Pacific Journal of Mathematics

Volume 256 No. 2 April 2012

PACIFIC JOURNAL OF MATHEMATICS

http://pacificmath.org

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

See inside back cover or pacificmath.org for submission instructions.

The subscription price for 2012 is US $420/year for the electronic version, and $485/year for print and electronic.
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 94704-0163, 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, PASCAL CNRS Index, Referativnyi Zhurnal, 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 94704-0163.

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

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. 256, No. 2, 2012

$\mathbb{O}$-OPERATORS ON ASSOCIATIVE ALGEBRAS AND ASSOCIATIVE YANG–BAXTER EQUATIONS

CHENGMING BAI, LI GUO AND XIANG NI

An $\mathbb{O}$-operator on an associative algebra is a generalization of a Rota–Baxter operator that plays an important role in the Hopf algebra approach of Connes and Kreimer to the renormalization of quantum field theory. It is also the associative analog of an $\mathbb{O}$-operator on a Lie algebra in the study of the classical Yang–Baxter equation. We introduce the concept of an extended $\mathbb{O}$-operator on an associative algebra whose Lie algebra analog has been applied to generalized Lax pairs and PostLie algebras. We study algebraic structures coming from extended $\mathbb{O}$-operators. Continuing the work of Aguiar deriving Rota–Baxter operators from the associative Yang–Baxter equation, we show that its solutions correspond to extended $\mathbb{O}$-operators through a duality. We also establish a relationship of extended $\mathbb{O}$-operators with the generalized associative Yang–Baxter equation.

1. Introduction

1a. ***Motivation.*** The interaction between studies in pure mathematics and mathematical physics has long been a rich source of inspirations that benefited both fields. One such instance can be found in the seminal work of Connes and Kreimer [Connes and Kreimer 2000; Kreimer 1999] on their Hopf algebra approach to the renormalization of quantum field theory. There a curious algebraic identity of linear operators appeared that turned out to be investigated concurrently in the contexts of operads, associative Yang–Baxter equation [Aguiar 2000a; 2000b; 2001], and commutative algebra [Guo and Keigher 2000a; 2000b; Guo 2000], under the name of the Baxter identity (later called the Rota–Baxter identity). It originated in the probability study of G. Baxter [1960] and was influenced by the combinatorial interests of G.-C. Rota [1969a; 1969b; 1995]. Connes and Kreimer's discovery of the connection between Rota–Baxter operators and quantum field theory inspired

C. Bai is supported by NSFC grants 10621101 and 10920161, NKBRPC grant 2006CB805905 and SRFDP grant 200800550015. L. Guo is supported by NSF grant DMS 0505445 and DMS-1001855 and thanks the Chern Institute of Mathematics at Nankai University for hospitality.
Li Guo is the corresponding author.
MSC2010: primary 16W99, 17A30; secondary 57R56.
Keywords: $\mathbb{O}$-operator, Rota–Baxter operator, Yang–Baxter equation, bimodule.

numerous studies to better understand the role played by the Rota–Baxter identity in quantum field theory renormalization, as well as in applying the idea of renormalization to study divergency in mathematics [Ebrahimi-Fard et al. 2004; Ebrahimi-Fard et al. 2006; Guo and Zhang 2008; Manchon and Paycha 2010].

In this paper we consider a generalization of the Rota–Baxter operator in the relative context called the $\mathbb{O}$-operator. It came from another connection between Rota–Baxter operators (on Lie algebras) and mathematical physics. In special cases, the Rota–Baxter identity for Lie algebras coincides with the operator form of the classical Yang–Baxter equation, named after the well-known physicists Yang [1967] and Baxter [1972]. The connection has its origin in the work of Semenov-Tyan-Shanskiĭ [1983] and its extension led to the concept of $\mathbb{O}$-operators [Bai 2007; Bordemann 1990; Kupershmidt 1999]. The relation defining an $\mathbb{O}$-operator was also called the Schouten curvature by Kosmann-Schwarzbach and Magri [1988], and is the algebraic version of the contravariant analog of the Cartan curvature of the Lie algebra-valued one-form on a Lie group.

Back to associative algebras, the first connection between Rota–Baxter operators and an associative analog of the classical Yang–Baxter equation was made by Aguiar [2000a; 2000b], who showed that a solution of the associative Yang–Baxter equation (AYBE) gives rise to a Rota–Baxter operator of weight zero.

Our study of this connection in this paper was motivated by the $\mathbb{O}$-operator approach to the classical Yang–Baxter equation, but we go beyond what was known in the Lie algebra case. On one hand, we generalize the concept of a Rota–Baxter operator to that of an $\mathbb{O}$-operator (of any weight)[1] and further to extended $\mathbb{O}$-operators. On the other hand, we investigate the operator properties of the associative Yang–Baxter equation motivated by the study in the Lie algebra case. Through this approach, we show that the operator property of solutions of the associative Yang–Baxter equation is to a large extent characterized by $\mathbb{O}$-operators. This generalization in the associative context, motivated by Lie algebra studies, has in turn motivated us to establish a similar generalization for Lie algebras and to apply it to generalized Lax pairs, classical Yang–Baxter equations and PostLie algebras [Bai et al. 2010b; 2011; Vallette 2007].

Our approach connects (extended) $\mathbb{O}$-operators to solutions of the AYBE and its generalizations, and therefore [Bai 2010] to the construction of antisymmetric infinitesimal bialgebras and their related Frobenius algebras. The latter plays an important role in topological field theory [Runkel et al. 2007]. In particular, we are able to reverse the connection made by Aguiar and derive, from a Rota–Baxter

[1] In the weight zero case, this has been considered by Uchino [2008] under the name "generalized Rota–Baxter operator". In the general case, the term "relative Rota–Baxter operator" is also used [Bai et al. 2010a].

operator of any weight, a solution of the AYBE and hence give an antisymmetric infinitesimal bialgebra. For further details, see [Bai et al. 2012, Section 4].

1b. *Rota–Baxter algebras and Yang–Baxter equations.*

Notation. In the rest of this paper, $\Bbbk$ denotes a field. By an algebra we mean an associative (not necessarily unitary) $\Bbbk$-algebra, unless otherwise stated.

Definition 1.1. Let R be a $\Bbbk$-algebra and let $\lambda \in \Bbbk$ be given. If a $\Bbbk$-linear map $P : R \to R$ satisfies the *Rota–Baxter relation*

$$P(x)P(y) = P(P(x)y) + P(xP(y)) + \lambda P(xy) \quad \text{for all } x, y \in R, \tag{1-1}$$

then P is called a *Rota–Baxter operator of weight* λ and (R, P) is called a *Rota–Baxter algebra of weight* λ.

For simplicity, we will only discuss the case of Rota–Baxter operators of weight zero in the introduction.

Relation (1-1) still makes sense when R is replaced by a $\Bbbk$-module with any binary operation. If the binary operation is the Lie bracket and if the Lie algebra is equipped with a nondegenerate symmetric invariant bilinear form, then a skew-symmetric solution of the *classical Yang–Baxter equation*

$$[r_{12}, r_{13}] + [r_{12}, r_{23}] + [r_{13}, r_{23}] = 0. \tag{1-2}$$

is just a Rota–Baxter operator of weight zero. We refer the reader to [Bai 2007; Ebrahimi-Fard 2002; Semenov-Tyan-Shanskiĭ 1983] for further details.

We will consider the following associative analog of the classical Yang–Baxter equation (1-2).

Definition 1.2. Let A be a $\Bbbk$-algebra. An element $r \in A \otimes A$ is called a *solution of the associative Yang–Baxter equation in* A if it satisfies the relation

$$r_{12}r_{13} + r_{13}r_{23} - r_{23}r_{12} = 0, \tag{1-3}$$

called the *associative Yang–Baxter equation (AYBE)*. Here, for $r = \sum_i a_i \otimes b_i \in A \otimes A$, we denote

$$r_{12} = \sum_i a_i \otimes b_i \otimes 1, \quad r_{13} = \sum_i a_i \otimes 1 \otimes b_i, \quad r_{23} = \sum_i 1 \otimes a_i \otimes b_i. \tag{1-4}$$

Both (1-3) and the associative analog

$$r_{13}r_{12} - r_{12}r_{23} + r_{23}r_{13} = 0 \tag{1-5}$$

of (1-2) were introduced by Aguiar [2000a; 2000b; 2001]. In fact, (1-3) is just (1-5) in the opposite algebra [Aguiar 2001]. When r is skew-symmetric it is easy to see that (1-3) comes from (1-5) under the operation $\sigma_{13}(x \otimes y \otimes z) = z \otimes y \otimes x$.

While (1-5) was emphasized by Aguiar in the works above, we will work with (1-3) for notational convenience and to be consistent with some of the earlier works on connections with antisymmetric infinitesimal bialgebras [Bai 2010] and associative D-bialgebras [Zhelyabin 1997].

Theorem 1.3 [Aguiar 2000b]. *Let A be a $\Bbbk$-algebra. If $r = \sum_i a_i \otimes b_i \in A \otimes A$ is a solution of* (1-5) *in A, the map*

$$P : A \to A, \quad x \mapsto \sum_i a_i x b_i$$

defines a Rota–Baxter operator of weight zero on A.

The theorem is obtained by replacing the tensor symbols in

$$r_{13}r_{12} - r_{12}r_{23} + r_{23}r_{13} = \sum_{i,j} a_i a_j \otimes b_j \otimes b_i - \sum_{i,j} a_i \otimes b_i a_j \otimes b_j + \sum_{i,j} a_j \otimes a_i \otimes b_i b_j = 0$$

by x and y in A.

1c. ***𝕆-operators.*** We will introduce an extended $\mathbb{O}$-operator as a generalization of a Rota–Baxter operator and the associative analog of an $\mathbb{O}$-operator on a Lie algebra. We then extend the connections of Rota–Baxter algebras with associative Yang–Baxter equations to those of $\mathbb{O}$-operators. This study is motivated by the relationship between $\mathbb{O}$-operator and the classical Yang–Baxter equation in Lie algebras [Bai 2007; Bai et al. 2010b; Bordemann 1990; Kupershmidt 1999]

Let $(A, \cdot)$ be a $\Bbbk$-algebra. Let (V, ℓ, r) be an A-bimodule, consisting of a compatible pair of a left A-module (V, ℓ) given by $\ell : A \to \mathrm{End}_{\Bbbk}(V)$ and a right A-module (V, r) given by $r : A \to \mathrm{End}_{\Bbbk}(V)$; see Section 2a for the precise definition. Fix a $\kappa \in \Bbbk$. A pair (α, β) of linear maps $\alpha, \beta : V \to A$ is called an *extended $\mathbb{O}$-operator with modification β of mass κ* if

$$\kappa \ell(\beta(u))v = \kappa u r(\beta(v)) \quad \text{and}$$
$$\alpha(u) \cdot \alpha(v) - \alpha(l(\alpha(u))v + ur(\alpha(v))) = \kappa \beta(u) \cdot \beta(v) \quad \text{for all } u, v \in V.$$

When $\beta = 0$ or $\kappa = 0$, we obtain the concept of an *$\mathbb{O}$-operator α* satisfying

$$\alpha(u) \cdot \alpha(v) = \alpha(\ell(\alpha(u))v) + \alpha(ur(\alpha(v))) \quad \text{for all } u, v \in V. \tag{1-6}$$

When V is taken to be the A-bimodule (A, L, R), where $L, R : A \to \mathrm{End}_{\Bbbk}(A)$ are given by the left and right multiplications, an $\mathbb{O}$-operator $\alpha : V \to A$ of weight zero is just a Rota–Baxter operator of weight zero. To illustrate the close relationship between $\mathbb{O}$-operators and solutions of the AYBE (1-3), we give the following reformulation of a part of Corollary 3.6. See Section 3 for general cases.

Let $\mathbb{k}$ be a field whose characteristic is not 2. Let A be a $\mathbb{k}$-algebra that we for now assume to have finite dimension over $\mathbb{k}$. Let

$$\sigma : A \otimes A \to A \otimes A, a \otimes b \mapsto b \otimes a,$$

be the switch operator and let

$$t : \mathrm{Hom}_{\mathbb{k}}(A^*, A) \to \mathrm{Hom}_{\mathbb{k}}(A^*, A)$$

be the transpose operator. Then the natural bijection

$$\phi : A \otimes A \to \mathrm{Hom}_{\mathbb{k}}(A^*, \mathbb{k}) \otimes A \to \mathrm{Hom}_{\mathbb{k}}(A^*, A)$$

is compatible with the operators σ and t. Let $\mathrm{Sym}^2(A \otimes A)$ and $\mathrm{Alt}^2(A \otimes A)$ (respectively $\mathrm{Hom}_{\mathbb{k}}(A^*, A)_+$ and $\mathrm{Hom}_{\mathbb{k}}(A^*, A)_-$) be the eigenspaces for the eigenvalues 1 and -1 of σ on $A \otimes A$ (respectively of t on $\mathrm{Hom}_{\mathbb{k}}(A^*, A)$). Then we have a commutative diagram of bijective linear maps given by

$$\begin{array}{ccc} A \otimes A & \xrightarrow{\quad\phi\quad} & \mathrm{Hom}_{\mathbb{k}}(A^*, A) \\ \downarrow & & \downarrow \\ \mathrm{Alt}^2(A \otimes A) \oplus \mathrm{Sym}^2(A \otimes A) & \xrightarrow{\phi} & \mathrm{Hom}_{\mathbb{k}}(A^*, A)_- \oplus \mathrm{Hom}_{\mathbb{k}}(A^*, A)_+, \end{array} \tag{1-7}$$

which preserves the factorizations. Define $\mathrm{Hom}_{\mathrm{bim}}(A^*, A)_+$ to be the subset of $\mathrm{Hom}_{\mathbb{k}}(A^*, A)_+$ consisting of A-bimodule homomorphisms from A^* to A, both of which are equipped with the natural A-bimodule structures. Let

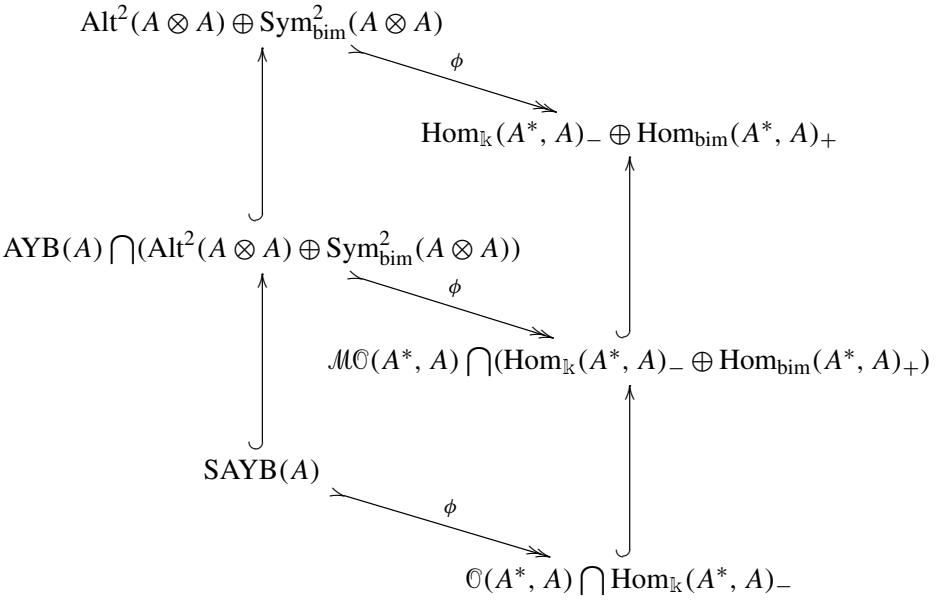
$$\mathrm{Sym}^2_{\mathrm{bim}}(A \otimes A) := \phi^{-1}(\mathrm{Hom}_{\mathrm{bim}}(A^*, A)_+) \subseteq \mathrm{Sym}^2(A \otimes A).$$

Then we have this (see Corollary 3.6):

Theorem 1.4. *An element $r = (r_-, r_+) \in \mathrm{Alt}^2(A \otimes A) \oplus \mathrm{Sym}^2_{\mathrm{bim}}(A \otimes A)$ is a solution of the AYBE* (1-3) *if and only if the pair $\phi(r) = (\phi(r)_-, \phi(r)_+) = (\phi(r_-), \phi(r_+))$ is an extended ©-operator with modification $\phi(r_+)$ of mass $\kappa = -1$. In particular, when r_+ is zero, an element $r = (r_-, 0) = r_- \in \mathrm{Alt}^2(A \otimes A)$ is a solution of the AYBE if and only if the pair $\phi(r) = (\phi(r)_-, 0) = \phi(r_-)$ is an ©-operator of weight zero given by* (1-6) *when (V, ℓ, r) is the dual bimodule (A^*, R^*, L^*) of (A, L, R).*

Let $\mathcal{M}©(A^*, A)$ denote the set of extended ©-operators (α, β) from A^* to A of mass $\kappa = -1$. Let $©(A^*, A)$ denote the set of ©-operators $\alpha : A^* \to A$ of weight 0. Let $\mathrm{AYB}(A)$ denote the set of solutions of the AYBE (1-3) in A. Let $\mathrm{SAYB}(A)$ denote the set of skew-symmetric solutions of the AYBE (1-3) in A. Then Theorem 1.4 means that the bijection in (1-7) restricts to bijections in the

following commutative diagram:



1d. ***Layout of the paper.*** In Section 2, we introduce extended $\mathbb{O}$-operators and study their connection with the associativity of certain products. Section 3 establishes the relationship of extended $\mathbb{O}$-operators with associative and extended associative Yang–Baxter equations. Section 4 introduces the concept of the generalized associative Yang–Baxter equation (GAYBE) and considers its relationship with extended $\mathbb{O}$-operators.

2. $\mathbb{O}$-operators and extended $\mathbb{O}$-operators

We give background notation in Section 2a before introducing the concept of an extended $\mathbb{O}$-operator in Section 2b. We then show in Section 2c and 2d that extended $\mathbb{O}$-operators can be characterized by the associativity of a multiplication derived from this operator.

2a. *Bimodules, A-bimodule $\Bbbk$-algebras and matched pairs of algebras.*

Definition 2.1. Let $(A, \cdot)$ be a $\Bbbk$-algebra.

(i) An *A-bimodule* is a $\Bbbk$-module V and linear maps $\ell, r : A \to \mathrm{End}_{\Bbbk}(V)$ such that (V, ℓ) defines a left A-module, (V, r) defines a right A-module and the two module structures on V are compatible in the sense that

$$(\ell(x)v)r(y) = \ell(x)(vr(y)) \quad \text{for all } x, y \in A, v \in V.$$

If we want more precision, we denote an A-bimodule V by the triple (V, ℓ, r).

(ii) A homomorphism between two A-bimodules (V_1, ℓ_1, r_1) and (V_2, ℓ_2, r_2) is a $\Bbbk$-linear map $g : V_1 \to V_2$ such that

$$g(\ell_1(x)v) = \ell_2(x)g(v) \quad \text{and} \quad g(vr_1(x)) = g(v)r_2(x) \quad \text{for all } x \in A,\ v \in V_1.$$

For a k-algebra A and $x \in A$, define the left and right actions

$$L(x) : A \to A, \quad L(x)y = xy \quad \text{and} \quad R(x) : A \to A,\ yR(x) = yx \quad \text{for all } y \in A.$$

Further define

$$L = L_A : A \to \mathrm{End}_{\Bbbk}(A), \quad x \mapsto L(x) \quad \text{and} \quad R = R_A : A \to \mathrm{End}_{\Bbbk}(A), \quad x \mapsto R(x).$$

Obviously, (A, L, R) is an A-bimodule.

For a $\Bbbk$-module V, let $V^* := \mathrm{Hom}_{\Bbbk}(V, \Bbbk)$ denote the dual $\Bbbk$-module. Denote the usual pairing between V^* and V by

$$\langle \cdot, \cdot \rangle : V^* \times V \to \Bbbk, \quad \langle u^*, v\rangle = u^*(v) \quad \text{for all } u^* \in V^* \text{ and } v \in V.$$

Proposition 2.2 [Bai 2010]. *Let A be a $\Bbbk$-algebra and let (V, ℓ, r) be an A-bimodule. Define the linear maps $\ell^*, r^* : A \to \mathrm{End}_{\Bbbk}(V^*)$ by*

$$\langle u^*\ell^*(x), v\rangle = \langle u^*, \ell(x)v\rangle \quad \textit{and} \quad \langle r^*(x)u^*, v\rangle = \langle u^*, vr(x)\rangle$$

for all $x \in A$, $u^* \in V^*$ and $v \in V$. Then (V^*, r^*, ℓ^*) is an A-bimodule, called the dual bimodule of (V, ℓ, r).

Let (A^*, R^*, L^*) denote the dual A-bimodule of the A-bimodule (A, L, R).

We next extend the concept of a bimodule to that of an A-bimodule algebra by replacing the $\Bbbk$-module V by a $\Bbbk$-algebra R.

Definition 2.3. Let $(A, \cdot)$ be a $\Bbbk$-algebra with multiplication $\cdot$ and let $(R, \circ)$ be a $\Bbbk$-algebra with multiplication $\circ$. Let $\ell, r : A \to \mathrm{End}_{\Bbbk}(R)$ be two linear maps. We call R (or the triple (R, ℓ, r) or the quadruple $(R, \circ, \ell, r)$) an *A-bimodule $\Bbbk$-algebra* if (R, ℓ, r) is an A-bimodule that is compatible with the multiplication $\circ$ on R. More precisely, we have, for all $x, y \in A$ and $v, w \in R$

$$\text{(2-1)} \qquad \ell(x \cdot y)v = \ell(x)(\ell(y)v), \qquad \ell(x)(v \circ w) = (\ell(x)v) \circ w,$$

$$\text{(2-2)} \qquad vr(x \cdot y) = (vr(x))r(y), \qquad (v \circ w)r(x) = v \circ (wr(x)),$$

$$\text{(2-3)} \qquad (\ell(x)v)r(y) = \ell(x)(vr(y)), \qquad (vr(x)) \circ w = v \circ (\ell(x)w).$$

Obviously, for any $\Bbbk$-algebra $(A, \cdot)$, the triple $(A, \cdot, L, R)$ is an A-bimodule $\Bbbk$-algebra. An A-bimodule $\Bbbk$-algebra R need not be a left or right A-algebra since we do not assume that $A \cdot 1$ is in the center of R. For example, the A-bimodule $\Bbbk$-algebra (A, L, R) is an A-algebra if and only if A is a commutative ring.

An A-bimodule $\Bbbk$-algebra is a special case of a matched pair as introduced in [Bai 2010]. It is easy to get the following result, which is a generalization of the

classical result between bimodule structures on V and semidirect product algebraic structures on $A \oplus V$.

Proposition 2.4. *If $(R, \circ, \ell, r)$ is an A-bimodule $\Bbbk$-algebra, then the direct sum $A \oplus R$ of vector spaces is turned into a $\Bbbk$-algebra (the semidirect sum) by defining multiplication in $A \oplus R$ by*

$$(x_1, v_1) * (x_2, v_2) = (x_1 \cdot x_2, \ell(x_1)v_2 + v_1 r(x_2) + v_1 \circ v_2)$$

for all $x_1, x_2 \in A$ and $v_1, v_2 \in R$.

We denote this algebra by $A \ltimes_{\ell,r} R$ or simply $A \ltimes R$.

2b. *Extended $\mathbb{O}$-operators.* We first define an $\mathbb{O}$-operator before introducing an extended $\mathbb{O}$-operator through an auxiliary operator.

Definition 2.5. Let $(A, \cdot)$ be a $\Bbbk$-algebra and $(R, \circ, \ell, r)$ be an A-bimodule $\Bbbk$-algebra. A linear map $\alpha : R \to A$ is called an *$\mathbb{O}$-operator of weight $\lambda \in \Bbbk$ associated to* $(R, \circ, \ell, r)$ if α satisfies

$$\alpha(u) \cdot \alpha(v) = \alpha(\ell(\alpha(u)v)) + \alpha(ur(\alpha(v))) + \lambda\alpha(u \circ v) \quad \text{for all } u, v \in V. \tag{2-4}$$

Remark 2.6. Under our assumption that $\Bbbk$ is a field, the nonzero weight can be normalized to weight 1. In fact, for a nonzero weight $\lambda \in \Bbbk$, if α is an $\mathbb{O}$-operator of weight λ associated to an A-bimodule $\Bbbk$-algebra $(R, \circ, \ell, r)$, then α is an $\mathbb{O}$-operator of weight 1 associated to $(R, \lambda\circ, \ell, r)$ and α/λ is an $\mathbb{O}$-operator of weight 1 associated to $(R, \circ, \ell, r)$.

When the multiplication on the A-bimodule $\Bbbk$-algebra happens to be trivial, an $\mathbb{O}$-operator is just a generalized Rota–Baxter operator defined in [Uchino 2008]. Further, an $\mathbb{O}$-operator associated to (A, L, R) is just a Rota–Baxter operator on A. An $\mathbb{O}$-operator can be viewed as the relative version of a Rota–Baxter operator in that the domain and range of an $\mathbb{O}$-operator might be different. Thus an $\mathbb{O}$-operator is also called a relative Rota–Baxter operator.

We now further generalize the concept of an $\mathbb{O}$-operator.

Definition 2.7. Let $(A, \cdot)$ be a $\Bbbk$-algebra.

(i) Let $\kappa \in \Bbbk$ and let (V, ℓ, r) be an A-bimodule. A linear map (respectively an A-bimodule homomorphism) $\beta : V \to A$ is called a *balanced linear map of mass κ* (respectively *balanced A-bimodule homomorphism of mass κ*) if

$$\kappa\ell(\beta(u))v = \kappa ur(\beta(v)) \quad \text{for all } u, v \in V. \tag{2-5}$$

(ii) Let $\kappa, \mu \in \Bbbk$ and let $(R, \circ, \ell, r)$ be an A-bimodule $\Bbbk$-algebra. A linear map (respectively an A-bimodule homomorphism) $\beta : R \to A$ is called a *balanced*

linear map of mass (κ, μ) (respectively a *balanced A-bimodule homomorphism of mass* (κ, μ)) if (2-5) holds and

$$\mu\ell(\beta(u \circ v))w = \mu u r(\beta(v \circ w)) \quad \text{for all } u, v, w \in R. \tag{2-6}$$

Clearly, if $\kappa = 0$ and $\mu = 0$, then (2-5) and (2-6), respectively, impose no restriction. So any A-bimodule homomorphism is balanced of mass $(\kappa, \mu) = (0, 0)$. For a nonzero mass, we have the following examples.

Example 2.8. Let A be a $\Bbbk$-algebra.

(i) The identity map $\beta = \mathrm{id} : (A, L, R) \to A$ is a balanced A-bimodule homomorphism (of any mass (κ, μ)).

(ii) Any A-bimodule homomorphism $\beta : (A, L, R) \to A$ is balanced (of any mass (κ, μ)).

(iii) Let $r \in A \otimes A$ be symmetric. If r regarded as a linear map from (A^*, R^*, L^*) to A is an A-bimodule homomorphism, then r is a balanced A-bimodule homomorphism (of any mass κ). See Lemma 3.2.

We can now introduce our first main concept in this paper.

Definition 2.9. Let $(A, \cdot)$ be a $\Bbbk$-algebra and let $(R, \circ, \ell, r)$ be an A-bimodule $\Bbbk$-algebra.

(i) Let $\lambda, \kappa, \mu \in \Bbbk$. Fix a balanced A-bimodule homomorphism $\beta : (R, \ell, r) \to A$ of mass (κ, μ). A linear map $\alpha : R \to A$ is called an *extended $\mathbb{O}$-operator of weight λ with modification β of mass* (κ, μ) if, for all $u, v \in R$,

$$\alpha(u) \cdot \alpha(v) - \alpha(\ell(\alpha(u))v + u r(\alpha(v)) + \lambda u \circ v) = \kappa\beta(u) \cdot \beta(v) + \mu\beta(u \circ v). \tag{2-7}$$

(ii) We also let (α, β) denote an extended $\mathbb{O}$-operator α with modification β.

(iii) When (V, ℓ, r) is an A-bimodule, we regard V as an A-bimodule $\Bbbk$-algebra with the zero multiplication. Then λ and μ are irrelevant. We then call the pair (α, β) an *extended $\mathbb{O}$-operator with modification β of mass κ*.

We note that, when the modification β is the zero map (and hence κ and μ are irrelevant), then α is the $\mathbb{O}$-operator defined in Definition 2.5.

2c. ***Extended $\mathbb{O}$-operators and associativity.*** The study of classical Yang–Baxter equations often gives rise to the study of additional Lie structures derived from a given Lie algebra [Bai et al. 2010b; Semenov-Tyan-Shanskiĭ 1983]. Similar derived structures in an associative algebra have also appeared in the study of dendriform algebras and Rota–Baxter algebras [Aguiar 2000b; Bai 2010; Loday and Ronco 2004]. Here we study derived structures arising from $\mathbb{O}$-operators.

Let $(A, \cdot)$ be a $\Bbbk$-algebra and $(R, \circ, \ell, r)$ be an A-bimodule $\Bbbk$-algebra. Let $\delta_\pm : R \to A$ be two linear maps and $\lambda \in \Bbbk$. We now consider the associativity of the multiplication

$$u \diamond v := \ell(\delta_+(u))v + ur(\delta_-(v)) + \lambda u \circ v \quad \text{for all } u, v \in R, \tag{2-8}$$

and several other related multiplications. This will be applied in the Section 4.

Let the characteristic of the field $\Bbbk$ be different from 2. Set

$$\alpha := (\delta_+ + \delta_-)/2 \quad \text{and} \quad \beta := (\delta_+ - \delta_-)/2, \tag{2-9}$$

called the *symmetrizer* and *antisymmetrizer* of $\delta_\pm$ respectively. Note that $\delta_\pm$ can be recovered from α and β by $\delta_\pm = \alpha \pm \beta$.

Lemma 2.10. *Let $(A, \cdot)$ be a $\Bbbk$-algebra and $(R, \circ, \ell, r)$ be an A-bimodule $\Bbbk$-algebra. Let $\alpha : R \to A$ be a linear map and let λ be in $\Bbbk$. Then the operation given by*

$$u *_\alpha v := \ell(\alpha(u))v + ur(\alpha(v)) + \lambda u \circ v \quad \textit{for all } u, v \in R \tag{2-10}$$

is associative if and only if

$$\ell(\alpha(u) \cdot \alpha(v) - \alpha(u *_\alpha v))w = ur(\alpha(v) \cdot \alpha(w) - \alpha(v *_\alpha w)) \tag{2-11}$$

for all $u, v, w \in R$.

Proof. It is straightforward to check that, for any $u, v, w \in R$, we have

$$\begin{aligned}(u *_\alpha v) *_\alpha w - u *_\alpha (v *_\alpha w)& \\ = ur(\alpha(v) \cdot \alpha(w) - \alpha(v *_\alpha w)) - \ell(\alpha(u) \cdot \alpha(v) - \alpha(u *_\alpha v))w.& \quad \square\end{aligned}$$

Corollary 2.11. *Let $\Bbbk$ be a field of characteristic not equal to 2. Let $(A, \cdot)$ be a $\Bbbk$-algebra and $(R, \circ, \ell, r)$ be an A-bimodule $\Bbbk$-algebra. Let $\delta_\pm : R \to A$ be two linear maps and $\lambda \in \Bbbk$. Let α and β be their symmetrizer and antisymmetrizer defined by* (2-9)*. If β is a balanced linear map of mass $\kappa = 1$, that is,*

$$\ell(\beta(u))v = ur(\beta(v)) \quad \textit{for all } u, v \in R, \tag{2-12}$$

then the operation $\diamond$ in (2-8) *defines an associative product on R if and only if α satisfies* (2-11)*.*

Proof. The conclusion follows from Lemma 2.10 since in this case, for any $u, v \in R$,

$$u \diamond v = \ell(\delta_+(u))v + ur(\delta_-(v)) + \lambda u \circ v = \ell(\alpha(u))v + ur(\alpha(v)) + \lambda u \circ v. \quad \square$$

Obviously, if α is an $\mathbb{O}$-operator of weight λ associated to an A-bimodule $\Bbbk$-algebra $(R, \circ, \ell, r)$, then (2-11) holds. Thus the operation on R defined by (2-8) is associative.

Theorem 2.12. *Let $\Bbbk$ have characteristic not equal to 2. Let $(A, \cdot)$ be a $\Bbbk$-algebra and $(R, \circ, \ell, r)$ be an A-bimodule $\Bbbk$-algebra. Let $\delta_\pm : R \to A$ be two linear maps and $\lambda \in \Bbbk$. Let α and β be the symmetrizer and antisymmetrizer of $\delta_\pm$.*

(i) *Suppose that β is a balanced linear map of mass (κ, μ) and α satisfies* (2-7). *Then the product $*_\alpha$ is associative.*

(ii) *Suppose β is a balanced A-bimodule homomorphism of mass $(-1, \pm\lambda)$, that is, β satisfies* (2-5) *with $\kappa = -1$,* (2-6) *with $\mu = \pm\lambda$ and*

$$\beta(\ell(x)u) = x \cdot \beta(u) \quad \text{and} \quad \beta(ur(x)) = \beta(u) \cdot x \quad \text{for all } x \in A, u \in R. \tag{2-13}$$

Then α is an extended $\mathbb{O}$-operator of weight λ with modification β of mass $(\kappa, \mu) = (-1, \pm\lambda)$ if and only if $\delta_\pm$ is an $\mathbb{O}$-operator of weight 1 associated to a new A-bimodule $\Bbbk$-algebra $(R, \circ_\pm, \ell, r)$:

$$\delta_\pm(u) \cdot \delta_\pm(v) = \delta_\pm(\ell(\delta_\pm(u))v + ur(\delta_\pm(v)) + u \circ_\pm v) \quad \text{for all } u, v \in R, \tag{2-14}$$

where the associative products $\circ_\pm = \circ_{\lambda,\beta,\pm}$ on R are defined by

$$u \circ_\pm v = \lambda u \circ v \mp 2\ell(\beta(u))v \quad \text{for all } u, v \in R. \tag{2-15}$$

In item (i) we do not assume that β is an A-bimodule homomorphism. Thus α need not be an extended $\mathbb{O}$-operator.

Proof. (i) The conclusion follows from Lemma 2.10.

(ii) It is straightforward to show that $(R, \circ_\pm, \ell, r)$ equipped with the product $\circ_\pm$ is an A-bimodule $\Bbbk$-algebra. Moreover, for any $u, v \in R$,

$$\begin{aligned}
&(\alpha \pm \beta)(u) \cdot (\alpha \pm \beta)(v) - (\alpha \pm \beta)(\ell((\alpha \pm \beta)(u))v + ur((\alpha \pm \beta)(v)) + u \circ_\pm v) \\
&\quad = \alpha(u) \cdot \alpha(v) + \beta(u) \cdot \beta(v) - \alpha(\ell(\alpha(u))v + ur(\alpha(v)) + \lambda u \circ v) \mp \lambda\beta(u \circ v) \\
&\qquad \pm (\beta(u) \cdot \alpha(v) - \beta(ur(\alpha(v))) + \alpha(u) \cdot \beta(v) - \beta(\ell(\alpha(u))v)) \quad \text{by (2-12)} \\
&\quad = \alpha(u) \cdot \alpha(v) + \beta(u) \cdot \beta(v) - \alpha(\ell(\alpha(u))v + ur(\alpha(v)) + \lambda u \circ v) \mp \lambda\beta(u \circ v) \\
&\qquad\qquad \text{by (2-13).}
\end{aligned}$$

Therefore the conclusion holds. □

We close this section with an obvious corollary of Theorem 2.12 by taking $R = V$ with the zero multiplication.

Corollary 2.13. *Let A be a $\Bbbk$-algebra and (V, ℓ, r) be an A-bimodule. Let $\alpha, \beta : V \to A$ be two linear maps such that β is a balanced A-bimodule homomorphism. Then α is an extended $\mathbb{O}$-operator with modification β of mass $\kappa = -1$ if and only if $\alpha \pm \beta$ is an $\mathbb{O}$-operator of weight 1 associated to an A-bimodule $\Bbbk$-algebra $(V, \star_\pm, \ell, r)$, that is, for all $u, v \in V$,*

$$(\alpha \pm \beta)(u) \cdot (\alpha \pm \beta)(v) = (\alpha \pm \beta)(\ell((\alpha \pm \beta)(u))v + ur((\alpha \pm \beta)(v)) + u \star_\pm v),$$

where the associative algebra products $\star_\pm$ *on* V *are defined by*

$$u \star_\pm v = \mp 2\ell(\beta(u))v \quad \textit{for all } u, v \in V.$$

2d. ***The case of*** $\mathbb{O}$***-operators and Rota–Baxter operators.*** Suppose $(A, \cdot)$ is a $\Bbbk$-algebra. Then $(A, \cdot, L, R)$ is an A-bimodule $\Bbbk$-algebra. Theorem 2.12 can be easily restated in this case. But we are mostly interested in the case of $\mu = 0$ when (2-7) takes the form

$$\alpha(x)\cdot\alpha(y) - \alpha(\alpha(x)\cdot y + x\cdot\alpha(y) + \lambda x\cdot y) = \kappa\beta(x)\cdot\beta(y) \quad \text{for all } x, y \in A. \tag{2-16}$$

We list the following special cases for later reference. If $\lambda = 0$, then (2-16) gives

$$\alpha(x)\cdot\alpha(y) - \alpha(\alpha(x)\cdot y + x\cdot\alpha(y)) = \kappa\beta(x)\cdot\beta(y) \quad \text{for all } x, y \in A. \tag{2-17}$$

If in addition, $\beta = \mathrm{id}$, then (2-17) gives

$$\alpha(x)\cdot\alpha(y) - \alpha(\alpha(x)\cdot y + x\cdot\alpha(y)) = \kappa x\cdot y \quad \text{for all } x, y \in A. \tag{2-18}$$

If furthermore $\kappa = -1$, then (2-18) becomes

$$\alpha(x)\cdot\alpha(y) - \alpha(\alpha(x)\cdot y + x\cdot\alpha(y)) = -x\cdot y \quad \text{for all } x, y \in A. \tag{2-19}$$

By Lemma 2.10 and Theorem 2.12, we reach the following conclusion.

Corollary 2.14. *Let* $(A, \cdot)$ *be a* $\Bbbk$*-algebra. Let* $\alpha, \beta : A \to A$ *be two linear maps and* $\lambda \in \Bbbk$.

(i) *For any* $\kappa \in \Bbbk$, *let* β *be balanced of mass* $(\kappa, 0)$ *and let* α *be an extended* $\mathbb{O}$*-operator of weight* λ *with modification* β *of mass* $(\kappa, \mu) = (\kappa, 0)$, *that is,* α *satisfies* (2-16). *Then the product* $*_\alpha$ *on* A *is associative.*

(ii) *If* β *is an* A*-bimodule homomorphism, then* α *and* β *satisfy* (2-17) *for* $\kappa = -1$ *if and only if* $r_\pm = \alpha \pm \beta$ *is an* $\mathbb{O}$*-operator of weight* 1 *associated to a new* A*-bimodule* $\Bbbk$*-algebra* $(A, \star_\pm, L, R)$:

$$r_\pm(x)\cdot r_\pm(y) = r_\pm(r_\pm(x)\cdot y + x\cdot r_\pm(y) + x \star_\pm y) \quad \textit{for all } x, y \in A,$$

where the associative products $\star_\pm$ *on* A *are defined by*

$$x \star_\pm y = \mp 2\beta(x)\cdot y \quad \textit{for all } x, y \in A.$$

Let $(A, \cdot)$ be a $\Bbbk$-algebra and let $(A, \cdot, L, R)$ be the corresponding A-bimodule $\Bbbk$-algebra. In this case, $\beta = \mathrm{id}$ clearly satisfies the conditions of Theorem 2.12 and (2-7) takes the form

$$\alpha(x)\cdot\alpha(y) - \alpha(\alpha(x)\cdot y + x\cdot\alpha(y) + \lambda x\cdot y) = \hat{\kappa} x\cdot y \quad \text{for all } x, y \in A, \tag{2-20}$$

where $\hat{\kappa} = \kappa + \mu$. Thus we have the following consequence of Theorem 2.12.

Corollary 2.15. *Let $\hat{\kappa} = -1 \pm \lambda$. Then $\alpha : A \to A$ satisfies* (2-20) *if and only if $\alpha \pm 1$ is a Rota–Baxter operator of weight $\lambda \mp 2$.*

When $\lambda = 0$, this fact can be found in [Ebrahimi-Fard 2002]. As noted there, the Lie algebraic version of (2-20) in this case, namely (2-19), is the operator form of the modified classical Yang–Baxter equation [Semenov-Tyan-Shanskiĭ 1983].

3. Extended 𝕆-operators and EAYBE

Here we study the relationship between extended 𝕆-operators and associative Yang–Baxter equations. We start with introducing various concepts of the associative Yang–Baxter equation (AYBE) in Section 3a. We then establish connections between 𝕆-operators in different generalities and solutions of these variations of AYBE in different algebras. The relationship between 𝕆-operators on a $\Bbbk$-algebra A and solutions of AYBE in A is considered in Section 3b. We then consider in Section 3c the relationship between an extended 𝕆-operator and solutions of AYBE and extended AYBE in an extension algebra of A. We finally consider the special case of Frobenius algebras in Section 3d.

3a. ***Extended associative Yang–Baxter equations.*** We define variations of the associative Yang–Baxter equation to be satisfied by two tensors from an algebra. We then study the linear maps from these two tensors in preparation for the relationship between 𝕆-operators and solutions of these associative Yang–Baxter equations.

Let A be a $\Bbbk$-algebra. Let $r = \sum_i a_i \otimes b_i \in A \otimes A$. We continue to use the notations r_{12}, r_{13} and r_{23} defined in (1-4). We similarly define

$$r_{21} = \sum_i b_i \otimes a_i \otimes 1, \quad r_{31} = \sum_i b_i \otimes 1 \otimes a_i, \quad r_{32} = \sum_i 1 \otimes b_i \otimes a_i.$$

Equip $A \otimes A \otimes A$ with the product of the tensor algebra. In particular,

$$(a_1 \otimes a_2 \otimes a_3)(b_1 \otimes b_2 \otimes b_3) = (a_1 b_1) \otimes (a_2 b_2) \otimes (a_3 b_3) \quad \text{for all } a_i, b_i \in A,\ i = 1, 2, 3.$$

Definition 3.1. Fix $\varepsilon \in \Bbbk$.

(i) The equation

$$r_{12}r_{13} + r_{13}r_{23} - r_{23}r_{12} = \varepsilon(r_{13} + r_{31})(r_{23} + r_{32}) \tag{3-1}$$

is called the *extended associative Yang–Baxter equation of mass ε* (or *ε-EAYBE* in short).

(ii) Let A be a $\Bbbk$-algebra. An element $r \in A \otimes A$ is called a *solution of the ε-EAYBE in A* if it satisfies (3-1).

When $\varepsilon = 0$ or r is skew-symmetric in the sense that $\sigma(r) = -r$ for the switch operator $\sigma : A \otimes A \to A \otimes A$ (and hence $r_{13} = -r_{31}$), then the ε-EAYBE is the same as the AYBE in (1-3):

$$r_{12}r_{13} + r_{13}r_{23} - r_{23}r_{12} = 0. \tag{3-2}$$

Let A be a $\Bbbk$-algebra with finite $\Bbbk$-dimension. For $r \in A \otimes A$, define a linear map $F_r : A^* \to A$ by

$$\langle v, F_r(u)\rangle = \langle u \otimes v, r\rangle \quad \text{for all } u, v \in A^*. \tag{3-3}$$

This defines a bijective linear map $F : A \otimes A \to \mathrm{Hom}_{\Bbbk}(A^*, A)$ and thus allows us to identify r with F_r, which we still denote by r for simplicity of notation. Similarly define a linear map $r^t : A^* \to A$ by

$$\langle u, r^t(v)\rangle = \langle r, u \otimes v\rangle. \tag{3-4}$$

Obviously r is symmetric or skew-symmetric in $A \otimes A$ if and only if, as a linear map, $r = r^t$ or $r = -r^t$, respectively. Suppose that the characteristic of $\Bbbk$ is not 2 and define

$$\alpha = \alpha_r = (r - r^t)/2 \quad \text{and} \quad \beta = \beta_r = (r + r^t)/2, \tag{3-5}$$

called the *skew-symmetric part* and the *symmetric part* of r, respectively. Then $r = \alpha + \beta$ and $r^t = -\alpha + \beta$.

Lemma 3.2. *Let $(A, \cdot)$ be a $\Bbbk$-algebra with finite $\Bbbk$-dimension. Let $s \in A \otimes A$ be symmetric. Then the following conditions are equivalent.*

(i) *s is* invariant, *that is*,

$$(\mathrm{id} \otimes L(x) - R(x) \otimes \mathrm{id})s = 0 \quad \textit{for all } x \in A. \tag{3-6}$$

(ii) *s regarded as a linear map from (A^*, R^*, L^*) to A is balanced, that is*,

$$R^*(s(a^*))b^* = a^*L^*(s(b^*)) \quad \textit{for all } a^*, b^* \in A^*. \tag{3-7}$$

(iii) *s regarded as a linear map from (A^*, R^*, L^*) to A is an A-bimodule homomorphism, that is*,

$$s(R^*(x)a^*) = x \cdot s(a^*), \quad s(a^*L^*(x)) = s(a^*) \cdot x \quad \textit{for all } x \in A, a^* \in A^*. \tag{3-8}$$

Proof. (i) $\Longleftrightarrow$ (ii). Since $s \in A \otimes A$ is symmetric, for any $x \in A$, $a^*, b^* \in A^*$,

$$\begin{aligned}\langle(\mathrm{id} \otimes L(x) - R(x) \otimes \mathrm{id})s, a^* \otimes b^*\rangle &= \langle s, a^* \otimes L^*(x)b^*\rangle - \langle s, R^*(x)a^* \otimes b^*\rangle \\ &= \langle x \cdot s(a^*), b^*\rangle - \langle a^*, s(b^*) \cdot x\rangle \\ &= \langle R^*(s(a^*))b^* - a^*L^*(s(b^*)), x\rangle.\end{aligned}$$

So s is invariant if and only if s regarded as a linear map from (A^*, R^*, L^*) to A is balanced.

(i) $\Longleftrightarrow$ (iii). For any $x \in A$, $a^*, b^* \in A^*$,

$$\begin{aligned}\langle(\mathrm{id}\otimes L(x) - R(x)\otimes \mathrm{id})s, a^*\otimes b^*\rangle &= \langle s, a^*\otimes L^*(x)b^*\rangle - \langle s, R^*(x)a^*\otimes b^*\rangle\\ &= \langle x\cdot s(a^*) - s(R^*(x)a^*), b^*\rangle,\\ \langle(\mathrm{id}\otimes L(x) - R(x)\otimes \mathrm{id})s, a^*\otimes b^*\rangle &= \langle s, a^*\otimes L^*(x)b^*\rangle - \langle s, R^*(x)a^*\otimes b^*\rangle\\ &= \langle s(L^*(x)b^*) - s(b^*)\cdot x, a^*\rangle\end{aligned}$$

by the symmetry of $s \in A\otimes A$. So s is invariant if and only if s regarded as a linear map from (A^*, R^*, L^*) to A is an A-bimodule homomorphism. □

Remark 3.3. The invariant condition in item (i) also arises in the construction of a coboundary antisymmetric infinitesimal bialgebra in the sense of [Bai 2010]; see also [Bai et al. 2012].

3b. ***Extended ©-operators from EAYBE.*** We first state the following special case of Corollary 2.13.

Corollary 3.4. *Let $\Bbbk$ be a field of characteristic not equal to 2. Let A be a $\Bbbk$-algebra with finite $\Bbbk$-dimension and $r \in A\otimes A$. Let α and β be defined by* (3-5)*. Suppose β is a balanced A-bimodule homomorphism. These two statements are equivalent*:

(i) *The map α is an extended ©-operator with modification β of mass -1:*

$$\alpha(a^*)\cdot\alpha(b^*) - \alpha(R^*(\alpha(a^*))b^* + a^*L^*(\alpha(b^*))) = -\beta(a^*)\cdot\beta(b^*) \tag{3-9}$$

for all $a^*, b^* \in A^*$.

(ii) *The map r (respectively $-r^t$) is an ©-operator of weight 1 associated to a new A-bimodule $\Bbbk$-algebra $(A^*, \circ_+, R^*, L^*)$ (respectively $(A^*, \circ_-, R^*, L^*)$):*

$$r(a^*)\cdot r(b^*) = r(R^*(r(a^*))b^* + a^*L^*(r(b^*)) + a^*\circ_+ b^*) \tag{3-10}$$

for all $a^*, b^* \in A^*$, (respectively

$$\begin{aligned}(-r^t)(a^*)\cdot(-r^t)(b^*)&\\ = (-r^t)(R^*((-r^t)(a^*))b^* + a^*L^*((-r^t)(b^*)) + a^*\circ_- b^*),&\end{aligned} \tag{3-11}$$

for all $a^*, b^* \in A^*$), where the associative algebra products $\circ_\pm$ on A^* are defined by

$$a^*\circ_\pm b^* = \mp 2R^*(\beta(a^*))b^* \quad \text{for all } a^*, b^* \in A^*. \tag{3-12}$$

In the theory of integrable systems [Kosmann-Schwarzbach 1997; Semenov-Tyan-Shanskiĭ 1983], *modified classical Yang–Baxter equation* usually refers to (the Lie algebraic version of) (2-19) and (3-9).

The following theorem establishes a close relationship between extended $\mathbb{O}$-operators on a $\Bbbk$-algebra A and solutions of the AYBE in A.

Theorem 3.5. *Let $\Bbbk$ be a field of characteristic not equal to 2. Let A be a $\Bbbk$-algebra with finite $\Bbbk$-dimension and let $r \in A \otimes A$, which is identified as a linear map from A^* to A.*

(i) *Then r is a solution of the AYBE in A if and only if r satisfies*

$$r(a^*)\cdot r(b^*) = r(R^*(r(a^*))b^* - a^*L^*(r^t(b^*))) \quad \text{for all } a^*, b^* \in A^*. \tag{3-13}$$

(ii) *Define α and β by* (3-5)*. Suppose that the symmetric part β of r is invariant. Then r is a solution of EAYBE of mass $(\kappa+1)/4$:*

$$r_{12}r_{13} + r_{13}r_{23} - r_{23}r_{12} = \tfrac{1}{4}(\kappa+1)(r_{13}+r_{31})(r_{23}+r_{32})$$

if and only if α is an extended $\mathbb{O}$-operator with modification β of mass κ:

$$\alpha(a^*)\cdot\alpha(b^*) - \alpha(R^*(\alpha(a^*))b^* + a^*L^*(\alpha(b^*))) = \kappa\beta(a^*)\cdot\beta(b^*)$$

for all $a^*, b^* \in A^*$.

Proof. (i) Write $r = \sum_{i,j} u_i \otimes v_j$. For any $a^*, b^*, c^* \in A^*$, we have

$$\begin{aligned}
\langle r_{12}\cdot r_{13}, a^*\otimes b^*\otimes c^*\rangle &= \sum_{i,j}\langle u_i\cdot u_j, a^*\rangle\langle v_i, b^*\rangle\langle v_j, c^*\rangle \\
&= \sum_{j}\langle r^t(b^*)\cdot u_j, a^*\rangle\langle v_j, b^*\rangle = \langle r(a^*L^*(r^t(b^*))), c^*\rangle, \\
\langle r_{13}\cdot r_{23}, a^*\otimes b^*\otimes c^*\rangle &= \sum_{i,j}\langle u_i, a^*\rangle\langle u_j, b^*\rangle\langle v_i\cdot v_j, c^*\rangle \\
&= \sum_{j}\langle u_j, b^*\rangle\langle r(a^*)\cdot v_j, c^*\rangle = \langle r(a^*)\cdot r(b^*), c^*\rangle, \\
\langle -r_{23}\cdot r_{12}, a^*\otimes b^*\otimes c^*\rangle &= -\sum_{i,j}\langle u_i, a^*\rangle\langle u_j\cdot v_i, b^*\rangle\langle v_j, c^*\rangle \\
&= -\sum_{j}\langle u_j\cdot r(a^*), b^*\rangle\langle v_j, c^*\rangle = \langle -r(R^*(r(a^*))b^*), c^*\rangle.
\end{aligned}$$

Therefore r is a solution of the AYBE in A if and only if r satisfies (3-13).

(ii) By the proof of item (i), we see that, for any $a^*, b^*, c^* \in A^*$,

$$\begin{aligned}
&\langle \alpha(a^*)\cdot\alpha(b^*) - \alpha(R^*(\alpha(a^*))b^* + a^*L^*(\alpha(b^*))) - \kappa\beta(a^*)\cdot\beta(b^*), c^*\rangle \\
&\quad = \langle \alpha(a^*)\cdot\alpha(b^*) - \alpha(R^*(\alpha(a^*))b^* + a^*L^*(\alpha(b^*))) \\
&\qquad\qquad + \beta(a^*)\cdot\beta(b^*) - (\kappa+1)\beta(a^*)\cdot\beta(b^*), c^*\rangle \\
&\quad = \langle r_{12}\cdot r_{13} + r_{13}\cdot r_{23} - r_{23}\cdot r_{12}, a^*\otimes b^*\otimes c^*\rangle - (\kappa+1)\langle \beta_{13}\cdot\beta_{23}, a^*\otimes b^*\otimes c^*\rangle \\
&\quad = \langle r_{12}\cdot r_{13} + r_{13}\cdot r_{23} - r_{23}\cdot r_{12} - (\kappa+1)\tfrac{1}{2}(r_{13}+r_{31})\cdot\tfrac{1}{2}(r_{23}+r_{32}), a^*\otimes b^*\otimes c^*\rangle.
\end{aligned}$$

So r is a solution of the EAYBE of mass $(\kappa+1)/4$ if and only if α is an extended $\mathbb{O}$-operator with modification β of mass κ. □

In the case when $\kappa = -1$, we have this:

Corollary 3.6. *Let $\Bbbk$ be a field of characteristic not equal to 2. Let A be a $\Bbbk$-algebra with finite $\Bbbk$-dimension and let $r \in A\otimes A$. Define α and β by* (3-5).

(i) *If β is invariant, then the following conditions are equivalent.*
   (a) *r is a solution of the AYBE in A.*
   (b) *r satisfies* (3-10)*, that is, r is an $\mathbb{O}$-operator of weight* 1 *associated to the A-bimodule $\Bbbk$-algebra $(A^*, \circ_+, R^*, L^*)$, where A^* is equipped with the associative algebra structure $\circ_+$ defined by* (3-12)*. (With $-r^t$ instead of r, replace* (3-10) *by* (3-11) *and $\circ_+$ with $\circ_-$.)*
   (c) *α is an extended $\mathbb{O}$-operator with modification β of mass -1.*
   (d) *For any $a^*, b^* \in A^*$,*

$$(\alpha \pm \beta)(a^* * b^*) = (\alpha\pm\beta)(a^*)\cdot(\alpha\pm\beta)(b^*), \tag{3-14}$$

where

$$a^* * b^* = R^*(r(a^*))b^* - a^*L^*(r^t(b^*)) \quad \text{for all } a^*, b^* \in A^*.$$

(ii) *If r is skew-symmetric, then r is a solution of the AYBE in A if and only if $r : A^* \to A$ is an $\mathbb{O}$-operator of weight zero.*

Proof. If the symmetric part β of r is invariant, then by Lemma 3.2, for any $a^*, b^* \in A^*$, we have

$$\begin{aligned}
&r(a^*)\cdot r(b^*) - r(R^*(r(a^*))b^* - a^*L^*(r^t(b^*))) \\
&\quad = r(a^*)\cdot r(b^*) - r(R^*(r(a^*))b^* + a^*L^*(r(b^*)) - 2a^*L^*(\beta(b^*))) \\
&\quad = r(a^*)\cdot r(b^*) - r(R^*(r(a^*))b^* + a^*L^*(r(b^*)) + a^*\circ_+ b^*),
\end{aligned}$$

where the product $\circ_+$ is defined by (3-12). Therefore by Corollary 3.4, r is a solution of the AYBE if and only if item (b) or (c) holds. Moreover, since for any

$a^*, b^* \in A^*$, we have

$$R^*(r(a^*))b^* - a^*L^*(r^t(b^*)) = R^*(r(a^*))b^* + a^*L^*(r(b^*)) + a^* \circ_+ b^*$$
$$= R^*((-r^t)(a^*))b^* + a^*L^*((-r^t)(b^*)) + a^* \circ_- b^*,$$

(3-14) is just a reformulation of (3-10) and (3-11). So r is a solution of the AYBE if and only if item (c) holds.

(ii) This is the special case of item (i) when $\beta = 0$. □

3c. ***EAYBEs from extended $\mathbb{O}$-operators.*** We now establish the relationship between an extended $\mathbb{O}$-operator $\alpha : V \to A$ in general and the AYBE and EAYBE. For this purpose we prove that an extended $\mathbb{O}$-operator $\alpha : V \to A$ naturally gives rise to an extended $\mathbb{O}$-operator on a larger associative algebra $\mathscr{A}$ associated to the dual bimodule $(\mathscr{A}^*, R^*_{\mathscr{A}}, L^*_{\mathscr{A}})$. We first introduce some notation.

Definition 3.7. Let A be a $\Bbbk$-algebra and let (V, ℓ, r) be an A-bimodule, both with finite $\Bbbk$-dimension. Let (V^*, r^*, ℓ^*) be the dual A-bimodule and let $\mathscr{A} = A \ltimes_{r^*,\ell^*} V^*$. Identify a linear map $\gamma : V \to A$ as an element in $\mathscr{A} \otimes \mathscr{A}$ through the injective map

$$\mathrm{Hom}_{\Bbbk}(V, A) \cong A \otimes V^* \hookrightarrow \mathscr{A} \otimes \mathscr{A}. \tag{3-15}$$

Denote

$$\tilde{\gamma}_{\pm} := \gamma \pm \gamma^{21}, \tag{3-16}$$

where $\gamma^{21} = \sigma(\gamma) \in V^* \otimes A \subset \mathscr{A} \otimes \mathscr{A}$ with $\sigma : A \otimes V^* \to V^* \otimes A$, $a \otimes u^* \mapsto u^* \otimes a$ being the switch operator.

