Pacific Journal of Mathematics

Volume 219 No. 2 April 2005

PACIFIC JOURNAL OF MATHEMATICS

http://www.pjmath.org

Founded in 1951 by

E. F. Beckenbach (1906–1982)   F. Wolf (1904–1989)

See inside back cover or www.pjmath.org for submission instructions.

Regular subscription rate: $400.00 a year (10 issues). Special rate: $200.00 a year to individual members of supporting institutions.
Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94707-0163, U.S.A. Prior back issues are obtainable from Periodicals Service Company, 11 Main Street, Germantown, NY 12526-5635. The Pacific Journal of Mathematics is indexed by Mathematical Reviews, Zentralblatt MATH, Current Mathematical Publications and the Science Citation Index.

The Pacific Journal of Mathematics (ISSN 0030-8730) at the University of California, c/o Department of Mathematics, 969 Evans Hall, Berkeley, CA 94720-3840 is published monthly except July and August. 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 94707-0163.

PUBLISHED BY PACIFIC JOURNAL OF MATHEMATICS
at the University of California, Berkeley 94720-3840
A NON-PROFIT CORPORATION
Typeset in LaTeX

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

WEYL TRANSFORMS ASSOCIATED WITH A SINGULAR SECOND-ORDER DIFFERENTIAL OPERATOR

CYRINE BACCAR AND LAKHDAR TANNECH RACHDI

For a class of singular second-order differential operators Δ, we define and study the Weyl transforms W_σ associated with Δ, where σ is a symbol in S^m, for $m \in \mathbb{R}$. We give criteria in terms of σ for boundedness and compactness of the transform W_σ.

Introduction

Herman Weyl [1931] studied extensively the properties of pseudodifferential operators arising in quantum mechanics, regarding them as bounded linear operators on $L^2(\mathbb{R}^n)$, the space of square-integrable functions on $\mathbb{R}^n$ with respect to Lebesgue measure). M. W. Wong calls these operators, which are the subject of his book [Wong 1998], Weyl transforms.

Here we consider the second-order differential operator defined on $]0, +\infty[$ by

$$\Delta u = u'' + \frac{A'}{A} u' + \rho^2 u,$$

where A is a nonnegative function satisfying certain conditions and ρ is a nonnegative real number.

This operator plays an important role in analysis. For example, many special functions (orthogonal polynomials) are eigenfunctions of an operator of Δ type. The radial part of the Beltrami–Laplacian in a symmetric space is also of Δ type. Many aspects of such operators have been studied; we mention, in chronological order, [Chebli 1979; Trimèche 1981; Zeuner 1989; Xu 1994; Trimèche 1997; Nessibi et al. 1998]. In particular, the first two of these references investigate standard constructions of harmonic analysis, such as translation operators, convolution product, and Fourier transform, in connection with Δ.

Building on these results, we define and study the Weyl transforms associated with Δ, giving criteria for boundedness and compactness of these transforms. To obtain these results we first define the Fourier–Wigner transform associated with Δ, and establish an inversion formula.

MSC2000: 42A38, 65R10.
Keywords: Weyl transform, compact operator, Fourier–Wigner transform.

More precisely, in Section 1 we recall some properties of harmonic analysis for the operator Δ. In Section 2 we define the Fourier–Wigner transform associated with Δ, study some of its properties, and prove an inversion formula.

In Section 3 we introduce the Weyl transform W_σ associated with Δ, with σ a symbol in class S^m, for $m \in \mathbb{R}$, and we give its connection with the Fourier–Wigner transform. We prove that, for σ sufficiently smooth, W_σ is a compact operator from $L^2(d\nu)$ (the space of square-integrable functions with respect to the measure $d\nu(x) = A(x)\,dx$) into itself.

In Section 4 we define W_σ for σ in a certain space $L^p(d\nu \otimes d\gamma)$, with $p \in [1, 2]$, and we establish that W_σ is again a compact operator.

In Section 5 we define W_σ for σ in another function space, and use this to prove in Section 6 that for $p > 2$ there exists a function σ in the L^p space corresponding to that of Section 4, with the property that the Weyl transform W_σ is not bounded on $L^2(d\nu)$.

1. The operator Δ

We consider the second-order differential operator Δ defined on $]0, +\infty[$ by

$$\Delta u = u'' + \frac{A'}{A} u' + \rho^2 u,$$

where ρ is a nonnegative real number and

$$A(x) = x^{2\alpha+1} B(x), \qquad \alpha > -\tfrac{1}{2}, \tag{1–1}$$

for B a positive, even, infinitely differentiable function on $\mathbb{R}$ such that $B(0) = 1$. Moreover we assume that A and B satisfy the following conditions:

(i) A is increasing and $\lim_{x\to+\infty} A(x) = +\infty$.

(ii) $\dfrac{A'}{A}$ is decreasing and $\lim_{x\to+\infty} \dfrac{A'(x)}{A(x)} = 2\rho$.

(iii) There exists a constant $\delta > 0$ such that

$$\frac{B'(x)}{B(x)} = D(x) \exp(-\delta x) \quad \text{if } \rho = 0,$$
$$\frac{A'(x)}{A(x)} = 2\rho + D(x) \exp(-\delta x) \quad \text{if } \rho > 0,$$

where D is an infinitely differentiable function on $]0, +\infty[$, bounded and with bounded derivatives on all intervals $[x_0, +\infty[$, for $x_0 > 0$.

This operator was studied in [Chebli 1979; Nessibi et al. 1998; Trimèche 1981], and the following results have been established:

(I) For all $\lambda \in \mathbb{C}$, the equation

$$\begin{cases} \Delta u = -\lambda^2 u \\ u(0) = 1, \ u'(0) = 0 \end{cases} \tag{1–2}$$

admits a unique solution, denoted by φ_λ, with the following properties:

- φ_λ satisfies the *product formula*

$$\varphi_\lambda(x)\varphi_\lambda(y) = \int_0^\infty \varphi_\lambda(z)\, w(x, y, z) A(z)\, dz \quad \text{for } x, y \geq 0; \tag{1–3}$$

  where $w(x, y, \cdot)$ is a measurable positive function on $[0, +\infty[$, with support in $[|x-y|, x+y]$, satisfying

$$\begin{gathered} \int_0^\infty w(x, y, z) A(z)\, dz = 1, \\ w(x, y, z) = w(y, x, z) \quad \text{for } z \geq 0, \\ w(x, y, z) = w(x, z, y) \quad \text{for } z > 0; \end{gathered}$$

- for $x \geq 0$, the function $\lambda \mapsto \varphi_\lambda(x)$ is analytic on $\mathbb{C}$;
- for $\lambda \in \mathbb{C}$, the function $x \mapsto \varphi_\lambda(x)$ is even and infinitely differentiable on $\mathbb{R}$;
- $|\varphi_\lambda(x)| \leq 1$ for all $\lambda \in \mathbb{R}$ and $x \in \mathbb{R}$;
- for $x > 0$, and $\lambda > 0$ we have

$$\varphi_\lambda(x) = \frac{1}{\sqrt{B(x)}} j_\alpha(\lambda x) + A^{-1/2}(x)\theta_\lambda(x), \tag{1–4}$$

  where j_α is defined by $j_\alpha(0) = 1$ and $j_\alpha(s) = 2^\alpha \Gamma(\alpha+1) s^{-\alpha} J_\alpha(s)$ if $s \neq 0$ (with J_α the Bessel function of first kind), and the function θ_λ satisfies

$$|\theta_\lambda(x)| \leq \frac{c_1}{\lambda^{\alpha+\frac{3}{2}}} \left(\int_0^x |Q(s)|\, ds \right) \exp\left(\frac{c_2}{\lambda} \int_0^x |Q(s)|\, ds \right)$$

  with c_1, c_2 positive constants and Q the function defined on $]0, +\infty[$ by

$$Q(x) = \frac{\frac{1}{4} - \alpha^2}{x^2} + \frac{1}{4}\left(\frac{A'(x)}{A(x)}\right)^2 + \frac{1}{2}\left(\frac{A'(x)}{A(x)}\right)' - \rho^2. \tag{1–5}$$

(II) For nonzero $\lambda \in \mathbb{C}$, the equation $\Delta u = -\lambda^2 u$ has a solution Φ_λ satisfying

$$\Phi_\lambda(x) = A^{-1/2}(x) \exp(i\lambda x) V(x, \lambda),$$

with $\lim_{x\to+\infty} V(x, \lambda) = 1$. Consequently there exists a function (spectral function)

$$\lambda \mapsto c(\lambda),$$

such that

$$\varphi_\lambda = c(\lambda)\Phi_\lambda + c(-\lambda)\Phi_{-\lambda} \quad \text{for nonzero } \lambda \in \mathbb{C}.$$

Moreover there exist positive constants k_1, k_2, k_3 such that

$$k_1 |\lambda|^{\alpha+1/2} \leq |c(\lambda)|^{-1} \leq k_2 |\lambda|^{\alpha+1/2} \tag{1–6}$$

for all λ such that $\operatorname{Im}\lambda \leq 0$ and $|\lambda| \geq k_3$.

Notation. We denote by

- $d\nu(x)$ the measure defined on $[0, +\infty[$ by
$$d\nu(x) = A(x)\,dx;$$
- $L^p(d\nu)$, for $1 \leq p \leq +\infty$, the space of measurable functions on $[0, +\infty[$ satisfying
$$\|f\|_{p,\nu} := \left(\int_0^{+\infty} |f(x)|^p d\nu(x)\right)^{1/p} < +\infty \quad \text{for } 1 \leq p < +\infty,$$
$$\|f\|_{\infty,\nu} := \operatorname*{ess\,sup}_{x\in[0,+\infty[} |f(x)| < +\infty;$$
- $d\gamma(\lambda)$ the measure defined on $[0, +\infty[$ by
$$d\gamma(\lambda) = \frac{d\lambda}{2\pi\,|c(\lambda)|^2};$$
- $L^p(d\gamma)$, for $1 \leq p \leq +\infty$, the space of measurable functions on $[0, +\infty[$ satisfying $\|f\|_{p,\gamma} < +\infty$;
- $D_*(\mathbb{R})$ the space of even, infinitely differentiable functions on $\mathbb{R}$, with compact support;
- $\mathbb{H}_*(\mathbb{C})$ the space of even analytic functions on $\mathbb{C}$, rapidly decreasing of exponential type.

Definition 1.1. The *translation operator* associated with Δ is defined on $L^1(d\nu)$ by

$$\mathcal{T}_x f(y) = \int_0^{+\infty} f(z)w(x, y, z)\,d\nu(z) \quad \text{for } x, y \geq 0,$$

where w is defined in (1–3). The *convolution product* associated with Δ is defined by

$$(f * g)(x) = \int_0^{+\infty} \mathcal{T}_x f(y)\,g(y)\,d\nu(y) \quad \text{for } f, g \in L^1(d\nu).$$

Properties of translation and convolution.

- The translation operator satisfies
$$\mathcal{T}_x\varphi_\lambda(y) = \varphi_\lambda(x)\varphi_\lambda(y).$$
- Let $f \in L^1(d\nu)$. Then
$$\int_0^{+\infty} \mathcal{T}_x f(y)\,d\nu(y) = \int_0^{+\infty} f(y)\,d\nu(y) \quad \text{for } x \in [0, +\infty[$$
and
$$\|\mathcal{T}_x f\|_{1,\nu} \le \|f\|_{1,\nu}.$$
- Let $f \in L^p(d\nu)$ with $1 \le p \le +\infty$. For all $x \in [0, +\infty[$, the function $\mathcal{T}_x f$ belongs to $L^p(d\nu)$ and
$$\|\mathcal{T}_x f\|_{p,\nu} \le \|f\|_{p,\nu}.$$
- For $f, g \in L^1(d\nu)$ the function $f * g$ also lies in $L^1(d\nu)$. The convolution product is commutative and associative.
- For $f \in L^1(d\nu)$ and $g \in L^p(d\nu)$, with $1 \le p < +\infty$, the function $f * g$ lies in $L^p(d\nu)$ and we have
$$\|f * g\|_{p,\nu} \le \|f\|_{1,\nu}\|g\|_{p,\nu}.$$
- For f, g even and continuous on $\mathbb{R}$, with supports
$$\operatorname{supp} f \subset [-a, a] \quad \text{and} \quad \operatorname{supp} g \subset [-b, b],$$
the function $f * g$ is continuous on $\mathbb{R}$ and
$$\operatorname{supp}(f * g) \subset [-a-b, a+b]. \tag{1–7}$$

Definition 1.2. The *Fourier transform* associated with the operator Δ is defined on $L^1(d\nu)$ by
$$\mathcal{F}f(\lambda) = \int_0^{+\infty} f(x)\varphi_\lambda(x)\,d\nu(x) \quad \text{for } \lambda \in \mathbb{R}.$$

Properties of the Fourier transform.

- For $f \in L^1(d\nu)$ such that $\mathcal{F}f \in L^1(d\gamma)$, we have the inversion formula
$$f(x) = \int_0^{+\infty} \mathcal{F}f(\lambda)\varphi_\lambda(x)\,d\gamma(\lambda) \quad \text{for a.e. } x \in [0, +\infty[. \tag{1–8}$$
- For $f \in L^1(d\nu)$,
$$\mathcal{F}(\mathcal{T}_x f)(\lambda) = \varphi_\lambda(x)\mathcal{F}f(\lambda) \quad \text{for all } x \in [0, +\infty[\text{ and } \lambda \in \mathbb{R}.$$

- For $f, g \in L^1(d\nu)$,

$$\mathscr{F}(f * g)(\lambda) = \mathscr{F}f(\lambda)\mathscr{F}g(\lambda). \quad \text{for all } \lambda \in [0, +\infty[.$$

- $\mathscr{F}$ can be extended to an isometric isomorphism from $L^2(d\nu)$ onto $L^2(d\gamma)$. This means that

$$\|\mathscr{F}f\|_{2,\gamma} = \|f\|_{2,\nu} \quad \text{for } f \in L^2(d\nu), \tag{1–9}$$

$$\|\mathscr{F}^{-1}f\|_{2,\nu} = \|f\|_{2,\gamma} \quad \text{for } f \in L^2(d\gamma). \tag{1–10}$$

Proposition 1.3. *Let f be in $L^p(d\nu)$, with $p \in [1, 2]$. Then $\mathscr{F}f$ belongs to $L^{p'}(d\gamma)$, with $1/p + 1/p' = 1$, and*

$$\|\mathscr{F}f\|_{p',\gamma} \le \|f\|_{p,\nu}. \tag{1–11}$$

Proof. Since $|\varphi_\lambda(x)| \le 1$ for $\lambda \in \mathbb{R}$ and $x \in \mathbb{R}$, we get $\|\mathscr{F}f\|_{\infty,\gamma} \le \|f\|_{1,\nu}$. This, together with (1–9) and the Riesz–Thorin Theorem [Stein 1956; Stein and Weiss 1971], shows that for under the conditions of the proposition $\mathscr{F}f$ belongs to $L^{p'}(d\gamma)$ and satisfies (1–11). □

From [Chebli 1979], the Fourier transform $\mathscr{F}$ is a topological isomorphism from $D_*(\mathbb{R})$ onto $\mathbb{H}_*(\mathbb{C})$ (see page 204 for notation). The inverse mapping is given by

$$\mathscr{F}^{-1}f(x) = \int_0^{+\infty} f(\lambda)\varphi_\lambda(x)\, d\gamma(\lambda) \quad \text{for } x \in \mathbb{R}. \tag{1–12}$$

2. Fourier–Wigner transform associated with Δ

Definition 2.1. The Fourier–Wigner transform associated with the operator Δ is the mapping V defined on $D_*(\mathbb{R}) \times D_*(\mathbb{R})$ by

$$V(f, g)(x, \lambda) = \int_0^{+\infty} f(y)\mathscr{T}_x g(y)\varphi_\lambda(y)\, d\nu(y) \quad \text{for } (x, \lambda) \in \mathbb{R} \times \mathbb{R}.$$

Remark. The transform V can also be written in the forms

$$V(f, g)(x, \lambda) = \mathscr{F}(f\mathscr{T}_x g)(\lambda) = \varphi_\lambda f * g(x). \tag{2–1}$$

Notation. We denote by

- $D_*(\mathbb{R}^2)$ the space of infinitely differentiable functions on $\mathbb{R}^2$, even with respect to each variable, with compact support;
- $S_*(\mathbb{R}^2)$ the space of infinitely differentiable functions on $\mathbb{R}^2$, even with respect to each variable, rapidly decreasing together with all their derivatives;

- $L^p(d\nu \otimes d\nu)$, for $1 \le p \le +\infty$, the space of measurable functions on the product $[0, +\infty[\times [0, +\infty[$ satisfying

$$\| f \|_{p,\nu\otimes\nu} := \left(\int_0^{+\infty} \int_0^{+\infty} | f(x, y)|^p \, d\nu(x) \, d\nu(y) \right)^{1/p} < +\infty \quad \text{for } 1 \le p < +\infty,$$

$$\| f \|_{\infty,\nu\otimes\nu} := \operatorname*{ess\,sup}_{x,y\in[0,+\infty[} | f(x, y)| < +\infty;$$

- $L^p(d\nu \otimes d\gamma)$, for $1 \le p \le +\infty$, the space similarly defined (with $d\nu(x)\, d\gamma(y)$ in the integrand).

Proposition 2.2. (i) *The Fourier–Wigner transform V is a bilinear mapping from $D_*(\mathbb{R}) \times D_*(\mathbb{R})$ into $S_*(\mathbb{R}^2)$.*

(ii) *For $p \in]1, 2]$ and p' such that $1/p + 1/p' = 1$, we have*

$$\| V(f, g)\|_{p',\nu\otimes\gamma} \le \| f \|_{p,\nu} \, \| g\|_{p',\nu}.$$

The transform V can be extended to a continuous bilinear operator, denoted also by V, from $L^p(d\nu) \times L^{p'}(d\nu)$ into $L^{p'}(d\nu \otimes d\gamma)$.

Proof. (i) Let F be the function defined on $\mathbb{R}^2$ by $F(x, y) = f(y)\mathcal{T}_x g(y)$. It's clear that $F \in D_*(\mathbb{R}^2)$, and we have

$$V(f, g)(x, \lambda) = I \otimes \mathcal{F}(F)(x, \lambda),$$

where I is the identity operator. This and the fact that $\mathcal{F}$ is a topological isomorphism from $D_*(\mathbb{R})$ onto $\mathbb{H}_*(\mathbb{C})$ imply (i).

(ii) This follows from the first equality in (2–1) together with Proposition 1.3, Minkowski's inequality for integrals [Folland 1984, p.186], and the fact that

$$\|\mathcal{T}_x g\|_{p',\nu} \le \|g\|_{p',\nu} \quad \text{for } x \in \mathbb{R}. \qquad \square$$

Theorem 2.3. *For $f, g \in D_*(\mathbb{R})$, we have*

$$\mathcal{F} \otimes \mathcal{F}^{-1} \left(V(f, g)\right) (\mu, \lambda) = \varphi_\mu(\lambda) \, f(\lambda) \mathcal{F}g(\mu) \quad \textit{for } \mu, \lambda \in \mathbb{R}.$$

Proof. Using Definition 2.1 and Fubini's Theorem we have, for all $\mu, \lambda \in \mathbb{R}$,

$$\begin{aligned}\mathcal{F} \otimes \mathcal{F}^{-1} \left(V(f, g)\right) (\mu, \lambda) &= \int_0^{+\infty}\int_0^{+\infty} V(f, g)(x, y)\, \varphi_\mu(x) \varphi_y(\lambda)\, d\nu(x)\, d\gamma(y) \\ &= \int_0^{+\infty}\int_0^{+\infty} \mathcal{F}(f \mathcal{T}_x g)(y)\, \varphi_\mu(x) \varphi_y(\lambda)\, d\nu(x)\, d\gamma(y) \\ &= \int_0^{+\infty} \varphi_\mu(x) \left(\int_0^{+\infty} \mathcal{F}(f \mathcal{T}_x g)(y) \varphi_y(\lambda)\, d\gamma(y) \right) d\nu(x).\end{aligned}$$

From (1–8) we deduce

$$\begin{aligned}\mathscr{F}\otimes\mathscr{F}^{-1}\,(V(f,g))\,(\mu,\lambda) &= \int_0^{+\infty}\varphi_\mu(x)\,f(\lambda)\,\mathscr{T}_x g(\lambda)\,d\nu(x)\\ &= f(\lambda)\,\mathscr{F}(\mathscr{T}_\lambda g)(\mu) = f(\lambda)\,\varphi_\mu(\lambda)\,\mathscr{F}g(\mu). \qquad\square\end{aligned}$$

Corollary 2.4. *For all* $f, g \in D_*(\mathbb{R})$, *we have*

$$\begin{aligned}&\int_0^{+\infty}\mathscr{F}\otimes\mathscr{F}^{-1}\,(V(f,g))\,(\mu,\lambda)\,d\nu(\lambda) = \mathscr{F}f(\mu)\,\mathscr{F}g(\mu) \quad \textit{for } \mu\in\mathbb{R},\\ &\int_0^{+\infty}\mathscr{F}\otimes\mathscr{F}^{-1}\,(V(f,g))\,(\mu,\lambda)\,d\gamma(\mu) = f(\lambda)\,g(\lambda) \qquad \textit{for } \lambda\in\mathbb{R}.\end{aligned}$$

Proof. Theorem 2.3 gives

$$\begin{aligned}\int_0^{+\infty}\mathscr{F}\otimes\mathscr{F}^{-1}\,(V(f,g))\,(\mu,\lambda)\,d\nu(\lambda) &= \int_0^{+\infty}\varphi_\mu(\lambda)\,f(\lambda)\mathscr{F}g(\mu)\,d\nu(\lambda)\\ &= \mathscr{F}f(\mu)\,\mathscr{F}g(\mu) \quad \text{for } \mu\in\mathbb{R},\\ \int_0^{+\infty}\mathscr{F}\otimes\mathscr{F}^{-1}\,(V(f,g))\,(\mu,\lambda)\,d\gamma(\mu) &= \int_0^{+\infty}\varphi_\mu(\lambda)\,f(\lambda)\mathscr{F}g(\mu)\,d\gamma(\mu)\\ &= f(\lambda)\int_0^{+\infty}\varphi_\mu(\lambda)\mathscr{F}g(\mu)\,d\gamma(\mu)\\ &= f(\lambda)\,g(\lambda) \quad \text{for } \lambda\in\mathbb{R}. \qquad\square\end{aligned}$$

Theorem 2.5. *Let* $f, g \in L^1(d\nu)\cap L^2(d\nu)$ *be such that* $c = \int_0^{+\infty} g(x)\,d\nu(x) \neq 0$. *Then*

$$\mathscr{F}f(\lambda) = \frac{1}{c}\int_0^{+\infty} V(f,g)(x,\lambda)\,d\nu(x) \quad \textit{for } \lambda\in\mathbb{R}.$$

Proof. From Definition 2.1, we have

$$\int_0^{+\infty} V(f,g)(x,\lambda)\,d\nu(x) = \int_0^{+\infty}\left(\int_0^{+\infty} f(y)\,\mathscr{T}_x g(y)\,\varphi_\lambda(y)\,d\nu(y)\right)d\nu(x)$$

for all $\lambda\in\mathbb{R}$. The result follows from Fubini's Theorem and the equality

$$\int_0^{+\infty}\mathscr{T}_x g(y)\,d\nu(y) = \int_0^{+\infty} g(x)\,d\nu(x) = c. \qquad\square$$

Corollary 2.6. *With the hypothesis of Theorem 2.5, if* $\mathscr{F}f \in L^1(d\gamma)$, *we have the following inversion formula for the Fourier–Wigner transform* V:

$$f(x) = \frac{1}{c}\int_0^{+\infty}\varphi_\mu(x)\left(\int_0^{+\infty} V(f,g)(y,\mu)\,d\nu(y)\right)d\gamma(\mu) \quad \textit{for a.e. } x\in\mathbb{R}.$$

3. The Weyl transform associated with Δ

We now introduce the Weyl transform and relate it to the Fourier–Wigner transform. To do this, we must define the class of pseudodifferential operators [Wong 1998].

Definition 3.1. Let $m \in \mathbb{R}$. We define S^m to be the set of all infinitely differentiable functions σ on $\mathbb{R} \times \mathbb{R}$, even with respect to each variable, and such that for all $p, q \in \mathbb{N}$, there exists a positive constant $C_{p,q,m}$ satisfying

$$\left|\left(\frac{\partial}{\partial x}\right)^p \left(\frac{\partial}{\partial y}\right)^q \sigma(x, y)\right| \le C_{p,q,m}(1 + y^2)^{m-q}.$$

Definition 3.2. For $m \in \mathbb{R}$ and $\sigma \in S^m$, we define the operator H_σ on $D_*(\mathbb{R}) \times D_*(\mathbb{R})$ by

$$H_\sigma(f, g)(\lambda) = \int_0^{+\infty}\left(\int_0^{+\infty} \sigma(x, y)\varphi_y(\lambda)V(f, g)(x, y)\,d\nu(x)\right) d\gamma(y), \tag{3–1}$$

for all $\lambda \in \mathbb{R}$, and we put

$$\mathbb{H}_\sigma(f, g) = H_\sigma(f, g)(0). \tag{3–2}$$

Proposition 3.3. *Define $\sigma \in S^m$ by $\sigma(x, y) = -y^2$ for $x, y \in \mathbb{R}$. Then, for all $f, g \in D_*(\mathbb{R})$, we have*

$$H_\sigma(f, g)(\lambda) = c\,\Delta f(\lambda) \quad \text{for } \lambda \in \mathbb{R},$$

where $c = \int_0^{+\infty} g(x)\,d\nu(x)$.

Proof. From (3–1), we have

$$H_\sigma(f, g)(\lambda) = \int_0^{+\infty}\left(\int_0^{+\infty} -y^2\varphi_y(\lambda)V(f, g)(x, y)\,d\nu(x)\right) d\gamma(y) \text{for } \lambda \in \mathbb{R}.$$

Using Definition 2.1 we obtain

$$H_\sigma(f, g)(\lambda) = \int_0^{+\infty}\left(\int_0^{+\infty} -y^2\varphi_y(\lambda)\left(\int_0^{+\infty} f(z)\,\mathcal{T}_x g(z)\,\varphi_y(z)\,d\nu(z)\right) d\nu(x)\right) d\gamma(y)$$

for $\lambda \in \mathbb{R}$. From Fubini's Theorem, we get

$$\begin{aligned} H_\sigma(f, g)(\lambda) & \\ &= \int_0^{+\infty} -y^2\varphi_y(\lambda)\left(\int_0^{+\infty} f(z)\varphi_y(z)\left(\int_0^{+\infty} \mathcal{T}_z g(x)\,d\nu(x)\right) d\nu(z)\right) d\gamma(y) \\ &= c\int_0^{+\infty} -y^2\varphi_y(\lambda)\left(\int_0^{+\infty} f(z)\varphi_y(z)\,d\nu(z)\right) d\gamma(y) \\ &= c\int_0^{+\infty} -y^2\varphi_y(\lambda)\,\mathcal{F}f(y)\,d\gamma(y). \end{aligned}$$

But, for all $y \in \mathbb{R}$, $-y^2 \mathscr{F} f(y) = \mathscr{F}(\Delta f)(y)$. We complete the proof using the inversion formula (1–8). □

Definition 3.4. Let $\sigma \in S^m$; $m < -\alpha - 1$. The *Weyl transform* associated with Δ is the mapping W_σ defined on $D_*(\mathbb{R})$ by

$$W_\sigma(f)(\lambda) = \int_0^{+\infty} \left(\int_0^{+\infty} \varphi_y(\lambda) \sigma(x, y) \mathscr{T}_\lambda f(x) \, d\nu(x) \right) d\gamma(y) \quad \text{for } \lambda \in \mathbb{R}.$$

Notation. We denote by

- $S_*(\mathbb{R})$ the space of even, infinitely differentiable functions on $\mathbb{R}$, rapidly decreasing together with all their derivatives.
- $S_*^2(\mathbb{R}) = \varphi_0 S_*(\mathbb{R})$, where φ_0 is the solution of (1–2) with $\lambda = 0$.

For $\rho = 0$ these two spaces coincide [Trimèche 1997]. The Fourier transform $\mathscr{F}$ is a topological isomorphism from $S_*^2(\mathbb{R})$ onto $S_*(\mathbb{R})$, whose inverse is given by (1–12).

Lemma 3.5. *For $\sigma \in D_*(\mathbb{R}^2)$, the function k defined by*

$$k(x, y) = \int_0^{+\infty} \varphi_\lambda(x) \mathscr{T}_x \big(\sigma(\,\cdot\,, \lambda) \big)(y) \, d\gamma(\lambda) \quad \textit{for } x, y \in \mathbb{R}$$

belongs to $L^p(d\nu \otimes d\nu)$, for all $p \in [2, +\infty[$.

Proof. The defining equation of k can be rewritten $k(x, y) = \mathscr{T}_x(G(\,\cdot\,, x))(y)$, where

$$G(x, y) = I \otimes \mathscr{F}^{-1}(\sigma)(x, y) \quad \text{for } x, y \in \mathbb{R},$$

for I the identity operator. It follows that, for all $p \in [2, +\infty[$,

$$\begin{aligned} \int_0^{+\infty} \int_0^{+\infty} |k(x, y)|^p d\nu(x) \, d\nu(y) &= \int_0^{+\infty} \left(\int_0^{+\infty} \big| \mathscr{T}_x(G(\,\cdot\,, x)(y)) \big|^p d\nu(y) \right) d\nu(x) \\ &\le \int_0^{+\infty} \left(\int_0^{+\infty} |G(y, x)|^p d\nu(y) \right) d\nu(x) \\ &\le \int_0^{+\infty} \left(\int_0^{+\infty} \big| I \otimes \mathscr{F}^{-1}(\sigma)(y, x) \big|^p d\nu(y) \right) d\nu(x). \end{aligned}$$

We distinguish two cases, $p = 2$ and $p \in]2, +\infty[$, the case $p = +\infty$ being trivial. For $p = 2$,

$$\int_0^{+\infty} \int_0^{+\infty} |k(x, y)|^2 \, d\nu(x) \, d\nu(y) \le \int_0^{+\infty} \left(\int_0^{+\infty} \big| \mathscr{F}^{-1}(\sigma(x, \cdot\,)(y)) \big|^2 \, d\nu(x) \right) d\nu(y).$$

From (1–10) we deduce that

$$\int_0^{+\infty} \int_0^{+\infty} |k(x, y)|^2 \, d\nu(x) \, d\nu(y) \le \int_0^{+\infty} \left(\int_0^{+\infty} |\sigma(y, x)|^2 \, d\gamma(y) \right) d\nu(y) < +\infty,$$

because σ belongs to $D_*(\mathbb{R}^2)$. The case $p \in]2, +\infty[$ is more complex. From the hypotheses on Δ, we deduce that, as $x \to +\infty$,

$$A(x) \sim \begin{cases} x^{2\alpha+1} & \text{if } \rho = 0, \\ \exp(2\rho x) & \text{if } \rho > 0. \end{cases} \tag{3–3}$$

- For $\rho = 0$, recall that $\mathscr{F}$ is an isomorphism from $S_*(\mathbb{R})$ onto itself. Thus $I \otimes \mathscr{F}^{-1}(\sigma)$ belongs to $S_*(\mathbb{R}^2)$, and the asymptotics (3–3) implies

$$\int_0^{+\infty}\int_0^{+\infty} |k(x, y)|^p d\nu(x)\, d\nu(y) \leq \int_0^{+\infty}\left(\int_0^{+\infty} \left|I \otimes \mathscr{F}^{-1}(\sigma)(y, x)\right|^p d\nu(x)\right) d\nu(y) < +\infty. \tag{3–4}$$

- For $\rho > 0$, we have from [Trimèche 1997, p. 99]

$$|\varphi_\lambda(x)| \leq \varphi_0(x) \leq m(1 + x) \exp(-\rho x) \quad \text{for all } \lambda \in \mathbb{R} \text{ and } x \geq 0,$$

  where m is a positive constant. Then

$$\left|I \otimes \mathscr{F}^{-1}(\sigma)(y, x)\right| \leq m(1 + x) \exp(-\rho x) \int_0^{+\infty} |\sigma(y, z)|\, d\nu(z).$$

  Since σ belongs to $D_*(\mathbb{R}^2)$, there exists a positive constant M such that

$$\int_0^{+\infty} |\sigma(y, z)|\, d\nu(z) \leq M \quad \text{for } \ y \geq 0,$$

  which implies that

$$\left|I \otimes \mathscr{F}^{-1}(\sigma)(y, x)\right| \leq mM(1 + x) \exp(-\rho x).$$

  This, together with the asymptotics (3–3), implies the validity of the same bound (3–4) as in the previous case. □

Theorem 3.6. *Let $\sigma \in D_*(\mathbb{R}^2)$ and $f \in D_*(\mathbb{R})$.*

(i) $W_\sigma(f)(x) = \int_0^{+\infty} k(x, y) f(y)\, d\nu(y)$ *for all* $x \in \mathbb{R}$.

(ii) $\|W_\sigma(f)\|_{p',\nu} \leq \|k\|_{p',\nu\otimes\nu} \|f\|_{p,\nu}$ *for* $p \in [1, 2]$ *and* p' *such that* $1/p + 1/p' = 1$.

(iii) W_σ *can be extended to a bounded operator from* $L^p(d\nu)$ *into* $L^{p'}(d\nu)$. *In particular,* $W_\sigma : L^2(d\nu) \to L^2(d\nu)$ *is a Hilbert–Schmidt operator, hence compact.*

Proof. (i) From Definition 3.4, we have, for all $x \in \mathbb{R}$;

$$\begin{aligned} W_\sigma(f)(x) &= \int_0^{+\infty} \varphi_y(x)\left(\int_0^{+\infty} \sigma(z,y)\mathcal{T}_x f(z)\,d\nu(z)\right) d\gamma(y) \\ &= \int_0^{+\infty} \varphi_y(x)\left(\int_0^{+\infty} f(z)\mathcal{T}_x[\sigma(.,y)](z)\,d\nu(z)\right) d\gamma(y) \end{aligned}.$$

From Fubini's Theorem, we get, for all $x \in \mathbb{R}$,

$$\begin{aligned} W_\sigma(f)(x) &= \int_0^{+\infty} f(z)\left(\int_0^{+\infty} \varphi_y(x)\mathcal{T}_x[\sigma(.,y)](z)\,d\gamma(y)\right) d\nu(z) \\ &= \int_0^{+\infty} f(z)k(x,z)\,d\nu(z). \end{aligned}$$

(ii) Follows from (i), Hölder's inequality, and Lemma 3.5.

(iii) Since $k \in L^2(d\nu \otimes d\nu)$, the mapping

$$W_\sigma : L^2(d\nu) \longrightarrow L^2(d\nu)$$

is a Hilbert–Schmidt operator, and so compact. □

Theorem 3.7. *Let $m < -\alpha - 1$ and $\sigma \in S^m$. For all $f, g \in D_*(\mathbb{R})$,*

$$\mathbb{H}_\sigma(f,g) = \int_0^{+\infty} f(x)W_\sigma g(x)\,d\nu(x). \tag{3–5}$$

Proof. Using (3–2) and Definition 2.1 we obtain

$$\begin{aligned} \mathbb{H}_\sigma(f,g) &= \int_0^{+\infty}\left(\int_0^{+\infty} \sigma(x,y)V(f,g)(x,y)\,d\nu(x)\right) d\gamma(y) \\ &= \int_0^{+\infty}\left(\int_0^{+\infty} \sigma(x,y)\left(\int_0^{+\infty} f(\lambda)\mathcal{T}_x g(\lambda)\varphi_y(\lambda)\,d\nu(\lambda)\right) d\nu(x)\right) d\gamma(y). \end{aligned}$$

From Fubini's theorem, we get

$$\begin{aligned} \mathbb{H}_\sigma(f,g) &= \int_0^{+\infty} f(\lambda)\left(\int_0^{+\infty} \varphi_y(\lambda)\left(\int_0^{+\infty} \sigma(x,y)\mathcal{T}_x g(\lambda)\,d\nu(x)\right) d\gamma(y)\right) d\nu(\lambda) \\ &= \int_0^{+\infty} f(\lambda)W_\sigma(g)(\lambda)\,d\nu(\lambda). \end{aligned}$$ □

4. The Weyl transform with symbol in $L^p(d\nu \otimes d\gamma)$, for $1 \le p \le 2$

In this section we show using (3–5) that, if $1 \le p \le 2$, the Weyl transform with symbol in $L^p(d\nu \otimes d\gamma)$ is a compact operator.

Notation. We denote by $\mathcal{B}(L^2(d\nu))$ the $\mathbb{C}^*$-algebra of bounded operators Ψ from $L^2(d\nu)$ into itself, equipped with the norm

$$\|\Psi\|_* = \sup_{\|f\|_{2,\nu}=1} \|\Psi(f)\|_{2,\nu}\,.$$

Theorem 4.1. *Let $\langle\,\cdot\,/\,\cdot\,\rangle$ denote the inner product in $L^2(d\nu)$. There exists a unique operator $Q : L^2(d\nu\otimes d\gamma) \to \mathcal{B}(L^2(d\nu))$, whose action we denote by $\sigma \mapsto Q_\sigma$, such that*

$$\langle Q_\sigma(g)/\bar{f}\rangle = \int_0^{+\infty}\biggl(\int_0^{+\infty}\sigma(x,y)V(f,g)(x,y)\,d\nu(x)\biggr)\,d\gamma(y) \quad \textit{for } f,g\in L^2(d\nu).$$

Furthermore, $\|Q_\sigma\|_* \le \|\sigma\|_{2,\nu\otimes\gamma}$.

Proof. Let $\sigma \in D_*(\mathbb{R}^2)$. For $g \in D_*(\mathbb{R})$, put $Q_\sigma(g) = W_\sigma(g)$. From Theorems 3.6 and 3.7, we obtain

$$\begin{aligned}\langle Q_\sigma(g)/\bar{f}\rangle &= \langle W_\sigma(g)/\bar{f}\rangle = \mathbb{H}_\sigma(f,g)\\ &= \int_0^{+\infty}\biggl(\int_0^{+\infty}\sigma(x,y)V(f,g)(x,y)\,d\nu(x)\biggr)\,d\gamma(y).\end{aligned}$$

On the other hand, from Proposition 2.2(ii), we have

$$\bigl|\langle Q_\sigma(g)/\bar{f}\rangle\bigr| \le \|\sigma\|_{2,\nu\otimes\gamma}\,\|f\|_{2,\nu}\,\|g\|_{2,\nu}\,.$$

Thus $Q_\sigma \in \mathcal{B}(L^2(d\nu))$ and

$$\|Q_\sigma\|_* \le \|\sigma\|_{2,\nu\otimes\gamma}\,. \tag{4–1}$$

Now consider $\sigma \in L^2(d\nu\otimes d\gamma)$. Let $(\sigma_k)_{k\in\mathbb{N}}$ be a sequence in $D_*(\mathbb{R}^2)$ such that $\|\sigma_k-\sigma\|_{2,\nu\otimes\gamma}$ approaches 0 as $k\to+\infty$. From (4–1) we have, for all $k,l\in\mathbb{N}$,

$$\|Q_{\sigma_k}-Q_{\sigma_l}\|_* \le \|\sigma_k-\sigma_l\|_{2,\nu\otimes\gamma} \le \|\sigma_k-\sigma\|_{2,\nu\otimes\gamma} + \|\sigma_l-\sigma\|_{2,\nu\otimes\gamma}.$$

Thus $(Q_{\sigma_k})_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{B}(L^2(d\nu))$. Let it converge to Q_σ. Clearly Q_σ is independent from the choice of $(\sigma_k)_{k\in\mathbb{N}}$, and we have

$$\|Q_\sigma\|_* = \lim_{k\to+\infty}\|Q_{\sigma_k}\|_* \le \lim_{k\to+\infty}\|\sigma_k\|_{2,\nu\otimes\gamma} = \|\sigma\|_{2,\nu\otimes\gamma}.$$

We consider first $f,g\in D_*(\mathbb{R})$. Then

$$\begin{aligned}\langle Q_\sigma(g)/\bar{f}\rangle &= \lim_{k\to+\infty}\langle Q_{\sigma_k}(g)/\bar{f}\rangle\\ &= \lim_{k\to+\infty}\int_0^{+\infty}\biggl(\int_0^{+\infty}\sigma_k(x,y)V(f,g)(x,y)\,d\nu(x)\biggr)\,d\gamma(y)\\ &= \int_0^{+\infty}\biggl(\int_0^{+\infty}\sigma(x,y)V(f,g)(x,y)\,d\nu(x)\biggr)\,d\gamma(y).\end{aligned}$$

Now let f, g be in $L^2(d\nu)$. Pick sequences $(f_k)_{k\in\mathbb{N}}$, and $(g_k)_{k\in\mathbb{N}}$ in $D_*(\mathbb{R})$ converging to f and g, respectively, in the $\|\cdot\|_{2,\nu}$-norm. Then

$$\begin{aligned}\langle Q_\sigma(g)/\bar{f}\rangle &= \lim_{k\to+\infty}\langle Q_\sigma(g_k)/\bar{f}_k\rangle\\ &= \lim_{k\to+\infty}\int_0^{+\infty}\left(\int_0^{+\infty}\sigma(x,y)V(f_k,g_k)(x,y)\,d\nu(x)\right)d\gamma(y)\\ &= \int_0^{+\infty}\left(\int_0^{+\infty}\sigma(x,y)V(f,g)(x,y)\,d\nu(x)\right)d\gamma(y). \qquad \square\end{aligned}$$

We now give an extension of Theorem 4.1 that will allow us to prove that for $1\le p\le 2$ the Weyl transform with symbol in $L^p(d\nu\otimes d\gamma)$, is a compact operator.

Theorem 4.2. *Let $p\in[1,2]$. There exists a unique bounded operator*

$$Q : L^p(d\nu\otimes d\gamma)\to \mathscr{B}(L^2(d\nu)),$$

whose action is denoted by $\sigma\to Q_\sigma$, such that

$$\langle Q_\sigma(g)/\bar{f}\rangle = \int_0^{+\infty}\left(\int_0^{+\infty}\sigma(x,y)V(f,g)(x,y)\,d\nu(x)\right)d\gamma(y) \quad \textit{for } f,g\in D_*(\mathbb{R}).$$

Moreover, $\|Q_\sigma\|_*\le\|\sigma\|_{p,\nu\otimes\gamma}$.

Proof. The case $p=2$ is given by Theorem 4.1. We turn to the case $p=1$. For $\sigma\in D_*(\mathbb{R}^2)$, we define Q_σ by

$$Q_\sigma(g)=W_\sigma(g)\quad \text{for } g\in D_*(\mathbb{R}).$$

From Theorems 3.6 and 3.7, we have, for $f\in D_*(\mathbb{R})$,

$$\langle Q_\sigma(g)/\bar{f}\rangle = \mathbb{H}_\sigma(f,g) = \int_0^{+\infty}\left(\int_0^{+\infty}\sigma(x,y)V(f,g)(x,y)\,d\nu(x)\right)d\gamma(y).$$

From Hölder's inequality we then obtain

$$\left|\langle Q_\sigma(g)/\bar{f}\rangle\right| \le \|\sigma\|_{1,\nu\otimes\gamma}\,\|V(f,g)\|_{\infty,\nu\otimes\gamma} \le \|\sigma\|_{1,\nu\otimes\gamma}\,\|f\|_{2,\nu}\,\|g\|_{2,\nu}\,.$$

This shows that $Q_\sigma\in\mathscr{B}(L^2(d\nu))$ and $\|Q_\sigma\|_*\le\|\sigma\|_{1,\nu\otimes\gamma}$.

We extend the definition of Q_σ and the two facts just proved to the case of $\sigma\in L^1(d\nu\otimes d\gamma)$, working as in the proof of Theorem 4.1.

Finally, the Riesz–Thorin Theorem [Stein 1956; Stein and Weiss 1971], allows us to generalize the same results from the cases $p=1$ and $p=2$ to all $p\in[1,2]$. $\square$

Theorem 4.3. *Let $p\in[1,2]$. For $\sigma\in L^p(d\nu\otimes d\gamma)$, the operator Q_σ from $L^2(d\nu)$ into itself is compact.*

Proof. Given $\sigma \in L^p(d\nu \otimes d\gamma)$, choose a sequence $(\sigma_k)_{k\in\mathbb{N}}$ in $D_*(\mathbb{R}^2)$ approximating σ in the $\|\cdot\|_{p,\nu\otimes\gamma}$-norm. The last assertion of Theorem 4.2 says that

$$\|Q_{\sigma_k} - Q_\sigma\|_* \le \|\sigma_k - \sigma\|_{p,\nu\otimes\gamma},$$

so Q_{σ_k} approaches Q_σ in $\mathcal{B}(L^2(d\nu))$. From Theorem 3.6 we know that $W_{\sigma_k} = Q_{\sigma_k}$ is compact for all $k \in \mathbb{N}$. The theorem then follows from the fact that the subspace $\mathcal{K}(L^2(d\nu))$ of $\mathcal{B}(L^2(d\nu))$ consisting of compact operators is a closed ideal of $\mathcal{B}(L^2(d\nu))$. □

5. The Weyl transform with symbol in $S'_{*,0}(\mathbb{R}^2)$

Notation. We denote by

- $S_{*,0}(\mathbb{R}^2)$ the subspace of $S_*(\mathbb{R}^2)$ consisting of functions with compact support with respect to the first variable;
- $S'_{*,0}(\mathbb{R}^2)$ the topological dual of $S_{*,0}(\mathbb{R}^2)$;
- $D'_*(\mathbb{R})$ the space of even distribution on $\mathbb{R}$. It is the topological dual of $D_*(\mathbb{R})$.

Definition 5.1. For $\sigma \in S'_{*,0}(\mathbb{R}^2)$ and $g \in D_*(\mathbb{R})$, we define the operator $W_\sigma(g)$ on $D_*(\mathbb{R})$ by

$$\big(W_\sigma(g)\big)(f) = \sigma(V(f,g)) \quad \text{for } f \in D_*(\mathbb{R}), \tag{5–1}$$

where V is the mapping from Definition 2.1. Clearly $W_\sigma(g)$ belongs to $D'_*(\mathbb{R})$.

Proposition 5.2. *Consider the distribution σ of $S'_{*,0}(\mathbb{R}^2)$ given by the constant function* 1. *For all $g \in D_*(\mathbb{R})$, we have*

$$W_\sigma(g) = c\delta,$$

where $c = \int_0^{+\infty} g(x)\,d\nu(x)$ and δ is the Dirac distribution at 0.

Proof. For $f, g \in D_*(\mathbb{R})$, we get

$$\big(W_\sigma(g)\big)(f) = \sigma(V(f,g)) = \int_0^{+\infty}\bigg(\int_0^{+\infty} V(f,g)(x,y)\,d\nu(x)\bigg)\,d\gamma(y).$$

But from the proof of Theorem 2.5, we have

$$\int_0^{+\infty} V(f,g)(x,y)\,d\nu(x) = c\mathcal{F}f(y) \quad \text{for } y \in \mathbb{R}.$$

Integrating both sides over $[0,+\infty[$ with respect to the measure $d\gamma$ and using (1–8), we obtain

$$\sigma(V(f,g)) = \big(W_\sigma(g)\big)(f) = c\int_0^{+\infty} \mathcal{F}f(y)\,d\gamma(y) = cf(0) = (c\delta,\ f). \qquad \square$$

Note that by Proposition 5.2, there exists $\sigma \in S'_{*,0}(\mathbb{R}^2)$, given by a function in $L^\infty(d\nu \otimes d\gamma)$, such that for all $g \in D_*(\mathbb{R})$ satisfying $c = \int_0^{+\infty} g(x)\, d\nu(x) \neq 0$, the distribution $W_\sigma(g)$ is *not* given by a function in $L^2(d\nu)$.

6. The Weyl transform with symbol in $L^p(d\nu \otimes d\gamma)$, for $2 < p < \infty$

Theorem 6.1. *Let $p \in]2, \infty[$. There exists a function $\sigma \in L^p(d\nu \otimes d\gamma)$ such that the Weyl transform W_σ defined by* (5–1) *is not a bounded linear operator on $L^2(d\nu)$.*

We break down the proof into two lemmas, of which the theorem is an immediate consequence.

Lemma 6.2. *Let $p \in]2, \infty[$. Suppose that for all $\sigma \in L^p(d\nu \otimes d\gamma)$, the Weyl transform W_σ given by* (5–1) *is a bounded linear operator on $L^2(d\nu)$. Then there exists a positive constant M such that*

$$\|W_\sigma\|_* \leq M\, \|\sigma\|_{p,\nu\otimes\gamma} \quad \text{for all } \sigma \in L^p(d\nu \otimes d\gamma). \tag{6–1}$$

Proof. Under the assumption of the lemma, there exists for each $\sigma \in L^p(d\nu \otimes d\gamma)$ a positive constant C_σ such that

$$\|W_\sigma(g)\|_{2,\nu} \leq C_\sigma \|g\|_{2,\nu} \quad \text{for } g \in L^2(d\nu).$$

Let $f, g \in D_*(\mathbb{R})$ be such that $\|f\|_{2,\nu} = \|g\|_{2,\nu} = 1$ and define a linear operator $Q_{f,g} : L^p(d\nu \otimes d\gamma) \to \mathbb{C}$ by

$$Q_{f,g}(\sigma) = \langle W_\sigma(g)/\bar{f}\rangle.$$

Then

$$\sup_{|f|_{2,\nu}=|g|_{2,\nu}=1} |Q_{f,g}(\sigma)| \leq C_\sigma.$$

By the Banach–Steinhaus theorem, the operator $Q_{f,g}$ is bounded on $L^p(d\nu \otimes d\gamma)$, so there exists $M > 0$ such that

$$\|Q_{f,g}\| = \sup_{\|\sigma\|_{p,\nu\otimes\gamma}=1} |Q_{f,g}(\sigma)| \leq M.$$

From this we deduce that for all $f, g \in D_*(\mathbb{R})$ and $\sigma \in L^p(d\nu \otimes d\gamma)$,

$$\left|\langle W_\sigma(g)/\bar{f}\rangle\right| \leq M \|\sigma\|_{p,\nu\otimes\gamma} \|f\|_{2,\nu}\, \|g\|_{2,\nu},$$

which implies (6–1). □

Lemma 6.3. *For $2 < p < \infty$, there is no positive constant M satisfying* (6–1).

Proof. Suppose there exists such an M. Let p' be such that $1/p + 1/p' = 1$. Then $p' \in]1, 2[$. We consider, for $f, g \in D_*(\mathbb{R})$, the function $V(f, g)$ of Definition 2.1. We have

$$\begin{aligned}\|V(f,g)\|_{p',\nu\otimes\gamma} &= \sup_{\|\sigma\|_{p,\nu\otimes\gamma}=1} \left| \int_0^{+\infty}\int_0^{+\infty} \sigma(x,y) V(f,g)(x,y)\, d\nu(x)\, d\gamma(y) \right| \\ &= \sup_{\|\sigma\|_{p,\nu\otimes\gamma}=1} \left| \langle W_\sigma(g)/\bar{f}\rangle \right| \le \sup_{\|\sigma\|_{p,\nu\otimes\gamma}=1} \|W_\sigma(g)\|_{2,\nu} \|f\|_{2,\nu},\end{aligned}$$

and consequently

$$\|V(f,g)\|_{p',\nu\otimes\gamma} \le M \|f\|_{2,\nu}\, \|g\|_{2,\nu}. \tag{6–2}$$

Now consider f, g in $L^2(d\nu)$. Choose sequences $(f_k)_{k\in\mathbb{N}}$ and $(g_k)_{k\in\mathbb{N}}$ in $D_*(\mathbb{R})$ approximating f and g in the $\|\cdot\|_{2,\nu}$-norm. By Proposition 2.2, the sequence $(V(f_k, g_k))_{k\in\mathbb{N}}$ converges to $V(f, g)$ in $L^{p'}(d\nu \otimes d\gamma)$, and thus we have extended (6–2) to all $f, g \in L^2(d\nu)$. We will exhibit an example where this leads to a contradiction.

Let f be an even, measurable function on $\mathbb{R}$, supported in $[-1, 1]$. We have

$$\left| V(f, f)(x, y) \right| \le |f| * |f|(x),$$

where $*$ is the convolution product (Definition 1.1). From (1–7), we deduce that for all $y \in \mathbb{R}$, the function $x \mapsto V(f, f)(x, y)$ is supported in $[-2, 2]$. Hölder's inequality gives

$$\begin{aligned}&\left(\int_0^{+\infty} \left| \int_0^2 V(f,f)(x,y)\, d\nu(x) \right|^{p'} d\gamma(y) \right)^{1/p'} \\ &\qquad \le \left(\int_0^2 d\nu(x) \right)^{1/p} \left(\int_0^{+\infty} \left(\int_0^2 |V(f,f)(x,y)|^{p'} d\nu(x) \right) d\gamma(y) \right)^{1/p'} \\ &\qquad = \left(\int_0^2 d\nu(x) \right)^{1/p} \|V(f,f)\|_{p',\nu\otimes\gamma} \le M \left(\int_0^2 d\nu(x) \right)^{1/p} \|f\|_{2,\nu}^2,\end{aligned}$$

the last inequality following from (6–2). This proves that the function

$$y \mapsto \int_0^{+\infty} V(f,f)(x,y)\, d\nu(x) = c\mathscr{F}f(y)$$

belongs to $L^{p'}(d\gamma)$; here $c = \int_0^{+\infty} f(x)\, d\nu(x)$. and we have used the proof of Theorem 2.5 for the equality on the right-hand side. Putting this together with the preceding inequality we see that, if $c \neq 0$, the function $\mathscr{F}f$ belongs to $L^{p'}(d\gamma)$ and

$$\|\mathscr{F}f\|_{p',\gamma} \le \frac{M}{|c|} \left(\int_0^2 d\nu(x) \right)^{1/p} \|f\|_{2,\nu}^2. \tag{6–3}$$

Now consider the particular function f given by

$$f(x) = \frac{|x|^r}{\sqrt{B(x)}}\,\mathbf{1}_{[-1,1]}(x)$$

where B is the function defined by (1–1) and $\mathbf{1}_{[-1,1]}$ is the characteristic function of the interval $[-1, 1]$. If $r > -(\alpha+1)$, this function belongs to $L^1(d\nu)\cap L^2(d\nu)$. From (1–4) we get

$$\begin{aligned}\mathscr{F}f(\lambda) &= \int_0^1 x^{r+2\alpha+1} j_\alpha(\lambda x)dx + \int_0^1 x^{r+\alpha+1/2}\theta_\lambda(x)dx\\ &= \frac{1}{\lambda^{r+2\alpha+2}}\int_0^\lambda x^{r+2\alpha+1} j_\alpha(x)dx + \int_0^1 x^{r+\alpha+1/2}\theta_\lambda(x)\,dx.\end{aligned}$$

Using the asymptotic expansion of the function j_α [Lebedev 1972; Watson 1944], given by

$$j_\alpha(x) = \frac{2^{\alpha+1/2}\Gamma(\alpha+1)}{\sqrt{\pi}\,x^{\alpha+1/2}}\left(\cos\left(x-\alpha\frac{\pi}{2}-\frac{\pi}{4}\right)+O\left(\frac{1}{x}\right)\right)\quad \text{as } x\to+\infty,$$

we deduce that for $-(\alpha+1) < r < -(\alpha+\frac{1}{2})$, the integral

$$a := \int_0^{+\infty} x^{r+2\alpha+1} j_\alpha(x)\,dx$$

exists and is finite, so

$$\frac{1}{\lambda^{r+2\alpha+2}}\int_0^\lambda x^{r+2\alpha+1} j_\alpha(x)\,dx \sim \frac{a}{\lambda^{r+2\alpha+2}}\quad \text{as } \lambda\to+\infty.$$

On the other hand, for $\lambda > 1$,

$$\left|\int_0^1 x^{r+\alpha+1/2}\theta_\lambda(x)dx\right| \le \frac{c_1}{\lambda^{\alpha+3/2}}\int_0^1 x^{r+\alpha+1/2}\Psi(x)\,dx,$$

where

$$\Psi(x) = \left(\int_0^x |Q(s)|\,ds\right)\exp\left(c_2\int_0^x |Q(s)|\,ds\right)\quad \text{for all } \; x>0$$

and Q is given by (1–5). Since $-(\alpha+1) < r < -(\alpha+\frac{1}{2})$, we deduce that

$$\mathscr{F}f(\lambda) \sim \frac{a}{\lambda^{r+2\alpha+2}}\quad \text{as } \lambda\to+\infty.$$

Using this and (1–6), it follows that there exist K, $R > 0$ such that

$$|\mathscr{F}f(\lambda)|^{p'}\frac{1}{2\pi|c(\lambda)|^2} \ge \frac{K}{\lambda^{p'(r+2\alpha+2)-2\alpha-1}}\quad \text{for } \lambda > R;$$

so for r such that $p'(r+2\alpha+2) < 2\alpha+2$, we get

$$\|\mathscr{F}f\|_{p',\gamma}^{p'} \geq \int_R^{+\infty} |\mathscr{F}f(\lambda)|^{p'} \frac{d\lambda}{2\pi\,|c(\lambda)|^2} \geq \int_R^{+\infty} \frac{K}{\lambda^{p'(r+2\alpha+2)-2\alpha-1}} d\lambda = +\infty.$$

This shows that the relation (6–3) is false if we choose r so as to satisfy simultaneously the conditions $r > -(\alpha+1)$, $r < -(\alpha+\frac{1}{2})$ and

$$r < -(2\alpha+2) + \frac{2\alpha+2}{p'}.$$

This contradiction proves the lemma and Theorem 6.1. □

References

[Chebli 1979] H. Chebli, "Théorème de Paley–Wiener associé à un opérateur différentiel singulier sur $(0, \infty)$", *J. Math. Pures Appl.* (9) **58**:1 (1979), 1–19. MR 80g:47050 Zbl 0394.34019

[Folland 1984] G. B. Folland, *Real analysis*, Pure and Applied Mathematics, Wiley, New York, 1984. MR 86k:28001 Zbl 0549.28001

[Lebedev 1972] N. N. Lebedev, *Special functions and their applications*, Dover, New York, 1972. MR 50 #2568 Zbl 0271.33001

[Nessibi et al. 1998] M. M. Nessibi, L. T. Rachdi, and K. Trimèche, "The local central limit theorem on the product of the Chébli–Trimèche hypergroup and the Euclidean hypergroup $\mathbb{R}^n$", *J. Math. Sci. (Calcutta)* **9**:2 (1998), 109–123. MR 99m:60017

[Stein 1956] E. M. Stein, "Interpolation of linear operators", *Trans. Amer. Math. Soc.* **83** (1956), 482–492. MR 18,575d Zbl 0072.32402

[Stein and Weiss 1971] E. M. Stein and G. Weiss, *Introduction to Fourier analysis on Euclidean spaces*, Princeton mathematical series **32**, Princeton University Press, Princeton, N.J., 1971. MR 46 #4102 Zbl 0232.42007

[Trimèche 1981] K. Trimèche, "Transformation intégrale de Weyl et théorème de Paley–Wiener associés à un opérateur différentiel singulier sur $(0, \infty)$", *J. Math. Pures Appl.* (9) **60**:1 (1981), 51–98. MR 83i:47058 Zbl 0416.44002

[Trimèche 1997] K. Trimèche, "Inversion of the Lions transmutation operators using generalized wavelets", *Appl. Comput. Harmon. Anal.* **4**:1 (1997), 97–112. MR 98d:34046 Zbl 0872.34059

[Watson 1944] G. N. Watson, *A treatise on the theory of Bessel functions*, second ed., Cambridge University Press, Cambridge, 1944. MR 6,64a Zbl 0063.08184

[Weyl 1931] H. Weyl, *The theory of groups and quantum mechanics*, Methuen, London, 1931. Reprinted by Dover, New York, 1950. Zbl 0041.56804 JFM 58.1374.01

[Wong 1998] M. W. Wong, *Weyl transforms*, Universitext, Springer, New York, 1998. MR 2000c:47098 Zbl 0908.44002

[Xu 1994] Z. Xu, *Harmonic analysis on Chébli–Trimèche hypergroups*, Ph.D. thesis, Murdoch University, Perth, Western Australia, 1994.

[Zeuner 1989] H. Zeuner, "The central limit theorem for Chébli–Trimèche hypergroups", *J. Theoret. Probab.* **2**:1 (1989), 51–63. MR 90e:60087 Zbl 0912.60008

Received November 20, 2003. Revised April 7, 2004.

CYRINE BACCAR
DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCES OF TUNIS
1060 TUNIS
TUNISIA

LAKHDAR TANNECH RACHDI
DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCES OF TUNIS
1060 TUNIS
TUNISIA
lakhdartannech.rachdi@fst.rnu.tn

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

DISTANCE AND BRIDGE POSITION

DAVID BACHMAN AND SAUL SCHLEIMER

J. Hempel's definition of the *distance* of a Heegaard surface generalizes to a notion of complexity for any knot that is in bridge position with respect to a Heegaard surface. Our main result is that the distance of a knot in bridge position is bounded above by twice the genus, plus the number of boundary components, of an essential surface in the knot complement. As a consequence knots constructed via sufficiently high powers of pseudo-Anosov maps have minimal bridge presentations which are thin.

1. Introduction

Hempel's definition [2001] of the *distance* of a Heegaard splitting is a natural measure of complexity, generalizing the standard notions of *reducibility* (distance zero), *weak reducibility* (distance at most one), and *strong irreducibility* (distance at least two). Hempel proves that there exist Heegaard splittings of arbitrarily high distance.

In his Ph.D. thesis, K. Hartshorn related the distance of a Heegaard splitting to the genus of any essential surface, thus refining work of T. Kobayashi [1988]:

Theorem [Hartshorn 1999]. *Let M be a closed, orientable, irreducible 3-manifold with Heegaard splitting F. Suppose M contains an orientable, incompressible surface S. Then the distance of F is bounded above by twice the genus of S.*

We introduce our results by recalling a generalization of the curve complex for surfaces with nonempty boundary. This allows us to translate Hempel's definition of distance for Heegaard splittings to a definition of distance for knots that are in bridge position with respect to a Heegaard surface [Morimoto and Sakuma 1991]. Our main result is a translation of Hartshorn's Theorem into this new context:

Theorem 5.1. *Let K be a knot in a closed, orientable 3-manifold M which is in bridge position with respect to a Heegaard surface F. Let S be a properly embedded, orientable, essential surface in M_K. Then the distance of K with respect to F is bounded above by twice the genus of S plus $|\partial S|$.*

MSC2000: 57M25, 57M27.
Keywords: Heegaard splitting, curve complex.

In the special case of a meridional disk we find that a stronger result holds; the distance of K with respect to F is zero. This follows from a variant of the Haken Lemma (see Lemma 4.1).

Although our proof contains Hartshorn's result as a special case ($K = \varnothing$), there is an interesting qualitative difference. Unlike Hartshorn, we make no minimality assumption on the way in which S intersects F. That is, *any generic position* of S with respect to F forces the bound on distance as stated in the theorem.

The main idea behind our proof is to simply count saddles. Let $d(K, F)$ denote the distance of K with respect to F. It is a standard technique in 3-manifold topology to use a Heegaard splitting F for a 3-manifold M to define a height function h on M. This, in turn, induces a height function on a surface S in M. With respect to this height function S will have maxima, minima, and saddles. The moral of the story is that each critical point of S either

(1) contributes at most 1 to $d(K, F)$ and exactly -1 to the Euler characteristic of S, or

(2) contributes nothing to $d(K, F)$ and nothing to the Euler characteristic of S.

Hence, the distance of K with respect to F would then be bounded by the negative of the Euler characteristic of S. Unfortunately, for Heegaard splittings the above classification isn't exactly correct. We find that there may be at most two special critical points that each contribute one to the distance of K, but nothing to the Euler characteristic of S. This gives the bound

$$d(K, F) \leq -\chi(S) + 2 = 2g(S) + |\partial S|.$$

We note that several authors have explicitly computed the distances of various classes of knots (using varying definitions of *distance*). See, for example, [Akiyoshi et al. 2000; Morimoto 1989; Saito 2004].

In the final section we present corollaries to Theorem 5.1. Among these are:

Corollary 6.1. *Suppose K is a knot in S^3 whose distance is $d(K, F)$ with respect to a bridge sphere F. Then the genus of K is at least $\frac{1}{2}\big(d(K, F) - 1\big)$.*

Corollary 6.2. *If K is a knot whose distance is at least* 3 *with respect to some Heegaard surface, the complement of K is hyperbolic.*

Finally, we define the *bridge link* associated to an element of the braid group B_{2n} to be the link obtained by gluing two trivial n-strand tangles by this element. By a construction essentially due to Kobayashi [1988], powers of certain pseudo-Anosov maps give associated bridge links with arbitrarily high distance. Suppose ϕ is such a map. Then it follows from Corollary 6.5 that for all sufficiently high powers of ϕ if the associated link is a knot, its minimal bridge presentation is thin.

A priori, bridge knots associated to high powers of pseudo-Anosov maps might have low bridge numbers. We conjecture that this is not in fact possible:

Conjecture. Suppose K is a knot whose distance is at least 2 with respect to some Heegaard surface F. Then the distance of K with respect to any other Heegaard surface is bounded above by $\chi(F - K) + 2$.

Compare this to the statement of Theorem 5.1. In the theorem we assert that the distance of a knot with respect to a Heegaard surface is bounded by two plus the Euler characteristic of an essential surface. In the conjecture we propose that distance is similarly bounded by a strongly irreducible surface.

2. Basic definitions

In this section we give the definitions that will be used throughout the paper. Let K be a knot in a closed, orientable 3-manifold, M. Let $M_K = M - N(K)$ where $N(K)$ denotes a regular neighborhood of K. For the remainder of this paper all surfaces S in M_K will be embedded, compact, and orientable with $S \cap \partial M_K \subset \partial S$.

Definition 2.1. A *cut surface* (see Figure 1) is either

(1) a disk $E \subset M_K$ such that $E \cap \partial M_K = \varnothing$,

(2) a bigon $E \subset M_K$ such that $E \cap \partial M_K$ is an arc, or

(3) an annulus $E \subset M_K$ with exactly one meridional boundary component on ∂M_K. In other words, $E \cap \partial M_K$ is a loop bounding a disk in $\overline{N(K)}$.

If E is a cut surface and $\gamma = \overline{\partial E - \partial M_K}$ we say that γ *bounds a cut surface*.

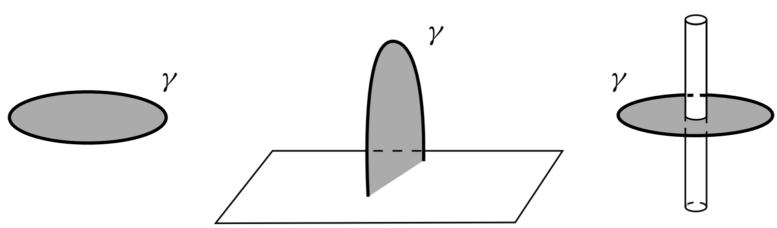

Figure 1. Disk, bigon, and meridional cut surfaces.

A properly embedded simple curve in S is *inessential* if it bounds a subsurface of S that is a cut surface, and *essential* otherwise.

Suppose γ bounds a cut surface E, that S is properly embedded in M_K, and that $S \cap E = \gamma$. We may then *surger* S along E by replacing a neighborhood of γ in S with two parallel copies of E. If γ is essential in S we say E is a *compression* for S. In this case we also say γ *bounds a compression* for S.

A properly embedded surface $S \subset M_K$ is *essential* if first there are no curves on S which bound compressions in M_K and second ∂S (if nonempty) is not null-homotopic on ∂M_K. We also consider a 2-sphere to be essential if it does not bound a ball in M_K. This notion of essentialness is *not* identical to that of "super-incompressible" found in [Morgan and Bass 1984].

A *handlebody* is a 3-manifold homeomorphic to the closure of a regular neighborhood of a compact, connected graph in $\mathbb{R}^3$. If such a graph has no valence-one vertices and the corresponding handlebody has nonzero genus, the graph's image under such a homeomorphism is a *spine* of the handlebody. We will insist that the spine of a 3-ball be a single edge.

A closed surface F in M is a *Heegaard surface* of M if F separates M into two handlebodies. An arc properly embedded in H is *trivial* if it bounds a bigon in H. Suppose K is a knot in a 3-manifold M with Heegaard surface F. The knot K is in *bridge position with respect to* F [Morimoto and Sakuma 1991] if K meets each of the handlebodies bounded by F in a collection of trivial arcs. Such a position is sometimes referred to as a (g, b)-*presentation* of K, where $g = \text{genus}(F)$ and $2b = |K \cap F|$.

3. The arc complex

Following Hempel's definition [2001] of the distance of a Heegaard splitting, we now define the *distance* of a knot K that is in bridge position with respect to a Heegaard surface $F \subset M$. Set

$$M_K = M - N(K) \quad \text{and} \quad F_K = F \cap M_K.$$

Construct a 1-complex $\Gamma(F_K)$ as follows: for each proper isotopy class of essential curves in F_K there is a vertex of $\Gamma(F_K)$. There is an edge of $\Gamma(F_K)$ between two distinct vertices if and only if there are representatives of the corresponding isotopy classes which are disjoint. $\Gamma(F_K)$ is called the *arc complex* of F_K (see [Masur and Minsky 1999], for example).

Now, F_K divides M_K into two submanifolds, H and H'. Let V and V' denote the sets of vertices of $\Gamma(F_K)$ corresponding to curves that bound compressions in H and H', respectively. Then $d(K, F)$, the *distance of K with respect to F*, is defined to be the number of edges in the shortest path from V to V' in $\Gamma(F_K)$. As long as $\chi(F_K)$ is at most -2 this is well defined, since the arc complex is connected in those cases.

4. Lemmas

The following is a slight variant of the Haken Lemma [1968]. We assume familiarity with the proof of this result found in [Jaco 1980, Theorem II.7].

Lemma 4.1 (Haken). *Let K be a knot in a 3-manifold M which is in bridge position with respect to a Heegaard surface F. If M_K contains an essential 2-sphere or meridional disk then $d(K, F) = 0$.*

Proof. Among all essential 2-spheres and meridional disks in M_K choose one, S, meeting F_K minimally. Let H and H' denote the submanifolds of M_K bounded by F_K, with ∂S (if nonempty) contained in H. If $S \cap F_K = \varnothing$ then S lies entirely in H or H', a contradiction. It follows that $S \cap F_K$ is a nonempty set of loops that are essential on F_K. Hence, if S meets F_K in a single loop, the result follows.

Suppose then that $|S \cap F_K| > 1$. Let H^* denote one of H or H', where there is a component T of $S \cap H^*$ with $|\partial T - \partial S| \geq 2$. Choose a *basis* Λ for H^*, that is, a system of disks and bigons cutting H^* into a 3-ball. If $S \cap \Lambda$ contains any loops, surger S along these loops, innermost (on Λ) first. At least one component of the resulting surface is again an essential sphere or meridional disk. We continue to denote this surface by S.

Now reduce $|S \cap \Lambda|$ as follows. If any component of $(S \cap H^*) - \Lambda$ is a bigon, surger Λ along this surface. Some subcollection of the resulting set is again a basis, which we continue to denote by Λ. If not, choose a bigon of $\Lambda - S$, and use this to guide an isotopy of S (see the "isotopy of type A" in [Jaco 1980, p. 24]). Repeat this procedure until all components T of $S \cap H^*$ satisfy $|\partial T - \partial S| = 1$. Let S' denote the resulting surface.

It follows from the argument of [Jaco 1980, Lemma II.9] that if $H^* = H'$ then $|S' \cap F_K| < |S \cap F_K|$, and we have reached a contradiction. If $H^* = H$ then $|S' \cap F_K| \leq |S \cap F_K|$. If equality holds we repeat the preceding steps with S' replacing S and letting $H^* = H'$. This gives a surface S'' with $|S'' \cap F_K| < |S \cap F_K|$, a contradiction. □

Lemma 4.2. *Let K be a knot in a 3-manifold M which is in bridge position with respect to a Heegaard surface F. Suppose γ bounds two cut surfaces A and B with $A \cap B = \gamma$. Then A and B are both bigons, both annuli, or both disks, unless $d(K, F) = 0$.*

Proof. If A and B are of different types, their union is a meridional disk. The result now follows from Lemma 4.1. □

Lemma 4.3. *Let K be a knot in a 3-manifold M which is in bridge position with respect to a Heegaard surface F and let Q be any properly embedded surface in M_K. If there is a curve γ that is essential on Q and bounds a cut surface E in M_K then either there is a curve $\gamma' \subset E \cap Q$ that bounds a compression for Q, unless $d(K, F) = 0$.*

Proof. Let $\Lambda \subset E \cap Q$ be the collection of curves that are essential on Q. Let E' denote the closure of a component of $E - \Lambda$ that is a cut surface. Set $\gamma' = E' \cap \Lambda$.

Consider the set Θ of cut surfaces bounded by γ' such that the only curve of intersection with Q, essential on Q, is γ'. Note that E' is such a surface, so Θ is nonempty. Let E^* be an element of Θ with $|E^* \cap Q|$ minimal.

We now claim $E^* \cap Q = \gamma'$. Suppose not. Let E'' be a cut surface component of $E^* - Q$. The curve $\gamma'' = E'' \cap Q$ is inessential on Q and hence bounds two cut surfaces, $A \subset Q$ and E''. Note that $A \cap E'' = \gamma''$. By Lemma 4.2 we may obtain a new cut surface from E^* by replacing E'' with a push-off of A. This violates the minimality of $|E^* \cap Q|$. We conclude that E^* is a compression for Q, which finishes the proof. □

Lemma 4.4. *Let K be a knot in a 3-manifold M which is in bridge position with respect to a Heegaard surface F and let S be an essential surface in M_K. If we surger S along a disk or bigon cut surface then at least one of the remaining components is essential, unless $d(K, F) = 0$.*

Proof. By assumption there is a curve $\gamma \subset S$ that bounds a cut surface E', homeomorphic to a disk and such that $E' \cap S = \gamma$. Since S is essential, γ bounds a cut surface $E \subset S$. Surgering S along E' produces two surfaces, isotopic to $E \cup E'$ and $S' = (S - E) \cup E'$. Suppose S' is not essential. Let γ' bound a compression C for S'. As E' is homeomorphic to a disk we may properly isotope γ' off of E'. The curve γ' is now on S, and bounds the cut surface C. By Lemma 4.3 there is a compression C' for S, a contradiction. □

Lemma 4.5. *Let K be a knot in a 3-manifold M which is in bridge position with respect to a Heegaard surface F and let S be an essential surface in M_K. If we surger S along a cut surface then at least one of the remaining components is essential, unless $d(K, F) = 0$.*

Proof. By assumption there is a curve $\gamma \subset S$ which bounds a cut surface E' such that $E' \cap S = \gamma$. Since S is essential, γ bounds a cut surface E in S. Surgering S along E' then produces two surfaces, isotopic to $E \cup E'$ and $S' = (S - E) \cup E'$.

By Lemma 4.4 we may assume E' is an annulus. By Lemma 4.2 we may assume E is also an annulus. If $E \cup E'$ is essential, we are done. Otherwise there must be a compressing bigon B for $E \cup E'$ (since the core loop of $E \cup E'$ is not essential). Surgering $E \cup E'$ along B gives a disk D with $\partial D \subset \partial M_K$ bounding a disk $D' \subset \partial M_K$. If the sphere $D \cup D'$ is essential, the proof is complete by Lemma 4.1. Otherwise we conclude that $E \cup E'$, together with an annulus of ∂M_K, bounds a solid torus. If the interior of the solid torus is disjoint from S then S' is properly isotopic to S and we are done. If S meets the interior of the solid torus we may push it entirely into the solid torus. Now consider $B \cap S$. Some component of $B - S$ is then a cut surface for S. This cut surface is either a disk or a bigon. By Lemma 4.4 we may surger S along this cut surface and obtain another essential surface that meets B fewer times. Continuing in this way we obtain an essential surface inside the solid torus that misses B, and is hence contained in a ball. This is impossible. $\square$

5. Proof of the Main Theorem

We recall the statement.

Theorem 5.1. *Let K be a knot in a closed, orientable 3-manifold M which is in bridge position with respect to a Heegaard surface F. Let S be a properly embedded, orientable, essential surface in M_K. Then the distance of K with respect to F is bounded above by twice the genus of S plus $|\partial S|$.*

We now begin the proof. Throughout we assume that $d(K, F) > 0$ to avoid the special cases of the lemmas from Section 4. Let Σ_0 and Σ_1 denote spines of the handlebodies bounded by F. Let $h : M \to I$ denote a height function on M such that $h^{-1}(0) = \Sigma_0$ and $h^{-1}(1) = \Sigma_1$. We require that for every $t \in (0, 1)$ the surface $h^{-1}(t)$ is parallel to $F = h^{-1}\left(\frac{1}{2}\right)$. Because K is in bridge position with respect to F we may isotope K so that each arc of $K - F$ has one critical point with respect to h. Now pull each minimum down to Σ_0 and each maximum up to Σ_1. If $M = S^3$ and F is a sphere we may assume that K has at least two maxima and at least two minima. In this case Σ_0 and Σ_1 are edges, and we assume that the vertices $\partial\Sigma_0$ coincide with two minima of K and the vertices $\partial\Sigma_1$ coincide with two maxima.

Set $F(t) = h^{-1}(t) \cap M_K$. Let $H(t)$ be the closure of the component of $M_K - F(t)$ that contains Σ_0. Let $H'(t)$ be the closure of $M_K - H(t)$. Let ϵ_0 be chosen just larger than the radius of $N(K)$, but small enough so that S meets $H(\epsilon_0)$ and $H'(1 - \epsilon_0)$ in compressions for $F(\epsilon_0)$ and $F(1 - \epsilon_0)$. Then the surface $F(t)$ is homeomorphic to $F_K = F \cap M_K$ for every value of $t \in [\epsilon_0, 1 - \epsilon_0]$. Hence, the submanifold $\bigcup_{t=\epsilon_0}^{1-\epsilon_0} F(t)$ is homeomorphic to $F_K \times [\epsilon_0, 1 - \epsilon_0]$. Let π denote the composition of such a homeomorphism with projection onto the first factor. Hence, if γ is a curve on $F(t)$ for some $t \in [\epsilon_0, 1 - \epsilon_0]$, then $\pi(\gamma)$ is a curve on F_K.

We make two types of assumptions on the position of the essential surface S. Any surface whose position satisfies these assumptions we will say is in *standard position*. The first concerns how S meets ∂M_K and the second is a genericity assumption on the interior of S. Near the boundary of S we assume the following:

- Meridional boundary components are "level"; that is, if S has meridional boundary, there exists for each boundary component C of S a $t \in (\epsilon_0, 1 - \epsilon_0)$ such that $C \subset \partial F(t)$. We consider t a critical value for S if some boundary component of S is contained in $\partial F(t)$.
- If S does not have meridional boundary then for generic t and each component γ of $\partial S - F(t)$ the endpoints of γ lie on distinct boundary components of $F(t)$.

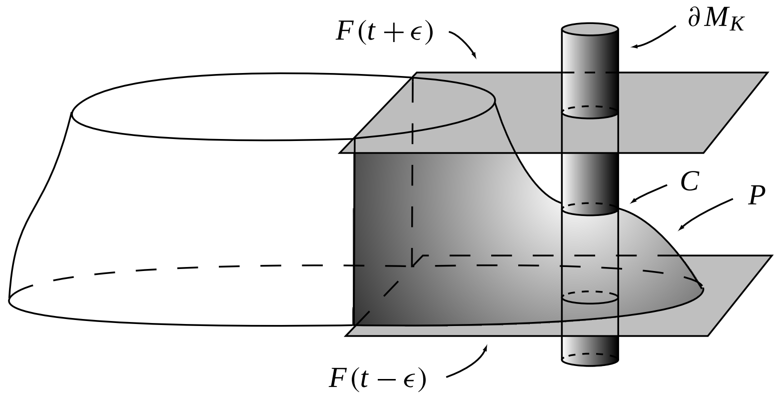

Figure 2. A piece of S between levels, $F(t-\epsilon)$ and $F(t+\epsilon)$, before and after a meridional boundary component, C.

These stipulations are possible since ∂S is not null-homotopic on ∂M_K. In the interior of M_K we assume the position of S is generic in the following sense:

- All critical points of $h|_S$ are maxima, minima, or saddles. We will refer to any such critical point whose height is between ϵ_0 and $1-\epsilon_0$ and to any meridional boundary component as a *critical submanifold* (of S).
- The heights of any two critical submanifolds of S are distinct.
- Suppose a meridional boundary component C of S happens at height t. Let P denote the closure of the component of $S - F(t \pm \epsilon)$ that has C as a boundary component. Then P is a once-punctured annulus with one boundary component on each of $F(t-\epsilon)$ and $F(t+\epsilon)$ (see Figure 2). (This uses the fact that ∂M_K is connected.)

Claim 5.2. *For each $t \in [\epsilon_0, 1-\epsilon_0]$ the submanifolds $H(t)$ and $H'(t)$ of M_K do not contain any essential surfaces.*

Proof. Choose a basis Λ of compressing disks and bigons in $H(t)$ that cut it into a ball. Suppose $D \in \Lambda$. Let D' be a cut surface component of $D - Q$, where Q is some essential surface in $H(t)$. By Lemma 4.4, compressing Q along D' yields an essential surface that meets D fewer times. Continuing in this way we produce an essential surface in $H(t)$ disjoint from Λ, and hence in a ball. This is impossible. $\square$

Definition 5.3. Let t_0 be the supremum of $t \in [\epsilon_0, 1-\epsilon_0]$ such that some curve in $S \cap F(t)$ bounds a compression for $F(t)$ in $H(t)$. (The compression for $F(t)$ need not be a subsurface of S.) Define t_1 likewise with infimum instead of supremum and $H'(t)$ instead of $H(t)$.

Claim 5.4. *The values t_0 and t_1 are well defined, and $t_0 > \epsilon_0$.*

Proof. To establish the claim it is enough to show that for some small $\epsilon > \epsilon_0$ there are curves in $S \cap F(\epsilon)$ and $S \cap F(1-\epsilon)$ that bound compressions for $F(\epsilon)$ and $F(1-\epsilon)$ in $H(\epsilon)$ and $H'(1-\epsilon)$, respectively.

