mathbb{C}^*)^r\} \cup \{(\boldsymbol{\lambda}^{-2}, \boldsymbol{\lambda}, \boldsymbol{\lambda}) \mid \boldsymbol{\lambda} \in (\mathbb{C}^*)^r\}.$$

Then $e(C_r) = 3(q-1)^r - 2\cdot 3^r$. The quotient $C_r/\Sigma_3 \cong (\mathbb{C}^*)^r$, so $e(C_r/\Sigma_3) = (q-1)^r$. And for $H = \langle(1,2)\rangle$ we have $C_r/H \cong \{(\boldsymbol{\lambda}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^{-2}) \mid \boldsymbol{\lambda} \in (\mathbb{C}^*)^r\} \cup \{(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{-2}, \boldsymbol{\lambda}) \mid \boldsymbol{\lambda} \in (\mathbb{C}^*)^r\}$, so $e(C_r/H) = 2(q-1)^r - 3^r$. Hence,

$$e_{\Sigma_3}(C_r) = (q-1)^r T + ((q-1)^r - 3^r)V.$$

For $B_r = B_r' - C_r$, we have

$$e_{\Sigma_3}(B_r) = \left(\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{6}(q-1)^{2r} + \tfrac{1}{3}(q^2+q+1)^r - (q-1)^r\right)T \tag{5}$$
$$+ \left(-\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{6}(q-1)^{2r} + \tfrac{1}{3}(q^2+q+1)^r\right)S$$
$$+ \left(\tfrac{1}{3}(q-1)^{2r} - \tfrac{1}{3}(q^2+q+1)^r - (q-1)^r + 3^r\right)V.$$

Hence Formula (3) and Lemma 5 imply

$$e(R_{31}) = aa' + bb' + cc'$$
$$= \left(\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{6}(q-1)^{2r} + \tfrac{1}{3}(q^2+q+1)^r - (q-1)^r\right)q^6$$
$$+ \left(-\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{6}(q-1)^{2r} + \tfrac{1}{3}(q^2+q+1)^r\right)q^3$$
$$+ \left(\tfrac{1}{3}(q-1)^{2r} - \tfrac{1}{3}(q^2+q+1)^r - (q-1)^r + 3^r\right)(q^5+q^4).$$

- R_{32} consists of representations of type

$$A_i = \begin{pmatrix} \lambda_i & 0 & 0 \\ 0 & \mu_i & a_i \\ 0 & 0 & \gamma_i \end{pmatrix},$$

with $\boldsymbol{a} \notin \langle \boldsymbol{\mu} - \boldsymbol{\gamma} \rangle$. The stabilizer is

$$H_{32} = \left\{ \begin{pmatrix} a & 0 & 0 \\ 0 & b & c \\ 0 & 0 & d \end{pmatrix} \right\}.$$

Hence,

$$e(R_{32}) = ((q-1)^{2r} - 3(q-1)^r + 2\cdot 3^r)(q^r - q)\frac{e(\mathrm{PGL}(3,\mathbb{C}))}{(q-1)^2 q}\cdot$$

- R_{33} consists of representations of type

$$A_i = \begin{pmatrix} \lambda_i & 0 & a_i \\ 0 & \mu_i & b_i \\ 0 & 0 & \gamma_i \end{pmatrix},$$

with $\boldsymbol{a} \notin \langle \boldsymbol{\lambda} - \boldsymbol{\gamma} \rangle$, $\boldsymbol{b} \notin \langle \boldsymbol{\mu} - \boldsymbol{\gamma} \rangle$. The stabilizer is

$$H_{33} = \left\{ \begin{pmatrix} a & 0 & b \\ 0 & c & d \\ 0 & 0 & e \end{pmatrix} \right\} \times \mathbb{Z}_2,$$

where $\mathbb{Z}_2$ permutes the eigenvalues λ_i, μ_i. Therefore,

$$R_{33} = \big(B_r \times (\mathbb{C}^r - \mathbb{C})^2 \times (\mathrm{PGL}(3, \mathbb{C})/H_{33})\big)/\mathbb{Z}_2.$$

By (5), we have that

$$\begin{aligned} e_H(B_r) = \big(\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{2}(q-1)^{2r} - 2(q-1)^r + 3^r\big)T \\ + \big(-\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{2}(q-1)^{2r} - (q-1)^r + 3^r\big)N, \end{aligned}$$

since under $H \subset \Sigma_3$, we have $T \mapsto T$, $S \mapsto N$, $V \mapsto T + N$. For the second factor, $e((\mathbb{C}^r - \mathbb{C})^2) = (q^r - q)^2$ and $e(\mathrm{Sym}^2(\mathbb{C}^r - \mathbb{C})) = q^{2r} - q^{r+1}$, so

$$e_H((\mathbb{C}^r - \mathbb{C})^2) = (q^{2r} - q^{r+1})T + (q^2 - q^{r+1})N.$$

Finally, $\mathrm{PGL}(3, \mathbb{C})/H_{33} \cong \mathbb{P}^2 \times \mathbb{P}^2 - \Delta$, by considering the first two columns of the matrix, where Δ is the diagonal. As $e(\mathbb{P}^2 \times \mathbb{P}^2 - \Delta) = (1+q+q^2)(q+q^2)$ and $e(\mathrm{Sym}^2 \mathbb{P}^2 - \overline{\Delta}) = q^4 + q^3 + q^2$, we have

$$e_H(\mathrm{PGL}(3, \mathbb{C})/H_{33}) = (q^4 + q^3 + q^2)T + (q^3 + q^2 + q)N.$$

Hence,

$$\begin{aligned} e(R_{33}) = \big(\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{2}(q-1)^{2r} - 2(q-1)^r + 3^r\big) \\ \times \big((q^{2r} - q^{r+1})(q^4 + q^3 + q^2) + (q^2 - q^{r+1})(q^3 + q^2 + q)\big) \\ + \big(-\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{2}(q-1)^{2r} - (q-1)^r + 3^r\big) \\ \times \big((q^{2r} - q^{r+1})(q^3 + q^2 + q) + (q^2 - q^{r+1})(q^4 + q^3 + q^2)\big). \end{aligned}$$

- R_{34} consists of representations of type

$$A_i = \begin{pmatrix} \lambda_i & a_i & b_i \\ 0 & \mu_i & 0 \\ 0 & 0 & \gamma_i \end{pmatrix},$$

with $\boldsymbol{a} \notin \langle \boldsymbol{\lambda} - \boldsymbol{\mu} \rangle$, $\boldsymbol{b} \notin \langle \boldsymbol{\lambda} - \boldsymbol{\gamma} \rangle$. The stabilizer is

$$H_{34} = \left\{ \begin{pmatrix} a & b & c \\ 0 & d & 0 \\ 0 & 0 & e \end{pmatrix} \right\}.$$

The computations are analogous to the case of R_{33}, so $e(R_{33}) = e(R_{34})$.

- R_{35} consists of representations of type

$$A_i = \begin{pmatrix} \lambda_i & a_i & b_i \\ 0 & \mu_i & c_i \\ 0 & 0 & \gamma_i \end{pmatrix},$$

with $\boldsymbol{a} \notin \langle \boldsymbol{\lambda} - \boldsymbol{\mu} \rangle$, $\boldsymbol{c} \notin \langle \boldsymbol{\mu} - \boldsymbol{\gamma} \rangle$. The stabilizer is

$$H_{35} = \left\{ \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} \right\}.$$

Hence

$$e(R_{35}) = ((q-1)^{2r} - 3(q-1)^r + 2 \cdot 3^r)(q^r - q)^2 q^r \frac{e(\mathrm{PGL}(3, \mathbb{C}))}{(q-1)^2 q^3} \cdot$$

All together, we have

$$\begin{aligned} e(R_3) = {} & \big(2 \cdot 3^r - 3(q-1)^r + (q-1)^{2r}\big) \\ & \qquad \times (q+1)(q^2+q+1)(q^r-q)(q^2+q^{2r}-q^{1+r}) \\ & + \big(2 \cdot 3^r - 2(q-1)^r + (q-1)^{2r} - (q^2-1)^r\big) \\ & \qquad \times q(q^2+q+1)(q^r-q)(q^r-q^2) \\ & + \big(2 \cdot 3^r - 4(q-1)^r + (q-1)^{2r} + (q^2-1)^r\big) \\ & \qquad \times q^2(q^2+q+1)(q^r-1)(q^r-q) \\ & + \tfrac{1}{6} q^3 \big((q-1)^{2r} - 3(q^2-1)^r + 2(q^2+q+1)^r \\ & \qquad\qquad + 2q(q+1)\big(3^{r+1} - 3(q-1)^r + (q-1)^{2r} - (q^2+q+1)^r\big)\big) \\ & + \tfrac{1}{6} q^6 \big(-6(q-1)^r + (q-1)^{2r} + 3(q^2-1)^r + 2(q^2+q+1)^r\big). \end{aligned}$$

Therefore,

$$\begin{aligned} e(\mathcal{R}^{\mathrm{red}}_{r,3}) = {} & \tfrac{1}{3}(q^2+q+1)^r (q-1)^2 q^3 (q+1) \\ & + (q^2+q+1)(2q^{2r}-q^2)(q-1)^{2r} q^r (q+1)^r \\ & - \tfrac{1}{3}(q-1)^{2r}(q+1)(q^2+q+1)(3q^{3r} - 3q^{r+2} + q^3), \end{aligned}$$

and so,

$$e(\mathcal{R}^{\mathrm{irr}}_{r,3}) = e(\mathcal{R}_{r,3}) - e(\mathcal{R}^{\mathrm{red}}_{r,3}) = e(\mathrm{SL}(3,\mathbb{C}))^r - e(\mathcal{R}^{\mathrm{red}}_{r,3}),$$

and consequently,

$$e(\mathcal{M}^{\mathrm{irr}}_{r,3}) = e(\mathcal{R}^{\mathrm{irr}}_{r,3})/e(\mathrm{PGL}(3,\mathbb{C})) = e(\mathrm{SL}(3,\mathbb{C}))^{r-1} - e(\mathcal{R}^{\mathrm{red}}_{r,3})/e(\mathrm{SL}(3,\mathbb{C})).$$

E-polynomial of the moduli of reducible representations. To compute $e(\mathcal{M}_{r,3})$, it remains to compute the moduli space of reducible representations $\mathcal{M}^{\mathrm{red}}_{r,3}$. This is formed by two strata:

(i) M_0 formed by semisimple representations which split into irreducible representations of ranks 1 and 2, that is, of the form:

$$A_i = \begin{pmatrix} \lambda_i^{-2} & 0 & 0 \\ 0 & & \\ 0 & \multicolumn{2}{c}{\lambda_i B_i} \end{pmatrix},$$

where $(B_1,\ldots,B_r) \in \mathcal{M}^{\mathrm{irr}}_{r,2}$. So $e(M_0) = (q-1)^r e(\mathcal{M}^{\mathrm{irr}}_{r,2})$.

(ii) M_1 formed by semisimple representations which split into three irreducible representations of rank 1. These are given by eigenvalues $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_r)$, $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_r)$, $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_r)$ in $(\mathbb{C}^*)^r$, where $\lambda_i\mu_i\gamma_i = 1$ for all $1 \le i \le r$. This is the space B'_r whose E-polynomial has been computed in (4). Thus

$$e(M_1) = e(B'_r/\Sigma_3) = \tfrac{1}{2}(q^2-1)^r + \tfrac{1}{6}(q-1)^{2r} + \tfrac{1}{3}(q^2+q+1)^r.$$

Finally, $e(\mathcal{M}^{\mathrm{red}}_{r,3}) = e(M_0) + e(M_1)$, and adding up everything we get

$$\begin{aligned} e(\mathcal{M}_{r,3}) &= e(\mathcal{M}^{\mathrm{irr}}_{r,3}) + e(\mathcal{M}^{\mathrm{red}}_{r,3}) \\ &= (q^8-q^6-q^5+q^3)^{r-1} + (q-1)^{2r-2}(q^{3r-3}-q^r) \\ &\quad + \tfrac{1}{6}(q-1)^{2r-2}q(q+1) + \tfrac{1}{2}(q^2-1)^{r-1}q(q-1) \\ &\quad + \tfrac{1}{3}(q^2+q+1)^{r-1}q(q+1) \\ &\quad - (q-1)^{r-1}q^{r-1}(q^2-1)^{r-1}(2q^{2r-2}-q). \end{aligned}$$

This completes the main part of the proof of Theorem 1. It remains to show that $e(\mathcal{M}_{r,3}) = e(\overline{\mathcal{M}}_{r,3})$, which we do in the following section.

Remark 8. By [Florentino and Lawton 2012], the singular locus of $\mathcal{M}_{r,3}$ is exactly the reducible locus (and so the smooth locus is its complement). Therefore, the above computation of M_0 and M_1 gives the E-polynomial of the singular locus of $\mathcal{M}_{r,3}$. Likewise, $e(\mathcal{M}^{\mathrm{irr}}_{r,3})$ is the E-polynomial of the smooth locus of $\mathcal{M}_{r,3}$. Moreover, by [Florentino and Lawton 2014], the abelian character variety $\mathcal{M}(\mathbb{Z}^r, \mathrm{SL}(3,\mathbb{C}))$ is exactly the diagonalizable representations in $\mathcal{M}_{r,3}$. The above computation of M_1 gives the E-polynomial of $\mathcal{M}(\mathbb{Z}^r, \mathrm{SL}(3,\mathbb{C}))$. In each case, setting $q=1$ gives the Euler characteristic of the corresponding space.

7. E-polynomials of character varieties for F_r, $r>1$, and PGL(3, $\mathbb{C}$)

In this final section, we focus on the space of representations

$$\begin{aligned}\overline{\mathcal{R}}_{r,3} &= \mathrm{Hom}(F_r, \mathrm{PGL}(3,\mathbb{C})) = \{\rho : F_r \to \mathrm{PGL}(3,\mathbb{C})\}\\ &= \{(A_1,\dots,A_r) \mid A_i \in \mathrm{PGL}(3,\mathbb{C})\} = \mathrm{PGL}(3,\mathbb{C})^r\end{aligned}$$

and the character variety

$$\overline{\mathcal{M}}_{r,3} = \mathrm{Hom}(F_r, \mathrm{PGL}(3,\mathbb{C}))/\!\!/\, \mathrm{PGL}(3,\mathbb{C}).$$

Let $\zeta = e^{2\pi\sqrt{-1}/3}$, and let $\mathbb{Z}_3 = \{1, \zeta, \zeta^2\}$ be the space of cubic roots of unity. Then $\mathrm{PGL}(3,\mathbb{C}) = \mathrm{SL}(3,\mathbb{C})/\mathbb{Z}_3$,

$$\overline{\mathcal{R}}_{r,3} = \mathcal{R}_{r,3}/(\mathbb{Z}_3)^r, \quad \text{and} \quad \overline{\mathcal{M}}_{r,3} = \mathcal{M}_{r,3}/(\mathbb{Z}_3)^r,$$

where $(\zeta^{a_1},\dots,\zeta^{a_r})$ acts as $(A_1,\dots,A_r) \mapsto (\zeta^{a_1}A_1,\dots,\zeta^{a_r}A_r)$. Clearly $\overline{\mathcal{R}}^{\mathrm{red}}_{r,3} = \mathcal{R}^{\mathrm{red}}_{r,3}/(\mathbb{Z}_3)^r$ and $\overline{\mathcal{R}}^{\mathrm{irr}}_{r,3} = \mathcal{R}^{\mathrm{irr}}_{r,3}/(\mathbb{Z}_3)^r$.

We know from Lemma 4 that $e(\mathrm{PGL}(3,\mathbb{C})) = e(\mathrm{SL}(3,\mathbb{C}))$. Let us see now that $e(\overline{\mathcal{R}}^{\mathrm{red}}_{r,3}) = e(\mathcal{R}^{\mathrm{red}}_{r,3})$. We stratify $\overline{\mathcal{R}}^{\mathrm{red}}_{r,3} = \bar{R}_0 \sqcup \bar{R}_1 \sqcup \bar{R}_2 \sqcup \bar{R}_3$, where $\bar{R}_i = R_i/(\mathbb{Z}_3)^r$ and the R_i, $i = 0,1,2,3$, have been defined in Section 6.

We now list the strata with the computation of their E-polynomials:

(i) $\bar{R}_0 = \bar{R}_{01} \cup \bar{R}_{02}$, where $\bar{R}_{0j} = R_{0j}/(\mathbb{Z}_3)^r$, $j=1,2$. To compute $e(\bar{R}_{01})$, recall that R_{01} is formed by representations $\rho = (A_1,\dots,A_r)$ with

$$A_i = \begin{pmatrix} \lambda_i^{-2} & b_i & c_i \\ 0 & \multicolumn{2}{c}{\multirow{2}{*}{$\lambda_i B_i$}} \\ 0 & & \end{pmatrix},$$

where $(B_1,\dots,B_r) \in \mathcal{R}^{\mathrm{irr}}_{2,r}$. The action of ζ^{a_i} on A_i is given by $(\lambda_i, b_i, c_i, B_i) \mapsto (\zeta^{a_i}\lambda_i, \zeta^{a_i}b_i, \zeta^{a_i}c_i, \zeta^{a_i}B_i)$. Note that $\mathbb{C}/\mathbb{Z}_3 \cong \mathbb{C}$ and $\mathbb{C}^*/\mathbb{Z}_3 \cong \mathbb{C}^*$, so the relevant cohomology is invariant. Therefore

$$e\big(((\mathbb{C}^*)^r \times \mathbb{C}^r \times \mathbb{C}^r \times \mathcal{R}^{\mathrm{irr}}_{2,r})/(\mathbb{Z}_3)^r\big) = e(\mathbb{C}^*)^r e(\mathbb{C})^r e(\mathbb{C})^r e(\mathcal{R}^{\mathrm{irr}}_{2,r}/(\mathbb{Z}_3)^r).$$

This means that $e(\bar{R}_{01}) = e(R_{01})$. Analogously $e(\bar{R}_{02}) = e(R_{02})$ and $e(\bar{R}_{01} \cap \bar{R}_{02}) = e(R_{01} \cap R_{02})$, so $e(\bar{R}_0) = e(R_0)$.

(ii) $\bar{R}_1 = R_1/(\mathbb{Z}_3)^r$. Note that R_1 is formed by 3^r copies of the same subvariety. Hence

$$e(\bar{R}_1) = \frac{e(R_1)}{3^r} = 1 + (1+q+q^2)(q^{3r+1} + q^{3r} - 2q^{2r+1} + q - 1).$$

(iii) $\bar{R}_2 = R_2/(\mathbb{Z}_3)^r$. Recall that R_2 is formed by matrices with eigenvalues $(\lambda_i, \lambda_i, \mu_i)$ where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_r) \in P = (\mathbb{C}^*)^r - \{1, \zeta, \zeta^2\}^r$. Now

$$\bar{P} = P/(\mathbb{Z}_3)^r \cong (\mathbb{C}^*)^r - \{(1, 1, \dots, 1)\},$$

so $e(\bar{P}) = (q-1)^r - 1$. It is more or less straightforward to see that $\bar{R}_2$ can be stratified by $\bar{R}_{2j} = R_{2j}/(\mathbb{Z}_3)^r$, $j = 1, 2, \dots, 9$. For each $\bar{R}_{2j}$ the computation of $e(\bar{R}_{2j})$ is the same as that of $e(R_{2j})$, but replacing $e(P) = (q-1)^r - 3^r$ by $e(\bar{P}) = (q-1)^r - 1$. Hence

$$e(\bar{R}_2) = ((q-1)^r - 1)(q^2+q+1)(3q^{3r+1} + 3q^{3r} - 2q^{2r+2} - 4q^{2r+1} + q^3).$$

(iv) $\bar{R}_3 = R_3/(\mathbb{Z}_3)^r$. We follow the lines of the computation of $e(R_3)$. The base for the space of eigenvalues is $\bar{B}_r = B_r/(\mathbb{Z}_3)^r$ with $e(\bar{B}_r) = (q-1)^{2r} - 3(q-1)^r + 2$.

- Let $\bar{R}_{31} = R_{31}/(\mathbb{Z}_3)^r \cong (\mathrm{PGL}(3,\mathbb{C})/D \times \bar{B}_r)/\Sigma_3$. If $\bar{B}'_r = B'_r/(\mathbb{Z}_3)^r$, then easily $e_{\Sigma_3}(\bar{B}'_r) = e_{\Sigma_3}(\bar{B}')^r = (q^2T + S - qV)^r = e_{\Sigma_3}(B'_r)$. For $\bar{C}_r = C_r/(\mathbb{Z}_3)^r$, we have instead that $e_{\Sigma_3}(\bar{C}_r) = (q-1)^r T + ((q-1)^r - 1)V$, so $\bar{B}_r = \bar{B}'_r - \bar{C}_r$ has

$$\begin{aligned} e_{\Sigma_3}(B_r) = {} & \big(\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{6}(q-1)^{2r} + \tfrac{1}{3}(q^2+q+1)^r - (q-1)^r\big)T \\ & + \big(-\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{6}(q-1)^{2r} + \tfrac{1}{3}(q^2+q+1)^r\big)S \\ & + \big(\tfrac{1}{3}(q-1)^{2r} - \tfrac{1}{3}(q^2+q+1)^r - (q-1)^r + 1\big)V, \end{aligned}$$

  and

$$\begin{aligned} e(\bar{R}_{31}) = {} & \big(\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{6}(q-1)^{2r} + \tfrac{1}{3}(q^2+q+1)^r - (q-1)^r\big)q^6 \\ & + \big(-\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{6}(q-1)^{2r} + \tfrac{1}{3}(q^2+q+1)^r\big)q^3 \\ & + \big(\tfrac{1}{3}(q-1)^{2r} - \tfrac{1}{3}(q^2+q+1)^r - (q-1)^r + 1\big)(q^5+q^4). \end{aligned}$$

- $\bar{R}_{32} = R_{32}/(\mathbb{Z}_3)^r$ has

$$e(\bar{R}_{32}) = ((q-1)^{2r} - 3(q-1)^r + 2)(q^r - q)\frac{e(\mathrm{PGL}(3,\mathbb{C}))}{(q-1)^2 q}\cdot$$

- $\bar{R}_{33} = R_{33}/(\mathbb{Z}_3)^r \cong \big(\bar{B}_r \times (\mathbb{C}^r - \mathbb{C})^2 \times (\mathrm{PGL}(3,\mathbb{C})/H_{33})\big)/\mathbb{Z}_2$, where $H = \mathbb{Z}_2$ acts by swapping the first two eigenvalues. Now

$$\begin{aligned} e_H(\bar{B}_r) = {} & \big(\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{2}(q-1)^{2r} - 2(q-1)^r + 1\big)T \\ & + \big(-\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{2}(q-1)^{2r} - (q-1)^r + 1\big)N, \end{aligned}$$

so

$$\begin{aligned}
e(\bar{R}_{33}) = \big(\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{2}(q-1)^{2r} - 2(q-1)^r + 1\big) \\
\times \big((q^{2r}-q^{r+1})(q^4+q^3+q^2) + (q^2-q^{r+1})(q^3+q^2+q)\big) \\
+ \big(-\tfrac{1}{2}(q^2-1)^r + \tfrac{1}{2}(q-1)^{2r} - (q-1)^r + 1\big) \\
\times \big((q^{2r}-q^{r+1})(q^3+q^2+q) + (q^2-q^{r+1})(q^4+q^3+q^2)\big).
\end{aligned}$$

- $\bar{R}_{34} = R_{34}/(\mathbb{Z}_3)^r$ has $e(\bar{R}_{34}) = e(\bar{R}_{33})$.
- $\bar{R}_{35} = R_{35}/(\mathbb{Z}_3)^r$ has

$$e(\bar{R}_{35}) = ((q-1)^{2r} - 3(q-1)^r + 2)(q^r-q)^2 q^r \, \frac{e(\mathrm{PGL}(3,\mathbb{C}))}{(q-1)^2 q^3} \cdot$$

All together, we have:

$$\begin{aligned}
e(\bar{R}_3) = {} & \big(2 - 3(q-1)^r + (q-1)^{2r}\big)(q+1)(q^2+q+1) \\
& \times (q^r-q)(q^2+q^{2r}-q^{r+1}) \\
& + \big(2 - 2(q-1)^r + (q-1)^{2r} - (q^2-1)^r\big) \\
& \times q(q^2+q+1)(q^r-q)(q^r-q^2) \\
& + \big(2 - 4(q-1)^r + (q-1)^{2r} + (q^2-1)^r\big) \\
& \times q^2(q^2+q+1)(q^r-1)(q^r-q) \\
& + \tfrac{1}{6} q^3 \big((q-1)^{2r} - 3(q^2-1)^r + 2(q^2+q+1)^r \\
& \quad + 2q(q+1)\big(3 - 3(q-1)^r + (q-1)^{2r} - (q^2+q+1)^r\big)\big) \\
& + \tfrac{1}{6} q^6 \big(-6(q-1)^r + (q-1)^{2r} + 3(q^2-1)^r + 2(q^2+q+1)^r\big).
\end{aligned}$$

Adding up all the contributions we get:

$$\begin{aligned}
e(\overline{\mathcal{R}}^{\mathrm{red}}_{r,3}) &= \tfrac{1}{3}(q^2+q+1)^r (q-1)^2 q^3 (q+1) \\
&\quad + (q^2+q+1)(2q^{2r}-q^2)(q-1)^{2r} q^r (q+1)^r \\
&\quad - \tfrac{1}{3}(q-1)^{2r}(q+1)(q^2+q+1)(3q^{3r} - 3q^{r+2} + q^3) \\
&= e(\mathcal{R}^{\mathrm{red}}).
\end{aligned}$$

From this $e(\overline{\mathcal{R}}^{\mathrm{irr}}_{r,3}) = e(\mathcal{R}^{\mathrm{irr}}_{r,3})$ and $e(\overline{\mathcal{M}}^{\mathrm{irr}}_{r,3}) = e(\mathcal{M}^{\mathrm{irr}}_{r,3})$.

The remaining thing to compute is $e(\overline{\mathcal{M}}^{\mathrm{red}}_{r,3})$. This is formed by two strata:

(i) $\overline{M}_0 = M_0/(\mathbb{Z}_3)^r \cong ((\mathbb{C}^*)^r \times \mathcal{M}^{\mathrm{irr}}_{r,2})/(\mathbb{Z}_3)^r$. Hence

$$e(\overline{M}_0) = (q-1)^r e(\overline{\mathcal{M}}^{\mathrm{irr}}_{r,2}) = e(M_0).$$

(ii) $\overline{M}_1 = M_1/(\mathbb{Z}_3)^r \cong ((\mathbb{C}^*)^r/(\mathbb{Z}_3)^r)/\Sigma_3 \cong (\mathbb{C}^*)^r/\Sigma_3$. So $e(\overline{M}_1) = e(M_1)$.

We get finally $e(\overline{\mathcal{M}}_{r,3}^{\mathrm{red}}) = e(\mathcal{M}_{r,3}^{\mathrm{red}})$. This concludes the proof of the equality $e(\overline{\mathcal{M}}_{r,3}) = e(\mathcal{M}_{r,3})$.