Lemma 3.8. *Let A be a $\Bbbk$-algebra and let (V, ℓ, r) be an A-bimodule, both with finite $\Bbbk$-dimension. Suppose that $\beta : V \to A$ is a linear map that is identified as an element in $\mathscr{A} \otimes \mathscr{A}$ by* (3-15)*. Define $\tilde{\beta}_+$ by* (3-16)*. Then $\tilde{\beta}_+$, identified as a linear map from $\mathscr{A}^*$ to $\mathscr{A}$, is a balanced $\mathscr{A}$-bimodule homomorphism from $(\mathscr{A}^*, R^*_{\mathscr{A}}, L^*_{\mathscr{A}})$ to $(\mathscr{A}, L_{\mathscr{A}}, R_{\mathscr{A}})$ if and only if $\beta : V \to A$ is a balanced A-bimodule homomorphism from (V, ℓ, r) to (A, L_A, R_A).*

Proof. For the linear map $\tilde{\beta}_+ : \mathscr{A}^* \to \mathscr{A}$, we have $\tilde{\beta}_+(a^*) = \beta^*(a^*)$ for $a^* \in A^*$ and $\tilde{\beta}_+(u) = \beta(u)$ for $u \in V$, where $\beta^* : A^* \to V^*$ is the dual linear map associated to β given by

$$\langle \beta^*(a^*), v\rangle = \langle a^*, \beta(v)\rangle \quad \text{for all } a^* \in A^*, v \in V.$$

First suppose that $\beta : (V, \ell, r) \to A$ is a balanced A-bimodule homomorphism. Let $b^* \in A^*$ and $v \in V$. Then

$$R^*_{\mathscr{A}}(\tilde{\beta}_+(a^* + u))(b^* + v)$$
$$= R^*_{\mathscr{A}}(\beta^*(a^*))b^* + R^*_{\mathscr{A}}(\beta^*(a^*))v + R^*_{\mathscr{A}}(\beta(u))b^* + R^*_{\mathscr{A}}(\beta(u))v,$$

and

$$(a^*+u)L^*_{\mathscr{A}}(\tilde{\beta}_+(b^*+v)) = a^*L^*_{\mathscr{A}}(\beta^*(b^*)) + a^*L^*_{\mathscr{A}}(\beta(v)) + uL^*_{\mathscr{A}}(\beta^*(b^*)) + uL^*_{\mathscr{A}}(\beta(v)).$$

On the other hand, for any $x \in A$, $w^* \in V^*$,

$$\begin{aligned}
\langle R^*_{\mathscr{A}}(\beta^*(a^*))b^* - a^*L^*_{\mathscr{A}}(\beta^*(b^*)), x\rangle &= \langle b^*, x\cdot\beta^*(a^*)\rangle - \langle a^*, \beta^*(b^*)\cdot x\rangle = 0,\\
\langle R^*_{\mathscr{A}}(\beta^*(a^*))b^* - a^*L^*_{\mathscr{A}}(\beta^*(b^*)), w^*\rangle &= \langle b^*, w^*\cdot\beta^*(a^*)\rangle - \langle a^*, \beta^*(b^*)\cdot w^*\rangle\\
&= 0,\\
\langle R^*_{\mathscr{A}}(\beta^*(a^*))v - a^*L^*_{\mathscr{A}}(\beta(v)), x\rangle &= \langle v, x\cdot\beta^*(a^*)\rangle - \langle a^*, \beta(v)\cdot x\rangle\\
&= \langle a^*, \beta(vr(x)) - \beta(v)\cdot x\rangle = 0,\\
\langle R^*_{\mathscr{A}}(\beta^*(a^*))v - a^*L^*_{\mathscr{A}}(\beta(v)), w^*\rangle &= \langle v, w^*\cdot\beta^*(a^*)\rangle - \langle a^*, \beta(v)\cdot w^*\rangle = 0,\\
\langle R^*_{\mathscr{A}}(\beta(u))b^* - uL^*_{\mathscr{A}}(\beta^*(b^*)), x\rangle &= \langle b^*, x\cdot\beta(u)\rangle - \langle u, \beta^*(b^*)\cdot x\rangle\\
&= \langle b^*, x\cdot\beta(u) - \beta(l(x)u)\rangle = 0,\\
\langle R^*_{\mathscr{A}}(\beta(u))b^* - uL^*_{\mathscr{A}}(\beta^*(b^*)), w^*\rangle &= \langle b^*, w^*\cdot\beta(u)\rangle - \langle u, \beta^*(b^*)\cdot w^*\rangle = 0,\\
\langle R^*_{\mathscr{A}}(\beta(u))v - uL^*_{\mathscr{A}}(\beta(v)), w^*\rangle &= \langle v, w^*\cdot\beta(u)\rangle - \langle u, \beta(v)\cdot w^*\rangle\\
&= \langle \ell(\beta(u))v - ur(\beta(v)), w^*\rangle = 0,\\
\langle R^*_{\mathscr{A}}(\beta(u))v - uL^*_{\mathscr{A}}(\beta(v)), x\rangle &= \langle v, x\cdot\beta(u)\rangle - \langle u, \beta(v)\cdot x\rangle = 0.
\end{aligned}$$

Therefore, $R^*_{\mathscr{A}}(\tilde{\beta}_+(a^*+u))(b^*+v) = (a^*+u)L^*_{\mathscr{A}}(\tilde{\beta}_+(b^*+v))$. Since $\tilde{\beta}_+ \in \mathscr{A}\otimes\mathscr{A}$ is symmetric, by Lemma 3.2, $\tilde{\beta}_+$ when identified as a linear map from $\mathscr{A}^*$ to $\mathscr{A}$ is a balanced $\mathscr{A}$-bimodule homomorphism from $(\mathscr{A}^*, R^*_{\mathscr{A}}, L^*_{\mathscr{A}})$ to $(\mathscr{A}, L_{\mathscr{A}}, R_{\mathscr{A}})$.

Conversely, if $\tilde{\beta}_+$ identified as a linear map from $\mathscr{A}^*$ to $\mathscr{A}$ is a balanced $\mathscr{A}$-bimodule homomorphism from $(\mathscr{A}^*, R^*_{\mathscr{A}}, L^*_{\mathscr{A}})$ to $(\mathscr{A}, L_{\mathscr{A}}, R_{\mathscr{A}})$, then

$$\begin{aligned}
R^*_{\mathscr{A}}(\tilde{\beta}_+(u))v = uL^*_{\mathscr{A}}(\tilde{\beta}_+(v)) \quad &\Longleftrightarrow \quad \ell(\beta(u))v = ur(\beta(v)),\\
\tilde{\beta}_+(R^*_{\mathscr{A}}(x)v) = x\cdot\tilde{\beta}_+(v) \quad &\Longleftrightarrow \quad \beta(\ell(x)v) = x\cdot\beta(v),\\
\tilde{\beta}_+(uL^*_{\mathscr{A}}(x)) = \tilde{\beta}_+(u)\cdot x \quad &\Longleftrightarrow \quad \beta(ur(x)) = \beta(u)\cdot x
\end{aligned}$$

for any $u, v \in V$, $x \in A$. So $\beta : (V, \ell, r) \to (A, L_A, R_A)$ is a balanced A-bimodule homomorphism. □

Theorem 3.9. *Let A be a $\Bbbk$-algebra and let (V, ℓ, r) be an A-bimodule, both with finite $\Bbbk$-dimension. Let $\alpha, \beta : V \to A$ be two $\Bbbk$-linear maps. Let $\tilde{\alpha}_-$ and $\tilde{\beta}_+$ be defined by* (3-15) *and identified as linear maps from $\mathscr{A}^*$ to $\mathscr{A}$. Then α is an extended ©-operator with modification β of mass κ if and only if $\tilde{\alpha}_-$ is an extended ©-operator with modification $\tilde{\beta}_+$ of mass κ.*

Proof. For any $a^* \in A^*$ and $v \in V$, we have $\tilde{\alpha}_-(a^*) = \alpha^*(a^*)$ and $\tilde{\alpha}_-(v) = -\alpha(v)$, where $\alpha^* : A^* \to V^*$ is the dual linear map of α. Suppose that α is an extended $\mathbb{O}$-operator with modification β of mass κ. Then for any $a^*, b^* \in A^*$ and $u, v \in V$, we have

$$\begin{aligned}
&\tilde{\alpha}_-(u+a^*)\cdot\tilde{\alpha}_-(v+b^*) - \tilde{\alpha}_-(R^*_{\mathscr{A}}(\tilde{\alpha}_-(u+a^*))(v+b^*) + (u+a^*)L^*_{\mathscr{A}}(\tilde{\alpha}_-(v+b^*)))\\
&= \alpha(u)\cdot\alpha(v) - \alpha(\ell(\alpha(u))v) - \alpha(ur(\alpha(v))) - r^*(\alpha(u))\alpha^*(b^*) + \alpha^*(R^*_{\mathscr{A}}(\alpha(u))b^*)\\
&\quad - \alpha^*(uL^*_{\mathscr{A}}(\alpha^*(b^*))) - \alpha^*(a^*)\ell^*(\alpha(v)) - \alpha^*(R^*_{\mathscr{A}}(\alpha^*(a^*))v) + \alpha^*(a^*L^*_{\mathscr{A}}(\alpha(v))).
\end{aligned}$$

On the other hand, for any $w \in V$ we have

$$\begin{aligned}
&\langle -r^*(\alpha(u))\alpha^*(b^*) + \alpha^*(R^*_{\mathscr{A}}(\alpha(u))b^*) - \alpha^*(uL^*_{\mathscr{A}}(\alpha^*(b^*))), w\rangle\\
&\qquad = \langle b^*, \alpha(w)\cdot\alpha(u) - \alpha(\ell(\alpha(w))u + wr(\alpha(u))\rangle = \langle b^*, \kappa\beta(w)\cdot\beta(u)\rangle\\
&\qquad = \langle b^*, \kappa\beta(wr(\beta(u)))\rangle = \langle \kappa r^*(\beta(u))\beta^*(b^*), w\rangle.
\end{aligned}$$

Therefore

$$-r^*(\alpha(u))\alpha^*(b^*) + \alpha^*(R^*_{\mathscr{A}}(\alpha(u))b^*) - \alpha^*(uL^*_{\mathscr{A}}(\alpha^*(b^*))) = \kappa r^*(\beta(u))\beta^*(b^*).$$

Similarly, we have

$$-\alpha^*(a^*)\ell^*(\alpha(v)) - \alpha^*(R^*_{\mathscr{A}}(\alpha^*(a^*))v) + \alpha^*(a^*L^*_{\mathscr{A}}(\alpha(v))) = \kappa\beta^*(a^*)\ell^*(\beta(v)).$$

So

$$\begin{aligned}
&\tilde{\alpha}_-(u+a^*)\cdot\tilde{\alpha}_-(v+b^*) - \tilde{\alpha}_-(R^*_{\mathscr{A}}(\tilde{\alpha}_-(u+a^*))(v+b^*)\\
&\qquad\qquad + (u+a^*)L^*_{\mathscr{A}}(\tilde{\alpha}_-(v+b^*)))\\
&= \kappa\beta(u)\cdot\beta(v) + \kappa r^*(\beta(u))\beta^*(b^*) + \kappa\beta^*(a^*)\ell^*(\beta(v))\\
&= \kappa\beta(u)\cdot\beta(v) + \kappa\beta(u)\cdot\beta^*(b^*) + \kappa\beta^*(a^*)\cdot\beta(v) = \kappa\tilde{\beta}_+(u+a^*)\tilde{\beta}_+(v+b^*).
\end{aligned}$$

If $\kappa = 0$, then this equation implies that $\tilde{\alpha}_-$ is an $\mathbb{O}$-operator of weight zero. If $\kappa \neq 0$, then β is a balanced A-bimodule homomorphism, which, according to Lemma 3.8, implies that $\tilde{\beta}_+$ from $(\mathscr{A}^*, R^*_{\mathscr{A}}, L^*_{\mathscr{A}})$ to $\mathscr{A}$ is a balanced $\mathscr{A}$-bimodule homomorphism. So $\tilde{\alpha}_-$ is an extended $\mathbb{O}$-operator with modification $\tilde{\beta}_+$ of mass κ.

Conversely, suppose $\tilde{\alpha}_-$ is an extended $\mathbb{O}$-operator with modification $\tilde{\beta}_+$ of mass κ. If $\kappa \neq 0$, then $\tilde{\beta}_+$ from $(\mathscr{A}^*, R^*_{\mathscr{A}}, L^*_{\mathscr{A}})$ to $\mathscr{A}$ is a balanced $\mathscr{A}$-bimodule homomorphism, which by Lemma 3.8 implies that β from (V, ℓ, r) to A is a balanced A-bimodule homomorphism. Moreover, for any $u, v \in V$ we have

$$\tilde{\alpha}_-(u)\cdot\tilde{\alpha}_-(v) - \tilde{\alpha}_-(R^*_{\mathscr{A}}(\tilde{\alpha}_-(u))v + uL^*_{\mathscr{A}}(\tilde{\alpha}_-(v))) = \kappa\tilde{\beta}_+(u)\tilde{\beta}_+(v), \tag{3-17}$$

which implies that α is an extended $\mathbb{O}$-operator with modification β of mass κ. If $\kappa = 0$, then (3-17) for $\kappa = 0$ implies that α is an $\mathbb{O}$-operator of weight zero. □

By Theorem 3.9, the results from the previous sections on $\mathbb{O}$-operators on A can be applied to general $\mathbb{O}$-operators.

Corollary 3.10. *Let A be a $\Bbbk$-algebra and let V be an A-bimodule, both with finite $\Bbbk$-dimension.*

(i) *Suppose the characteristic of the field $\Bbbk$ is not 2. Let $\alpha, \beta : V \to A$ be linear maps that are identified as elements in $(A \ltimes_{r^*,\ell^*} V^*) \otimes (A \ltimes_{r^*,\ell^*} V^*)$. Then α is an extended $\mathbb{O}$-operator with modification β of mass k if and only if $(\alpha - \alpha^{21}) \pm (\beta + \beta^{21})$ is a solution of the EAYBE of mass $(\kappa+1)/4$ in $A \ltimes_{r^*,\ell^*} V^*$.*

(ii) *Let $\alpha : V \to A$ be a linear map identified as an element in $(A \ltimes_{r^*,\ell^*} V^*) \otimes (A \ltimes_{r^*,\ell^*} V^*)$. Then α is an $\mathbb{O}$-operator of weight zero if and only if $\alpha - \alpha^{21}$ is a skew-symmetric solution of the AYBE in* (3-2) *in $A \ltimes_{r^*,\ell^*} V^*$. In particular, a linear map $P : A \to A$ is a Rota–Baxter operator of weight zero if and only if $r = P - P^{21}$ is a skew-symmetric solution of the AYBE in $A \ltimes_{R^*,L^*} A^*$.*

(iii) *Let $\alpha, \beta : V \to A$ be two linear maps identified as elements in $(A \ltimes_{r^*,\ell^*} V^*) \otimes (A \ltimes_{r^*,\ell^*} V^*)$. Then α is an extended $\mathbb{O}$-operator with modification β of mass -1 if and only if $(\alpha - \alpha^{21}) \pm (\beta + \beta^{21})$ is a solution of the AYBE in $A \ltimes_{r^*,\ell^*} V^*$.*

(iv) *Let $\alpha : A \to A$ be a linear map identified as an element in $(A \ltimes_{R^*,L^*} A^*) \otimes (A \ltimes_{R^*,L^*} A^*)$. Then α satisfies* (2-19) *if and only if $(\alpha - \alpha^{21}) \pm (\mathrm{id} + \mathrm{id}^{21})$ is a solution of the AYBE in $A \ltimes_{R^*,L^*} A^*$.*

(v) *Let $P : A \to A$ be a linear map identified as an element of $A \ltimes_{R^*,L^*} A^*$. Then P is a Rota–Baxter operator of weight $\lambda \neq 0$ if and only if $2/\lambda(P - P^{21}) + 2\,\mathrm{id}$ and $(2/\lambda)(P - P^{21}) - 2\,\mathrm{id}^{21}$ are both solutions of the AYBE in $A \ltimes_{R^*,L^*} A^*$.*

Proof. (i) This follows from Theorem 3.9 and Theorem 3.5.

(ii) This follows from Theorem 3.9 for $\kappa = 0$ (or $\beta = 0$) and Corollary 3.6.

(iii) This follows from Theorem 3.9 for $\kappa = -1$ and Corollary 3.6.

(iv) This follows from item (iii) in the case that $(V, r, \ell) = (A, L, R)$ and $\beta = \mathrm{id}$.

(v) By [Ebrahimi-Fard 2002] (see also the discussion after Corollary 2.15), P is a Rota–Baxter operator of weight $\lambda \neq 0$ if and only if $2P/\lambda + \mathrm{id}$ is an extended $\mathbb{O}$-operator with modification id of mass -1 from (A, L, R) to A, that is, $2P/\lambda + \mathrm{id}$ satisfies (2-19). Then the conclusion follows from item (iv). □

3d. *$\mathbb{O}$-operators and AYBE on Frobenius algebras.* Here we consider the relationship between $\mathbb{O}$-operators and solutions of the AYBE on Frobenius algebras.

Definition 3.11. (i) Let A be a $\Bbbk$-algebra and let $B(\cdot, \cdot) : A \otimes A \to \Bbbk$ be a nondegenerate bilinear form. Let $\varphi : A \to A^*$ denote the induced injective linear map defined by

$$B(x, y) = \langle \varphi(x), y \rangle \quad \text{for all } x, y \in A. \tag{3-18}$$

(ii) A *Frobenius* $\Bbbk$*-algebra* is a $\Bbbk$-algebra $(A, \cdot)$ together with a nondegenerate bilinear form $B(\cdot, \cdot) : A \otimes A \to \Bbbk$ that is invariant in the sense that

$$B(x \cdot y, z) = B(x, y \cdot z) \quad \text{for all } x, y, z \in A.$$

We use $(A, \cdot, B)$ to denote a Frobenius $\Bbbk$-algebra.

(iii) A Frobenius $\Bbbk$-algebra is called *symmetric* if

$$B(x, y) = B(y, x) \quad \text{for all } x, y \in A.$$

(iv) A linear map $\beta : A \to A$ is called *self-adjoint* with respect to a bilinear form B if for any $x, y \in A$, we have $B(\beta(x), y) = B(x, \beta(y))$, and *skew-adjoint* if $B(\beta(x), y) = -B(x, \beta(y))$.

A symmetric Frobenius $\Bbbk$-algebra is also simply called a *symmetric* $\Bbbk$*-algebra* [Brauer and Nesbitt 1937]. We will not use this term to avoid confusion with the symmetrization of the tensor algebra. Frobenius algebras have found applications in broad areas of mathematics and physics. See [Bai 2010; Yamagata 1996] for further details.

It is easy to get the following result.

Proposition 3.12 [Yamagata 1996]. *Let A be a symmetric Frobenius $\Bbbk$-algebra with finite $\Bbbk$-dimension. Then the A-bimodule (A, L, R) is isomorphic to the A-bimodule (A^*, R^*, L^*).*

The following statement gives a class of symmetric Frobenius algebras from symmetric, invariant tensors.

Corollary 3.13. *Let $(A, \cdot)$ be a $\Bbbk$-algebra with finite $\Bbbk$-dimension. Let $s \in A \otimes A$ be symmetric and invariant. Suppose that s regarded as a linear map from $A^* \to A$ is invertible. Then $s^{-1} : A \to A^*$ regarded as a bilinear form $B(\cdot, \cdot) : A \otimes A \to \Bbbk$ on A through* (3-18) *for $\varphi = s^{-1}$ is symmetric, nondegenerate and invariant. Thus $(A, \cdot, B)$ is a symmetric Frobenius algebra.*

Proof. Since s is symmetric and s regarded as a linear map from A^* to A is invertible, $B(\cdot, \cdot)$ is symmetric and nondegenerate. On the other hand, since s is invariant, (3-7) holds by Lemma 3.2. Thus, for any $x, y, z \in A$ and $a^* = s^{-1}(x)$, $b^* = s^{-1}(y)$ and $c^* = s^{-1}(z)$, we have

$$\begin{aligned} B(x \cdot y, z) &= \langle c^*, s(a^*) \cdot s(b^*) \rangle = \langle c^* L^*(s(a^*)), b^* \rangle \\ &= \langle R^*(s(c^*))a^*, b^* \rangle = \langle a^*, s(b^*) \cdot s(c^*) \rangle = B(x, y \cdot z), \end{aligned}$$

that is, $B(\cdot, \cdot)$ is invariant. So the conclusion follows. □

Lemma 3.14. *Let $(A, \cdot, B)$ be a symmetric Frobenius $\Bbbk$-algebra with finite $\Bbbk$-dimension. Suppose that $\beta : A \to A$ is an endomorphism of A that is self-adjoint with respect to B. Then $\tilde{\beta} = \beta\varphi^{-1} : A^* \to A$ regarded as an element of $A \otimes A$*

is symmetric, where $\varphi : A \to A^*$ is defined by (3-18)*. Moreover, β is a balanced A-bimodule homomorphism if and only if $\tilde{\beta}$ is.*

Proof. Since β is self-adjoint with respect to B, it is easy to show that $\tilde{\beta}$ regarded as an element of $A \otimes A$ is symmetric. Moreover, for any $a^*, b^* \in A^*$, $z \in A$ and $x = \varphi^{-1}(a^*)$, $y = \varphi^{-1}(b^*)$, we have

$$\begin{aligned}\langle R^*(\tilde{\beta}(a^*))b^*, z\rangle &= \langle R^*(\beta(x))\varphi(y), z\rangle = B(y, z\cdot\beta(x)),\\ \langle a^*L^*(\tilde{\beta}(b^*)), z\rangle &= \langle \varphi(x)L^*(\beta(y)), z\rangle = B(x, \beta(y)\cdot z)\\ &= B(\beta(y), z\cdot x) = B(y, \beta(z\cdot x)).\end{aligned}$$

Thus $\tilde{\beta}$ satisfies (3-7) if and only if $\beta(z\cdot x) = z\cdot\beta(x)$ for any $x, z \in A$. On the other hand,

$$\begin{aligned}\langle R^*(\tilde{\beta}(a^*))b^*, z\rangle &= \langle R^*(\beta(x))\varphi(y), z\rangle = B(y, z\cdot\beta(x))\\ &= B(\beta(x), y\cdot z) = B(x, \beta(y\cdot z)),\\ \langle a^*L^*(\tilde{\beta}(b^*)), z\rangle &= \langle \varphi(x)L^*(\beta(y)), z\rangle = B(x, \beta(y)\cdot z).\end{aligned}$$

Therefore, $\tilde{\beta}$ satisfies (3-7) if and only if $\beta(y\cdot z) = \beta(y)\cdot z$ for any $y, z \in A$. Hence β is an A-bimodule homomorphism if and only if $\tilde{\beta}$ is. □

If $\beta = \mathrm{id}$, the lemma above states that $\varphi^{-1} : A^* \to A$ is a balanced A-bimodule homomorphism.

Corollary 3.15. *Let $(A, \cdot, B)$ be a symmetric Frobenius $\Bbbk$-algebra of finite $\Bbbk$-dimension and let $\varphi : A \to A^*$ be the linear map defined by* (3-18)*. Suppose $\beta \in A \otimes A$ is symmetric. Then β regarded as a linear map from (A^*, R^*, L^*) to A is a balanced A-bimodule homomorphism if and only if $\hat{\beta} = \beta\varphi : A \to A$ is a balanced A-bimodule homomorphism.*

Proof. In fact, $\hat{\beta} = \beta\varphi$ is self-adjoint with respect to $B(\cdot, \cdot)$ since for any $x, y \in A$,

$$\begin{aligned}\langle \beta, \varphi(x)\otimes\varphi(y)\rangle = \langle \beta, \varphi(y)\otimes\varphi(x)\rangle &\iff \langle \beta(\varphi(x)), \varphi(y)\rangle = \langle \beta(\varphi(y)), \varphi(x)\rangle\\ &\iff B(\hat{\beta}(x), y) = B(\hat{\beta}(y), x).\end{aligned}$$

So the conclusion follows from Lemma 3.14. □

Theorem 3.16. *Let $\Bbbk$ be a field of characteristic not equal to 2. Let $(A, \cdot, B)$ be a symmetric Frobenius algebra of finite $\Bbbk$-dimension. Suppose that α and β are two endomorphisms of A and that β is self-adjoint with respect to B.*

(i) *α is an extended ©-operator with modification β of mass κ if and only if $\tilde{\alpha} := \alpha\circ\varphi^{-1} : A^* \to A$ is an extended ©-operator with modification $\tilde{\beta} := \beta\circ\varphi^{-1} : A^* \to A$ of mass κ, where the linear map $\varphi : A \to A^*$ is defined by* (3-18)*.*

(ii) *Suppose that in addition, α is skew-adjoint with respect to B. Then $\tilde{\alpha}$ regarded as an element of $A \otimes A$ is skew-symmetric and*

(a) *$r_\pm = \tilde{\alpha} \pm \tilde{\beta}$ regarded as an element of $A \otimes A$ is a solution of the EAYBE of mass $(\kappa+1)/4$ if and only if α is an extended $\mathbb{O}$-operator with modification β of mass k;*

(b) *if $\kappa = -1$, then $r_\pm = \tilde{\alpha} \pm \tilde{\beta}$ regarded as an element of $A \otimes A$ is a solution of the AYBE if and only if α is an extended $\mathbb{O}$-operator with modification β of mass -1; and*

(c) *if $\kappa = 0$, then $\tilde{\alpha}$ regarded as an element of $A \otimes A$ is a solution of the AYBE if and only if α is a Rota–Baxter operator of weight zero.*

Proof. (i) Since B is symmetric and invariant, for any $x, y, z \in A$, we have

$$\begin{aligned} B(x \cdot y, z) = B(x, y \cdot z) &\Longleftrightarrow \langle \varphi(x \cdot y), z\rangle = \langle \varphi(x), y \cdot z\rangle \\ &\Longleftrightarrow \varphi(xR(y)) = \varphi(x)L^*(y), \end{aligned} \tag{3-19}$$

$$\begin{aligned} B(z, x \cdot y) = B(y \cdot z, x) &\Longleftrightarrow \langle \varphi(z), x \cdot y\rangle = \langle \varphi(y \cdot z), x\rangle \\ &\Longleftrightarrow R^*(y)\varphi(z) = \varphi(L(y)z). \end{aligned} \tag{3-20}$$

On the other hand, since φ is invertible, for any $a^*, b^* \in A^*$, there exist $x, y \in A$ such that $\varphi(x) = a^*, \varphi(y) = b^*$. So according to (3-19) and (3-20), the equation

$$\tilde{\alpha}(a^*) \cdot \tilde{\alpha}(b^*) - \tilde{\alpha}(\varphi(\tilde{\alpha}(a^*) \cdot \varphi^{-1}(b^*) + \varphi^{-1}(a^*) \cdot \tilde{\alpha}(b^*))) = \kappa \tilde{\beta}(a^*) \cdot \tilde{\beta}(b^*),$$

is equivalent to

$$\tilde{\alpha}(a^*) \cdot \tilde{\alpha}(b^*) - \tilde{\alpha}(R^*(\tilde{\alpha}(a^*))b^* + a^*L^*(\tilde{\alpha}(b^*))) = \kappa \tilde{\beta}(a^*) \cdot \tilde{\beta}(b^*).$$

By Lemma 3.14, $\beta : A \to A$ is a balanced A-bimodule homomorphism if and only if $\tilde{\beta} : A^* \to A$ is. So α is an extended $\mathbb{O}$-operator with modification β of mass κ if and only if $\tilde{\alpha}$ is an extended $\mathbb{O}$-operator with modification $\tilde{\beta}$ of mass κ.

(i) If α is skew-adjoint with respect to B, then

$$\langle \alpha(x), \varphi(y)\rangle + \langle \varphi(x), \alpha(y)\rangle = 0 \quad \text{for all } x, y \in A.$$

Hence $\langle \tilde{\alpha}(a^*), b^*\rangle + \langle a^*, \tilde{\alpha}(b^*)\rangle = 0$ for any $a^*, b^* \in A^*$. So $\tilde{\alpha}$ regarded as an element of $A \otimes A$ is skew-symmetric.

By Theorem 3.5, item (a) holds. By Corollary 3.6, items (b) and (c) hold. □

Corollary 3.17. *Let $\Bbbk$ be a field of characteristic not equal to 2. Let A be a $\Bbbk$-algebra of finite $\Bbbk$-dimension and let $r \in A \otimes A$. Define $\alpha, \beta \in A \otimes A$ by* (3-5). *Then $r = \alpha + \beta$. Let $B : A \otimes A \to \Bbbk$ be a nondegenerate symmetric and invariant bilinear form. Define the linear map $\varphi : A \to A^*$ by* (3-18).

(i) *Suppose that $\beta \in A \otimes A$ is invariant. Then r is a solution of the EAYBE of mass $(\kappa+1)/4$ if and only if $\hat{\alpha} = \alpha\varphi : A \to A$ is an extended $\mathbb{O}$-operator with modification $\hat{\beta} = \beta\varphi : A \to A$ of mass k.*

(ii) *Suppose that $\beta \in A \otimes A$ is invariant. Then r is a solution of the AYBE if and only if $\hat{\alpha} = \alpha\varphi : A \to A$ is an extended $\mathbb{O}$-operator with modification $\hat{\beta} = \beta\varphi : A \to A$ of mass -1. If in addition, $\beta = 0$, that is, r is skew-symmetric, then r is a solution of the AYBE if and only if $\hat{\alpha} = \hat{r} = r\varphi : A \to A$ is a Rota–Baxter operator of weight zero.*

Proof. By the proof of Corollary 3.15, we show that $\hat{\beta} = \beta\varphi$ is self-adjoint with respect to $B(\cdot,\cdot)$ since $\beta \in A \otimes A$ is symmetric. Similarly, since $\alpha \in A \otimes A$ is skew-symmetric, $\hat{\alpha} = \alpha\varphi$ is skew-adjoint with respect to $B(\cdot,\cdot)$. So the conclusion follows from Theorem 3.16. □

4. Extended $\mathbb{O}$-operators and the generalized associative Yang–Baxter equation

We define the generalized associative Yang–Baxter equation and study its relationship with extended $\mathbb{O}$-operators.

4a. ***Generalized associative Yang–Baxter equation.*** We adapt the same notation as in Definition 3.1.

The following proposition (also see [Aguiar 2000a, Proposition 5.1]) is related to the construction of variations of bialgebras under the names of associative D-bialgebras [Zhelyabin 1997], balanced infinitesimal bialgebras (in the opposite algebras) [Aguiar 2001] and antisymmetric infinitesimal bialgebras [Bai 2010].

Proposition 4.1 [Aguiar 2000a; 2001; Bai 2010]. *Let A be a $\Bbbk$-algebra with finite $\Bbbk$-dimension and let $r \in A \otimes A$. Define $\Delta : A \to A \otimes A$ by*

$$\Delta(x) = (\mathrm{id} \otimes L(x) - R(x) \otimes \mathrm{id})r \quad \text{for all } x \in A. \tag{4-1}$$

Then

$$\Delta^* : A^* \otimes A^* \hookrightarrow (A \otimes A)^* \to A^* \tag{4-2}$$

defines an associative multiplication on A^* if and only if r is a solution of the equation

$$(\mathrm{id} \otimes \mathrm{id} \otimes L(x) - R(x) \otimes \mathrm{id} \otimes \mathrm{id})(r_{12}r_{13} + r_{13}r_{23} - r_{23}r_{12}) = 0 \quad \text{for all } x \in A. \tag{4-3}$$

Definition 4.2. Let A be a $\Bbbk$-algebra. Equation (4-3) is called the *generalized associative Yang–Baxter equation (GAYBE)*. An element $r \in A \otimes A$ satisfying (4-3) is called a *solution of the GAYBE in A*.

Lemma 4.3. *Let $(A, \cdot)$ be a $\Bbbk$-algebra with finite $\Bbbk$-dimension. Let $r \in A \otimes A$. The multiplication $*$ on A^* defined by* (4-2) *is also given by*

$$a^* * b^* = R^*(r(a^*))b^* - L^*(r^t(b^*))a^* \quad \text{for all } a^*, b^* \in A^*. \tag{4-4}$$

Proof. Let $\{e_1, \ldots, e_n\}$ be a basis of A and $\{e_1^*, \ldots, e_n^*\}$ be its dual basis. Suppose that $r = \sum_{i,j} a_{i,j} e_i \otimes e_j$ and $e_i \cdot e_j = \sum_k c_{i,j}^k e_k$. Then for any k and l, we have

$$\begin{aligned} e_k^* * e_l^* &= \sum_s \langle e_k^* \otimes e_l^*, \Delta(e_s)\rangle e_s^* \\ &= \sum_s \langle e_k^* \otimes e_l^*, (\mathrm{id} \otimes L(e_s) - R(e_s) \otimes \mathrm{id})r\rangle e_s^* \\ &= \sum_{s,t} (a_{k,t} c_{s,t}^l - c_{t,s}^k a_{t,l}) e_s^* = R^*(r(e_k^*))e_l^* - L^*(r^t(e_l^*))e_k^*. \end{aligned}$$

□

This lemma suggests that we apply the approach considered in Section 2b. More precisely, we take the A-bimodule $\Bbbk$-algebra $(R, \circ, \ell, r)$ to be (A^*, R^*, L^*) with the zero multiplication and set

$$\delta_+ = r \quad \text{and} \quad \delta_- = -r^t. \tag{4-5}$$

Assume that $\Bbbk$ has characteristic not equal to 2 and define

$$\alpha = (r - r^t)/2 \quad \text{and} \quad \beta = (r + r^t)/2, \tag{4-6}$$

that is, α and β are the *skew-symmetric part* and the *symmetric part* of r. So $r = \alpha + \beta$ and $r^t = -\alpha + \beta$.

Proposition 4.4. *Let $\Bbbk$ have characteristic not equal to 2. Let $(A, \cdot)$ be a $\Bbbk$-algebra with finite $\Bbbk$-dimension and $r \in A \otimes A$. Let α and β be given by* (4-6)*. Suppose that β is a balanced A-bimodule homomorphism, that is, β satisfies* (3-6)*. If α is an extended $\mathbb{O}$-operator with modification β of any mass $\kappa \in \Bbbk$, then the product defined by* (4-4) *defines a $\Bbbk$-algebra structure on A^* and r is a solution of the GAYBE.*

Proof. By applying Theorem 2.12 to the A-bimodule $\Bbbk$-algebra $(R, \circ, \ell, r)$ we constructed before the proposition, we see that the product defined by (4-4) is associative. Then r is a solution of the GAYBE by Lemma 4.3. □

Corollary 4.5. *Under the assumptions of Proposition 4.4, a solution of the EAYBE of any mass $\kappa \in \Bbbk$ is also a solution of the GAYBE.*

Proof. Let r be a solution of the EAYBE of mass κ. Define α and β by (4-6). Then by Theorem 3.5, α is an extended $\mathbb{O}$-operator with modification β of mass $4\kappa - 1$. Hence by Proposition 4.4, r is a solution of the GAYBE. □

4b. *GAYBE and extended ©-operators.* We now consider the operator form of GAYBE with emphasis on its relationship with extended ©-operators.

Lemma 4.6. *Let A be a $\Bbbk$-algebra and (V, ℓ, r) be a bimodule. Let $\alpha : V \to A$ be a linear map. Then the product*

$$u *_\alpha v := \ell(\alpha(u))v + ur(\alpha(v)) \quad \text{for all } u, v \in V, \tag{4-7}$$

defines a $\Bbbk$-algebra structure on V if and only if

$$\ell\big(\alpha(u)\cdot\alpha(v) - \alpha(u *_\alpha v)\big)w = ur\big(\alpha(v)\cdot\alpha(w) - \alpha(v *_\alpha w)\big) \quad \text{for all } u, v \in V. \tag{4-8}$$

Proof. It follows from Lemma 2.10 by setting $(R, \ell, r) = (V, \ell, r)$ and $\lambda = 0$. □

Theorem 4.7. *Let A be a $\Bbbk$-algebra and (V, ℓ, r) be an A-bimodule, both of finite dimension over $\Bbbk$. Let $\alpha : V \to A$ be a linear map. Using the same notation as in Definition 3.7, $\tilde{\alpha}_-$ identified as an element of $\mathcal{A} \otimes \mathcal{A}$ is a skew-symmetric solution of the GAYBE* (4-3) *if and only if* (4-8) *and the equations*

$$\alpha(u)\cdot\alpha(\ell(x)v) - \alpha(u *_\alpha (\ell(x)v)) = \alpha(ur(x))\cdot\alpha(v) - \alpha((ur(x)) *_\alpha v), \tag{4-9}$$

$$\alpha(u)\cdot\alpha(vr(x)) - \alpha(u *_\alpha (vr(x))) = (\alpha(u)\cdot\alpha(v))\cdot x - \alpha(u *_\alpha v)\cdot x, \tag{4-10}$$

$$\alpha(\ell(x)u)\cdot\alpha(v) - \alpha((\ell(x)u) *_\alpha v) = x\cdot(\alpha(u)\cdot\alpha(v)) - x\cdot\alpha(u *_\alpha v) \tag{4-11}$$

hold for any $u, v \in V$, $x \in A$.

Proof. By Proposition 4.1, Lemma 4.3 and Lemma 4.6, we see that $\tilde{\alpha}_- \in \mathcal{A} \otimes \mathcal{A}$ is a skew-symmetric solution of the GAYBE (4-3) if and only if for any $u, v, w \in V$ and $a^*, b^*, c^* \in A^*$,

$$\begin{aligned}
&R^*_{\mathcal{A}}\big(\tilde{\alpha}_-(u+a^*)\cdot\tilde{\alpha}_-(v+b^*) - \tilde{\alpha}_-(R^*_{\mathcal{A}}(\tilde{\alpha}_-(u+a^*))(v+b^*) \\
&\qquad + (u+a^*)L^*_{\mathcal{A}}(\tilde{\alpha}_-(v+b^*)))\big)(w+c^*) \\
&= (u+a^*)L^*_{\mathcal{A}}\big(\tilde{\alpha}_-(v+b^*)\cdot\tilde{\alpha}_-(w+c^*) - \tilde{\alpha}_-(R^*_{\mathcal{A}}(\tilde{\alpha}_-(v+b^*))(w+c^*) \\
&\qquad + (v+b^*)L^*_{\mathcal{A}}(\tilde{\alpha}_-(w+c^*)))\big),
\end{aligned}$$

By the proof of Theorem 3.9, the equation above is equivalent to

$$\begin{aligned}
&R^*_{\mathcal{A}}\big(\alpha(u)\cdot\alpha(v) - \alpha(\ell(\alpha(u))v) - \alpha(ur(\alpha(v))) - r^*(\alpha(u))\alpha^*(b^*) + \alpha^*(R^*_{\mathcal{A}}(\alpha(u))b^*) \\
&\qquad - \alpha^*(uL^*_{\mathcal{A}}(\alpha^*(b^*))) - \alpha^*(a^*)\ell^*(\alpha(v)) - \alpha^*(R^*_{\mathcal{A}}(\alpha^*(a^*))v) + \alpha^*(a^*L^*_{\mathcal{A}}(\alpha(v)))\big)w \\
&+ R^*_{\mathcal{A}}\big(\alpha(u)\cdot\alpha(v) - \alpha(\ell(\alpha(u))v) - \alpha(ur(\alpha(v))) - r^*(\alpha(u))\alpha^*(b^*) + \alpha^*(R^*_{\mathcal{A}}(\alpha(u))b^*) \\
&\qquad - \alpha^*(uL^*_{\mathcal{A}}(\alpha^*(b^*))) - \alpha^*(a^*)\ell^*(\alpha(v)) - \alpha^*(R^*_{\mathcal{A}}(\alpha^*(a^*))v) + \alpha^*(a^*L^*_{\mathcal{A}}(\alpha(v)))\big)c^* \\
&= uL^*_{\mathcal{A}}\big(\alpha(v)\cdot\alpha(w) - \alpha(\ell(\alpha(v))w) - \alpha(vr(\alpha(w))) - r^*(\alpha(v))\alpha^*(c^*) + \alpha^*(R^*_{\mathcal{A}}(\alpha(v))c^*) \\
&\qquad - \alpha^*(vL^*_{\mathcal{A}}(\alpha^*(c^*))) - \alpha^*(b^*)\ell^*(\alpha(w)) - \alpha^*(R^*_{\mathcal{A}}(\alpha^*(b^*))w) + \alpha^*(b^*L^*_{\mathcal{A}}(\alpha(w)))\big) \\
&+ a^*L^*_{\mathcal{A}}(\alpha(v)\cdot\alpha(w) - \alpha(\ell(\alpha(v))w) - \alpha(vr(\alpha(w))) - r^*(\alpha(v))\alpha^*(c^*) + \alpha^*(R^*_{\mathcal{A}}(\alpha(v))c^*) \\
&\qquad - \alpha^*(vL^*_{\mathcal{A}}(\alpha^*(c^*))) - \alpha^*(b^*)\ell^*(\alpha(w)) - \alpha^*(R^*_{\mathcal{A}}(\alpha^*(b^*))w) + \alpha^*(b^*L^*_{\mathcal{A}}(\alpha(w)))).
\end{aligned}$$

By suitable choices of $u, v, w \in V$ and $a^*, b^*, c^* \in A^*$, we find that this equation holds if and only if the following equations hold:

$$R^*_{\mathscr{A}}(\alpha(u)\cdot\alpha(v)-\alpha(\ell(\alpha(u))v+ur(\alpha(v))))w = uL^*_{\mathscr{A}}(\alpha(v)\cdot\alpha(w)-\alpha(\ell(\alpha(v))w+vr(\alpha(w)))) \quad \text{(take } a^*=b^*=c^*=0\text{)}, \tag{4-12}$$

$$R^*_{\mathscr{A}}(-r^*(\alpha(u))\alpha^*(b^*)+\alpha^*(R^*_{\mathscr{A}}(\alpha(u))b^*)-\alpha^*(uL^*_{\mathscr{A}}(\alpha^*(b^*))))w = uL^*_{\mathscr{A}}(-\alpha^*(b^*)\ell^*(\alpha(w))-\alpha^*(R^*_{\mathscr{A}}(\alpha^*(b^*))w)+\alpha^*(b^*L^*_{\mathscr{A}}(\alpha(w)))) \quad \text{(take } v=a^*=c^*=0\text{)}, \tag{4-13}$$

$$R^*_{\mathscr{A}}(-\alpha^*(a^*)\ell^*(\alpha(v))-\alpha^*(R^*_{\mathscr{A}}(\alpha^*(a^*))v)+\alpha^*(a^*L^*_{\mathscr{A}}(\alpha(v))))w = a^*L^*_{\mathscr{A}}(\alpha(v)\cdot\alpha(w)-\alpha(\ell(\alpha(v))w)-\alpha(vr(\alpha(w)))) \quad \text{(take } u=b^*=c^*=0\text{)}, \tag{4-14}$$

$$R^*_{\mathscr{A}}(\alpha(u)\cdot\alpha(v)-\alpha(\ell(\alpha(u))v)-\alpha(ur(\alpha(v))))c^* = uL^*_{\mathscr{A}}(-r^*(\alpha(v))\alpha^*(c^*)+\alpha^*(R^*_{\mathscr{A}}(\alpha(v))c^*)-\alpha^*(vL^*_{\mathscr{A}}(\alpha^*(c^*)))) \quad \text{(take } w=a^*=b^*=0\text{)}, \tag{4-15}$$

$$R^*_{\mathscr{A}}(-r^*(\alpha(u))\alpha^*(b^*)+\alpha^*(R^*_{\mathscr{A}}(\alpha(u))b^*)-\alpha^*(uL^*_{\mathscr{A}}(\alpha^*(b^*))))c^* = 0 \quad \text{(take } v=w=a^*=0\text{)}, \tag{4-16}$$

$$R^*_{\mathscr{A}}(-\alpha^*(a^*)\ell^*(\alpha(v))-\alpha^*(R^*_{\mathscr{A}}(\alpha^*(a^*))v)+\alpha^*(a^*L^*_{\mathscr{A}}(\alpha(v))))c^* = a^*L^*_{\mathscr{A}}(-r^*(\alpha(v))\alpha^*(c^*)+\alpha^*(R^*_{\mathscr{A}}(\alpha(v))c^*)-\alpha^*(vL^*_{\mathscr{A}}(\alpha^*(c^*)))) \quad \text{(take } u=w=b^*=0\text{)}, \tag{4-17}$$

$$a^*L^*_{\mathscr{A}}(-\alpha^*(b^*)\ell^*(\alpha(w))-\alpha^*(R^*_{\mathscr{A}}(\alpha^*(b^*))w)+\alpha^*(b^*L^*_{\mathscr{A}}(\alpha(w)))) = 0 \quad \text{(take } u=v=c^*=0\text{)}. \tag{4-18}$$

Thus we just need to prove

(i) (4-12) $\Longleftrightarrow$ (4-8), (ii) (4-13) $\Longleftrightarrow$ (4-9),

(iii) (4-14) $\Longleftrightarrow$ (4-10), (iv) (4-15) $\Longleftrightarrow$ (4-11),

(v) both sides of (4-17) equal zero, (vi) (4-16) and (4-18) hold.

The proofs of these statements are similar. So we just prove that (4-13) holds if and only if (4-9) holds. Let *LHS* and *RHS* denote the left-hand side and right-hand side of (4-13). Then for any $x \in A$ and $s^* \in V^*$, we have

$$\langle LHS, s^*\rangle = \langle RHS, s^*\rangle = 0.$$

Further

$$\begin{aligned}
\langle LHS, x\rangle &= \langle w, -r^*(x)(r^*(\alpha(u))\alpha^*(b^*)) \\
&\qquad + r^*(x)\alpha^*(R^*_{\mathcal{A}}(\alpha(u))b^*) - r^*(x)\alpha^*(uL^*_{\mathcal{A}}(\alpha^*(b^*)))\rangle \\
&= \langle -\alpha((wr(x))r(\alpha(u))) + \alpha(wr(x))\cdot\alpha(u), b^*\rangle \\
&\qquad - \langle \alpha^*(b^*)\cdot\alpha(wr(x)), u\rangle \\
&= \langle -\alpha((wr(x))r(\alpha(u))) + \alpha(wr(x))\cdot\alpha(u) - \alpha(\ell(\alpha(wr(x)))u), b^*\rangle, \\
\langle RHS, x\rangle &= \langle u, -(\alpha^*(b^*)\ell^*(\alpha(w)))\ell^*(x) \\
&\qquad - \alpha^*(R^*_{\mathcal{A}}(\alpha^*(b^*))w)\ell^*(x) + \alpha^*(b^*L^*_{\mathcal{A}}(\alpha(w)))\ell^*(x)\rangle \\
&= \langle -\alpha(\ell(\alpha(w))(\ell(x)u)), b^*\rangle \\
&\qquad - \langle \alpha(\ell(x)u)\cdot\alpha^*(b^*), w\rangle + \langle \alpha(w)\cdot\alpha(\ell(x)w), b^*\rangle \\
&= \langle -\alpha(\ell(\alpha(w))(\ell(x)u)) - \alpha(wr(\alpha(\ell(x)u))) + \alpha(w)\cdot\alpha(\ell(x)u), b^*\rangle.
\end{aligned}$$

So (4-13) holds if and only if (4-9) holds. □

Equations (4-9)–(4-11) in Theorem 4.7 can be regarded as an operator form of GAYBE. To get a more manageable form, we restrict below to the case of extended $\mathbb{O}$-operators.

Corollary 4.8. *Let $(A, \cdot)$ be a $\Bbbk$-algebra with finite $\Bbbk$-dimension.*

(i) *Let $(R, \circ, \ell, r)$ be an A-bimodule $\Bbbk$-algebra with finite $\Bbbk$-dimension. Let $\alpha, \beta : R \to A$ be two linear maps such that α is an extended $\mathbb{O}$-operator of weight λ with modification β of mass (κ, μ), that is, β is an A-bimodule homomorphism and the conditions* (2-5) *and* (2-6) *in Definition 2.7 hold, and α and β satisfy* (2-7)*. Then $\alpha - \alpha^{21}$, when identified as an element of $(A \ltimes_{r^*,\ell^*} R^*)\otimes(A \ltimes_{r^*,\ell^*} R^*)$, is a skew-symmetric solution of the GAYBE* (4-3) *if and only if*

$$\lambda\ell(\alpha(u\circ v))w = \lambda ur(\alpha(v\circ w)) \qquad \text{for all } u, v, w \in R, \tag{4-19}$$

$$\lambda\alpha(u(vr(x))) = \lambda\alpha(u\circ v)\cdot x \qquad \text{for all } u, v \in R, x \in A, \tag{4-20}$$

$$\lambda\alpha((\ell(x)u)\circ v) = \lambda x\cdot\alpha(u\circ v) \qquad \text{for all } u, v \in R, x \in A. \tag{4-21}$$

In particular, when $\lambda = 0$, that is, α is an extended $\mathbb{O}$-operator of weight zero with modification β of mass (κ, μ), then $\alpha - \alpha^{21}$ identified as an element of $(A \ltimes_{r^*,\ell^*} R^*)\otimes(A \ltimes_{r^*,\ell^*} R^*)$ is a skew-symmetric solution of the GAYBE (4-3).

(ii) *Let $(R, \circ, \ell, r)$ be an A-bimodule $\Bbbk$-algebra with finite $\Bbbk$-dimension. Let $\alpha : R \to A$ be an $\mathbb{O}$-operator of weight λ. Then $\alpha - \alpha^{21}$ identified as an element of $(A \ltimes_{r^*,\ell^*} R^*) \otimes (A \ltimes_{r^*,\ell^*} R^*)$ is a skew-symmetric solution of the GAYBE if and only if* (4-19)–(4-21) *hold.*

(iii) *Let (V, ℓ, r) be a bimodule of A with finite $\Bbbk$-dimension. Let $\alpha, \beta : V \to A$ be two linear maps such that α is an extended $\mathbb{O}$-operator with modification β of mass κ. Then $\alpha - \alpha^{21}$ identified as an element of $(A \ltimes_{r^*,\ell^*} V^*) \otimes (A \ltimes_{r^*,\ell^*} V^*)$ is a skew-symmetric solution of the GAYBE.*

(iv) *Let $\alpha : A \to A$ be a linear endomorphism of A. Suppose that α satisfies* (2-18). *Then $\alpha - \alpha^{21}$ identified as an element of $(A \ltimes_{R^*,L^*} A^*) \otimes (A \ltimes_{R^*,L^*} A^*)$ is a skew-symmetric solution of the GAYBE.*

(v) *Let $(R, \circ, \ell, r)$ be an A-bimodule $\Bbbk$-algebra of finite $\Bbbk$-dimension. Let $\alpha, \beta : R \to A$ be two linear maps such that α is an extended $\mathbb{O}$-operator with modification β of mass $(\kappa, \mu) = (0, \mu)$, that is, β is an A-bimodule homomorphism and the condition* (2-6) *in Definition 2.7 holds, and α and β satisfy*

$$\alpha(u) \cdot \alpha(v) - \alpha(\ell(\alpha(u))v + ur(\alpha(v))) = \mu\beta(u \circ v) \quad \text{for all } u, v \in R.$$

Then $\alpha - \alpha^{21}$ identified as an element of $(A \ltimes_{r^*,\ell^*} R^*) \otimes (A \ltimes_{r^*,\ell^*} R^*)$ is a skew-symmetric solution of the GAYBE.

Proof. (i) Since α is an extended $\mathbb{O}$-operator of weight λ with modification β of mass (κ, μ), by Theorem 4.7, $\alpha - \alpha^{21}$ identified as an element of $(A \ltimes_{r^*,\ell^*} R^*) \otimes (A \ltimes_{r^*,\ell^*} R^*)$ is a skew-symmetric solution of the GAYBE (4-3) if and only if

$$\begin{aligned}
(4\text{-}22)\quad & -\lambda\ell(\alpha(u \circ v))w + \kappa\ell(\beta(u) \cdot \beta(v))w + \mu\ell(\beta(u \circ v))w \\
& \quad = -\lambda ur(\alpha(v \circ w)) + \kappa ur(\beta(v) \cdot \beta(w)) + \mu ur(\beta(v \circ w)), \\
(4\text{-}23)\quad & -\lambda\alpha((ur(x)) \circ v) + \kappa\beta(ur(x)) \cdot \beta(v) + \mu\beta((ur(x)) \circ v) \\
& \quad = -\lambda\alpha(u \circ (l(x)v)) + \kappa\beta(u) \cdot \beta(\ell(x)v) + \mu\beta(u \circ (\ell(x)v)), \\
(4\text{-}24)\quad & -\lambda\alpha(u \circ (vr(x))) + \kappa\beta(u) \cdot \beta(vr(x)) + \mu\beta(u \circ (vr(x))) \\
& \quad = -\lambda\alpha(u \circ v) \cdot x + \kappa(\beta(u) \cdot \beta(v)) \cdot x + \mu\beta(u \circ v) \cdot x, \\
(4\text{-}25)\quad & -\lambda\alpha((\ell(x)u) \circ v) + \kappa\beta(\ell(x)u) \cdot \beta(v) + \mu\beta((\ell(x)u) \circ v) \\
& \quad = -\lambda x \cdot \alpha(u \circ v) + \kappa x \cdot (\beta(u) \cdot \beta(v)) + \mu x \cdot \beta(u \circ v)
\end{aligned}$$

for any $u, v \in R$, $x \in A$. Since β is an A-bimodule homomorphism and the conditions (2-5) and (2-6) in Definition 2.7 hold, we have (4-19) holds if and only if (4-22) holds, (4-20) holds if and only if (4-24) holds, (4-21) holds if and only if (4-25) holds and (4-23) holds automatically.

(ii) This follows from item (i) by setting $\kappa = \mu = 0$.

(iii) This follows from item (i) by setting $\lambda = \mu = 0$.

(iv) This follows from item (iii) for $(V, \ell, r) = (A, L, R)$ and $\beta = \mathrm{id}$.

(v) This follows from item (i) by setting $\lambda = \kappa = 0$. □

References

[Aguiar 2000a] M. Aguiar, "Infinitesimal Hopf algebras", pp. 1–29 in *New trends in Hopf algebra theory* (La Falda, 1999), edited by N. Andruskiewitsch et al., Contemp. Math. **267**, Amer. Math. Soc., Providence, RI, 2000. MR 2001k:16066 Zbl 0982.16028

[Aguiar 2000b] M. Aguiar, "Pre-Poisson algebras", *Letters Math. Phys.* **54**:4 (2000), 263–277. MR 2002k:17041 Zbl 1032.17038

[Aguiar 2001] M. Aguiar, "On the associative analog of Lie bialgebras", *J. Algebra* **244**:2 (2001), 492–532. MR 2003c:17035 Zbl 0991.16033

[Bai 2007] C. Bai, "A unified algebraic approach to the classical Yang–Baxter equation", *J. Phys. A* **40**:36 (2007), 11073–11082. MR 2009b:17014 Zbl 1118.17008

[Bai 2010] C. Bai, "Double constructions of Frobenius algebras, Connes cocycles and their duality", *J. Noncommut. Geom.* **4**:4 (2010), 475–530. MR 2012b:17045 Zbl 05797230

[Bai et al. 2010a] C. Bai, L. Guo, and X. Ni, "©-operators on associative algebras and dendriform algebras", preprint, 2010. arXiv 1003.2432

[Bai et al. 2010b] C. Bai, L. Guo, and X. Ni, "Nonabelian generalized Lax pairs, the classical Yang–Baxter equation and PostLie algebras", *Comm. Math. Phys.* **297** (2010), 553–596. MR 2011i:17032 Zbl 1206.17020

[Bai et al. 2011] C. Bai, L. Guo, and X. Ni, "Generalizations of the classical Yang–Baxter equation and ©-operators", *J. Math. Phys.* **52**:6 (2011), 063515. MR 2841771

[Bai et al. 2012] C. Bai, L. Guo, and X. Ni, "©-operators on associative algebras, associative Yang–Baxter equations and dendriform algebras: a survey", pp. 10–51 in *Quantized algebra and physics* (Tianjin, 2009), edited by M.-L. Ge et al., Nankai Series Pure Appl. Math. Theor. Phys. **8**, World Scientific, Singapore, 2012.

[Baxter 1960] G. Baxter, "An analytic problem whose solution follows from a simple algebraic identity", *Pacific J. Math.* **10**:3 (1960), 731–742. MR 22 #9990 Zbl 0095.12705

[Baxter 1972] R. J. Baxter, "One-dimensional anisotropic Heisenberg chain", *Ann. Physics* **70**:2 (1972), 323–337. MR 45 #8147

[Bordemann 1990] M. Bordemann, "Generalized Lax pairs, the modified classical Yang–Baxter equation, and affine geometry of Lie groups", *Comm. Math. Phys.* **135**:1 (1990), 201–216. MR 91k:58049 Zbl 0714.58025

[Brauer and Nesbitt 1937] R. Brauer and C. Nesbitt, "On the regular representations of algebras", *Proc. Nat. Acad. Sci. USA* **23**:4 (1937), 236–240. JFM 63.0091.06

[Connes and Kreimer 2000] A. Connes and D. Kreimer, "Renormalization in quantum field theory and the Riemann–Hilbert problem, I: The Hopf algebra structure of graphs and the main theorem", *Comm. Math. Phys.* **210**:1 (2000), 249–273. MR 2002f:81070 Zbl 1032.81026

[Ebrahimi-Fard 2002] K. Ebrahimi-Fard, "Loday-type algebras and the Rota–Baxter relation", *Lett. Math. Phys.* **61**:2 (2002), 139–147. MR 2004b:17003 Zbl 1035.17001

[Ebrahimi-Fard et al. 2004] K. Ebrahimi-Fard, L. Guo, and D. Kreimer, "Spitzer's identity and the algebraic Birkhoff decomposition in pQFT", *J. Phys. A* **37**:45 (2004), 11037–11052. MR 2006b:81332 Zbl 1062.81113

[Ebrahimi-Fard et al. 2006] K. Ebrahimi-Fard, L. Guo, and D. Manchon, "Birkhoff type decompositions and the Baker–Campbell–Hausdorff recursion", *Comm. Math. Phys.* **267**:3 (2006), 821–845. MR 2008c:17020 Zbl 1188.17020

[Guo 2000] L. Guo, "Properties of free Baxter algebras", *Adv. Math.* **151**:2 (2000), 346–374. MR 2001f:16048 Zbl 0964.16028

[Guo and Keigher 2000a] L. Guo and W. Keigher, "Baxter algebras and shuffle products", *Adv. Math.* **150**:1 (2000), 117–149. MR 2001g:05015 Zbl 0947.16013

[Guo and Keigher 2000b] L. Guo and W. Keigher, "On free Baxter algebras: completions and the internal construction", *Adv. Math.* **151**:1 (2000), 101–127. MR 2001c:16046 Zbl 0964.16027

[Guo and Zhang 2008] L. Guo and B. Zhang, "Renormalization of multiple zeta values", *J. Algebra* **319**:9 (2008), 3770–3809. MR 2009b:11155 Zbl 1165.11071

[Kosmann-Schwarzbach 1997] Y. Kosmann-Schwarzbach, "Lie bialgebras, Poisson Lie groups and dressing transformations", pp. 104–170 in *Integrability of nonlinear systems* (Pondicherry, 1996), edited by Y. Kosmann-Schwarzbach et al., Lecture Notes in Phys. **495**, Springer, Berlin, 1997. MR 99m:58092 Zbl 1078.37517

[Kosmann-Schwarzbach and Magri 1988] Y. Kosmann-Schwarzbach and F. Magri, "Poisson–Lie groups and complete integrability, I: Drinfeld bialgebras, dual extensions and their canonical representations", *Ann. Inst. H. Poincaré Phys. Théor.* **49**:4 (1988), 433–460. MR 91a:17018 Zbl 0667.16005

[Kreimer 1999] D. Kreimer, "Chen's iterated integral represents the operator product expansion", *Adv. Theor. Math. Phys.* **3**:3 (1999), 627–670. MR 2003b:81128 Zbl 0971.81093 arXiv hep-th/9901099

[Kupershmidt 1999] B. A. Kupershmidt, "What a classical r-matrix really is", *J. Nonlinear Math. Phys.* **6**:4 (1999), 448–488. MR 2001a:17023 Zbl 1015.17015

[Loday and Ronco 2004] J.-L. Loday and M. Ronco, "Trialgebras and families of polytopes", pp. 369–398 in *Homotopy theory: relations with algebraic geometry, group cohomology, and algebraic K-theory*, edited by P. Goerss and S. Priddy, Contemp. Math. **346**, Amer. Math. Soc., Providence, RI, 2004. MR 2006e:18016 Zbl 1065.18007 arXiv math/0205043

[Manchon and Paycha 2010] D. Manchon and S. Paycha, "Nested sums of symbols and renormalized multiple zeta values", *Int. Math. Res. Not.* **2010**:24 (2010), 4628–4697. MR 2012a:81184 Zbl 1206.11108

[Rota 1969a] G.-C. Rota, "Baxter algebras and combinatorial identities, I", *Bull. Amer. Math. Soc.* **75**:2 (1969), 325–329. MR 39 #5387 Zbl 0192.33801

[Rota 1969b] G.-C. Rota, "Baxter algebras and combinatorial identities, II", *Bull. Amer. Math. Soc.* **75**:2 (1969), 330–334. MR 39 #5387 Zbl 0319.05008

[Rota 1995] G.-C. Rota, "Baxter operators: An introduction", pp. 504–512 in *Gian-Carlo Rota on combinatorics*, edited by J. P. S. Kung, Birkhäuser, Boston, 1995. MR 1392973 Zbl 0841.01031

[Runkel et al. 2007] I. Runkel, J. Fjelstad, J. Fuchs, and C. Schweigert, "Topological and conformal field theory as Frobenius algebras", pp. 225–247 in *Categories in algebra, geometry and mathematical physics*, edited by A. Davydov et al., Contemp. Math. **431**, Amer. Math. Soc., Providence, RI, 2007. MR 2009d:81324 Zbl 1154.18006 arXiv math/0512076

[Semenov-Tyan-Shanskiĭ 1983] M. A. Semenov-Tyan-Shanskiĭ, "What is a classical r-matrix?", *Funktsional. Anal. i Prilozhen.* **17**:4 (1983), 17–33. In Russian; translated in *Funct. Anal. Appl.* **17**:4 (1983), 259–272. MR 85i:58061 Zbl 0535.58031

[Uchino 2008] K. Uchino, "Quantum analogy of Poisson geometry, related dendriform algebras and Rota–Baxter operators", *Lett. Math. Phys.* **85**:2-3 (2008), 91–109. MR 2010a:17003 Zbl 05544981

[Vallette 2007] B. Vallette, "Homology of generalized partition posets", *J. Pure Appl. Algebra* **208**:2 (2007), 699–725. MR 2007m:18010 Zbl 1109.18002

[Yamagata 1996] K. Yamagata, "Frobenius algebras", pp. 841–887 in *Handbook of algebra*, vol. 1, edited by M. Hazewinkel, North-Holland, Amsterdam, 1996. MR 97k:16022 Zbl 0879.16008

[Yang 1967] C. N. Yang, "Some exact results for the many-body problem in one dimension with repulsive delta-function interaction", *Phys. Rev. Lett.* **19**:23 (1967), 1312–1315. MR 41 #6480 Zbl 0152.46301

[Zhelyabin 1997] V. N. Zhelyabin, "Jordan bialgebras and their relation to Lie bialgebras", *Algebra i Logika* **36**:1 (1997), 3–25. In Russian; translated in *Algebra and Logic* **36**:1 (1997), 1–15. MR 98d:17036 Zbl 0935.17014

Received April 14, 2011. Revised August 5, 2011.