There are essentially two cases. Suppose first the essential surface S is closed, or has meridional boundary. If $S \cap \Sigma_0 = \varnothing$ then S can be properly isotoped entirely into $H'(\epsilon)$, violating Claim 5.2. We conclude that $S \cap \Sigma_0 \neq \varnothing$. $F(\epsilon) \cap S$ then contains a loop that bounds a compression for $F(\epsilon)$ in $H(\epsilon)$. On the other hand, if S has nonempty, nonmeridional boundary then $F(\epsilon) \cap S$ contains an arc that bounds a bigon compression in $H(\epsilon)$. This proves that t_0 is well defined and $t_0 > \epsilon > \epsilon_0$. A symmetric argument shows t_1 is well defined. □

Claim 5.5. *The value of t_0 is less than $1-\epsilon_0$.*

Proof. Suppose $t_0 = 1-\epsilon_0$. Let $\epsilon > \epsilon_0$ be small enough that $1-\epsilon$ is greater than the height of the highest critical submanifold. As $t_0 = 1-\epsilon_0$ there is a curve α of $F(1-\epsilon) \cap S$ that is essential in $F(1-\epsilon)$ but bounds a compression in $H(1-\epsilon)$.

Recall that the boundary of S has been isotoped into standard position. It follows that the components of $S \cap H'(1-\epsilon)$ are all disks and bigons. Hence, α bounds compressions for $F(1-\epsilon)$ on both sides and $d(K, F) = 0$. □

Claim 5.6. *If $t_0 = t_1 < 1-\epsilon_0$ then $d(K, F) = 1$.*

Proof. If $t_0 = t_1 < 1-\epsilon_0$ then for all sufficiently small ϵ there is a curve of $S \cap F(t_0+\epsilon)$ bounding a compression in $H'(t)$ and a curve of $S \cap F(t_0-\epsilon)$ bounding a compression in $H(t)$. But for ϵ sufficiently small the curves of $\pi(S \cap F(t_0+\epsilon))$ can be made disjoint from the curves of $\pi\big(S \cap F(t_0-\epsilon)\big)$, because F and S are orientable. This is basically identical to [Gabai 1987, Lemma 4.4]. □

Henceforth we assume that $\epsilon_0 < t_0 < t_1 < 1-\epsilon_0$.

Claim 5.7. *If $t_* \in (t_0, t_1)$ is a critical value then for sufficiently small ϵ the curves of $\pi(F(t_*-\epsilon) \cap S)$ are at a distance of at most one from the curves of $\pi(F(t_*+\epsilon) \cap S)$.*

Proof. As in the proof of Claim 5.6, the curves of $\pi(S \cap F(t_*+\epsilon))$, for ϵ sufficiently small, can be made disjoint from the curves of $\pi(S \cap F(t_*-\epsilon))$. The result follows unless either of these are collections of inessential curves, and hence are not represented in $\Gamma(F_K)$. However, if this is the case then all curves of $S \cap F(t_*+\epsilon)$ (say) are inessential on S. By Lemma 4.5 a sequence of surgeries produces an essential surface disjoint from $F(t_*+\epsilon)$, contradicting Claim 5.2. □

Claim 5.8. *A component of $F(t) \cap S$ that is inessential on $F(t)$ is inessential on S.*

Proof. This follows directly from Lemma 4.3. □

Now let $t \in [\epsilon_0, 1-\epsilon_0]$ be a regular value of $h|_S$. Pick a component γ of $F(t) \cap S$. The curve γ is *mutually essential* if it is essential on both $F(t)$ and S, *mutually inessential* if it is inessential on both and *mutual* if it is mutually essential or mutually inessential. Finally, γ is *special* if it is inessential on S but essential on $F(t)$. There are three kinds of special curves: loops that bound disks on S, loops that cobound (with ∂S) annuli in S, and arcs isotopic (via bigons) into ∂S.

Claim 5.9. *Suppose t is a regular value of $h|_S$ in $[t_0, t_1]$. Every curve of $F(t) \cap S$ is mutual.*

Proof. Pick a regular value $t \in [\epsilon_0, 1-\epsilon_0]$. By Claim 5.8 we may assume that there is a special curve γ in $F(t) \cap S$. By definition, γ is essential on $F(t)$ but inessential on S. It follows that a component E of $S - \gamma$ is a cut surface. By Lemma 4.3 there is a curve of $E \cap F(t)$ that bounds a compression for $F(t)$. This compression either lies in $H(t)$ or in $H'(t)$. Since $E \cap F(t) \subset S \cap F(t)$ we conclude $t \notin [t_0, t_1]$. □

Claim 5.10. *If α is an arc component of $F(t) \cap S$ and $h(\alpha) = t \in (t_0, t_1)$ then α is mutually essential.*

Proof. By Claim 5.9 the only other possibility is that α is mutually inessential. In this case $\partial\alpha$ is the boundary of some arc γ of $\partial S - F(t)$. Also, $\partial\gamma = \partial\alpha$ lies on the same component of $\partial F(t)$. This violates our assumption that S is in standard position. □

In $h^{-1}([t_0, t_1])$ we see the usual four types of critical submanifolds for S: maxima, minima, saddles, and meridional boundary components. Suppose a critical submanifold happening at height t is a saddle or meridional boundary component. Let P be the closure of the component of $S - F(t \pm \epsilon)$ that contains the critical submanifold. We call P a *horizontal neighborhood* (in S) of the critical submanifold. Let $\partial_\pm P = P \cap F(t \pm \epsilon)$. We say the critical submanifold at t is *special* if there is some component of $\partial_\pm P$ that is special. If the critical submanifold at t is not special, we say it is *inessential* if some component of the closure of $S - P$ is a disk and *essential* otherwise. If the critical submanifold at t is inessential, Claim 5.10 implies that there is a mutually inessential loop component of $\partial_\pm P$ that bounds a disk in S.

Claim 5.11. *Suppose $t_* \in [t_0, t_1]$. If there is a special critical submanifold at t_* then $t_* = t_0$ or t_1.*

Proof. By definition, if a special critical submanifold happens at t_* there is a special curve α in $S \cap F(t_* - \epsilon)$ or $S \cap F(t_* + \epsilon)$. Assuming the former, Claim 5.9 implies $t_* - \epsilon \notin [t_0, t_1]$. Hence $t_* = t_0$. If, on the other hand, $\alpha \subset F(t_* + \epsilon)$, we deduce $t_* = t_1$. □

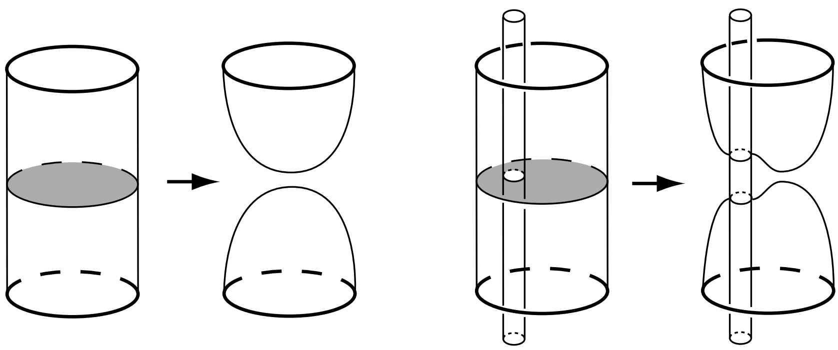

Figure 3. Constructing S' from S. On the left two new critical values are created. On the right four are created.

Lemma 5.12. *Let t_- and t_+ be regular values in $[t_0, t_1]$ such that every saddle and every meridional boundary component of S in $h^{-1}(t_-, t_+)$ is inessential. Then $\pi(F(t_-) \cap S)$ and $\pi(F(t_+) \cap S)$ share a vertex in $\Gamma(F_K)$.*

Proof. Let $\{t_i\}$ be the critical values of $h|_S$ lying in $[t_-, t_+]$. Choose r_i slightly greater than the t_i and let $R = \{r_i\} \cup \{t_- + \epsilon\}$.

For every $r \in R$ surger S in the following way. If $S \cap F(r)$ contains mutually inessential curves, some such curve bounds a cut surface in $F(r)$. Surger S along this cut surface. After a sequence of such surgeries we obtain from S a surface that meets $F(r)$ only in mutually essential curves, for all $r \in R$.

Set $M' = h^{-1}([t_-, t_+])$. Let S' be the intersection of the surgered surface with M'. Note that $h|_{S'}$, the height function restricted to S', has either two or four new critical values for every surgery performed. See Figure 3.

We say a surface V is *vertical* in M' if $V = \pi^{-1}(\alpha) \cap M'$, where α is a properly embedded one-manifold in F_K. A vertical surface V is either a disk or an annulus. We need the following claim to prove the lemma:

Claim 5.13. *Each component S'' of S' is either*

- *a sphere or a meridional annulus, or*
- *properly isotopic into $F(t_-)$ or $F(t_+)$, or*
- *properly isotopic to a vertical surface V with $\pi(V)$ essential in F_K.*

Proof. If $h|_{S''}$ has no critical values, S'' is isotopic to a vertical annulus or disk. In this case $S'' \cap \partial M'$ must be essential by the construction of S'. Note that this kind of situation is the desired conclusion of the lemma at hand. If $h|_{S''}$ has only critical values of even index (and no meridional boundary components) then S'' is a boundary parallel disk or a sphere.

We now assume that S'' contains a critical submanifold which is not a max or min. The component S'' either contains a saddle or meridional boundary component of S, or it does not. Suppose the latter. It follows that S'' is either a meridional annulus or a boundary parallel annulus (with one boundary component on ∂M_K).

Now suppose that S'' contains a saddle or meridional boundary component of S at height t_*. Let P be the closure of the component of $S'' - F(t_* \pm \epsilon)$ that contains this critical submanifold. (Note that P is also a subsurface of S since ϵ is very small.) Recall that P is the horizontal neighborhood of the critical submanifold. Let $\partial_\pm P = P \cap F(t_* \pm \epsilon)$. Since every critical submanifold of $S \cap M'$ is inessential, at least one loop component of $\partial_\pm P$ bounds a disk in S (see the comment preceding Claim 5.11).

Now suppose that S'' contains a meridional boundary component of S at height t_*. Let P be the corresponding horizontal neighborhood. Let $\partial_\pm P = C_1 \cup C_2$, where C_1 bounds a disk D in S. Hence, $D \cup P \subset S$ is a cut annulus and we see that C_2 is also inessential in S. By Claim 5.9 the C_i are inessential in $F(t_* \pm \epsilon)$. It now follows from Lemma 4.2 that C_1 bounds a disk in $F(t_* \pm \epsilon)$ while C_2 bounds a cut annulus in $F(t_* \mp \epsilon)$. Thus S'' is a meridional annulus.

We now assume that S'' contains no meridional boundary components of S, and hence contains a saddle. Suppose some such saddle has a horizontal neighborhood P such that two components of $\partial_\pm P$ are inessential. It follows that all three components are inessential. If two bound disks, all three do. Therefore, by Lemma 4.2, S'' is a sphere. If one bounds a disk and the other two bound cut annuli then S'' is a meridional annulus.

Finally, we assume that S'' contains no meridional boundary components and that every saddle x has a horizontal neighborhood P_x with exactly one component γ_x of $\partial_\pm P_x$ inessential, bounding a disk in S (see Figure 4). By Claim 5.9 and Lemma 4.2 it follows that γ_x bounds a disk in S''. Hence S'' is either a union of disks or a union of annuli. In the first case S'' is isotopic to a vertical disk. In

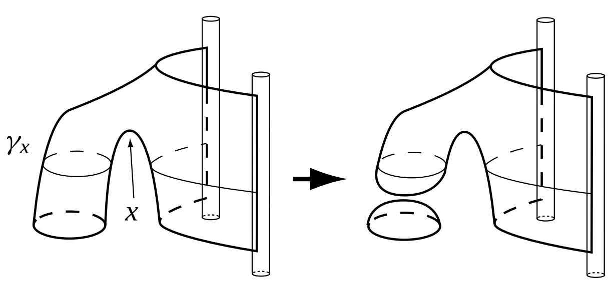

Figure 4. Surgery near a saddle whose horizontal neighborhood has exactly one inessential boundary component.

the latter case S'' is either isotopic to a vertical annulus or is a boundary parallel annulus. $\square$

To complete the proof of the Lemma 5.12, suppose no component of S' meets both boundary components of M'. By Claim 5.13, every component of S' meeting $F(t_-)$ is boundary parallel in M'. Isotope $F(t_-)$ across these boundary parallelisms to obtain a surface F' that intersects the surface S only in mutually inessential curves. Some component of $F' - S$ is then a cut surface, which we may use to surger S. By Lemma 4.5 we obtain an essential surface that meets F' in fewer curves. Continuing in this fashion we obtain an essential surface disjoint from F', violating Claim 5.2.

We conclude that there is a component $S'' \subset S'$ meeting both $F(t_-)$ and $F(t_+)$. By Claim 5.13, this S'' must be isotopic to a vertical annulus or vertical disk with essential boundary. The lemma is thus proved. $\square$

We now complete the proof of Theorem 5.1. Note that when $t \in [t_0, t_1]$ is a regular value, $\pi(F(t) \cap S)$ is a properly embedded 1-manifold in F_K (recall that $F_K = F \cap M_K$). The distance between the loops and arcs of $\pi(F(t_0 - \epsilon) \cap S)$ and of $\pi(F(t_1 + \epsilon) \cap S)$ in $\Gamma(F_K)$ is an upper bound for the distance $d(K, F)$. By Lemma 5.12 and Claim 5.7 this number is bounded by the number of essential critical submanifolds, e, plus the number of special critical submanifolds. By Claim 5.11 this latter number is at most two. We therefore conclude $d(K, F) \le e + 2$.

We now bound the Euler characteristic of S. Suppose an essential critical submanifold happens at t_* and let P be its horizontal neighborhood in S. Note that in all cases $\chi(P) = -1$. (When P has vertical boundary compute its Euler characteristic by doubling across the vertical boundary and taking half of the Euler characteristic of the resulting surface. See, for example, the surface on the left in Figure 4.) By the definition of an essential critical submanifold $\partial P - \partial S$ is essential in S. We conclude that $\chi(S) \le -e$.

Putting these facts together we conclude that

$$d(K, F) \le e + 2 \le -\chi(S) + 2 = -(2 - 2g(S) - |\partial S|) + 2 = 2g(S) + |\partial S|.$$

6. Applications

We now present a few quick corollaries to Theorem 5.1.

Corollary 6.1. *Suppose K is a knot in S^3 whose distance is $d(K, F)$ with respect to a bridge sphere F. Then the genus of K is at least $\frac{1}{2}\big(d(K, F) - 1\big)$.*

Proof. The genus of K is defined to be the smallest genus of all orientable spanning surfaces for K. Such a spanning surface is essential and has exactly one boundary component. Hence, an immediate application of Theorem 5.1 implies $d(K, F) \le 2g(K) + 1$. $\square$

Corollary 6.2. *If K is a knot whose distance is at least* 3 *with respect to some Heegaard surface, the complement of K is hyperbolic of finite volume.*

Proof. If the distance is greater than two, M_K is irreducible, atoroidal, anannular, and has incompressible boundary. It follows from Thurston's geometrization theorem for Haken manifolds that M_K is hyperbolic of finite volume. □

Definition 6.3. Suppose M is obtained by removing a neighborhood of a knot K in S^3 and gluing in a new solid torus to the resulting boundary component. Then we say that M was obtained by *Dehn surgery* on K.

Corollary 6.4. *Suppose K is a knot in S^3 whose distance is $d(K, F)$ with respect to a bridge sphere F. If a manifold M obtained by Dehn surgery on K contains an incompressible torus T, then $\left|\partial(T \cap M_K)\right|$ is at least $d(K, F) - 2$.*

Proof. Choose T so as to minimize $|T \cap K|$ in M. Let $T_K = T \cap M_K$. follows from the minimality assumption that T_K is essential. Theorem 5.1 says that $d(K, F)$ is bounded above by twice the genus of T_K plus $|\partial T_K|$. But T is a torus, so the genus of T_K is one. □

Corollary 6.5. *Suppose K is a knot in S^3 whose distance with respect to some bridge sphere is greater than its bridge number. Then a minimal bridge presentation for K is thin.*

Proof. Let F be a bridge sphere for which $d(K, F) \geq |K \cap F|$. If thin position for K does not equal bridge position then by [Thompson 1997] there is a planar, meridional, essential surface S in the complement of K with fewer boundary components than $|K \cap F|$. Hence, by Theorem 5.1 the distance $d(K, F)$ is at most $|\partial S| \leq |K \cap F|$. □

References

[Akiyoshi et al. 2000] H. Akiyoshi, M. Sakuma, M. Wada, and Y. Yamashita, "Ford domains of punctured torus groups and two-bridge knot groups", *Sūrikaisekikenkyūsho Kōkyūroku* (= *Suri Kaiseki Kenkyujo kokyuroku*) **1163** (2000), 67–77. MR 1799369 Zbl 0969.57505

[Gabai 1987] D. Gabai, "Foliations and the topology of 3-manifolds, III", *J. Differential Geom.* **26**:3 (1987), 479–536. MR 89a:57014b Zbl 0639.57008

[Haken 1968] W. Haken, "Some results on surfaces in 3-manifolds", pp. 39–98 in *Studies in Modern Topology*, Math. Assoc. Amer., Buffalo and Prentice-Hall, Englewood Cliffs (NJ), 1968. MR 36 #7118 Zbl 0194.24902

[Hartshorn 1999] K. Hartshorn, *Heegaard splittings: the distance complex and the stabilization conjecture*, Ph.D. thesis, University of California, Berkeley, 1999.

[Hempel 2001] J. Hempel, "3-manifolds as viewed from the curve complex", *Topology* **40**:3 (2001), 631–657. MR 2002f:57044 Zbl 0985.57014

[Jaco 1980] W. Jaco, *Lectures on three-manifold topology*, CBMS Regional Conference Series in Mathematics **43**, Amer. Math. Soc., Providence, 1980. MR 81k:57009 Zbl 0433.57001

[Kobayashi 1988] T. Kobayashi, "Heights of simple loops and pseudo-Anosov homeomorphisms", pp. 327–338 in *Braids* (Santa Cruz, CA, 1986), edited by J. S. Birman and A. Libgober, Contemp. Math. **78**, Amer. Math. Soc., Providence, RI, 1988. MR 89m:57015 Zbl 0663.57010

[Masur and Minsky 1999] H. A. Masur and Y. N. Minsky, "Geometry of the complex of curves, I: Hyperbolicity", *Invent. Math.* **138**:1 (1999), 103–149. MR 2000i:57027 Zbl 0941.32012

[Morgan and Bass 1984] J. Morgan and H. Bass (editors), *The Smith conjecture* (New York, 1979), edited by J. Morgan and H. Bass, Pure and Applied Mathematics **112**, Academic Press, Orlando, FL, 1984. MR 86i:57002 Zbl 0599.57001

[Morimoto 1989] K. Morimoto, "On minimum genus Heegaard splittings of some orientable closed 3-manifolds", *Tokyo J. Math.* **12**:2 (1989), 321–355. MR 91b:57020 Zbl 0714.57007

[Morimoto and Sakuma 1991] K. Morimoto and M. Sakuma, "On unknotting tunnels for knots", *Math. Ann.* **289**:1 (1991), 143–167. MR 92e:57015 Zbl 0697.57002

[Saito 2004] T. Saito, "Genus one 1-bridge knots as viewed from the curve complex", *Osaka J. Math.* **41**:2 (2004), 427–454. MR 2069094 Zbl 02111464

[Thompson 1997] A. Thompson, "Thin position and bridge number for knots in the 3-sphere", *Topology* **36**:2 (1997), 505–507. MR 97m:57013 Zbl 0867.57009

Received August 29, 2003. Revised October 2, 2003.

DAVID BACHMAN
PITZER COLLEGE
1050 MILLS AVE
CLAREMONT, CA 91711
bachman@pitzer.edu

SAUL SCHLEIMER
MATHEMATICS DEPARTMENT
UNIVERSITY OF ILLINOIS AT CHICAGO
CHICAGO, IL 60607
saul@math.uic.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

ON THE BEHAVIOUR OF ∞-HARMONIC FUNCTIONS ON SOME SPECIAL UNBOUNDED DOMAINS

TILAK BHATTACHARYA

We study nonnegative ∞-harmonic functions defined on unbounded domains, in particular the half-space and the exterior of the unit closed ball. We prove that if such a function u vanishes continuously on the boundary then in the first case u is affine, and in the second case u is radial and linear. We also discuss growth rates in an infinite strip.

1. Introduction and statements of results

We study nonnegative ∞-harmonic functions on unbounded domains with special geometry, in particular the half-space and the exterior of the unit closed ball. We consider functions that vanish on the boundaries while their behaviour at infinity is left unspecified. One may view this work as a step towards understanding the kind of growth rates possible for infinity-harmonic functions on unbounded domains. An analogous result appears in [Crandall et al. 2001], where it is shown that an ∞-harmonic function bounded below by a plane is affine. This is related to the conjecture that globally Lipschitz ∞-harmonic functions on $\mathbb{R}^n$ are affine; however we do not attempt to prove this. The restriction on the sign plays a strong role in this work and has been critical in obtaining estimates for growth rates. It is unclear what happens if this restriction is removed.

Let $u = u(x)$ be an ∞-harmonic function defined on a (possibly unbounded) domain $\Omega \subset \mathbb{R}^n$, for $n \geq 2$. That is, u solves

$$\Delta_\infty u(x) = \sum_{i,j=1}^{n} \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j} \frac{\partial^2 u}{\partial x_i \partial x_j} = 0 \quad \text{for } x \in \Omega$$

in the viscosity sense. We refer to [Bhattacharya 2002; 2004, Crandall and Evans 2001; Crandall et al. 1992; 2001] for definitions. For the most part we assume that

MSC2000: 35J60, 35J70.
Keywords: ∞-harmonic functions, viscosity solutions, growth rates, unbounded domains.
This research is partially supported by a fund from the NSERC.

$u(x) \geq 0$ for $x \in \Omega$, that the boundary $\partial\Omega$ is smooth, and that u is continuous up to $\partial\Omega$. Let O denote the origin in $\mathbb{R}^n$, and for $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, set

$$|x| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Theorem 1.1 (The infinite half-space). *Let $\Omega = \{x \in \mathbb{R}^n : x_n > 0\}$ be the infinite half-space. Suppose $u(x) \geq 0$ is ∞-harmonic in Ω and vanishes continuously on the hyperplane $\{x_n = 0\}$. Then either $u(x) = 0$ for all $x \in \Omega$, or there exists a positive constant $K > 0$ such that $u(x) = K x_n$ for all $x \in \Omega$.*

In this case, the sign restriction leads to linear growth rate in Ω. This also holds when Ω is the exterior of a ball. In both cases, linear growth rate implies global Lipschitz continuity. The truth of the conjecture mentioned earlier would then imply Theorem 1.1. Solutions with unrestricted sign may have faster growth rates as demonstrated by the well known example

$$u(x, y) = |x|^{4/3} - |y|^{4/3}$$

on $\mathbb{R}^2$, in the half planes bounded by $|x| = |y|$. It is not clear whether a growth rate faster than $\frac{4}{3}$ is possible in general, or whether the imposition of a growth rate of $\frac{4}{3}$ would imply that u is of this type.

Let $B(1, O)$ be the unit open ball in $\mathbb{R}^n$, centered at O, and let $\Omega = \mathbb{R}^n \setminus \bar{B}(1, O)$, where $\bar{B}(1, O)$ denotes the closure of $B(1, O)$.

Theorem 1.2 (The exterior of a ball). *Let $u \geq 0$ be ∞-harmonic in Ω. Suppose that u vanishes continuously on $\partial B(1, O)$. Then either $u(x) = 0$ for all $x \in \Omega$, or there exists a positive constant K such that $u(x) = K(|x| - 1)$ for all $x \in \Omega$.*

While solutions are globally Lipschitz continuous, Theorem 1.2 would not follow from the conjecture mentioned earlier. It is unclear if a faster growth rate is possible when the sign restriction is removed. It would also be interesting to know if Theorems 1.1 and 1.2 would follow for solutions with unrestricted sign but with linear growth rate.

We also discuss the case of the infinite strip $\{x : 0 < x_n < 1\}$ and show that any nontrivial solution $u(r)$ grows faster than any integral power of r, where r is the distance from the x_n-axis. However, it is not clear if nontrivial solutions exist (see Section 5). In this work, we make considerable use of the properties proven in [Bhattacharya 2002; 2004; Crandall et al. 2001]. The devices mostly used are monotonicity, the Harnack inequality, comparison, cone comparison and the boundary Harnack inequality for flat boundaries. For discussion see [Bhattacharya 2002; Crandall et al. 2001]. Also see [Aronsson et al. 2004; Bhattacharya et al. 1989] for more information of the origins of such questions and issues related to ∞-harmonic functions.

We have divided our work as follows. In Section 2, we introduce notations needed in our work, and recall some preliminary results about ∞-harmonic functions. We prove Theorem 1.1 in Section 3, and the proof of Theorem 1.2 appears in Section 4. Finally, in Section 5, we present a short discussion in the case of the infinite strip.

2. Notations and Preliminaries

Let O be the origin in $\mathbb{R}^n$, let $\overline{U}$ denote the closure of a set U in $\mathbb{R}^n$, and let $\vec{e}_n$ be the unit vector parallel to the positive x_n-axis. Let $B(r, P)$ denote the open ball in $\mathbb{R}^n$ with center P and radius $r > 0$, let $\Omega(r, P)$ denote the intersection $\Omega \cap B(r, P)$, let $\partial(r, P)$ be $\partial\Omega \cap \overline{B}(r, P)$ and let $E(r, P)$ be $\partial B(r, P) \cap \overline{\Omega}$. For $A = (A_1, A_2, A_3, \dots, A_n) \in \mathbb{R}^n$, let $x_n(A) = A_n$, let $A' = (A_1, A_2, A_3, \dots, A_{n-1})$, let

$$|P - Q|_{n-1} = \sqrt{\sum_{i=1}^{n-1} (P_i - Q_i)^2}$$

and, for $t \in \mathbb{R}$, let $A + t\vec{e}_n = (A', A_n + t)$. For $P \in \mathbb{R}^n$, let $C(r, P)$ denote the cylinder $\{x \in \mathbb{R}^n : P_n < x_n < P_n + 2r, |x - P|_{n-1} < r\}$. Thus $C(r, P)$ has length $2r$ and radius r, and its axis is parallel to the x_n-axis. Let $F(r, P)$ denote the flat face $\{x \in \mathbb{R}^n : x_n = P_n, |x - P|_{n-1} < r\}$ of $C(r, P)$ which lies in the hyperplane $x_n = P_n$. We study the problem

$$\begin{aligned} \Delta_\infty u(x) &= 0 \quad \text{for } x \in \Omega, \\ u(x) &= 0 \quad \text{for } x \in \partial\Omega. \end{aligned}$$

We assume that $u(x) \geq 0$ for $x \in \Omega$ unless otherwise stated, and that $\partial\Omega$ will be smooth and u continuous up to $\partial\Omega$. It is well known that u is locally Lipschitz continuous in Ω (see [Bhattacharya 2002; Crandall et al. 2001; Jensen 1993; Lindqvist and Manfredi 1995]) and has the cone comparison property, and we make considerable use of these facts throughout this work. We now list a set of facts about ∞-harmonic functions.

We use the following version of the Harnack inequality [Bhattacharya 2002; 2004; Lindqvist and Manfredi 1995]: let $u > 0$ be ∞-harmonic in Ω, and let $\delta > 0$ be such that the set $\Omega_\delta = \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) \geq \delta\}$ is not empty. If P and Q are points in Ω_δ and the segment $PQ \subset \Omega_\delta$, then $u(P) \geq e^{-|P-Q|/\delta} u(Q)$. If P is joined to Q by a smooth path in Ω_δ, with arc length $l(P, Q)$ then

$$u(P) \geq e^{-l(P,Q)/\delta} u(Q). \tag{1}$$

Monotonicity plays a crucial role here [Bhattacharya 2002, Lemma 3.6; 2004, Lemma 3]. Let $u \geq 0$ be ∞-harmonic in Ω, and $B(r, z) \subset \Omega$. For $x \in B(r, z)$, define $d(x) = r - |x - z| = \operatorname{dist}(x, \partial B(r, z))$; if $\vec{\eta}$ is a unit vector and $0 \leq s < t < r$, then

$$\frac{u(z)}{r}=\frac{u(z)}{d(z)}\le\frac{u(z+s\vec{\eta})}{d(z+s\vec{\eta})}=\frac{u(z+s\vec{\eta})}{r-s}\le\frac{u(z+t\vec{\eta})}{r-t}=\frac{u(z+t\vec{\eta})}{d(z+t\vec{\eta})}. \tag{2}$$

We will need a different version of (2). We take $u=0$ on $\partial\Omega$, and for $z\in\mathbb{R}^n$ we define $M(r,z)=\sup_{\Omega(r,z)}u(x)=\sup_{\partial\Omega(r,z)}u(x)$.

Lemma 2.1. *Let $u\ge 0$ be ∞-harmonic in Ω and $u|_{\partial\Omega}=0$; suppose $z\in\mathbb{R}^n$, and $r>0$ is such that $\Omega(r,z)$ is not empty. Let $x, y\in\Omega(r,z)$ be on the same radial line through z, with $|x-z|<|y-z|<r$, and suppose that $u(x)\le l+(M(r,z)-l)|x-z|/r$ for some $l\in\mathbb{R}$, and all $x\in\Omega(r,z)$. Then*

$$\frac{M(r,z)-l}{r}\le\frac{M(r,z)-u(x)}{r-|x-z|}\le\frac{M(r,z)-u(y)}{r-|y-z|}.$$

If $z\in\overline{\Omega}$ this holds with $u(z)$ in place of l.

Proof. Fix $x\in\Omega$, set $B(r,x,z)=B(r-|x-z|,x)$ and $O(x,z,r)=B(r,x,z)\cap\Omega$. For $w\in O(x,z,r)$ define

$$\omega(w)=u(x)+\frac{(M(r,z)-u(x))|w-x|}{r-|x-z|}.$$

Then $u\le\omega$ on $\partial B(r,x,z)\cap\overline{\Omega}$ and $\partial\Omega\cap\overline{B}(r,x,z)$, and $u(x)=\omega(x)$. By comparison, $u\le\omega$ in $O(x,z,r)$ [Barles and Busca 2001; Bhattacharya 2002; Crandall et al. 2001; Jensen 1993]. Note that $O(x,z,r)\subset\Omega(r,z)$. The first inequality follows trivially, and the second follows by taking $w=y$. Let $z\in\overline{\Omega}$ and define

$$v(x)=u(z)+\frac{(M(r,z)-u(z))|x-z|}{r}$$

in $\Omega(r,z)$. By comparison, $u(x)\le v(x)$ in $\Omega(r,z)$ and the claim follows. □

We recall the boundary Harnack inequality [Bhattacharya 2002]. Let $P\in\mathbb{R}^n$ and $s>0$. Suppose that $u_1,u_2>0$ are ∞-harmonic in $C(8s,P)$, and vanish continuously on $F(8s,P)$. Then there exist constants M_1 and M_2, independent of s and u_i, such that for all $x\in C(s,P)$,

$$M_1\frac{u_1(z)}{u_2(z)}\le\frac{u_1(x)}{u_2(x)}\le M_2\frac{u_1(z)}{u_2(z)}, \tag{3}$$

where $z=(P', P_n+2s)$. We now assume that Ω is unbounded and show that nonconstant ∞-harmonic functions, with unrestricted sign, have at least linear growth. If $u\ge 0$ and has linear growth in Ω then Lemma 2.3 implies global Lipschitz continuity.

Lemma 2.2. *Let u be ∞-harmonic in Ω such that $u|_{\partial\Omega}=0$. Fix $z\in\mathbb{R}^n$ and $t\ge 0$, and define $\delta=\mathrm{dist}(z,\Omega)$. Then*

(i) *$M(r, z)$ is convex in r, for all $r \geq \delta$, and*

(ii) $M(r, z) \geq M(\delta, z) + \big(M(t, z) - M(\delta, z)\big)\Big(\dfrac{r-\delta}{t-\delta}\Big)$ *for $r > t > \delta$.*

Proof. Set $\rho(x) = |x - z|$ and choose a, b and c such that $\delta < a < c < b$. As the intersection $\partial B(a, z) \cap \partial\Omega$ is not empty, by the maximum principle,

$$0 \leq M(a, z) \leq M(c, z) \leq M(b, z).$$

Define $v(x) = M(a, z) + \big(M(b, z) - M(a, z)\big)(\rho(x) - a)/(b - a) \geq 0$ for all x in $\Omega(b, z) \setminus \overline{\Omega}(a, z)$. Clearly, $u \leq v$ on $\partial B(a, z) \cap \overline{\Omega}$ and on $\partial B(b, z) \cap \overline{\Omega}$, and $u = 0 \leq v$ on $\partial\Omega \cap (\overline{B}(b, z) \setminus B(a, z))$. By the cone comparison, $u(x) \leq v(x)$ for all $x \in \Omega(b, z) \setminus \overline{\Omega}(a, z)$. Hence

$$\sup_{|x-z|=c} u(x) = M(c, z) \leq M(a, z) + \big(M(b, z) - M(a, z)\big)(c - a)/(b - a),$$

and convexity follows. Since $\big(M(r, z) - M(a, z)\big)/(r - a)$ increases as r increases, selecting $a = \delta$ and $r > t > \delta$, a simple rearrangement yields part (ii). Note that $M(\delta, z) = 0$ if $z \in \mathbb{R}^n \setminus \overline{\Omega}$, and $M(\delta, z) = M(0, z) = u(z)$ if $z \in \overline{\Omega}$. □

Lemma 2.3. *Let $u \geq 0$ be ∞-harmonic in Ω and $u|_{\partial\Omega} = 0$. If, for some $z \in \Omega$, some $C > 0$ and some $a > 0$, $M(r, z) \leq Cr$ for all $r \geq a$, then u is globally Lipschitz continuous in Ω, with Lipschitz constant C.*

Proof. Fix $x, y \in \Omega$. For $\rho > 0$, let $v(w) = u(x) + \big(M(\rho, x) - u(x)\big)\,(|w - x|/\rho)$ in $\Omega(\rho, x)$. Then $u \leq v = M(\rho, x)$, on $\partial B(\rho, x) \cap \Omega$. Also, $u = 0 \leq v$ on $\partial\Omega \cap B(\rho, x)$ and $u(x) = v(x)$. By comparison, $u \leq v$ in $\Omega(\rho, x)$. By the maximum principle, $M(\rho, x) \leq M(\rho + |x - z|, z) \leq C(\rho + |x - z|)$. Set $w = y$ and $\rho > |x - y|$, then

$$\begin{aligned} \frac{u(y) - u(x)}{|x - y|} &\leq \frac{M(\rho, x) - u(x)}{\rho} \\ &\leq \frac{M(\rho + |x - z|, z) - u(x)}{\rho} \leq C\left(1 + \frac{|x - z|}{\rho} - \frac{u(x)}{\rho}\right). \end{aligned}$$

The claim follows by letting $\rho \to \infty$. □

3. The infinite half-space

Here $\Omega \subset \mathbb{R}^n$ is the half-space $H = \{x \in \mathbb{R}^n : x_n > 0\}$ and H_0 is the hyperplane $x_n = 0$. Also $u \geq 0$ is ∞-harmonic in H and vanishes continuously on H_0. By the Harnack inequality, $u > 0$ in H. We will prove that $u(x) = Cx_n$ in H for some $C > 0$. This will be the consequence of several lemmas. For $P \in \mathbb{R}^n$, it follows that $\partial(r, P) = \partial\Omega \cap \overline{B}(r, P) = H_0 \cap \overline{B}(r, P)$ and $E(r, P) = \partial B(r, P) \cap \overline{\Omega} = \partial B(r, P) \cap \overline{H}$. Thus $\partial\Omega(r, P) = \partial(r, P) \cup E(r, P)$. If $\Omega(r, P)$ is not empty, then by the maximum principle

(4) $$M(r, P) = \sup_{\Omega(r,P)} u(x) = \sup_{\partial\Omega(r,P)} u(x) = \sup_{E(r,P)} u(x) > 0.$$

If $P \in H$, then $M(0, P) = u(P)$. We introduce some additional notation. For $S \in H_0$, let

$$T(S) = \{x : x = (S_1, S_2, \dots, S_{n-1}, t) = (S', t) \quad \text{for } t \geq 0\}$$

be the straight ray in H, starting at S and parallel to the x_n-axis. Set $u(S+t) = u(S + t\vec{e}_n)$ for $t > 0$, and let $B(\theta, S+\theta)$ be the ball of radius $\theta > 0$ centered at $S + \theta\vec{e}_n$.

Lemma 3.1. *Let $u > 0$ be ∞-harmonic in H such that $u|_{H_0} = 0$, and let $x_n > 0$. Then for every $S \in H_0$*

(i) $u(S+x_n)/x_n$ *is decreasing in x_n and* $\lim_{x_n \uparrow \infty} u(S+x_n)/x_n = L(S, \infty) < \infty$,

(ii) $0 < \lim_{x_n \downarrow 0} u(S+x_n)/x_n = L(S, 0) < \infty$, *and*

(iii) $0 < L(S, \infty) \leq L(S, 0)$.

Moreover, there is a positive number L such that $L(S, \infty) = L$ for all $S \in H_0$.

Proof. Let $S \in H_0$ and, for $x_n > 0$, consider the ball $B = B(x_n, S+x_n)$. If $0 < y_n < x_n$ then $S+y_n$ and $S+x_n$ lie in $T(S)$. Also $y_n = \operatorname{dist}(S+y_n, \partial B)$ and $x_n = \operatorname{dist}(x_n, \partial B)$. Monotonicity (2) implies that $u(S+x_n)/x_n \leq u(S+y_n)/y_n$. Thus the first assertion follows and implies the second. Except for the finiteness of $L(S, 0)$, the third assertion follows from the first two. To show that $L(S, 0)$ is finite, consider the function $v(x) = M(1, S)|x - S|$ in $\Omega(1, S)$. Clearly $v(x) \geq u(x)$ on $\partial\Omega(1, S)$. By comparison, $u(x) \leq v(x)$ in $\Omega(1, S)$ and $u(S+x_n) \leq M(1, S)x_n$ for $0 \leq x_n \leq 1$. Thus $0 < L(S, 0) \leq M(1, S) < \infty$. We now show that the $L(S, \infty)$ are all equal. Take $x_n > |S|$, then $S+x_n \in B(x_n, O+x_n)$ and $\operatorname{dist}(S+x_n, \partial B(x_n, O+x_n)) = x_n - |S|$. By (2), $u(O + x_n)/x_n \leq u(S + x_n)/(x_n - |S|)$. Then $L(O, \infty) \leq L(S, \infty)$ by letting $x_n \to \infty$. Switch S with O to get equality. We employ the boundary Harnack inequality (3) to show that $L > 0$. We select $u_1(x) = u(x)$ and $u_2(x) = x_n$. For all $s > 0$, the cylinder $C(8s, O)$ is contained in H; (3) then implies that

$$M_1 \frac{u(O + 2s\vec{e}_n)}{2s} \leq \frac{u(x)}{x_n} \leq M_2 \frac{u(O + 2s\vec{e}_n)}{2s} \quad \text{for all } x \in C(s, O).$$

Take $s > 0$ large and fix $t \in (0, s)$. The preceding inequalities yield, for $x = O + t\vec{e}_n$,

$$M_1 L = \lim_{s \to \infty} M_1 \frac{u(O + 2s\vec{e}_n)}{2s} \leq \frac{u(O + t\vec{e}_n)}{t} \leq \lim_{s \to \infty} M_2 \frac{u(O + 2s\vec{e}_n)}{2s} = M_2 L.$$

Letting $t \to 0$, it follows $M_1 L \leq L(O, 0) \leq M_2 L$. It is clear that this estimate holds for every $S \in H$. □

Remark 3.2. Lemma 3.1 implies that $M_1 L \le b = \sup_{S \in H_0} L(S, 0) \le M_2 L < \infty$. By the first part of Lemma 3.1, we have $0 < L x_n \le u(x) \le b x_n$ for $x \in H$.

This remark and Lemma 2.3 imply that u is globally Lipschitz continuous in H. Thus there exists $K > 0$, independent of $x, y \in H$, such that

$$|u(x) - u(y)| \le K|x - y| \text{ for } x, y \in H. \tag{5}$$

We now study u on infinite strips in H. For $a > 0$, define $H_a = \{x : x_n = a\}$, $H(a) = \{x : 0 < x_n < a\}$, and $\partial H(a) = H_0 \cup H_a$. Define

$$\mu(a) = \sup\nolimits_{H_a} u(x) > 0 \quad \text{and} \quad F(a) = \sup\nolimits_{H(a)} u(x) > 0. \tag{6}$$

By Remark 3.2 and (5), $\mu(a)$ is bounded, and $F(a)$ is bounded and increasing.

Lemma 3.3. *Let $u > 0$ be as in Lemma 3.1. If $\mu(a)$ and $F(a)$ are as defined in* (6),

$$\mu(a) = F(a) \quad \textit{and} \quad \mu(a) = \Lambda a,$$

where $\Lambda = \mu(1)$ and $a > 0$ is arbitrary.

Proof. By the Harnack inequality (1), $F(a)$ cannot be attained in the interior of $H(a)$. If $F(a) > \mu(a)$, then there is a sequence $\{P_m\}_{m=1}^{\infty}$ such that $0 < x_n(P_m) < a$ for all m, $|P_m| \to \infty$ and $u(P_m) \to F(a) > \mu(a)$. We argue by contradiction. For each m, let $Q_m = (P'_m, a) \in H_a$, then $u(Q_m) \le \mu(a)$. By (5), we see that $u(P_m) - \mu(a) \le u(P_m) - u(Q_m) \le K\big(a - x_n(P_m)\big)$ for all m. Thus, for large m,

$$a - x_n(P_m) \ge \frac{u(P_m) - u(Q_m)}{K} \ge \frac{3}{4}\,\frac{F(a) - \mu(a)}{K} = A > 0. \tag{7}$$

For $0 < \theta < 1$, let $R_m(\theta) = (P'_m, a - \theta A) \in H_{a-\theta A} \subset H(a)$. From (5), we see that $u(P_m) \le K x_n(P_m)$ and so, for all m,

$$\frac{F(a)}{K} \le \liminf_{k\to\infty} x_n(P_k) \le \limsup_{k\to\infty} x_n(P_k) \le a - A < x_n(R_m(\theta)) = a - \theta A. \tag{8}$$

Fix θ; then (7) and (8) imply that, for large m,

$$0 < x_n(P_m) \le a - A < x_n(R_m(\theta)) = a - \theta A,$$

$$\limsup_{k\to\infty} \big|P_k - R_k(\theta)\big| = \limsup_{k\to\infty} \big(x_n(R_k(\theta)) - x_n(P_k)\big) \le (a - \theta A) - F(a)/K,$$

and

$$\min\big(\operatorname{dist}(P_m, \partial H(a)), \operatorname{dist}(R_m(\theta), \partial H(a))\big) \ge \min(\theta A, F(a)/K) = B > 0.$$

By applying the Harnack inequality to the function $v(x) = F(a) - u(x)$, we now see that

$$
\begin{aligned}
0 &= \lim_{m\to\infty} \big(F(a) - u(P_m)\big) \\
&\geq \exp\left(\frac{-\big((a-\theta A) - F(a)/K\big)}{B}\right) \lim_{m\to\infty} \big(F(a) - u(R_m(\theta))\big) \geq 0.
\end{aligned}
$$

Thus $u(R_m(\theta)) \to F(a)$ as $m \to \infty$. We show that this, together with (5), leads to a contradiction. Let $X_m(\theta) = (R_m(\theta)', a)$; then (7) holds. That is, for large m,

$$
\theta A = a - x_n(R_m(\theta)) \geq \frac{u(R_m(\theta)) - u(X_m(\theta))}{K} \geq \frac{3}{4}\,\frac{F(a) - \mu(a)}{K} = A > 0.
$$

Thus $F(a) = \mu(a)$.

We now prove that $\mu(a) = \Lambda a$. Let $b > a$; we show that $\mu(a)/a \geq \mu(b)/b$. By the first part of Lemma 3.1, $u(S + a\vec{e}_n)/a \geq u(S + b\vec{e}_n)/b$ for all $S \in H_0$. Now take the supremum of both sides. We claim that $\mu(a)$ is convex in a. Let $S \in H_0$ and, for $0 \leq s < t$, set $r = (s+t)/2$. Consider

$$
v_{s,t}(x) = \mu(s) + \big(\mu(t) - \mu(s)\big)\,\frac{|x-S| - s}{t - s} \geq 0,
$$

for all $x \in \Omega(t, S) \setminus \overline{\Omega}(s, S)$. Using the equality $F(a) = \mu(a)$ and the cone comparison we see that $u \leq v_{s,t}$ in $\Omega(t, S) \setminus \overline{\Omega}(s, S)$. Now we take $x = S + r\vec{e}_n$ to see that $\mu(r) = \sup_{S\in H_0} u(S + r\vec{e}_n) \leq \frac{1}{2}\big(\mu(s) + \mu(t)\big)$. Convexity follows. Since $\mu(0) = 0$, we see that $\mu(a)/a$ is both increasing and decreasing as a function of a — in other words, it is constant. $\square$

Proof of Theorem 1.1. It is clear that Theorem 1.1 would follow if we could show that $\Lambda = L$. For $Q \in H_0$ and $r > 0$, set $P_Q(r) = (Q', r)$. For $0 \leq \varepsilon < \Lambda$, let $Q = Q(\varepsilon) \in H_0$ be such that $u(P_{Q(\varepsilon)}(1)) \geq \Lambda - \varepsilon = \mu(1) - \varepsilon$. We fix ε and Q, and suppress the argument ε. By Lemma 3.1(i) we have $u(P_Q(1)) \leq u(Q + x_n)/x_n$ for $0 < x_n < 1$. Thus

$$
u(P_Q(x_n)) = u(Q + x_n) \geq (\Lambda - \varepsilon)x_n \quad \text{for } 0 < x_n \leq 1. \tag{9}
$$

Since $M(0, Q) = 0$, Lemma 2.2(i) and equation (4) imply that $M(r, Q)/r$ is increasing. From (9)

$$
M(1, Q) \geq u(P_Q(1)) \geq \Lambda - \varepsilon \quad \text{and} \quad M(r, Q) \geq (\Lambda - \varepsilon)r \quad \text{for } r \geq 1. \tag{10}
$$

For $r > 0$, define $t = t(\varepsilon, r)$ by $\mu(t) = \Lambda t = (\Lambda - \varepsilon)r$. Let $T = T(r, \varepsilon) \in \partial B(r, Q)$ be such that $u(T) = M(r, Q)$. By Lemma 3.3 and (10), $u(x) < \mu(t)$ for all $x \in H(t)$; moreover

$$
T \in \partial B(r, Q) \cap \{x : x_n \geq t\} \quad \text{and} \quad x_n(T) \geq t = \left(1 - \frac{\varepsilon}{\Lambda}\right) r. \tag{11}
$$

Let $\Upsilon = \Upsilon(\varepsilon, Q)$ be the interior of the cone with vertex Q, axis parallel to $\vec{e}_n$ and half-angle $\theta = \theta(\varepsilon, Q) = \cos^{-1}(1-\varepsilon/\Lambda)$. Clearly (11) implies that T lies in the intersection $\overline{\Upsilon} \cap \partial B(r, Q)$. Since the points $P_Q(r)$ and T lie on $\partial B(r, Q)$, the arc length of $P_Q(r)T$, along a great circle, is at most θr. The distance to H_0 is at least t. Applying the Harnack inequality (1) to $u(P_Q(r))$ and $u(T) = M(r, Q)$, and using (10), we see that for $r > 1$,

$$u(P_Q(r)) \geq \exp\Big(-\frac{\theta}{1-\varepsilon/\Lambda}\Big) M(r, Q) \geq \exp\left(-\frac{\cos^{-1}(1-\varepsilon/\Lambda)}{1-\varepsilon/\Lambda}\right)(\Lambda-\varepsilon)r,$$

and so

$$L = \lim_{r\uparrow\infty} \frac{u(Q+r\vec{e}_n)}{r} = \lim_{r\uparrow\infty} \frac{u(P_Q(r))}{r} \geq \exp\left(-\frac{\cos^{-1}(1-\varepsilon/\Lambda)}{1-\varepsilon/\Lambda}\right)(\Lambda-\varepsilon).$$

Since this holds for all $\varepsilon > 0$, it follows that $L \geq \Lambda$ and $u(x) = \Lambda x_n$ for $x_n > 0$. $\square$

4. The exterior of a ball

Let $\Omega = \mathbb{R}^n \setminus \overline{B}(1, O)$, and assume that $u > 0$ and $u|_{\partial\Omega} = 0$. We prove that $u = K(|x|-1)$ for some $K > 0$. For $r > 1$ set

$$\mu(r) = \sup_{|x|=r} u(x) = \sup_{B(r,O)\setminus B(1,O)} u(x),$$
$$m(r) = \inf_{|x|=r} u(x) > 0.$$

(On the first line we have used the maximum principle.) Clearly, $\mu(1) = m(1) = 0$. Let S^{n-1} be the unit sphere in $\mathbb{R}^n$, and for $t > 1$ and $\omega \in S^{n-1}$ set $\Delta(t, \omega) = u(t\omega)/(t-1)$.

Lemma 4.1. *Let $u > 0$ be ∞-harmonic in $\mathbb{R}^n \setminus B(1, O)$ and $u|_{\partial B(1,O)} = 0$. Let μ, m and Δ be as defined above. Then*

(i) *$u(t\omega)/(t-1)$ decreases as t increases, and $\lim_{t\to\infty} u(t\omega)/(t-1) = L(\omega) > 0$ for all $\omega \in S^{n-1}$;*

(ii) *$\mu(2) = \mu(t)/(t-1)$ for all $t > 1$, and $L(\omega) \leq \mu(2)$ for all $\omega \in S^{n-1}$;*

(iii) *$m(t)/(t-1)$ decreases as t increases, and $L(\omega) \geq \lim_{t\to\infty} m(t)/(t-1) \geq e^{-\pi}\mu(2) > 0$, for all $\omega \in S^{n-1}$;*

(iv) *there exists $K > 0$ such that, if $\min(t_1,t_2) > 4$, $\max(t_1,t_2) < 10\min(t_1,t_2)$, and $\alpha = \cos^{-1}\langle\omega_1, \omega_2\rangle$ for $\omega_1, \omega_2 \in S^{n-1}$, then*

$$\big|\Delta(t_1, \omega_1) - \Delta(t_2, \omega_2)\big| \leq K\Big(\frac{|t_1-t_2|}{t_2-1} + \alpha\Big).$$

Proof. Parts (i), (ii) and (iii) follow from (2) and are interrelated. Fix $\omega \in S^{n-1}$; for $t > 1$, consider the ball $B(t-1, t\omega)$. If $1 < s < t$ then $\operatorname{dist}\big(s\omega, \partial B(t-1, t\omega)\big) = s-1$, and (2) implies $u(s\omega)/(s-1) \ge u(t\omega)/(t-1)$. Thus $L(\omega) \ge 0$ exists (for positivity see part (iii)). Taking the supremum over ω on both sides, we see that $\mu(t)/(t-1)$ decreases as t increases. Similarly, $m(t)/(t-1)$ decreases as t increases. By Lemma 2.2, $\mu(t)$ is convex in $t \ge 1$ and since $\mu(1) = 0$, $\mu(t)/(t-1)$ increases as t increases. Thus $\mu(t)$ is linear in $t-1$ and part (ii) follows. Let $P(t), Q(t) \in \partial B(t, O)$ be such that $u(P(t)) = \mu(t)$ and $u(Q(t)) = m(t)$. The arc length of $P(t)Q(t)$ along a great circle does not exceed πt, and $\operatorname{dist}\big(P(t)Q(t), \partial B(1, O)\big) = t-1$. Applying the Harnack inequality (1),

$$\frac{u(t\omega)}{t-1} \ge \frac{m(t)}{t-1} = \frac{u(Q(t))}{t-1} \ge \frac{u(P(t))}{t-1} \exp\left(\frac{-\pi t}{t-1}\right) = \mu(2) \exp\left(\frac{-\pi t}{t-1}\right).$$

Part (iii) follows by letting $t \to \infty$. To see (iv), fix ω_1 and ω_2 in S^{n-1} and let $0 \le \alpha = \cos^{-1}\langle \omega_1, \omega_2\rangle \le \pi$. Take $\min(t_1, t_2) > 4$ and $t_1 \le t_2 \le 10t_1$. The distance from $t_1\omega_1$ to $t_2\omega_2$ is estimated by going from $t_1\omega_1$ to $t_1\omega_2$ along a great circle, and then from $t_1\omega_2$ to $t_2\omega_2$. Setting $\delta = |t_1 - t_2| + t_1\alpha$ and $d = t_1 - 1$ (the distance to the boundary), the Harnack inequality implies $u(t_1\omega_1) \le e^{\delta/d} u(t_2\omega_2)$. Set $J = \max\big(u(t_1\omega_1), u(t_2\omega_2)\big)$; then, for some $K_1 = K_1(\mu(2)) > 0$ and $K = K(\mu(2)) > 0$,

$$\begin{aligned}\Delta(t_1, \omega_1) - \Delta(t_2, \omega_2) &= \frac{u(t_1\omega_1)(t_2-t_1)}{(t_1-1)(t_2-1)} + \frac{u(t_1\omega_1) - u(t_2\omega_2)}{t_2-1}\\ &\le \mu(2)\,\frac{|t_2-t_1|}{t_2-1} + J\,\frac{e^{\delta/d}-1}{t_2-1}\\ &\le K_1\left(\frac{|t_2-t_1|}{t_2-1} + \frac{\delta}{d}\right) \le K\left(\frac{|t_2-t_1|}{t_2-1} + \alpha\right),\end{aligned}$$

which proves part (iv). □

Remark 4.2. From Lemma 4.1, if $L = \inf_{\omega \in S^{n-1}} L(\omega)$, then $L \le u(x)/(|x|-1) \le \mu(2)$ for all $x \in \Omega$. Also $|L(\omega_1) - L(\omega_2)| \le \mu(2)(e^{\alpha} - 1) \le C|\omega_1 - \omega_2|$.

Remark 4.3. As in Section 3, there is a ray through O on which $\mu(t)$ is attained for every $t > 1$. To see this, let $P(t) \in \partial B(t, O)$ be such that $u(P(t)) = \mu(t)$, and let $\omega(t) = P(t)/|P(t)|$. Since S^{n-1} is compact, there is a sequence $\{t_m\}_{m=1}^{\infty}$ so that $t_m \uparrow \infty$, $\omega(t_m) \to \omega_0$, and $\theta_m = \cos^{-1}\langle \omega(t_m), \omega_0\rangle \to 0$ as $m \uparrow \infty$. Setting $Q_m = t_m\omega_0$, the Harnack inequality (1) and Lemma 4.1 imply that

$$\mu(2) = \frac{u(P(t_m))}{t_m-1} \ge \frac{u(Q_m)}{t_m-1} \ge \frac{u(P(t_m))}{t_m-1} \exp\left(\frac{-\theta_m t_m}{t_m-1}\right) = \mu(2) \exp\left(\frac{-\theta_m t_m}{t_m-1}\right)$$

for all m. Clearly $L(\omega_0) = \mu(2)$ and the claim follows.

We now prove that $u(t\omega) = L(\omega)(t-1)$ for all $\omega \in S^{n-1}$ and all $t > 1$ (see Lemma 4.5). This depends on a comparison result, Lemma 4.4, involving u and a scaled version of u, and uses the fact that the $\Delta(t, \omega)$, for different values of

ω, are comparable at large values of t. Now some notation for Lemma 4.4: fix $P \in \partial B(1, O)$. For $\omega \in S^{n-1}$, let $R_1 = R_1(\omega)$ be the ray $\{O + s\omega, s \geq 0\}$, and $R_2 = R_2(\omega)$ be the ray $\{P + s\omega, s \geq 0\}$; also set $Q = Q(\omega) = O + \omega \in \partial B(1, O)$. For $x \in \Omega$, let $\omega = \omega(x) = (x - P)/|x - P|$, then $x = P + |x - P|\omega(x) \in R_2(w(x))$. We define $y = y(x, P) = Q(\omega(x)) + (x - P)$, so $|y - Q| = |x - P|$. We scale x as follows: for $\theta > 1$, set $x_\theta = x_\theta(P) = P + \theta(x - P)$ and $y_\theta = y_\theta(x, P) = Q + \theta(x - P)$. Then $y, y_\theta \in R_1(\omega(x))$ and $x, x_\theta \in R_2(\omega(x))$, and $y - x = y_\theta - x_\theta = Q - P$; thus $|y - x| = |y_\theta - x_\theta| \leq 2$. Now set $u_\theta(x) = u_\theta(x, P) = u(x_\theta) = u\bigl(P + \theta(x - P)\bigr)$. Clearly for fixed P and $\theta > 0$, $u_\theta(x)$ is ∞-harmonic.

Lemma 4.4. *Let $u > 0$ be as in Lemma 4.1 and $P \in \partial B(1, O)$. For $\theta > 1$ and $x \in \Omega$, let $\omega(x)$, R_1, R_2, Q, x_θ, and y_θ be as defined above. Set $u_\theta(x) = u_\theta(x, P) = u(x_\theta)$; if $1 < s < \theta$, then $u_\theta(x) \geq su(x)$ for all $x \in \Omega$. Then $u_\theta(x) \geq \theta u(x)$ for all $x \in \Omega$.*

Proof. This is done in four steps. Fix P and $\theta > 1$. We show that there exists $\rho > 1$ such that $u_\theta(x) \geq su(x)$ for all $x \in \partial B(r, O)$ and $r \geq \rho$. Comparison will then imply the lemma.

Step 1: Properties of u_θ. Clearly the set

$$Z_\theta = \{x : u_\theta(x) = 0\} = \Bigl\{|x - (1 - 1/\theta)P| = 1/\theta\Bigr\}$$

lies in $\bar{B}(1, O)$. Thus $u_\theta(x) \geq 0$ on $\partial B(1, O)$. Since $|Q| = |P| = 1$, we have

(a) $\theta|x - P| - 1 \leq |x_\theta| \leq \theta|x - P| + 1$,

(b) $|y_\theta| = \theta|x - P| + 1$, and

(c) $|x - P| - 1 \leq |x| \leq |x - P| + 1$.

Thus $\operatorname{dist}(y_\theta, \partial B(1, O)) = |y_\theta| - 1 \approx \operatorname{dist}(x_\theta, \partial B(1, O)) = |x_\theta| - 1$ when $|x|$ is large. From (a), (b) and the Harnack inequality,

$$u(y_\theta)\ \exp\left(-\frac{|y_\theta - x_\theta|}{\theta|x - P|}\right) \leq u_\theta(x) = u(x_\theta) \leq u(y_\theta)\ \exp\left(\frac{|y_\theta - x_\theta|}{\theta|x - P| - 2}\right).$$

Fix ω and select $x, x_\theta \in R_2(\omega)$, and $y, y_\theta \in R_1(\omega)$. Divide by $|y_\theta| - 1$ and note that (a), (b), (c) and Lemma 4.1 imply that

$$\lim_{x\to\infty} \frac{u_\theta(x)}{|x| - 1} = \theta L(\omega) \quad \text{and} \quad \lim_{x\to\infty} \frac{u(x)}{|x| - 1} = L(\omega), \tag{12}$$

since $|y_\theta - x_\theta| = |P - Q|$. The second conclusion follows by working similarly with u, x and y.

Step 2. Fix $1 < s < \theta$ and let $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ be such that $\varepsilon_1 + \varepsilon_2 < L(\theta - s)$, where $L = \inf_\omega L(\omega)$. Then $\theta L(\omega) - \varepsilon_1 > sL(\omega) + \varepsilon_2$ for all $\omega \in S^{n-1}$. From (12) there is a $\rho = \rho(\varepsilon_1, \varepsilon_2, s, \omega) > 0$, such that for all $x \in R_2(\omega)$ with $|x| \geq \rho$,

$$\frac{u_\theta(x)}{|x|-1} > \theta L(\omega) - \varepsilon_1 > sL(\omega) + \varepsilon_2 > \frac{su(x)}{|x|-1}. \tag{13}$$

In Step 3, we show that there is a ρ such that (13) holds independently of the choice of $\omega \in S^{n-1}$.

Step 3. We first show that there is a $\rho > 0$ such that the first inequality in (13) holds for all ω. Recall that $\omega = \omega(v) = (v - P)/|v - P|$ for $v \in \partial B(r, O)$. We prove that the quantity

$$D(\theta, r, v, \omega, \varepsilon_1) = \frac{u_\theta(v)}{|v|-1} - (\theta L(\omega) - \varepsilon_1)$$

is continuous in ω and positive for large r, for all $v \in \partial B(r, O)$. Let $\omega_1 \in S^{n-1}$ and let x lie in $R_2(\omega_1) \cap \partial B(r, O)$, with $r > 20$; take ω_2 close to ω_1 and let z lie in $R_2(\omega_2) \cap \partial B(r, O)$. By Remark 4.2, $\theta|L(\omega_2) - L(\omega_1)|$ is small. Clearly, $\max\bigl(|z-P|, |x-P|\bigr) \leq r+1$. Noting that $z_\theta = P + \theta(z - P)$ and $x_\theta = P + \theta(x - P)$, we see that $|z_\theta - x_\theta| = \theta|x - z| \leq \theta(r+5)\alpha$, where $\alpha = \cos^{-1}\langle\omega_1, \omega_2\rangle$. From (a) and (c) in Step 1, we see that, for large r,

$$\theta(r-1) \approx \max(|x_\theta| - 1, |z_\theta| - 1) \geq \min(|x_\theta| - 1, |z_\theta| - 1) \approx \theta(r-1).$$

By the Harnack inequality, $u_\theta(z) = u(z_\theta) \geq u_\theta(x)\exp\bigl(-\alpha(r+5)/(r-1)\bigr)$. Thus $u_\theta(x) \leq u_\theta(z)e^{25\alpha/19}$. By Remark 4.2, $u_\theta(v) = u(v_\theta) \leq \mu(2)(|v_\theta| - 1)$. Thus (12) yields

$$\frac{|u_\theta(x) - u_\theta(z)|}{|x|-1} \leq \frac{\sup(u_\theta(x), u_\theta(z))}{|x|-1}(e^{25\alpha/19} - 1) \leq \theta\mu(2)(e^{25\alpha/19} - 1). \tag{14}$$

This and Remark 4.2 yield, for some $C > 0$ independent of r,

$$\bigl|D(\theta, r, x, \omega_1, \varepsilon_1) - D(\theta, r, z, \omega_2, \varepsilon_1)\bigr| \leq C\theta\alpha \tag{15}$$

for $x, z \in \partial B(r, O)$ and $r > 20$. Fix ω_1, and let ρ be such that $D(\theta, r, x, \omega_1, \varepsilon_1) > \frac{1}{2}\varepsilon_1$ for all $r \geq \rho$ and $x \in R_2(\omega_1) \cap \partial B(r, O)$. By (15), in a fixed small ω-neighborhood, positivity of D persists. The conclusion follows by the compactness of S^{n-1}.

We now discuss the second inequality in (12). Let $x \in R_2(\omega_1) \cap \partial B(r, O)$ and $z \in R_2(\omega_2) \cap \partial B(r, O)$ with α small. Clearly

$$|u(x) - u(z)| \leq \max\bigl(u(x), u(z)\bigr)e^{(r+5)\alpha/(|r-1)}.$$

Applying Remark 4.2, Lemma 4.1 and selecting $r \geq \rho$ to be large, we see that

$$\left|\left(sL(\omega_1)+\varepsilon_2-\frac{su(x)}{|x|-1}\right)-\left(sL(\omega_2)+\varepsilon_2-\frac{su(z)}{|z|-1}\right)\right|$$
$$\leq s\frac{|u(x)-u(z)|}{|x|-1}+s|L(\omega_1)-L(\omega_2)| \leq s\mu(2)(e^{2\alpha}-1)+Ks\alpha.$$

We again use the compactness of S^{n-1}.

Step 4. From Step 1, $u_\theta \geq 0 = su$ on $\partial B(1, O)$. From Step 3, with ε_1 and ε_2 as in Step 2, and for all $r \geq \rho(\varepsilon_1, \varepsilon_2)$, we see that $u_\theta > su$ in $\partial B(r, O)$. By comparison, $u_\theta \geq su$ in $B(\rho, O) \setminus \bar{B}(1, O)$. This holds in all of Ω, and for all $1 < s < \theta$. Thus $u_\theta(x) \geq \theta u(x)$ for all $x \in \Omega$. □

Next we show that Lemma 4.4 implies that u is linear along rays through O.

Lemma 4.5. *For every $\omega \in S^{n-1}$, let $T(\omega)$ be the ray $\{O+s\omega : s \geq 0\}$, and let $P = O+\omega$. Let $\theta > 1$, and let u and u_θ be as in Lemma 4.4. Then $u(x) = L(\omega)(|x|-1)$ for all $x \in T(\omega)\cap\Omega$.*

Proof. Fix $x, y \in T(\omega)\cap\Omega$ with $|x-P| < |y-P|$, and define $\theta = |y-P|/|x-P|$. Then $y = P+\theta(x-P) = x_\theta$, and so $u_\theta(x) = u(x_\theta) = u(y)$. By Lemma 4.4,

$$u(y) = u(x_\theta) \geq \theta u(x) = \frac{|y-P|}{|x-P|}u(x), \quad \text{hence} \quad \frac{u(y)}{|y-P|} \geq \frac{u(x)}{|x-P|}.$$

Since $|x-P| = |x|-1$ and $|y-P| = |y|-1$, Lemma 4.1(i) implies equality in these equations. Since x and y are arbitrary, $u(z) = L(\omega)(|z|-1)$ for all $z \in T(\omega)$. □

We set $\Omega_a = \{x \in \Omega : u(x) < a\}$ for $a > 0$ and show, using Lemma 4.5, that $\bar{B}(1, O)\cup\Omega_a$ is convex. For $\omega \in S^{n-1}(O)$ and $t > 1$, set $Q = Q(t, \omega) = O+t\omega$. Define the hyperplane $H_t = H_t(\omega) = \{x : \langle x-Q, \omega\rangle = 0\}$, and the half-planes $H_t^+ = H_t^+(\omega) = \{x : \langle x-Q, \omega\rangle > 0\}$ and $H_t^- = H_t^-(\omega) = \{x : \langle x-Q, \omega\rangle < 0\}$. Then $\bar{B}(1, O) \subset H_t^-(\omega)$. For $a > 0$, let $t(a) = t(a, \omega) = 1+a/L(\omega)$, and let $Q_a = Q_a(\omega) = O+t(a)\omega$.

Lemma 4.6. *For $a > 0$, let $S(a) = \Omega_a \cup \bar{B}(1, O)$. Then*

(i) $u(x) \geq L(\omega)(\langle x, \omega\rangle - 1) \geq L(\omega)(t-1)$ *for all* $x \in H_t^+(\omega)$,

(ii) $S(a) = \bigcap_{\omega\in S^{n-1}} H_{t(a)}^-(\omega)$ *and* $H_{t(a)}(\omega)$ *is a supporting hyperplane to* $S(a)$ *at* Q_a, *and* $OQ_a \perp H_{t(a)(\omega)}$, *for all* $\omega \in S^{n-1}$. *Clearly,* $S(a)$ *is convex.*

Proof. By Lemma 4.5, $u(Q) = L(\omega)(|Q|-1)$ for all $\omega \in S^{n-1}$. Also, Q_a lies in $H_{t(a)}(\omega)$, and $u(Q_a) = a$. To prove part (i), set $R(\omega) = \{O+s\omega, s > 0\}$ and let $r \geq t$. Fix $x \in H_r(\omega) \subset H_t^+(\omega)$, and choose $P \in R(\omega)$, with $|P|$ large, such that

$|P|-1>|P-x|$. Then x lies in $B(|P|-1, P)$, and by applying monotonicity (2) along the ray Px, we see that

$$L(\omega)=u(P)/(|P|-1)\leq \lim_{|P|\to\infty} u(x)/(|P|-1-|x-P|)=u(x)/(r-1)$$

for $P\in R(\omega)$. Since $r=\langle x,\omega\rangle>t$, part (i) follows. We now prove part (ii). Fix ω. Then, by Lemma 4.5, $u(t\omega)<u(Q_a)=a$ whenever $0\leq t<t(a)=1+a/L(\omega)$. If $A(a)=\bigcup_{\omega\in S^{n-1}}\{t\omega:1<t<t(a)\}$, then $u(x)<a$ for all $x\in A(a)$. We show that $\Omega_a=A(a)$. Clearly, $A(a)\subset\Omega_a$, so suppose $x\notin A(a)$, and set $\omega=x/|x|$. Then $x=s\omega$ for some $s\geq t(a,\omega)$, and $x\in H_{t(a)}(\omega)\cup H^+_{t(a)}(\omega)$. By part (i), $u(x)\geq a$ and hence $A(a)=\Omega_a$. Also $A(a)\cap(H_{t(a)}(\omega)\cup H^+_{t(a)}(\omega))=\varnothing$ for all ω, implying that $A(a)\subset H^-_{t(A)}(\omega)$. As $x\notin A(a)$ implies that $x\notin H^-_{t(a)}(\omega)$ for some ω, $A(a)=\bigcap_\omega\big(H^-_{t(a)}\setminus\bar{B}(1,O)\big)$. By part (i), $\partial A(a)\cap H_{t(a)}(\omega)=Q_a(\omega)$. Thus $S(a)$ is convex and $\partial S(a)=\bigcup_\omega\{Q_a(\omega)\}=\bigcup_\omega\{\big(1+a/L(\omega)\big)\omega\}$. Clearly $H_{t(a)}$ is the supporting hyperplane at every $Q_a\in\partial S(a)$. By the definition of $H_{t(a)}$ it follows that $OQ_a(\omega)\perp H_{t(a)}(\omega)$ for all $\omega\in S^{n-1}$. □

We now show that Lemma 4.6 implies that Ω_a is a ball.

Proof of Theorem 1.2. Let $F:\mathbb{R}^+\times S^{n-1}\to\mathbb{R}^n$ by $F(a,\omega)=O+(1+a/L(\omega))\omega$. Then by Lemmas 4.1 and 4.6, for $a>0$ fixed, F is a bijective Lipschitz map, and $F(S^{n-1})=\partial\Omega_a$. Thus $\partial\Omega_a$ is connected and $\vec{F}(\omega)\perp H_{t(a)}(\omega)$. Let $\omega_1,\omega_2\in S^{n-1}$, then $Q_1=Q_a(\omega_1)$ and $Q_2=Q_a(\omega_2)$ lie on $\partial\Omega_a$. Let Π be the two-dimensional plane containing O, ω_1 and ω_2, and C be $\partial B(1,O)\cap\Pi$. Note that Q_1 and Q_2 lie in Π. Let $\tau(s)\in\partial B(1,O)\cap\Pi$ be a smooth parametrization of C such that $\tau(0)=\omega_1$ and $\tau(1)=\omega_2$. The curve $\sigma(s)=F(\tau(s))=\big(1+a/L(\tau(s))\big)\tau(s)$ in $\Pi\cap\partial\Omega_a$ is Lipschitz continuous in s, and $\sigma(0)=Q_1$ and $\sigma(1)=Q_2$. Let $s_0\in[0,1]$ be a point of differentiability of $\sigma(s)$. Call $\Sigma(s)=H_{t(a)}(\tau(s))$; by Lemma 4.6, $\Sigma(s_0)$ is the supporting hyperplane at $\sigma(s_0)$. Furthermore, $\Sigma(s_0)$ is perpendicular to $\sigma(s_0)$, and $\sigma(s)\in H^-_{t(a)}(\tau(s_0))\cap\Pi$ for all s. Since s_0 is a point of differentiability, a simple argument shows that $\sigma'(s_0)$ lies in $\Sigma(s_0)\cap\Pi$, and hence $\sigma(s_0)\perp\sigma'(s_0)$. Thus $|\sigma(s_0)|'=0$. Since this holds for almost every $s\in[0,1]$, Lipschitz continuity implies that $|Q_2|=|\sigma(1)|=|\sigma(s)|=|\sigma(0)|=|Q_1|$. Thus Ω_a is a ball and $L(\omega)=C$ for all $\omega\in S^{n-1}$. The remainder of the proof follows from Lemma 4.5. □

5. The infinite strip $\{0<x_n<1\}$

Let Ω be the infinite strip $\{x:0<x_n<1\}$, let $H(0)=\{x:x_n=0\}$, and let $H(1)=\{x:x_n=1\}$. We assume that u is ∞-harmonic, that $u\geq 0$ in Ω, and that u vanishes continuously on $H(0)$ and $H(1)$. For $r>0$, define $D(r)$ to be

$\{x : |x'|_{n-1} < r,\ 0 < x_n < 1\}$, where $x' = (x_1, x_2, \ldots, x_{n-1})$ and

$$|x'|_{n-1} = \sqrt{\textstyle\sum_{i=1}^{n-1} x_i^2}.$$

Set $M(r) = \sup_{D(r)} u(x)$, with the understanding that $M(0) = \sup_{\{x'=0\}\cap\Omega} u(x)$. We set $L(r) = \{x \in D(r) : |x'| = r, 0 \le x_n \le 1\}$, the lateral boundary of the cylinder $D(r)$. By the maximum principle, $M(r)$ is attained only on $L(r)$. Let $J(r) \in L(r)$ be such that $M(r) = u(J(r))$. Let $C(r, P)$ denote the truncated cylinder $\{x : |x' - P'|_{n-1} < r, P_n < x_n < P_n + 2r\}$. The function u_e is the extension of u to all of $\mathbb{R}^n$ defined as follows. Set

$$u_e(x', x_n) = \begin{cases} u(x', x_n) & \text{for } \quad 0 \le x_n \le 1, \\ -u(x', -x_n) & \text{for } -1 \le x_n \le 0, \end{cases}$$

and extend periodically with period 2. Then u_e is ∞-harmonic in $\mathbb{R}^n$; see [Bhattacharya 2002].

Step 1. We first observe that there exists a universal constant $K > 0$ such that

$$\min\bigl(x_n(J(r)), 1 - x_n(J(r))\bigr) \ge K \quad \text{for all } r > 0. \tag{16}$$

Let $T = T(r) \in L(r) \cap H(0)$ and consider the cylinder $C(\frac{1}{2}, T) \subset \Omega$. Since $x_n(T) = 0$ and $u > 0$ in $C(\frac{1}{2}, T)$, the boundary Harnack inequality (3) with $s = \frac{1}{16}$, $u_1 = u$, $u_2 = x_n$ and $z = T + \frac{1}{8}\vec{e}_n = (T', \frac{1}{8})$ yields

$$M_1 \frac{u(z)}{1/8} \le \frac{u(x)}{x_n} \le M_2 \frac{u(z)}{1/8} \quad \text{for all } x \in C(\tfrac{1}{16}, T). \tag{17}$$

Let $P = (T', \frac{1}{2})$. Since $|z - P|/z_n = 3$, the Harnack inequality implies that $u(z)e^{-3} \le u(P) \le u(z)e^3$. Thus (17) with new constants M_1 and M_2 yields

$$M_1 u(P) \le \frac{u(x)}{x_n} \le M_2 u(P) \quad \text{for all } x \in C(\tfrac{1}{16}, T). \tag{18}$$

Let $E(T) = \bigl\{x : |x - T|_{n-1} < \frac{1}{16},\ 0 < x_n < \frac{1}{2}\bigr\}$; if $x \in E(T) \setminus C(\frac{1}{16}, T)$ then $|x - P|/x_n \le 16$ and

$$u(P)e^{-16} \le u(x) \le u(P)e^{16}.$$

Then (18), with new M_1 and M_2, implies that

$$M_1 u(P) \le \frac{u(x)}{x_n} \le M_2 u(P) \quad \text{for all } x \in E(T),$$

since $C(\frac{1}{16}, T) \subset E(T)$. From this we get

$$M_1 u(P) x_n(J(r)) \le M(r) \le M_2 u(P) x_n(J(r)),$$

since $J(r) \in L(r) \cap E(J'(r), 0)$. Dividing by $u(P)$ we see that $x_n(J(r)) \geq 1/M_2$. We argue similarly for dist$(J(r), H(1))$, and (16) follows. Note that by the Harnack inequality, $M(r) = u(J(r)) \leq e^{(r+1)/K} u(J(0)) = e^{(r+1)/K} M(0)$. Hence $M(r)$ cannot grow faster than the exponential rate.

Step 2. We now show that $M(r)$ is at least of the order $r^{c \log r}$, for large r and for some $c > 0$. We work with $u_e(x)$; for $r > 0$, let $T(r)$ denote the line through $J(r)$ parallel to the x_n-axis. Clearly,

$$\sup_{\{x:|x'|_{n-1}<r\}} u_e(x) = M(r) \quad \text{and} \quad \inf_{\{x:|x'|_{n-1}<r\}} u_e(x) = -M(r).$$

Let $F(r) = \big(J'(r), 2 - x_n(J(r))\big)$. Then $u(F(r)) = -M(r)$, since u_e arises from the odd reflection of u about $x_n = 1$. Note that $|J(r) - F(r)| \leq 2(1 - K) = \delta$. Since $M(2r) - u_e(x) \geq 0$ in $\{x : |x'|_{n-1} < 2r\}$, applying the Harnack inequality to $u_e(J(r))$ and $u_e(F(r))$, we see that $M(2r) - M(r) \geq e^{-\delta/r}\big(M(2r) + M(r)\big)$, and hence that

$$M(2r) \geq \frac{e^{\delta/r} + 1}{e^{\delta/r} - 1} M(r) \quad \text{for } r > 0. \tag{19}$$

We employ iteration noticing that $(e^{\delta/r} + 1)/(e^{\delta/r} - 1) \uparrow \infty$ as r increases. Let $\xi > 0$, select $R = R(\xi) > 0$ such that $(e^{\delta/r} + 1)/(e^{\delta/r} - 1) > \xi$ for all $r > R$. Then (19) implies that $M(2^m R) \geq \xi^m M(R)$ and $M(r) \geq (r/R)^{\log \xi / \log 2} M(R)/\xi$. Also $M(2^{m+1}\delta) \geq M(2^m \delta)(e^{1/2^m} + 1)/(e^{1/2^m} - 1)$. Take N large, so that $e^{1/2^k} - 1 \leq 2/2^k$ for $k \geq N$. Starting an iteration from N, we get

$$\begin{aligned} M(2^{m+1}\delta) &\geq \left(\prod_{k=N}^{m} \left(1 + \frac{2}{e^{1/2^k} - 1}\right)\right) M(2^N \delta) \geq \left(\prod_{k=N}^{m} (1 + 2^k) \right) M(2^N \delta) \\ &= \left(\prod_{k=N}^{m} 2^k \right)\left(\prod_{k=N}^{m} (1 + 2^{-k}) \right) M(2^N \delta) \geq C(N) 2^{m^2/2} M(2^N \delta). \end{aligned}$$

Since $M(r)$ is increasing, the right side is of the order $r^{c \log r}$, for some universal $c > 0$.

References

[Aronsson et al. 2004] G. Aronsson, M. G. Crandall, and P. Juutinen, "A tour of the theory of absolutely minimizing functions", *Bull. Amer. Math. Soc.* (*N.S.*) **41**:4 (2004), 439–505. MR MR2083637 Zbl 02108961

[Barles and Busca 2001] G. Barles and J. Busca, "Existence and comparison results for fully nonlinear degenerate elliptic equations without zeroth-order term", *Comm. Partial Differential Equations* **26**:11-12 (2001), 2323–2337. MR 2002k:35078 Zbl 0997.35023

[Bhattacharya 2002] T. Bhattacharya, "On the properties of ∞-harmonic functions and an application to capacitary convex rings", *Electron. J. Diff. Equations* 101 (2002), 22 pp. MR 2003j:35126 Zbl 1037.35028

[Bhattacharya 2004] T. Bhattacharya, "On the behaviour of ∞-harmonic functions near isolated points", *Nonlinear Anal.* **58**:3-4 (2004), 333–349. MR 2073529 Zbl 1053.31003

[Bhattacharya et al. 1989] T. Bhattacharya, E. DiBenedetto, and J. Manfredi, "Limits as $p \to \infty$ of $\Delta_p u_p = f$ and related extremal problems", *Rend. Sem. Mat. Univ. Politec. Torino* special issue (1989), 15–68. MR 93a:35049

[Crandall and Evans 2001] M. G. Crandall and L. C. Evans, "A remark on infinity harmonic functions", pp. 123–129 in *Proceedings of the USA-Chile Workshop on Nonlinear Analysis* (Viña del Mar and Valparaiso, 2000), Electron. J. Differ. Equ. Conf. **6**, Southwest Texas State Univ., San Marcos, TX, 2001. MR 2001j:35076 Zbl 0964.35061

[Crandall et al. 1992] M. G. Crandall, H. Ishii, and P.-L. Lions, "User's guide to viscosity solutions of second order partial differential equations", *Bull. Amer. Math. Soc.* (*N.S.*) **27**:1 (1992), 1–67. MR 92j:35050 Zbl 0755.35015

[Crandall et al. 2001] M. G. Crandall, L. C. Evans, and R. F. Gariepy, "Optimal Lipschitz extensions and the infinity Laplacian", *Calc. Var. Partial Differential Equations* **13**:2 (2001), 123–139. MR 2002h:49048 Zbl 0996.49019

[Jensen 1993] R. Jensen, "Uniqueness of Lipschitz extensions: minimizing the sup norm of the gradient", *Arch. Rational Mech. Anal.* **123**:1 (1993), 51–74. MR 94g:35063 Zbl 0789.35008

[Lindqvist and Manfredi 1995] P. Lindqvist and J. J. Manfredi, "The Harnack inequality for ∞-harmonic functions", *Electron. J. Diff. Equations* 5 (1995), 5 pp. MR 96b:35025 Zbl 0818.35033

Received October 21 2003. Revised May 3 2004.