Remark 9. There is an arithmetic argument communicated to us by S. Mozgovoy to prove that $e(\overline{\mathcal{M}}_{r,n}) = e(\mathcal{M}_{r,n})$ for n odd. It goes as follows: find infinitely many primes p such that $p-1$ and n are coprime (by Dirichlet's theorem on arithmetic progressions); then $\mathrm{SL}(n, \mathbb{F}_p) \to \mathrm{PGL}(n, \mathbb{F}_p)$ is bijective and one gets a bijection between corresponding character varieties over $\mathbb{F}_p$. So the count number of points of $\mathcal{M}_{r,n}$ and $\overline{\mathcal{M}}_{r,n}$ over $\mathbb{F}_p$ coincide, and hence the E-polynomials coincide.

However this argument cannot be used for even n. Despite this, the E-polynomials for the $\mathrm{SL}(2, \mathbb{C})$-character varieties of free groups do equal those of $\mathrm{PGL}(2, \mathbb{C})$. We expect to address the case of $\mathrm{SL}(4, \mathbb{C})$ in future work.

Acknowledgments

We are grateful to S. Mozgovoy and M. Reineke for providing us with a copy of [Mozgovoy and Reineke 2015], and also for giving us an explicit formula for $e\big(\mathcal{M}(F_r, \mathrm{PGL}(3, \mathbb{C}))\big)$ for checking against our polynomials. We also thank the referee for helping improve the exposition of this article. Lawton was supported by the Simons Foundation Collaboration grant 245642, and the U.S. NSF grant DMS 1309376. Muñoz was partially supported by Project MICINN (Spain) MTM2010-17389. We also acknowledge support from U.S. NSF grants DMS 1107452, 1107263, 1107367 "RNMS: GEometric structures And Representation varieties" (the GEAR Network).

References

[Cavazos and Lawton 2014] S. Cavazos and S. Lawton, "E-polynomial of $\mathrm{SL}_2(\mathbb{C})$-character varieties of free groups", *Int. J. Math.* **25**:6 (2014), 1450058. MR 3225582 Zbl 1325.14065

[Deligne 1971] P. Deligne, "Théorie de Hodge, II", *Inst. Hautes Études Sci. Publ. Math.* 40 (1971), 5–57. MR 58 #16653a Zbl 0219.14007

[Deligne 1974] P. Deligne, "Théorie de Hodge, III", *Inst. Hautes Études Sci. Publ. Math.* 44 (1974), 5–77. MR 58 #16653b Zbl 0237.14003

[Florentino and Lawton 2009] C. Florentino and S. Lawton, "The topology of moduli spaces of free group representations", *Math. Ann.* **345**:2 (2009), 453–489. MR 2010h:14075 Zbl 1200.14093

[Florentino and Lawton 2012] C. Florentino and S. Lawton, "Singularities of free group character varieties", *Pacific J. Math.* **260**:1 (2012), 149–179. MR 3001789 Zbl 1264.14064

[Florentino and Lawton 2014] C. Florentino and S. Lawton, "Topology of character varieties of Abelian groups", *Topology Appl.* **173** (2014), 32–58. MR 3227204 Zbl 1300.14045

[Lawton 2007] S. Lawton, "Generators, relations and symmetries in pairs of 3×3 unimodular matrices", *J. Algebra* **313**:2 (2007), 782–801. MR 2008k:16039 Zbl 1119.13004

[Lawton 2008] S. Lawton, "Minimal affine coordinates for $\mathrm{SL}(3, \mathbb{C})$ character varieties of free groups", *J. Algebra* **320**:10 (2008), 3773–3810. MR 2009j:20060 Zbl 1157.14030

[Lawton 2010] S. Lawton, "Algebraic independence in SL(3, $\mathbb{C}$) character varieties of free groups", *J. Algebra* **324**:6 (2010), 1383–1391. MR 2011g:20071 Zbl 1209.14036

[Logares and Muñoz 2014] M. Logares and V. Muñoz, "Hodge polynomials of the SL(2, $\mathbb{C}$)-character variety of an elliptic curve with two marked points", *Int. J. Math.* **25**:14 (2014), 1450125. MR 3306833 Zbl 1316.14025

[Logares et al. 2013] M. Logares, V. Muñoz, and P. E. Newstead, "Hodge polynomials of SL(2, $\mathbb{C}$)-character varieties for curves of small genus", *Rev. Mat. Complut.* **26**:2 (2013), 635–703. MR 3068615

[Martínez and Muñoz 2015a] J. Martínez and V. Muñoz, "E-polynomial of SL(2, $\mathbb{C}$)-character varieties of complex curves of genus 3", preprint, 2015, Available at http://www.math.sci.osaka-u.ac.jp/ojm/pdf/3905.pdf. To appear in *Osaka J. Math.*

[Martínez and Muñoz 2015b] J. Martínez and V. Muñoz, "E-polynomials of the SL(2, $\mathbb{C}$)-character varieties of surface groups", *Int. Math. Res. Not.* **2015** (online publication June 2015).

[Mozgovoy and Reineke 2015] S. Mozgovoy and M. Reineke, "Arithmetic of character varieties of free groups", *Int. J. Math.* **26**:12 (2015), 1550100.

[Sikora 2013] A. S. Sikora, "Generating sets for coordinate rings of character varieties", *J. Pure Appl. Algebra* **217**:11 (2013), 2076–2087. MR 3057078 Zbl 1200.14093

Received March 24, 2015. Revised July 29, 2015.

SEAN LAWTON
DEPARTMENT OF MATHEMATICAL SCIENCES
GEORGE MASON UNIVERSITY
4400 UNIVERSITY DRIVE
FAIRFAX, VA 22030
UNITED STATES

slawton3@gmu.edu

VICENTE MUÑOZ
FACULTAD DE MATEMÁTICAS
UNIVERSIDAD COMPLUTENSE DE MADRID
PLAZA DE CIENCIAS 3
28040 MADRID
SPAIN

vicente.munoz@mat.ucm.es

PACIFIC JOURNAL OF MATHEMATICS
Vol. 282, No. 1, 2016
dx.doi.org/10.2140/pjm.2016.282.203

THE BLUM–HANSON PROPERTY FOR $\mathcal{C}(K)$ SPACES

PASCAL LEFÈVRE AND ÉTIENNE MATHERON

We show that if K is a compact metrizable space, then the Banach space $\mathcal{C}(K)$ has the so-called Blum–Hanson property exactly when K has finitely many accumulation points. We also show that the space $\ell_\infty(\mathbb{N}) = \mathcal{C}(\beta\mathbb{N})$ does not have the Blum–Hanson property.

1. Introduction

The following intriguing result is usually referred to as the *Blum–Hanson theorem* (see [Blum and Hanson 1960] and [Jones and Kuftinec 1971]): if T is a linear operator on a Hilbert space H with $\|T\| \le 1$, and if $x \in H$ is such that $T^n x \to 0$ weakly as $n \to \infty$, then the sequence $(T^n x)$ is *strongly mixing*, which means that every subsequence of $(T^n x)$ converges to 0 in the Cesàro sense; in other words,

$$\lim_{K\to\infty} \left\| \frac{1}{K} \sum_{i=1}^{K} T^{n_i} x \right\| = 0$$

for any increasing sequence of integers (n_i). (The terminology "strongly mixing" comes from [Berend and Bergelson 1986].)

Accordingly, a Banach space X is said to have the *Blum–Hanson property* if the Blum–Hanson theorem holds true on X; that is, if T is a linear operator on X such that $\|T\| \le 1$, then every weakly null T-orbit is strongly mixing. For example, it was shown rather recently in [Müller and Tomilov 2007] that $\ell_p(\mathbb{N})$ has the Blum–Hanson property for any $p \in [1, \infty)$. On the other hand, it is known since [Akcoglu et al. 1974] that $\mathcal{C}(\mathbb{T}^2)$, the space of all continuous real-valued functions on the torus $\mathbb{T}^2$, does not have this property. Further results and references can be found in [Lefèvre et al. 2015].

In this short note, we address the Blum–Hanson property for $\mathcal{C}(K)$ spaces. Our main result is the following.

Theorem 1.1. *Let K be a metrizable compact space. Then $\mathcal{C}(K)$ has the Blum–Hanson property if and only if K has finitely many accumulation points.*

MSC2010: primary 46E15, 47A35; secondary 46B25.
Keywords: Blum–Hanson property, spaces of continuous functions, Stone–Čech compactification.

This will be proved in the next section. In Section 3, we obtain in much the same way one nonmetrizable result, namely that the space $\ell_\infty(\mathbb{N}) = \mathcal{C}(\beta\mathbb{N})$ fails the Blum–Hanson property. Our two results can be put together to get a single theorem on the Blum–Hanson property for spaces of bounded continuous functions, which is done in Section 4. We conclude the paper by stating explicitly the "general principle" underlying our proofs.

2. Proof of Theorem 1.1

For the "if" part of the proof, we will make use of a result from [Lefèvre et al. 2015] which is stated as Lemma 2.1 below.

Let X be a Banach space. For any $x \in X$ and $t \in \mathbb{R}^+$, set

$$r_X(t,x) := \sup\Big\{\limsup_{n\to\infty} \|x + ty_n\|\Big\},$$

where the supremum is taken over all weakly null sequences $(y_n) \subset X$ with $\|y_n\| \le 1$.

Since $r_X(t,x)$ is 1-Lipschitz with respect to t, the quantity $r_X(t,x) - t$ is nonincreasing and hence it has a limit as $t \to \infty$, possibly equal to $-\infty$. Actually, this limit is nonnegative if X does not have the Schur property, i.e., there is at least one weakly null sequence in X which is not norm null.

For the needs of the present paper only, we shall say that the Banach space X has *property* (P) if, for every weakly null sequence $(x_k) \subset X$, it holds that

$$\lim_{k\to\infty} \lim_{t\to\infty} (r_X(t,x_k) - t) = 0.$$

The result we need is the following lemma; for the proof, see the remark just after Theorem 2.1 in [Lefèvre et al. 2015].

Lemma 2.1. *Property* (P) *implies the Blum–Hanson property.*

An extreme example of a space with property (P) is $X := c_0(\mathbb{N})$. Indeed, if $x \in c_0$ and if (z_n) is a weakly null sequence in c_0, then

$$\limsup_{n\to\infty} \|x + z_n\|_\infty = \max(\|x\|_\infty, \limsup \|z_n\|_\infty).$$

It follows that

$$(*) \qquad r_{c_0}(t,x) = \max(\|x\|, t),$$

so that $r_{c_0}(t,x) - t = 0$ whenever $t \ge \|x\|$, for any $x \in c_0$.

Let us also note the following useful stability property, whose proof is straightforward.

Remark 2.2. If $X_1, \dots, X_N$ are Banach spaces with property (P), then the ℓ_∞ direct sum $X_1 \oplus \cdots \oplus X_N$ also has (P).

We can now start the proof of Theorem 1.1.

Proof of Theorem 1.1. Let us denote by K' the set of all accumulation points of K. We may assume that $K' \neq \varnothing$, since otherwise K is finite and hence $\mathcal{C}(K)$ is finite-dimensional.

(a) Assume first that K' is finite, say $K' = \{a_1, \dots, a_N\}$, and let us show that $X := \mathcal{C}(K)$ has the Blum–Hanson property.

One may write $K = K_1 \cup \cdots \cup K_N$, where the K_i are pairwise disjoint compact sets and $K_i' = \{a_i\}$. Then $\mathcal{C}(K)$ is isometric to the ℓ_∞ direct sum $\mathcal{C}(K_1) \oplus \cdots \oplus \mathcal{C}(K_N)$, and each $\mathcal{C}(K_i)$ is isometric to the space c of all convergent sequences of real numbers. Therefore (by Lemma 2.1 and Remark 2.2), it is enough to show that the space c has property (P).

We view c as the space $\mathcal{C}(\mathbb{N} \cup \{\infty\})$, so that c_0 is identified with the subspace of all $f \in \mathcal{C}(\mathbb{N} \cup \{\infty\})$ such that $f(\infty) = 0$. We have to show that if (f_k) is a weakly null sequence in c, then

$$\lim_{k\to\infty} \lim_{t\to\infty} (r_c(t, f_k) - t) = 0.$$

Observe first that since $f_k(\infty) \to 0$ as $k \to \infty$, one can find a (weakly null) sequence $(\tilde{f}_k) \subset c$ such that $\tilde{f}_k \in c_0$ for all k and $\|\tilde{f}_k - f_k\|_\infty \to 0$: just set $\tilde{f}_k := f_k - f_k(\infty)\mathbf{1}$.

Let (g_n) be a weakly null sequence in c with $\|g_n\|_\infty \leq 1$. As above, choose a (weakly null) sequence $(\tilde{g}_n) \subset c$ such that $\|\tilde{g}_n - g_n\|_\infty \to 0$ and $\tilde{g}_n \in c_0$ for all n. Since $\|g_n\|_\infty \leq 1$, we may also assume that $\|\tilde{g}_n\|_\infty \leq 1$ for all n. Then, since f_k and the $\tilde{g}_n$ are living in c_0, we get from $(*)$ above that, for any $t \in \mathbb{R}^+$ and for each $k \in \mathbb{N}$,

$$\limsup_{n\to\infty} \|\tilde{f}_k + t\tilde{g}_n\|_\infty \leq r_{c_0}(t, \tilde{f}_k) = \max(\|\tilde{f}_k\|_\infty, t).$$

By the triangle inequality, it follows that

$$\limsup_{n\to\infty} \|f_k + tg_n\|_\infty \leq \|\tilde{f}_k - f_k\|_\infty + \max(\|\tilde{f}_k\|_\infty, t)$$

for each $k \in \mathbb{N}$ and all $t \geq 0$. This being true for any weakly null sequence (g_n) with $\|g_n\|_\infty \leq 1$, we conclude that

$$\lim_{t\to\infty} (r_c(f_k, t) - t) \leq \|\tilde{f}_k - f_k\|_\infty$$

for each $k \in \mathbb{N}$, and hence that

$$\lim_{k\to\infty} \lim_{t\to\infty} (r_c(t, f_k) - t) = 0.$$

(b) Now assume that K' is infinite. Since K is metrizable, it follows that K contains a compact set S of the form

$$S = \bigcup_{k=1}^{\infty} \big[\{s_{i,k} : i \in \mathbb{N}\} \cup \{s_{\infty,k}\} \big] \cup \{s_{\infty,\infty}\},$$

where all the points involved are distinct and

- $s_{i,k} \to s_{\infty,k}$ as $i \to \infty$ for each fixed $k \geq 1$;
- $s_{\infty,k} \to s_{\infty,\infty}$ as $k \to \infty$;
- the sets $S_k := \{s_{i,k} : i \in \mathbb{N}\} \cup \{s_{\infty,k}\}$ accumulate to $\{s_{\infty,\infty}\}$, i.e., they are eventually contained in any neighborhood of $s_{\infty,\infty}$.

Thus, we have $S' = \{s_{\infty,k} : k \geq 1\} \cup \{s_{\infty,\infty}\}$ and $S'' = \{s_{\infty,\infty}\}$.

The key point is now to construct a special continuous map $\theta : S \to S$ and to consider the associated *composition operator* C_θ acting on $\mathcal{C}(S)$. This is the same strategy as in [Akcoglu et al. 1974], in our setting.

Fact 2.3. *One can construct a continuous map $\theta : S \to S$ such that, denoting by θ^n the iterates of θ, the following properties hold true*:

(i) $\theta^n(s) \to s_{\infty,\infty}$ *pointwise on S as $n \to \infty$;*

(ii) *there exists an open neighborhood V of $s_{\infty,\infty}$ in S such that*

$$\sup_{s \in S} \#\{n \in \mathbb{N} : \theta^n(s) \notin V\} = \infty.$$

Proof. We define the map θ as follows:

$$\begin{aligned} \theta(s_{\infty,\infty}) &= s_{\infty,\infty}, & \\ \theta(s_{i,k}) &= s_{i,k-1} & \text{if } k \geq 2, \\ \theta(s_{\infty,k}) &= s_{\infty,k-1} & \text{if } k \geq 2, \\ \theta(s_{i,1}) &= s_{i-1,i-1} & \text{if } i \geq 2, \\ \theta(s_{\infty,1}) &= s_{\infty,\infty}, & \\ \theta(s_{1,1}) &= s_{\infty,\infty}. & \end{aligned}$$

It is clear that θ is continuous at each point $s_{\infty,k}$, $k \geq 2$. Moreover, since $s_{i-1,i-1} \to s_{\infty,\infty}$ as $i \to \infty$, the map θ is also continuous at $s_{\infty,1}$ and at $s_{\infty,\infty}$. Since all other points of S are isolated, it follows that θ is continuous on S.

An examination of the orbits of θ reveals that, for any $s \in S$, we have $\theta^n(s) = s_{\infty,\infty}$ for all but finitely many $n \in \mathbb{N}$. Indeed, if $s = s_{\infty,k}$ for some $k \in \mathbb{N}$, then

$$\mathrm{Orb}(s, \theta) = \{s_{\infty,k}, s_{\infty,k-1}, \ldots, s_{\infty,1}, s_{\infty,\infty}\},$$

whereas if $s = s_{i,k}$ for some $(i, k) \in \mathbb{N} \times \mathbb{N}$, then

$$\mathrm{Orb}(s,\theta) = \{s_{i,k}, s_{i,k-1}, \ldots, s_{i,1}, s_{i-1,i-1}, \ldots, s_{i-1,1}, s_{i-2,i-2}, \ldots, s_{1,2}, s_{1,1}, s_{\infty,\infty}\}.$$

So property (i) is satisfied.

Set $V := S \setminus S_1$, where $S_1 = \{s_{i,1} : i \in \mathbb{N}\} \cup \{s_{\infty,1}\}$. This is an open (actually clopen) neighborhood of $s_{\infty,\infty}$ in S. For any $N \in \mathbb{N}$, the orbit of $s_N := s_{N,1}$ contains exactly N points of $S \setminus V = S_1$, namely $s_{N,1}, s_{N-1,1}, \dots, s_{1,1}$. So property (ii) is satisfied as well. □

Fact 2.4. *The space $\mathcal{C}(S)$ does not have the Blum–Hanson property.*

Proof. Let $\theta : S \to S$ be given by Fact 2.3, and let $C_\theta : \mathcal{C}(S) \to \mathcal{C}(S)$ be the composition operator associated with θ,

$$C_\theta u = u \circ \theta \quad \text{for all } u \in \mathcal{C}(S).$$

By property (i) above, we see that $C_\theta^n u \to u(s_{\infty,\infty})\mathbf{1}$ weakly as $n \to \infty$, for every $u \in \mathcal{C}(S)$.

Let us choose a function $f \in \mathcal{C}(S)$ such that $f(s_{\infty,\infty}) = 0$ and $f \equiv 1$ on $F := S \setminus V$, where V satisfies (ii). Then $C_\theta^n f \to 0$ weakly. On the other hand, since $f \equiv 1$ on F, it follows from (ii) that one can find points $s \in S$ such that $\#\{n \in \mathbb{N} : C_\theta^n f(s) = 1\}$ is arbitrarily large. So we have

$$\frac{1}{\#I} \left\| \sum_{n \in I} C_\theta^n f \right\|_\infty \geq 1$$

for finite sets $I \subset \mathbb{N}$ with arbitrarily large cardinality. From this, it is a simple matter to deduce that the sequence $(C_\theta^n f)$ is not strongly mixing, which concludes the proof of Fact 2.4. □

It is now easy to conclude the proof of Theorem 1.1, by using the following trivial observation.

Fact 2.5. *Let X be a Banach space, and let Z be a closed subspace of X. Assume that Z is* 1-complemented *in X, i.e., there is a linear projection $\pi : X \to Z$ such that $\|\pi\| = 1$. If Z fails the Blum–Hanson property, then so does X.*

Proof. If $T : Z \to Z$ and $z \in Z$ witness that Z fails the Blum–Hanson property, then $\widetilde{T} := T \circ \pi : X \to Z \subset X$ and z witness that X also does. □

It is well known that since K is metrizable, there is an isometric linear extension operator $J : \mathcal{C}(S) \to \mathcal{C}(K)$. This is a classical result due to Dugundji [1951]. So the space $\mathcal{C}(S)$ is isometric to a 1-complemented subspace of $\mathcal{C}(K)$, namely $Z := J[\mathcal{C}(S)]$. By Fact 2.5, this concludes the proof of Theorem 1.1. □

Remark 2.6. The above proof shows that the space $\mathcal{C}(S)$ fails the Blum–Hanson property in a very special way. Namely, there exists a composition operator C_θ on $\mathcal{C}(S)$ all of whose orbits are weakly convergent and such that some weakly null orbit is not strongly mixing. As shown in [Akcoglu et al. 1974], the same is true for the space $\mathcal{C}(\mathbb{T}^2)$. On the other hand, it is observed in [Lefèvre et al. 2015] that this is *not* so in the space $\mathcal{C}([0,1])$, for the following reason: if $\theta : [0,1] \to [0,1]$ is a continuous map and if the iterates θ^n converge pointwise to some continuous map $\alpha : [0,1] \to [0,1]$, then the convergence is in fact uniform.

Remark 2.7. Our proof gives the following more precise result: if K has finitely many accumulation points, then $\mathcal{C}(K)$ has property (P); and, otherwise, one can find an operator T on $\mathcal{C}(K)$ with $\|T\| \leq 1$ such that *all T-orbits are weakly convergent* and some weakly null orbit is not strongly mixing.

3. A nonmetrizable example

We have been unable to show without the metrizability assumption on K that $\mathcal{C}(K)$ fails the Blum–Hanson property if K has infinitely many accumulation points. Note that metrizability was used twice in the proof of Theorem 1.1: to ensure that if K' is infinite then K contains the special compact set S; and for the existence of an isometric (linear) extension operator $J : \mathcal{C}(S) \to \mathcal{C}(K)$.

It is well known that the linear extension theorem may fail in the nonmetrizable case (see, e.g., [Pełczyński 1964, Remark 2.3]). The simplest way to see this is to observe that if there exists a linear extension operator $J : \mathcal{C}(S) \to \mathcal{C}(K)$ then, denoting by $R : \mathcal{C}(K) \to \mathcal{C}(S)$ the canonical restriction map, the operator $\pi := JR$ is a continuous projection on $\mathcal{C}(K)$ with kernel $I(S) := \{f \in \mathcal{C}(K) : f_{|S} = 0\}$, so $I(S)$ is a complemented subspace of $\mathcal{C}(K)$. But this may fail for some pairs (K, S); for example, one may take $(K, S) = (\beta\mathbb{N}, \beta\mathbb{N} \setminus \mathbb{N})$, where $\beta\mathbb{N}$ is the Stone–Čech compactification of $\mathbb{N}$, since $\mathcal{C}(K) = \ell_\infty(\mathbb{N})$ and $I(\beta\mathbb{N} \setminus \mathbb{N}) = c_0(\mathbb{N})$.

It may also happen that a compact set K has infinitely many accumulation points and yet does not contain any compact set like S. For example, this holds for $K = \beta\mathbb{N}$ because there are no nontrivial convergent sequences in $\beta\mathbb{N}$. However, in this (very) special case it is possible to adapt the proof of Theorem 1.1 to obtain the following result.

Proposition 3.1. *The space $\ell_\infty(\mathbb{N}) = \mathcal{C}(\beta\mathbb{N})$ does not have the Blum–Hanson property.*

Proof. It will be more convenient to view ℓ_∞ as $\ell_\infty(\mathbb{N} \times \mathbb{N}) = \mathcal{C}(\beta(\mathbb{N} \times \mathbb{N}))$.

Let $\theta : \mathbb{N} \times \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ be essentially the same map as in the proof of Theorem 1.1 but ignoring the limit points:

$$\begin{aligned} \theta(i,k) &= (i, k-1) && \text{if } k \geq 2, \\ \theta(i,1) &= (i-1, i-1) && \text{if } i \geq 2, \\ \theta(1,1) &= (1,1). \end{aligned}$$

We denote by C_θ the associated composition operator acting on $\ell_\infty = \ell_\infty(\mathbb{N} \times \mathbb{N})$:

$$C_\theta f(i,k) = f(\theta(i,k)) \quad \text{for every } (i,k) \in \mathbb{N} \times \mathbb{N}.$$

Set $f := \mathbf{1}_F \in \ell_\infty(\mathbb{N}\times\mathbb{N})$, where $F = \{(i,1) : i \geq 1\} \setminus \{(1,1)\} = \{(i,1) : i \geq 2\}$. Exactly as in the proof of Theorem 1.1, one checks that the sequence $(C_\theta^n f)$ is not strongly mixing in $\ell_\infty(\mathbb{N}\times\mathbb{N})$. So it is enough to show that, on the other hand, $C_\theta^n f \to 0$ weakly in $\ell_\infty(\mathbb{N}\times\mathbb{N})$.

Viewing $\ell_\infty(\mathbb{N}\times\mathbb{N})$ as $\mathcal{C}(\beta(\mathbb{N}\times\mathbb{N}))$, we have to show that $C_\theta^n f(\mathcal{U}) \to 0$ for every ultrafilter $\mathcal{U}$ on $\mathbb{N}\times\mathbb{N}$. Let us fix such an ultrafilter $\mathcal{U}$.

Since $C_\theta^n f = C_\theta^n \mathbf{1}_F = \mathbf{1}_{\theta^{-n}(F)}$ when considered as an element of $\ell_\infty(\mathbb{N}\times\mathbb{N})$, we have, for any $n \in \mathbb{N}$,

$$C_\theta^n f(\mathcal{U}) = \begin{cases} 1 & \text{if } \theta^{-n}(F) \in \mathcal{U}, \\ 0 & \text{if } \theta^{-n}(F) \notin \mathcal{U}. \end{cases}$$

So we need to prove that if n is large enough, then $\theta^{-n}(F) \notin \mathcal{U}$.

If we set $S_1 := \mathbb{N}\times\{1\}$, then $\theta^{-n}(S_1) \cap S_1$ is finite for every $n \in \mathbb{N}$. This is readily checked from the definition of θ. Indeed, for each $s = (i,1) \in S_1$, the first $n \in \mathbb{N}$ such that $\theta^n(s) \in S_1$ is at least equal (in fact, exactly equal) to i; so for each fixed n there are at most n points $s \in S_1$ such that $\theta^n(s) \in S_1$.

Since $F \subset S_1$ and θ is finite-to-one, it follows that $\theta^{-n}(F) \cap \theta^{-n'}(F)$ is finite whenever $n \neq n'$.

Now, assume without loss of generality that $\theta^{-n}(F) \in \mathcal{U}$ for more than one $n \in \mathbb{N}$. Then, by what we have just observed, $\mathcal{U}$ contains a finite set. Hence, $\mathcal{U}$ is a principal ultrafilter, defined by some point $s_0 \in \mathbb{N}\times\mathbb{N}$. On the other hand, we know from the definition of the map θ that $\theta^n(s_0) = (1,1)$ for all but finitely many $n \in \mathbb{N}$. Since $(1,1) \notin F$, it follows that $\theta^{-n}(F) \notin \mathcal{U}$ for all but finitely many n. □

Corollary 3.2. *The space $L_\infty = L_\infty(0,1)$ does not have the Blum–Hanson property. Likewise, if H is an infinite-dimensional Hilbert space, then the space $\mathcal{B}(H)$ of all bounded operators on H does not have the Blum–Hanson property.*

Proof. This is clear from Proposition 3.1, since these two spaces contain a 1-complemented isometric copy of ℓ_∞. □

4. Further remarks

For any topological space E, let us denote by $\mathcal{C}_b(E)$ the space of all real-valued, bounded continuous functions on E. Combining Theorem 1.1 and Proposition 3.1, we obtain the following result.