CHENGMING BAI
CHERN INSTITUTE OF MATHEMATICS AND LPMC
NANKAI UNIVERSITY
TIANJIN 300071
CHINA

baicm@nankai.edu.cn

LI GUO
SCHOOL OF MATHEMATICS AND STATISTICS
LANZHOU UNIVERSITY
LANZHOU, GANSU 730000
CHINA

and

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
RUTGERS UNIVERSITY
216 SMITH HALL
101 WARREN STREET
NEWARK, NJ 07102
UNITED STATES

liguo@rutgers.edu

XIANG NI
DEPARTMENT OF MATHEMATICS
CALTECH
PASADENA, CA 91125
UNITED STATES

xni@caltech.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 256, No. 2, 2012

BOTANY OF IRREDUCIBLE AUTOMORPHISMS OF FREE GROUPS

THIERRY COULBOIS AND ARNAUD HILION

We give a classification of iwip (i.e., fully irreducible) outer automorphisms of the free group, by discussing the properties of their attracting and repelling trees.

1. Introduction

An outer automorphism Φ of the free group F_N is *fully irreducible* (abbreviated as *iwip*) if no positive power Φ^n fixes a proper free factor of F_N. Being an iwip is one (in fact the most important) of the analogs for free groups of being pseudo-Anosov for mapping classes of hyperbolic surfaces. Another analog of pseudo-Anosov is the notion of an atoroidal automorphism: an element $\Phi \in \mathrm{Out}(F_N)$ is *atoroidal* or *hyperbolic* if no positive power Φ^n fixes a nontrivial conjugacy class. Bestvina and Feighn [1992] and Brinkmann [2000] proved that Φ is atoroidal if and only if the mapping torus $F_N \rtimes_\Phi \mathbb{Z}$ is Gromov-hyperbolic.

Pseudo-Anosov mapping classes are known to be "generic" elements of the mapping class group (in various senses). Rivin [2008] and Sisto [2011] recently proved that, in the sense of random walks, generic elements of $\mathrm{Out}(F_N)$ are atoroidal iwip automorphisms.

Bestvina and Handel [1992] proved that iwip automorphisms have the key property of being represented by (absolute) train-track maps.

A pseudo-Anosov element f fixes two projective classes of measured foliations $[(\mathscr{F}^+, \mu^+)]$ and $[(\mathscr{F}^-, \mu^-)]$:

$$(\mathscr{F}^+, \mu^+) \cdot f = (\mathscr{F}^+, \lambda\mu^+) \quad \text{and} \quad (\mathscr{F}^-, \mu^-) \cdot f = (\mathscr{F}^-, \lambda^{-1}\mu^-),$$

where $\lambda > 1$ is the expansion factor of f. Alternatively, considering the dual $\mathbb{R}$-trees T^+ and T^-, we get:

$$T^+ \cdot f = \lambda T^+ \quad \text{and} \quad T^- \cdot f = \lambda^{-1} T^-.$$

We now discuss the analogous situation for iwip automorphisms. The group of outer automorphisms $\mathrm{Out}(F_N)$ acts on the *outer space* CV_N and its boundary

MSC2010: 20E05, 20E08, 20F65, 57R30.
Keywords: free group automorphisms, real trees, laminations, iwip.

$\partial \mathrm{CV}_N$. Recall that the compactified outer space $\overline{\mathrm{CV}}_N = \mathrm{CV}_N \cup \partial \mathrm{CV}_N$ is made up of (projective classes of) $\mathbb{R}$-trees with an action of F_N by isometries which is minimal and very small. See [Vogtmann 2002] for a survey on outer space. An iwip outer automorphism Φ has north-south dynamics on $\overline{\mathrm{CV}}_N$: it has a unique attracting fixed tree $[T_\Phi]$ and a unique repelling fixed tree $[T_{\Phi^{-1}}]$ in the boundary of outer space (see [Levitt and Lustig 2003]):

$$T_\Phi \cdot \Phi = \lambda_\Phi T_\Phi \text{ and } T_{\Phi^{-1}} \cdot \Phi = \frac{1}{\lambda_{\Phi^{-1}}} T_{\Phi^{-1}},$$

where $\lambda_\Phi > 1$ is the *expansion factor* of Φ (i.e., the exponential growth rate of nonperiodic conjugacy classes).

Contrary to the pseudo-Anosov setting, the expansion factor λ_Φ of Φ is typically different from the expansion factor $\lambda_{\Phi^{-1}}$ of Φ^{-1}. More generally, qualitative properties of the fixed trees T_Φ and $T_{\Phi^{-1}}$ can be fairly different. This is the purpose of this paper to discuss and compare the properties of Φ, T_Φ and $T_{\Phi^{-1}}$.

First, the free group, F_N, may be realized as the fundamental group of a surface S with boundary. It is part of folklore that, if Φ comes from a pseudo-Anosov mapping class on S, then its limit trees T_Φ and $T_{\Phi^{-1}}$ live in the Thurston boundary of Teichmüller space: they are dual to a measured foliation on the surface. Such trees T_Φ and $T_{\Phi^{-1}}$ are called *surface trees* and such an iwip outer automorphism Φ is called *geometric* (in this case S has exactly one boundary component).

The notion of surface trees has been generalized (see for instance [Bestvina 2002]). An $\mathbb{R}$-tree which is transverse to measured foliations on a finite CW-complex is called *geometric*. It may fail to be a surface tree if the complex fails to be a surface.

If Φ does not come from a pseudo-Anosov mapping class and if T_Φ is geometric then Φ is called *parageometric*. For a parageometric iwip Φ, Guirardel [2005] and Handel and Mosher [2007] proved that the repelling tree $T_{\Phi^{-1}}$ is not geometric. So we have that, Φ comes from a pseudo-Anosov mapping class on a surface with boundary if and only if both trees T_Φ and $T_{\Phi^{-1}}$ are geometric. Moreover in this case both trees are indeed surface trees.

In [Coulbois and Hilion 2010] we introduced a second dichotomy for trees in the boundary of outer space with dense orbits. For a tree T, we consider its *limit set* $\Omega \subseteq \overline{T}$ (where $\overline{T}$ is the metric completion of T). The limit set Ω consists of points of $\overline{T}$ with at least two pre-images by the map $\mathfrak{Q} : \partial F_N \to \hat{T} = \overline{T} \cup \partial T$ introduced in [Levitt and Lustig 2003]; see Section 4A. We are interested in the two extremal cases: A tree T in the boundary of outer space with dense orbits is of *surface type* if $T \subseteq \Omega$ and T is of *Levitt type* if Ω is totally disconnected. As the terminology suggests, a surface tree is of surface type. Trees of Levitt type where discovered by Levitt [1993].

Combining together the two sets of properties, we introduced in [Coulbois and Hilion 2010] the following definitions. A tree T in $\partial \mathrm{CV}_N$ with dense orbits is

- a *surface tree* if it is both geometric and of surface type;
- *Levitt* if it is geometric and of Levitt type;
- *pseudo-surface* if it is not geometric and of surface type;
- *pseudo-Levitt* if it is not geometric and of Levitt type

The following theorem is the main result of this paper.

Theorem 5.2. *Let Φ be an iwip outer automorphism of F_N. Let T_Φ and $T_{\Phi^{-1}}$ be its attracting and repelling trees. Then exactly one of the following occurs*

(1) *The trees T_Φ and $T_{\Phi^{-1}}$ are surface trees. Equivalently, Φ is geometric.*

(2) *The tree T_Φ is Levitt (i.e., geometric and of Levitt type), and the tree $T_{\Phi^{-1}}$ is pseudo-surface (i.e., nongeometric and of surface type). Equivalently, Φ is parageometric.*

(3) *The tree $T_{\Phi^{-1}}$ is Levitt (i.e., geometric and of Levitt type), and the tree T_Φ is pseudo-surface (i.e., nongeometric and of surface type). Equivalently, Φ^{-1} is parageometric.*

(4) *The trees T_Φ and $T_{\Phi^{-1}}$ are pseudo-Levitt (nongeometric and of Levitt type).*

Case (1) *corresponds to toroidal iwips whereas cases* (2), (3) *and* (4) *corresponds to atoroidal iwips. In case* (4) *the automorphism Φ is called pseudo-Levitt.*

Gaboriau, Jaeger, Levitt and Lustig [Gaboriau et al. 1998] introduced the notion of an *index* $\mathrm{ind}(\Phi)$, computed from the rank of the fixed subgroup and from the number of attracting fixed points of the automorphisms φ in the outer class Φ. Another index for a tree T in $\overline{\mathrm{CV}}_N$ has been defined and studied by Gaboriau and Levitt [1995]; we call it the *geometric index* $\mathrm{ind}_{\mathrm{geo}}(T)$. Finally in [Coulbois and Hilion 2010] we introduced and studied the $\mathbb{Q}$*-index* $\mathrm{ind}_{\mathbb{Q}}(T)$ of an $\mathbb{R}$-tree T in the boundary of outer space with dense orbits. The two indices $\mathrm{ind}_{\mathrm{geo}}(T)$ and $\mathrm{ind}_{\mathbb{Q}}(T)$ describe qualitative properties of the tree T [Coulbois and Hilion 2010]. We define these indices and recall our botanical classification of trees in Section 4A.

The key to prove Theorem 5.2 is this:

Propositions 4.2 and 4.4. *Let Φ be an iwip outer automorphism of F_N. Let T_Φ and $T_{\Phi^{-1}}$ be its attracting and repelling trees. Replacing Φ by a suitable power, we have*

$$2\,\mathrm{ind}(\Phi) = \mathrm{ind}_{\mathrm{geo}}(T_\Phi) = \mathrm{ind}_{\mathbb{Q}}(T_{\Phi^{-1}}).$$

We prove this proposition in Sections 4B and 4C.

To study limit trees of iwip automorphisms, we need to state that they have the strongest mixing dynamical property, which is called *indecomposability*.

Theorem 2.1. *Let $\Phi \in \mathrm{Out}(F_N)$ be an iwip outer automorphism. The attracting tree T_Φ of Φ is indecomposable.*

The proof of this theorem is quite independent of the rest of the paper and is the purpose of Section 2. The proof relies on a key property of iwip automorphisms: they can be represented by (absolute) train-track maps.

2. Indecomposability of the attracting tree of an iwip automorphism

Following [Guirardel 2008], a (projective class of) $\mathbb{R}$-tree $T \in \overline{\mathrm{CV}}_N$ is *indecomposable* if for all nondegenerate arcs I and J in T, there exists finitely many elements $u_1, \dots, u_n$ in F_N such that

$$J \subseteq \bigcup_{i=1}^{n} u_i I \tag{2-1}$$

and

$$\forall i = 1, \dots, n-1, \quad u_i I \cap u_{i+1} I \text{ is a nondegenerate arc.} \tag{2-2}$$

The main purpose of this section is to prove this result:

Theorem 2.1. *Let $\Phi \in \mathrm{Out}(F_N)$ be an iwip outer automorphism. The attracting tree T_Φ of Φ is indecomposable.*

Before proving this theorem in Section 2C, we collect the results we need from [Bestvina and Handel 1992] and [Gaboriau et al. 1998].

2A. *Train-track representative of Φ.* The rose R_N is the graph with one vertex $*$ and N edges. Its fundamental group $\pi_1(R_N, *)$ is naturally identified with the free group F_N. A *marked graph* is a finite graph G with a homotopy equivalence $\tau : R_N \to G$. The marking τ induces an isomorphism

$$\tau_* : F_N = \pi_1(R_N, *) \xrightarrow{\cong} \pi_1(G, v_0),$$

where $v_0 = \tau(*)$.

A homotopy equivalence $f : G \to G$ defines an outer automorphism of F_N. Indeed, if a path m from v_0 to $f(v_0)$ is given, $a \mapsto mf(a)m^{-1}$ induces an automorphism φ of $\pi_1(G, v_0)$, and thus of F_N through the marking. Another path m' from v_0 to $f(v_0)$ gives rise to another automorphism φ' of F_N in the same outer class Φ.

A *topological representative* of $\Phi \in \mathrm{Out}(F_N)$ is an homotopy equivalence $f : G \to G$ of a marked graph G, such that

(i) f maps vertices to vertices,

(ii) f is locally injective on any edge, and

(iii) f induces Φ on $F_N \cong \pi_1(G, v_0)$.

Let $e_1, \ldots, e_p$ be the edges of G (an orientation is arbitrarily given on each edge, and e^{-1} denotes the edge e with the reverse orientation). The *transition matrix* of the map f is the $p \times p$ nonnegative matrix M with (i, j)-entry equal to the number of times the edge e_i occurs in $f(e_j)$ (we say that a path (or an edge) w of a graph G *occurs* in a path u of G if it is w or its inverse w^{-1} is a subpath of u).

A topological representative $f : G \to G$ of Φ is a *train-track map* if, moreover,

(iv) for all $k \in \mathbb{N}$, the restriction of f^k on any edge of G is locally injective, and

(v) any vertex of G has valence at least 3.

According to [Bestvina and Handel 1992, Theorem 1.7], an iwip outer automorphism Φ can be represented by a train-track map, with a primitive transition matrix M (i.e., there exists $k \in \mathbb{N}$ such all the entries of M^k are strictly positive). Thus the Perron–Frobenius theorem applies. In particular, M has a real dominant eigenvalue $\lambda > 1$ associated to a strictly positive eigenvector $u = (u_1, \ldots, u_p)$. Indeed, λ is the expansion factor of Φ: $\lambda = \lambda_\Phi$. We turn the graph G to a metric space by assigning the length u_i to the edge e_i (for $i = 1, \ldots, p$). Since, with respect to this metric, the length of $f(e_i)$ is λ times the length of e_i, we can assume that, on each edge, f is linear of ratio λ.

We define the set $\mathscr{L}_2(f)$ of paths w of combinatorial length 2 (i.e., $w = ee'$, where e, e' are edges of G, $e^{-1} \neq e'$) which occurs in some $f^k(e_i)$ for some $k \in \mathbb{N}$ and some edge e_i of G:

$$\mathscr{L}_2(f) = \{ee' : \exists e_i \text{ edge of } G,\ \exists k \in \mathbb{N} \text{ such that } ee' \text{ is a subpath of } f^k({e_i}^{\pm 1})\}.$$

Since the transition matrix M is primitive, there exists $k \in \mathbb{N}$ such that for any edge e of G, for any $w \in \mathscr{L}_2(f)$, w occurs in $f^k(e)$.

Let v be a vertex of G. The *Whitehead graph* $\mathscr{W}_v$ of v is the unoriented graph defined as follows:

- The vertices of $\mathscr{W}_v$ are the edges of G with v as terminal vertex.
- There is an edge in $\mathscr{W}_v$ between e and e' if $e'e^{-1} \in \mathscr{L}_2(f)$.

As remarked in [Bestvina et al. 1997, Section 2], if $f : G \to G$ is a train-track representative of an iwip outer automorphism Φ, any vertex of G has a connected Whitehead graph. We summarize the previous discussion:

Proposition 2.2. *Let $\Phi \in \mathrm{Out}(F_N)$ be an iwip outer automorphism. There exists a train-track representative $f : G \to G$ of Φ, with primitive transition matrix M and connected Whitehead graphs of vertices. The edge e_i of G is isometric to the segment $[0, u_i]$, where $u = (u_1, \ldots, u_p)$ is a Perron–Frobenius eigenvector of M. The map f is linear of ratio λ on each edge e_i of G.*

Remark 2.3. Let $f : G \to G$ be a train-track map, with primitive transition matrix M and connected Whitehead graphs of vertices. Then for any path $w = ab$ in G of

combinatorial length 2, there exist $w_1 = a_1 b_1, \dots, w_q = a_q b_q \in \mathcal{L}_2(f)$ (a, b, a_i, b_i edges of G) such that

- $a_{i+1} = b_i^{-1}$, $i \in \{1, \dots, q-1\}$, and
- $a = a_1$ and $b = b_q$.

2B. *Construction of T_Φ.* Let $\Phi \in \mathrm{Out}(F_N)$ be an iwip automorphism, and let T_Φ be its attracting tree. Following [Gaboriau et al. 1998], we recall a concrete construction of the tree T_Φ.

We start with a train-track representative $f : G \to G$ of Φ as in Proposition 2.2. The universal cover $\tilde{G}$ of G is a simplicial tree, equipped with a distance d_0 obtained by lifting the distance on G. The fundamental group F_N acts by deck transformations, and thus by isometries, on $\tilde{G}$. Let $\tilde{f}$ be a lift of f to $\tilde{G}$. This lift $\tilde{f}$ is associated to a unique automorphism φ in the outer class Φ, characterized by

$$\forall u \in F_N, \forall x \in \tilde{G}, \quad \varphi(u)\tilde{f}(x) = \tilde{f}(ux). \tag{2-3}$$

For $x, y \in \tilde{G}$ and $k \in \mathbb{N}$, we define:

$$d_k(x, y) = \frac{d_0(\tilde{f}^k(x), \tilde{f}^k(y))}{\lambda^k}.$$

The sequence of distances d_k is decreasing and converges to a pseudo-distance d_∞ on $\tilde{G}$. Identifying points x, y in $\tilde{G}$ which have distance $d_\infty(x, y)$ equal to 0, we obtain the tree T_Φ. The free group F_N still acts by isometries on T_Φ. The quotient map $p : \tilde{G} \to T_\Phi$ is F_N-equivariant and 1-Lipschitz. Moreover, for any edge e of $\tilde{G}$, for any $k \in \mathbb{N}$, the restriction of p to $f^k(e)$ is an isometry. Through p the map $\tilde{f}$ factors to a homothety H of T_Φ, of ratio λ_Φ:

$$\forall x \in \tilde{G}, \quad H(p(x)) = p(\tilde{f}(x)).$$

Property (2-3) leads to

$$\forall u \in F_N, \forall x \in T_\Phi, \quad \varphi(u)H(x) = H(ux). \tag{2-4}$$

2C. *Indecomposability of T_Φ.* We say that a path (or an edge) w of the graph G occurs in a path u of the universal cover $\tilde{G}$ of G if w has a lift $\tilde{w}$ that occurs in u.

Lemma 2.4. *Let I be a nondegenerate arc in T_Φ. There exists an arc I' in $\tilde{G}$ and an integer k such that*

- *$p(I') \subseteq I$, and*
- *any element of $\mathcal{L}_2(f)$ occurs in $H^k(I')$.*

Proof. Let $I \subset T_\Phi$ be a nondegenerate arc. There exists an edge e of $\tilde{G}$ such that $I_0 = p(e) \cap I$ is a nondegenerate arc: $I_0 = [x, y]$. We choose $k_1 \in \mathbb{N}$ such that $d_\infty(H^{k_1}(x), H^{k_1}(y)) > L$ where

$$L = 2 \max\{u_i = |e_i| \mid e_i \text{ edge of } G\}.$$

Let x', y' be the points in e such that $p(x') = x$, $p(y') = y$, and let I' be the arc $[x', y']$. Since p maps $f^{k_1}(e)$ isometrically into T_Φ, we obtain that

$$d_0(f^{k_1}(x'), f^{k_1}(y')) \geq L.$$

Hence there exists an edge e' of $\tilde{G}$ contained in $[f^{k_1}(x'), f^{k_1}(y')]$. Moreover, for any $k_2 \in \mathbb{N}$, the path $f^{k_2}(e')$ isometrically injects in $[H^{k_1+k_2}(x), H^{k_1+k_2}(y)]$. We take k_2 big enough so that any path in $\mathscr{L}_2(f)$ occurs in $f^{k_2}(e')$. Then $k = k_1 + k_2$ is suitable. □

Proof of Theorem 2.1. Let I, J be two nontrivial arcs in T_Φ. We have to prove that I and J satisfy properties (2-1) and (2-2). Since H is a homeomorphism, and because of (2-4), we can replace I and J by $H^k(I)$ and $H^k(J)$, accordingly, for some $k \in \mathbb{N}$.

We consider an arc I' in $\tilde{G}$ and an integer $k \in \mathbb{N}$ as given by Lemma 2.4. Let x, y be the endpoints of the arc $H^k(J)$: $H^k(J) = [x, y]$. Let x', y' be points in $\tilde{G}$ such that $p(x') = x$, $p(y') = y$, and let J' be the arc $[x', y']$. According to Remark 2.3, there exist $w_1, \ldots, w_n$ such that

- w_i is a lift of some path in $\mathscr{L}_2(f)$,
- $J' \subseteq \bigcup_{i=1}^n w_i$, and
- $w_i \cap w_{i+1}$ is an edge.

Since Lemma 2.4 ensures that any element of $\mathscr{L}_2(f)$ occurs in $H^k(I')$, we deduce that $H^k(I)$ and $H^k(J)$ satisfy properties (2-1) and (2-2). □

3. Index of an outer automorphism

An automorphism φ of the free group F_N extends to a homeomorphism $\partial\varphi$ of the boundary at infinity ∂F_N. We denote by $\mathrm{Fix}(\varphi)$ the fixed subgroup of φ. It is a finitely generated subgroup of F_N and thus its boundary $\partial\mathrm{Fix}(\varphi)$ naturally embeds in ∂F_N. Elements of $\partial\mathrm{Fix}(\varphi)$ are fixed by $\partial\varphi$ and they are called *singular*. Non-singular fixed points of $\partial\varphi$ are called *regular*. A fixed point X of $\partial\varphi$ is *attracting* (resp. *repelling*) if it is regular and if there exists an element u in F_N such that $\varphi^n(u)$ (resp. $\varphi^{-n}(u)$) converges to X. The set of fixed points of $\partial\varphi$ is denoted by $\mathrm{Fix}(\partial\varphi)$.

Following Nielsen, fixed points of $\partial\varphi$ have been classified by Gaboriau, Jaeger, Levitt and, Lustig:

Proposition 3.1 [Gaboriau et al. 1998, Proposition 1.1]. *Let φ be an automorphism of the free group F_N, and X a fixed point of $\partial\varphi$. Exactly one of the following occurs*:

(1) *X is in the boundary of the fixed subgroup of φ.*

(2) *X is attracting.*

(3) *X is repelling.* □

We denote by $\mathrm{Att}(\varphi)$ the set of attracting fixed points of $\partial\varphi$. The fixed subgroup $\mathrm{Fix}(\varphi)$ acts on the set $\mathrm{Att}(\varphi)$ of attracting fixed points.

In [Gaboriau et al. 1998] the following *index* of the automorphism φ is defined:

$$\mathrm{ind}(\varphi) = \tfrac{1}{2}\#(\mathrm{Att}(\varphi)/\mathrm{Fix}(\varphi)) + \mathrm{rank}(\mathrm{Fix}(\varphi)) - 1$$

If φ has a trivial fixed subgroup, the above definition is simpler:

$$\mathrm{ind}(\varphi) = \tfrac{1}{2}\#\mathrm{Att}(\varphi) - 1.$$

Let u be an element of F_N and let i_u be the corresponding inner automorphism of F_N:

$$\forall w \in F_N, \quad i_u(w) = uwu^{-1}.$$

The inner automorphism i_u extends to the boundary of F_N as left multiplication by u:

$$\forall X \in \partial F_N, \quad \partial i_u(X) = uX.$$

The group $\mathrm{Inn}(F_N)$ of inner automorphisms of F_N acts by conjugacy on the automorphisms in an outer class Φ. Following Nielsen, two automorphisms, $\varphi, \varphi' \in \Phi$ are *isogredient* if they are conjugated by some inner automorphism i_u:

$$\varphi' = i_u \circ \varphi \circ i_{u^{-1}} = i_{u\varphi(u)^{-1}} \circ \varphi.$$

In this case, the actions of $\partial\varphi$ and $\partial\varphi'$ on ∂F_N are conjugate by the left multiplication by u. In particular, a fixed point X' of $\partial\varphi'$ is a translate $X' = uX$ of a fixed point X of $\partial\varphi$. Two isogredient automorphisms have the same index: this is the index of the isogrediency class. An isogrediency class $[\varphi]$ is *essential* if it has positive index: $\mathrm{ind}([\varphi]) > 0$. We note that essential isogrediency classes are principal in the sense of [Feighn and Handel 2011], but the converse is not true.

The *index* of the outer automorphism Φ is the sum, over all essential isogrediency classes of automorphisms φ in the outer class Φ, of their indices, or alternatively:

$$\mathrm{ind}(\Phi) = \sum_{[\varphi]\in\Phi/\mathrm{Inn}(F_N)} \max(0; \mathrm{ind}(\varphi)).$$

We adapt the notion of *forward rotationless outer automorphism* of [Feighn and Handel 2011] to our purpose. We denote by $\mathrm{Per}(\varphi)$ the set of elements of F_N fixed

by some positive power of φ:

$$\mathrm{Per}(\varphi) = \bigcup_{n\in\mathbb{N}^*} \mathrm{Fix}(\varphi^n);$$

and by $\mathrm{Per}(\partial\varphi)$ the set of elements of ∂F_N fixed by some positive power of $\partial\varphi$:

$$\mathrm{Per}(\partial\varphi) = \bigcup_{n\in\mathbb{N}^*} \mathrm{Fix}(\partial\varphi^n).$$

Definition 3.2. An outer automorphism $\Phi \in \mathrm{Out}(F_N)$ is FR if:

(FR1) for any automorphism $\varphi \in \Phi$, $\mathrm{Per}(\varphi) = \mathrm{Fix}(\varphi)$ and $\mathrm{Per}(\partial\varphi) = \mathrm{Fix}(\partial\varphi)$, and

(FR2) if ψ is an automorphism in the outer class Φ^n for some $n > 0$, with $\mathrm{ind}(\psi)$ positive, then there exists an automorphism φ in Φ such that $\psi = \varphi^n$.

Proposition 3.3. *Let $\Phi \in \mathrm{Out}(F_N)$. There exists $k \in \mathbb{N}^*$ such that Φ^k is FR.*

Proof. By [Levitt and Lustig 2000, Theorem 1] there exists a power Φ^k with (FR1). An automorphism $\varphi \in \mathrm{Aut}(F_N)$ with positive index $\mathrm{ind}(\varphi) > 0$ is principal in the sense of [Feighn and Handel 2011, Definition 3.1]. Thus our property (FR2) is a consequence of the forward rotationless property of [loc. cit., Definition 3.13]. By [loc. cit., Lemma 4.43] there exists a power $\Phi^{k\ell}$ which is forward rotationless and thus which satisfies (FR2). □

4. Indices

4A. ***Botany of trees.*** We recall in this section the classification of trees in the boundary of outer space, given in [Coulbois and Hilion 2010].

Gaboriau and Levitt [1995] introduced an index for a tree T in $\overline{\mathrm{CV}}_N$, we call it the *geometric index* and denote it by $\mathrm{ind}_{\mathrm{geo}}(T)$. It is defined using the valence of the branch points, of the $\mathbb{R}$-tree T, with an action of the free group by isometries:

$$\mathrm{ind}_{\mathrm{geo}}(T) = \sum_{[P]\in T/F_N} \mathrm{ind}_{\mathrm{geo}}(P).$$

where the local index of a point P in T is

$$\mathrm{ind}_{\mathrm{geo}}(P) = \#(\pi_0(T \smallsetminus \{P\})/\mathrm{Stab}(P)) + 2\,\mathrm{rank}(\mathrm{Stab}(P)) - 2.$$

Gaboriau and Levitt proved that the geometric index of a geometric tree is equal to $2N-2$ and that for any tree in the compactification of outer space $\overline{\mathrm{CV}}_N$ the geometric index is bounded above by $2N-2$. Moreover, they proved that the trees in $\overline{\mathrm{CV}}_N$ with geometric index equal to $2N-2$ are precisely the geometric trees.

If, moreover, T has dense orbits, Levitt and Lustig [2003; 2008] defined the map $\mathfrak{D} : \partial F_N \to \hat{T}$, characterized as follows:

Proposition 4.1. *Let T be an $\mathbb{R}$-tree in $\overline{\mathrm{CV}}_N$ with dense orbits. There exists a unique map $\mathcal{Q} : \partial F_N \to \hat{T}$ such that for any sequence $(u_n)_{n\in\mathbb{N}}$ of elements of F_N which converges to $X \in \partial F_N$, and any point $P \in T$, if the sequence of points $(u_n P)_{n\in\mathbb{N}}$ converges to a point $Q \in \hat{T}$, then $\mathcal{Q}(X) = Q$. Moreover, $\mathcal{Q}$ is onto.*

Let us consider the case of a tree T dual to a measured foliation $(\mathcal{F}, \mu)$ on a hyperbolic surface S with boundary (T is a surface tree). Let $\tilde{\mathcal{F}}$ be the lift of $\mathcal{F}$ to the universal cover $\tilde{S}$ of S. The boundary at infinity of $\tilde{S}$ is homeomorphic to ∂F_N. On the one hand, a leaf ℓ of $\tilde{\mathcal{F}}$ defines a point in T. On the other hand, the ends of ℓ define points in ∂F_N. The map $\mathcal{Q}$ precisely sends the ends of ℓ to the point in T. The Poincaré–Lefschetz index of the foliation $\mathcal{F}$ can be computed from the cardinal of the fibers of the map $\mathcal{Q}$. This leads to the following definition of the $\mathcal{Q}$-index of an $\mathbb{R}$-tree T in a more general context.

Let T be an $\mathbb{R}$-tree in $\overline{\mathrm{CV}}_N$ with dense orbits. The $\mathcal{Q}$-index of the tree T is defined by

$$\mathrm{ind}_{\mathcal{Q}}(T) = \sum_{[P]\in\hat{T}/F_N} \max(0; \mathrm{ind}_{\mathcal{Q}}(P)),$$

where the local index of a point P in T is

$$\mathrm{ind}_{\mathcal{Q}}(P) = \#(\mathcal{Q}_r^{-1}(P)/\mathrm{Stab}(P)) + 2\,\mathrm{rank}(\mathrm{Stab}(P)) - 2$$

with $\mathcal{Q}_r^{-1}(P) = \mathcal{Q}^{-1}(P) \smallsetminus \partial\mathrm{Stab}(P)$ the regular fiber of P.

Levitt and Lustig [2003] proved that points in ∂T have exactly one pre-image by $\mathcal{Q}$. Thus, only points in $\overline{T}$ contribute to the $\mathcal{Q}$-index of T.

We proved in [Coulbois and Hilion 2010] that the $\mathcal{Q}$-index of an $\mathbb{R}$-tree in the boundary of outer space with dense orbits is bounded above by $2N-2$. And it is equal to $2N-2$ if and only if it is of surface type.

The botanical classification in [Coulbois and Hilion 2010] of a tree T with a minimal very small indecomposable action of F_N by isometries is as follows:

	geometric $\mathrm{ind}_{\mathrm{geo}}(T) = 2N-2$	not geometric $\mathrm{ind}_{\mathrm{geo}}(T) < 2N-2$
Surface type: $\mathrm{ind}_{\mathcal{Q}}(T) = 2N-2$	surface	pseudo-surface
Levitt type: $\mathrm{ind}_{\mathcal{Q}}(T) < 2N-2$	Levitt	pseudo-Levitt

The following remark is not necessary for the sequel of the paper, but may help the reader's intuition.

Remark. In [Coulbois et al. 2008a; 2008b], in collaboration with Lustig, we defined and studied the dual lamination of an $\mathbb{R}$-tree T with dense orbits:

$$L(T) = \{(X, Y) \in \partial^2 F_N \mid \mathcal{Q}(X) = \mathcal{Q}(Y)\}.$$

The $\mathfrak{Q}$-index of T can be interpreted as the index of this dual lamination.

Using the dual lamination, with Lustig [Coulbois et al. 2009], we defined the compact heart $K_A \subseteq \overline{T}$ (for a basis A of F_N). We proved that the tree T is completely encoded by a system of partial isometries $S_A = (K_A, A)$. We also proved that the tree T is geometric if and only if the compact heart K_A is a finite tree (that is to say the convex hull of finitely many points). In [Coulbois and Hilion 2010] we used the Rips machine on the system of isometries S_A to get the bound on the $\mathfrak{Q}$-index of T. In particular, an indecomposable tree T is of Levitt type if and only if the Rips machine never halts.

4B. ***Geometric index.*** As in Section 2B, an iwip outer automorphism Φ has an expansion factor $\lambda_\Phi > 1$, an attracting $\mathbb{R}$-tree T_Φ in $\partial \mathrm{CV}_N$. For each automorphism φ in the outer class Φ there is a homothety H of the metric completion $\bar{T}_\Phi$, of ratio λ_Φ, such that

$$\forall P \in \overline{T}_\Phi,\ \forall u \in F_N, \quad H(uP) = \varphi(u)H(P). \tag{4-1}$$

In addition, the action of Φ on the compactification of Culler and Vogtmann's outer space has north-south dynamics and the projective class of T_Φ is the attracting fixed point [Levitt and Lustig 2003]. Of course the attracting trees of Φ and Φ^n ($n > 0$) are equal.

For the attracting tree T_Φ of the iwip outer automorphism Φ, the geometric index is well understood.

Proposition 4.2 [Gaboriau et al. 1998, Section 4]. *Let Ψ be an iwip outer automorphism. There exists a power $\Phi = \Psi^k$ ($k > 0$) of Ψ such that*

$$2\,\mathrm{ind}(\Phi) = \mathrm{ind}_{\mathrm{geo}}(T_\Phi),$$

where T_Φ is the attracting tree of Φ (and of Ψ). □

4C. $\mathfrak{Q}$***-index.*** Let Φ be an iwip outer automorphism of F_N. Let T_Φ be its attracting tree. The action of F_N on T_Φ has dense orbits.

Let φ an automorphism in the outer class Φ. The homothety H associated to φ extends continuously to an homeomorphism of the boundary at infinity of T_Φ which we still denote by H. We get from Proposition 4.1 and identity (4-1):

$$\forall X \in \partial F_N, \quad \mathfrak{Q}(\partial\varphi(X)) = H(\mathfrak{Q}(X)). \tag{4-2}$$

We are going to prove that the $\mathfrak{Q}$-index of T_Φ is twice the index of Φ^{-1}. As mentioned in the introduction for geometric automorphisms both these numbers are equal to $2N-2$ and thus we restrict to the study of nongeometric automorphisms. For the rest of this section we assume that Φ is nongeometric. This will be used in two ways:

- The action of F_N on T_Φ is free.
- For any φ in the outer class Φ, all the fixed points of φ in ∂F_N are regular.

Let C_H be the center of the homothety H. The following Lemma is essentially contained in [Gaboriau et al. 1998], although the map $\mathfrak{Q}$ is not used there.

Lemma 4.3. *Let $\Phi \in \mathrm{Out}(F_N)$ be a FR nongeometric iwip outer automorphism. Let T_Φ be the attracting tree of Φ. Let $\varphi \in \Phi$ be an automorphism in the outer class Φ, and let H be the homothety of T_Φ associated to φ, with C_H its center. The $\mathfrak{Q}$-fiber of C_H is the set of repelling points of φ.*

Proof. Let $X \in \partial F_N$ be a repelling point of $\partial\varphi$. By definition there exists an element $u \in F_N$ such that the sequence $(\varphi^{-n}(u))_n$ converges towards X. By (4-1),

$$\varphi^{-n}(u)C_H = \varphi^{-n}(u)H^{-n}(C_H) = H^{-n}(uC_H).$$

The homothety H^{-1} is strictly contracting and therefore the sequence of points $(\varphi^{-n}(u)C_H)_n$ converges towards C_H. By Proposition 4.1 we get that $\mathfrak{Q}(X) = C_H$.

Conversely let $X \in \mathfrak{Q}^{-1}(C_H)$ be a point in the $\mathfrak{Q}$-fiber of C_H. Using the identity (4-2), $\partial\varphi(X)$ is also in the $\mathfrak{Q}$-fiber. The $\mathfrak{Q}$-fiber is finite by [Coulbois and Hilion 2010, Corollary 5.4], X is a periodic point of $\partial\varphi$. Since Φ satisfies property (FR1), X is a fixed point of $\partial\varphi$. From [Gaboriau et al. 1998, Lemma 3.5], attracting fixed points of $\partial\varphi$ are mapped by $\mathfrak{Q}$ to points in the boundary at infinity ∂T_Φ. Thus X has to be a repelling fixed point of $\partial\varphi$. □

Proposition 4.4. *Let $\Phi \in \mathrm{Out}(F_N)$ be a FR nongeometric iwip outer automorphism. Let T_Φ be the attracting tree of Φ. Then*

$$2\,\mathrm{ind}(\Phi^{-1}) = \mathrm{ind}_{\mathfrak{Q}}(T_\Phi).$$

Proof. To each automorphism φ in the outer class Φ is associated a homothety H of T_Φ and the center C_H of this homothety. As the action of F_N on T_Φ is free, two automorphisms are isogredient if and only if the corresponding centers are in the same F_N-orbit.

The index of Φ^{-1} is the sum over all essential isogrediency classes of automorphism φ^{-1} in Φ^{-1} of the index of φ^{-1}. For each of these automorphisms the index $2\,\mathrm{ind}(\varphi^{-1})$ is equal by Lemma 4.3 to the contribution $\#\mathfrak{Q}^{-1}(C_H)$ of the orbit of C_H to the $\mathfrak{Q}$ index of T_Φ.

Conversely, let now P be a point in $\overline{T}_\Phi$ with at least three elements in its $\mathfrak{Q}$-fiber. Let φ be an automorphism in Φ and let H be the homothety of T_Φ associated to φ. For any integer n, the $\mathfrak{Q}$-fiber $\mathfrak{Q}^{-1}(H^n(P)) = \partial\varphi^n(\mathfrak{Q}^{-1}(P))$ of $H^n(P)$ also has at least three elements. By [Coulbois and Hilion 2010, Theorem 5.3] there are finitely many orbits of such points in T_Φ and thus we can assume that $H^n(P) = wP$ for some $w \in F_N$ and some integer $n > 0$. Then P is the center of the homothety

$w^{-1}H^n$ associated to $i_{w^{-1}} \circ \varphi^n$. Since Φ satisfies property (FR2), P is the center of a homothety uH associated to $i_u \circ \varphi$ for some $u \in F_N$. This concludes the proof of the equality of the indices. □

This proposition can alternatively be deduced from the techniques of [Handel and Mosher 2011].

5. Botanical classification of irreducible automorphisms

Theorem 5.1. *Let Φ be an iwip outer automorphism of F_N. Let T_Φ and $T_{\Phi^{-1}}$ be its attracting and repelling trees. Then, the $\mathcal{Q}$-index of the attracting tree is equal to the geometric index of the repelling tree*:

$$\mathrm{ind}_{\mathcal{Q}}(T_\Phi) = \mathrm{ind}_{\mathrm{geo}}(T_{\Phi^{-1}}).$$

Proof. First, if Φ is geometric, then the trees T_Φ and $T_{\Phi^{-1}}$ have maximal geometric indices $2N-2$. On the other hand the trees T_Φ and $T_{\Phi^{-1}}$ are surface trees and thus their $\mathcal{Q}$-indices are also maximal:

$$\mathrm{ind}_{\mathrm{geo}}(T_\Phi) = \mathrm{ind}_{\mathcal{Q}}(T_\Phi) = \mathrm{ind}_{\mathrm{geo}}(T_{\Phi^{-1}}) = \mathrm{ind}_{\mathcal{Q}}(T_{\Phi^{-1}}) = 2N-2.$$

We now assume that Φ is not geometric and we can apply Propositions 4.2 and 4.4 to get the desired equality. □

From Theorem 5.1 and from the characterization of geometric and surface-type trees by the maximality of the indices we get

Theorem 5.2. *Let Φ be an iwip outer automorphism of F_N. Let T_Φ and $T_{\Phi^{-1}}$ be its attracting and repelling trees. Then exactly one of the following occurs*:

(1) *T_Φ and $T_{\Phi^{-1}}$ are surface trees.*

(2) *T_Φ is Levitt and $T_{\Phi^{-1}}$ is pseudo-surface.*

(3) *$T_{\Phi^{-1}}$ is Levitt and T_Φ is pseudo-surface.*

(4) *T_Φ and $T_{\Phi^{-1}}$ are pseudo-Levitt.*

Proof. The trees T_Φ and $T_{\Phi^{-1}}$ are indecomposable by Theorem 2.1 and thus they are either of surface type or of Levitt type by [Coulbois and Hilion 2010, Proposition 5.14]. Recall, from [Gaboriau and Levitt 1995] (see also [Coulbois and Hilion 2010, Theorem 5.9] or [Coulbois et al. 2009, Corollary 6.1]) that T_Φ is geometric if and only if its geometric index is maximal:

$$\mathrm{ind}_{\mathrm{geo}}(T_\Phi) = 2N-2.$$

From [Coulbois and Hilion 2010, Theorem 5.10], T_Φ is of surface type if and only if its $\mathcal{Q}$-index is maximal:

$$\mathrm{ind}_{\mathcal{Q}}(T_\Phi) = 2N-2.$$

The theorem now follows from Theorem 5.1. □

Let $\Phi \in \mathrm{Out}(F_N)$ be an iwip outer automorphism.

The outer automorphism Φ is *geometric* if both its attracting and repelling trees T_Φ and $T_{\Phi^{-1}}$ are geometric. This is equivalent to saying that Φ is induced by a pseudo-Anosov homeomorphism of a surface with boundary; see [Guirardel 2005] and [Handel and Mosher 2007]. This is case (1) of Theorem 5.2.

The outer automorphism Φ is *parageometric* if its attracting tree T_Φ is geometric but its repelling tree $T_{\Phi^{-1}}$ is not. This is case (2) of Theorem 5.2.

The outer automorphism Φ is *pseudo-Levitt* if both its attracting and repelling trees are not geometric. This is case (4) of Theorem 5.2

We now bring expansion factors into play. An iwip outer automorphism Φ of F_N has an expansion factor $\lambda_\Phi > 1$: it is the exponential growth rate of (nonfixed) conjugacy classes under iteration of Φ.

If Φ is geometric, the expansion factor of Φ is equal to the expansion factor of the associated pseudo-Anosov mapping class and thus $\lambda_\Phi = \lambda_{\Phi^{-1}}$.

Handel and Mosher [2007] proved that if Φ is a parageometric outer automorphism of F_N then $\lambda_\Phi > \lambda_{\Phi^{-1}}$ (see also [Behrstock et al. 2010]). Examples are also given by Gautero [2007].

For pseudo-Levitt outer automorphisms of F_N nothing can be said on the comparison of the expansion factors of the automorphism and its inverse. On one hand, Handel and Mosher [2007, Introduction] gave an explicit example of a nongeometric automorphism with $\lambda_\Phi = \lambda_{\Phi^{-1}}$: thus this automorphism is pseudo-Levitt. On the other hand, there are examples of pseudo-Levitt automorphisms with $\lambda_\Phi > \lambda_{\Phi^{-1}}$. Let $\varphi \in \mathrm{Aut}(F_3)$ be the automorphism such that

$$\begin{array}{lll} \varphi : a \mapsto b & \text{and} & \varphi^{-1} : a \mapsto c \\ \quad b \mapsto ac & & \quad b \mapsto a \\ \quad c \mapsto a & & \quad c \mapsto c^{-1}b \end{array}$$

Let Φ be its outer class. Then Φ^6 is FR, has index $\mathrm{ind}(\Phi^6) = \frac{3}{2} < 2$. The expansion factor is $\lambda_\Phi \simeq 1{,}3247$. The outer automorphism Φ^{-3} is FR, has index $\mathrm{ind}(\Phi^{-3}) = \frac{1}{2} < 2$. The expansion factor is $\lambda_{\Phi^{-1}} \simeq 1{,}4655 > \lambda_\Phi$. The computation of these two indices can be achieved using the algorithm of [Jullian 2009].

Now that we have classified outer automorphisms of F_N into four categories, questions of genericity naturally arise. In particular, is a generic outer automorphism of F_N iwip, pseudo-Levitt and with distinct expansion factors? This was suggested in [Handel and Mosher 2007], in particular for statistical genericity: given a set of generators of $\mathrm{Out}(F_N)$ and considering the word metric associated

to it, is it the case that

$$\lim_{k\to\infty} \frac{\#(\text{pseudo-Levitt iwip with } \lambda_\Phi \neq \lambda_{\Phi^{-1}}) \cap B(k))}{\#B(k)} = 1,$$

where $B(k)$ is the ball of radius k, centered at 1, in $\mathrm{Out}(F_N)$?

5A. ***Botanical memo.*** In this section we give a glossary of our classification of automorphisms for the working mathematician.

For a FR iwip outer automorphism Φ of F_N, we used 6 indices which are related in the following way:

$$\begin{aligned} 2\,\mathrm{ind}(\Phi) &= \mathrm{ind}_{\mathrm{geo}}(T_\Phi) \quad = \mathrm{ind}_{\mathbb{Q}}(T_{\Phi^{-1}}), \\ 2\,\mathrm{ind}(\Phi^{-1}) &= \mathrm{ind}_{\mathrm{geo}}(T_{\Phi^{-1}}) = \mathrm{ind}_{\mathbb{Q}}(T_\Phi). \end{aligned}$$

All these indices are bounded above by $2N-2$. We sum up our Theorem 5.2 in the following table.

Automorphisms		Trees		Indices
Φ geometric $\Updownarrow$ Φ^{-1} geometric	$\Leftrightarrow$	T_Φ and $T_{\Phi^{-1}}$ geometric $\Updownarrow$ T_Φ surface $\Updownarrow$ $T_{\Phi^{-1}}$ surface	$\Leftrightarrow$	$\mathrm{ind}(\Phi) = \mathrm{ind}(\Phi^{-1}) = N-1$
Φ parageometric	$\Leftrightarrow$	$\begin{cases} T_\Phi \text{ geometric} \\ \text{and} \\ T_{\Phi^{-1}} \text{ nongeometric} \end{cases}$ $\Updownarrow$ T_Φ Levitt $\Updownarrow$ $T_{\Phi^{-1}}$ pseudo-surface	$\Leftrightarrow$	$\begin{cases} \mathrm{ind}(\Phi) = N-1 \\ \text{and} \\ \mathrm{ind}(\Phi^{-1}) < N-1 \end{cases}$
Φ pseudo-Levitt $\Updownarrow$ Φ^{-1} pseudo-Levitt	$\Leftrightarrow$	$T_\Phi, T_{\Phi^{-1}}$ nongeometric $\Updownarrow$ T_Φ pseudo-Levitt $\Updownarrow$ $T_{\Phi^{-1}}$ pseudo-Levitt	$\Leftrightarrow$	$\begin{cases} \mathrm{ind}(\Phi) < N-1 \\ \text{and} \\ \mathrm{ind}(\Phi^{-1}) < N-1 \end{cases}$

Acknowledgements

We thank Martin Lustig for his constant interest in our work and the referee for the suggested improvements.

References

[Behrstock et al. 2010] J. Behrstock, M. Bestvina, and M. Clay, "Growth of intersection numbers for free group automorphisms", *J. Topol.* **3**:2 (2010), 280–310. MR 2011j:20072 Zbl 1209.20031

[Bestvina 2002] M. Bestvina, "$\mathbb{R}$-trees in topology, geometry, and group theory", pp. 55–91 in *Handbook of geometric topology*, edited by R. J. Daverman and R. B. Sher, North-Holland, Amsterdam, 2002. MR 2003b:20040 Zbl 0998.57003

[Bestvina and Feighn 1992] M. Bestvina and M. Feighn, "A combination theorem for negatively curved groups", *J. Differential Geom.* **35**:1 (1992), 85–101. MR 93d:53053 Zbl 0724.57029

[Bestvina and Handel 1992] M. Bestvina and M. Handel, "Train tracks and automorphisms of free groups", *Ann. of Math.* (2) **135**:1 (1992), 1–51. MR 92m:20017 Zbl 0757.57004

[Bestvina et al. 1997] M. Bestvina, M. Feighn, and M. Handel, "Laminations, trees, and irreducible automorphisms of free groups", *Geom. Funct. Anal.* **7**:2 (1997), 215–244. MR 98c:20045 Zbl 0884.57002

[Brinkmann 2000] P. Brinkmann, "Hyperbolic automorphisms of free groups", *Geom. Funct. Anal.* **10**:5 (2000), 1071–1089. MR 2001m:20061 Zbl 0970.20018

[Coulbois and Hilion 2010] T. Coulbois and A. Hilion, "Rips induction: index of the dual lamination of an $\mathbb{R}$-tree", preprint, 2010. arXiv 1002.0972

[Coulbois et al. 2008a] T. Coulbois, A. Hilion, and M. Lustig, "$\mathbb{R}$-trees and laminations for free groups. I. Algebraic laminations", *J. Lond. Math. Soc.* (2) **78**:3 (2008), 723–736. MR 2010e:20038

[Coulbois et al. 2008b] T. Coulbois, A. Hilion, and M. Lustig, "$\mathbb{R}$-trees and laminations for free groups. II. The dual lamination of an $\mathbb{R}$-tree", *J. Lond. Math. Soc.* (2) **78**:3 (2008), 737–754. MR 2010h:20056 Zbl 1198.20023

[Coulbois et al. 2009] T. Coulbois, A. Hilion, and M. Lustig, "$\mathbb{R}$-trees, dual laminations and compact systems of partial isometries", *Math. Proc. Cambridge Philos. Soc.* **147**:2 (2009), 345–368. MR 2010m:20031 Zbl 05617513

[Feighn and Handel 2011] M. Feighn and M. Handel, "The recognition theorem for $\mathrm{Out}(F_n)$", *Groups Geom. Dyn.* **5**:1 (2011), 39–106. MR 2012b:20061 Zbl 05973328

[Gaboriau and Levitt 1995] D. Gaboriau and G. Levitt, "The rank of actions on **R**-trees", *Ann. Sci. École Norm. Sup.* (4) **28**:5 (1995), 549–570. MR 97c:20039 Zbl 0835.20038

[Gaboriau et al. 1998] D. Gaboriau, A. Jaeger, G. Levitt, and M. Lustig, "An index for counting fixed points of automorphisms of free groups", *Duke Math. J.* **93**:3 (1998), 425–452. MR 99f:20051 Zbl 0946.20010

[Gautero 2007] F. Gautero, "Combinatorial mapping-torus, branched surfaces and free group automorphisms", *Ann. Sc. Norm. Super. Pisa Cl. Sci.* (5) **6**:3 (2007), 405–440. MR 2008m:20043 Zbl 1173.20017

[Guirardel 2005] V. Guirardel, "Cœur et nombre d'intersection pour les actions de groupes sur les arbres", *Ann. Sci. École Norm. Sup.* (4) **38**:6 (2005), 847–888. MR 2007e:20055

[Guirardel 2008] V. Guirardel, "Actions of finitely generated groups on $\mathbb{R}$-trees", *Ann. Inst. Fourier* (*Grenoble*) **58**:1 (2008), 159–211. MR 2009g:20044 Zbl 1187.20020

[Handel and Mosher 2007] M. Handel and L. Mosher, "Parageometric outer automorphisms of free groups", *Trans. Amer. Math. Soc.* **359**:7 (2007), 3153–3183. MR 2008c:20045 Zbl 1120.20042

[Handel and Mosher 2011] M. Handel and L. Mosher, *Axes in outer space*, Mem. Amer. Math. Soc. **1004**, Amer. Math. Soc., Providence, RI, 2011. MR 2858636 Zbl 05947410 arXiv math/0605355

[Jullian 2009] Y. Jullian, *Représentations géométriques des systèmes dynamiques substitutifs par substitutions d'arbre*, PhD thesis, Université Aix-Marseille II, 2009.

[Levitt 1993] G. Levitt, "La dynamique des pseudogroupes de rotations", *Invent. Math.* **113**:3 (1993), 633–670. MR 94k:58158 Zbl 0791.58055

[Levitt and Lustig 2000] G. Levitt and M. Lustig, "Periodic ends, growth rates, Hölder dynamics for automorphisms of free groups", *Comment. Math. Helv.* **75**:3 (2000), 415–429. MR 2002g:20047 Zbl 0965.20026

[Levitt and Lustig 2003] G. Levitt and M. Lustig, "Irreducible automorphisms of F_n have north-south dynamics on compactified outer space", *J. Inst. Math. Jussieu* **2**:1 (2003), 59–72. MR 2004a:20046 Zbl 1034.20038

[Levitt and Lustig 2008] G. Levitt and M. Lustig, "Automorphisms of free groups have asymptotically periodic dynamics", *J. Reine Angew. Math.* **619** (2008), 1–36. MR 2009m:20034 Zbl 1157.20017

[Rivin 2008] I. Rivin, "Walks on groups, counting reducible matrices, polynomials, and surface and free group automorphisms", *Duke Math. J.* **142**:2 (2008), 353–379. MR 2009m:20077 Zbl 1207.20068

[Sisto 2011] A. Sisto, "Contracting elements and random walks", preprint, 2011. arXiv 1112.2666

[Vogtmann 2002] K. Vogtmann, "Automorphisms of free groups and outer space", *Geom. Dedicata* **94** (2002), 1–31. MR 2004b:20060 Zbl 1017.20035

Received January 16, 2012. Revised March 20, 2012.

THIERRY COULBOIS
LATP
UNIVERSITÉ D'AIX–MARSEILLE
AVENUE DE L'ESCADRILLE NORMANDIE–NIÉMEN
13013 MARSEILLE
FRANCE
thierry.coulbois@univ-amu.fr

ARNAUD HILION
LATP
UNIVERSITÉ D'AIX–MARSEILLE
AVENUE DE L'ESCADRILLE NORMANDIE–NIÉMEN
13013 MARSEILLE
FRANCE
arnaud.hilion@univ-amu.fr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 256, No. 2, 2012

A NOTE ON INVERSE CURVATURE FLOWS IN ASYMPTOTICALLY ROBERTSON–WALKER SPACETIMES

CLAUS GERHARDT

We prove that the leaves of the rescaled curvature flow considered in earlier work converge to the graph of a constant function.

1. Introduction

In [Gerhardt 2004] and [Gerhardt 2006a, Chapter 7] we considered the inverse mean curvature flow in a Lorentzian manifold $N = N^{n+1}$ which we called an asymptotically Robertson–Walker space, and which is defined by the following conditions:

Definition 1.1. A cosmological spacetime N, $\dim N = n+1$, is said to be *asymptotically Robertson–Walker* (ARW) with respect to the future, if a future end of N, N_+, can be written as a product $N_+ = [a, b) \times \mathcal{S}_0$, where $\mathcal{S}_0$ is a compact Riemannian space, and there exists a future directed time function $\tau = x^0$ such that the metric in N_+ can be written as

$$d\check{s}^2 = e^{2\widetilde{\psi}}\left\{-\left(dx^0\right)^2 + \sigma_{ij}(x^0, x)\, dx^i dx^j\right\}, \tag{1-1}$$

where $\mathcal{S}_0$ corresponds to $x^0 = a$, $\widetilde{\psi}$ is of the form

$$\widetilde{\psi}(x^0, x) = f(x^0) + \psi(x^0, x), \tag{1-2}$$

and we assume that there exists a positive constant c_0 and a smooth Riemannian metric $\overline{\sigma}_{ij}$ on $\mathcal{S}_0$ such that

$$\lim_{\tau\to b} e^{\psi} = c_0 \quad \text{and} \quad \lim_{\tau\to b} \sigma_{ij}(\tau, x) = \overline{\sigma}_{ij}(x), \tag{1-3}$$

and

$$\lim_{\tau\to b} f(\tau) = -\infty. \tag{1-4}$$

This work was supported by the DFG.

MSC2010: 35J60, 53C21, 53C44, 53C50, 58J05.

Keywords: Lorentzian manifold, mass, cosmological spacetime, general relativity, inverse curvature flow, ARW spacetimes.

Without loss of generality we shall assume $c_0 = 1$. Then N is ARW with respect to the future if the metric is close to the Robertson–Walker metric

$$d\bar{s}^2 = e^{2f}\left\{-\left(dx^0\right)^2 + \overline{\sigma}_{ij}(x)\,dx^i dx^j\right\} \tag{1-5}$$

near the singularity $\tau = b$. By *close* we mean that the derivatives of arbitrary order with respect to space and time of the conformal metric $e^{-2f}\breve{g}_{\alpha\beta}$ in (1-1) should converge to the corresponding derivatives of the conformal limit metric in (1-5) when x^0 tends to b. We emphasize that in our terminology Robertson–Walker metric does not imply that $(\overline{\sigma}_{ij})$ is a metric of constant curvature, it is only the spatial metric of a warped product.

We assume, furthermore, that f satisfies the following five conditions:

$$-f' > 0. \tag{1-6}$$

There exists $\omega \in \mathbb{R}$ such that

$$n + \omega - 2 > 0 \quad \text{and} \quad \lim_{\tau\to b} |f'|^2 e^{(n+\omega-2)f} = m > 0. \tag{1-7}$$

Set $\tilde{\gamma} = \frac{1}{2}(n+\omega-2)$, then the limit

$$\lim_{\tau\to b}\left(f'' + \tilde{\gamma}|f'|^2\right) \tag{1-8}$$

exists and

$$|D_\tau^m(f'' + \tilde{\gamma}|f'|^2)| \le c_m|f'|^m \qquad \text{for all } m \ge 1, \tag{1-9}$$

as well as

$$|D_\tau^m f| \le c_m|f'|^m \qquad \text{for all } m \ge 1. \tag{1-10}$$

We call N a *normalized* ARW spacetime if

$$\int_{\mathcal{G}_0} \sqrt{\det \overline{\sigma}_{ij}} = |S^n|. \tag{1-11}$$

Remark 1.2. (i) If these assumptions are satisfied, then we proved in [Gerhardt 2004] that the range of τ is finite, hence, we shall assume without loss of generality that $b = 0$, that is,

$$a < \tau < 0. \tag{1-12}$$

(ii) Any ARW spacetime can be normalized as one easily checks. For normalized ARW spaces the constant m in (1-7) is defined uniquely and can be identified with the mass of N, see [Gerhardt 2006b].

(iii) In view of the assumptions on f the mean curvature of the coordinate slices $M_\tau = \{x^0 = \tau\}$ tends to ∞ if τ goes to zero.

(iv) ARW spaces satisfy a strong volume decay condition, see [Gerhardt 2008, Definition 0.1].

(v) Similarly one can define N to be ARW with respect to the past. In this case the singularity would lie in the past, correspond to $\tau = 0$, and the mean curvature of the coordinate slices would tend to $-\infty$.