TILAK BHATTACHARYA
DEPARTMENT OF MATHEMATICS
PURDUE UNIVERSITY
WEST LAFAYETTE, IN 47907
tbhatta@math.purdue.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

TRANSVERSAL HOLOMORPHIC SECTIONS AND LOCALIZATION OF ANALYTIC TORSIONS

HUITAO FENG AND XIAONAN MA

We prove a Bott-type residue formula twisted by $\bigwedge(\mathbb{V}^*)$ with a holomorphic vector bundle $\mathbb{V}$, and relate certain analytic torsions on the total manifold to the analytic torsions on the zero set of a holomorphic section of $\mathbb{V}$.

Introduction

Beasley and Witten [2003], studying half-linear models, have described a compactification on any Calabi–Yau threefold Y that is a complete intersection in a compact toric variety X. In particular, a remarkable cancellation involving the instanton effect [Beasley and Witten 2003, (1.3)], involving certain determinants of the $\bar{\partial}$-operator, was derived directly from a residue theorem. One would like to understand its implications in mathematics, for example in Gromov–Witten theory. Bershadsky, Cecotti, Ooguri and Vafa [Bershadsky et al. 1993; 1994] predicted that the analytic torsion of Ray–Singer will play a role regarding the genus-1 Gromov–Witten invariant. Thus we naturally try to understand the results about analytic torsion first.

As an application of [Bismut and Lebeau 1991] and the localization formula (1–3) in this paper, we were able to relate certain analytic torsions on the total manifold with the zero set of a holomorphic transversal section of $\mathbb{V}$, generalizing [Bismut 2004, Theorem 6.6] and [Zhang n.d.] with $\mathbb{V} = TX$ therein. We expect our formula will be useful for understanding [Beasley and Witten 2003, (1.3)] from a mathematical point of view.

This paper is organized as follows. In Section 1 we prove a Bott-type residue formula. In Section 2 we get a localization formula for Quillen metrics. In Section 3 we get a localization formula for analytic torsions under extra conditions. In Section 4, for the reader's convenience, we write down six intermediate results, corresponding to [Bismut and Lebeau 1991, Theorems 6.4-6.9].

MSC2000: 58J52, 32L10, 58J20, 32C35, 57R20.
Keywords: analytic torsion, characteristic classes, characteristic numbers, residue formula.
Feng was partially supported by the NNSF of China (10271059).

1. A Bott-type residue formula

In this section, along the lines of [Bismut 1986, §1], we give a Bott-type residue formula (1–3) by assuming that the holomorphic section is transversal; compare to [Beasley and Witten 2003, (2.32), (2.34)].

Let X be a compact complex manifold with $\dim X = n$ and let $\mathbb{V}$ be a holomorphic vector bundle on X with $\dim \mathbb{V} = l$. We assume that the line bundles $\det TX$ and $\det \mathbb{V}$ are holomorphically isomorphic. We fix a holomorphic isomorphism $\phi : \det \mathbb{V}^* \simeq \det T^*X$, which is clearly unique up to a constant. Thus ϕ defines a map from the $\mathbb{Z}_2$-graded tensor product $\bigwedge(\overline{T^*X}) \widehat{\otimes} \bigwedge(\mathbb{V}^*)$ to $\bigwedge(\overline{T^*X}) \widehat{\otimes} \bigwedge^{\max}(T^*X) \subset \bigwedge(T^*_{\mathbb{R}}X) \otimes_{\mathbb{R}} \mathbb{C}$. We can define the integral of an element α of $\Omega(X, \bigwedge(\mathbb{V}^*))$, the set of smooth sections of $\bigwedge(\overline{T^*X}) \widehat{\otimes} \bigwedge(\mathbb{V}^*)$ on X, by

$$\int_X \alpha = \int_X \phi(\alpha).$$

Let v be a holomorphic section of $\mathbb{V}$ on X. Assume that v vanishes on a complex manifold $Y \subset X$. Then $\nabla v|_Y : TX|_Y \to \mathbb{V}|_Y$ mapping U to $\nabla_U v$ does not depend on the choice of a connection ∇ on $\mathbb{V}$, and $\nabla_U v|_Y = 0$ for $U \in TY$. Let N be the normal bundle to Y in X. Assume also that $\nabla v|_Y : N \to \mathbb{V}|_Y$ is injective, and there is a holomorphic vector subbundle $\mathbb{V}_1$ on Y such that

$$\mathbb{V}|_Y = \mathbb{V}_1 \oplus \operatorname{Im} \nabla v|_Y. \tag{1–1}$$

Let $P^{\mathbb{V}_1}$ and $P^{\operatorname{Im}\nabla v}$ be the natural projections from $\mathbb{V}$ onto $\mathbb{V}_1$ and $\operatorname{Im}\nabla v|_Y$.

Let $i(v)$ be the standard contraction operator acting on $\bigwedge(\mathbb{V}^*)$. A natural question, posed in [Beasley and Witten 2003, §2], is how to express $\int_X \alpha$ using the local data near the zero set Y of v for a $(\bar\partial^X + i(v))$-closed form α, that is, a form satisfying $(\bar\partial^X + i(v))\alpha = 0$.

First we recall an idea due to Bismut [Bismut 1986]; see also [Zhang 1990].

Proposition 1.1. *Let $\alpha \in \Omega(X, \bigwedge(\mathbb{V}^*))$ be a $(\bar\partial^X + i(v))$-closed form. Then*

$$\int_X \alpha = \int_X e^{-(\bar\partial^X + i(v))\omega/t}\alpha \quad \textit{for any } \omega \in \Omega(X, \textstyle\bigwedge(\mathbb{V}^*)) \textit{ and } t > 0.$$

Proof. For any $\omega \in \Omega(X, \bigwedge(\mathbb{V}^*))$

$$\int_X \bar\partial^X \omega = \int_X \phi(\bar\partial^X \omega) = \int_X \bar\partial^X \phi(\omega) = \int_X d\phi(\omega) = 0. \tag{1–2}$$

From $(\bar\partial^X + i(v))^2 = 0$ and $(\bar\partial^X + i(v))\alpha = 0$, we have

$$\frac{\partial}{\partial s} \int_X e^{-s(\bar\partial^X + i(v))\omega}\alpha = -\int_X (\bar\partial^X + i(v))\big(\omega\, e^{-s(\bar\partial^X + i(v))\omega}\alpha\big) = 0,$$

and the desired equality follows. □

Recall that $\nabla v|_Y : N \to \operatorname{Im} \nabla v|_Y$ is an isomorphism that induces isomorphisms of holomorphic line bundles $\phi_N = (\det \nabla v|_Y)^* : \det(\operatorname{Im} \nabla v|_Y)^* \to \det N^*$ and $\phi_Y = \phi|_Y / \big((\det \nabla v|_Y)^*\big) : \det \mathbb{V}_1^* \to \det T^*Y$. These two isomorphisms make the integral $\int_N$ along the normal bundle N and $\int_Y$ well defined.

Let $h^{\mathbb{V}}$ be a Hermitian metric on $\mathbb{V}$ such that $\mathbb{V}_1$ and $\operatorname{Im} \nabla v|_Y$ are orthogonal on Y. Let g_1^N be a Hermitian metric on N such that $\nabla_{\cdot} v|_Y : N \to \operatorname{Im} \nabla v|_Y$ is an isometry. Let $R^{\mathbb{V}}$ be the curvature of the holomorphic Hermitian connection $\nabla^{\mathbb{V}}$ on $(\mathbb{V}, h^{\mathbb{V}})$. Let $j : Y \to X$ be the natural embedding, and $\{Y_j\}_j$ the connected components of Y. On Y, define

$$R_v^{\mathbb{V}} = -(\nabla_{\cdot} v)^{-1} P^{\operatorname{Im} \nabla v} R^{\mathbb{V}}(\,\cdot\,, j_* \cdot) P^{\mathbb{V}_1} \cdot \in \overline{T^*Y} \mathbin{\widehat{\otimes}} \mathbb{V}_1^* \otimes \operatorname{End} N.$$

$R_v^{\mathbb{V}}$ is well defined since $P^{\operatorname{Im} \nabla v} R^{\mathbb{V}}(j_* \cdot, j_* \cdot) P^{\mathbb{V}_1} = 0$. Thus, for $U \in TY$, $W \in \mathbb{V}_1$, $u_1, u_2 \in N$,

$$\big\langle R_v^{\mathbb{V}}(\overline{U}, W) u_1, u_2 \big\rangle_{g_1^N} = -\big\langle R^{\mathbb{V}}(u_1, \overline{U}) W, \nabla_{u_2} v \big\rangle = \big\langle W, R^{\mathbb{V}}(\overline{u_1}, U) \nabla_{u_2} v \big\rangle.$$

Certainly $\det_N((1 + R_v^{\mathbb{V}})/2\pi i)$ is $\bar\partial^Y$-closed.

The following result verifies a formula of Beasley and Whitney [2003, (2.32), (2.34)] and generalizes corresponding results in [Zhang 1990], [Liu 1995] and [Bott 1967].

Theorem 1.2. *For any $(\bar\partial^X + i(v))$-closed form $\alpha \in \Omega(X, \bigwedge(\mathbb{V}^*))$,*

$$\int_X \alpha = \sum_j \int_{Y_j} \frac{(-1)^{(l-n)(n-\dim Y_j)} \alpha}{\det_N\big((1 + R_v^{\mathbb{V}})/(-2\pi i)\big)}. \tag{1–3}$$

Proof. Set

$$S = \langle\,\cdot\,, v \rangle_{h^{\mathbb{V}}} \in C^\infty(X, \mathbb{V}^*).$$

By Proposition 1.1, for any $t \in \,]0, +\infty[$,

$$\int_X \alpha = \int_X e^{-\frac{1}{2t}(\bar\partial^X + i(v))S} \alpha = \int_X e^{-\frac{1}{2t}(\bar\partial^X S + |v|^2)} \alpha. \tag{1–4}$$

Thus, as $t \to 0$, the integral $\int_X \alpha$ is asymptotically equal to $\int_{\mathcal{U}} e^{-\frac{1}{2t}(\bar\partial^X S + |v|^2)} \alpha$ for any neighborhood $\mathcal{U}$ of Y.

Take $y \in Y$. Since Y is a complex submanifold, we can find holomorphic coordinates $\{z_i\}_{i=1}^n$ of a neighborhood U of y such that y corresponds to 0 and $\{(\partial/\partial z_i)(0)\}_{i=m+1}^n$ is an orthonormal basis of (N, g_1^N), and, moreover,

$$U \cap Y = \{p \in U, z_{m+1}(p) = \cdots = z_n(p) = 0\}.$$

Let $\{\mu_k\}_{k=1}^{l'}$ and $\{\mu_k\}_{k=l'+1}^{l}$ be holomorphic frames for $\mathbb{V}_1$ and $\operatorname{Im} \nabla v|_Y$ on $U \cap Y$, with

$$\nabla^{\mathbb{V}}_{\partial/\partial z_k(0)} v = \mu_k(0) \quad \text{for } l' + 1 \le k \le l,$$

and for $z' = (z_1, \dots, z_m)$, $z'' = (z_{m+1}, \dots, z_n)$, $z = (z', z'')$, define $\mu_k(z)$ by parallel transport of $\mu_k(z', 0)$ with respect to $\nabla^{\mathbb{V}}$ along the curve $u \mapsto (z', uz'')$. Identify $\mathbb{V}_z$ with $\mathbb{V}_{(z',0)}$ by identifying $\mu_k(z)$ with $\mu_k(z', 0)$. Denote by $W_y(\varepsilon)$ the ε-neighborhood of y in the normal space N. Then

$$\text{(1–5)}\quad \int_{Y\cap U}\int_{W_y(\varepsilon)} e^{-\frac{1}{2t}(\bar\partial^X S+|v|^2)}\alpha = \int_{Y\cap U}\int_{z\in W_y(\varepsilon/\sqrt{t})} e^{-\frac{1}{2t}(|v(\sqrt{t}z)|^2+(\bar\partial^X S)(\sqrt{t}z))} t^{n-m}\alpha(y,\sqrt{t}z).$$

Define $z = \sum_j z_j(\partial/\partial z_j)$ and $\bar z = \sum_j \bar z_j(\partial/\partial \bar z_j)$. The tautological vector field is $Z = z + \bar z$. Then, for $z \in N_y$,

$$\frac{1}{2t}|v(\sqrt{t}z)|^2 = \frac{1}{2}|\nabla^{\mathbb{V}}_z v|^2 + O(\sqrt{t}) = \frac{1}{2}|z|^2 + O(\sqrt{t})$$

and

$$\bar\partial^X S = \sum_{k=1}^{l} \langle \mu_k, \nabla^{\mathbb{V}}_{\cdot} v\rangle \mu^k.$$

From now on, set $z = (0, z'')$ and $Z = z + \bar z$. Since $\nabla^{\mathbb{V}}_Z \mu_k(0) = 0$, we know that

$$\begin{aligned}\text{(1–6)}\quad \frac{1}{2t}\bar\partial^X S(\sqrt{t}z) &= \frac{1}{2t}\sum_{k=1}^{l} \langle \mu_k, \nabla^{\mathbb{V}}_{\cdot} v\rangle(\sqrt{t}z)\mu^k(0)\\ &= \frac{1}{2t}\sum_{k=1}^{l}\Big(\langle \mu_k, \nabla^{\mathbb{V}}_{\cdot} v\rangle(0) + \sqrt{t}\,\langle \mu_k, \nabla^{\mathbb{V}}_Z\nabla^{\mathbb{V}}_{\cdot} v\rangle(0)\\ &\qquad + \frac{t}{2}\big(\langle \nabla^{\mathbb{V}}_Z\nabla^{\mathbb{V}}_Z\mu_k, \nabla^{\mathbb{V}}_{\cdot} v\rangle + \langle \mu_k, \nabla^{\mathbb{V}}_Z\nabla^{\mathbb{V}}_Z\nabla^{\mathbb{V}}_{\cdot} v\rangle\big)(0) + O(t^{3/2})\Big)\mu^k(0).\end{aligned}$$

Because of the factor t^{n-m} in (1–5), it should be clear that in the limit, only those monomials in the vertical form

$$d\bar z_{m+1}\wedge\cdots\wedge d\bar z_n \,\widehat{\otimes}\, \mu^{l'+1}\wedge\cdots\wedge\mu^l$$

whose weight is exactly t^{m-n} should be kept. Now,

$$\nabla^{\mathbb{V}}_Z\nabla^{\mathbb{V}}_{\partial/\partial z_j} v = R^{\mathbb{V}}\Big(Z, \frac{\partial}{\partial z_j}\Big)v + \nabla^{\mathbb{V}}_{\partial/\partial z_j}\nabla^{\mathbb{V}}_Z v - 1_{[m,n]}(j)\nabla^{\mathbb{V}}_{\partial/\partial z_j} v,$$

$$\nabla^{\mathbb{V}}_{\bar z}\nabla^{\mathbb{V}}_{\partial/\partial z_j} v(0) = R^{\mathbb{V}}\Big(\bar z, \frac{\partial}{\partial z_j}\Big)v + \nabla^{\mathbb{V}}_{\partial/\partial z_j}\nabla^{\mathbb{V}}_{\bar z} v = 0,$$

where $1_{[m,n]}$ is the characteristic function of the interval $[m,n]$. Note that $\nabla^{\mathbb{V}} = \nabla^{\mathbb{V}_1} \oplus \nabla^{\operatorname{Im}\nabla v}$ on Y and that

$$\langle \mu_k, \nabla^{\mathbb{V}}_z \nabla^{\mathbb{V}}_{\partial/\partial z_j} v\rangle(0) = 0 \quad \text{for } 1 \le j \le m, 1 \le k \le l'.$$

It follows that in the expression

$$\frac{1}{2\sqrt{t}} \langle \mu_k, \nabla^{\mathbb{V}}_Z \nabla^{\mathbb{V}}_{\cdot} v\rangle(0) \mu^k(0)$$

a nonzero contribution can only appear in the term

$$\frac{1}{2\sqrt{t}} \biggl(\sum_{j=1}^{m} \sum_{k=l'+1}^{l} + \sum_{j=m+1}^{n} \sum_{k=1}^{l'} \biggr) \langle \mu_k, \nabla^{\mathbb{V}}_z \nabla^{\mathbb{V}}_{\partial/\partial z_j} v\rangle(0)\, d\bar{z}^j \otimes \mu^k(0). \tag{1–7}$$

Similarly, in the last term of (1–6), the only term with a nonzero contribution is

$$\frac{1}{4} \sum_{j=1}^{m} \sum_{k=1}^{l'} \Bigl(\langle \nabla^{\mathbb{V}}_Z \nabla^{\mathbb{V}}_Z \mu_k, \nabla^{\mathbb{V}}_{\partial/\partial z_j} v\rangle(0) + \langle \mu_k, \nabla^{\mathbb{V}}_Z \nabla^{\mathbb{V}}_Z \nabla^{\mathbb{V}}_{\partial/\partial z_j} v\rangle(0) \Bigr)\, d\bar{z}^j \otimes \mu^k(0).$$

But for $1 \le j \le m$, both $\nabla^{\mathbb{V}}_{\partial/\partial z_j} v(0)$ and $\nabla^{\mathbb{V}}_{\partial/\partial z_j} \nabla^{\mathbb{V}}_{\bar{z}} \nabla^{\mathbb{V}}_z v(0) = \nabla^{\mathbb{V}}_{\partial/\partial z_j} (R^{\mathbb{V}}(\bar{z}, z) v)(0)$ vanish, since $v = 0$ on Y. Thus, for $1 \le j \le m$,

$$\nabla^{\mathbb{V}}_Z \nabla^{\mathbb{V}}_Z \nabla^{\mathbb{V}}_{\partial/\partial z_j} v(0) = 2R^{\mathbb{V}}\Bigl(\bar{z}, \frac{\partial}{\partial z_j}\Bigr) \nabla^{\mathbb{V}}_z v(0) + \nabla^{\mathbb{V}}_{\partial/\partial z_j} \nabla^{\mathbb{V}}_z \nabla^{\mathbb{V}}_z v(0).$$

By the preceding discussion, as $t \to 0$, in (1–5), we should replace $\frac{1}{2t} \bar{\partial}^X S(y, \sqrt{t} z)$ by the 2-form

$$\begin{aligned} \frac{1}{2} \sum_{k=1}^{l} \langle \mu_k, \nabla^{\mathbb{V}}_{\cdot} v\rangle (0) \mu^k(0) \;+\; \sqrt{t} \times \text{expression (1–7)} \\ + \frac{1}{2} \sum_{j=1}^{m} \sum_{k=1}^{l'} \Bigl\langle \mu_k, R^{\mathbb{V}}\Bigl(\bar{z}, \frac{\partial}{\partial z_j}\Bigr) \nabla^{\mathbb{V}}_z v + \nabla^{\mathbb{V}}_{\partial/\partial z_j} \nabla^{\mathbb{V}}_z \nabla^{\mathbb{V}}_z v \Bigr\rangle (0)\, d\bar{z}^j \otimes \mu^k(0). \end{aligned}$$

Set $\beta_Y = d\bar{z}_1 \cdots d\bar{z}_m \wedge \mu^1(0) \cdots \mu^{l'}(0)$, $\beta_N = d\bar{z}_{m+1} \cdots d\bar{z}_n \wedge \mu^{l'+1}(0) \cdots \mu^l(0)$, $\phi(\mu^1(0) \cdots \mu^l(0)) = f\, dz_1 \cdots dz_n$. Then

$$\phi_Y(\mu^1(0) \cdots \mu^{l'}(0)) \phi_N(\mu^{l'+1}(0) \cdots \mu^l(0)) = f\, dz_1 \cdots dz_n.$$

Thus

$$\begin{aligned} \phi(\beta_Y \wedge \beta_N) &= (-1)^{l'(n-m)} f\, d\bar{z}_1 \cdots d\bar{z}_n \wedge dz_1 \cdots dz_n \\ &= (-1)^{(l'-m)(n-m)} \phi_Y(\beta_Y) \phi_N(\beta_N). \end{aligned}$$

Now, observing that $\int_{\mathbb{C}} \bar{z}^i e^{-|z|^2}\, dz\, d\bar{z} = 0$ for $i > 0$ and that $\nabla^{\mathbb{V}}_{\cdot} v : (N, g_1^N) \to \bigl(\operatorname{Im}\nabla v, h^{\operatorname{Im}\nabla v}\bigr)$ is an isometry and $l - l' = n - m$, we find that the limit of (1–4) as $t \to 0$ is the sum over j of

$$\int_{Y_j} (-1)^{(l-n)(n-m)} j^*\alpha \int_N \exp\Big(-\frac{1}{2}\sum_{k=1}^{l} \big\langle \mu_k, \nabla^{\mathbb{V}}_{\cdot} v\big\rangle(0)\,\mu^k(0) \\ -\tfrac{1}{2}\big\langle \cdot\,, P^{\mathbb{V}_1}R^{\mathbb{V}}(\bar z, j_*\cdot)\nabla^{\mathbb{V}}_z v\big\rangle(0) - \tfrac{1}{2}|\nabla^{\mathbb{V}}_z v|^2\Big). \tag{1–8}$$

The second integrand in this expression can be rewritten as

$$\exp\Big(-\frac{1}{2}\sum_{i=1}^{n-m} d\bar z_{m+i}\wedge \mu^{l'+i}(0) + \tfrac{1}{2}\big\langle R^{\mathbb{V}}(z, j_*\cdot)P^{\mathbb{V}_1}\cdot, \nabla^{\mathbb{V}}_z v\big\rangle(0) - \tfrac{1}{2}|z|^2\Big) \\ = \exp\Big(\tfrac{1}{2}\big\langle (\nabla^{\mathbb{V}} v)^{-1} R^{\mathbb{V}}(z, j_*\cdot)P^{\mathbb{V}_1}\cdot, z\big\rangle - \tfrac{1}{2}|z|^2\Big)\big(\tfrac{1}{2}\big)^{l-l'} dz_{m+1}\,d\bar z_{m+1}\cdots dz_n\,d\bar z_n.$$

Thus the expression in (1–8) is equal to

$$\int_{Y_j} \frac{(-1)^{(l-n)(n-m)}\alpha}{\det_N\big((1+R^{\mathbb{V}}_v)/(-2\pi i)\big)},$$

which leads to (1–3). □

2. Localization of Quillen metrics via a transversal section

Let X be a compact complex manifold of dimension n. Let $\mathbb{V}$ and ξ be holomorphic vector bundles on X with $\dim \mathbb{V} = m$, and let v be a holomorphic section of $\mathbb{V}$. Assume that v vanishes on a complex manifold $Y \subset X$ and satisfies (1–1). Then we have a complex of holomorphic vector bundles on X,

$$0 \to \textstyle\bigwedge^m(\mathbb{V}^*) \xrightarrow{i(v)} \bigwedge^{m-1}(\mathbb{V}^*) \xrightarrow{i(v)} \cdots \xrightarrow{i(v)} \bigwedge^1(\mathbb{V}^*) \xrightarrow{i(v)} \bigwedge^0(\mathbb{V}^*) \to 0. \tag{2–1}$$

Let $\big(\Omega(X, \bigwedge(\mathbb{V}^*)\otimes\xi),\ \bar\partial^X\big)$ be the Dolbeault complex associated to the holomorphic vector bundle $\bigwedge(\mathbb{V}^*)\otimes\xi$. Let $\mathcal{H}_v(X, \bigwedge(\mathbb{V}^*)\otimes\xi)$ be the hypercohomologies of the bicomplex $\big(\Omega(X, \bigwedge(\mathbb{V}^*)\otimes\xi), \bar\partial^X,\ i(v)\big)$. Let $j : Y \to X$ be the obvious embedding. Now the pullback map j^* induces naturally a map of complexes

$$j^* : \big(\Omega(X, \textstyle\bigwedge(\mathbb{V}^*)\otimes\xi),\ \bar\partial^X + i(v)\big) \to \big(\Omega(Y, \bigwedge(\mathbb{V}_1^*)\otimes\xi),\ \bar\partial^Y\big). \tag{2–2}$$

Theorem 2.1. *The map j^* is a quasi-isomorphism of complexes. In particular, j^* induces an isomorphism*

$$\mathcal{H}_v(X, \textstyle\bigwedge(\mathbb{V}^*)\otimes\xi) \simeq H(Y, \bigwedge(\mathbb{V}_1^*)\otimes\xi). \tag{2–3}$$

Proof. In [Feng 2003] there is an analytic proof of this theorem when $\mathbb{V} = TX$. There we used the twisted vector bundle $\bigwedge(T^*X)$ and here $\bigwedge(\mathbb{V}^*)$ takes its place; the proof works just the same. For an algebraic proof, we can modify the proof of [Bismut 2004, Theorem 5.1]. □

Let N^X, N_H^X be the number operators on $\bigwedge(T^*X)$, $\bigwedge(\mathbb{V}^*)$ corresponding to multiplication by p on $\bigwedge^p(T^*X)$, $\bigwedge^p(\mathbb{V}^*)$; do the same replacing X by Y and $\mathbb{V}^*$ by $\mathbb{V}_1^*$. Then $N^X - N_H^X$ and $N^Y - N_H^Y$ define $\mathbb{Z}$-gradings on $\Omega(X, \bigwedge(\mathbb{V}^*) \otimes \xi)$ and $\Omega(Y, \bigwedge(\mathbb{V}_1^*) \otimes \xi)$, which in turn induce $\mathbb{Z}$-gradings on $\mathscr{H}_v(X, \bigwedge(\mathbb{V}^*) \otimes \xi)$ and $H(Y, \bigwedge(\mathbb{V}_1^*) \otimes \xi)$, respectively. The isomorphism j^* preserves these $\mathbb{Z}$-gradings.

From [Bismut and Lebeau 1991, (1.24)], we define the complex lines $\lambda_v(\mathbb{V}^*)$ and $\lambda(\mathbb{V}_1^*)$ by

$$\lambda_v(\mathbb{V}^*) = \bigotimes_{p=-m}^{n} \left(\det \mathscr{H}_v^p(X, \textstyle\bigwedge(\mathbb{V}^*) \otimes \xi)\right)^{(-1)^{p+1}},$$

$$\lambda(\mathbb{V}_1^*) = \bigotimes_{p=0}^{n} \bigotimes_{q=0}^{m} \left(\det H^p(Y, \textstyle\bigwedge^q(\mathbb{V}_1^*) \otimes \xi)\right)^{(-1)^{p+q+1}}.$$

By (2–3), we have a canonical isomorphism of complex lines

$$\lambda_v(\mathbb{V}^*) \simeq \lambda(\mathbb{V}_1^*).$$

Let ρ be the nonzero section of $\lambda(\mathbb{V}_1^*)^{-1} \otimes \lambda_v(\mathbb{V}^*)$ associated with this canonical isomorphism.

Let g^{TX} be a Kähler metric on TX. We identify N with the bundle orthogonal to TY in $TX|_Y$. Let g^{TY} and g^N be the metrics on TY and N induced by g^{TX}. Let h^ξ be a Hermitian metric on ξ. Let $h^{\mathbb{V}}$ be a metric on $\mathbb{V}$ such that $\mathbb{V}_1$ and $\operatorname{Im} \nabla v|_Y$ are orthogonal on Y and $\nabla v|_Y : N \to \operatorname{Im} \nabla v|_Y$ is an isometry.

Let dv_X be the Riemannian volume form on (X, g^{TX}). Let $\langle \cdot, \cdot \rangle_0$ be the metric on $\bigwedge(\overline{T^*X}) \widehat{\otimes} \bigwedge(\mathbb{V}^*) \otimes \xi$ induced by $g^{TX}, h^{\mathbb{V}}, h^\xi$. The Hermitian product on $\Omega(X, \bigwedge(\mathbb{V}^*) \otimes \xi)$ is defined by

$$\langle \alpha, \alpha' \rangle = \frac{1}{(2\pi)^n} \int_X \langle \alpha, \alpha' \rangle_0 \, dv_X \quad \text{for } \alpha, \alpha' \in \Omega(X, \textstyle\bigwedge(\mathbb{V}^*) \otimes \xi). \tag{2–4}$$

Let $\bar\partial^{X*}$ and $v^*\wedge = i(v)^*$ be the adjoint of $\bar\partial^X$ and $i(v)$ with respect to $\langle \cdot, \cdot \rangle$. Set

$$V = i(v) + i(v)^*, \quad D^X = \bar\partial^X + \bar\partial^{X*}.$$

By Hodge theory,

$$\mathscr{H}_v(X, \textstyle\bigwedge(\mathbb{V}^*) \otimes \xi) \simeq \operatorname{Ker}(D^X + V). \tag{2–5}$$

Denote by P be the operator of orthogonal projection from $\Omega(X, \bigwedge(\mathbb{V}^*) \otimes \xi)$ onto $\ker(D^X + V)$ and set $P^\perp = 1 - P$. Let $h^{\mathscr{H}_v}$ be the L^2-metric on $\mathscr{H}_v(X, \bigwedge(\mathbb{V}^*) \otimes \xi)$ induced by the L^2-product (2–4) via the isomorphism (2–5). Define in the same way a Hermitian product on $\Omega(Y, \bigwedge(\mathbb{V}_1^*) \otimes \xi)$ associated to $g^{TY}, h^{\mathbb{V}_1}, h^\xi$. Let $\bar\partial^{Y*}$ be the adjoint of $\bar\partial^Y$, and $h^{H(Y, \wedge(\mathbb{V}_1^*) \otimes \xi)}$ the corresponding L^2-metric on

$H(Y, \bigwedge(\mathbb{V}_1^*) \otimes \xi)$. Set

$$D^Y = \bar{\partial}^Y + \bar{\partial}^{Y*}.$$

Let Q be the orthogonal projection operator from $\Omega(Y, \bigwedge(\mathbb{V}_1^*) \otimes \xi)$ on $\operatorname{Ker} D^Y$, and $Q^\perp = 1 - Q$. Let $|\cdot|_{\lambda_v(\mathbb{V}^*)}$ and $|\cdot|_{\lambda(\mathbb{V}^*)}$ be the L^2-metrics on $\lambda_v(\mathbb{V}^*)$ and $\lambda(\mathbb{V}^*)$ induced by $h^{\mathscr{H}_v}$ and $h^{H(Y, \wedge(\mathbb{V}_1^*) \otimes \xi)}$. Following [Bismut and Lebeau 1991, (1.49)], let

$$\theta_v^X(s) = -\operatorname{Tr}_s\big((N^X - N_H^X)((D^X + V)^2)^{-s} P^\perp\big).$$

Then $\theta_v^X(s)$ extends to a meromorphic function of $s \in \mathbb{C}$, which is holomorphic at $s = 0$.

The Quillen metric $\|\cdot\|_{\lambda_v(\mathbb{V}^*)}$ on the line $\lambda_v(\mathbb{V}^*)$ is defined by

$$\|\cdot\|_{\lambda_v(\mathbb{V}^*)} = |\cdot|_{\lambda_v(\mathbb{V}^*)} \exp\left(-\frac{1}{2}\frac{\partial \theta_v^X}{\partial s}(0)\right).$$

In the same way, the function

$$\theta^Y(s) = -\operatorname{Tr}_s\big((N^Y - N_H^Y)(D^{Y,2})^{-s} Q^\perp\big)$$

extends to a meromorphic function of $s \in \mathbb{C}$, holomorphic at $s = 0$. The Quillen metric $\|\cdot\|_{\lambda(\mathbb{V}_1^*)}$ on the line $\lambda(\mathbb{V}_1^*)$ is defined by

$$\|\cdot\|_{\lambda(\mathbb{V}_1^*)} = |\cdot|_{\lambda(\mathbb{V}_1^*)} \exp\left(-\frac{1}{2}\frac{\partial \theta^Y}{\partial s}(0)\right).$$

Let $\|\cdot\|_{\lambda(\mathbb{V}_1^*)^{-1} \otimes \lambda_v(\mathbb{V}^*)}$ be the Quillen metric on $\lambda(\mathbb{V}_1^*)^{-1} \otimes \lambda_v(\mathbb{V}^*)$ induced by $\|\cdot\|_{\lambda_v(\mathbb{V}^*)}$ and $\|\cdot\|_{\lambda(\mathbb{V}_1^*)}$ as in [Bismut and Lebeau 1991, §1e].

The purpose of this section is to give a formula for $\|\rho\|^2_{\lambda(\mathbb{V}_1^*)^{-1} \otimes \lambda_v(\mathbb{V}^*)}$. Now we introduce some notations.

For a holomorphic Hermitian vector bundle (E, h^E) on X, we denote by $\operatorname{Td}(E)$, $\operatorname{ch}(E)$, $c_{\max}(E)$ the Todd class, Chern character, and top Chern class of E, and by $\operatorname{Td}(E, h^E)$, $\operatorname{ch}(E, h^E)$, $c_{\max}(E, h^E)$ the Chern–Weil representatives of $\operatorname{Td}(E)$, $\operatorname{ch}(E)$, $c_{\max}(E)$ with respect to the holomorphic Hermitian connection ∇^E on (E, h^E).

Let δ_Y be the current of integration on Y. By [Bismut 1992, Theorem 3.6], a current $\tilde{c}_{\max}(\mathbb{V}, h^{\mathbb{V}})$ on X is well defined by the holomorphic section v (which induces an embedding $v : X \to \mathbb{V}$), and this current satisfies

$$\frac{\bar{\partial}\partial}{2\pi i}\tilde{c}_{\max}(\mathbb{V}, h^{\mathbb{V}}) = c_{\max}(\mathbb{V}_1, h^{\mathbb{V}_1})\delta_Y - c_{\max}(\mathbb{V}, h^{\mathbb{V}}). \tag{2–6}$$

Let $\widetilde{\operatorname{Td}}(TY, TX, g^{TX|_Y})$ be the Bott–Chern current on Y associated to the exact sequence

$$0 \to TY \to TX|_Y \to N \to 0 \tag{2–7}$$

constructed in [Bismut et al. 1988a, §1f], which satisfies

$$\frac{\bar\partial\partial}{2\pi i}\,\widetilde{\mathrm{Td}}(TY, TX, g^{TX|_Y}) = \mathrm{Td}(TX|_Y, g^{TX|_Y}) - \mathrm{Td}(TY, g^{TY})\,\mathrm{Td}(N, g^N).$$

Finally, let $R(x)$ be the power series introduced in [Gillet and Soulé 1991], which is such that if $\zeta(s)$ is the Riemann zeta function, then

$$R(x) = \sum_{\substack{n\geq 1\\ n\ \mathrm{odd}}}\left(\sum_{j=1}^{n}\frac{1}{j}\zeta(-n) + 2\frac{\partial\zeta}{\partial s}(-n)\right)\frac{x^n}{n!}.$$

We identify R with the corresponding additive genus. We also set

$$\mathrm{ch}(\textstyle\bigwedge^*(\mathbb{V}_1^*)) = \displaystyle\sum_i (-1)^i\,\mathrm{ch}(\textstyle\bigwedge^i(\mathbb{V}_1^*)),$$

and denote by $\mathrm{ch}(\bigwedge^*(\mathbb{V}_1^*), h^{\bigwedge^*(\mathbb{V}_1^*)})$ its Chern–Weil representative.

Theorem 2.2. *The Quillen metric* $\|\rho\|^2_{\lambda(\mathbb{V}_1^*)^{-1}\otimes\lambda_v(\mathbb{V}^*)}$ *is given by the exponential of*

$$\begin{aligned}(2\text{–}8)\quad & -\int_X \mathrm{Td}(TX, g^{TX})\,\mathrm{Td}^{-1}(\mathbb{V}, h^{\mathbb{V}})\tilde{c}_{\max}(\mathbb{V}, h^{\mathbb{V}})\,\mathrm{ch}(\xi, h^{\xi})\\ &+\int_Y \mathrm{Td}^{-1}(N, g^N)\,\widetilde{\mathrm{Td}}(TY, TX|_Y, g^{TX|_Y})\,\mathrm{ch}(\textstyle\bigwedge^*(\mathbb{V}_1^*), h^{\bigwedge^*(\mathbb{V}_1^*)})\,\mathrm{ch}(\xi, h^{\xi})\\ &-\int_Y \mathrm{Td}(TY)\,R(N)\,\mathrm{ch}(\textstyle\bigwedge^*(\mathbb{V}_1^*))\,\mathrm{ch}(\xi).\end{aligned}$$

Proof. Set

$$(2\text{–}9)\qquad T(\textstyle\bigwedge(\mathbb{V}^*), h^{\bigwedge(\mathbb{V}^*)}) = \mathrm{Td}^{-1}(\mathbb{V}, h^{\mathbb{V}})\tilde{c}_{\max}(\mathbb{V}, h^{\mathbb{V}}).$$

By the same argument as in [Bismut et al. 1990, Theorem 3.17], the current

$$T(\textstyle\bigwedge(\mathbb{V}^*), h^{\bigwedge(\mathbb{V}^*)})$$

is exactly the current on X associated to (2–1) (evaluated modulo irrelevant ∂ or $\bar\partial$ coboundaries).

Now, from the choice of our metric $h^{\mathbb{V}}$, the analogue of [Bismut and Lebeau 1991, Definition 1.21, assumption (A)] is satisfied for the complex (2–1). Then we verify that as far as local index theoretic computations are concerned, the situation is exactly the same as in [Bismut and Lebeau 1991]. Because of the quasi-isomorphism of Theorem 2.1, there are no "small" eigenvalues of the operator $D + TV$ when $T \to +\infty$. In Section 3, we write down the intermediate results corresponding to [Bismut and Lebeau 1991, §6c]. Comparing to [Bismut and Lebeau 1991, §§6c–6e], the proof of Theorem 2.2 is complete. □

Remark 2.3. Assume that Y consists only discrete points; then $l \geq n$ and the last two terms of (2–8) are zero. In this case, if $n = l$, then (2–1) is a resolution of $j_*(\mathcal{O}_Y)$ and Theorem 2.2 is a direct consequence of [Bismut and Lebeau 1991, Theorem 0.1]. By [Bismut 1992, Theorem 3.2, Definition 3.5], $\tilde{c}_{\max}(\mathbb{V}, h^{\mathbb{V}})$ is zero if $l > n+1$.

3. L^2 metrics on $H_v(X, \bigwedge(\mathbb{V}^*))$ and localization

We keep the assumptions and notations of Section 2.

Let g^{TX} be a Kähler metric on TX, and let g^{TY}, g^N be the metrics on TY, N induced by g^{TX}. Let $h^{\mathbb{V}}$ be a metric on $\mathbb{V}$ such that $\mathbb{V}_1$ and $\operatorname{Im} \nabla v|_Y$ are orthogonal on Y and $\nabla v|_Y : (N, g^N) \to \operatorname{Im} \nabla v|_Y$ is an isometry.

Let $\phi_1 : \det \mathbb{V}_1^* \to \det T^*Y$ be a nonzero holomorphic section. Let $h_1^{\mathbb{V}}$ be a metric on $\mathbb{V}$ such that on Y, $\mathbb{V}_1$ and $\operatorname{Im} \nabla v|_Y$ are orthogonal and

$$|\phi|_{\det \mathbb{V} \otimes \det T^*X, 1} = |\phi_1|_{\det \mathbb{V}_1 \otimes \det T^*Y, 1} = 1,$$

where $|\cdot|_{\det \mathbb{V} \otimes \det T^*X, 1}$ and $|\cdot|_{\det \mathbb{V}_1 \otimes \det T^*Y, 1}$ are the norms on the holomorphic line bundles $\det \mathbb{V} \otimes \det T^*X$ and $\det \mathbb{V}_1 \otimes \det T^*Y$ induced by $h_1^{\mathbb{V}}$ and g^{TX}.

We will add a subscript 1 to denote the objects induced by $h_1^{\mathbb{V}}$. For

$$\beta \in \textstyle\bigwedge^p(\overline{T^*X}) \mathbin{\widehat{\otimes}} \bigwedge^q(\mathbb{V}^*),$$

we define $*_{\mathbb{V},1}\beta \in \bigwedge^{n-p}(\overline{T^*X}) \mathbin{\widehat{\otimes}} \bigwedge^{l-q}(\mathbb{V}^*)$ by

$$\langle \alpha, \beta \rangle_1 \, \phi^{-1}(dv_X) = \alpha \wedge *_{\mathbb{V},1}\beta.$$

It's useful to write down a local expression for $*_{\mathbb{V},1}$. if $\{w^i\}_{i=1}^n$ and $\{\mu^i\}_{i=1}^l$, are orthonormal bases of T^*X and $(\mathbb{V}^*, h_1^{\mathbb{V}})$, then

$$dv_X = (-1)^{n(n+1)/2}(\sqrt{-1})^n \overline{w}^1 \wedge \cdots \wedge \overline{w}^n \mathbin{\widehat{\otimes}} w^1 \wedge \cdots \wedge w^n$$

and $\phi^{-1}(w^1 \wedge \cdots \wedge w^n) = f \, \mu^1 \wedge \cdots \wedge \mu^l$ with $|f| = 1$. If

$$\beta = \overline{w}^1 \wedge \cdots \wedge \overline{w}^p \mathbin{\widehat{\otimes}} \mu^1 \wedge \cdots \wedge \mu^q,$$

then

$$*_{\mathbb{V},1}\beta = (-1)^{(n-p)q+n(n+1)/2}(\sqrt{-1})^n f \, \overline{w}^{p+1} \wedge \cdots \wedge \overline{w}^n \mathbin{\widehat{\otimes}} \mu^{q+1} \wedge \cdots \wedge \mu^l.$$

Thus $*_{\mathbb{V},1} *_{\mathbb{V},1} \beta = (-1)^{(p+q)(n+l+1)}\beta$, for any $\beta \in \bigwedge^p(\overline{T^*X}) \widehat{\otimes} \bigwedge^q(\mathbb{V}^*)$. Combining this with (1–2), we find that

$$\bar{\partial}^{X*}\beta = (-1)^{p+q+1} *_{\mathbb{V},1}^{-1} \bar{\partial}^X *_{\mathbb{V},1} \beta, \quad (i(v))^*\beta = (-1)^{p+q+1} *_{\mathbb{V},1}^{-1} i(v) *_{\mathbb{V},1} \beta.$$

Thus the antilinear map $*_{\mathbb{V},1}$ is an isometry from $\left(\mathcal{H}_v(X, \bigwedge(\mathbb{V}^*)), h_1^{\mathcal{H}_v}\right)$ to itself.

The bilinear form

$$\alpha, \beta \in \mathscr{H}_v(X, \wedge(\mathbb{V}^*)) \mapsto \frac{1}{(2\pi)^n} \int_X \alpha \wedge \beta \tag{3–1}$$

is nondegenerate; indeed, $\alpha \in \mathscr{H}_v(X, \wedge(\mathbb{V}^*))$ implies $*_{\mathbb{V},1}\alpha \in \mathscr{H}_v(X, \wedge(\mathbb{V}^*))$, so $\alpha \neq 0$ implies

$$\int_X \alpha \wedge *_{\mathbb{V},1}\alpha > 0.$$

Thus the metric $|\cdot|_{\lambda_v(\mathbb{V}^*),1}$ on $\lambda_v(\mathbb{V}^*)$ only depends on the nondegenerate bilinear form (3–1) on $\mathscr{H}_v(X, \wedge(\mathbb{V}^*))$, which is metric-independent.

Recall the definition of $\det \nabla v|_Y$ from Section 1. Now,

$$\frac{\phi|_Y/((\det \nabla v|_Y)^*)}{\phi_1}$$

is a holomorphic function on Y. Since Y is compact, this function is locally constant. Then we have the following extension of [Bismut 2004, Theorem 5.7].

Theorem 3.1.

$$\log\big(|\rho|_{\lambda(\mathbb{V}_1^*)^{-1}\otimes\lambda_v(\mathbb{V}^*),1}\big)^2 = \int_Y \mathrm{Td}(TY)\,\mathrm{ch}(\wedge(\mathbb{V}_1^*)) \log \left|\frac{\phi|_Y/((\det \nabla v|_Y)^*)}{\phi_1}\right|. \tag{3–2}$$

Proof. We use ϕ_1 to define the integral $\int_Y \gamma$ for $\gamma \in H(Y, \wedge(\mathbb{V}_1^*))$. Since

$$|\phi_1|_{\det \mathbb{V}_1 \otimes \det T^*Y,1} = 1,$$

following the same considerations as above, we find that the antilinear operator $*_{\mathbb{V}_1,1}$ maps $H(Y, \wedge(\mathbb{V}_1^*))$ into itself isometrically. Therefore, to evaluate the left-hand side of (3–2), we only need to compare the bilinear forms (3–1) with

$$a, b \in H(Y, \wedge(\mathbb{V}_1^*)) \mapsto \frac{1}{(2\pi)^m} \int_Y a \wedge b.$$

Let $A_v \in \mathrm{End}^{\mathrm{even}} H(Y, \wedge(\mathbb{V}_1^*))$ be given by

$$a \to \frac{(-1)^{(l-n)(n-m)} a}{(2\pi)^{n-m} \det_N\big((1+R_v^{\mathbb{V}})/(-2\pi i)\big)} \frac{\phi|_Y/((\det \nabla v|_Y)^*)}{\phi_1}. \tag{3–3}$$

Set

$$\det A_v = \frac{\det A_v|_{H^{\mathrm{even}}(Y,\wedge(\mathbb{V}_1^*))}}{\det A_v|_{H^{\mathrm{odd}}(Y,\wedge(\mathbb{V}_1^*))}};$$

then

$$\big(|\rho|_{\lambda(\mathbb{V}_1^*)^{-1}\otimes\lambda_v(\mathbb{V}^*),1}\big)^2 = |\det A_v|.$$

Now, A_v is a degree-increasing operator in $H(Y, \wedge(\mathbb{V}_1^*))$. Therefore it acts like a triangular matrix whose diagonal part is just multiplication by the locally constant function $\dfrac{\phi|_Y/((\det\nabla v|_Y)^*)}{\phi_1}$. Using (3–3), we get

$$\det A_v = \Big(\frac{\phi|_Y/((\det\nabla v|_Y)^*)}{\phi_1}\Big)^{\chi(Y,\wedge(\mathbb{V}_1^*))}.$$

But $\chi(Y,\wedge(\mathbb{V}_1^*)) = \int_Y \mathrm{Td}(TY)\,\mathrm{ch}(\wedge(\mathbb{V}_1^*))$; thus we get (3–2). □

Let g_1^N be the metric on N such that $\nabla v|_Y : (N, g_1^N) \to (\mathrm{Im}(\nabla v), h_1^{\mathrm{Im}(\nabla v)})$ is an isometry. Let $\widetilde{\mathrm{Td}}^{-1}(N, g^N, g_1^N)$ be the Bott–Chern class constructed in [Bismut et al. 1988a, §1f] such that

$$\frac{\bar\partial\partial}{2\pi i}\,\widetilde{\mathrm{Td}}^{-1}(N, g^N, g_1^N) = \mathrm{Td}^{-1}(N, g_1^N) - \mathrm{Td}^{-1}(N, g^N).$$

Finally, we can compute the analytic torsion on the total manifold via the zero set of a transversal section v.

Theorem 3.2. *If $h_1^{\mathbb{V}_1} = h^{\mathbb{V}_1}$ on Y, then*

$$\begin{aligned}(3\text{–}4)\quad -\frac{\partial\theta^X_{v,1}}{\partial s}(0) + \frac{\partial\theta^Y}{\partial s}(0) = &-\int_X \mathrm{Td}(TX, g^{TX})\,\mathrm{Td}^{-1}(\mathbb{V}, h_1^{\mathbb{V}})\tilde c_{\max}(\mathbb{V}, h_1^{\mathbb{V}})\\ &+\int_Y\Big(\mathrm{Td}^{-1}(N, g^N)\,\widetilde{\mathrm{Td}}\big(TY, TX|_Y, g^{TX|_Y}\big)\\ &\qquad + \mathrm{Td}(TX, g^{TX})\,\widetilde{\mathrm{Td}}^{-1}(N, g^N, g_1^N)\Big)\,\mathrm{ch}\big(\wedge^*(\mathbb{V}_1^*), h^{\wedge^*(\mathbb{V}_1^*)}\big)\\ &-\int_Y \mathrm{Td}(TY)\,\mathrm{ch}(\wedge^*(\mathbb{V}_1^*))\Big(R(N) + \log\Big|\frac{\phi|_Y/((\det\nabla v|_Y)^*)}{\phi_1}\Big|\Big).\end{aligned}$$

Proof. Since $h_1^{\mathbb{V}_1} = h^{\mathbb{V}_1}$, we have $|\cdot|_{\lambda(\mathbb{V}_1^*)} = |\cdot|_{\lambda(\mathbb{V}_1^*),1}$ and $\|\cdot\|_{\lambda(\mathbb{V}_1^*)} = \|\cdot\|_{\lambda(\mathbb{V}_1^*),1}$. Let $\widetilde{\mathrm{ch}}\big(\wedge(\mathbb{V}^*), h_1^{\wedge(\mathbb{V}^*)}, h^{\wedge(\mathbb{V}^*)}\big)$ be the Bott–Chern class constructed in [Bismut et al. 1988a, §1f], so that

$$\frac{\bar\partial\partial}{2\pi i}\,\widetilde{\mathrm{ch}}\big(\wedge(\mathbb{V}^*), h_1^{\wedge(\mathbb{V}^*)}, h^{\wedge(\mathbb{V}^*)}\big) = \mathrm{ch}\big(\wedge(\mathbb{V}^*), h^{\wedge(\mathbb{V}^*)}\big) - \mathrm{ch}\big(\wedge(\mathbb{V}^*), h_1^{\wedge(\mathbb{V}^*)}\big).$$

Then by the anomaly formula [Bismut et al. 1988b, Theorem 1.23],

$$\log\Big(\frac{\|\cdot\|^2_{\lambda_v(\mathbb{V}^*)}}{\|\cdot\|^2_{\lambda_v(\mathbb{V}^*),1}}\Big) = \int_X \mathrm{Td}(TX, g^{TX})\,\widetilde{\mathrm{ch}}\big(\wedge(\mathbb{V}^*), h_1^{\wedge(\mathbb{V}^*)}, h^{\wedge(\mathbb{V}^*)}\big).$$

By [Bismut et al. 1990, Theorem 2.5],

$$\begin{aligned}(3\text{–}5)\quad & T\big(\wedge(\mathbb{V}^*), h^{\wedge(\mathbb{V}^*)}\big) - T\big(\wedge(\mathbb{V}^*), h_1^{\wedge(\mathbb{V}^*)}\big)\\ &= \mathrm{ch}\big(\wedge^*(\mathbb{V}_1^*), h^{\wedge^*(\mathbb{V}_1^*)}\big)\widetilde{\mathrm{Td}}^{-1}(N, g_1^N, g^N)\delta_Y - \widetilde{\mathrm{ch}}\big(\wedge(\mathbb{V}^*), h_1^{\wedge(\mathbb{V}^*)}, h^{\wedge(\mathbb{V}^*)}\big).\end{aligned}$$

By (2–9), Theorems 2.2 and 3.1, and the preceding equations, the proof of Theorem 3.2 is complete. □

Remark 3.3. If Y consists only of discrete points and $n = l$, then $\phi_1 = \mathrm{Id}$. In this case let $g^{\det N}$ and $g_1^{\det N}$ be the metrics on $\det N = \det TX$ induced by g^N and g_1^N. By Remark 2.3 and Theorem 3.2,

$$-\frac{\partial \theta^X_{v,1}}{\partial s}(0) = -\int_X \mathrm{Td}(TX, g^{TX})\,\mathrm{Td}^{-1}(\mathbb{V}, h_1^{\mathbb{V}})\,\tilde{c}_{\max}(\mathbb{V}, h_1^{\mathbb{V}}) \\ + \sum_{p\in Y}\big(\tfrac{1}{2}\log(g^{\det N}/g_1^{\det N}) - \log|\phi/(\det \nabla v|_Y)^*|\big).$$

Remark 3.4. If $\mathbb{V} = TX$ and v is a holomorphic Killing vector field, (3–4) is a special case of [Bismut 1992, Theorems 6.2 and 7.7]. In this case, $h_1^{\mathbb{V}} = g^{TX}$, and on Y, we have a holomorphic and orthogonal splitting $TX|_Y = TY \oplus N$. Thus $\widetilde{\mathrm{Td}}(TY, TX|_Y, g^{TX|_Y}) = 0$. To compute $\widetilde{\mathrm{Td}^{-1}}(N, g^N, g_1^N)$, note that $g_1^N = g^N((\nabla v)\cdot, (\nabla v)\cdot)$, as $A = (\nabla v)^*(\nabla v)$ is positive and self-adjoint; thus $(A)^s$ is well defined for $s \in [0, 1]$. Taking $g_s^N = g^N((A)^s\cdot, \cdot)$, we obtain by [Bismut et al. 1988a, Theorem 1.30]

$$\widetilde{\mathrm{Td}^{-1}}(N, g^N, g_1^N) = \int_0^1 \big\langle (\mathrm{Td}^{-1})'(N, g_s^N), \log A\big\rangle\, ds.$$

But ∇v is holomorphic, so the curvature R_s^N associated to the holomorphic connection on (N, g_s^N) is $R_s^N = R^N$ for $s \in [0, 1]$. Thus

$$\widetilde{\mathrm{Td}^{-1}}(N, g^N, g_1^N) = \big\langle (\mathrm{Td}^{-1})'(N, g^N), \log A\big\rangle. \tag{3–6}$$

Now

$$\mathrm{Td}(TX, g^{TX})\, T\big(\textstyle\bigwedge(T^*X), h^{\wedge(T^*X)}\big) = \tilde{c}_{\max}(TX, g^{TX}) \tag{3–7}$$

is an $(n-1, n-1)$-form on X.

In this case, we get easily the special case of [Bismut 2004, Theorem 4.15] directly from [Ray and Singer 1973] by using Poincaré duality:

$$\frac{\partial \theta^Y}{\partial s}(0) = 0. \tag{3–8}$$

From (3–4), (3–6), (3–7), and the vanishing of the constant terms of $R(N)$ and $\frac{\mathrm{Td}'}{\mathrm{Td}}(N, g^N) - \frac{1}{2}$, we get

$$-\frac{\partial \theta^X_{v,1}}{\partial s}(0) = \int_Y c_{\max}(TY)\bigg(R(N) - \Big\langle \frac{\mathrm{Td}'}{\mathrm{Td}}(N, g^N) - \tfrac{1}{2}, \log A\Big\rangle\bigg) = 0. \tag{3–9}$$

4. Appendix: six intermediate results

In this section, to help readers understand how to obtain Theorem 2.2, we write down the corresponding intermediate results from [Bismut and Lebeau 1991, Theorems 6.4-6.9].

Let $\nabla^{\wedge(\mathbb{V}^*)}$ be the connection on $\wedge(\mathbb{V}^*)$ induced by $\nabla^{\mathbb{V}^*}$. Set $C_u = \nabla^{\wedge(\mathbb{V}^*)} + \sqrt{u}V$. Let $\mathcal{B}^2_{T^2}$ and $\operatorname{Tr}_s\big(N^Y_H \exp(-\mathcal{B}^2_{T^2})\big)$ be the operator and the generalized trace associated to the complex (2–7) as in [Bismut and Lebeau 1991, §5]. Let Φ be the homomorphism from $\wedge^{\text{even}}(T^*_{\mathbb{R}}X)$ into itself which to $\alpha \in \wedge^{2p}(T^*_{\mathbb{R}}X)$ associates $(2\pi i)^{-p}\alpha$.

Theorem 4.1. *For any $u_0 > 0$, there exists $C > 0$ such that for $u \ge u_0$, $T \ge 1$,*

$$\Big|\operatorname{Tr}_s\big(N^X_H e^{-u(D^X+TV)^2}\big) - \operatorname{Tr}_s\big((\tfrac{1}{2}\dim N + N^Y_H)e^{-uD^{Y,2}}\big)\Big| \le \frac{C}{\sqrt{T}},$$

$$\Big|\operatorname{Tr}_s\big((N^X - N^X_H)e^{-u(D^X+TV)^2}\big) - \operatorname{Tr}_s\big((N^Y - N^Y_H)e^{-uD^{Y,2}}\big)\Big| \le \frac{C}{\sqrt{T}}.$$

Theorem 4.2. *Let $\tilde{P}_T$ be the orthogonal projection operator from $\Omega(X, \wedge(\mathbb{V}^*)\otimes\xi)$ to $\operatorname{Ker}(D^X + TV)$. There exist $c > 0$ and $C > 0$ such that, for any $u \ge 1$ and $T \ge 1$,*

$$\Big|\operatorname{Tr}_s\big((N^X - N^X_H)e^{-u(D^X+TV)^2}\big) - \operatorname{Tr}_s\big((N^X - N^X_H)\tilde{P}_T\big)\Big| \le c e^{-Cu},$$

Theorem 4.3. *There exist $C > 0$ and $\gamma \in]0, 1]$ such that, for any $u \in]0, 1]$ and $0 \le T \le 1/u$,*

$$\Big|\operatorname{Tr}_s\big(N^X_H e^{-(uD^X+TV)^2}\big) - \int_X \operatorname{Td}(TX, g^{TX})\Phi\operatorname{Tr}_s\big(N^X_H e^{-C^2_{T^2}}\big)\Big| \le C(u(1+T))^\gamma.$$

There exists a constant $C' > 0$ such that for $u \in]0, 1]$ and $0 \le T \le 1$,

$$\Big|\operatorname{Tr}_s\Big(N^X_H e^{-(uD^X+TV)^2}\Big) - \operatorname{Tr}_s\Big(N^X_H e^{-(uD^X)^2}\Big)\Big| \le C'T.$$

Theorem 4.4. *For any $T > 0$,*

$$\lim_{u\to 0}\operatorname{Tr}_s\big(N^X_H e^{-(uD^X+(T/u)V)^2}\big) = \int_Y \Phi\operatorname{Tr}_s\big(N^Y_H e^{-\mathcal{B}^2_{T^2}}\big)\operatorname{ch}\big(\wedge(\mathbb{V}^*_1), h^{\wedge(\mathbb{V}^*_1)}\big)\operatorname{ch}(\xi, h^\xi).$$

Theorem 4.5. *There exist $C > 0$ and $\delta \in]0, 1]$ such that, for any $u \in]0, 1]$ and $T \ge 1$,*

$$\Big|\operatorname{Tr}_s\big(N^X_H e^{-(uD^X+(T/u)V)^2}\big) - \operatorname{Tr}_s\big((\tfrac{1}{2}\dim N + N^Y_H)e^{-uD^{Y,2}}\big)\Big| \le \frac{C}{T^\delta}.$$

Let $|\cdot|^2_{\lambda_v(\mathbb{V}^*),T}$ be the L^2-metric on $\lambda_v(\mathbb{V}^*)$ induced by g^{TX}, $T^2h^{\mathbb{V}}$ as in (2–5).

Theorem 4.6. *As* $T \to +\infty$,

$$\log\left(\frac{|\cdot|^2_{\lambda_v(\mathbb{V}^*),T}}{|\cdot|^2_{\lambda_v(\mathbb{V}^*)}}\right) = -\log|\rho|^2_{\lambda(\mathbb{V}_1^*)^{-1}\otimes\lambda_v(\mathbb{V}^*)} + \mathrm{Tr}_s\big((\dim N + 2N_H^Y)Q\big)\log T + O\Big(\frac{1}{T}\Big).$$

Acknowledgements

Feng thanks Jean-Pierre Bourguignon and the IHES, where part of this research was performed, for their hospitality. Thanks also to K. Liu for drawing our attention to [Beasley and Witten 2003].

References

[Beasley and Witten 2003] C. Beasley and E. Witten, "Residues and world-sheet instantons", 2003. hep-th/0304115

[Bershadsky et al. 1993] M. Bershadsky, S. Cecotti, H. Ooguri, and C. Vafa, "Holomorphic anomalies in topological field theories", *Nuclear Phys. B* **405**:2-3 (1993), 279–304. MR 94j:81254 Zbl 1039.81550

[Bershadsky et al. 1994] M. Bershadsky, S. Cecotti, H. Ooguri, and C. Vafa, "Kodaira–Spencer theory of gravity and exact results for quantum string amplitudes", *Comm. Math. Phys.* **165**:2 (1994), 311–427. MR 95f:32029 Zbl 0815.53082

[Bismut 1986] J.-M. Bismut, "Localization formulas, superconnections, and the index theorem for families", *Comm. Math. Phys.* **103**:1 (1986), 127–166. MR 87f:58147 Zbl 0602.58042

[Bismut 1992] J.-M. Bismut, "Bott–Chern currents, excess normal bundles and the Chern character", *Geom. Funct. Anal.* **2**:3 (1992), 285–340. MR 94a:58206 Zbl 0776.32007

[Bismut 2004] J.-M. Bismut, "Holomorphic and de Rham torsion", *Compos. Math.* **140**:5 (2004), 1302–1356. MR 2081158 Zbl 02110378

[Bismut and Lebeau 1991] J.-M. Bismut and G. Lebeau, "Complex immersions and Quillen metrics", *Inst. Hautes Études Sci. Publ. Math.* **74** (1991), 1–297. MR 94a:58205 Zbl 0784.32010

[Bismut et al. 1988a] J.-M. Bismut, H. Gillet, and C. Soulé, "Analytic torsion and holomorphic determinant bundles, I: Bott–Chern forms and analytic torsion", *Comm. Math. Phys.* **115**:1 (1988), 49–78. MR 89g:58192a Zbl 0651.32017

[Bismut et al. 1988b] J.-M. Bismut, H. Gillet, and C. Soulé, "Analytic torsion and holomorphic determinant bundles, III: Quillen metrics on holomorphic determinants", *Comm. Math. Phys.* **115**:2 (1988), 301–351. MR 89g:58192c Zbl 0651.32017

[Bismut et al. 1990] J.-M. Bismut, H. Gillet, and C. Soulé, "Complex immersions and Arakelov geometry", pp. 249–331 in *The Grothendieck Festschrift*, vol. I, Progr. Math. **86**, Birkhäuser, Boston, 1990. MR 92a:14019 Zbl 0744.14015

[Bott 1967] R. Bott, "A residue formula for holomorphic vector-fields", *J. Differential Geometry* **1** (1967), 311–330. MR 38 #730 Zbl 0179.28801

[Feng 2003] H. Feng, "Holomorphic equivariant cohomology via a transversal holomorphic vector field", *Internat. J. Math.* **14**:5 (2003), 499–514. MR 2004j:32022 Zbl 1050.32013

[Gillet and Soulé 1991] H. Gillet and C. Soulé, "Analytic torsion and the arithmetic Todd genus", *Topology* **30**:1 (1991), 21–54. MR 92d:14015 Zbl 0787.14005

[Liu 1995] K. Liu, "Holomorphic equivariant cohomology", *Math. Ann.* **303**:1 (1995), 125–148. MR 97f:32041 Zbl 0835.14006

[Ray and Singer 1973] D. B. Ray and I. M. Singer, "Analytic torsion for complex manifolds", *Ann. of Math.* (2) **98** (1973), 154–177. MR 52 #4344 Zbl 0267.32014

[Zhang 1990] W. Zhang, "A remark on a residue formula of Bott", *Acta Math. Sinica* (*N.S.*) **6**:4 (1990), 306–314. MR 91j:58153 Zbl 0738.32007

[Zhang n.d.] W. Zhang, "Equivariant Dolbeault complex and total Quillen metrics", preprint.

Received July 25, 2003. Revised February 13, 2004.

HUITAO FENG
COLLEGE OF MATHEMATICAL SCIENCES
NANKAI UNIVERSITY
300071, TIANJIN
CHINA
fht@nankai.edu.cn

XIAONAN MA
CENTRE DE MATHÉMATIQUES
UMR 7640 DU CNRS
ÉCOLE POLYTECHNIQUE
91128 PALAISEAU CEDEX
FRANCE
ma@math.polytechnique.fr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

CONVEXITY OF THE FIGURE EIGHT SOLUTION TO THE THREE-BODY PROBLEM

TOSHIAKI FUJIWARA AND RICHARD MONTGOMERY

The Newtonian three-body problem with equal masses has a remarkable solution where the bodies chase each other around a planar curve having the qualitative shape and symmetries of a figure eight. Here we prove that each lobe of this curve is convex.

1. Introduction

The figure eight is a recently discovered periodic solution to the Newtonian three-body problem in which three equal masses traverse a single closed planar curve in the form of an 8 (Figure 1). See [Moore 1993; Chenciner and Montgomery 2000]. The curve has one self-intersection, the origin, which divides it into two symmetric lobes. In [Chenciner and Montgomery 2000] it was proved that each lobe is star-shaped. Here we prove the lobes are convex. (A computer proof based on interval arithmetic appears in [Kapela and Zgliczyński 2003].)

Theorem 1. *Each lobe of the eight solution is a convex curve.*

In the final section we describe how the theorem generalizes to prove the convexity of eights for many three-body potentials besides Newton's.

2. Preliminaries

We present a number of properties of the eight established in [Chenciner and Montgomery 2000] and three assertions relating mechanics and plane geometry. The convexity proof relies on these properties and assertions.

Center of Mass. Write $q_1(t), q_2(t), q_3(t)$ for the location of the three masses in the plane at time t. At each time t we have $q_1(t)+q_2(t)+q_3(t)=0$.

MSC2000: 70F07.
Keywords: figure-eight, convexity, Newtonian potential.
The authors thank AIM/ARCC for funding a workshop in celestial mechanics where the authors met.

Symmetry. Write $R_y(x, y) = (-x, y)$ for the reflection about the y axis. Then the eight solution enjoys the following symmetries:

$$\big(q_1(t), q_2(t), q_3(t)\big) = \big(R_y\big(q_3(t - \tfrac{1}{6}T)\big), R_y\big(q_1(t - \tfrac{1}{6}T)\big), R_y\big(q_2(t - \tfrac{1}{6}T)\big)\big),$$
$$\big(q_1(t), q_2(t), q_3(t)\big) = \big(-q_1(-t), -q_3(-t), -q_2(-t)\big).$$

The right-hand side of these equations defines transformations s and σ on the space of all T-periodic loops. These transformations generate an action of the dihedral group

$$D_6 = \{s, \sigma \mid s^6 = 1,\ \sigma^2 = 1,\ s\sigma = \sigma s^{-1}\},$$

the symmetry group of a regular hexagon, which is consequently a symmetry group of the eight.

Invariance under $s^2 \in D_6$ implies that $\big(s^2(q_1, q_2, q_2)\big)(t) = \big(q_1(t), q_2(t), q_3(t)\big)$. Setting $q = q_1$ this last equation reads

$$q_1(t) = q(t), \quad q_2(t) = q\big(t + \tfrac{1}{3}T\big), \quad q_3(t) = q\big(t + \tfrac{2}{3}T\big). \tag{1}$$

A *choreography* is a three-body solution satisfying (1). The curve $q(t)$ is the curve of the eight whose lobes are the subject of Theorem 1.

The D_6-invariance of the figure eight implies that it is completely determined by the three arcs $q_1\big([-\frac{1}{12}T, 0]\big)$, $q_2\big([-\frac{1}{12}T, 0]\big)$, $q_3\big([-\frac{1}{12}T, 0]\big)$ swept out by the three masses over the time interval $[-\frac{1}{12}T, 0]$. *To prove Theorem 1 it is enough to prove that the curvatures of these three arcs are never zero* (with the exception of the point $q_1(0)$, the self-intersection point of the eight, which is taken to be the origin).

A configuration (q_1, q_2, q_3) satisfying $q_1 + q_2 + q_3 = 0$ is called an *Euler configuration* if one of the q_i vanishes. Then necessarily the other two masses q_j, q_k are of the form $\zeta, -\zeta$, so that the entire configuration (q_1, q_2, q_3) is collinear with mass i at the origin located at the midpoint of the segment defined by the other two masses j and k. Upon translating time if necessary, and relabeling the masses, we can insist that at time 0 the configuration is an Euler configuration with mass 1 at the origin and 3 in the first quadrant, as indicated in Figure 1. At the initial time $t = -\frac{1}{12}T$ the three masses form an isosceles triangle, with mass 2 at the vertex and lying on the negative x-axis.

The eight minimizes the usual action of mechanics (integral of the kinetic minus potential energy) among all T-periodic loops enjoying D_6 symmetry. Equivalently [Chenciner and Montgomery 2000] the path $\big(q_1(t), q_2(t), q_3(t)\big)$ of the eight over the fundamental time interval $\big[-\frac{1}{12}T, 0\big]$ minimizes the action among all paths starting at time $-\frac{1}{12}T$ in an isosceles configuration with 2 being the vertex and ending at time 0 in an Euler configuration with 1 being the origin. An important consequence of minimization, proved in [Chenciner and Montgomery 2000,

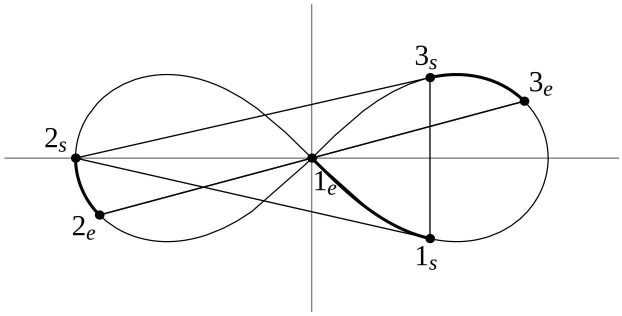

Figure 1. The eight. The labels 1_s and 1_e represent the location of mass 1 at $t = -\frac{1}{12}T$ and $t = 0$, and likewise for 2 and 3.

pp. 896–897], is that there are no times in the fundamental domain besides the endpoints at which the configuration is either collinear or isosceles. It follows that, for all $t \in \left(-\frac{1}{12}T, 0\right)$,

$$r_{13} < r_{12} < r_{23} \tag{2}$$

and

$$q_1 \wedge q_2 = q_2 \wedge q_3 = q_3 \wedge q_1 < 0, \tag{3}$$

where $r_{ij} = |q_i - q_j|$ is the distance between masses i and j and we write

$$(x, y) \wedge (u, v) = xv - yu$$

for planar vectors (x, y) and (u, v). We call equation (2) the *distance ordering inequality*.

Initial and final velocities. At the Euler time, $t = 0$, the velocities of 2 and 3 are antiparallel to the velocity of 1 and half its size. See Figure 1. This follows from the action minimization of the eight. At the isosceles time $t = -\frac{1}{12}T$, the velocity of 2 is vertical, pointing down, and the velocities of 1 and 3 are such that their tangent lines pass through 2. This follows from the three-tangents theorem and the angular momentum property, both of which are described below.

Angular momentum and star-shapedness. Write

$$\ell_j = q_j \wedge \dot{q}_j$$

for the angular momentum of the j-th particle. Action minimization of the eight implies that its total angular momentum is zero:

$$\ell_1 + \ell_2 + \ell_3 = 0$$

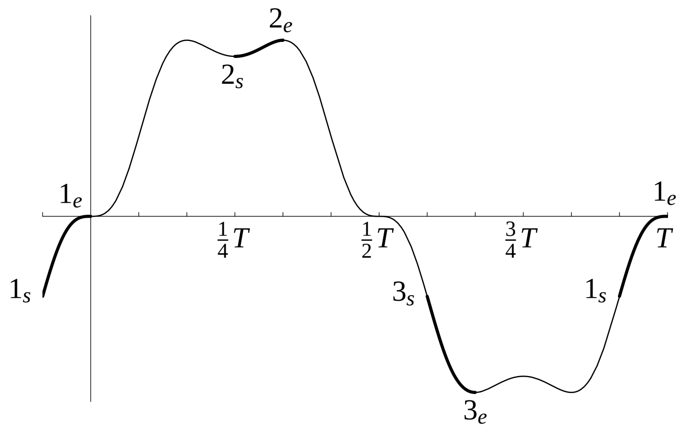

Figure 2. $\ell(t)$ versus t.

of the eight. Newton's equations imply (see [Chenciner and Montgomery 2000, p. 896])

$$\dot{\ell}_3 = \left(\frac{1}{r_{13}^3} - \frac{1}{r_{23}^3}\right)(q_1 \wedge q_2)$$

for all time. Upon taking account the distance inequality (2) and (3) we find that $\dot{\ell}_3 < 0$ on the arc 3. Similarly,

$$\dot{\ell}_1 > 0, \qquad \dot{\ell}_2 > 0, \qquad \dot{\ell}_3 < 0.$$

We use the notation 1_s to indicate body 1 at the starting time $t = -\frac{1}{12}T$, etc. By the symmetry, $\ell_{1_s} = \ell_{3_s} = -2\ell_{2_s} < 0$. (The inequalities $\ell_{1_s} < 0$ and $\ell_{1_e} = 0$ are consistent with $\dot{\ell}_1 > 0$.) Also $\ell_{2_s} > 0$ and $\dot{\ell}_2 > 0$ imply $\ell_{2_e} = -\ell_{3_e} > 0$. (See Figure 2.) Therefore over the interior $(-\frac{1}{12}T, 0)$ of our fundamental domain we have

$$\ell_1 < 0, \qquad \ell_2 > 0, \qquad \ell_3 < 0.$$

More generally, set

$$\ell = q \wedge \dot{q}$$

as q varies over the eight. It follows that on the right lobe ($x > 0$) we have

$$\ell < 0 \quad \text{for } x > 0.$$

(See Figure 2.)

A curve in the plane is called *star-shaped* with respect to the origin if every ray from the origin intersects the curve at most once. For a smooth curve, this is equivalent to the assertion that, when written in polar coordinates as $(r(t), \theta(t))$, the function $\theta(t)$ is strictly monotone and does not vary by more than 2π. Since

$\ell = r^2\dot{\theta}$ the star-shapedness of a curve (such as one lobe of the eight) which lies in the half-plane $x > 0$ is thus equivalent to $\ell \neq 0$.

The three-tangents theorem. The following theorem can be found in [Fujiwara et al. 2003], where it was used to establish the existence of a choreographic three-body lemniscate for a non-Newtonian potential.

Theorem 2 (Three tangents). *Let $(q_1(t), q_2(t), q_3(t))$ be three planar curves whose total linear and total angular momentum are zero. Then the three instantaneous tangent lines to these three curves are coincident — they all three intersect in the same (time-dependent) point or are parallel.*

Proof. Fix the time t. Because $\dot{q}_1 + \dot{q}_2 + \dot{q}_3 = 0$, translating all the q_i in the same fixed direction does not change the condition of having zero angular momentum. So, without loss of generality, we can choose the origin to be the point of intersection of the tangent lines to q_1 and q_2 at time t. Because the point $q_1(t)$ lies along the line through the origin in the direction $\dot{q}_1$ we have $q_1(t) \wedge \dot{q}_1(t) = 0$. Similarly $q_2(t) \wedge \dot{q}_2(t) = 0$. But the total angular momentum is zero so we must have $q_3(t) \wedge \dot{q}_3(t) = 0$ which asserts that the line tangent to the curve of q_3 at t also passes through the origin. □

The proof also works for unequal masses m_1, m_2, m_3. Simply use the correct mass-weighted formulae for linear and angular momentum.

The splitting lemma. We will use the following splitting lemma in several places in the proof. A line in the plane divides the plane into three pieces: two open half-planes and the line itself. We say that a point lies *strictly on one side* of the line if it lies in one of the open half-planes. We say that this line *splits* the points A and B of the plane if the two points lie in opposite open half-planes.

Lemma 1. *Let $(q_1(t), q_2(t), q_3(t))$ be a planar solution to Newton's three-body equation with attractive $1/r$ potential. Suppose that at time t_* the arc $q_i(t)$ of mass i has an inflection point and nonzero speed. Then the tangent line ℓ to this arc at time t_* must either* (A) *split the other two masses $q_j(t_*)$ and $q_k(t_*)$ or* (B) *all three masses must lie on this tangent line.*

Proof. Suppose, to the contrary, that either both $q_j(t_*)$ and $q_k(t_*)$ lie strictly on one side of ℓ, or that one lies on ℓ while the other lies strictly on one side. According to Newton's equations the acceleration $\ddot{q}_i(t_*)$ is a linear combination of $q_j(t_*) - q_i(t_*)$ and $q_k(t_*) - q_i(t_*)$ and the coefficients of this linear combination are positive. Thus, translating ℓ and the configuration of masses back to the origin by subtracting $q_i(t_*)$, we see that this acceleration lies strictly on one side of the line through 0 spanned by the velocity $\dot{q}_i(t_*)$. Consequently, the acceleration and velocity of $q_i(t)$ are *linearly independent* at t_*. But the condition of being an inflection point is precisely that the acceleration and velocity be linearly dependent. □

The same proof works if the Newtonian potential $-\sum_{i<j} m_i m_j / r_{ij}$ is replaced by any potential $V = \sum_{i<j} f(r_{ij})$, where $df/dr > 0$.

A Convexity Proposition. A parametrization t of a curve C is *nondengenerate* if the derivative $dC(t)/dt$ is never zero. A smooth, possibly self-intersecting curve is called *locally convex* if its curvature never vanishes.

Proposition. *Let C be a smooth locally convex planar curve parametrized by a nondegenerate parameter t. Let $\ell(t)$ be the tangent to C at $C(t)$. Let m be a line not intersecting C. Let $P(t)$ be the point of intersection of $\ell(t)$ and m. Then $P(t)$ moves on the line m always in the same direction, for all t such that $P(t)$ is finite.*

Proof. We can take m to be the y-axis. If C is parametrized by $(x(t), y(t))$, the line $\ell(t)$ is given by $\{(x(t), y(t)) + \lambda(\dot{x}(t), \dot{y}(t)) : \lambda \in \mathbb{R}\}$, and it intersects m at $P(t) = (0, p(t))$, where

$$p = -\frac{x(t)\dot{y}(t) - y(t)\dot{x}(t)}{\dot{x}(t)}.$$

Differentiation and the definition of the curvature κ yield

$$\frac{dp}{dt} = -\frac{v^3 x}{\dot{x}^2}\kappa,$$

where $v = \sqrt{\dot{x}^2 + \dot{y}^2}$ is the curve's speed. The factors v, x, κ are never zero by assumption (in the case of x because C avoids m); therefore they have constant sign. Thus dp/dt has constant sign wherever defined. □

3. To each mass its own quadrant

A crucial ingredient in the proof of Theorem 1 is that each mass "stays in its own quadrant" during the time interval $(-\frac{1}{12}T, 0)$. Initially 3 is in the first quadrant, 1 is in the fourth, and 2 is on the x-axis between the second and third quadrants, moving into the third. Hence, for a short time interval $(-\frac{1}{12}T, -\frac{1}{12}T + \epsilon)$, mass 3 lies in the first quadrant, 1 in the fourth, and 2 in the third.

Lemma 2. *Over the time interval $(-\frac{1}{12}T, 0)$ body* 1 *lies in the fourth quadrant, body* 2 *lies in the third, and body* 3 *lies in the first.*

Proof. Suppose one of the masses leaves its initial quadrant before time 0. It must exit along the boundary of this quadrant. It cannot exit through the origin, as this would imply an Euler configuration and the only Euler configuration occurs at the endpoint of the interval.

We argue individually that each mass cannot be the first to exit. Suppose that 2 exits first (perhaps simultaneously with another). It cannot leave crossing the x-axis, as this would contradict star-shapedness of the lobe it lies on. Neither can it exit through the y-axis, for then its x-coordinate would be zero, and, because collinearity of the three masses is excluded, at least one of 1 and 3 would not be exiting at the same time and so would have a positive x-coordinate. Thus the sum of the x-coordinates of the masses would be positive, contradicting that the center of mass is at the origin.

Mass 1 cannot leave first. For it cannot leave through the x-axis, as this would again contradict star-shapedness. It cannot leave through the y-axis as this would violate the distance ordering $r_{13} < r_{12} < r_{23}$ guaranteed by (2). To see this violation, write the exit point for mass 1 as $(0, y_1)$, with $y_1 < 0$. Then the other masses must be at $(-x, y_2)$ and (x, y_3) with $x > 0$ (since the configuration cannot be collinear) and $y_2 < 0$, $y_3 > 0$. We have $r_{13}^2 = x^2 + (y_3 - y_1)^2$ and $r_{12}^2 = x^2 + (y_2 - y_1)^2$. But $y_3 > 0,\ 0 > y_1, y_2$, and $y_1 + y_2 + y_3 = 0$, so

$$y_3 - y_1 = -2y_1 - y_2 = 2|y_1| + |y_2|,$$

while $|y_2 - y_1| < |y_2| + |y_1|$, so that $(y_3 - y_1)^2 > (y_2 - y_1)^2$ and $r_{13} > r_{12}$, contradicting the distance ordering.

Mass 3 cannot leave first. It cannot exit across the x-axis, for if it did the center of mass of the system would have a negative y-coordinate. It cannot leave across the y-axis, for this would contradict star-shapedness. □

4. Proof of Theorem 1

We refer to the arc swept out by mass j during the the time interval $\left[-\frac{1}{12}T, 0\right]$ as arc j, and write κ_j for its curvature. We must show that $\kappa_1 \leq 0$ with $\kappa_1 < 0$ for $t \neq 0$, that $\kappa_2 > 0$ and that $\kappa_3 < 0$.

Convexity of arc 1. We begin by showing that $\ddot{y}_1 > 0$ along arc 1. Since each mass stays in its own quadrant, we have $y_3 - y_1 > 0$; moreover $r_{13} < r_{12}$ by (2). Thus

$$\begin{aligned}\ddot{y}_1 &= (y_3 - y_1)/r_{13}^3 + (y_2 - y_1)/r_{12}^3 \\ &> (y_3 - y_1)/r_{12}^3 + (y_2 - y_1)/r_{12}^3 \\ &= -3y_1/r_{12}^3 > 0.\end{aligned}$$

Next we show that $\dot{y}_1 > 0$ along the arc. From the fact that $\ddot{y}_1 > 0$, it suffices to show that $\dot{y}_1 > 0$ at the initial point of arc 1, the isosceles point. By the three-tangents theorem and the fact that $\ell_1 < 0$ it follows that at the isosceles point $\dot{q}_1$ points from q_1 to the vertex q_2, so that $\dot{y}_1 > 0$.

We have seen that $\ell_1 < 0$ while $\dot{\ell}_1 > 0$ along the arc. Combining these inequalities, we see that $\dot{\ell}_1 \dot{y}_1 - \ell_1 \ddot{y}_1 > 0$ holds along the arc. On the other hand, expanding the angular momentum, we get $\dot{\ell}_1 \dot{y}_1 - \ell_1 \ddot{y}_1 = (x_1 \ddot{y}_1 - y_1 \ddot{x}_1)\dot{y}_1 - (x_1 \dot{y}_1 - y_1 \dot{x}_1)\ddot{y}_1 = y_1(\dot{x}_1 \ddot{y}_1 - \dot{y}_1 \ddot{x}_1) = y_1 v_1^3 \kappa$. Thus $y_1 v_1^3 \kappa_1 > 0$. Since $y_1 < 0$, $v_1 > 0$ we have $\kappa_1 < 0$.