Theorem 4.1. *If T is a metrizable topological space, then $\mathcal{C}_b(T)$ has the Blum–Hanson property exactly when T is compact and has finitely many accumulation points.*

Proof. By Theorem 1.1, it is enough to show that if $\mathcal{C}_b(T)$ has the Blum–Hanson property, then T is compact. Now, if T is not compact, it contains a countably infinite closed discrete set S (thanks to the metrizability assumption). By Dugundji's extension theorem, $\mathcal{C}_b(T)$ then contains a 1-complemented isometric copy of $\mathcal{C}_b(S)$. Since $\mathcal{C}_b(S)$ is isometric to $\ell_\infty(\mathbb{N})$, it follows from Proposition 3.1 that $\mathcal{C}_b(T)$ does not have the Blum–Hanson property. □

To conclude this paper, and since this may be useful elsewhere, we isolate the following kind of criterion for detecting the failure of the Blum–Hanson property in $\mathcal{C}_b(T)$ for a not necessarily metrizable topological space T.

Lemma 4.2. *Let T be a Hausdorff topological space. Assume that there exists a subset S of T which is normal as a topological space, such that the following properties hold true.*

(1) *One can find a continuous map $\theta : S \to S$ and a point $a \in S$ such that*
   (i) *$\theta^n(s) \to a$ pointwise on S as $n \to \infty$;*
   (ii) *there exists an open neighborhood V of a such that*
$$\sup_{s\in S} \#\{n \in \mathbb{N} : \theta^n(s) \notin V\} = \infty;$$
   (iii) *there exists a further open neighborhood W of a with $\overline{W} \subset V$ such that, for any infinite set $\boldsymbol{N} \subset \mathbb{N}$, one can find $n_1, \dots, n_p \in \boldsymbol{N}$ such that the set $\theta^{-n_1}(S \setminus W) \cap \cdots \cap \theta^{-n_p}(S \setminus W)$ is finite.*

(2) *There is a linear isometric extension operator $J : \mathcal{C}_b(S) \to \mathcal{C}_b(T)$.*

Then, one can conclude that the space $\mathcal{C}_b(T)$ fails the Blum–Hanson property.

Proof. By (2), it is enough to show that $\mathcal{C}_b(S)$ does not have the Blum–Hanson property. This will of course be done by considering the composition operator $C_\theta : \mathcal{C}_b(S) \to \mathcal{C}_b(S)$.

Since $\overline{W} \subset V$ by (iii) and since S is normal, one can choose a function $f \in \mathcal{C}_b(S)$ such that $f \equiv 0$ on $\overline{W}$ and $f \equiv 1$ on $F := S \setminus V$. By condition (ii) in (1), the sequence $(C_\theta^n f)$ is not strongly mixing; so we just need to check that $C_\theta^n f \to 0$ weakly in $\mathcal{C}_b(S)$.

Being Hausdorff and normal, the space S is completely regular; so the space $\mathcal{C}_b(S)$ is canonically isometric with $\mathcal{C}(\beta S)$, where βS is the Stone–Čech compactification of S. The latter can be described as the space of all *z-ultrafilters* on S, i.e., maximal filters of zero sets for functions in $\mathcal{C}_b(S)$, or, equivalently (since S is normal), maximal filters of closed subsets of S; see [Gillman and Jerison 1960]. Therefore, what we have to do is to show that
$$\lim_{n\to\infty}\Big[\lim_{\mathcal{U}} f(\theta^n(s))\Big] = 0 \quad \text{for any } z\text{-ultrafilter } \mathcal{U} \text{ on } S.$$

If $\mathcal{U}$ is a "principal" z-ultrafilter defined by some $s_0 \in S$, i.e., $\mathcal{U}$ is convergent with limit s_0, then $\lim_{\mathcal{U}} f(\theta^n(s)) = f(\theta^n(s_0))$ for all n, so the result is clear since $f(\theta^n(s_0)) \to f(a) = 0$ as $n \to \infty$ by (i).

Now, let us assume that $\mathcal{U}$ is not principal. Then $\mathcal{U}$ does not contain any finite set. Indeed, if a maximal filter of closed sets contains a finite union of closed sets $F_1 \cup \cdots \cup F_N$, then it has to contain one of the F_i by maximality; so, if $\mathcal{U}$ were to contain a finite set, then it would contain a singleton and hence would be principal in a trivial way. By (iii), it follows that $\theta^{-n}(S \setminus W) \notin \mathcal{U}$ for all but finitely many $n \in \mathbb{N}$; and since $\mathcal{U}$ is a maximal filter of closed sets, this implies that $\theta^{-n}(\overline{W}) \in \mathcal{U}$ for all but finitely many n. Since $f \equiv 0$ on $\overline{W}$, it follows that $\lim_{\mathcal{U}} f(\theta^n(s)) = 0$ for all but finitely many n, which concludes the proof. $\square$

Remark 4.3. This lemma would be much neater if condition (iii) above could be dispensed with; but we don't know how to prove the lemma without it. The proof of Theorem 1.1 shows that when S is compact, (i) and (ii) alone are enough for $\mathcal{C}(S)$ to fail the Blum–Hanson property. At the other extreme, the proof of Proposition 3.1 shows that when S is discrete (and infinite), one can find a map $\theta : S \to S$ satisfying (i), (ii) and a property stronger than (iii).

Remark 4.4. When S is compact, condition (iii) actually follows from (i). Indeed, let W be any open neighborhood of a, and assume that (iii) fails for W and some infinite set $\boldsymbol{N} \subset \mathbb{N}$. Then, by compactness we have $\bigcap_{n \in \boldsymbol{N}} \theta^{-n}(S \setminus W) \neq \varnothing$. But if $s \in \bigcap_{n \in \boldsymbol{N}} \theta^{-n}(S \setminus W)$ then $\theta^n(s)$ does not tend to a as $n \to \infty$, which contradicts (i).

References

[Akcoglu et al. 1974] M. A. Akcoglu, J. P. Huneke, and H. Rost, "A counter example to the Blum Hanson theorem in general spaces", *Pacific J. Math.* **50** (1974), 305–308. MR 50 #2947 Zbl 0252.47006

[Berend and Bergelson 1986] D. Berend and V. Bergelson, "Mixing sequences in Hilbert spaces", *Proc. Amer. Math. Soc.* **98**:2 (1986), 239–246. MR 87j:47012 Zbl 0611.47021

[Blum and Hanson 1960] J. R. Blum and D. L. Hanson, "On the mean ergodic theorem for subsequences", *Bull. Amer. Math. Soc.* **66** (1960), 308–311. MR 22 #9572 Zbl 0096.09005

[Dugundji 1951] J. Dugundji, "An extension of Tietze's theorem", *Pacific J. Math.* **1** (1951), 353–367. MR 13,373c Zbl 0043.38105

[Gillman and Jerison 1960] L. Gillman and M. Jerison, *Rings of continuous functions*, Van Nostrand, Princeton, NJ, 1960. Reprinted in Graduate Texts in Mathematics **43**, Springer, New York, 1976. MR 22 #6994 Zbl 0093.30001

[Jones and Kuftinec 1971] L. K. Jones and V. Kuftinec, "A note on the Blum–Hanson theorem", *Proc. Amer. Math. Soc.* **30** (1971), 202–203. MR 43 #6742 Zbl 0218.28012

[Lefèvre et al. 2015] P. Lefèvre, É. Matheron, and A. Primot, "Smoothness, asymptotic smoothness and the Blum–Hanson property", *Israel J. Math.* (online publication November 2015).

[Müller and Tomilov 2007] V. Müller and Y. Tomilov, "Quasisimilarity of power bounded operators and Blum–Hanson property", *J. Funct. Anal.* **246**:2 (2007), 385–399. MR 2009a:47003 Zbl 1127.47011

[Pełczyński 1964] A. Pełczyński, "On simultaneous extension of continuous functions: a generalization of theorems of Rudin–Carleson and Bishop", *Studia Math.* **24**:3 (1964), 285–304. MR 30 #5184a Zbl 0145.16204

Received December 7, 2014.

PASCAL LEFÈVRE
LABORATOIRE DE MATHÉMATIQUES DE LENS
UNIVERSITÉ D'ARTOIS
RUE JEAN SOUVRAZ S.P. 18
62307 LENS
FRANCE

pascal.lefevre@univ-artois.fr

ÉTIENNE MATHERON
LABORATOIRE DE MATHÉMATIQUES DE LENS
UNIVERSITÉ D'ARTOIS
RUE JEAN SOUVRAZ S.P. 18
62307 LENS
FRANCE

etienne.matheron@univ-artois.fr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 282, No. 1, 2016
dx.doi.org/10.2140/pjm.2016.282.213

CROSSED PRODUCT ALGEBRAS AND DIRECT INTEGRAL DECOMPOSITION FOR LIE SUPERGROUPS

KARL-HERMANN NEEB AND HADI SALMASIAN

For every finite dimensional Lie supergroup $(G, \mathfrak{g})$, we define a C^*-algebra $\mathcal{A} := \mathcal{A}(G, \mathfrak{g})$ and show that there exists a canonical bijective correspondence between unitary representations of $(G, \mathfrak{g})$ and nondegenerate $*$-representations of $\mathcal{A}$. The proof of existence of such a correspondence relies on a subtle characterization of smoothing operators of unitary representations previously studied by Neeb, Salmasian, and Zellner.

For a broad class of Lie supergroups, which includes nilpotent as well as classical simple ones, we prove that the associated C^*-algebra is CCR. In particular, we obtain the uniqueness of direct integral decomposition for unitary representations of these Lie supergroups.

1. Introduction

Unitary representations of Lie supergroups play an important role in the mathematical theory of supersymmetric quantum mechanics. One distinguished example of the role of these unitary representations is the classification of free relativistic superparticles (see [Ferrara et al. 1981] and [Salam and Strathdee 1974]), where a super analogue of the little group method of Mackey and Wigner is used.

Although the super version of the Mackey–Wigner method was used in the physics literature as early as the 1970s, the problem of mathematical validity of this method in the context of supergroups was not addressed until less than a decade ago. This was done in [Carmeli et al. 2006], where the authors remedy this issue by laying the mathematically rigorous foundations of the analytic theory of unitary representations of Lie supergroups, using the equivalence of categories between the category of Lie supergroups and the category of *Harish-Chandra pairs* [Deligne and Morgan 1999, Section 3.8; Kostant 1977, Section 3.2]. The Harish-Chandra pair description of Lie supergroups will be explained in Definition 2.1 below.

MSC2010: 17B15, 22E45, 47L65.

Keywords: crossed product algebras, unitary representations, Lie supergroups, Harish-Chandra pairs, direct integral decomposition, CCR algebras.

The groundwork laid in [Carmeli et al. 2006] has spawned research on the harmonic analysis of Lie supergroups. In particular, in [Salmasian 2010] the irreducible unitary representations of a nilpotent Lie supergroup are classified using an extension of Kirillov's orbit method (see also [Neeb and Salmasian 2011, Section 8]). For Lie supergroups corresponding to basic classical Lie superalgebras [Musson 2012, Definition 1.14], the irreducible unitary representations are indeed highest weight modules [Neeb and Salmasian 2011, Section 7], and therefore they are completely classified by the work done in [Jakobsen 1994].

The goal of this paper is to systematically study disintegration of arbitrary unitary representations of Lie supergroups into direct integrals of irreducible representations. To this end, for every finite dimensional Lie supergroup $(G, \mathfrak{g})$ we construct a C^*-algebra $\mathcal{A} := \mathcal{A}(G, \mathfrak{g})$ whose nondegenerate $*$-representations are in bijective correspondence with unitary representations of $(G, \mathfrak{g})$. The C^*-algebra $\mathcal{A}$ is obtained as the completion of a crossed product $*$-algebra $\mathcal{A}^\circ$ that is associated to the action of G by left translation on the convolution algebra of test functions. Here, indeed, it will be more convenient to replace G by a slightly larger group $G_{\boldsymbol{\varepsilon}} \cong G \times \{1, \boldsymbol{\varepsilon}\}$, as the action of the extra element $\boldsymbol{\varepsilon}$ will automatically keep track of the $\mathbb{Z}_2$-grading of the representation space. Starting from a unitary representation $(\pi, \rho^\pi, \mathscr{H})$ of $(G, \mathfrak{g})$, we obtain a representation of $\mathcal{A}$ by first extending $(\pi, \rho^\pi, \mathscr{H})$ canonically to $\mathcal{A}^\circ$, and then uniquely to a nondegenerate $*$-representation $(\hat{\pi}, \mathscr{H})$ of $\mathcal{A}$ by continuity. Nevertheless, the construction of a representation of $(G, \mathfrak{g})$ from a representation $(\hat{\pi}, \mathscr{H})$ of $\mathcal{A}$ is more subtle, because the standard method of extending $(\hat{\pi}, \mathscr{H})$ to the multiplier algebra $M(\mathcal{A})$ does not work. Indeed, the Lie superalgebra $\mathfrak{g}$ does not act on $\mathcal{A}$ by multipliers. To circumvent this issue, we use the extension of $(\hat{\pi}, \mathscr{H})$ to the multiplier algebra $M(\mathcal{A}^\circ)$, and use the fact that G and $\mathfrak{g}$ act on $\mathcal{A}^\circ$ through $M(\mathcal{A}^\circ)$. To complete the construction of the unitary representation of $(G, \mathfrak{g})$, we need to show that the action of $\mathfrak{g}$ is indeed defined on $\mathscr{H}^\infty$. To this end, we prove that $\hat{\pi}(\mathcal{A}^\circ)\mathscr{H} = \mathscr{H}^\infty$, where $\mathscr{H}^\infty$ denotes the space of smooth vectors of the action of G on $\mathscr{H}$. The proof of the latter statement requires the Dixmier–Malliavin theorem [1978] and a subtle result from [Neeb et al. 2015, Theorem 2.11] on the characterization of smoothing operators of unitary representations, that is, operators $A : \mathscr{H} \to \mathscr{H}$ which map $\mathscr{H}$ into $\mathscr{H}^\infty$.

By the standard machinery of C^*-algebras [Dixmier 1974], statements on the existence and uniqueness of disintegration of nondegenerate $*$-representations of $\mathcal{A}$ can be transformed to similar statements on direct integral decompositions of unitary representations of $(G, \mathfrak{g})$. To obtain uniqueness of disintegration, it suffices to know that $\mathcal{A}$ is CCR, that is, the image of every irreducible $*$-representation of $\mathcal{A}$ lies in the algebra of compact operators. (Such C^*-algebras are sometimes called *liminal*.) We prove that $\mathcal{A}$ is CCR for a broad class of Lie supergroups, which includes nilpotent Lie supergroups as well as those which correspond to classical simple Lie superalgebras (see [Musson 2012, Section 1.3]). Therefore, for the aforementioned classes of Lie supergroups, one obtains uniqueness of disintegration of unitary representations.

This article is organized as follows. Section 2 is devoted to definitions and basic properties of unitary representations that will be used in the rest of the paper. In Section 3 we define the crossed product $*$-algebra $\mathcal{A}^\circ$. In Section 4 we construct the C^*-algebra $\mathcal{A} := \mathcal{A}(G, \mathfrak{g})$ as the completion of the crossed product algebra $\mathcal{A}^\circ$. In Section 5 we prove that under the G-action on $\mathcal{A}$, orbit maps of elements of $\mathcal{A}^\circ$ are smooth. In Section 6 we describe the canonical bijective correspondence between unitary representations of the Lie supergroup $(G, \mathfrak{g})$ and the (ungraded) nondegenerate $*$-representations of $\mathcal{A}$. Finally, in Section 7 we give our liminality results for C^*-algebras of a broad class of Lie supergroups, including the nilpotent and classical simple ones.

2. Basic definitions

We begin with a rapid review of Lie supergroups (from the Harish-Chandra pair viewpoint) and their unitary representations. For a more elaborate reference, see [Carmeli et al. 2006].

Throughout this paper, $\mathbb{Z}/2\mathbb{Z} := \{\bar{0}, \bar{1}\}$ denotes the field with two elements. If $V = V_{\bar{0}} \oplus V_{\bar{1}}$ is a $\mathbb{Z}/2\mathbb{Z}$-graded vector space, then the parity of a homogeneous element $x \in V$ is denoted by $|x| \in \mathbb{Z}/2\mathbb{Z}$.

Definition 2.1. A *Lie supergroup* is an ordered pair $(G, \mathfrak{g})$ with the following properties.

(i) G is a Lie group.

(ii) $\mathfrak{g} = \mathfrak{g}_{\bar{0}} \oplus \mathfrak{g}_{\bar{1}}$ is a Lie superalgebra over $\mathbb{R}$.

(iii) $\mathfrak{g}_{\bar{0}}$ is the Lie algebra of G.

(iv) There exists a group homomorphism $\mathrm{Ad} : G \to \mathrm{Aut}(\mathfrak{g})$, defining a smooth action $G \times \mathfrak{g} \to \mathfrak{g}$, such that

$$\mathrm{Ad}(g)x = \mathsf{d}c_g(\mathbf{1})(x) \quad \text{and} \quad \mathsf{d}\,\mathrm{Ad}_y(\mathbf{1})(x) = [x, y]$$

for every $x \in \mathfrak{g}_{\bar{0}}$, $y \in \mathfrak{g}$, and $g \in G$, where the map $c_g : G \to G$ is defined by $c_g(g') := gg'g^{-1}$ and $\mathrm{Ad}_y : G \to \mathfrak{g}$ is defined by $\mathrm{Ad}_y(g) := \mathrm{Ad}(g)y$.

We assume that $\dim \mathfrak{g} < \infty$ and that the component group G/G° is finite. The Lie supergroup $(G, \mathfrak{g})$ is called *connected* if G is a connected Lie group.

Remark 2.2. Here we should clarify that the condition given in Definition 2.1(iv) is identical to the ones given in our previous papers [Neeb and Salmasian 2013b, Definition 4.6.3(iv)] and [Neeb and Salmasian 2013a, Definition 7.1(iv)]. More precisely, in those two papers we tacitly assume that Ad is an extension of the adjoint action of G on $\mathfrak{g}_{\bar{0}}$.

Let $(\pi, \mathscr{H})$ be a unitary representation of a Lie group G. For $x \in \mathrm{Lie}(G)$ and $v \in \mathscr{H}$, we set

$$\mathrm{d}\pi(x)v := \lim_{t \to 0} \frac{1}{t}(\pi(e^{tx})v - v),$$

whenever the limit exists. Here $e^{tx} := \exp(tx)$ denotes the exponential map of G.

Definition 2.3. Let $(G, \mathfrak{g})$ be a Lie supergroup. A *unitary representation* of $(G, \mathfrak{g})$ is a triple $(\pi, \rho^\pi, \mathscr{H})$ which satisfies the following properties.

(i) $\mathscr{H}$ has a $\mathbb{Z}/2\mathbb{Z}$-grading, that is, $\mathscr{H} = \mathscr{H}_{\bar{0}} \oplus \mathscr{H}_{\bar{1}}$, and $(\pi, \mathscr{H})$ is a smooth unitary representation of G such that $\pi(g)$ preserves the $\mathbb{Z}/2\mathbb{Z}$-grading of $\mathscr{H}$ for every $g \in G$.

(ii) $\rho^\pi : \mathfrak{g} \to \mathrm{End}_{\mathbb{C}}(\mathscr{H}^\infty)$ is a representation of the Lie superalgebra $\mathfrak{g}$, where $\mathscr{H}^\infty = \mathscr{H}^\infty_{\bar{0}} \oplus \mathscr{H}^\infty_{\bar{1}}$ is the subspace consisting of all $v \in \mathscr{H}$ for which the orbit map $G \to \mathscr{H}$, $g \mapsto \pi(g)v$ is smooth.

(iii) $\rho^\pi(x) = \mathrm{d}\pi(x)|_{\mathscr{H}^\infty}$ for every $x \in \mathfrak{g}_{\bar{0}}$.

(iv) For every $x \in \mathfrak{g}_{\bar{1}}$, the operator $e^{-\pi i/4}\rho^\pi(x)$ is symmetric. That is,

$$-i\rho^\pi(x) \subseteq \rho^\pi(x)^*.$$

(v) $\pi(g)\rho^\pi(x)\pi(g)^{-1} = \rho^\pi(\mathrm{Ad}(g)x)$ for every $g \in G$ and every $x \in \mathfrak{g}_{\bar{1}}$.

Remark 2.4. By [Neeb and Salmasian 2013a, Proposition 6.13], the condition given in Definition 2.3(v) follows from the weaker condition that, for every element of the component group G/G°, there exists a coset representative $g \in G$ such that

$$\pi(g)\rho^\pi(x)\pi(g^{-1}) = \rho^\pi(\mathrm{Ad}(g)x) \quad \text{for every } x \in \mathfrak{g}.$$

Remark 2.5. As in [Neeb and Salmasian 2013b, Definition 6.7.1], a unitary representation $(\pi, \rho^\pi, \mathscr{H})$ is called *cyclic* if there exists a vector $v \in \mathscr{H}^\infty_{\bar{0}}$ such that $\pi(G)\rho^\pi(\boldsymbol{U}(\mathfrak{g}_{\mathbb{C}}))v$ spans a dense subspace of $\mathscr{H}$, where $\mathfrak{g}_{\mathbb{C}} := \mathfrak{g} \otimes_{\mathbb{R}} \mathbb{C}$ and where $\boldsymbol{U}(\mathfrak{g}_{\mathbb{C}})$ denotes the universal enveloping algebra of $\mathfrak{g}_{\mathbb{C}}$. A standard Zorn lemma argument shows that every unitary representation can be written as a direct sum of representations which are cyclic up to parity change. Furthermore, in [Neeb and Salmasian 2013b, Theorem 6.7.5] a Gelfand–Naimark–Segal construction is given which results in a correspondence between cyclic unitary representations and positive definite superfunctions of $(G, \mathfrak{g})$.

Let $(\pi, \rho^\pi, \mathscr{H})$ be a unitary representation of $(G, \mathfrak{g})$. We equip the space $\mathscr{H}^\infty$ with the Fréchet topology induced by the seminorms $v \mapsto \|\mathrm{d}\pi(D)v\|$, for all $D \in \boldsymbol{U}(\mathfrak{g}_{\bar{0}})$. This topology makes $\mathscr{H}^\infty$ a Fréchet space.

Proposition 2.6. *For every $x \in \mathfrak{g}$, the map $\rho^\pi(x) : \mathscr{H}^\infty \to \mathscr{H}^\infty$ is continuous.*

Proof. Continuity for $x \in \mathfrak{g}_{\bar{0}}$ is standard, and therefore we will assume that $x \in \mathfrak{g}_{\bar{1}}$. We need to prove that, for every $x_1, \dots, x_\ell \in \mathfrak{g}_{\bar{0}}$, for $D := x_1 \cdots x_\ell \in \boldsymbol{U}(\mathfrak{g}_{\bar{0}})$, the map

$$\mathscr{H}^\infty \to \mathbb{R}, \quad v \mapsto \|\mathrm{d}\pi(D)\rho^\pi(x)v\| \tag{1}$$

is continuous at $0 \in \mathscr{H}^\infty$. First assume that $\ell = 0$, so that $D = 1 \in \boldsymbol{U}(\mathfrak{g}_{\bar{0}})$. In this case, continuity of (1) follows from the inequality

$$\|\rho^\pi(x)v\|^2 = |\langle v, \rho^\pi(x)^2 v\rangle| \le \tfrac{1}{2}\|v\| \cdot \|\mathrm{d}\pi([x, x])v\|$$

and the definition of the topology of $\mathscr{H}^\infty$. To prove continuity of (1) for $\ell \ge 1$, we use the relation

$$x_1 \cdots x_\ell x = x x_1 \cdots x_\ell + \sum_{i=1}^{\ell} x_1 \cdots x_{i-1}[x_i, x]x_{i+1} \cdots x_\ell$$

and induction on ℓ. $\square$

Definition 2.7. A *multiplier* of an associative algebra $\mathcal{A}$ is a pair (λ, ρ) of linear maps $\mathcal{A} \to \mathcal{A}$ which satisfy the relations

$$\lambda(ab) = \lambda(a)b, \quad \rho(ab) = a\rho(b), \quad \text{and} \quad a\lambda(b) = \rho(a)b$$

for every $a, b \in \mathcal{A}$.

If $\mathcal{A}$ is a $*$-algebra, then the multipliers of $\mathcal{A}$ form a $*$-algebra, denoted by $M(\mathcal{A})$, with multiplication and involution defined by

$$(\lambda, \rho)(\lambda', \rho') := (\lambda\lambda', \rho'\rho) \quad \text{and} \quad (\lambda, \rho)^* := (\rho^*, \lambda^*), \tag{2}$$

where $\lambda^*(a) := \lambda(a^*)^*$ and $\rho^*(a) = \rho(a^*)^*$.

3. The crossed product $*$-algebra $\mathcal{A}^\circ$

Fix a Lie supergroup $(G, \mathfrak{g})$. Set $G_{\boldsymbol{\varepsilon}} := G \times \{\mathbf{1}, \boldsymbol{\varepsilon}\}$ such that $\boldsymbol{\varepsilon}^2 = \mathbf{1}$, and define $\mathrm{Ad}(\boldsymbol{\varepsilon})x := (-1)^{|x|}x$ for every homogeneous $x \in \mathfrak{g}$. We endow $G_{\boldsymbol{\varepsilon}}$ with the product topology. Clearly $(G_{\boldsymbol{\varepsilon}}, \mathfrak{g})$ is also a Lie supergroup. Let $\mathcal{D}(G_{\boldsymbol{\varepsilon}})$ be the convolution algebra of test functions (i.e., smooth compactly supported complex-valued functions) on $G_{\boldsymbol{\varepsilon}}$. The convolution on $\mathcal{D}(G_{\boldsymbol{\varepsilon}})$ is defined by

$$(f_1 \star f_2)(g') := \int_{G_{\boldsymbol{\varepsilon}}} f_1(g) f_2(g^{-1}g')\, dg,$$

where dg is the left-invariant Haar measure. The $*$-algebra structure is given by the involution

$$\check{f}(g) := \Delta(g)^{-1}\,\overline{f(g^{-1})},$$

where $g \mapsto \Delta(g)$ is the modular function satisfying $d(gg') = \Delta(g')\,dg$. From now on, we set

$$\mathrm{L}_g f(g') := f(g^{-1}g') \quad \text{and} \quad \mathsf{R}_x f(g) := \lim_{t\to 0} \frac{1}{t}(\mathrm{L}_{e^{tx}} f(g) - f(g)),$$

for $x \in \mathfrak{g}_{\bar{0}},\ g, g' \in G_{\boldsymbol{\varepsilon}},\ f \in \mathcal{D}(G_{\boldsymbol{\varepsilon}})$, and $t \in \mathbb{R}$.