We assume that N satisfies the timelike convergence condition. Consider the future end N_+ of N and let $M_0 \subset N_+$ be a spacelike hypersurface with positive mean curvature $\check{H}_{|M_0} > 0$ with respect to the past directed normal vector $\check{\nu}$—we shall explain in Section 2 why we use the symbols $\check{H}$ and $\check{\nu}$ and not the usual ones H and ν. Then, as we have proved in [Gerhardt 2008], the inverse mean curvature flow

$$\dot{x} = -\check{H}^{-1}\,\check{\nu} \tag{1-13}$$

with initial hypersurface M_0 exists for all time, is smooth, and runs straight into the future singularity.

If we express the flow hypersurfaces $M(t)$ as graphs over $\mathcal{S}_0$

$$M(t) = \operatorname{graph} u(t, \cdot), \tag{1-14}$$

then one of the main results in our former paper was:

Theorem 1.3. (i) *Let N satisfy the above assumptions, then the range of the time function x^0 is finite, that is, we may assume that $b = 0$. Set*

$$\tilde{u} = ue^{\gamma t}, \tag{1-15}$$

where $\gamma = \frac{1}{n}\tilde{\gamma}$, then there are positive constants c_1, c_2 such that

$$-c_2 \le \tilde{u} \le -c_1 < 0, \tag{1-16}$$

and $\tilde{u}$ converges in $C^\infty(\mathcal{S}_0)$ to a smooth function, if t goes to infinity. We shall also denote the limit function by $\tilde{u}$.

(ii) *Let $\check{g}_{ij}$ be the induced metric of the leaves $M(t)$, then the rescaled metric*

$$e^{\frac{2}{n}t}\check{g}_{ij} \tag{1-17}$$

converges in $C^\infty(\mathcal{S}_0)$ to

$$(\tilde{\gamma}^2 m)^{\frac{1}{\tilde{\gamma}}}(-\tilde{u})^{\frac{2}{\tilde{\gamma}}}\overline{\sigma}_{ij}. \tag{1-18}$$

(iii) *The leaves $M(t)$ get more umbilical if t tends to infinity, namely,*

$$\check{H}^{-1}\left|\check{h}_i^j - \tfrac{1}{n}\check{H}\delta_i^j\right| \le ce^{-2\gamma t}. \tag{1-19}$$

In case $n+\omega-4>0$, *we even get a better estimate*

$$\left|\check{h}_i^j-\tfrac{1}{n}\check{H}\delta_i^j\right|\le ce^{-\frac{1}{2n}(n+\omega-4)t}. \tag{1-20}$$

The results for the mean curvature flow have recently also been proved for other inverse curvature flows, where the mean curvature is replaced by a curvature function F of class (K^*) homogeneous of degree 1, which includes the n-th root of the Gaussian curvature, see Kröner [2011].

In this note we want to prove that the functions in (1-15) converge to a constant. This result will also be valid when, instead of the mean curvature, other curvature functions F homogeneous of degree one will be considered satisfying

$$F(1,\dots,1)=n \tag{1-21}$$

provided the rescaled functions in (1-15) can be estimated as in (1-16) and converge in $C^3(\mathcal{S}_0)$. For simplicity we shall formulate the result only for the solution in Theorem 1.3, but it will be apparent from the proof that the result is also valid for different curvature functions.

Theorem 1.4. *The functions* $\tilde{u}$ *in* (1-15) *converge to a constant.*

2. Proof of Theorem 1.4

When we proved the convergence results for the inverse mean curvature flow in [Gerhardt 2004], we considered the flow hypersurfaces to be embedded in N equipped with the conformal metric

$$d\bar{s}^2=-\left(dx^0\right)^2+\sigma_{ij}(x^0,x)\,dx^i dx^j. \tag{2-1}$$

Though, formally, we have a different ambient space we still denote it by the same symbol N and distinguish only the metrics $\check{g}_{\alpha\beta}$ and $\bar{g}_{\alpha\beta}$

$$\check{g}_{\alpha\beta}=e^{2\widetilde{\psi}}\bar{g}_{\alpha\beta} \tag{2-2}$$

and the corresponding geometric quantities of the hypersurfaces $\check{h}_{ij}$, $\check{g}_{ij}$, $\check{\nu}$, respectively h_{ij}, g_{ij}, ν, and so on.

The second fundamental forms $\check{h}_i^j$ and h_i^j are related by

$$e^{\widetilde{\psi}}\check{h}_i^j=h_i^j+\widetilde{\psi}_\alpha\nu^\alpha\delta_i^j \tag{2-3}$$

and, if we define F by

$$F=e^{\widetilde{\psi}}\check{H}, \tag{2-4}$$

then

$$F=H-n\tilde{v}f'+n\psi_\alpha\nu^\alpha, \tag{2-5}$$

where

$$\tilde{v} = v^{-1}, \tag{2-6}$$

and

$$v^2 = 1 - \sigma^{ij} u_u u_j \equiv 1 - |Du|^2. \tag{2-7}$$

The evolution equation can be written as

$$\dot{x} = -F^{-1} \nu, \tag{2-8}$$

since

$$\check{\nu} = e^{-\tilde{\psi}} \nu. \tag{2-9}$$

The flow (2-8) can also be considered to comprise more general curvature functions F by assuming that $F = F\big(\check{h}^i_j\big)$, where $\check{h}^i_j$ is an abbreviation for the right-hand side of (2-3). Stipulating that indices of tensors will be raised or lowered with the help of the metric

$$g_{ij} = -u_i u_j + \sigma_{ij}, \tag{2-10}$$

we may also consider F to depend on

$$\check{h}_{ij} = h_{ij} - \tilde{v} f' g_{ij} + \psi_\alpha \nu^\alpha g_{ij} \tag{2-11}$$

and we define accordingly

$$F^{ij} = \frac{\partial F}{\partial \check{h}_{ij}}. \tag{2-12}$$

Now, let us prove Theorem 1.4. We use the relation

$$\tilde{v}^2 = 1 + \|Du\|^2 = 1 + g^{ij} u_i u_j \tag{2-13}$$

and shall prove that

$$\lim_{t\to\infty} \big(\|Du\|^2\big)' e^{2\gamma t} = 2\gamma \Delta\tilde{u}\, \tilde{u}, \tag{2-14}$$

where

$$\tilde{u} = \lim_{t\to\infty} u e^{\gamma t}, \tag{2-15}$$

as well as

$$\lim_{t\to\infty} \big(\tilde{v}^2\big)' e^{2\gamma t} = -2\gamma \|D\tilde{u}\|^2 \tag{2-16}$$

yielding

$$-\Delta\tilde{u}\, \tilde{u} = \|D\tilde{u}\|^2 \tag{2-17}$$

on the compact limit hypersurface M. Since $\tilde{u}$ is strictly negative we then conclude

$$\int_M \|D\tilde{u}\|^2 \tilde{u}^{-1} = 0, \tag{2-18}$$

hence $\|D\tilde{u}\| = 0$.

Let us first derive (2-14). Using

$$\dot{g}_{ij} = -2F^{-1}h_{ij}, \tag{2-19}$$

see [Gerhardt 2006a, Lemma 2.3.1], where we write $g_{ij} = g_{ij}(t,\xi)$, $\xi = (\xi^i)$ are local coordinates for $\mathcal{S}_0$, and where

$$\dot{g}_{ij} = \frac{\partial g_{ij}}{\partial t} = \dot{u}_i u_j + u_i \dot{u}_j + \dot{\sigma}_{ij}\dot{u}, \tag{2-20}$$

and $\dot{\sigma}_{ij}$ is defined by

$$\dot{\sigma}_{ij} = \frac{\partial \sigma_{ij}}{\partial u}, \tag{2-21}$$

we deduce

$$\begin{aligned}\big(\|Du\|^2\big)' = \big(g^{ij}u_i u_j\big)' &= 2g^{ij}\dot{u}_i u_j - \dot{g}_{ij}u^i u^j \\ &= 2F^{-1}H + g^{ij}\dot{\sigma}_{ij}\dot{u} - \dot{g}_{ij}u^i u^j \\ &= 2F^{-1}H + \tilde{v}F^{-1}g^{ij}\dot{\sigma}_{ij} + 2F^{-1}h_{ij}u^i u^j \\ &= 2F^{-1}H + \tilde{v}F^{-1}\sigma^{ij}\dot{\sigma}_{ij} + \tilde{v}^3F^{-1}\dot{\sigma}_{ij}\check{u}^i\check{u}^j + 2F^{-1}h_{ij}u^i u^j,\end{aligned} \tag{2-22}$$

where we used the relation

$$g^{ij} = \sigma^{ij} + \tilde{v}^2\check{u}^i\check{u}^j \tag{2-23}$$

and where $\check{u}^i$ is defined by

$$\check{u}^i = \sigma^{ij}u_j. \tag{2-24}$$

The last two terms on the right-hand side of (2-22) are an $o\big(e^{-2\gamma t}\big)$, thus we have

$$\big(\|Du\|^2\big)' = 2F^{-1}\big(H + \tilde{v}\tfrac{1}{2}\sigma^{ij}\dot{\sigma}_{ij}\big) + o\big(e^{-2\gamma t}\big). \tag{2-25}$$

On the other hand,

$$h_{ij}\tilde{v} = -u_{ij} + \bar{h}_{ij}, \tag{2-26}$$

where $\bar{h}_{ij}$ is the second fundamental form of the slices $\{x^0 = \text{const}\}$

$$\bar{h}_{ij} = -\tfrac{1}{2}\dot{\sigma}_{ij} \tag{2-27}$$

and we infer

$$H\tilde{v} = -\Delta u + g^{ij}\bar{h}_{ij} = -\Delta u + \bar{H} + \tilde{v}^2\bar{h}_{ij}\check{u}^i\check{u}^j. \tag{2-28}$$

Combining (2-22), (2-27) and (2-28) we obtain

$$\begin{aligned}\big(\|Du\|^2\big)' &= 2F^{-1}\big(H - \tilde{v}\bar{H}\big) + o\big(e^{-2\gamma t}\big)\\ &= 2F^{-1}\big(H - \bar{H}\big) + o\big(e^{-2\gamma t}\big)\\ &= -2F^{-1}\Delta u + o\big(e^{-2\gamma t}\big).\end{aligned} \tag{2-29}$$

In view of [Gerhardt 2006a, Lemma 7.3.4], the estimates for h_{ij}, u, and ψ, and the homogeneity of F, we have

$$\lim_{t\to\infty} F(-u) = n\tilde{\gamma}^{-1} = \gamma^{-1}, \tag{2-30}$$

hence we deduce

$$\lim_{t\to\infty}\big(\|Du\|^2\big)' e^{2\gamma t} = 2\gamma\,\Delta\tilde{u}\,\tilde{u}. \tag{2-31}$$

Let us now differentiate $\tilde{v}^2$. From the relation

$$\tilde{v} = \eta_\alpha \nu^\alpha, \quad (\eta_\alpha) = (-1, 0, \dots, 0), \tag{2-32}$$

we infer

$$\dot{\tilde{v}} = \eta_{\alpha\beta}\nu^\alpha\dot{x}^\beta + \eta_\alpha\dot{\nu}^\alpha = -F^{-1}\eta_{\alpha\beta}\nu^\alpha\nu^\beta + \big(F^{-1}\big)_k u^k, \tag{2-33}$$

where we used

$$\dot{\nu} = \big(-F^{-1}\big)^k x_k, \tag{2-34}$$

see [Gerhardt 2006a, Lemma 2.3.2]. The first term on the right-hand side of (2-33) is an $o\big(e^{-2\gamma t}\big)$ in view of the asymptotic behavior of an ARW space, see the definition of *close* in Definition 1.1, while

$$\begin{aligned}&\big(F^{-1}\big)_k\\ &\quad = -F^{-2}F^{ij}\big\{h_{ij;k} - \tilde{v}_k f' g_{ij} - \tilde{f}'' u_k g_{ij} + \psi_{\alpha\beta}\nu^\alpha x^\beta_k g_{ij} + \psi_\alpha x^\alpha_l h^l_k g_{ij}\big\},\end{aligned} \tag{2-35}$$

where we applied the Weingarten equation to derive the last term on the right-hand side. Therefore, we infer

$$\lim_{t\to\infty}\big(F^{-1}\big)_k u^k e^{2\gamma t} = \|D\tilde{u}\|^2\tfrac{1}{n}\lim\frac{f''}{|f'|^2} = -\frac{\tilde{\gamma}}{n}\|D\tilde{u}\|^2 = -\gamma\|D\tilde{u}\|^2, \tag{2-36}$$

in view of (1-8) and the definition of γ in Theorem 1.3, and we deduce further

$$\lim_{t\to\infty}\big(\tilde{v}^2\big)' e^{2\gamma t} = -2\gamma\|D\tilde{u}\|^2, \tag{2-37}$$

hence the limit function $\tilde{u}$ satisfies

$$\|D\tilde{u}\|^2 = -\Delta\tilde{u}\,\tilde{u} \tag{2-38}$$

completing the proof of Theorem 1.4.

Remark 2.1. We believe that this method of proof will also work for other curvature flows driven by extrinsic curvatures, in Riemannian or Lorentzian manifolds, to prove that the leaves of the rescaled curvature flows converge to the graph of a constant function.

Indeed, applying this method we proved in [Gerhardt 2011, Lemma 6.12] that the rescaled curvature flow converges to a sphere.

References

[Gerhardt 2004] C. Gerhardt, "The inverse mean curvature flow in ARW spaces—transition from big crunch to big bang", preprint, 2004. arXiv math/0403485

[Gerhardt 2006a] C. Gerhardt, *Curvature problems*, Series in Geometry and Topology **39**, International Press, Somerville, MA, 2006. MR 2007j:53001 Zbl 1131.53001

[Gerhardt 2006b] C. Gerhardt, "The mass of a Lorentzian manifold", *Adv. Theor. Math. Phys.* **10**:1 (2006), 33–48. MR 2006m:53109 Zbl 1104.83019

[Gerhardt 2008] C. Gerhardt, "The inverse mean curvature flow in cosmological spacetimes", *Adv. Theor. Math. Phys.* **12**:6 (2008), 1183–1207. MR 2009i:53059 Zbl 1153.83016

[Gerhardt 2011] C. Gerhardt, "Inverse curvature flows in hyperbolic space", *J. Diff. Geom.* **89**:3 (2011), 487–527. arXiv 1101.2578

[Kröner 2011] H. Kröner, "The inverse F-curvature flow in ARW spaces", preprint, 2011. arXiv 1106.4703

Received July 5, 2011. Revised July 6, 2011.

CLAUS GERHARDT
INSTITUTE OF APPLIED MATHEMATICS
RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
D-69120 HEIDELBERG
GERMANY

gerhardt@math.uni-heidelberg.de
http://www.math.uni-heidelberg.de/studinfo/gerhardt/

PACIFIC JOURNAL OF MATHEMATICS
Vol. 256, No. 2, 2012

TOTAL CURVATURE OF GRAPHS AFTER MILNOR AND EULER

ROBERT GULLIVER AND SUMIO YAMADA

We define a new notion of total curvature, called *net total curvature*, for finite graphs embedded in $\mathbb{R}^n$, and investigate its properties. Two guiding principles are given by Milnor's way of measuring using a local Crofton-type formula, and by considering the double cover of a given graph as an Eulerian circuit. The strength of combining these ideas in defining the curvature functional is that it allows us to interpret the singular/noneuclidean behavior at the vertices of the graph as a superposition of vertices of a 1-dimensional manifold, so that one can compute the total curvature for a wide range of graphs by contrasting local and global properties of the graph utilizing the integral geometric representation of the curvature. A collection of results on upper/lower bounds of the total curvature on isotopy/homeomorphism classes of embeddings is presented, which in turn demonstrates the effectiveness of net total curvature as a new functional measuring complexity of spatial graphs in differential-geometric terms.

1. Introduction: curvature of a graph

The celebrated Fáry–Milnor theorem states that a curve in $\mathbb{R}^n$ of total curvature at most 4π is unknotted.

As a key step in his proof, John Milnor [1950] showed that for a smooth Jordan curve Γ in $\mathbb{R}^3$, the total curvature equals half the integral over $e \in S^2$ of the number $\mu(e)$ of local maxima of the linear height function $\langle e, \cdot \rangle$ along Γ. This equality can be regarded as a Crofton-type representation formula of total curvature where the order of integrations over the curve and the unit tangent sphere (the space of directions) are reversed. The Fáry–Milnor theorem follows, since total curvature less than 4π implies there is a unit vector $e_0 \in S^2$ so that $\langle e_0, \cdot \rangle$ has a unique local maximum, and therefore that this linear function is increasing on an interval of Γ and decreasing on the complement. Without changing the pointwise value of this height function, Γ can be topologically untwisted to a standard embedding of S^1

Supported in part by JSPS Grant-in-aid for Scientific Research No. 17740030. The authors thank the Korea Institute for Advanced Study for invitations.
MSC2010: 05C99, 53A04, 57M25, 57N45.
Keywords: spatial graphs, total curvature, Milnor.

into $\mathbb{R}^3$. The Fenchel theorem, that any curve in $\mathbb{R}^3$ has total curvature at least 2π, also follows from Milnor's key step, since for all $e \in S^2$, the linear function $\langle e, \cdot \rangle$ assumes its maximum somewhere along Γ, implying $\mu(e) \geq 1$. Milnor's proof is independent of the proof of Istvan Fáry, published earlier [1949], which takes a different approach.

We would like to extend the methods of Milnor's seminal paper, replacing the simple closed curve by a finite *graph* Γ in $\mathbb{R}^3$. Γ consists of a finite number of points, the *vertices*, and a finite number of simple arcs, the *edges*, each of which has as its endpoints one or two of the vertices. We shall assume Γ is connected. The *degree* of a vertex q is the number $d(q)$ of edges which have q as an endpoint. (Another word for degree is "valence".) We remark that it is technically not needed that the dimension n of the ambient space equals three. All the arguments can be generalized to higher dimensions, although in higher dimensions ($n \geq 4$) there are no nontrivial knots, and any two homeomorphic graphs are isotopic.

The key idea in generalizing total curvature for curves to total curvature for graphs is to consider the Euler circuits, namely, parametrizations by S^1, of the *double* cover of the graph. We note that given a graph of even degree, there can be several Euler circuits, or ways to trace it without lifting the pen. A topological vertex of a graph of degree d is a singularity, in that the graph is not locally Euclidean. However by considering an Euler circuit of the double of the graph, the vertex becomes locally the intersection point of d paths. We will show (Corollary 3.7) that at the vertex, each path through it has a (signed) measure-valued curvature, and the absolute value of the sum of those measures is well-defined, independent of the choice of the Euler circuit of the double cover. We define (Definition 2.1) the *net total curvature* (NTC) of a piecewise C^2 graph to be the sum of the total curvature of the smooth arcs and the contributions from the vertices as described.

This notion of net total curvature is substantially different from the total curvature, denoted TC, as defined by Taniyama [1998]. (Taniyama writes τ for TC.) See Section 2 below.

This is consistent with known results for the vertices of degree $d = 2$; with vertices of degree three or more, this definition helps facilitate a new Crofton-type representation formula (Theorem 3.13) for total curvature of graphs, where the total curvature is represented as an integral over the unit sphere. Recall that the vertex is now seen as d distinct points on an Euler circuit. The way we pick up the contribution of the total curvature at the vertices identifies the d distinct points, and thus the $2d$ unit tangent spheres on a circuit. As Crofton's formula in effect reverses the order of integrations — one over the circuit, the other over the space of tangent directions — the sum of the d exterior angles at the vertex is incorporated in the integral over the unit sphere. On the other hand the integrand of the integral over the unit sphere counts the number of net local maxima of the height function

along an axis, where net local maximum means the number of local maxima minus the number of local minima at these d points of the Euler circuit. This establishes a correspondence between the differential geometric quantity (net total curvature) and the differential topological quantity (average number of maxima) of the graph, as stated in Theorem 3.13 below.

In Section 2, we compare several definitions for total curvature of graphs which have appeared in the recent literature. In Section 3, we introduce the main tool (Lemma 3.5) which in a sense reduces the computation of NTC to counting intersections with planes.

Milnor's treatment [1950] of total curvature also contained an important topological extension. Namely, in order to define total curvature, the curve needs only to be *continuous*. This makes the total curvature a geometric quantity defined on any homeomorphic image of S^1. In this article, we first define net total curvature (Definition 2.1) on piecewise C^2 graphs, and then extend the definition to continuous graphs (Definition 2.3.) In analogy to Milnor, we approximate a given continuous graph by a sequence of polygonal graphs. In showing the monotonicity of the total curvature (Proposition 4.1) under the refining process of approximating graphs we use our representation formula (Theorem 3.13) applied to the polygonal graphs.

Consequently the Crofton-type representation formula is also extended to cover continuous graphs (Theorem 4.9). Additionally, we are able to show that continuous graphs with finite total curvature (NTC or TC) are tame. We say that a graph is *tame* when it is isotopic to an embedded polyhedral graph.

In sections 5 through 8, we characterize NTC with respect to the geometry and the topology of the graph. Proposition 5.5 shows the subadditivity of NTC under the union of graphs which meet in a finite set. In Section 6, the concept of bridge number is extended from curves to graphs, in terms of which the minimum of NTC can be explicitly computed, provided the graph has at most one vertex of degree > 3. In Section 7, Theorem 7.1 gives a lower bound for NTC in terms of the width of an isotopy class. The infimum of NTC is computed for specific graph types: the two-vertex graphs θ_m, the "ladder" L_m, the "wheel" W_m, the complete graph K_m on m vertices and the complete bipartite graph $K_{m,n}$.

Finally we prove a result (Theorem 8.5) which gives a Fenchel type lower bound ($\geq 3\pi$) for total curvature of a theta graph (an image of the graph consisting of a circle with an arc connecting a pair of antipodal points), and a Fáry–Milnor type upper bound ($< 4\pi$) to imply the theta graph is isotopic to the standard embedding. A similar result was given by Taniyama [1998], referring to TC. In contrast, for graphs of the type of K_m ($m \geq 4$), the infimum of NTC in the isotopy class of a polygon on m vertices is also the infimum for a sequence of distinct isotopy classes (Corollary 8.3).

Many of the results in our earlier preprint [Gulliver and Yamada 2008] have been incorporated into the present paper.

We thank Yuya Koda for his comments regarding Proposition 6.1, and Jaigyoung Choe and Rob Kusner for their comments about Theorem 8.5, especially about the sharp case $\mathrm{NTC}(\Gamma) = 3\pi$ of the lower bound estimate.

2. Definitions of total curvature

The first difficulty, in extending the results of Milnor's classic paper, is to understand the contribution to total curvature at a vertex of degree $d(q) \geq 3$. We first consider the well-known case:

Definition of total curvature for curves. For a smooth closed curve Γ, the total curvature is

$$\mathscr{C}(\Gamma) = \int_\Gamma |\vec{k}|\, ds,$$

where s denotes arc length along Γ and $\vec{k}$ is the curvature vector. If $x(s) \in \mathbb{R}^3$ denotes the position of the point measured at arc length s along the curve, then $\vec{k} = \frac{d^2x}{ds^2}$. For a piecewise smooth curve, that is, a graph with vertices $q_1, \ldots, q_N$ having always degree $d(q_i) = 2$, the total curvature is readily generalized to

$$\mathscr{C}(\Gamma) = \sum_{i=1}^{N} \mathrm{c}(q_i) + \int_{\Gamma_{\mathrm{reg}}} |\vec{k}|\, ds, \tag{2-1}$$

where the integral is taken over the separate C^2 edges of Γ without their endpoints; and where $\mathrm{c}(q_i) \in [0, \pi]$ is the exterior angle formed by the two edges of Γ which meet at q_i. That is, $\cos(\mathrm{c}(q_i)) = \langle T_1, -T_2 \rangle$, where $T_1 = \frac{dx}{ds}(q_i^+)$ and $T_2 = -\frac{dx}{ds}(q_i^-)$ are the unit tangent vectors at q_i pointing into the two edges which meet at q_i. The exterior angle $\mathrm{c}(q_i)$ is the correct contribution to total curvature, since any sequence of smooth curves converging to Γ in C^0, with C^1 convergence on compact subsets of each open edge, includes a small arc near q_i along which the tangent vector changes from near $\frac{dx}{ds}(q_i^-)$ to near $\frac{dx}{ds}(q_i^+)$. The greatest lower bound of the contribution to total curvature of this disappearing arc along the smooth approximating curves equals $\mathrm{c}(q_i)$.

Note that $\mathscr{C}(\Gamma)$ is well defined for an *immersed* curve Γ.

Definitions of total curvature for graphs. When we turn our attention to a *graph* Γ, we find the above definition for curves (degree $d(q) = 2$) does not generalize in any obvious way to higher degree (see [Gulliver 2007]). The ambiguity of the general formula (2-1) is resolved if we specify the replacement for $\mathrm{c}(0)$ when Γ is the cone over a finite set $\{T_1, \ldots, T_d\}$ in the unit sphere S^2.

The earliest notion of total curvature of a graph appears in the context of the first variation of length of a graph, which we call *variational total curvature*, and is called the *mean curvature* of the graph in [Allard and Almgren 1976]: we shall write VTC. The contribution to VTC at a vertex q of degree 2, with unit tangent vectors T_1 and T_2, is $\mathrm{vtc}(q) = |T_1 + T_2| = 2\sin(c(q)/2)$. At a nonstraight vertex q of degree 2, $\mathrm{vtc}(q)$ is less than the exterior angle $\mathrm{c}(q)$. For a vertex of degree d, the contribution is $\mathrm{vtc}(q) = |T_1 + \cdots + T_d|$.

A rather natural definition of total curvature of graphs was given in [Taniyama 1998]. We have called this *maximal total curvature* $\mathrm{TC}(\Gamma)$ in [Gulliver 2007]. The contribution to total curvature at a vertex q of degree d is

$$\mathrm{tc}(q) := \sum_{1 \le i < j \le d} \arccos\langle T_i, -T_j \rangle.$$

In the case $d(q) = 2$, the sum above has only one term, the exterior angle $\mathrm{c}(q)$ at q. Since the length of the Gauss image of a curve in S^2 is the total curvature of the curve, $\mathrm{tc}(q)$ may be interpreted as adding to the Gauss image in $\mathbb{R}P^2$ of the edges, a complete great-circle graph on $T_1(q), \ldots, T_d(q)$, for each vertex q of degree d. Note that the edge between two vertices does not measure the distance in $\mathbb{R}P^2$ but its supplement.

In [Gulliver and Yamada 2006], studying the density of an area-minimizing two-dimensional rectifiable set Σ spanning Γ, we found that it was very useful to apply the Gauss–Bonnet formula to the cone over Γ with a point p of Σ as vertex. The relevant notion of total curvature in that context is *cone total curvature* $\mathrm{CTC}(\Gamma)$, defined using $\mathrm{ctc}(q)$ as the replacement for $\mathrm{c}(q)$ in (2-1):

$$\mathrm{ctc}(q) := \sup_{e \in S^2} \left\{ \sum_{i=1}^{d} \left(\frac{\pi}{2} - \arccos\langle T_i, e \rangle \right) \right\}. \tag{2-2}$$

Note that in the case $d(q) = 2$, the supremum above is assumed at vectors e lying in the smaller angle between the tangent vectors T_1 and T_2 to Γ, so that $\mathrm{ctc}(q)$ is then the exterior angle $\mathrm{c}(q)$ at q. The main result of [Gulliver and Yamada 2006] is that 2π times the area density of Σ at any of its points is at most equal to $\mathrm{CTC}(\Gamma)$. The same result had been proven by Eckholm, White and Wienholtz for the case of a simple closed curve [Ekholm et al. 2002]. Taking Σ to be the branched immersion of the disk given by Douglas [1931] and Radó [1933], it follows that if $\mathscr{C}(\Gamma) \le 4\pi$, then Σ is embedded, and therefore Γ is unknotted. Thus [Ekholm et al. 2002] provided an independent proof of the Fáry–Milnor theorem. However, $\mathrm{CTC}(\Gamma)$ may be small for graphs which are far from the simplest isotopy types of a graph Γ.

In this paper, we introduce the notion of *net total curvature* $\mathrm{NTC}(\Gamma)$, which is the appropriate definition for generalizing — *to graphs* — Milnor's approach to

isotopy and total curvature of *curves*. For each unit tangent vector T_i at q, where $1 \le i \le d = d(q)$, let $\chi_i : S^2 \to \{-1, +1\}$ be equal to -1 on the hemisphere with center at T_i, and $+1$ on the opposite hemisphere (modulo sets of zero Lebesgue measure). We then define

$$\mathrm{ntc}(q) := \frac{1}{4} \int_{S^2} \left[\sum_{i=1}^{d} \chi_i(e) \right]^+ dA_{S^2}(e). \tag{2-3}$$

We note that the function $\sum_{i=1}^{d} \chi_i(e)$ is odd, hence the quantity above can be written as

$$\mathrm{ntc}(q) := \frac{1}{8} \int_{S^2} \left| \sum_{i=1}^{d} \chi_i(e) \right| dA_{S^2}(e).$$

as well. In the case $d(q) = 2$, the integrand of (2-3) is positive (and equals 2) only on the set of unit vectors e which have negative inner products with both T_1 and T_2, ignoring e in sets of measure zero. This set is bounded by great semicircles orthogonal to T_1 and to T_2, and has spherical area equal to twice the exterior angle. So in this case, $\mathrm{ntc}(q)$ is the exterior angle. Thus, in the special case where Γ is a piecewise smooth curve, the following quantity $\mathrm{NTC}(\Gamma)$ coincides with total curvature, as well as with $\mathrm{TC}(\Gamma)$ and $\mathrm{CTC}(\Gamma)$:

Definition 2.1. We define the *net total curvature* of a piecewise C^2 graph Γ with vertices $\{q_1, \ldots, q_N\}$ as

$$\mathrm{NTC}(\Gamma) := \sum_{i=1}^{N} \mathrm{ntc}(q_i) + \int_{\Gamma_{\mathrm{reg}}} |\vec{k}|\, ds. \tag{2-4}$$

For the sake of simplicity, elsewhere in this paper, we consider the ambient space to be $\mathbb{R}^3$. However the definition of the net total curvature can be generalized for a graph in $\mathbb{R}^n$ by defining the vertex contribution in terms of an average over S^{n-1}:

$$\mathrm{ntc}(q) := \pi \fint_{S^{n-1}} \left[\sum_{i=1}^{d} \chi_i(e) \right]^+ dA_{S^{n-1}}(e),$$

which is consistent with the definition (2-3) of ntc when $n = 3$.

Recall that Milnor defines the total curvature of a continuous simple closed curve C as the supremum of the total curvature of all polygons inscribed in C. By analogy, we define net total curvature of a *continuous* graph Γ to be the supremum of the net total curvature of all polygonal graphs P suitably inscribed in Γ as follows.

Definition 2.2. For a given continuous graph Γ, we say a polygonal graph $P \subset \mathbb{R}^3$ is Γ-*approximating*, provided that its topological vertices (those of degree $\neq 2$) are

exactly the topological vertices of Γ, and having the same degrees; and that the arcs of P between two topological vertices correspond one-to-one to the edges of Γ between those two vertices.

Note that if P is a Γ-approximating polygonal graph, then P is homeomorphic to Γ. According to the statement of Proposition 4.1, whose proof will be given in the next section, if P and $\widetilde{P}$ are Γ-approximating polygonal graphs, and $\widetilde{P}$ is a refinement of P, then $\mathrm{NTC}(\widetilde{P}) \geq \mathrm{NTC}(P)$. Here $\widetilde{P}$ is said to be a refinement of P provided the set of vertices of P is a subset of the vertices of $\widetilde{P}$. Assuming Proposition 4.1 for the moment, we can generalize the definition of the total curvature to nonsmooth graphs.

Definition 2.3. Define the *net total curvature* of a continuous graph Γ by

$$\mathrm{NTC}(\Gamma) := \sup_P \mathrm{NTC}(P)$$

where the supremum is taken over all Γ-approximating polygonal graphs P.

For a polygonal graph P, applying Definition 2.1,

$$\mathrm{NTC}(P) := \sum_{i=1}^{N} \mathrm{ntc}(q_i),$$

where $q_1, \ldots, q_N$ are the vertices of P.

Definition 2.3 is consistent with Definition 2.1 in the case of a piecewise C^2 graph Γ. Namely, as Milnor showed [1950, p. 251], the total curvature $\mathscr{C}(\Gamma_0)$ of a smooth curve Γ_0 is the supremum of the total curvature of inscribed polygons, which gives the required supremum for each edge. At a vertex q of the piecewise-C^2 graph Γ, as a sequence P_k of Γ-approximating polygons become arbitrarily fine, a vertex q of P_k (and of Γ) has unit tangent vectors converging in S^2 to the unit tangent vectors to Γ at q. It follows that for $1 \leq i \leq d(q)$, $\chi_i^{P_k} \to \chi_i^{\Gamma}$ in measure on S^2, and therefore $\mathrm{ntc}_{P_k}(q) \to \mathrm{ntc}_{\Gamma}(q)$.

3. Crofton-type representation formula for total curvature

We would like to explain how the net total curvature $\mathrm{NTC}(\Gamma)$ of a piecewise C^2 graph Γ is related to more familiar notions of total curvature. Recall that Γ has an Euler circuit if and only if its vertices all have even degree, by a theorem of Euler. An Euler circuit is a closed, connected path which traverses each edge of Γ exactly once. Of course, we do not have the hypothesis of even degree. We can attain that hypothesis by passing to the *double* $\widetilde{\Gamma}$ of Γ: $\widetilde{\Gamma}$ is the graph with the same vertices as Γ, but with two copies of each edge of Γ. Then at each vertex q, the degree as a vertex of $\widetilde{\Gamma}$ is $\widetilde{d}(q) = 2\,d(q)$, which is even. By Euler's theorem, there is an Euler circuit Γ' of $\widetilde{\Gamma}$, which may be thought of as a closed path which traverses

each edge of Γ exactly *twice*. Now at each of the points $\{q_1, \ldots, q_d\}$ along Γ' which are mapped to $q \in \Gamma$, we may consider the exterior angle $c(q_i)$. The sum of these exterior angles, however, depends on the choice of the Euler circuit Γ'. For example, if Γ is the union of the x-axis and the y-axis in Euclidean space $\mathbb{R}^3$, then one might choose Γ' to have four right angles, or to have four straight angles, or something in between, with completely different values of total curvature. In order to form a version of total curvature at a vertex q which only depends on the original graph Γ and not on the choice of Euler circuit Γ', it is necessary to consider some of the exterior angles as partially balancing others. In the example just considered, where Γ is the union of two orthogonal lines, two opposing right angles will be considered to balance each other completely, so that $\mathrm{ntc}(q) = 0$, regardless of the choice of Euler circuit of the double.

It will become apparent that the connected character of an Euler circuit of $\widetilde{\Gamma}$ is not required for what follows. Instead, we shall refer to a *parametrization* Γ' of the double $\widetilde{\Gamma}$, which is a mapping from a 1-dimensional manifold without boundary, not necessarily connected; the mapping is assumed to cover each edge of $\widetilde{\Gamma}$ once.

The nature of $\mathrm{ntc}(q)$ is clearer when it is localized on S^2, analogously to [Milnor 1950]. In the case $d(q) = 2$, Milnor observed that the exterior angle at the vertex q equals half the area of those $e \in S^2$ such that the linear function $\langle e, \cdot \rangle$, restricted to Γ, has a local maximum at q. In our context, we may describe $\mathrm{ntc}(q)$ as one-half the integral over the sphere of the number of *net local maxima*, which is half the difference of local maxima and local minima. Along the parametrization Γ' of the double of Γ, the linear function $\langle e, \cdot \rangle$ may have a local maximum at some of the vertices $q_1, \ldots, q_d$ over q, and may have a local minimum at others. In our construction, each local minimum balances against one local maximum. If there are more local minima than local maxima, the number $\mathrm{nlm}(e, q)$, the net number of local maxima, will be negative; however, our definition uses only the positive part $[\mathrm{nlm}(e, q)]^+$.

We need to show that

$$\int_{S^2} [\mathrm{nlm}(e, q)]^+ \, dA_{S^2}(e)$$

is independent of the choice of parametrization, and in fact is equal to $2\,\mathrm{ntc}(q)$; this will follow from another way of computing $\mathrm{nlm}(e, q)$ (see Corollary 3.7).

Definition 3.1. Let a parametrization Γ' of the double of Γ be given. Then a vertex q of Γ corresponds to a number of vertices $q_1, \ldots, q_d$ of Γ', where d is the degree $d(q)$ of q as a vertex of Γ. Choose $e \in S^2$. If $q \in \Gamma$ is a local extremum of $\langle e, \cdot \rangle$, then we consider q as a vertex of degree $d(q) = 2$. Let $\mathrm{lmax}(e, q)$ be the number of local maxima of $\langle e, \cdot \rangle$ along Γ' at the points $q_1, \ldots, q_d$ over q, and similarly let $\mathrm{lmin}(e, q)$ be the number of local minima. We define the number of

net local maxima of $\langle e, \cdot \rangle$ at q to be

$$\mathrm{nlm}(e, q) = \tfrac{1}{2}[\mathrm{lmax}(e, q) - \mathrm{lmin}(e, q)].$$

Remark 3.2. The definition of $\mathrm{nlm}(e, q)$ appears to depend not only on Γ but on a choice of the parametrization Γ' of the double of Γ: $\mathrm{lmax}(e, q)$ and $\mathrm{lmin}(e, q)$ may depend on the choice of Γ'. However, we shall see in Corollary 3.6 below that the number of *net* local maxima $\mathrm{nlm}(e, q)$ is in fact independent of Γ'.

Remark 3.3. We have included the factor $\frac{1}{2}$ in the definition of $\mathrm{nlm}(e, q)$ in order to agree with the difference of the numbers of local maxima and minima along a parametrization of Γ itself, if $d(q)$ is even.

We shall *assume* for the rest of this section that a unit vector e has been chosen, and that the linear height function $\langle e, \cdot \rangle$ has only a finite number of critical points along Γ; this excludes e belonging to a subset of S^2 of measure zero. We shall also assume that the graph Γ is subdivided to include among the vertices all critical points of the linear function $\langle e, \cdot \rangle$, with degree $d(q) = 2$ if q is an interior point of one of the topological edges of Γ.

Definition 3.4. Choose a unit vector e. At a point $q \in \Gamma$ of degree $d = d(q)$, let the *up-degree* $d^+ = d^+(e, q)$ be the number of edges of Γ with endpoint q on which $\langle e, \cdot \rangle$ exceeds $\langle e, q \rangle$, the height of q. Similarly, let the *down-degree* $d^-(e, q)$ be the number of edges along which $\langle e, \cdot \rangle$ is less than its value at q. Note that $d(q) = d^+(e, q) + d^-(e, q)$, for almost all e in S^2.

Lemma 3.5 (combinatorial lemma). *For all $q \in \Gamma$ and for almost all $e \in S^2$,*

$$\mathrm{nlm}(e, q) = \tfrac{1}{2}[d^-(e, q) - d^+(e, q)].$$

Proof. Let a parametrization Γ' of the double of Γ be chosen, with respect to which $\mathrm{lmax}(e, q)$ and $\mathrm{lmin}(e, q)$ are defined. Recall the assumption above, that Γ has been subdivided so that along each edge, the linear function $\langle e, \cdot \rangle$ is strictly monotone.

Consider a vertex q of Γ, of degree $d = d(q)$. Then Γ' has $2d$ edges with an endpoint among the points $q_1, \ldots, q_d$ which are mapped to $q \in \Gamma$. On $2d^+$, resp. $2d^-$ of these edges, $\langle e, \cdot \rangle$ is greater resp. less than $\langle e, q \rangle$. But for each $1 \leq i \leq d$, the parametrization Γ' has exactly two edges which meet at q_i. Depending on the up/down character of the two edges of Γ' which meet at q_i, $1 \leq i \leq d$, we can count:

(+) If $\langle e, \cdot \rangle$ is greater than $\langle e, q \rangle$ on both edges, then q_i is a local minimum point; there are $\mathrm{lmin}(e, q)$ of these among $q_1, \ldots, q_d$.
(-) If $\langle e, \cdot \rangle$ is less than $\langle e, q \rangle$ on both edges, then q_i is a local maximum point; there are $\mathrm{lmax}(e, q)$ of these.

(0) In all remaining cases, the linear function $\langle e, \cdot \rangle$ is greater than $\langle e, q \rangle$ along one edge and less along the other, in which case q_i is not counted in computing $\operatorname{lmax}(e, q)$ nor $\operatorname{lmax}(e, q)$; there are $d(q) - \operatorname{lmax}(e, q) - \operatorname{lmin}(e, q)$ of these.

Now count the individual edges of Γ':

(+) There are $\operatorname{lmin}(e, q)$ pairs of edges, each of which is part of a local minimum, both of which are counted among the $2d^+(e, q)$ edges of Γ' with $\langle e, \cdot \rangle$ greater than $\langle e, q \rangle$.

(-) There are $\operatorname{lmax}(e, q)$ pairs of edges, each of which is part of a local maximum; these are counted among the number $2d^-(e, q)$ of edges of Γ' with $\langle e, \cdot \rangle$ less than $\langle e, q \rangle$. Finally,

(0) there are $d(q) - \operatorname{lmax}(e, q) - \operatorname{lmin}(e, q)$ edges of Γ' which are not part of a local maximum or minimum, with $\langle e, \cdot \rangle$ greater than $\langle e, q \rangle$; and an equal number of edges with $\langle e, \cdot \rangle$ less than $\langle e, q \rangle$.

Thus, the total number of these edges of Γ' with $\langle e, \cdot \rangle$ greater than $\langle e, q \rangle$ is

$$2d^+ = 2 \operatorname{lmin} + (d - \operatorname{lmax} - \operatorname{lmin}) = d + \operatorname{lmin} - \operatorname{lmax}.$$

Similarly,

$$2d^- = 2 \operatorname{lmax} + (d - \operatorname{lmax} - \operatorname{lmin}) = d + \operatorname{lmax} - \operatorname{lmin}.$$

Subtracting gives the conclusion:

$$\operatorname{nlm}(e, q) := \frac{\operatorname{lmax}(e, q) - \operatorname{lmin}(e, q)}{2} = \frac{d^-(e, q) - d^+(e, q)}{2}. \qquad \square$$

Corollary 3.6. *The number of net local maxima* $\operatorname{nlm}(e, q)$ *is independent of the choice of parametrization* Γ' *of the double of* Γ.

Proof. Given a direction $e \in S^2$, the up-degree and down-degree $d^{\pm}(e, q)$ at a vertex $q \in \Gamma$ are defined independently of the choice of Γ'. $\square$

Corollary 3.7. *For any* $q \in \Gamma$, *we have* $\operatorname{ntc}(q) = \frac{1}{2} \int_{S^2} [\operatorname{nlm}(e, q)]^+ \, dA_{S^2}$.

Proof. Consider $e \in S^2$. In the definition (2-3) of $\operatorname{ntc}(q)$, $\chi_i(e) = \pm 1$ whenever $\pm\langle e, T_i \rangle < 0$. But the number of $1 \le i \le d$ with $\pm\langle e, T_i \rangle < 0$ equals $d^{\mp}(e, q)$, so that

$$\sum_{i=1}^{d} \chi_i(e) = d^-(e, q) - d^+(e, q) = 2 \operatorname{nlm}(e, q)$$

by Lemma 3.5, for almost all $e \in S^2$. $\square$

Definition 3.8. For a graph Γ in $\mathbb{R}^3$ and $e \in S^2$, define the *multiplicity at* e as

$$\mu(e) = \mu_\Gamma(e) = \sum \{\operatorname{nlm}^+(e, q) : q \text{ a vertex of } \Gamma \text{ or a critical point of } \langle e, \cdot \rangle\}.$$

Note that $\mu(e)$ is a half-integer. Note also that in the case when Γ is a curve, or equivalently, when $d(q) \equiv 2$, $\mu(e)$ is exactly the integer $\mu(\Gamma, e)$, the number of local maxima of $\langle e, \cdot \rangle$ along Γ as defined in [Milnor 1950, p. 252].

Corollary 3.9. *For almost all $e \in S^2$ and for any parametrization Γ' of the double of Γ, $\mu_\Gamma(e) \le \frac{1}{2}\mu_{\Gamma'}(e)$.*

Proof. We have

$$\mu_\Gamma(e) = \tfrac{1}{2}\sum_q \left[\operatorname{lmax}_{\Gamma'}(e,q) - \operatorname{lmin}_{\Gamma'}(e,q)\right] \le \tfrac{1}{2}\sum_q \operatorname{lmax}_{\Gamma'}(e,q) = \tfrac{1}{2}\mu_{\Gamma'}. \quad \square$$

If, in place of the positive part, we sum $\operatorname{nlm}(e,q)$ itself over q located above a plane orthogonal to e, we find a useful quantity:

Corollary 3.10. *For almost all $s_0 \in \mathbb{R}$ and almost all $e \in S^2$,*

$$2\sum \left\{\operatorname{nlm}(e,q) : \langle e, q\rangle > s_0\right\} = \#(e, s_0),$$

the cardinality of the fiber $\{p \in \Gamma : \langle e, p\rangle = s_0\}$.

Proof. If $s_0 > \max_{p\in\Gamma}\langle e, p\rangle$, then $\#(e, s_0) = 0$. Now proceed downward, using Lemma 3.5 by induction. $\square$

Note that the fiber cardinality of Corollary 3.10 is also the value obtained for curves, where the more general nlm may be replaced by the number of local maxima [Milnor 1950].

Remark 3.11. In analogy with Corollary 3.10, we expect that an appropriate generalization of NTC to curved polyhedral complexes of dimension ≥ 2 will in the future allow computation of the homology of level sets and sublevel sets of a (generalized) Morse function in terms of a generalization of $\operatorname{nlm}(e,q)$.

Corollary 3.12. *The multiplicity of a graph in direction $e \in S^2$ may also be computed as $\mu(e) = \frac{1}{2}\sum_{q\in\Gamma} |\operatorname{nlm}(e,q)|$.*

Proof. It follows from Corollary 3.10 with $s_0 < \min_\Gamma \langle e, \cdot\rangle$ that $\sum_{q\in\Gamma} \operatorname{nlm}(e,q) = 0$, which is the difference of positive and negative parts. The sum of these parts is $\sum_{q\in\Gamma} |\operatorname{nlm}(e,q)| = 2\mu(e)$. $\square$

It was shown in Theorem 3.1 of [Milnor 1950] that, in the case of curves, $\mathcal{C}(\Gamma) = \frac{1}{2}\int_{S^2} \mu(e)\, dA_{S^2}$, where Milnor refers to Crofton's formula. We may now extend this result to *graphs*:

Theorem 3.13. *For a (piecewise C^2) graph Γ mapped into $\mathbb{R}^3$, the net total curvature has the representation*

$$\operatorname{NTC}(\Gamma) = \frac{1}{2}\int_{S^2} \mu(e)\, dA_{S^2}(e).$$

Proof. We have $\mathrm{NTC}(\Gamma) = \sum_{j=1}^{N} \mathrm{ntc}(q_j) + \int_{\Gamma_{\mathrm{reg}}} |\vec{k}|\, ds$, where $q_1, \ldots, q_N$ are the vertices of Γ, including local extrema as vertices of degree $d(q_j) = 2$, and where $\mathrm{ntc}(q) := \frac{1}{4}\int_{S^2} \left[\sum_{i=1}^{d} \chi_i(e)\right]^+ dA_{S^2}(e)$ by the definition (2-3) of $\mathrm{ntc}(q)$. Applying Milnor's result to each C^2 edge, we have $\mathscr{C}(\Gamma_{\mathrm{reg}}) = \frac{1}{2}\int_{S^2} \mu_{\Gamma_{\mathrm{reg}}}(e)\, dA_{S^2}$. But $\mu_\Gamma(e) = \mu_{\Gamma_{\mathrm{reg}}}(e) + \sum_{j=1}^{N} \mathrm{nlm}^+(\mathrm{e}, \mathrm{q_j})$, and the theorem follows. □

Corollary 3.14. *If $f : \Gamma \to \mathbb{R}^3$ is piecewise C^2 but is not an embedding, then the net total curvature* $\mathrm{NTC}(\Gamma)$ *is well defined, using the right-hand side of the conclusion of Theorem 3.13. Moreover,* $\mathrm{NTC}(\Gamma)$ *has the same value when some or all of the points of self-intersection of* Γ *are redefined as vertices.*

For $e \in S^2$, we use the notation $p_e : \mathbb{R}^3 \to e\mathbb{R}$ for the orthogonal projection $\langle e, \cdot \rangle$. We sometimes identify $\mathbb{R}$ with the one-dimensional subspace $e\mathbb{R}$ of $\mathbb{R}^3$.

Corollary 3.15. *If* $\{\Gamma\}$ *is any homeomorphism type of graphs, then the infimum* $\mathrm{NTC}(\{\Gamma\})$ *of net total curvature among mappings $f : \Gamma \to \mathbb{R}^n$ is assumed by a mapping $f_0 : \Gamma \to \mathbb{R}$.*

For any isotopy class $[\Gamma]$ *of embeddings $f : \Gamma \to \mathbb{R}^3$, the infimum* $\mathrm{NTC}([\Gamma])$ *of net total curvature is assumed by a mapping $f_0 : \Gamma \to \mathbb{R}$ in the closure of the given isotopy class.*

Conversely, if $f_0 : \Gamma \to \mathbb{R}$ is in the closure of a given isotopy class $[\Gamma]$ *of embeddings into $\mathbb{R}^3$, then for all $\delta > 0$ there is an embedding $f : \Gamma \to \mathbb{R}^3$ in that isotopy class with* $\mathrm{NTC}(f) \le \mathrm{NTC}(f_0) + \delta$.

Proof. Let $f : \Gamma \to \mathbb{R}^3$ be any piecewise smooth mapping. By Corollary 3.14 and Corollary 3.10, the net total curvature of the projection $p_e \circ f : \Gamma \to \mathbb{R}$ of f onto the line in the direction of almost any $e \in S^2$ is given by $2\pi\,\mu(e) = \pi(\mu(e) + \mu(-e))$. It follows from Theorem 3.13 that $\mathrm{NTC}(\Gamma)$ is the average of $2\pi\,\mu(e)$ over e in S^2. But the half-integer-valued function $\mu(e)$ is lower semicontinuous almost everywhere, as may be seen using Definition 3.1. Let $e_0 \in S^2$ be a point where μ attains its essential infimum. Then $\mathrm{NTC}(\Gamma) \ge \pi\,\mu(e_0) = \mathrm{NTC}(p_{e_0} \circ f)$. But $(p_{e_0} \circ f)e_0$ is the limit as $\varepsilon \to 0$ of the map f_ε whose projection in the direction e_0 is the same as that of f and is multiplied by ε in all orthogonal directions. Since f_ε is isotopic to f, $(p_{e_0} \circ f)e_0$ is in the closure of the isotopy class of f.

Conversely, given $f_0 : \Gamma \to \mathbb{R}$ in the closure of a given isotopy class, let f be an embedding in that isotopy class uniformly close to $f_0\, e_0$; f_ε as constructed above converges uniformly to f_0 as $\varepsilon \to 0$, and $\mathrm{NTC}(f_\varepsilon) \to \mathrm{NTC}(f_0)$. □

Definition 3.16. We call a mapping $f : \Gamma \to \mathbb{R}^n$ *flat* (or NTC-flat) if $\mathrm{NTC}(f) = \mathrm{NTC}(\{\Gamma\})$, the minimum value for the topological type of Γ, among all ambient dimensions n.

Corollary 3.15 above shows that for any Γ, there is a flat mapping $f : \Gamma \to \mathbb{R}$.

Proposition 3.17. *Consider a piecewise C^2 mapping $f_1 : \Gamma \to \mathbb{R}$. There is a mapping $f_0 : \Gamma \to \mathbb{R}$ which is monotonic along the topological edges of Γ, has values at topological vertices of Γ arbitrarily close to those of f_1, and has* $\mathrm{NTC}(f_0) \leq \mathrm{NTC}(f_1)$.

Proof. Any piecewise C^2 mapping $f_1 : \Gamma \to \mathbb{R}$ may be approximated uniformly by mappings with a finite set of local extreme points, using the compactness of Γ. Thus, we may assume without loss of generality that f_1 has only finitely many local extreme points. Note that for a mapping $f : \Gamma \to \mathbb{R} = \mathbb{R}e$, $\mathrm{NTC}(f) = 2\pi\mu(e)$: hence, we only need to compare $\mu_{f_0}(e)$ with $\mu_{f_1}(e)$.

If f_1 is not monotonic on a topological edge E, then it has a local extremum at a point z in the interior of E. For concreteness, we shall assume z is a local maximum point; the case of a local minimum is similar. Write v, w for the endpoints of E. Let v_1 be the closest local minimum point to z on the interval of E from z to v (or $v_1 = v$ if there is no local minimum point between), and let w_1 be the closest local minimum point to z on the interval from z to w (or $w_1 = w$). Let $E_1 \subset E$ denote the interval between v_1 and w_1. Then E_1 is an interval of a topological edge of Γ, having end points v_1 and w_1 and containing an interior point z, such that f_1 is monotone increasing on the interval from v_1 to z, and monotone decreasing on the interval from z to w_1. By switching v_1 and w_1 if needed, we may assume that $f_1(v_1) < f_1(w_1) < f_1(z)$.

Let f_0 equal f_1 except on the interior of the interval E_1, and map E_1 monotonically to the interval of $\mathbb{R}$ between $f_1(v_1)$ and $f_1(w_1)$. Then for $f_1(w_1) < s < f_1(z)$, the cardinality $\#(e, s)_{f_0}$ equals $\#(e, s)_{f_1} - 2$. For s in all other intervals of $\mathbb{R}$, this cardinality is unchanged. Therefore, $\mathrm{nlm}_{f_1}(w_1) = \mathrm{nlm}_{f_0}(w_1) - 1$, by Lemma 3.5. This implies that $\mathrm{nlm}^+_{f_1}(w_1) \geq \mathrm{nlm}^+_{f_0}(w_1) - 1$. Meanwhile, $\mathrm{nlm}_{f_1}(z) = 1$, a term which does not appear in the formula for μ_{f_0} (see Definition 3.8). Thus $\mu_{f_0} \leq \mu_{f_1}$, and $\mathrm{NTC}(f_0) \leq \mathrm{NTC}(f_1)$.

Proceeding inductively, we remove each local extremum in the interior of any edge of Γ, without increasing NTC. □

4. Representation formula for nowhere-smooth graphs

Recall that, while defining the total curvature for continuous graphs in Section 2, we needed the monotonicity of $\mathrm{NTC}(P)$ under refinement of *polygonal* graphs P. We are now ready to prove this.

Proposition 4.1. *Let P and $\widetilde{P}$ be polygonal graphs in $\mathbb{R}^3$, having the same topological vertices, and homeomorphic to each other. Suppose that every vertex of P is also a vertex of $\widetilde{P}$: $\widetilde{P}$ is a refinement of P. Then for almost all $e \in S^2$, the multiplicity $\mu_{\widetilde{P}}(e) \geq \mu_P(e)$. As a consequence,* $\mathrm{NTC}(\widetilde{P}) \geq \mathrm{NTC}(P)$.

Proof. We may assume, as an induction step, that $\widetilde{P}$ is obtained from P by replacing the edge having endpoints q_0, q_2 with two edges, one having endpoints q_0, q_1 and the other having endpoints q_1, q_2. Choose $e \in S^2$. We consider various cases:

If the new vertex q_1 satisfies $\langle e, q_0\rangle < \langle e, q_1\rangle < \langle e, q_2\rangle$, then $\mathrm{nlm}_{\widetilde{P}}(e, q_i) = \mathrm{nlm}_P(e, q_i)$ for $i = 0, 2$ and $\mathrm{nlm}_{\widetilde{P}}(e, q_1) = 0$, hence $\mu_{\widetilde{P}}(e) = \mu_P(e)$.

If $\langle e, q_0\rangle < \langle e, q_2\rangle < \langle e, q_1\rangle$, then $\mathrm{nlm}_{\widetilde{P}}(e, q_0) = \mathrm{nlm}_P(e, q_0)$ and $\mathrm{nlm}_{\widetilde{P}}(e, q_1) = 1$. The vertex q_2 requires more careful counting: the up- and down-degree satisfy $d^{\pm}_{\widetilde{P}}(e, q_2) = d^{\pm}_P(e, q_2) \pm 1$, so that by Lemma 3.5, $\mathrm{nlm}_{\widetilde{P}}(e, q_2) = \mathrm{nlm}_P(e, q_2) - 1$. Meanwhile, for each of the polygonal graphs, $\mu(e)$ is the sum over q of $\mathrm{nlm}^+(e, q)$, so the change from $\mu_P(e)$ to $\mu_{\widetilde{P}}(e)$ depends on the value of $\mathrm{nlm}_P(e, q_2)$:

(a) If $\mathrm{nlm}_P(e, q_2) \le 0$, then $\mathrm{nlm}^{\pm}_{\widetilde{P}}(e, q_2) = \mathrm{nlm}^{+}_P(e, q_2) = 0$.
(b) If $\mathrm{nlm}_P(e, q_2) = \frac{1}{2}$, then $\mathrm{nlm}^{+}_{\widetilde{P}}(e, q_2) = \mathrm{nlm}^{+}_P(e, q_2) - \frac{1}{2}$.
(c) If $\mathrm{nlm}_P(e, q_2) \ge 1$, then $\mathrm{nlm}^{+}_{\widetilde{P}}(e, q_2) = \mathrm{nlm}^{+}_P(e, q_2) - 1$.

Since the new vertex q_1 does not appear in P, recalling that $\mathrm{nlm}_{\widetilde{P}}(e, q_1) = 1$, we have $\mu_{\widetilde{P}}(e) - \mu_P(e) = +1, +\frac{1}{2}$ or 0 in the respective cases (a), (b) or (c). In any case, $\mu_{\widetilde{P}}(e) \ge \mu_P(e)$.

The reverse inequality $\langle e, q_1\rangle < \langle e, q_2\rangle < \langle e, q_0\rangle$ may be reduced to the case just above by replacing $e \in S^2$ with $-e$, since $\mu_P(-e) = -\mu_P(e)$ for any polyhedral graph P. Then, depending whether $\mathrm{nlm}_P(e, q_2)$ is ≤ -1, $= -\frac{1}{2}$ or ≥ 0, we find that $\mu_{\widetilde{P}}(e) - \mu_P(e) = \mathrm{nlm}^{\pm}_{\widetilde{P}}(e, q_2) - \mathrm{nlm}^{+}_P(e, q_2) = 0, \frac{1}{2}$, or 1. In any case, $\mu_{\widetilde{P}}(e) \ge \mu_P(e)$.

These arguments are unchanged if q_0 is switched with q_2. This covers all cases except those in which equality occurs between $\langle e, q_i\rangle$ and $\langle e, q_j\rangle$ $(i \ne j)$. The set of such unit vectors e form a set of measure zero in S^2. The conclusion $\mathrm{NTC}(\widetilde{P}) \ge \mathrm{NTC}(P)$ now follows from Theorem 3.13. □

We remark here that this step of proving the monotonicity for the nowhere-smooth case differs from Milnor's argument for the total curvature of curves, where it was shown by two applications of the triangle inequality for spherical triangles.

Milnor extended his results for piecewise smooth curves to continuous curves in [Milnor 1950]; we shall carry out an analogous extension to continuous graphs.

Definition 4.2. We say a point $q \in \Gamma$ is *critical* relative to $e \in S^2$ when q is a topological vertex of Γ or when $\langle e, \cdot\rangle$ is not monotone in any open interval of Γ containing q.

Note that at some points of a differentiable curve, $\langle e, \cdot\rangle$ may have derivative zero but still not be considered a critical point relative to e by our definition. This is appropriate to the C^0 category. For a continuous graph Γ, when $\mathrm{NTC}(\Gamma)$ is finite, we shall show that the number of critical points is finite for almost all e in S^2 (see Lemma 4.7 below).