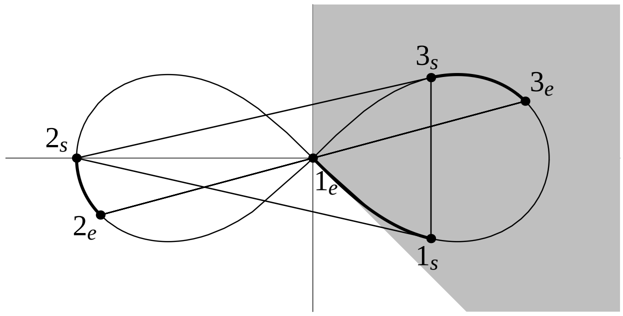

Figure 3. Region for bodies 1 and 3.

Convexity of arc 2. Assume, by way of contradiction, that there exists an inflection point $\kappa_2 = 0$ on arc 2. Let a be the last inflection point on arc 2 — the one whose time t is closest to 0. From the initial conditions at $t = -\frac{1}{12}T, 0$ described above we also know that $\kappa_2 > 0$ at the points 2_s and 2_e. By continuity, $\kappa_2 > 0$ near both of these points. Then $\kappa_2 > 0$ on the arc from a to 2_e.

We already know that arc 1 is convex ($\kappa_1 < 0$) and we also know that body 3 moves in the first quadrant. It follows that bodies 1 and 3 must lie within the shaded region in the Figure 3.

Consider the Gauss map (hodograph) of arc 2. This is the map that assigns to a point of arc 2 the unit tangent to arc 2, $\dot{q}_2/|\dot{q}_2|$, at that point.

By Newton's equation and the fact that $x_1 - x_2$ and $x_3 - x_2$ are positive we have $\ddot{x}_2 > 0$ on the entire arc 2. Since $\dot{x}_2 = 0$ at 2_s, this implies that $\dot{x}_2 > 0$ on the open arc of 2, from 2_s to 2_e, and so in particular $\dot{x}_2 > 0$ at a. Since $\kappa_2 > 0$ on the arc $a \to 2_e$, the vector $\dot{q}_2/|\dot{q}_2|$ must approach 2_e from the point a monotonically counterclockwise. Therefore the point a lies on the arc between the points 2_s and 2_e on the right half of the circle as shown in the Gauss map (Figure 4).

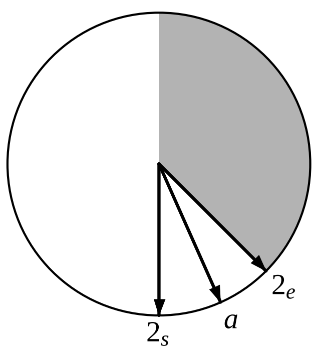

Figure 4. Gauss map of the unit tangent vector $\dot{q}_2/|\dot{q}_2|$.

But then the tangent line to arc 2 at a cannot split the points 1 and 3, which, according to the splitting lemma (Lemma 1), contradicts the assumption that a is an inflection point.

Thus we have proved that arc 2 has no inflection points, that is, $\kappa_2 > 0$.

Convexity of arc 3. Assume, by way of contradiction, that there are inflection points on arc 3. Let b be the first such point, the one for which the time t is closest to $-\frac{1}{12}T$. Then, by the splitting lemma (Lemma 1), the tangent line to arc 3 at b must split bodies 1 and 2. In order to do that, the line must have passed earlier through either body 1 or body 2. We argue that both passings are impossible.

The tangent line to arc 3 cannot pass through body 1. For, by the three-tangent theorem, at the instant this happened, the tangent line from the body 2 would also pass through the body 1. We have already proved that $\kappa_2 > 0$ on the arc 2. Thus the tangent line from the body 2 never pass through the body 1 in this interval. (See Figures 3 and 4.) This is a contradiction.

The tangent line to arc 3 cannot pass through body 2. For if it did, by the three-tangents theorem (Theorem 2), the tangent line to 1's curve would also pass through body 2 at the same instant. To see that this latter passing is impossible, start by joining the endpoints 2_s and 2_e of arc 2 by a straight line m (see Figure 5). Arc 2 lies completely on one side of this line, by convexity.

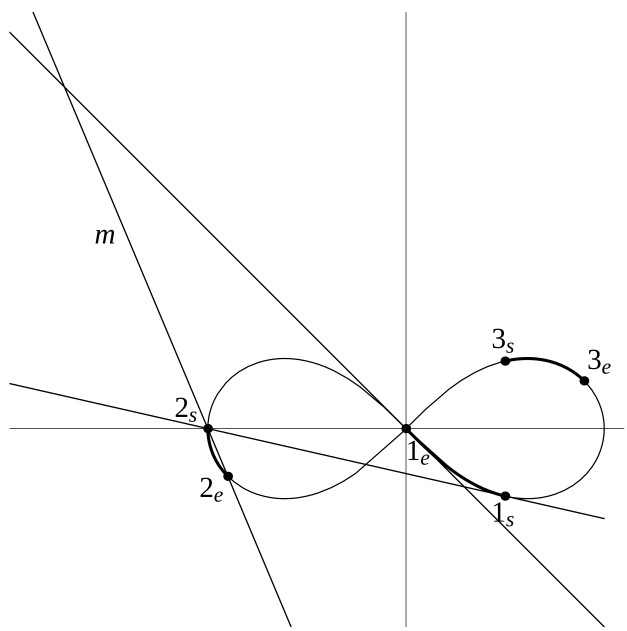

Figure 5. Line m and tangent lines to arc 1 at $t = -\frac{1}{12}T$ and $t = 0$.

We now apply the Proposition on page 276 to our situation. At the final points 1_e and 2_e, the tangents to 1 and 2 are parallel, so the intersection of m with 1's tangent lies in the massless quadrant $x < 0$, $y > 0$. At the initial point s the intersection point of m and arc 1's tangent is 2_s. We claim that for $-\frac{1}{12}T < t < 0$ the moving intersection point of 1's tangent with m always lies in the empty quadrant. This follows from the convexity of 1 and 2: the tangent at 1 rotates clockwise, while m stays fixed. When 1 finally reaches the endpoint 1_e its tangent is parallel to 2_e's, which in turn lies 'earlier' on the clockface than m (by 2's convexity). So 1's tangent can never have been tangent to m, and hence the intersection point remains finite, in the empty quadrant.

Now recall that we are trying to show that the tangent to 1 cannot pass through point 2. To do so it would have to cross line m between 2_s and 2_e, which is in the quadrant of arc 2, and hence it is impossible that this tangent passes through 2.

Therefore, we have proved that there is no inflection point on the arc 3. In other word, $\kappa_3 < 0$ on the arc 3.

Putting together the convexity of all three arcs we obtain Theorem 1.

5. Convexity for other potentials

Theorem 1 holds for the figure eight solution of other potentials. Indeed, our proof only depended on the properties of the eight listed in Section 2 and a monotonicity property of the Newtonian potential discussed below.

To be precise, we need to define what we mean by an eight. Let

$$V = V(r_{12}, r_{23}, r_{31})$$

be a three-body potential depending only on the interparticle distances r_{ij} and invariant under interchange of the masses. Then the symmetry group D_6 of the eight acts on solutions to the corresponding Newton equation, taking solutions to solutions, and so we can speak of D_6-*invariant solutions*.

A planar solution to the Newton's equation for V is called an *eight solution* if

(i) it is invariant under the D_6 symmetries,

(ii) on the interior of each fundamental domain $\left(m\frac{1}{12}T, (m+1)\frac{1}{12}T\right)$, for $m = 0, \pm 1, \pm 2, \ldots$, the configuration is never collinear and never isosceles, and

(iii) the solution has no collisions.

Such a solution will necessarily be a planar choreography (see (1) on page 272), and so the three masses travel a single planar curve. Condition (i) implies that the center of mass is 0 and that the angular momentum is zero. If, in addition, our potential V has the form

$$V = \sum_{i<j} f(r_{ij}),$$

where

(iv) $df/dr > 0$ (attractive two-body potential) and

(v) $g(r) := r^{-1} df/dr$ is a strictly monotone decreasing function of r,

then all properties and inequalities used in this paper hold.

Indeed, return to the starting point, the distance ordering inequality (2). At $t = -\frac{1}{12}T$ we have $r_{23} = r_{12}$, and at $t = 0$ we have $r_{12} = r_{31} < r_{23} = 2r_{12}$. By property (ii), the possible distance orderings on the time interval $\left(-\frac{1}{12}T, 0\right)$ are $r_{31} < r_{12} < r_{23}$ or $r_{12} < r_{31} < r_{23}$. Consider the equation for $\dot{\ell}_1$,

$$\dot{\ell}_1 = \big(g(r_{21}) - g(r_{31})\big)(q_2 \wedge q_3),$$

for a monotone decreasing function $g(r)$. We have $\dot{\ell}_1 > 0$ for the first ordering and $\dot{\ell}_1 < 0$ for the second ordering. But, since $\ell_1 < 0$ at $t = -\frac{1}{12}T$ and $\ell_1 = 0$ at $t = 0$, the value of $\dot{\ell}_1$ must be positive. So we must have the first ordering, namely, equation (2). Then all equalities and inequalities in this paper hold. Thus:

Theorem 3. *Let V be a three-body potential of the form $V = \sum_{i<j} f(r_{ij})$ where f satisfies* (iv) *and* (v) *above, and admitting an eight solution as defined by* (i)–(iii) *above. Then each lobe of this eight for V is convex.*

The theorem begs the question, do eight solutions exist for any potentials besides Newton's? Recall from [Chenciner and Montgomery 2000, pp. 896–897] that if a solution that satisfies (i) and (ii) is known to minimize the action associated to V among all paths satisfying (i), and if that solution is not identically collinear, then automatically the solution satisfies (ii). The power law potentials

$$V_a = (a)^{-1}(r_{12}^a + r_{23}^a + r_{31}^a),$$

for $a \le -2$ admit such collision-free action minimizing solutions, and consequently they admit eight solutions. Moreover, the proof of [Chenciner and Montgomery 2000], specific to $a = -1$, is based on strict inequalities, and hence is valid for a range of exponents $-1-\epsilon_1 < a < -1+\epsilon_2$ for ϵ_1, ϵ_2 positive numbers. Numerical evidence presented in [Chenciner et al. 2002] suggests that eights exist for all power laws V_a, where $a < 0$. (These eights are dynamically stable only in a neighborhood of the Newtonian potential $a = -1$.)

Corollary. *For the power law potentials V_a with $a \le -2$ or with a in some open interval about -1, there exist eight solutions and each lobe of these eight solutions is convex.*

6. Unicity

Showing the unicity of the Newtonian eight remains an open problem [Chenciner 2003]. Our work here drastically reduces the candidate eights, and hence the scope of nonunicity, to those eights with convex lobes. It might allow a handhold towards surmounting the unicity problem. If our reader will allow us to fantasize in this direction, imagine two distinct Newtonian eights, both enjoying (i) D_6 symmetry, (ii) the same period, and (iii) having the same minimum value for the action. Connect these two eights by a family of eights having (i) and (ii), and having convex lobes. Apply the min-max procedure to extract out of such a family a third eight that is variationally unstable, meaning that the Hessian of the action there has a negative direction. Now establish a contradiction between the existence of the negative mode and the convexity of the lobe of this third eight. Such a program, or a similar one, could conceivably lead to a proof of unicity of the eight.

Acknowledgments

We express our sincere thanks to Alain Chenciner, Hiroshi Fukuda and Hiroshi Ozaki for discussions.

Note added in proof

For the power law potentials V_a, Barutello, Ferrario and Terracini [Barutello et al. 2004] have proved existence of eights for all $a < 0$; see the proof following Proposition (4.15) on p. 19. Montgomery [2004] has proved the uniqueness of the eight for $a = -2$.

References

[Barutello et al. 2004] V. Barutello, D. L. Ferrario, and S. Terracini, “Symmetry groups of the planar 3-body problem and action-minimizing trajectories”, 2004. math.DS/0404514

[Chenciner 2003] A. Chenciner, “Some facts and more questions about the Eight”, pp. 77–88 in *Topological methods, variational methods and their applications: Proceedings of the ICM 2002 Satellite Conference on Nonlinear Functional Analysis* (Taiyuan, China, 2002), edited by H. Brézis et al., World Scientific, River Edge, NJ, 2003. MR 2004m:70012 Zbl 02079192

[Chenciner and Montgomery 2000] A. Chenciner and R. Montgomery, “A remarkable periodic solution of the three-body problem in the case of equal masses”, *Ann. of Math.* (2) **152**:3 (2000), 881–901. MR 2001k:70010 Zbl 0987.70009

[Chenciner et al. 2002] A. Chenciner, J. Gerver, R. Montgomery, and C. Simó, “Simple choreographic motions of N bodies: a preliminary study”, pp. 287–308 in *Geometry, mechanics, and dynamics*, edited by P. Newton et al., Springer, New York, 2002. MR 2003f:70019 Zbl 01820129

[Fujiwara et al. 2003] T. Fujiwara, H. Fukuda, and H. Ozaki, “Choreographic three bodies on the lemniscate”, *J. Phys. A* **36**:11 (2003), 2791–2800. MR 2004b:70020 Zbl 02071994

[Kapela and Zgliczyński 2003] T. Kapela and P. Zgliczyński, “The existence of simple choreographies for the N-body problem: a computer-assisted proof”, *Nonlinearity* **16**:6 (2003), 1899–1918. MR 2004h:70019 Zbl 02028748

[Montgomery 2004] R. Montgomery, “Fitting hyperbolic pants to a three-body problem”, 2004. math.DS/0405014

[Moore 1993] C. Moore, "Braids in classical dynamics", *Phys. Rev. Lett.* **70**:24 (1993), 3675–3679. MR 94d:58055 Zbl 1050.37522

Received September 2, 2003. Revised November 5, 2003.

TOSHIAKI FUJIWARA
COLLEGE OF LIBERAL ARTS AND SCIENCES
KITASATO UNIVERSITY
KITASATO 1-15-1, SAGAMIHARA
KANAGAWA 228-8555
JAPAN
fujiwara@clas.kitasato-u.ac.jp

RICHARD MONTGOMERY
MATHEMATICS DEPARTMENT
UC SANTA CRUZ
SANTA CRUZ, CA 95064
UNITED STATES
rmont@count.ucsc.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

BOSONIC REALIZATIONS OF HIGHER-LEVEL TOROIDAL LIE ALGEBRAS

NAIHUAN JING, KAILASH MISRA AND SHAOBIN TAN

We construct realizations for the 2-toroidal Lie algebra associated with the Lie algebra A_1 using vertex operators based on bosonic fields. In particular our construction realizes higher-level representations of the 2-toroidal algebra for any given pair of levels (k_0, k_1) with $k_0 \neq 0$. We also construct a smaller module of level $(k_0, 0)$ for the toroidal algebra from the Fock space using certain screening vertex operator, and this later representation generalizes the higher-level construction of the affine Lie algebra $\widehat{\mathfrak{sl}}_2$.

1. Introduction

Toroidal Lie algebras are a natural generalization of the affine Kac–Moody algebras introduced by Moody, Rao and Yokonuma [Moody et al. 1990]. Let $A = \mathbb{C}[s, s^{-1}, t, t^{-1}]$ be the ring of Laurent polynomials in commuting variables. By definition a 2-toroidal Lie algebra is a perfect central extension of the iterated loop algebra $\mathfrak{g} \otimes A$, where $\mathfrak{g}$ is a finite-dimensional simple Lie algebra over $\mathbb{C}$.

Let Ω_A/dA be the Kähler differentials of A modulo the exact forms. The universal central extension of the iterated loop algebra is given by

$$T(\mathfrak{g}) = (\mathfrak{g} \otimes A) \oplus \Omega_A/dA.$$

Any 2-toroidal Lie algebra is a homomorphic image of this toroidal Lie algebra. The center of $T(\mathfrak{g})$ is Ω_A/dA, which is a infinite-dimensional vector space. The Laurent polynomial ring A induces a natural $\mathbb{Z}^2$-gradation on $T(\mathfrak{g})$. For the center we have $\Omega_A/dA = \bigoplus_{\sigma\in\mathbb{Z}^2} \mathscr{Z}(\mathfrak{g})_\sigma$, with $\dim \mathscr{Z}_\sigma = 1$ if $\sigma \neq (0, 0)$ and 2 if $\sigma = (0, 0)$. We denote by c_0 and c_1 the two standard degree-zero central elements in the toroidal Lie algebra $T(\mathfrak{g})$. A module of $T(\mathfrak{g})$ is called a level-(k_0, k_1) module if the standard center (c_0, c_1) acts as (k_0, k_1) for some complex numbers k_0 and k_1. Here we study the level-(k_0, k_1) modules for $k_0 \neq 0$.

MSC2000: 17B65, 17B69,

Keywords: toroidal Lie algebra, vertex operator, bosonic realization.

Jing acknowledges the partial support from the NSF and NSA. Misra acknowledges the partial support from the NSA. Tan was supported by the National Science Foundation of China, No. 10371100.

In various constructions of the affine Lie algebra $\widehat{\mathfrak{sl}}_2$ the free field representation is of particular use in its applications. Wakimoto [1986] and Feigen and Frenkel [1988] first gave a general construction for the general case, and later Nemenschansky [1989] gave an invariant form in the special case. Though the two forms can be interchanged by a nontrivial map, we realized that the later form is better for our purpose in the toroidal cases. The operators in question have the form $e^A(B+C)$, where A, B, C are generating functions of the scaled Heisenberg operators. One of the nice things is that all root generators in the toroidal algebra associated with the Lie algebra $\mathfrak{sl}_2$ can be represented by this type of vertex operators. In our construction we have fully used this simplicity and make all calculations in a uniform manner.

As we mentioned earlier, toroidal algebras are generalizations of finite-dimensional Lie algebras, like affine Lie algebras. This similarity is constantly kept in mind as we study their structure and representation theory. Some other basic references related to our work include [Berman and Billig 1999; Eswara Rao and Moody 1994; Fabbri and Moody 1994; Larsson 1999; Moody et al. 1990; Tan 1999]. Our aim in this paper is to give a higher-level representation for the simplest nontrivial example: the 2-toroidal Lie algebra. Our construction generalizes previous work on higher-level representations of the affine Lie algebra $\widehat{\mathfrak{sl}}_2$.

In Section 2 we define the toroidal Lie algebra and state the MRY-presentation [Moody et al. 1990] of the toroidal algebra in terms of generators and relations. The algebra structure is expressed in terms of formal power series identities. We also state some results in this section to be used later. In Section 3 we start with a finite-rank lattice with a symmetric bilinear form and define a Fock space and some vertex operators, which in turn give representations of the toroidal Lie algebra of type A_1, and also a level-$(k_0, 0)$ module with $k_0 \neq 0$ for the double affine algebra of type A_1. In Section 4 we study the structure of the Fock space for the toroidal Lie algebra by using certain screening vertex operators, thus generalizing the higher-level representation of the affine algebra $\widehat{\mathfrak{sl}}_2$ to the toroidal Lie algebra.

2. Toroidal Lie algebras

Let $\mathfrak{sl}_2$ be the 3-dimensional simple Lie algebra over the complex numbers and

$$A = \mathbb{C}[s, s^{-1}, t, t^{-1}]$$

the ring of Laurent polynomials in commuting variables. We consider the iterated loop algebra

$$\mathfrak{g} = \mathfrak{sl}_2 \otimes A.$$

A toroidal Lie algebra of type A_1 is a perfect central extension of the iterated loop algebra $\mathfrak{g}$, which is often an infinite-dimensional central extension. Let Ω_A be the

A-module of differentials with differential mapping $d : A \to \Omega_A$, such that

$$d(f_1 f_2) = (df_1) f_2 + f_1(df_2) \quad \text{for all } f_1,\ f_2 \text{ in } A.$$

Let $\overline{} : \Omega_A \to \Omega_A/dA$ be the canonical linear map for which $\overline{df} = 0$ for all $f \in A$. Endow the vector space

$$T(A_1) := (\mathfrak{sl}_2 \otimes A) \oplus \Omega_A/dA$$

with the bracket operation defined by

$$[x \otimes f_1, y \otimes f_2] = [x, y] \otimes f_1 f_2 + (x, y)\, \overline{f_2 df_1},$$

for $x, y \in \mathfrak{sl}_2$, f_1, $f_2 \in A$, where $(\cdot, \cdot)$ is the trace form and Ω_A/dA is central. From [Moody et al. 1990] we know that $T(A_1)$ is a perfect Lie algebra and is the universal central extension of the iterated loop algebra $\mathfrak{sl}_2 \otimes A$. Therefore any toroidal Lie algebra of type A_1 is a homomorphic image of $T(A_1)$. The gradation of the polynomial ring A gives a natural $\mathbb{Z}^2$-gradation to the toroidal Lie algebra

$$T(A_1) := \bigoplus_{\sigma \in \mathbb{Z}^2} T(A_1)_\sigma,$$

where $T(A_1)_\sigma$ is spanned by $x \otimes s^{m_0} t^{m_1}$, $\overline{s^{m_0} t^{m_1} s^{-1} ds}$ and $\overline{s^{m_0} t^{m_1} t^{-1} dt}$ for $\sigma = (m_0, m_1) \in \mathbb{Z}^2$ and $x \in \mathfrak{sl}_2$. The condition $\overline{df} = 0$ for all $f \in A$ implies that $m_0 \overline{s^{m_0} t^{m_1} s^{-1} ds} + m_1 \overline{s^{m_0} t^{m_1} t^{-1} dt} = 0$ for all $m_0, m_1 \in \mathbb{Z}$. Therefore the dimension of $T(A_1)_\sigma$ is 4 if $\sigma \neq (0, 0)$ and 5 if $\sigma = (0, 0)$. In particular, $T(A_1)_{(0,0)}$ is spanned by $x \otimes 1$ for $x \in \mathfrak{sl}_2$, and central elements $\overline{s^{-1} ds}$, $\overline{t^{-1} dt}$. We denote these two degree-zero central elements by c_0 and c_1.

The most interesting quotient algebra of the toroidal Lie algebra $T(A_1)$ is the double affine algebra denoted by $T_0(A_1)$, that is, the toroidal Lie algebra of type A_1 with a two-dimensional center. The double affine algebra is the quotient of $T(A_1)$ modulo all the central elements with degree other than zero. In fact, $T_0(A_1)$ has the realization

$$T_0(A_1) = (\mathfrak{sl}_2 \otimes A) \oplus \mathbb{C}c_0 \oplus \mathbb{C}c_1$$

with the Lie product

$$[x \otimes f_1, y \otimes f_2] = [x, y] \otimes f_1 f_2 + \Phi(f_2 \partial_s f_1) c_0 + \Phi(f_2 \partial_t f_1) c_1$$

for all $x, y \in \mathfrak{sl}_2$ and f_1, $f_2 \in A$, where Φ is the linear functional on A defined by

$$\Phi(s^k t^m) = \begin{cases} 0, & \text{if } (k, m) \neq (0, 0) \\ 1, & \text{if } (k, m) = (0, 0) \end{cases}$$

for all $k, m \in \mathbb{Z}$.

Definition 2.1. If M is a module for a toroidal Lie algebra of type A_1, we call M a level-(k_0, k_1) module for some complex numbers k_0, k_1 if the degree-zero central elements c_0, c_1 act on M as constants k_0, k_1.

In this paper we give a concrete construction for a level-(k_0, k_1) module with $k_0 \neq 0$ for the toroidal Lie algebra $T(A_1)$ and for the double affine algebra $T_0(A_1)$.

Let $\{x_\pm, h\}$ be the standard basis of $\mathfrak{sl}_2$. Also let $(a_{ij})_{2\times 2}$ be the generalized Cartan matrix of the affine algebra $A_1^{(1)}$ and

$$Q := \mathbb{Z}\alpha_0 + \mathbb{Z}\alpha_1$$

its root lattice. The toroidal Lie algebra $T(A_1)$ has a presentation [Moody et al. 1990] with generators $¢$, $\alpha_i(k)$ and $x_k(\pm\alpha_i)$, for $k \in \mathbb{Z}$ and $i = 0, 1$, and the following relations, for $k, m \in \mathbb{Z}$ and $i, j = 0, 1$:

(R0) $\big[¢, \alpha_i(k)\big] = 0 = \big[¢, x_k(\pm\alpha_i)\big]$;

(R1) $\big[\alpha_i(k), \alpha_j(m)\big] = k a_{ij}\delta_{k+m,0}¢$;

(R2) $\big[\alpha_i(k), x_m(\pm\alpha_j)\big] = \pm a_{ij} x_{k+m}(\pm\alpha_j)$;

(R3) $\big[x_k(\alpha_i), x_m(-\alpha_j)\big] = -\delta_{ij}\,\big\{\alpha_i(k+m) + k\delta_{k+m,0}¢\big\}$;

(R4) $\big[x_k(\alpha_i), x_m(\alpha_i)\big] = 0 = \big[x_k(-\alpha_i), x_m(-\alpha_i)\big]$;

$$(\operatorname{ad} x_0(\alpha_i))^3\, x_m(\alpha_j) = 0 \text{ if } i \neq j; \quad (\operatorname{ad} x_0(-\alpha_i))^3\, x_m(-\alpha_j) = 0 \text{ if } i \neq j.$$

The Lie algebra isomorphism ψ between the two presentations of $T(A_1)$ is given by

$$\begin{aligned} ¢ &\mapsto \overline{s^{-1}ds}, \\ x_m(\pm\alpha_1) &\mapsto \pm x_\pm \otimes s^m, \\ x_m(\pm\alpha_0) &\mapsto \pm x_\mp \otimes s^m t^{\pm 1}, \\ \alpha_1(k) &\mapsto h \otimes s^k, \\ \alpha_0(k) &\mapsto -h \otimes s^k + \overline{s^k t^{-1} dt}. \end{aligned}$$

Therefore, the degree-zero central elements are $c_0 = ¢$ and $c_1 = \delta(0)$, where $\delta = \alpha_0 + \alpha_1$ is the null root in Q. We will identify the two presentations of the toroidal Lie algebra $T(A_1)$ via this isomorphism ψ.

Following [Moody et al. 1990], we introduce a $\mathbb{Z} \times Q$-gradation on $T(A_1)$ by assigning $\deg ¢ = (0, 0)$, $\deg \alpha_i(k) = (k, 0)$, $\deg x_k(\pm\alpha_i) = (k, \pm\alpha_i)$, with $i = 0, 1$ and $k \in \mathbb{Z}$. We denote by T_k^α the subspace of $T(A_1)$ spanned by the elements with degree (k, α) for $k \in \mathbb{Z}$, $\alpha \in Q$. Then, under the isomorphism ψ, we have $\psi^{-1}(\overline{s^k t^{-1} dt}) = \delta(k) \in T_k^0$ and $\psi^{-1}(\overline{s^k t^r s^{-1} ds}) \in T_k^{r\delta}$.

Let $z, w, z_1, z_2, \ldots$ be formal variables. We define formal power series with coefficients from the toroidal Lie algebra $T(A_1)$:

$$\alpha_i(z) = \sum_{n\in\mathbb{Z}} \alpha_i(n) z^{-n-1},$$
$$x(\pm\alpha_i, z) = \sum_{n\in\mathbb{Z}} x_n(\pm\alpha_i) z^{-n-1},$$

for $i = 0, 1$. Then the Lie algebra structure of $T(A_1)$ can be expressed in terms of the following power series identities:

(R0′) $\big[\not{c}, \alpha_i(z)\big] = 0 = \big[\not{c}, x(\pm\alpha_i, z)\big]$;

(R1′) $\big[\alpha_i(z), \alpha_j(w)\big] = a_{ij} z^{-1} \partial_w \delta(\frac{w}{z}) \not{c}$;

(R2′) $\big[\alpha_i(z), x(\pm\alpha_j, w)\big] = \pm a_{ij} x(\pm\alpha_j, w) z^{-1} \delta(\frac{w}{z})$;

(R3′) $\big[x(\alpha_i, z), x(-\alpha_j, w)\big] = -\delta_{ij} \left\{\alpha_i(w) z^{-1} \delta(\frac{w}{z}) + z^{-1} \partial_w \delta(\frac{w}{z}) \not{c}\right\}$;

(R4′) $\big[x(\alpha_i, z), x(\alpha_i, w)\big] = 0 = \big[x(-\alpha_i, z), x(-\alpha_i, w)\big]$;

$$(\operatorname{ad} x(\alpha_i, z_1))(\operatorname{ad} x(\alpha_i, z_2))(\operatorname{ad} x(\alpha_i, z_3)) x(\alpha_j, z_4) = 0 \quad \text{if } i \neq j;$$
$$(\operatorname{ad} x(-\alpha_i, z_1))(\operatorname{ad} x(-\alpha_i, z_2))(\operatorname{ad} x(-\alpha_i, z_3)) x(-\alpha_j, z_4) = 0 \quad \text{if } i \neq j.$$

Finally, we recall a result from [Moody et al. 1990] that will be used in the next section.

Proposition 2.2. *Suppose $\mathscr{L}$ is a Lie algebra over $\mathbb{C}$ graded by $\mathbb{Z} \otimes Q$, and $\phi : T(A_1) \to \mathscr{L}$ is a surjective graded homomorphism of Lie algebras such that*

(i) *ϕ is injective on T_n^α for all $n \in \mathbb{Z}$ and real root α,*

(ii) *$\phi(\delta(k)) \neq 0$ for all k and $\phi|_{\mathbb{C}\delta(0)+\mathbb{C}\not{c}}$ is injective, and*

(iii) *for all nonzero integers k, m,*

$$\phi\big([x_m(\alpha_1 + k\delta), x_0(-\alpha_1)] - [x_0(\alpha_1 + k\delta), x_m(-\alpha_1)]\big) \neq 0,$$
$$\phi\big([x_1(\alpha_1 + k\delta), x_{-1}(-\alpha_1)] - [x_{-1}(\alpha_1 + k\delta), x_1(-\alpha_1)]\big) \neq 0.$$

Then ϕ is an isomorphism, where $x_m(\pm\alpha_1 + k\delta) := \psi^{-1}(\pm x_\pm \otimes s^m t^k)$.

Proposition 2.3. *Suppose $\mathscr{L}$ is a Lie algebra over $\mathbb{C}$ graded by $\mathbb{Z} \otimes Q$, and $\phi : T(A_1) \to \mathscr{L}$ is a surjective graded homomorphism of Lie algebras such that*

(i) *ϕ is injective on T_n^α for all $n \in \mathbb{Z}$ and real root α,*

(ii) *$\phi(\delta(k)) = 0$ for all $k \neq 0$ and $\phi|_{\mathbb{C}\delta(0)+\mathbb{C}\not{c}}$ is injective, and*

(iii) *for all nonzero integers k, m,*

$$\phi\big([x_m(\alpha_1 + k\delta), x_0(-\alpha_1)] - [x_0(\alpha_1 + k\delta), x_m(-\alpha_1)]\big) = 0,$$
$$\phi\big([x_1(\alpha_1 + k\delta), x_{-1}(-\alpha_1)] - [x_{-1}(\alpha_1 + k\delta), x_1(-\alpha_1)]\big) = 0,$$

Then $\mathscr{L}$ is isomorphic to the double affine algebra $T_0(A_1)$.

Proof. We only need to show that the set of nonzero-degree central elements of the toroidal Lie algebra $T(A_1)$ is in the kernel of ϕ. Indeed, under the isomorphism ψ of the toroidal Lie algebras, we see that $\delta(k) = \psi^{-1}(\overline{s^k t^{-1} dt})$ and

$$\big[x_m(\alpha_1 + k\delta), x_0(-\alpha_1)\big] - \big[x_0(\alpha_1 + k\delta), x_m(-\alpha_1)\big] = -m\psi^{-1}(\overline{s^m t^k s^{-1} ds}),$$
$$\big[x_1(\alpha_1 + k\delta), x_{-1}(-\alpha_1)\big] - \big[x_{-1}(\alpha_1 + k\delta), x_1(-\alpha_1)\big] = -2\psi^{-1}(\overline{t^k s^{-1} ds}),$$

but, from [Moody et al. 1990], the elements $\overline{s^p t^q s^{-1} ds}$, $\overline{s^p t^{-1} dt}$ and $\overline{s^{-1} ds}$ for $(p, q) \in \mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$ form a basis of the center for the toroidal Lie algebra $T(A_1)$. The assumption implies that the nonzero-degree central elements $\psi^{-1}(\overline{s^p t^q s^{-1} ds})$ and $\psi^{-1}(\overline{s^q t^{-1} dt})$ are in the kernel of the homomorphism ϕ for

$$(p, q) \in \mathbb{Z} \times (\mathbb{Z} \setminus \{0\}). \qquad \square$$

3. Representations of the toroidal algebra

In this section we give two bosonic realizations for the toroidal Lie algebra $T(A_1)$. Let k_0 be a fixed complex number with $k_0 \neq 0$, and Γ a finite rank lattice with a symmetric $\mathbb{C}$-valued $\mathbb{Z}$-bilinear form $(\cdot\,,\cdot)$. We extend the form to a $\mathbb{C}$-bilinear form on the vector space $H = \mathbb{C} \otimes_{\mathbb{Z}} \Gamma$. Let Γ_0 be a fixed integral sublattice of Γ. We define

$$\Gamma_0^\star = \{\alpha \in H; (\alpha, \Gamma_0) \subset \mathbb{Z}\}.$$

Then $\Gamma_0 \subset \Gamma_0^\star$. Let

$$\mathscr{H} = \langle h(n), ¢ | h \in H, n \in \mathbb{Z}\rangle,$$

with $H = \mathbb{C} \otimes_{\mathbb{Z}} \Gamma$, be the affinization of the vector space H, defined with the Lie product

$$[\alpha(m), \beta(n)] = m(\alpha, \beta)\delta_{m+n,0}¢$$

for $m, n \in \mathbb{Z}$, $\alpha, \beta \in \Gamma$, and $¢$ central. We define the Fock space

$$V := \mathbb{C}[\Gamma_0^\star] \otimes S(\mathscr{H}^-),$$

where $S(\mathscr{H}^-)$ is the symmetric algebra on $\mathscr{H}^- := \langle h(n) \mid n < 0\rangle$, and

$$\mathbb{C}[\Gamma_0^\star] = \bigoplus_{\alpha \in \Gamma_0^\star} \mathbb{C}e^\alpha$$

is the group algebra on the additive subgroup $\Gamma_0^\star$ of the vector space H. Then V has a natural module structure for the Lie algebra $\mathscr{H}$ and the group algebra $\mathbb{C}[\Gamma_0^\star]$ with the actions defined by making $¢$ act as k_0, $h(-n)$ act as multiplication, and $h(n)$ act as a partial differential operator, for $n > 0$, $h \in H$, so that

$$[\alpha(m), \beta(n)] = mk_0(\alpha, \beta)\delta_{m+n,0}$$

for all $\alpha, \beta \in H$ and $m, n \in \mathbb{Z}$. Moreover $\alpha(0)$ acts as a partial differential operator on $\mathbb{C}[\Gamma_0^\star]$ for which $[\alpha(0), e^\beta] = (\alpha, \beta)e^\beta$. Therefore $\alpha(0).\beta = (\alpha, \beta)$ for $\alpha, \beta \in H$.

With a formal variable z, and $\alpha, \beta \in H$, we define fields

$$\begin{aligned}
\alpha(z) &= \sum_{n\in\mathbb{Z}} \alpha(n) z^{-n-1}, \\
\alpha(z)_+ &= \sum_{n<0} \alpha(n) z^{-n-1}, \\
\overline{\beta(z)} &= \beta + \beta(0)\log z - \sum_{n\neq 0} \frac{\beta(n)}{n} z^{-n}, \\
\overline{\beta(z)}_+ &= \beta - \sum_{n<0} \frac{\beta(n)}{n} z^{-n}.
\end{aligned}$$

It is easy to see that $\partial_z \overline{\beta(z)} = \beta(z)$ and $\partial_z \overline{\beta(z)}_+ = \beta(z)_+$. For

$$A, B \in \{\alpha(z), \overline{\beta(z)} \mid \alpha, \beta \in H\},$$

we define $\langle A, B\rangle = [A, B_+]$. Then it is easy to show (see [Frenkel et al. 1988]) that $\langle \overline{\alpha(z)}, \overline{\beta(w)}\rangle = (\alpha, \beta)\log(z - w)$ for $\alpha, \beta \in H$, which then implies

$$\begin{aligned}
\langle \alpha(z), \overline{\beta(w)}\rangle &= (\alpha, \beta)(z-w)^{-1}, \\
\langle \overline{\alpha(z)}, \beta(w)\rangle &= -(\alpha, \beta)(z-w)^{-1}, \\
\langle \alpha(z), \beta(w)\rangle &= (\alpha, \beta)(z-w)^{-2},
\end{aligned}$$

where the formal power series in z and w are understood to be expanded in the second variable w.

Define the usual normal ordering : : as in [Frenkel et al. 1988]. Then we have for $\alpha \in H$

$$:\alpha(z)\beta(w): = \alpha(z)\beta(w) - \langle \alpha(z), \beta(w)\rangle,$$

and, for $\alpha \in \Gamma_0$,

$$:e^{\overline{\alpha(z)}}: = e^\alpha z^{\alpha(0)} \exp\Big(-\sum_{n<0} \frac{\alpha(n)}{n} z^{-n}\Big) \exp\Big(-\sum_{n>0} \frac{\alpha(n)}{n} z^{-n}\Big).$$

It is clear that the vertex operators $:e^{\overline{\alpha(z)}}:$, for $\alpha \in \Gamma_0$, can be formally expanded as a power series in z for which the coefficients are well defined operators acting on the Fock space V.

We will need the following result in the study of the bosonic realizations for the toroidal Lie algebra $T(A_1)$; see [Jing and Lyerly 1999].

Lemma 3.1. *Let $P_i(z)$, $Q_i(w)$, for $i = 1, 2$, be fields such that the contractions $\langle P_i, Q_j\rangle$ commute with all fields $P_i(z)$, $Q_i(w)$. Then*

$$:e^{P_1}P_2::e^{Q_1}Q_2: = :e^{P_1}P_2e^{Q_1}Q_2:e^{\langle P_1,Q_1\rangle} + :e^{P_1}P_2e^{Q_1}:e^{\langle P_1,Q_1\rangle}\langle P_1, Q_2\rangle$$
$$+:e^{P_1}e^{Q_1}Q_2:e^{\langle P_1,Q_1\rangle}\langle P_2, Q_1\rangle + :e^{P_1}e^{Q_1}:e^{\langle P_1,Q_1\rangle}(\langle P_2, Q_2\rangle + \langle P_1, Q_2\rangle\langle P_2, Q_1\rangle).$$

For $\alpha, \beta \in \Gamma_1$, we have, from [Frenkel et al. 1988], the identity

$$:e^{\overline{\alpha(z)}}::e^{\overline{\beta(w)}}: = :e^{\overline{\alpha(z)}}e^{\overline{\beta(w)}}:(z-w)^{(\alpha,\beta)}.$$

Inductively one can show, for $\beta_1, \dots, \beta_k \in \Gamma_0$, the following Wick theorem

$$:e^{\overline{\beta_1(z_1)}}:\cdots:e^{\overline{\beta_k(z_k)}}: = :e^{\overline{\beta_1(z_1)}}\cdots e^{\overline{\beta_k(z_k)}}:\prod_{i<j}(z_i-z_j)^{(\beta_i,\beta_j)}.$$

Corollary 3.2. *For $\alpha, \beta \in \Gamma_0$ and $\gamma, \tau \in H$, suppose $(\alpha, \beta) = 0$. Then*

$$\left[:e^{\overline{\alpha(z)}}\gamma(z):, :e^{\overline{\beta(w)}}\tau(w):\right] = :e^{\overline{(\alpha+\beta)(z)}}A(z):z^{-1}\delta\left(\frac{w}{z}\right) + B:e^{\overline{(\alpha+\beta)(z)}}:z^{-1}\partial_w\delta\left(\frac{w}{z}\right),$$

where $A = (\gamma, \beta)\tau - (\alpha, \tau)\gamma - B\beta \in H$ and $B = (\gamma, \tau) - (\alpha, \tau)(\gamma, \beta) \in \mathbb{C}$.

To give our first representation of the toroidal Lie algebra $T(A_1)$ we consider the lattice

$$\Gamma := \frac{1}{k_0}(\mathbb{Z}a_0 \oplus \mathbb{Z}a_1 \oplus \mathbb{Z}b \oplus \mathbb{Z}r),$$

with a symmetric bilinear form determined by

$$(b, b) = -2k_0, \quad (r, r) = 2(k_0 + 2), \quad (a_i, a_j) = k_0a_{ij} \quad \text{for } i, j = 0, 1,$$

the others being zero. Let $\Gamma_0 = \frac{1}{k_0}(\mathbb{Z}(a_0 - b) + \mathbb{Z}(a_1 + b))$, which is clearly an integral sublattice of Γ. On the corresponding Fock space $V := \mathbb{C}[\Gamma_0^\star] \otimes S(\mathcal{H}^-)$, we define vertex operators

$$X_0(\pm\alpha_1, z) = \tfrac{1}{2}:e^{\pm\frac{1}{k_0}\overline{(a_1+b)(z)}}(b(z) \mp r(z)):$$
$$X_0(\pm\alpha_0, z) = \tfrac{1}{2}:e^{\pm\frac{1}{k_0}\overline{(a_0-b)(z)}}(b(z) \pm r(z)):,$$

where α_0, α_1 are the simple roots of the affine Lie algebra $A_1^{(1)}$.

Theorem 3.3. *Let k_0 be any nonzero complex number. Then on the Fock space V we have a representation for the toroidal Lie algebra $T(A_1)$. The homomorphism is given by $\not{c} \mapsto k_0$, $\alpha_i(z) \mapsto a_i(z)$, $x(\pm\alpha_i, z) \mapsto X_0(\pm\alpha_i, z)$, for $i = 0, 1$.*

Proof. We first write the vertex operators in the form

$$X_0(\pm\alpha_i, z) = \tfrac{1}{2}:e^{\pm\frac{1}{k_0}\overline{(a_i-\epsilon_ib)(z)}}(b(z) \pm \epsilon_ir(z)):,$$

where $\epsilon_i = (-1)^i$ for $i = 0, 1$. We will now show that the operators $a_i(z)$ and $X_0(\pm\alpha_i, z)$ satisfy the relations (R0′)–(R4′) of the toroidal Lie algebra $T(A_1)$. In fact, (R0′) and (R1′) are obvious. For (R2′) we have

$$\begin{aligned}[a_i(z), X_0(\pm\alpha_j, w)] &= \tfrac{1}{2}[:a_i(z):, :e^{\pm\frac{1}{k_0}\overline{(a_j-\epsilon_j b)(z)}}(b(z) \pm \epsilon_j r(z)):] \\ &= \tfrac{1}{2}:e^{\pm\frac{1}{k_0}\overline{(a_j-\epsilon_j b)(z)}}A(z):z^{-1}\delta\Big(\frac{w}{z}\Big),\end{aligned}$$

where $A = (a_i, \pm\frac{1}{k_0}(a_j - \epsilon_j b))(b \pm \epsilon_j r) = \pm a_{ij}(b \pm \epsilon_j r)$. Therefore

$$\begin{aligned}[a_i(z), X_0(\pm\alpha_j, w)] &= \pm\tfrac{1}{2}a_{ij}:e^{\pm\frac{1}{k_0}\overline{(a_j-\epsilon_j b)(z)}}(b(z) \pm \epsilon_j r(z)):z^{-1}\delta\Big(\frac{w}{z}\Big) \\ &= \pm a_{ij}X_0(\pm\alpha_j, z)z^{-1}\delta\Big(\frac{w}{z}\Big),\end{aligned}$$

which is the required relation. To prove relation (R3′) we have

$$\begin{aligned}&[X_0(\alpha_i, z), X_0(\alpha_j, w)] \\ &\quad= \tfrac{1}{4}\Big[:e^{\frac{1}{k_0}\overline{(a_i-\epsilon_i b)(z)}}(b(z) + \epsilon_i r(z)):, :e^{-\frac{1}{k_0}\overline{(a_j-\epsilon_j b)(w)}}(b(w) - \epsilon_j r(w)):\Big] \\ &\quad= \tfrac{1}{4}\Big(:e^{\frac{1}{k_0}\overline{(a_i-a_j-\epsilon_i b+\epsilon_j b)(z)}}A(z):z^{-1}\delta\big(\tfrac{w}{z}\big) + B:e^{\frac{1}{k_0}\overline{(a_i-a_j-\epsilon_i b+\epsilon_j b)(z)}}:z^{-1}\partial_w\delta\big(\tfrac{w}{z}\big)\Big),\end{aligned}$$

where, by applying Corollary 3.2,

$$\begin{aligned}B &= (b + \epsilon_i r, b - \epsilon_j r) - \Big(\frac{a_i - \epsilon_i b}{k_0}, b - \epsilon_j r\Big)\Big(b + \epsilon_i r, -\frac{a_j - \epsilon_j b}{k_0}\Big) \\ &= -2k_0 - 2\epsilon_i\epsilon_j k_0, \\ A &= \Big(b + \epsilon_i r, -\frac{a_j - \epsilon_j b}{k_0}\Big)(b - \epsilon_j r) \\ &\qquad - \Big(\frac{a_i - \epsilon_i b}{k_0}, b - \epsilon_j r\Big)(b + \epsilon_i r) - \Big(-2k_0 - 2\epsilon_i\epsilon_j k_0\Big)\Big(-\frac{a_j - \epsilon_j b}{k_0}\Big) \\ &= -2(1 + \epsilon_i\epsilon_j)a_j.\end{aligned}$$

Therefore, we get

$$\begin{aligned}[X_0(\alpha_i, z), X_0(-\alpha_j, w)] &= -\tfrac{1}{2}(1 + \epsilon_i\epsilon_j)\Big(:e^{\frac{1}{k_0}\overline{(a_i-a_j-\epsilon_i b+\epsilon_j b)(z)}}a_j(z):z^{-1}\delta\big(\tfrac{w}{z}\big) \\ &\qquad + k_0:e^{\frac{1}{k_0}\overline{(a_i-a_j-\epsilon_i b+\epsilon_j b)(z)}}:z^{-1}\partial_w\delta\big(\tfrac{w}{z}\big)\Big) \\ &= -\delta_{ij}\left(a_j(z)z^{-1}\delta\big(\tfrac{w}{z}\big) + k_0 z^{-1}\partial_w\delta\big(\tfrac{w}{z}\big)\right),\end{aligned}$$

as required.

(R4′) contains two types of relations. We give only the proof for the "positive" case. The "negative" case can be proved similarly.

$$
\begin{aligned}
&[X_0(\alpha_i,z),\,X_0(\alpha_j,w)]\\
&=\tfrac{1}{4}\big[:e^{\frac{1}{k_0}\overline{(a_i-\epsilon_i b)(z)}}(b(z)+\epsilon_i r(z)):,\ :e^{\frac{1}{k_0}\overline{(a_j-\epsilon_j b)(w)}}(b(w)+\epsilon_j r(w)):\big]\\
&=\frac{1}{4}\Big(:e^{\frac{1}{k_0}\overline{(a_i+a_j-\epsilon_i b-\epsilon_j b)(z)}}A(z):z^{-1}\delta\big(\tfrac{w}{z}\big)+B:e^{\frac{1}{k_0}\overline{(a_i+a_j-\epsilon_i b-\epsilon_j b)(z)}}:z^{-1}\partial_w\delta\big(\tfrac{w}{z}\big)\Big),
\end{aligned}
$$

where, by applying Corollary 3.2,

$$
\begin{aligned}
B&=(b+\epsilon_i r,b+\epsilon_j r)-\Big(\frac{a_i-\epsilon_i b}{k_0},b+\epsilon_j r\Big)\Big(b+\epsilon_i r,\,\frac{a_j-\epsilon_j b}{k_0}\Big)=2k_0(\epsilon_i\epsilon_j-1),\\
A&=\Big(b+\epsilon_i r,\,\frac{a_j-\epsilon_j b}{k_0}\Big)(b+\epsilon_j r)\\
&\qquad-\Big(\frac{a_i-\epsilon_i b}{k_0},b+\epsilon_j r\Big)\Big(b+\epsilon_i r\Big)-2k_0(\epsilon_i\epsilon_j-1)\Big(\frac{a_j-\epsilon_j b}{k_0}\Big)\\
&=2(1-\epsilon_i\epsilon_j)a_j.
\end{aligned}
$$

Therefore $[X_0(\alpha_i,z),\,X_0(\alpha_i,w)]=0$ and, for $i\neq j$,

$$
\begin{aligned}
&[X_0(\alpha_i,z),\,X_0(\alpha_j,w)]\\
&\qquad=\ :e^{\frac{1}{k_0}\overline{(a_i+a_j)(z)}}a_j(z):z^{-1}\delta\Big(\frac{w}{z}\Big)-k_0:e^{\frac{1}{k_0}\overline{(a_i+a_j)(z)}}:z^{-1}\partial_w\delta\Big(\frac{w}{z}\Big).
\end{aligned}
$$

Clearly, for $i\neq j$, the vertex operator $X_0(\alpha_i,z)$ commutes with

$$:e^{\frac{1}{k_0}\overline{(a_i+a_j)}(z)}:.$$

Therefore to complete the proof of relation (R4′) we only need to show the identity

$$\big[X_0(\alpha_i,z_1),[X_0(\alpha_i,z_2),\,:e^{\frac{1}{k_0}\overline{(a_i+a_j)}(z_3)}a_j(z_3):]\big]=0 \tag{1}$$

for $i\neq j$. Indeed,

$$
\begin{aligned}
&\big[X_0(\alpha_i,z),\,:e^{\frac{1}{k_0}\overline{(a_i+a_j)}(w)}a_j(w):\big]\\
&\qquad=\tfrac{1}{2}\big[:e^{\frac{1}{k_0}\overline{(a_i-\epsilon_i b)}(z)}(b+\epsilon_i r)(z):,\ :e^{\frac{1}{k_0}\overline{(a_i+a_j)}(w)}a_j(w):\big]\\
&\qquad=\tfrac{1}{2}\Big(:e^{\frac{1}{k_0}\overline{(2a_i+a_j-\epsilon_i b)(z)}}A(z):z^{-1}\delta\big(\tfrac{w}{z}\big)+B:e^{\frac{1}{k_0}\overline{(2a_i+a_j-\epsilon_i b)(z)}}:z^{-1}\partial_w\delta\big(\tfrac{w}{z}\big)\Big),
\end{aligned}
$$

where, by applying Corollary 3.2,

$$B=(b+\epsilon_i r,a_j)-\Big(\frac{a_i-\epsilon_i b}{k_0},a_j\Big)\Big(b+\epsilon_i r,\,\frac{a_i+a_j}{k_0}\Big)=0$$

and

$$A=\Big(b+\epsilon_i r,\,\frac{a_i+a_j}{k_0}\Big)a_j-\Big(\frac{a_i-\epsilon_i b}{k_0},a_j\Big)(b+\epsilon_i r)=2(b+\epsilon_i r);$$

that is

$$\big[X_0(\alpha_i, z),\, :e^{\frac{1}{k_0}\overline{(a_i+a_j)}(w)} a_j(w):\big] = :e^{\frac{1}{k_0}\overline{(2a_i+a_j-\epsilon_i b)}(z)}(b+\epsilon_i r):z^{-1}\delta\Big(\frac{w}{z}\Big).$$

Therefore (1) is reduced to the identity

$$\big[X_0(\alpha_i, z),\, :e^{\frac{1}{k_0}\overline{(2a_i+a_j-\epsilon_i b)}(w)}(b+\epsilon_i r)(w):\big] = 0$$

for $i \neq j$. The left side is equal to

$$\begin{aligned}&\tfrac{1}{2}\big[:e^{\frac{1}{k_0}\overline{(a_i-\epsilon_i b)}(z)}(b+\epsilon_i r)(z):,\, :e^{\frac{1}{k_0}\overline{(2a_i+a_j-\epsilon_i b)}(w)}(b+\epsilon_i r)(w):\big]\\ &\quad= \tfrac{1}{2}\left(:e^{\frac{1}{k_0}\overline{(3a_i+a_j-2\epsilon_i b)(z)}}A(z):z^{-1}\delta\big(\tfrac{w}{z}\big) + B:e^{\frac{1}{k_0}\overline{(3a_i+a_j-2\epsilon_i b)(z)}}:z^{-1}\partial_w\delta\big(\tfrac{w}{z}\big)\right),\end{aligned}$$

where, by applying Corollary 3.2,

$$B = (b+\epsilon_i r, b+\epsilon_i r) - \Big(\frac{a_i-\epsilon_i b}{k_0}, b+\epsilon_i r\Big)\Big(b+\epsilon_i r, \frac{2a_i+a_j-\epsilon_i b}{k_0}\Big) = 0$$

and

$$A = \Big(b+\epsilon_i r, \frac{2a_i+a_j-\epsilon_i b}{k_0}\Big)(b+\epsilon_i r) - \Big(\frac{a_i-\epsilon_i b}{k_0}, b+\epsilon_i r\Big)(b+\epsilon_i r) = 0,$$

giving the desired identity. □

From the construction of the representation for the toroidal Lie algebra given in the previous theorem, it is easy to see that the operators $\alpha_1(k)+\alpha_0(k)$ act on the Fock space V trivially for all positive integers k, which in turn implies that the central elements $\psi(\delta(k))$ act as the zero operator for $k > 0$. Therefore the representation is not faithful. Indeed, the quotient space $V(0)$ of the Fock space

$$\mathbb{C}[\Gamma_0^*] \otimes S(\mathcal{H}^-)$$

defines a representation for the double affine Lie algebra $T_0(A_1)$, which is isomorphic to the Lie algebra $T(A_1)$ modulo all central elements of degree other then zero (see Section 2).

Corollary 3.4. *The vector space $V(0)$ is endowed with a representation of the double affine Lie algebra $T_0(A_1)$ with level-$(k_0, 0)$, under the formula given before Theorem 3.3.*

We will study this module structure again in the next section.

To give a faithful representation of the toroidal Lie algebra, we consider the rank-six lattice

$$\Gamma := \frac{1}{k_0}(\mathbb{Z}a_0 \oplus \mathbb{Z}a_1 \oplus \mathbb{Z}b \oplus \mathbb{Z}c \oplus \mathbb{Z}d) \oplus \frac{1}{k_0+2}\mathbb{Z}r,$$

with the symmetric bilinear form determined by

$$(b,b)=-2k_0,\quad (r,r)=2(k_0+2),\quad (c,d)=k_0,\quad (a_i,a_j)=k_0a_{ij}\quad \text{for } i,\, j=0,1,$$

all others being zero. Then

$$\Gamma_0 := \frac{1}{k_0}\mathbb{Z}(a_0-b)+\frac{1}{k_0}\mathbb{Z}(a_1+b)+\frac{1}{k_0}\mathbb{Z}c$$

is clearly an integral sublattice of Γ. Let $\Gamma_0^\star$ be the corresponding additive subgroup of $H=\mathbb{C}\otimes_{\mathbb{Z}}\Gamma$, and V the corresponding Fock space.

We also modify the vertex operators from the previous theorem to the form

$$X(\pm\alpha_1,z)=\tfrac{1}{2}:e^{\pm\frac{1}{k_0}\overline{(a_1+b)(z)}}(b(z)\mp r(z)):,$$
$$X(\pm\alpha_0,z)=\tfrac{1}{2}:e^{\pm\frac{1}{k_0}\overline{(a_0-b+c)(z)}}(b(z)\pm r(z)):.$$

Theorem 3.5. *The coefficient operators of the vertex operators $a_i(z),\ X(\pm\alpha_i,z)$, for $i=0,1$, acting on the Fock space V, generate a Lie algebra $\mathscr{L}(A_1)$ isomorphic to the toroidal Lie algebra $T(A_1)$, the isomorphism begin given by the linear map ϕ defined by*

$$\begin{aligned} \not{c} &\mapsto k_0,\\ \alpha_1(z) &\mapsto a_1(z),\\ \alpha_0(z) &\mapsto a_0(z)+c(z),\\ x(\pm\alpha_i,z) &\mapsto X(\pm\alpha_i,z)\quad \text{for } i=0,1. \end{aligned}$$

Therefore, on the Fock space V, we have a faithful representation of the toroidal Lie algebra $T(A_1)$.

Proof. We first need to show that the surjective mapping ϕ defines a Lie algebra homomorphism from $T(A_1)$ to $\mathscr{L}(A_1)$. It suffices to show that the vertex operators $a_i(z)$, $X(\pm\alpha_i,z)$ satisfy the corresponding power series identities (R0′)–(R4′). The argument is just as in the proof of Theorem 3.3, and we omit it for brevity's sake.

We next use Proposition 2.2 to show that the mapping ϕ is indeed an injective homomorphism. For $\alpha=\mu_1a_0+\mu_2a_1+\mu_3b+\mu_4c\in\Gamma_0^\star$ with $\mu_i\in\frac{1}{k_0}\mathbb{Z}$, let

$$e^\alpha\otimes\lambda_1(-n_1)\cdots\lambda_k(-n_k)\in V.$$

We define a $\mathbb{Z}\times Q$-gradation on the Fock space V by setting

$$\deg\bigl(e^\alpha\otimes\lambda_1(-n_1)\cdots\lambda_k(-n_k)\bigr)=(n_1+\cdots+n_k,\,k_0\mu_1\alpha_0+k_0\mu_2\alpha_1).$$

With this gradation, the operator $a(n)$, for $a\in H$, is a homogeneous operator of degree $(-n,0)$. Moreover, if the vertex operator $X(\pm\alpha_i,z)$ is formally expanded into power series as

$$X(\pm\alpha_i, z) = \sum_{m\in\mathbb{Z}} X_m(\pm\alpha_i) z^{-m-1},$$

the coefficient operator $X_m(\pm\alpha_i)$ is a homogeneous operator of degree $(-m, \pm\alpha_i)$. Thus the map ϕ is a $(\mathbb{Z}\times Q)$-graded Lie algebra homomorphism. To finish the proof of this theorem, we need only show that ϕ satisfies the three conditions of Proposition 2.2.

Recall the notation $x_m(\pm\alpha_1 + k\delta) = \psi^{-1}(\pm s^m t^k \otimes x_\pm)$, where $\delta = \alpha_0 + \alpha_1$ is the null root in Q. Let

$$x(\alpha, z) = \sum_{m\in\mathbb{Z}} x_m(\alpha) z^{-m-1} \qquad \text{for } \alpha = \pm\alpha_1 + k\delta.$$

Then it is easy to show that $\phi : x(\alpha, z) \mapsto X(\alpha, z)$, where $\alpha = \pm\alpha_1 + k\delta$, and

$$X(\pm\alpha_1 + k\delta, z) = \tfrac{1}{2} {:} e^{\pm\frac{1}{k_0}\overline{(a_1+b+k(a_0+a_1)+kc)(z)}} (b(z) \mp r(z)) {:} \, .$$

Applying Corollary 3.2 again we have

$$\begin{aligned}\big[X(\alpha_1 + k\delta, z),\, X(-\alpha_1 - k\delta, w)\big]& \\ = -k_0 z^{-1}\partial_w \delta\Big(\frac{w}{z}\Big) &- (a_1 + k(a_0 + a_1) + kc)(z) z^{-1}\delta\Big(\frac{w}{z}\Big).\end{aligned}$$

This gives

$$\big[X_m(\alpha_1 + k\delta),\, X_{-m}(-\alpha_1 - k\delta)\big] = -a_1(0) - k(a_0 + a_1)(0) - kc(0) - mk_0,$$

which is clearly a nonzero operator for any $m, k \in \mathbb{Z}$. Thus ϕ is injective on the one-dimensional subspace $T_m^\alpha = \mathbb{C}x_m(\alpha)$ for any real root $\alpha = \pm\alpha_1 + k\delta$ and $k, m \in \mathbb{Z}$. Moreover,

$$\phi(\delta(k)) = a_0(k) + a_1(k) + c(k)$$

is also a nonzero operator, and ϕ is clearly injective on $\mathbb{C}\delta(0) + \mathbb{C}\rlap{/}c$.

Finally, we need to show that, for $m, k \neq 0$,

$$\begin{aligned}\big[X_m(\alpha_1 + k\delta),\, X_0(-\alpha_1)\big] - \big[X_0(\alpha_1 + k\delta),\, X_m(-\alpha_1)\big] \neq 0,\\ \big[X_1(\alpha_1 + k\delta),\, X_{-1}(-\alpha_1)\big] - \big[X_{-1}(\alpha_1 + k\delta),\, X_1(-\alpha_1)\big] \neq 0.\end{aligned} \tag{2}$$

By Corollary 3.2,

$$\begin{aligned}\big[X(\alpha_1 + k\delta, z),\, X(-\alpha_1, w)\big] + \big[X(-\alpha_1, z),\, X(\alpha_1 + k\delta, w)\big]& \\ = -2k_0 {:} e^{\frac{1}{k_0}\overline{(ka_0+ka_1+kc)(z)}} {:} z^{-1}\partial_w\delta\Big(\frac{w}{z}\Big)& \\ + k {:} e^{\frac{1}{k_0}\overline{(ka_0+ka_1+kc)(z)}} (a_0(z) + a_1(z) + c(z)) {:} z^{-1}\delta\Big(\frac{w}{z}\Big)&,\end{aligned}$$

which gives

$$\big[X(\alpha_1+k\delta,z),\,X_0(-\alpha_1)\big]-\big[X_0(\alpha_1+k\delta),\,X(-\alpha_1,z)\big]$$
$$=k:e^{\frac{1}{k}{}_0\overline{(ka_0+ka_1+kc)(z)}}(a_0(z)+a_1(z)+c(z)):.$$

To see that the coefficient of z^{-m-1} in the expression on the right is nonzero for $m\neq 0$, we notice that

$$\big[:e^{\frac{1}{k_0}\overline{(ka_0+ka_1+kc)(z)}}(a_0(z)+a_1(z)+c(z)):,\,:e^{-\frac{1}{k_0}\overline{(ka_0+ka_1+kc)(z)}}d(z):\big]$$
$$=k_0z^{-1}\partial_w\delta\Big(\frac{w}{z}\Big).$$

The coefficient of z^{-m-1} on the right-hand side of the previous identity is k_0mw^{m-1}, which is nonzero whenever $m\neq 0$. This proves the first line in (2), while the second can be proved by a similar argument which is omitted here. Therefore ϕ is an isomorphism of Lie algebras. □

Corollary 3.6. *For any fixed $k_1\in\mathbb{Z}$, define*

$$V(k_1)=e^{k_1d+\Gamma_0}\otimes S(\mathcal{H}^-).$$

Then the vector space $V(k_1)$ is endowed with a representation of the toroidal Lie algebra $T(A_1)$ with level-(k_0,k_1).

4. Module structure

We now define a smaller module from our Fock space representation via the so-called screening operator. We will only consider the case when $c=0$.

For given $j_0,\,j_1,\,l_1,\,l_2\in\mathbb{C}$ with $j_0+j_1\in\mathbb{Z}\frac{k_0}{2}$, set

$$v_{j_0,j_1,l_1,l_2}:=e^{j_0\frac{a_0}{k_0}}e^{j_1\frac{a_1}{k_0}}e^{l_1\frac{b}{k_0}}e^{-l_2\frac{r}{k_0+2}}.$$

We define the Fock space F_{j_0,j_1,l_1,l_2} to be the space $S(\mathcal{H}^-)v_{j_0,j_1,l_1,l_2}$. Then the vertex operators $X(\pm\alpha_i,z)$ are well defined on F_{j_0,j_1,l_1,l_2}, provided that $2(j_1-l_1)$ and $2(j_0+l_1)$ are integers. It is clear that the vertex operators satisfy

$$X(\pm\alpha_0,z):F_{j_0,j_1,l_1,l_2}\longrightarrow F_{j_0\pm1,j_1,l_1\mp1,l_2},$$
$$X(\pm\alpha_1,z):F_{j_0,j_1,l_1,l_2}\longrightarrow F_{j_0,j_1\pm1,l_1\pm1,l_2}.$$

Introduce a screening operator $S_0\colon F_{j_0,j_1,l_1,l_2}\to F_{j_0,j_1,l_1+\frac{k_0}{2},l_2+\frac{k_0+2}{2}}$ by setting

$$S(z)=:e^{\frac{1}{2}\overline{(b(z)-r(z))}}:=\sum_n S_nz^{-n-1}.$$

This is well defined provided that $l_1-l_2\in\mathbb{Z}$.

Proposition 4.1.

$$\{X(\alpha_1,z),S(w)\}=\frac{\partial}{\partial w}\left(:e^{\overline{\frac{1}{k_0},(a_1+b)(w)+\frac{1}{2}(b-r)(w)}}:\frac{1}{z-w}\right),$$

$$\{X(-\alpha_1,z),S(w)\}=0,$$

$$\{X(-\alpha_0,z),S(w)\}=0,$$

$$\{X(\alpha_0,z),S(w)\}=\frac{\partial}{\partial w}\left(:e^{\overline{\frac{1}{k_0}(a_0+b)(w)+\frac{1}{2}(b-r)(w)}}:\frac{1}{z-w}\right).$$

Proof. Let

$$\phi(\alpha_1,z)=\phi(-\alpha_0,z):=\tfrac{1}{2}:e^{\overline{\frac{1}{k_0}b(z)}}(b(z)-r(z)):,$$

$$\phi(\alpha_0,z)=\phi(-\alpha_1,z):=\tfrac{1}{2}:e^{\overline{-\frac{1}{k_0}b(z)}}(b(z)+r(z)):$$

be the parafermions. It follows from Lemma 3.1 that

$$\begin{aligned}\phi(\alpha_1,z)S(w)&\\ &\sim\tfrac{1}{2}:e^{\overline{\frac{1}{k_0}b(z)}}(b-r)(z)e^{\overline{\frac{1}{2}(b-r)(w)}}:\frac{1}{z-w}+\tfrac{1}{2}:e^{\overline{\frac{1}{k_0}b(z)+\frac{1}{2}(b-r)(w)}}:\frac{2}{(z-w)^2}\\ &\sim\frac{\partial}{\partial w}\Big(:e^{\overline{\frac{1}{k_0}b(w)+\frac{1}{2}(b-r)(w)}}:\frac{1}{z-w}\Big).\end{aligned}$$

□

Let d be the zero mode of $S(z)$: $d=\int S(z)dz$. It is easy to check that the anticommutator $\{S(z),S(z)\}=0$, thus d gives rise to a complex of vector spaces:

$$\cdots\longrightarrow F_{j_0,j_1,l_1-\frac{k_0}{2},l_2-\frac{k_0+2}{2}}\longrightarrow F_{j_0,j_1,l_1,l_2}\longrightarrow$$
$$F_{j_0,j_1,l_1+\frac{k_0}{2},l_2+\frac{k_0+2}{2}}\longrightarrow F_{j_0,j_1,l_1+k_0,l_2+k_0+2}\longrightarrow\cdots$$

We can define the restricted $T(A)$-submodule using Proposition 4.1. Given l we define a $T(A)$-submodule

$$F_l=\bigoplus_{j_1\in l+\mathbb{Z},\,j_0\in -l+\mathbb{Z}}\ker\Big(d:F_{j_0,j_1,j_1,l}\to F_{j_0,j_1,j_1+\frac{k_0}{2},l+\frac{k_0+2}{2}}\Big).$$

Theorem 4.2. *The operator d commutes or anticommutes with elements of the toroidal algebra $T(A_1)$ and $d^2=0$. Moreover we have the long exact sequence*

$$0\longrightarrow F_l\longrightarrow\bigoplus_{j_0,j_1}F_{j_0,j_1,j_1,l}\longrightarrow\bigoplus_{j_0,j_1}F_{j_0,j_1,j_1+\frac{k_0}{2},l+\frac{k_0+2}{2}}$$
$$\longrightarrow\bigoplus_{j_0,j_1}F_{j_0,j_1,j_1+k_0,l+k_0+2}\longrightarrow\cdots,$$

where the maps from $\bigoplus_{j_0,j_1}F_{j_0,j_1,j_1,l}$ onward are $\bigoplus d$ and the summations run through $j_0\in -l+\mathbb{Z}$ and $j_1\in l+\mathbb{Z}$.

Proof. We introduce the operator $S^\star(z) = e^{\overline{-\frac{1}{2}(b(z)-r(z))}} = \sum_n S_n^* z^{-n}$, and set $d^\star = S_0^\star$. It is easy to see that $\{S(z), S^\star(w)\} = 1$. Hence $dd^\star + d^\star d = 1$, and we already knew that $d^2 = 0$. Thus the following long sequence of vector spaces is exact:

$$0 \to \ker_{F_{j_0,j_1,j_1,l}} d \to F_{j_0,j_1,j_1,l} \to F_{j_0,j_1,j_1+\frac{k_0}{2},l+\frac{k_0+2}{2}} \to F_{j_0,j_1,j_1+k_0,l+k_0+2} \to \cdots.$$

Taking the direct sum we obtain Theorem 4.2. □

Since $a_0(n) + a_1(n)$ acts trivially we can modulo the relation and define

$$\tilde{F}_l = F_l/(a_0(n) + a_1(n); -n \in \mathbb{N});$$

then it is also a $T(A_1)$-module and the results in Proposition 4.1 obviously hold for the module $\tilde{F}_l$. If we further modulo $a_1(0) + a_0(0)$ we will obtain the Verma module for the affine Lie algebra generically.

Using the exact sequence we can compute the character for the module $\tilde{F}_l$ as follows.

Theorem 4.3. *The character of the $T(A_1)$-module $\tilde{F}_l$ is given by*

$$\mathrm{ch}(\tilde{F}_l) = \sum_{s=0}^{\infty}(-1)^s \frac{\sum_{\alpha\in\overline{Q}} e^{-l\frac{r}{k_0+2}+s(\frac{k_0+2}{2}r+\frac{k_0}{2}b)}e^{\alpha}}{\prod(e^{-\delta_1})\prod(e^{-\delta_b})\prod(e^{-\delta_r})}$$

where

$$\prod(x) = \prod_{m>0}(1-x^m) \quad \textit{and} \quad \overline{Q} = \frac{1}{k_0}(\mathbb{Z}\alpha_1 + \mathbb{Z}b).$$ □

References

[Berman and Billig 1999] S. Berman and Y. Billig, "Irreducible representations for toroidal Lie algebras", *J. Algebra* **221**:1 (1999), 188–231. MR 2000k:17004 Zbl 0942.17016

[Eswara Rao and Moody 1994] S. Eswara Rao and R. V. Moody, "Vertex representations for n-toroidal Lie algebras and a generalization of the Virasoro algebra", *Comm. Math. Phys.* **159**:2 (1994), 239–264. MR 94m:17028 Zbl 0808.17018

[Fabbri and Moody 1994] M. A. Fabbri and R. V. Moody, "Irreducible representations of Virasoro-toroidal Lie algebras", *Comm. Math. Phys.* **159**:1 (1994), 1–13. MR 95d:17031 Zbl 0796.17024

[Feĭgin and Frenkel' 1988] B. L. Feĭgin and È. V. Frenkel', "A family of representations of affine Lie algebras", *Uspekhi Mat. Nauk* **43**:5 (1988), 227–228. In Russian; translated in *Russian Math. Surveys* **43**:5 (1988), 221–222. MR 89k:17016 Zbl 0657.17013

[Frenkel et al. 1988] I. Frenkel, J. Lepowsky, and A. Meurman, *Vertex operator algebras and the Monster*, Pure and Applied Mathematics **134**, Academic Press, Boston, 1988. MR MR996026 (90h:17026) Zbl 0674.17001

[Jing and Lyerly 1999] N. Jing and C. M. Lyerly, "Level two vertex representations of $G_2^{(1)}$", *Comm. Algebra* **27**:9 (1999), 4355–4362. MR 2000g:17035 Zbl 1007.17019

[Larsson 1999] T. A. Larsson, "Lowest-energy representations of non-centrally extended diffeomorphism algebras", *Comm. Math. Phys.* **201**:2 (1999), 461–470. MR 2000c:17042 Zbl 0936.17025

[Moody et al. 1990] R. V. Moody, S. E. Rao, and T. Yokonuma, "Toroidal Lie algebras and vertex representations", *Geom. Dedicata* **35**:1-3 (1990), 283–307. MR 91i:17032 Zbl 0704.17011

[Nemeschansky 1989] D. Nemeschansky, "Feĭgin–Fuchs representation of $\widehat{\mathrm{su}}(2)_k$ Kac–Moody algebra", *Phys. Lett. B* **224**:1-2 (1989), 121–124. MR 90g:17026

[Tan 1999] S. Tan, "Principal construction of the toroidal Lie algebra of type A_1", *Math. Z.* **230**:4 (1999), 621–657. MR 2001d:17030 Zbl 0932.17028

[Wakimoto 1986] M. Wakimoto, "Fock representations of the affine Lie algebra $A_1^{(1)}$", *Comm. Math. Phys.* **104**:4 (1986), 605–609. MR 87m:17011 Zbl 0587.17009

Received October 2, 2002. Revised March 18, 2004.

NAIHUAN JING
DEPARTMENT OF MATHEMATICS
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NC 27695
UNITED STATES

FACULTY OF MATHEMATICS
HUBEI UNIVERSITY
WUHAN, HUBEI 430064
CHINA

jing@math.ncsu.edu

KAILASH MISRA
DEPARTMENT OF MATHEMATICS
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NC 27695
UNITED STATES

misra@math.ncsu.edu

SHAOBIN TAN
DEPARTMENT OF MATHEMATICS
XIAMEN UNIVERSITY
XIAMEN, FUJIAN 361005
CHINA

tans@jingxian.xmu.edu.cn

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

THE KERNEL OF Burau(4) $\otimes \mathbb{Z}_p$ IS ALL PSEUDO-ANOSOV

SANG JIN LEE AND WON TAEK SONG

The kernel of Burau(4) $\otimes \mathbb{Z}_p$, the reduced Burau representation with coefficients in $\mathbb{Z}_p$ of the 4-braid group B_4, consists only of pseudo-Anosov braids.

1. Introduction

Given two pseudo-Anosov homeomorphisms with distinct invariant measured foliations, some powers of their isotopy classes generate a rank two free subgroup of the mapping class group of the surface [Long 1986]. This construction gives an example of all pseudo-Anosov subgroup of the mapping class group. A positive answer is given in [Whittlesey 2000] to the natural question of the existence of all pseudo-Anosov *normal* subgroups by showing that the Brunnian mapping classes on a sphere with at least five punctures are neither periodic nor reducible. Not every Brunnian n-braid maps to a Brunnian mapping class on an $(n+1)$-punctured sphere. One can however show that a nontrivial Brunnian n-braid should be pseudo-Anosov for $n \geq 3$, by adapting the arguments in [Whittlesey 2000].

In this note we show that the kernel of Burau(4) $\otimes \mathbb{Z}_p$, the reduced Burau representation with coefficients in $\mathbb{Z}_p$ of the 4-braid group B_4, consists only of pseudo-Anosov braids. Our result also implies that the kernel of Burau(4), if nontrivial, is all pseudo-Anosov. By [Cooper and Long 1997; 1998], Burau(4) $\otimes \mathbb{Z}_p$ for $p = 2, 3$ is not faithful. It is straightforward to check that there exist non-Brunnian braids in the kernels, hence giving new examples of all pseudo-Anosov normal subgroups of B_4 that are not contained in the example of Whittlesey.

For the proof, assume that we are given a nontrivial 4-braid that is not pseudo-Anosov. If it is periodic, it is conjugate to a rigid rotation [Brouwer 1919], whose Burau action is clearly nontrivial. If it is reducible, then in many ways it is similar to a 3-braid so that its Burau action is fairly predictable, for which case an automaton that records the polynomial degrees suffices to prove faithfulness. Our argument is similar to that of the ping-pong lemma. We construct an automaton whose states are disjoint subsets of $\mathbb{Z}_p[t, t^{-1}]^3$ and whose arrows are braid actions that map the subsets into the subsets.

MSC2000: 20F36, 57M60.
Keywords: braid group, Burau representation, all pseudo-Anosov.
Lee's research was supported by the faculty research fund of Konkuk University in 2003.

For braids with more than four strands, this approach immediately faces obstacles. Since Burau(4) $\otimes \mathbb{Z}_2$ is not faithful, the kernel of Burau(5) $\otimes \mathbb{Z}_2$ contains reducible braids. Taking other representations or taking intersection with other subgroups to get rid of such reducible braids then makes the proof more difficult.

We remark that the present result is a byproduct of working on the faithfulness question of Burau(4) [Moody 1991; 1993; Long and Paton 1993; Bigelow 1999].

2. No periodic or reducible braids

The n-braid group B_n consists of the mapping classes on the n-punctured disk. The center of B_n is the infinite cyclic group generated by the Dehn twist along the boundary. A braid is called *periodic* if some of its powers are contained in the center. A braid is called *reducible* if it is represented by a disk homeomorphism that fixes a collection of disjoint essential curves. If a braid is neither periodic nor reducible, the Nielsen–Thurston classification of surface homeomorphisms [Thurston 1988; Fathi et al. 1979] implies that it is represented by a pseudo-Anosov homeomorphism. Such a braid is called *pseudo-Anosov*. A subgroup of B_n is called *all pseudo-Anosov* if its nontrivial elements are all pseudo-Anosov.

The n-braid group B_n has the presentation

$$B_n = \left\langle \sigma_1, \ldots, \sigma_{n-1} \;\middle|\; \begin{array}{ll} \sigma_i\sigma_j = \sigma_j\sigma_i, & |i-j| \geq 2 \\ \sigma_i\sigma_j\sigma_i = \sigma_j\sigma_i\sigma_j, & |i-j| = 1 \end{array} \right\rangle$$

The reduced Burau representation

$$\rho_n = \mathrm{Burau}(n) : B_n \to \mathrm{GL}_{n-1}(\mathbb{Z}[t, t^{-1}])$$

is defined by the action on the first homology of the cyclic cover of the punctured disk. For the purpose of this note, it suffices to define ρ_4 by the three matrices

$$\rho_4(\sigma_1) = \begin{pmatrix} -t & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \rho_4(\sigma_2) = \begin{pmatrix} 1 & t & 0 \\ 0 & -t & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad \rho_4(\sigma_3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & t \\ 0 & 0 & -t \end{pmatrix}.$$

We use the convention that B_4 acts on $\mathbb{Z}[t, t^{-1}]^3$ from the right. We denote by $\boldsymbol{v} *_\rho \beta$, or more simply by $\boldsymbol{v} * \beta$, the matrix multiplication $\boldsymbol{v}\rho(\beta)$ for a row vector $\boldsymbol{v}$, a representation ρ and a braid β. For example, $(f, g, h) *_{\rho_4} \sigma_1 = (-tf + g, g, h)$ for $f, g, h \in \mathbb{Z}[t, t^{-1}]$.

Theorem 1. *The kernel of* $(\rho_4 \otimes \mathbb{Z}_p) : B_4 \to \mathrm{GL}_3(\mathbb{Z}_p[t, t^{-1}])$ *for* $p \geq 2$ *does not contain a nontrivial periodic or reducible braid. In particular if* $\rho_4 \otimes \mathbb{Z}_p$ *is not faithful, its kernel is an all pseudo-Anosov normal subgroup of* B_4.

The proof will involve several lemmas.

Lemma 2. $\rho_n \otimes \mathbb{Z}_p$ *is faithful for periodic braids.*

Proof. If $\beta \in B_n$ is a periodic n-braid, then it is represented by a rigid rotation on the punctured disk [Brouwer 1919] so that it is conjugate to $(\sigma_{n-1} \cdots \sigma_2\sigma_1)^k$ or to $(\sigma_{n-1} \cdots \sigma_2\sigma_1\sigma_1)^k$ for some $k \in \mathbb{Z}$. Since $\det\big((\rho_n \otimes \mathbb{Z}_p)(\beta)\big) = (-t)^{e(\beta)}$, where the exponent sum $e(\beta)$ is $k(n-1)$ or kn, we see that if β is in the kernel of $\rho_n \otimes \mathbb{Z}_p$, then $k = 0$ and β is trivial. □

Let $\Delta_3 = \sigma_1\sigma_2\sigma_1 \in B_3$ and $\Delta_4 = \sigma_1\sigma_2\sigma_1\sigma_3\sigma_2\sigma_1 \in B_4$ be the square roots of the generator of the center of B_3 and B_4, respectively. For a Laurent polynomial $f(t) = \sum_m a_m t^m$, define $\deg f = \max\{m : a_m \neq 0\}$. By convention we define $\deg f = -\infty$ if $f = 0$.

Lemma 3. *$\rho_3 \otimes \mathbb{Z}_p$ is faithful.*

Proof. Let $\rho = \rho_3 \otimes \mathbb{Z}_p$ be the reduced Burau representation of B_3 with coefficients in $\mathbb{Z}_p$. It is given by the matrices

$$\rho(\sigma_1) = \begin{pmatrix} -t & 0 \\ 1 & 1 \end{pmatrix}, \qquad \rho(\sigma_2) = \begin{pmatrix} 1 & t \\ 0 & -t \end{pmatrix}.$$

Suppose that $\rho(\beta)$ is trivial for some nontrivial 3-braid β. By Lemma 2, it is either reducible or pseudo-Anosov. If β is reducible, it is conjugate to $\Delta_3^{2m}\sigma_1^k$ for some integers k and m, which is an arbitrary 3-braid with an invariant curve standardly embedded in the disk enclosing the first two punctures as in Figure 1, right. Since $\rho(\beta)$ is trivial,

$$\rho(\Delta_3^{2m}\sigma_1^k) = t^{3m} \begin{pmatrix} (-t)^k & 0 \\ * & 1 \end{pmatrix}$$

must be the identity matrix. So $m = 0$ and $k = 0$ hence β is trivial, which contradicts the assumption.