Set $\mathfrak{g}_{\mathbb{C}} := \mathfrak{g} \otimes_{\mathbb{R}} \mathbb{C}$. For every $g \in G_{\boldsymbol{\varepsilon}}$, let $\alpha_g : \boldsymbol{U}(\mathfrak{g}_{\mathbb{C}}) \to \boldsymbol{U}(\mathfrak{g}_{\mathbb{C}})$ denote the automorphism that is canonically induced by $\mathrm{Ad}(g) : \mathfrak{g} \to \mathfrak{g}$. Our next goal is to define a *crossed product* $*$-algebra $\mathcal{A}^\circ = \mathcal{A}^\circ(G, \mathfrak{g})$. As a complex vector space,

$$\mathcal{A}^\circ := \boldsymbol{U}(\mathfrak{g}_{\mathbb{C}}) \otimes \mathcal{D}(G_{\boldsymbol{\varepsilon}}).$$

We identify $\mathcal{A}^\circ$ with a subspace of the vector space of $\boldsymbol{U}(\mathfrak{g}_{\mathbb{C}})$-valued functions on $G_{\boldsymbol{\varepsilon}}$ in the canonical way. Using this identification, we define a multiplication and a complex conjugation on $\mathcal{A}^\circ$ by the relations

$$(D_1 \otimes f_1)(D_2 \otimes f_2)(g') := \int_{G_{\boldsymbol{\varepsilon}}} f_1(g) f_2(g^{-1}g') D_1 \alpha_g(D_2)\,dg \tag{3}$$

and

$$(D \otimes f)^*(g) := \Delta(g^{-1})\overline{f(g^{-1})}\alpha_g(D^\dagger), \tag{4}$$

where the map $x \mapsto x^\dagger$ is the antilinear antiautomorphism of $\boldsymbol{U}(\mathfrak{g}_{\mathbb{C}})$ uniquely defined by

$$x^\dagger := \begin{cases} -x & \text{if } x \in \mathfrak{g}_{\bar{0}}, \\ -ix & \text{if } x \in \mathfrak{g}_{\bar{1}}. \end{cases} \tag{5}$$

In particular,

$$(D_1 \otimes f_1)(1 \otimes f_2) = D_1 \otimes (f_1 \star f_2).$$

Every $g \in G_{\boldsymbol{\varepsilon}}$ yields a multiplier (λ_g, ρ_g) of $\mathcal{A}^\circ$ by setting

$$\lambda_g(D \otimes f) := \alpha_g(D) \otimes \mathrm{L}_g f \quad \text{and} \quad \rho_g(D \otimes f) := D \otimes \Delta(g^{-1})\mathrm{R}_{g^{-1}} f, \tag{6}$$

where $\mathrm{R}_g f(g') := f(g'g)$. The algebra $\mathcal{A}^\circ$ is not necessarily unital. Nevertheless, we have the following lemma.

Lemma 3.1. *Every $a \in \mathcal{A}^\circ$ can be written as a finite sum $a = a_1b_1 + \cdots + a_mb_m$, where $a_k, b_k \in \mathcal{A}^\circ$ for $1 \le k \le m$. In other words, $\mathcal{A}^\circ = \mathcal{A}^\circ\mathcal{A}^\circ$.*

Proof. This follows from the more general result of [Alldridge 2014, Proposition 2.15], but we give a direct and simple argument. It is enough to prove the statement for $a = D \otimes f \in \mathcal{A}^\circ$. By the Dixmier–Malliavin theorem [1978], we can write $f = f_1 \star h_1 + \cdots + f_m \star h_m$, where $f_1, \ldots, f_m, h_1, \ldots, h_m \in \mathcal{D}(G_{\boldsymbol{\varepsilon}})$. It follows that $D \otimes f = (D \otimes f_1)(1 \otimes h_1) + \cdots + (D \otimes f_m)(1 \otimes h_m)$. $\square$

Every unitary representation $(\pi, \rho^\pi, \mathscr{H})$ of $(G, \mathfrak{g})$ extends to a unitary representation of $(G_{\boldsymbol{\varepsilon}}, \mathfrak{g})$ by setting $\pi(\boldsymbol{\varepsilon})v = (-1)^{|v|}v$ for every homogeneous $v \in \mathscr{H}$. From now on, we assume that every unitary representation of $(G, \mathfrak{g})$ has been extended to $(G_{\boldsymbol{\varepsilon}}, \mathfrak{g})$ as indicated above.

Fix a unitary representation $(\pi, \rho^\pi, \mathscr{H})$ of $(G, \mathfrak{g})$. Let $D \otimes f \in \mathcal{A}^\circ$, and as usual set

$$\pi(f) := \int_{G_{\boldsymbol{\varepsilon}}} f(g)\pi(g)\,dg.$$

Note that $\|\pi(f)\| \leq \|f\|_{L^1}$. By Gårding's theorem we know that $\pi(f)\mathscr{H} \subseteq \mathscr{H}^\infty$, so that the linear map

$$\rho^\pi(D)\pi(f) : \mathscr{H} \to \mathscr{H}$$

is well-defined.

Proposition 3.2. *Let $D \otimes f \in \mathcal{A}^\circ$. There exists a constant $M_{D\otimes f} > 0$ such that*

$$\|\rho^\pi(D)\pi(f)\| \leq M_{D\otimes f}$$

for every unitary representation $(\pi, \rho^\pi, \mathscr{H})$ of $(G, \mathfrak{g})$.

Proof. Note that $\pi(g)\pi(f) = \pi(\mathsf{L}_g f)$ for $g \in G_{\boldsymbol{\varepsilon}}$ and $f \in \mathcal{D}(G_{\boldsymbol{\varepsilon}})$ and that for $x \in \mathfrak{g}_{\bar{0}}$ we have $\lim_{t\to 0}\left\|\frac{1}{t}(\mathsf{L}_{e^{tx}} f - f) - \mathsf{R}_x f\right\|_{L^1} = 0$. Thus, for every $v \in \mathscr{H}$, we obtain that

$$\begin{aligned} \mathrm{d}\pi(x)\pi(f)v &= \lim_{t\to 0}\frac{1}{t}(\pi(e^{tx})\pi(f)v - \pi(f)v) \\ &= \lim_{t\to 0}\frac{1}{t}(\pi(\mathsf{L}_{e^{tx}} f)v - \pi(f)v) = \pi(\mathsf{R}_x f)v. \end{aligned} \tag{7}$$

By induction, from (7) it follows that

$$\mathrm{d}\pi(D)\pi(f)v = \pi(\mathsf{R}_D f)v \quad \text{for } D \in \boldsymbol{U}(\mathfrak{g}_{\bar{0}}),\ f \in \mathcal{D}(G_{\boldsymbol{\varepsilon}}),\ \text{and } v \in \mathscr{H}. \tag{8}$$

If $x \in \mathfrak{g}_{\bar{1}}$, then from (7) it also follows that

$$\begin{aligned} \|\rho^\pi(x)\pi(f)v\|^2 &= \langle\rho^\pi(x)\pi(f)v, \rho^\pi(x)\pi(f)v\rangle \\ &= \tfrac{1}{2}|\langle\rho^\pi([x,x])\pi(f)v, \pi(f)v\rangle| \\ &\leq \tfrac{1}{2}\|\rho^\pi([x,x])\pi(f)v\| \cdot \|\pi(f)v\| \\ &\leq \tfrac{1}{2}\|\mathsf{R}_{[x,x]}f\|_{L^1} \cdot \|f\|_{L^1} \cdot \|v\|^2. \end{aligned} \tag{9}$$

Similarly, if $x_1, \ldots, x_d \in \mathfrak{g}_{\bar{1}}$ for some $d > 1$, then

$$\begin{aligned} &\|\rho^\pi(x_1)\cdots\rho^\pi(x_d)\pi(f)v\|^2 \\ &\quad= \langle\rho^\pi(x_1)\cdots\rho^\pi(x_d)\pi(f)v, \rho^\pi(x_1)\cdots\rho^\pi(x_d)\pi(f)v\rangle \\ &\quad\leq \tfrac{1}{2}\|\rho^\pi(x_2)\cdots\rho^\pi(x_d)\pi(f)v\| \cdot \|\rho^\pi([x_1,x_1])\rho^\pi(x_2)\cdots\rho^\pi(x_d)\pi(f)v\|. \end{aligned} \tag{10}$$

Furthermore,

$$\begin{aligned}(11)\quad \rho^\pi([x_1,x_1])\rho^\pi(x_2)\cdots\rho^\pi(x_d)\pi(f)v \\ = \sum_{j=2}^{d} \rho^\pi(x_2)\cdots\rho^\pi([[x_1,x_1],x_j])\cdots\rho^\pi(x_d)\pi(f)v \\ + \rho^\pi(x_2)\cdots\rho^\pi(x_d)\pi(\mathsf{R}_{[x_1,x_1]}f)v.\end{aligned}$$

By the Poincaré-Birkhoff-Witt theorem, it is enough to prove the statement of the proposition when

$$D = y_1\cdots y_{\ell'}x_1\cdots x_\ell,$$

where $x_1,\ldots,x_\ell \in \mathfrak{g}_{\bar{0}}$ and $y_1,\ldots,y_{\ell'} \in \mathfrak{g}_{\bar{1}}$. From (8), (9), (10), and (11), and by induction on ℓ', it follows that $\|\rho^\pi(D)\pi(f)\|$ is bounded above by a constant which is expressible in terms of the L^1-norms of derivatives of f. □

Remark 3.3. The crossed product algebra $\mathcal{A}^\circ$ is also considered in [Alldridge 2014] and [Alldridge et al. 2013]. Here we have introduced two new gadgets: the involution $x \mapsto x^\dagger$, and the extension by $\boldsymbol{\varepsilon}$ which we will use below to keep track of the $\mathbb{Z}/2\mathbb{Z}$-grading of representations. The C^*-algebra $\mathcal{A} := \mathcal{A}(G,\mathfrak{g})$, which will be considered in Section 4, is closely related but not identical to the Fréchet algebra $|\Omega|_c(G)$ defined in [Alldridge 2014]. The reader should note that Lemma 4.1 and the results of Section 6 are analogous to the results of Section 2 and, in particular, Proposition 2.15 of that reference. However, our results cannot be obtained as direct consequences of Alldridge's, because a few technical issues arise that one needs to circumvent. In order to address these technical issues, and for the reader's convenience, we provide detailed proofs.

4. The C^*-algebra $\mathcal{A} := \mathcal{A}(G,\mathfrak{g})$

For a given unitary representation $(\pi,\rho^\pi,\mathscr{H})$ of $(G,\mathfrak{g})$, we define the linear map $\hat{\pi} : \mathcal{A}^\circ \to B(\mathscr{H})$ by setting

$$(12)\qquad \hat{\pi}(D\otimes f) := \rho^\pi(D)\pi(f) \quad \text{for every } D\otimes f \in \mathcal{A}^\circ,$$

and then extending $\hat{\pi}$ to $\mathcal{A}^\circ$ by linearity. Consider the seminorm on $\mathcal{A}^\circ$ defined by

$$(13)\qquad \|a\| := \sup_{(\pi,\rho^\pi,\mathscr{H})} \|\hat{\pi}(a)\|,$$

where the supremum is taken over all unitary equivalence classes of cyclic unitary representations $(\pi,\rho^\pi,\mathscr{H})$ of $(G,\mathfrak{g})$. From Proposition 3.2 it follows that $\|a\| < \infty$.

Lemma 4.1. *$\hat{\pi}$ is a $*$-representation of $\mathcal{A}^\circ$.*

Proof. This is the analogue of [Alldridge 2014, Lemma 2.14]. First we prove that $\hat{\pi}(ab) = \hat{\pi}(a)\hat{\pi}(b)$ for every $a,b \in \mathcal{A}^\circ$. It is enough to assume that $a =$

$D_1 \otimes f_1$ and $b = D_2 \otimes f_2$. Choose $\eta_1, \dots, \eta_r \in \mathcal{C}^\infty(G_\varepsilon)$ and $E_1, \dots, E_r \in \boldsymbol{U}(\mathfrak{g}_\mathbb{C})$ such that $\alpha_g(D_2) = \sum_{i=1}^r \eta_i(g) E_i$. Fix $v \in \mathscr{H}$ and set $w := \pi(f_2)v$. Then $w \in \mathscr{H}^\infty$ and therefore the map $G_\varepsilon \to \mathscr{H}^\infty$, $g \mapsto \pi(g)w$ is smooth [Poulsen 1972, Proposition 2.1]. Using Proposition 2.6 we obtain

$$\begin{aligned}\hat{\pi}(a)\hat{\pi}(b)v &= \rho^\pi(D_1)\int_{G_\varepsilon} f_1(g)\pi(g)\rho^\pi(D_2)\pi(f_2)v\,dg\\ &= \rho^\pi(D_1)\sum_{i=1}^r \rho^\pi(E_i)\pi(\eta_i f_1)\pi(f_2)v\\ &= \hat{\pi}\biggl(\sum_{i=1}^r D_1E_i \otimes (\eta_i f_1 * f_2)\biggr)v\\ &= \hat{\pi}(ab)v.\end{aligned}$$

The equality $\hat{\pi}(a)^* = \hat{\pi}(a^*)$ can be verified by a similar calculation, using the relation

$$\langle \rho^\pi(D\otimes f)v, w\rangle = \langle \rho^\pi(D)\pi(f)v, w\rangle = \langle v, \pi(\check{f})\rho^\pi(D^\dagger)w\rangle,$$

where $\check{f}(g) = \Delta(g)^{-1}\overline{f(g^{-1})}$. □

We are now ready to define $\mathcal{A} := \mathcal{A}(G, \mathfrak{g})$. From Lemma 4.1 it follows that the map $a \mapsto a^*$ is an isometry of $\mathcal{A}^\circ$ and that $\|aa^*\| = \|a\|^2$. Set $\mathcal{A}^\circ_- := \{a \in \mathcal{A}^\circ : \|a\| = 0\}$ and let $\mathcal{A}$ denote the completion of the quotient $\mathcal{A}^\circ/\mathcal{A}^\circ_-$ with respect to its induced norm. It is straightforward to check that $\mathcal{A}$ is a C^*-algebra.

Lemma 4.2. *Let $f \in \mathcal{D}(G_\varepsilon)$ and $D \otimes h \in \mathcal{A}^\circ$. Then the map*

$$\gamma_{f,D,h} : G_\varepsilon \to \mathcal{A}, \quad g \mapsto f(g)\alpha_g(D) \otimes \mathrm{L}_g h$$

is continuous and

$$\int_{G_\varepsilon} \gamma_{f,D,h}(g)\,dg = (1 \otimes f)(D \otimes h). \tag{14}$$

Proof. Choose $E_1, \dots, E_r \in \boldsymbol{U}(\mathfrak{g}_\mathbb{C})$ and $\eta_1, \dots, \eta_r \in \mathcal{C}^\infty(G_\varepsilon)$ such that we have $\alpha_g(D) = \sum_{i=1}^r \eta_i(g)E_i$ for every $g \in G_\varepsilon$. Then $\gamma_{f,D,h}(g) = \sum_{i=1}^r E_i \otimes f(g)\eta_i(g)\mathrm{L}_g h$. Next we prove that, for every $1 \le i \le r$, the map $G_\varepsilon \to \mathcal{A}$, $g \mapsto E_i \otimes \mathrm{L}_g h$ is continuous. Since we can replace h by $\mathrm{L}_g h$, it suffices to prove continuity at $\mathbf{1} \in G_\varepsilon$. To this end, we need to show that

$$\lim_{g\to\mathbf{1}}\biggl(\sup_{(\pi,\rho^\pi,\mathscr{H})} \|\rho^\pi(E_i)\pi(\mathrm{L}_g h - h)\|\biggr) = 0.$$

By an argument similar to the proof of Proposition 3.2, the latter statement can be reduced to showing that $\lim_{g\to\mathbf{1}} \|\mathrm{R}_D(\mathrm{L}_g h - h)\|_{L^1} = 0$ for every $D \in \boldsymbol{U}(\mathfrak{g}_{\bar{0}})$. This is straightforward.

Next we prove (14). From (3) it follows that

$$(15)\qquad (1\otimes f)(D\otimes h)(g') = \sum_{i=1}^{r}\int_{G_{\varepsilon}} f(g)(\mathrm{L}_g h)(g')\eta_i(g)E_i\,dg$$
$$= \sum_{i=1}^{r}(f\eta_i \star h)(g')E_i.$$

The left-regular representation of G_{ε} on $L^1(G_{\varepsilon})$ is strongly continuous, and its integrated representation is given by convolution, that is, $\int_{G_{\varepsilon}} f(g)\mathrm{L}_g h\,dg = f \star h$ for every $f, h \in L^1(G_{\varepsilon})$. We can now finish the proof by an argument similar to the one for [Alldridge 2014, (2.13)]. More precisely, we have

$$(16)\qquad \int_{G_{\varepsilon}} \gamma_{f,D,h}(g)\,dg = \sum_{i=1}^{r} E_i \otimes \int_{G_{\varepsilon}} f(g)\eta_i(g)\mathrm{L}_g h\,dg = \sum_{i=1}^{r} E_i \otimes ((f\eta_i)\star h).$$

Equality (14) now follows from (15) and (16). □

5. Multipliers of $\mathcal{A}$ and $\mathcal{A}^{\circ}$

For every $g \in G_{\varepsilon}$, let (λ_g, ρ_g) be the multiplier of $\mathcal{A}^{\circ}$ that is defined in (6). It is straightforward to verify that λ_g and ρ_g are isometries of $\mathcal{A}^{\circ}$, and therefore the multiplier (λ_g, ρ_g) extends uniquely to a multiplier of $\mathcal{A}$. For every $g \in G_{\varepsilon}$, the map

$$(17)\qquad \eta_G(g) : \mathcal{A} \to \mathcal{A}, \quad a \mapsto \lambda_g(a)$$

is an isometry and we have $\eta_G(gg') = \eta_G(g)\eta_G(g')$.

Proposition 5.1. *For every $a \in \mathcal{A}^{\circ}$, the map*

$$G \to \mathcal{A}, \quad g \mapsto \eta_G(g)a$$

is smooth.

Proof. It suffices to prove that the orbit map of every $D \otimes f \in \mathcal{A}^{\circ}$ is smooth. Set

$$\hat{\pi}_u := \bigoplus_{(\pi,\rho^{\pi},\mathscr{H})} \hat{\pi} \quad\text{and}\quad (\pi_u, \mathscr{H}_u) := \bigoplus_{(\pi,\rho^{\pi},\mathscr{H})} (\pi, \mathscr{H}),$$

where the direct sums are over unitary equivalence classes of cyclic unitary representations of $(G, \mathfrak{g})$. Then $(\pi_u, \mathscr{H}_u)$ is a smooth unitary representation of G_{ε}, and the map $\hat{\pi}_u : \mathcal{A}^{\circ} \to B(\mathscr{H}_u)$ is an isometry. Furthermore, $\hat{\pi}_u(\eta_G(g)a) = \pi_u(g)\hat{\pi}_u(a)$

for $g \in G_\varepsilon$ and $a \in \mathcal{A}^\circ$. Consequently, to complete the proof it suffices to show that, for every $a \in \mathcal{A}^\circ$, the map

$$G \to B(\mathscr{H}_u), \quad g \mapsto \pi_u(g)\hat{\pi}_u(a)$$

is smooth. The latter statement is a consequence of [Neeb et al. 2015, Theorem 2.11]. Indeed, by the same theorem we need to verify that $\hat{\pi}_u(a)\mathscr{H} \subseteq \mathscr{H}^\infty$. Without loss of generality we can assume $a = D \otimes f$, and therefore $\hat{\pi}_u(a) = \rho^{\pi_u}(D)\pi_u(f)$. From Gårding's theorem and Definition 2.3(ii), we obtain $\hat{\pi}_u(a)\mathscr{H} \subseteq \mathscr{H}^\infty$. □

By [Fell and Doran 1988b, Propositions VIII.1.11 and VIII.1.18], every multiplier of $\mathcal{A}$ is bounded and the multipliers of $\mathcal{A}$ form a unital Banach $*$-algebra $M(\mathcal{A})$ with multiplication and complex conjugation defined in (2) and the norm defined by $\|(\lambda, \rho)\| := \max\{\|\lambda\|, \|\rho\|\}$. Furthermore, the multipliers (λ_g, ρ_g) for $g \in G_\varepsilon$ are unitary, that is,

$$(\lambda_g, \rho_g)(\lambda_g, \rho_g)^* = 1 \in M(\mathcal{A}). \tag{18}$$

6. Nondegenerate ∗-representations of $\mathcal{A}$

In this section we prove that the category of unitary representations of $(G, \mathfrak{g})$ is isomorphic to the category of nondegenerate (in the sense of [Fell and Doran 1988a, Definition V.1.7]) $*$-representations of the C^*-algebra $\mathcal{A} = \mathcal{A}(G, \mathfrak{g})$.

Proposition 6.1. *Let $(\pi, \rho^\pi, \mathscr{H})$ be a unitary representation of a Lie supergroup $(G, \mathfrak{g})$. Then the $*$-representation $\hat{\pi}$ defined in Lemma 4.1 extends in a unique way to a nondegenerate $*$-representation*

$$\hat{\pi} : \mathcal{A} \to B(\mathscr{H}).$$

Proof. From (13) and Remark 2.5 it follows that $\|\hat{\pi}(a)\| \le \|a\|$ for every $a \in \mathcal{A}^\circ$. The existence and uniqueness of the extension $\hat{\pi} : \mathcal{A} \to B(\mathscr{H})$ now follows immediately. Nondegeneracy of $\hat{\pi}$ follows from the equality $\hat{\pi}(1 \otimes f) = \pi(f)$ for $f \in \mathcal{D}(G_\varepsilon)$. □

We now give a construction of a unitary representation of $(G, \mathfrak{g})$ from a nondegenerate $*$-representation $\hat{\pi} : \mathcal{A} \to B(\mathscr{H})$ of $\mathcal{A}$. By [Fell and Doran 1988b, Propositions VIII.1.11 and VIII.1.12], there exists a unique extension of $\hat{\pi}$ to a $*$-representation $\hat{\pi} : M(\mathcal{A}) \to B(\mathscr{H})$ of the multiplier algebra $M(\mathcal{A})$ satisfying

$$\hat{\pi}((\lambda, \rho))\hat{\pi}(a) = \hat{\pi}(\lambda(a)) \quad \text{for } (\lambda, \rho) \in M(\mathcal{A}) \text{ and } a \in \mathcal{A}. \tag{19}$$

Set

$$\pi(g) := \hat{\pi}((\lambda_g, \rho_g)) \quad \text{for every } g \in G_\varepsilon. \tag{20}$$

From (18) it follows that

$$\pi(g^{-1}) = \hat{\pi}((\lambda_{g^{-1}}, \rho_{g^{-1}})) = \hat{\pi}((\lambda_g, \rho_g)^*) = \hat{\pi}((\lambda_g, \rho_g))^* = \pi(g)^*,$$

that is, the operators $\pi(g)$ are unitary. Furthermore, from (19) it follows that the subspace $\mathscr{H}^\circ := \hat{\pi}(\mathcal{A}^\circ)\mathscr{H}$ is invariant under $\pi(g)$ for every $g \in G_{\boldsymbol{\varepsilon}}$.

Lemma 6.2. *For every $v \in \mathscr{H}$, the map $G_{\boldsymbol{\varepsilon}} \to \mathscr{H}$, $g \mapsto \pi(g)v$ is smooth if and only if $v \in \mathscr{H}^\circ$. In particular, $(\pi, \mathscr{H})$ is a smooth unitary representation of G.*

Proof. The statement and proof are analogous to [Alldridge 2014, Proposition 2.15], but there are subtle technical differences, and therefore we provide a detailed argument. First we show that, for every $v \in \mathscr{H}^\circ$, the orbit map $G_{\boldsymbol{\varepsilon}} \to \mathscr{H}$, $g \mapsto \pi(g)v$ is smooth. Assume that $v = \hat{\pi}(a)w$ for $a \in \mathcal{A}^\circ$ and $w \in \mathscr{H}$. Then

$$\pi(g)v = \hat{\pi}(\eta_G(g)a)w,$$

where $\eta_G(g) : \mathcal{A} \to \mathcal{A}$ is defined in (17). Since the map $\mathcal{A} \to \mathscr{H}$, $a \mapsto \hat{\pi}(a)w$ is continuous and linear, Proposition 5.1 implies that the orbit map $g \mapsto \pi(g)v$ is smooth.

Next we observe that $\mathscr{H}^\circ$ is a dense subspace of $\mathscr{H}$, because $\mathcal{A}^\circ$ is a dense subspace of $\mathcal{A}$. Therefore, the representation $(\pi, \mathscr{H})$ is smooth.

Finally, we prove that every smooth vector of $(\pi, \mathscr{H})$ belongs to $\mathscr{H}^\circ$. By the Dixmier–Malliavin theorem, it is enough to show that

$$\pi(f) = \hat{\pi}(1 \otimes f) \quad \text{for every } f \in \mathcal{D}(G_{\boldsymbol{\varepsilon}}), \tag{21}$$

where $\pi(f)v := \int_{G_{\boldsymbol{\varepsilon}}} f(g)\pi(g)v\,dg$ for $v \in \mathscr{H}$. Since both sides of (21) are bounded operators and $\mathscr{H}^\circ$ is dense in $\mathscr{H}$, it is enough to prove that

$$\pi(f)\hat{\pi}(D \otimes h)v = \hat{\pi}((1 \otimes f)(D \otimes h))v \quad \text{for } D \otimes h \in \mathcal{A}^\circ \text{ and } v \in \mathscr{H}.$$

Let $\gamma_{f,D,h}$ be defined as in Lemma 4.2. From (20) and (19) it follows that, for every $v \in \mathscr{H}$,

$$\begin{aligned}\pi(f)\hat{\pi}(D \otimes h)v &= \int_{G_{\boldsymbol{\varepsilon}}} f(g)\pi(g)\hat{\pi}(D \otimes h)v\,dg \\ &= \int_{G_{\boldsymbol{\varepsilon}}} \hat{\pi}(\gamma_{f,D,h}(g))v\,dg = \hat{\pi}\left(\int_{G_{\boldsymbol{\varepsilon}}} \gamma_{f,D,h}(g)\,dg\right)v,\end{aligned}$$

and from (14) it follows that $\pi(f)\hat{\pi}(D \otimes h)v = \hat{\pi}((1 \otimes f)(D \otimes h))v$. □

Set

$$\pi^\circ(a) := \hat{\pi}(a)|_{\mathscr{H}^\circ} \quad \text{for every } a \in \mathcal{A}^\circ.$$

From Lemmas 3.1 and 6.2 it follows that $(\pi^\circ, \mathscr{H}^\circ)$ is a nondegenerate $*$-representation of $\mathcal{A}^\circ$ in the sense defined in [Fell and Doran 1988a, Definition IV.3.17]. Therefore, by [Fell and Doran 1988b, Proposition VIII.1.9] there exists a unique extension of π° to a $*$-representation $\pi^\circ : M(\mathcal{A}^\circ) \to \mathrm{End}_{\mathbb{C}}(\mathscr{H}^\circ)$ satisfying

$$\pi^\circ((\lambda, \rho))\pi^\circ(a) = \pi^\circ(\lambda(a)) \quad \text{for } (\lambda, \rho) \in M(\mathcal{A}^\circ) \text{ and } a \in \mathcal{A}^\circ.$$

From the latter equality, Lemma 3.1, and (19), it follows that $\pi^\circ((\lambda_g, \rho_g)) = \pi(g)|_{\mathscr{H}^\circ}$ for every $g \in G_{\boldsymbol{\varepsilon}}$.