Lemma 4.3. *Let Γ be a continuous, finite graph in $\mathbb{R}^3$, and choose a sequence $\widehat{P}_k$ of Γ-approximating polygonal graphs with $\mathrm{NTC}(\Gamma) = \lim_{k\to\infty} \mathrm{NTC}(\widehat{P}_k)$. Then for each $e \in S^2$, there is a refinement P_k of $\widehat{P}_k$ such that $\lim_{k\to\infty} \mu_{P_k}(e)$ exists in $[0, \infty]$.*

Proof. First, for each k in sequence, we refine $\widehat{P}_k$ to include all vertices of $\widehat{P}_{k-1}$. Then for all $e \in S^2$, $\mu_{\widehat{P}_k}(e) \geq \mu_{\widehat{P}_{k-1}}(e)$, by Proposition 4.1. Second, we refine $\widehat{P}_k$ so that the arc of Γ corresponding to each edge of $\widehat{P}_k$ has diameter $\leq 1/k$. Third, given a particular $e \in S^2$, for each edge $\widehat{E}_k$ of $\widehat{P}_k$, we add 0, 1 or 2 points from Γ as vertices of $\widehat{P}_k$ so that $\max_{\widehat{E}_k}\langle e, \cdot\rangle = \max_E \langle e, \cdot\rangle$ where E is the closed arc of Γ corresponding to $\widehat{E}_k$; and similarly so that $\min_{\widehat{E}_k}\langle e, \cdot\rangle = \min_E \langle e, \cdot\rangle$. Write P_k for the result of this three-step refinement. Note that all vertices of P_{k-1} appear among the vertices of P_k. Then by Proposition 4.1,

$$\mathrm{NTC}(\widehat{P}_k) \leq \mathrm{NTC}(P_k) \leq \mathrm{NTC}(\Gamma),$$

so we still have $\mathrm{NTC}(\Gamma) = \lim_{k\to\infty} \mathrm{NTC}(P_k)$.

Now compare the values of $\mu_{P_k}(e) = \sum_{q\in P_k} \mathrm{nlm}_{\mathrm{P_k}}{}^{+}(e, q)$ with the same sum for P_{k-1}. Since P_k is a refinement of P_{k-1}, Proposition 4.1 gives $\mu_{P_k}(e) \geq \mu_{P_{k-1}}(e)$.

Therefore the values $\mu_{P_k}(e)$ are nondecreasing in k, which implies they are either convergent or properly divergent; in the latter case we write $\lim_{k\to\infty} \mu_{P_k}(e) = \infty$. □

Definition 4.4. For a continuous graph Γ, define the *multiplicity* at $e \in S^2$ as $\mu_\Gamma(e) := \lim_{k\to\infty} \mu_{P_k}(e) \in [0, \infty]$, where P_k is a sequence of Γ-approximating polygonal graphs, refined with respect to e, as given in Lemma 4.3.

Remark 4.5. Note that any two Γ-approximating polygonal graphs have a common refinement. Hence, from the proof of Lemma 4.3, any two choices of sequences $\{\widehat{P}_k\}$ of Γ-approximating polygonal graphs lead to the same value $\mu_\Gamma(e)$.

Lemma 4.6. *Let Γ be a continuous, finite graph in $\mathbb{R}^3$. Then $\mu_\Gamma : S^2 \to [0, \infty]$ takes its values in the half-integers, or $+\infty$. Now assume $\mathrm{NTC}(\Gamma) < \infty$. Then μ_Γ is integrable, hence finite almost everywhere on S^2, and*

$$\mathrm{NTC}(\Gamma) = \frac{1}{2}\int_{S^2} \mu_\Gamma(e)\, dA_{S^2}(e). \tag{4-1}$$

For almost all $e \in S^2$, a sequence P_k of Γ-approximating polygonal graphs, converging uniformly to Γ, may be chosen (depending on e) so that each local extreme point q of $\langle e, \cdot\rangle$ along Γ occurs as a vertex of P_k for sufficiently large k.

Proof. Given $e \in S^2$, let $\{P_k\}$ be the sequence of Γ-approximating polygonal graphs from Lemma 4.3. If $\mu_\Gamma(e)$ is finite, then $\mu_{P_k}(e) = \mu_\Gamma(e)$ for k sufficiently large, a half-integer.

Suppose $\mathrm{NTC}(\Gamma) < \infty$. Then the half-integer-valued functions μ_{P_k} are nonnegative, integrable on S^2 with bounded integrals since $\mathrm{NTC}(P_k) < \mathrm{NTC}(\Gamma) < \infty$, and monotone increasing in k. Thus for almost all $e \in S^2$, $\mu_{P_k}(e) = \mu_\Gamma(e)$ for k sufficiently large.

Since the functions μ_{P_k} are nonnegative and pointwise nondecreasing almost everywhere on S^2, it now follows from the monotone convergence theorem that

$$\int_{S^2} \mu_\Gamma(e)\, dA_{S^2}(e) = \lim_{k\to\infty} \int_{S^2} \mu_{P_k}(e)\, dA_{S^2}(e) = 2\mathrm{NTC}(\Gamma).$$

Finally, the polygonal graphs P_k have maximum edge length $\to 0$. For almost all $e \in S^2$, $\langle e, \cdot \rangle$ is not constant along any open arc of Γ, and $\mu_\Gamma(e)$ is finite. Given such an e, choose $l = l(e)$ sufficiently large that $\mu_{P_k}(e) = \mu_\Gamma(e)$ and $\mu_{P_k}(-e) = \mu_\Gamma(-e)$ for all $k \geq l$. Then for $k \geq l$, along any edge E_k of P_k with corresponding arc E of Γ, the maximum and minimum values of $\langle e, \cdot \rangle$ along E occur at the endpoints, which are also the endpoints of E_k. Otherwise, as P_k is further refined, new interior local maximum and local minimum points of E would each contribute a new, positive value to $\mu_{P_k}(e)$ or $\mu_{P_k}(-e)$, respectively, as k increases. Since the diameter of the corresponding arc E of Γ tends to zero as $k \to \infty$, any local maximum or local minimum of $\langle e, \cdot \rangle$ must become an endpoint of some edge of P_k for k sufficiently large, and for $k \geq l$ in particular. □

Our next lemma focuses on the regularity of a graph Γ, originally only assumed continuous, provided it has finite net total curvature, or another notion of total curvature of a graph which includes the total curvature of the edges.

Lemma 4.7. *Let Γ be a continuous, finite graph in $\mathbb{R}^3$, with $\mathrm{NTC}(\Gamma) < \infty$. Then Γ has continuous one-sided unit tangent vectors $T_1(p)$ and $T_2(p)$ at each point p, not a topological vertex. If p is a vertex of degree d, then each of the d edges which meet at p have well-defined unit tangent vectors at p: $T_1(p), \ldots, T_d(p)$. For almost all $e \in S^2$,*

$$\mu_\Gamma(e) = \sum_q \{\mathrm{nlm}(e, q)\}^+, \tag{4-2}$$

where the sum is over the finite number of topological vertices of Γ and critical points q of $\langle e, \cdot \rangle$ along Γ. Further, for each q, $\mathrm{nlm}(e, q) = \frac{1}{2}[d^-(e, q) - d^+(e, q)]$. All of these critical points which are not topological vertices are local extrema of $\langle e, \cdot \rangle$ along Γ.

Proof. We have seen in the proof of Lemma 4.6 that for almost all $e \in S^2$, the linear function $\langle e, \cdot \rangle$ is not constant along any open arc of Γ, and by Lemma 4.3 there is a sequence $\{P_k\}$ of Γ-approximating polygonal graphs with $\mu_\Gamma(e) = \mu_{P_k}(e)$ for k sufficiently large. We have further shown that each local maximum point of $\langle e, \cdot \rangle$ is a vertex of P_k, possibly of degree two, for k large enough. Recall that

$\mu_{P_k}(e) = \sum_q \mathrm{nlm}^+_{P_k}(e, q)$. Thus, each local maximum point q for $\langle e, \cdot \rangle$ along Γ provides a nonnegative term $\mathrm{nlm}^+_{P_k}(e, q)$ in the sum for $\mu_{P_k}(e)$. Fix such an integer k.

Consider a point $q \in \Gamma$ which is not a topological vertex of Γ but is a critical point of $\langle e, \cdot \rangle$. We shall show, by an argument similar to one used in [van Rooij 1965], that q must be a local extreme point. As a first step, we show that $\langle e, \cdot \rangle$ is monotone on a sufficiently small interval on either side of q. Choose an ordering of the closed edge E of Γ containing q, and consider the interval E_+ of points $\geq q$ with respect to this ordering. Suppose that $\langle e, \cdot \rangle$ is not monotone on any subinterval of E_+ with q as endpoint. Then in any interval (q, r_1) there are points $p_2 > q_2 > r_2$ so that the numbers $\langle e, p_2\rangle$, $\langle e, q_2\rangle$, $\langle e, r_2\rangle$ are not monotone. It follows by an induction argument that there exist decreasing sequences $p_n \to q$, $q_n \to q$, and $r_n \to q$ of points of E_+ such that for each n, $r_{n-1} > p_n > q_n > r_n > q$, but the value $\langle e, q_n\rangle$ lies outside of the closed interval between $\langle e, p_n\rangle$ and $\langle e, r_n\rangle$. As a consequence, there is a local extremum $s_n \in (r_n, p_n)$. Since $r_{n-1} > p_n$, the s_n are all distinct, $1 \leq n < \infty$. But by Lemma 4.6, all local extreme points, specifically s_n, of $\langle e, \cdot \rangle$ along Γ occur among the *finite* number of vertices of P_k, a contradiction. This shows that $\langle e, \cdot \rangle$ is monotone on an interval to the right of q. An analogous argument shows that $\langle e, \cdot \rangle$ is monotone on an interval to the left of q.

Recall that for a *critical point* q relative to e, $\langle e, \cdot \rangle$ is not monotone on any neighborhood of q. Since $\langle e, \cdot \rangle$ is monotone on an interval on either side, the sense of monotonicity must be opposite on the two sides of q. Therefore every critical point q along Γ for $\langle e, \cdot \rangle$, which is not a topological vertex, is a local extremum.

We have chosen k large enough that $\mu_\Gamma(e) = \mu_{P_k}(e)$. Then for any edge E_k of P_k, the function $\langle e, \cdot \rangle$ is monotone along the corresponding arc E of Γ, as well as along E_k. Also, E and E_k have common end points. It follows that for each $t \in \mathbb{R}$, the cardinality $\#(e, t)$ of the fiber $\{q \in \Gamma : \langle e, q\rangle = t\}$ is the same for P_k as for Γ. We may see from Lemma 3.5 applied to P_k that for each vertex or critical point q, $\mathrm{nlm}_{P_k}(e, q) = \frac{1}{2}[d^-_{P_k}(e, q) - d^+_{P_k}(e, q)]$; but $\mathrm{nlm}(e, q)$ and $d^\pm(e, q)$ have the *same* values for Γ as for P_k. The formula $\mu_\Gamma(e) = \sum_q \{\mathrm{nlm}_\Gamma(e, q)\}^+$ now follows from the corresponding formula for P_k, for almost all $e \in S^2$.

Consider an open interval E of Γ with endpoint q. We have just shown that for almost all $e \in S^2$, $\langle e, \cdot \rangle$ is monotone on a subinterval with endpoint q. Choose a sequence p_l from E, $p_l \to q$, and write

$$T_l := \frac{p_l - q}{|p_l - q|} \in S^2.$$

Then $\lim_{l\to\infty} T_l$ exists. Otherwise, since S^2 is compact, there are subsequences $\{T_{m_n}\}$ and $\{T_{k_n}\}$ with $T_{m_n} \to T'$ and $T_{k_n} \to T'' \neq T'$. But for an open set of $e \in S^2$,

$\langle e, T' \rangle < 0 < \langle e, T'' \rangle$. For such e, $\langle e, q_{m_n} \rangle < \langle e, q \rangle < \langle e, q_{k_n} \rangle$ for $n \gg 1$. That is, as $p \to q$, $p \in E$, $\langle e, p \rangle$ assumes values above and below $\langle e, q \rangle$ infinitely often, contradicting monotonicity on an interval starting at q for almost all $e \in S^2$.

This shows that Γ has one-sided tangent vectors $T_1(q), \ldots, T_d(q)$ at each point $q \in \Gamma$ of degree $d = d(q)$ ($d = 2$ if q is not a topological vertex). Further, as $k \to \infty$, $T_i^{P_k}(q) \to T_i^{\Gamma}(q)$, $1 \le i \le d(q)$, since edges of P_k have diameter $\le \frac{1}{k}$.

The remaining conclusions follow readily. □

Corollary 4.8. *Let Γ be a continuous, finite graph in $\mathbb{R}^3$, with* $\mathrm{NTC}(\Gamma) < \infty$. *Then for each point q of Γ, the contribution at q to net total curvature is given by* (2-3), *where for $e \in S^2$, $\chi_i(e) =$ the sign of $\langle -T_i(q), e \rangle$, $1 \le i \le d(q)$.* (Here, if q is not a topological vertex, we understand $d = 2$.)

Proof. According to Lemma 4.7, for $1 \le i \le d(q)$, $T_i(q)$ is defined and tangent to an edge E_i of Γ, which is continuously differentiable at its end point q. If P_n is a sequence of Γ-approximating polygonal graphs with maximum edge length tending to 0, the corresponding unit tangent vectors $T_i^{P_n}(q) \to T_i^{\Gamma}(q)$ as $n \to \infty$. For each P_n, we have

$$\mathrm{ntc}^{P_n}(q) = \frac{1}{4} \int_{S^2} \left[\sum_{i=1}^{d} {\chi_i}^{P_n}(e) \right]^+ dA_{S^2}(e),$$

and ${\chi_i}^{P_n} \to {\chi_i}^{\Gamma}$ in measure on S^2. Hence, the integrals for P_n converge to those for Γ, which is (2-3). □

We are ready to state the formula for net total curvature, by localization on S^2, a generalization of Theorem 3.13:

Theorem 4.9. *For a continuous graph Γ, the net total curvature* $\mathrm{NTC}(\Gamma) \in (0, \infty]$ *has the representation*

$$\mathrm{NTC}(\Gamma) = \frac{1}{4} \int_{S^2} \mu(e)\, dA_{S^2}(e),$$

where, for almost all $e \in S^2$, the multiplicity $\mu(e)$ is a positive half-integer or $+\infty$, given as the finite sum (4-2).

Proof. If $\mathrm{NTC}(\Gamma)$ is finite, the theorem follows from Lemma 4.6 and Lemma 4.7.

Choose $e \in S^2$. Suppose $\mathrm{NTC}(\Gamma) = \sup \mathrm{NTC}(P_k)$ is infinite, where P_k is a refined sequence of polygonal graphs as in Lemma 4.3. Then $\mu_\Gamma(e)$ is the nondecreasing limit of $\mu_{P_k}(e)$ for all $e \in S^2$. Thus $\mu_\Gamma(e) \ge \mu_{P_k}(e)$ for all e and k, and $\mu_\Gamma(e) = \mu_{P_k}(e)$ for $k \ge l(e)$. This implies that $\mu_\Gamma(e)$ is a positive half-integer or ∞. Since $\mathrm{NTC}(\Gamma)$ is infinite, the integral

$$\mathrm{NTC}(P_k) = \frac{1}{2} \int_{S^2} \mu_{P_k}(e)\, dA_{S^2}(e)$$

is arbitrarily large as $k \to \infty$, but for each k is less than or equal to

$$\frac{1}{2}\int_{S^2} \mu_\Gamma(e)\, dA_{S^2}(e).$$

Therefore this latter integral equals ∞, and thus equals NTC(Γ). □

We turn our attention next to the tameness of graphs of finite total curvature.

Proposition 4.10. *Let n be a positive integer, and write Z for the set of nth roots of unity in $\mathbb{C} = \mathbb{R}^2$. Given a continuous one-parameter family S_t, $0 \le t < 1$, of sets of n points in $\mathbb{R}^2$, there exists a continuous one-parameter family $\Phi_t : \mathbb{R}^2 \to \mathbb{R}^2$ of homeomorphisms with compact support such that $\Phi_t(S_t) = Z$, $0 \le t < 1$.*

Proof. It is well known that there is an isotopy $\Phi_0 : \mathbb{R}^2 \to \mathbb{R}^2$ such that $\Phi_0(S_0) = Z$ and $\Phi_0 = \mathrm{id}$ outside of a compact set. This completes the case $t_0 = 0$ of the following continuous induction argument.

Suppose that $[0, t_0] \subset [0, 1)$ is a subinterval such that there exists a continuous one-parameter family $\Phi_t : \mathbb{R}^2 \to \mathbb{R}^2$ of homeomorphisms with compact support, with $\Phi_t(S_t) = Z$ for all $0 \le t \le t_0$. We shall extend this property to an interval $[0, t_0+\delta]$. Write $B_\varepsilon(Z)$ for the union of balls $B_\varepsilon(\zeta_i)$ centered at the n roots of unity $\zeta_1, \ldots \zeta_n$. For $\varepsilon < \sin\frac{\pi}{n}$, these balls are disjoint. We may choose $0 < \delta < 1 - t_0$ such that $\Phi_{t_0}(S_t) \subset B_\varepsilon(Z)$ for all $t_0 \le t \le t_0+\delta$. Write the points of S_t as $x_i(t)$, $1 \le i \le n$, where $\Phi_{t_0}(x_i(t)) \in B_\varepsilon(\zeta_i)$. For each $t \in [t_0, t_0+\delta]$, each of the balls $B_\varepsilon(\zeta_i)$ may be mapped onto itself by a homeomorphism ψ_t, varying continuously with t, such that ψ_{t_0} is the identity, ψ_t is the identity near the boundary of $B_\varepsilon(\zeta_i)$ for all $t \in [t_0, t_0+\delta]$, and $\psi_t(\Phi_{t_0}(x_i(t))) = \zeta_i$ for all such t. For example, we may construct ψ_t so that for each $y \in B_\varepsilon(\zeta_i)$, $y - \psi_t(y)$ is parallel to $\Phi_{t_0}(x_i(t)) - \zeta_i$. We now define $\Phi_t = \psi_t \circ \Phi_{t_0}$ for each $t \in [t_0, t_0 + \delta]$.

As a consequence, we see that there is no maximal interval $[0, t_0] \subset [0, 1)$ such that there is a continuous one-parameter family $\Phi_t : \mathbb{R}^2 \to \mathbb{R}^2$ of homeomorphisms with compact support with $\Phi_t(S_t) = Z$, for all $0 \le t \le t_0$. Thus, this property holds for the entire interval $0 \le t < 1$. □

In the following theorem, the total curvature of a graph may be understood in terms of any definition which includes the total curvature of edges and which is continuous as a function of the unit tangent vectors at each vertex. This includes net total curvature, TC of [Taniyama 1998] and CTC of [Gulliver and Yamada 2006].

Theorem 4.11. *Suppose $\Gamma \subset \mathbb{R}^3$ is a continuous graph with finite total curvature. Then for any $\varepsilon > 0$, Γ is isotopic to a Γ-approximating polygonal graph P with edges of length at most ε, whose total curvature is less than or equal to that of Γ.*

Proof. Since Γ has finite total curvature, by Lemma 4.7, at each topological vertex of degree d the edges have well-defined unit tangent vectors $T_1, \ldots, T_d$, which are each the limit as $\varepsilon \to 0$ of the unit tangent vectors to the corresponding edges of P. If at each vertex the unit tangent vectors $T_1, \ldots, T_d$ are distinct, then any sufficiently fine Γ-approximating polygonal graph will be isotopic to Γ; this easier case is proven.

We thus consider n edges $E_1, \ldots, E_n$ ending at a vertex q, with common unit tangent vectors $T_1 = \cdots = T_n$. Choose orthogonal coordinates (x, y, z) for $\mathbb{R}^3$ so that this common tangent vector $T_1 = \cdots = T_n = (0, 0, -1)$ and $q = (0, 0, 1)$. For some $\varepsilon > 0$, in the slab $1 - \varepsilon \leq z \leq 1$, the edges $E_1, \ldots, E_n$ project one-to-one onto the z-axis. After rescaling about q by a factor $\geq 1/\varepsilon$, the edges $E_1, \ldots, E_n$ form a braid B of n strands in the slab $0 \leq z < 1$ of $\mathbb{R}^3$, plus the point $q = (0, 0, 1)$. Each strand E_i has q as an endpoint, and the coordinate z is strictly monotone along E_i, $1 \leq i \leq n$. Write $S_t = B \cap \{z = t\}$. Then S_t is a set of n distinct points in the plane $\{z = t\}$ for each $0 \leq t < 1$. By Proposition 4.10, there are homeomorphisms Φ_t of the plane $\{z = t\}$ for each $0 \leq t < 1$, isotopic to the identity in that plane, continuous as a function of t, such that $\Phi_t(S_t) = Z \times \{t\}$, where Z is the set of n-th roots of unity in the (x, y)-plane, and Φ_t is the identity outside of a compact set of the plane $\{z = t\}$.

We may suppose that S_t lies in the open disk of radius $a(1 - t)$ of the plane $\{z = t\}$, for some (arbitrarily small) constant $a > 0$. We modify Φ_t, first replacing its values with $(1 - t)\Phi_t$ inside the disk of radius $a(1 - t)$. We then modify Φ_t outside the disk of radius $a(1 - t)$, such that Φ_t is the identity outside the disk of radius $2a(1 - t)$.

Having thus modified the homeomorphisms Φ_t of the planes $\{z = t\}$, we may now define an isotopy Φ of $\mathbb{R}^3$ by mapping each plane $\{z = t\}$ to itself by the homeomorphism $\Phi_0^{-1} \circ \Phi_t$, $0 \leq t < 1$; and extend to the remaining planes $\{z = t\}$, $t \geq 1$ and $t < 0$, by the identity. Then the closure of the image of the braid B is the union of line segments from $q = (0, 0, 1)$ to the n points of S_0 in the plane $\{z = 0\}$. Since each Φ_t is isotopic to the identity in the plane $\{z = t\}$, Φ is isotopic to the identity of $\mathbb{R}^3$.

This procedure may be carried out in disjoint sets of $\mathbb{R}^3$ surrounding each unit vector which occurs as tangent vector to more than one edge at a vertex of Γ. Outside these sets, we inscribe a polygonal arc in each edge of Γ to obtain a Γ-approximating polygonal graph P. By Definition 2.3, P has total curvature less than or equal to the total curvature of Γ. □

Artin and Fox [1948] introduced the notion of *tame* and *wild* knots in $\mathbb{R}^3$; the extension to graphs is the following:

Definition 4.12. We say that a graph in $\mathbb{R}^3$ is *tame* if it is isotopic to a polyhedral graph; otherwise, it is *wild*.

Milnor [1950] proved that curves of finite total curvature are tame. More generally, we have

Corollary 4.13. *A continuous graph $\Gamma \subset \mathbb{R}^3$ of finite total curvature is tame.*

Proof. This is an easy consequence of Theorem 4.11, since the Γ-approximating polygonal graph P is isotopic to Γ. □

Observation 4.14. *Tameness does not imply finite total curvature.*

For a well-known example, let $\Gamma \subset \mathbb{R}^2$ be the continuous curve

$$\{(x, h(x)) : x \in [-1, 1]\},$$

where

$$h(x) = -\frac{x}{\pi} \sin \frac{\pi}{x} \quad \text{for } x \neq 0$$

and $h(0) = 0$. This function has a sequence of zeroes $\pm 1/n \to 0$ as $n \to \infty$. The total curvature of Γ between $(0, 1/n)$ and $(0, 1/(n+1))$ converges to π as $n \to \infty$. Thus $\mathcal{C}(\Gamma) = \infty$.

On the other hand, $h(x)$ is continuous on $[-1, 1]$, from which it readily follows that Γ is tame.

5. On vertices of small degree

We will now illustrate some properties of net total curvature $\mathrm{NTC}(\Gamma)$ in a few relatively simple cases, and make some observations regarding $\mathrm{NTC}(\{\Gamma\})$, the minimum net total curvature for the homeomorphism type of a graph $\Gamma \subset \mathbb{R}^n$ (see Definition 3.16 above).

Minimum curvature for given degree.

Proposition 5.1. *If a vertex q has odd degree, then $\mathrm{ntc}(q) \geq \pi/2$. If $d(q) = 3$, then equality holds if and only if the three tangent vectors T_1, T_2, T_3 at q are coplanar but do not lie in any open half-plane. If q has even degree $2m$, then the minimum value of $\mathrm{ntc}(q)$ is 0. Moreover, the equality $\mathrm{ntc}(q) = 0$ only occurs when $T_1(q), \ldots, T_{2m}(q)$ form m opposite pairs.*

Proof. Let q have odd degree $d(q) = 2m+1$. Then from Lemma 3.5, for any $e \in S^2$, we see that $\mathrm{nlm}(e, q)$ is one of the half-integers $\pm\frac{1}{2}, \ldots, \pm\frac{2m+1}{2}$. In particular, $|\mathrm{nlm}(e, q)| \geq \frac{1}{2}$. Corollary 3.7 and the proof of Corollary 3.12 show that

$$\mathrm{ntc}(q) = \frac{1}{4} \int_{S^2} \big|\mathrm{nlm}(e, q)\big| \, dA_{S^2}.$$

Therefore $\mathrm{ntc}(q) \geq \pi/2$.

If the degree $d(q) = 3$, then $|\mathrm{nlm}(e, q)| = \frac{1}{2}$ if and only if both $d^+(q)$ and $d^-(q)$ are nonzero, that is, q is not a local extremum for $\langle e, \cdot \rangle$. If $\mathrm{ntc}(q) = \pi/2$, then

this must be true for almost every direction $e \in S^2$. Thus, the three tangent vectors must be coplanar, and may not lie in an open half-plane.

If $d(q) = 2m$ is even and equality $\mathrm{ntc}(q) = 0$ holds, then the formula above for $\mathrm{ntc}(q)$ in terms of $|\mathrm{nlm}(e, q)|$ would require $\mathrm{nlm}(e, q) \equiv 0$, and hence $d^+(e, q) = d^-(e, q) = m$ for almost all $e \in S^2$: whenever e rotates so that the plane orthogonal to e passes T_i, another tangent vector T_j must cross the plane in the opposite direction, for almost all e, which implies $T_j = -T_i$. □

Observation 5.2. *If a vertex q of odd degree $d(q) = 2p + 1$, has the minimum value* $\mathrm{ntc}(q) = \pi/2$, *and a hyperplane $P \subset \mathbb{R}^n$ contains an even number of the tangent vectors at q, and no others, then these tangent vectors form opposite pairs.*

The proof is seen by fixing any $(n-2)$-dimensional subspace L of P and rotating P by a small positive or negative angle δ to a hyperplane P_δ containing L. Since P_δ must have k of the vectors $T_1, \ldots, T_{2p+1}$ on one side and $k + 1$ on the other side, for some $0 \le k \le p$, by comparing $\delta > 0$ with $\delta < 0$ it follows that exactly half of the tangent vectors in P lie nonstrictly on each side of L. The proof may be continued as in the last paragraph of the proof of Proposition 5.1. In particular, any two independent tangent vectors T_i and T_j share the 2-plane they span with a third, the three vectors not lying in any open half-plane: in fact, the third vector needs to lie in any hyperplane containing T_i and T_j.

For example, a flat $K_{5,1}$ in $\mathbb{R}^3$ must have five straight segments, two being opposite; and the remaining three being coplanar but not in any open half-plane. This includes the case of four coplanar line segments, since the four must be in opposite pairs, and either opposing pair may be considered as coplanar with the fifth segment.

Nonmonotonicity of* NTC *for subgraphs.

Observation 5.3. *If Γ_0 is a subgraph of a graph Γ, then* $\mathrm{NTC}(\Gamma_0)$ *might not be* $\le \mathrm{NTC}(\Gamma)$.

For a simple polyhedral example, we may consider the "butterfly" graph Γ in the plane with six vertices: $q_0^\pm = (0, \pm 1)$, $q_1^\pm = (1, \pm 3)$, and $q_2^\pm = (-1, \pm 3)$. Γ has seven edges: three vertical edges L_0, L_1 and L_2 are the line segments L_i joining q_i^- to q_i^+. Four additional edges are the line segments from $q_0^\pm$ to $q_1^\pm$ and from $q_0^\pm$ to $q_2^\pm$, which form the smaller angle 2α at $q_0^\pm$, where $\tan\alpha = 1/2$, so that $\alpha < \pi/4$.

The subgraph Γ_0 will be Γ minus the interior of L_0. Then $\mathrm{NTC}(\Gamma_0) = \mathscr{C}(\Gamma_0) = 6\pi - 8\alpha$. However, $\mathrm{NTC}(\Gamma) = 4(\pi - \alpha) + 2(\pi/2) = 5\pi - 4\alpha$, which is less than $\mathrm{NTC}(\Gamma_0)$. □

The monotonicity property, which Observation 5.3 shows fails for $\mathrm{NTC}(\Gamma)$, is a virtue of Taniyama's total curvature $\mathrm{TC}(\Gamma)$.

Net total curvature ≠ cone total curvature ≠ Taniyama's total curvature. It is not difficult to construct three unit vectors T_1, T_2, T_3 in $\mathbb{R}^3$ such that the values of $\mathrm{ntc}(q)$, $\mathrm{ctc}(q)$ and $\mathrm{tc}(q)$, with these vectors as the $d(q) = 3$ tangent vectors to a graph at a vertex q, have different values. For example, we may take T_1, T_2 and T_3 to be three unit vectors in a plane, making equal angles $2\pi/3$. According to Proposition 5.1, we have the contribution to net total curvature $\mathrm{ntc}(q) = \pi/2$. But the contribution to cone total curvature is $\mathrm{ctc}(q) = 0$. Namely,

$$\mathrm{ctc}(q) := \sup_{e \in S^2} \sum_{i=1}^{3} \Big(\frac{\pi}{2} - \arccos\langle T_i, e \rangle \Big).$$

In this supremum, we may choose e to be normal to the plane of T_1, T_2 and T_3, and $\mathrm{ctc}(q) = 0$ follows. Meanwhile, $\mathrm{tc}(q)$ is the sum of the exterior angles formed by the three pairs of vectors, each equal to $\pi/3$, so that $\mathrm{tc}(q) = \pi$.

A similar computation for degree d and coplanar vectors making equal angles gives $\mathrm{ctc}(q) = 0$, and $\mathrm{tc}(q) = \frac{\pi}{2} \lfloor \frac{1}{2}(d-1)^2 \rfloor$ (floor function), while $\mathrm{ntc}(q) = \pi/2$ for d odd, $\mathrm{ntc}(q) = 0$ for d even. This example indicates that $\mathrm{tc}(q)$ may be significantly larger than $\mathrm{ntc}(q)$. In fact, we have

Observation 5.4. *If a vertex q of a graph Γ has degree $d = d(q) \geq 2$, then*

$$\mathrm{tc}(q) \geq (d-1)\,\mathrm{ntc}(q).$$

This follows from the definition (2-3) of $\mathrm{ntc}(q)$. Let $T_1, \ldots, T_d$ be the unit tangent vectors at q. The exterior angle between T_i and T_j is

$$\arccos\langle -T_i, T_j \rangle = \frac{1}{4} \int_{S^2} (\chi_i + \chi_j)^+ \, dA_{S^2}.$$

The contribution $\mathrm{tc}(q)$ at q to total curvature $\mathrm{TC}(\Gamma)$ equals the sum of these integrals over all $1 \leq i < j \leq d$. The sum of the integrands is

$$\sum_{1 \leq i < j \leq d} (\chi_i + \chi_j)^+ \geq \Big[\sum_{1 \leq i < j \leq d} (\chi_i + \chi_j) \Big]^+ = (d-1) \Big[\sum_{i=1}^{d} \chi_i \Big]^+.$$

Integrating over S^2 and dividing by 4, we have $\mathrm{tc}(q) \geq (d-1)\mathrm{ntc}(q)$. □

Conditional additivity of net total curvature under taking union. Observation 5.3 shows the failure of monotonicity of NTC for subgraphs due to the cancellation phenomena at each vertex. The following subadditivity statement specifies the necessary and sufficient condition for the additivity of net total curvature under taking union of graphs.

Proposition 5.5. *Given two graphs* Γ_1 *and* $\Gamma_2 \subset \mathbb{R}^n$ *with* $\Gamma_1 \cap \Gamma_2 = \{p_1, \dots, p_N\}$, *the net total curvature of* $\Gamma = \Gamma_1 \cup \Gamma_2$ *obeys the subadditivity law*

$$\begin{aligned}(5\text{-}1)\quad \mathrm{NTC}(\Gamma) &= \mathrm{NTC}(\Gamma_1) + \mathrm{NTC}(\Gamma_2) \\ &\quad + \frac{1}{2}\sum_{j=1}^{N} \int_{S^2} [\mathrm{nlm}^+_{\Gamma}(e, p_j) - \mathrm{nlm}^+_{\Gamma_1}(e, p_j) - \mathrm{nlm}^+_{\Gamma_2}(e, p_j)]\, dA_{S^2} \\ &\le \mathrm{NTC}(\Gamma_1) + \mathrm{NTC}(\Gamma_2).\end{aligned}$$

In particular, *additivity holds if and only if*

$$\mathrm{nlm}_{\Gamma_1}(e, p_j)\, \mathrm{nlm}_{\Gamma_2}(e, p_j) \ge 0$$

for all points p_j *of* $\Gamma_1 \cap \Gamma_2$ *and almost all* $e \in S^2$.

Proof. The edges of Γ and vertices other than $p_1, \dots, p_N$ are edges and vertices of Γ_1 or of Γ_2, so we only need to consider the contribution at the vertices $p_1, \dots, p_N$ to $\mu(e)$ for $e \in S^2$ (see Definition 3.8). The subadditivity follows from the general inequality $(a+b)^+ \le a^+ + b^+$ for any real numbers a and b. Namely, let $a := \mathrm{nlm}_{\Gamma_1}(e, p_j)$ and $b := \mathrm{nlm}_{\Gamma_2}(e, p_j)$, so that $\mathrm{nlm}_{\Gamma}(e, p_j) = a + b$, as follows from Lemma 3.5. Now integrate both sides of the inequality over S^2, sum over $j = 1, \dots, N$ and apply Theorem 3.13.

As for the equality case, suppose that $ab \ge 0$. We then note that either $a > 0$ and $b > 0$, or $a < 0$ and $b < 0$, or $a = 0$, or $b = 0$. In all four cases, we have $a^+ + b^+ = (a+b)^+$. Applied with $a = \mathrm{nlm}_{\Gamma_1}(e, p_j)$ and $b = \mathrm{nlm}_{\Gamma_2}(e, p_j)$, assuming that $\mathrm{nlm}_{\Gamma_1}(e, p_j)\mathrm{nlm}_{\Gamma_2}(e, p_j) \ge 0$ holds for all $j = 1, \dots, N$ and almost all $e \in S^2$, this implies that $\mathrm{NTC}(\Gamma_1 \cup \Gamma_2) = \mathrm{NTC}(\Gamma_1) + \mathrm{NTC}(\Gamma_2)$.

To show that the equality $\mathrm{NTC}(\Gamma_1 \cup \Gamma_2) = \mathrm{NTC}(\Gamma_1) + \mathrm{NTC}(\Gamma_2)$ implies the inequality $\mathrm{nlm}_{\Gamma_1}(e, p_j)\mathrm{nlm}_{\Gamma_2}(e, p_j) \ge 0$ for all $j = 1, \dots, N$ and for almost all $e \in S^2$, we suppose, to the contrary, that there is a set U of positive measure in S^2, such that for some vertex p_j in $\Gamma_1 \cap \Gamma_2$, whenever e is in U, the inequality $ab < 0$ is satisfied, where $a = \mathrm{nlm}_{\Gamma_1}(e, p_j)$ and $b = \mathrm{nlm}_{\Gamma_2}(e, p_j)$. Then for e in U, a and b are of opposite signs. Let U_1 be the part of U where $a < 0 < b$ holds: we may assume U_1 has positive measure, otherwise exchange Γ_1 with Γ_2. On U_1, we have

$$(a+b)^+ < b^+ = a^+ + b^+.$$

Recall that $a + b = \mathrm{nlm}_{\Gamma}(e, p_j)$. Hence the inequality between half-integers

$$\mathrm{nlm}^+_{\Gamma}(e, p_j) < \mathrm{nlm}^+_{\Gamma_1}(e, p_j) + \mathrm{nlm}^+_{\Gamma_2}(e, p_j)$$

is valid on the set U_1, which has positive measure. This, in turn, implies that $\mathrm{NTC}(\Gamma_1 \cup \Gamma_2) < \mathrm{NTC}(\Gamma_1) + \mathrm{NTC}(\Gamma_2)$, contradicting the assumption of equality. □

One-point union of graphs.

Proposition 5.6. *If the graph* Γ *is the one-point union of graphs* Γ_1 *and* Γ_2, *where the points* p_1 *chosen in* Γ_1 *and* p_2 *chosen in* Γ_2 *are not topological vertices*, *then the minimum* NTC *among all mappings is subadditive*, *and the minimum* NTC *minus* 2π *is superadditive*:

$$\mathrm{NTC}(\{\Gamma_1\}) + \mathrm{NTC}(\{\Gamma_2\}) - 2\pi \leq \mathrm{NTC}(\{\Gamma\}) \leq \mathrm{NTC}(\{\Gamma_1\}) + \mathrm{NTC}(\{\Gamma_2\}).$$

Further, *if the points* $p_1 \in \Gamma_1$ *and* $p_2 \in \Gamma_2$ *may appear as extreme points on mappings of minimum* NTC, *then the minimum net total curvature among all mappings*, *minus* 2π, *is additive*:

$$\mathrm{NTC}(\{\Gamma\}) = \mathrm{NTC}(\{\Gamma_1\}) + \mathrm{NTC}(\{\Gamma_2\}) - 2\pi.$$

Proof. Write $p \in \Gamma$ for the identified points $p_1 = p_2 = p$.

Choose flat mappings $f_1 : \Gamma_1 \to \mathbb{R}$ and $f_2 : \Gamma_2 \to \mathbb{R}$, adding constants so that the chosen points $p_1 \in \Gamma_1$ and $p_2 \in \Gamma_2$ have $f_1(p_1) = f_2(p_2) = 0$. Further, by Proposition 3.17, we may assume that f_1 and f_2 are strictly monotone on the edges of Γ_1 and Γ_2 containing p_1 and p_2, respectively. Let $f : \Gamma \to \mathbb{R}$ be defined as f_1 on Γ_1 and as f_2 on Γ_2. Then at the common point of Γ_1 and Γ_2, $f(p) = 0$, and f is continuous. But since f_1 and f_2 are monotone on the edges containing p_1 and p_2, $\mathrm{nlm}_{\Gamma_1}(p_1) = 0 = \mathrm{nlm}_{\Gamma_2}(p_2)$, so we have $\mathrm{NTC}(\{\Gamma\}) \leq \mathrm{NTC}(f) = \mathrm{NTC}(f_1) + \mathrm{NTC}(f_2) = \mathrm{NTC}(\{\Gamma_1\}) + \mathrm{NTC}(\{\Gamma_2\})$ by Proposition 5.5.

Next, for all $g : \Gamma \to \mathbb{R}$, we show that $\mathrm{NTC}(g) \geq \mathrm{NTC}(\{\Gamma_1\}) + \mathrm{NTC}(\{\Gamma_2\}) - 2\pi$. Given g, write g_1, g_2 for the restriction of g to Γ_1, Γ_2. Then

$$\mu_g(e) = \mu_{g_1}(e) - \mathrm{nlm}^+_{g_1}(p_1) + \mu_{g_2}(e) - \mathrm{nlm}^+_{g_2}(p_2) + \mathrm{nlm}^+_g(p).$$

Now for any real numbers a and b, the difference $(a+b)^+ - (a^+ + b^+)$ is equal to $\pm a$, $\pm b$ or 0, depending on the various signs. Let $a = \mathrm{nlm}_{g_1}(p_1)$ and $b = \mathrm{nlm}_{g_2}(p_2)$. Then since p_1 and p_2 are not topological vertices of Γ_1 and Γ_2, respectively, we have $a, b \in \{-1, 0, +1\}$ and $a + b = \mathrm{nlm}_g(p)$ by Lemma 3.5. In any case, we have

$$\mathrm{nlm}^+_g(p) - \mathrm{nlm}^+_{g_1}(p_1) - \mathrm{nlm}^+_{g_2}(p_2) \geq -1.$$

Thus, $\mu_g(e) \geq \mu_{g_1}(e) + \mu_{g_2}(e) - 1$, and multiplying by 2π, $\mathrm{NTC}(g) \geq \mathrm{NTC}(g_1) + \mathrm{NTC}(g_2) - 2\pi \geq \mathrm{NTC}(\{\Gamma_1\}) + \mathrm{NTC}(\{\Gamma_2\}) - 2\pi$.

Finally, assume p_1 and p_2 are extreme points for flat mappings $f_1 : \Gamma_1 \to \mathbb{R}$ and $f_2 : \Gamma_2 \to \mathbb{R}$. We may assume that $f_1(p_1) = 0 = \min f_1(\Gamma_1)$ and $f_2(p_2) = 0 = \max f_2(\Gamma_2)$. Then $\mathrm{nlm}_{f_2}(p_2) = 1$ and $\mathrm{nlm}_{f_1}(p_1) = -1$, and hence using Lemma 3.5, $\mathrm{nlm}_f(p) = 0$. So $\mu_f(e) = \mu_{f_1}(e) - \mathrm{nlm}^+_{f_1}(p_1) + \mu_{f_2}(e) - \mathrm{nlm}^+{}_{f_2}(p_2) + \mathrm{nlm}^+_f(p) = \mu_{f_1}(e) + \mu_{f_2}(e) - 1$. Multiplying by 2π, we have

$$\mathrm{NTC}(\{\Gamma\}) \leq \mathrm{NTC}(f) = \mathrm{NTC}(\{\Gamma_1\}) + \mathrm{NTC}(\{\Gamma_2\}) - 2\pi. \qquad \square$$

6. Net total curvature for degree 3

Simple description of net total curvature.

Proposition 6.1. *For any graph* Γ *and any parametrization* Γ' *of its double*,

$$\mathrm{NTC}(\Gamma) \le \tfrac{1}{2}\mathscr{C}(\Gamma').$$

If Γ *is a trivalent graph, that is, having vertices of degree at most three, then* $\mathrm{NTC}(\Gamma) = \frac{1}{2}\mathscr{C}(\Gamma')$ *for any parametrization* Γ' *that does not immediately repeat any edge of* Γ.

Proof. The first conclusion follows from Corollary 3.9.

Now consider a trivalent graph Γ. Observe that Γ' would be forced to immediately repeat any edge which ends in a vertex of degree 1; thus, we may assume that Γ has only vertices of degree 2 or 3. Since Γ' covers each edge of Γ twice, we need only show, for every vertex q of Γ, having degree $d = d(q) \in \{2, 3\}$, that

$$2\,\mathrm{ntc}_\Gamma(q) = \sum_{i=1}^{d} \mathrm{c}_{\Gamma'}(q_i), \tag{6-1}$$

where $q_1, \ldots, q_d$ are the vertices of Γ' over q. If $d = 2$, since Γ' does not immediately repeat any edge of Γ, we have $\mathrm{ntc}_\Gamma(q) = \mathrm{c}_{\Gamma'}(q_1) = \mathrm{c}_{\Gamma'}(q_2)$, so (6-1) clearly holds. For $d = 3$, write both sides of (6-1) as integrals over S^2, using the definition (2-3) of $\mathrm{ntc}_\Gamma(q)$. Since Γ' does not immediately repeat any edge, the three pairs of tangent vectors $\{T_1^{\Gamma'}(q_j), T_2^{\Gamma'}(q_j)\}$, $1 \le j \le 3$, comprise all three pairs taken from the triple $\{T_1^\Gamma(q), T_2^\Gamma(q), T_3^\Gamma(q)\}$. We need to show that

$$2\int_{S^2} [\chi_1 + \chi_2 + \chi_3]^+ \, dA_{S^2} = \int_{S^2} [\chi_1 + \chi_2]^+ \, dA_{S^2} + \int_{S^2} [\chi_2 + \chi_3]^+ \, dA_{S^2} + \int_{S^2} [\chi_3 + \chi_1]^+ \, dA_{S^2},$$

where at each direction $e \in S^2$, $\chi_j(e) = \pm 1$ is the sign of $\langle -e, T_j^\Gamma(q)\rangle$. But the integrands are equal at almost every point e of S^2:

$$2[\chi_1 + \chi_2 + \chi_3]^+ = [\chi_1 + \chi_2]^+ + [\chi_2 + \chi_3]^+ + [\chi_3 + \chi_1]^+,$$

as may be confirmed by cases: $6 = 6$ if $\chi_1 = \chi_2 = \chi_3 = +1$; $2 = 2$ if exactly one of the χ_i equals -1, and $0 = 0$ in the remaining cases. □

Simple description of net total curvature fails, $d \ge 4$.

Observation 6.2. *We have seen in Proposition 6.1 that for graphs with vertices of degree* ≤ 3, *if a parametrization* Γ' *of the double* $\widetilde{\Gamma}$ *of* Γ *does not immediately repeat any edge of* Γ, *then* $\mathrm{NTC}(\Gamma) = \frac{1}{2}\mathscr{C}(\Gamma')$, *the total curvature in the usual*

sense of the link Γ'. *A natural suggestion would be that for general graphs* $\Gamma \subset \mathbb{R}^3$, $\mathrm{NTC}(\Gamma)$ *might be half the infimum of total curvature of all such parametrizations* Γ' *of the double. However, in some cases, we have the strict inequality* $\mathrm{NTC}(\Gamma) < \inf_{\Gamma'} \frac{1}{2}\mathrm{NTC}(\Gamma')$.

In light of Proposition 6.1, we choose an example of a vertex q of degree four, and consider the local contributions to NTC for $\Gamma = K_{1,4}$ and for Γ', which is the union of four arcs.

Suppose that for a small positive angle α, ($\alpha \leq 1$ radian would suffice) the four unit tangent vectors at q are $T_1 = (1, 0, 0)$; $T_2 = (0, 1, 0)$; $T_3 = (-\cos\alpha, 0, \sin\alpha)$; and $T_4 = (0, -\cos\alpha, -\sin\alpha)$. Write the exterior angles as $\theta_{ij} = \pi - \arccos\langle T_i, T_j\rangle$. Then $\inf_{\Gamma'} \frac{1}{2}\mathcal{C}(\Gamma') = \theta_{13} + \theta_{14} = 2\alpha$. However, $\mathrm{ntc}(q)$ is strictly less than 2α. This may be seen by writing $\mathrm{ntc}(q)$ as an integral over S^2, according to the definition (2-3), and noting that cancellation occurs between two of the four lune-shaped sectors. □

Minimum NTC for trivalent graphs. Using the relation $\mathrm{NTC}(\Gamma) = \frac{1}{2}\mathrm{NTC}(\Gamma')$ between the net total curvature of a given trivalent graph Γ and the total curvature for a nonreversing double cover Γ' of the graph, we can determine the minimum net total curvature of a trivalent graph embedded in $\mathbb{R}^n$, whose value is then related to the Euler characteristic of the graph $\chi(\Gamma) = -k/2$.

First we introduce the following definition.

Definition 6.3. For a given graph Γ and a mapping $f : \Gamma \to \mathbb{R}$, let the *extended bridge number* $B(f)$ be one-half the number of local extrema. Write $B(\{\Gamma\})$ for the minimum of $B(f)$ among all mappings $f : \Gamma \to \mathbb{R}$. For a given isotopy type $[\Gamma]$ of embeddings into $\mathbb{R}^3$, let $B([\Gamma])$ be one-half the minimum number of local extrema for a mapping $f : \Gamma \to \mathbb{R}$ in the closure of the isotopy class $[\Gamma]$.

For an integer $m \geq 3$, let θ_m be the graph with two vertices q^+, q_- and m edges, each of which has q^+ and q^- as its two endpoints. Then $\theta = \theta_3$ has the form of the lower-case Greek letter θ.

Remark 6.4. For a curve, the number of local maxima equals the number of local minima. The minimum number of local maxima is called the *bridge number*, and equals the number of local minima. This is consistent with our Definition 6.3 of the extended bridge number. Of course, for curves, the minimum bridge number among all isotopy classes $B(\{S^1\}) = 1$, and only $B([S^1])$ is of interest for a specific isotopy class $[S^1]$. For certain graphs, the minimum numbers of local maxima and local minima may not occur at the same time for any mapping: see the example of Observation 6.9 below. For isotopy classes of θ-graphs, Goda [1997] has given a definition of an integer-valued bridge index which is similar in spirit to the definition above.

Theorem 6.5. *If Γ is a trivalent graph, and if $f_0 : \Gamma \to \mathbb{R}$ is monotone on topological edges and has the minimum number $2B(\{\Gamma\})$ of local extrema, then*

$$\mathrm{NTC}(f_0) = \mathrm{NTC}(\{\Gamma\}) = \pi\big(2B(\{\Gamma\}) + k/2\big),$$

where k is the number of topological vertices of Γ. For a given isotopy class $[\Gamma]$,

$$\mathrm{NTC}([\Gamma]) = \pi\big(2B([\Gamma]) + k/2\big).$$

Proof. Recall that $\mathrm{NTC}(\{\Gamma\})$ denotes the infimum of $\mathrm{NTC}(f)$ among $f : \Gamma \to \mathbb{R}^3$ or among $f : \Gamma \to \mathbb{R}$, as may be seen from Corollary 3.15.

We first consider a mapping $f_1 : \Gamma \to \mathbb{R}$ with the property that any local maximum or local minimum points of f_1 are interior points of topological edges. Then all topological vertices v, since they have degree $d(v) = 3$ and $d^{\pm}(v) \neq 0$, have $\mathrm{nlm}(v) = \pm 1/2$, by Proposition 5.1. Let Λ be the number of local maximum points of f_1, V the number of local minimum points, λ^+ the number of vertices with $\mathrm{nlm} = +1/2$, and λ^- the number of vertices with $\mathrm{nlm} = -1/2$. Then $\lambda^+ + \lambda^- = k$, the total number of vertices, and $\Lambda + V \geq 2B(\{\Gamma\})$. Hence, applying Corollary 3.12,

$$\mu = \frac{1}{2}\sum_v |\mathrm{nlm}(v)| = \frac{1}{2}\Big(\Lambda + V + \frac{\lambda^+ + \lambda^-}{2}\Big) \geq B(\{\Gamma\}) + k/4, \tag{6-2}$$

with equality if and only if $\Lambda + V = 2B(\{\Gamma\})$.

We next consider any mapping $f_0 : \Gamma \to \mathbb{R}$ in general position: in particular, the critical values of f_0 are isolated. In a similar fashion to the proof of Proposition 3.17, we shall replace f_0 with a mapping whose local extrema are not topological vertices. Specifically, if f_0 assumes a local maximum at any topological vertex v, then, since $d(v) = 3$, $\mathrm{nlm}_{f_0}(v) = 3/2$. f_0 may be isotoped in a small neighborhood of v to $f_1 : \Gamma \to \mathbb{R}$ so that near v, the local maximum occurs at an interior point q of one of the three edges with endpoint v, and thus $\mathrm{nlm}_{f_1}(q) = 1$; while the up-degree $d^+_{f_1}(v) = 1$ and the down-degree $d^-_{f_1}(v) = 2$, so that $\mathrm{nlm}_{f_1}(v)$ is now $\frac{1}{2}$. Thus, $\mu_{f_1}(e) = \mu_{f_0}(e)$. Similarly, if f_0 assumes a local minimum at a topological vertex w, then f_0 may be isotoped in a neighborhood of w to $f_1 : \Gamma \to \mathbb{R}$ so that the local minimum of f_1 near w occurs at an interior point of any of the three edges with endpoint w, and $\mu_{f_1}(e) = \mu_{f_0}(e)$. Then any local extreme points of f_1 are interior points of topological edges. Thus, we have shown that $\mu_{f_0}(e) \geq B(\{\Gamma\}) + k/4$, with equality if f_1 has exactly $2B(\{\Gamma\})$ as its number of local extrema, which holds if and only if f_0 has the minimum number $2B(\{\Gamma\})$ of local extrema. Thus

$$\mathrm{NTC}(\{\Gamma\}) = 2\pi\,\mu_{f_0}(e) = 2\pi\big(B(\{\Gamma\}) + k/4\big) = \pi\big(2B(\{\Gamma\}) + k/2\big).$$

Similarly, for a given isotopy class $[\Gamma]$ of embeddings into $\mathbb{R}^3$, we may choose $f_0 : \Gamma \to \mathbb{R}$ in the closure of the isotopy class, deform f_0 to a mapping f_1 in the

closure of $[\Gamma]$ having no topological vertices as local extrema and count $\mu_{f_0}(e) = \mu_{f_1}(e) \geq B([\Gamma]) + k/4$, with equality if f_0 has the minimum number $2B([\Gamma])$ of local extrema. This shows that $\mathrm{NTC}([\Gamma]) = \pi\big(2B([\Gamma]) + k/2\big)$. □

Remark 6.6. An example geometrically illustrating the lower bound is given by the dual graph Γ^* of the one-skeleton Γ of a triangulation of S^2, with the $\{\infty\}$ not coinciding with any of the vertices of Γ^*. The Koebe–Andreev–Thurston (see [Stephenson 2003]) theorem says that there is a circle packing that realizes the vertex set of Γ^* as the set of centers of the circles. The so realized Γ^*, stereographically projected to $\mathbb{R}^2 \subset \mathbb{R}^3$, attains the lower bound of Theorem 6.5 with $B(\{\Gamma^*\}) = 1$, namely $\mathrm{NTC}([\Gamma]) = \pi(2 + \frac{k}{2}) = \pi(2 - \chi(\Gamma^*))$, where k is the number of vertices.

Corollary 6.7. *If Γ is a trivalent graph with k topological vertices, and $f_0 : \Gamma \to \mathbb{R}$ is a mapping in general position, having Λ local maximum points and V local minimum points, then*

$$\mu_{f_0}(e) = \tfrac{1}{2}(\Lambda + V) + \tfrac{1}{4}k \geq B(\{\Gamma\}) + \tfrac{1}{4}k.$$

Proof. Follows immediately from the proof of Theorem 6.5: f_0 and f_1 have the same number of local maximum or minimum points. □

An interesting trivalent graph is L_m, the "ladder of m rungs" obtained from two unit circles in parallel planes by adding m line segments ("rungs") perpendicular to the planes, each joining one vertex on the first circle to another vertex on the second circle. For example, L_4 is the 1-skeleton of the cube in $\mathbb{R}^3$. Note that L_m may be embedded in $\mathbb{R}^2$, and that the bridge number $B(\{L_m\}) = 1$. Since L_m has $2m$ trivalent vertices, we may apply Theorem 6.5 to compute the minimum NTC for the type of L_m:

Corollary 6.8. *The minimum net total curvature* $\mathrm{NTC}(\{L_m\})$ *for graphs of the type of L_m equals $\pi(2+m)$.*

Observation 6.9. *For certain connected trivalent graphs Γ containing cut points, the minimum extended bridge number $B(\{\Gamma\})$ may be greater than* 1.

Example. Let Γ be the union of three disjoint circles C_1, C_2, C_3 with three edges E_i connecting a point $p_i \in C_i$ with a fourth vertex p_0, which is not in any of the C_i, and which is a *cut point* of Γ: the number of connected components of $\Gamma \setminus p_0$ is greater than for Γ. Given $f : \Gamma \to \mathbb{R}$, after a permutation of $\{1, 2, 3\}$, we may assume there is a minimum point $q_1 \in C_1 \cup E_1$ and a maximum point $q_3 \in C_3 \cup E_3$. If q_1 and q_3 are both in $C_1 \cup E_1$, we may choose C_2 arbitrarily in what follows. Restricted to the closed set $C_2 \cup E_2$, f assumes either a maximum or a minimum at a point $q_2 \neq p_0$. Since $q_2 \neq p_0$, q_2 is also a local maximum or a local minimum for f on Γ. That is, q_1, q_2, q_3 are all local extrema. In the notation of the proof

of Theorem 6.5, we have the number of local extrema $V + \Lambda \geq 3$. Therefore $B(\{\Gamma\}) \geq \frac{3}{2}$, and $\mathrm{NTC}(\{\Gamma\}) \geq \pi(3 + k/2) = 5\pi$.

The reader will be able to construct similar trivalent examples with $B(\{\Gamma\})$ arbitrarily large. □

In contrast to the results of Theorem 6.5 and of Theorem 6.11, below, for trivalent or nearly trivalent graphs, the minimum of NTC for a given graph type cannot be computed merely by counting vertices, but depends in a more subtle way on the topology of the graph:

Observation 6.10. *When* Γ *is not trivalent, the minimum* $\mathrm{NTC}(\{\Gamma\})$ *of net total curvature for a connected graph* Γ *with* $B(\{\Gamma\}) = 1$ *is not determined by the number of vertices and their degrees.*

Example. We shall construct two planar graphs S_m and R_m having the same number of vertices, all of degree 4.

Choose an integer $m \geq 3$ and take the image of the embedding f_ε of the "sine wave" S_m to be the union of the polar-coordinate graphs $C_\pm \subset \mathbb{R}^2$ of two functions: $r = 1 \pm \varepsilon \sin(m\theta)$. S_m has $4m$ edges; and $2m$ vertices, all of degree 4, at $r = 1$ and $\theta = \pi/m, 2\pi/m, \ldots, 2\pi$. For $0 < \varepsilon < 1$, $f_\varepsilon(S_m) = C_+ \cup C_-$ is the union of two smooth cycles. For small positive ε, C_+ and C_- are convex. The $2m$ vertices all have $\mathrm{nlm}(q) = 0$, so

$$\mathrm{NTC}(f_\varepsilon) = \mathrm{NTC}(C_+) + \mathrm{NTC}(C_-) = 2\pi + 2\pi.$$

Therefore $\mathrm{NTC}(\{S_m\}) \leq \mathrm{NTC}(f_\varepsilon) = 4\pi$.

For the other graph type, let the "ring graph" $R_m \subset \mathbb{R}^2$ be constructed by adding m disjoint small circles C_i, each crossing one large circle C at two points v_{2i-1}, v_{2i}, $1 \leq i \leq m$. Then R_m has $4m$ edges. We construct R_m so that the $2m$ vertices $v_1, v_2, \ldots, v_{2m}$, appear in cyclic order around C. Then R_m has the same number $2m$ of vertices as does S_m, all of degree 4. At each vertex v_j, we have $\mathrm{nlm}(v_j) = 0$, so in this embedding, $\mathrm{NTC}(R_m) = 2\pi(m+1)$. We shall show that $\mathrm{NTC}(f_1) \geq 2\pi m$ for any $f_1 : R_m \to \mathbb{R}^3$. According to Corollary 3.15, it is enough to show for every $f : R_m \to \mathbb{R}$ that $\mu_f \geq m$. We may assume f is monotone on each topological edge, according to Proposition 3.17. Depending on the order of $f(v_{2i-2})$, $f(v_{2i-1})$ and $f(v_{2i})$, $\mathrm{nlm}(v_{2i-1})$ might equal ± 1 or ± 2, but cannot be 0, as follows from Lemma 3.5, since the unordered pair $\{d^-(v_{2i-1}), d^+(v_{2i-1})\}$ may only be $\{1, 3\}$ or $\{0, 4\}$. Similarly, v_{2i} is connected by three edges to v_{2i-1} and by one edge to v_{2i+1}. For the same reasons, $\mathrm{nlm}(v_{2i})$ might equal ± 1 or ± 2, and cannot $= 0$. So $|\mathrm{nlm}(v_j)| \geq 1$, $1 \leq j \leq 2m$, and thus by Corollary 3.12, $\mu = \frac{1}{2}\sum_j |\mathrm{nlm}(v_j)| \geq m$. Therefore the minimum of net total curvature $\mathrm{NTC}(\{R_m\}) \geq 2m\pi$, which is greater than $\mathrm{NTC}(\{S_m\}) \leq 4\pi$, since $m \geq 3$.