If β is pseudo-Anosov, it is conjugate to $P(\sigma_1^{-1}, \sigma_2)\Delta_3^{2k}$ where P is a positive word on two letters [Murasugi 1974; Song et al. 2002]. By taking inverse or conjugation by Δ_3 if necessary, we can assume that $P(\sigma_1^{-1}, \sigma_2)$ starts with σ_2. In other words, β or β^{-1} is conjugate to $\alpha = \sigma_2 Q(\sigma_1^{-1}, \sigma_2)\Delta_3^{2k}$ for some positive word Q. The ρ-actions of σ_1^{-1}, σ_2 and Δ_3^2 on $\mathbb{Z}_p[t, t^{-1}]^2$ are given as follows: for $\boldsymbol{v} = (f, g) \in \mathbb{Z}_p[t, t^{-1}]^2$,

$$\boldsymbol{v} * \sigma_1^{-1} = (-t^{-1}(f-g), g), \quad \boldsymbol{v} * \sigma_2 = (f, t(f-g)) \quad \text{and} \quad \boldsymbol{v} * \Delta_3^2 = (t^3 f, t^3 g).$$

Consider the subset $V_0 = \big\{(f, g) \in \mathbb{Z}_p[t, t^{-1}]^2 \mid \deg f < \deg g\big\}$. It is easy to check that V_0 is invariant under the action of σ_1^{-1}, σ_2 and Δ_3^2. Let $\boldsymbol{v}_0 = (1, 0)$. Then $\boldsymbol{v}_0 * \sigma_2 = (1, t) \in V_0$, so that $\boldsymbol{v}_0 * \alpha = (1, t) * Q(\sigma_1^{-1}, \sigma_2)\Delta_3^{2k} \in V_0$. Since $\boldsymbol{v}_0 \notin V_0$, we have $\boldsymbol{v}_0 * \alpha \neq \boldsymbol{v}_0$, which contradicts the assumption that β is in the kernel of ρ. □

Figure 1. Up to homeomorphisms on a 4-punctured disk, there are only two essential curves.

Proof of Theorem 1. Let $\rho = \rho_4 \otimes \mathbb{Z}_p$ be the reduced Burau representation of B_4 with coefficients in $\mathbb{Z}_p$. Assume $\rho(\beta)$ is trivial for some nontrivial 4-braid $\beta \in B_4$. The braid β is either reducible or pseudo-Anosov by Lemma 2. We need to show that β is not reducible.

Suppose that β is reducible. By taking some power of β if necessary, we may assume that β is represented by a homeomorphism that fixes an essential simple closed curve C. By applying a conjugation by a braid that sends C to one of the curves in Figure 1, we assume that C is one of the two standardly embedded curves and the homeomorphism representing β fixes C.

Let C be the curve enclosing the first three punctures as Figure 1, left. Then β can be written as $\beta = \Delta_4^{2m} W(\sigma_1, \sigma_2)$ for an integer m and a word W on two letters. Observing that the ρ-action by a 3-braid leaves the third coordinate invariant, i.e., $(f, g, h) * W(\sigma_1, \sigma_2) = (f_1, g_1, h)$, we have $(0, 0, 1) * \beta = (f, g, t^{4m})$ for some $f, g \in \mathbb{Z}_p[t, t^{-1}]$. Since $\rho(\beta)$ is trivial, we obtain $m = 0$, which in turn implies that β is in $\langle \sigma_1, \sigma_2 \rangle = B_3 \subset B_4$. The faithfulness of $\rho_3 \otimes \mathbb{Z}_p$ by Lemma 3 leads to a contradiction.

Now assume that C contains the first two punctures as Figure 1, right. The 4-braids represented by homeomorphisms that fix C form a subgroup of B_4 generated by σ_1, $x = \sigma_2 \sigma_1^2 \sigma_2$ and $y = \sigma_3$. Since σ_1 commutes with both x and y, we write

$$\beta = \sigma_1^k W(x, y)$$

for an integer k and a word W on two letters.

By using the relations $xyxy = yxyx$, $(xyxy)\sigma_1^2 = \Delta_4^2$ and that $xyxy$ commutes with x, y and σ_1, we rewrite β into another form by which we will track $(0, 0, 1) * \beta$.

By replacing x^{-1} with $(yxy)(xyxy)^{-1}$ and y^{-1} with $(xyx)(xyxy)^{-1}$ and then collecting $(xyxy)^{\pm 1}$ to the left, we have $W(x, y) = (xyxy)^m P(x, y)$ for some integer m and a positive word P on two letters. We can assume that we have moved $(xyxy)$ to the left as many as possible so that neither $xyxy$ nor $yxyx$ occurs in P as a subword. We have

$$\beta = \sigma_1^k (xyxy)^m P(x, y) = \Delta_4^{2m} \sigma_1^{k-2m} P(x, y).$$

We claim that P contains both x and y as a subword. If P does not contain y, i.e., $P = x^l$ for some $l \geq 0$, then $\beta = \Delta_4^{2m} \sigma_1^{k-2m} x^l = \Delta_4^{2m} \sigma_1^{k-2m} (\sigma_2 \sigma_1^2 \sigma_2)^l$ fixes the curve in Figure 1, left. By the previous argument β is trivial. If P does not contain x, i.e., $P = y^l$ for some $l \geq 0$, then $\beta = \Delta_4^{2m}\sigma_1^{k-2m}y^l$. From the equalities

$$(0, 0, 1) * \beta = (0, 0, (-t)^{4m+l}), \qquad (1, 0, 0) * \beta = ((-t)^{4m+(k-2m)}, 0, 0),$$

we deduce $l = -4m$ and $k = -2m$. The exponent sum $e(\beta) = 12m + (k-2m) + l = 4m$ should equal zero since $\rho(\beta)$ is trivial. Therefore we have $m = l = k = 0$, which implies that β is trivial.

Next, since x and y both commute with σ_1 and Δ_4^2, by applying a conjugation we may assume that P starts with y and ends with x. In Figure 2, left, we construct an automaton that accepts a positive word in x, y without any occurrence of $xyxy$ and $yxyx$. Arbitrary paths following the arrows give words accepted by the automaton. Now we have

$$\beta = \Delta_4^{2m}\sigma_1^{k-2m}Q(x, y, xy, yx, yxy, xyx)$$

for some positive word Q accepted by the automaton in Figure 2, left. Note that Q starts with one of y, yxy, yx and ends with one of x, xyx, yx. In other words, Q is represented by a path starting at the state Y and ending at the state X.

We replace xyx by $y^{-1}(xyxy)$, yxy by $x^{-1}(xyxy)$ and then collect all $(xyxy)$'s to the left to obtain

$$\beta = \Delta_4^{2m_1}\sigma_1^{k_1}Q(x, y, xy, yx, x^{-1}, y^{-1})$$

for some k_1 and m_1.

Consider the subsets of $\mathbb{Z}_p[t, t^{-1}]^3$ given by

$$V_X = \{(f, g, h) \in \mathbb{Z}_p[t, t^{-1}]^3 \mid \deg g > \deg f,\ \deg g \geq \deg h\},$$
$$V_Y = \{(f, g, h) \in \mathbb{Z}_p[t, t^{-1}]^3 \mid \deg h > \deg f,\ \deg h > \deg g\}.$$

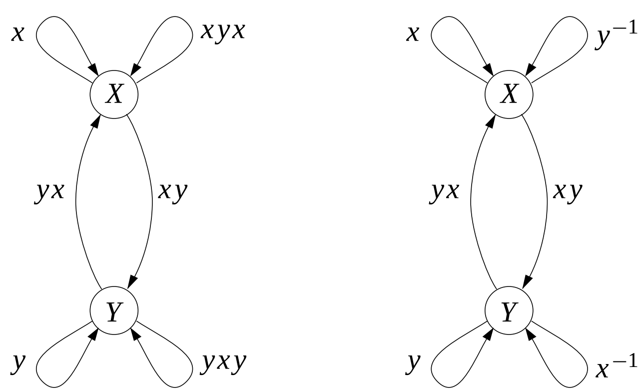

Figure 2. Left: an automaton that accepts exactly those words not containing $xyxy$ or $yxyx$. Right: see next page.

The ρ-action of each arrow of the automaton in Figure 2, right, is given as follows. Let $\boldsymbol{v} = (f, g, h) \in \mathbb{Z}_p[t, t^{-1}]^3$ be an arbitrary vector.

$$\begin{aligned}
\boldsymbol{v} * x &= \big(tf + (t^2 - t)g + (1-t)h,\ t^3 g + (1-t^2)h,\ h\big),\\
\boldsymbol{v} * y &= \big(f,\ g,\ tg - th\big),\\
\boldsymbol{v} * (xy) &= \big(tf + (t^2 - t)g + (1-t)h,\ t^3 g + (1-t^2)h,\ t^4 g - t^3 h\big),\\
\boldsymbol{v} * (yx) &= \big(tf + (t^2 - t)h,\ tg + (t^3 - t)h,\ tg - th\big),\\
\boldsymbol{v} * x^{-1} &= \big(t^{-1} f + (t^{-3} - t^{-2})g + (t^{-2} - t^{-3})h,\ t^{-3} g + (t^{-2} - t^{-3})h,\ h\big),\\
\boldsymbol{v} * y^{-1} &= \big(f,\ g,\ g - t^{-1}h\big).
\end{aligned}$$

Then it is routine to check from these formulae that

$$\begin{aligned}
&V_X * x \subset V_X, \quad V_X * y^{-1} \subset V_X, \quad V_X * (xy) \subset V_Y,\\
&V_Y * y \subset V_Y, \quad V_Y * x^{-1} \subset V_Y, \quad V_Y * (yx) \subset V_X.
\end{aligned}$$

These relations are compatible with the automaton in Figure 2, right. If a path starts at Y and ends at X then the ρ-action of its braid word maps V_Y into V_X. So we have $V_Y * Q \subset V_X$ for $Q = Q(x, y, xy, yx, x^{-1}, y^{-1})$.

Since $(0, 0, t^{4m_1}) \in V_Y$, we have

$$\begin{aligned}
(0, 0, 1) * \beta &= (0, 0, 1) * \Delta_4^{2m_1} \sigma_1^{k_1} Q\\
&= (0, 0, t^{4m_1}) * \sigma_1^{k_1} Q\\
&= (0, 0, t^{4m_1}) * Q,
\end{aligned}$$

which lies in V_X. Since $(0, 0, 1) \in V_Y$ and $V_X \cap V_Y = \varnothing$, the condition $(0, 0, 1) * \beta \in V_X$ implies that $\rho(\beta)$ is nontrivial. $\square$

We remark that the group generated by x and y is the Artin group of Coxeter type B_2 and that $xyxy = yxyx$ is the defining relation of the subgroup generated by x and y. So the subgroup generated by x, y and σ_1 is the direct product of the infinite cyclic subgroup generated by σ_1 and the subgroup generated by x and y.

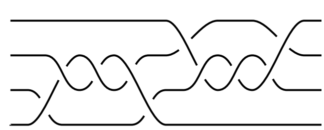

Figure 3. The braid $\sigma_1^{-1}\sigma_2^3\sigma_1\sigma_3\sigma_2^{-3}\sigma_3^{-1}$, whose fourth power is in the kernel of $\text{Burau}(4) \otimes \mathbb{Z}_2$.

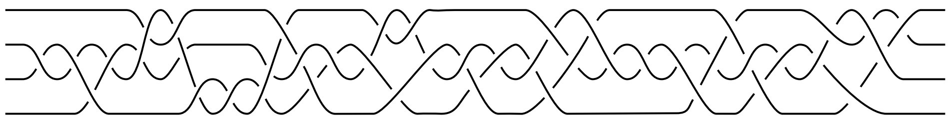

Figure 4. A braid in the kernel of Burau(4) $\otimes \mathbb{Z}_3$.

3. Non-Brunnian elements

Cooper and Long [1997] obtained a presentation of the image of $\rho_4 \otimes \mathbb{Z}_2$. As a corollary, $\rho_4 \otimes \mathbb{Z}_2$ is not faithful. The same authors computed in [Cooper and Long 1998] a presentation of a group containing the image of $\rho_4 \otimes \mathbb{Z}_3$ as a finite index subgroup and gave a nontrivial braid in the kernel explicitly. In this section we show that the examples of Cooper and Long are not Brunnian.

Let $\alpha_k = (\sigma_1^{-1}\sigma_2^{k}\sigma_1\sigma_3\sigma_2^{-k}\sigma_3^{-1})^4$ for $k \neq 0$. (See Figure 3 for the expression in parentheses, with $k = 3$.) The braid α_k comes from the fourth relation of [Cooper and Long 1997, Theorem 1.4] and is in the kernel of $\beta_4 \otimes \mathbb{Z}_2$. α_k is not Brunnian because we obtain σ_1^{4k} by forgetting the second and the fourth strands.

Now let α be the braid

$$\sigma_2^2\sigma_1\sigma_2^{-2}\sigma_3^{-2}\sigma_2\sigma_1^{-3}\sigma_2^{-1}\sigma_3\sigma_2^{-1}\sigma_1\sigma_2^2\sigma_3^{-2}\sigma_1^{-1}\sigma_2^{-2} \cdot \sigma_1\sigma_2^{-2}\sigma_1\sigma_3\sigma_2^{-1}\sigma_3\sigma_2^3\sigma_1\sigma_2^{-1}\sigma_3\sigma_2^{-1}\sigma_1\sigma_2^{-2}\sigma_1\sigma_3^2\sigma_2\sigma_3^{-1},$$

as in Figure 4. It is conjugate to the braid given by [Cooper and Long 1998] as a nontrivial element of ker Burau(4) $\otimes \mathbb{Z}_3$. It is easy to see that α is not Brunnian. If we forget the fourth strand from α as Figure 5, we get a nontrivial 3-braid

$$\begin{aligned}\alpha' &= \sigma_2^2\sigma_1\sigma_2^{-1}\sigma_1^{-3}\sigma_2^{-1}\sigma_1^2\sigma_2^{-2}\sigma_1^{-2}\sigma_2^3\sigma_1^{-1}\sigma_2\sigma_1^{-1}\sigma_2^2 \\ &= (\sigma_2\sigma_1^{-1}\sigma_2\sigma_1^{-1}\sigma_2^2)^3\Delta_3^{-2}.\end{aligned}$$

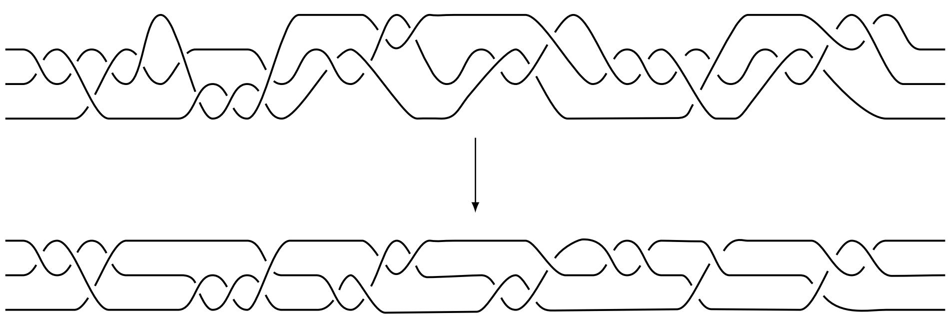

Figure 5. Forgetting the fourth strand.

References

[Bigelow 1999] S. Bigelow, "The Burau representation is not faithful for $n = 5$", *Geom. Topol.* **3** (1999), 397–404. MR 2001j:20055 Zbl 0942.20017

[Brouwer 1919] L. E. J. Brouwer, "Über die periodischen Transformationen der Kugel", *Math. Ann.* **80** (1919), 39–41. JFM 47.0527.01

[Cooper and Long 1997] D. Cooper and D. D. Long, "A presentation for the image of Burau(4) $\otimes$ Z_2", *Invent. Math.* **127**:3 (1997), 535–570. MR 97m:20050 Zbl 0913.57009

[Cooper and Long 1998] D. Cooper and D. D. Long, "On the Burau representation modulo a small prime", pp. 127–138 in *The Epstein birthday schrift*, Geom. Topol. Monogr. **1**, Geom. Topol. Publ., Coventry, 1998. MR 99k:20077 Zbl 0923.20030

[Fathi et al. 1979] A. Fathi, F. Laudenbach, and V. Poénaru, *Travaux de Thurston sur les surfaces*, Astérisque **66**, Société Math. de France, Paris, 1979. MR 82m:57003 Zbl 0731.57001

[Long 1986] D. D. Long, "A note on the normal subgroups of mapping class groups", *Math. Proc. Cambridge Philos. Soc.* **99**:1 (1986), 79–87. MR 87c:57009 Zbl 0584.57008

[Long and Paton 1993] D. D. Long and M. Paton, "The Burau representation is not faithful for $n \geq 6$", *Topology* **32**:2 (1993), 439–447. MR 94c:20071 Zbl 0810.57004

[Moody 1991] J. A. Moody, "The Burau representation of the braid group B_n is unfaithful for large n", *Bull. Amer. Math. Soc.* (*N.S.*) **25**:2 (1991), 379–384. MR 92b:20041 Zbl 0751.57005

[Moody 1993] J. A. Moody, "The faithfulness question for the Burau representation", *Proc. Amer. Math. Soc.* **119**:2 (1993), 671–679. MR 93k:57019 Zbl 0796.57004

[Murasugi 1974] K. Murasugi, *On closed* 3*-braids*, Memoirs Amer. Math. Soc. **151**, American Mathematical Society, Providence, 1974. MR 50 #8496 Zbl 0327.55001

[Song et al. 2002] W. T. Song, K. H. Ko, and J. E. Los, "Entropies of braids", *J. Knot Theory Ramifications* **11**:4 (2002), 647–666. MR 1 915 500 Zbl 1010.57004

[Thurston 1988] W. P. Thurston, "On the geometry and dynamics of diffeomorphisms of surfaces", *Bull. Amer. Math. Soc.* (*N.S.*) **19**:2 (1988), 417–431. MR 89k:57023 Zbl 0674.57008

[Whittlesey 2000] K. Whittlesey, "Normal all pseudo-Anosov subgroups of mapping class groups", *Geom. Topol.* **4** (2000), 293–307. MR 2001j:57022 Zbl 0962.57007

Received March 24, 2004. Revised June 6, 2004.

SANG JIN LEE
DEPARTMENT OF MATHEMATICS
KONKUK UNIVERSITY
GWANGJIN-GU, SEOUL 143-701
KOREA
sangjin@konkuk.ac.kr

WON TAEK SONG
SCHOOL OF MATHEMATICS
KOREA INSTITUTE FOR ADVANCED STUDY
207-43 CHEONGNYANGNI 2-DONG
DONGDAEMUN-GU, SEOUL 130-722
KOREA
cape@kias.re.kr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

ON A SPECIAL CLASS OF FIBRATIONS AND KÄHLER RIGIDITY

NICKOLAS J. MICHELACAKIS

Let $\mathcal{A}\mathcal{B}^n$ be the class of torsion-free, discrete groups that contain a normal, at most n-step, nilpotent subgroup of finite index. We give sufficient conditions for the fundamental group of a fibration $F \to T \to B$, with base B an infra-nilmanifold, to belong to $\mathcal{A}\mathcal{B}^n$. Manifolds of this kind may, for example, appear as thin ends of nonpositively curved manifolds. We prove that if, in addition, we require that T be Kähler, then T possesses a flat Riemannian metric and the fundamental group $\pi_1(T)$ is necessarily a Bieberbach group. Further, we prove that a torsion-free, virtually polycyclic group that can be realised as the fundamental group of a compact, Kähler $\mathrm{K}(\pi, 1)$-manifold is necessarily Bieberbach.

1. Introduction

Torsion-free, discrete, cocompact subgroups of the group of affine motions of $\mathbb{R}^n$ were first studied by Bieberbach in 1912, and more recently by Charlap; they are called *Bieberbach groups*. They correspond precisely to the fundamental groups of compact manifolds endowed with a flat Riemannian metric [Charlap 1965], and such manifolds are finitely covered by flat tori [Bieberbach 1911].

L. Auslander [1960] and Lee and Raymond [1985] turned their attention to *almost-Bieberbach groups*, that is, torsion-free, discrete, cocompact subgroups of $G \rtimes C$, with C a maximal, compact subgroup of $\operatorname{Aut} G$ for G a simply connected, nilpotent Lie group. They succeeded in generalising much of Bieberbach's work. Malcev's equivalence [1949] shows that torsion-free, finitely generated, nilpotent groups correspond precisely to the fundamental groups of *nilmanifolds*, that is, compact manifolds of the form $M = G/N$, where G is a simply connected, nilpotent Lie group, and N a discrete subgroup. Theorem 3.2 shows that almost-Bieberbach groups correspond to *infra-nilmanifolds*, compact manifolds of the form G/Γ with G as above and Γ a discrete subgroup of $G \rtimes C$, where

MSC2000: primary 22E40; secondary 32Q15, 14R20.
Keywords: affinely flat manifold, (almost)-crystallographic, (almost)-Bieberbach group, (almost)-torsion-free, (virtually) polycyclic group, nilpotent Lie group, discrete cocompact subgroups, lattice, Malcev completion, cohomology of groups, complex (Kähler) structure, group action, group representation, flat Riemannian manifold, (infra)-nilmanifold.

C is a maximal compact subgroup of $\operatorname{Aut} G$. We denote by $\mathcal{AB}^n$ the class of almost-Bieberbach groups whose maximal, normal, nilpotent subgroup is at most n-step nilpotent. We shall say that a group Γ admits an n-step almost-Bieberbach structure if and only if $\Gamma \in \mathcal{AB}^n$ and its maximal normal nilpotent subgroup is n-step nilpotent.

(We know from [Gromov 1981] and [Wolf 1968] that, among finitely generated groups, virtually nilpotent groups are precisely those groups that have polynomial growth. For details and precise definitions, see those works or [Tits 1981].)

We employ algebraic methods to study closed manifolds that fibre over infra-nilmanifolds. If $F \to T \to B$ is such a fibration, where F, T and B are all acyclic, the long homotopy exact sequence reduces to a group extension of the form

$$1 \longrightarrow \pi_1(F) \longrightarrow \pi_1(T) \longrightarrow \pi_1(B) \longrightarrow 1.$$

Manifolds of this type appear as thin ends of geometrically finite hyperbolic manifolds, which are an interesting subclass of nonpositively curved manifolds. More specifically, Apanasov and Xie [1997] proved that if $\Gamma \subset \mathcal{H}_n \rtimes U(n-1)$ is a torsion-free discrete group acting on the Heisenberg group $\mathcal{H}_n := \mathbb{C}^{n-1} \times \mathbb{R}$, the orbit space $\mathcal{H}_n / \Gamma$ is a Heisenberg manifold of zero Euler characteristic and a vector bundle over a compact manifold. Further, this compact manifold is finitely covered by a nilmanifold which is either a torus or a torus bundle over a torus. This generalises earlier results on almost flat manifolds concerning lattices in $\mathcal{H}_n \rtimes U(n-1)$ [Gromov 1978; Buser and Karcher 1981].

As mentioned above, groups in $\mathcal{AB}^n$ correspond to infra-nilmanifolds. In Section 2 we study extensions of the form

$$1 \longrightarrow G \longrightarrow \Gamma \longrightarrow K \longrightarrow 1,$$

with $K \in \mathcal{AB}^n$, to provide sufficient conditions under which Γ belongs to $\mathcal{AB}^n$. In particular, Proposition 2.2 guarantees the existence of an almost-Bieberbach structure on Γ provided G is a normal subgroup of Γ in a precise way. Proposition 2.4 does the same provided G lies in $\mathcal{AB}^n$ and the action of K on G respects some suitable minimal conditions.

In Section 3 we use the Johnson–Rees characterisation of fundamental groups of flat, Kähler [Johnson and Rees 1991], and projective [Johnson 1990] manifolds, and apply the Benson–Gordon theorem [1988] for the existence of a Kähler structure on a compact nilmanifold to show, in Theorem 3.3, that the existence of a Kähler structure on a special fibration as above implies the existence of a flat Riemann metric on T. In particular, in $\mathcal{AB}^n$, the classes of fundamental groups of Kähler and projective manifolds coincide, as shown in Corollary 3.4. Further, as a consequence of the Lefschetz hyperplane theorem and Bertini's theorem, this is a subclass of the class of fundamental groups of compact, closed, nonsingular projective surfaces.

Finally, in Section 4, we use a structure theorem concerning virtually polycyclic groups, proved in [Dekimpe and Igodt 1994], together with the results in [Arapura and Nori 1999], to prove, in Theorem 4.2, that a torsion-free, virtually polycyclic group can be realised as the fundamental group of a $\mathrm{K}(\pi, 1)$-compact, Kähler manifold if and only if it is Bieberbach of a special kind, namely, its operator homomorphism is essentially complex.

2. Group extensions

A group N is said to be *nilpotent* if its upper central series

$$1 = N_0 \lhd N_1 = Z(N) \lhd N_2 \lhd \cdots,$$

defined by $N_{i+1}/N_i = Z(N/N_i)$, is finite. If n is the smallest integer such that $N_n = N$, then N is said to be *n-step nilpotent*. We shall say that a finitely generated, torsion-free group Γ admits an (n-step) *almost-Bieberbach* group structure if it can be written as an extension of a finitely generated, (n-step) nilpotent group N by a finite group Φ. Notice that, given such a torsion-free, finitely generated, nilpotent group, its quotients N_{i+1}/N_i are of a special form, namely $N_{i+1}/N_i \cong \mathbb{Z}^{i_j}$.

Lemma 2.1. *Let Γ fit in an extension*

$$0 \longrightarrow \mathbb{Z}^m \longrightarrow \Gamma \xrightarrow{p} G \longrightarrow 1,$$

where the torsion-free group G has an n-step nilpotent, normal subgroup N of finite index and $\mathbb{Z}^m$ a trivial N-module. Then $\Gamma \in \mathscr{AB}^{n+1}$.

Proof. Let G be defined by the extension

$$1 \longrightarrow N \longrightarrow G \longrightarrow \Phi \longrightarrow 1,$$

with N n-step nilpotent, Φ finite and $\phi : \Phi \to \operatorname{Out} N$ the operator homomorphism. Consider $\overline{\Gamma} := p^{-1}(N)$. Then the extension

$$0 \longrightarrow \mathbb{Z}^m \longrightarrow \overline{\Gamma} \xrightarrow{p} N \longrightarrow 1 \tag{2–1}$$

is central, which implies that $\overline{\Gamma}$ is at most $(n+1)$-step nilpotent. The proof is completed by the observation that $\overline{\Gamma} = p^{-1}(N) \lhd \Gamma$ and $\Gamma/\overline{\Gamma} \cong (\Gamma/\mathbb{Z}^m)/(\overline{\Gamma}/\mathbb{Z}^m) \cong G/N \cong \Phi$. Notice that Γ is torsion-free since so are $\mathbb{Z}^m$ and G. □

We now turn our attention to the fibre of the fibration $F \to T \to B$ to prove the following:

Proposition 2.2. *Let Γ be a torsion-free extension*

$$1 \longrightarrow K \longrightarrow \Gamma \xrightarrow{p} G \longrightarrow 1 \tag{2–2}$$

of a finitely generated group K by a group G admitting an n-step nilpotent, almost-Bieberbach structure such that $\operatorname{Im} c$ is finite, where $c : \Gamma \to \operatorname{Aut} K$ denotes the conjugation map. Then $\Gamma \in \mathcal{AB}^{n+1}$.

Proof. Since G admits an n-step almost-Bieberbach structure there is a short exact sequence

$$1 \longrightarrow N \longrightarrow G \longrightarrow \Phi \longrightarrow 1$$

where N is n-step nilpotent, Φ is finite, and $\phi : \Phi \to \operatorname{Out} N$ is the operator homomorphism. Let $\hat{\Gamma} := p^{-1}(N)$. Then $\hat{\Gamma}$ fits in a short exact sequence

$$1 \longrightarrow K \longrightarrow \hat{\Gamma} \xrightarrow{p} N \longrightarrow 1,$$

where we denote by $\bar{c}$ the restriction of the conjugation map $c : \Gamma \to \operatorname{Aut} K$ to $\hat{\Gamma}$. Let $\bar{\Gamma} := \operatorname{Ker} \bar{c}$, which is nonempty since Γ is infinite. Then the extension

$$1 \longrightarrow \bar{\Gamma} \cap K \longrightarrow \bar{\Gamma} \longrightarrow p(\bar{\Gamma}) \longrightarrow 1$$

is central, with $p(\bar{\Gamma}) \lhd N$, and therefore itself nilpotent. This means that $\bar{\Gamma} \cap K$ is a finitely generated, torsion-free, abelian group and $\bar{\Gamma}$ is at most $(n+1)$-step nilpotent. The proof is completed by observing that the normal subgroup $\bar{\Gamma}$ of $\hat{\Gamma}$ has finite index in $\hat{\Gamma}$, since $\operatorname{Im} \bar{c}$ is finite. □

The group $\operatorname{Aut} K$, for K a Bieberbach group, is not necessarily finite. For an example, see [Charlap 1986, p. 219]. It does, then, make sense to check what happens if the fibre admits an n-step almost-Bieberbach structure. But first:

Proposition 2.3. *Let Γ be a torsion-free extension*

$$1 \longrightarrow K \longrightarrow \Gamma \xrightarrow{p} \mathbb{Z}^n \longrightarrow 0$$

of a Bieberbach group K by a free abelian group of rank n, such that $\mathbb{Z}^m \subseteq Z(\Gamma)$, where $\mathbb{Z}^m$ is the translation subgroup of K and $Z(\Gamma)$ the center of Γ. Then $\Gamma \in \mathcal{AB}^2$.

Proof. First observe that $\mathbb{Z}^m \lhd \Gamma$, since $\mathbb{Z}^m \leq Z(\Gamma)$. We therefore have a short exact sequence

$$1 \longrightarrow K/\mathbb{Z}^m \longrightarrow \Gamma/\mathbb{Z}^m \xrightarrow{p} \mathbb{Z}^n \longrightarrow 0 \tag{2–3}$$

where $K/\mathbb{Z}^m$ is isomorphic to F, the finite holonomy group of K. We distinguish two cases:

(i) Assume that (2–3) is a central extension. Then choose $Q := (\mathbb{Z}^{n-1} \times k\mathbb{Z}) \lhd \mathbb{Z}^n$ of index $k = |F|$, with $|F|$ the exponent of F. Let $\Gamma' := p^{-1}(Q)$. Then Γ' fits in a short exact sequence

$$1 \longrightarrow F \longrightarrow \Gamma' \xrightarrow{p} Q \longrightarrow 0$$

that splits as a direct product. By construction, $\Gamma' = (F \times Q) \lhd \Gamma/\mathbb{Z}^m$ is of finite index. Let $q : \Gamma \to \Gamma/\mathbb{Z}^m$ be the identification map. The free abelian group Q imbeds as a normal subgroup of $\Gamma' \lhd \Gamma/\mathbb{Z}^m$ and so also, because $\operatorname{Aut}(F \times Q) = \operatorname{Aut} F \times \operatorname{Aut} Q$, as a normal subgroup of $\Gamma/\mathbb{Z}^m$. Let $\tilde{\Gamma} := q^{-1}(Q)$ and $\hat{\Gamma} := q^{-1}(\Gamma')$; then $Q \cong \tilde{\Gamma}/\mathbb{Z}^m$ and $F \times Q \cong \hat{\Gamma}/\mathbb{Z}^m$. Since $Q \leq \mathbb{Z}^n$, it acts on $\mathbb{Z}^m$ in the same way as $\mathbb{Z}^n$, namely trivially. So $\tilde{\Gamma}$ is 2-step nilpotent normal in Γ. One can further check that its index $|\Gamma/\tilde{\Gamma}|$ in Γ is finite, because $|\Gamma/\tilde{\Gamma}| = |\Gamma/\hat{\Gamma}| \cdot |\hat{\Gamma}/\tilde{\Gamma}| = |\mathbb{Z}^n/Q| \cdot |F|$. This completes the proof in this case.

(ii) Assume that the sequence (2–3) is not central, and let $c : \Gamma/\mathbb{Z}^m \to \operatorname{Aut} F$ be the conjugation map. Since F is finite and $\Gamma/\mathbb{Z}^m$ infinite, the kernel of c is nontrivial. Let $\bar{\Gamma} := \operatorname{Ker} c \lhd \Gamma/\mathbb{Z}^m$, let $\bar{F} := F \cap \bar{\Gamma}$, and let $\bar{Q} := p(\bar{\Gamma})$. Then the extension

$$1 \longrightarrow \bar{F} \longrightarrow \bar{\Gamma} \longrightarrow \bar{Q} \longrightarrow 0$$

with $\bar{Q} \lhd \mathbb{Z}^n$ (so that $\bar{Q} \cong \mathbb{Z}^\rho$ for some $\rho \leq n$) belongs to the previous case. The result now follows, since $\bar{\Gamma}$ has finite index in Γ. □

Proposition 2.4. *Let Γ be a torsion-free extension*

$$1 \longrightarrow K \longrightarrow \Gamma \xrightarrow{p} G \longrightarrow 1, \tag{2–4}$$

where K and G admit m-step and n-step almost-Bieberbach structures, respectively. If $Z(L/L_i) \subseteq Z(\Gamma/L_i)$, where $\{L_i\}_i$ is the upper central series of an m-step nilpotent, normal subgroup L of finite index in K, then $\Gamma \in \mathcal{AB}^{n+m}$.

Proof. We first check inductively that $L_i \lhd \Gamma$. This is clear for $i = 1$. Assume it is true for i and let $q_i : L \to L/L_i$ be the identification map, where

$$L_{i+1} = q^{-1}\big(Z(L/L_i)\big),$$

so that $L_{i+1}/L_i \cong Z(L/L_i)$. Then $L_{i+1}/L_i \lhd Z(\Gamma/L_i)$ and $L_{i+1} \lhd \Gamma$. The rest of the proof also follows by induction, first on m and then on n. The group G is of the form

$$1 \longrightarrow N \longrightarrow G \longrightarrow \Phi \longrightarrow 1,$$

where N is n-step nilpotent and Φ finite. By letting $\hat{\Gamma} := p^{-1}(N)$ we get a sequence

$$1 \longrightarrow K \longrightarrow \hat{\Gamma} \longrightarrow N \longrightarrow 1.$$

The case $m = n = 1$ follows from Proposition 2.3. Assuming the theorem is true for some m and $n = 1$, we shall show it is true for $m + 1$ and $n = 1$. If K is of the form

$$1 \longrightarrow L \longrightarrow K \longrightarrow F \longrightarrow 1,$$

where L is m-step nilpotent and F finite, consider $L_1 = Z(L)$. The conditions of the theorem ensure that $L_1 = Z(L) \cong \mathbb{Z}^\rho \lhd \Gamma$ for some positive integer ρ. This gives a short exact sequence

$$1 \longrightarrow K/\mathbb{Z}^\rho \longrightarrow \hat{\Gamma}/\mathbb{Z}^\rho \longrightarrow \mathbb{Z}^\nu \longrightarrow 0,$$

with $\nu > 0$. Then $\{L_i/L_1\}_i$ is the upper central series of L/L_1 and $\hat{\Gamma}/\mathbb{Z}^\rho$ admits an $(n+1)$-step almost-Bieberbach structure by the induction hypothesis. $\hat{\Gamma}$ fits into a central short exact sequence

$$0 \longrightarrow \mathbb{Z}^\rho \longrightarrow \hat{\Gamma} \longrightarrow \hat{\Gamma}/\mathbb{Z}^\rho \longrightarrow 1.$$

Lemma 2.1 now applies to prove that $\hat{\Gamma}$, and therefore Γ, admit an $(n+2)$-step almost-Bieberbach structure. Now assume the theorem is true for all m and n up to a certain value. We complete the proof by showing it holds for all m and $n + 1$. If $\{N_i\}_i$ is the upper central series of some $(n+1)$-step nilpotent N, define $\overline{\Gamma} := p^{-1}(N_n)$. Then $\overline{\Gamma}$ fits in

$$1 \longrightarrow K \longrightarrow \overline{\Gamma} \longrightarrow N_n \longrightarrow 1$$

and admits an $(n+m)$-step almost-Bieberbach structure. Also there is a positive integer μ such that the sequence

$$1 \longrightarrow \overline{\Gamma} \longrightarrow \hat{\Gamma} \longrightarrow \overline{\Gamma}/\hat{\Gamma} \cong \mathbb{Z}^\mu \longrightarrow 0$$

is exact. The induction argument on the fibre implies that $\hat{\Gamma} \in \mathcal{AB}^{n+m+1}$, and so $\Gamma \in \mathcal{AB}^{n+m+1}$ too. □

3. Almost-Bieberbach groups and Kähler structures

Let $\operatorname{Aut} G$ denote the group of automorphisms of a simply connected Lie group G. We shall be concerned with discrete subgroups Γ of $\operatorname{Aut} G$ that act properly discontinuously on G.

A group Γ is said to be *crystallographic* if it is a cocompact, discrete subgroup of $\mathbb{R}^n \rtimes O(n) \subset \operatorname{Aff}(\mathbb{R}^n)$, where $O(n)$ is the maximal compact subgroup of $\operatorname{GL}(n, \mathbb{R})$ and $\operatorname{Aff}(\mathbb{R}^n)$ is the group of Euclidean motions of $\mathbb{R}^n$. It is a *Bieberbach* crystallographic group if it is torsion-free as well. Bieberbach groups are precisely the fundamental groups of compact, complete Riemannian manifolds that are flat (locally isometric to Euclidean space), as first proved in [Bieberbach 1911]. An alternative characterisation of flat Riemannian manifolds is that in such manifolds, transition maps can be extended to elements of $\mathbb{R}^n \rtimes O(n)$. Charlap [1965] classified these manifolds, up to connection-preserving diffeomorphisms, by associating to a manifold M a short exact sequence

$$1 \longrightarrow \Lambda \longrightarrow G \longrightarrow \Phi \longrightarrow 1$$

in which the holonomy group Φ of M is finite and $\Lambda \cong \mathbb{Z}^n$ is the translation subgroup of $\Gamma \cong \pi_1(M)$, a torsion-free, discrete, cocompact subgroup of $\mathbb{R}^n \rtimes O(n) \subset \mathrm{Aff}(\mathbb{R}^n)$.

More generally, if G is a simply connected, nilpotent Lie group, we consider a maximal compact subgroup $C \subseteq \mathrm{Aut}\, G$. A cocompact, discrete subgroup Γ of $G \rtimes C$ is called an *almost-crystallographic* group, and if torsion-free it is called *almost-Bieberbach*. The quotient G/Γ is called an *infra-nilmanifold*, and if $\Gamma \subseteq G$ it is a *nilmanifold*.

Most of Bieberbach's work has been generalised to the nilpotent case in [Auslander 1960] and [Lee and Raymond 1985]:

Theorem 3.1 (Auslander). *Let $\Gamma \subseteq G \rtimes \mathrm{Aut}\, G$ be an almost-crystallographic group, where G is a connected, simply connected, nilpotent Lie group. Then $(\Gamma \cap G) \lhd \Gamma$ is a cocompact lattice in G, and $\Gamma/(\Gamma \cap G)$ is finite.*

Parts of the statement of the following theorem can already be found in [Lee and Raymond 1985]. We simplify the proof.

Theorem 3.2. *Γ is almost-crystallographic if and only if it is of the form*

$$1 \longrightarrow N \longrightarrow \Gamma \longrightarrow \Phi \longrightarrow 1,$$

with N finitely generated, torsion-free, maximal nilpotent, and Φ finite.

Proof. If $\Gamma \subseteq G \rtimes \mathrm{Aut}\, G$ is an almost-Bieberbach group, Theorem 3.1 says that $N = \Gamma \cap G$ is a maximal nilpotent, normal subgroup of Γ of finite index, and finitely generated because it is a discrete subgroup of the nilpotent group G. To prove the converse, given an extension like the one in the statement of the theorem, with abstract kernel $\phi : \Phi \to \mathrm{Out}\, N$, consider the extension of the Malcev completion $\mathcal{N}$ of N,

$$1 \longrightarrow \mathcal{N} \longrightarrow S(\Gamma) \longrightarrow \Phi \longrightarrow 1,$$

with abstract kernel $\psi : \Phi \xrightarrow{\phi} \mathrm{Out}\, N \to \mathrm{Out}\, \mathcal{N}$. The claim is that there is exactly one extension of $\mathcal{N}$ by Φ, namely $\mathcal{N} \hookrightarrow \mathcal{N} \rtimes_{\hat{\psi}} \Phi$, where $\hat{\psi} : \Phi \to \mathrm{Aut}\, \mathcal{N}$ is a lifting morphism of ψ.

Since $Z(\mathcal{N})$ is a vector space and Φ is finite, $H^3(\Phi, Z(\mathcal{N}))$ vanishes and by [Mac Lane 1963, Theorem 8.7] the abstract kernel $[\Phi, \mathcal{N}, \psi]$ has an extension. Furthermore, [Mac Lane 1963, Theorem 8.8] says that this extension is unique because the set $H^2(\Phi, Z(\mathcal{N}))$ parametrizing all congruence classes of such extensions is null, for the same reason. So we know that there is precisely one extension $\mathcal{N} \hookrightarrow \Gamma \twoheadrightarrow \Phi$. If we can further show that ψ has a lifting morphism $\hat{\psi} : \Phi \to \operatorname{Aut} \mathcal{N}$, then $\Gamma \cong \mathcal{N} \rtimes_{\hat{\psi}} \Phi$. To this end, we apply induction on the nilpotency class of $\mathcal{N}$. If $\mathcal{N}$ is 1-step nilpotent then $\mathcal{N} \cong Z(\mathcal{N})$ and the result is obvious. If $\mathcal{N}$ is c-step nilpotent, consider the inverse image under the natural projection $q : \operatorname{Aut} \mathcal{N} \to \operatorname{Out} \mathcal{N}$ of the finite group $\psi(\Phi)$. This gives birth to a short exact sequence $\operatorname{Inn} \mathcal{N} \hookrightarrow q^{-1}(\psi(\Phi)) \twoheadrightarrow \psi(\Phi)$ with $\operatorname{Inn} \mathcal{N} \cong \mathcal{N}/Z(\mathcal{N})$ fulfilling the induction hypothesis. We can thus find a splitting morphism $s : \psi(\Psi) \to q^{-1}(\psi(\Phi)) < \operatorname{Aut} \mathcal{N}$. But now, $s \circ \psi$ is the lifting we were looking for, completing the proof. We thus have the commutative diagram

$$\begin{array}{ccccccccc}
1 & \longrightarrow & N & \longrightarrow & \Gamma & \longrightarrow & \Phi & \longrightarrow & 1 \\
 & & {\scriptstyle \imath}\big\downarrow & & {\scriptstyle \jmath}\big\downarrow & & \big\| & & \\
1 & \longrightarrow & \mathcal{N} & \longrightarrow & \mathcal{N} \rtimes \Phi & \longrightarrow & \Phi & \longrightarrow & 1
\end{array}$$

The map $\jmath$, with $\jmath(n, g) = (\imath(n), g)$, embeds Γ as a discrete, cocompact subgroup of the disconnected Lie group $S(\Gamma)$, proving the theorem. □

Given a short exact sequence $\mathbb{Z}^{2n} \hookrightarrow \Gamma \twoheadrightarrow \Phi$ with operator homomorphism $\phi : \Phi \to \operatorname{Aut} \mathbb{Z}^{2n}$, we say ϕ is *essentially complex* if there is a *complex structure* for the Φ-module $\mathbb{Z}^{2n} \otimes \mathbb{R}$, that is, a map $t \in \operatorname{End}_{\mathbb{R}[\Phi]}(\mathbb{Z}^{2n} \otimes \mathbb{R})$ such that $t^2 = -1$. In other words, $\phi : \Phi \to \operatorname{Aut} \mathbb{Z}^{2n}$ is essentially complex if $\operatorname{Im} \phi \subseteq \mathrm{GL}_{\mathbb{C}}\big((\mathbb{Z}^{2n} \otimes \mathbb{R})^t\big)$, with

$$\mathrm{GL}_{\mathbb{C}}\big((\mathbb{Z}^{2n} \otimes \mathbb{R})^t\big) := \{m \in \mathrm{GL}_{\mathbb{R}}(\mathbb{Z}^{2n} \otimes \mathbb{R}) \text{ such that } mt = tm\}.$$

Theorem 3.3. *Let Γ be the torsion-free extension*

$$1 \longrightarrow N \longrightarrow \Gamma \longrightarrow \Phi \longrightarrow 1,$$

where N is a torsion-free, finitely generated maximal nilpotent group and Φ is a finite group. Then there is a compact Kähler $K(\Gamma, 1)$-manifold M if and only if $N \cong \mathbb{Z}^{2n}$ and the operator homomorphism $\phi : \Phi \to \operatorname{Aut} N$ is essentially complex.

Proof. By Theorem 3.2 there is a connected, simply connected Lie group G such that Γ is a torsion-free, discrete, cocompact subgroup of $G \rtimes \operatorname{Aut} G$. Since M is a $\mathrm{K}(\Gamma, 1)$-manifold, its universal covering is homeomorphically equivalent to G and $M \cong G/\Gamma$. The hypotheses on N say that G contains $N \hookrightarrow \Gamma$ as a discrete cocompact subgroup. Then $\hat{M} \cong G/N$ is a compact $\mathrm{K}(N, 1)$-nilmanifold that covers M in a finite, unramified way. Because the Kähler condition is local, the fact that M admits a Kähler structure implies that $\hat{M}$ also admits a Kähler structure. The Benson–Gordon theorem says that this can happen only if $N \cong \mathbb{Z}^{2n}$, forcing the finite cover $\hat{M}$ of M to be holomorphically equivalent to the complex torus $\mathbb{C}^n/\mathbb{Z}^{2n}$. The converse is settled by [Johnson and Rees 1991, Theorem 3.1]. $\square$

Let $\mathcal{B}_{\mathcal{K}}$ be the class of groups that can be realised as fundamental groups of compact, Kähler manifolds whose underlying Riemannian structure is flat, and $\mathcal{B}_{\mathcal{P}} \subseteq \mathcal{B}_{\mathcal{K}}$ the subclass consisting of groups that can be realised as fundamental groups of complex projective varieties. Let $\mathcal{AB}_{\mathcal{K}}$ denote the class of groups that can be realised as fundamental groups of compact nilmanifolds; that is, compact manifolds of the form G/Γ, where G is a simply connected, nilpotent Lie group and Γ a discrete subgroup admitting a Kähler structure, and let $\mathcal{AB}_{\mathcal{P}} \subseteq \mathcal{AB}_{\mathcal{K}}$ be the subclass consisting of groups that can be realised as fundamental groups of complex projective nilvarieties.

Corollary 3.4. (1) $\mathcal{AB}_{\mathcal{K}} \equiv \mathcal{B}_{\mathcal{K}} \equiv \mathcal{B}_{\mathcal{P}} \equiv \mathcal{AB}_{\mathcal{P}}$.

(2) *Every group in $\mathcal{AB}_{\mathcal{K}}$ is the fundamental group of a smooth, compact, complex algebraic surface.*

Proof. (1) The first equality follows directly from Theorem 3.3 and [Johnson and Rees 1991, Theorem 3.1]. The second is [Johnson 1990, Corollary 4.3], while the third stems from the first two together with the inclusion $\mathcal{AB}_{\mathcal{P}} \subseteq \mathcal{AB}_{\mathcal{K}}$.

(2) If M is a smooth projective manifold, then by Bertini's theorem there is a smooth hyperplane section $M_{(n-1)}$. By the Lefschetz hyperplane theorem [Milnor 1963], $\pi_l(M, M_{(n-1)}) = 0$ for $l < n$, so M and $M_{(n-1)}$ have isomorphic fundamental groups if $n \geq 3$. $\square$

We now combine Proposition 2.2, Proposition 2.4 and Theorem 3.3:

Theorem 3.5. *If the Kähler manifold T is the total space of a fibration $F \to T \to B$ over an infra-nilmanifold B with aspherical fibre F and if the short exact sequence*

$$1 \longrightarrow \pi_1(F) \longrightarrow \pi_1(T) \longrightarrow \pi_1(B) \longrightarrow 1$$

of their respective fundamental groups satisfies the conditions of either Proposition 2.2 or Proposition 2.4, then T admits a flat Riemannian metric.

4. Virtually polycyclic groups and Kähler rigidity

An *affinely flat* manifold is an n-manifold endowed with an atlas whose transition maps can be extended to elements of $\mathrm{Aff}(\mathbb{R}^n) = \mathbb{R}^n \rtimes \mathrm{GL}(n, \mathbb{R})$. A torsion-free group Γ is *virtually polycyclic* if it has a subgroup Γ_0 of finite index which is polycyclic, that is, one that admits a finite composition series $\Gamma_0 \supseteq \Gamma_1 \supseteq \Gamma_2 \supseteq \cdots \supseteq \Gamma_n = 1$ such that $\Gamma_i/\Gamma_{i+1} \cong \mathbb{Z}$ for all i. The number n is an invariant, called the *rank* of Γ. Groups in $\mathcal{AB}^n$ are obviously virtually polycyclic. Auslander [1964] has conjectured that the fundamental group of a compact, complete, affinely flat manifold has to be virtually polycyclic. Milnor [1977] has shown that torsion-free, virtually polycyclic groups can be realised as fundamental groups of complete affinely flat manifolds. On the other hand, Johnson [1976] has proved that torsion-free, virtually polycyclic groups can be realised as fundamental groups of compact $K(\pi, 1)$-manifolds. However, contrary to the Bieberbach case, Benoist [1992] has given an example of a 10-step nilpotent group of rank 11, proving that it is not always possible to do both!

If Γ is a virtually polycyclic group, the *Fitting group* of Γ, denoted $\mathrm{Fitt}(\Gamma)$, is the unique maximal normal subgroup of Γ. The *closure* $\overline{\mathrm{Fitt}(\Gamma)}$ of the Fitting group of a group Γ is the maximal normal subgroup of Γ containing $\mathrm{Fitt}(\Gamma)$ as a normal subgroup of finite index. The basic property of $\overline{\mathrm{Fitt}(\Gamma)}$ is that it leaves the quotient $\Gamma/\overline{\mathrm{Fitt}(\Gamma)}$ with no finite, normal subgroup in it — in other words, *almost-torsion-free*. In [Dekimpe and Igodt 1994] it is proved that if Γ is a finitely generated virtually nilpotent group then Γ is almost-torsion-free if and only if $\overline{\mathrm{Fitt}(\Gamma)}$ is almost-crystallographic.

If N is a torsion-free, finitely generated, c-step nilpotent group, then to any extension

$$N \hookrightarrow \Gamma \overset{p}{\twoheadrightarrow} Q$$

with abstract kernel $\psi : Q \to \mathrm{Out}\, N$ we can inductively associate c morphisms $\psi_i : Q \to \mathrm{Aut}(N_{i+1}/N_i)$, where $N_{i+1}/N_i = Z(N/N_i)$. Now if $q \in \Gamma$ is such that $p(q)$ has finite order in Q, and $\langle q, N\rangle$ is nilpotent, then $p(q) \in \bigcap_1^c \mathrm{Ker}\,\psi_i$. Conversely, if $q \in \Gamma$ is such that $p(q) \in \bigcap_1^c \mathrm{Ker}\,\psi_i$, then $\langle q, N\rangle$ is nilpotent in Γ.

We shall use the following lemma, which is half of [Dekimpe and Igodt 1994, Theorem 2.2]. For completeness, we write a proof here.

Lemma 4.1. *Let* Γ *be a virtually polycyclic group. If* Γ *is almost-torsion-free,* $\mathrm{Fitt}(\Gamma)$ *is torsion-free maximal nilpotent in* Γ.

Proof. Since Γ is polycyclic-by-finite, $\mathrm{Fitt}(\Gamma)$ is finitely generated nilpotent. Therefore its torsion set is a finite characteristic subgroup of $\mathrm{Fitt}(\Gamma)$, and thus normal in Γ, and hence trivial since Γ is almost torsion-free. So, Γ fits in an extension

$$1 \longrightarrow \mathrm{Fitt}(\Gamma) \longrightarrow \Gamma \overset{p}{\longrightarrow} Q \longrightarrow 1 \tag{4–1}$$

with $\mathrm{Fitt}(\Gamma)$ torsion-free and Q abelian-by-finite, say $A \hookrightarrow Q \overset{j}{\twoheadrightarrow} F$. Now let $q \in \Gamma$ be such that $N := \langle q, \mathrm{Fitt}(\Gamma)\rangle$ is nilpotent, and look at $p(N)$. If $p(N) \cap A \neq \{1\}$ then $p^{-1}(p(N) \cap A)$ is normal in Γ since $(p(N) \cap A) \lhd A$ is nilpotent as a subgroup of N. Thus, $p^{-1}(p(N) \cap A) \subseteq \mathrm{Fitt}(\Gamma)$ and $p(N) \cap A = \{1\}$, a contradiction. We deduce that $p(N) \cong j(p(N)) \subseteq F$, and hence that $p(q)$ is of finite order in Q. The discussion preceding the theorem shows that $p(q) \in \bigcap_1^c(\psi_i) \cap p(N)$, where ψ_i are the morphisms associated with (4–1), which is a finite group since

F is finite; therefore $q \in \overline{\mathrm{Fitt}(\Gamma)}$. But since Γ is almost torsion-free, $\overline{\mathrm{Fitt}(\Gamma)}$ is almost crystallographic and $\mathrm{Fitt}(\overline{\mathrm{Fitt}(\Gamma)}) = \mathrm{Fitt}(\Gamma)$ is maximal nilpotent in $\overline{\mathrm{Fitt}(\Gamma)}$, implying $q \in \mathrm{Fitt}(\Gamma)$, a contradiction. □

Theorem 4.2. *Let Γ be a torsion-free, virtually polycyclic group. Then Γ can be realised as the fundamental group of a $K(\pi, 1)$ compact, Kähler manifold if and only if Γ is Bieberbach with essentially complex operator homomorphism.*

Proof. The converse is the second half of Theorem 3.3. For the direct statement, observe that since Γ is torsion-free, it is almost-torsion-free. Thus, by Lemma 4.1, $\mathrm{Fitt}(\Gamma)$ is torsion-free maximal nilpotent in Γ, and Γ fits in a short exact sequence of the form

$$1 \longrightarrow \mathrm{Fitt}(\Gamma) \longrightarrow \Gamma \xrightarrow{p} Q \longrightarrow 1,$$

where Q is abelian-by-finite. Since Γ is Kähler, by [Arapura and Nori 1999], there exists a nilpotent subgroup $\Delta \subseteq \Gamma$ of finite index. But Δ is necessarily contained in $\mathrm{Fitt}(\Gamma)$, so Q is finite, and Theorem 3.3 completes the proof with $N = \mathrm{Fitt}(\Gamma)$. □

Provided that the Auslander conjecture is true, Theorem 4.2 would immediately imply:

Conjecture 4.3. *If a Kähler manifold T is the total space of a fibration $F \to T \to B$ where both the base B and the fibre F are infra-nilmanifolds, then T admits a Riemann flat structure.*

References

[Apanasov and Xie 1997] B. Apanasov and X. Xie, "Geometrically finite complex hyperbolic manifolds", *Internat. J. Math.* **8**:6 (1997), 703–757. MR 99e:57051 Zbl 0912.53027

[Arapura and Nori 1999] D. Arapura and M. Nori, "Solvable fundamental groups of algebraic varieties and Kähler manifolds", *Compositio Math.* **116**:2 (1999), 173–188. MR 2000k:14018 Zbl 0971.14020

[Auslander 1960] L. Auslander, "Bieberbach's theorems on space groups and discrete uniform subgroups of Lie groups", *Ann. of Math.* (2) **71** (1960), 579–590. MR 22 #12161 Zbl 0099.25602

[Auslander 1964] L. Auslander, "The structure of complete locally affine manifolds", *Topology* **3**:suppl. 1 (1964), 131–139. MR 28 #4463 Zbl 0136.43102

[Auslander and Johnson 1976] L. Auslander and F. E. A. Johnson, "On a conjecture of C. T. C. Wall", *J. London Math. Soc.* (2) **14**:2 (1976), 331–332. MR 54 #11341 Zbl 0364.22008

[Benoist 1992] Y. Benoist, "Une nilvariété non affine", *C. R. Acad. Sci. Paris Sér. I Math.* **315**:9 (1992), 983–986. MR 93j:22008 Zbl 0776.57010

[Benson and Gordon 1988] C. Benson and C. S. Gordon, "Kähler and symplectic structures on nilmanifolds", *Topology* **27**:4 (1988), 513–518. MR 90b:53042 Zbl 0672.53036

[Bieberbach 1911] L. Bieberbach, "Über die Bewegungsgruppen der Euklidischen Räume (erste Abh.)", *Math. Ann.* **70** (1911), 297–336. JFM 42.0144.02

[Buser and Karcher 1981] P. Buser and H. Karcher, *Gromov's almost flat manifolds*, Astérisque **81**, Société Mathématique de France, Paris, 1981. MR 83m:53070 Zbl 0459.53031

[Charlap 1965] L. S. Charlap, “Compact flat riemannian manifolds, I”, *Ann. of Math.* (2) **81** (1965), 15–30. MR 30 #543 Zbl 0132.16506

[Charlap 1986] L. S. Charlap, *Bieberbach groups and flat manifolds*, Universitext, Springer, New York, 1986. MR 88j:57042 Zbl 0608.53001

[Dekimpe and Igodt 1994] K. Dekimpe and P. Igodt, “The structure and topological meaning of almost-torsion free groups”, *Comm. Algebra* **22**:7 (1994), 2547–2558. MR 95f:20057 ZBL 0802.20041

[Gromov 1978] M. a. Gromov, “Almost flat manifolds”, *J. Differential Geom.* **13**:2 (1978), 231–241. MR 80h:53041 Zbl 0432.53020

[Gromov 1981] M. Gromov, “Groups of polynomial growth and expanding maps”, *Inst. Hautes Études Sci. Publ. Math.* **53** (1981), 53–73. MR 83b:53041 Zbl 0474.20018

[Johnson 1990] F. E. A. Johnson, “Flat algebraic manifolds”, pp. 73–91 in *Geometry of low-dimensional manifolds* (Durham, 1989), vol. 1, edited by S. K. Donaldson and C. B. Thomas, London Math. Soc. Lecture Note Ser. **150**, Cambridge Univ. Press, Cambridge, 1990. MR 93k:32064 Zbl 0837.57021

[Johnson and Rees 1991] F. E. A. Johnson and E. G. Rees, “Kähler groups and rigidity phenomena”, *Math. Proc. Cambridge Philos. Soc.* **109**:1 (1991), 31–44. MR 91i:58040 Zbl 0736.53058

[Lee and Raymond 1985] K. B. Lee and F. Raymond, “Rigidity of almost crystallographic groups”, pp. 73–78 in *Combinatorial methods in topology and algebraic geometry* (Rochester, NY, 1982), edited by J. R. Harper and R. Mandelbaum, Contemp. Math. **44**, Amer. Math. Soc., Providence, RI, 1985. MR 87d:57026 Zbl 0575.57026

[Mac Lane 1963] S. Mac Lane, *Homology*, Grundlehren der math. Wissenschaften **114**, Springer, Berlin, 1963. MR 28 #122 Zbl 0133.26502

[Malcev 1949] A. I. Mal’cev, “On a class of homogeneous spaces”, *Izvestiya Akad. Nauk. SSSR. Ser. Mat.* **13** (1949), 9–32. In Russian; English translation in *Amer. Math. Soc. Translations* **39**, 1951, 1–33. MR 10,507d

[Milnor 1963] J. Milnor, *Morse theory*, Annals of Mathematics Studies **51**, Princeton University Press, Princeton, N.J., 1963. MR 29 #634 Zbl 0108.10401

[Milnor 1977] J. Milnor, “On fundamental groups of complete affinely flat manifolds”, *Advances in Math.* **25**:2 (1977), 178–187. MR 56 #13130 Zbl 0364.55001

[Tits 1981] J. Tits, “Groupes à croissance polynomiale (d’après M. Gromov et al.)”, pp. 176–188 in *Bourbaki Seminar*, 1980/81, Lecture Notes in Math. **901**, Springer, Berlin, 1981. MR 83i:53065 Zbl 0507.20015

[Wolf 1968] J. A. Wolf, “Growth of finitely generated solvable groups and curvature of Riemaniann manifolds”, *J. Differential Geometry* **2** (1968), 421–446. MR 40 #1939 Zbl 0207.51803

Received June 25 2003. Revised June 5 2004.

NICKOLAS J. MICHELACAKIS
MATHEMATICS AND STATISTICS DEPARTMENT
UNIVERSITY OF CYPRUS
P.O. BOX 20537
NICOSIA 1678
CYPRUS

njm@ucy.ac.cy

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

UPPER BOUNDS FOR THE SPECTRAL RADIUS OF THE $n \times n$ HILBERT MATRIX

PETER OTTE

We derive upper bounds for the spectral radius of the $n \times n$ Hilbert matrix. The key idea is to write the Hilbert matrix as integral operator with positive kernel function and then to use a Wielandt-type min-max principle for the spectral radius. Choosing special trial functions yields a new bound that improves the best bound known heretofore.

1. Introduction

The spectral asymptotics of the Hilbert matrix has attracted a lot of interest concerning both the lowest and the largest eigenvalue. Here we shall focus on the spectral radius ρ_n of the $n \times n$ Hilbert matrix for which we shall prove, particularly, the bound

$$\rho_n \leq 2w_n \arcsin \frac{1}{w_n} \quad \text{with} \quad w_n := 2\left(\frac{(n!)^2}{(2n)!}\right)^{1/2n}, \quad n \in \mathbb{N}. \tag{1}$$

This improves, at least for large values of n, Cassels' bound, given in (5) below, which is the best hitherto known. Numerical computations suggest that (1) is actually better for all n except $n = 1, 2$.

We base the proof of (1) upon relating the Hilbert matrix to an integral operator H_n whose spectral radius can be expressed by a min-max principle for operators having positive kernel functions:

$$\rho_n = \inf_{\varphi \in M} \sup_{0<x<1} \frac{(H_n\varphi)(x)}{\varphi(x)} \tag{2}$$

where M is some set of appropriate trial functions. For the sake of completeness we shall prove (2) without recourse to the general theory. In the matrix case the above min-max principle is due to Wielandt [1950] and related to the enclosure result of Collatz [1942]. It has been generalized in many directions; see [Friedland 1990; Marek 1966; Schaefer 1984], for example.

MSC2000: 15A42, 15A60, 47G10.
Keywords: Hilbert matrix, Hilbert inequality, spectral radius, Wielandt min-max principle, integral operator.

To derive estimates we pick $\varphi(x) := (1-x)^{\gamma}$ in (2), with $-1 < \gamma < 0$. We restrict ourselves to the case $\gamma = -\frac{1}{2}$, for which the calculations are manageable, and obtain (1).

Hilbert was the first one to investigate spectral properties of the matrix named after him. In his lectures he showed his double series theorem stating that ρ_n stays finite as $n \to \infty$; this was first published by Weyl [1908] (see also [Wiener 1910]). The concrete inequality

$$\rho_n \leq \pi \tag{3}$$

is due to Schur [1911]. This is the optimal constant that does not depend on the dimension n. However, if we do want the bound to depend on n it is possible to strengthen (3). Frazer [1946] obtained

$$\rho_n \leq n \sin \frac{\pi}{n} \quad \text{for } n \geq 2, \tag{4}$$

by refining a method of Fejér and Riesz [1921], which they used to prove what is now called the Fejér–Riesz inequality for analytic functions. Equation (4) was later rediscovered by Hsiang [1957] and Yahya [1965], and was eventually improved by Cassels [1948] to

$$\rho_n \leq 2 \arctan \sqrt{2n}. \tag{5}$$

Finally, it might be instructive to look at the asymptotic expansion of ρ_n. The first asymptotic result was obtained by Taussky [1949] by computing the quadratic form with special trial vectors having components $c_k := 1/\sqrt{k}$; it was

$$\rho_n = \pi + O\left(\frac{1}{\ln n}\right).$$

The exact asymptotic behaviour

$$\rho_n = \pi - \frac{\pi^5}{2 \ln^2 n} + O\left(\frac{\ln \ln n}{\ln^3 n}\right)$$

was determined by de Bruijn and Wilf [1962], who compared the matrix operator with an integral operator whose spectral asymptotics can be derived from general results of Widom [1958] (see also [Widom 1961]).

2. Estimates for the spectral radius

We start by relating the Hilbert matrix

$$A_n := \left(\frac{1}{j+k+1}\right)_{j,k=0,\dots,n-1} \tag{6}$$

to the integral operator $H_n : C[0,1] \to C[0,1]$ having the kernel function

$$K_n(xy) := \sum_{j=0}^{n-1} (xy)^j = \frac{1-(xy)^n}{1-xy}. \tag{7}$$

For $n = \infty$ this operator was used by Magnus [1950] to study the spectrum of the infinite Hilbert matrix. We let H_n act on $C[0,1]$ because we want to have sufficiently many trial functions at hand. As we hoped, H_n has (almost) the same spectrum as A_n. In particular, they have the same spectral radius, henceforth denoted by ρ_n.

Lemma 1. *Let $C[0,1]$ be equipped with the usual maximum norm. Then $H_n : C[0,1] \to C[0,1]$ is a bounded linear operator. The respective spectra of the Hilbert matrix A_n and the integral operator H_n are the same apart from* 0. *Their common spectral radius ρ_n can be expressed by*

$$\rho_n = \inf_{\varphi \in M} \sup_{0<x<1} \frac{(H_n\varphi)(x)}{\varphi(x)}, \text{ where } M := \left\{\varphi \in L^1[0,1] \,\middle|\, \varphi > 0,\ \tfrac{1}{\varphi} \in C[0,1]\right\}. \tag{8}$$

Proof. It is clear from the definition and (7) that H_n is linear and bounded. Also (7) shows that H_n has n-dimensional range spanned by the monomials x^k, for $k = 0, \dots, n-1$, which implies that the spectrum of H_n consists only of eigenvalues. To each $c \in \mathbb{C}^n$ we associate $\varphi_c \in C[0,1]$ in the natural way:

$$c = (c_0, \dots, c_{n-1}) \in \mathbb{C}^n \;\longleftrightarrow\; \varphi_c(x) = \sum_{j=0}^{n-1} c_j x^j. \tag{9}$$

The statement on the spectra then follows from

$$\begin{aligned}(H_n\varphi_c)(x) &= \int_0^1 \sum_{j=0}^{n-1} (xy)^j \sum_{k=0}^{n-1} c_k y^k \, dy \\ &= \sum_{j,k=0}^{n-1} c_k x^j \int_0^1 y^{j+k} \, dy = \sum_{j=0}^{n-1} x^j \sum_{k=0}^{n-1} \frac{1}{j+k+1} c_k.\end{aligned}$$

Note that H_n must have a kernel and A_n does not.

To prove Formula (8) we recall from the Perron–Frobenius Theorem that, since A_n has positive entries, ρ_n is an eigenvalue of A_n and hence of H_n. Let v be the corresponding eigenfunction. Writing down the eigenvalue equation for v and dividing by $\varphi \in M$ yields

$$\rho_n \frac{v(x)}{\varphi(x)} = \int_0^1 K_n(xy) \frac{\varphi(y)}{\varphi(x)} \frac{v(y)}{\varphi(y)} \, dy.$$

This shows that $v/\varphi \in C[0,1]$ is an eigenfunction of the operator $H_{n,\varphi}$ with kernel

$$K_{n,\varphi}(xy) := K_n(xy)\frac{\varphi(y)}{\varphi(x)},$$

whence $\rho_n \le \rho(H_{n,\varphi})$, the spectral radius of $H_{n,\varphi}$. Since $\rho(H_{n,\varphi}) \le \|H_{n,\varphi}\|_\infty$ we conclude that

$$\rho_n \le \|H_{n,\varphi}\|_\infty = \sup_{0<x<1}\int_0^1 K_{n,\varphi}(xy)\,dy$$

where we have used $\varphi(x) > 0$, $K_n(xy) \ge 0$, and thus $K_{n,\varphi}(xy) \ge 0$. To show equality in (8) we once again invoke the Perron–Frobenius Theorem, according to which the eigenvector of A_n belonging to ρ_n can be chosen to have positive components, whence we can, via (9), likewise choose the eigenfunction $v > 0$. In particular, $v \in M$. □

We use Lemma 1 to estimate the spectral radius from above by cleverly choosing trial functions in (8):

$$r_n(x) := \frac{(H_n\varphi)(x)}{\varphi(x)} = \frac{1}{\varphi(x)}\int_0^1 \frac{1-(xy)^n}{1-xy}\varphi(y)\,dy \quad \text{for } n \in \mathbb{N}. \tag{10}$$

To get an idea of what the φ's should look like we cast r_n into a form more amenable to further investigation. The crucial point is to evaluate the integral

$$J_n(x) := \int_0^1 \frac{y^n}{1-xy}\varphi(y)\,dy.$$

We start by differentiating with respect to x:

$$J_n'(x) = \int_0^1 \frac{y^{n+1}}{(1-xy)^2}\varphi(y)\,dy = \frac{1}{x}\int_0^1 \frac{y^n}{(1-xy)^2}\varphi(y)\,dy - \frac{1}{x}J_n(x). \tag{11}$$

The explicitly written integral on the right can also be produced by integration by parts, which we perform in such a way that $\varphi(1)$ is omitted because our trial functions will have a singularity at $x = 1$:

$$\begin{aligned} J_n(x) &= \left[(y-1)\frac{y^n}{1-xy}\varphi(y)\right]_0^1 \\ &\qquad - \int_0^1 (y-1)\left(\left(\frac{ny^{n-1}}{1-xy} + \frac{xy^n}{(1-xy)^2}\right)\varphi(y) + \frac{y^n}{1-xy}\varphi'(y)\right)dy \\ &= \delta_n\varphi(0) + nJ_{n-1}(x) - nJ_n(x) + (x-1)\int_0^1 \frac{y^n}{(1-xy)^2}\varphi(y)\,dy \\ &\qquad + J_n(x) + \tilde{J}_n(x), \end{aligned}$$

with $\delta_n := \delta_{n,0}$ the Kronecker delta and

$$\tilde{J}_n(x) := \int_0^1 \frac{y^n}{1-xy}(1-y)\varphi'(y)\,dy.$$

Hence we can eliminate the integral in question from (11):

$$J_n'(x) = \frac{1}{x(1-x)}\big(\delta_n\varphi(0) + nJ_{n-1}(x) - nJ_n(x) + \tilde{J}_n(x)\big) - \frac{1}{x}J_n(x). \tag{12}$$

To eliminate the annoying J_{n-1} we observe that

$$J_n(x) = \int_0^1 \frac{y^n}{1-xy}\varphi(y)\,dy = \frac{1}{x}J_{n-1}(x) - \frac{1}{x}\int_0^1 y^{n-1}\varphi(y)\,dy,$$

and therewith (12) becomes

$$J_n'(x) = \frac{\delta_n}{x(1-x)}\varphi(0) - \frac{n+1}{x}J_n(x) + \frac{\kappa_n}{x(1-x)} + \frac{1}{x(1-x)}\tilde{J}_n(x) \tag{13}$$

where we have put

$$\kappa_0 := 0, \qquad \kappa_n := n\int_0^1 y^{n-1}\varphi(y)\,dy \quad \text{for } n \in \mathbb{N}. \tag{14}$$

We are going to express

$$\Phi_n(x) := \frac{x^n}{\varphi(x)}J_n(x) \tag{15}$$

by dint of (13) through a differential equation:

$$\begin{aligned}
\Phi_n'(x) &= -\frac{\varphi'(x)}{\varphi^2(x)}x^n J_n(x) + \frac{nx^{n-1}}{\varphi(x)}J_n(x) + \frac{x^n}{\varphi(x)}J_n'(x)\\
&= -\left(\frac{\varphi'(x)}{\varphi(x)} + \frac{1}{x}\right)\Phi_n(x) + \frac{x^{n-1}}{(1-x)\varphi(x)}\tilde{J}_n(x) + \frac{x^{n-1}}{(1-x)\varphi(x)}(\delta_n\varphi(0)+\kappa_n).
\end{aligned}$$

At this point we fix our trial function φ in such a way that

$$(1-x)\varphi'(x) = -\gamma\varphi(x), \tag{16}$$

that is, $\varphi(x) = (1-x)^\gamma$ with some $\gamma \in \mathbb{R}$ to be specified later, whereby $\tilde{J}_n$ becomes a multiple of J_n, and we arrive at a differential equation for Φ_n:

$$\Phi_n'(x) = -(\gamma+1)\frac{1}{x}\Phi_n(x) + \frac{x^{n-1}}{(1-x)^{1+\gamma}}\left(\delta_n + \kappa_n\right). \tag{17}$$

This is equivalent to

$$(x^{1+\gamma}\Phi_n(x))' = \frac{x^{n+\gamma}}{(1-x)^{1+\gamma}}(\delta_n + \kappa_n),$$

which we can solve immediately for Φ_n:

$$\Phi_n(x) = \frac{\delta_n + \kappa_n}{x^{1+\gamma}} \int_0^x \frac{\xi^{n+\gamma}}{(1-\xi)^{1+\gamma}}\, d\xi \quad \text{for } n \in \mathbb{N}_0. \tag{18}$$

In particular, we can now see that γ must satisfy $-1 < \gamma < 0$ in order to yield well defined integrals and to have $\varphi \in M$ in (8). We summarize our calculations.

Theorem 2. *The spectral radius ρ_n of the $n \times n$ Hilbert matrix A_n as in* (6) *can be estimated by*

$$\rho_n \le \inf_{0<\alpha<1} \sup_{0<x<1} \frac{1}{x^{1-\alpha}} \int_0^x \frac{1-\kappa_n \xi^n}{\xi^\alpha (1-\xi)^{1-\alpha}}\, d\xi, \tag{19}$$

where

$$\kappa_n = \frac{n!}{(n-\alpha)(n-1-\alpha)\cdots(1-\alpha)} \quad \textit{for } n \in \mathbb{N}. \tag{20}$$

Proof. Put $\alpha := -\gamma$ and use in turn the min-max principle (8) and the definitions of r_n and Φ_n as in (10) and (15), with φ being chosen according to (16) to obtain

$$\rho_n \le \inf_{0<\alpha<1} \sup_{0<x<1} r_n(x) = \inf_{0<\alpha<1} \sup_{0<x<1} \big(\Phi_0(x) - \Phi_n(x)\big).$$

Then (19) follows directly from the representation (18) of Φ_n.

For φ as in (16) the integral in (14) is Euler's beta function. Hence,

$$\kappa_n = nB(n, 1-\alpha) = \frac{n\,\Gamma(n)\,\Gamma(1-\alpha)}{\Gamma(n+1-\alpha)}$$

wherefrom we deduce (20). □

The optimal way to derive bounds on ρ_n would be to determine the maximum of the function r_n exactly. Unfortunately, this turns out to be rather complicated, and we content ourselves with narrowing the region where the maximum must lie.

Corollary 3. *The spectral radius ρ_n of the $n \times n$ Hilbert matrix A_n can be estimated by*

$$\rho_n \le \inf_{0<\alpha<1} \kappa_n^{(1-\alpha)/n} \int_0^{1/\kappa_n^{1/n}} \frac{1}{\xi^\alpha (1-\xi)^{1-\alpha}}\, d\xi, \tag{21}$$

which in the case $\alpha = \frac{1}{2}$ specializes to

$$\rho_n \le 2w_n \arcsin \frac{1}{w_n} \quad \textit{with } w_n := \kappa_n^{1/2n} = 2\left(\frac{(n!)^2}{(2n)!}\right)^{1/2n}. \tag{22}$$

Proof. When $1-\kappa_n\xi^n \le 0$ the function r_n is decreasing, whence the maximum must lie in the interval $[0, x_0]$ for x_0 the unique zero of the integrand in (19):

$$1-\kappa_n x_0^n = 0, \quad \text{i.e.,} \quad x_0 = 1/\kappa_n^{1/n}.$$

We conclude

$$\sup_{0<x<1} r_n(x) = \sup_{0<x\le x_0} r_n(x) \le \sup_{0<x\le x_0} \Phi_0(x) = \Phi_0(x_0)$$

because the $\Phi_n(x)$ are nonnegative and Φ_0 increases.

For $\alpha = \frac{1}{2}$ the w_n are easily obtained from (22) and (20) and the integral in (21) can be evaluated by the change of variables $\xi = s^2$. □

Finally, we shall check that our estimate (22) is indeed better than (5). Using some familiar formulae for arctan and arcsin we obtain

$$\begin{aligned}\arctan\sqrt{2n} - w_n \arcsin\frac{1}{w_n} &= \arctan\sqrt{2n} - \arcsin\frac{1}{w_n} - (w_n-1)\arcsin\frac{1}{w_n} \\ &\ge \arctan\sqrt{2n} - \arctan\frac{1}{\sqrt{w_n^2-1}} - \frac{\pi}{2}(w_n-1) \\ &= \arctan\frac{\sqrt{2n(w_n^2-1)}-1}{\sqrt{w_n^2-1}+\sqrt{2n}} - \frac{\pi}{2}(w_n-1).\end{aligned}$$

Now the asymptotics of the middle binomial coefficient yields

$$w_n \sim 2\left(\frac{\sqrt{\pi n}}{4^n}\right)^{1/2n} \sim n^{1/4n} \quad \text{as } n\to\infty,$$

which implies immediately $\lim_{n\to\infty} w_n = 1$, and further

$$n(w_n-1) \sim n(n^{1/4n}-1) = n(e^{(\ln n)/4n}-1) \sim \tfrac{1}{4}\ln n \quad \text{as } n\to\infty.$$

Therefore, for large n,

$$\frac{\sqrt{2n(w_n^2-1)}-1}{\sqrt{w_n^2-1}+\sqrt{2n}} \ge \frac{1}{4}\frac{\sqrt{2n(w_n^2-1)}}{\sqrt{2n}} \ge \frac{1}{4}\sqrt{w_n-1}.$$

Noting $\arctan x \ge cx$ for small x with some constant $c>0$ and using the monotonicity of arctan we conclude

$$\arctan\sqrt{2n} - w_n\arcsin\frac{1}{w_n} \ge \frac{c}{4}\sqrt{w_n-1} - \frac{\pi}{2}(w_n-1) > 0$$

for large values of n. With some care it should be possible to show the statement for smaller values of n, too.

3. Remarks

We suggest some topics that might be worth further investigation.

(1) In order to derive from Theorem 2 a bound that can be computed more or less explicitly we did not determine in (19) the maximum of the function r_n exactly. Thus, the first possibility to strengthen (22) is to study the maximum of r_n.

(2) Also for computational reasons we fixed the exponent $\alpha = \frac{1}{2}$ in (21). However, numerical computations suggest that $\alpha = \frac{1}{2}$ is generally not the optimal choice and that other values of α give much more accurate estimates. According to a theorem of Čebyšev the integral in (21) can be evaluated for any $0 < \alpha < 1$ in closed form by means of elementary functions. It is not clear whether these elementary functions allow for an efficient minimizing procedure.

(3) A vaguer idea is to pick other trial functions than $(1-x)^{-\alpha}$. Our method will work as long as we arrive at a differential equation for Φ_n as in (17).

(4) Since Wielandt's min-max principle is accompanied by a max-min principle, one can also think of deriving lower bounds for the spectral radius in which case; however, completely different trial functions are needed.

Acknowledgments

I thank Robert Denk and Hubert Kalf for stimulating discussions and their steady interest in this subject.

References

[de Bruijn and Wilf 1962] N. G. de Bruijn and H. S. Wilf, "On Hilbert's inequality in *n* dimensions", *Bull. Amer. Math. Soc.* **68** (1962), 70–73. MR 31 #291 Zbl 0105.04201

[Cassels 1948] J. W. S. Cassels, "An elementary proof of some inequalities", *J. London Math. Soc.* **23** (1948), 285–290. MR 10,434g Zbl 0034.18501

[Collatz 1942] L. Collatz, "Einschließungssatz für die charakteristischen Zahlen von Matrizen", *Math. Z.* **48** (1942), 221–226. MR 5,30d Zbl 0027.00604

[Fejér and Riesz 1921] L. Fejér and F. Riesz, "Über einige funktionentheoretische Ungleichungen", *Math. Z.* **11** (1921), 305–314. JFM 48.0327.01

[Frazer 1946] H. Frazer, "Note on Hilbert's inequality", *J. London Math. Soc.* **21** (1946), 7–9. MR 8,259h Zbl 0060.14903

[Friedland 1990] S. Friedland, "Characterizations of the spectral radius of positive operators", *Linear Algebra Appl.* **134** (1990), 93–105. MR 92d:47054 Zbl 0707.15005

[Hsiang 1957] F. C. Hsiang, "An inequality for finite sequences", *Math. Scand.* **5** (1957), 12–14. MR 19,955a Zbl 0078.24701

[Magnus 1950] W. Magnus, "On the spectrum of Hilbert's matrix", *Amer. J. Math.* **72** (1950), 699–704. MR 12,836e Zbl 0041.23805

[Marek 1966] I. Marek, "Spektrale Eigenschaften der $\mathcal{K}$-positiven Operatoren und Einschließungssätze für den Spektralradius", *Czechoslovak Math. J.* **16 (91)** (1966), 493–517. MR 36 #711 Zbl 0152.33701

[Schaefer 1984] H. H. Schaefer, "A minimax theorem for irreducible compact operators in L^p-spaces", *Israel J. Math.* **48**:2-3 (1984), 196–204. MR 86g:47042 Zbl 0576.47020

[Schur 1911] I. Schur, "Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen", *J. Reine Angew. Math.* **140** (1911), 1–28. JFM 42.0367.01

[Taussky 1949] O. Taussky, "A remark concerning the characteristic roots of the finite segments of the Hilbert matrix", *Quart. J. Math., Oxford Ser.* **20** (1949), 80–83. MR 11,16a Zbl 0036.01302

[Weyl 1908] H. Weyl, *Singuläre Integralgleichungen mit besonderer Berücksichtigung des Fourierschen Integraltheorems*, Inauguraldissertation, Universität Göttingen, 1908.

[Widom 1958] H. Widom, "On the eigenvalues of certain Hermitian operators", *Trans. Amer. Math. Soc.* **88** (1958), 491–522. MR 20 #4782 Zbl 0101.09202

[Widom 1961] H. Widom, "Extreme eigenvalues of translation kernels", *Trans. Amer. Math. Soc.* **100** (1961), 252–262. MR 25 #2420 Zbl 0197.10903

[Wielandt 1950] H. Wielandt, "Unzerlegbare, nicht negative Matrizen", *Math. Z.* **52** (1950), 642–648. MR 11,710g Zbl 0035.29101

[Wiener 1910] F. Wiener, "Elementarer Beweis eines Reihensatzes von Herrn Hilbert", *Math. Ann.* **68** (1910), 361–366. JFM 41.0391.04

[Yahya 1965] Q. A. M. M. Yahya, "On the generalisation of Hilbert's inequality", *Amer. Math. Monthly* **72** (1965), 518–520. MR 32 #4235 Zbl 0125.30101

Received October 20, 2003. Revised March 19, 2004.