For every $x \in \mathfrak{g}$, let $(\lambda_x, \rho_x) \in M(\mathcal{A}^\circ)$ be the multiplier defined by

$$\lambda_x(D \otimes f) := xD \otimes f \quad \text{and} \quad \rho_x(D \otimes f)(g) := f(g)D\alpha_g(x) \quad \text{for } g \in G_{\boldsymbol{\varepsilon}}.$$

It is straightforward to verify that $(\lambda_x, \rho_x)^* = (\lambda_{x^\dagger}, \rho_{x^\dagger})$ for every $x \in \mathfrak{g}$, where $x^\dagger$ is defined as in (5). For every $x \in \mathfrak{g}$, we define a linear map

$$\rho^\pi(x) : \mathscr{H}^\circ \to \mathscr{H}^\circ, \quad v \mapsto \pi^\circ((\lambda_x, \rho_x))v. \tag{22}$$

Since $\pi(\boldsymbol{\varepsilon})^2 = \mathbf{1}$, we obtain a $\mathbb{Z}/2\mathbb{Z}$-grading $\mathscr{H} = \mathscr{H}_{\bar{0}} \oplus \mathscr{H}_{\bar{1}}$ by the ± 1 eigenspaces of $\pi(\boldsymbol{\varepsilon})$, i.e.,

$$\mathscr{H}_{\bar{0}} := \{v \in \mathscr{H} : \pi(\boldsymbol{\varepsilon})v = v\} \quad \text{and} \quad \mathscr{H}_{\bar{1}} := \{v \in \mathscr{H} : \pi(\boldsymbol{\varepsilon})v = -v\}.$$

Since $\pi(\boldsymbol{\varepsilon})$ leaves $\mathscr{H}^\circ$ invariant, the $\mathbb{Z}/2\mathbb{Z}$-grading of $\mathscr{H}$ induces a $\mathbb{Z}/2\mathbb{Z}$-grading $\mathscr{H}^\circ = \mathscr{H}^\circ_{\bar{0}} \oplus \mathscr{H}^\circ_{\bar{1}}$ on $\mathscr{H}^\circ$. We now prove the following proposition.

Proposition 6.3. *$(\pi, \rho^\pi, \mathscr{H})$ is a unitary representation of $(G, \mathfrak{g})$.*

Proof. Every $g \in G$ commutes with $\boldsymbol{\varepsilon}$, and therefore $\pi(g)$ preserves the $\mathbb{Z}/2\mathbb{Z}$-grading of $\mathscr{H}$. If $x \in \mathfrak{g}_{\bar{1}}$, then $(\lambda_x, \rho_x)^* = (\rho_{-ix}, \lambda_{-ix})$ in $M(\mathcal{A}^\circ)$, and it follows that the operator $e^{-\pi i/4}\rho^\pi(x)$ is symmetric. For every $x \in \mathfrak{g}$ and $g \in G_{\boldsymbol{\varepsilon}}$, we have

$$(\lambda_g, \rho_g)(\lambda_x, \rho_x)(\lambda_{g^{-1}}, \rho_{g^{-1}}) = (\lambda_{\mathrm{Ad}(g)x}, \rho_{\mathrm{Ad}(g)x}),$$

and consequently

$$\pi(g)\rho^\pi(x)\pi(g)^{-1} = \rho^\pi(\mathrm{Ad}(g)x). \tag{23}$$

In particular, from (23) for $g = \boldsymbol{\varepsilon}$, it follows that $\rho^\pi(x) \in \mathrm{End}_{\mathbb{C}}(\mathscr{H}^\circ)_{\bar{0}}$ for $x \in \mathfrak{g}_{\bar{0}}$ and $\rho^\pi(x) \in \mathrm{End}_{\mathbb{C}}(\mathscr{H}^\circ)_{\bar{1}}$ for $x \in \mathfrak{g}_{\bar{1}}$. The relation $\rho^\pi([x, y]) = [\rho^\pi(x), \rho^\pi(y)]$ for $x, y \in \mathfrak{g}$ follows from the corresponding relation in the multiplier algebra $M(\mathcal{A}^\circ)$. Finally, we prove that $\rho^\pi(x) = \mathrm{d}\pi(x)|_{\mathscr{H}^\circ}$ for every $x \in \mathfrak{g}_{\bar{0}}$. Fix $a \in \mathcal{A}^\circ$ and $v \in \mathscr{H}^\circ$, and set

$$\phi_{a,t} := \frac{1}{t}(\pi(e^{tx})\pi^\circ(a)v - \pi^\circ(a)v) - \rho^\pi(x)\pi^\circ(a)v \in \mathscr{H}.$$

Then $\phi_{a,t} = \pi^\circ(a_t)v$, where $a_t := \frac{1}{t}(\lambda_{e^{tx}}(a) - a) - \lambda_x(a) \in \mathcal{A}^\circ$. To complete the proof, we need to show that $\lim_{t\to 0} \|\phi_{a,t}\| = 0$. But

$$\|\phi_{a,t}\| = \|\pi^\circ(a_t)v\| = \|\hat{\pi}(a_t)v\| \le \|a_t\| \cdot \|v\|,$$

and therefore it suffices to show that $\lim_{t\to 0} \|a_t\| = 0$. Without loss of generality we can assume that $a = D \otimes f$. From the definition of the norm of $\mathcal{A}$ we obtain

$$\|a_t\| = \sup_{(\sigma, \rho^\sigma, \mathscr{K})} \left\| \frac{1}{t}(\sigma(e^{tx})\rho^\sigma(D)\sigma(f) - \rho^\sigma(D)\sigma(f)) - \rho^\sigma(xD)\sigma(f) \right\|,$$

where the supremum is taken over all unitary equivalence classes of cyclic unitary representations $(\sigma, \rho^\sigma, \mathscr{K})$ of $(G, \mathfrak{g})$. Now fix a unitary representation $(\sigma, \rho^\sigma, \mathscr{K})$ and a vector $v \in \mathscr{H}_\sigma$ such that $\|v\| = 1$. By Taylor's theorem,

$$\begin{aligned}\sigma(e^{tx})\rho^\sigma(D)\sigma(f)v \\ &= \rho^\sigma(D)\sigma(f)v + t\rho^\sigma(xD)\sigma(f)v + \frac{1}{2}\int_0^t (t-s)\sigma(e^{sx})\rho^\sigma(x^2D)\sigma(f)v\,ds.\end{aligned}$$

Proposition 3.2 implies that there exists a constant $M > 0$, independent of $(\sigma, \rho^\sigma, \mathscr{K})$, such that $\|\rho^\sigma(x^2D)\sigma(f)\| \le M$. It follows that $\|a_t\| \le \frac{1}{2}M \cdot |t|$, and consequently $\lim_{t\to 0}\|a_t\| = 0$. □

Recall that the morphisms in the two categories of unitary representations of $(G, \mathfrak{g})$, and nondegenerate $*$-representations of $\mathcal{A} = \mathcal{A}(G, \mathfrak{g})$, are bounded linear intertwining operators.

Theorem 6.4. *The correspondences of Propositions 6.1 and 6.3 result in an isomorphism between the category of unitary representations of $(G, \mathfrak{g})$ and the category of nondegenerate $*$-representations of $\mathcal{A} = \mathcal{A}(G, \mathfrak{g})$.*

Proof. **Step 1.** First we verify that the correspondences of Propositions 6.1 and 6.3 are mutual inverses. Let $(\pi, \rho^\pi, \mathscr{H})$ be a unitary representation of $(G, \mathfrak{g})$. Let $\hat{\pi}$ be the $*$-representation of $\mathcal{A}$ constructed by Proposition 6.1, and let $(\bar{\pi}, \rho^{\bar{\pi}}, \mathscr{H})$ be the unitary representation of $(G, \mathfrak{g})$ constructed from $\hat{\pi}$ by Proposition 6.3. For $D \otimes f \in \mathcal{A}^\circ$ and $g \in G_{\boldsymbol{\varepsilon}}$,

$$\begin{aligned}\bar{\pi}(g)\hat{\pi}(D\otimes f) &= \hat{\pi}(\lambda_g(D\otimes f)) = \hat{\pi}(\alpha_g(D)\otimes \mathrm{L}_g f) \\ &= \rho^\pi(\alpha_g(D))\pi(\mathrm{L}_g f) = \pi(g)\rho^\pi(D)\pi(f) = \pi(g)\hat{\pi}(D\otimes f).\end{aligned}$$

Since $\pi(g)$ and $\bar{\pi}(g)$ are bounded operators and $\hat{\pi}$ is nondegenerate, we obtain $\pi(g) = \bar{\pi}(g)$ for $g \in G_{\boldsymbol{\varepsilon}}$. Let $\mathscr{H}^\infty$ denote the space of smooth vectors of $(\pi, \mathscr{H})$. For $x \in \mathfrak{g}$, $D \otimes f \in \mathcal{A}^\circ$, and $w \in \mathscr{H}$,

$$\begin{aligned}\rho^{\bar{\pi}}(x)\hat{\pi}(D\otimes f)w &= \hat{\pi}(xD\otimes f)w = \rho^\pi(xD)\pi(f)w \\ &= \rho^\pi(x)\rho^\pi(D)\pi(f)w = \rho^\pi(x)\hat{\pi}(D\otimes f)w.\end{aligned} \tag{24}$$

By the Dixmier–Malliavin theorem, $\mathscr{H}^\infty = \hat{\pi}(\mathcal{A}^\circ)\mathscr{H}$. Therefore, (24) implies that $\rho^{\bar{\pi}}(x) = \rho^\pi(x)$ for every $x \in \mathfrak{g}$.

Conversely, let $\hat{\pi}\colon \mathcal{A} \to B(\mathscr{H})$ be a nondegenerate $*$-representation. Let $(\pi, \rho^\pi, \mathscr{H})$ be the unitary representation of $(G, \mathfrak{g})$ corresponding to $\hat{\pi}$ by Proposition 6.3, and let $\hat{\pi}' : \mathcal{A} \to B(\mathscr{H})$ be the $*$-representation corresponding to $(\pi, \rho^\pi, \mathscr{H})$ by

Proposition 6.1. For $D_1 \otimes h_1 \in \mathcal{A}^\circ$ and $w \in \mathscr{H}$, we obtain by Lemma 4.2 that

$$\begin{aligned}
\pi(f)\hat{\pi}(D_1 \otimes h_1)w &= \int_{G_\varepsilon} f(g)\pi(g)\hat{\pi}(D_1 \otimes h_1)w\,dg \\
&= \int_{G_\varepsilon} \hat{\pi}(\gamma_{f,D_1,h_1}(g))w\,dg \\
&= \hat{\pi}\left(\int_{G_\varepsilon} \gamma_{f,D_1,h_1}(g)\,dg\right)w \\
&= \hat{\pi}((1 \otimes f)(D_1 \otimes h_1))w.
\end{aligned} \tag{25}$$

Now set $a := D_1 \otimes h_1$. From (25) it follows that, for every $D \otimes f \in \mathcal{A}^\circ$,

$$\begin{aligned}
\hat{\pi}'(D \otimes f)\hat{\pi}(a)w &= \rho^\pi(D)\pi(f)\hat{\pi}(a)w \\
&= \rho^\pi(D)\hat{\pi}((1 \otimes f)a)w = \hat{\pi}(D \otimes f)\hat{\pi}(a)w.
\end{aligned}$$

Nondegeneracy of $\hat{\pi}$ and boundedness of the operators $\hat{\pi}'(D \otimes f)$ and $\hat{\pi}(D \otimes f)$ imply that $\hat{\pi}'(D\otimes f) = \hat{\pi}(D\otimes f)$. Since $\mathcal{A}^\circ$ is dense in $\mathcal{A}$, the equality $\hat{\pi}'(a) = \hat{\pi}(a)$ holds for every $a \in \mathcal{A}$.

Step 2. To complete the proof, we need to show that the correspondences of Propositions 6.1 and 6.3 are compatible with morphisms in the two categories. Suppose that $(\pi, \rho^\pi, \mathscr{H})$ and $(\sigma, \rho^\sigma, \mathscr{K})$ are two unitary representations of $(G, \mathfrak{g})$, and let $\hat{\pi} : \mathcal{A} \to B(\mathscr{H})$ and $\hat{\sigma} : \mathcal{A} \to B(\mathscr{K})$ be the $*$-representations of $\mathcal{A}$ constructed from $(\pi, \rho^\pi, \mathscr{H})$ and $(\sigma, \rho^\sigma, \mathscr{K})$ by Proposition 6.1. If $T : \mathscr{H} \to \mathscr{K}$ is a $(G, \mathfrak{g})$-intertwining operator, then it is easy to verify that T commutes with the action of $\mathcal{A}^\circ$ on $\mathscr{H}$ and $\mathscr{K}$, and therefore, by a continuity argument, T commutes with the action of $\mathcal{A}$ on $\mathscr{H}$ and $\mathscr{K}$ as well.

Conversely, assume that $T : \mathscr{H} \to \mathscr{K}$ commutes with the actions of $\mathcal{A}$ on $\mathscr{H}$ and $\mathscr{K}$. First note that, for every $a \in \mathcal{A}$ and every $(\lambda, \rho) \in M(\mathcal{A})$,

$$T\hat{\pi}((\lambda, \rho))\hat{\pi}(a) = T\hat{\pi}(\lambda(a)) = \hat{\sigma}(\lambda(a))T = \hat{\sigma}((\lambda, \rho))\hat{\sigma}(a)T = \hat{\sigma}((\lambda, \rho))T\hat{\pi}(a).$$

Since $\hat{\pi}(\mathcal{A})\mathscr{H}$ is a dense subspace of $\mathscr{H}$, it follows that

$$T\hat{\pi}((\lambda, \rho)) = \hat{\sigma}((\lambda, \rho))T \quad \text{for every } (\lambda, \rho) \in M(\mathcal{A}).$$

Setting $(\lambda, \rho) := (\lambda_g, \rho_g)$ in the last relation, we obtain

$$T\pi(g) = \sigma(g)T \quad \text{for every } g \in G_\varepsilon, \tag{26}$$

and in particular $T\mathscr{H}^\infty \subseteq \mathscr{K}^\infty$. From (26) for $g = \boldsymbol{\varepsilon}$, it follows that T preserves the $\mathbb{Z}/2\mathbb{Z}$-grading of $\mathscr{H}$. Now for $(\lambda, \rho) \in M(\mathcal{A}^\circ)$, $a \in \mathcal{A}^\circ$, and $v \in \mathscr{H}^\infty$, using

Lemma 6.2 we obtain that

$$\begin{aligned} T\pi^\circ(\lambda,\rho)\pi^\circ(a)v &= T\pi^\circ(\lambda(a))v = \sigma^\circ(\lambda(a))Tv \\ &= \sigma^\circ(\lambda,\rho)\sigma^\circ(a)Tv = \sigma^\circ(\lambda,\rho)T\pi^\circ(a)v. \end{aligned}$$

Thus Lemmas 6.2 and 3.1 imply that $T\pi^\circ(\lambda,\rho)w = \sigma^\circ(\lambda,\rho)Tw$ for every $w \in \mathscr{H}^\infty$. Setting $(\lambda,\rho) := (\lambda_x, \rho_x)$ for $x \in \mathfrak{g}$, we obtain $T\rho^\pi(x) = \rho^\sigma(x)T$. Therefore, T is a $(G,\mathfrak{g})$-intertwining map from $(\pi,\rho^\pi,\mathscr{H})$ to $(\sigma,\rho^\sigma,\mathscr{K})$. □

7. Unique direct integral decompositions

For a unitary representation $(\pi,\rho^\pi,\mathscr{H})$ of a Lie supergroup $(G,\mathfrak{g})$, it is desirable to have a decomposition as a direct integral of irreducible unitary representations. From Theorem 6.4 it follows that the problem of existence and uniqueness of such a direct integral decomposition can be reduced to the same problem for the associated C^*-algebra $\mathcal{A} = \mathcal{A}(G,\mathfrak{g})$.

In this section we prove that existence and uniqueness of direct integral decompositions hold for two general classes of Lie supergroups, which include nilpotent and basic classical Lie supergroups.

Recall that a C^*-algebra $\mathcal{A}$ is called *CCR* if $\hat{\pi}(\mathcal{A}) \subseteq K(\mathscr{H})$ for every irreducible $*$-representation $\hat{\pi} : \mathcal{A} \to B(\mathscr{H})$, where $K(\mathscr{H}) \subseteq B(\mathscr{H})$ denotes the subspace of compact operators. It is well known that, for C^*-algebras which are CCR, existence and uniqueness of direct integral decompositions hold.

A unitary representation $(\pi,\mathscr{H})$ of a Lie group G is called *completely continuous* if $\pi(f) \in K(\mathscr{H})$ for every $f \in \mathcal{D}(G)$.

Theorem 7.1. *Let $(G,\mathfrak{g})$ be a Lie supergroup such that, for every irreducible unitary representation $(\pi,\rho^\pi,\mathscr{H})$ of $(G,\mathfrak{g})$, the unitary representation $(\pi,\mathscr{H})$ of G is completely continuous. Then the C^*-algebra $\mathcal{A} = \mathcal{A}(G,\mathfrak{g})$ is CCR.*

Proof. Let $\hat{\pi} : \mathcal{A} \to B(\mathscr{H})$ be an irreducible $*$-representation of $\mathcal{A}$. Since $K(\mathscr{H})$ is a closed ideal of $B(\mathscr{H})$ and $\|\hat{\pi}(a)\| \le \|a\|$ for every $a \in \mathcal{A}$, it suffices to prove that $\hat{\pi}(D \otimes f) \in K(\mathscr{H})$ for every $D \otimes f \in \mathcal{A}^\circ$. Let $(\pi,\rho^\pi,\mathscr{H})$ be the unitary representation of $(G,\mathfrak{g})$ that corresponds to $\hat{\pi}$. Theorem 6.4 implies that $(\pi,\rho^\pi,\mathscr{H})$ is irreducible. The Dixmier–Malliavin theorem implies that there exist $f_1,\cdots,f_r,h_1,\ldots,h_r \in \mathcal{D}(G_\varepsilon)$ such that $f = \sum_{i=1}^r f_i \star h_i$. Thus

$$\hat{\pi}(D\otimes f) = \sum_{i=1}^r \hat{\pi}(D\otimes f_i)\hat{\pi}(1\otimes h_i) = \sum_{i=1}^r \hat{\pi}(D\otimes f_i)\pi(h_i).$$

From the assumption of the theorem it follows that $\pi(h_i) \in K(\mathscr{H})$ for $1 \le i \le r$. Consequently, $\hat{\pi}(D\otimes f) \in K(\mathscr{H})$. □

As in [Salmasian 2010], a Lie supergroup $(G, \mathfrak{g})$ is called *nilpotent* if $\mathfrak{g}$ is a nilpotent Lie superalgebra.

Theorem 7.2. *Let $(G, \mathfrak{g})$ be a connected nilpotent Lie supergroup. Then the C^*-algebra $\mathcal{A} = \mathcal{A}(G, \mathfrak{g})$ is CCR.*

Proof. From [Salmasian 2010, Corollary 6.1.1] it follows that the restriction of every irreducible unitary representation $(\pi, \rho^\pi, \mathscr{H})$ of $(G, \mathfrak{g})$ to G is a direct sum of finitely many irreducible unitary representations. Since every nilpotent Lie group is CCR [Fell 1962], the unitary representation $(\pi, \mathscr{H})$ is completely continuous. Therefore, Theorem 7.1 applies. □

Recall from [Neeb and Salmasian 2011] that a Lie supergroup $(G, \mathfrak{g})$ is called *$\star$-reduced* if for every nonzero $x \in \mathfrak{g}$ there exists a unitary representation $(\pi, \rho^\pi, \mathscr{H})$ of $(G, \mathfrak{g})$ such that $\rho^\pi(x) \neq 0$.

Theorem 7.3. *Let $(G, \mathfrak{g})$ be a connected Lie supergroup which is $\star$-reduced and satisfies $\mathfrak{g}_{\bar{0}} = [\mathfrak{g}_{\bar{1}}, \mathfrak{g}_{\bar{1}}]$. Then the C^*-algebra $\mathcal{A} = \mathcal{A}(G, \mathfrak{g})$ is CCR.*

Proof. We show that the hypotheses of Theorem 7.1 are satisfied.

Step 1. From [Neeb and Salmasian 2011, Theorem 7.3.2] it follows that there exists a compactly embedded (in the sense of [Neeb 2000, Definition VII.1.1]) Cartan subalgebra $\mathfrak{t} \subseteq \mathfrak{g}_{\bar{0}}$ and a positive system $\Delta^+ = \{\alpha_1, \ldots, \alpha_r\}$ of $\mathfrak{t}$-roots of $\mathfrak{g}$, such that the space $\mathscr{H}^\mathfrak{t}$ of $\mathfrak{t}$-finite smooth vectors in $\mathscr{H}$ is a dense subspace of $\mathscr{H}$. Furthermore, $\mathscr{H}^\mathfrak{t}$ is an irreducible $\mathfrak{g}$-module which is a direct sum of $\mathfrak{t}$-weight spaces with weights of the form

$$\lambda - \sum_{i=1}^{r} n_i \alpha_i, \qquad \text{where } n_i \in \mathbb{N} \cup \{0\} \text{ for every } 1 \leq i \leq r. \tag{27}$$

Since $\boldsymbol{U}(\mathfrak{g})$ is a finitely generated $\boldsymbol{U}(\mathfrak{g}_{\bar{0}})$-module, the irreducible (hence cyclic) $\boldsymbol{U}(\mathfrak{g})$-module $\mathscr{H}^\mathfrak{t}$ is a finitely generated $\boldsymbol{U}(\mathfrak{g}_{\bar{0}})$-module. Since $\boldsymbol{U}(\mathfrak{g}_{\bar{0}})$ is a Noetherian ring [Dixmier 1974, Corollary 2.3.8], $\mathscr{H}^\mathfrak{t}$ is a Noetherian $\boldsymbol{U}(\mathfrak{g}_{\bar{0}})$-module.

Step 2. We prove that $(\pi, \mathscr{H})$ is a direct sum of finitely many irreducible unitary representations of G. Assume the contrary. Then we can write $\mathscr{H} = \bigoplus_{\ell=1}^\infty \mathscr{H}_\ell$ such that each $\mathscr{H}_\ell$ is a G-invariant closed subspace of $\mathscr{H}$. From the inclusion $\bigoplus_{i=1}^\infty \mathscr{H}_\ell^\mathfrak{t} \subseteq \mathscr{H}^\mathfrak{t}$ it follows that as a $\boldsymbol{U}(\mathfrak{g}_{\bar{0}})$-module, $\mathscr{H}^\mathfrak{t}$ is not Noetherian. This contradicts Step 1.

Step 3. From (27) and Step 2 it follows that $(\pi, \mathscr{H})$ is a direct sum of finitely many irreducible highest weight (in the sense of [Neeb 2000, Definition X.2.9]) unitary representations of G. From [Neeb 2000, Theorem X.4.10] it follows that every irreducible highest weight unitary representation of G is CCR. Thus $(\pi, \mathscr{H})$ is also a CCR unitary representation of G. □

Remark 7.4. Here it would be helpful to the reader to make a correction to [Neeb and Salmasian 2011, Theorem 7.3.2]. The proof of the theorem uses the bijective correspondence between G-invariant closed subspaces of $\mathscr{H}$ and $\mathfrak{g}_{\bar{0}}$-invariant subspaces of analytic vectors in $\mathscr{H}$. Such a bijection holds only when G is connected. Therefore, connectedness of G should be added in the statement of the theorem. It is plausible to expect that Theorem 7.3 holds when G has finitely many connected components.

Remark 7.5. Let $(G, \mathfrak{g})$ be a connected Lie supergroup such that $\mathfrak{g}$ is a real form of a classical simple Lie superalgebra (see [Musson 2012, Section 1.3]). That is, we assume that $\mathfrak{g} \otimes_{\mathbb{R}} \mathbb{C}$ is isomorphic to one of the Lie superalgebras of type $\mathfrak{sl}(m|n)$ for $m > n \geq 0$, $\mathfrak{psl}(m|m)$ for $m \geq 1$, $\mathfrak{osp}(m|2n)$ for $m, n \geq 0$, $D(2, 1; \alpha)$ for $\alpha \neq 0, -1$, $\mathfrak{p}(n)$ for $n \geq 1$, $\mathfrak{q}(n)$ for $n \geq 1$, $G(3)$, or $F(4)$. Assume that $(G, \mathfrak{g})$ has nontrivial unitary representations. (A complete list of these Lie supergroups can be obtained from [Neeb and Salmasian 2011, Theorem 6.2.1].) It is then straightforward to verify that $(G, \mathfrak{g})$ satisfies the hypotheses of Theorem 7.3, and therefore the C^*-algebra $\mathcal{A} = \mathcal{A}(G, \mathfrak{g})$ is CCR.

Acknowledgements

During the completion of this project, Salmasian was supported by an NSERC Discovery Grant and the Emerging Field Project "Quantum Geometry" of FAU Erlangen-Nürnberg. The authors thank the referee for useful comments and for pointing out the connection between the recent articles [Alldridge 2014] and [Alldridge et al. 2013] and our work.

References

[Alldridge 2014] A. Alldridge, "Fréchet globalisations of Harish–Chandra supermodules", preprint, 2014.