(A more detailed analysis shows that $\mathrm{NTC}(\{S_m\}) = 4\pi$ and $\mathrm{NTC}(\{R_m\}) = 2\pi(m+1)$.) □

Finally, we may extend the methods of proof for Theorem 6.5 to allow *one* vertex of higher degree:

Theorem 6.11. *If Γ is a graph with one vertex w of degree $d(w) = m \geq 3$, all other vertices being trivalent, and if w shares edges with m distinct trivalent vertices, then $\mathrm{NTC}(\{\Gamma\}) = \pi\big(2B(\{\Gamma\}) + k/2\big)$, where k is the number of vertices of Γ having odd degree. For a given isotopy class $[\Gamma]$, $\mathrm{NTC}([\Gamma]) \geq \pi\big(2B([\Gamma]) + k/2\big)$.*

Proof. Consider any mapping $g : \Gamma \to \mathbb{R}$ in general position. If m is even, then $|\mathrm{nlm}_g(w)| \geq 0$; if m is odd, then $|\mathrm{nlm}_g(w)| \geq \frac{1}{2}$, by Proposition 5.1. If some topological vertex is a local extreme point, then as in the proof of Theorem 6.5, g may be modified without changing $\mathrm{NTC}(g)$ so that all $\Lambda + V$ local extreme points are interior points of edges, with $\mathrm{nlm} = \pm 1$. By Corollary 3.12, we have $\mu_g(e) = \frac{1}{2}\sum |\mathrm{nlm}(v)| \geq \frac{1}{2}\big(\Lambda + V + k/2\big) \geq B(\{\Gamma\}) + k/4$. This shows that

$$\mathrm{NTC}(\{\Gamma\}) \geq \pi\big(2B(\{\Gamma\}) + k/2\big).$$

Now let $f_0 : \Gamma \to \mathbb{R}$ be monotone on topological edges and have the minimum number $2B(\{\Gamma\})$ of local extreme points (see Proposition 3.17). As in the proof of Theorem 6.5, f_0 may be modified without changing $\mathrm{NTC}(f_0)$ so that all $2B(\{\Gamma\})$ local extreme points are interior points of edges. f_0 may be further modified so that the distinct vertices $v_1, \ldots, v_m$ which share edges with w are balanced: $f(v_j) < f(w)$ for half of the $j = 1, \ldots, m$, if m is even, or for half of $m+1$, if m is odd. Having chosen $f(v_j)$, we define f along the (unique) edge from w to v_j to be monotone, for $j = 1, \ldots, m$. Therefore if m is even, then $\mathrm{nlm}_f(w) = 0$; and if m is odd, then $\mathrm{nlm}_f(w) = \frac{1}{2}$, by Lemma 3.5. We compute

$$\mu_f(e) = \tfrac{1}{2}\sum |\mathrm{nlm}(v)| = \tfrac{1}{2}(\Lambda + V + k/2) = B(\{\Gamma\}) + k/4.$$

We conclude that $\mathrm{NTC}(\{\Gamma\}) = \pi\big(2B(\{\Gamma\}) + k/2\big)$.

For a given isotopy class $[\Gamma]$, the proof is analogous to the above. Choose a mapping $g : \Gamma \to \mathbb{R}$ in the closure of $[\Gamma]$, and modify g without leaving the closure of the isotopy class. Choose $f : \Gamma \to \mathbb{R}$ which has the minimum number $2B([\Gamma])$ of local extreme points, and modify it so that topological vertices are not local extreme points. In contrast to the proof of Theorem 6.5, a balanced arrangement of vertices may not be possible in the given isotopy class. In any case, if m is even, then $|\mathrm{nlm}_f(w)| \geq 0$; and if m is odd, $|\mathrm{nlm}_f(w)| \geq \frac{1}{2}$, by Proposition 5.1. Thus applying Corollary 3.12, we find $\mathrm{NTC}([\Gamma]) \geq \pi\big(2B([\Gamma]) + k/2\big)$. □

Observation 6.12. *When all vertices of* Γ *are trivalent except* w, $d(w) \geq 4$, *and when* w *shares more than one edge with another vertex of* Γ, *then in certain cases*, $\mathrm{NTC}(\{\Gamma\}) > \pi\left(2B(\{\Gamma\}) + \frac{k}{2}\right)$, *where* k *is the number of vertices of odd degree.*

Example. Choose Γ to be the one-point union of Γ_1, Γ_2 and Γ_3, where $\Gamma_i = \theta = \theta_3$, $i = 1, 2, 3$, and the point w_i chosen from Γ_i is one of its two vertices v_i, w_i. Then the identified point $w = w_1 = w_2 = w_3$ of Γ has $d(w) = 9$, and each of the other three vertices v_1, v_2, v_3 has degree 3.

Choose a flat map $f : \Gamma \to \mathbb{R}$. We may assume that f is monotone on each edge, applying Proposition 3.17. If $f(v_1) < f(v_2) < f(w) < f(v_3)$, then $d^+(w) = 3$, $d^-(w) = 6$, so $\mathrm{nlm}(w) = \frac{3}{2}$, while v_i is a local extreme point, so $\mathrm{nlm}(v_i) = \pm\frac{3}{2}$, $1 = 1, 2, 3$. This gives $\mu = 3$. The case where $f(v_1) < f(w) < f(v_2) < f(v_3)$ is similar. If w is an extreme point of f, then $\mathrm{nlm}(w) = \pm\frac{9}{2}$ and $\mu \geq \frac{9}{2} > 3$, contradicting flatness of f. This shows that $\mathrm{NTC}(\{\Gamma\}) = \mathrm{NTC}(f) = 6\pi$.

On the other hand, we may show as in Observation 6.9 that $B(\{\Gamma\}) = \frac{3}{2}$. All four vertices have odd degree, so $k = 4$, and $\pi\left(2B(\{\Gamma\}) + k/2\right) = 5\pi$. □

Let W_m denote the "wheel" of m spokes, consisting of a cycle C containing m vertices $v_1, \ldots, v_m$ (the "rim"), a central vertex w (the "hub") not on C, and edges E_i (the "spokes") connecting w to v_i, $1 \leq i \leq m$.

Corollary 6.13. *The minimum net total curvature* $\mathrm{NTC}(\{W_m\})$ *for graphs in* $\mathbb{R}^3$ *homeomorphic to* W_m *equals* $\pi(2 + \lceil m/2 \rceil)$.

Proof. We have one "hub" vertex w with $d(w) = m$, and all other vertices have degree 3. Observe that the bridge number $B(\{W_m\}) = 1$. According to Theorem 6.11, we have $\mathrm{NTC}(\{W_m\}) = \pi\left(2B(\{W_m\}) + k/2\right)$, where k is the number of vertices of odd degree: $k = m$ if m is even, or $k = m + 1$ if m is odd: $k = 2\lceil m/2 \rceil$. Thus $\mathrm{NTC}(\{W_m\}) = \pi\left(2 + \lceil m/2 \rceil\right)$. □

7. Lower bounds of net total curvature

The *width* of an isotopy class $[\Gamma]$ of embeddings of a graph Γ into $\mathbb{R}^3$ is the minimum among representatives of the class of the maximum number of points of the graph meeting a family of parallel planes. More precisely, we write

$$\mathrm{width}([\Gamma]) := \min_{f:\Gamma\to\mathbb{R}^3 \mid f\in[\Gamma]} \min_{e\in S^2} \max_{s\in\mathbb{R}} \#(e, s).$$

For any homeomorphism type $\{\Gamma\}$ define $\mathrm{width}(\{\Gamma\})$ to be the minimum over isotopy types.

Theorem 7.1. *Let* Γ *be a graph, and consider an isotopy class* $[\Gamma]$ *of embeddings* $f : \Gamma \to \mathbb{R}^3$. *Then*

$$\mathrm{NTC}([\Gamma]) \geq \pi\ \mathrm{width}([\Gamma]).$$

As a consequence,

$$\mathrm{NTC}(\{\Gamma\}) \geq \pi \ \mathrm{width}(\{\Gamma\}).$$

Moreover, if for some $e \in S^2$, *an embedding* $f : \Gamma \to \mathbb{R}^3$ *and* $s_0 \in \mathbb{R}$, *the integers* $\#(e, s)$ *are increasing in* s *for* $s < s_0$ *and decreasing for* $s > s_0$, *then* $\mathrm{NTC}([\Gamma]) = \#(e, s_0)\,\pi$.

Proof. Choose an embedding $g : \Gamma \to \mathbb{R}^3$ in the given isotopy class, with

$$\min_{e \in S^2} \max_{s \in \mathbb{R}} \#(e, s) = \mathrm{width}([\Gamma]).$$

There exist $e \in S^2$ and $s_0 \in \mathbb{R}$ with $\#(e, s_0) = \max_{s \in \mathbb{R}} \#(e, s) = \mathrm{width}([\Gamma])$. Replace e if necessary by a nearby point in S^2 so that the values $g(v_i)$, $i = 1, \ldots, m$ are distinct. Next do cylindrical shrinking: without changing $\#(e, s)$ for $s \in \mathbb{R}$, shrink the image of g in directions orthogonal to e by a factor $\delta > 0$ to obtain a family $\{g_\delta\}$ from the same isotopy class $[\Gamma]$, with $\mathrm{NTC}(g_\delta) \to \mathrm{NTC}(g_0)$, where we may identify $g_0 : \Gamma \to \mathbb{R}e \subset \mathbb{R}^3$ with $p_e \circ g = p_e \circ g_\delta : \Gamma \to \mathbb{R}$. But

$$\mathrm{NTC}(p_e \circ g) = \frac{1}{2} \int_{S^2} \mu(u)\, dA_{S^2}(u) = 2\pi\ \mu(e),$$

since for $p_u \circ p_e \circ g$, the local maximum and minimum points are the same as for $p_e \circ g$ if $\langle e, u \rangle > 0$ and reversed if $\langle e, u \rangle < 0$ (recall that $\mu(-e) = \mu(e)$).

We write the topological vertices and the local extrema of g_0 as $v_1, \ldots, v_m$. Let the indexing be chosen so that $g_0(v_i) < g_0(v_{i+1})$, $i = 1, \ldots, m-1$. Now estimate $\mu(e)$ from below: using Lemma 3.5 and Corollary 3.10,

$$\mu(e) = \sum_{i=1}^{m} \mathrm{nlm}^+_{g_0}(e, v_i) \geq \sum_{i=k+1}^{m} \mathrm{nlm}_{g_0}(e, v_i) = \tfrac{1}{2}\#(e, s) \tag{7-1}$$

for any s, $g_0(v_k) < s < g_0(v_{k+1})$. This shows that $\mu(e) \geq \frac{1}{2}\mathrm{width}([\Gamma])$, and therefore

$$\mathrm{NTC}(g) \geq \mathrm{NTC}(g_0) = 2\pi\ \mu(e) \geq \pi\ \mathrm{width}([\Gamma]).$$

Now suppose that the integers $\#(e, s)$ are increasing in s for $s < s_0$ and decreasing for $s > s_0$. Then for $g_0(v_i) > s_0$, we have $\mathrm{nlm}(e, g_0(v_i)) \geq 0$ by Lemma 3.5, and the inequality (7-1) becomes equality at $s = s_0$. □

Lemma 7.2. *For an integer* l, *the minimum width of the complete graph* K_{2l} *on* $2l$ *vertices is* $\mathrm{width}(\{K_{2l}\}) = l^2$; *for* $2l+1$ *vertices*, $\mathrm{width}(\{K_{2l+1}\}) = l(l+1)$.

Proof. Write E_{ij} for the edge of K_m joining v_i to v_j, $1 \leq i < j \leq m$, and suppose $g : K_m \to \mathbb{R}$ has distinct values at the vertices: $g(v_1) < g(v_2) < \cdots < g(v_m)$.

Then for any $g(v_k) < s < g(v_k + 1)$, there are $k(m-k)$ edges E_{ij} with $i \leq k < j$; each of these edges has at least one interior point mapping to s, which shows that $\#(e, s) \geq k(m-k)$. If m is even: $m = 2l$, these lower bounds have the maximum

value l^2 when $k=l$. If m is odd: $m=2l+1$, these lower bounds have the maximum value $l(l+1)$ when $k=l$ or $k=l+1$. This shows that the width of $K_{2l} \geq l^2$ and the width of $K_{2l+1} \geq l(l+1)$. On the other hand, equality holds for the piecewise linear embedding of K_m into $\mathbb{R}$ with vertices in general position and straight edges E_{ij}, which shows that $\text{width}(\{K_{2l}\}) = l^2$ and $\text{width}(\{K_{2l+1}\}) = l(l+1)$. □

Proposition 7.3. *For all $g : K_m \to \mathbb{R}$, $\text{NTC}(g) \geq \pi\, l^2$ if $m = 2l$ is even; and $\text{NTC}(g) \geq \pi\, l(l+1)$ if $m = 2l+1$ is odd. Equality holds for an embedding of K_m into $\mathbb{R}$ with vertices in general position and monotone on each edge; therefore $\text{NTC}(\{K_{2l}\}) = \pi\, l^2$, and $\text{NTC}(\{K_{2l+1}\}) = \pi\, l(l+1)$.*

Proof. The lower bound on $\text{NTC}(\{K_m\})$ follows from Theorem 7.1 and Lemma 7.2.

Now suppose $g : K_m \to \mathbb{R}$ is monotone on each edge, and number the vertices of K_m so that for all i, $g(v_i) < g(v_{i+1})$. Then as in the proof of Lemma 7.2, $\#(e,s) = k(m-k)$ for $g(v_k) < s < g(v_{k+1})$. These cardinalities are increasing for $0 \leq k \leq l$ and decreasing for $l+1 < k < m$. Thus, if $g(v_l) < s_0 < g(v_{l+1})$, then by Theorem 7.1, $\text{NTC}([\Gamma]) = \#(e,s_0)\,\pi = l(m-l)\,\pi$, as claimed. □

Let $K_{m,n}$ be the complete bipartite graph with $m+n$ vertices divided into two sets: v_i, $1 \leq i \leq m$ and w_j, $1 \leq j \leq n$, having one edge E_{ij} joining v_i to w_j, for each $1 \leq i \leq m$ and $1 \leq j \leq n$.

Proposition 7.4. $\text{NTC}(\{K_{m,n}\}) = \lceil mn/2 \rceil\, \pi$.

Proof. $K_{m,n}$ has vertices $v_1, \ldots, v_m$ of degree $d(v_i) = n$ and vertices $w_1, \ldots, w_n$ of degree $d(w_j) = m$. Consider a mapping $g : K_{m,n} \to \mathbb{R}$ in general position, so that the $m+n$ vertices of $K_{m,n}$ have distinct images. We wish to show $\mu(e) = \mu_g(e) \geq mn/4$, if m or n is even, or $(mn+1)/4$, if both m and n are odd.

For this purpose, according to Proposition 3.17, we may first reduce $\mu(e)$ or leave it unchanged by replacing g with a mapping (also called g) which is monotone on each edge E_{ij} of $K_{m,n}$. The values of $\text{nlm}(w_j)$ and of $\text{nlm}(v_i)$ are now determined by the order of the vertex images $g(v_1), \ldots, g(v_m), g(w_1), \ldots, g(w_n)$. Since $K_{m,n}$ is symmetric under permutations of $\{v_1, \ldots, v_m\}$ and permutations of $\{w_1, \ldots, w_n\}$, we shall assume that $g(v_i) < g(v_{i+1})$, $i = 1, \ldots, m-1$ and $g(w_j) < g(w_{j+1})$, $j = 1, \ldots, n-1$. For $i = 1, \ldots, m$ we write k_i for the largest index j such that $g(w_j) < g(v_i)$. Then $0 \leq k_1 \leq \cdots \leq k_m \leq n$, and these integers determine $\mu(e)$. According to Lemma 3.5, $\text{nlm}(v_i) = k_i - n/2$, $i = 1, \ldots, m$. For $j \leq k_1$ and for $j \geq k_m + 1$, we have $\text{nlm}(w_j) = \pm m/2$; for $k_1 < j \leq k_2$ and for $k_{m-1} < j \leq k_m$, we find $\text{nlm}(w_j) = \pm\big(m/2 - 1\big)$; and so on until we find $\text{nlm}(w_j) = 0$ on the middle interval $k_p < j \leq k_{p+1}$, if $m = 2p$ is even; or, if $m = 2p+1$ is odd, $\text{nlm}(w_j) = -\frac{1}{2}$ for $k_p < j \leq k_{p+1}$ and $\text{nlm}(w_j) = +\frac{1}{2}$ for the other middle interval $k_{p+1} < j \leq k_{p+2}$. Thus according to Lemma 3.5 and Corollary 3.12, if $m = 2p$ is

even,

(7-2)

$$\begin{aligned}2\mu(e) &= \sum_{i=1}^{m} \left|\operatorname{nlm}(v_i)\right| + \sum_{j=1}^{n} \left|\operatorname{nlm}(w_j)\right| = \sum_{i=1}^{m} \left|k_i - \frac{n}{2}\right| + (k_1 + n - k_m)\frac{m}{2} \\ &\quad + (k_2 - k_1 + k_m - k_{m-1})\left[\frac{m}{2} - 1\right] + \cdots \\ &\quad + (k_p - k_{p-1} + k_{p+2} - k_{p+1})\left[\frac{m}{2} - (p-1)\right] + (k_{p+1} - k_p)[0] \\ &= \sum_{i=1}^{m} \left|k_i - \frac{n}{2}\right| + \frac{mn}{2} + \sum_{i=1}^{p} k_i - \sum_{i=p+1}^{m} k_i \\ &= \frac{mn}{2} + \sum_{i=1}^{p} \left[\left|k_i - \frac{n}{2}\right| + \left(k_i - \frac{n}{2}\right)\right] + \sum_{i=p+1}^{m} \left[\left|k_i - \frac{n}{2}\right| - \left(k_i - \frac{n}{2}\right)\right].\end{aligned}$$

Note that formula (7-2) assumes its minimum value $2\mu(e) = mn/2$ when

$$k_1 \le \cdots \le k_p \le n/2 \le k_{p+1} \le \cdots \le k_m.$$

If $m = 2p + 1$ is *odd*, then

$$\begin{aligned}2\mu(e) &= \sum_{i=1}^{m} \left|k_i - \frac{n}{2}\right| + (k_1 + n - k_m)\frac{m}{2} + (k_2 - k_1 + k_m - k_{m-1})\left[\frac{m}{2} - 1\right] + \cdots \\ &\quad + (k_{p+3} - k_{p+2})\left[\frac{m}{2} - (p-1)\right] + (k_{p+2} - k_p)\left[\frac{1}{2}\right] = \\ &= \sum_{i=1}^{m} \left|k_i - \frac{n}{2}\right| + \frac{mn}{2} + \sum_{i=1}^{p} k_i - \sum_{i=p+2}^{m} k_i \\ &= \frac{mn}{2} + \sum_{i=1}^{p} \left[\left|k_i - \frac{n}{2}\right| + \left(k_i - \frac{n}{2}\right)\right] \\ &\qquad + \sum_{i=p+2}^{m} \left[\left|k_i - \frac{n}{2}\right| - \left(k_i - \frac{n}{2}\right)\right] + \left|k_{p+1} - \frac{n}{2}\right|.\end{aligned} \tag{7-3}$$

Observe that formula (7-3) has the minimum value $2\mu(e) = \frac{1}{2}mn$ when n is even and $k_1 \le \cdots \le k_p \le \frac{1}{2}n = k_{p+1} \le \cdots \le k_m$. If n as well as m is odd, then the last term $\left|k_{p+1} - \frac{1}{2}n\right|$ is at least $\frac{1}{2}$, and the minimum value of $2\mu(e)$ is $\frac{1}{2}(mn+1)$, attained if and only if $k_1 \le \cdots \le k_p \le \frac{1}{2}n \le k_{p+2} \le \cdots \le k_m$.

This shows that for either parity of m or of n, $\mu(e) \ge \frac{1}{4}mn$. If n and m are both odd, we have the stronger inequality $\mu(e) \ge \frac{1}{4}(mn+1)$. We may summarize these conclusions as $2\mu(e) \ge \left\lceil \frac{1}{2}mn \right\rceil$, and therefore as in the proof of Corollary 3.15, $\operatorname{NTC}(\{K_{m,n}\}) \ge \left\lceil \frac{1}{2}mn \right\rceil \pi$, as we wished to show.

By abuse of notation, write the formula (7-2) or (7-3) as $\mu(k_1, \ldots, k_m)$.

To show the inequality in the opposite direction, we need to find a mapping $f : K_{m,n} \to \mathbb{R}$ with $\mathrm{NTC}(f) = \frac{1}{2}mn\,\pi$ (m or n even) or $\mathrm{NTC}(f) = \frac{1}{2}(mn+1)\,\pi$ (m and n odd). The above computation suggests choosing f with $f(v_1), \ldots, f(v_m)$ together in the middle of the images of the w_j. Write $n = 2l$ if n is even, or $n = 2l + 1$ if n is odd. Choose values $f(w_1) < \cdots < f(w_l) < f(v_1) < \cdots < f(v_m) < f(w_{l+1}) < \cdots < f(w_n)$, and extend f monotonically to each of the mn edges E_{ij}. From formulas (7-2) and (7-3), we have $\mu_f(e) = \mu(l, \ldots, l) = \frac{1}{4}mn$, if m or n is even; or $\mu_f(e) = \mu(l, \ldots, l) = \frac{1}{4}(mn + 1)$, if m and n are odd. □

Recall that θ_m is the graph with two vertices q^+, q_- and m edges.

Corollary 7.5. $\mathrm{NTC}(\{\theta_m\}) = m\,\pi$.

Proof. θ_m is homeomorphic to the complete bipartite graph $K_{m,2}$, and by the proof of Proposition 7.4, we find $\mu(e) \geq \frac{1}{2}m$ for almost all $e \in S^2$, and hence $\mathrm{NTC}(\{K_{m,2}\}) = m\,\pi$. □

8. Fáry–Milnor type isotopy classification

Recall the Fáry–Milnor theorem, which states that if the total curvature of a Jordan curve Γ in $\mathbb{R}^3$ is less than or equal to 4π, then Γ is unknotted. As we have demonstrated above, there are a collection of graphs whose values of the minimum total net curvatures are known. It is natural to hope when the net total curvature is small, in the sense of being in a specific interval to the right of the minimal value, that the isotopy type of the graph is restricted, as is the case for curves: $\Gamma = S^1$. The following proposition and corollaries, however, tell us that results of the Fáry–Milnor type *cannot* be expected to hold for more general graphs.

Proposition 8.1. *If Γ is a graph in $\mathbb{R}^3$ and if $C \subset \Gamma$ is a cycle, such that for some $e \in S^2$, $p_e \circ C$ has at least two local maximum points, then for each positive integer q, there is a nonisotopic embedding $\widetilde{\Gamma}_q$ of Γ in which C is replaced by a knot not isotopic to C, with $\mathrm{NTC}(\widetilde{\Gamma}_q)$ as close as desired to $\mathrm{NTC}(p_e \circ \Gamma)$.*

Proof. It follows from Corollary 3.15 that the one-dimensional graph $p_e \circ \Gamma$ may be replaced by an embedding $\widehat{\Gamma}$ into a small neighborhood of the line $\mathbb{R}e$ in $\mathbb{R}^3$, with arbitrarily small change in its net total curvature. Since $p_e \circ C$ has at least two local maximum points, there is an interval of $\mathbb{R}$ over which $p_e \circ C$ contains an interval which is the image of four oriented intervals J_1, J_2, J_3, J_4 appearing in that cyclic order around the oriented cycle C. Consider a plane presentation of Γ by orthogonal projection into a generic plane containing the line $\mathbb{R}e$. Choose an integer $q \in \mathbb{Z}$, $|q| \geq 3$. We modify $\widehat{\Gamma}$ by wrapping its interval J_1 q times around J_3 and returning, passing over any other edges of Γ, including J_2 and J_4, which it encounters along the way. The new graph in $\mathbb{R}^3$ is called $\widetilde{\Gamma}_q$. Then, if C was the

unknot, the cycle $\widetilde{C}_q$ which has replaced it is a $(2, q)$-torus knot (see [Lickorish 1997]). In any case, $\widetilde{C}_q$ is not isotopic to C, and therefore $\widetilde{\Gamma}_q$ is not isotopic to Γ.

As in the proof of Theorem 7.1, let $g_\delta : \mathbb{R}^3 \to \mathbb{R}^3$ be defined by cylindrical shrinking, so that g_1 is the identity and $g_0 = p_e$. Then $p_e \circ \widetilde{\Gamma}_q = g_0(\widetilde{\Gamma}_q)$, and for $\delta > 0$, $g_\delta(\widetilde{\Gamma}_q)$ is isotopic to $\widetilde{\Gamma}_q$. But $\mathrm{NTC}(g_\delta) \to \mathrm{NTC}(g_0)$ as $\delta \to 0$. □

Corollary 8.2. *If $e = e_0 \in S^{n-1}$ minimizes* $\mathrm{NTC}(p_e \circ \Gamma)$, *and there is a cycle $C \subset \Gamma$ so that $p_{e_0} \circ C$ has two (or more) local maximum points, then there is a sequence of nonisotopic embeddings $\widetilde{\Gamma}_q$ of Γ with* $\mathrm{NTC}(\widetilde{\Gamma}_q)$ *less than, or as close as desired, to* $\mathrm{NTC}(\Gamma)$, *in which C is replaced by its connected sum with a $(2, q)$-torus knot.*

Corollary 8.3. *If Γ is an embedding of K_m into $\mathbb{R}^3$, linear on each topological edge of K_m, $m \geq 4$, then there is a sequence of nonisotopic embeddings $\widetilde{\Gamma}_q$ of Γ with* $\mathrm{NTC}([\widetilde{\Gamma}_q])$ *as close as desired to* $\mathrm{NTC}([\Gamma])$, *in which an unknotted cycle C of Γ is replaced by a $(2, q)$-torus knot.*

Proof. According to Corollary 8.2, we only need to construct an isotopy of K_m with the minimum value of NTC, such that there is a cycle C so that $p_e \circ C$ has two local maximum points, where $\mu(e)$ is a minimum among $e \in S^2$.

Choose $g : K_m \to \mathbb{R}$ which is monotone on each edge of K_m, and has distinct values at vertices. Then according to Proposition 7.3, we have $\mathrm{NTC}(g) = \mathrm{NTC}(\{K_m\})$. Number the vertices $v_1, \ldots, v_m$ so that $g(v_1) < g(v_2) < \cdots < g(v_m)$. Write E_{ji} for the edge E_{ij} with the reverse orientation, $i \neq j$. Then the cycle C formed in sequence from E_{13}, E_{32}, E_{24} and E_{41} has local maximum points at v_3 and v_4, and covers the interval $\Big(g(v_2), g(v_3)\Big) \subset \mathbb{R}$ four times. Since C is formed out of four straight edges, it is unknotted. The procedure of Corollary 8.2 replaces C with a $(2, q)$-torus knot, with an arbitrarily small increase in NTC. □

Note that Corollary 8.2 gives a set of conditions for those graph types where a Fáry–Milnor type isotopy classification might hold. In particular, we consider one of the simpler homeomorphism types of graphs, the *theta graph* $\theta = \theta_3 = K_{3,2}$ (cf. description following Definition 6.3). The *standard theta graph* is the isotopy class in $\mathbb{R}^3$ of a plane circle plus a diameter. We have seen in Corollary 7.5 that the minimum of net total curvature for a theta graph is 3π. On the other hand note that in the range $3\pi \leq \mathrm{NTC}(\Gamma) < 4\pi$, for e in a set of positive measure of S^2, $p_e(\Gamma)$ cannot have two local maximum points. In Theorem 8.5 below, we shall show that a theta graph Γ with $\mathrm{NTC}(\Gamma) < 4\pi$ is isotopically standard.

We may observe that there are nonstandard theta graphs in $\mathbb{R}^3$. For example, the union of two edges might be knotted. Moreover, as S. Kinoshita has shown, there are θ-graphs in $\mathbb{R}^3$, not isotopic to a planar graph, such that each of the three cycles formed by deleting one edge is unknotted [Kinoshita 1972].

We begin with a well-known property of curves, whose proof we give for the sake of completeness.

Lemma 8.4. *Let $C \subset \mathbb{R}^3$ be homeomorphic to S^1, and not a convex planar curve. Then there is a nonempty open set of planes $P \subset \mathbb{R}^3$ which each meet C in at least four points.*

Proof. For $e \in S^2$ and $t \in \mathbb{R}$ write the plane $P_t^e = \{x \in \mathbb{R}^3 : \langle e, x\rangle = t\}$.

If C is not planar, then there exist four noncoplanar points p_1, p_2, p_3, p_4, numbered in order around C. Note that no three of the points can be collinear. Let an oriented plane P_0 be chosen to contain p_1 and p_3 and rotated until both p_2 and p_4 are above P_0 strictly. Write e_1 for the unit normal vector to P_0 on the side where p_2 and p_3 lie, so that $P_0 = P^{e_1}_{t_0=0}$. Then the set $P_t \cap C$ contains at least four points, for $t_0 = 0 < t < \delta_1$, with some $\delta_1 > 0$, since each plane $P_t = P_t^{e_1}$ meets each of the four open arcs between the points p_1, p_2, p_3, p_4. This conclusion remains true, for some $0 < \delta < \delta_1$, when the normal vector e_1 to P_0 is replaced by any nearby $e \in S^2$, and t is replaced by any $0 < t < \delta$.

If C is planar but nonconvex, then there exists a plane $P_0 = P_0^{e_1}$, transverse to the plane containing C, which supports C and touches C at two distinct points, but does not include the arc of C between these two points. Consider disjoint open arcs of C on either side of these two points and including points not in P_0. Then for $0 < t < \delta \ll 1$, the set $P_t \cap C$ contains at least four points, since the planes $P_t = P_t^{e_1}$ meet each of the four disjoint arcs. Here once again e_1 may be replaced by any nearby unit vector e, and the plane P_t^e will meet C in at least four points, for t in a nonempty open interval $t_1 < t < t_1 + \delta$. □

Using the notion of net total curvature, we may extend the theorems of Fenchel [1929] as well as the Fáry–Milnor theorem, for curves homeomorphic to S^1, to graphs homeomorphic to the theta graph. An analogous result is given by Taniyama [1998], who showed that the minimum of TC for polygonal θ-graphs is 4π, and that any θ-graph Γ with $\mathrm{TC}(\Gamma) < 5\pi$ is isotopically standard.

Theorem 8.5. *Suppose $f : \theta \to \mathbb{R}^3$ is a continuous embedding, $\Gamma = f(\theta)$. Then* $\mathrm{NTC}(\Gamma) \geq 3\pi$. *If* $\mathrm{NTC}(\Gamma) < 4\pi$, *then Γ is isotopic in $\mathbb{R}^3$ to the planar theta graph. Moreover,* $\mathrm{NTC}(\Gamma) = 3\pi$ *if and only if the graph is a planar convex curve plus a straight chord.*

Proof. We consider first the case when $f : \theta \to \mathbb{R}^3$ is piecewise C^2.

(1) We have shown the *lower bound* 3π for $\mathrm{NTC}(f)$, where $f : \theta \to \mathbb{R}^n$ is any piecewise C^2 mapping, since $\theta = \theta_3$ is one case of Corollary 7.5, with $m = 3$.

(2) We show next that if there is a cycle C in a graph Γ (a subgraph homeomorphic to S^1) which satisfies the conclusion of Lemma 8.4, then $\mu(e) \geq 2$ for e in a nonempty open set of S^2. Namely, for $t_0 < t < t_0 + \delta$, a family of planes P_t^e meets C, and therefore meets Γ, in at least four points. This is equivalent to saying that the cardinality $\#(e, t) \geq 4$. This implies, by Corollary 3.10, that $\sum\{\mathrm{nlm}(e, q) :$

$p_e(q) > t_0\} \geq 2$. Thus, since $\mathrm{nlm}^+(e,q) \geq \mathrm{nlm}(e,q)$, using Definition 3.8, we have $\mu(e) \geq 2$.

Now consider the *equality* case of a theta graph Γ with $\mathrm{NTC}(\Gamma) = 3\pi$. As we have seen in the proof of Proposition 7.4 with $m = 3$ and $n = 2$, the multiplicity $\mu(e) \geq \frac{3}{2} = \frac{1}{4}mn$ for almost all $e \in S^2$, while the integral of $\mu(e)$ over S^2 equals $2\,\mathrm{NTC}(\Gamma) = 6\pi$ by Theorem 3.13, implying $\mu(e) = 3/2$ almost everywhere on S^2. Thus, the conclusion of Lemma 8.4 is impossible for any cycle C in Γ. By Lemma 8.4, all cycles C of Γ must be planar and convex.

Now Γ consists of three arcs a_1, a_2 and a_3, with common endpoints q^+ and q^-. As we have just shown, the three Jordan curves $\Gamma_1 := a_2 \cup a_3$, $\Gamma_2 := a_3 \cup a_1$ and $\Gamma_3 := a_1 \cup a_2$ are each planar and convex. It follows that Γ_1, Γ_2 and Γ_3 lie in a common plane. In terms of the topology of this plane, one of the three arcs a_1, a_2 and a_3 lies in the middle between the other two. But the middle arc, say a_2, must be a line segment, as it needs to be a shared piece of two curves Γ_1 and Γ_3 bounding disjoint convex open sets in the plane. The conclusion is that Γ is a planar, convex Jordan curve Γ_2, plus a straight chord a_2, whenever $\mathrm{NTC}(\Gamma) = 3\pi$.

(3) We next turn our attention to the *upper bound* of NTC, to imply that a θ-graph is isotopically standard: we shall assume that $g : \theta \to \mathbb{R}^3$ is an embedding in general position with $\mathrm{NTC}(g) < 4\pi$, and write $\Gamma = g(\theta)$. By Theorem 3.13, since S^2 has area 4π, the average of $\mu(e)$ over S^2 is less than 2, and it follows that there exists a set of positive measure of $e_0 \in S^2$ with $\mu(e_0) < 2$. Since $\mu(e_0)$ is a half-integer, and since $\mu(e) \geq \frac{3}{2}$, as we have shown in part (1) of this proof, we have $\mu(e_0) = \frac{3}{2}$ exactly.

From Corollary 6.7 applied to $p_{e_0} \circ g : \theta \to \mathbb{R}$, we find $\mu_g(e_0) = \frac{1}{2}(\Lambda + V) + \frac{k}{4}$, where Λ is the number of local maximum points, V is the number of local minimum points and $k = 2$ is the number of vertices, both of degree 3. Thus, $\frac{3}{2} = \frac{1}{2}(\Lambda + V) + \frac{1}{2}$, so that $\Lambda + V = 2$. This implies that the local maximum/minimum points are unique, and must be the unique global maximum/minimum points $p_{\max}$ and $p_{\min}$ (which may be one of the two vertices $q^{\pm}$). Then $p_{e_0} \circ g$ is monotone along edges except at the points $p_{\max}$, $p_{\min}$ and $q^{\pm}$.

Introduce Euclidean coordinates (x, y, z) for $\mathbb{R}^3$ so that e_0 is in the increasing z-direction. Write $t_{\max} = p_{e_0} \circ g(p_{\max}) = \langle e_0, p_{\max} \rangle$ and $t_{\min} = \langle e_0, p_{\min} \rangle$ for the maximum and minimum values of z along $g(\theta)$. Write $t^{\pm}$ for the value of z at $g(q^{\pm})$, where we may assume $t_{\min} \leq t^- < t^+ \leq t_{\max}$.

We construct a "model" standard θ-curve $\widehat{\Gamma}$ in the (x, z)-plane, as follows. $\widehat{\Gamma}$ will consist of a circle C plus the straight chord of C, joining $\hat{q}^-$ to $\hat{q}^+$ (points to be chosen). Choose C so that the maximum and minimum values of z on C equal $t_{\max}$ and $t_{\min}$. Write $\hat{p}_{\max}$ and $\hat{p}_{\min}$ for the maximum and minimum points of z along C. Choose $\hat{q}^+$ as a point on C where $z = t^+$. There may be two nonequivalent choices

for $\hat{q}^-$ as a point on C where $z = t^-$: we choose so that $\hat{p}_{\max}$ and $\hat{p}_{\min}$ are in the same or different topological edge of $\widehat{\Gamma}$, where $p_{\max}$ and $p_{\min}$ are in the same or different topological edge, respectively, of Γ. Note that there is a homeomorphism from Γ to $\widehat{\Gamma}$ which preserves z.

We now proceed to extend this homeomorphism to an isotopy. For $t \in \mathbb{R}$, write P_t for the plane $\{z = t\}$. As in the proof of Proposition 4.10, there is a continuous 1-parameter family of homeomorphisms $\Phi_t : P_t \to P_t$ such that $\Phi_t(\Gamma \cap P_t) = \widehat{\Gamma} \cap P_t$; Φ_t is the identity outside a compact subset of P_t; and Φ_t is isotopic to the identity of P_t, uniformly with respect to t. Defining $\Phi : \mathbb{R}^3 \to \mathbb{R}^3$ by $\Phi(x, y, z) := \Phi_z(x, y)$, we have an isotopy of Γ with the model graph $\widehat{\Gamma}$.

(4) Finally, consider an embedding $g : \theta \to \mathbb{R}^3$ which is only *continuous*, and write $\Gamma = g(\theta)$.

It follows from Theorem 4.11 that for any θ-graph Γ of finite net total curvature, there is a Γ-approximating polygonal θ-graph P isotopic to Γ, with $\mathrm{NTC}(P) \leq \mathrm{NTC}(\Gamma)$ and as close as desired to $\mathrm{NTC}(\Gamma)$.

If a θ-graph Γ would have $\mathrm{NTC}(\Gamma) < 3\pi$, then the Γ-approximating polygonal graph P would also have $\mathrm{NTC}(P) < 3\pi$, in contradiction to what we have shown for piecewise C^2 theta graphs in part (1) above. This shows that $\mathrm{NTC}(\Gamma) \geq 3\pi$.

If equality $\mathrm{NTC}(\Gamma) = 3\pi$ holds, then $\mathrm{NTC}(P) \leq \mathrm{NTC}(\Gamma) = 3\pi$, so that by the equality case part (2) above, $\mathrm{NTC}(P)$ must equal 3π, and P must be a convex planar curve plus a chord. But this holds for *all* Γ-approximating polygonal graphs P, implying that Γ itself must be a convex planar curve plus a chord.

Finally, If $\mathrm{NTC}(\Gamma) < 4\pi$, then $\mathrm{NTC}(P) < 4\pi$, implying by part (3) above that P is isotopic to the standard θ-graph. But Γ is isotopic to P, and hence is isotopically standard. □

References

[Allard and Almgren 1976] W. K. Allard and F. J. Almgren, Jr., "The structure of stationary one dimensional varifolds with positive density", *Invent. Math.* **34**:2 (1976), 83–97. MR 54 #13694 Zbl 0339.49020

[Douglas 1931] J. Douglas, "Solution of the problem of Plateau", *Trans. Amer. Math. Soc.* **33**:1 (1931), 263–321. MR 1501590 Zbl 0001.14102

[Ekholm et al. 2002] T. Ekholm, B. White, and D. Wienholtz, "Embeddedness of minimal surfaces with total boundary curvature at most 4π", *Ann. of Math.* (2) **155**:1 (2002), 209–234. MR 2003f: 53010 Zbl 1017.53013

[Fáry 1949] I. Fáry, "Sur la courbure totale d'une courbe gauche faisant un nœud", *Bull. Soc. Math. France* **77** (1949), 128–138. MR 11,393h

[Fenchel 1929] W. Fenchel, "Über Krümmung und Windung geschlossener Raumkurven", *Math. Ann.* **101**:1 (1929), 238–252. MR 1512528

[Fox and Artin 1948] R. H. Fox and E. Artin, "Some wild cells and spheres in three-dimensional space", *Ann. of Math.* (2) **49** (1948), 979–990. MR 10,317g Zbl 0033.13602

[Goda 1997] H. Goda, "Bridge index for theta curves in the 3-sphere", *Topology Appl.* **79**:3 (1997), 177–196. MR 98f:57004 Zbl 0888.57003

[Gulliver 2007] R. Gulliver, "Total curvature of graphs in space", *Pure Appl. Math. Q.* **3**:3, part 2 (2007), 773–783. MR 2008k:53005 Zbl 1146.53002

[Gulliver and Yamada 2006] R. Gulliver and S. Yamada, "Area density and regularity for soap film-like surfaces spanning graphs", *Math. Z.* **253**:2 (2006), 315–331. MR 2006m:53012 Zbl 1087.49031

[Gulliver and Yamada 2008] R. Gulliver and S. Yamada, "Total curvature and isotopy of graphs in R^3", preprint, 2008. arXiv 0806.0406

[Kinoshita 1972] S. Kinoshita, "On elementary ideals of polyhedra in the 3-sphere", *Pacific J. Math.* **42** (1972), 89–98. MR 47 #1042 Zbl 0239.55002

[Lickorish 1997] W. B. R. Lickorish, *An introduction to knot theory*, Graduate Texts in Mathematics **175**, Springer, New York, 1997. MR 98f:57015 Zbl 0886.57001

[Milnor 1950] J. W. Milnor, "On the total curvature of knots", *Ann. of Math.* (2) **52** (1950), 248–257. MR 12,273c Zbl 0037.38904

[Radó 1933] T. Radó, *On the problem of Plateau*, Ergebnisse der Math. (2) **2**, 1933. Reprinted Springer, New York, 1971. MR 49 #9718 Zbl 0211.13803

[van Rooij 1965] A. C. M. van Rooij, "The total curvature of curves", *Duke Math. J.* **32** (1965), 313–324. MR 31 #674 Zbl 0171.41702

[Stephenson 2003] K. Stephenson, "Circle packing: a mathematical tale", *Notices Amer. Math. Soc.* **50**:11 (2003), 1376–1388. MR 2004h:52030 Zbl 1047.52016

[Taniyama 1998] K. Taniyama, "Total curvature of graphs in Euclidean spaces", *Differential Geom. Appl.* **8**:2 (1998), 135–155. MR 99m:05050 Zbl 0924.53004

Received April 30, 2011. Revised July 8, 2011.

ROBERT GULLIVER
SCHOOL OF MATHEMATICS
UNIVERSITY OF MINNESOTA
127 VINCENT HALL
206 CHURCH ST. SE
MINNEAPOLIS 55455
UNITED STATES

gulliver@math.umn.edu
http://www.math.umn.edu/∼gulliver

SUMIO YAMADA
MATHEMATICAL INSTITUTE
TOHOKU UNIVERSITY
AOBA
SENDAI 980-8578
JAPAN

yamada@math.tohoku.ac.jp
http://www.math.tohoku.ac.jp/∼yamada

PACIFIC JOURNAL OF MATHEMATICS
Vol. 256, No. 2, 2012

ENTIRE SOLUTIONS OF DONALDSON'S EQUATION

WEIYONG HE

We construct infinitely many special entire solutions to Donaldson's equation. We also prove a Liouville type theorem for entire solutions of Donaldson's equation. We believe that all entire solutions of Donaldson's equation have the form of the examples constructed in the paper.

1. Introduction

Donaldson [2010] introduced an interesting differential operator when he set up a geometric structure for the space of volume forms on compact Riemannian manifolds. The Dirichlet problems for Donaldson's operator are considered in [He 2008; Chen and He 2011]. In this note we shall consider this operator on Euclidean spaces.

For $(t, x) \in \Omega \subset \mathbb{R} \times \mathbb{R}^n$ $(n \geq 1)$, let $u(t, x)$ be a smooth function such that $\triangle u > 0$, $u_{tt} > 0$. We use ∇u, $\triangle u$ to denote derivatives with respect to x and $u_t = \partial_t u$, $u_{tt} = \partial_t^2 u$ to denote derivatives with respect to t. Define a differential operator Q by

$$Q(D^2u) = u_{tt}\triangle u - |\nabla u_t|^2.$$

This operator is strictly elliptic when $u_{tt} > 0$, $\triangle u > 0$ and $Q(D^2u) > 0$. When $n = 1$, then

$$Q(D^2u) = u_{tt}u_{xx} - u_{xt}^2$$

is a real Monge–Ampère operator. When $n = 2$, Q can be viewed as a special case of the complex Monge–Ampère operator. In the x direction, we identify $\mathbb{R}^2 = \mathbb{C}$ with a coordinate w. In the t direction, we take a product by $\mathbb{R}$ with a coordinate s and let $z = t + \sqrt{-1}s$. We extend u on $\mathbb{R} \times \mathbb{R}^2$ to $\mathbb{R}^4 = \mathbb{C}^2$ by $u(z, w) = u(t, x)$. Then

$$Q(D^2u) = 4(u_{z\bar{z}}u_{w\bar{w}} - u_{z\bar{w}}u_{w\bar{z}})$$

is a complex Monge–Ampère operator.

The author is supported in part by a start-up grant of the University of Oregon and by NSF grant DMS 1005392.

MSC2010: 35J60, 35J96.

Keywords: Donaldson's equation, entire solution, Liouville type theorem.

In this paper we shall consider entire solutions $u : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ of

$$Q(D^2u) = 1. \tag{1-1}$$

One celebrated result, proved by Jörgens [1954] in dimension 2 and by Calabi [1958] and Pogorelov [1978] in higher dimensions, is that the only convex solutions of the real Monge–Ampère equation

$$\det(f_{ij}) = 1 \tag{1-2}$$

on the whole of $\mathbb{R}^n$ are the obvious ones: quadratic functions.

Theorem 1.1 (Calabi, Jörgens, Pogorelov). *Let f be a global convex viscosity solution of* (1-2) *on the whole of $\mathbb{R}^n$. Then f has to be a quadratic function.*

One can also ask similar questions for the complex Monge–Ampère equations for plurisubharmonic functions. Let $v : \mathbb{C}^n \to \mathbb{R}$ be a strictly plurisubharmonic function such that $(v_{i\bar{j}}) > 0$, which satisfies

$$\det(v_{i\bar{j}}) = 1. \tag{1-3}$$

The analogous results to Theorem 1.1 for the complex Monge–Ampère equation (1-3) or Donaldson's equation (1-1) ($n > 1$) are not known. For the complex Monge–Ampère equation, LeBrun [1991] investigated the Euclidean Taub–NUT metric constructed by Hawking [1977] and proved that it is a Kähler Ricci-flat metric on $\mathbb{C}^2$ but a nonflat metric. His example provides a nontrivial entire solution of the complex Monge–Ampère equation. We shall construct infinitely many solutions for Donaldson's equation (1-1), which are nontrivial solutions in the sense that u_{tt} is constant, but $\triangle u$, ∇u_t are both not constant. However, when $n = 2$, the Kähler metrics corresponding to these examples are the Euclidean metric on $\mathbb{C}^2$. We shall prove a Liouville type theorem for Donaldson's equation (1-1), which says u_{tt} has to be constant provided some restrictions on u_{tt}. Our proof relies on a transformation introduced by Donaldson [2010]. We then ask if all solutions of (1-1) satisfy that u_{tt} is constant; this would characterize all entire solutions of (1-1) if confirmed.

2. Examples of entire solutions

In this section we shall construct infinitely many nontrivial solutions of (1-1) and (1-3). First we consider (1-1). Let $u_{tt} = 2a$ for some $a > 0$; also let $u(0, x) = g(x)$ and $u_t(0, x) = b(x)$. Then

$$u(t, x) = at^2 + tb(x) + g(x). \tag{2-1}$$

If u solves (1-1), then

$$2a(t\triangle b + \triangle g) - |\nabla b|^2 = 1.$$

It follows that

$$\triangle b = 0 \quad \text{and} \quad \triangle g = \frac{1}{2a}\big(1 + |\nabla b|^2\big).$$

So we shall construct the examples as follows. Let $b = b(x_1, x_2, \ldots, x_n)$ be a harmonic function in $\mathbb{R}^n$. Define

$$h(x) = \frac{1 + |\nabla b|^2}{2a}.$$

Consider the following equation for $g(x)$:

$$\triangle g = h(x). \tag{2-2}$$

We can write $g = b^2(x)/4a + f$ for some function f such that $\triangle f = 1/2a$. We can summarize our results above as follows.

Theorem 2.1. *Let u be the form of* (2-1) *such that b is a harmonic function and g satisfies* (2-2). *Then u is an entire solution of* (1-1). *Moreover, any entire solution of* (1-1) *with $u_{tt} =$ constant has the form of* (2-1).

When $n = 2$, these examples also provide solutions of the complex Monge–Ampère equation (1-3). Actually, let $u(z, w) : \mathbb{C}^2 \to \mathbb{R}$ be a solution of (1-3). If $u_{z\bar{z}} = a$ for some constant $a > 0$, it is not hard to derive that

$$u(z, w) = az\bar{z} + f(z, \bar{z}) + zb(w, \overline{w}) + \bar{z}\bar{b}(w, \overline{w}) + g(w, \overline{w}) \tag{2-3}$$

such that

$$\frac{\partial^2 f}{\partial z\, \partial \bar{z}} = \frac{\partial^2 b}{\partial w\, \partial \overline{w}} = 0 \quad \text{and} \quad \frac{\partial^2 g}{\partial w\, \partial \overline{w}} = \frac{1}{a}\left(1 + \left|\frac{\partial b}{\partial \overline{w}}\right|^2\right).$$

However these examples are all trivial solutions of the complex Monge–Ampère equation in the sense that the corresponding Kähler metrics are flat. For simplicity, we can assume $a = 1$. Since $\partial^2 b/\partial w\, \partial \overline{w} = 0$, we can assume that b is holomorphic or antiholomorphic. If b is holomorphic, then the corresponding Kähler metric is just $dz \otimes d\bar{z} + dw \otimes d\overline{w}$. If b is antiholomorphic, we can set $b(w, \overline{w}) = c(\overline{w})$ and $\bar{b}(w, \overline{w}) = c(w)$. The corresponding Kähler metric is given by

$$dz \otimes d\bar{z} + c_{\overline{w}}\, dz \otimes d\overline{w} + c_w\, d\bar{z} \otimes dw + g_{w\overline{w}}\, dw \otimes d\overline{w}$$
$$= d(z + c(w)) \otimes d(\bar{z} + c(\overline{w})) + dw \otimes d\overline{w}.$$

Then under the holomorphic transformation $(z, w) \to (z + c(w), w)$ it is clear that the Kähler metric is actually flat.

3. A theorem of Liouville type

In this section we shall prove a Liouville type result for solutions of (1-1). We shall describe a transformation introduced by Donaldson [2010], which relates the solutions of (1-1) with harmonic functions. Using this transformation, Theorem 3.1 follows from the standard Liouville theorem for positive harmonic functions.

Theorem 3.1. *Let u be a solution of* (1-1) *with $u_{tt} > 0$. For any $x \in \mathbb{R}^n$, if $u_{tt}(t,x)\,dt^2$ defines a complete metric on $\mathbb{R}\times\{x\}$, then u_{tt} is constant. In particular, it has the form of* (2-1) *such that b is a harmonic function and g satisfies* (2-2)*.*

Proof. For any x fixed, let $z = u_t(t,x)$. Then $\Phi : (t,x) \to (z,x)$ gives a transformation since $u_{tt} > 0$ and the Jacobian of Φ is always positive. In particular, $\Phi : \mathbb{R}\times\mathbb{R}^n \to \text{Image}\,\Phi \subset \mathbb{R}\times\mathbb{R}^n$ is a diffeomorphism. When $u_{tt}(x,t)\,dt^2$ is a complete metric on $\mathbb{R}\times\{x\}$ for all x, then Image $\Phi = \mathbb{R}\times\mathbb{R}^n$. To see this, we note that for any x fixed, then

$$z(t,x) = u_t(0,x) + \int_0^t u_{ss}(s,x)\,ds.$$

Hence if $u_{tt}(t,x)\,dt^2$ is complete, the map $z : t \to z(t,x)$ satisfies $z(\mathbb{R}) = \mathbb{R}$. For x fixed, there exists a unique $t = t(z,x)$ such that $z = u_t(t,x)$. Define a function $\theta(z,x) = t(z,x)$. We claim that θ is a harmonic function in $\mathbb{R}\times\mathbb{R}^n$. The identity $z = u_t(\theta,x)$ implies

$$\frac{\partial\theta}{\partial x_i}u_{tt} + u_{tx_i} = 0 \quad \text{and} \quad u_{tt}\frac{\partial\theta}{\partial z} = 1.$$

It then follows that

$$u_{tt}\frac{\partial^2\theta}{\partial x_i^2} + 2u_{ttx_i}\frac{\partial\theta}{\partial x_i} + u_{ttt}\left(\frac{\partial\theta}{\partial x_i}\right)^2 + u_{tx_ix_i} = 0 \quad \text{and} \quad u_{tt}\frac{\partial^2\theta}{\partial z^2} + \frac{u_{ttt}}{u_{tt}^2} = 0.$$

We compute, if u solves (1-1),

$$\begin{aligned}
\triangle_{(z,x)}\theta &= \frac{\partial^2\theta}{\partial z^2} + \sum_i \frac{\partial^2\theta}{\partial x_i^2} \\
&= \frac{1}{u_{tt}}\left(-\frac{u_{ttt}}{u_{tt}^2} - \triangle u_t + 2\sum_i \frac{u_{ttx_i}u_{tx_i}}{u_{tt}} - \sum_i \frac{u_{ttt}u_{tx_i}^2}{u_{tt}^2}\right) \\
&= \frac{1}{u_{tt}}\left(-\frac{u_{ttt}}{u_{tt}^2}\left(1 + \sum u_{tx_i}^2\right) - \triangle u_t + 2\sum_i \frac{u_{ttx_i}u_{tx_i}}{u_{tt}}\right) \\
&= \frac{-1}{u_{tt}}\left(\frac{u_{ttt}\triangle u}{u_{tt}} + \triangle u_t - 2\sum_i \frac{u_{ttx_i}u_{tx_i}}{u_{tt}}\right) = \frac{-1}{u_{tt}^2}\partial_t\left(\triangle u u_{tt} - |\nabla u_t|^2\right) = 0.
\end{aligned}$$

On the other hand, $\partial\theta/\partial z = 1/u_{tt} > 0$. Hence $\partial\theta/\partial z$ is a positive harmonic function on $\mathbb{R}\times\mathbb{R}^n$. It follows that $\partial\theta/\partial z$ is constant, and so u_{tt} is constant. □

One could classify all solutions of (1-1) if one could prove that u_{tt} does not decay too fast to zero when $|t|\to\infty$, such that $u_{tt}\,dt^2$ defines a complete metric on a line. This motivates the following:

Problem 3.2. *Do all solutions of* (1-1) *with* $u_{tt} > 0$ *satisfy* $u_{tt} = constant$?

Acknowledgement

I am grateful to Professors Xiuxiong Chen and Jingyi Chen for constant support and encouragements. I benefitted from conversations with Professor Pengfei Guan about the complex Monge–Ampère equations and I would like to thank him. I also thank the referee for valuable comments and suggestions.

References

[Calabi 1958] E. Calabi, "Improper affine hyperspheres of convex type and a generalization of a theorem by K. Jörgens", *Michigan Math. J.* **5** (1958), 105–126. MR 21 #5219 Zbl 0113.30104

[Chen and He 2011] X. Chen and W. He, "The space of volume forms", *Int. Math. Res. Not.* **2011**:5 (2011), 967–1009. MR 2012d:58016 Zbl 1218.58008

[Donaldson 2010] S. K. Donaldson, "Nahm's equations and free-boundary problems", pp. 71–91 in *The many facets of geometry*, edited by O. García-Prada et al., Oxford Univ. Press, Oxford, 2010. MR 2011h:58023 Zbl 1211.58015

[Hawking 1977] S. W. Hawking, "Gravitational instantons", *Phys. Lett. A* **60**:2 (1977), 81–83. MR 57 #4965

[He 2008] W. He, "The Donaldson equation", preprint, 2008. arXiv 0810.4123

[Jörgens 1954] K. Jörgens, "Über die Lösungen der Differentialgleichung $rt - s^2 = 1$", *Math. Ann.* **127** (1954), 130–134. MR 15,961e Zbl 0055.08404

[LeBrun 1991] C. LeBrun, "Complete Ricci-flat Kähler metrics on $\mathbf{C}^n$ need not be flat", pp. 297–304 in *Several complex variables and complex geometry, Part 2* (Santa Cruz, CA, 1989), edited by E. Bedford et al., Proc. Sympos. Pure Math. **52**, Amer. Math. Soc., Providence, RI, 1991. MR 93a:53038 Zbl 0739.53053

[Pogorelov 1978] A. V. Pogorelov, *The Minkowski multidimensional problem*, Scripta Series in Mathematics **5**, V. H. Winston & Sons, Washington, D.C., 1978. Translated from the Russian by Vladimir Oliker. MR 57 #17572 Zbl 0387.53023

Received November 1, 2011. Revised January 27, 2012.

WEIYONG HE
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OREGON
EUGENE 97403
OREGON
UNITED STATES
whe@uoregon.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 256, No. 2, 2012

ENERGY IDENTITY AND REMOVABLE SINGULARITIES OF MAPS FROM A RIEMANN SURFACE WITH TENSION FIELD UNBOUNDED IN L^2

YONG LUO

We prove removable singularity results for maps with bounded energy from the unit disk B of $\mathbb{R}^2$ centered at the origin to a closed Riemannian manifold whose tension field is unbounded in $L^2(B)$ but satisfies the following condition:

$$\left(\int_{B_t\setminus B_{t/2}} |\tau(u)|^2\right)^{\frac{1}{2}} \le C_1\left(\frac{1}{t}\right)^a$$

for some $0 < a < 1$ and any $0 < t < 1$, where C_1 is a constant independent of t.

We will also prove that if a sequence $\{u_n\}$ has uniformly bounded energy and satisfies

$$\left(\int_{B_t\setminus B_{t/2}} |\tau(u_n)|^2\right)^{\frac{1}{2}} \le C_2\left(\frac{1}{t}\right)^a$$

for some $0 < a < 1$ and any $0 < t < 1$, where C_2 is a constant independent of n and t, then the energy identity holds for this sequence and there will be no neck formation during the blow up process.

1. Introduction

Let (M, g) be a Riemannian manifold and (N, h) a Riemannian manifold without boundary. For a $W^{1,2}(M, N)$ map u, the energy density of u is defined by

$$e(u) = \tfrac{1}{2}|\nabla u|^2 = \mathrm{Tr}_g(u^*h),$$

where u^*h is the pullback of the metric tensor h.

The energy functional of the mapping u is defined as

$$E(u) = \int_M e(u)\, dV.$$

The author is supported by the DFG collaborative Research Center SFB/Transregio 71.
MSC2010: 35B44.
Keywords: harmonic maps, energy identity.

A map $u \in C^1(M, N)$ is called a harmonic map if it is a critical point of the energy.

By the Nash embedding theorem, N can be isometrically embedded into a Euclidean space $\mathbb{R}^K$ for some positive integer K. Then (N, h) can be viewed as a submanifold of $\mathbb{R}^K$, and a map $u \in W^{1,2}(M, N)$ is a map in $W^{1,2}(M, \mathbb{R}^K)$ whose image lies on N. The space $C^1(M, N)$ should be understood in the same way. In this sense we have the following Euler–Lagrangian equation for harmonic maps.

$$\Delta u = A(u)(\nabla u, \nabla u).$$

The tension field of a map u, $\tau(u)$, is defined by

$$\tau(u) = \Delta u - A(u)(\nabla u, \nabla u),$$

where A is the second fundamental form of N in $\mathbb{R}^K$. So u is a harmonic map if and only if $\tau(u) = 0$.

Notice that, when M is a Riemann surface, the functional $E(u)$ is conformal invariant. Harmonic maps are of special interest in this case. Consider a harmonic map u from a Riemann surface M to N. Recall that Sacks and Uhlenbeck, in a fundamental paper [1981], established the well-known removable singularity theorem by using a class of piecewise smooth harmonic functions to approximate the weak harmonic map. Li and Wang [2010] gave a slightly different proof of the following removable singularity theorem.

Theorem 1.1 [Li and Wang 2010]. *Let B be the unit disk in $\mathbb{R}^2$ centered at the origin. If $u : B \setminus \{0\} \to N$ is a $W^{2,2}_{\mathrm{loc}}(B \setminus \{0\}, N) \cap W^{1,2}(B, N)$ map and u satisfies*

$$\tau(u) = g \in L^2(B, \mathbb{R}^K),$$

then u can be extended to a map belonging to $W^{2,2}(B, N)$.