PETER OTTE
RUHR-UNIVERSITÄT BOCHUM
FAKULTÄT FÜR MATHEMATIK
UNIVERSITÄTSSTRASSE 150
D-44780 BOCHUM
GERMANY

peter.otte@ruhr-uni-bochum.de

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

UNRAMIFIED HILBERT MODULAR FORMS, WITH EXAMPLES RELATING TO ELLIPTIC CURVES

JUDE SOCRATES AND DAVID WHITEHOUSE

We give a method to explicitly determine the space of unramified Hilbert cusp forms of weight two, together with the action of Hecke, over a totally real number field of even degree and narrow class number one. In particular, one can determine the eigenforms in this space and compute their Hecke eigenvalues to any reasonable degree. As an application we compute this space of cusp forms for $\mathbb{Q}(\sqrt{509})$, and determine each eigenform in this space which has rational Hecke eigenvalues. We find that not all of these forms arise via base change from classical forms. To each such eigenform f we attach an elliptic curve with good reduction everywhere whose L-function agrees with that of f at every place.

1. Introduction

In general, finding unramified cuspidal representations for a given group is a difficult problem. If one tries to tackle this problem using the trace formula, for example, one usually needs to shrink the discrete group and hence allow some ramification. In this paper we are concerned with computing the space of unramified Hilbert cusp forms for a totally real field of even degree.

Let F be a totally real number field of narrow class number one and of even degree over $\mathbb{Q}$. In Section 2 we explain how, by results of Jacquet, Langlands and Shimizu, the construction of the space of Hilbert cusp forms of weight 2 (i.e., of weight $(2, \dots, 2)$) and full level for F can be done on the quaternion algebra $\boldsymbol{B}$ over F that is ramified precisely at the infinite places of F. In fact the space of such cusp forms can be identified with a certain space of functions on the set of equivalence classes of ideals for a maximal order in $\boldsymbol{B}$.

In Sections 3 and 4 we extend the definition of Θ-series and Brandt matrices, as found in [Pizer 1980a], to this case. We show that each simultaneous eigenvector for the family of modified Brandt matrices corresponds to a Hilbert cusp form that is an eigenvector for all the Hecke operators. In order to compute the Brandt matrices, and hence the space of cusp forms, we need to be able to find representatives for

MSC2000: 11F41, 11G05.
Keywords: Hilbert modular forms, elliptic curves, everywhere good reduction.

all the ideal classes for a maximal order; we outline our strategy to find these representatives in Section 5.

Next we specialize to the case of a real quadratic field of narrow class number one. In Section 6, using a result of Pizer, we give an explicit formula for the type number of $\boldsymbol{B}$ and the class number of a maximal order in $\boldsymbol{B}$. In Section 7 we give defining relations for the quaternion algebra $\boldsymbol{B}$ over a real quadratic field $\mathbb{Q}(\sqrt{m})$, and when $m \equiv 5 \bmod 8$ we give a maximal order in this algebra.

We now turn to our application to elliptic curves. To any Hilbert modular newform $\boldsymbol{f}$ over a totally real field F, having weight 2, level $\mathfrak{n}$ and rational Hecke eigenvalues, one expects to be able to attach an elliptic curve $E_{\boldsymbol{f}}$ that is defined over F, has conductor $\mathfrak{n}$ and whose L-function agrees with that of $\boldsymbol{f}$ at all places of F. This is known if F has odd degree over $\mathbb{Q}$ or if the automorphic representation associated to $\boldsymbol{f}$ belongs to the discrete series at some finite place (see [Blasius 2004, 1.7.1]).

Conjecture 1.1. *Let F be a totally real number field of even degree over $\mathbb{Q}$. To each unramified Hilbert modular eigenform $\boldsymbol{f}$ over F having weight 2 and rational Hecke eigenvalues one can attach an elliptic curve $E_{\boldsymbol{f}}$ defined over F with good reduction everywhere, such that the L-functions of $E_{\boldsymbol{f}}$ and $\boldsymbol{f}$ agree at each place of F.*

When $\boldsymbol{f}$ is the base change of a classical modular form one can sometimes attach an elliptic curve to $\boldsymbol{f}$ as in Conjecture 1.1; see [Shimura 1971, 7.7]. Also, by [Blasius 2004], this conjecture is true under the hypothesis of the Hodge conjecture. In this paper we establish this conjecture for $F = \mathbb{Q}(\sqrt{509})$. The reason for this choice of field is, as we shall see, that there exist eigenforms that do not arise via base change from $\mathrm{GL}_2(\mathbb{Q})$, nor are they CM forms since $h^+(F) = 1$. To our knowledge this provides the first verification of this conjecture in the case that not all forms arise by base change; see [Blasius 2004, 1.7.3].

We now outline the verification of Conjecture 1.1 for $F = \mathbb{Q}(\sqrt{509})$. In Section 8 we give representatives for the ideal classes in $\boldsymbol{B}$ from which we are able to compute the Brandt matrices and therefore the eigenvalues of the unramified eigenforms of weight 2. We find that there are three eigenforms whose Hecke eigenvalues all lie in $\mathbb{Q}$. In Section 9 we give the equations for the three elliptic curves over F that are attached to our three eigenforms. These elliptic curves already exist in the literature [Cremona 1992; Pinch 1982].

In Section 10 we prove Conjecture 1.1 for $\mathbb{Q}(\sqrt{509})$. One of our forms is a base change of a classical form given in [Cremona 1992]. In this case one knows, by work of Shimura, that an elliptic curve is attached to this form. Now we take $\boldsymbol{f}$ to be one of the forms that is not a base change from $\mathbb{Q}$ and we take E to be the elliptic curve (or its Galois conjugate) defined over F given in [Pinch 1982].

By work of Taylor, building on work of Carayol and Wiles, and independently by Blasius and Rogawski, there exists for each rational prime ℓ an ℓ-adic representation

$$\sigma_{\boldsymbol{f},\ell} : \mathrm{Gal}(\bar{F}/F) \to \mathrm{GL}_2(\mathbb{Q}_\ell)$$

which is unramified outside ℓ. If $\mathfrak{p}$ is a prime of F not dividing ℓ and $\mathrm{Fr}_\mathfrak{p}$ is a Frobenius element at $\mathfrak{p}$, we have $\mathrm{Tr}\,\sigma_{\boldsymbol{f},\ell}(\mathrm{Fr}_\mathfrak{p}) = a_{\boldsymbol{f}}(\mathfrak{p})$, the eigenvalue of $\boldsymbol{f}$ with respect to the $\mathfrak{p}$-th Hecke operator, and $\det\sigma_{\boldsymbol{f},\ell}(\mathrm{Fr}_\mathfrak{p}) = N\mathfrak{p}$. Similarly, for each rational prime ℓ we have a representation

$$\sigma_{E,\ell} : \mathrm{Gal}(\bar{F}/F) \to \mathrm{GL}_2(\mathbb{Q}_\ell)$$

given by the action of Galois on the ℓ-adic Tate module of E. Since E has good reduction everywhere, $\sigma_{E,\ell}$ is unramified outside ℓ and for each prime $\mathfrak{p}$ not dividing ℓ, we have $\mathrm{Tr}\,\sigma_{E,\ell}(\mathrm{Fr}_\mathfrak{p}) = a_E(\mathfrak{p})$ and $\det\sigma_{E,\ell}(\mathrm{Fr}_\mathfrak{p}) = N\mathfrak{p}$.

The verification of Conjecture 1.1 for $\boldsymbol{f}$ will therefore be complete if we can show, for some prime ℓ, that these two representations are equivalent. For this we take $\ell = 2$ and use a result of Faltings and Serre proved in [Livné 1987]. We cannot apply this result directly since it requires the traces of all Frobenius elements to be even, which is not the case here. So we begin by showing that the extensions of F cut out by the kernels of the mod 2 representations obtained from the eigenform and the elliptic curve are the same. Having identified these extensions, we can apply the theorem of Faltings and Serre to show that these two representations are equivalent when restricted to this extension of F. Using Frobenius reciprocity we conclude that these representations of $\mathrm{Gal}(\bar{F}/F)$ are equivalent.

This work was begun by the first author in his PhD thesis [Socrates 1993], which gave a construction of the space of cusp forms for a real quadratic field of narrow class number one. The cusp form $\boldsymbol{f}$ above and the elliptic curve E were shown there to have the same L-factors at all primes generated by a totally positive element $a + b\theta$ with $1 \le a \le 64$, where $\theta = \frac{1}{2}(1 + \sqrt{509})$.

This work was completed by the second author, who extended the methods of [Socrates 1993] to any totally real field of narrow class number one with even degree over $\mathbb{Q}$, adapted the result of Faltings and Serre, and independently computed the necessary eigenvalues given in Table 4.

2. Construction of the space of cusp forms

Throughout this paper F will be a totally real number field of narrow class number one and of even degree over $\mathbb{Q}$. We denote by R the ring of integers in F, by F^+ the set of totally positive elements in F, and likewise for R^+. We now explain how one can construct the space of cusp forms for F of weight 2 and full level.

Let $\boldsymbol{B}/F$ be the unique (up to isomorphism) quaternion algebra that is ramified only at the infinite places of F. We now give some definitions.

An *R-lattice* (or *ideal*) V in $\boldsymbol{B}$ is a finitely generated R-submodule of $\boldsymbol{B}$ such that $V \otimes_R F \cong \boldsymbol{B}$. An element $\boldsymbol{b} \in \boldsymbol{B}$ is *integral* (or *an integer*) if $R[\boldsymbol{b}]$ is an R-lattice in $\boldsymbol{B}$. An *order* in $\boldsymbol{B}$ is a ring $\mathbb{O}$ consisting of integers and containing R such that $F\mathbb{O} = \boldsymbol{B}$. A *left ideal* I for an order $\mathbb{O}$ is an R-lattice for which $\mathbb{O}I \subset I$. Two left $\mathbb{O}$-ideals I_1 and I_2 are said to be *right equivalent* if $I_1 = I_2\boldsymbol{b}$ for some $\boldsymbol{b} \in \boldsymbol{B}^\times$. Similarly, two orders $\mathbb{O}_1$ and $\mathbb{O}_2$ are *of the same type* if $\mathbb{O}_1 = \boldsymbol{b}\mathbb{O}_2\boldsymbol{b}^{-1}$ for some $\boldsymbol{b} \in \boldsymbol{B}^\times$. The number H of right equivalence classes of left $\mathbb{O}$-ideals is called the *class number of* $\mathbb{O}$ and the number T of type classes of maximal orders of $\boldsymbol{B}$ is called the *type number of* $\boldsymbol{B}$. Both numbers are finite (for any order $\mathbb{O}$).

We now fix a maximal order $\mathbb{O}$ in $\boldsymbol{B}$. Let $\boldsymbol{G} = \boldsymbol{B}^\times$ viewed as an algebraic group over F. Since $\boldsymbol{B}$ only ramifies at the infinite places of F for each finite prime $\mathfrak{p}$ we have

$$\boldsymbol{B} \otimes_F F_\mathfrak{p} \cong M_2(F_\mathfrak{p}).$$

Moreover we can choose these isomorphisms so as to give an isomorphism of $\mathbb{O}_\mathfrak{p} = \mathbb{O} \otimes R_\mathfrak{p}$ with $M_2(R_\mathfrak{p})$. Clearly each of these isomorphisms gives rise to an isomorphism of $\boldsymbol{G}(F_\mathfrak{p})$ with $\mathrm{GL}_2(F_\mathfrak{p})$ under which $\mathbb{O}_\mathfrak{p}^\times$ corresponds to $\mathrm{GL}_2(R_\mathfrak{p})$.

We construct the double coset space

$$X = M_{\boldsymbol{G}} \backslash \boldsymbol{G}(\boldsymbol{A}_F^f)/\boldsymbol{G}(F),$$

where $\boldsymbol{A}_F^f$ is the ring of finite adèles and $M_{\boldsymbol{G}} = \prod_{\mathfrak{p}<\infty} \mathrm{GL}_2(R_\mathfrak{p})$ is a maximal compact open subgroup of $\boldsymbol{G}(\boldsymbol{A}_F^f)$. We note that $M_{\boldsymbol{G}}$, as a subgroup of $\boldsymbol{G}(\boldsymbol{A}_F^f)$, depends on the choice of $\mathbb{O}$ and hence so does X. The set X can be identified with the right equivalence classes of left $\mathbb{O}$-ideals in the following way. Given $(x_\mathfrak{p}) \in \boldsymbol{G}(\boldsymbol{A}_F^f)$, consider the open compact subset $\prod_\mathfrak{p} \mathbb{O}_\mathfrak{p} x_\mathfrak{p}$ in $\boldsymbol{B} \otimes \boldsymbol{A}_F^f$. Taking the intersection of $\prod_\mathfrak{p} \mathbb{O}_\mathfrak{p} x_\mathfrak{p}$ with $\boldsymbol{B}$, embedded diagonally in $\boldsymbol{B} \otimes \boldsymbol{A}_F^f$, yields a left $\mathbb{O}$-ideal. Conversely, given a left $\mathbb{O}$-ideal I one recovers an element of $\boldsymbol{G}(\boldsymbol{A}_F^f)$ by choosing, for each prime $\mathfrak{p}$, a generator of the principal left $\mathbb{O}_\mathfrak{p}$-ideal $\mathbb{O}_\mathfrak{p} I$.

We denote by S the space

$$S = \{\, \boldsymbol{f} : X \to \mathbb{C}\}/\{ \text{ constant functions on } X \,\}.$$

There is a natural definition of Hecke operators on this space, as follows. Let $\pi_\mathfrak{p}$ be a uniformizer for $R_\mathfrak{p}$ and let $g_\mathfrak{p} \in \boldsymbol{G}(\boldsymbol{A}_F^f)$ be such that the $\mathfrak{p}$-th component of $g_\mathfrak{p}$ is

$$\begin{pmatrix} \pi_\mathfrak{p} & 0 \\ 0 & 1 \end{pmatrix}$$

and is the identity otherwise. Since $\mathrm{GL}_2(R_\mathfrak{p})$ is open and compact in $\mathrm{GL}_2(F_\mathfrak{p})$, we have $M_{\boldsymbol{G}} g_\mathfrak{p} M_{\boldsymbol{G}} = \coprod_{i=1}^n M_{\boldsymbol{G}} g_i$. A classical result states that we can choose the set

$\{g_i\}$ to be

$$\left\{\begin{pmatrix}\pi_\mathfrak{p} & 0\\ \alpha & 1\end{pmatrix} : \alpha \in R/\mathfrak{p}\right\} \bigcup \left\{\begin{pmatrix}1 & 0\\ 0 & \pi_\mathfrak{p}\end{pmatrix}\right\}.$$

Define, for $\boldsymbol{f} \in S$ and $h \in \boldsymbol{G}(\boldsymbol{A}_F^f)$,

$$\big(\boldsymbol{T}_\mathfrak{p}(\boldsymbol{f})\big)(h) = \sum_{i=1}^{n} \boldsymbol{f}(g_i h).$$

This gives a well-defined action on S, which is independent of the choices of the g_i and also of $\pi_\mathfrak{p}$.

Let $\mathcal{S}$ be the $\mathbb{C}$-vector space of holomorphic Hilbert cusp forms over F of weight 2 and full level. Then $\mathcal{S}$ is a multiplicity-free direct sum of simultaneous 1-dimensional Hecke eigenspaces. A similar decomposition holds for S. By [Gelbart and Jacquet 1979], there is a Hecke-equivariant isomorphism between $\mathcal{S}$ and S.

Our goal now is to give a method to compute the action of the Hecke operators on the space S. This will be done by constructing Brandt matrices $B(\xi)$ and modified Brandt matrices $B'(\xi)$, which are families of rational matrices indexed by $\xi \in R^+$. These are objects that were first defined over $\mathbb{Q}$ and later used to construct cusp forms for congruence subgroups of $\mathrm{SL}2(\mathbb{Z})$.

3. Θ-series of an ideal

The notion and construction of a Θ-series for an ideal in a quaternion algebra is discussed in several papers, including [Pizer 1976; 1980a; 1980b; Gross 1987]. In this section we extend these definitions to ideals in a totally definite quaternion algebra $\boldsymbol{B}$ defined over F.

Let J be an ideal in the totally definite quaternion algebra $\boldsymbol{B}$. Let nr denote the reduced norm from $\boldsymbol{B}$ to F. The norm of any nonzero element in $\boldsymbol{B}$ is totally positive. We denote by $\mathrm{nr}(J)_+$ a totally positive generator of $\mathrm{nr}(J)$, the fractional ideal of F generated by the norms of the elements in J. For any $\beta \in J$ we define

$$\mathcal{N}_J(\beta) = \mathrm{nr}(\beta)/\,\mathrm{nr}(J)_+.$$

We define the Θ-series of J for $\tau \in \mathcal{H}^{\mathrm{Hom}(F,\mathbb{R})}$ by

$$\Theta_J(\tau) = \sum_{\beta\in J} \exp(\tau \mathcal{N}_J(\beta)) = \sum_{\xi\in R^+} c_{\xi,J} \exp(\tau\xi),$$

where $c_{\xi,J}$ is the number of elements β in J with $\mathcal{N}_J(\beta) = \xi$. This sum converges since composing $\mathcal{N}_J$ with the trace map from F to $\mathbb{Q}$ gives a positive definite quadratic form on J as a $\mathbb{Z}$-lattice.

Proposition 3.1. *The definition of $c_{\xi,J}$ is independent of the choice of* $\mathrm{nr}(J)_+$.

Proof. Any two choices for $\mathrm{nr}(J)_+$ will differ by a totally positive unit v. Since F has narrow class number one, $v = u^2$ for some unit u. Thus multiplication by $u \in R^\times$ gives a bijection between the set of elements in J of norm $\xi\,\mathrm{nr}(J)_+$ and those of norm $\xi v\,\mathrm{nr}(J)_+$. □

We note that if $J' = \gamma_1 J \gamma_2$ with $\gamma_i \in \boldsymbol{B}^\times$ the Θ-series of J and J' are identical. The proof in [Pizer 1980a, Proposition 2.17] holds in this case.

Suppose that we are given an ideal J in terms of a basis over R. We give an effective algorithm to determine the $c_{\xi,J}$. Let $\{\beta_1, \ldots, \beta_4\}$ be a basis for J over R and let $\{\omega_1, \ldots, \omega_n\}$ be a basis for R over $\mathbb{Z}$. We can write $\beta \in J$ uniquely as

$$\beta = \sum_{i=1}^{4} \sum_{j=1}^{n} x_{ij} \omega_j \beta_i$$

with $x_{ij} \in \mathbb{Z}$. Then $\mathcal{N}_J(\beta)$ is a totally positive element of R, provided $\beta \neq 0$, and composing $\mathcal{N}_J$ with the trace map from F to $\mathbb{Q}$ gives a positive definite quadratic form in the $\{x_{ij}\}$. Therefore, given a basis of an ideal J and $M \in \mathbb{R}$ we can use [Cohen 1993, Algorithm 2.7.7] to compute $c_{\xi,J}$ for all $\xi \in R^+$ with $\mathrm{Tr}\,\xi \leq M$.

4. Brandt matrices and eigenforms

Brandt matrices were classically constructed from a complete set of representatives of left $\mathbb{O}$-ideal classes of an Eichler order $\mathbb{O}$ of $\boldsymbol{B}'$, a definite quaternion algebra over $\mathbb{Q}$ with $\mathrm{Ram}(\boldsymbol{B}') = \{\infty, p\}$. For such a $\boldsymbol{B}'$, [Pizer 1980a; 1980b] show that terms appearing in a so-called Brandt matrix series are actually modular forms (for $\mathbb{Q}$) of a given weight and level p. In this section we extend these definitions to a totally definite quaternion algebra $\boldsymbol{B}$ defined over F. We then give an adelic construction of the Brandt matrices and show that each eigenvector for the family of modified Brandt matrices corresponds to a cusp form.

Let $\mathbb{O}$ be a maximal order in $\boldsymbol{B}$ and $\{I_1, \ldots, I_H\}$ a complete (ordered) set of representatives of distinct left $\mathbb{O}$-ideal classes. For each k let

$$\mathbb{O}_r(I_k) = \{\boldsymbol{b} \in \boldsymbol{B} : I_k \boldsymbol{b} \subset I_k\}$$

denote the right order of I_k; this is another maximal order in $\boldsymbol{B}$. The inverse of I_k is defined by

$$I_k^{-1} = \{\boldsymbol{b} \in \boldsymbol{B} : I_k \boldsymbol{b} I_k \subset I_k\}.$$

Then, for each k, the elements $I_k^{-1} I_1, \ldots, I_k^{-1} I_H$ represent the left $\mathbb{O}_r(I_k)$-ideal classes.

In the notation of Section 3, let

$$e_j = e(I_j) = c_{1,\mathbb{O}_r(I_j)},$$

which is simply the number of elements of norm 1 in the order $\mathcal{O}_r(I_j)$. We define $b_{i,j}(0) = 1/e_j$ and for $\xi \in R^+$

$$b_{i,j}(\xi) = \frac{1}{e_j} c_{\xi, I_j^{-1} I_i},$$

which is $1/e_j$ times the number of elements in the left $\mathcal{O}_r(I_j)$-ideal $I_j^{-1} I_i$ of norm $\xi \operatorname{nr}(I_i)_+ / \operatorname{nr}(I_j)_+$. Now define the *$\xi$-th Brandt matrix for $\mathcal{O}$* by

$$B(\xi, \mathcal{O}) = (b_{i,j}(\xi)).$$

The construction of $B(\xi, \mathcal{O})$ is well defined up to conjugation by a permutation matrix. Moreover, if $\mathcal{O}'$ is another maximal order, the matrices $B(\xi, \mathcal{O})$ and $B(\xi, \mathcal{O}')$ are conjugate by a permutation matrix independent of ξ. In view of this, we shall denote by $B(\xi) = B(\xi, \mathcal{O})$ the ξ-th Brandt matrix, for some fixed maximal order $\mathcal{O}$.

The following properties of the Brandt matrices are stated in [Pizer 1980a] and proved there for quaternion algebras over $\mathbb{Q}$. The proofs carry over for the Brandt matrices defined above.

Theorem 4.1. (1) $e_j b_{i,j}(\xi) = e_i b_{j,i}(\xi)$.

(2) $\sum_{j=1}^{H} b_{i,j}(\xi)$ *is independent of* i. *Denote this value by* $b(\xi)$. *Then* $b(\xi)$ *is the number of integral left* $\mathcal{O}$ *ideals of norm* ξ.

(3) *The Brandt matrices generate a commutative semisimple ring.*

Define the $H \times H$ matrix A by

$$A = \begin{pmatrix} 1 & e_1/e_2 & e_1/e_3 & \dots & e_1/e_H \\ 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

Then for $\xi \in R^+$ or $\xi = 0$ we have

$$AB(\xi)A^{-1} = \begin{pmatrix} b(\xi) & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & B'(\xi) & \\ 0 & & & \end{pmatrix}$$

This is proved in [Pizer 1980a], with the proof carrying over here. The submatrix $B'(\xi)$ will be called the *ξ-th modified Brandt matrix*.

We now show that each simultaneous eigenvector for the family of modified Brandt matrices corresponds to a cusp form. Shimizu [1965] constructed a representation of the Hecke algebra acting on the space of automorphic forms, and in [Hijikata et al. 1989, Chapter 5] it is shown that this can be used to provide another construction of Brandt matrices. We follow the discussion in this latter source, simplifying it for the case that we are interested in.

Fix a maximal order $\mathbb{O}$ in $\boldsymbol{B}$. Let $\boldsymbol{G}$ be the multiplicative group $\boldsymbol{B}^{\times}$, viewed as an algebraic group over F. Every left $\mathbb{O}$-ideal is of the form $\mathbb{O}\tilde{a}$ for some $\tilde{a} \in \boldsymbol{G}(\boldsymbol{A}_F)$. Let

$$\mathcal{U} = \mathcal{U}(\mathbb{O}) = \{\tilde{u} = (u_{\mathfrak{p}}) \in \boldsymbol{G}(\boldsymbol{A}_F) : u_{\mathfrak{p}} \in \mathbb{O}_{\mathfrak{p}}^{\times} \text{ for all } \mathfrak{p} < \infty\}.$$

Since $\tilde{\alpha}\mathcal{U}\tilde{\alpha}^{-1}$ is commensurable with $\mathcal{U}$ for all $\tilde{\alpha} \in \boldsymbol{G}(\boldsymbol{A}_F)$, we can define the usual Hecke ring $R(\mathcal{U}, \boldsymbol{G}(\boldsymbol{A}_F))$; see [Shimura 1971]. Put

$$\mathcal{U}(\boldsymbol{A}_F) = \{\tilde{u} = (u_{\mathfrak{p}}) \in \boldsymbol{I}_F : u_{\mathfrak{p}} \in R_{\mathfrak{p}}^{\times} \text{ for all } \mathfrak{p} < \infty\},$$

where $\boldsymbol{I}_F$ is the group of idèles of F. For $\xi \in R^+$, denote by $\boldsymbol{T}(\xi)$ the element of $R(\mathcal{U}, \boldsymbol{G}(\boldsymbol{A}_F))$ which is the sum of all double cosets $\mathcal{U}\tilde{a}\mathcal{U}$ such that $a_{\mathfrak{p}} \in \mathbb{O}_{\mathfrak{p}}$ for all $\mathfrak{p} < \infty$ and $\mathrm{nr}(\tilde{a}) \in \xi\mathcal{U}(\boldsymbol{A}_F)$.

Denote by $\mathcal{M} = \mathcal{M}_2(\mathbb{O})$ the space of continuous $\mathbb{C}$-valued functions f on $\boldsymbol{G}(\boldsymbol{A}_F)$, satisfying

$$f(u\tilde{a}\boldsymbol{b}) = f(\tilde{a}) \quad \text{for all } u \in \mathcal{U},\ \tilde{a} \in \boldsymbol{G}(\boldsymbol{A}_F), \text{ and } \boldsymbol{b} \in \boldsymbol{G}(F).$$

We define a representation of $R(\mathcal{U}, \boldsymbol{G}(\boldsymbol{A}_F))$ on $\mathcal{M}$ as follows. For

$$\mathcal{U}y\mathcal{U} \in R(\mathcal{U}, \boldsymbol{G}(\boldsymbol{A}_F)),$$

let $\mathcal{U}y\mathcal{U} = \bigcup_i \mathcal{U}y_i$ be its decomposition into disjoint right cosets. Now write

$$\rho(\mathcal{U}y\mathcal{U})\, f(\tilde{a}) = \sum_i f(y_i\tilde{a})$$

and extend ρ to $R(\mathcal{U}, \boldsymbol{G}(\boldsymbol{A}_F))$ by linearity. It is shown in [Hijikata et al. 1989, p. 31] that this representation is independent of the choice of a maximal order, in the sense that, if $\mathbb{O}'$ is another maximal order, there is an isomorphism between $R(\mathcal{U}, \boldsymbol{G}(\boldsymbol{A}_F))$ and $R(\mathcal{U}', \boldsymbol{G}(\boldsymbol{A}_F))$ preserving the Hecke operators $\boldsymbol{T}(\xi)$, and also an isomorphism $\mathcal{M}_2(\mathbb{O})$ and $\mathcal{M}_2(\mathbb{O}')$, such that the representation of $R(\mathcal{U}, \boldsymbol{G}(\boldsymbol{A}_F))$ on $\mathcal{M}_2(\mathbb{O})$ induced by these isomorphisms is equivalent to the original representation of $R(\mathcal{U}, \boldsymbol{G}(\boldsymbol{A}_F))$ on $\mathcal{M}_2(\mathbb{O})$.

If H is the class number of $\mathbb{O}$, we have

$$\boldsymbol{G}(\boldsymbol{A}_F) = \bigcup_{\lambda=1}^{H} \mathcal{U}\tilde{x}_{\lambda}\boldsymbol{G}(F).$$

Note that the $I_{\lambda} = \mathbb{O}\tilde{x}_{\lambda}$ give a complete set of representatives of left $\mathbb{O}$-ideal classes. Since the elements of $\mathcal{M}$ are determined by their values at the x_{λ}, the map

$$f \mapsto (f_1, \dots, f_H) \tag{1}$$

gives an isomorphism of $\mathcal{M}$ with $\mathbb{C}^H = \mathbb{C}_1 \oplus \cdots \oplus \mathbb{C}_H$, where each $\mathbb{C}_i$ is a copy of $\mathbb{C}$. We can use the isomorphism (1) to give a matrix representation for ρ. For $\xi \in R^+$, let

$$B(\xi) = (\rho_{i,j}(\xi))_{i,j=1\ldots H},$$

where multiplication by $\rho_{i,j}(\xi) : \mathbb{C}_j \to \mathbb{C}_i$ is the composition of the injection of $\mathbb{C}_j$ into $\mathbb{C}^H$, the inverse of map (1), $\rho(\mathcal{U}\xi\mathcal{U})$, map (1), and the projection of $\mathbb{C}^H$ into $\mathbb{C}_i$. The following is proved in [Hijikata et al. 1989, Proposition 5.1], with the proof carrying over here.

Proposition 4.2. *The definition of $B(\xi)$ yields the same matrix as the Brandt matrices defined above, assuming that we use the same maximal order $\mathbb{O}$ and set of left $\mathbb{O}$-ideal representatives I_λ.*

We shall now make explicit the isomorphism as Hecke modules between the spaces of Hilbert modular cusp forms and $\mathbb{C}$-valued functions on the finite set X modulo constant functions, which was mentioned in Section 2. We will follow the construction of Hida [1988], which is also discussed in [Taylor 1989]. As before, we shall be interested only in the weight 2, full level case.

Having fixed isomorphisms between $\boldsymbol{G}(F_\mathfrak{p})$ and $\mathrm{GL}_2(F_\mathfrak{p})$ as in Section 2 we set

$$U = M_{\boldsymbol{G}} = \prod_{\mathfrak{p}<\infty} \mathrm{GL}_2(R_\mathfrak{p}),$$

an open subgroup of the finite part of the adelization of $\mathbb{O}$. Denote by $S(U)$ the space of $\mathbb{C}$-valued functions on X, the set of right equivalence classes of left $\mathbb{O}$-ideals. Via the identification of X as a double coset space, $S(U)$ is just the space $\mathcal{M}_2(\mathbb{O})$ defined above. The Hecke action on $S(U)$ is that given in Section 2. Let $\mathrm{inv}(U)$ be the subspace of $S(U)$ consisting of functions of the form $f \circ \mathrm{nr}$, where nr is the reduced norm map

$$\mathrm{nr} : \boldsymbol{G}(\boldsymbol{A}_F^f) \to \boldsymbol{I}_F^f$$

and f is an appropriate $\mathbb{C}$-valued function on $\boldsymbol{I}_F^f$, the finite idèles of F. The map nr, when restricted to the image of $\boldsymbol{B}^\times$, surjects into the totally positive elements of F (this is the Theorem of Norms in [Vignéras 1980, p. 80]). Hence we can view $\mathrm{inv}(U)$ as consisting of functions of the form

$$\boldsymbol{G}(\boldsymbol{A}_F^f) \xrightarrow{\mathrm{nr}} \boldsymbol{I}_F^f \longrightarrow \mathcal{U}(R_\mathfrak{p})\backslash \boldsymbol{I}_F^f/F^+ \xrightarrow{\cong} \mathrm{Cl}^+(F) \longrightarrow \mathbb{C},$$

where $\mathrm{Cl}^+(F)$ is the ray class group of F. Since we are assuming that $h^+(F) = 1$, $\mathrm{inv}(U)$ is the space of constant functions on X.

The Hecke operators certainly fix $\mathrm{inv}(U)$. Thus, in order to examine the Hecke action on the space of cusp forms, we must decompose $S(U)$ into a direct sum of $\mathrm{inv}(U)$ and a space $S_2(U)$ preserved by the Hecke algebra.

We describe the Hecke action on $\mathrm{inv}(U)$. Let $\boldsymbol{T}_{\mathfrak{p}}$ be the $\mathfrak{p}$-th Hecke operator and f the function which is 1 on all elements of X. In Section 2 we saw the decomposition of

$$\Big(\prod_{\mathfrak{p}<\infty} \mathrm{GL}_2(R_{\mathfrak{p}})\Big) g_{\mathfrak{p}} \Big(\prod_{\mathfrak{p}<\infty} \mathrm{GL}_2(R_{\mathfrak{p}})\Big)$$

into disjoint right cosets. Note, though, that in this decomposition we also obtain exactly the elements in $\boldsymbol{G}(\boldsymbol{A}_F)$ that yield, upon multiplying to the right of $\mathbb{O}$, the set of integral left $\mathbb{O}$-ideals of norm $\mathfrak{p}$. Thus $\boldsymbol{T}_{\mathfrak{p}}(f)$ is the function with constant value equal to the number of such ideals.

We have seen that the matrix A transforms the Brandt matrices into two blocks consisting of a 1×1 cell containing $b(\xi)$ and the modified Brandt matrix $B'(\xi)$. And in Theorem 4.1 we noted that $b(\xi)$ is precisely the number of integral left $\mathbb{O}$-ideals of norm ξ. Thus we have:

Proposition 4.3. *Let $\{\boldsymbol{v}_i\}$ be a basis for $\mathbb{C}^{H-1}$ consisting of eigenvectors for all the modified Brandt matrices. Then each $\boldsymbol{v}_i$ corresponds to a* (*normalized*) *holomorphic Hilbert modular eigenform $\boldsymbol{f}_i$ of weight* 2 *and full level whose eigenvalue with respect to the $\mathfrak{p}$-th Hecke operator is precisely the eigenvalue of $\boldsymbol{v}_i$ with respect to $B'(\pi)$, where π is a totally positive generator of $\mathfrak{p}$.*

To find a basis of $\mathbb{C}^{H-1}$ of simultaneous eigenvectors for all the modified Brandt matrices one computes the matrices $B'(\xi)$, ordered by the trace of ξ, and successively decomposes the space $\mathbb{C}^{H-1}$ into simultaneous eigenspaces until one is left with one-dimensional eigenspaces.

It is, of course, desirable to know which of these forms do not arise by base change. Suppose that $F/\mathbb{Q}$ is a cyclic extension with Galois group G. Then G acts on the set of eigenforms via permutation of the primes of F. And one knows that a form does not arise by base change from an intermediate field if and only if its Galois orbit has order equal to the degree of the extension $F/\mathbb{Q}$. Using this one can then determine precisely which forms arise via base change once one has found a basis of $\mathbb{C}^{H-1}$ of simultaneous eigenvectors of the $B'(\xi)$ using the procedure described above. In the case that $F/\mathbb{Q}$ is solvable there are added complications to determining which forms don't arise by base change coming from the existence of Galois fixed Hecke characters that do not descend; see [Rajan 2002].

5. Finding type and ideal class representatives

In order to use Proposition 4.3 to compute the space of cusp forms we need to be able to find representatives for the ideal classes of a maximal order $\mathbb{O}$ in $\boldsymbol{B}$. In this section we give a strategy to find these representatives.

We continue with $\boldsymbol{B}$, the quaternion algebra over F ramified only at the infinite places of F, and we take $\mathbb{O}$ to be a maximal order in $\boldsymbol{B}$. It is easy to manufacture ideals of $\mathbb{O}$ when they are of a particular form. Let $\alpha \in \boldsymbol{B} \setminus F$. Then $K = F(\alpha)$ is a quadratic extension of F contained in $\boldsymbol{B}$. Let I be an ideal in the ring of integers S of K. Then $J = \mathbb{O}I$ is a left ideal of $\mathbb{O}$. Moreover we have $\mathrm{nr}(J) = N_{K/F}(I)$, since $1 \in \mathbb{O}$. Clearly, if I and I' are in the same ideal class in K then J and J' are in the same left $\mathbb{O}$-ideal class.

We will now see that to find representatives of left $\mathbb{O}$-ideal classes it suffices to consider ideals of the form $\mathbb{O}I$ as in the construction above.

Proposition 5.1. *Every left $\mathbb{O}$-ideal class of a maximal order $\mathbb{O}$ contains an ideal of the form $\mathbb{O}I$, where I is an ideal in a field extension $K = F(\boldsymbol{b})$ contained in $\boldsymbol{B}$.*

Proof. The left $\mathbb{O}$-ideal classes are in bijection with

$$X = M_{\boldsymbol{G}} \backslash \boldsymbol{G}(\boldsymbol{A}_F^f)/\boldsymbol{G}(F),$$

as stated in Section 2. Since this is a finite set, there is a finite set of primes S such that $\boldsymbol{G}(\boldsymbol{A}_F^f) = M_{\boldsymbol{G}} \boldsymbol{B}_S^\times \boldsymbol{G}(F)$, where $\boldsymbol{B}_S = \prod_{\mathfrak{p} \in S} \boldsymbol{B}_{\mathfrak{p}}$. Now

$$i_S(\boldsymbol{B}) := \{(\boldsymbol{b}, \ldots, \boldsymbol{b}) \in \boldsymbol{B}_S : \boldsymbol{b} \in \boldsymbol{B}\}$$

is dense in $\boldsymbol{B}_S$; hence $i_S(\boldsymbol{B}^\times)$ is dense in $\boldsymbol{B}_S^\times$. Since $M_{\boldsymbol{G}}$ is open in $\boldsymbol{G}(\boldsymbol{A}_F^f)$ we have by strong approximation $\boldsymbol{G}(\boldsymbol{A}_F^f) = M_{\boldsymbol{G}} i_S(\boldsymbol{B}^\times)\boldsymbol{G}(F)$. Thus every $\beta \in \boldsymbol{G}(\boldsymbol{A}_F^f)$ is of the form $\beta = \mu i_S(\boldsymbol{b})\boldsymbol{b}_0$ for some $\mu \in M_{\boldsymbol{G}}$ and $\boldsymbol{b}, \boldsymbol{b}_0 \in \boldsymbol{B}^\times$. Under the local-global correspondence, then, the left $\mathbb{O}$-ideal $\mathbb{O}\beta$ is in the same class as $\mathbb{O}i_S(\boldsymbol{b})$, where $i_S(\boldsymbol{b})$ can be viewed as a fractional ideal in $F(\boldsymbol{b})$. □

We now outline the algorithm for finding representatives for left $\mathbb{O}$-ideal classes.

1. Determine the class number H. (This can be done; see [Pizer 1973]. We will make this explicit in the case of a quadratic field in Section 6 below.)
2. Initialize the list of representatives of left $\mathbb{O}$-ideal classes to $L = \{\mathbb{O}\}$.
3. Find an element $\alpha \in \boldsymbol{B}$ such that the ring of integers of $K = F[\alpha]$ is exactly $R[\alpha]$.
4. Determine $h = h(K)$ and $S = \{I_1 \ldots I_h\}$, ideal representatives for the class group of K,

   OR

   Generate a large list $S = \{I_i\}$ of prime ideals of K.
5. Now, for $I_i \in S$, do:
   (a) Find a basis for $J_i = \mathbb{O}I_i$.
   (b) Determine if J_i is in the same class as any of the ideals in L obtained so far. If not, add J to L, and keep a note of α and I_i.

6. Stop if H representatives have been found; otherwise resume from Step 3.

We would like to know how to determine if two left $\mathbb{O}$-ideals belong to different ideal classes, which is step 5(b) of the algorithm. In Section 3 we saw that the Θ-series gives a necessary test for two ideals to be in the same class. We now give a necessary and sufficient condition for two ideals to be in the same class.

Proposition 5.2. *Let I and J be left $\mathbb{O}$-ideals for an Eichler order $\mathbb{O}$. Then I and J belong to the same left ideal class if and only if there is an $\alpha \in M = \bar{J} I$ (where $\bar{J}$ denotes the conjugate ideal of J) such that* $\mathrm{nr}(\alpha) = \mathrm{nr}(I)\,\mathrm{nr}(J)$, *i.e., with* $\mathcal{N}_M(\alpha) = 1$.

This is proved in [Pizer 1980a], with the proof valid for any quaternion algebra over a number field. To use this proposition we will need to construct a basis for M, then compute the normalized norm $\mathcal{N}_M$ as in Section 3.

6. Computing *T* and *H*

We now specialize to the case of a real quadratic field $F = \mathbb{Q}(\sqrt{m})$ of narrow class number one. As is well known, this condition implies that either $m = 2$ or m is prime and congruent to 1 mod 4. In this section we give an explicit formula for the type number of $\boldsymbol{B}$ and the class number of a maximal order $\mathbb{O}$ in $\boldsymbol{B}$. The most important tool will be the main theorem in [Pizer 1973], which we restate here:

Theorem 6.1 (Pizer). *Let F be a totally real number field of degree n over $\mathbb{Q}$, and let R be its ring of integers. Let $\boldsymbol{B}$ be a positive definite quaternion algebra over F. Let q_1 be the product of the finite primes in F that ramify in $\boldsymbol{B}$ and q_2 a finite product of distinct finite primes of F such that $(q_1, q_2) = 1$. Then the type number $T_{q_1q_2}$ of Eichler orders of level q_1q_2 in $\boldsymbol{B}$ is*

$$T_{q_1q_2} = \frac{1}{2^e h(F)}\left(M + \frac{1}{2}\sum_{\mathcal{G}_a \in C} E_{q_1q_2}(\mathcal{G}_a)\frac{h(\mathcal{G}_a)}{w(\mathcal{G}_a)}\right), \tag{2}$$

where

- *e is the number of primes dividing q_1q_2;*
- *M is Eichler's mass, given by*

$$M = \frac{2h(F)\zeta_F(2)\,\mathrm{disc}(F)^{3/2}}{(2\pi)^{2n}} \prod_{\mathfrak{p}\mid q_1}(N(\mathfrak{p})-1)\prod_{\mathfrak{p}\mid q_2}(N(\mathfrak{p})+1),$$

  where ζ_F is the zeta function of F;
- *$h(\mathcal{G}_a)$ is the ideal class number of locally principal $\mathcal{G}_a$-fractional ideals;*
- *$w(\mathcal{G}_a)$ is the index of the group of units of R in the group of units in $\mathcal{G}_a$;*

- $E_{q_1q_2}(\mathcal{S}_a) = \prod_{\mathfrak{p} \mid q_1} \Bigl(1 - \Bigl\{\frac{\mathcal{S}_a}{\mathfrak{p}}\Bigr\}\Bigr) \prod_{\mathfrak{p} \mid q_2} \Bigl(1 + \Bigl\{\frac{\mathcal{S}_a}{\mathfrak{p}}\Bigr\}\Bigr)$;
- *C is the collection of all orders defined by the following procedure*:

1. *Let $e_1, \dots, e_s$ be a compete set of representatives of $U \bmod U^2$, where U are the units of R;*
2. *let $d_1, \dots, d_k$ be a complete set of integral ideal representatives of*

$$E.\,\mathrm{Fr}(F)^2 \bmod (\mathrm{Pr}(F)^2),$$

where E is the subgroup of $\mathrm{Fr}(F)$ (*the divisor group of F*) *generated by all the $\mathfrak{p}$ which divide q_1q_2, and* $\mathrm{Pr}(F)$ *is the subgroup of principal divisors of* $\mathrm{Fr}(F)$.

3. *Let $n_1, \dots, n_t$ be a set of all elements of R such that*

   (a) $(n_j) = d_{j'}$ *for some j' with $1 \le j' \le k$, and*
   (b) $(n_i) \ne (n_j)$ *for $i \ne j$.*

4. *Consider the collection of all polynomials over R of the form*

$$f_{\mu,\rho,\tau}(x) = x^2 - \tau x + n_\mu e_\rho \quad \text{with } 1 \le \rho \le s \text{ and } 1 \le \mu \le t,$$

where

   (a) *$f_{\mu,\rho,\tau}$ is irreducible over F,*
   (b) *$F[x]/f_{\mu,\rho,\tau}(x)$ cannot be embedded in any F_{∞_i}, $i = 1, \dots, n$,*
   (c) *$\mathfrak{p}^{s_\mathfrak{p}} \mid \tau$ for all $\mathfrak{p} < \infty$, where $s_\mathfrak{p} = \bigl[\frac{1}{2} v_\mathfrak{p}(n_\mu)\bigr]$ (floor function), and*
   (d) *if $v_\mathfrak{p}(n_\mu)$ is odd then $\mathfrak{p}^{s_\mathfrak{p}+1} \mid \tau$.*

5. *Let a be a root of some $f_{\mu,\rho,\tau}$ and for each $f_{\mu,\rho,\tau}$ choose only one root. Then $C = \{\mathcal{S}_a : \mathcal{S}_a$ is an order of $F(a)\}$ such that*

   (a) *$R[a] \subset \mathcal{S}_a$, and*
   (b) *if $\mathfrak{p} < \infty$ then $a\pi_\mathfrak{p}^{-s_\mathfrak{p}} \in \mathcal{S}_{a,\mathfrak{p}}$, where $s_\mathfrak{p} = \bigl[\frac{1}{2} v_\mathfrak{p}(N(a))\bigr]$.*

We now use this theorem of Pizer to derive a more explicit formula for the algebra $\boldsymbol{B}$ over any real quadratic field of narrow class number one.

Theorem 6.2. *Let $m \equiv 1 \bmod 4$ be a positive squarefree number greater than 5, set $F = \mathbf{Q}(\sqrt{m})$ and let R be the ring of integers in F. Assume that F has narrow class number one. Let $\boldsymbol{B}$ be the totally definite quaternion algebra which is unramified at all the finite primes of F. Then the type number T of $\boldsymbol{B}$ is given by*

$$T = \frac{1}{48m} \sum_{u=1}^{m} \left(\frac{u}{m}\right) u^2 + \tfrac{1}{8} h\bigl(\mathbb{Q}(\sqrt{-m})\bigr) + \tfrac{1}{6} h\bigl(\mathbb{Q}(\sqrt{-3m})\bigr).$$

For completeness we note that if $m = 5$ the type number of $\boldsymbol{B}$ is 1; see [Socrates 1993, Theorem 5.2].

Proof of Theorem 6.2. We proceed to determine the quantities in Theorem 6.1. We have $h(F) = 1$. Since $\boldsymbol{B}$ is unramified at all finite primes, $q_1 = 1$ and for maximal orders $q_2 = 1$. Thus $e = 0$ and the two products in the definition of Eichler's mass M are both empty. Since $m \equiv 1 \bmod 4$ we have $\operatorname{disc} F = m$ and

$$M = \frac{2\zeta_F(2)m^{3/2}}{(2\pi)^4} = \frac{m^{3/2}}{8\pi^4}\zeta_F(2).$$

We shall further simplify M by explicitly calculating $\zeta_F(2)$. Our method will be that of [Leopoldt 1958], which uses generalized Bernoulli numbers; see also [Neukirch 1999, Chapter VII]. Define the n-th Bernoulli number, B_n, by

$$\frac{te^t}{e^t - 1} = \sum_{n \geq 0} B_n \frac{t^n}{n!}.$$

For a character χ mod f, define $B_{n,\chi}$ by

$$\sum_{u=1}^{f} \chi(u) \frac{te^{ut}}{e^{ft} - 1} = \sum_{n \geq 0} B_{n,\chi} \frac{t^n}{n!}.$$

For $F = \mathbf{Q}(\sqrt{m})$, with $m > 0$, define

$$B_{n,F} = \prod_{\chi} B_{n,\chi},$$

where the product runs over the characters mod $d = |\operatorname{disc} F| = m$ that correspond to characters of $\operatorname{Gal}(F/\mathbb{Q})$. Hence this product involves only the trivial character and χ the Legendre symbol mod m. Thus $B_{n,F} = B_n B_{n,\chi}$. In [Leopoldt 1958] it is shown that

$$\zeta_F(n) = \frac{(2\pi)^{2n}\sqrt{d}\,B_{n,F}}{4d^n(n!)^2}$$

if n is a positive even integer. Thus $M = \frac{1}{48}B_{2,\chi}$, since $B_2 = \frac{1}{6}$. Now

$$B_{2,\chi} = \frac{1}{m} \sum_{u=1}^{m} \left(\frac{u}{m}\right) u^2$$

and hence

$$M = \frac{1}{48m} \sum_{u=1}^{m} \left(\frac{u}{m}\right) u^2.$$

Now we proceed with the rest of the algorithm. The product defining

$$E_{q_1q_2}(\mathcal{S}_a) = E_1(\mathcal{S}_a)$$

is also empty regardless of $\mathcal{S}_a$, so $E_1(\mathcal{S}_a) = 1$. Equation (2) then becomes

$$T = M + \frac{1}{2}\sum_{\mathcal{S}_a \in C} \frac{h(\mathcal{S}_a)}{w(\mathcal{S}_a)}.$$

We now follow the algorithm to find the collection C.

1. Since $U = \langle -1\rangle\langle u\rangle$, where u is a fundamental unit of F and $U^2 = \langle u^2\rangle$, we get $s = 4$, and a set of representatives for U mod U^2 is given by $\{\pm 1, \pm u\}$.

2. Since $q_1q_2 = 1$ and $\mathrm{Fr}(F) = \mathrm{Pr}(F)$, we have $k = 1$, $E = (1)$ and $\{(1)\}$ is a complete set of representatives for $E.\mathrm{Fr}(F)^2$ mod $\mathrm{Pr}(F)^2$.

3. From step 2, we can take $t = 1$ and $n = n_1 = 1$.

4. We shall call the polynomials obtained in this step *contributing polynomials*, and denote this set by Ψ. Since $\mu = 1 = t$ and $n = n_1 = 1$ we shall use the abbreviation

$$f_{\rho,\tau}(x) = x^2 - \tau x + e_\rho.$$

Since $v_\mathfrak{p}(n) = 0$ for any $v_\mathfrak{p}$, we have $s_\mathfrak{p} = 0$ for every finite $\mathfrak{p}$, so condition 4(c) is always satisfied by any τ. Condition 4(d) is vacuous. Now we look at condition 4(b). Since F is totally real this condition requires that the discriminant

$$\Delta(f_{\rho,\tau}) = \tau^2 - 4e_\rho$$

of $f_{\rho,\tau}$ be totally negative. But for any τ, $\Delta(f_{-1,\tau})$ and $\Delta(f_{-u,\tau})$ are always positive, since $u > 0$. Hence we need only consider $f_{1,\tau}$ and $f_{u,\tau}$. But $N_{F/\mathbb{Q}}(u) = -1$ tells us that $\sigma(u) < 0$, where σ is the nontrivial element of $\mathrm{Gal}(F/\mathbb{Q})$. So $\sigma(\tau)^2 - 4\sigma(u) > 0$ for any τ. Thus only $e_\rho = 1$ remains. We further abbreviate

$$f_\tau(x) = x^2 - \tau x + 1.$$

Our problem is therefore to find all $\tau = a + b\theta \in R$, where $\theta = \frac{1}{2}(1+\sqrt{m})$, such that $\tau^2 - 4 < 0$ and $\sigma(\tau)^2 - 4 < 0$, i.e., such that

$$-2 < a + b\theta, a + b - b\theta < 2.$$

Thus we see that we necessarily need $-4 < (2\theta - 1)b < 4$, which is $-4 < \sqrt{m}b < 4$. Hence if $m > 16$ then $b = 0$ is the only possible value. In this case, $\tau = a = 0, \pm 1$. Note that these three values yield a contributing f_τ. On the other hand, if $m < 16$

the only possible value for m is 13 and in this case we must have $b = 0$ or ± 1. But for $b = 1$ we must have

$$\frac{-5+\sqrt{13}}{2} < a < \frac{3-\sqrt{13}}{2}$$

and there are no such integers a. On the other hand if $b = -1$ then we must have

$$\frac{-3+\sqrt{13}}{2} < a < \frac{5-\sqrt{13}}{2}$$

and again there are no such integers. Clearly condition 4(a), irreducibility, is satisfied by all the f_τ above since the roots are imaginary. We summarize step 4 in the following result:

Lemma 6.3. *Assume the hypotheses in Theorem 6.2. The only contributing polynomials in Ψ are f_τ with $\tau = 0, \pm 1$.*

The roots of these polynomials and the fields they generate over $\mathbf{Q}(\sqrt{m})$ are shown below.

τ	Roots a_τ, a'_τ of f_τ	$F(a_\tau)$
0	ζ_4, ζ_4^3	$\mathbb{Q}(\sqrt{m}, \zeta_4)$
1	ζ_6, ζ_6^5	$\mathbb{Q}(\sqrt{m}, \zeta_6)$
-1	ζ_6^2, ζ_6^3	$\mathbb{Q}(\sqrt{m}, \zeta_6)$

5. We proceed to the last step of the algorithm: finding the orders $\mathcal{S}_a$. Condition 5(a) says that $R[a_\tau]$ must be contained in $\mathcal{S}_a$. However, we find that $R[a_\tau]$ is the maximal order of $F(a_\tau)$.

Lemma 6.4. *Let m be as in Theorem 6.2, R the ring of integers of $\mathbf{Q}(\sqrt{m})$ and u a fundamental unit in R.*

(1) *The ring of integers of $\mathbb{Q}(\sqrt{m}, \zeta_4)$ is $R[\zeta_4]$ and $R[\zeta_4]^\times = \langle \zeta_4 \rangle \langle u \rangle$.*

(2) *The ring of integers of $\mathbb{Q}(\sqrt{m}, \zeta_6)$ is $R[\zeta_6]$ and $R[\zeta_6]^\times = \langle \zeta_6 \rangle \langle u \rangle$.*

Proof. (a) Let S denote the ring of integers in $K = \mathbb{Q}(\sqrt{m}, \zeta_4)$. Then by [Marcus 1977, Ex. 42, p. 51] we have $S = R[\zeta_4]$. Now let $\alpha = \omega_1 + \omega_2 \zeta_4 \in S$. We compute

$$N_{K/\mathbb{Q}}(\alpha) = N_{F/\mathbb{Q}}(\omega_1)^2 + (\omega_1^\sigma \omega_2)^2 + (\omega_1 \omega_2^\sigma)^2 + N_{F/\mathbb{Q}}(\omega_2)^2$$

where σ is the nontrivial element of $\mathrm{Gal}(F/\mathbb{Q})$. We deduce that α is a unit if and only if either $\omega_1 = 0$ and $N_{F/\mathbb{Q}}(\omega_2) = \pm 1$ or $\omega_2 = 0$ and $N_{F/\mathbb{Q}}(\omega_1) = \pm 1$. The result now follows.

(b) Let S denote the ring of integers in $K = \mathbb{Q}(\sqrt{m}, \zeta_6)$. Then by [Marcus 1977, Ex. 42, p. 51] we have $S = R[\zeta_6]$, since $3 \nmid m$ as $\mathbb{Q}(\sqrt{m})$ has narrow class number one. Let $\alpha = a + b\theta + c\zeta_6 + d\theta\zeta_6 \in S$, where $\theta = \frac{1}{2}(1 + \sqrt{m})$. We compute

$$16 N_{K/\mathbb{Q}}(\alpha) = N_{F/\mathbb{Q}}(\omega_1)^2 + 3((\omega_1\omega_2^\sigma)^2 + (\omega_1^\sigma\omega_2)^2) + 9 N_{F/\mathbb{Q}}(\omega_2)^2,$$

where $\omega_1 = 2a + c + (2b + d)\theta$, $\omega_2 = c + d\theta$ and σ is the nontrivial element of $\mathrm{Gal}(F/\mathbb{Q})$. Assume that $\alpha \in S^\times$. Then we have $N_{F/\mathbb{Q}}(\omega_2) = 0$ or ± 1. If $N_{F/\mathbb{Q}}(\omega_2) = 0$ then $\alpha \in R^\times$. Now assume that $N_{F/\mathbb{Q}}(\omega_2) = \pm 1$. In this case we must have $N_{F/\mathbb{Q}}(\omega_1) = \pm 1$ since $N_{F/\mathbb{Q}}(\omega_1) \equiv N_{F/\mathbb{Q}}(\omega_2) \bmod 2$ rules out the possibility that $N_{F/\mathbb{Q}}(\omega_1) = \pm 2$. So we can write $\omega_1 = \pm u^r$ and $\omega_2 = \pm u^s$. Now α is a unit if and only if

$$16 = 1 + 3\big((\omega_1\omega_2^\sigma)^2 + (\omega_1^\sigma\omega_2)^2\big) + 9,$$

that is, if and only if $2 = u^{2(r-s)} + u^{-2(r-s)}$. This is true if and only if $r = s$. We deduce that if $N_{F/\mathbb{Q}}(\omega_2) = \pm 1$, then α is a unit in S if and only if

$$\alpha = \frac{\omega_1 - \omega_2}{2} + \omega_2\zeta_6 = u^r \zeta_6^k$$

with $k \in \{1, 2, 4, 5\}$. The result follows. □

Lemma 6.5. *The set of orders C consists of the rings of integers $\mathcal{S}$ of the extensions $F(a_\tau)$ where a_τ is a chosen root of a contributing polynomial f_τ as determined by Lemma 6.3.*

Proof. Only condition 5(b) needs to be verified. Our computations show that all of the roots a_τ of f_τ are roots of unity and $N_{F(a_\tau)/F}(a_\tau) = 1$. Thus $s_\mathfrak{p} = 0$ for every $\mathfrak{p}$ and $a_\tau \in \mathcal{S}_{a_\tau,\mathfrak{p}}$ is always satisfied. □

Hence, equation (2) becomes

$$T = M + \frac{1}{2} \sum_{\mathcal{S}_{a_\tau} \in C} \frac{h(\mathcal{S}_{a_\tau})}{w(\mathcal{S}_{a_\tau})}.$$

We now study the contributions from the biquadratic fields $\mathbb{Q}(\sqrt{m}, \sqrt{-1})$ and $\mathbb{Q}(\sqrt{m}, \sqrt{-3})$ to this sum. For this we need the following result of Hasse [1952].

Proposition 6.6. *Let m_1, m_2 be negative squarefree integers and set $m_0 = m_1 m_2$. For each i we set $F_i = \mathbb{Q}(\sqrt{m_i})$, w_i the number of roots of unity in F_i, h_i the order of the class group of F_i. Let $K = \mathbb{Q}(\sqrt{m_1}, \sqrt{m_2})$, h the order of the class group of K, w the number of roots of unity in K and u the fundamental unit in K. Let u_0 be the fundamental unit of F_0. Then*

$$h = \frac{w}{w_1 w_2} h_0 h_1 h_2 \frac{\log u_0}{\log |u|}.$$

From this proposition we get:

(1) For $\mathbb{Q}(\sqrt{m}, \sqrt{-1})$: Let $m_1 = -1$, $m_2 = -m$, $m_0 = m$, $K = \mathbb{Q}(\sqrt{-1}, \sqrt{-m})$. Hence $h_0 = 1$, by hypothesis. It is well known that the class group order of $\mathbb{Q}(\sqrt{-1})$ is 1, and the only roots of unity are powers of $\sqrt{-1}$, i.e., $h_1 = 1$, $w_1 = 4$. Also, the only roots of unity in $\mathbb{Q}(\sqrt{-m})$, with $m \neq 1, 3$, are ± 1, so $w_2 = 2$. Then $w = 4$ and $u_0 = u$. Thus we obtain $h = \frac{1}{2}h(\sqrt{-m})$.

(2) For $\mathbb{Q}(\sqrt{m}, \sqrt{-3})$: Let $m_1 = -3$, $m_2 = -3m$, $m_0 = 9m$, $K = \mathbb{Q}(\sqrt{-3}, \sqrt{-3m})$. Similarly, it is know that the class group order of $\mathbb{Q}(\sqrt{-3})$ is 1, and the only roots of unity are powers of ζ_6, i.e., $h_1 = 1$, $w_1 = 6$. Then, $w = 6$ and $u_0 = u$. Again $w_2 = 2$ and we obtain $h = \frac{1}{2}h(\sqrt{-3m})$.

Next, $[\mathcal{G}^\times : U] = 2$ and 3, respectively, for $\mathbb{Q}(\sqrt{m}, \sqrt{-1})$ and $\mathbb{Q}(\sqrt{m}, \sqrt{-3})$. We can now finish proving Theorem 6.2. The field $\mathbb{Q}(\sqrt{m}, \sqrt{-3})$ contributes twice in the sum (for $\tau = 1, -1$), so equation (2) becomes

$$\begin{aligned} T &= M + \frac{1}{2}\left(\frac{h\big(\mathbb{Q}(\sqrt{m}, \sqrt{-1})\big)}{2} + 2\frac{h\big(\mathbb{Q}(\sqrt{m}, \sqrt{-3})\big)}{3}\right) \\ &= M + \tfrac{1}{8}h\big(\mathbb{Q}(\sqrt{-m})\big) + \tfrac{1}{6}h\big(\mathbb{Q}(\sqrt{-3m})\big) \end{aligned}$$

and this completes the proof of Theorem 6.2. □

We can also determine H. Following the proof of Theorem 6.1 given in [Pizer 1973], we see that

$$T_{q_1 q_2} = \frac{1}{2^e h(F)}\left(H_{q_1 q_2} + \frac{1}{2}\sum_{\mathcal{G}_a \in C_2} E_{q_1 q_2}(\mathcal{G}_a)\frac{h(\mathcal{G}_a)}{w(\mathcal{G}_a)}\right), \tag{3}$$

where $C_2 = C - C_1$ and $C_1 = \{\mathcal{G}_a \in C \mid (N(a)) = (1)\}$. That is, a is a root of $f_{\mu,\varrho,\tau}(x)$ with $(n_\mu) = (1)$. From this we have:

Proposition 6.7. *Let m be a positive squarefree integer*, $F = \mathbf{Q}(\sqrt{m})$, *with* $h(F) = 1$, *and* $\boldsymbol{B}$ *the unique quaternion algebra with* $\mathrm{Ram}(\boldsymbol{B}) = \{\infty_1, \infty_2\}$. *Then* $H = T$. *Consequently, if* $I_1, \dots, I_H$ *is a complete set of representatives of distinct left* $\mathbb{O}$-*ideal classes for a fixed maximal order* $\mathbb{O}$, *then the corresponding right orders* $\mathbb{O}_r(I_1), \dots, \mathbb{O}_r(I_H)$ *form a complete set of distinct representatives of maximal orders of different types.*

Proof. We have $h(F) = 1$, $q_1 = q_2 = 1$, $2^e = 1$ and $n_\mu = n_1 = 1$ in the algorithm to find C. Thus $C_2 = \varnothing$. Substitute these in (3) to get the result. □

7. The algebra $\boldsymbol{B}$ and a maximal order $\mathbb{O}$

In this section we obtain defining relations for $\boldsymbol{B}$, the positive definite quaternion algebra over $F = \mathbb{Q}(\sqrt{m})$ that is ramified precisely at the infinite places of F. We also find a basis over R for a maximal order $\mathbb{O}$ in $\boldsymbol{B}$ when $m \equiv 5 \bmod 8$.

Definition 7.1. Over a field K of characteristic not equal to two, let (a, b) for $a, b \in K^\times$ denote the quaternion algebra over K with basis $\{1, i, j, k\}$ and relations $k = ij$, $i^2 = a$, $j^2 = b$ and $ij = -ji$.

Proposition 7.2. *Let $m \not\equiv 1 \bmod 8$ be a positive squarefree integer. Then $\boldsymbol{B} = (-1, -1)$ is the unique quaternion algebra defined over $\mathbb{Q}(\sqrt{m})$ that is ramified precisely at the infinite places of $\mathbb{Q}(\sqrt{m})$.*

Proof. It is clear that $\boldsymbol{B} = (-1, -1)$ is positive definite. We shall show that at every finite prime $\mathfrak{p}$ of F the algebra $\boldsymbol{B}_\mathfrak{p} = \boldsymbol{B} \otimes_F F_\mathfrak{p}$ is the matrix algebra. Let $\boldsymbol{B}'$ be the quaternion algebra over $\mathbb{Q}$ given by $\boldsymbol{B}' = (-1, -1)$. Then $\boldsymbol{B} = \boldsymbol{B}' \otimes_\mathbb{Q} F$. As is well known, $\mathrm{Ram}(\boldsymbol{B}') = \{2, \infty\}$. Hence $\boldsymbol{B}$ is split at every prime $\mathfrak{p}$ of F not lying above 2. Since $m \not\equiv 1 \bmod 8$ there is only one prime in F above 2. But now $\mathrm{Ram}(\boldsymbol{B})$ has even cardinality and contains the two infinite places of F and hence $\boldsymbol{B}$ must be unramified at the prime of F above 2. $\square$

In the case that $F = \mathbb{Q}(\sqrt{m})$ has narrow class number 1 and $m \equiv 1 \bmod 8$ one can take $\boldsymbol{B}'$ to be the quaternion algebra over $\mathbb{Q}$ ramified precisely at $\{m, \infty\}$. By [Pizer 1980a, Proposition 5.1], one has $\boldsymbol{B}' = (-m, -q)$ where q is a prime with $q \equiv 3 \bmod 4$ and $\left(\frac{m}{q}\right) = -1$. The same argument as above shows that $\boldsymbol{B} = \boldsymbol{B}' \otimes_\mathbb{Q} F$. We now give a maximal order $\mathbb{O}$ in $\boldsymbol{B}$ when $m \equiv 5 \bmod 8$.

Proposition 7.3. *Let $m \equiv 5 \bmod 8$ be a positive squarefree integer. Let $F = \mathbb{Q}(\sqrt{m})$ with ring of integers $R = \mathbb{Z}[\theta]$, where $\theta = \frac{1}{2}(1 + \sqrt{m})$. Let $\boldsymbol{B} = (-1, -1)$. Then $\mathbb{O} = R\left[\delta_1, \delta_2, j, k\right]$ is a maximal order in $\boldsymbol{B}$, where $\delta_1 = \frac{1}{2}(1 + i + j + k)$ and $\delta_2 = \frac{1}{2}(i + \theta j + (1 + \theta)k)$.*

Proof. It is clear that $\mathbb{O}$ is a full lattice in $\boldsymbol{B}$. It is simple, but tedious, to check that $\mathbb{O}$ is a ring and that every element of $\mathbb{O}$ is integral. Finally one can check that $\mathbb{O}$ is maximal by computing its discriminant. For all the details see [Socrates 1993, Theorem 4.2]. $\square$

8. Cusp form calculations

We now compute the space of cusp forms for the field $F = \mathbb{Q}(\sqrt{509})$. From Theorem 6.2 and Proposition 6.7 we compute that the class number for $\boldsymbol{B}$ is 24. We will give representatives for each of the 24 ideal classes, which will then enable us to compute the necessary Brandt matrices using the algorithm from Section 3.

In the algorithm of Section 5, we first find suitable α. The α that eventually led us to distinct ideal classes were i together with

$$\alpha_1 = \tfrac{1}{2} + 5i + \tfrac{1}{2}(1+\theta)j + \bigl(1 - \tfrac{1}{2}\theta\bigr)k = \delta_1 + 9\delta_2 - 4\theta j - (4+5\theta)k$$
$$\bigl(\mathrm{nr}(\alpha_1) = 90, \quad k_1 = -359, \quad h(\mathbb{Q}(\sqrt{-359})) = 19\bigr)$$

and

$$\alpha_2 = \tfrac{1}{2} + (4 - \tfrac{1}{2}\theta)i + 2j + \tfrac{1}{2}(7+\theta)k = \delta_1 + (7-\theta)\delta_2 + (65-3\theta)j + (63-2\theta)k$$
$$\bigl(\mathrm{nr}(\alpha_2) = 96, \quad k_2 = -383, \quad h(\mathbb{Q}(\sqrt{-383})) = 17\bigr).$$

Let $K = F(\alpha_i)$. Note that $R\bigl[\sqrt{k_i}\,\bigr]$ has index 2 in the ring of integers of K. We set $\alpha_i' = 2\alpha_i - 1$, which satisfies $x^2 - k_i = 0$. Since F has class number one, we will be interested only in prime ideals of F that split in K. If $x^2 - k_i$ splits into two distinct factors $(x-\beta_1)(x-\beta_2)$ modulo the prime ideal $\mathfrak{p} = (a+b\theta)$ of F, then as an ideal in K

$$\mathfrak{p} = (a+b\theta, \alpha_i' - \beta_1)\,(a+b\theta, \alpha_i' - \beta_2)$$

and it suffices to consider only one of the ideals I on the right, as they belong to the same K-ideal class. Moreover we have $\mathrm{nr}(\mathbb{O}I) = (a+b\theta)$.

Since the class number of $\mathbb{O}$ is rather large, we first used the Θ-series of $\mathbb{O}I$ for various prime ideals I in the extensions $K = F(\alpha_i)$ above. We computed the Θ-series of these ideals up to $30+2\theta$. Using this method we found 23 of the 24 ideal classes. These ideals, together with the initial coefficients of their Θ-series, are listed in the tables below.

After a lengthy search that did not yield another ideal with a distinct Θ-series, we switched to using the necessary and sufficient conditions of Proposition 5.2. Let I be an ideal in S, the ring of integers in some $F(\alpha_i)$. Assume that the initial coefficients of the Θ-series of $\mathbb{O}I$ are the same as those of one of the left ideals above, say J_s. Construct a basis for $I' = I^{-1}J_s$ and construct

$$\mathcal{N}_{I'}(\alpha) = \Psi_1(X) + \Psi_2(X)\theta,$$

with Ψ_1 in Hermite normal form. Proposition 5.2 then says that $\mathbb{O}I$ is actually in a *different* class as J_s if and only if $a_{1,1}$, the leading term of Ψ_1, is greater than 1. (Note that $1+b\theta$ is totally positive if and only if $b=0$). Using this condition, we quickly determined that we could take $J_{24} = \mathbb{O}I_{24}$ with

$$I_{24} = \bigl(46+5\theta,\ 334-10i-(1+\theta)j+(-2+\theta)k\bigr)$$

a prime ideal in $F(\alpha_1)$ dividing 829.

Now that we have concrete representatives of left ideal classes, we are able to construct explicitly the first few Brandt matrices $B(\xi)$ and the modified Brandt matrices $B'(\xi)$ using the algorithm [Cohen 1993, Algorithm 2.7.7] mentioned at the end of Section 3. This involves computing the Θ-series of the 300 ideals $J_r^{-1}J_s$, $r \geq s$, due to the symmetry properties in Theorem 4.1. We also computed the characteristic polynomials of the $B'(\xi)$ and factored them over $\mathbb{Q}$. We found that the characteristic polynomial of $B'(19+\theta)$ has three distinct rational roots and an irreducible factor of degree 20. Hence, although $\mathbb{C}^{23}$ has a basis of eigenvectors for all the $B'(\xi)$, only three eigenvectors have eigenvalues that are all rational. The three rational eigenvectors are

$$\boldsymbol{v}_1 = (0, 0, 0, 0, 1, 0, -2, -1, 1, 1, 0, -2, 0, 0, -3, 1, 0, 0, 0, -1, 2, 0, 2),$$

$$\boldsymbol{v}_2 = (0, 0, 0, 0, -1, 0, 2, 1, -1, 1, 0, 2, 0, 0, -2, -1, 0, 0, 0, 1, -2, 0, 3),$$

$$\boldsymbol{v}_3 = (45, 45, 25, 60, 23, 40, 34, 27, 18, 28, 30, 19, 35, 20, 31, 28, 20, 15, 25, 37, 51, 40, 31).$$

We let $\boldsymbol{f}_1$, $\boldsymbol{f}_2$ and $\boldsymbol{f}_3$ denote the forms corresponding to the vectors $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$ by Proposition 4.3. The initial Fourier coefficients of these forms are tabulated in Table 3. From this table we note that $\boldsymbol{f}_1 = \boldsymbol{f}_2^{\sigma}$, where σ is the nontrivial element of $\mathrm{Gal}(F/\mathbb{Q})$, while $\boldsymbol{f}_3 = \boldsymbol{f}_3^{\sigma}$ and hence $\boldsymbol{f}_3$ is the base change of a classical form. That none of these forms are CM forms follows from the following proposition.

Proposition 8.1. *Let $\boldsymbol{f}$ be a Hilbert eigenform of full level for a totally real number field F of narrow class number one. Then $\boldsymbol{f}$ is not a CM form.*

Proof. Recall that $\boldsymbol{f}$ is a CM form if and only if there exists a quadratic character ε corresponding to an imaginary quadratic extension K/F such that $\boldsymbol{f} = \boldsymbol{f} \otimes \varepsilon$. So suppose we have $\boldsymbol{f} = \boldsymbol{f} \otimes \varepsilon$ for such a character ε. Let π denote the cuspidal representation of $\mathrm{GL}_2(\boldsymbol{A}_F)$ corresponding to π. Then we have $\pi \cong \pi \otimes (\varepsilon \circ \det)$. By a theorem of Labesse and Langlands [1979] we have an equality of L-series $L(\pi, s) = L(\chi, s)$ for some grössencharakter χ of K, and it is known that $\operatorname{cond} \pi = N_{K/F}(\operatorname{cond} \chi) \operatorname{disc}(K/F)$. Since π is assumed to be unramified it follows that K/F is an unramified extension. But this is impossible since F has narrow class number one. $\square$

9. The elliptic curves

In this section we give equations for the elliptic curves that we will show are attached to the forms $\boldsymbol{f}_1$, $\boldsymbol{f}_2$ and $\boldsymbol{f}_3$ of the previous section.

Let E_3 be the elliptic curve given by the Weierstrass equation

$$y^2 + (1+\theta)xy + (1+\theta)y = x^3 + (-4051846 + 343985\theta)x + 4312534180 - 366073300\theta.$$

This curve is found in [Cremona 1992] and is a $\mathbb{Q}$-curve (that is, it is isogenous to its Galois conjugate). Let E_1 denote the elliptic curve given by the Weierstrass equation

$$y^2 - xy - \theta y = x^3 + (2+2\theta)x^2 + (162+3\theta)x + 71 + 34\theta.$$

This elliptic curve is in a table found in [Pinch 1982], among other curves that have good reduction everywhere over certain quadratic fields. We show below that E_1 is not F-isogenous to its Galois conjugate. This is also noted (without proof) in [Cremona 1992]. We take E_2 to be the curve E_1^σ, where σ is the nontrivial element of $\mathrm{Gal}(F/\mathbb{Q})$.

Proposition 9.1. *The elliptic curve E_1 is not isogenous over F to its Galois conjugate.*

I_i	K	$a_i + b_i\theta$	γ_i	$I_i \mid p \in \mathbb{Z}$
I_1	F	1		
I_2	$F(\alpha_1)$	61	$-23+46\theta-10i-(1+\theta)j+(-2+\theta)k$	61
I_3	$F(\alpha_1)$	$45+4\theta$	$81-10i-(1+\theta)j+(-2+\theta)k$	173
I_4	$F(\alpha_1)$	149	$45-10i-(1+\theta)j+(-2+\theta)k$	149
I_5	$F(\alpha_1)$	$53+5\theta$	$34-10i-(1+\theta)j+(-2+\theta)k$	101
I_6	$F(\alpha_1)$	79	$6-10i-(1+\theta)j+(-2+\theta)k$	79
I_7	$F(\alpha_1)$	53	$-22+44\theta-10i-(1+\theta)j+(-2+\theta)k$	53
I_8	$F(\alpha_2)$	$23+2\theta$	$32+(-8+\theta)i-4j-(7+\theta)k$	67
I_9	$F(\alpha_1)$	$9+\theta$	$14-10i-(1+\theta)j+(-2+\theta)k$	37
I_{10}	$F(\alpha_1)$	$10+\theta$	$7-10i-(1+\theta)j+(-2+\theta)k$	17
I_{11}	$F(\alpha_1)$	$184+17\theta$	$22-10i-(1+\theta)j+(-2+\theta)k$	281
I_{12}	$F(\alpha_1)$	$107+10\theta$	$33-10i-(1+\theta)j+(-2+\theta)k$	181
I_{13}	$F(\alpha_2)$	47	$-18+36\theta+(-8+\theta)i-4j-(7+\theta)k$	47
I_{14}	$F(\alpha_1)$	31	$-1+2\theta-10i-(1+\theta)j+(-2+\theta)k$	31
I_{15}	$F(\alpha_1)$	$32+3\theta$	$3-10i-(1+\theta)j+(-2+\theta)k$	23
I_{16}	$F(\alpha_1)$	131	$54-10i-(1+\theta)j+(-2+\theta)k$	131
I_{17}	$F(\alpha_1)$	59	$-14+28\theta-10i-(1+\theta)j+(-2+\theta)k$	59
I_{18}	$F(\alpha_2)$	61	$-26+52\theta+(-8+\theta)i-4j-(7+\theta)k$	61
I_{19}	$F(i)$	$31+3\theta$	$34+i$	89
I_{20}	$F(\alpha_1)$	$75+7\theta$	$15-10i-(1+\theta)j+(-2+\theta)k$	73
I_{21}	$F(\alpha_1)$	13	$-3+6\theta-10i-(1+\theta)j+(-2+\theta)k$	13
I_{22}	$F(\alpha_1)$	157	$-6+12\theta-10i-(1+\theta)j+(-2+\theta)k$	157
I_{23}	$F(i)$	$11+\theta$	$2+i$	5

Table 1. Prime ideals $I_i = (a_i + b_i\theta, \gamma_i)$, where the $\mathbb{O}I_i$ have distinct Θ-series.

Proof. If E_1 and E_1^σ are isogenous, the local factors of the L-series of E_1 and E_1^σ will be the same for all primes of F. Let $\mathfrak{p} = (5, 1 + 2\theta)$ denote one of the prime ideals of F above 5. We have an isomorphism of $R/\mathfrak{p}$ with $\mathbb{Z}/5$ that maps θ to

	1	2	3	4	5	6	7	8	9	10	11	$11+\theta$	$12-\theta$	12
J_1	24	24	96	24	144	96	192	24	312	144	288	0	0	96
J_2	0	24	0	24	0	96	0	24	0	144	0	0	0	96
J_3	0	0	24	0	0	24	0	0	96	0	0	0	0	24
J_4, J_5	0	0	0	24	0	0	0	24	0	0	0	0	0	96
J_6, J_7, J_8	0	0	0	0	24	0	0	0	0	24	0	0	0	0
J_9	0	0	0	0	0	24	0	0	0	0	24	0	0	24
J_{10}	0	0	0	0	0	24	0	0	0	0	0	0	0	48
J_{11}	0	0	0	0	0	0	24	0	0	24	0	0	0	24
J_{12}	0	0	0	0	0	0	24	0	0	24	0	0	0	0
J_{13}, J_{14}	0	0	0	0	0	0	24	0	0	0	48	0	0	0
J_{15}	0	0	0	0	0	0	0	24	24	0	0	0	0	48
J_{16}	0	0	0	0	0	0	0	24	0	24	24	0	0	24
J_{17}	0	0	0	0	0	0	0	24	0	24	48	0	0	0
J_{18}	0	0	0	0	0	0	0	24	0	48	0	0	0	0
J_{19}	0	0	0	0	0	0	0	48	0	0	0	0	0	0
J_{20}	0	0	0	0	0	0	0	0	24	48	48	0	0	0
J_{21}	0	0	0	0	0	0	0	0	48	24	0	0	0	24
J_{22}	0	0	0	0	0	0	0	0	48	0	48	0	0	0
J_{23}	0	0	0	0	0	0	0	0	0	48	48	24	24	48

	$12+\theta$	$13-\theta$	13	$13+\theta$	$14-\theta$	14	$14+\theta$	$15-\theta$	15	$15+\theta$
J_4, J_5	0	0	0	0	0	0	0	0	0	0
J_6	0	0	24	0	0	0	0	0	96	0
J_7	0	0	0	0	0	24	0	0	144	24
J_8	0	0	0	0	0	48	0	0	96	0
J_{13}	0	0	48	0	0	24	48	48	0	0
J_{14}	0	0	24	24	24	48	0	0	24	24

	$16-\theta$	16	$16+\theta$	$17-\theta$	17	$17+\theta$	$18-\theta$	18	$18+\theta$
J_4	0	24	0	0	48	0	0	24	24
J_5	0	48	0	0	0	0	0	0	0

Table 2. Beginning coefficients c_{ξ, J_i} of the Θ-series of J_1 to J_{23}. More are given for ideals whose early coefficients agree.

ξ	3	7	11+θ	12−θ	12+θ	13−θ	13	14+θ	15−θ	15+θ	16−θ
$\xi \mid p$	3	7	5	5	29	29	13	83	83	113	113
$\boldsymbol{v}_1$	−4	−6	3	−2	0	10	1	14	9	11	6
$\boldsymbol{v}_2$	−4	−6	−2	3	10	0	1	9	14	6	11
$\boldsymbol{v}_3$	1	9	−2	−2	−5	−5	26	14	14	11	11

Table 3. Eigenvalues for simultaneous rational eigenvectors for $B'(\xi)$.

2 mod 5. Then the equation for the reduced curve $\widetilde{E}_1$ over $\mathbb{Z}/5$ has affine equation

$$\widetilde{E}_1 : y^2 + 4xy + 3y = x^3 + x^2 + 3x + 4.$$

and we compute that $\widetilde{E}_1(R/\mathfrak{p})$ has order 8. Similarly, the reduction of the curve E_1^σ has equation

$$\widetilde{E}_1^\sigma : y^2 + 4xy + y = x^3 + 4x + 2.$$

and we compute that $\widetilde{E}_1^\sigma(R/\mathfrak{p})$ has order 3. Therefore E_1 is not isogenous to E_1^σ. □

Finally we check that our curves E_1, E_2 and E_3 do not possess potential complex multiplication. We first remark that $h^+(F) = 1$. Our conclusion about these curves now follows from:

Proposition 9.2. *Let K be a totally real number field of narrow class number one. Let E/K be an elliptic curve that has good reduction everywhere. Then E does not possess potential complex multiplication.*

Proof. Suppose $E(\mathbb{C})$ has CM defined over the field $\mathbb{Q}(\sqrt{n})$, where $n < 0$. Consider the field $L = K(\sqrt{n})$. Then E and its complex multiplications are defined over L. Consider the ℓ-adic representation given by the action of Galois on the ℓ-adic Tate module of E/L

$$\sigma_\ell : \mathrm{Gal}(\overline{\mathbb{Q}}/L) \to \mathrm{GL}_2(\mathbb{Q}_\ell).$$

We construct another representation

$$\begin{aligned} \sigma_\ell^{[\rho]} : \mathrm{Gal}(\overline{\mathbb{Q}}/L) &\to \mathrm{GL}_2(\mathbb{Q}_\ell), \\ \tau &\mapsto \sigma_\ell(\rho\tau\rho^{-1}), \end{aligned}$$

where $\rho \in \mathrm{Gal}(\overline{\mathbb{Q}}/K)$ is nontrivial when restricted to L. Now, since E is actually defined over K, this σ_ℓ extends to a representation $\tilde{\sigma}_\ell$ of $\mathrm{Gal}(\overline{\mathbb{Q}}/K)$. However,

$$\tilde{\sigma}_\ell(\rho\tau\rho^{-1}) = \tilde{\sigma}_\ell(\rho)\tilde{\sigma}_\ell(\tau)\tilde{\sigma}_\ell(\rho)^{-1} = \tilde{\sigma}_\ell(\rho)\sigma_\ell(\tau)\tilde{\sigma}_\ell(\rho)^{-1},$$

and hence $\sigma_\ell^{[\rho]} \cong \sigma_\ell$.