[Alldridge et al. 2013] A. Alldridge, J. Hilgert, and M. Laubinger, "Harmonic analysis on Heisenberg–Clifford Lie supergroups", *J. Lond. Math. Soc.* (2) **87**:2 (2013), 561–585. MR 3046286 Zbl 06160946

[Carmeli et al. 2006] C. Carmeli, G. Cassinelli, A. Toigo, and V. S. Varadarajan, "Unitary representations of super Lie groups and applications to the classification and multiplet structure of super particles", *Comm. Math. Phys.* **263**:1 (2006), 217–258. MR 2006m:22028 Zbl 1124.22007

[Deligne and Morgan 1999] P. Deligne and J. W. Morgan, "Notes on supersymmetry (following Joseph Bernstein)", pp. 41–97 in *Quantum fields and strings: A course for mathematicians, I* (Princeton, NJ, 1996/1997), edited by P. Deligne et al., Amer. Math. Soc., Providence, RI, 1999. MR 2001g:58007 Zbl 1170.58302

[Dixmier 1974] J. Dixmier, *Algèbres enveloppantes*, Cahiers Scientifiques **37**, Gauthier-Villars, Paris, 1974. MR 58 #16803a Zbl 0308.17007

[Dixmier and Malliavin 1978] J. Dixmier and P. Malliavin, "Factorisations de fonctions et de vecteurs indéfiniment différentiables", *Bull. Sci. Math.* (2) **102**:4 (1978), 307–330. MR 80f:22005 Zbl 0392.43013

[Fell 1962] J. M. G. Fell, "A new proof that nilpotent groups are CCR", *Proc. Amer. Math. Soc.* **13** (1962), 93–99. MR 24 #A3238 Zbl 0105.09602

[Fell and Doran 1988a] J. M. G. Fell and R. S. Doran, *Representations of *-algebras, locally compact groups, and Banach *-algebraic bundles, I: Basic representation theory of groups and algebras*, Pure and Applied Mathematics **125**, Academic Press, Boston, 1988. MR 90c:46001 Zbl 0652.46050

[Fell and Doran 1988b] J. M. G. Fell and R. S. Doran, *Representations of *-algebras, locally compact groups, and Banach *-algebraic bundles, II: Banach *-algebraic bundles, induced representations, and the generalized Mackey analysis*, Pure and Applied Mathematics **126**, Academic Press, Boston, 1988. MR 90c:46002 Zbl 0652.46051

[Ferrara et al. 1981] S. Ferrara, C. A. Savoy, and B. Zumino, "General massive multiplets in extended supersymmetry", *Phys. Lett. B* **100**:5 (1981), 393–398. MR 82f:81085

[Jakobsen 1994] H. P. Jakobsen, *The full set of unitarizable highest weight modules of basic classical Lie superalgebras*, vol. 111, Mem. Amer. Math. Soc. **532**, Amer. Math. Soc., Providence, RI, 1994. MR 95c:17013 Zbl 0811.17002

[Kostant 1977] B. Kostant, "Graded manifolds, graded Lie theory, and prequantization", pp. 177–306 in *Differential geometrical methods in mathematical physics* (Bonn, 1975), edited by K. Bleuler and A. Reetz, Lecture Notes in Math. **570**, Springer, Berlin, 1977. MR 58 #28326 Zbl 0358.53024

[Musson 2012] I. M. Musson, *Lie superalgebras and enveloping algebras*, Graduate Studies in Mathematics **131**, Amer. Math. Soc., Providence, RI, 2012. MR 2906817 Zbl 1255.17001

[Neeb 2000] K.-H. Neeb, *Holomorphy and convexity in Lie theory*, de Gruyter Expositions in Mathematics **28**, Walter de Gruyter, Berlin, 2000. MR 2001j:32020 Zbl 0936.22001

[Neeb and Salmasian 2011] K.-H. Neeb and H. Salmasian, "Lie supergroups, unitary representations, and invariant cones", pp. 195–239 in *Supersymmetry in mathematics and physics*, edited by S. Ferrara et al., Lecture Notes in Math. **2027**, Springer, Heidelberg, 2011. MR 2012m:22022 Zbl 1275.22011

[Neeb and Salmasian 2013a] K.-H. Neeb and H. Salmasian, "Differentiable vectors and unitary representations of Fréchet–Lie supergroups", *Math. Z.* **275**:1-2 (2013), 419–451. MR 3101815 Zbl 1277.22020

[Neeb and Salmasian 2013b] K.-H. Neeb and H. Salmasian, "Positive definite superfunctions and unitary representations of Lie supergroups", *Transform. Groups* **18**:3 (2013), 803–844. MR 3084335 Zbl 1276.22012

[Neeb et al. 2015] K.-H. Neeb, H. Salmasian, and C. Zellner, "Smoothing operators and C* algebras for infinite dimensional Lie groups", preprint, 2015. arXiv 1505.02659

[Poulsen 1972] N. S. Poulsen, "On C^∞-vectors and intertwining bilinear forms for representations of Lie groups", *J. Functional Analysis* **9** (1972), 87–120. MR 46 #9239 Zbl 0237.22013

[Salam and Strathdee 1974] A. Salam and J. Strathdee, "Unitary representations of super-gauge symmetries", *Nuclear Phys.* **B80** (1974), 499–505. MR 50 #12029

[Salmasian 2010] H. Salmasian, "Unitary representations of nilpotent super Lie groups", *Comm. Math. Phys.* **297**:1 (2010), 189–227. MR 2011f:22011 Zbl 1192.22003

Received June 4, 2015. Revised July 30, 2015.

KARL-HERMANN NEEB
DEPARTMENT MATHEMATIK
FAU ERLANGEN-NÜRNBERG
CAUERSTR. 11
D-91058 ERLANGEN
GERMANY
neeb@mi.uni-erlangen.de

HADI SALMASIAN
DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF OTTAWA
585 KING EDWARD AVENUE
OTTAWA, ON K1N 6N5
CANADA
hsalmasi@uottawa.ca

PACIFIC JOURNAL OF MATHEMATICS
Vol. 282, No. 1, 2016
dx.doi.org/10.2140/pjm.2016.282.233

ASSOCIATED PRIMES OF LOCAL COHOMOLOGY MODULES OVER REGULAR RINGS

TONY J. PUTHENPURAKAL

Let R be an excellent regular ring of dimension d containing a field K of characteristic zero. Let I be an ideal in R. We show that $\operatorname{Ass} H_I^{d-1}(R)$ is a finite set. As an application, we show that if I is an ideal of height g with height $Q = g$ for all minimal primes of I then for all but finitely many primes $P \supseteq I$ with height $P \geq g+2$, the topological space $\operatorname{Spec}^\circ(R_P/IR_P)$ is connected. We also show that to prove a conjecture of Lyubeznik (regarding finiteness of associate primes for local cohomology modules) for all excellent regular rings of dimension $\leq d$ containing a field of characteristic zero, it suffices to prove $\operatorname{Ass}_S H_J^{g+1}(S)$ is finite for *all* ideals J in S of height g (here $0 \leq g \leq d$), where S is an excellent regular domain of dimension $\leq d$ containing an uncountable field of characteristic zero.

1. Introduction

Throughout this paper R is a commutative Noetherian ring. If M is an R-module and if I is an ideal in R, we denote by $H_I^i(M)$ the i-th local cohomology module of M with respect to I.

The following conjecture is due to Lyubeznik [2002]:

Conjecture 1.1. *If R is a regular ring, then each local cohomology module $H_I^i(R)$ has finitely many associated prime ideals.*

There are many cases where this conjecture is true: for regular rings R of prime characteristic [Huneke and Sharp 1993], for regular local and affine rings of characteristic zero [Lyubeznik 1993], and for unramified regular local rings of mixed characteristic [Lyubeznik 2000]. It is also true for smooth $\mathbb{Z}$-algebras [Bhatt et al. 2014].

Lyubeznik [2002] especially asked whether Conjecture 1.1 is valid for a regular ring R containing a field of characteristic zero. It is easy to give examples where existing techniques, to show finiteness of associate primes of local cohomology modules, fail:

MSC2010: primary 13D45; secondary 13D02, 13H10.
Keywords: local cohomology, associate primes, D-modules.

Example 1.2. (a) Let $(S, \mathfrak{m})$ be a complete local domain of dimension $d \geq 2$ containing a field of characteristic zero. Assume S is not regular. Let the singular locus be defined by the ideal J. Notice $J \neq 0$. Let $x \in J$ be nonzero. Set $R = S_x$. Then R is a domain of dimension $d-1$; see [Matsumura 1989, Lemma 1, p. 247]. Also clearly R is regular.

(b) Let T be a regular domain as above containing a field K of characteristic zero. Let $f \in K[X_1, \dots, X_n]$ be a smooth polynomial. Then $R = T[X_1, \dots, X_n]/(f)$ is a regular ring.

In both the examples above, we do not know whether $\operatorname{Ass}_R(H^i_I(R))$ is a finite set for all ideals I of R.

In all the essential cases where finiteness of associated primes is known, the local cohomology modules of R have some additional global structure. If R is of characteristic p then local cohomology modules have the structure of F-modules [Lyubeznik 1997]. In characteristic zero, for complete local rings and smooth affine algebras over algebraically closed fields, local cohomology modules have an appropriate D-module structure. For smooth $\mathbb{Z}$-algebras, Bhatt et al. [2014] use a rather clever mixture of D-module and F-module theory. For a general regular ring containing a field of characteristic zero, there is no obvious structure that local cohomology modules satisfy that we can exploit to prove finiteness of associate primes.

For the rest of the paper, assume that R contains a field of characteristic zero. For simplicity, we assume that $\dim R = d$ is finite. By Grothendieck's vanishing theorem, $H^i_I(R) = 0$ for all $i > d$; see [Brodmann and Sharp 1998, Theorem 6.1.2]. In general, for a Noetherian ring R of dimension d, the set $\operatorname{Ass}_R(H^d_I(R))$ is finite; see [Brodmann et al. 2000, Remark 3.11] (also see [Marley 2001, Proposition 2.3]). If R is a regular ring of dimension d and I is an ideal in R then using the Hartshorne–Lichtenbaum theorem (cf. [Iyengar et al. 2007, Theorem 14.1]), it is easy to prove that
$$\operatorname{Ass}_R(H^d_I(R)) = \{P \mid P \in \operatorname{Min} R/I \text{ and } \operatorname{height} P = d\}.$$

The following is the main result of this paper:

Theorem 1.3. *Let R be an excellent regular ring of dimension d containing a field of characteristic zero. Let I be an ideal in R. Then $\operatorname{Ass}_R(H^{d-1}_I(R))$ is a finite set.*

The main idea of this paper is that it is fruitful to look at the following relative situation: Let $I \supseteq J$ be ideals in a Noetherian ring R. We have a natural map $\theta^i_{I,J} \colon H^i_I(R) \to H^i_J(R)$ for each $i \geq 0$. Fix $i_0 \geq 0$. Set
$$\mathcal{C}^{i_0}_R = \{(I, J) \mid I \supset J \text{ and } \sharp \operatorname{Ass}_R(\operatorname{image} \theta^{i_0}_{I,J}) = \infty\}.$$

Here $\sharp S$ denotes the number of elements in a set S. If $\operatorname{Ass}_R H^{i_0}_I(R)$ is infinite then $(I, I) \in \mathcal{C}^{i_0}_R$. Conversely, if $(I, J) \in \mathcal{C}^{i_0}_R$ then $\operatorname{Ass}_R H^{i_0}_J(R)$ is infinite. We partially order $\mathcal{C}^{i_0}_R$ as follows: set $(I, J) \preceq (I', J')$ if $I' \supseteq I$ and $J' \supseteq J$. It is easy to see that every ascending chain in $\mathcal{C}_R^{i_0}$ stabilizes. If $\mathcal{C}_R^{i_0}$ is nonempty then its maximal elements have some peculiar properties; see Lemma 2.10.

The following result is a crucial ingredient in the proof of Theorem 1.3.

Theorem 1.4. *Let R be an excellent regular domain of dimension d containing an uncountable field of characteristic zero. Assume for some i_0 that the set $\mathcal{C}_R^{i_0}$ is nonempty. Let (I, J) be a maximal element in $\mathcal{C}_R^{i_0}$. Then there exists a multiplicatively closed set S of R such that in the ring $A = S^{-1}R$ we have*

(a) $(S^{-1}I, S^{-1}J)$ *is a maximal element in* $\mathcal{C}_A^{i_0}$,

(b) $\operatorname{height} S^{-1}I = i_0$,

(c) $S^{-1}I = P_1 \cap P_2 \cap \cdots \cap P_r$, *where P_i is a prime in A of height i_0*,

(d) $\operatorname{Ass} \operatorname{image} \theta^{i_0}_{S^{-1}I, S^{-1}J} \supseteq \text{m-Spec}(A)$,

(e) $\operatorname{height} \mathfrak{m} = \operatorname{height} \mathfrak{m}'$ *for* $\mathfrak{m}, \mathfrak{m}' \in \text{m-Spec}(A)$,

(f) $\text{m-Spec}(A)$ *is a countably infinite set.*

Furthermore, if $\operatorname{Ass}_R(H^r_L(R))$ *is a finite set for all $r < i_0$ and for all ideals L of R then* $\operatorname{height} S^{-1}J = i_0 - 1$.

The assumptions that R is a domain and contains an uncountable field are mild hypotheses; see Section 3. The assumption on the excellence of R is satisfied by most examples. As an easy consequence of Theorem 1.4, we get the following significant simplification of Lyubeznik's conjecture.

Theorem 1.5. *The following are equivalent*:

(i) *Lyubeznik's conjecture has a positive answer for all excellent regular rings of dimension $\leq d$ containing a field of characteristic zero.*

(ii) *For all excellent regular domains R of dimension $\leq d$ containing an uncountable field of characteristic zero,* $\operatorname{Ass}_R(H^{g+1}_J(R))$ *is a finite set for* ***all*** *ideals J of height g, with $1 \leq g \leq d$.*

The following are applications of Theorem 1.3:

Corollary 1.6. *Let R be an excellent regular ring of dimension $d \leq 4$ containing a field K. Then for any ideal I, we have* $\operatorname{Ass} H^i_I(R)$ *is a finite set for all $i \geq 0$.*

If M is an R-module then set

$$\operatorname{Ass}^i_R(M) = \{P \mid P \in \operatorname{Ass} M \text{ and } \operatorname{height} P = i\}.$$

Using the Hartshorne–Lichtenbaum theorem, we get that if R is regular and I is an ideal in R then

$$\bigcup_{i \geq 0} \operatorname{Ass}^i_R(H^i_I(R)) = \operatorname{Min} R/I;$$

see Corollary 8.2.

Corollary 1.7. *Let R be an excellent regular ring of dimension d containing a field of characteristic zero. Let I be an ideal in R. Then*

$$\bigcup_{i\geq 0} \operatorname{Ass}_R^{i+1}(H_I^i(R)) \text{ is a finite set.}$$

We need a description of primes which appear in Corollary 1.7. To do so, we first make the following definition: recall if $(A, \mathfrak{m})$ is a local ring then $\operatorname{Spec}^\circ(A) = \operatorname{Spec}(A) \setminus \{\mathfrak{m}\}$ considered as a subspace of $\operatorname{Spec}(A)$.

Definition 1.8. Let $(A, \mathfrak{m})$ be a local ring and let I be an ideal in A. We say $\operatorname{Spec}^\circ(A/I)$ is *absolutely connected* if for every flat local map $(A, \mathfrak{m}) \to (B, \mathfrak{n})$ with $\mathfrak{m}B = \mathfrak{n}$, B complete and $B/\mathfrak{n}$ algebraically closed, $\operatorname{Spec}^\circ(B/IB)$ is connected.

Remark 1.9. It is easy to see that if $\operatorname{Spec}^\circ(A/I)$ is absolutely connected then it is connected.

As a consequence of a result of Ogus [1973, Corollary 2.11] (also see [Huneke and Lyubeznik 1990, Theorem 1.1]), we get this:

Proposition 1.10. *Let R be an excellent regular ring of dimension d containing a field of characteristic zero and let I be an ideal in R of height g. Assume* height $Q = g$ *for all minimal primes Q of I. Then*

$$\bigcup_{i\geq 0} \operatorname{Ass}_R^{i+1}(H_I^i(R)) = \big\{P \bigm| P \supseteq I,\ \text{height } P \geq g+2, \textit{ and } \operatorname{Spec}^\circ(R_P/I_P) \textit{ is } \text{NOT } \textit{absolutely connected}\big\}.$$

As an immediate application we get this:

Corollary 1.11. *Assume the hypotheses of Proposition 1.10 hold. For all but finitely many primes $P \supseteq I$ with* height $P \geq g+2$, *we get that* $\operatorname{Spec}^\circ(R_P/I_P)$ *is absolutely connected. In particular,* $\operatorname{Spec}^\circ(R_P/I_P)$ *is connected.*

Here is an overview of the contents of the paper. In Section 2 we discuss a few general results that we need. In the next section we discuss the flat extension $R \to R[[X]]_X$. We need it as an essential technique in our paper requires an uncountable field contained in R. In Section 4 we discuss countable prime avoidance. We also give a construction which is used several times in our paper. In Section 5 we prove Theorem 1.4. In the next section we prove our main result, Theorem 1.3. In Section 7 we prove the simplicity of a D-module. This is needed in the proof of Theorem 1.3. In Section 8 we prove Corollary 1.6. Finally in Section 9 we give proofs of Corollary 1.7 and Proposition 1.10.

2. Generalities

In this section we prove some general results. Some of them are perhaps already known to the experts. However, we prove them as we do not have a reference.

We first prove the following general result:

Proposition 2.1. *Let R be a Noetherian ring and let I be an ideal in R. Let M be a finitely generated R-module. Then $\operatorname{Ass}_R(H^i_I(M))$ is a countable set.*

To prove the proposition, we need the following notion.

Definition 2.2. We say an R-module E is *countably generated* if there exists a countable set of elements $\{e_n\}_{n\geq 1}$ which generate E as an R-module.

The following lemma is useful:

Lemma 2.3. *Let R be a Noetherian ring and let E be a countably generated R-module. Then $\operatorname{Ass}_R(E)$ is a countable set.*

Proof. Let $\{e_n\}_{n\geq 1}$ generate E as an R-module. For $m \geq 1$, let D_m be the R-submodule of E generated by $e_1, \dots, e_m$. Clearly $D_m \subseteq D_{m+1}$ for all $m \geq 1$, $E = \bigcup_{m\geq 1} D_m$ and $\bigcup_{m\geq 1} \operatorname{Ass}_R(D_m) \subseteq \operatorname{Ass}_R(E)$. If $P \in \operatorname{Ass}_R(E)$ then $P = (0\colon u)$ for some $u \in E$. Say $u \in D_r$. Then $P \in \operatorname{Ass}_R(D_r)$. Thus

$$\bigcup_{m\geq 1} \operatorname{Ass}_R(D_m) = \operatorname{Ass}_R(E).$$

Since R is Noetherian and D_m is a finitely generated R-module, $\operatorname{Ass}_R(D_m)$ is a finite set for all $m \geq 1$. It follows that $\operatorname{Ass}_R(E)$ is a countable set. $\square$

Proof of Proposition 2.1. Fix $i \geq 0$. Then

$$H^i_I(M) = \varinjlim \operatorname{Ext}^i_R(R/I^n, M).$$

In particular $H^i_I(M)$ is a quotient of $\bigoplus_{n\geq 1} \operatorname{Ext}^i_R(R/I^n, M)$. It follows that $H^i_I(M)$ is countably generated. The result now follows from Lemma 2.3. $\square$

2.4. We need to compare R-linear maps $f, g\colon E \to F$, where E, F are R-modules.

Definition 2.5. Let $f, g\colon E \to F$ be R-linear maps. We say $f \cong g$ if there exist isomorphisms $\alpha\colon E \to E$ and $\beta\colon F \to F$ such that the following diagram commutes:

$$\begin{array}{ccc} E & \xrightarrow{f} & F \\ \downarrow{\scriptstyle\alpha} & & \downarrow{\scriptstyle\beta} \\ E & \xrightarrow{g} & F \end{array}$$

The following result is clear:

Proposition 2.6. *Let R be a ring and let E, F be R-modules. Let $f, g\colon E \to F$ be R-linear. If $f \cong g$ then we have the following isomorphisms of R-modules:*

$$\ker f \cong \ker g, \quad \operatorname{image} f \cong \operatorname{image} g \quad \textit{and} \quad \operatorname{coker} f \cong \operatorname{coker} g. \qquad \square$$

2.7. Let R be a Noetherian ring and let I, J be ideals of R with $I \supseteq J$. Let Γ_I, Γ_J be the I-torsion and J-torsion functors respectively. Let M be an R-module. Let us recall the construction of the natural maps

$$\theta^i_{I,J}(M) \colon H^i_I(M) \to H^i_J(M)$$

for all $i \geq 0$:

Let $\mathbb{E}$ be an injective resolution of M. Then note that we have a natural morphism of complexes $\theta \colon \Gamma_I(\mathbb{E}) \to \Gamma_J(\mathbb{E})$. Taking cohomology, we obtain our natural maps $\theta^i_{I,J}(M)$ for all $i \geq 0$.

The following result will be used several times.

Lemma 2.8. *Let $R \to S$ be a flat map of Noetherian rings. Let I, J be ideals of R with $I \supseteq J$. Let M be an R-module. Then for all $i \geq 0$, we have*

$$\theta^i_{I,J}(M) \otimes S \cong \theta^i_{IS,JS}(M \otimes S).$$

Proof. Let $\mathbb{E}$ be an injective resolution of M. Then $\mathbb{E} \otimes S$ is Γ_{KS}-acyclic for any ideal K of R; see [Brodmann and Sharp 1998, Theorem 4.1.9]. Furthermore, we have a natural equivalence of functors $\Gamma_K \otimes S \cong \Gamma_{KS}$; see [loc. cit., Proposition 4.3.1]. Thus we have a commutative diagram of complexes

$$\begin{tikzcd} \Gamma_I(\mathbb{E}) \otimes S \arrow[r, "\theta(M)\otimes S"] \arrow[d, "\alpha"] & \Gamma_J(\mathbb{E}) \otimes S \arrow[d, "\beta"] \\ \Gamma_{IS}(\mathbb{E} \otimes S) \arrow[r, "\theta(M\otimes S)"] & \Gamma_{JS}(\mathbb{E} \otimes S) \end{tikzcd}$$

where α and β are isomorphisms of complexes. The result follows. ☐

2.9. Let R be a Noetherian ring. Let $I \supseteq J$ be ideals in R. We have a natural map $\theta^i_{I,J} \colon H^i_I(R) \to H^i_J(R)$ for each $i \geq 0$. Fix $i_0 \geq 0$. Set

$$\mathcal{C}^{i_0}_R = \{(I, J) \mid I \supset J \text{ and } \sharp \operatorname{Ass}_R(\operatorname{image} \theta^{i_0}_{I,J}) = \infty\}.$$

Here $\sharp S$ denotes the number of elements in a set S. We partially order $\mathcal{C}^{i_0}_R$ as follows: set $(I, J) \preceq (I', J')$ if $I' \supseteq I$ and $J' \supseteq J$. It is easy to see that every ascending chain in $\mathcal{C}^{i_0}_R$ stabilizes. If $\mathcal{C}^{i_0}_R$ is nonempty then its maximal element has a peculiar property which we now describe:

Lemma 2.10. *Assume the hypotheses of 2.9 hold. Assume $(I, J) \in \mathcal{C}^{i_0}_R$ is a maximal element. Let S be a multiplicatively closed subset of R. If $(S^{-1}I, S^{-1}J) \in \mathcal{C}^{i_0}_{S^{-1}R}$ then*

(1) *$(S^{-1}I, S^{-1}J)$ is a maximal element in $\mathcal{C}^{i_0}_{S^{-1}R}$,*
(2) *$S^{-1}I \cap R = I$ and $S^{-1}J \cap R = J$.*

To prove the lemma, we need the following easy result.

Proposition 2.11. *Assume the hypotheses of Lemma 2.10 hold. If* $(K, L) \in \mathcal{C}^{i_0}_{S^{-1}R}$ *then* $(K \cap R, L \cap R) \in \mathcal{C}^{i_0}_{R}$.

Proof. Set $K_1 = K \cap R$ and $L_1 = L \cap R$. We note that $S^{-1}K_1 = K$ and $S^{-1}L_1 = L$. By Propositions 2.8 and 2.6, we get that

$$S^{-1}(\text{image}\, \theta^{i_0}_{K_1,L_1}(R)) \cong \text{image}\, \theta^{i_0}_{K,L}(S^{-1}R).$$

It follows that

$$\sharp\, \operatorname{Ass}_R(\text{image}\, \theta^{i_0}_{K_1,L_1}(R)) = \infty.$$

So $(K_1, L_1) \in \mathcal{C}^{i_0}_{R}$. ☐

Proof of Lemma 2.10. (1) Suppose $(S^{-1}I, S^{-1}J) \preceq (K, L)$ for some $(K, L) \in \mathcal{C}^{i_0}_{S^{-1}R}$. Then by Proposition 2.11, we get $(K \cap R, L \cap R) \in \mathcal{C}^{i_0}_{R}$. Notice that $(I, J) \preceq (K \cap R, L \cap R)$. By the maximality of (I, J) in $\mathcal{C}^{i_0}_{R}$, we get that $I = K \cap R$ and $J = L \cap R$. So $(S^{-1}I, S^{-1}J) = (K, L)$.

(2) Set $K = S^{-1}I$ and $L = S^{-1}J$. Then by Proposition 2.11 and our hypotheses, $(K \cap R, L \cap R) \in \mathcal{C}^{i_0}_{R}$. Notice that $(I, J) \preceq (K \cap R, L \cap R)$. So by the maximality of (I, J) in $\mathcal{C}^{i_0}_{R}$, we get $I = K \cap R$ and $J = L \cap R$. ☐

2.12. ***Product of rings.*** Assume $R = R_1 \times R_2 \times \cdots \times R_n$. If R is Noetherian then each R_i is Noetherian. Note that an ideal I in R is of the form $I_1 \times I_2 \times \cdots \times I_n$, where I_j is an ideal in R_j. Further note that P is a prime in R if and only if P_i is a prime ideal in R_i for some i and $P_j = R_j$ for $j \neq i$. Thus $\operatorname{Spec}(R)$ is a disjoint union of $\operatorname{Spec}(R_1), \ldots, \operatorname{Spec}(R_n)$.

The following result is easy to verify.

Proposition 2.13. *Let* M_i *be* R_i *modules. Then* $M = M_1 \times \cdots \times M_n$ *is an* R*-module. Furthermore*,

$$\operatorname{Ass}_R(M) = \bigcup_{i=1}^{n} \operatorname{Ass}_{R_i}(M_i).$$

Remark 2.14. In the above proposition, a prime P of R_i is identified with the following prime of R:

$$R_1 \times \cdots \times R_{i-1} \times P \times R_{i+1} \times \cdots \times R_n.$$

The following result takes a little work. However, it is completely elementary and so we skip the proof.

Proposition 2.15. *Assume the hypotheses of 2.12 hold. For each* $i \geq 0$, *we have an isomorphism*

$$H^i_I(R) \cong H^i_{I_1}(R_1) \times H^i_{I_2}(R_2) \times \cdots \times H^i_{I_n}(R_n).$$

The following remark will be used often.

Remark 2.16. Let R be a regular ring. Let $\{P_1, \dots, P_n\}$ be minimal primes of R. Set $R_i = R/P_i$. Then

$$R \cong R_1 \times R_2 \times \cdots \times R_n;$$

see [Matsumura 1989, Exercise 9.11]. Notice that the R_i are regular domains.

If R is excellent then each R_i is excellent. Furthermore, if $\dim R = d$ then it is easy to see that $\dim R_i \le d$ for all i and $\dim R_m = d$ for some m.

By Propositions 2.15 and 2.13, it follows that if

$$I = I_1 \times \cdots \times I_n$$

then $\operatorname{Ass} H^i_I(R)$ is finite if and only if $\operatorname{Ass} H^i_{I_j}(R_j)$ is finite for all j. Thus for the questions we are interested in, it suffices to assume that R is a domain.

3. The flat extension $R \to R[[X]]_X$

In our arguments we need to assume that R contains an uncountable field. When this is not the case, we consider the flat extension $R \to R[[X]]_X$. Set $S = R[[X]]$ and let $T = S_X = R[[X]]_X$, i.e., the ring obtained by inverting X.

Remark 3.1. (i) If R contains a countable field K then note that $K[[X]]$ is a subring of S and so $K[[X]]_X$ is a subring of T. The field $K[[X]]_X$ is uncountable. Thus T contains an uncountable field.

(ii) If R is regular then so is S; see [Bruns and Herzog 1993, Theorem 2.2.13]. Therefore T is also regular.

(iii) Let R be an excellent regular ring of finite dimension containing a field of characteristic zero. As $S = R[[X]]$ is regular, it is universally catenary. Also $(X) \subseteq \operatorname{rad} S$ and $S/(X) = R$ is excellent. So by [Rotthaus 1980], we get that S is excellent. It follows that T is excellent.

The following proposition gives information about the behavior of primes when we pass from R to T.

Proposition 3.2. *Assume that the hypotheses above hold. Let $\mathfrak{p}$ be a prime in R. Then*:

(i) *$\mathfrak{p}T$ is a prime in T.*

(ii) *T is a faithfully flat extension of R.*

(iii) $\mathfrak{p}T \cap R = \mathfrak{p}$.

(iv) $\operatorname{height} \mathfrak{p}T = \operatorname{height} \mathfrak{p}$.

(v) *Let P be a prime ideal in T. If $P \cap R = \mathfrak{p}$ then $P = \mathfrak{p}T$.*

(vi) $\dim T = \dim R$.

Proof. (i) Clearly $\mathfrak{p}S$ is a prime in S. Also $X \notin \mathfrak{p}S$. As $\mathfrak{p}T$ is localization of $\mathfrak{p}S$, we get that it is prime in T.