In this direction we will prove the following result:

Proposition 1.2. *Let B be the unit disk in $\mathbb{R}^2$ centered at the origin. If*

$$u : B \setminus \{0\} \to N$$

is a $W^{2,2}_{\mathrm{loc}}(B \setminus \{0\}, N) \cap W^{1,2}(B, N)$ map and u satisfies

$$\left(\int_{B_t \setminus B_{t/2}} |\tau(u)|^2 \right)^{\frac{1}{2}} \leq C \left(\frac{1}{t} \right)^a$$

for some $0 < a < 1$ and any $0 < t < 1$, where C is a constant independent of t, then there exists some $s > 1$ such that

$$\nabla u \in L^{2s}(B).$$

A direct corollary of this result is the following removable singularity theorem:

Theorem 1.3. *Assume that* $u \in W^{2,2}_{\mathrm{loc}}(B \setminus \{0\}, N) \cap W^{1,2}(B, N)$ *and* u *satisfies*

$$\left(\int_{B_t \setminus B_{t/2}} |\tau(u)|^2\right)^{\frac{1}{2}} \leq C\left(\frac{1}{t}\right)^a$$

for some $0 < a < 1$ *and any* $0 < t < 1$, *where* C *is a constant independent of* t. *Then we have*

$$u \in \bigcap_{1<p<\frac{2}{1+a}} W^{2,p}(B, N).$$

Consider a sequence of maps $\{u_n\}$ from a Riemann surface M to N with uniformly bounded energy. Clearly $\{u_n\}$ converges to u weakly in $W^{1,2}(M, N)$ for some $u \in W^{1,2}(M, N)$, but in general it may not converge strongly in $W^{1,2}(M, N)$ to u, and the falling of the strong convergence is due to the energy concentration at finite points. Jost [1987] and Parker [1996] independently proved that, when $\tau(u_n) = 0$, that is, u_n are harmonic maps, the lost energy is exactly the sum of the energy of the bubbles. Recall that Sacks and Uhlenbeck [1981] proved that the bubbles for such a sequence are harmonic spheres defined as nontrivial harmonic maps from S^2 to N. This result is called energy identity. Furthermore they proved that there is no neck formation during the blow up process, that is, the bubble tree convergence holds true.

For the case when $\tau(u_n)$ is bounded in L^2, that is, $\{u_n\}$ is an approximated harmonic map sequence, the energy identity was proved for N is a sphere by Qing [1995], and for the general target manifold N by Ding and Tian [1995] and, independently, by Wang [1996]. Qing and Tian [1997] proved that there is no neck formation during the blow up process; see also [Lin and Wang 1998]. For the heat flow of harmonic maps, related results can also be found in [Topping 2004a; 2004b]. For the case where the target manifold is a sphere, the energy identity and bubble tree convergence were proved by Lin and Wang [2002] for sequences with tension fields uniform bounded in L^p, for any $p > 1$. In fact, they proved this result under a scaling invariant condition which can be deduced from the uniform boundness of the tension field in L^p.

By virtue of Fanghua Lin and Changyou Wang's result, it is natural to ask the following question.

Question. Let $\{u_n\}$ be a sequence from a closed Riemann surface to a closed Riemannian manifold with tension field uniformly bounded in L^p for some $p > 1$. Do energy identity and bubble tree convergence results hold true during blowing up for such a sequence?

Remark 1.4. Parker [1996] constructed a sequence from a Riemann surface whose tension field is uniformly bounded in L^1, in which the energy identity fails.

Theorem 1.5 [Li and Zhu 2010]. *Let $\{u_n\}$ be a sequence of maps from B to N in $W^{1,2}(B, N)$ with tension field $\tau(u_n)$, where B is the unit disk of $\mathbb{R}^2$ centered at the origin. If*

(I) $\|u_n\|_{W^{1,2}(B)}+\|\tau(u_n)\|_{W^{1,p}(B)} \le \Lambda$ *for some* $p \ge \frac{6}{5}$, *and*

(II) $u_n \to u$ *strongly in* $W^{1,2}_{\mathrm{loc}}(B \setminus \{0\}, N)$ *as* $n \to \infty$,

there exists a subsequence of $\{u_n\}$ (still denoted by $\{u_n\}$) and some nonnegative integer k such that, for any $i = 1, \dots, k$, there are some points x_n^i, positive numbers r_n^i, and a nonconstant harmonic sphere ω^i (viewed as a map from $\mathbb{R}^2 \cup \{\infty\} \to N$) such that:

(1) $x_n^i \to 0$ *and* $r_n^i \to 0$ *as* $n \to \infty$;

(2) $\lim_{n\to\infty}\left(\frac{r_n^i}{r_n^j}+\frac{r_n^j}{r_n^i}+\frac{|x_n^i - x_n^j|}{r_n^i + r_n^j}\right) = \infty$ *for any* $i \ne j$;

(3) ω^i *is the weak limit or strong limit of* $u_n(x_n^i + r_n^i x)$ *in* $W^{1,2}_{\mathrm{loc}}(\mathbb{R}^2, N)$;

(4) *Energy identity*:

$$\lim_{n\to\infty} E(u_n, B) = E(u, B) + \sum_{i=1}^{k} E(\omega^i, \mathbb{R}^2);$$

(5) *Necklessness*: *the image* $u(B) \bigcup_{i=1}^{k} \omega^i(\mathbb{R}^2)$ *is a connected set.*

Lemma 1.6. *Suppose $\tau(u)$ satisfies*

$$\left(\int_{B_t \setminus B_{t/2}} |\tau(u)|^2\right)^{\frac{1}{2}} \le C\left(\frac{1}{t}\right)^a,$$

for some $0 < a < \frac{2}{3}$ and any $0 < t < 1$. Then $\tau(u)$ is bounded in $L^p(B)$ for some $p \ge \frac{6}{5}$.

Proof. We have

$$\begin{aligned}\int_{B_{2-k+1}\setminus B_{2-k}} |\tau(u)|^p &\le C(2^{-k})^{2-p}\|\tau(u)\|^p_{L^2(B_{2-k+1}\setminus B_{2-k})}\\ &\le C(2^{-k})^{2-p-ap}.\end{aligned}$$

Hence

$$\int_B |\tau(u)|^p \le C\sum_{k=1}^{\infty}(2^{-k})^{2-p-ap}.$$

When $0 < a < \frac{2}{3}$, we can choose some $p \geq \frac{6}{5}$ such that $2 - p - ap > 0$, and so

$$\sum_{k=1}^{\infty} (2^{-k})^{2-p-ap} \leq C,$$

which implies that $\tau(u)$ is bounded in $L^p(B)$ for some $p \geq \frac{6}{5}$. □

Thus Theorem 1.5 holds for sequences $\{u_n\}$ satisfying the following conditions.

(I) $\|u_n\|_{W^{1,2}(B)} \leq \Lambda$ and $(\int_{B_t \setminus B_{t/2}} |\tau(u_n)|^2)^{\frac{1}{2}} \leq C(\frac{1}{t})^a$ for some $0 < a < \frac{2}{3}$ and any $0 < t < 1$, where C is independent of n and t, and

(II) $u_n \to u$ strongly in $W^{1,2}_{\mathrm{loc}}(B \setminus \{0\}, N)$ as $n \to \infty$.

With the help of this observation, we find the following theorem.

Theorem 1.7. *Let $\{u_n\}$ be a sequence of maps from B to N in $W^{1,2}(B, N)$ with tension field $\tau(u_n)$, where B is the unit disk of $\mathbb{R}^2$ centered at the origin. If*

(I) $\|u_n\|_{W^{1,2}(B)} \leq \Lambda$ *and*

$$\left(\int_{B_t \setminus B_{t/2}} |\tau(u_n)|^2\right)^{\frac{1}{2}} \leq C\left(\frac{1}{t}\right)^a$$

for some $0 < a < 1$ and any $0 < t < 1$, where C is independent of n and t, and

(II) $u_n \to u$ *strongly in* $W^{1,2}_{\mathrm{loc}}(B \setminus \{0\}, N)$ *as* $n \to \infty$,

then there exists a subsequence of $\{u_n\}$ (still denoted by $\{u_n\}$) and some nonnegative integer k such that, for any $i = 1, \ldots, k$, there are some points x_n^i, positive numbers r_n^i, and a nonconstant harmonic sphere ω^i (which is viewed as a map from $\mathbb{R}^2 \cup \{\infty\} \to N$), such that:

(1) $x_n^i \to 0$, $r_n^i \to 0$ *as* $n \to \infty$;

(2) $\lim_{n\to\infty} \left(\frac{r_n^i}{r_n^j} + \frac{r_n^j}{r_n^i} + \frac{|x_n^i - x_n^j|}{r_n^i + r_n^j}\right) = \infty$ *for any* $i \neq j$;

(3) ω^i *is the weak limit or strong limit of* $u_n(x_n^i + r_n^i x)$ *in* $W^{1,2}_{\mathrm{loc}}(\mathbb{R}^2, N)$;

(4) *Energy identity*: $\lim_{n\to\infty} E(u_n, B) = E(u, B) + \sum_{i=1}^k E(\omega^i, \mathbb{R}^2)$;

(5) *Neckless*: *the image* $u(B) \bigcup_{i=1}^k \omega^i(\mathbb{R}^2)$ *is a connected set.*

Remark 1.8. When

$$\left(\int_{B_t \setminus B_{t/2}} |\tau(u_n)|^2\right)^{\frac{1}{2}} \leq C\left(\frac{1}{t}\right)^a$$

for some $0 < a < 1$ and any $0 < t < 1$, where C is independent of n and t, we can deduce that $\tau(u_n)$ is uniformly bounded in $L^p(B)$ for any $p < 2/(1+a)$, and when $a \to 1$, $p \to 1$. Hence our condition is stronger than the condition that the tension

field is bounded in L^p for some $p > 1$, and this result suggests that we probably have a positive answer to the Question on page 367.

Organization of the paper. In Section 2 we quote and prove several important results. In Section 3 we prove the removable singularity result. Theorem 1.7 is proved in Section 4. Throughout the paper, the letter C is used to denote positive constants which vary from line to line. We do not always distinguish between sequences and their subsequences.

2. The ϵ-regularity lemma and the Pohozaev inequality

This section contains a well-known small energy regularity lemma for approximated harmonic maps and a version of the Pohozaev inequality, which will be important later. We assume that the disk $B \subseteq \mathbb{R}^2$ is the unit disk centered at the origin, which has the standard flat metric.

Lemma 2.1. *Suppose that $u \in W^{2,2}(B, N)$ and $\tau(u) = g \in L^2(B, \mathbb{R}^K)$. Then there exists an $\varepsilon_0 > 0$ such that if $\int_B |\nabla u|^2 \le \varepsilon_0^2$, we have*

$$\|u - \bar{u}\|_{W^{2,2}(B_{1/2})} \le C(\|\nabla u\|_{L^2(B)} + \|g\|_{L^2(B)}). \tag{2-1}$$

Here $\bar{u}$ is the mean value of u over $B_{1/2}$.

Proof. We can find a complete proof of this lemma in [Ding and Tian 1995]. □

Using the standard elliptic estimates and the embedding theorems, we can derive from the above lemma that

Corollary 2.2. *Under the assumptions of Proposition 1.2, we have*

$$\begin{aligned} \operatorname{Osc}_{B_{2r}\setminus B_r} u &\le C(\|\nabla u\|_{L^2(B_{4r}\setminus B_{r/2})} + r\|g\|_{L^2(B_{4r}\setminus B_{r/2})}) \\ &\le C(\|\nabla u\|_{L^2(B_{4r}\setminus B_{r/2})} + r^{1-a}). \end{aligned} \tag{2-2}$$

Lemma 2.3 (Pohozaev inequality). *Under the assumptions of Proposition 1.2, for $0 < t_2 < t_1 < 1$,*

$$\int_{\partial(B_{t_1}\setminus B_{t_2})} r\Big(\Big|\frac{\partial u}{\partial r}\Big|^2 - \tfrac{1}{2}|\nabla u|^2\Big) \le t_1 \|\nabla u\|_{L^2(B_{t_1}\setminus B_{t_2})} \|g\|_{L^2(B_{t_1}\setminus B_{t_2})}. \tag{2-3}$$

Proof. Multiplying both sides of the equation $\tau(u) = g$ by $r(\partial u/\partial r)$, we get

$$\int_{B_{t_1}\setminus B_{t_2}} r\frac{\partial u}{\partial r} \triangle u = \int_{B_{t_1}\setminus B_{t_2}} r\frac{\partial u}{\partial r} g.$$

Integrating by parts, we get

$$\int_{B_{t_1}\setminus B_{t_2}} r\frac{\partial u}{\partial r} \triangle u \, dx = \int_{\partial(B_{t_1}\setminus B_{t_2})} r\Big|\frac{\partial u}{\partial r}\Big|^2 - \int_{B_{t_1}\setminus B_{t_2}} \nabla\Big(r\frac{\partial u}{\partial r}\Big)\nabla u \, dx$$

and

$$\begin{aligned}\int_{B_{t_1}\setminus B_{t_2}} \nabla\Big(r\frac{\partial u}{\partial r}\Big)\nabla u\,dx &= \int_{B_{t_1}\setminus B_{t_2}} \nabla\Big(x^k\frac{\partial u}{\partial x^k}\Big)\nabla u\,dx\\ &= \int_{B_{t_1}\setminus B_{t_2}} |\nabla u|^2 + \int_{t_2}^{t_1}\int_0^{2\pi} \frac{r}{2}\frac{\partial}{\partial r}|\nabla u|^2 r\,d\theta\,dr\\ &= \int_{B_{t_1}\setminus B_{t_2}} |\nabla u|^2 + \tfrac{1}{2}\int_{\partial(B_{t_1}\setminus B_{t_2})} |\nabla u|^2 r - \int_{B_{t_1}\setminus B_{t_2}} |\nabla u|^2\\ &= \tfrac{1}{2}\int_{\partial(B_{t_1}\setminus B_{t_2})} |\nabla u|^2 r.\end{aligned}$$

This implies the conclusion of the lemma. □

Corollary 2.4. *Under the assumptions of Proposition 1.2, we have*

$$\frac{\partial}{\partial t}\int_{B_t\setminus B_{t/2}} \Big|\frac{\partial u}{\partial r}\Big|^2 - \tfrac{1}{2}|\nabla u|^2 \le C\|\nabla u\|_{L^2(B_t\setminus B_{t/2})} t^{-a}. \tag{2-4}$$

Proof. In the previous lemma, let $t_1 = t$ and $t_2 = t/2$. Then

$$\begin{aligned}\frac{\partial}{\partial t}\int_{B_t\setminus B_{t/2}} \Big|\frac{\partial u}{\partial r}\Big|^2 - \tfrac{1}{2}|\nabla u|^2 &= \int_{\partial B_t}\Big(\Big|\frac{\partial u}{\partial r}\Big|^2 - \tfrac{1}{2}|\nabla u|^2\Big) - \tfrac{1}{2}\int_{\partial B_{t/2}}\Big(\Big|\frac{\partial u}{\partial r}\Big|^2 - \tfrac{1}{2}|\nabla u|^2\Big)\\ &\le \|g\|_{L^2(B_t\setminus B_{t/2})}\|\nabla u\|_{L^2(B_t\setminus B_{t/2})}\\ &\le C\|\nabla u\|_{L^2(B_t\setminus B_{t/2})} t^{-a}.\end{aligned}$$

□

Corollary 2.5. *Under the assumptions of Proposition 1.2,*

$$\int_{B_t\setminus B_{t/2}} \Big|\frac{\partial u}{\partial r}\Big|^2 - \tfrac{1}{2}|\nabla u|^2 \le C\|\nabla u\|_{L^2(B_t)} t^{1-a}. \tag{2-5}$$

Proof. Integrating both sides of the inequality (2-4) from 0 to t and noting that $\|\nabla u\|_{L^2(B_s\setminus B_{s/2})} \le \|\nabla u\|_{L^2(B_t)}$ for any $s \le t$, we get (2-5). □

3. Removal of singularities

We now discuss the removal of singularities of a class of approximated harmonic maps from the unit disk of $\mathbb{R}^2$ to a closed Riemannian manifold N.

Lemma 3.1. *Assume that u satisfies the assumptions of Proposition 1.2. Then there are constants $\lambda > 0$ and $C > 0$ such that*

$$\int_{B_r} |\nabla u|^2 \le C r^{\lambda} \tag{3-1}$$

for r small enough.

Proof. Because we only need to prove the lemma for r small, we can assume that $E(u, B) < \varepsilon_0$. Let $u^*(r) : (0, 1) \to \mathbb{R}^K$ be a curve defined by

$$u^*(r) = \frac{1}{2\pi} \int_0^{2\pi} u(r, \theta)\, d\theta.$$

Then

$$\frac{\partial u^*}{\partial r} = \frac{1}{2\pi} \int_0^{2\pi} \frac{\partial u}{\partial r}\, d\theta.$$

On the one hand, we have

$$\begin{aligned}\int_{B_{2-k_t}\setminus B_{2-k-1_t}} \nabla u \nabla(u - u^*) &\geq \int_{B_{2-k_t}\setminus B_{2-k-1_t}} \left(|\nabla u|^2 - \left|\frac{\partial u}{\partial r}\right|^2\right) \\ &\geq \tfrac{1}{2} \int_{B_{2-k_t}\setminus B_{2-k-1_t}} |\nabla u|^2 - C(2^{-k}t)^{1-a},\end{aligned}$$

where the second inequality makes use of (2-5).

Summing k from 0 to infinity, we get

$$\int_{B_t} \nabla u \nabla(u - u^*) \geq \tfrac{1}{2} \int_{B_t} |\nabla u|^2 - Ct^{1-a}.$$

On the other hand,

$$\begin{aligned}&\int_{B_{2-k_t}\setminus B_{2-k-1_t}} \nabla u \nabla(u - u^*) \\ &\quad = -\int_{B_{2-k_t}\setminus B_{2-k-1_t}} (u - u^*)\Delta u + \int_{\partial(B_{2-k_t}\setminus B_{2-k-1_t})} \frac{\partial u}{\partial r}(u - u^*) \\ &\quad = -\int_{B_{2-k_t}\setminus B_{2-k-1_t}} (u - u^*)(\tau(u) - A(u)(\nabla u, \nabla u)) + \int_{\partial(B_{2-k_t}\setminus B_{2-k-1_t})} \frac{\partial u}{\partial r}(u - u^*).\end{aligned}$$

Hence, by summing k from 0 to infinity, we get

$$\begin{aligned}&\int_{B_t} \nabla u \nabla(u - u^*) \\ &\quad \leq \sum_{k=0}^{\infty} \|u - u^*\|_{L^\infty(B_{2-k_t}\setminus B_{2-k-1_t})} \left(\|A\|_{L^\infty} \int_{B_{2-k_t}\setminus B_{2-k-1_t}} |\nabla u|^2 + C(2^{-k}t)^{1-a}\right) \\ &\qquad\qquad + \int_{\partial B_t} \frac{\partial u}{\partial r}(u - u^*) \\ &\quad \leq \epsilon \int_{B_t} |\nabla u|^2 + Ct^{1-a} + \int_{\partial B_t} \frac{\partial u}{\partial r}(u - u^*).\end{aligned}$$

Note that we used Corollary 2.2 and ensured that ϵ is small by letting t be small.

Note that

$$\left|\int_{\partial B_t}\frac{\partial u}{\partial r}(u-u^*)\right| \le \left(\int_{\partial B_t}\left|\frac{\partial u}{\partial r}\right|^2\right)^{\frac{1}{2}}\left(\int_{\partial B_t}|u-u^*|^2\right)^{\frac{1}{2}}$$
$$\le \left(\int_0^{2\pi} t^2\left|\frac{\partial u}{\partial r}\right|^2 d\theta\right)^{\frac{1}{2}}\left(\int_0^{2\pi}\left|\frac{\partial u}{\partial \theta}\right|^2 d\theta\right)^{\frac{1}{2}}$$
$$\le \tfrac{1}{2}\int_0^{2\pi}\left(\left|\frac{\partial u}{\partial \theta}\right|^2 + t^2\left|\frac{\partial u}{\partial r}\right|^2\right)d\theta = \frac{t}{2}\int_{\partial B_t}|\nabla u|^2.$$

Combining the two sides of the inequalities and letting ϵ be small (we can do this by letting t be small), we conclude that there is a constant $\lambda \in (0,1)$ such that

$$\lambda\int_{B_t}|\nabla u|^2 \le t\int_{\partial B_t}|\nabla u|^2 + Ct^{1-a}.$$

Set $f(t) = \int_{B_t}|\nabla u|^2$. Then we get the ordinary differential inequality

$$\left(\frac{f(t)}{t^\lambda}\right)' \ge -Ct^{-\lambda-a}.$$

Letting λ be small enough that $\lambda + a < 1$, we get

$$f(t) = \int_{B_t}|\nabla u|^2 \le Ct^\lambda$$

for t small enough. □

Proof of Proposition 1.2. Let $r_k = 2^{-k}$ and $v_k(x) = u(r_k x)$. Then

$$\left(\int_{B_2\setminus B_1}|\nabla v_k|^{2s}\right)^{\frac{1}{2s}} \le C\|v_k - \bar{v_k}\|_{W^{2,2}(B_2\setminus B_1)}$$
$$\le \left(\int_{B_4\setminus B_{1/2}}|\nabla v_k|^2\right)^{\frac{1}{2}} + C\left(\int_{B_{4r_k}\setminus B_{r_k/2}} r_k^2|\tau|^2\right)^{\frac{1}{2}}.$$

Therefore we deduce that

$$\int_{B_2\setminus B_1}|\nabla v_k|^{2s} \le C\left(\int_{B_4\setminus B_{1/2}}|\nabla v_k|^2\right)^s + C\left(\int_{B_{4r_k}\setminus B_{r_k/2}} r_k^2|\tau|^2\right)^s$$
$$\le C\left(\int_{B_4\setminus B_{1/2}}|\nabla v_k|^2\right)^s + Cr_k^{2s(1-a)}.$$

Note that when k is large enough,

$$\int_{B_{4r_k}\setminus B_{r_k/2}}|\nabla u|^2 \le 1.$$

Hence

$$r_k^{2s-2}\int_{B_{2r_k}\setminus B_{r_k}}|\nabla u|^{2s}\le C\left(\int_{B_{4r_k}\setminus B_{r_k/2}}|\nabla u|^2\right)^s+Cr_k^{2s(1-a)}$$
$$\le C\int_{B_{4r_k}\setminus B_{r_k/2}}|\nabla u|^2+Cr_k^{2s(1-a)}.$$

This implies that

$$\int_{B_{2r_k}\setminus B_{r_k}}|\nabla u|^{2s}\le Cr_k^{2-2s}r_k^{\lambda}+Cr_k^{2-2sa}.$$

Now choose $s>1$ such that $2s-2<\lambda/2$ and $2-2sa>0$. There exists a positive integer k_0 such that when $k\ge k_0$,

$$\int_{B_{2-k+1}\setminus B_{2-k}}|\nabla u|^{2s}\le C(2^{(-\lambda/2)k}+2^{-k(2-2sa)}).$$

Therefore $\int_{B_r}|\nabla u|^{2s}\le C\sum_{k=k_0}^{\infty}(2^{(-\lambda/2)k}+2^{-k(2-2sa)})\le C$ for any $r\le 2^{-k_0+1}$, which completes the proof. □

Proof of Theorem 1.3. Note that

$$\int_{B_{2-k}\setminus B_{2-k-1}}|\tau(u)|^p\le C(2^{-k})^{2-p}\left(\int_{B_{2-k}\setminus B_{2-k-1}}|\tau(u)|^2\right)^{p/2}\le C(2^{-k})^{2-p-pa}.$$

Summing over k from 0 to infinity, we deduce that $\int_B|\tau(u)|^p\le C$ for $p<2/(1+a)$.

Recall that we have proved that $\nabla u\in L^{2s}(B)$ for some $s>1$. Hence, by standard elliptic estimates and the bootstrap argument, we can deduce that

$$u\in\bigcap_{1<p<\frac{2}{1+a}}W^{2,p}(B,N).$$ □

4. The bubble tree structure

Energy identity. Assume that $\{u_n\}$ is a uniformly bounded sequence in $W^{1,2}(B,N)$ and that there exists a constant C, independent of n and t, such that

$$\left(\int_{B_t\setminus B_{t/2}}|\tau(u_n)|^2\right)^{\frac{1}{2}}\le C\left(\frac{1}{t}\right)^a$$

for some $0<a<1$ and any $0<t<1$. In this section, we will prove the energy identity for this sequence. For convenience, we will assume that there is only one bubble ω, which is the strong limit of $u_n(r_n.)$ in $W^{1,2}_{\mathrm{loc}}(\mathbb{R}^2,N)$. Under this assumption we can deduce the following by a standard blowup argument.

Lemma 4.1. *For any $\epsilon > 0$, there exist R and δ such that*

$$\int_{B_{2\lambda}\setminus B_\lambda} |\nabla u_n|^2 \le \epsilon^2 \quad \textit{for any} \quad \lambda \in \left(\frac{Rr_n}{2}, 2\delta\right). \tag{4-1}$$

Proof of the energy identity. For a given $R > 0$, we have

$$\lim_{n\to\infty}\int_B |\nabla u_n|^2 = \lim_{n\to\infty}\int_{B\setminus B_\delta} |\nabla u_n|^2 + \lim_{n\to\infty}\int_{B_\delta\setminus B_{Rr_n}} |\nabla u_n|^2 + \lim_{n\to\infty}\int_{B_{Rr_n}} |\nabla u_n|^2,$$

$$\lim_{\delta\to 0}\lim_{n\to\infty}\int_{B\setminus B_\delta} |\nabla u_n|^2 = \int_B |\nabla u|^2, \quad \text{and} \quad \lim_{R\to\infty}\lim_{n\to\infty}\int_{B_{Rr_n}} |\nabla u_n|^2 = \int_{\mathbb{R}^2} |\nabla \omega|^2,$$

Hence, to prove the energy identity, we only need to prove that

$$\lim_{R\to\infty}\lim_{\delta\to 0}\lim_{n\to\infty}\int_{B_\delta\setminus B_{Rr_n}} |\nabla u_n|^2 = 0. \tag{4-2}$$

The proof is a little similar to the proof in the previous section. We assume that $\delta = 2^{m_n} Rr_n$, where m_n is a positive integer.

On the one hand, we have

$$\begin{aligned}\int_{B_{2^k Rr_n}\setminus B_{2^{k-1}Rr_n}} \nabla u_n \nabla(u_n - u_n^*) &\ge \int_{B_{2^k Rr_n}\setminus B_{2^{k-1}Rr_n}} \left(|\nabla u_n|^2 - \left|\frac{\partial u_n}{\partial r}\right|^2\right) \\ &\ge \tfrac{1}{2}\int_{B_{2^k Rr_n}\setminus B_{2^{k-1}Rr_n}} |\nabla u_n|^2 - C(2^k Rr_n)^{1-a}.\end{aligned}$$

This implies that

$$\int_{B_\delta\setminus B_{Rr_n}} \nabla u_n \nabla(u_n - u_n^*) \ge \tfrac{1}{2}\int_{B_\delta\setminus B_{Rr_n}} |\nabla u_n|^2 - C\delta^{1-a}.$$

On the other hand, we have

$$\begin{aligned}&\int_{B_{2^k Rr_n}\setminus B_{2^{k-1}Rr_n}} \nabla u_n \nabla(u_n - u_n^*) \\ &\quad = -\int_{B_{2^k Rr_n}\setminus B_{2^{k-1}Rr_n}} (u_n - u_n^*)\Delta u_n + \int_{\partial(B_{2^k Rr_n}\setminus B_{2^{k-1}Rr_n})} \frac{\partial u_n}{\partial r}(u_n - u_n^*) \\ &\quad = -\int_{B_{2^k Rr_n}\setminus B_{2^{k-1}Rr_n}} (u_n - u_n^*)(\tau(u_n) - A(u_n)(\nabla u_n, \nabla u_n)) \\ &\qquad\qquad + \int_{\partial(B_{2^k Rr_n}\setminus B_{2^{k-1}Rr_n})} \frac{\partial u_n}{\partial r}(u_n - u_n^*).\end{aligned}$$

Summing from 1 to m_n, we deduce that

$$\begin{aligned}
&\int_{B_\delta\setminus B_{Rr_n}} \nabla u_n \nabla(u_n - u_n^*) \\
&\le \sum_{k=1}^{m_n} \|u_n - u_n^*\|_{L^\infty(B_{2^k Rr_n}\setminus B_{2^{k-1}Rr_n})} \left(\|A\|_{L^\infty} \int_{B_{2^k Rr_n}\setminus B_{2^{k-1}Rr_n}} |\nabla u_n|^2 + C(2^k Rr_n)^{1-a} \right) \\
&\qquad + \int_{\partial(B_\delta\setminus B_{Rr_n})} \frac{\partial u_n}{\partial r}(u_n - u_n^*) \\
&\le \epsilon \int_{B_\delta\setminus B_{Rr_n}} |\nabla u_n|^2 + C\delta^{1-a} + \int_{\partial(B_\delta\setminus B_{Rr_n})} \frac{\partial u_n}{\partial r}(u_n - u_n^*).
\end{aligned}$$

Comparing the two sides, we get

$$(1-2\epsilon)\int_{B_\delta\setminus B_{Rr_n}} |\nabla u_n|^2 \le C\delta^{1-a} + 2\int_{\partial(B_\delta\setminus B_{Rr_n})} \frac{\partial u_n}{\partial r}(u_n - u_n^*).$$

As for the boundary terms, we have

$$\begin{aligned}
\int_{\partial B_\delta} \frac{\partial u_n}{\partial r}(u_n - u_n^*) &\le \left(\int_{\partial B_\delta} \left| \frac{\partial u_n}{\partial r} \right|^2 \right)^{\frac{1}{2}} \left(\int_{\partial B_\delta} |u_n - u_n^*|^2 \right)^{\frac{1}{2}} \\
&\le \left(\int_0^{2\pi} \delta^2 \left| \frac{\partial u_n}{\partial r}\, d\theta \right|^2 \right)^{\frac{1}{2}} \left(\int_0^{2\pi} \left| \frac{\partial u_n}{\partial \theta} \right|^2 d\theta \right)^{\frac{1}{2}} \\
&\le \frac{1}{2}\int_0^{2\pi} \delta^2 \left| \frac{\partial u_n}{\partial r} d\theta \right|^2 + \left| \frac{\partial u_n}{\partial \theta} \right|^2 d\theta = \frac{\delta^2}{2}\int_0^{2\pi} |\nabla u_n|^2 d\theta.
\end{aligned}$$

Now, by the trace embedding theorem, we have

$$\begin{aligned}
\int_0^{2\pi} |\nabla u_n(\cdot,\delta)|^2 \delta d\theta &= \int_{\partial B_\delta} |\nabla u_n(\cdot,\delta)|^2 dS_\delta \\
&\le C\delta \|\nabla u_n\|^2_{W^{1,2}(B_{3\delta/2}\setminus B_{\delta/2})} \\
&\le C\delta \|u_n - \bar{u}_n\|^2_{W^{2,2}(B_{3\delta/2}\setminus B_{\delta/2})} \\
&\le C\delta \left(\frac{1}{\delta}\|\nabla u_n\|^2_{L^2(B_{2\delta})} + \|\tau(u_n)\|^2_{L^2(B_{2\delta}\setminus B_{\delta/4})} \right) \\
&\le C\delta^{1-2a},
\end{aligned}$$

for δ small. From this we deduce that

$$\int_{\partial B_\delta} \frac{\partial u_n}{\partial r}(u_n - u_n^*) \le C\delta^{2(1-a)}.$$

Similarly we get

$$\int_{\partial B_{Rr_n}} \frac{\partial u_n}{\partial r}(u_n - u_n^*) \le C(Rr_n)^{2(1-a)},$$

for n big enough. Therefore

$$(1-2\epsilon)\int_{B_\delta\setminus B_{Rr_n}}|\nabla u_n|^2 \le C\delta^{1-a}+C\delta^{2(1-a)}+C(Rr_n)^{2(1-a)},$$

which clearly implies (4-2), and we are done. □

Necklessness. In this part we prove that there is no neck between the base map u and the bubble ω, that is, the C^0 compactness of the sequence modulo bubbles.

Proof. We only need to prove that

$$\lim_{R\to\infty}\lim_{\delta\to 0}\lim_{n\to\infty}\operatorname{Osc}_{B_\delta\setminus B_{Rr_n}}u_n=0. \tag{4-3}$$

Again we assume that $\delta=2^{m_n}Rr_n$ and let $Q(t)=B_{2^{t+t_0}Rr_n}\setminus B_{2^{t_0-t}Rr_n}$. Similarly to the proof of the previous part, we can get

$$\begin{aligned}&(1-2\epsilon)\int_{Q(k)}|\nabla u_n|^2\\ &\quad\le 2^{k+t_0}Rr_n\int_{\partial B_{2^{k+t_0}Rr_n}}|\nabla u_n|^2+2^{t_0-k}Rr_n\int_{\partial B_{2^{t_0-k}Rr_n}}|\nabla u_n|^2+C(2^{k+t_0}Rr_n)^{1-a}.\end{aligned}$$

Set $f(t)=\int_{Q(t)}|\nabla u_n|^2$. Then we have

$$(1-2\epsilon)f(t)\le(1-2\epsilon)f(k+1)\le\frac{1}{\log 2}f'(k+1)+C(2^{k+t_0}Rr_n)^{1-a}$$

for $k\le t\le k+1$.

Note that

$$\begin{aligned}&f'(k+1)-f'(t)\\ &\quad=\int_{\partial(B_{2^{k+1+t_0}Rr_n}\setminus B_{2^{t+t_0}Rr_n})}\frac{\partial u_n}{\partial r}(u_n-u_n^*)+\int_{\partial(B_{2^{t_0-t}Rr_n}\setminus B_{2^{t_0-k-1}Rr_n})}\frac{\partial u_n}{\partial r}(u_n-u_n^*)\\ &\quad\le C(2^{t+t_0}Rr_n)^{2(1-a)}.\end{aligned}$$

Therefore

$$(1-2\epsilon)f(t)\le\frac{1}{\log 2}f'(t)+C(2^{t+t_0}Rr_n)^{1-a}. \tag{4-4}$$

It follows that

$$\begin{aligned}(2^{-(1-2\epsilon)t}f(t))'&=2^{-(1-2\epsilon)t}f'(t)-(1-2\epsilon)2^{-(1-2\epsilon)t}f(t)\log 2\\ &\ge -C2^{(1-a-(1-2\epsilon))t}(2^{t_0}Rr_n)^{1-a}.\end{aligned}$$

Integrating from 1 to L, we get

$$\begin{aligned} 2^{-(1-2\epsilon)L} f(L) - 2^{-(1-2\epsilon)} f(1) &\geq -C \int_1^L 2^{(1-a-(1-2\epsilon))t} (2^{t_0} R r_n)^{1-a} \\ &= -C \frac{2^{(1-a-(1-2\epsilon))t}}{\log 2(1-a-(1-2\epsilon))} \Big|_1^L (2^{t_0} R r_n)^{1-a} \\ &\geq -C (2^{t_0} R r_n)^{1-a}. \end{aligned}$$

Therefore we have

$$f(1) \leq f(L) 2^{-(1-2\epsilon)(L-1)} + C (2^{t_0} R r_n)^{1-a}. \tag{4-5}$$

Now let $t_0 = i$ and $D_i = B_{2^{i+1} R r_n} \setminus B_{2^i R r_n}$. Then we have $f(1) = \int_{D_i \cup D_{i-1}} |\nabla u_n|^2$, and the inequality holds true for L satisfying

$$Q(L) \subseteq B_\delta \setminus B_{R r_n} = B_{2^{m_n} R r_n} \setminus B_{R r_n}.$$

In other words, L should satisfy $i - L \geq 0$ and $i + L \leq M_n$.

(I) If $i \leq \frac{1}{2} m_n$, let $L = i$. Then

$$f(1) = \int_{D_i \cup D_{i-1}} |\nabla u_n|^2 \leq C E^2(u_n, B_\delta \setminus B_{R r_n}) 2^{-(1-2\epsilon)i} + C (2^i R r_n)^{1-a}.$$

(II) If $i > \frac{1}{2} m_n$, let $L = m_n - i$. Then

$$f(1) = \int_{D_i \cup D_{i-1}} |\nabla u_n|^2 \leq C E^2(u_n, B_\delta \setminus B_{R r_n}) 2^{-(1-2\epsilon)(m_n - i)} + C (2^i R r_n)^{1-a}.$$

Hence we have

$$\begin{aligned} \sum_{i=1}^{m_n} E(u_n, D_i) &\leq C E(u_n, B_\delta \setminus B_{R r_n}) \Bigg(\sum_{i \leq \frac{1}{2} m_n} 2^{-i(1-2\epsilon)/2} + \sum_{i > \frac{1}{2} m_n} 2^{-(m_n - i) 1 - 2\epsilon/(2)} \Bigg) \\ &\qquad + C \sum_{i=1}^{m_n} (2^i R r_n)^{(1-a)/2} \\ &\leq C E(u_n, B_\delta \setminus B_{R r_n}) + C \delta^{(1-a)/2}. \end{aligned}$$

Thus we get

$$\begin{aligned} \mathrm{Osc}_{B_\delta \setminus B_{R r_n}} u_n &\leq C \sum_{i=1}^{m_n} (E(u_n, D_i) + (2^i R r_n)^{1-a}) \\ &\leq C E(u_n, B_\delta \setminus B_{R r_n}) + C \delta^{(1-a)/2}. \end{aligned}$$

Clearly this implies (4-3), as needed. □

Acknowledgements

I thank Professor Guofang Wang for pointing out many typing errors, and Professor Youde Wang for bringing [Li and Wang 2010] to my attention.

My interest in this kind of problem began at a class given by Professor Yuxiang Li at Tinghua University, and I had many useful discussions with him.

I am thankful to the referee for detailed comments, which have made this paper more readable.

References

[Ding and Tian 1995] W. Ding and G. Tian, "Energy identity for a class of approximate harmonic maps from surfaces", *Comm. Anal. Geom.* **3**:3-4 (1995), 543–554. MR 97e:58055 Zbl 0855.58016

[Jost 1987] J. Jost, "Two-dimensional geometric variational problems", pp. 1094–1100 in *Proceedings of the International Congress of Mathematicians, Vol. 1* (Berkeley, Calif., 1986), edited by A. M. Gleason, Amer. Math. Soc., Providence, RI, 1987. MR 89g:58045 Zbl 0850.58010

[Li and Wang 2010] Y. Li and Y. Wang, "Bubbling location for sequences of approximate f-harmonic maps from surfaces", *Internat. J. Math.* **21**:4 (2010), 475–495. MR 2011g:58028 Zbl 1190.35067

[Li and Zhu 2010] J. Li and X. Zhu, "Energy identity for the maps from a surface with tension field bounded in L^p", preprint, 2010. arXiv 1205.2978

[Lin and Wang 1998] F. Lin and C. Wang, "Energy identity of harmonic map flows from surfaces at finite singular time", *Calc. Var. Partial Differential Equations* **6**:4 (1998), 369–380. MR 99k:58047 Zbl 0908.58008

[Lin and Wang 2002] F. Lin and C. Wang, "Harmonic and quasi-harmonic spheres. II", *Comm. Anal. Geom.* **10**:2 (2002), 341–375. MR 2003d:58029 Zbl 1042.58005

[Parker 1996] T. H. Parker, "Bubble tree convergence for harmonic maps", *J. Differential Geom.* **44**:3 (1996), 595–633. MR 98k:58069 Zbl 0874.58012

[Qing 1995] J. Qing, "On singularities of the heat flow for harmonic maps from surfaces into spheres", *Comm. Anal. Geom.* **3**:1-2 (1995), 297–315. MR 97c:58154 Zbl 0868.58021

[Qing and Tian 1997] J. Qing and G. Tian, "Bubbling of the heat flows for harmonic maps from surfaces", *Comm. Pure Appl. Math.* **50**:4 (1997), 295–310. MR 98k:58070 Zbl 0879.58017

[Sacks and Uhlenbeck 1981] J. Sacks and K. Uhlenbeck, "The existence of minimal immersions of 2-spheres", *Ann. of Math.* (2) **113**:1 (1981), 1–24. MR 82f:58035 Zbl 0462.58014

[Topping 2004a] P. Topping, "Repulsion and quantization in almost-harmonic maps, and asymptotics of the harmonic map flow", *Ann. of Math.* (2) **159**:2 (2004), 465–534. MR 2005g:58029 Zbl 1065.58007

[Topping 2004b] P. Topping, "Winding behaviour of finite-time singularities of the harmonic map heat flow", *Math. Z.* **247**:2 (2004), 279–302. MR 2004m:53120 Zbl 1067.53055

[Wang 1996] C. Wang, "Bubble phenomena of certain Palais–Smale sequences from surfaces to general targets", *Houston J. Math.* **22**:3 (1996), 559–590. MR 98h:58053 Zbl 0879.58019

Received February 16, 2011. Revised December 2, 2011.

YONG LUO
MATHEMATISCHES INSTITUT, ALBERT-LUDWIGS-UNIVERSITÄT
ECKERSTRASSE 1
79104 FREIBURG
GERMANY
yong.luo@math.uni-freiburg.de

PACIFIC JOURNAL OF MATHEMATICS
Vol. 256, No. 2, 2012

QUOTIENTS BY ACTIONS OF THE DERIVED GROUP OF A MAXIMAL UNIPOTENT SUBGROUP

DMITRI I. PANYUSHEV

Let U be a maximal unipotent subgroup of a connected semisimple group G and U' the derived group of U. If X is an affine G-variety, then the algebra of U'-invariants, $k[X]^{U'}$, is finitely generated and the quotient morphism $\pi : X \to X/\!\!/U' = \mathsf{Spec}\, k[X]^{U'}$ is well-defined. In this article, we study properties of such quotient morphisms, e.g. the property that all the fibres of π are equidimensional. We also establish an analogue of the Hilbert-Mumford criterion for the null-cones with respect to U'-invariants.

Introduction

The ground field $\Bbbk$ is algebraically closed and of characteristic zero. Let G be a semisimple algebraic group with Lie algebra $\mathfrak{g}$. Fix a maximal unipotent subgroup $U \subset G$ and a maximal torus T of the Borel subgroup $B = N_G(U)$. Set $U' = (U, U)$. Let X be an irreducible affine variety acted upon by G. The algebra of covariants (or, U-invariants) $\Bbbk[X]^U$ is a classical and important object in Invariant Theory. It is known that $\Bbbk[X]^U$ is finitely generated and has many other useful properties and applications, see e.g. [9, Ch. 3, § 3]. For a factorial conical variety X with rational singularities, there are interesting relations between the Poincaré series of the graded algebras $\Bbbk[X]$ and $\Bbbk[X]^U$, see [3], [12, Ch. 5]. Similar results for U'-invariants are obtained in [14].

A surprising observation that stems from [14] is that, to a great extent, the theory of U'-invariants is parallel to that of U-invariants. In this article, we elaborate on further aspects of this parallelism. Our main object is the quotient $\pi_{X,U'} : X \to X/\!\!/U' = \mathsf{Spec}(\Bbbk[X]^{U'})$. Specifically, we are interested in the property that $X/\!\!/U'$ is an affine space and/or the morphism $\pi_{X,U'}$ is equidimensional (i.e., all the fibres of $\pi_{X,U'}$ have the same dimension). Our ultimate goal is to prove for U' an analogue of the Hilbert–Mumford criterion and to provide a classification of the irreducible representations V of simple algebraic groups G such that $\Bbbk[V]$ is a free $\Bbbk[V]^{U'}$-module. We also develop some theory for U'-actions on the affine prehomogeneous

At the author's request, this article did not undergo any editorial changes.
MSC2010: 14L30, 17B20, 22E46.
Keywords: semisimple algebraic group, quotient, equidimensional morphism, invariant.

horospherical varieties of G ($\mathcal{S}$-*varieties* in terminology of [22]). As $U' = \{1\}$ for $G = SL_2$, one sometimes has to assume that G has no simple factors SL_2.

If X has a G-fixed point, say x_0, then the fibre of $\pi_{X,U'}$ containing x_0 is called the *null-cone*, and we denote it by $\mathfrak{N}_{U'}(X)$. (The null-cone $\mathfrak{N}_H(X)$ can be defined for any subgroup $H \subset G$ such that $\Bbbk[X]^H$ is finitely generated.) If G has no simple factors SL_2 nor SL_3, then the canonical affine model of $\Bbbk[G/U']$ constructed in [14, Sect. 2] consists of unstable points in the sense of GIT, and using this property we give a characterisation of $\mathfrak{N}_{U'}(X)$ in terms of one-parameter subgroups of T. We call it the *Hilbert–Mumford criterion for* U'. This is inspired by similar results of Brion for U-invariants [3, Sect. IV]. It is easily seen that $\mathfrak{N}_{U'}(X) \subset \mathfrak{N}_G(X)$. Therefore $G{\cdot}\mathfrak{N}_{U'}(X) \subset \mathfrak{N}_G(X)$. Using the Hilbert–Mumford criterion for U' we prove that $G{\cdot}\mathfrak{N}_{U'}(X) = \mathfrak{N}_G(X)$ whenever G has no simple factors SL_n. This should be compared with the result of Brion [3] that $G{\cdot}\mathfrak{N}_U(X) = \mathfrak{N}_G(X)$ for all G.

The $\mathcal{S}$-varieties are in one-to-one correspondence with the finitely generated monoids $\mathfrak{S}$ in the monoid $\mathfrak{X}_+$ of dominant weights, and the $\mathcal{S}$-variety corresponding to $\mathfrak{S} \subset \mathfrak{X}_+$ is denoted by $\mathcal{C}(\mathfrak{S})$. We give exhaustive answers to three natural problems related to the actions of U' on $\mathcal{S}$-varieties. A set of fundamental weights M is said to be *sparse* if the corresponding nodes of the Dynkin diagram are disjoint and, moreover, there does not exist any node (not in M) that is adjacent to two nodes from M. Our results are:

a) $\Bbbk[\mathcal{C}(\mathfrak{S})]^{U'}$ is a polynomial algebra *if and only if* the monoid $\mathfrak{S}$ is generated by a set of fundamental weights;

b) $\Bbbk[\mathcal{C}(\mathfrak{S})]^{U'}$ is a polynomial algebra and $\pi_{\mathcal{C}(\mathfrak{S}),U'}$ is equidimensional *if and only if* the monoid $\mathfrak{S}$ is generated by a sparse set of fundamental weights;

c) the morphism $\pi_{\mathcal{C}(\mathfrak{S}),U'}$ is equidimensional *if and only if* the convex polyhedral cone $\mathbb{R}^+\mathfrak{S}$ is generated by a sparse set of fundamental weights. (In particular, the cone $\mathbb{R}^+\mathfrak{S}$ is simplicial.)

Part a) is rather easy, while parts b) and c) require technical details related to the Bruhat decomposition of the flag variety associated with $\mathcal{C}(\mathfrak{S})$. If $\mathfrak{S}$ has one generator, say λ, and $\mathsf{R}(\lambda)$ is a simple G-module with highest weight λ, then $\mathcal{C}(\mathfrak{S})$ is the closure of the orbit of highest weight vectors in the dual G-module $\mathsf{R}(\lambda)^*$. Such a variety is denoted by $\mathcal{C}(\lambda)$. As in [22], we say that $\mathcal{C}(\lambda)$ is an HV-*variety*. Our results for HV-varieties are more complete. For instance, we compute the homological dimension of $\mathcal{C}(\lambda)/\!\!/U'$ and prove that $\mathfrak{N}_{U'}(\mathcal{C}(\lambda))$ is always of codimension 2 in $\mathcal{C}(\lambda)$. The criterion of part b) is then transformed into a sufficient condition applicable to a wider class of affine varieties:

Theorem 0.1. *Suppose that G acts on an irreducible affine variety X such that (1) $\Bbbk[X]^U$ is a polynomial algebra and (2) the weights of free generators are*

fundamental, pairwise distinct, and form a sparse set. Then $\Bbbk[X]^{U'}$ *is also polynomial, of Krull dimension* $2\dim X/\!\!/U$*, and the quotient* $\pi_{X,U'} : X \to X/\!\!/U'$ *is equidimensional.*

This exploits the theory of "contractions of actions" of G [15] and can be regarded as a continuation of our work in [13, Sect. 5], where the equidimensionality problem was considered for quotient morphism by U. For instance, under the hypotheses of Theorem 0.1, the morphism $\pi_{X,U}$ is also equidimensional.

In [14], we obtained a classification of the irreducible representations of simple algebraic groups such that $\Bbbk[V]^{U'}$ is a polynomial algebra. Now, using Theorem 0.1 and some ad hoc arguments, we extract from that list the representations having the additional property that $\pi_{V,U'}$ is equidimensional. The resulting list is precisely the list of representations such that $\Bbbk[V]$ is a free $\Bbbk[V]^{U'}$-module (such G-representations are said to be U'-*cofree*).

This work is organized as follows. Section 1 contains auxiliary results on $\mathcal{S}$-varieties [22], U'-invariants [14], and equidimensional morphisms. In Section 2, we consider U'-actions on the HV-varieties. Section 3 is devoted to the U'-actions on arbitrary $\mathcal{S}$-varieties. Here we prove results of items a) and b) above (Theorems 3.2, 3.4, and 3.7). In Section 4, we prove the general equidimensionality criterion for $\mathcal{S}$-varieties (item c)). The Hilbert–Mumford criterion for U' and relations between two null-cones are discussed in Section 5. In Section 6, we prove Theorem 0.1 and obtain the classification of U'-cofree representations of G.

Notation. If an algebraic group Q acts regularly on an irreducible affine variety X, then X is called a Q-variety and

- $Q_x = \{q \in Q \mid q{\cdot}x = x\}$ is the *stabiliser* of $x \in X$;
- $\Bbbk[X]^Q$ is the algebra of Q-invariant polynomial functions on X. If $\Bbbk[X]^Q$ is finitely generated, then $X/\!\!/Q := \mathsf{Spec}\,(\Bbbk[X]^Q)$, and the *quotient morphism* $\pi_Q = \pi_{X,Q} : X \to X/\!\!/Q$ is the mapping associated with the embedding $\Bbbk[X]^Q \hookrightarrow \Bbbk[X]$.

Throughout, G is a semisimple simply-connected algebraic group, $W = N_G(T)/T$ is the Weyl group, $B = TU$, and $r = \mathsf{rk}\,G$. Then

– Δ is the root system of (G, T), $\Pi = \{\alpha_1, \dots, \alpha_r\} \subset \Delta$ are the simple roots corresponding to U, and $\varpi_1, \dots, \varpi_r$ are the corresponding fundamental weights.

– The character group of T is denoted by $\mathfrak{X}$. All roots and weights are regarded as elements of the r-dimensional real vector space $\mathfrak{X}_{\mathbb{R}} := \mathfrak{X} \otimes \mathbb{R}$.

– $(\ ,\ )$ is a W-invariant symmetric non-degenerate bilinear form on $\mathfrak{X}_{\mathbb{R}}$ and $s_i \in W$ is the reflection corresponding to α_i. For any $\lambda \in \mathfrak{X}_+$, let λ^* denote the highest weight of the dual G-module, i.e., $\mathsf{R}(\lambda)^* \simeq \mathsf{R}(\lambda^*)$. The μ-weight space of $\mathsf{R}(\lambda)$ is denoted by $\mathsf{R}(\lambda)_\mu$.

We refer to [21] for standard results on root systems and representations of semisimple algebraic groups.

1. Recollections

1.1. ***Horospherical varieties with a dense orbit.*** A G-variety X is said to be *horospherical* if the stabiliser of any $x \in X$ contains a maximal unipotent subgroup of G. Following [22], affine horospherical varieties with a dense G-orbit are called $\mathcal{S}$*-varieties*. Let $\mathfrak{S}$ be a finitely generated monoid in $\mathfrak{X}_+$ and $\{\lambda_1, \dots, \lambda_m\}$ the minimal set of generators of $\mathfrak{S}$. Let $v_{-\lambda_i} \in \mathsf{R}(\lambda_i^*)$ be a lowest weight vector. Set $\boldsymbol{v} = (v_{-\lambda_1}, \dots, v_{-\lambda_m})$ and consider

$$\mathscr{C}(\mathfrak{S}) := \overline{G{\cdot}\boldsymbol{v}} \subset \mathsf{R}(\lambda_1^*) \oplus \dots \oplus \mathsf{R}(\lambda_m^*).$$

Clearly, $\mathscr{C}(\mathfrak{S})$ is an $\mathcal{S}$-variety; conversely, each $\mathcal{S}$-variety is obtained in this way [22]. Write $\langle \mathfrak{S} \rangle$ for the linear span of $\mathfrak{S}$ in $\mathfrak{X}_{\mathbb{R}}$ and set $\mathsf{rk}\, \mathfrak{S} = \dim_{\mathbb{R}} \langle \mathfrak{S} \rangle$. Let $L_{\mathfrak{S}}$ be the Levi subgroup such that $T \subset L_{\mathfrak{S}}$ and the roots of $L_{\mathfrak{S}}$ are those orthogonal to $\lambda_1, \dots, \lambda_m$. Then $P_{\mathfrak{S}} = L_{\mathfrak{S}} N_{\mathfrak{S}}$ is the standard parabolic subgroup, with unipotent radical $N_{\mathfrak{S}} \subset U$.

Theorem 1.1 ([22]). *The affine variety $\mathscr{C}(\mathfrak{S})$ has the following properties:*

1. *The algebra $\Bbbk[\mathscr{C}(\mathfrak{S})]$ is a multiplicity free G-module. More precisely, $\Bbbk[\mathscr{C}(\mathfrak{S})] = \bigoplus_{\lambda \in \mathfrak{S}} \mathsf{R}(\lambda)$ and this decomposition is a multigrading, i.e., $\mathsf{R}(\lambda)\mathsf{R}(\mu) = \mathsf{R}(\lambda + \mu)$;*
2. *The G-orbits in $\mathscr{C}(\mathfrak{S})$ are in a one-to-one correspondence with the faces of the convex polyhedral cone in $\mathfrak{X}_{\mathbb{R}}$ generated by $\mathfrak{S}$;*
3. *$\mathscr{C}(\mathfrak{S})$ is normal if and only if $\mathbb{Z}\mathfrak{S} \cap \mathbb{Q}^+\mathfrak{S} = \mathfrak{S}$;*
4. $\dim \mathscr{C}(\mathfrak{S}) = \dim G/P_{\mathfrak{S}} + \mathsf{rk}\, \mathfrak{S}$.

If $\mathfrak{S} = \mathbb{N}\lambda$, then we write $\mathscr{C}(\lambda)$, $P_\lambda, \dots$ in place of $\mathscr{C}(\mathbb{N}\lambda)$, $P_{\mathbb{N}\lambda}, \dots$. The variety $\mathscr{C}(\lambda)$ is the closure of the G-orbit of highest weight vectors in $\mathsf{R}(\lambda^*)$. Such varieties are called HV*-varieties*; they are always normal. Recall that a G-variety X is *spherical*, if B has a dense orbit in X. Since $B{\cdot}\boldsymbol{v}$ is dense in $\mathscr{C}(\mathfrak{S})$, all $\mathcal{S}$-varieties are spherical. By [15, Theorem 10]), a normal spherical variety has rational singularities and therefore is Cohen-Macaulay. In particular, if $\mathfrak{S}$ is a free monoid, then $\mathscr{C}(\mathfrak{S})$ has rational singularities.

1.2. ***Generalities on U'-invariants.*** We recall some results of [14] and thereby fix relevant notation. We regard $\mathfrak{X}$ as a poset with respect to the *root order* "$\preccurlyeq$". This means that $\nu \preccurlyeq \mu$ if $\mu - \nu$ is a non-negative integral linear combination of simple roots. For any $\lambda \in \mathfrak{X}_+$, we fix a simple G-module $\mathsf{R}(\lambda)$ and write $\mathcal{P}(\lambda)$ for the set of T-weights of $\mathsf{R}(\lambda)$. Then $(\mathcal{P}(\lambda), \preccurlyeq)$ is a finite poset and λ is its unique maximal element. Let $e_i \in \mathfrak{u} = \operatorname{Lie} U$ be a root vector corresponding to $\alpha_i \in \Pi$. Then $(e_1, \dots, e_r)$ is a basis for $\operatorname{Lie}(U/U')$.

The subspace of U'-invariants in $\mathsf{R}(\lambda)$ has a nice description. Since $\mathsf{R}(\lambda)^{U'}$ is acted upon by B/U', it is T-stable. Hence $\mathsf{R}(\lambda)^{U'} = \bigoplus_{\mu\in\mathcal{I}_\lambda} \mathsf{R}(\lambda)^{U'}_\mu$, where $\mathcal{I}_\lambda$ is a subset of $\mathcal{P}(\lambda)$.

Theorem 1.2 ([14, Theorem 1.6]). *Suppose that* $\lambda = \sum_{i=1}^r a_i\varpi_i \in \mathfrak{X}_+$. *Then*

(1) $\mathcal{I}_\lambda = \{\lambda - \sum_{i=1}^r b_i\alpha_i \mid 0 \leqslant b_i \leqslant a_i \ \ \forall i\}$;

(2) $\dim \mathsf{R}(\lambda)^{U'}_\mu = 1$ *for all* $\mu \in \mathcal{I}_\lambda$, *i.e.,* $\mathsf{R}(\lambda)^{U'}$ *is a multiplicity free* T*-module;*

(3) A nonzero U'*-invariant of weight* $\lambda - \sum_{i=1}^r a_i\alpha_i$, *say* $\boldsymbol{f}$, *is a cyclic vector of the* U/U'*-module* $\mathsf{R}(\lambda)^{U'}$. *That is, the vectors* $\{(\prod_{i=1}^r e_i^{b_i})(\boldsymbol{f}) \mid 0 \leqslant b_i \leqslant a_i \ \ \forall i\}$ *form a basis for* $\mathsf{R}(\lambda)^{U'}$.

It follows from (1) and (2) that $\dim \mathsf{R}(\lambda)^{U'} = \prod_{i=1}^r (a_i + 1)$. In particular, $\dim \mathsf{R}(\varpi_i)^{U'} = 2$. The weight spaces $\mathsf{R}(\varpi_i)_{\varpi_i}$ and $\mathsf{R}(\varpi_i)_{\varpi_i-\alpha_i}$ are one-dimensional, and we fix corresponding nonzero weight vectors f_i, $\tilde{f}_i$ such that $e_i(\tilde{f}_i) = f_i$. That is, $\tilde{f}_i$ is a cyclic vector of $\mathsf{R}(\varpi_i)^{U'}$.

The biggest $\mathcal{S}$-variety corresponds to the monoid $\mathfrak{S} = \mathfrak{X}_+$. Here

$$\Bbbk[G/U] = \Bbbk[\mathscr{C}(\mathfrak{X}_+)] = \bigoplus_{\lambda\in\mathfrak{X}_+} \mathsf{R}(\lambda),$$

and the multiplicative structure of $\Bbbk[\mathscr{C}(\mathfrak{X}_+)]$ together with Theorem 1.2 imply

Theorem 1.3 (cf. [14, Theorem 1.8]). *The algebra of* U'*-invariants* $\Bbbk[\mathscr{C}(\mathfrak{X}_+)]^{U'}$ *is freely generated by* $f_1, \tilde{f}_1, \ldots, f_r, \tilde{f}_r$. *Therefore, any basis for the* $2r$*-dimensional vector space* $\bigoplus_{i=1}^r \mathsf{R}(\varpi_i)^{U'}$ *yields a free generating system for* $\Bbbk[\mathscr{C}(\mathfrak{X}_+)]^{U'}$.

The algebra $\Bbbk[G/U]$ is sometimes called the *flag algebra* for G, because it can be realized as the multi-homogeneous coordinate ring of the flag variety G/B. More generally, we have

Theorem 1.4. *If* $\mathfrak{S}$ *is generated by some fundamental weights, say* $\{\varpi_i \mid i \in M\}$, *then any basis for* $\bigoplus_{i\in M} \mathsf{R}(\varpi_i)^{U'}$ *yields a free generating system for* $\Bbbk[\mathscr{C}(\mathfrak{S})]^{U'}$.