Since E has CM over L, the representation σ_ℓ is abelian, so $\sigma_\ell = \chi_\ell \oplus \chi'_\ell$ for some characters χ_ℓ, χ'_ℓ of H. It can easily be seen from such a decomposition that, in the obvious notation,

$$\sigma_\ell^{[\rho]} = \chi_\ell^{[\rho]} \oplus \chi_\ell'^{[\rho]}$$

as well. Now, χ_ℓ corresponds to a weight 1 grössencharakter ψ of L, and $\chi_\ell = \chi_\ell^{[\rho]}$ if and only if $\psi(z) = \psi(\bar{z})$ for all $z \in L_\infty^* = \mathbb{C}^*$. But $\psi(z) = z^{-1}$ and $\psi(\bar{z}) = \bar{z}^{-1}$, hence $\psi(z) \neq \psi(\bar{z})$, so $\chi_\ell \neq \chi_\ell^{[\rho]}$. Thus $\chi'_\ell = \chi_\ell^{[\rho]}$, and so $\sigma_\ell = \chi_\ell \oplus \chi_\ell^{[\rho]}$, hence $\tilde{\sigma}_\ell = \mathrm{Ind}_H^G(\chi_\ell)$. Since the degree of χ_ℓ is 1, we get the formula

$$\operatorname{cond} \tilde{\sigma}_\ell = N_{L/K}(\operatorname{cond} \chi_\ell) \operatorname{disc}(L/K)$$

for the conductor of $\tilde{\sigma}_\ell$; see [Martinet 1977]. Recall that E has good reduction everywhere, so every $\tilde{\sigma}_\ell$ is unramified at all the primes of K not dividing ℓ. Since $\tilde{\sigma}_\ell$ is ramified at all the primes which divide $\operatorname{cond}(\tilde{\sigma}_\ell)$, we see that $\operatorname{disc}(L/K)$ must be the unit ideal. Thus L is an unramified finite abelian extension of K. But since $h^+(K) = 1$ this implies that $K = L$ which is impossible since $n < 0$. □

10. Matching the elliptic curves to the cusp forms

Continuing with the notation of the previous section we have $F = \mathbb{Q}(\sqrt{509})$, R the ring of integers in F and $\theta = \frac{1}{2}(1 + \sqrt{509})$.

We begin by showing that the curve E_3 is attached to the form $\boldsymbol{f}_3$. The curve E_3 is equal to the curve A' that arises from Shimura's construction [1971, 7.7]. This curve is constructed from a pair of eigenforms $\{f_1, f_2\}$ in $S_2(\Gamma_0(509), \chi)$ where χ is the quadratic character of $(\mathbb{Z}/509\mathbb{Z})^\times$. These forms are constructed in [Cremona 1992]. Furthermore we know that

$$L(E_3, s) = L(f_1, s) L(f_2, s).$$

The base change of f_1 to $\mathrm{GL}_2(F)$ will be a form with rational coefficients of full level, trivial character and weight 2. Hence we see that $\boldsymbol{f}_3$ is the base change of the form f_1 and we have

$$L(E_3, s) = L(\boldsymbol{f}_3, s).$$

Let E_1 be as in Section 9. Since E_1 has good reduction everywhere, the 2-adic representation on the Tate module of E_1,

$$\sigma_1 : \mathrm{Gal}(\bar{F}/F) \to \mathrm{GL}_2(\mathbb{Q}_2),$$

is unramified outside the prime ideal $2R$ of F. For each prime ideal $\mathfrak{p}$ of F outside $2R$ we have

$$\operatorname{Tr} \sigma_1(\mathrm{Fr}_\mathfrak{p}) = a(E_1)_\mathfrak{p},$$

where $\mathrm{Fr}_{\mathfrak{p}}$ denotes a Frobenius element at $\mathfrak{p}$ and $a(E_1)_{\mathfrak{p}}$ denotes the $\mathfrak{p}$-th Fourier coefficient of E_1. Moreover $\det \sigma_1(\mathrm{Fr}_{\mathfrak{p}}) = N\mathfrak{p}$.

Let $\boldsymbol{f}_1$ denote the unramified cusp form given in Section 8 above. By [Taylor 1989] and [Blasius and Rogawski 1993] there exists a 2-dimensional representation

$$\sigma_2 : \mathrm{Gal}(\bar{F}/F) \to \mathrm{GL}_2(\mathbb{Q}_2)$$

unramified outside the prime ideal $2R$ of F and such that for each prime ideal $\mathfrak{p}$ of F outside $2R$ we have

$$\mathrm{Tr}\, \sigma_2(\mathrm{Fr}_{\mathfrak{p}}) = a(\boldsymbol{f}_1)_{\mathfrak{p}}$$

where again $\mathrm{Fr}_{\mathfrak{p}}$ denotes a Frobenius element at $\mathfrak{p}$ and $a(\boldsymbol{f}_1)_{\mathfrak{p}}$ denotes the $\mathfrak{p}$-th Fourier coefficient of $\boldsymbol{f}_1$. Moreover we have $\det \sigma_2(\mathrm{Fr}_{\mathfrak{p}}) = N\mathfrak{p}$.

To prove that E_1 is attached to the form $\boldsymbol{f}_1$ we must show that the representations σ_1 and σ_2 are equivalent. For this we will use the following result of Faltings and Serre as stated and proved in [Livné 1987].

Theorem 10.1. *Let K be a global field, S a finite set of primes of K, and E a finite extension of $\mathbb{Q}_2$. Denote the maximal ideal in the ring of integers of E by $\mathfrak{p}$ and the compositum of all quadratic extensions of K unramified outside S by K_S. Suppose*

$$\rho_1, \rho_2 : \mathrm{Gal}(\overline{K}/K) \to \mathrm{GL}_2(E)$$

are continuous representations, unramified outside S, and furthermore satisfying:

1. $\mathrm{Tr}\, \rho_1 \equiv \mathrm{Tr}\, \rho_2 \equiv 0 \bmod \mathfrak{p}$ *and* $\det \rho_1 \equiv \det \rho_2 \bmod \mathfrak{p}$.
2. *There exists a set T of primes of K, disjoint from S, for which*
   - *the image of the set $\{\mathrm{Fr}_t : t \in T\}$ in the $\mathbb{Z}/2\mathbb{Z}$-vector space $\mathrm{Gal}(K_S/K)$ is noncubic*;
   - $\mathrm{Tr}\, \rho_1(\mathrm{Fr}_t) = \mathrm{Tr}\, \rho_2(\mathrm{Fr}_t)$ *and* $\det \rho_1(\mathrm{Fr}_t) = \det \rho_2(\mathrm{Fr}_t)$ *for all $t \in T$.*

Then ρ_1 and ρ_2 have isomorphic semi-simplifications.

A subset S of the $\mathbb{Z}/2\mathbb{Z}$-vector space $\mathrm{Gal}(K_S/K)$ is said to be noncubic if every homogeneous polynomial of degree three that vanishes on S vanishes on all of $\mathrm{Gal}(K_S/K)$. In particular $\mathrm{Gal}(K_S/K)$ is itself noncubic and we will apply this theorem with T chosen such that the image of $\{\mathrm{Fr}_t : t \in T\}$ in $\mathrm{Gal}(K_S/K)$ is the whole space.

As we can see from Table 3, we cannot apply this result immediately since the traces of Frobenius are not all even. Therefore for each i we let $\bar{\sigma}_i$ denote the mod 2 representations obtained from σ_i and let L_i denote the extension of F cut out by $\bar{\sigma}_i$. We begin by showing that we can identify these two extensions.

Matching L_1 ***and*** L_2**.** We know that $L_1 = F(E_1[2])$. Hence L_1 is the splitting field of the polynomial

$$g(x) = 4x^3 + (9+8\theta)x^2 + (648+14\theta)x + 411 + 137\theta$$

and is an S_3-extension of F unramified outside of $2R$. Moreover the quadratic extension of F contained in L_1 is $F(\sqrt{u})$, where $u = 442 + 41\theta$ is a fundamental unit of F.

We now consider L_2. We know that L_2 is an extension of F that is unramified outside of $2R$. Moreover since some of the $a(\boldsymbol{f}_1)_{\mathfrak{p}}$'s are odd we know that L_2 is either a normal cubic extension of F or else is an S_3 extension. By the next lemma we deduce that L_2/F must be an S_3 extension.

Lemma 10.2. *There are no normal cubic extensions of F unramified outside of $2R$.*

Proof. Suppose that L/F is such an extension. Let $\mathfrak{f}(L/F)$ denote the conductor of F. By [Cohen 2000, Corollary 3.5.12] we deduce that $\mathfrak{f}(L/F)$ divides $2R$. But now using Pari [Cohen et al. 2004] we compute that the ray class group for the modulus $2R\infty_1\infty_2$, where ∞_i denote the infinite places of F, is trivial. Therefore no such extension L of F exists. □

Let F_1 be the unique quadratic extension of F contained in L_2. We let $u = 442 + 41\theta$ be the fundamental unit of F. Since F_1 is unramified outside 2 we know that F_1 must be one of the fields

$$F(\sqrt{-1}),\ F(\sqrt{u}),\ F(\sqrt{2}),\ F(\sqrt{-u}),\ F(\sqrt{-2}),\ F(\sqrt{2u}) \text{ or } F(\sqrt{-2u}).$$

Let $\mathfrak{p}$ be a prime of F and let $\mathfrak{P}$ be a prime of F_1 above $\mathfrak{p}$. We note that if $a(\boldsymbol{f}_1)_{\mathfrak{p}}$ is odd then $f(\mathfrak{P}/\mathfrak{p}) = 1$. We use this criterion to eliminate all the above quadratic extension of F except for $F(\sqrt{u})$. Taking $\mathfrak{p} = (11+\theta)R$ eliminates the fields $F(\sqrt{2})$, $F(\sqrt{-2})$, $F(\sqrt{2u})$ and $F(\sqrt{-2u})$. While taking $\mathfrak{p} = (15-\theta)R$ eliminates the fields $F(\sqrt{-1})$ and $F(\sqrt{-u})$. Therefore we have $F_1 = F(\sqrt{u})$.

Lemma 10.3. *There is a unique normal cubic extension of F_1 which is unramified outside of $2R_1$, where R_1 denotes the ring of integers in F_1.*

Proof. We note that $2R_1 = \mathfrak{p}^2$, where $\mathfrak{p}$ is the unique prime of F_1 above 2. Suppose that L/F_1 is such an extension. Let $\mathfrak{f}(L/F_1)$ denote the conductor of L/F_1. By [Cohen 2000, Corollary 3.5.12] we deduce that $\mathfrak{f}(L/F_1)$ divides $\mathfrak{p}$. Using Pari we compute that the order of the ray class group for the modulus $\mathfrak{p}\infty_1\infty_2$, where ∞_i denote the real places of F_1, is three, from which we deduce that L is unique. □

Since both L_1 and L_2 contain $F(\sqrt{u})$, we deduce that $L_1 = L_2$.

Application of Faltings and Serre. Let K denote a fixed cubic extension of F contained in $L = L_1 = L_2$. We now apply Theorem 10.1 to the representations $\sigma_1|_K$ and $\sigma_2|_K$. We note that these representations satisfy the conditions of the theorem.

Now $K = F(\alpha)$, where α satisfies the equation

$$\psi_2(x) = 4x^3 + (9+8\theta)x^2 + (648+14\theta)x + 411 + 137\theta$$

over F. Using $\theta^2 - \theta - 127 = 0$ we find that α satisfies the equation

$$m(x) = 64x^6 + 416x^5 - 10940x^4 - 30552x^3 + 550476x^2 + 560056x - 8633740$$

over $\mathbb{Q}$. In fact we can write $K = \mathbb{Q}(\beta)$, where β satisfies the equation

$$x^6 - 25x^4 - 46x^3 + 29x^2 + 66x + 20.$$

Using Pari we find that K has class number one and $\mathbb{O}_K^\times \cong \{\pm 1\} \times \mathbb{Z}^4$ with fundamental units given by

$$\begin{aligned}
u_1 &= \tfrac{1}{34}\beta^5 + \tfrac{3}{17}\beta^4 - \tfrac{23}{34}\beta^3 - \tfrac{92}{17}\beta^2 - \tfrac{293}{34}\beta - \tfrac{47}{17},\\
u_2 &= \tfrac{7}{102}\beta^5 - \tfrac{13}{51}\beta^4 - \tfrac{31}{34}\beta^3 + \tfrac{19}{51}\beta^2 + \tfrac{91}{102}\beta + \tfrac{11}{51},\\
u_3 &= \tfrac{10}{51}\beta^5 - \tfrac{8}{51}\beta^4 - \tfrac{71}{17}\beta^3 - \tfrac{361}{51}\beta^2 + \tfrac{79}{51}\beta + \tfrac{199}{51},\\
u_4 &= \tfrac{106}{51}\beta^5 - \tfrac{44}{51}\beta^4 - \tfrac{875}{17}\beta^3 - \tfrac{3745}{51}\beta^2 + \tfrac{4693}{51}\beta + \tfrac{5047}{51}.
\end{aligned}$$

Now the ideal $2R_K$ factors as $\mathfrak{p}_1\mathfrak{p}_2^2$. A generator for $\mathfrak{p}_1$ is given by

$$a_1 = \tfrac{4}{51}\beta^5 + \tfrac{7}{51}\beta^4 - \tfrac{42}{17}\beta^3 - \tfrac{277}{51}\beta^2 + \tfrac{205}{51}\beta + \tfrac{304}{51}$$

and a generator for $\mathfrak{p}_2$ is given by

$$a_2 = \tfrac{16}{51}\beta^5 - \tfrac{23}{51}\beta^4 - \tfrac{117}{17}\beta^3 - \tfrac{292}{51}\beta^2 + \tfrac{667}{51}\beta + \tfrac{349}{51}.$$

Let K_S denote the compositum of all quadratic extensions of K which are unramified outside of $S = \{\mathfrak{p}_1, \mathfrak{p}_2\}$. Then K_S is the compositum of the fields

$$K(\sqrt{-1}),\ K(\sqrt{u_1}),\ K(\sqrt{u_2}),\ K(\sqrt{u_3}),\ K(\sqrt{u_4}),\ K(\sqrt{a_1}) \text{ and } K(\sqrt{a_2}).$$

Using Pari we can find a set T of primes in K such that

$$\mathrm{Gal}(K_S/K) = \{\mathrm{Fr}_{\mathfrak{P}} \in \mathrm{Gal}(K_S/K) : \mathfrak{P} \in T\}$$

where $\mathrm{Fr}_{\mathfrak{P}}$ denotes the Frobenius element in $\mathrm{Gal}(K_S/K)$ at $\mathfrak{P}$. Let T_0 denote the primes of F generated by the elements of F in the left hand column of Table 4. Then we can take T to be the set of primes in K above those in T_0.

$\xi = a + b\theta$	p	$a(\mathfrak{p})$
3	inert	-4
7	inert	-6
$11+\theta$	5	3
$12-\theta$	5	-2
$12+\theta$	29	0
$13-\theta$	29	10
$14+\theta$	83	14
$15-\theta$	83	9
$15+\theta$	113	11
$16-\theta$	113	6
$17+\theta$	179	0
$18-\theta$	179	25
19	inert	-12
$20+\theta$	293	16
$21-\theta$	293	26
$22+\theta$	379	-20
$23-\theta$	379	20
$23+2\theta$	67	-7
$25-2\theta$	67	8
$25+\theta$	523	36
$26-\theta$	523	11
$25+2\theta$	167	22
$27-2\theta$	167	-8
$29+\theta$	743	44
$30-\theta$	743	-36
31	inert	-18
$32+\theta$	929	40
$33-\theta$	929	10
$33+2\theta$	647	18
$35-2\theta$	647	43
$34+\theta$	1063	4
$35-\theta$	1063	-1
$37+\theta$	1279	-20
$38-\theta$	1279	25
$37+3\theta$	337	-28
$40-3\theta$	337	2
$39+2\theta$	1091	60
$41-2\theta$	1091	0
41	inert	-18
$41+3\theta$	661	-20
$44-3\theta$	661	-10
$45+2\theta$	1607	42
$47-2\theta$	1607	57
$50+\theta$	2423	-24
$51-\theta$	2423	-69
$51+4\theta$	773	-24
$55-4\theta$	773	-4
$54+\theta$	2843	-6
$55-\theta$	2843	-61

$\xi = a + b\theta$	p	$a(\mathfrak{p})$
$54+5\theta$	11	3
$59-5\theta$	11	-2
$55+4\theta$	1213	-46
$59-4\theta$	1213	34
$56+5\theta$	241	2
$61-5\theta$	241	-8
$57+5\theta$	359	-6
$62-5\theta$	359	9
59	inert	-22
$60+\theta$	3533	6
$61-\theta$	3533	-84
$62+\theta$	3779	30
$63-\theta$	3779	0
$62+3\theta$	2887	-73
$65-3\theta$	2887	62
$65+2\theta$	3847	82
$67-2\theta$	3847	32
$66+5\theta$	1511	-8
$71-5\theta$	1511	-13
$67+3\theta$	3547	-68
$70-3\theta$	3547	2
$68+5\theta$	1789	-34
$73-5\theta$	1789	-14
$69+2\theta$	4391	130
$71-2\theta$	4391	75
$71+3\theta$	4111	-35
$74-3\theta$	4111	100
$71+5\theta$	2221	-18
$76-5\theta$	2221	-53
$74+5\theta$	2671	-72
$79-5\theta$	2671	-12
$76+3\theta$	4861	70
$79-3\theta$	4861	-30
79	inert	-32
$79+5\theta$	3461	-2
$84-5\theta$	3461	-57
$79+6\theta$	2143	24
$85-6\theta$	2143	-56
$79+7\theta$	571	-20
$86-7\theta$	571	10
$82+3\theta$	5827	-28
$85-3\theta$	5827	22
$84+\theta$	7013	-6
$85-\theta$	7013	-16
$85+3\theta$	6337	78
$88-3\theta$	6337	48
$85+6\theta$	3163	86
$91-6\theta$	3163	-4
$91+3\theta$	7411	-55
$94-3\theta$	7411	100

$\xi = a + b\theta$	p	$a(\mathfrak{p})$
$92+\theta$	8429	100
$93-\theta$	8429	-110
$95+2\theta$	8707	-28
$97-2\theta$	8707	182
$95+6\theta$	5023	76
$101-6\theta$	5023	86
$95+8\theta$	1657	28
$103-8\theta$	1657	-22
$100+3\theta$	9157	98
$103-3\theta$	9157	73
$101+4\theta$	8573	66
$105-4\theta$	8573	-79
$105+\theta$	11003	116
$106-\theta$	11003	36
$108+5\theta$	9029	-54
$113-5\theta$	9029	96
$109+\theta$	11863	-66
$110-\theta$	11863	24
$109+6\theta$	7963	-16
$115-6\theta$	7963	59
$110+3\theta$	11287	208
$113-3\theta$	11287	178
$111+8\theta$	5081	-30
$119-8\theta$	5081	90
$112+5\theta$	9929	-146
$117-5\theta$	9929	-96
$113+5\theta$	10159	76
$118-5\theta$	10159	56
$114+5\theta$	10391	98
$119-5\theta$	10391	-117
$122+\theta$	14879	0
$123-\theta$	14879	-75
$122+3\theta$	14107	152
$125-3\theta$	14107	32
$124+11\theta$	1373	34
$135-11\theta$	1373	-6
$137+11\theta$	4909	20
$148-11\theta$	4909	40
$139+6\theta$	15583	-156
$145-6\theta$	15583	4
$143+7\theta$	15227	108
$150-7\theta$	15227	-122
$164+7\theta$	21821	-210
$171-7\theta$	21821	-150
$169+11\theta$	15053	-6
$180-11\theta$	15053	-106
$171+10\theta$	18251	-68
$181-10\theta$	18251	-198
$178+5\theta$	29399	56
$183-5\theta$	29399	96

Table 4. $a(\mathfrak{p})$'s for the elliptic curve E_1 and cusp form $\boldsymbol{f}_1$.

For each $\mathfrak{p} \in T_0$ we have computed $a(E_1)_{\mathfrak{p}}$ and $a(\boldsymbol{f}_1)_{\mathfrak{p}}$ and found that they are the same. Hence we deduce that for all $\mathfrak{P} \in T$ we have

$$\operatorname{Tr} \sigma_1(\mathrm{Fr}_{\mathfrak{P}}) = \operatorname{Tr} \sigma_2(\mathrm{Fr}_{\mathfrak{P}}).$$

Thus by Theorem 10.1 σ_1 and σ_2 are isomorphic.

End of Proof. We have proved in the previous subsection that $\sigma_1|_K$ is isomorphic to $\sigma_2|_K$ and therefore that $\sigma_1|_L$ is isomorphic to $\sigma_2|_L$. We note that since E_1 does not possess potential complex multiplication by Proposition 9.2 so $\sigma_1|_L$ and hence $\sigma_2|_L$ are both irreducible. Then by Frobenius reciprocity we know that $\sigma_1|_{F_1}$ is isomorphic to $\sigma_2|_{F_1} \otimes \chi$ for some character χ of $\mathrm{Gal}(\bar{F}/F_1)$ trivial on $\mathrm{Gal}(\bar{F}/L)$. Let $\mathfrak{p} = (11+\theta)R$ then $a(E_1)_{\mathfrak{p}}$ is odd and $\mathfrak{p}$ splits in F_1. Let $\mathfrak{P}$ be a prime of F_1 above $\mathfrak{p}$ and let $\mathrm{Fr}_{\mathfrak{P}}$ be a Frobenius element at $\mathfrak{P}$ in $\mathrm{Gal}(\bar{F}/F_1)$. Then

$$\operatorname{Tr}(\sigma_1|_{F_1}(\mathrm{Fr}_{\mathfrak{P}})) = a(E_1)_{\mathfrak{p}} = \operatorname{Tr}(\sigma_2|_{F_1}(\mathrm{Fr}_{\mathfrak{P}}))$$

and hence $\chi(\mathrm{Fr}_{\mathfrak{P}}) = 1$. But since $\mathfrak{P}$ is inert in L we deduce that χ must be trivial. Therefore we have $\sigma_1|_{F_1} = \sigma_2|_{F_1}$. Now using Frobenius reciprocity again we deduce that σ_1 is isomorphic to $\sigma_2 \otimes \delta$ for some character δ of $\mathrm{Gal}(\bar{F}/F)$ trivial on $\mathrm{Gal}(\bar{F}/F_1)$. If we take $\mathfrak{p} = (12-\theta)R$ then $\mathfrak{p}$ is inert in F_1. Now

$$\operatorname{Tr}(\sigma_1(\mathrm{Fr}_{\mathfrak{p}})) = a(E_1)_{\mathfrak{p}} = \operatorname{Tr}(\sigma_2(\mathrm{Fr}_{\mathfrak{p}}))$$

and hence $\delta(\mathrm{Fr}_{\mathfrak{p}}) = 1$. We deduce that δ is trivial and hence that $\sigma_1 = \sigma_2$.

Thus we conclude that E_1 is attached to the form $\boldsymbol{f}_1$. It immediately follows that the curve E_2 is attached to the form $\boldsymbol{f}_2$. The verification of Conjecture 1.1 for $F = \mathbb{Q}(\sqrt{509})$ is now complete.

Remark. We found after this work was completed that one could use [Skinner and Wiles 1999, Theorem A] to prove that the curve E_1 is modular. Here one uses that the Galois representation on the 5-adic Tate module of E_1 is residually reducible. However, our method can, in principle, be used in situations where their results do not apply. Moreover, our interest in this problem arises from attaching elliptic curves to unramified Hilbert modular forms, for which one needs to be able to determine the space of cusp forms. Furthermore, it appears that our method of computing the space of cusp forms can be extended to higher weight, where eigenforms with rational Hecke eigenvalues should correspond to certain other geometric objects.

Acknowledgments

Both authors thank their advisor, Dinakar Ramakrishnan, for his support and guidance through this work. They also thank Don Blasius for comments on an earlier version of this paper, Barry Mazur for his encouragement and the referee for a thorough report that led to several improvements in the exposition.

References

[Blasius 2004] D. Blasius, "Elliptic curves, Hilbert modular forms, and the Hodge conjecture", pp. 83–103 in *Contributions to automorphic forms, geometry, and number theory*, edited by H. Hida et al., Johns Hopkins Univ. Press, Baltimore, MD, 2004. MR MR2058605

[Blasius and Rogawski 1993] D. Blasius and J. D. Rogawski, "Motives for Hilbert modular forms", *Invent. Math.* **114**:1 (1993), 55–87. MR 94i:11033 Zbl 0829.11028

[Cohen 1993] H. Cohen, *A course in computational algebraic number theory*, Graduate Texts in Mathematics **138**, Springer, Berlin, 1993. MR 94i:11105 Zbl 0786.11071

[Cohen 2000] H. Cohen, *Advanced topics in computational number theory*, Graduate Texts in Mathematics **193**, Springer, New York, 2000. MR 2000k:11144 Zbl 0977.11056

[Cohen et al. 2004] H. Cohen et al., "PARI/GP version 2.1.3", software, Bordeaux, 2004, Available at http://pari.math.u-bordeaux.fr.

[Cremona 1992] J. E. Cremona, "Modular symbols for $\Gamma_1(N)$ and elliptic curves with everywhere good reduction", *Math. Proc. Cambridge Philos. Soc.* **111**:2 (1992), 199–218. MR 93e:11065 Zbl 0752.11022

[Gelbart and Jacquet 1979] S. Gelbart and H. Jacquet, "Forms of $\mathrm{GL}(2)$ from the analytic point of view", pp. 213–251 in *Automorphic forms, representations and L-functions* (Corvallis, OR, 1977), vol. 1, edited by A. Borel and W. Casselman, Proc. Sympos. Pure Math. **33**, Amer. Math. Soc., Providence, RI, 1979. MR 81e:10024 Zbl 0409.22013

[Gross 1987] B. H. Gross, "Heights and the special values of L-series", pp. 115–187 in *Number theory* (Montreal, 1985), edited by H. Kisilevsky and J. Labute, CMS Conf. Proc. **7**, Amer. Math. Soc., Providence, RI, 1987. MR 89c:11082 Zbl 0623.10019

[Hasse 1952] H. Hasse, *Über die Klassenzahl abelscher Zahlkörper*, Akademie-Verlag, Berlin, 1952. MR 14,141a Zbl 0046.26003

[Hida 1988] H. Hida, "On p-adic Hecke algebras for GL_2 over totally real fields", *Ann. of Math.* (2) **128**:2 (1988), 295–384. MR 89m:11046 Zbl 0658.10034

[Hijikata et al. 1989] H. Hijikata, A. K. Pizer, and T. R. Shemanske, *The basis problem for modular forms on $\Gamma_0(N)$*, Mem. Amer. Math. Soc. **418**, Amer. Math. Soc., Providence, RI, 1989. MR 90d:11056 Zbl 0689.10034

[Labesse and Langlands 1979] J.-P. Labesse and R. P. Langlands, "L-indistinguishability for $\mathrm{SL}(2)$", *Canad. J. Math.* **31**:4 (1979), 726–785. MR 81b:22017 Zbl 0421.12014

[Leopoldt 1958] H.-W. Leopoldt, "Eine Verallgemeinerung der Bernoullischen Zahlen", *Abh. Math. Sem. Univ. Hamburg* **22** (1958), 131–140. MR 19,1161e Zbl 0080.03002

[Livné 1987] R. Livné, "Cubic exponential sums and Galois representations", pp. 247–261 in *Current trends in arithmetical algebraic geometry* (Arcata, CA, 1985), edited by K. A. Ribet, Contemp. Math. **67**, Amer. Math. Soc., Providence, RI, 1987. MR 88g:11032 Zbl 0621.14019

[Marcus 1977] D. A. Marcus, *Number fields*, Springer, New York, 1977. MR 56 #15601 ZBL 0383.12001

[Martinet 1977] J. Martinet, "Character theory and Artin L-functions", pp. 1–87 in *Algebraic number fields: L-functions and Galois properties* (Durham, 1975), edited by A. Fröhlich, Academic Press, London, 1977. MR 56 #5502 Zbl 0359.12015

[Neukirch 1999] J. Neukirch, *Algebraic number theory*, Grundlehren der Math. Wissenschaften **322**, Springer, Berlin, 1999. MR 2000m:11104 Zbl 0956.11021

[Pinch 1982] R. G. E. Pinch, *Elliptic curves over number fields*, Ph.D. thesis, Oxford University, 1982.

[Pizer 1973] A. K. Pizer, “Type numbers of Eichler orders”, *J. Reine Angew. Math.* **264** (1973), 76–102. MR 49 #2650 Zbl 0274.12008

[Pizer 1976] A. Pizer, “The representability of modular forms by theta series”, *J. Math. Soc. Japan* **28**:4 (1976), 689–698. MR 54 #10154 Zbl 0344.10012

[Pizer 1980a] A. Pizer, “An algorithm for computing modular forms on $\Gamma_0(N)$”, *J. Algebra* **64**:2 (1980), 340–390. MR 83g:10020 Zbl 0433.10012

[Pizer 1980b] A. Pizer, “Theta series and modular forms of level p^2M”, *Compositio Math.* **40**:2 (1980), 177–241. MR 81k:10040 Zbl 0416.10021

[Rajan 2002] C. S. Rajan, “On the image and fibres of solvable base change”, *Math. Res. Lett.* **9**:4 (2002), 499–508. MR 2003g:11054 Zbl 01886044

[Shimizu 1965] H. Shimizu, “On zeta functions of quaternion algebras”, *Ann. of Math.* (2) **81** (1965), 166–193. MR 30 #1998 Zbl 0201.37903

[Shimura 1971] G. Shimura, *Introduction to the arithmetic theory of automorphic functions*, Publ. Math. Soc. Japan **11**, Iwanami Shoten, Tokyo, 1971. MR 47 #3318 Zbl 0221.10029

[Skinner and Wiles 1999] C. M. Skinner and A. J. Wiles, “Residually reducible representations and modular forms”, *Inst. Hautes Études Sci. Publ. Math.* **89** (1999), 5–126. MR 2002b:11072 Zbl 1005.11030

[Socrates 1993] J. Socrates, *The quaternionic bridge between elliptic curves and Hilbert modular forms*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1993.

[Taylor 1989] R. Taylor, “On Galois representations associated to Hilbert modular forms”, *Invent. Math.* **98**:2 (1989), 265–280. MR 90m:11176 Zbl 0705.11031

[Vignéras 1980] M.-F. Vignéras, *Arithmétique des algèbres de quaternions*, Lecture Notes in Mathematics **800**, Springer, Berlin, 1980. MR 82i:12016 Zbl 0422.12008

Received January 6, 2004. Revised July 1, 2004.

JUDE SOCRATES
DIVISION OF MATHEMATICS
PASADENA CITY COLLEGE
PASADENA, CA 91106

jtsocrates@paccd.cc.ca.us

DAVID WHITEHOUSE
MATHEMATICS 253-37
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CA 91125

dw@caltech.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

INDECOMPOSABILITY OF FREE GROUP FACTORS OVER NONPRIME SUBFACTORS AND ABELIAN SUBALGEBRAS

MARIUS B. ŞTEFAN

We use the free entropy defined by D. Voiculescu to prove that the free group factors cannot be decomposed as closed linear spans of noncommutative monomials in elements of nonprime subfactors or abelian $*$-subalgebras, if the degrees of monomials have an upper bound depending on the number of generators. The resulting estimates for the hyperfinite and abelian dimensions of free group factors settle in the affirmative a conjecture of L. Ge and S. Popa (for infinitely many generators).

1. Introduction

L. Ge and S. Popa [1998] defined for a given type II_1-factor $\mathcal{M}$ the two quantities

$$\ell_h(\mathcal{M}) = \min\{ f \in \mathbb{N} \mid \exists \text{ hyperfinite } \mathcal{R}_1, \dots, \mathcal{R}_f \subset \mathcal{M} \text{ s.t. } \overline{\mathrm{sp}}^w \mathcal{R}_1\mathcal{R}_2 \cdots \mathcal{R}_f = \mathcal{M}\},$$
$$\ell_a(\mathcal{M}) = \min\{ f \in \mathbb{N} \mid \exists \text{ abelian } \mathcal{A}_1, \dots, \mathcal{A}_f \subset \mathcal{M} \text{ s.t. } \overline{\mathrm{sp}}^w \mathcal{A}_1\mathcal{A}_2 \cdots \mathcal{A}_f = \mathcal{M}\}$$

(the min being ∞ if $\mathcal{M}$ cannot be generated as stated) and conjectured that

$$\ell_h(\mathcal{L}(\mathbb{F}_n)) = \ell_a(\mathcal{L}(\mathbb{F}_n)) = \infty \quad \text{for } n \geq 2,$$

where $\mathcal{L}(\mathbb{F}_n)$ is the type II_1-factor associated to the free group with n generators.

We use the concept of free entropy introduced by D. Voiculescu in his breakthrough paper [1994] to prove that the conjecture mentioned above is true at least partially (for $n = \infty$) that is, $\ell_h(\mathcal{L}(\mathbb{F}_n)), \ell_a(\mathcal{L}(\mathbb{F}_n)) \geq \left[\frac{n-2}{2}\right] + 1$ for $4 \leq n \leq \infty$. Actually, our result is more general and it states that the free group factor with n generators cannot be asymptotically generated (Definitions 3.2 and 4.2) as

$$\lim_{\omega\to 0}{}^{\|\cdot\|_2} \sum_{\substack{1\leq j_1,\dots,j_{t+1}\leq f\\ 1\leq t\leq d}} \mathcal{N}^\omega_{j_1}\mathfrak{X}^\omega\mathcal{N}^\omega_{j_2}\mathfrak{X}^\omega \cdots \mathcal{N}^\omega_{j_t}\mathfrak{X}^\omega\mathcal{N}^\omega_{j_{t+1}}$$

or

$$\lim_{\omega\to 0}{}^{\|\cdot\|_2} \sum_{\substack{1\leq j_1,\dots,j_{t+1}\leq f\\ 1\leq t\leq d}} \mathcal{A}^\omega_{j_1}\mathfrak{X}^\omega\mathcal{A}^\omega_{j_2}\mathfrak{X}^\omega \cdots \mathcal{A}^\omega_{j_t}\mathfrak{X}^\omega\mathcal{A}^\omega_{j_{t+1}}$$

MSC2000: primary 46L54; secondary 46L35.
Keywords: free entropy, free group factors.

if the $\mathcal{N}_1^\omega, \ldots, \mathcal{N}_f^\omega$ (for all ω) are nonprime subfactors, the $\mathcal{A}_1^\omega, \ldots, \mathcal{A}_f^\omega$ are abelian $*$-subalgebras, the $\mathcal{X}^\omega \subset \mathcal{L}(\mathbb{F}_n)$ are subsets containing p self-adjoint elements, and $f, d \geq 1$ are integers such that $n \geq p + 2f + 1$. Note that $\mathcal{L}(\mathbb{F}_n)$ admits decompositions of this sort if we allow $d = \infty$, for example if $\mathcal{X}^\omega = \mathcal{X} = \{1\}$, $f = n$, $\mathcal{N}_1^\omega = \mathcal{N}_1, \ldots, \mathcal{N}_n^\omega = \mathcal{N}_n$ are n distinct copies of the hyperfinite type II_1-factor $\mathcal{R}$ and $\mathcal{A}_1^\omega = \mathcal{A}_1, \ldots, \mathcal{A}_n^\omega = \mathcal{A}_n$ are n distinct copies of $L^\infty([0, 1])$ (since $\mathcal{L}(\mathbb{F}_n)$ is both the free product of n copies of $\mathcal{R}$ and the free product of n copies of $L^\infty([0, 1])$; see [Voiculescu et al. 1992]). The indecomposability of $\mathcal{L}(\mathbb{F}_n)$ as $\overline{\mathrm{sp}}^w \mathcal{N}\mathcal{X}\mathcal{N}$ implies the primeness of its finite-index subfactors; more generally, all subfactors of finite index in the interpolated free group factors of Dykema [1994] and Rădulescu [1994] are prime [Ştefan 1998]. Indeed, according to V. Jones [1983], if $\mathcal{N}$ is a subfactor of finite index in $\mathcal{M}$ then $\mathcal{M}$ decomposes as $\mathcal{N}e\mathcal{N}$, where e is the Jones projection. In particular, the indecomposability properties of $\mathcal{L}(\mathbb{F}_n)$ over nonprime subfactors and abelian subalgebras are preserved to its subfactors of finite index. Recall that the Haagerup approximation property [Haagerup 1978/79] is another property preserved to the free group subfactors. A first example of a prime II_1-factor (with a nonseparable predual, though) was given by Popa [1983] and then Ge [1998] proved (with a free entropy estimate) that the free group factor $\mathcal{L}(\mathbb{F}_n)$ is prime for all n with $2 \leq n < \infty$, thus answering a question from [Popa 1995].

Our results are based on estimates of free entropy, that is, estimates of volumes of various sets of matrix approximants (matricial microstates). Voiculescu [1996] pioneered this technique in his proof of the absence of Cartan subalgebras in the free group factors. Subsequently, Ge [1997] and Dykema [1997] were able to prove that the free group factors do not have abelian subalgebras of finite multiplicity.

The paper has four parts. In Section 2 we prove the first estimate of free entropy and recover a result of Voiculescu [1994]: if a free family of m self-adjoint noncommutative random variables can be generated by noncommutative power series by another family of n self-adjoint noncommutative random variables, then $n \geq m$ (Theorem 2.3). However, we show that the assumption of freeness from [Voiculescu 1994] is not essential and can be dropped. As a consequence, the number of self-adjoint generators with finite entropy that generate a $*$-algebra $\mathcal{A}$ *algebraically*, is constant. In Section 3 we prove the indecomposability of $\mathcal{L}(\mathbb{F}_n)$ (and of its subfactors of finite index) over nonprime subfactors (Theorem 3.5), and in Section 4 the indecomposability over abelian subalgebras (Theorem 4.4).

We give next a short account of Voiculescu's free probability theory [Voiculescu 1990; Voiculescu et al. 1992] and of his original concept of free entropy [Voiculescu 1994; 1996]. A type II_1-factor $\mathcal{M}$ endowed with its unique normalized, faithful, normal trace τ is sometimes called a W^*-probability space. The trace τ determines the 2-norm on $\mathcal{M}$ by the formula $\|x\|_2 = \tau(x^*x)^{1/2}$, for all $x \in M$, and the completion of $\mathcal{M}$ with respect to $\|\cdot\|_2$ is denoted $L^2(\mathcal{M}, \tau)$. An element $x \in \mathcal{M}$ is a semicircular element if it is self-adjoint and if its distribution is given by the semicircle law:

$$\tau(x^k) = \frac{2}{\pi}\int_{-1}^{1} t^k\sqrt{1-t^2}\,dt \quad \text{for all } k \in \mathbb{N}.$$

A family $(\mathcal{A}_i)_{i\in I}$ of unital $*$-subalgebras of $\mathcal{M}$ is a free family if the conditions $n \in \mathbb{N}$, $i_1, \ldots, i_n \in I$, $i_1 \neq i_2 \neq \cdots \neq i_n$, $x_k \in \mathcal{A}_{i_k}$ and $\tau(x_k) = 0$ for $1 \leq k \leq n$ imply $\tau(x_1x_2\cdots x_n) = 0$. A set $\{x_i\}_{i\in I} \subset \mathcal{M}$ is free if the family $(*\text{-alg}\{1, x_i\})_{i\in I}$ is free. A free set $\{x_i\}_{i\in I} \subset \mathcal{M}$ consisting of semicircular elements is called a semicircular system. If $\mathbb{F}_n$ is the free group with n generators ($2 \leq n \leq \infty$) then $\mathcal{L}(\mathbb{F}_n)$ denotes the von Neumann algebra generated by the left regular representation $\lambda : \mathbb{F}_n \to \mathcal{B}(l^2(\mathbb{F}_n))$; see [Murray and von Neumann 1943]. $\mathcal{L}(\mathbb{F}_n)$ is a factor of type II_1 — the free group factor on n generators. It has a canonical trace $\tau(\,\cdot\,) = (\,\cdot\,\delta_e, \delta_e)$, where $\{\delta_g\}_{g\in\mathbb{F}_n}$ is the standard orthonormal basis in $l^2(\mathbb{F}_n)$. Every $\mathcal{L}(\mathbb{F}_n)$ is generated as a von Neumann algebra by a semicircular system with n elements [Voiculescu et al. 1992]. We denote by $\mathcal{M}_k^{\mathrm{sa}} = \mathcal{M}_k^{\mathrm{sa}}(\mathbb{C})$ the set of $k \times k$ self-adjoint complex matrices and by τ_k its unique normalized trace. τ_k induces the 2-norm $\|\cdot\|_2 : \mathcal{M}_k^{\mathrm{sa}} \to \mathbb{R}_+$ and the euclidean norm $\|\cdot\|_e := \sqrt{k}\|\cdot\|_2$. If B is a measurable subset of an m-dimensional (real) manifold, $\mathrm{vol}_m(B)$ denotes the Lebesgue measure of B. The free entropy $\chi(x_1, \ldots, x_n)$ of a finite family of self-adjoint elements was introduced in [Voiculescu 1994], but we will recall the definition of the modified free entropy [Voiculescu 1996], which is better suited for applications. For self-adjoint elements $x_1, \ldots, x_{n+m} \in \mathcal{M}$ one defines first the set of matricial microstates: Fixing $R, \epsilon > 0$ and $p, k \in \mathbb{N}$ we define $\Gamma_R(x_1, \ldots, x_n : x_{n+1}, \ldots, x_{n+m}; p, k, \epsilon)$ to be the set

$$\Big\{(A_1, \ldots, A_n) \in (\mathcal{M}_k^{\mathrm{sa}})^n \;\Big|\; \text{there exist } A_{n+1}, \ldots, A_{n+m} \in \mathcal{M}_k^{\mathrm{sa}} \\ \text{such that } \|A_j\| \leq R \text{ and } \big|\tau(x_{i_1}\cdots x_{i_q}) - \tau_k(A_{i_1}\cdots A_{i_q})\big| < \epsilon \\ \text{for all } q = 1, \ldots, p \text{ and all } j, i_1, \ldots, i_q \in \{1, \ldots, n+m\}\Big\}.$$

Next we define

$$\chi_R(x_1, \ldots, x_n : x_{n+1}, \ldots, x_{n+m}; p, k, \epsilon) := \\ \log\big(\mathrm{vol}_{nk^2}(\Gamma_R(x_1, \ldots, x_n : x_{n+1}, \ldots, x_{n+m}; p, k, \epsilon))\big),$$

$$\chi_R(x_1, \ldots, x_n : x_{n+1}, \ldots, x_{n+m}; p, \epsilon) := \\ \limsup_{k\to\infty}\Big(\frac{1}{k^2}\chi_R(x_1, \ldots, x_n : x_{n+1}, \ldots, x_{n+m}; p, k, \epsilon) + \frac{n}{2}\log k\Big),$$

$$\chi_R(x_1, \ldots, x_n : x_{n+1}, \ldots, x_{n+m}) := \\ \inf\big\{\chi_R(x_1, \ldots, x_n : x_{n+1}, \ldots, x_{n+m}; p, \epsilon) \;\big|\; p \in \mathbb{N}, \epsilon > 0\big\},$$

$$\chi(x_1, \ldots, x_n : x_{n+1}, \ldots, x_{n+m}) := \sup\big\{\chi_R(x_1, \ldots, x_n : x_{n+1}, \ldots, x_{n+m}) \;\big|\; R > 0\big\}.$$

When taking the last sup it suffices to assume $0 < R \le \max\{\|x_1\|, \dots, \|x_{n+m}\|\}$ rather than $0 < R < \infty$ [Voiculescu 1994; 1996]. The quantity

$$\chi(x_1, \dots, x_n : x_{n+1}, \dots, x_{n+m})$$

is the free entropy of $x_1, \dots, x_n$ in the presence of $x_{n+1}, \dots, x_{n+m}$. If $m = 0$, it is simply called the free entropy of $x_1, \dots, x_n$ and written $\chi(x_1, \dots, x_n)$. If $\{x_{n+1}, \dots, x_{n+m}\} \subset \{x_1, \dots, x_n\}''$ we have

$$\chi(x_1, \dots, x_n : x_{n+1}, \dots, x_{n+m}) = \chi(x_1, \dots, x_n);$$

see [Voiculescu 1996]. For a single self-adjoint element $x = x^* \in \mathcal{M}$ one has:

$$\chi(x) = \tfrac{3}{4} + \tfrac{1}{2}\log 2\pi + \iint \log|s-t|\, d\mu(s)\, d\mu(t),$$

where μ is the distribution of x; see [Voiculescu 1994]. If $x_1, \dots, x_n$ are n self-adjoint free elements of $\mathcal{M}$ then $\chi(x_1, \dots, x_n) = \chi(x_1) + \dots + \chi(x_n)$ [Voiculescu 1994]. The converse is also true [Voiculescu 1997], provided that $\chi(x_i) > -\infty$ for $1 \le i \le n$. In particular, the free entropy of a finite semicircular system is finite; hence the free group factor $\mathcal{L}(\mathbb{F}_n)$ has a system of generators with finite free entropy for $2 \le n < \infty$.

2. Noncommutative power series and free entropy

The main result of this section is that if a (not necessarily free) family of m self-adjoint noncommutative random variables with finite free entropy can be generated as noncommutative power series by another family of n self-adjoint noncommutative random variables, then $n \ge m$. In other words, a finite system with finite free entropy has minimal cardinality among all finite systems of self-adjoint elements that are equivalent under the noncommutative analytic functional calculus. Thus, we recover Voiculescu's result [1994], with the observation that our approach does not require the assumption of freeness.

We review first a few facts concerning the theory of systems of algebraic equations [van der Waerden 1949], necessary in the proof of Lemma 2.1. If $g_1, \dots, g_n$ are forms in n variables, there exists a polynomial (the resolvent) in their coefficients, $R(g_1, \dots, g_n)$, with the property that $R(g_1, \dots, g_n) = 0$ if and only if the system $g_1(\xi_1, \dots, \xi_n) = \dots = g_n(\xi_1, \dots, \xi_n) = 0$ has a nontrivial solution. If $h_1, \dots, h_{n-1}$ are $n-1$ forms in n variables and

$$h_n(u)(\xi_1, \dots, \xi_n) := u_1\xi_1 + \dots + u_n\xi_n,$$

then $R_u(h_1, \dots, h_{n-1}) := R(h_1, \dots, h_{n-1}, h_n(u))$ (the u-resolvent) is either identically 0, or a form of degree $\deg h_1 \times \dots \times \deg h_{n-1}$ in $u = (u_1, \dots, u_n)$. In the first case, the system $h_1 = \dots = h_{n-1} = 0$ has infinitely many solutions

$[(\xi_1, \dots, \xi_n)] \in \mathbb{P}\mathbb{C}^{n-1}$; in the second, all the solutions $[(\xi_1, \dots, \xi_n)] \in \mathbb{P}\mathbb{C}^{n-1}$ are given by the factorization of $R_u(h_1, \dots, h_{n-1})$ (and thus, the system admits at most $\deg h_1 \times \cdots \times \deg h_{n-1}$ solutions, as predicted by Bézout's Theorem).

Let $f_1, \dots, f_n \in \mathbb{R}[\Xi_1, \dots, \Xi_n]$ be n polynomials in n indeterminates, of degrees $d_1, \dots, d_n$, respectively. For $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ define

$$F_{i,a_i}(\xi_1, \dots, \xi_{n+1}) = \xi_{n+1}^{d_i}\left(f_i\Big(\frac{\xi_1}{\xi_{n+1}}, \dots, \frac{\xi_n}{\xi_{n+1}}\Big) - a_i \right) \qquad \text{for } i = 1, \dots, n.$$

Bézout's Theorem implies that the system of equations

$$f_1(\xi_1, \dots, \xi_n) = a_1, \dots, f_n(\xi_1, \dots, \xi_n) = a_n$$

admits at most $d_1 \dots d_n$ solutions $(\xi_1, \dots, \xi_n) \in \mathbb{C}^n$ if $R_u(F_{1,a_1}, \dots, F_{n,a_n}) \not\equiv 0$. Note also that the set

$$S_u(f_1, \dots, f_n) := \left\{(a_1, \dots, a_n) \in \mathbb{R}^n \mid R_u(F_{1,a_1}, \dots, F_{n,a_n}) \not\equiv 0\right\}$$

is either open and dense in $\mathbb{R}^n$, or empty.

We proceed now with Lemma 2.1, which gives an upper bound for the Lebesgue measure of the intersection of an algebraically parameterized manifold embedded in $\mathbb{R}^m$ with the unit ball of $\mathbb{R}^m$. This lemma will be of further use in estimating the volumes of various sets of matricial microstates that will appear as sets of points within a given distance from such manifolds.

Lemma 2.1. *For integers $n \le m$ and polynomials $f_1, \dots, f_m \in \mathbb{R}[\Xi_1, \dots, \Xi_n]$ define $f = (f_1, \dots, f_m) : \mathbb{R}^n \to \mathbb{R}^m$. If the polynomials*

$$\det\Big(\frac{\partial f_J}{\partial \xi}\Big)$$

are not identically 0 *for all multiindices $J \in \{(i_1, \dots, i_n) \mid 1 \le i_1 < \cdots < i_n \le m\}$ and if $S_u = S_u(f_1, \dots, f_n) \ne \varnothing$, then*

$$\int_{f^{-1}\left(\overline{B(0,1)}\right)} \left(\sum_{|J|=n} \det^2\Big(\frac{\partial f_J}{\partial \xi}\Big)\right)^{1/2} d\xi \le \binom{m}{n} \cdot C \cdot \mathrm{vol}_n(B(0,1)), \tag{1}$$

where $C = C(\deg f) = \max\{\deg f_{i_1} \times \cdots \times \deg f_{i_n} \mid 1 \le i_1 < \cdots < i_n \le m\}$ and $B(0,1) = B_n(0,1)$ is the unit ball in $\mathbb{R}^n$.

Proof. We consider first the case $m = n$. Let S denote the set of all irregular values of f, that is,

$$S = f\left(\{\xi \in \mathbb{R}^n \mid \mathrm{rank}(df_\xi) < n\}\right).$$

It suffices to show that (1) holds with $f^{-1}\big(\overline{B(0,1)} \setminus S_\epsilon\big)$ replacing $f^{-1}\big(\overline{B(0,1)}\big)$, where S_ϵ is an arbitrary open set that contains $S \cup (\mathbb{R}^n \setminus S_u)$. For any

$$a = (a_1, \dots, a_n) \in \mathrm{Range}\, f \cap \overline{B(0,1)} \setminus S_\epsilon$$

the set $f^{-1}(\{a\})$ has at most $C = \deg f_1 \times \cdots \times \deg f_n$ elements, say $f^{-1}(\{a\}) = \{b_1, \ldots, b_{p(a)}\}$ for some $1 \leq p(a) \leq C$. There exist an open ball $B_a \ni a$ and open neighborhoods $V_1^a \ni b_1, \ldots, V_{p(a)}^a \ni b_{p(a)}$ such that B_a and V_i^a are diffeomorphic via f for $1 \leq i \leq p(a)$ and $f^{-1}(B_a) = \bigcup_{i=1}^{p(a)} V_i^a$. Since it is compact, we can cover Range $f \cap \overline{B(0,1)} \setminus S_\epsilon$ with a finite set of such open balls $B_{a_1}, \ldots, B_{a_k}$. This covering determines a finite partition of Range $f \cap \overline{B(0,1)} \setminus S_\epsilon$, say $W_1, \ldots, W_t$. For each $1 \leq j \leq t$ choose a unique $1 \leq l = l(j) \leq k$ such that $W_j \subset B_{a_l}$ and $f^{-1}(W_j) = T_{j1} \cup \cdots \cup T_{jp(a_l)}$, where $T_{ji} \subset V_i^{a_l}$ and W_j and T_{ji} are diffeomorphic via f for all $1 \leq i \leq p(a_l)$. We have

$$\begin{aligned}
\int_{f^{-1}\left(\overline{B(0,1)}\setminus S_\epsilon\right)} \left|\det\left(\frac{\partial f}{\partial \xi}\right)\right| d\xi &= \sum_{j=1}^{t} \int_{f^{-1}(W_j)} \left|\det\left(\frac{\partial f}{\partial \xi}\right)\right| d\xi \\
&= \sum_{j=1}^{t} \sum_{i=1}^{p(a_{l(j)})} \int_{T_{ji}} \left|\det\left(\frac{\partial f}{\partial \xi}\right)\right| d\xi \\
&= \sum_{j=1}^{t} \sum_{i=1}^{p(a_{l(j)})} \operatorname{vol}_n(W_j) \\
&\leq C \sum_{j=1}^{t} \operatorname{vol}_n(W_j) = C \cdot \operatorname{vol}_n\left(\overline{B(0,1)} \setminus S_\epsilon\right).
\end{aligned}$$

In the case $m > n$ one has the estimates

$$\begin{aligned}
\int_{f^{-1}(\overline{B(0,1)})} \left(\sum_{|J|=n} \det{}^2\left(\frac{\partial f_J}{\partial \xi}\right)\right)^{1/2} d\xi &\leq \int_{f^{-1}(\overline{B(0,1)})} \sum_{|J|=n} \left|\det\left(\frac{\partial f_J}{\partial \xi}\right)\right| d\xi \\
&\leq \sum_{|J|=n} \int_{f_J^{-1}(\overline{B(0,1)})} \left|\det\left(\frac{\partial f_J}{\partial \xi}\right)\right| d\xi \\
&\leq \binom{m}{n} \cdot C \cdot \operatorname{vol}_n(B(0,1)). \qquad \square
\end{aligned}$$

Lemma 2.1 will be used in the proof of Proposition 2.2. The $k \times k$ matricial microstates of $x_1, \ldots, x_m$ are points within euclidean distance $2\omega\sqrt{mk}$ from the range of a polynomial function in the matricial microstates of $y_1, \ldots, y_n$ provided that each x_i is within $\|\cdot\|_2$-distance ω from noncommutative polynomials in $y_1, \ldots, y_n$.

Proposition 2.2. *Let $P_1, \ldots, P_m \in \mathbb{C}\langle Y_1, \ldots, Y_n\rangle$ be complex polynomials in n noncommutative self-adjoint variables. Assume that $(\mathcal{M}, \tau)$ is a II_1-factor and that $\{x_1, \ldots, x_m\} \subset \mathcal{M}$ is a finite set of self-adjoint generators of $\mathcal{M}$. Set*

$$a = \max\left\{\|x_1\|_2 + 1, \ldots, \|x_m\|_2 + 1\right\}$$

and $d = \max\{\deg P_1, \dots, \deg P_m\}$. *If* $\{y_1, \dots, y_n\} \subset \mathcal{M}$ *is another finite set of self-adjoint generators of* $\mathcal{M}$ *with* $n < m$ *and such that*

$$\|x_i - P_i(y_1, \dots, y_n)\|_2 < \omega \quad \textit{for all } i = 1, \dots, m \textit{ and some } \omega \in (0, a],$$

then

$$\chi(x_1, \dots, x_m) \leq C(m, n, a) + (m-n)\log\omega + n\log d,$$

where $C(m, n, a)$ *is a constant that depends only on* m, n, a.

Proof. Replacing each P_i by $\frac{1}{2}(P_i + P_i^*)$ if necessary, we can assume that $P_i = P_i^*$ for $i = 1, \dots, m$. Given $R > 0$, $\epsilon > 0$ and an integer $p \geq 1$, consider

$$(A_1, \dots, A_m) \in \Gamma_R(x_1, \dots, x_m : y_1, \dots, y_n;\, p, k, \epsilon).$$

If p is large enough and $\epsilon > 0$ is sufficiently small, one can find matrices $B_1, \dots, B_n$ in $\mathcal{M}_k^{\mathrm{sa}}$ such that $\|B_1\|, \dots, \|B_n\| \leq R$ and

$$\|A_i - P_i(B_1, \dots, B_n)\|_2 < \omega \quad \text{for } i = 1, \dots, m,$$

or, equivalently,

$$\|A_i - P_i(B_1, \dots, B_n)\|_e < \omega\sqrt{k} \quad \text{for } i = 1, \dots, m.$$

With the identifications $g = (g_1, \dots, g_{mk^2}) : (\mathcal{M}_k^{\mathrm{sa}})^n \cong \mathbb{R}^{nk^2} \to (\mathcal{M}_k^{\mathrm{sa}})^m \cong \mathbb{R}^{mk^2}$, $(B_1, \dots, B_n) = (\xi_1, \dots, \xi_{nk^2}) \in \mathbb{R}^{nk^2}$, and

$$g(B_1, \dots, B_n) = \big(P_1(B_1, \dots, B_n), \dots, P_m(B_1, \dots, B_n)\big),$$

the previous inequalities imply

$$\big\|(A_i)_{1\leq i\leq m} - g(\xi_1, \dots, \xi_{nk^2})\big\|_e < \omega\sqrt{mk}.$$

At the cost of introducing an additional variable $\xi_{nk^2+1} \in \mathbb{R}$, we can assume that the components of g are mk^2 homogeneous polynomial functions in the variables $\xi_1, \dots, \xi_{nk^2+1}$, all having degrees at most d.

Now let $f_1, \dots, f_{mk^2}$ be arbitrary homogeneous polynomial functions in $\xi_1, \dots, \xi_{nk^2+1}$, such that $\deg f_j = \deg g_j$ for $j = 1, \dots, mk^2$. For every multiindex $J = (j_1, \dots, j_{nk^2+1})$ with $1 \leq j_1 < \cdots < j_{nk^2+1} \leq mk^2$, saying that $S_u(f_{j_1}, \dots, f_{j_{nk^2+1}})$ is empty is equivalent to saying that the coefficients of $f_{j_1}, \dots, f_{j_{nk^2+1}}$ satisfy a certain system of algebraic equations. Hence the set

$$\Omega_1 = \Big\{ f = (f_1, \dots, f_{mk^2}) \;\Big|\; \deg f_j = \deg g_j \text{ for } j = 1, \dots, mk^2, \\ S_u(f_{j_1}, \dots, f_{j_{nk^2+1}}) \neq \varnothing \text{ for all } J = (j_1, \dots, j_{nk^2+1}) \Big\}$$

is open and dense in its natural ambient linear space. Similarly, the set

$$\Omega_2 = \Big\{ f = (f_1, \dots, f_{mk^2}) \,\Big|\, \deg f_j = \deg g_j \text{ for } j = 1, \dots, mk^2, \\ \det\Big(\frac{\partial f_J}{\partial \xi}\Big) \not\equiv 0 \text{ for all } J = (j_1, \dots, j_{nk^2+1}) \Big\}$$

is also open and dense in the same linear space.

The matrix df_ξ has $\binom{mk^2}{nk^2+1}$ minors of dimension $(nk^2+1)\times(nk^2+1)$ and all these minors have a nontrivial common zero only if a certain system of algebraic equations in the coefficients of $f_1, \dots, f_{mk^2}$ has a solution [van der Waerden 1949]. Not all the polynomials appearing in this system are identically equal to 0. It follows that the set

$$\Omega_3 = \big\{ f = (f_1, \dots, f_{mk^2}) \,\big|\, \deg f_j = \deg g_j \text{ for } j = 1, \dots, mk^2, \\ \operatorname{rank}(df_\xi) = nk^2 + 1 \ \forall \xi \in \mathbb{R}^{nk^2+1} \setminus \{0\} \big\}$$

contains a subset that is open and dense in the linear space previously considered. Therefore there exists an element $f \in \Omega_1 \cap \Omega_2 \cap \Omega_3$ such that

$$\big\| f(\xi_1, \dots, \xi_{nk^2+1}) - g(\xi_1, \dots, \xi_{nk^2+1}) \big\|_e < \omega\sqrt{mk} \quad \text{if } |\xi_i| \le R \text{ for } 1 \le i \le nk^2+1;$$

hence $\big\|(A_i)_{1\le i\le m} - f(\xi_1, \dots, \xi_{nk^2+1})\big\|_e < 2\omega\sqrt{mk}$. The function f satisfies the hypotheses of Lemma 2.1 and its components are homogeneous polynomials. It has the property that $\operatorname{dist}_e\big((A_i)_{1\le i\le m}, \operatorname{Range} f\big) < 2\omega\sqrt{mk}$ and it does not depend on the system $(A_i)_{1\le i\le m}$.

We have $\|(A_1, \dots, A_m)\|_e \le a\sqrt{mk}$ (if $\epsilon > 0$ is small enough); hence the set of matricial microstates $(A_1, \dots, A_m)$ of $(x_1, \dots, x_m)$ such that

$$\operatorname{dist}_e\big((A_1, \dots, A_m), \operatorname{Range} f\big) < 2\omega\sqrt{mk}$$

is contained in the (mk^2, nk^2+1)-tube of radius $2\omega\sqrt{mk}$ around

$$\operatorname{Range} f \cap B_{mk^2}\big(0, (a+2\omega)\sqrt{mk}\,\big).$$

If B is a small ball in $\mathbb{R}^{nk^2+1} \setminus \{0\}$ and if $V_B(2\omega\sqrt{mk})$ denotes the (mk^2, nk^2+1)-tube of radius $2\omega\sqrt{mk}$ around $f(B)$, the formula for volumes of tubes [Weyl 1939] implies

$$\operatorname{vol}_{mk^2}\big(V_B(2\omega\sqrt{mk})\big) = \operatorname{vol}_{mk^2-nk^2-1}\big(B_{mk^2-nk^2-1}(0,1)\big) \\ \cdot \sum_{\substack{e \text{ even} \\ 0\le e\le nk^2+1}} \frac{\big(2\omega\sqrt{mk}\,\big)^{e+mk^2-nk^2-1} k_{B,e}}{(mk^2-nk^2+1)(mk^2-nk^2+3)\cdots(mk^2-nk^2-1+e)}.$$

With the notations from [Weyl 1939] one has $k_{B,e} = \int_{f(B)} H_e ds$ and

$$H_e = \frac{1}{2^e (e/2)!} \sum_{\sigma \in \Sigma_e} \operatorname{sgn} \sigma \sum_{\alpha_1,\dots,\alpha_e=1}^{nk^2+1} H_{\alpha_1\alpha_2}^{\alpha_{\sigma(1)}\alpha_{\sigma(2)}} H_{\alpha_3\alpha_4}^{\alpha_{\sigma(3)}\alpha_{\sigma(4)}} \dots,$$

where $H_{\alpha\beta}^{\lambda\mu}$ denotes the Riemann tensor of $f(B)$. Assuming without loss of generality that $\deg f_j = d$ for $j = 1, \dots, mk^2$, one can verify that each $H_{\alpha\beta}^{\lambda\mu}(f(\xi))$ is a sum of quotients of homogeneous polynomials where all numerators have degree $6(d-1)(nk^2+1)-2d$ and all denominators have degree $6(d-1)(nk^2+1)$. Hence H_e is a rational function in ξ and in the coefficients of $f(\xi)$. Due to its intrinsic nature, H_e is independent of the embedding of Range f in $\mathbb{R}^{mk^2+1}$; in particular it is invariant under orthogonal transformations in $\mathbb{R}^{mk^2+1}$. Since there exist sufficiently many polynomials $f(\xi)$ such that Range f is flat, this entails $H_e = 0$ for even e such that $2 \le e \le nk^2 + 1$. Therefore the volume of the $(mk^2, nk^2 + 1)$-tube of radius $2\omega\sqrt{mk}$ around $f(B)$ is

$$\operatorname{vol}_{mk^2} V_B(2\omega\sqrt{mk}) = (\operatorname{vol}_{mk^2-nk^2-1} B_{mk^2-nk^2-1}(0,1))\,(2\omega\sqrt{mk})^{mk^2-nk^2-1} \int_{f(B)} ds,$$

and with Lemma 2.1 and the inequality

$$\frac{1}{\Gamma\Big(1+\dfrac{nk^2+1}{2}\Big)} \cdot \frac{1}{\Gamma\Big(1+\dfrac{mk^2-nk^2-1}{2}\Big)} \le \frac{2^{mk^2/2}}{\Gamma\Big(1+\dfrac{mk^2}{2}\Big)} \tag{2}$$

we obtain the estimate

$$\begin{aligned}
&\operatorname{vol}_{mk^2} \Gamma_R(x_1, \dots, x_m : y_1, \dots, y_n;\, p, k, \epsilon) \\
&\quad \le \binom{mk^2}{nk^2+1} \cdot C(d) \cdot \operatorname{vol}_{nk^2+1} B(0,\, (a+2\omega)\sqrt{mk}) \\
&\qquad\qquad \cdot \operatorname{vol}_{mk^2-nk^2-1} B(0,1) \cdot (2\omega\sqrt{mk})^{mk^2-nk^2-1} \\
&\quad = \binom{mk^2}{nk^2+1} \cdot C(d) \cdot \pi^{(nk^2+1)/2} \\
&\qquad \cdot \frac{(a+2\omega)^{nk^2+1} (mk)^{(nk^2+1)/2} \pi^{(mk^2-nk^2-1)/2} (2\omega)^{mk^2-nk^2-1} (mk)^{(mk^2-nk^2-1)/2}}{\Gamma\Big(1+\dfrac{nk^2+1}{2}\Big)\,\Gamma\Big(1+\dfrac{mk^2-nk^2-1}{2}\Big)} \\
&\quad \le \binom{mk^2}{nk^2+1} \cdot C(d) \cdot \frac{\pi^{mk^2/2} (mk)^{mk^2/2} 2^{mk^2/2} (3a)^{nk^2+1} (2\omega)^{mk^2-nk^2-1}}{\Gamma\Big(1+\dfrac{mk^2}{2}\Big)}.
\end{aligned}$$

The last inequality implies further

$$\begin{aligned}
&\chi_R(x_1,\dots,x_m : y_1,\dots,y_n;\ p,k,\epsilon)\\
&\quad\le \frac{1}{k^2}\log\binom{mk^2}{nk^2+1}+\frac{1}{k^2}\log C(d)+\frac{m}{2}\log\pi+\Big(\frac{3m}{2}-n\Big)\log 2+n\log(3a)\\
&\qquad\qquad+\frac{m}{2}\log(mk)+(m-n)\log\omega-\frac{1}{k^2}\log\Gamma\Big(1+\frac{mk^2}{2}\Big)+\frac{m}{2}\log k+o(1).
\end{aligned}$$

Note that

$$\frac{1}{k^2}\log\Gamma\Big(1+\frac{mk^2}{2}\Big)=\frac{m}{2}\log\frac{mk^2}{2e}+o(1),$$

$C(d)\le d^{nk^2+1}$ and $\frac{1}{k^2}\log\binom{mk^2}{nk^2+1}=m\log m-n\log n-(m-n)\log(m-n)+o(1)$; therefore

$$\begin{aligned}
&\chi_R(x_1,\dots,x_m : y_1,\dots,y_n;\ p,k,\epsilon)\\
&\quad\le m\log m-n\log n+n\log d-(m-n)\log(m-n)+\frac{m}{2}\log\pi\\
&\qquad\quad+\Big(\frac{3m}{2}-n\Big)\log 2+n\log(3a)+\frac{m}{2}\log m+\frac{m}{2}\log k+(m-n)\log\omega\\
&\qquad\quad-\frac{m}{2}\log\frac{m}{2e}-m\log k+\frac{m}{2}\log k+o(1)\\
&\quad=C(m,n,a)+(m-n)\log\omega+n\log d+o(1).
\end{aligned}$$

By taking the appropriate limits after k, p, ϵ, we finally obtain

$$\chi_R(x_1,\dots,x_m : y_1,\dots,y_n)\le C(m,n,a)+(m-n)\log\omega+n\log d,$$

and since $R>0$ is arbitrary,

$$\chi(x_1,\dots,x_m : y_1,\dots,y_n)\le C(m,n,a)+(m-n)\log\omega+n\log d.$$

Now recall that $\{x_1,\dots,x_m\}$ is a system of generators of $\mathcal{M}$; hence $\chi(x_1,\dots,x_m)=\chi(x_1,\dots,x_m : y_1,\dots,y_n)$. □

Let $Y_1,\dots,Y_n$ be noncommutative indeterminates and let

$$P(Y_1,\dots,Y_n)=\sum_{k=0}^{\infty}\sum_{1\le i_1,\dots,i_k\le n} a_{i_1\cdots i_k}Y_{i_1}\cdots Y_{i_k}$$

be a noncommutative power series in $Y_1,\dots,Y_n$, with complex coefficients. Following [Voiculescu 1994], we say that $R>0$ is a radius of convergence of P if

$$\sum_{k=0}^{\infty}\sum_{1\le i_1,\dots,i_k\le n}|a_{i_1\cdots i_k}|R^k<\infty.$$

It is well known from the theory of power series that if $0 < R_0 < R$, then

$$\sum_{k=q+1}^{\infty} \sum_{1\le i_1,\dots,i_k\le n} |a_{i_1\cdots i_k}| R_0^k = O\left(\left(\frac{R_0}{R}\right)^{q+1}\right).$$

The next result is basically [Voiculescu 1994, Corollary 6.12], with the observation that the freeness of $\{x_1, \dots, x_m\}$ assumed there has been dropped.

Theorem 2.3. *Let $x_1, \dots, x_m$ and $y_1, \dots, y_n$ be self-adjoint noncommutative random variables in a II_1-factor $(\mathcal{M}, \tau)$ such that $y_1, \dots, y_n \in \{x_1, \dots, x_m\}''$ and $\chi(x_1, \dots, x_m) > -\infty$. If $x_i = P_i(y_1, \dots, y_n)$ for $i = 1, \dots, m$, where the P_i are noncommutative power series having a common radius of convergence $R > b = \max\{\|y_1\|, \dots, \|y_n\|\}$, then $n \ge m$.*

Proof. Suppose that $m > n$. For $1 \le i \le m$, x_i is a noncommutative power series of $y_1, \dots, y_n$, i.e.,

$$x_i = \sum_{k=0}^{\infty} \sum_{1\le i_1,\dots,i_k\le n} a^{(i)}_{i_1\cdots i_k} y_{i_1} \cdots y_{i_k}.$$

For every integer $q \ge 0$, $P_{i,q}(y_1, \dots, y_n) := \sum_{k=0}^{q} \sum_{1\le i_1,\dots,i_k\le n} a^{(i)}_{i_1\cdots i_k} y_{i_1} \cdots y_{i_k}$ is a noncommutative polynomial of degree at most q, and

$$\begin{aligned}\bigl\|x_i - P_{i,q}(y_1, \dots, y_n)\bigr\|_2 &= \Bigl\|\sum_{k=q+1}^{\infty} \sum_{1\le i_1,\dots,i_k\le n} a^{(i)}_{i_1\cdots i_k} y_{i_1} \cdots y_{i_k}\Bigr\|_2 \\ &\le \sum_{k=q+1}^{\infty} \sum_{1\le i_1,\dots,i_k\le n} |a^{(i)}_{i_1\cdots i_k}| b^k = O\left(\left(\frac{b}{R}\right)^{q+1}\right).\end{aligned}$$

The estimate of free entropy from Proposition 2.2 implies

$$\chi(x_1, \dots, x_m) \le C(m, n, a) + (m-n)\log\left(\frac{b}{R}\right)^{q+1} + n\log q + O(1)$$

and letting q tend to ∞, one obtains $\chi(x_1, \dots, x_m) = -\infty$, a contradiction. □

Let $\mathcal{N}$ be a $*$-algebra in a W^*-probability space $(\mathcal{M}, \tau)$. Suppose that $\mathcal{N}$ is finitely generated and let $\{x_1, \dots, x_m\}$ be a system of self-adjoint generators. Let also $\{y_1, \dots, y_n\}$ be another set of self-adjoint elements that generate $\mathcal{N}$ *algebraically* as a $*$-algebra. In particular, there exist noncommutative polynomials $(P_i)_{1\le i\le m}$ such that $x_i = P_i(y_1, \dots, y_n)$ for $i = 1, \dots, \le m$. In this context, Corollary 2.4 is an immediate consequence of Theorem 2.3.

Corollary 2.4. *If $\chi(x_1, \dots, x_m) > -\infty$ and $*$-alg$\{y_1, \dots, y_n\} = *$-alg$\{x_1, \dots, x_m\}$ then $n \ge m$, so any 2 systems of self-adjoint elements with finite free entropy that generate $\mathcal{N}$ algebraically as a $*$-algebra have the same cardinality.*

Voiculescu [1998] proved that the modified free entropy dimension [Voiculescu 1996] of a finite set of self-adjoint elements that generate algebraically a $*$-algebra $\mathcal{N}$ is independent of the set of generators. It is still an open question whether the free entropy dimension is a von Neumann algebra invariant. Voiculescu [1999] also showed that sets of generators satisfying sequential commutation in certain property T factors have modified free entropy dimension ≤ 1. L. Ge and J. Shen ([2000]) proved then that the estimate $\delta_0 \leq 1$ is true for any set of generators, as long as the factor has one set of generators satisfying sequential commutation. Recall from [Voiculescu 1996] the definition of the modified free entropy dimension:

$$\delta_0(x_1, \ldots, x_m) = m + \limsup_{\omega \to 0} \frac{\chi(x_1 + \omega s_1, \ldots, x_1 + \omega s_m : s_1, \ldots, s_m)}{|\log \omega|},$$

where $\{s_1, \ldots, s_m\}$ is a semicircular system free from $\{x_1, \ldots, x_m\}$. In general one has $\delta_0(x_1, \ldots, x_m) \leq m$, and also $0 \leq \delta_0(x_1, \ldots, x_m)$ if $\{x_1, \ldots, x_m\} \subset \mathcal{L}(\mathbb{F}_p)$ for some p. Considering two sets $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$ of self-adjoint elements that generate algebraically the $*$-algebra $\mathcal{N}$ and noticing that $\{y_1, \ldots, y_n\} \subset \{x_1 + \omega s_1, \ldots, x_1 + \omega s_m, s_1, \ldots, s_m\}''$, one has

$$\begin{aligned}\delta_0(x_1, \ldots, x_m) &= m + \limsup_{\omega \to 0} \frac{\chi(x_1 + \omega s_1, \ldots, x_1 + \omega s_m : s_1, \ldots, s_m, y_1, \ldots, y_n)}{|\log \omega|} \\ &\leq m + \limsup_{\omega \to 0} \frac{\chi(x_1 + \omega s_1, \ldots, x_1 + \omega s_m : y_1, \ldots, y_n)}{|\log \omega|}.\end{aligned}$$

Also, $\left\|x_i + \omega s_i - P_i(y_1, \ldots, y_n)\right\| = \|\omega s_i\| \leq \omega$ for $i = 1, \ldots, m$, and with Proposition 2.2 we obtain

$$\begin{aligned}\delta_0(x_1, \ldots, x_m) &\leq m + \limsup_{\omega \to 0} \frac{C(m, n, a) + (m - n)\log \omega + n \log d}{|\log \omega|} \\ &\leq m + n - m = n,\end{aligned}$$

where $a = \max\{\|x_1\|_2 + 1, \ldots, \|x_m\|_2 + 1, \|y_1\|_2 + 1, \ldots, \|y_n\|_2 + 1\}$ and $d = \max\{\deg P_i \mid 1 \leq i \leq m\}$. In particular, if there exists a set $\{y_1, \ldots, y_n\}$ with $\delta_0(y_1, \ldots, y_n) = n$ which generates $\mathcal{N}$ algebraically, then

$$\sup\left\{\delta_0(x_1, \ldots, x_m) \;\middle|\; *\text{-alg}\{x_1, \ldots, x_m\} = \mathcal{N}\right\} = n.$$

3. Indecomposability over nonprime subfactors

In this section we prove that the free group factor $\mathcal{L}(\mathbb{F}_n)$ does not admit an asymptotic decomposition of the form

$$\lim_{\omega \to 0}{}^{\|\cdot\|_2} \sum_{\substack{1 \leq j_1, \ldots, j_{t+1} \leq f \\ 1 \leq t \leq d}} \mathcal{N}^{\omega}_{j_1} \mathcal{Z}^{\omega} \mathcal{N}^{\omega}_{j_2} \mathcal{Z}^{\omega} \cdots \mathcal{N}^{\omega}_{j_t} \mathcal{Z}^{\omega} \mathcal{N}^{\omega}_{j_{t+1}},$$

where (for each ω) $\mathcal{X}^\omega \subset \mathcal{L}(\mathbb{F}_n)$ is a subset with p self-adjoint elements, $\mathcal{N}_1^\omega, \dots, \mathcal{N}_f^\omega$ are nonprime subfactors of $\mathcal{L}(\mathbb{F}_n)$, the integer d is at least 1, and $n \geq p+2f+1$. A nonprime II_1-factor is just a factor isomorphic to the tensor product of two factors of type II_1. For free group subfactors one has the following: if $n \geq p+2f+2$ and $\mathcal{P} \subset \mathcal{L}(\mathbb{F}_n)$ is a subfactor of finite index, then $\mathcal{P}$ does not admit such an asymptotic decomposition either. In particular, the hyperfinite dimension of $\mathcal{L}(\mathbb{F}_n)$ is at least $\left[\frac{n-2}{2}\right]+1$ and that of $\mathcal{P}$ is at least $\left[\frac{n-3}{2}\right]+1$. For $n=\infty$ this settles a conjecture of Ge and Popa [1998]: the hyperfinite dimension of free group factors is infinite. The definitions of hyperfinite dimension and of asymptotic decomposition over nonprime subfactors are given next.

Definition 3.1 [Ge and Popa 1998]. If $\mathcal{M}$ is a type II_1-factor, the hyperfinite dimension of $\mathcal{M}$, denoted $\ell_h(\mathcal{M})$, is by definition the smallest positive integer $f \in \mathbb{N}$ with the property that there exist hyperfinite subalgebras $\mathcal{R}_1, \dots, \mathcal{R}_f \subset \mathcal{M}$ such that $\overline{\mathrm{sp}}^w \mathcal{R}_1\mathcal{R}_2 \cdots \mathcal{R}_f = \mathcal{M}$. If there is no such positive integer, $\ell_h(\mathcal{M}) = +\infty$.

Definition 3.2. A type II_1-factor $\mathcal{M}$ admits an asymptotic decomposition over nonprime subfactors if, for any $n \geq 1$, any $x_1, \dots, x_n \in \mathcal{M}$, and any $\omega > 0$, there exist nonprime subfactors $\mathcal{N}_1^\omega = \mathcal{N}_1(x_1, \dots, x_n; \omega), \dots, \mathcal{N}_f^\omega = \mathcal{N}_f(x_1, \dots, x_n; \omega)$ of $\mathcal{M}$ and also a set $\mathcal{X}^\omega = \mathcal{X}(x_1, \dots, x_n; \omega) \subset \mathcal{M}$ containing p self-adjoint elements, such that

$$\mathrm{dist}_{\|\cdot\|_2}\Bigg(x_j, \sum_{\substack{1\leq j_1,\dots,j_{t+1}\leq f\\ 1\leq t\leq d}} \mathcal{N}_{j_1}^\omega \mathcal{X}^\omega \mathcal{N}_{j_2}^\omega \mathcal{X}^\omega \cdots \mathcal{N}_{j_t}^\omega \mathcal{X}^\omega \mathcal{N}_{j_{t+1}}^\omega\Bigg) < \omega \quad \text{for } j = 1, \dots, n.$$

In this situation we write

$$\mathcal{M} = \lim_{\omega\to 0}{}^{\|\cdot\|_2} \sum_{\substack{1\leq j_1,\dots,j_{t+1}\leq f\\ 1\leq t\leq d}} \mathcal{N}_{j_1}^\omega \mathcal{X}^\omega \mathcal{N}_{j_2}^\omega \mathcal{X}^\omega \cdots \mathcal{N}_{j_t}^\omega \mathcal{X}^\omega \mathcal{N}_{j_{t+1}}^\omega.$$

If $\mathcal{L}(\mathbb{F}_n)$ admitted an asymptotic decomposition over nonprime subfactors as in this definition, the situation described in Proposition 3.4 (with $\mathcal{M} = \mathcal{L}(\mathbb{F}_n)$) would take place for arbitrary $\omega > 0$, since any II_1-factor is generated by its projections of given trace ($\frac{1}{2}$, for example). The following is a result from [Ge 1998, p. 155] (see also [Kadison and Ringrose 1986, Exercise 12.4.11]); we include a proof for completeness.