(ii) T is clearly a flat R-algebra. Let $\mathfrak{m}$ be a maximal ideal of R. Then $\mathfrak{m}T$ is a prime ideal of T. In particular, $\mathfrak{m}T \neq T$. So T is faithfully flat [Matsumura 1989, Theorem 7.2].

(iii) We have

$$\mathfrak{p}T \cap R = \mathfrak{p}T \cap S \cap R = \mathfrak{p}S \cap R = \mathfrak{p}.$$

(iv) By [loc. cit., Theorem 15.1], we get

$$\operatorname{height} \mathfrak{p}T = \operatorname{height} \mathfrak{p} + \dim T_{\mathfrak{p}T}/\mathfrak{p}T_{\mathfrak{p}T}.$$

As $\mathfrak{p}T_{\mathfrak{p}T}$ is the maximal ideal in $T_{\mathfrak{p}T}$, we get the required result.

(v) Again by [loc. cit., Theorem 15.1], we get

$$\operatorname{height} P = \operatorname{height} \mathfrak{p} + \dim T_P/\mathfrak{p}T_P.$$

Set $\kappa(\mathfrak{p}) = R_{\mathfrak{p}}/\mathfrak{p}R_{\mathfrak{p}}$, the residue field of $R_{\mathfrak{p}}$. Then note that $T_P/\mathfrak{p}T_P$ is a localization of $\kappa(\mathfrak{p})[[X]]$ at a multiplicatively closed set containing X. It follows that $T_P/\mathfrak{p}T_P$ is a field. So $\operatorname{height} P = \operatorname{height} \mathfrak{p}$. Now P contains the prime ideal $\mathfrak{p}T$ which by (iii) also has height $= \operatorname{height} \mathfrak{p}$. So $P = \mathfrak{p}T$.

(vi) This follows easily from (iv) and (v). □

We need Theorem 23.3 from [Matsumura 1989]. Unfortunately there is a typographical error in the statement of this theorem, so we state it here.

Theorem 3.3. *Let $\varphi \colon A \to B$ be a homomorphism of Noetherian rings, and let E be an A-module and G a B-module. Suppose that G is flat over A; then we have the following:*

(i) *If $\mathfrak{p} \in \operatorname{Spec} A$ and $G/\mathfrak{p}G \neq 0$ then*

$$^{a}\varphi(\operatorname{Ass}_B(G/\mathfrak{p}G)) = \operatorname{Ass}_A(G/\mathfrak{p}G) = \{\mathfrak{p}\}.$$

(ii)
$$\operatorname{Ass}_B(E \otimes_A G) = \bigcup_{\mathfrak{p} \in \operatorname{Ass}_A(E)} \operatorname{Ass}_B(G/\mathfrak{p}G).$$

Remark 3.4. In [Matsumura 1989], $\operatorname{Ass}_A(E \otimes G)$ is typed instead of $\operatorname{Ass}_B(E \otimes G)$. Also note that $^{a}\varphi(P) = P \cap A$ for $P \in \operatorname{Spec} B$.

3.5. Let M be an R-module. Set

$$\operatorname{Ass}^i_R(M) = \{P \mid P \in \operatorname{Ass}_R(M) \text{ and } \operatorname{height} P = i\}.$$

We now state the main result of this section.

Theorem 3.6. *Let R be a Noetherian ring and let M be an R-module. Set $S = R[[X]]$ and $T = S_X$. Then*:

(1) *The mapping defined by*

$$\psi \colon \operatorname{Ass}_R(M) \to \operatorname{Ass}_T(M \otimes_R T),$$
$$\mathfrak{p} \to \mathfrak{p}T,$$

is a bijection.

(2) *ψ maps $\operatorname{Ass}^i_R(M)$ bijectively to $\operatorname{Ass}^i_T(M \otimes_R T)$.*

Proof. (1) By Theorem 3.3, we get that

$$\operatorname{Ass}_T(M \otimes_R T) = \bigcup_{\mathfrak{p} \in \operatorname{Ass}_R(M)} \operatorname{Ass}_T(T/\mathfrak{p}T).$$

By Proposition 3.2(i), we get that $\mathfrak{p}T$ is a prime ideal in T. So $\operatorname{Ass}_T T/\mathfrak{p}T = \{\mathfrak{p}T\}$. Thus

$$\operatorname{Ass}_T(M \otimes_R T) = \{\mathfrak{p}T \mid \mathfrak{p} \in \operatorname{Ass}_R(M)\}.$$

So the map ψ is well-defined and surjective. By Proposition 3.2(iii), we get that it is injective.

(2) This follows from (1) and Proposition 3.2(iv). □

An immediate corollary is this:

Corollary 3.7. *Assume that the hypotheses of Theorem 3.6 hold. Then*:

(1) *$\operatorname{Ass}_R M$ is an infinite set if and only if $\operatorname{Ass}_T M \otimes_R T$ is an infinite set.*

(2) *$\operatorname{Ass}^i_R M$ is an infinite set if and only if $\operatorname{Ass}^i_T M \otimes_R T$ is an infinite set.* □

4. Countable prime avoidance

4.1. ***Setup.*** In this section, R is a Noetherian ring containing an uncountable field K. We describe a construction which we will use often. The essential ingredient is the following well-known countable avoidance. We give a proof due to lack of a suitable reference.

Lemma 4.2. *Assume the hypotheses of 4.1 hold. Let $\{I_n\}_{n \geq 1}$ be ideals in R and let J be another ideal in R. If $J \subseteq \bigcup_{n \geq 1} I_n$ then $J \subseteq I_m$ for some m.*

Proof. Let $J = (x_1, \dots, x_c)$. Let $V = Kx_1 + Kx_2 + \cdots + Kx_c$. Then V is a finite dimensional K-vector space. Also, clearly $V = \bigcup_{n \geq 1} V \cap I_n$. As K is an uncountable field, we get that $V \cap I_m = V$ for some m. Thus $x_i \in I_m$ for all i. So $J \subseteq I_m$. □

4.3. ***Construction.*** Assume that the hypotheses of 4.1 hold and that $\{\mathfrak{p}_n\}_{n\geq 1}$ is a sequence of primes in R. Also assume that $\mathfrak{p}_i \not\subseteq \mathfrak{p}_j$ for $i \neq j$. Consider the multiplicatively closed set

$$S = R \setminus \bigcup_{n\geq 1} \mathfrak{p}_n.$$

Set $T = S^{-1}R$. The following result gives information about primes in T.

Proposition 4.4. *Assume the hypotheses of 4.3 hold. We have*:

(1) *If $\mathfrak{p}T$ is a prime in T, where $\mathfrak{p}$ is a prime in R, then $\mathfrak{p} \subseteq \mathfrak{p}_n$ for some n.*

(2) *$\mathfrak{p}_i T \not\subseteq \mathfrak{p}_j T$ for $i \neq j$.*

(3) *$\mathfrak{p}_n T$ are distinct maximal ideals of T.*

(4) m-Spec$(T) = \{\mathfrak{p}_n T\}_{n\geq 1}$.

Proof. (1) We have $\mathfrak{p} \cap S = \varnothing$. So $\mathfrak{p} \subseteq \bigcup_{n\geq 1} \mathfrak{p}_n$. By Lemma 4.2, we get that $\mathfrak{p} \subseteq \mathfrak{p}_n$ for some n.

(2) If $\mathfrak{p}_i T \subseteq \mathfrak{p}_j T$ for some $i \neq j$, then intersecting with R, we get $\mathfrak{p}_i \subseteq \mathfrak{p}_j$ for some $i \neq j$; a contradiction.

(3) Let $\mathfrak{p}_n T \subseteq P$ for some prime P of T. Say $P = \mathfrak{p}T$, where $\mathfrak{p}$ is a prime in R. By (1) we get that $\mathfrak{p}T \subseteq \mathfrak{p}_m T$ for some m. By (2) we get $m = n$. So $P = \mathfrak{p}_n T$. Thus $\mathfrak{p}_n T$ is a maximal ideal in T. That they are all distinct follows from (2).

(4) This follows from (1) and (3). □

We will need the following intersection result:

Proposition 4.5. *Assume that the hypotheses of 4.3 hold. Let Λ be any subset of $\{\mathfrak{p}_n\}_{n\geq 1}$ (possibly infinite). Set*

$$U = \bigcap_{\mathfrak{p}\in\Lambda} \mathfrak{p} \quad \textit{and} \quad V = \bigcap_{\mathfrak{p}\in\Lambda} \mathfrak{p}T.$$

Then $UT = V$.

Proof. Clearly $UT \subseteq V$ Let $\xi = a/s \in V$ and $\mathfrak{p} \in \Lambda$. As $\xi \in \mathfrak{p}T$, we get that $\xi = r/s_1$, where $r \in \mathfrak{p}$. It follows that there exists $s' \in S$ such that $s's_1 a \in \mathfrak{p}$. As $s's_1 \notin \mathfrak{p}$, we get that $a \in \mathfrak{p}$. Thus $a \in U$. Therefore $\xi \in UT$, and hence $V = UT$. □

5. Proof of Theorem 1.4

Proof of Theorem 1.4. Suppose $\mathcal{C}_R^{i_0}$ is nonempty for some i_0. Let (I, J) be a maximal element in $\mathcal{C}_R^{i_0}$. It follows from Proposition 2.1 that $\mathrm{Ass}_R(\text{image}\,\theta^{i_0}_{I,J})$ is a countably infinite set. As dim R is finite, we can choose an infinite subset $\{\mathfrak{p}_n\}_{n\geq 1}$

of $\operatorname{Ass}_R(\text{image}\, \theta^{i_0}_{I,J})$ such that $\operatorname{height} \mathfrak{p}_r = \operatorname{height} \mathfrak{p}_s$ for all r, s. Clearly the primes $\{\mathfrak{p}_n\}_{n\geq 1}$ are mutually incomparable. Set

$$S = R \setminus \bigcup_{n\geq 1} \mathfrak{p}_n \quad \text{and} \quad A = S^{-1}R.$$

By Proposition 4.4, we get $\text{m-Spec}(A) = \{\mathfrak{p}_n A\}_{n\geq 1}$. Furthermore, by construction, $\operatorname{height} \mathfrak{p}_n A = \operatorname{height} \mathfrak{p}_n$ is constant. Also by Lemma 2.8, we get that

$$\operatorname{Ass}_A(\text{image}\, \theta^{i_0}_{IA,JA}) \supseteq \text{m-Spec}(A).$$

By Lemma 2.10, we also get that (IA, JA) is a maximal element of $\mathcal{C}^{i_0}_A$.

Suppose, if possible, that $g = \operatorname{height} IA < i_0$.

Claim 1: $\operatorname{Ass}_A H^{i_0}_{IA}(A)$ is an infinite set.

Suppose, if possible, that $\operatorname{Ass}_A H^{i_0}_{IA}(A) = \{Q_1, Q_2, \dots, Q_s\}$ is a finite set. Notice that $\operatorname{height} Q_i \geq i_0 > g$ for all i. Note that IA is a radical ideal. Let $IA = P_1 \cap P_2 \cap \dots \cap P_r$ for some primes P_j. Say $\operatorname{height} P_1 = g$. Choose $x_i \in Q_i \setminus P_1$. Set $x = x_1 x_2 \cdots x_s$. Then $x \in Q_i$ for all i. Also $x \notin P_1$, so $x \notin IA$.

We have an exact sequence

$$\cdots \to H^{i_0}_{IA+(x)}(A) \xrightarrow{\theta^{i_0}_{IA+(x),IA}} H^{i_0}_{IA}(A) \to (H^{i_0}_{IA}(A))_x \to \cdots.$$

As $x \in Q_i$ for all i, we get that $(H^{i_0}_{IA}(A))_x = 0$. Thus $\theta^{i_0}_{IA+(x),IA}$ is surjective. Notice that

$$\theta^{i_0}_{IA+(x),JA} = \theta^{i_0}_{IA,JA} \circ \theta^{i_0}_{IA+(x),IA}.$$

As $\theta^{i_0}_{IA+(x),IA}$ is surjective, it follows that $\operatorname{Ass}_A(\text{image}\, \theta^{i_0}_{IA+(x),JA})$ is an infinite set. So

$$(IA + (x), JA) \in \mathcal{C}^{i_0}_A.$$

Also $x \notin IA$. This contradicts the maximality of (IA, JA) in $\mathcal{C}^{i_0}_A$. Thus $\operatorname{Ass}_A H^{i_0}_{IA}(A)$ is an infinite set.

Now let Λ be an *infinite* subset of $\operatorname{Ass}_A(H^{i_0}_{IA}(A))$. Set

$$W_\Lambda = \bigcap_{P\in\Lambda} P.$$

Clearly $IA \subseteq W_\Lambda$.

Claim 2: $IA = W_\Lambda$.

Suppose, if possible, that there exists $x \in W_\Lambda \setminus IA$. We have an exact sequence

$$\cdots \to H^{i_0}_{IA+(x)}(A) \xrightarrow{\theta^{i_0}_{IA+(x),IA}} H^{i_0}_{IA}(A) \xrightarrow{\pi_{i_0}} (H^{i_0}_{IA}(A))_x \to \cdots.$$

For $P \in \Lambda$, let $P = (0 \colon a_P)$ for some $a_P \in H^{i_0}_{IA}(A)$. Clearly $\pi_{i_0}(a_P) = 0$ for all $P \in \Lambda$. It follows that

$$\Lambda \subseteq \operatorname{Ass}_A(\text{image}(\theta^{i_0}_{IA+(x),IA})).$$

Thus $(IA+(x), IA)\in\mathcal{C}_A^{i_0}$. Also $x\notin IA$. This contradicts the maximality of (IA, JA) in $\mathcal{C}_A^{i_0}$. Thus Claim 2 is true.

Let

$$L_j=\{P \mid P\in \operatorname{Ass}_A(H_{IA}^{i_0}(A)) \text{ and height } P=j\}.$$

As $\dim A$ is finite, we get by Claim 1 that L_j is infinite for some j. Choose j_0 to be the maximum j with L_j infinite. Set

$$\Lambda=L_{j_0}\setminus\{\mathfrak{p}\mid \mathfrak{p} \text{ a minimal prime of } IA \text{ and height}\,\mathfrak{p}=j_0\}.$$

Clearly Λ is an infinite set. Set

$$T=A\setminus\bigcup_{P\in\Lambda}P \quad\text{and}\quad B=T^{-1}A.$$

By Proposition 4.4, we get that

$$\text{m-Spec}(B)=\{PB\mid P\in\Lambda\}.$$

It follows from Claim 2 and Proposition 4.5 that

$$IB=\bigcap_{P\in\Lambda}PB.$$

Thus IB is the Jacobson radical of B. Note we also get that $\operatorname{Ass}_B H_{IB}^{i_0}(B)$ contains m-Spec(B).

Since R is excellent, A is excellent and hence B is excellent. Therefore $\operatorname{Reg}(B/IB)$ is an open set. As IB is a radical ideal, we get that $\operatorname{Reg}(B/IB)$ is *nonempty*. Since $\operatorname{rad}(B/IB)=0$, it follows that there exists a maximal ideal $\mathfrak{m}$ of B with $(B/IB)_\mathfrak{m}$ a regular local ring. As B is a regular ring, $B_\mathfrak{m}$ is a regular ring. Also note that by our construction, $IB_\mathfrak{m}\neq\mathfrak{m}B_\mathfrak{m}$. As $B_\mathfrak{m}/IB_\mathfrak{m}$ is regular, we get that $IB_\mathfrak{m}$ is generated by part of a regular system of parameters. Say $IB_\mathfrak{m}=(x_1,\dots,x_c)$ and $c<\dim B_\mathfrak{m}$. We now note that

$$H_{IB_\mathfrak{m}}^j(B_\mathfrak{m})=0 \text{ for } j\neq c \quad\text{and}\quad \operatorname{Ass} H_{IB_\mathfrak{m}}^c(B_\mathfrak{m})=\{(x_1,\dots,x_c)\}.$$

It follows that $\mathfrak{m}\notin\operatorname{Ass}_B H_{IB}^{i_0}(B)$; a contradiction. Thus height $IA=i_0$.

Let $IA=P_1\cap P_2\cap\cdots\cap P_s\cap Q_1\cap Q_2\cap\cdots\cap Q_l$, where height $P_j=i_0$ for all j and height $Q_j\geq i_0+1$ for all j. Set $K=P_1\cap\cdots\cap P_s$. It is well known that the natural map

$$\theta_{K,IA}^{i_0}\colon H_K^{i_0}(A)\to H_{IA}^{i_0}(A) \quad \text{is an isomorphism.}$$

It follows that $(K, JA)\in\mathcal{C}_A^{i_0}$. By the maximality of (IA, JA) in $\mathcal{C}_A^{i_0}$, we get that $K=IA$.

Now assume $\operatorname{Ass}_R(H_L^r(R))$ is a finite set for all $r<i_0$ and for all ideals L of R. In particular, $\operatorname{Ass}_R(H_J^{i_0-1}(R))$ is a finite set. So $\operatorname{Ass}_A(H_{JA}^{i_0-1}(A))$ is a finite set. Suppose, if possible, that height $JA<i_0-1$. Let

$$\operatorname{Ass}_A(H_{JA}^{i_0-1}(A))=\{Q_1,Q_2,\dots,Q_s\}.$$

Notice height $Q_j \geq i_0 - 1$ for all j. Let $JA = P_1 \cap P_2 \cdots \cap P_r$, where height $P_1 =$ height $JA < i_0 - 1$. Choose $y_j \in Q_j \setminus P_1$. Also choose $x \in I \setminus P_1$. Set $t = y_1 y_2 \cdots y_s x$. Then $t \in Q_j$ for all j. We note that

$$(H^{i_0-1}_{JA}(A))_t = 0.$$

Also note that as $t \notin P_1$, we get $t \notin JA$.

Thus we have a commutative diagram

$$\begin{tikzcd}
0 \arrow[r] & H^{i_0}_{IA+(t)}(A) \arrow[r, "\cong"] \arrow[d, "\theta^{i_0}_{IA,JA+(t)}"] & H^{i_0}_{IA}(A) \arrow[r] \arrow[d, "\theta^{i_0}_{IA,JA}"] & 0 \\
0 \arrow[r] & H^{i_0}_{JA+(t)}(A) \arrow[r, "\theta^{i_0}_{JA+(t),JA}"] & H^{i_0}_{JA}(A) \arrow[r, "\pi"] & (H^{i_0}_{JA}(A))_t
\end{tikzcd}$$

We note that as $t \in I$, we get $t \in \operatorname{rad}(A)$. Let $\mathfrak{m}$ be a maximal ideal in A and let $(0\colon a_{\mathfrak{m}}) = \mathfrak{m}$, where $a_{\mathfrak{m}} \in \text{image}\, \theta^{i_0}_{IA,JA}$. Clearly $\pi(a_{\mathfrak{m}}) = 0$. Thus it follows that $\mathfrak{m} \in \operatorname{Ass}_A(H^{i_0}_{JA+(t)}(A))$. A simple diagram chase shows that, in fact,

$$\mathfrak{m} \in \operatorname{Ass}_A(\text{image}\, \theta^{i_0}_{IA,JA+(t)}).$$

Thus $(IA, JA + (t)) \in \mathcal{C}^{i_0}_A$. This contradicts the maximality of (IA, JA) in $\mathcal{C}^{i_0}_A$. Therefore height $JA = i_0 - 1$. □

As an immediate consequence, we get this:

Proof of Theorem 1.5. (i) $\Longrightarrow$ (ii): This is obvious.

(ii) $\Longrightarrow$ (i): By Theorem 1.4, it follows that Lyubeznik's conjecture holds for all excellent regular domains of dimension $\leq d$ containing an uncountable field of characteristic zero. By results in Section 3, it follows that Lyubeznik's conjecture holds for all excellent regular domains of dimension $\leq d$ containing a field of characteristic zero. By Remark 2.16, the result holds for all excellent regular rings of dimension $\leq d$ containing a field of characteristic zero. □

6. Proof of Theorem 1.3

In this section we prove Theorem 1.3. We will need the following result (see [Matsumura 1989, proof of Theorem 31.1]).

Lemma 6.1. *Let $\{Q_n\}_{n\geq 1}$ be an infinite family of primes in a Noetherian ring T, and let P be another prime ideal in T with $P \subseteq Q_n$ for all n. Suppose that* height$(Q_n/P) = 1$ *for all n. Then*

$$\bigcap_{n\geq 1} Q_n = P.$$

We also need Corollary 7.8. As the techniques to prove it are totally different, we postpone the proof of Corollary 7.8 to the next section.

Proof of Theorem 1.3. By Remark 2.16, we may assume that R is a domain. By results in Section 3, we may assume that R contains an uncountable field of characteristic zero.

Suppose, if possible, that for some ideal K of R, we have $\operatorname{Ass}_R(H^{d-1}_K(R))$ is an infinite set. So $(K, K) \in \mathcal{C}^{d-1}_R$. Thus $\mathcal{C}^{d-1}_R \neq \varnothing$. Let (I, J) be a maximal element in $\mathcal{C}^{d-1}_R$. We now do the construction as in Theorem 1.4. Then in the ring $A = S^{-1}R$, we have

(a) (IA, JA) is a maximal element in $\mathcal{C}^{d-1}_A$,

(b) $\operatorname{height} IA = d - 1$,

(c) $IA = P_1 \cap P_2 \cap \cdots \cap P_r$, where P_i is a prime in A of height $d - 1$,

(d) $\operatorname{Ass}_A(\operatorname{image} \theta^{d-1}_{IA,JA}) \supseteq \text{m-Spec}(A)$,

(e) $\text{m-Spec}(A)$ is a countably infinite set.

We now note that $\dim A = d$. This is so since for any Noetherian ring T of dimension n and an ideal L of T, we have $\operatorname{Ass}_T(H^n_L(T))$ is a finite set. Thus $\dim A \neq d - 1$. Furthermore, by Grothendieck's vanishing theorem, it is not possible that $\dim A < d - 1$. Thus $\dim A = d$. Again by Theorem 1.4, we get that $\operatorname{height} \mathfrak{m} = d$ for all maximal ideals of A. We note that $IA \subseteq \operatorname{rad}(A)$.

Claim 1: There exists a localization B of A such that

(1) $\dim B = d$,

(2) $\operatorname{height} \mathfrak{m} = d$ for all maximal ideals $\mathfrak{m}$ of B,

(3) $\text{m-Spec}(B)$ is a countably infinite set,

(4) $\operatorname{Ass}_B(\operatorname{image} \theta^{d-1}_{IB,JB}) \supseteq \text{m-Spec}(B)$,

(5) IB is a prime ideal of height $d - 1$,

(6) $IB = \operatorname{rad}(B)$.

To prove the claim, we recall that $IA = P_1 \cap P_2 \cap \cdots \cap P_r$, where P_i is a prime ideal in A of height $d - 1$. We consider two cases.

Case 1: $r = 1$. Then $IA = P_1$ is a prime ideal of height $d - 1$. Also $P_1 \subseteq \mathfrak{m}$ for all maximal ideals $\mathfrak{m}$ of A. As $\operatorname{height} \mathfrak{m} = d$ for all maximal ideals of A and as $\text{m-Spec}(A)$ is a countably infinite set, by Lemma 6.1 we get $IA = P_1 = \operatorname{rad}(A)$. Thus we can take B to be A.

Case 2: $r \geq 2$. Consider the sets

$$Y_i = \{\mathfrak{m} \mid \mathfrak{m} \in \text{m-Spec}(A) \text{ and } \mathfrak{m} \supseteq P_i\}.$$

As $IA \subseteq \operatorname{rad}(A)$, we get $Y_1 \cup Y_2 \cup \cdots \cup Y_r = \text{m-Spec}(A)$. So there exists i such that Y_i is an infinite set. After relabeling, we may assume $i = 1$. Set

$$T_1 = A \setminus \bigcup_{\mathfrak{m} \in Y_1} \mathfrak{m} \quad \text{and} \quad A_1 = T_1^{-1} A.$$

Let $IA_1 = Q_1 \cap Q_2 \cap \cdots \cap Q_s$ be an *irredundant* primary decomposition of IA_1 with $Q_1 = P_1 A_1$. (Note as IA_1 is a radical ideal, all Q_i are prime ideals.) We note that height $Q_1 = d - 1$. Furthermore, height $\mathfrak{n} = d$ for each maximal ideal of A_1. Also by Proposition 4.4, $\sharp\, \text{m-Spec}(A_1) = \sharp Y_1 = \infty$. As $Q_1 \subseteq \mathfrak{m}$ for each maximal ideal of A_1, by Lemma 6.1 we get $Q_1 = \operatorname{rad}(A_1)$.

We also note that $\operatorname{Ass}_{A_1}(\text{image } \theta^{d-1}_{IA_1, JA_1}) \supseteq \text{m-Spec}(A_1)$ and $IA_1 \subseteq \operatorname{rad}(A_1)$. For $j \geq 2$, consider the sets

$$Y'_j = \{\mathfrak{m} \mid \mathfrak{m} \in \text{m-Spec}(A_1) \text{ and } \mathfrak{m} \supseteq Q_j\}.$$

Claim 2: Y'_j is a finite set for all $j \geq 2$.

Suppose, if possible, that Y_j is an infinite set for some j. Then by Lemma 6.1, we get

$$Q_j = \bigcap_{\mathfrak{m} \in Y'_j} \mathfrak{m} \supseteq \operatorname{rad}(A_1) = Q_1.$$

This contradicts the fact that $Q_1 \cap \cdots \cap Q_s$ is an *irredundant* primary decomposition of IA_1. Thus Claim 2 is proved.

Now set

$$\Lambda = \text{m-Spec}(A_1) \setminus \bigcup_{j \geq 2} Y'_j, \quad T_2 = A_1 \setminus \bigcup_{\mathfrak{m} \in \Lambda} \mathfrak{m} \quad \text{and} \quad B = T_2^{-1} A_1.$$

It is easy to prove that B satisfies all the assertions in Claim 1.

We now note that B is excellent. So $\operatorname{Reg}(B/IB)$ is an open set. As IB is a prime ideal, we get that $\operatorname{Reg}(B/IB)$ is *nonempty*. Since $\operatorname{rad}(B/IB) = 0$, it follows that there exists a maximal ideal $\mathfrak{m}$ of B with $(B/IB)_\mathfrak{m}$ a regular local ring. As B is a regular ring, $B_\mathfrak{m}$ is a regular ring. Also note that height $IB_\mathfrak{m} = d - 1$ and height $\mathfrak{m} B_\mathfrak{m} = d$. As $B_\mathfrak{m}/IB_\mathfrak{m}$ is regular, we get that $IB_\mathfrak{m}$ is generated by part of a regular system of parameters. Say $IB_\mathfrak{m} = (x_1, \ldots, x_{d-1})$ and say $\mathfrak{m} B_\mathfrak{m} = (x_1, \ldots, x_{d-1}, x_d)$. Let k be the residue field of $B_\mathfrak{m}$. Set $C = \widehat{B_\mathfrak{m}}$. Note that $C = k[[x_1, \ldots, x_d]]$ and $IC = (x_1, \ldots, x_{d-1})$. Let $\mathfrak{n} = (x_1, \ldots, x_d)C$ be the maximal ideal of C. By our assumption, $\mathfrak{n} \in \operatorname{Ass}_C(\text{image } \theta^{d-1}_{IC, JC})$. However, this contradicts Corollary 7.8. Thus $\operatorname{Ass}_R(H^{d-1}_K(R))$ is a finite set for all ideals K of R. □

7. Simplicity of a local cohomology module

The references for this section are [Björk 1979; Lyubeznik 1993]. Let $\mathcal{O} = K[[X_1, \ldots, X_n]]$, where K is a field of characteristic zero. Let D be the ring of K-linear differential operators on D. By the work of Lyubeznik, it is known that if I is an ideal in $\mathcal{O}$ then $H^i_I(\mathcal{O})$ are finitely generated D-modules for all $i \geq 0$. The main goal of this section is to prove that if $P_g = (X_1, \dots, X_g)$ for $1 \leq g \leq n$ then $H^g_{P_g}(\mathcal{O})$ is simple as a D-module. This result is known in the polynomial case and is easy to prove. A few experts we asked also claimed that this fact is known even in the case of power series rings. However, none of them had a reference for this result. As this result is crucial for us; we give a proof of this fact.

7.1. Consider the $\sum$-filtration where $\sum_k = \{Q \in D \mid Q = \sum_{|\alpha| \leq k} q_\alpha(X)\partial^\alpha\}$ is the set of differential operators of order $\leq k$. The associated graded ring $\operatorname{gr} D = \sum_0 \oplus \sum_1 / \sum_0 \oplus \cdots$ is isomorphic to the polynomial ring $\mathcal{O}[\zeta_1, \dots, \zeta_n]$, where ζ_i is the image of ∂_i in $\sum_1 / \sum_0$.

7.2. Let M be a finitely generated D-module. We consider filtrations $\mathcal{F}$ of K-linear subspaces of M with the property that $\mathcal{F}_i = 0$ for $i < 0$, $\mathcal{F}_i \subseteq \mathcal{F}_{i+1}$, $M = \bigcup_{i \geq 0} \mathcal{F}_i$ and $\sum_i \mathcal{F}_j \subseteq \mathcal{F}_{i+j}$. Note that $\operatorname{gr}_{\mathcal{F}} M = \bigoplus_{i \geq 0} \mathcal{F}_i / \mathcal{F}_{i-1}$ is a $\operatorname{gr} D$-module. We say $\mathcal{F}$ is a *good filtration* of M if $\operatorname{gr}_{\mathcal{F}} M$ is finitely generated as a $\operatorname{gr} D$-module. We note that every finitely generated D-module M has a good filtration.

7.3. Let $\mathcal{M}$ be the unique graded maximal ideal of $\operatorname{gr} D$. Let $\mathcal{F}$ be a good filtration on M. Then note that $\dim \operatorname{gr}_{\mathcal{F}} M = \dim(\operatorname{gr}_{\mathcal{F}} M)_{\mathcal{M}}$; see [Bruns and Herzog 1993, Theorem 1.5.8]. Let $e(\operatorname{gr}_{\mathcal{F}} M)$ be the multiplicity of $(\operatorname{gr}_{\mathcal{F}} M)_{\mathcal{M}}$ with respect to the maximal ideal $\mathcal{M}(\operatorname{gr} D)_{\mathcal{M}}$ of the regular local ring $(\operatorname{gr} D)_{\mathcal{M}}$. Let $\mathcal{F}, \mathcal{G}$ be two good filtrations on M. Then by [Björk 1979, Lemma 6.2, Chapter 2], we get that

$$\dim \operatorname{gr}_{\mathcal{F}} M = \dim \operatorname{gr}_{\mathcal{G}} M \quad \text{and} \quad e(\operatorname{gr}_{\mathcal{F}} M) = e(\operatorname{gr}_{\mathcal{G}} M).$$

Thus if $\mathcal{F}$ is a good filtration on M then we can set

$$\dim M = \dim \operatorname{gr}_{\mathcal{F}} M \quad \text{and} \quad e_{\mathcal{M}}(M) = e(\operatorname{gr}_{\mathcal{F}} M).$$

7.4. It is well known that if M is a nonzero D-module then $\dim M \geq n$. If $\dim M = n$ or if $M = 0$ then we say M is a *holonomic* D-module. By the work of Lyubeznik, it is known that if I is an ideal in $\mathcal{O}$ then $H^i_I(\mathcal{O})$ is a holonomic D-module for all $i \geq 0$.

We need the following lemma and its corollary:

Lemma 7.5. *Let M be a holonomic D-module. If $M \neq 0$ and M is not simple then $e_{\mathcal{M}}(M) \geq 2$.*

Proof. As M is not simple, it has a proper nonzero submodule K. Set $C = M/K$. Then K, C are *nonzero* holonomic D-modules.

Let $\mathcal{F}$ be a good filtration on M. Set $\overline{\mathcal{F}} =$ quotient filtration on C and let $\mathcal{G} = \{\mathcal{F}_n \cap K\}_{n \geq 0}$ be the induced filtration on K. Then we have an exact sequence of graded gr D-modules,

$$(*) \qquad 0 \to \operatorname{gr}_{\mathcal{G}} K \to \operatorname{gr}_{\mathcal{F}} M \to \operatorname{gr}_{\overline{\mathcal{F}}} C \to 0.$$

Thus $\operatorname{gr}_{\mathcal{G}} K$ and $\operatorname{gr}_{\overline{\mathcal{F}}} C$ are finitely generated gr D modules. So $\overline{\mathcal{F}}$ is a good filtration of C and $\mathcal{G}$ is a good filtration of K.

All the modules in equation $(*)$ have dimension n. Computing multiplicities we get

$$e_{\mathcal{M}}(M) = e_{\mathcal{M}}(K) + e_{\mathcal{M}}(C) \geq 2. \qquad \square$$

As an immediate corollary, we obtain this:

Corollary 7.6. *Let M be a holonomic D-module. If $e_{\mathcal{M}}(M) = 1$ then M is a simple D-module.*

The main result of this section is this:

Lemma 7.7. *Let $\mathcal{O} = K[[X_1, \dots, X_n]]$ and let $P_g = (X_1, \dots, X_g)$ for $g \geq 1$. Then $e_{\mathcal{M}}(H^g_{P_g}(\mathcal{O})) = 1$. Thus $H^g_{P_g}(\mathcal{O})$ is a simple D-module.*

Proof. We first consider the case when $g = n$. So $P_n = \mathfrak{m} = (X_1, \dots, X_n)$. In this case it is well known that $H^n_{\mathfrak{m}}(\mathcal{O}) \cong D/D\mathfrak{m} = K[\overline{\partial_1}, \dots, \overline{\partial_n}]$ as D-modules. Let $\mathcal{F}_i = \left\{\sum_{|\alpha| \leq i} a_\alpha \bar{\partial}^\alpha \mid a_\alpha \in K\right\}$. Let $Q \in \sum_i$ be a differential operator of order $\leq i$. Then notice that Q can also be written as $Q = \sum_{|\alpha| \leq i} \partial^\alpha a_\alpha$. Set $a_\alpha = c_\alpha + t_\alpha$, where $a_\alpha \in K$ and $t_\alpha \in \mathfrak{m}$. Then notice that the image of Q in $D/D\mathfrak{m}$ is $\sum_{|\alpha| \leq i} c_\alpha \bar{\partial}^\alpha$. Thus $\mathcal{F}$ is the quotient filtration of $\sum$. Therefore $\mathcal{F}$ is a good filtration on $D/D\mathfrak{m}$. Also note that

$$\operatorname{gr}_{\mathcal{F}} D/D\mathfrak{m} = \operatorname{gr} D/\mathfrak{m} \operatorname{gr} D = K[\zeta_1, \dots, \zeta_n].$$

Clearly $e_{\mathcal{M}}(D/D\mathfrak{m}) = 1$. Also note that $X_i \mathcal{F}_\nu \subseteq \mathcal{F}_{\nu-1}$ for all $\nu \geq 0$.

We now consider the case when $g < n$. Set $S = K[[X_1, \dots, X_g]]$ and $\mathfrak{n} = (X_1, \dots, X_g)$. Let D' be the ring of K-linear differential operators on S. Let $M = H^g_{\mathfrak{n}}(S) = D'/D'\mathfrak{n}$. Set $N = H^g_{P_g}(\mathcal{O}) = M \otimes_S \mathcal{O}$. We note that the D-module structure on N is given by $\partial_1, \dots, \partial_g$ acting on M and ∂_i acting on $\mathcal{O}$ for $i > g$. Also note that for $r \in \mathcal{O}$, $m \in M$ and $t \in \mathcal{O}$, we have $r \cdot (m \otimes t) = m \otimes rt$. Let $\mathcal{F}$ be the D'-filtration on M as discussed earlier. Set

$$\Omega_\nu = \left\{ \sum_{\text{finite sum}} m_\alpha \otimes r_\alpha \;\middle|\; m_\alpha \in \mathcal{F}_\nu \text{ and } r_\alpha \in \mathcal{O} \right\}.$$

Let $\xi = a\partial_1^{a_1} \cdots \partial_g^{a_g} \cdots \partial_n^{a_n} \in \sum_i$. So $\sum_{k=1}^n a_k \leq i$. In particular, $\sum_{k=1}^g a_k \leq i$. Let $m \otimes r \in \Omega_\nu$. Then

$$\xi \cdot (m \otimes r) = \left(a(\partial_1^{a_1} \cdots \partial_g^{a_g} m)\right) \otimes \left(a(\partial_{g+1}^{a_{g+1}} \cdots \partial_n^{a_n} r)\right) \in \Omega_{\nu+i}.$$

Thus Ω is a filtration on N compatible with the $\sum$-filtration on D. Therefore $\operatorname{gr}_\Omega N$ is a $\operatorname{gr} D$-module.

We first assert that $\operatorname{gr}_\Omega N$ is generated by $\overline{1 \otimes 1} \in \Omega_0$. Let $\xi = m \otimes r \in \Omega_\nu$. Then note that if $m = \sum_\alpha a_\alpha \bar{\partial}^\alpha$ then $m = Q \cdot 1$, where $Q = \sum_\alpha a_\alpha \partial^\alpha$. Thus $\xi = rQ \cdot (1 \otimes 1)$. Therefore we have an obvious surjective map $\varphi\colon \operatorname{gr} D \to \operatorname{gr}_\Omega N$ which takes 1 to $\overline{1 \otimes 1}$. Thus Ω is a good filtration on N.

Let $\xi = [m \otimes r] \in \Omega_\nu / \Omega_{\nu-1}$. For $i \le g$, we have $X_i m \in \mathcal{F}_{\nu-1}$. Thus $X_i(m \otimes r) = m \otimes X_i r = X_i m \otimes r \in \Omega_{\nu-1}$. Therefore $X_i \xi = 0$ for all $i \le g$. Also notice that for $i > g$, we have $\zeta_i \xi = [m \otimes \partial_i(r)]$. But $m \otimes \partial_i r \in \Omega_\nu$. As degree $\zeta_i = 1$, we get that $\zeta_i \xi = 0$. Thus φ factors to a surjective map $\bar{\varphi}\colon \mathcal{O}/(X_1, \dots, X_g)[\zeta_1, \dots, \zeta_g] \to \operatorname{gr}_\Omega N$.

As $\operatorname{gr}_\Omega N$ has dimension n, it follows that $\bar{\varphi}$ is in fact an isomorphism. Thus $\operatorname{gr}_\Omega N \cong \mathcal{O}/(X_1, \dots, X_g)[\zeta_1, \dots, \zeta_g]$ and clearly it has multiplicity one. $\square$

We need the following result in the proof of our main result.

Corollary 7.8. *Let $\mathcal{O} = K[[X_1, \dots, X_n]]$ and let $P = (X_1, \dots, X_g)$, where $g < n$. Set $\mathfrak{m} = (X_1, \dots, X_n)$. Let $J \subseteq P$. Let $\theta^g\colon H^g_P(\mathcal{O}) \to H^g_J(\mathcal{O})$ be the natural map. Then* $\operatorname{Ass} \operatorname{image} \theta^g$ *is either empty or equal to $\{P\}$. In particular* $\mathfrak{m} \notin \operatorname{Ass} \operatorname{image} \theta^g$.

Proof. We note that any injective D-module is also an injective $\mathcal{O}$-module. Thus we can take an injective resolution of $\mathcal{O}$ as a D-module and note that computing θ^i with this resolution, we get that $\theta^i\colon H^i_P(\mathcal{O}) \to H^i_J(\mathcal{O})$ is D-linear. By Lemma 7.7, we get that $N = H^g_P(\mathcal{O})$ is a simple D-module. Thus image $\theta^g \cong N$ or it is zero. The result follows. $\square$

8. Small dimensions

In this section we prove several elementary results regarding local cohomology modules of regular rings of dimension ≤ 4. We note that the results for dimension ≤ 3 are probably well known to the experts. However, as we are unable to find a reference, we give a proof for these cases too.

We first prove the following general result.

Proposition 8.1. *Let R be a regular ring of dimension d. Let I be an ideal in R. Then*

$$\operatorname{Ass}_R(H^d_I(R)) = \{P \mid P \in \operatorname{Min} R/I \text{ and } \operatorname{height} P = d\}.$$

Proof. We may assume I is a radical ideal. Suppose $P \in \operatorname{Ass}_R(H^d_I(R))$. Then as height $P \ge d = \dim R$, we get that P is a maximal ideal of R. Note that $PR_P \in \operatorname{Ass}_{R_P}(H^d_{IR_P}(R_P))$ and thus $P\widehat{R_P} \in \operatorname{Ass}_{\widehat{R_P}}(H^d_{I\widehat{R_P}}(\widehat{R_P}))$.

Notice that $\widehat{R_P}$ is a domain and so by the Hartshorne–Lichtenbaum theorem (see [Iyengar et al. 2007, Theorem 14.1]), we get that $I\widehat{R_P}$ is a zero-dimensional ideal in $\widehat{R_P}$. It follows that IR_P is a zero -dimensional ideal in R_P. As IR_P is a radical ideal, we get that $IR_P = PR_P$. The result follows. $\square$

If M is an R-module then set

$$\operatorname{Ass}^i_R(M) = \{P \mid P \in \operatorname{Ass} M \text{ and } \operatorname{height} P = i\}.$$

As an easy corollary to Proposition 8.1, we get this:

Corollary 8.2. *Let R be a regular ring of dimension d. Let I be an ideal in R. Then*

$$\bigcup_{i\geq 0} \operatorname{Ass}^i_R(H^i_I(R)) = \operatorname{Min} R/I.$$

Proof. We may assume I is a radical ideal. Let $I = P_1 \cap \cdots \cap P_r$. Notice that $IR_{P_i} = P_i R_{P_i}$ for each i and R_{P_i} is a regular local ring of height $c_i = \operatorname{height} P_i$. Notice that $P_i \in \operatorname{Ass}^{c_i}_R(H^{c_i}_I(R))$. Thus

$$\operatorname{Min} R/I \subseteq \bigcup_{i\geq 0} \operatorname{Ass}^i_R(H^i_I(R)).$$

Conversely let $P \in \operatorname{Ass}^i_R(H^i_I(R))$. We localize at P. Then R_P is a regular local ring of dimension i. The result now follows from Proposition 8.1. □

Remark 8.3. If I is an ideal of height g then it is well known that

$$\operatorname{Ass} H^g_I(R) = \{P \mid P \supseteq I, \operatorname{height} P = g\}.$$

We now prove this:

Lemma 8.4. *Let R be a regular ring and let I be an ideal of height one. Then $\operatorname{Ass}_R H^i_I(R)$ is finite for $i = 1, 2$.*

Proof. For $i = 1$, the result follows from Remark 8.3.

For $i = 2$, we first consider the case when I has a primary decomposition $I = Q_1 \cap \cdots \cap Q_r$, where height $\sqrt{Q_i} = 1$ for all i. We claim that $H^j_I(R) = 0$ for all $j \geq 2$. Suppose this is not true. Let $P \in \operatorname{Ass} H^j_I(R)$ for some $j \geq 2$. We localize at P. We now note that IR_P is a principal ideal; see [Bruns and Herzog 1993, Exercise 2.2.28]. So $H^s_{IR_P}(R_P) = 0$ for all $s \geq 2$, a contradiction. Therefore $H^j_I(R) = 0$ for all $j \geq 2$. Thus our assertion holds in this special case.

Now let I be a general ideal of height one. Then $I = J \cap K$, where K has height ≥ 2 and J is an ideal of height one. Furthermore, J is of the special kind discussed above. So $H^j_J(R) = 0$ for $j \geq 2$. By the Mayer–Vietoris sequence (see [Iyengar et al. 2007, Theorem 15.1]) and noting that height$(J + K) \geq 3$, we have an exact sequence

$$0 \to H^2_K(R) \to H^2_I(R) \to H^3_{J+K}(R).$$

As height $K \geq 2$ and height$(J + K) \geq 3$, the result follows from Remark 8.3. □

An easy consequence of the above results is the following:

Corollary 8.5. *Let R be a regular ring of dimension $d \leq 3$. Then for any ideal I, we have* $\operatorname{Ass} H^i_I(R)$ *is a finite set for all $i \geq 0$.*

Proof. By Remark 2.16, we may assume that R is a domain. We have nothing to show for $d = 0$. The assertion for $d = 1$ follows from Proposition 8.1. For $d = 2$, the result follows from Lemma 8.4 and Remark 8.3.

Now consider the case when $d = 3$. If height $I = 1$ then the result follows from Lemma 8.4 and Proposition 8.1. If height $I = 2$, the result follows from Remark 8.3 and Proposition 8.1. If height $I = 3$, the result follows from Proposition 8.1. □

We now give a proof of Corollary 1.6.

Proof. By Corollary 8.5, we may assume $\dim R = 4$. By the results of [Huneke and Sharp 1993], the result holds when $\operatorname{char} K = p > 0$. Thus we may assume that $\operatorname{char} K = 0$. By Remark 2.16, we may assume that R is a domain.

If height $I = 1$, then the result follows from Lemma 8.4, Theorem 1.3 and Proposition 8.1. If height $I = 2$, the result follows from Remark 8.3, Theorem 1.3 and Proposition 8.1. If height $I = 3$, the result follows from Remark 8.3 and Proposition 8.1. If height $I = 4$, the result follows from Proposition 8.1. □

9. Proofs of Corollary 1.7 and Proposition 1.10

In this final section we give an application of Theorem 1.3. Throughout R will denote a regular ring of dimension d containing a field of characteristic zero.

Proof of Corollary 1.7. Suppose, if possible, that

$$\bigcup_{i \geq 0} \operatorname{Ass}^{i+1}_R(H^i_I(R)) \quad \text{is an infinite set.}$$

As $H^i_I(R) = 0$ for $i > d$, we get that $\operatorname{Ass}^{i+1}_R(H^i_I(R))$ is infinite for some $i \leq d$. By Corollary 3.7, we may assume that R contains an uncountable field. Suppose $\operatorname{Ass}^{i+1}_R(H^i_I(R)) = \{\mathfrak{p}_n\}_{n \geq 1}$. Consider the ring

$$A = S^{-1}R, \quad \text{where } S = R \setminus \bigcup_{n \geq 1} \mathfrak{p}_n.$$

Then A is a regular ring of dimension $i + 1$. Furthermore, $\operatorname{Ass}_A H^i_{IA}(A)$ is an infinite set. This contradicts Theorem 1.3. □

Recall that if $(A, \mathfrak{m})$ is a local ring then $\operatorname{Spec}^\circ(A) = \operatorname{Spec}(A) \setminus \{\mathfrak{m}\}$ considered as a subspace of $\operatorname{Spec}(A)$.

Definition 9.1. Let $(A, \mathfrak{m})$ be a local ring and let I be an ideal in A. We say $\operatorname{Spec}^\circ(A/I)$ is *absolutely connected* if for every flat local map $(A, \mathfrak{m}) \to (B, \mathfrak{n})$ with $\mathfrak{m}B = \mathfrak{n}$ and $B/\mathfrak{n}$ algebraically closed, $\operatorname{Spec}^\circ(B/IB)$ is connected.

Remark 9.2. It is easy to see that if $\operatorname{Spec}^\circ(A/I)$ is absolutely connected then it is connected.

Proof of Proposition 1.10. First assume that $P \supseteq I$, height $P = c \geq g+2$ and $\operatorname{Spec}^\circ(R_P/I_P)$ is not absolutely connected. So there is a flat extension $(B, \mathfrak{n})$ of R_P such that B is complete, $PB = \mathfrak{n}$, $B/\mathfrak{n}$ algebraically closed and $\operatorname{Spec}^\circ(B/IB)$ is disconnected. We note that B is a complete regular ring of dimension c. Also note that as $\dim R_P/I_P \geq 2$, we have that $\dim B/IB \geq 2$. By the result of Ogus [1973, Corollary 2.11], we have that $H^{c-1}_{IB}(B) \neq 0$. So $H^{c-1}_{IR_P}(R_P) \neq 0$. Therefore $P \in \operatorname{Supp} H^{c-1}_I(R)$. We claim that P is a minimal prime of $H^{c-1}_I(R)$. Suppose there exists $Q \in \operatorname{Supp} H^{c-1}_I(R)$ and $Q \subsetneq P$. Then height $Q \leq c-1$. By Grothendieck's vanishing theorem, we have height $Q \geq c-1$. So height $Q = c-1$. By our assumption, we have that $\dim R_Q/IR_Q \geq 1$. So $\dim \widehat{R_Q}/I\widehat{R_Q} \geq 1$. By the Hartshorne–Lichtenbaum theorem, we have $H^{c-1}_{I\widehat{R_Q}}(\widehat{R_Q}) = 0$. Thus

$$(H^{c-1}_I(R))_Q = H^{c-1}_{IR_Q}(R_Q) = 0,$$

a contradiction. Therefore P is a minimal prime of $H^{c-1}_I(R)$ and so belongs to $\operatorname{Ass}^c(H^{c-1}_I(R))$.

Conversely assume $P \in \operatorname{Ass}^c(H^{c-1}_I(R))$. So $H^{c-1}_{IR_P}(R_P) \neq 0$. By Remark 8.3, we get that $c-1 \geq g+1$. So $c \geq g+2$. Let $R_P \to B$ be any local flat extension with $(B, \mathfrak{n})$ complete, $PB = \mathfrak{n}$ and $B/\mathfrak{n}$ algebraically closed. We note that by our assumptions, $\dim R_P/IR_P \geq 2$. As $R_P/IR_P \to B/IB$ is flat local map with fiber a field, we have that $\dim B/IB = \dim R_P/IR_P \geq 2$. Also by faithful flatness, $H^{c-1}_{IB}(B) \neq 0$. Thus again by the same result of Ogus, $\operatorname{Spec}^\circ(B/IB)$ is disconnected. Therefore $\operatorname{Spec}^\circ(R_P/I_P)$ is not absolutely connected. $\square$

Remark 9.3. The above result is also true in characteristic p with the same proof. The reason is that Ogus' result is true in characteristic p; see [Peskine and Szpiro 1973, Theorem III.5.5].

References

[Bhatt et al. 2014] B. Bhatt, M. Blickle, G. Lyubeznik, A. K. Singh, and W. Zhang, "Local cohomology modules of a smooth $\mathbb{Z}$-algebra have finitely many associated primes", *Invent. Math.* **197**:3 (2014), 509–519. MR 3251828 Zbl 1318.13025

[Björk 1979] J.-E. Björk, *Rings of differential operators*, North-Holland Mathematical Library **21**, North-Holland, Amsterdam, 1979. MR 82g:32013 Zbl 0499.13009

[Brodmann and Sharp 1998] M. P. Brodmann and R. Y. Sharp, *Local cohomology: an algebraic introduction with geometric applications*, Cambridge Studies in Advanced Mathematics **60**, Cambridge University Press, 1998. MR 99h:13020 Zbl 0903.13006

[Brodmann et al. 2000] M. P. Brodmann, C. Rotthaus, and R. Y. Sharp, "On annihilators and associated primes of local cohomology modules", *J. Pure Appl. Algebra* **153**:3 (2000), 197–227. MR 2002b:13027 Zbl 0968.13010

[Bruns and Herzog 1993] W. Bruns and J. Herzog, *Cohen–Macaulay rings*, Cambridge Studies in Advanced Mathematics **39**, Cambridge University Press, 1993. 2nd edition published in 1998. MR 95h:13020 Zbl 0788.13005

[Huneke and Lyubeznik 1990] C. L. Huneke and G. Lyubeznik, "On the vanishing of local cohomology modules", *Invent. Math.* **102**:1 (1990), 73–93. MR 91i:13020 Zbl 0717.13011

[Huneke and Sharp 1993] C. L. Huneke and R. Y. Sharp, "Bass numbers of local cohomology modules", *Trans. Amer. Math. Soc.* **339**:2 (1993), 765–779. MR 93m:13008 Zbl 0785.13005

[Iyengar et al. 2007] S. B. Iyengar, G. J. Leuschke, A. Leykin, C. Miller, E. Miller, A. K. Singh, and U. Walther, *Twenty-four hours of local cohomology*, Graduate Studies in Mathematics **87**, American Mathematical Society, Providence, RI, 2007. MR 2009a:13025 Zbl 1129.13001

[Lyubeznik 1993] G. Lyubeznik, "Finiteness properties of local cohomology modules (an application of D-modules to commutative algebra)", *Invent. Math.* **113**:1 (1993), 41–55. MR 94e:13032 Zbl 0795.13004

[Lyubeznik 1997] G. Lyubeznik, "F-modules: applications to local cohomology and D-modules in characteristic $p > 0$", *J. Reine Angew. Math.* **491** (1997), 65–130. MR 99c:13005 Zbl 0904.13003

[Lyubeznik 2000] G. Lyubeznik, "Finiteness properties of local cohomology modules for regular local rings of mixed characteristic: the unramified case", *Comm. Algebra* **28**:12 (2000), 5867–5882. MR 2002b:13028 Zbl 0981.13008

[Lyubeznik 2002] G. Lyubeznik, "A partial survey of local cohomology", pp. 121–154 in *Local cohomology and its applications* (Guanajuato, 1999), edited by G. Lyubeznik, Lecture Notes in Pure and Applied Mathematics **226**, Dekker, New York, 2002. MR 2003b:14006 Zbl 1061.14005

[Marley 2001] T. Marley, "The associated primes of local cohomology modules over rings of small dimension", *Manuscripta Math.* **104**:4 (2001), 519–525. MR 2002h:13027 Zbl 0987.13009

[Matsumura 1989] H. Matsumura, *Commutative ring theory*, 2nd ed., Cambridge Studies in Advanced Mathematics **8**, Cambridge University Press, 1989. MR 90i:13001 Zbl 0666.13002

[Ogus 1973] A. Ogus, "Local cohomological dimension of algebraic varieties", *Ann. of Math.* (2) **98** (1973), 327–365. MR 58 #22059 Zbl 0308.14003

[Peskine and Szpiro 1973] C. Peskine and L. Szpiro, "Dimension projective finie et cohomologie locale: applications à la démonstration de conjectures de M. Auslander, H. Bass et A. Grothendieck", *Inst. Hautes Études Sci. Publ. Math.* **42** (1973), 47–119. MR 51 #10330 Zbl 0268.13008

[Rotthaus 1980] C. Rotthaus, "Zur Komplettierung ausgezeichneter Ringe", *Math. Ann.* **253**:3 (1980), 213–226. MR 82c:13023 Zbl 0444.13007

Received June 10, 2015. Revised August 20, 2015.

TONY J. PUTHENPURAKAL
DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY BOMBAY
POWAI
MUMBAI 400 076
INDIA
tputhen@gmail.com

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 282 No. 1 May 2016