Lemma 3.3. *Any type* II_1*-factor* $\mathcal{M}$ *with separable predual is generated by a countable family of projections of given trace.*

Proof. Every II_1-factor with separable predual is generated by a countable family of abelian subalgebras, so there exist abelian subalgebras $\mathcal{A}_1, \mathcal{A}_2, \dots$ of $\mathcal{M}$ generating $\mathcal{M}$ as a von Neumann algebra. If necessary, one can replace each $\mathcal{A}_n$ by a maximal abelian subalgebra of $\mathcal{M}$ containing it, hence $\mathcal{A}_n$ can assume to be a maximal abelian subalgebra of $\mathcal{M}$ for $1 \le n < \infty$. Being a maximal abelian subalgebra of a type II_1-factor, each $\mathcal{A}_n$ has no atoms and thus it is generated by a countable subset of projections of given trace. $\square$

Proposition 3.4. *Let $z_1, \dots, z_p$ be self-adjoint elements of a II_1-factor $\mathcal{M}$ and let $(\mathcal{N}_v)_{1\le v\le f}$ be a family of subfactors of $\mathcal{M}$. Assume that $\mathcal{N}_v = \mathcal{R}_1^{(v)} \vee \mathcal{R}_2^{(v)} \simeq \mathcal{R}_1^{(v)} \otimes \mathcal{R}_2^{(v)}$, where $\mathcal{R}_1^{(v)}, \mathcal{R}_2^{(v)}$ are II_1-factors, and assume that $x_1, \dots, x_n$ are self-adjoint generators of $\mathcal{M}$. Assume moreover that there exist projections $p_1^{(v)}, \dots, p_{r_v}^{(v)} \in \mathcal{R}_2^{(v)}$ and $q_1^{(v)}, \dots, q_{s_v}^{(v)} \in \mathcal{R}_1^{(v)}$ of trace $\frac{1}{2}$ and complex noncommutative polynomials $(\phi_j)_{1\le j\le n}$ of degree at most d (where $d \ge 1$ is fixed) in the variables $(z_u)_{1\le u\le p}$ such that*

$$\left\| x_j - \phi_j \left((p_i^{(v)})_{\substack{1\le i\le r_v\\ 1\le v\le f}}, (q_l^{(v)})_{\substack{1\le l\le s_v\\ 1\le v\le f}}, (z_u)_{1\le u\le p} \right) \right\|_2 < \omega \quad \text{for } j = 1, \dots, n, \tag{3}$$

where $\omega \in (0, a]$ is a given positive number, and such that in all the monomials of each ϕ_j the projections $p_i^{(v)}, q_l^{(v)}$ and $p_k^{(w)}, q_s^{(w)}$ are separated by some z_u if $v \ne w$. Then

$$\chi(x_1, \dots, x_n) \le C(n, p, a, d, f) + (n - p - 2f)\log \omega, \tag{4}$$

where $a = \max\left\{ \|x_j\|_2 + 1 \;\middle|\; 1 \le j \le n \right\}$ and $C(n, p, a, d, f)$ is a constant that depends only on n, p, a, d, f.

Proof. All variables involved are self-adjoint, so we can assume that $\phi_j = \phi_j^*$ for $j = 1, \dots, n$. Fix an integer $k_0 \ge 1$ and let $R > 0$. Suppose $\mathcal{M}_{k_0}(\mathbb{C}) \cong \mathcal{M}_1^{(v)} \subset \mathcal{R}_1^{(v)}$ and $\mathcal{M}_{k_0}(\mathbb{C}) \cong \mathcal{M}_2^{(v)} \subset \mathcal{R}_2^{(v)}$, and let $\{e_{jl}^{(v)}\}_{j,l}, \{f_{jl}^{(v)}\}_{j,l}$ be matrix units for $\mathcal{M}_1^{(v)}$ and $\mathcal{M}_2^{(v)}$ respectively. If

$$\left((A_j)_{1\le j\le n}, (G_i^{(v)})_{\substack{1\le i\le r_v\\ 1\le v\le f}}, (H_l^{(v)})_{\substack{1\le l\le s_v\\ 1\le v\le f}}, \{E_{jl}^{(v)}\}_{j,l,v}, \{F_{jl}^{(v)}\}_{j,l,v}, (Z_u)_{1\le u\le p} \right)$$

is an arbitrary microstate in the set of matricial microstates

$$\Gamma_R\left((x_j)_{1\le j\le n}, (p_i^{(v)})_{\substack{1\le i\le r_v\\ 1\le v\le f}}, (q_l^{(v)})_{\substack{1\le l\le s_v\\ 1\le v\le f}}, \{e_{jl}^{(v)}\}_{j,l,v}, \{f_{jl}^{(v)}\}_{j,l,v}, (z_u)_{1\le u\le p}; m, k, \epsilon \right)$$

and if m is large and ϵ is small enough, then

$$\left\| A_j - \phi_j \left((G_i^{(v)})_{\substack{1\le i\le r_v\\ 1\le v\le f}}, (H_l^{(v)})_{\substack{1\le l\le s_v\\ 1\le v\le f}}, (Z_u)_{1\le u\le p} \right) \right\|_2 < \omega \quad \text{for } j = 1, \dots, n.$$

Let $\delta > 0$ and write $k = k_0^2 t + w$ for some integers w, t with $0 \le w \le k_0^2 - 1$. If m, ϵ are suitably chosen, there exist $\mathcal{M}_1^{(v)} \cong \tilde{\mathcal{M}}_1^{(v)} \subset \mathcal{M}_k(\mathbb{C})$, $\mathcal{M}_2^{(v)} \cong \tilde{\mathcal{M}}_2^{(v)} \subset \mathcal{M}_k(\mathbb{C})$ (not necessarily unital inclusions) and matrix units $\{\tilde{E}_{jl}^{(v)}\}_{j,l,v} \subset \tilde{\mathcal{M}}_1^{(v)}$, $\{\tilde{F}_{jl}^{(v)}\}_{j,l,v} \subset \tilde{\mathcal{M}}_2^{(v)}$ such that

$$\|\tilde{E}_{jl}^{(v)} - E_{jl}^{(v)}\|_2 < \delta \quad \text{and} \quad \|\tilde{F}_{jl}^{(v)} - F_{jl}^{(v)}\|_2 < \delta \quad \text{for } j, l = 1, \dots, k_0,$$

and $\tilde{\mathcal{M}}_1^{(v)} \subset \big(\tilde{\mathcal{M}}_2^{(v)}\big)' \cap \mathcal{M}_k(\mathbb{C})$. The relative commutants of $\tilde{\mathcal{M}}_1^{(v)}$ and $\tilde{\mathcal{M}}_2^{(v)}$ in $\mathcal{M}_k(\mathbb{C})$ satisfy

$$\big(\tilde{\mathcal{M}}_1^{(v)}\big)' \cap \mathcal{M}_k(\mathbb{C}) \cong \big(\mathcal{M}_{k_0}(\mathbb{C}) \otimes 1 \otimes \mathcal{M}_t(\mathbb{C})\big) \oplus \mathcal{M}_w(\mathbb{C}),$$
$$\big(\tilde{\mathcal{M}}_2^{(v)}\big)' \cap \mathcal{M}_k(\mathbb{C}) \cong \big(1 \otimes \mathcal{M}_{k_0}(\mathbb{C}) \otimes \mathcal{M}_t(\mathbb{C})\big) \oplus \mathcal{M}_w(\mathbb{C}).$$

Let

$$\eta^{(v)}\big(x, \{e_{jl}^{(v)}\}_{j,l}\big) := \frac{1}{k_0} \sum_{j,l=1}^{k_0} e_{jl}^{(v)} x e_{lj}^{(v)} \in \mathbb{C}\langle X_1, \dots, X_{k_0^2+1}\rangle$$

be the polynomial in k_0^2+1 indeterminates that gives the conditional expectation $E_{(\mathcal{M}_1^{(v)})' \cap \mathcal{M}} : \mathcal{M} \to (\mathcal{M}_1^{(v)})' \cap \mathcal{M}$, that is,

$$E_{(\mathcal{M}_1^{(v)})' \cap \mathcal{M}}(x) = \eta^{(v)}\big(x, \{e_{jl}^{(v)}\}_{j,l}\big).$$

Then $G_1^{(v,1)} := \eta^{(v)}\big(G_1^{(v)}, \{\tilde{E}_{jl}^{(v)}\}_{j,l}\big) \in \big(\tilde{\mathcal{M}}_1^{(v)}\big)' \cap \mathcal{M}_k(\mathbb{C})$, and since

$$p_1^{(v)} = E_{(\mathcal{M}_1^{(v)})' \cap \mathcal{M}}(p_1^{(v)}) = \eta^{(v)}\big(p_1^{(v)}, \{e_{jl}^{(v)}\}_{j,l}\big),$$

it follows that

$$\Big|\tau_k\big((G_1^{(v,1)})^l\big) - \tau\big((p_1^{(v)})^l\big)\Big| < \delta_1 \quad \text{for all } \ l = 1, \dots, m_1$$

for any given δ_1, m_1, provided that ϵ, δ are small and m is large enough. For suitable m_1, δ_1 there exists a projection $P_1^{(v,1)} \in \big(\tilde{\mathcal{M}}_1^{(v)}\big)' \cap \mathcal{M}_k(\mathbb{C})$ of rank $\big[\frac{k_0 t + w}{2}\big]$ such that $\big\|P_1^{(v,1)} - G_1^{(v,1)}\big\|_2 < \delta_2$. Then $\big\|G_1^{(v)} - P_1^{(v,1)}\big\|_2 \le \big\|G_1^{(v)} - G_1^{(v,1)}\big\|_2 + \big\|G_1^{(v,1)} - P_1^{(v,1)}\big\|_2 < 2\delta_2$, since $\big\|G_1^{(v)} - G_1^{(v,1)}\big\|_2 < \delta_2$ for convenient m, ϵ, δ. With this procedure we can find projections

$$P_1^{(v,1)}, \dots, P_{r_v}^{(v,1)} \in \big(\tilde{\mathcal{M}}_1^{(v)}\big)' \cap \mathcal{M}_k(\mathbb{C}) \quad \text{and} \quad Q_1^{(v,1)}, \dots, Q_{s_v}^{(v,1)} \in \big(\tilde{\mathcal{M}}_2^{(v)}\big)' \cap \mathcal{M}_k(\mathbb{C}),$$

all of rank $\big[\frac{k_0 t + w}{2}\big]$, such that $\big\|G_i^{(v)} - P_i^{(v,1)}\big\|_2 < 2\delta_2$ and $\big\|H_j^{(v)} - Q_j^{(v,1)}\big\|_2 < 2\delta_2$ for all indices i, j, v. Moreover,

$$\Big\|A_j - \phi_j\Big((P_i^{(v,1)})_{\substack{1 \le i \le r_v \\ 1 \le v \le f}}, (Q_l^{(v,1)})_{\substack{1 \le l \le s_v \\ 1 \le v \le f}}, (Z_u)_{1 \le u \le p}\Big)\Big\|_2 < \omega \quad \text{for all } \ j = 1, \dots, n$$

if we choose a sufficiently small $\delta_2 > 0$. Let $\mathcal{G}_1^{(v)}(k) \subset \big(\tilde{\mathcal{M}}_1^{(v)}\big)' \cap \mathcal{M}_k(\mathbb{C})$ and $\mathcal{G}_2^{(v)}(k) \subset \big(\tilde{\mathcal{M}}_2^{(v)}\big)' \cap \mathcal{M}_k(\mathbb{C})$ be two fixed copies of the Grassmann manifold $\mathcal{G}\big(k_0 t + w, \big[\frac{k_0 t + w}{2}\big]\big)$ of projections in $\mathcal{M}_{k_0 t + w}(\mathbb{C})$ of rank $\big[\frac{k_0 t + w}{2}\big]$. There exists a unitary $U^{(v)} \in \mathcal{U}(k)$ such that

$$U^{(v)} P_1^{(v,1)} U^{(v)*}, \dots, U^{(v)} P_{r_v}^{(v,1)} U^{(v)*} \in \mathcal{G}_1^{(v)}(k),$$
$$U^{(v)} Q_1^{(v,1)} U^{(v)*}, \dots, U^{(v)} Q_{s_v}^{(v,1)} U^{(v)*} \in \mathcal{G}_2^{(v)}(k).$$

The previous inequality becomes

$$\Big\| A_j - \phi_j\Big(\big(U^{(v)}P_i^{(v,1)}U^{(v)*}\big)_{\substack{1\le i\le r_v\\ 1\le v\le f}}, \big(U^{(v)}Q_l^{(v,1)}U^{(v)*}\big)_{\substack{1\le l\le s_v\\ 1\le v\le f}}, (Z_u)_{1\le u\le p}, \big(\operatorname{Re}U^{(v)}, \operatorname{Im}U^{(v)}\big)_{1\le v\le f}\Big)\Big\|_2 < \omega$$

for all $j = 1, \dots, n$. The euclidean norm on $\mathcal{M}_k^{\mathrm{sa}}$ induces a $\mathcal{U}(k_0t+w)$-invariant metric on the manifold $\mathcal{G}\Big(k_0t+w, \Big[\frac{k_0t+w}{2}\Big]\Big)$, and if $\{P_a\}_{a\in A(k)}$ is a minimal θ-net in the manifold with respect to this metric, it follows from [Szarek 1982] that $|A(k)| \le (Ch_k/\theta)^{g_k}$, where C is a universal constant, $g_k = 2\big[\frac{k_0t+w}{2}\big]\cdot\big(k_0t+w-\big[\frac{k_0t+w}{2}\big]\big)$ is the dimension of $\mathcal{G}\big(k_0t+w, \big[\frac{k_0t+w}{2}\big]\big)$ and $h_k \le \sqrt{2k}$ is the diameter of the Grassmann manifold $\mathcal{G}\big(k_0t+w, \big[\frac{k_0t+w}{2}\big]\big)$ in $\mathcal{M}_k^{\mathrm{sa}}$. There exist $\alpha := (a_1^{(v)}, \dots, a_{r_v}^{(v)})_{1\le v\le f}$ and $\beta := (b_1^{(v)}, \dots, b_{s_v}^{(v)})_{1\le v\le f}$ with entries from $A(k)$ such that

$$\big\|P_{a_i^{(v)}}^{(v)} - U^{(v)}P_i^{(v,1)}U^{(v)*}\big\|_e \le \theta \quad \text{and} \quad \big\|P_{b_l^{(v)}}^{(v)} - U^{(v)}Q_l^{(v,1)}U^{(v)*}\big\|_e \le \theta$$

for $1 \le i \le r_v,\ 1 \le l \le s_v,\ 1 \le v \le f$. In particular, the polynomials $(\phi_j)_{1\le j\le n}$ are Lipschitz functions; hence there exists a constant $D = D\big((\phi_j)_{1\le j\le n}, R\big) > 0$ (note that $|\alpha| = r_1 + \cdots + r_f$ and $|\beta| = s_1 + \cdots + s_f$) such that

$$\begin{aligned}\big\|\phi_j(V_1, \dots, V_{|\alpha|+|\beta|+p+2f}) &- \phi_j(W_1, \dots, W_{|\alpha|+|\beta|+p+2f})\big\|_e \\ &\le D\,\big\|(V_1, \dots, V_{|\alpha|+|\beta|+p+2f}) - (W_1, \dots, W_{|\alpha|+|\beta|+p+2f})\big\|_e\end{aligned}$$

for all $1 \le j \le n$ and all

$$V_1, \dots, V_{|\alpha|+|\beta|+p+2f}, W_1, \dots, W_{|\alpha|+|\beta|+p+2f} \in \{V \in M_k \mid \|V\| \le R\}.$$

We then have

$$\begin{aligned}\Big\|A_j &- \phi_j\Big((P_a)_{a\in\alpha}, (P_b)_{b\in\beta}, (Z_u)_{1\le u\le p}, \big(\operatorname{Re}U^{(v)}, \operatorname{Im}U^{(v)}\big)_{1\le v\le f}\Big)\Big\|_e \\ &< \omega\sqrt{k} + D\,\Big\|\Big(\big(U^{(v)}P_i^{(v,1)}U^{(v)*}\big)_{\substack{1\le i\le r_v\\ 1\le v\le f}}, \big(U^{(v)}Q_l^{(v,1)}U^{(v)*}\big)_{\substack{1\le l\le s_v\\ 1\le v\le f}}, \\ &\qquad\qquad (Z_u)_{1\le u\le p}, \big(\operatorname{Re}U^{(v)}, \operatorname{Im}U^{(v)}\big)_{1\le v\le f}\Big) \\ &\qquad - \Big((P_a)_{a\in\alpha}, (P_b)_{b\in\beta}, (Z_u)_{1\le u\le p}, \big(\operatorname{Re}U^{(v)}, \operatorname{Im}U^{(v)}\big)_{1\le v\le f}\Big)\Big\|_e \\ &< \omega\sqrt{k} + D\theta\sqrt{|\alpha|+|\beta|} \\ &= 2\omega\sqrt{k},\end{aligned}$$

if we choose

$$\theta := \frac{\omega}{D}\sqrt{\frac{k}{|\alpha|+|\beta|}}.$$

Define $F_{\alpha,\beta} : (\mathcal{M}_k^{\mathrm{sa}})^{p+2f} \to (\mathcal{M}_k^{\mathrm{sa}})^n$ by

$$F_{\alpha,\beta}\Big((W_u)_{1\le u\le p}, \big(W_1^{(v)}, W_2^{(v)}\big)_{1\le v\le f}\Big) = \Big(\phi_j\big((P_a)_{a\in\alpha}, (P_b)_{b\in\beta}, (W_u)_{1\le u\le p}, (W_1^{(v)}, W_2^{(v)})_{1\le v\le f}\big)\Big)_{1\le j\le n},$$

and note that $\operatorname{dist}_e\big((A_j)_{1\le j\le n}, \operatorname{Range}F_{\alpha,\beta}\big) < 2\omega\sqrt{nk}$. Note also that all the components of $F_{\alpha,\beta}$ are polynomial functions of degrees at most $3d+2$. Now use Lemma 2.1 as in the proof of Proposition 2.2 to obtain the estimates

$$\begin{aligned}
&\operatorname{vol}_{nk^2}\Gamma_R\Big((x_j)_{1\le j\le n} : (p_i^{(v)})_{\substack{1\le i\le r_v\\ 1\le v\le f}}, (q_l^{(v)})_{\substack{1\le l\le s_v\\ 1\le v\le f}}, \{e_{jl}^{(v)}\}_{j,l,v}, \{f_{jl}^{(v)}\}_{j,l,v}, (z_u)_{1\le u\le p}; m, k, \epsilon\Big)\\
&\le \bigg(\Big(\frac{Ch_k}{\theta}\Big)^{g_k}\bigg)^{|\alpha|+|\beta|}\cdot\binom{nk^2}{(p+2f)k^2}\cdot C(d)\\
&\qquad\cdot \operatorname{vol}_{(p+2f)k^2} B\big(0, (a+2\omega)\sqrt{nk}\,\big)\cdot \operatorname{vol}_{nk^2-(p+2f)k^2} B\big(0, 2\omega\sqrt{nk}\,\big)\\
&= \bigg(\frac{CDh_k}{\omega}\sqrt{\frac{|\alpha|+|\beta|}{k}}\bigg)^{(|\alpha|+|\beta|)g_k}\cdot\binom{nk^2}{(p+2f)k^2}\cdot C(d)\\
&\qquad\cdot\frac{(\pi nk)^{(p+2f)k^2/2}(2\omega+a)^{(p+2f)k^2}}{\Gamma\Big(1+\frac{(p+2f)k^2}{2}\Big)}\cdot\frac{(\pi nk)^{(nk^2-(p+2f)k^2)/2}(2\omega)^{nk^2-(p+2f)k^2}}{\Gamma\Big(1+\frac{nk^2-(p+2f)k^2}{2}\Big)}.
\end{aligned}$$

This estimate, inequality (2) on page 373, and the inequalities $h_k\le\sqrt{2k},\ 0<\omega\le a$,

$$\begin{aligned}
g_k &= 2\Big[\frac{k_0t+w}{2}\Big]\Big(k_0t+w-\Big[\frac{k_0t+w}{2}\Big]\Big)\\
&\le 2\,\frac{k_0t+w}{2}\Big(k_0t+w-\frac{k_0t+w}{2}\Big) = \frac{(k_0t+w)^2}{2} = \frac{(k+k_0w-w)^2}{2k_0^2},
\end{aligned}$$

together with $C(d)\le(3d+2)^{(p+2f)k^2}$, imply

$$\begin{aligned}
&\operatorname{vol}_{nk^2}\Gamma_R\Big((x_j)_{1\le j\le n} : (p_i^{(v)})_{\substack{1\le i\le r_v\\ 1\le v\le f}}, (q_l^{(v)})_{\substack{1\le l\le s_v\\ 1\le v\le f}}, \{e_{jl}^{(v)}\}_{j,l,v}, \{f_{jl}^{(v)}\}_{j,l,v}, (z_u)_{1\le u\le p}; m, k, \epsilon\Big)\\
&\le \bigg(\frac{CD\sqrt{2(|\alpha|+|\beta|)}}{\omega}\bigg)^{\frac{(k+k_0w-w)^2}{2k_0^2}(|\alpha|+|\beta|)}\\
&\qquad\cdot\frac{2^{nk^2/2}(\pi nk)^{nk^2/2}(3a)^{(p+2f)k^2}(2\omega)^{(n-p-2f)k^2}}{\Gamma\Big(1+\frac{nk^2}{2}\Big)}\binom{nk^2}{(p+2f)k^2}(3d+2)^{(p+2f)k^2}.
\end{aligned}$$

Therefore

$$\frac{1}{k^2}\chi_R\Big((x_j)_{1\le j\le n}:(p_i^{(v)})_{\substack{1\le i\le r_v\\1\le v\le f}},(q_l^{(v)})_{\substack{1\le l\le s_v\\1\le v\le f}},\{e_{jl}^{(v)}\}_{j,l,v},\{f_{jl}^{(v)}\}_{j,l,v},\\(z_u)_{1\le u\le p};m,k,\epsilon\Big)+\frac{n}{2}\log k\\ \le C(n,p,a,d,f)+n\log k+\frac{|\alpha|+|\beta|}{2k_0^2}\Big(1+\frac{k_0w-w}{k}\Big)^2\log\frac{CD\sqrt{2(|\alpha|+|\beta|)}}{\omega}\\ +(n-p-2f)\log\omega-\frac{1}{k^2}\log\Gamma\Big(1+\frac{nk^2}{2}\Big)+\frac{1}{k^2}\log\binom{nk^2}{(p+2f)k^2}.$$

Use the asymptotics

$$\frac{1}{k^2}\log\binom{nk^2}{(p+2f)k^2}\\ =n\log n-(p+2f)\log(p+2f)-(n-p-2f)\log(n-p-2f)+o(1)$$

and Stirling's formula

$$\frac{1}{k^2}\log\Gamma\Big(1+\frac{nk^2}{2}\Big)=\frac{n}{2}\log\frac{nk^2}{2e}+o(1)$$

to conclude that

$$\chi_R\Big((x_j)_{1\le j\le n}:(p_i^{(v)})_{\substack{1\le i\le r_v\\1\le v\le f}},(q_l^{(v)})_{\substack{1\le l\le s_v\\1\le v\le f}},\{e_{jl}^{(v)}\}_{j,l,v},\{f_{jl}^{(v)}\}_{j,l,v},\\(z_u)_{1\le u\le p};m,\epsilon\Big)\\ \le\frac{|\alpha|+|\beta|}{2k_0^2}\log(CD\sqrt{2(|\alpha|+|\beta|)})+C(n,p,a,d,f)\\ +\Big(n-p-2f-\frac{|\alpha|+|\beta|}{2k_0^2}\Big)\log\omega. \tag{5}$$

The last inequality shows that the free entropy of $\{x_1,\dots,x_n\}$ does not exceed $C(n,p,a,d,f)+(n-p-2f)\log\omega$, since k_0 is an arbitrary integer, R is an arbitrary positive number and $x_1,\dots,x_n$ generate M. □

3.1. *Hyperfinite dimension of free group factors.*

Theorem 3.5. *If $n\ge p+2f+1$, the free group factor $\mathcal{L}(\mathbb{F}_n)$ cannot be asymptotically decomposed as*

$$\lim_{\omega\to0}{}^{\|\cdot\|_2}\sum_{\substack{1\le j_1,\dots,j_{t+1}\le f\\1\le t\le d}}\mathcal{N}_{j_1}^{\omega}\mathfrak{X}^{\omega}\mathcal{N}_{j_2}^{\omega}\mathfrak{X}^{\omega}\cdots\mathcal{N}_{j_t}^{\omega}\mathfrak{X}^{\omega}\mathcal{N}_{j_{t+1}}^{\omega},$$

where (for each ω) $\mathfrak{X}^{\omega}\subset\mathcal{L}(\mathbb{F}_n)$ contains p self-adjoint elements, $\mathcal{N}_1^{\omega},\dots,\mathcal{N}_f^{\omega}$ are nonprime subfactors of $\mathcal{L}(\mathbb{F}_n)$, and $d\ge1$ is an integer.

Proof. Suppose first that $\infty > n \geq p + 2f + 1$ and consider a semicircular system $\{x_1, \ldots, x_n\}$ that generates $\mathscr{L}(\mathbb{F}_n)$ as a von Neumann algebra. If there were a decomposition as in the theorem, one could find for every $\omega > 0$ noncommutative polynomials and projections as in Proposition 3.4 satisfying the inequalities (3). But then the estimate of the free entropy (4) would imply that $\chi(x_1, \ldots, x_n) = -\infty$ as ω tends to 0, a contradiction.

If $n = \infty$ then $\mathscr{L}(\mathbb{F}_\infty)$ is generated by an infinite semicircular system $\{x_t\}_{t\geq 1}$. If we fix an integer $k \geq p + 2f + 1$, we can approximate $x_1, \ldots, x_k$ by polynomials $(\phi_j)_{1\leq j\leq k}$ as in (3), getting the estimate of the modified free entropy (5) with k instead of n. Taking m, $1/\epsilon$, R, $k_0 \to \infty$ and $\omega \to 0$ in this estimate, one obtains

$$\chi\Big((x_j)_{1\leq j\leq k} : (p_i^{(v)})_{\substack{1\leq i\leq r_v\\ 1\leq v\leq f}}, (q_l^{(v)})_{\substack{1\leq l\leq s_v\\ 1\leq v\leq f}}, \{e_{jl}^{(v)}\}_{j,l,v}, \{f_{jl}^{(v)}\}_{j,l,v}, (z_u)_{1\leq u\leq p}\Big) < \chi(x_1, \ldots, x_k),$$

where $(p_i^{(v)})_{1\leq i\leq r_v, 1\leq v\leq f}$, $(q_l^{(v)})_{1\leq l\leq s_v, 1\leq v\leq f}$, $\{e_{jl}^{(v)}\}_{j,l,v}$, $\{f_{jl}^{(v)}\}_{j,l,v}$, and $(z_u)_{1\leq u\leq p}$ are as in Proposition 3.4. If $\mathscr{A}_t$ denotes the von Neumann algebra $\{x_1, \ldots, x_t\}''$ and E_t the conditional expectation onto it, then

$$\Big((x_j)_{1\leq j\leq k}, (E_t(p_i^{(v)}))_{\substack{1\leq i\leq r_v\\ 1\leq v\leq f}}, (E_t(q_l^{(v)}))_{\substack{1\leq l\leq s_v\\ 1\leq v\leq f}}, \{E_t(e_{jl}^{(v)})\}_{j,l,v}, \{E_t(f_{jl}^{(v)})\}_{j,l,v}, (E_t(z_u))_{1\leq u\leq p}\Big)_{t\geq 1}$$

converges in distribution as $t \to \infty$ to

$$\Big((x_j)_{1\leq j\leq k}, (p_i^{(v)})_{\substack{1\leq i\leq r_v\\ 1\leq v\leq f}}, (q_l^{(v)})_{\substack{1\leq l\leq s_v\\ 1\leq v\leq f}}, \{e_{jl}^{(v)}\}_{j,l,v}, \{f_{jl}^{(v)}\}_{j,l,v}, (z_u)_{1\leq u\leq p}\Big).$$

Therefore

$$\chi\Big((x_j)_{1\leq j\leq k} : (E_t(p_i^{(v)}))_{\substack{1\leq i\leq r_v\\ 1\leq v\leq f}}, (E_t(q_l^{(v)}))_{\substack{1\leq l\leq s_v\\ 1\leq v\leq f}}, \{E_t(e_{jl}^{(v)})\}_{j,l,v}, \{E_t(f_{jl}^{(v)})\}_{j,l,v}, (E_t(z_u))_{1\leq u\leq p}\Big) < \chi(x_1, \ldots, x_k)$$

for some large integer $t > k$. But this leads to a contradiction:

$$\begin{aligned}\chi(x_1, \ldots, x_t) &= \chi\Big((x_j)_{1\leq j\leq t} : (E_t(p_i^{(v)}))_{\substack{1\leq i\leq r_v\\ 1\leq v\leq f}}, (E_t(q_l^{(v)}))_{\substack{1\leq l\leq s_v\\ 1\leq v\leq f}}, \{E_t(e_{jl}^{(v)})\}_{j,l,v}, \{E_t(f_{jl}^{(v)})\}_{j,l,v}, (E_t(z_u))_{1\leq u\leq p}\Big)\\ &\leq \chi\Big((x_j)_{1\leq j\leq k} : (E_t(p_i^{(v)}))_{\substack{1\leq i\leq r_v\\ 1\leq v\leq f}}, (E_t(q_l^{(v)}))_{\substack{1\leq l\leq s_v\\ 1\leq v\leq f}}, \{E_t(e_{jl}^{(v)})\}_{j,l,v}, \{E_t(f_{jl}^{(v)})\}_{j,l,v}, (E_t(z_u))_{1\leq u\leq p}\Big)\\ &\quad + \chi(x_{k+1}, \ldots, x_t)\\ &< \chi(x_1, \ldots, x_k) + \chi(x_{k+1}, \ldots, x_t) = \chi(x_1, \ldots, x_t). \qquad \square\end{aligned}$$

Corollary 3.6. *If $\mathcal{P} \subset \mathcal{L}(\mathbb{F}_n)$ is a subfactor of finite index and if $n \geq p + 2f + 2$, then $\mathcal{P}$ cannot be asymptotically decomposed as*

$$\lim_{\omega \to 0} {}^{\|\cdot\|_2} \sum_{\substack{1 \leq j_1, \dots, j_{t+1} \leq f \\ 1 \leq t \leq d}} \mathcal{N}_{j_1}^{\omega} \mathfrak{X}^{\omega} \mathcal{N}_{j_2}^{\omega} \mathfrak{X}^{\omega} \cdots \mathcal{N}_{j_t}^{\omega} \mathfrak{X}^{\omega} \mathcal{N}_{j_{t+1}}^{\omega},$$

where (for each ω) $\mathfrak{X}^{\omega}$ contains p self-adjoint elements of $\mathcal{P}$, the $\mathcal{N}_1^{\omega}, \dots, \mathcal{N}_f^{\omega}$ are nonprime subfactors of $\mathcal{P}$, and $d \geq 1$ is an integer.

Proof. Since $\mathcal{P} \subset \mathcal{L}(\mathbb{F}_n)$ is a subfactor of finite index, $\mathcal{L}(\mathbb{F}_n)$ can be obtained from $\mathcal{P}$ with the basic construction [Jones 1983; Jones and Sunder 1997]: there exists a subfactor $\mathfrak{Q} \subset \mathcal{P}$ such that $\mathcal{L}(\mathbb{F}_n) = \langle \mathcal{P}, e_{\mathfrak{Q}} \rangle$, where $e_{\mathfrak{Q}}$ is the Jones projection associated to the inclusion $\mathfrak{Q} \subset \mathcal{P}$. But $\langle \mathcal{P}, e_{\mathfrak{Q}} \rangle = \mathcal{P} e_{\mathfrak{Q}} \mathcal{P}$ [Jones and Sunder 1997]; hence $\mathcal{L}(\mathbb{F}_n)$ can be decomposed as $\mathcal{P} e_{\mathfrak{Q}} \mathcal{P}$. Now apply Theorem 3.5. □

Corollary 3.7. *If $n \geq p+2f+1$, the free group factor $\mathcal{L}(\mathbb{F}_n)$ cannot be decomposed as*

$$\overline{\mathrm{sp}}^{w} \sum_{\substack{1 \leq j_1, \dots, j_{t+1} \leq f \\ 1 \leq t \leq d}} \mathcal{N}_{j_1} \mathfrak{X} \mathcal{N}_{j_2} \mathfrak{X} \cdots \mathcal{N}_{j_t} \mathfrak{X} \mathcal{N}_{j_{t+1}},$$

where $\mathfrak{X} \subset \mathcal{L}(\mathbb{F}_n)$ contains p self-adjoint elements, $\mathcal{N}_1, \dots, \mathcal{N}_f$ are nonprime subfactors of $\mathcal{L}(\mathbb{F}_n)$, and $d \geq 1$ is an integer. Moreover, if $\mathcal{P} \subset \mathcal{L}(\mathbb{F}_n)$ is a subfactor of finite index and if $n \geq p + 2f + 2$, then $\mathcal{P}$ also cannot be decomposed as

$$\overline{\mathrm{sp}}^{w} \sum_{\substack{1 \leq j_1, \dots, j_{t+1} \leq f \\ 1 \leq t \leq d}} \mathcal{N}_{j_1} \mathfrak{X} \mathcal{N}_{j_2} \mathfrak{X} \cdots \mathcal{N}_{j_t} \mathfrak{X} \mathcal{N}_{j_{t+1}},$$

for any subset $\mathfrak{X}$ containing p self-adjoint elements of $\mathcal{P}$, any $\mathcal{N}_1, \dots, \mathcal{N}_f$ nonprime subfactors of $\mathcal{P}$, and any integer $d \geq 1$.

Proof. This follows from Theorem 3.5 and Corollary 3.6, with $\mathfrak{X}^{\omega} = \mathfrak{X}$, $\mathcal{N}_1^{\omega} = \mathcal{N}_1$, $\dots$, $\mathcal{N}_f^{\omega} = \mathcal{N}_f$. □

Corollary 3.8 settles a conjecture from [Ge and Popa 1998] in the case $n = \infty$. Recall that for a type II_1-factor $\mathcal{M}$ one defines

$$\ell_h(\mathcal{M}) = \min\{ f \in \mathbb{N} \mid \exists \text{ hyperfinite } \mathcal{R}_1, \dots, \mathcal{R}_f \subset M \text{ s.t. } \overline{\mathrm{sp}}^{w} \mathcal{R}_1 \mathcal{R}_2 \cdots \mathcal{R}_f = \mathcal{M} \}.$$

Note that the definition of hyperfinite dimension is given in terms of hyperfinite subalgebras. If one defined the hyperfinite dimension in terms of hyperfinite subfactors instead of hyperfinite subalgebras, the proof of Corollary 3.8 would have followed immediately from Corollary 3.7. But with Definition 3.1, we need the asymptotic indecomposability result from Theorem 3.5.

Corollary 3.8. $\ell_h(\mathcal{L}(\mathbb{F}_n)) \geq \left[\frac{n-2}{2}\right] + 1$ *for* $4 \leq n \leq \infty$.

Proof. If $\ell_h(\mathscr{L}(\mathbb{F}_n)) \le \left[\frac{n-2}{2}\right]$, then $\mathscr{L}(\mathbb{F}_n) = \overline{\mathrm{sp}}^w \mathscr{R}_1 \mathscr{R}_2 \cdots \mathscr{R}_f$ for some hyperfinite subalgebras $\mathscr{R}_1, \ldots, \mathscr{R}_f$ and some integer f with $n \ge 2f+2$. Let $m \ge 1$, $y_1, \ldots, y_m \in \mathscr{L}(\mathbb{F}_n)$ and $\omega > 0$ be fixed. There exist finite dimensional subalgebras $\mathscr{B}_v^\omega = \mathscr{B}_v(y_1, \ldots, y_m; \omega) \subset \mathscr{R}_v$, for $1 \le v \le f$, such that

$$\mathrm{dist}_{\|\cdot\|_2}\big(y_j, \mathscr{B}_1^\omega \mathscr{B}_2^\omega \cdots \mathscr{B}_f^\omega\big) < \omega \quad \text{for } 1 \le j \le m.$$

Each finite dimensional subalgebra $\mathscr{B}_v^\omega$ is contained in a copy of the hyperfinite II_1-factor, say $\mathscr{B}_v^\omega \subset \mathscr{R}_v^\omega = \mathscr{R}_v^\omega(y_1, \ldots, y_m; \omega) \subset \mathscr{L}(\mathbb{F}_n)$. Consequently,

$$\mathrm{dist}_{\|\cdot\|_2}\big(y_j, \mathscr{R}_1^\omega \mathscr{R}_2^\omega \cdots \mathscr{R}_f^\omega\big) < \omega \quad \text{for } 1 \le j \le m;$$

hence $\mathscr{L}(\mathbb{F}_n)$ admits an asymptotic decomposition of the form

$$\lim_{\omega \to 0}{}^{\|\cdot\|_2} \mathscr{R}_1^\omega \mathscr{R}_2^\omega \cdots \mathscr{R}_f^\omega,$$

contradicting Theorem 3.5 since $\mathscr{R}_1^\omega, \ldots, \mathscr{R}_f^\omega$ are nonprime and $n \ge 2f+2$. □

Corollary 3.9. *If $\mathscr{P} \subset \mathscr{L}(\mathbb{F}_n)$ is a subfactor of finite index and $5 \le n \le \infty$, then $\ell_h(\mathscr{P}) \ge \left[\frac{n-3}{2}\right] + 1$.*

Proof. Follows from Corollary 3.6. □

4. Indecomposability over abelian subalgebras

Another estimate of free entropy is used to prove that the free group factor $\mathscr{L}(\mathbb{F}_n)$ does not admit an asymptotic decomposition of the form

$$\lim_{\omega \to 0}{}^{\|\cdot\|_2} \sum_{\substack{1 \le j_1, \ldots, j_{t+1} \le f \\ 1 \le t \le d}} \mathscr{A}_{j_1}^\omega \mathscr{X}^\omega \mathscr{A}_{j_2}^\omega \mathscr{X}^\omega \cdots \mathscr{A}_{j_t}^\omega \mathscr{X}^\omega \mathscr{A}_{j_{t+1}}^\omega,$$

where (for each ω) the $\mathscr{A}_1^\omega, \ldots, \mathscr{A}_f^\omega$ are abelian subalgebras of $\mathscr{L}(\mathbb{F}_n)$, $\mathscr{X}^\omega \subset \mathscr{L}(\mathbb{F}_n)$ is a subset with p self-adjoint elements, $d \ge 1$ is an arbitrary integer, and $n \ge p + 2f + 1$. Similarly, for free group subfactors one has the following: if $n \ge p + 2f + 2$ and $\mathscr{P} \subset \mathscr{L}(\mathbb{F}_n)$ is a subfactor of finite index, then $\mathscr{P}$ does not admit such an asymptotic decomposition either. In particular, the abelian dimension of $\mathscr{L}(\mathbb{F}_n)$ is $\ge \left[\frac{n-2}{2}\right] + 1$ and the abelian dimension of $\mathscr{P}$ is $\ge \left[\frac{n-3}{2}\right] + 1$. For $n = \infty$ this proves the second part of Ge and Popa's conjecture [Ge and Popa 1998]: the abelian dimension of free group factors is infinite. The definitions of abelian dimension and asymptotic decomposition over abelian subalgebras are given next.

Definition 4.1 [Ge and Popa 1998]. If $\mathscr{M}$ is a II_1-factor, the abelian dimension of $\mathscr{M}$, denoted $\ell_a(\mathscr{M})$, is defined as the smallest positive integer $f \in \mathbb{N}$ with the property that there exist abelian subalgebras $\mathscr{A}_1, \ldots, \mathscr{A}_f \subset \mathscr{M}$ such that $\overline{\mathrm{sp}}^w \mathscr{A}_1 \mathscr{A}_2 \cdots \mathscr{A}_f = \mathscr{M}$. If there is no such positive integer, $\ell_a(\mathscr{M}) = +\infty$.

Definition 4.2. A type II_1-factor $\mathcal{M}$ admits an asymptotic decomposition over abelian subalgebras if, for any $n \geq 1$, any $x_1, \dots, x_n \in \mathcal{M}$, and any $\omega > 0$, there exist abelian $*$-subalgebras $\mathcal{A}_1^\omega = \mathcal{A}_1(x_1, \dots, x_n; \omega), \dots, \mathcal{A}_f^\omega = \mathcal{A}_f(x_1, \dots, x_n; \omega)$ of $\mathcal{M}$ and also a set $\mathcal{Z}^\omega = \mathcal{Z}(x_1, \dots, x_n; \omega) \subset \mathcal{M}$ containing p self-adjoint elements, such that

$$\operatorname{dist}_{\|\cdot\|_2}\Biggl(x_j, \sum_{\substack{1\leq j_1,\dots,j_{t+1}\leq f\\ 1\leq t\leq d}} \mathcal{A}_{j_1}^\omega \mathcal{Z}^\omega \mathcal{A}_{j_2}^\omega \mathcal{Z}^\omega \cdots \mathcal{A}_{j_t}^\omega \mathcal{Z}^\omega \mathcal{A}_{j_{t+1}}^\omega\Biggr) < \omega \quad \text{for } 1 \leq j \leq n.$$

In this situation we write

$$\mathcal{M} = \lim_{\omega\to 0}{}^{\|\cdot\|_2} \sum_{\substack{1\leq j_1,\dots,j_{t+1}\leq f\\ 1\leq t\leq d}} \mathcal{A}_{j_1}^\omega \mathcal{Z}^\omega \mathcal{A}_{j_2}^\omega \mathcal{Z}^\omega \cdots \mathcal{A}_{j_t}^\omega \mathcal{Z}^\omega \mathcal{A}_{j_{t+1}}^\omega.$$

Proposition 4.3 gives an estimate of the free entropy of a (finite) system of generators of a II_1-factor $\mathcal{M}$ that can be asymptotically decomposed as

$$\lim_{\omega\to 0}{}^{\|\cdot\|_2} \sum_{\substack{1\leq j_1,\dots,j_{t+1}\leq f\\ 1\leq t\leq d}} \mathcal{A}_{j_1}^\omega \mathcal{Z}^\omega \mathcal{A}_{j_2}^\omega \mathcal{Z}^\omega \cdots \mathcal{A}_{j_t}^\omega \mathcal{Z}^\omega \mathcal{A}_{j_{t+1}}^\omega.$$

As in the statement of Proposition 3.4, the approximations in the $\|\cdot\|_2$-norm (6) hold for every $\omega > 0$ if the II_1-factor can be decomposed as above.

Proposition 4.3. *Let $z_1, \dots, z_p$ be self-adjoint elements of a II_1-factor $\mathcal{M}$ and let $(\mathcal{A}_v)_{1\leq v\leq f}$ be a family of abelian subalgebras of $\mathcal{M}$. Let $x_1, \dots, x_n$ be self-adjoint generators of $\mathcal{M}$ and assume that there exist projections $p_1^{(v)}, \dots, p_{r_v}^{(v)} \in \mathcal{A}_v$ and complex noncommutative polynomials $(\phi_j)_{1\leq j\leq n}$ of degree at most d (where $d \geq 1$ is fixed) in the variables $(z_u)_{1\leq u\leq p}$ such that*

$$\Bigl\| x_j - \phi_j\Bigl((p_i^{(v)})_{\substack{1\leq i\leq r_v\\ 1\leq v\leq f}}, (z_u)_{1\leq u\leq p}\Bigr)\Bigr\|_2 < \omega \quad \textit{for } j = 1, \dots, n, \tag{6}$$

where $\omega \in (0, a]$ is a given positive number, and such that in all monomials of every ϕ_j the projections $p_i^{(v)}$ and $p_k^{(w)}$ are separated by some z_u if $v \neq w$. Then

$$\chi(x_1, \dots, x_n) \leq C(n, p, a, d, f) + (n-p-2f)\log\omega, \tag{7}$$

where $a = \max\bigl\{\|x_j\|_2 + 1 \mid 1 \leq j \leq n\bigr\}$ and $C(n, p, a, d, f)$ is a constant that depends only on n, p, a, d, f.

Proof. As in the proof of Proposition 3.4 we can assume that $\phi_j = \phi_j^*$ for $1 \leq j \leq n$, and fix $R > 0$. Consider an arbitrary element

$$\Bigl((B_j)_{1\leq j\leq n}, (P_i^{(v)})_{\substack{1\leq i\leq r_v\\ 1\leq v\leq f}}, (Z_u)_{1\leq u\leq p}\Bigr)$$

of

$$\Gamma_R\Big((x_j)_{1\le j\le n},\, (p_i^{(v)})_{\substack{1\le i\le r_v\\ 1\le v\le f}},\, (z_u)_{1\le u\le p};\, m, k, \epsilon\Big)$$

for some large integers m, k and small $\epsilon > 0$. Possibly after further restricting m and ϵ, we can find mutually orthogonal projections $Q_1^{(v)}, \ldots, Q_{r_v}^{(v)} \in \mathcal{M}_k^{\mathrm{sa}}$ with rank $Q_i^{(v)} = \big[\tau(p_i^{(v)})k\big]$ for $i = 1, \ldots, r_v$, such that

$$\Big\| B_j - \phi_j\Big((Q_i^{(v)})_{\substack{1\le i\le r_v\\ 1\le v\le f}},\, (Z_u)_{1\le u\le p}\Big)\Big\|_2 < \omega \quad \text{for all } 1 \le j \le n.$$

If $S_1^{(v)}, \ldots, S_r^{(v)} \in \mathcal{M}_k^{\mathrm{sa}}$ are fixed, mutually orthogonal projections with rank $S_i^{(v)} = \big[\tau(p_i^{(v)})k\big]$ for every $1 \le i \le r_v$, then there exists a unitary $U^{(v)} \in \mathcal{U}(k)$ such that $Q_i^{(v)} = U^{(v)*} S_i U^{(v)}$ for every $1 \le i \le r_v$. The previous inequality becomes

$$\Big\| B_j - \phi_j\Big((S_i^{(v)})_{\substack{1\le i\le r_v\\ 1\le v\le f}},\, (Z_u)_{1\le u\le p},\, \big(\operatorname{Re} U^{(v)}, \operatorname{Im} U^{(v)}\big)_{1\le v\le f}\Big)\Big\|_2 < \omega,$$

and all the components of ϕ_j are polynomials of degrees $\le 3d+2$ in the last $p+2f$ variables. Reasoning as in the last part of the proof of Proposition 3.4 we can easily obtain now the estimate $\chi(x_1, \ldots, x_n) \le C(n, p, a, d, f) + (n-p-2f)\log\omega$. □

Abelian dimension of free group factors.

Theorem 4.4. *If $n \ge p + 2f + 1$, the free group factor $\mathcal{L}(\mathbb{F}_n)$ does not admit an asymptotic decomposition of the form*

$$\lim_{\omega\to 0}{}^{\|\cdot\|_2} \sum_{\substack{1\le j_1,\ldots,j_{t+1}\le f\\ 1\le t\le d}} \mathcal{A}_{j_1}^{\omega}\mathfrak{X}^{\omega}\mathcal{A}_{j_2}^{\omega}\mathfrak{X}^{\omega}\cdots\mathcal{A}_{j_t}^{\omega}\mathfrak{X}^{\omega}\mathcal{A}_{j_{t+1}}^{\omega},$$

where each subset $\mathfrak{X}^{\omega}$ contains p self-adjoint elements, $\mathcal{A}_1^{\omega}, \ldots, \mathcal{A}_f^{\omega} \subset \mathcal{L}(\mathbb{F}_n)$ are abelian $*$-subalgebras and $d \ge 1$ is an integer.

Proof. Apply Proposition 4.3 in the same manner that Proposition 3.4 was used in the proof of Theorem 3.5. □

Corollary 4.5. *If $\mathcal{P} \subset \mathcal{L}(\mathbb{F}_n)$ is a subfactor of finite index and if $n \ge p + 2f + 2$, then $\mathcal{P}$ cannot be asymptotically decomposed as*

$$\lim_{\omega\to 0}{}^{\|\cdot\|_2} \sum_{\substack{1\le j_1,\ldots,j_{t+1}\le f\\ 1\le t\le d}} \mathcal{A}_{j_1}^{\omega}\mathfrak{X}^{\omega}\mathcal{A}_{j_2}^{\omega}\mathfrak{X}^{\omega}\cdots\mathcal{A}_{j_t}^{\omega}\mathfrak{X}^{\omega}\mathcal{A}_{j_{t+1}}^{\omega},$$

where each subset $\mathfrak{X}^{\omega}$ contains p self-adjoint elements of $\mathcal{P}$, the $\mathcal{A}_1^{\omega}, \ldots, \mathcal{A}_f^{\omega} \subset \mathcal{P}$ are abelian $*$-subalgebras, and $d \ge 1$ is an integer.

Proof. This is a direct consequence of Theorem 4.4 and of decomposition $\mathcal{L}(\mathbb{F}_n) = \mathcal{P}e_{\mathcal{Q}}\mathcal{P}$ (see the proof of Corollary 3.6). □

Corollary 4.6. *If $n \geq p+2f+1$, the free group factor $\mathscr{L}(\mathbb{F}_n)$ cannot be decomposed as*

$$\overline{\mathrm{sp}}^w \sum_{\substack{1\leq j_1,\ldots,j_{t+1}\leq f\\ 1\leq t\leq d}} \mathscr{A}_{j_1}\mathscr{X}\mathscr{A}_{j_2}\mathscr{X}\cdots\mathscr{A}_{j_t}\mathscr{X}\mathscr{A}_{j_{t+1}},$$

where $\mathscr{X} \subset \mathscr{L}(\mathbb{F}_n)$ contains p self-adjoint elements, $\mathscr{A}_1, \ldots, \mathscr{A}_f$ are abelian $*$-subalgebras of $\mathscr{L}(\mathbb{F}_n)$, and $d \geq 1$ is an integer. Moreover, if $\mathscr{P} \subset \mathscr{L}(\mathbb{F}_n)$ is a subfactor of finite index and if $n \geq p+2f+2$, then $\mathscr{P}$ also cannot be decomposed as

$$\overline{\mathrm{sp}}^w \sum_{\substack{1\leq j_1,\ldots,j_{t+1}\leq f\\ 1\leq t\leq d}} \mathscr{A}_{j_1}\mathscr{X}\mathscr{A}_{j_2}\mathscr{X}\cdots\mathscr{A}_{j_t}\mathscr{X}\mathscr{A}_{j_{t+1}},$$

for any subset $\mathscr{X}$ containing p self-adjoint elements of $\mathscr{P}$, any $\mathscr{A}_1, \ldots, \mathscr{A}_f$ abelian $*$-subalgebras of $\mathscr{P}$, and any integer $d \geq 1$.

Proof. Apply Theorem 4.4 and Corollary 4.5 for $\mathscr{X}^\omega = \mathscr{X}$, $\mathscr{A}_1^\omega = \mathscr{A}_1, \ldots, \mathscr{A}_f^\omega = \mathscr{A}_f$. □

Corollary 4.7 settles the second part of the conjecture of Ge and Popa [1998], in the case $n = \infty$. As a reminder, $\ell_a(\mathscr{M})$ is defined as

$$\min\left\{ f \in \mathbb{N} \mid \exists \text{ abelian } *\text{-algebras } \mathscr{A}_1, \ldots, \mathscr{A}_f \subset \mathscr{M} \text{ s.t. } \overline{\mathrm{sp}}^w \mathscr{A}_1\mathscr{A}_2\cdots\mathscr{A}_f = \mathscr{M}\right\}$$

for every type II_1-factor $\mathscr{M}$.

Corollary 4.7. $\ell_a(\mathscr{L}(\mathbb{F}_n)) \geq \left[\frac{n-2}{2}\right]+1$ *for* $4 \leq n \leq \infty$.

Proof. This follows from the first part of Corollary 4.6 with $\mathscr{X} = \{1\}$. □

Corollary 4.8. *If $\mathscr{P} \subset \mathscr{L}(\mathbb{F}_n)$ is a subfactor of finite index and $5 \leq n \leq \infty$, then* $\ell_a(\mathscr{P}) \geq \left[\frac{n-3}{2}\right]+1$.

Proof. Apply the second part of Corollary 4.6. □

Remark 4.9. One can combine both indecomposability properties of $\mathscr{L}(\mathbb{F}_n)$ into a single statement: if $n \geq p+2f+1$, the free group factor $\mathscr{L}(\mathbb{F}_n)$ does not admit an asymptotic decomposition of the form

$$\lim_{\omega\to 0}{}^{\|\cdot\|_2} \sum_{\substack{1\leq j_1,\ldots,j_{t+1}\leq f\\ 1\leq t\leq d}} \mathscr{M}_{j_1}^\omega\mathscr{X}^\omega\mathscr{M}_{j_2}^\omega\mathscr{X}^\omega\cdots\mathscr{M}_{j_t}^\omega\mathscr{X}^\omega\mathscr{M}_{j_{t+1}}^\omega,$$

where each subset $\mathscr{X}^\omega$ contains p self-adjoint elements, each $\mathscr{M}_1^\omega, \ldots, \mathscr{M}_f^\omega \subset \mathscr{L}(\mathbb{F}_n)$ is either a nonprime subfactor or an abelian $*$-subalgebra and $d \geq 1$ is an integer.

Acknowledgment

This paper has been circulating in preprint form since July 1999. The author thanks F. Rădulescu for suggestions and many helpful conversations.

References

[Dykema 1994] K. Dykema, "Interpolated free group factors", *Pacific J. Math.* **163**:1 (1994), 123–135. MR 95c:46103 Zbl 0791.46038

[Dykema 1997] K. J. Dykema, "Two applications of free entropy", *Math. Ann.* **308**:3 (1997), 547–558. MR 98g:46099 Zbl 0927.46035

[Ge 1997] L. Ge, "Applications of free entropy to finite von Neumann algebras", *Amer. J. Math.* **119**:2 (1997), 467–485. MR 98a:46074 Zbl 0871.46031

[Ge 1998] L. Ge, "Applications of free entropy to finite von Neumann algebras, II", *Ann. of Math.* (2) **147**:1 (1998), 143–157. MR 99c:46068 Zbl 0924.46050

[Ge and Popa 1998] L. Ge and S. Popa, "On some decomposition properties for factors of type II_1", *Duke Math. J.* **94**:1 (1998), 79–101. MR 99j:46070 Zbl 0947.46042

[Ge and Shen 2000] L. Ge and J. Shen, "Free entropy and property T factors", *Proc. Natl. Acad. Sci. USA* **97**:18 (2000), 9881–9885. MR 2001g:46145 Zbl 0960.46042

[Haagerup 1978/79] U. Haagerup, "An example of a nonnuclear C^*-algebra, which has the metric approximation property", *Invent. Math.* **50**:3 (1978/79), 279–293. MR 80j:46094 Zbl 0408.46046

[Jones 1983] V. F. R. Jones, "Index for subfactors", *Invent. Math.* **72**:1 (1983), 1–25. MR 84d:46097 Zbl 0508.46040

[Jones and Sunder 1997] V. Jones and V. S. Sunder, *Introduction to subfactors*, London Mathematical Society Lecture Note Series **234**, Cambridge University Press, Cambridge, 1997. MR 98h:46067 Zbl 0903.46062

[Kadison and Ringrose 1986] R. V. Kadison and J. R. Ringrose, *Fundamentals of the theory of operator algebras*, vol. II, Pure and Applied Mathematics **100-2**, Academic Press, New York, 1986. MR 85j:46099 Zbl 0518.46046

[Murray and von Neumann 1943] F. J. Murray and J. von Neumann, "On rings of operators, IV", *Ann. of Math.* (2) **44** (1943), 716–808. MR 5,101a Zbl 0060.26903

[Popa 1983] S. Popa, "Singular maximal abelian $*$-subalgebras in continuous von Neumann algebras", *J. Funct. Anal.* **50**:2 (1983), 151–166. MR 84e:46065 Zbl 0526.46059

[Popa 1995] S. Popa, "Free-independent sequences in type II_1 factors and related problems", pp. 187–202 in *Recent advances in operator algebras* (Orléans, 1992), Astérisque **232**, Soc. Math. de France, Paris, 1995. MR 97b:46080 Zbl 0840.46039

[Rădulescu 1994] F. Rădulescu, "Random matrices, amalgamated free products and subfactors of the von Neumann algebra of a free group, of noninteger index", *Invent. Math.* **115**:2 (1994), 347–389. MR 95c:46102 Zbl 0861.46038

[Ştefan 1998] M. B. Ştefan, "The primality of subfactors of finite index in the interpolated free group factors", *Proc. Amer. Math. Soc.* **126**:8 (1998), 2299–2307. MR 98j:46067 Zbl 0896.46043

[Szarek 1982] S. J. Szarek, "Nets of Grassmann manifold and orthogonal group", pp. 169–185 in *Proceedings of research workshop on Banach space theory* (Iowa City, 1981), edited by B.-L. Lin, Univ. Iowa, Iowa City, IA, 1982. MR 85h:58021 Zbl 0526.53047

[Voiculescu 1990] D. Voiculescu, "Circular and semicircular systems and free product factors", pp. 45–60 in *Operator algebras, unitary representations, enveloping algebras, and invariant theory* (Paris, 1989), edited by A. Connes et al., Progr. Math. **92**, Birkhäuser, Boston, 1990. MR 92e:46124 Zbl 0744.46055

[Voiculescu 1994] D. Voiculescu, "The analogues of entropy and of Fisher's information measure in free probability theory, II", *Invent. Math.* **118**:3 (1994), 411–440. MR 96a:46117 Zbl 0820.60001

[Voiculescu 1996] D. Voiculescu, "The analogues of entropy and of Fisher's information measure in free probability theory, III: The absence of Cartan subalgebras", *Geom. Funct. Anal.* **6**:1 (1996), 172–199. MR 96m:46119 Zbl 0856.60012

[Voiculescu 1997] D. Voiculescu, "The analogues of entropy and of Fisher's information measure in free probability theory, IV: Maximum entropy and freeness", pp. 293–302 in *Free probability theory* (Waterloo, ON, 1995), edited by D. V. Voiculescu, Fields Inst. Commun. **12**, Amer. Math. Soc., Providence, RI, 1997. MR 97j:46071 Zbl 0960.46040

[Voiculescu 1998] D. Voiculescu, "A strengthened asymptotic freeness result for random matrices with applications to free entropy", *International Math. Research Notices* **1998**:1 (1998), 41–63. MR 2000d:46080 Zbl 0895.60004

[Voiculescu 1999] D. Voiculescu, "Free entropy dimension ≤ 1 for some generators of property T factors of type II_1", *J. Reine Angewandte Mathematik* **514** (1999), 113–118. MR 2000k:46092 Zbl 0959.46047

[Voiculescu et al. 1992] D. V. Voiculescu, K. J. Dykema, and A. Nica, *Free random variables*, CRM Monograph Series **1**, American Mathematical Society, Providence, RI, 1992. MR 94c:46133 Zbl 0795.46049

[van der Waerden 1949] B. L. van der Waerden, *Modern Algebra*, vol. I, Frederick Ungar Pub. Co., New York, 1949. MR 10,587b Zbl 0039.00902

[Weyl 1939] H. Weyl, "On the volume of tubes", *Amer. J. Math.* **61** (1939), 461–472. ZBL 0021.35503

Received October 24, 2003.

MARIUS B. ŞTEFAN
MATHEMATICS DEPARTMENT
UCLA
LOS ANGELES, CA 90095-1555
stefan@math.ucla.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 219, No. 2, 2005

STABLE REFLEXIVE SHEAVES ON SMOOTH PROJECTIVE 3-FOLDS

PETER VERMEIRE

Motivated by Hartshorne's work on curves in $\mathbb{P}^3$, we study the properties of reflexive rank-2 sheaves on smooth projective threefolds.

1. Introduction

We work over an algebraically closed field of characteristic 0.

There has been a tremendous amount of interest in recent years in the study of curves on Calabi–Yau threefolds, and especially on the general quintic in $\mathbb{P}^4$. In this note, motivated by Hartshorne's work [1978; 1980] on curves in $\mathbb{P}^3$, we study the properties of reflexive rank-2 sheaves on smooth projective threefolds.

Some similar results are obtained in [Ballico and Miró-Roig 1997] for Fano threefolds (and somewhat more generally). The greatest advantage of our results is the determination of explicit effective bounds for the third Chern class, c_3, of a reflexive sheaf (Theorem 14) and of explicit bounds for vanishing of higher cohomology and the existence of global sections (Corollary 13). In Section 3 we write out these bounds for the case of a smooth threefold hypersurface of degree d.

We refer the reader to [Hartshorne 1980] for basic properties of reflexive sheaves. Recall the following *Serre correspondence* for reflexive sheaves (the referenced result is only for $\mathbb{P}^3$, but as noted in [Hartshorne 1978, 1.1.1] the general case follows immediately from the proof):

Theorem 1 [Hartshorne 1980, 4.1]. *Let X be a smooth projective threefold, M an invertible sheaf with $H^1(X, M^*) = H^2(X, M^*) = 0$. There is a one-to-one correspondence between*

(1) *pairs $(\mathscr{F}, s)$, where $\mathscr{F}$ is a rank-2 reflexive sheaf on X with $\bigwedge^2\mathscr{F} = M$ and $s \in \Gamma(\mathscr{F})$ is a section whose zero set has codimension 2, and*

(2) *pairs (Y, ξ), where Y is a closed Cohen–Macaulay curve in X, generically a local complete intersection, and $\xi \in \Gamma(Y, \omega_Y \otimes \omega_X^* \otimes M^*)$ is a section that generates the sheaf $\omega_Y \otimes \omega_X^* \otimes M^*$ except at finitely many points.*

MSC2000: 14J60.
Keywords: reflexive sheaves, Serre correspondence, Chern classes.

Furthermore, $c_3(\mathscr{F}) = 2p_a(Y) - 2 - c_2(\mathscr{F})c_1(\omega_X) - c_2(\mathscr{F})c_1(\mathscr{F})$. $\square$

The case where $\mathscr{F}$ is locally free corresponds the curve Y being a local complete intersection. Furthermore $\omega_Y \otimes \omega_X^* \otimes M^* \cong \mathscr{O}_Y$, ξ is a nonzero section and $c_3(\mathscr{F}) = 0$. In this case we say Y is *subcanonical*.

Example 2. Suppose $X \subset \mathbb{P}^4$ is a smooth hypersurface of degree d, $Y \subset X$ a smooth rational curve. Then Y is the zero locus of a section of some rank two vector bundle V if and only if Y is a line or a plane conic in the embedding given by $\mathscr{O}_X(1)$. If Y is a line, then $\bigwedge^2 V = \mathscr{O}_X(3-d)$; if Y is a plane conic, then $\bigwedge^2 V = \mathscr{O}_X(4-d)$.

Example 3. Suppose $X \subset \mathbb{P}^4$ is a smooth hypersurface of degree d, $Y \subset X$ a smooth elliptic curve. Then Y is the zero locus of a section of some rank two vector bundle V with $\bigwedge^2 V = \mathscr{O}_X(5-d)$.

Finally, we recall some basic formulae:

Proposition 4. *Let $\mathscr{F}$ be a coherent sheaf of rank r on a smooth threefold X. Then*

$$\begin{aligned}\chi(X,\mathscr{F}) = \tfrac{1}{6}c_1(\mathscr{F})^3 - \tfrac{1}{2}c_1(\mathscr{F})c_2(\mathscr{F}) - \tfrac{1}{2}c_1(X)c_2(\mathscr{F}) + \tfrac{1}{4}c_1(X)c_1(\mathscr{F})^2 \\ + \tfrac{1}{12}c_1(X)^2c_1(\mathscr{F}) + \tfrac{1}{12}c_2(X)c_1(\mathscr{F}) + \tfrac{r}{24}c_1(X)c_2(X) + \tfrac{1}{2}c_3(\mathscr{F}).\end{aligned}$$

Note also that if $\mathscr{F}$ has rank two and L is an invertible sheaf, then

(1) $c_1(\mathscr{F} \otimes L) = c_1(\mathscr{F}) + 2c_1(L)$,

(2) $c_2(\mathscr{F} \otimes L) = c_2(\mathscr{F}) + c_1(L)c_1(\mathscr{F}) + c_1(L)^2$,

(3) $c_3(\mathscr{F} \otimes L) = c_3(\mathscr{F})$.

2. Stability and Boundedness

Definition 5. Let L be a very ample line bundle on a smooth projective variety X. A reflexive coherent sheaf $\mathscr{F}$ on X is L*-semistable* if for every coherent subsheaf $\mathscr{F}'$ of $\mathscr{F}$ with $0 < \operatorname{rank} \mathscr{F}' < \operatorname{rank} \mathscr{F}$, we have $\mu(\mathscr{F}', L) \le \mu(\mathscr{F}, L)$, where

$$\mu(\mathscr{F}, L) = \frac{c_1(\mathscr{F}).[L]^{\dim X - 1}}{(\operatorname{rank}\mathscr{F})\,[L]^{\dim X}}.$$

If the inequality is strict, $\mathscr{F}$ is L*-stable*. Note that if $\operatorname{rank}\mathscr{F} = 2$, it suffices to take $\mathscr{F}'$ invertible.

Definition 6. We say that a reflexive sheaf $\mathscr{F}$ is *normalized with respect to* L if $-1 < \mu(\mathscr{F}, L) \le 0$. As L is typically fixed, we usually say simply that $\mathscr{F}$ is *normalized*. Note that since $\mu(\mathscr{F} \otimes L, L) = \mu(\mathscr{F}, L) + 1$, there exists, for any fixed $\mathscr{F}$, a unique $k \in \mathbb{Z}$ such that $\mathscr{F} \otimes L^k$ is normalized with respect to L.

For a fixed X, our goal is to give a bound on $c_3(\mathscr{F})$ in terms of $c_1(\mathscr{F})$ and $c_2(\mathscr{F})$. Note that the formula for c_3 in Theorem 1 gives:

Lemma 7. *Let X be a smooth threefold, L a very ample line bundle, $\mathscr{F}$ a rank two reflexive sheaf, $\bigwedge^2 \mathscr{F} = M$ a line bundle with $H^1(M^*) = H^2(M^*) = 0$. If $s \in \Gamma(\mathscr{F})$ is a section whose zero locus is a curve, then*

$$c_3(\mathscr{F}) \leq d^2 - 3d - c_2(\mathscr{F})c_1(\omega_X) - c_2(\mathscr{F})c_1(\mathscr{F})$$

where $d = c_2(\mathscr{F})c_1(L)$.

Proof. In light of Theorem 1, we need only note that the degree of the curve section in the embedding given by L is $d = c_2(\mathscr{F})c_1(L)$. The fact that $2p_a(Y) - 2 \leq d^2 - 3d$ is just the bound coming from the degree of a plane curve. □

The idea now is: given a very ample line bundle L, bound the twist of $\mathscr{F}$ by L^r needed to produce a section, and then use the bound in Lemma 7. First note the following elementary result:

Lemma 8. *Let $\mathscr{F}$ be a reflexive sheaf on a smooth projective variety X with a very ample line bundle L. If either*

(1) *$\mathscr{F}$ is L-stable and $\mu(\mathscr{F}, L) \leq 0$ or*

(2) *$\mathscr{F}$ is L-semistable and $\mu(\mathscr{F}, L) < 0$*

then $H^0(X, \mathscr{F}) = 0$.

Proof. Suppose otherwise that $\mathscr{F}$ has a section $\mathcal{O}_X \to \mathscr{F}$. Dualizing, we get a surjection $\mathscr{F}^* \to \mathscr{I}_Y \subset \mathcal{O}_X$; dualizing again we have $0 \to \mathscr{I}_Y^* \to \mathscr{F}$, but $\mathscr{I}_Y^*$ is invertible and $H^0(X, \mathscr{I}_Y^*) = \operatorname{Hom}_{\mathcal{O}_X}(\mathscr{I}_Y, \mathcal{O}_X) \neq 0$. Hence $\mu(\mathscr{I}_Y^*, L) \geq 0$ and the result follows. □

The main technical result is:

Proposition 9. *Let X be a smooth projective threefold with very ample line bundle L and with $\operatorname{Pic} X = \mathbb{Z}L$. Let $\mathscr{F}$ be a normalized L-semistable rank-2 reflexive sheaf, and D be a general member of the linear system $|L|$. Assume that the general member of the linear system $|L \otimes \mathcal{O}_D|$ is not rational, and that $m < 0$ is an integer satisfying*

$$2m < 3\mu(\Theta_X, L) - 2\mu(\mathscr{F}, L) - 2.$$

Then $H^0(\mathscr{F}_D(mD)) = 0$.

Remark 10. The assumption that the general member of the linear system $|L \otimes \mathcal{O}_D|$ is not rational can be dropped if we require that

$$2m < 3\mu(\Theta_X, L) - 2\mu(\mathscr{F}, L) - 4.$$

As this would impact all further estimates, we have chosen to add the extra hypothesis rather than explicitly keeping track of the two separate cases. The interested reader will have little trouble altering the bounds in subsequent arguments in cases where this is of interest (say a threefold quadric hypersurface).

Proof of Proposition 9. We proceed by contradiction. Let m be the smallest integer, if one exists, satisfying the inequality and such that $H^0(\mathscr{F}_D(mD))$ is nonzero for the general, hence every, member of $|L|$. We will show $m \geq 0$.

Fix a smooth member D such that $\mathscr{F}_D$ is locally free. The proposed section yields a sequence

$$0 \to \mathbb{O}_D \to \mathscr{F}_D(mD) \to \mathscr{I}_Z(2mD) \otimes \bigwedge^2 \mathscr{F} \to 0,$$

where $Z \subset D$ is zero-dimensional of length $c_2(\mathscr{F}(mD))$. Choose a smooth curve C in the system $|L_D|$ (i.e., in the class of $D.[L]$) with $Z \cap C$ empty. Tensoring the sequence above by $\mathbb{O}_C$ yields an extension of line bundles.

The class of the extension lies in

$$\mathrm{Ext}^1_{\mathbb{O}_C}\big(\mathbb{O}_C, \mathbb{O}_C(-2mD) \otimes \bigwedge^2 \mathscr{F}^*\big) = H^1\big(C, \mathbb{O}_C(-2mD) \otimes \bigwedge^2 \mathscr{F}^*\big).$$

Note that $K_C = K_X \otimes \mathbb{O}_C(2D)$. Now, as the inequality in the hypotheses is easily seen to be equivalent to

$$-2m[L]^3 - c_1(\mathscr{F}).[L]^2 > K_X.[L]^2 + 2[L]^3 = 2g(C) - 2$$

the extension group vanishes, hence

$$\mathscr{F}_C(mD) = \mathbb{O}_C \oplus \big[\mathbb{O}_C(2mD) \otimes \bigwedge^2 \mathscr{F}\big]$$

and $h^0(C, \mathscr{F}_C(mD)) = 1$. By minimality of m, we see also that

$$h^0(D, \mathscr{F}_D(mD)) = 1.$$

Now blow up X along C, and consider $\pi : \mathrm{Bl}_C(X) \to X$. We have a morphism $f : \mathrm{Bl}_C(X) \to \mathbb{P}^1$ given by the pencil of divisors in $|L|$ containing C. It is easy to see that for *every* one of these divisors, $h^0(\mathscr{F}_D(mD)) = 1$. Then, because $\pi^*\mathscr{F}$ is reflexive, $f_*\pi^*\mathscr{F}(mD)$ is invertible [Hartshorne 1980, 1.4,1.7]. However, we know that $H^0(\mathscr{F}_D(mD)) \to H^0(\mathscr{F}_C(mD))$ is an isomorphism and therefore $f_*\pi^*\mathscr{F}(mD) \cong f_*\pi^*\mathscr{F}_C(mD) \cong \mathbb{O}_{\mathbb{P}^1}$, where the last isomorphism follows directly from the splitting of $\mathscr{F}_C$.

Consequently, $H^0(X, \mathscr{F}(mD)) \neq 0$ and so $m \geq 0$ by Lemma 8, contradicting the assumption that m is negative. □

Corollary 11. *With notation and hypotheses as is Proposition 9, if*

$$2k > max\big\{0, 2 + 2\mu(\mathscr{F}, L) - 3\mu(\Theta_X, L)\big\},$$

then $H^2\big(D, K_D \otimes \mathscr{F}_D^*(kD)\big) = 0$ *for the general member* D. *If, furthermore,* k *is such that*

$$(6k^2+6k+2)-(6k+3)\big(2\mu(\mathscr{F}, L)+3\mu(\Theta_X, L)\big) \geq \frac{\big(6c_2(\mathscr{F})-c_1(X)^2-3c_1(\mathscr{F})^2-3c_1(\mathscr{F})c_1(X)-c_2(X)\big)[L]}{[L]^3},$$

then $H^0\big(D, K_D \otimes \mathscr{F}_D^*(kD)\big) \neq 0$.

Proof. We can choose D smooth and so that $\mathscr{F}_D$ is locally free. Then

$$h^2\big(D, K_D \otimes \mathscr{F}_D^*(kD)\big) = h^0\big(D, \mathscr{F}_D(-kD)\big),$$

which is zero by Proposition 9.

Because of the vanishing of H^2 above, the second part follows directly from a computation of the Euler characteristic. □

Corollary 12. *With notation and hypotheses as is Proposition 9 there exists a constant* ρ *depending on* $c_1(\mathscr{F})$, $c_2(\mathscr{F})$, $c_1(L)$ *and* $c_i(\Theta_X)$ *such that if* $r \geq \rho$ *then* $H^1\big(D, K_D \otimes \mathscr{F}_D^*(rD)\big) = 0$.

Proof. By the previous corollary, there is a constant depending on the above parameters such that if k is larger than that constant, then $K_D \otimes \mathscr{F}_D^*(kD)$ has a section. Choosing the *smallest* such integer k we have a sequence

$$0 \to \mathbb{O}_D \to K_D \otimes \mathscr{F}_D^*(kD) \to \mathscr{I}_Z(2kD) \otimes K_D^2 \otimes \textstyle\bigwedge^2 \mathscr{F}^* \to 0,$$

where, as above, $Z \subset D$ is zero-dimensional of length

$$\ell = c_2(\mathscr{F}_D) - \big(c_1(K_D)+kc_1(\mathbb{O}(D))\big)c_1(\mathscr{F}_D) + \big(c_1(K_D)+kc_1(\mathbb{O}(D))\big)^2.$$

Let $\alpha \in \mathbb{Z}$ be such that $K_X^2 \otimes \bigwedge^2 \mathscr{F}^* = L^\alpha$. Because D is a smooth surface, $H^1(D, \mathbb{O}(pD)) = 0$ for $p \geq 3c_1(L)^3 - 5$ (by [Bertram et al. 1991, 1.10], for instance). Further, by the standard uniform regularity result [Mumford 1966, p.103], $H^1\big(D, \mathscr{I}_Z((2k+t)D) \otimes K_D^2 \otimes \bigwedge^2 \mathscr{F}^*\big)$ vanishes for $t \geq \ell - 2k - \alpha - 2$ and $t \geq 3c_1(L)^3 - 7 - 2k - \alpha$.

Consequently, $H^1\big(D, K_D \otimes \mathscr{F}_D^*(rD)\big) = 0$ for

$$r \geq max\,\{\ell - k - \alpha - 2,\ 3c_1(L)^3 - 7 - k - \alpha\}. \qquad \square$$

Corollary 13. *With notation and hypotheses as in Proposition 9, there exists an integer* ρ_2 *depending on* $c_1(\mathscr{F})$, $c_2(\mathscr{F})$, $c_1(L)$ *and* $c_i(\Theta_X)$ *such that if* $r \geq \rho_2$ *then* $H^0(X, K_X \otimes \mathscr{F}^* \otimes L^r) \neq 0$.

Proof. The vanishing of H^1 and H^2 on D described in the corollaries above gives $H^2(X, K_X \otimes \mathscr{F}^* \otimes L^r) = 0$. The result now follows by another Euler characteristic argument (see Proposition 4). □

Theorem 14. *Let X be a smooth projective threefold with very ample line bundle L and with* $\operatorname{Pic} X = \mathbb{Z}L$. *Let $\mathscr{F}$ be an L-semistable rank-2 reflexive sheaf. Then there exists an integer C depending on $c_1(\mathscr{F})$, $c_2(\mathscr{F})$, $c_1(L)$ and $c_i(\Theta_X)$ such that $C \geq c_3(\mathscr{F})$.*

Proof. As $c_3(\mathscr{F})$ is unaffected by twisting by a line bundle, we may assume that $\mathscr{F}$ is normalized. The preceding results apply and we can take a section of $K_X \otimes \mathscr{F}^* \otimes L^k$ for some k, bounded as in Corollary 13. We then have an exact sequence

$$0 \to \mathscr{O}_X \to K_X \otimes \mathscr{F}^* \otimes L^k \to \mathscr{I}_Y \otimes K_X^2 \otimes L^{2k} \otimes \textstyle\bigwedge^2 \mathscr{F}^* \to 0,$$

where $Y \subset X$ is a curve. Computing Euler characteristics gives

$$2p_a(Y) - 2 = d_1 d_2 + c_3(\mathscr{F}) + c_1(\omega_X) d_2,$$

where

$$d_1 = c_1(K_X \otimes \mathscr{F}^* \otimes L^k) = -c_1(\mathscr{F}) - 2c_1(X) + 2kc_1(L)$$

and

$$\begin{aligned} d_2 &= c_2(K_X \otimes \mathscr{F}^* \otimes L^k) \\ &= c_2(\mathscr{F}) + c_1(\mathscr{F})c_1(X) - kc_1(\mathscr{F})c_1(L) + c_1(X)^2 - 2kc_1(X)c_1(L) + k^2 c_1(L)^2. \end{aligned}$$

In the embedding determined by L, the degree of the curve Y is precisely $d_2 c_1(L)$. This implies $d_2 c_1(L)(d_2 c_1(L) - 3) \geq 2p_a(Y) - 2$ and so

$$d_2 c_1(L)(d_2 c_1(L) - 3) - d_1 d_2 - d_2 c_1(\omega_X) \geq c_3(\mathscr{F}). \qquad \square$$

3. Explicit bounds

Let X be a smooth hypersurface in $\mathbb{P}^4$ of degree $d > 2$, and $\mathscr{F}$ a rank two L-semistable reflexive sheaf. In this case, we have $K_X = \mathscr{O}_X(d-5)$; since L-semistability is independent of the choice of L, we take $L = \mathscr{O}(1)$. Note that $[L]^3 = d$, that $c_2(\Theta_X) = (10 - 5d + d^2)c_1(L)^2$, and that $\mu(\Theta_X, L) = \frac{1}{2}(5-d)$. Further, if $\mathscr{F}$ is normalized then $\mu(\mathscr{F}, L) = 0$ or $\mu(\mathscr{F}, L) = -\frac{1}{2}$. We explicitly compute the bound in the case $\mu(\mathscr{F}, L) = 0$, the other case being exactly analogous, though a bit more notationally cluttered. For notational convenience we let $S = c_2(\mathscr{F})c_1(L)$.

The first bound in Corollary 11 becomes

$$k > \max\left\{0, \tfrac{1}{4}(3d - 11)\right\},$$

so here it suffices to take $k > 0$ if $d < 5$ and $k > \frac{1}{4}(3d-11)$ if $d \geq 5$.

The second bound in Corollary 11 becomes

$$(6k^2 + 6k + 2) + \frac{(6k+3)(3d-15)}{2} - \frac{(6S - 35d + 15d^2 - 2d^3)}{d} > 0;$$

hence

$$k > \frac{-3d^2+13d+\sqrt{11d^4-150d^3+391d^2+48dS}}{4d}$$

when $S \geq \frac{1}{48}(-11d^3+150d^2-391d)$, otherwise the second bound in Corollary 11 is unnecessary.

In Corollary 12, note that $K_D = \mathbb{O}_D(d-4)$ and that the bound for p is irrelevant since the vanishing holds already for $p=0$. The length of Z is at most

$$S + d(d-4+k)^2,$$

so for the vanishing $H^1\big(D, K_D \otimes \mathscr{F}_D^*(rD)\big) = 0$ we need

$$2r \geq S + d(d-4+k)^2 - 2(d-4).$$

In Corollary 13, we compute the Euler characteristic of $\mathscr{F}^*(m)$ and take $m \geq r$ such that $\chi(K_X \otimes \mathscr{F}^*(m)) - \frac{1}{2}(c_3(\mathscr{F})) > 0$. We have

$$\chi(K_X \otimes \mathscr{F}^*(m)) - \tfrac{1}{2}(c_3(\mathscr{F})) \\ = \tfrac{1}{12}(2m+d-5)(d^3+2md^2-5d^2-10md+10d+2m^2d-6S);$$

hence we need

$$m > \frac{-d^2+5d+\sqrt{5d^2-d^4+12dS}}{2d}.$$

As before, this bound is irrelevant unless $S \geq \frac{1}{12}(d^3-5d)$.

For example, in the case of the quintic we obtain

(1) for $S \geq 13$:

$$256c_3(\mathscr{F}) < \big(320S^2+80S\sqrt{60S-525}-4004S-540\sqrt{60S-525}+11955\big) \\ \times\big(320S^2+80S\sqrt{60S-525}-4068S-548\sqrt{60S-525}+12339\big);$$

(2) for $S < 13$:

$$16c_3(\mathscr{F}) < (5S^2+184S+1620)(5S^2+180S+1536).$$

Acknowledgments

I thank Sheldon Katz for bringing this approach to the study of curves on projective threefolds to my attention, as well as for providing me with help in using Schubert [Katz and Stromme n.d.].

References

[Ballico and Miró-Roig 1997] E. Ballico and R. M. Miró-Roig, “Rank 2 stable vector bundles on Fano 3-folds of index 2”, *J. Pure Appl. Algebra* **120**:3 (1997), 213–220. MR 98h:14050 ZBL 0907.14017

[Bertram et al. 1991] A. Bertram, L. Ein, and R. Lazarsfeld, “Vanishing theorems, a theorem of Severi, and the equations defining projective varieties”, *J. Amer. Math. Soc.* **4**:3 (1991), 587–602. MR 92g:14014 Zbl 0762.14012

[Hartshorne 1978] R. Hartshorne, “Stable vector bundles of rank 2 on $\mathbf{P}^3$”, *Math. Ann.* **238**:3 (1978), 229–280. MR 80c:14011 Zbl 0411.14002

[Hartshorne 1980] R. Hartshorne, “Stable reflexive sheaves”, *Math. Ann.* **254**:2 (1980), 121–176. MR 82b:14011 Zbl 0431.14004

[Katz and Stromme n.d.] S. Katz and S. A. Stromme, “Schubert: a Maple package for intersection theory”, Available at http://www.mi.uib.no/~stromme/schubert/.

[Mumford 1966] D. Mumford, *Lectures on curves on an algebraic surface*, Annals of Mathematics Studies **59**, Princeton Univ. Press, Princeton, 1966. MR 35 #187

Received September 15, 2003. Revised July 24, 2004.

PETER VERMEIRE
DEPARTMENT OF MATHEMATICS
CENTRAL MICHIGAN UNIVERSITY
MOUNT PLEASANT, MI 48859
UNITED STATES
verme1pj@cmich.edu

CONTENTS

Volume 219, no. 1 and no. 2

Guidelines for Authors

Authors may submit manuscripts for publication to any of the editors. Submission of a manuscript acknowledges that the paper is original and has not been submitted elsewhere. Information regarding the preparation of manuscripts is provided below. For further information, 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 submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Electronic submissions are strongly encouraged; the editors' email addresses are listed inside the front cover. Papers submitted in hard copy should be sent in triplicate and authors should keep a copy.

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 Web site 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 219 No. 2 April 2005

0030-8730(200504)219:2;1-H