Proof. As in the proof of [14, Theorem 1.8], one observes that, for $\lambda = \sum_{i\in M} a_i\varpi_i$, the monomials $\{\prod_{i\in M} f_i^{b_i} \tilde{f}_i^{a_i-b_i} \mid 0 \leqslant b_i \leqslant a_i\}$ form a basis for the space $\mathsf{R}(\lambda)^{U'}$. [Another way is to consider the natural embedding $\mathscr{C}(\mathfrak{S}) \hookrightarrow \mathscr{C}(\mathfrak{X}_+)$ [22] and the surjective homomorphism $\Bbbk[\mathscr{C}(\mathfrak{X}_+)]^{U'} \to \Bbbk[\mathscr{C}(\mathfrak{S})]^{U'}$.] □

Given $\lambda \in \mathfrak{X}_+$, we always consider a basis for $\mathsf{R}(\lambda)^{U'}$ generated by a cyclic vector and elements $e_i \in \mathfrak{g}_{\alpha_i}$, i.e., a basis $\{f_\mu \in \mathsf{R}(\lambda)_\mu \mid \mu \in \mathcal{I}_\lambda\}$ such that

$$e_i(f_\mu) = \begin{cases} f_{\mu+\alpha_i}, & \mu+\alpha_i \in \mathcal{I}_\lambda, \\ 0, & \mu+\alpha_i \notin \mathcal{I}_\lambda. \end{cases}$$

However, for the fundamental G-modules $\mathsf{R}(\varpi_i)$, we write f_i in place of f_{ϖ_i} and $\tilde{f}_i$ in place of $f_{\varpi_i-\alpha_i}$.

1.3. ***Equidimensional morphisms and conical varieties.*** Let $\pi : X \to Y$ be a dominant morphism of irreducible algebraic varieties. We say that π *is equidimensional at* $y \in Y$ if all irreducible components of $\pi^{-1}(y)$ are of dimension $\dim X - \dim Y$. Then π is said to be *equidimensional* if it is equidimensional at any $y \in \pi(X)$. By a result of Chevalley [6, Ch. 5, n.5, Prop. 3], if $y = \pi(x)$ is a normal point, π is equidimensional at y, and $\Omega \subset X$ is a neighbourhood of x, then $\pi(\Omega)$ is a neighbourhood of y. Consequently, an equidimensional morphism to a normal variety is open.

An affine variety X is said to be *conical* if $\Bbbk[X]$ is $\mathbb{N}$-graded, $\Bbbk[X] = \bigoplus_{n\geqslant 0} \Bbbk[X]_n$, and $\Bbbk[X]_0 = \Bbbk$. Then the point x_0 corresponding to the maximal ideal $\bigoplus_{n\geqslant 1} \Bbbk[X]_n$ is called the *vertex*. Geometrically, this means that X is equipped with an action of the multiplicative group $\Bbbk^\times$ such that $\{x_0\}$ is the only closed $\Bbbk^\times$-orbit in X.

Lemma 1.5. *Suppose that both X and Y are conical, and $\pi : X \to Y$ is dominant and $\Bbbk^\times$-equivariant. (Then $\pi(x_0) =: y_0$ is the vertex in Y.) If Y is normal and π is equidimensional at y_0, then π is onto and equidimensional.*

This readily follows from the above-mentioned result of Chevalley and standard inequalities for the dimension of fibres.

Remark 1.6. As $\mathfrak{S}$ lies in an open half-space of $\mathfrak{X}_{\mathbb{R}}$, taking a suitable $\mathbb{N}$-specialisation of the multi-grading of $\Bbbk[\mathscr{C}(\mathfrak{S})]$ shows that $\mathscr{C}(\mathfrak{S})$ is conical and the origin in $\mathsf{R}(\lambda_1^*) \oplus \dots \oplus \mathsf{R}(\lambda_m^*)$ is its vertex. This implies that $\mathscr{C}(\mathfrak{S})/\!\!/U'$ is conical, too. We will apply the above lemma to the study of equidimensional quotient maps $\pi : \mathscr{C}(\mathfrak{S}) \to \mathscr{C}(\mathfrak{S})/\!\!/U'$. It is important that such π appears to be onto.

The idea of applying Chevalley's result to the study of equidimensional quotients (by U) is due to Vinberg and Gindikin [20].

2. Actions of U' on HV-varieties

Let $\mathscr{C}(\lambda) = \overline{G{\cdot}v_{-\lambda}} \subset \mathsf{R}(\lambda^*)$ be an HV-variety. The algebra $\Bbbk[\mathscr{C}(\lambda)]$ is $\mathbb{N}$-graded and its component of degree n is $\mathsf{R}(n\lambda)$. Since $\mathscr{C}(\lambda)$ is normal, $\mathscr{C}(\lambda)/\!\!/U'$ is normal, too.

Theorem 2.1. *$\mathscr{C}(\lambda)/\!\!/U'$ is an affine space if and only if λ is a fundamental weight.*

Proof. 1) Suppose that λ is not fundamental, i.e., $\lambda = \dots + a\varpi_i + b\varpi_j + \dots$ with $a, b \geqslant 1$.

• If $i \neq j$, then $\mathsf{R}(\lambda)^{U'}$ contains linearly independent vectors f_λ, $f_{\lambda-\alpha_i}$, $f_{\lambda-\alpha_j}$, $f_{\lambda-\alpha_i-\alpha_j}$ that occur in any minimal generating system, since $\Bbbk[\mathscr{C}(\lambda)]_1 \simeq \mathsf{R}(\lambda)$. Using the relations $e_i(f_{\lambda-\alpha_i-\alpha_j}) = f_{\lambda-\alpha_j}$, etc., one easily verifies that

$$p = f_\lambda f_{\lambda-\alpha_i-\alpha_j} - f_{\lambda-\alpha_i} f_{\lambda-\alpha_j}$$

is a U-invariant function on $\mathscr{C}(\lambda)$, of degree 2. The only highest weight in degree 2 is 2λ. Since the weight of p is not 2λ, we must have $p \equiv 0$, and this is a non-trivial relation.

• If $i = j$, then the coefficient of ϖ_i is at least 2 and we consider vectors f_λ, $f_{\lambda-\alpha_i}$, $f_{\lambda-2\alpha_i} \in \mathsf{R}(\lambda)^{U'}$. Then $\tilde{p} = 2f_\lambda f_{\lambda-2\alpha_i} - f_{\lambda-\alpha_i}^2$ is a U-invariant function of degree 2 and weight $2(\lambda-\alpha_i)$, and this yields the relation $\tilde{p} = 0$ in $\Bbbk[\mathscr{C}(\lambda)]^{U'}$.

2) If $\lambda = \varpi_i$, then $\dim \mathsf{R}(\varpi_i)^{U'} = 2$ and $\mathscr{C}(\varpi_i)/\!\!/U' \simeq \mathbb{A}^2$ by Theorem 1.4. □

For an affine variety X, let $\operatorname{edim} X$ denote the minimal number of generators of $\Bbbk[X]$ and $\mathsf{hd}(X)$ the homological dimension of $\Bbbk[X]$. If $\Bbbk[X]$ is a graded Cohen-Macaulay algebra, then $\mathsf{hd}(X) = \operatorname{edim} X - \dim X$ [17, Ch. IV].

Theorem 2.2. *If* $\lambda = \sum_{i=1}^r a_i\varpi_i \in \mathfrak{X}_+$, *then*

(i) $\dim \mathscr{C}(\lambda)/\!\!/U' = 1 + \#\{j \mid a_j \neq 0\}$;

(ii) *the graded algebra* $\Bbbk[\mathscr{C}(\lambda)]^{U'}$ *is generated by functions of degree one, i.e., by the space* $\mathsf{R}(\lambda)^{U'}$, *and* $\operatorname{edim} \mathscr{C}(\lambda)/\!\!/U' = \prod_{i=1}^r (a_i+1)$.

Proof. (i) Recall that $P_\lambda = L_\lambda N_\lambda$ is the standard parabolic subgroup associated with $\mathscr{C}(\lambda)$ and the simple roots of L_λ are those orthogonal to λ. Set $k = \#\{j \mid a_j \neq 0\}$. Then $\mathsf{srk}\, L_\lambda := \mathsf{rk}\,(L_\lambda, L_\lambda) = \mathsf{rk}\, G - k$ and $\dim \mathscr{C}(\lambda) = \dim N_\lambda + 1$. Since $U\cdot(\Bbbk v_{-\lambda})$ is dense in $\mathscr{C}(\lambda)$, $U(L_\lambda) := U \cap L_\lambda$ is a generic stabiliser for the U-action on $\mathscr{C}(\lambda)$. By [14, Lemma 2.5], the minimal dimension of stabilisers for the U'-action on $\mathscr{C}(\lambda)$ equals $\dim(U(L_\lambda) \cap U') = \dim U(L_\lambda) - \mathsf{srk}\, L_\lambda$. Consequently,

$$\dim \mathscr{C}(\lambda)/\!\!/U' = \dim \mathscr{C}(\lambda) - \dim U' + \min_{x\in\mathscr{C}(\lambda)} \dim U'_x =$$
$$= \dim N_\lambda + 1 - (\dim U - \mathsf{rk}\, G) + (\dim U(L_\lambda) - \mathsf{srk}\, L_\lambda) = 1 + \mathsf{rk}\, G - \mathsf{srk}\, L_\lambda = 1 + k.$$

(ii) By Theorem 1.2, $\dim \mathsf{R}(\lambda)^{U'} = \prod_{i=1}^r (a_i+1)$, which shows that $\operatorname{edim} \mathscr{C}(\lambda)/\!\!/U' \geqslant \prod_{i=1}^r (a_i+1)$. Therefore, it suffices to prove that the graded algebra $\Bbbk[\mathscr{C}(\lambda)]^{U'}$ is generated by elements of degree 1. The weights of U'-invariants of degree n are

$$\mathscr{I}_{n\lambda} = \{n\lambda - \sum_i b_i\alpha_i \mid b_i = 0, 1, \ldots, na_i\}.$$

In particular,

$$\mathscr{I}_\lambda = \{\lambda - \sum_i b_i\alpha_i \mid b_i = 0, 1, \ldots, a_i\}.$$

Obviously, each element of $\mathscr{I}_{n\lambda}$ is a sum of n elements of $\mathscr{I}_\lambda$. Since $\mathsf{R}(n\lambda)^{U'}$ is a multiplicity free T-module, this space is spanned by products of n elements of $\mathsf{R}(\lambda)^{U'}$. □

Corollary 2.3. *We have* $\mathsf{hd}(\mathscr{C}(\lambda)/\!\!/U') = \prod_{i=1}^{r}(1+a_i)-1-\#\{j \mid a_j \neq 0\}$. *Therefore,*

- $\mathsf{hd}(\mathscr{C}(\lambda)/\!\!/U') = 0$ *if and only if* λ *is fundamental;*
- $\mathsf{hd}(\mathscr{C}(\lambda)/\!\!/U') = 1$ *if and only if* $\lambda = \varpi_i + \varpi_j$ *or* $2\varpi_i$.

Proof. As it was mentioned above, the HV-varieties have rational singularities. In view of [14, Theorem 2.3], $\mathscr{C}(\lambda)/\!\!/U'$ also has rational singularities and in particular is Cohen-Macaulay. Hence $\mathsf{hd}(\mathscr{C}(\lambda)/\!\!/U') = \mathrm{edim}\,\mathscr{C}(\lambda)/\!\!/U' - \dim \mathscr{C}(\lambda)/\!\!/U'$. □

Remark 2.4. 1) As above, $k = \mathsf{rk}\,G - \mathsf{srk}\,L_\lambda$ and hence $\dim \mathscr{C}(\lambda)/\!\!/U' = k+1$. Another consequence of Theorems 1.2 and 2.2 is that $\mathscr{C}(\lambda)/\!\!/U'$ is a toric variety with respect to $\Bbbk^\times \times T$, where $\Bbbk^\times$ acts on $\mathsf{R}(\lambda^*)$ (and hence on $\mathscr{C}(\lambda)$) by homotheties. Note that the T-action on $\mathscr{C}(\lambda)/\!\!/U'$ has a non-effectivity kernel of dimension $\mathsf{rk}\,G - k$. The quotient morphism $\pi_{\mathscr{C}(\lambda),U'}$ has the following description. Let $\mathsf{ann}(\mathsf{R}(\lambda)^{U'})$ be the annihilator of $\mathsf{R}(\lambda)^{U'}$ in $\mathsf{R}(\lambda^*)$. Then $(\mathsf{R}(\lambda)^{U'})^* = \mathsf{R}(\lambda^*)/\mathsf{ann}(\mathsf{R}(\lambda)^{U'})$ and $\pi_{\mathscr{C}(\lambda),U'}$ is the restriction to $\mathscr{C}(\lambda)$ of the projection $\mathsf{R}(\lambda^*) \to (\mathsf{R}(\lambda)^{U'})^*$. Thus, $\mathscr{C}(\lambda)/\!\!/U'$ is embedded in the vector space $(\mathsf{R}(\lambda)^{U'})^*$. Consequently, $\mathbb{P}(\mathscr{C}(\lambda)/\!\!/U') \subset \mathbb{P}((\mathsf{R}(\lambda)^{U'})^*)$ is a normal toric variety with respect to T. As is well-known, a projective toric T-variety can be described via a convex polytope in $\mathfrak{X}_{\mathbb{Q}}$ [7, 5.8]. The polytope corresponding to $\mathbb{P}(\mathscr{C}(\lambda)/\!\!/U')$ is the convex hull of $\mathscr{I}_\lambda$. It is a k-dimensional parallelepiped, in particular, a simple polytope. It follows that the corresponding complete fan is simplicial. Therefore the complex cohomology of $\mathbb{P}(\mathscr{C}(\lambda)/\!\!/U')$ satisfies Poincaré duality and has a number of other good properties, see [7, § 14].

2) Along with the toric structure (i.e., a dense T-orbit), the projective variety $\mathbb{P}(\mathscr{C}(\lambda)/\!\!/U')$ also has a dense orbit of the commutative unipotent group U/U'.

3. Actions of U' on arbitrary $\mathscr{S}$-varieties

Let $\mathscr{C}(\mathfrak{S})$ be an $\mathscr{S}$-variety. In this section, we answer the following questions:

– When is $\mathscr{C}(\mathfrak{S})/\!\!/U'$ an affine space?

– Suppose that $\mathscr{C}(\mathfrak{S})/\!\!/U'$ is an affine space. When is $\pi_{\mathscr{C}(\mathfrak{S}),U'}$ equidimensional?

We begin with a formula for $\dim \mathscr{C}(\mathfrak{S})/\!\!/U'$, which generalises Theorem 2.2(i).

Proposition 3.1. $\dim \mathscr{C}(\mathfrak{S})/\!\!/U' = \mathsf{rk}\,\mathfrak{S} + (\mathsf{rk}\,G - \mathsf{srk}\,L_{\mathfrak{S}})$.

Proof. By Theorem 1.1, $\dim \mathscr{C}(\mathfrak{S}) = \dim N_{\mathfrak{S}} + \mathsf{rk}\,\mathfrak{S}$ and $\dim \mathscr{C}(\mathfrak{S})/\!\!/U = \mathsf{rk}\,\mathfrak{S}$. This readily implies that $U(L_{\mathfrak{S}}) := U \cap L_{\mathfrak{S}}$ is a generic stabiliser for the U-action on $\mathscr{C}(\mathfrak{S})$. By [14, Lemma 2.5], the minimal dimension of stabilisers for the U'-action on $\mathscr{C}(\mathfrak{S})$ equals $\dim(U(L_{\mathfrak{S}}) \cap U') = \dim U(L_{\mathfrak{S}}) - \mathsf{srk}\,L_{\mathfrak{S}}$. Consequently,

$$\dim \mathscr{C}(\mathfrak{S})/\!\!/U' = \dim \mathscr{C}(\mathfrak{S}) - \dim U' + \min_{x \in \mathscr{C}(\mathfrak{S})} \dim U'_x =$$

$$= \dim N_{\mathfrak{S}} + \mathsf{rk}\,\mathfrak{S} - (\dim U - \mathsf{rk}\,G) + (\dim U(L_{\mathfrak{S}}) - \mathsf{srk}\,L_{\mathfrak{S}}) = \mathsf{rk}\,\mathfrak{S} + (\mathsf{rk}\,G - \mathsf{srk}\,L_{\mathfrak{S}}).$$

Here we use the fact that U is a semi-direct product of $N_{\mathfrak{S}}$ and $U(L_{\mathfrak{S}})$. □

Remark. Note that $\mathsf{rk}\,\mathfrak{S} \leqslant \mathsf{rk}\,G - \mathsf{srk}\,L_{\mathfrak{S}}$, and the equality here is equivalent to the fact that the space $\langle\mathfrak{S}\rangle$ has a basis that consists of fundamental weights.

Theorem 3.2. *Let $\mathfrak{S} \subset \mathfrak{X}_+$ be an arbitrary finitely generated monoid. Then $\mathscr{C}(\mathfrak{S})/\!\!/U'$ is an affine space if and only if $\mathfrak{S}$ is generated by fundamental weights.*

Proof. 1) Suppose that $\mathscr{C}(\mathfrak{S})/\!\!/U'$ is an affine space. If λ is a generator of $\mathfrak{S}$, then any generating system of $\Bbbk[\mathscr{C}(\mathfrak{S})]^{U'}$ contains a basis for $\mathsf{R}(\lambda)^{U'}$. Arguing as in the proof of Theorem 2.1, we conclude that λ must be a fundamental weight. [Another way is to use Proposition 3.1 and the inequality $\dim \mathscr{C}(\mathfrak{S})/\!\!/U' \geqslant 2\mathsf{rk}\,\mathfrak{S}$.]

2) The converse is contained in Theorem 1.4. □

In the rest of this section, we only consider monoids generated by fundamental weights. Fix a numbering of the simple roots (fundamental weights). For any $M \subset \{1, 2, \dots, r\}$, let $\mathscr{C}(M)$ denote the $\mathcal{S}$-variety corresponding to the monoid $\mathfrak{S} = \sum_{i\in M} \mathbb{N}\varpi_i$. Our aim is to characterise the subsets M having the property that $\pi_{U'} : \mathscr{C}(M) \to \mathscr{C}(M)/\!\!/U'$ is equidimensional. The origin (vertex) is the only G-fixed point of $\mathscr{C}(M)$ and the corresponding fibre of $\pi_{U'}$ (the *null-cone*) is denoted by $\mathfrak{N}_{U'}(M)$.

Recall that $\Bbbk[\mathscr{C}(M)]$ is a graded Cohen-Macaulay ring and $\Bbbk[\mathscr{C}(M)]^{U'}$ is a polynomial algebra freely generated by $\{f_i, \tilde{f}_i \mid i \in M\}$ (Theorem 1.4). Therefore, $\pi_{U'}$ is equidimensional *if and only if* the functions $\{f_i, \tilde{f}_i \mid i \in M\}$ form a regular sequence in $\Bbbk[\mathscr{C}(M)]$ *if and only if* $\dim \mathfrak{N}_{U'}(M) = \dim \mathscr{C}(M) - 2(\#M)$ [16, § 17].

Definition 1. A subset $M \subset \{1, \dots, r\}$ is said to be *sparse*, if 1) the roots α_i with $i \in M$ are pairwise orthogonal, i.e., disjoint in the Dynkin diagram; 2) there are no $i, j \in M$ and no $k \notin M$ such that $(\alpha_k, \alpha_i) < 0$ and $(\alpha_k, \alpha_j) < 0$, i.e., α_k is adjacent to both α_i and α_j.
Accordingly, we say that a certain set of fundamental weights (simple roots) is *sparse*.

Clearly, if M is sparse and $J \subset M$, then J is also sparse.

Lemma 3.3. *Let $\alpha_{i_1}, \dots, \alpha_{i_l}$ be a sequence of different simple roots such that $\alpha_{i_j}, \alpha_{i_{j+1}}$ are adjacent for $j = 1, 2, \dots, l-1$). Then $\mu := \varpi_{i_1} - \sum_{j=1}^{l} \alpha_{i_j}$ is a weight of $\mathsf{R}(\varpi_{i_1})$ and $\dim \mathsf{R}(\varpi_{i_1})_\mu = 1$.*

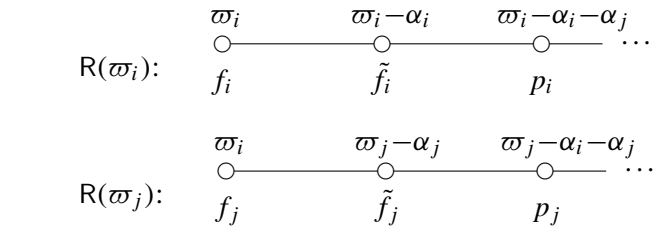
Proof. The first assertion is easily proved by induction on l. The second assertion follows from [1, Prop. 2.2] □

Theorem 3.4. *If the quotient $\pi_{U'} : \mathscr{C}(M) \to \mathscr{C}(M)/\!\!/U'$ is equidimensional, then M is sparse.*

Proof. As we already know, $\Bbbk[\mathscr{C}(M)]^{U'}$ is freely generated by the functions $\{f_i, \tilde{f}_i \mid i \in M\}$. Assuming that M is not sparse, we point out certain relations in $\Bbbk[\mathscr{C}(M)]$,

which show that these free generators do not form a regular sequence. There are two possibilities for that.

- Suppose first that α_i and α_j are adjacent simple roots for some $i, j \in M$. Then $\lambda_{ij} := \varpi_i + \varpi_j - \alpha_i - \alpha_j$ is dominant. Consider upper parts of the Hasse diagrams of weight posets for $\mathsf{R}(\varpi_i)$ and $\mathsf{R}(\varpi_j)$:

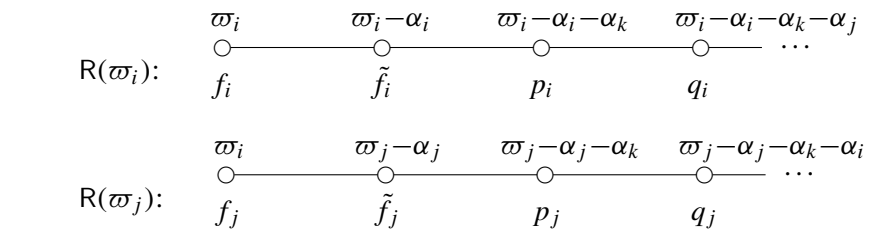


In these figures, each node depicts a weight space, and we put the weight over the node and a weight vector under the node. There can be other edges incident to the node $\varpi_i - \alpha_i$ (if there exist other simple roots adjacent to α_i), but we do not need them. By Lemma 3.3, the weight spaces $\mathsf{R}(\varpi_i)_{\varpi_i}$, $\mathsf{R}(\varpi_i)_{\varpi_i-\alpha_i}$, and $\mathsf{R}(\varpi_i)_{\varpi_i-\alpha_i-\alpha_j}$ are one-dimensional. Here f_i, $\tilde{f}_i$, and p_i are normalised such that $e_i(\tilde{f}_i) = f_i$ and $e_j(p_i) = \tilde{f}_i$; and likewise for $\mathsf{R}(\varpi_j)$. Note also that $e_i(p_i) = 0$, since $\varpi_i - \alpha_j$ is not a weight of $\mathsf{R}(\varpi_i)$. It is then easily seen that

$$f_i \otimes p_j - \tilde{f}_i \otimes \tilde{f}_j + p_i \otimes f_j$$

is a U-invariant of weight λ_{ij} in $\mathsf{R}(\varpi_i) \otimes \mathsf{R}(\varpi_j)$. However, only the Cartan component of $\mathsf{R}(\varpi_i) \otimes \mathsf{R}(\varpi_j)$ survives in the algebra $\Bbbk[\mathcal{C}(M)]$, i.e., in the product $\mathsf{R}(\varpi_i)\cdot\mathsf{R}(\varpi_j)$. Consequently, $f_i p_j - \tilde{f}_i \tilde{f}_j + p_i f_j = 0$ in $\Bbbk[\mathcal{C}(M)]$. This means that $(f_i, f_j, \tilde{f}_i, \tilde{f}_j)$ is not a regular sequence in $\Bbbk[\mathcal{C}(M)]$.

- Yet another possibility is that there are $k \notin M$ and $i, j \in M$ such that α_k is adjacent to both α_i and α_j. Here one verifies that $\tilde{\lambda}_{ij} := \varpi_i + \varpi_j - \alpha_i - \alpha_k - \alpha_j$ is dominant. In this situation, we need larger fragments of the weight posets:



Here all the weight spaces are one-dimensional by Lemma 3.3, and we follow the same conventions as above. Additionally, we assume that $e_j(q_i) = p_i$. Note that $e_k(q_i) = 0$ and $e_i(q_i) = 0$, since neither $\varpi_i - \alpha_i - \alpha_j$ nor $\varpi_i - \alpha_k - \alpha_j$ is a weight of $\mathsf{R}(\varpi_i)$. (And likewise for $\mathsf{R}(\varpi_j)$.) Then $f_i \otimes q_j - \tilde{f}_i \otimes p_j + p_i \otimes \tilde{f}_j - q_i \otimes f_j$ is a

U-invariant of weight $\tilde{\lambda}_{ij}$, and hence

$$f_i q_j - \tilde{f}_i p_j + p_i \tilde{f}_j - q_i f_j = 0 \tag{3·1}$$

in $\Bbbk[\mathscr{C}(M)]$ for the same reason as above. This again implies that $(f_i, f_j, \tilde{f}_i, \tilde{f}_j)$ is not a regular sequence in $\Bbbk[\mathscr{C}(M)]$. □

Example 3.5. Let $\mathfrak{g} = \mathfrak{sl}_4$ and $M = \{1, 3\}$ in the usual numbering of Π. Then $\dim \mathsf{R}(\varpi_1) = \dim \mathsf{R}(\varpi_3) = 4$ and $\dim \mathscr{C}(M) = 7$. In this case, the above 4-node fragments provide the whole weight posets. Therefore, $\mathsf{R}(\varpi_1) = \langle f_1, \tilde{f}_1, p_1, q_1\rangle$, $\mathsf{R}(\varpi_3) = \langle f_3, \tilde{f}_3, p_3, q_3\rangle$, and (3·1) with $(i, j) = (1, 3)$ is the equation of the hypersurface $\mathscr{C}(M)$. Since $\dim \mathscr{C}(M)/\!\!/U' = 4$ and $\mathfrak{N}_{U'}(M) \supset \langle p_1, q_1, p_3, q_3\rangle$, the morphism $\pi_{U'}$ is not equidimensional.

To prove the converse to Theorem 3.4, we need some preparations. Recall that the partial order "$\preccurlyeq$" is defined in 1.2. We also write $\nu \prec \mu$ if $\nu \preccurlyeq \mu$ and $\mu \neq \nu$.

Lemma 3.6. *Suppose that M is sparse and $w \in W$ has the property that $w(\varpi_i) \prec \varpi_i - \alpha_i$ for all $i \in M$. Then $\ell(w) \geqslant 2\cdot\#(M)$.*

Proof. Since $w(\varpi_i) \prec \varpi_i$, any reduced decomposition of w contains s_i. Furthermore, since $w(\varpi_i) \prec \varpi_i - \alpha_i$, there exists a node i' adjacent to i such that $w(\varpi_i) \preccurlyeq \varpi_i - \alpha_i - \alpha_{i'}$. Therefore, w must also contain the reflection $s_{i'}$. Because M is sparse, all the reflections $\{s_i, s_{i'} \mid i \in M\}$ are different. Thus, $\ell(w) \geqslant 2\cdot\#(M)$. □

For any $I \subset \Pi$, we consider the following objects. Let $P_I = L_I N_I$ be the standard parabolic subgroup of G. Here L_I is the Levi subgroup whose set of simple roots is I and N_I is the unipotent radical of P_I. Then $P_I^- = L_I N_I^-$ is the opposite parabolic subgroup of G. We also need the factorisation

$$W = W^I \times W_I,$$

where W_I is the subgroup generated by $\{s_i \mid \alpha_i \in I\}$ and W^I is the set of representatives of minimal length for W/W_I [8, 1.10]. It is also true that $W^I = \{w \in W \mid w(\alpha_i) \in \Delta^+ \ \forall \alpha_i \in I\}$ [8, 5.4]. If $I = \{\alpha \in \Pi \mid (\alpha, \lambda) = 0\}$ for some $\lambda \in \mathfrak{X}_+$, then we write P_λ, W_λ, W^λ, etc.

For each $w \in W$, we fix a representative, $\dot{w}$, in $N_G(T)$. As is well-known, the U-orbits in G/P_I^- can be parametrised by W^I, and letting $\mathbb{O}(w) = U\dot{w}P_I^- \subset G/P_I^-$ $(w \in W^I)$, we have $G/P_I^- = \sqcup_{w\in W^I}\mathbb{O}(w)$ and $\operatorname{codim}\mathbb{O}(w) = \ell(w)$.

Theorem 3.7. *If $M \subset \{1, \dots, r\}$ is sparse, then the quotient $\pi_{U'} : \mathscr{C}(M) \to \mathscr{C}(M)/\!\!/U'$ is equidimensional.*

Proof. Set $m = \#M$ and $I = \Pi \setminus \{\alpha_i \mid i \in M\}$. Consider $\boldsymbol{v} = \sum_{i\in M} v_{-\varpi_i} \in \bigoplus_{i\in M} \mathsf{R}(\varpi_i^*)$. As explained in Subsection 1.1, then $\mathscr{C}(M) \simeq \overline{G\cdot\boldsymbol{v}}$ and $\dim \mathscr{C}(M) =$

$\dim G/P_I^- + m$. We also have $\dim \mathcal{C}(M)/\!\!/U' = 2m$. Therefore, our goal is to prove that $\dim \mathfrak{N}_{U'}(M) \leqslant \dim G/P_I^- - m$.

Set $V = \overline{T\cdot \boldsymbol{v}} = \bigoplus_{i\in M} \Bbbk v_{-\varpi_i}$. It is an m-dimensional subspace of $\bigoplus_{i\in M} \mathsf{R}(\varpi_i^*)$, which is contained in $\mathcal{C}(M)$ and is P_I^--stable. Recall that $G\times_{P_I^-} V$ is a homogeneous vector bundle on G/P_I^-. A typical element of it is denoted by $g * v$, where $g \in G$ and $v = \sum_{i\in M} v_i \in V$. Our main tool for estimating $\dim \mathfrak{N}_{U'}(M)$ is the following diagram:

$$\begin{array}{ccc} G\times_{P_I^-} V & \xrightarrow{\ \tau\ } & \mathcal{C}(M) \\ \big\downarrow \phi & & \big\downarrow \pi_{U'} \\ G/P_I^- & & \mathcal{C}(M)/\!\!/U' \end{array}$$

where $\phi(g*v) := gP_I^-$ and $\tau(g*v) := g\cdot v$. Note that $\mathfrak{N}_{U'}(M)$ is B-stable, and hence so is $\tau^{-1}(\mathfrak{N}_{U'}(M))$. It is easily seen that the morphism τ is birational and therefore it is an equivariant resolution of singularities of $\mathcal{C}(M)$.

Let $n \in U$ and $w \in W^I$. As $\Bbbk[\mathcal{C}(M)]^{U'}$ is generated by $\{f_i, \tilde{f}_i \mid i \in M\}$, we have

(3·2)

$$\phi^{-1}(n\dot{w}P_I^-) \cap \tau^{-1}(\mathfrak{N}_{U'}(M)) = \{n\dot{w} * v \mid f_i(n\dot{w}\cdot v) = 0,\ \tilde{f}_i(n\dot{w}\cdot v) = 0 \quad \forall i \in M\}.$$

Here f_i (resp. $\tilde{f}_i$) is regarded as the coordinate of $v_{-\varpi_i} \in \mathsf{R}(\varpi_i^*)$ (resp. $v_{-\varpi_i+\alpha_i} \in \mathsf{R}(\varpi_i^*)$). Note that $f_i(n\dot{w}\cdot v)$ depends only on the component v_i of v, and v_i is proportional to $v_{-\varpi_i}$. Let us simplify condition (3·2). Since f_i is actually a U-invariant, we have $f_i(n\dot{w}\cdot v_i) = f_i(\dot{w}\cdot v_i)$. Next, $\tilde{f}_i$ is invariant with respect to a subgroup of codimension 1 in U. Namely, consider the decomposition $U = U^{\alpha_i}U_{\alpha_i} \simeq U^{\alpha_i} \times U_{\alpha_i}$, where U_{α_i} is the root subgroup and U^{α_i} is the unipotent radical of the minimal parabolic subgroup associated with α_i. If $n_i \in U_{\alpha_i}$ and $\tilde{n} \in U^{\alpha_i}$, then $\tilde{n}\cdot\tilde{f}_i = \tilde{f}_i$ and $n_i^{-1}\cdot\tilde{f}_i = \tilde{f}_i + c_i f_i$ for some $c_i = c_i(n_i) \in \Bbbk$. Hence for $n = \tilde{n}n_i \in U$, we have

$$\tilde{f}_i(n\dot{w}\cdot v_i) = \tilde{f}_i(n_i\dot{w}\cdot v_i) = (n_i^{-1}\cdot\tilde{f}_i)(\dot{w}\cdot v_i) = \tilde{f}_i(\dot{w}\cdot v_i) + f_i(\dot{w}\cdot v_i)c_i\,.$$

Therefore, (3·2) reduces to the following:

(3·3)

$$\phi^{-1}(n\dot{w}P_I^-) \cap \tau^{-1}(\mathfrak{N}_{U'}(M)) = \{n\dot{w} * v \mid f_i(\dot{w}\cdot v_i) = 0,\ \tilde{f}_i(\dot{w}\cdot v_i) = 0 \quad \forall i \in M\}.$$

Thus, the dimension of this intersection does not depend on $n \in U$; it depends only on $w \in W^I$, i.e., on $\mathcal{O}(w) \subset G/P_I^-$. We can make (3·3) more precise by using the partition of $\mathcal{C}(M)$ into (finitely many) G-orbits. For any subset $J \subset M$, let $\boldsymbol{v}_J = \sum_{i\in J} v_{-\varpi_i} \in V$. Then $\{\boldsymbol{v}_J \mid J \subset M\}$ is a complete set of representatives of the G-orbits in $\mathcal{C}(M)$ (Theorem 1.1(2)). Set $\overset{\circ}{V}_J = G\cdot\boldsymbol{v}_J \cap V = T\cdot\boldsymbol{v}_J$. It is an open

subset of a $(\#J)$-dimensional vector space. Then

$$\phi^{-1}(n\dot{w}P_I^-)\cap\tau^{-1}(\mathfrak{N}_{U'}(M)\cap G{\cdot}\boldsymbol{v}_J)$$
$$=\{n\dot{w}*v\mid v\in\overset{\circ}{V}_J,\ \ f_i(\dot{w}{\cdot}v_i)=0,\ \ \tilde{f}_i(\dot{w}{\cdot}v_i)=0\ \forall i\in M\}.$$

This set is non-empty if and only if $\dot{w}\cdot v_{-\varpi_i}$ has the trivial projection to $\langle v_{-\varpi_i}, v_{-\varpi_i+\alpha_i}\rangle\subset \mathsf{R}(\varpi_i^*)$ for all $i\in J$, i.e., $w(\varpi_i)\prec\varpi_i-\alpha_i$ for all $i\in J$. In this case the dimension of this set equals $\dim\overset{\circ}{V}_J=\#J$. Consequently, if $\phi^{-1}(\mathbb{O}(w))\cap\tau^{-1}(\mathfrak{N}_{U'}(M)\cap G{\cdot}\boldsymbol{v}_J)\neq\varnothing$, then

$$w(\varpi_i)\prec\varpi_i-\alpha_i\ \text{ for all }\ i\in J\ \text{ and}$$
$$\dim\Big(\phi^{-1}(\mathbb{O}(w))\cap\tau^{-1}(\mathfrak{N}_{U'}(M)\cap G{\cdot}\boldsymbol{v}_J)\Big)=\#J+\dim\mathbb{O}(w).$$

By Lemma 3.6, $\ell(w)\geqslant 2{\cdot}\#J$. Therefore,

$$\dim\Big(\phi^{-1}(\mathbb{O}(w))\cap\tau^{-1}(\mathfrak{N}_{U'}(M)\cap G{\cdot}\boldsymbol{v}_J)\Big)=$$
$$\#J-\operatorname{codim}\mathbb{O}(w)+\dim G/P_I^-=\#J-\ell(w)+\dim G/P_I^-\leqslant\dim G/P_I^--\#J.$$

This is an upper bound for the dimension of the pullback in $G\times_{P_I^-}V$ of a subset of $\mathfrak{N}_{U'}(M)$. If $\boldsymbol{v}_J$ is not generic, i.e., $J\neq M$, then $\dim\tau^{-1}(\boldsymbol{v}_J)>0$ and the actual subset of $\mathfrak{N}_{U'}(M)$ has smaller dimension. More precisely, set $\tilde{I}=\{\alpha_i\mid i\notin J\}$. Then $\tilde{I}\supset I$ and $\tau^{-1}(\boldsymbol{v}_J)\simeq P_{\tilde{I}}^-/P_I^-$. Since $\mathsf{srk}\,(L_{\tilde{I}})=\mathsf{srk}\,(L_I)+(m-\#J)$, we have $\dim\tau^{-1}(\boldsymbol{v}_J)\geqslant m-\#J$. Thus, for all $w\in W^I$ and $J\subset M$, we have

$$\dim\Big(\tau\big(\phi^{-1}(\mathbb{O}(w))\big)\cap\mathfrak{N}_{U'}(M)\cap G{\cdot}\boldsymbol{v}_J\Big)\leqslant$$
$$\dim G/P_I^--\#J-(m-\#J)=\dim G/P_I^--m,$$

and therefore $\dim\mathfrak{N}_{U'}(M)\leqslant\dim G/P_I^--m$. □

Remark 3.8. A "dual" approach is to consider the P_I-stable subspace $\tilde{V}=\bigoplus_{i\in M}\Bbbk v_{\varpi_i^*}\subset\bigoplus_{i\in M}\mathsf{R}(\varpi_i^*)$ and the map $G\times_{P_I}\tilde{V}\to\mathfrak{C}(M)$. Then one has to work with U_--orbits in G/P_I and U_--invariants in $\Bbbk[\mathfrak{C}(M)]$, but all dimension estimates remain the same. Such an approach is realised in [13, Sect. 5], where the equidimensionality problem is considered for the actions of U on $\mathcal{S}$-varieties.

Combining Theorems 3.2, 3.4, and 3.7, we obtain the general criterion:

Theorem 3.9. *For a finitely generated monoid $\mathfrak{S}\subset\mathfrak{X}_+$, the following conditions are equivalent:*

(i) *$\mathfrak{C}(\mathfrak{S})/\!\!/U'$ is an affine space and $\pi_{\mathfrak{C}(\mathfrak{S}),U'}:\mathfrak{C}(\mathfrak{S})\to\mathfrak{C}(\mathfrak{S})/\!\!/U'$ is equidimensional;*

(ii) *$\mathfrak{S}$ is generated by a sparse set of fundamental weights.*

4. Equidimensional quotients by U'

In this section, the quotient morphism for the $\mathcal{S}$-variety $\mathscr{C}(\mathfrak{S})$ will be denoted by $\pi_{\mathfrak{S},U'}$. Similarly, for the HV-variety $\mathscr{C}(\lambda)$, we use notation $\pi_{\lambda,U'}$. Our goal is to characterise the monoids $\mathfrak{S}$ such that $\pi_{\mathfrak{S},U'}:\mathscr{C}(\mathfrak{S})\to\mathscr{C}(\mathfrak{S})/\!\!/U'$ is equidimensional (i.e., without assuming that $\mathscr{C}(\mathfrak{S})/\!\!/U'$ is an affine space). We assume that $U'\neq\{1\}$, i.e., G is not a product of several SL_2.

First, we consider the case of HV-varieties.

Theorem 4.1. *For any $\lambda\in\mathfrak{X}_+$, the null-cone $\mathfrak{N}_{U'}(\mathscr{C}(\lambda))$ is of codimension 2 in $\mathscr{C}(\lambda)$.*

Proof. As in the proof of Theorem 3.7, we work with the diagram

$$\begin{array}{ccc} G\times_{P_\lambda^-}V & \xrightarrow{\ \tau\ } & \mathscr{C}(\lambda) \\ \big\downarrow\phi & & \big\downarrow\pi_{\lambda,U'} \\ G/P_\lambda^- & & \mathscr{C}(\lambda)/\!\!/U', \end{array}$$

where $V=\Bbbk v_{-\lambda}$, $\phi(g*v):=gP_\lambda^-$ and $\tau(g*v):=g{\cdot}v$. Note that P_λ^- is just the stabiliser of the line $V\subset\mathsf{R}(\lambda^*)$. For simplicity, we write $\mathfrak{N}_{U'}(\lambda)$ in place of $\mathfrak{N}_{U'}(\mathscr{C}(\lambda))$.

Since $\mathfrak{N}_{U'}(\lambda)$ is U-stable, $\phi(\tau^{-1}(\mathfrak{N}_{U'}(\lambda)))$ is a union of U-orbits. Recall that $\Bbbk[\mathscr{C}(\lambda)]^{U'}$ is generated by the space $\mathsf{R}(\lambda)^{U'}$, and the corresponding set of T-weights is $\mathcal{J}_\lambda$.

We point out a $w\in W^\lambda$ such that the U-orbit $\mathbb{O}(w)\subset G/P_\lambda^-$ is of codimension 2 and $\phi^{-1}(\mathbb{O}(w))\subset\tau^{-1}(\mathfrak{N}_{U'}(\lambda))$. Suppose that $(\lambda,\alpha_1^\vee)=a_1\geqslant 1$ and α_1 is a simple root of a simple component of G of rank $\geqslant 2$. Let α_2 be a simple root adjacent to α_1 in the Dynkin diagram. Take $w=s_2s_1$. Regardless of the value of (λ,α_2), it is true that $w\in W^\lambda$ and $\ell(w)=2$. We have

$$s_2s_1(\lambda)=\lambda-a_1\alpha_1-(a_2-a_1(\alpha_1,\alpha_2^\vee))\alpha_2\preccurlyeq\lambda-a_1\alpha_1-(a_1+a_2)\alpha_2,$$

where $a_2=(\lambda,\alpha_2^\vee)$. Hence $s_2s_1(\lambda)\notin\mathcal{J}_\lambda$. It follows that $\dot{s_2}\dot{s_1}(v_{-\lambda})\in\mathfrak{N}_{U'}(\lambda)$ and

$$\tau(\phi^{-1}(\mathbb{O}(w)))=U{\cdot}(\dot{s_2}\dot{s_1}(V))\in\mathfrak{N}_{U'}(\lambda).$$

Thus, $w=s_2s_1$ is the required element. Since τ is injective outside the zero section of ϕ, it is still true that $\operatorname{codim}_{\mathscr{C}(\lambda)}\tau(\phi^{-1}(\mathbb{O}(w)))=2$. This proves that $\operatorname{codim}\mathfrak{N}_{U'}(\lambda)\leqslant 2$.

On the other hand, the similar argument shows that if $w\in W^\lambda$ and $\ell(w)=1$ (i.e., $w=s_i$, where $(\alpha_i,\lambda)\neq 0$), then $\dot{w}{\cdot}v_{-\lambda}\notin\mathfrak{N}_{U'}(\lambda)$. Therefore, $\operatorname{codim}\mathfrak{N}_{U'}(\lambda)=2$. □

Corollary 4.2. *Suppose that $U'\neq\{1\}$. Then $\pi_{\lambda,U'}:\mathscr{C}(\lambda)\to\mathscr{C}(\lambda)/\!\!/U'$ is equidimensional if and only if $\lambda=a_i\varpi_i$ for some i. In particular, if the action of G on $\mathscr{C}(\lambda)$ is effective and $\pi_{\lambda,U'}$ is equidimensional, then G is simple.*

Proof. It follows from Theorem 2.2(i) that $\dim \mathscr{C}(\lambda)/\!\!/U' = 2$ if and only if $\lambda = a_i \varpi_i$. □

Now, we turn to considering general monoids $\mathcal{S} \subset \mathfrak{X}_+$. For any $S \subset \mathfrak{X}$, let $\mathsf{con}(S)$ denote the closed cone in $\mathfrak{X}_{\mathbb{R}}$ generated by S.

Lemma 4.3. *Suppose that we are given two monoids $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\mathsf{con}(\mathcal{S}_1) = \mathsf{con}(\mathcal{S}_2)$. Then $\pi_{\mathcal{S}_1,U'}$ is equidimensional if and only if $\pi_{\mathcal{S}_2,U'}$ is.*

Proof. It suffices to treat the case in which $\mathcal{S}_2 = \mathsf{con}(\mathcal{S}_1) \cap \mathfrak{X}_+$. Then $\Bbbk[\mathscr{C}(\mathcal{S}_2)]$ is a finite $\Bbbk[\mathscr{C}(\mathcal{S}_1)]$-module [22, Prop. 4]. Consider the commutative diagram

$$\begin{array}{ccc} \mathscr{C}(\mathcal{S}_2) & \xrightarrow{\psi} & \mathscr{C}(\mathcal{S}_1) \\ \downarrow \pi_{\mathcal{S}_2,U'} & & \downarrow \pi_{\mathcal{S}_1,U'} \\ \mathscr{C}(\mathcal{S}_2)/\!\!/U' & \xrightarrow{\psi/\!\!/U'} & \mathscr{C}(\mathcal{S}_1)/\!\!/U'. \end{array}$$

Here ψ is finite, and it suffices to prove that $\psi/\!\!/U'$ is also finite, i.e., that $\Bbbk[\mathscr{C}(\mathcal{S}_2)]^{U'}$ is a finite $\Bbbk[\mathscr{C}(\mathcal{S}_1)]^{U'}$-module. By the "transfer principle" ([2, Ch. 1], [15, § 3]), we have

$$\Bbbk[X]^{U'} \simeq (\Bbbk[X] \otimes \Bbbk[G/U'])^G$$

for any affine G-variety X. Hence, one has to prove that $(\Bbbk[\mathcal{S}_2] \otimes \Bbbk[G/U'])^G$ is a finite $(\Bbbk[\mathcal{S}_1] \otimes \Bbbk[G/U'])^G$-module, which readily follows from the fact that $\Bbbk[G/U']$ is finitely generated and G is reductive. □

Theorem 4.4. *The quotient morphism $\pi_{\mathcal{S},U'}$ is equidimensional if and only if $\mathsf{con}(\mathcal{S})$ is generated by a sparse set of fundamental weights.*

Proof. 1) The "if" part readily follows from Lemma 4.3 and Theorem 3.7.

2) Suppose that $\pi_{\mathcal{S},U'} : \mathscr{C}(\mathcal{S}) \to \mathscr{C}(\mathcal{S})/\!\!/U'$ is equidimensional. By Lemma 4.3, it suffices to consider the case in which $\mathcal{S} = \mathsf{con}(\mathcal{S}) \cap \mathfrak{X}_+$. Then $\mathscr{C}(\mathcal{S})$ is normal (see Theorem 1.1(3)). Consider an arbitrary edge, $\mathsf{con}(\lambda)$, of $\mathsf{con}(\mathcal{S})$. It is assumed that $\lambda \in \mathcal{S}$ is a primitive element of $\mathfrak{X}_+$. By [22, Prop. 7], the HV-variety $\mathscr{C}(\lambda)$ is a subvariety of $\mathscr{C}(\mathcal{S})$. On the other hand, $\Bbbk[\mathscr{C}(\lambda)] = \bigoplus_{n \geqslant 0} \mathsf{R}(n\lambda)$ is a G-stable subalgebra of $\Bbbk[\mathscr{C}(\mathcal{S})] = \bigoplus_{\mu \in \mathcal{S}} \mathsf{R}(\mu)$. This yields the chain of G-equivariant maps

$$\mathscr{C}(\lambda) \hookrightarrow \mathscr{C}(\mathcal{S}) \xrightarrow{r} \mathscr{C}(\lambda).$$

Here the composite map is the identity, i.e., r is a G-equivariant retraction. Furthermore, passage to the subalgebras of U'-invariants (= quotient varieties) yields the maps

$$\mathscr{C}(\lambda)/\!\!/U' \hookrightarrow \mathscr{C}(\mathcal{S})/\!\!/U' \xrightarrow{r/\!\!/U'} \mathscr{C}(\lambda)/\!\!/U',$$

which shows that $r/\!\!/U'$ is a retraction, too. This also shows that both r and $r/\!\!/U'$ are onto. Consider the commutative diagram

$$\begin{array}{ccccc}
\mathcal{C}(\lambda) & \hookrightarrow & \mathcal{C}(\mathfrak{S}) & \xrightarrow{\ r\ } & \mathcal{C}(\lambda) \\
\downarrow{\scriptstyle \pi_{\lambda,U'}} & & \downarrow{\scriptstyle \pi_{\mathfrak{S},U'}} & & \downarrow{\scriptstyle \pi_{\lambda,U'}} \\
\mathcal{C}(\lambda)/\!\!/U' & \hookrightarrow & \mathcal{C}(\mathfrak{S})/\!\!/U' & \xrightarrow{r/\!\!/U'} & \mathcal{C}(\lambda)/\!\!/U'
\end{array}$$

As $\mathcal{C}(\mathfrak{S})$ is normal, the same is true for $\mathcal{C}(\mathfrak{S})/\!\!/U'$. Since $\pi_{\mathfrak{S},U'}$ is equidimensional and both $\mathcal{C}(\mathfrak{S})$ and $\mathcal{C}(\mathfrak{S})/\!\!/U'$ are conical, it follows from Lemma 1.5 that $\pi_{\mathfrak{S},U'}$ is onto. Therefore, $\pi_{\lambda,U'}$ is onto as well. Furthermore, $\pi_{\lambda,U'} = \pi_{\mathfrak{S},U'}|_{\mathcal{C}(\lambda)}$, since $\mathcal{C}(\lambda)$ is a G-stable subvariety of $\mathcal{C}(\mathfrak{S})$. This shows that $\pi_{\mathfrak{S},U'}(\mathcal{C}(\lambda))$ is a closed subset of $\mathcal{C}(\mathfrak{S})/\!\!/U'$.

Let $Y \subset \mathcal{C}(\mathfrak{S})$ be an irreducible component of $\pi^{-1}_{\mathfrak{S},U'}(\pi_{\mathfrak{S},U'}(\mathcal{C}(\lambda)))$ that contains $\mathcal{C}(\lambda)$ and maps dominantly to $\pi_{\mathfrak{S},U'}(\mathcal{C}(\lambda))$. Consider the commutative diagram

$$\begin{array}{ccc}
Y & \xrightarrow{\ r|_Y\ } & \mathcal{C}(\lambda) \\
& {\scriptstyle \pi_{\mathfrak{S},U'}|_Y}\searrow \quad \swarrow{\scriptstyle \pi_{\mathfrak{S},U'}|_{\mathcal{C}(\lambda)}} & \\
& \pi_{\mathfrak{S},U'}(\mathcal{C}(\lambda)) &
\end{array}$$

By the very construction of Y, the morphism $r|_Y$ is onto and $\pi_{\mathfrak{S},U'}|_Y$ is equidimensional. It follows that $\pi_{\mathfrak{S},U'}|_{\mathcal{C}(\lambda)}$ is also equidimensional. Consequently, $\pi_{\lambda,U'} = \pi_{\mathfrak{S},U'}|_{\mathcal{C}(\lambda)}$ is equidimensional and, by Corollary 4.2, $\lambda = \varpi_i$ for some i (recall that λ is supposed to be primitive). Thus, the edges of $\mathsf{con}(\mathfrak{S})$ are generated by fundamental weights. Finally, by Theorem 3.4, the corresponding set of fundamental weights is sparse. □

Remark 4.5. Our proof of the "only if" part exploits ideas of Vinberg and Wehlau for the equidimensional quotients by G (see [23, Theorem 8.2] and [24, Prop. 2.6]).

Remark 4.6. We can prove a general equidimensionality criterion for the quotients of $\mathcal{S}$-varieties by U. This topic will be considered in a forthcoming publication.

5. The Hilbert–Mumford criterion for U'

Let X be an irreducible affine G-variety and $x_0 \in X^G$. For any $H \subset G$, define the *null-cone* with respect to H and x_0 as

$$\mathfrak{N}_H(X) = \{x \in X \mid F(x) = F(x_0) \quad \forall F \in \Bbbk[X]^H\}.$$

If $\Bbbk[X]^H$ is finitely generated, then $\mathfrak{N}_H(X)$ can be regarded as the fibre of $\pi_{X,H}$ containing x_0. Below, we give a characterisation of $\mathfrak{N}_{U'}(X)$ via one-parameter

subgroups (1-PS for short) of T. This is inspired by Brion's description of null-cones for U-invariants [3, Sect. IV]. Recall that the Hilbert–Mumford criterion for G asserts that

$x \in \mathfrak{N}_G(X)$ if and only if there is a 1-PS $\tau : \Bbbk^\times \to G$ such that $\lim_{t\to 0} \tau(t)\cdot x = x_0$ (cf. [9, III.2], [23, § 5.3]). By [14, Theorem 2.2], there is the canonical affine model of the homogeneous space G/U', that is, an affine pointed G-variety $(\overline{G/U'}, \boldsymbol{p})$ such that

- $G_{\boldsymbol{p}} = U'$;
- $G\cdot\boldsymbol{p}$ is dense in $\overline{G/U'}$;
- $\Bbbk[\overline{G/U'}] = \Bbbk[G]^{U'}$.

Here $\boldsymbol{p} = (f_1, \tilde{f}_1, \dots, f_r, \tilde{f}_r)$ is a direct sum of weight vectors in $2\mathsf{R}(\varpi_1) \oplus \cdots \oplus 2\mathsf{R}(\varpi_r)$, with weights $\varpi_i, \varpi_i - \alpha_i$ $(1 \leqslant i \leqslant r)$. If G has no simple factors SL_2, SL_3, then all these weights belong to an open half-space of $\mathfrak{X}_\mathbb{R}$ (see the proof of [14, Prop. 1.9]). In this case, $\boldsymbol{p}$ is unstable and $\overline{G/U'}$ contains the origin in $2\mathsf{R}(\varpi_1) \oplus \cdots \oplus 2\mathsf{R}(\varpi_r)$. Let $\tau : \Bbbk^\times \to T$ be a 1-PS. Using the canonical pairing between $\mathfrak{X}$ and the set of 1-PS of T, we will regard τ as an element of $\mathfrak{X}_\mathbb{R}$. Let us say that τ is *U'-admissible*, if $(\tau, \varpi_i) > 0$ and $(\tau, \varpi_i - \alpha_i) > 0$ for all i; that is, if $\lim_{t\to 0} \tau(t)\cdot\boldsymbol{p} = 0$. Since $\Bbbk[\overline{G/U'}] = \Bbbk[G]^{U'}$, one has the isomorphism

$$\Bbbk[X \times \overline{G/U'}]^G = (\Bbbk[X] \otimes \Bbbk[G]^{U'})^G \xrightarrow{\sim} \Bbbk[X]^{U'} \tag{5·1}$$

that takes $\tilde{F}(\cdot,\cdot) \in \Bbbk[X \times \overline{G/U'}]^G$ to $F(\cdot) = \tilde{F}(\cdot, \boldsymbol{p}) \in \Bbbk[X]^{U'}$.

Theorem 5.1. *Suppose that G has no simple factors SL_2, SL_3. Then the following conditions are equivalent:*

(i) *$x \in \mathfrak{N}_{U'}(X)$, i.e., $F(x) = F(x_0)$ for all $F \in \Bbbk[X]^{U'}$;*

(ii) *there is $u \in U$ and a U'-admissible 1-PS $\tau : \Bbbk^\times \to T$ such that* $\lim_{t\to 0} \tau(t)u\cdot x = x_0$.

Proof. (i) $\Rightarrow$ (ii). Suppose that $x \in \mathfrak{N}_{U'}(X)$. Then $\tilde{F}(x, \boldsymbol{p}) = F(x) = F(x_0) = \tilde{F}(x_0, \boldsymbol{p})$. Since $\boldsymbol{p}$ is unstable in $\overline{G/U'}$, we have $\tilde{F}(x_0, \boldsymbol{p}) = \tilde{F}(x_0, 0)$. Thus, $\tilde{F}(x, \boldsymbol{p}) = \tilde{F}(x_0, 0)$ for all $\tilde{F} \in (\Bbbk[X] \otimes \Bbbk[G]^{U'})^G$, i.e., $(x, \boldsymbol{p}) \in \mathfrak{N}_G(X \times \overline{G/U'})$. By the Hilbert–Mumford criterion for G, there is a 1-PS $\nu : \Bbbk^\times \to G$ such that $\nu(t)\cdot(x, \boldsymbol{p}) \xrightarrow[t\to 0]{} (x_0, 0)$.

By a result of Grosshans [10, Cor. 1] (see also [3, IV.1]), we may assume that $\nu(\Bbbk^\times) \subset B$. Then there is $u \in U$ such that $\tau(t) := u\nu(t)u^{-1} \in T$. Therefore,

$$\tau(t)u\cdot(x, \boldsymbol{p}) \xrightarrow[t\to 0]{} (x_0, 0).$$

Note that $u\cdot\boldsymbol{p}$ $(u \in U)$ does not differ much from $\boldsymbol{p}$. Namely, each component f_i remains intact, whereas $\tilde{f}_i$ is replaced with $\tilde{f}_i + c_i f_i$ for some $c_i \in \Bbbk$. This means

that $\tau(t)u\cdot\boldsymbol{p} \underset{t\to 0}{\longrightarrow} 0$ if and only if $\tau(t)\cdot\boldsymbol{p} \underset{t\to 0}{\longrightarrow} 0$. That is, τ is actually U'-admissible and $\lim_{t\to 0}\tau(t)u\cdot x = x_0$.
(ii) $\Rightarrow$ (i). Suppose that $F \in \Bbbk[X]^{U'}$ and $\tilde{F}$ is the corresponding G-invariant in $\Bbbk[X\times\overline{G/U'}]$. Then $F(x) = \tilde{F}(x, \boldsymbol{p}) = \tilde{F}(\tau(t)u\cdot x, \tau(t)u\cdot\boldsymbol{p})$. Since $u\cdot\boldsymbol{p}$ is a linear combination of weight vectors with the same weights and τ is U'-admissible, we have $\lim_{t\to 0}\tau(t)u\cdot\boldsymbol{p} = 0$. Hence $F(x) = \tilde{F}(x_0, 0) = \tilde{F}(x_0, \boldsymbol{p}) = F(x_0)$. □

Remark 5.2. Our Theorem 5.1 is similar to Theorem 5 in [3] on null-cones for U-invariants. The only difference is that we end up with a smaller class of admissible 1-PS.

Obviously, there are inclusions $\mathfrak{N}_{U'}(X) \subset \mathfrak{N}_U(X) \subset \mathfrak{N}_G(X)$ and hence

$$G\cdot\mathfrak{N}_{U'}(X) \subset G\cdot\mathfrak{N}_U(X) \subset \mathfrak{N}_G(X).$$

It is proved in [3, Théorème 6(ii)] that actually $G\cdot\mathfrak{N}_U(X) = \mathfrak{N}_G(X)$. Below, we investigate the similar problem for U'.

Recall that $\mathsf{con}(S)$ is the closed cone in $\mathfrak{X}_{\mathbb{R}}$ generated by S. If $K \subset \mathfrak{X}_{\mathbb{R}}$ is a closed cone, then $K^{\perp}$ denotes the dual cone and K^{o} denotes the relative interior of K. By the very definition, the cone generated by the U'-admissible 1-PS is open, and its closure is dual to $\mathsf{con}(\{\varpi_i, \varpi_i - \alpha_i \mid i = 1, \dots, r\})$. By [14, Theorem 4.2], we have

$$\mathsf{con}(\{\varpi_i, \varpi_i - \alpha_i \mid i = 1, \dots, r\})^{\perp} = \mathsf{con}(\Delta^{+} \setminus \Pi).$$

Hence the cone generated by the U'-admissible 1-PS equals $\mathsf{con}(\Delta^{+} \setminus \Pi)^{o}$.

Theorem 5.3. *Suppose that G has no simple factors of type SL. Then*
1) $\mathsf{con}(\varpi_1, \dots, \varpi_r) \subset \mathsf{con}(\Delta^{+} \setminus \Pi)$,
2) $G\cdot\mathfrak{N}_{U'}(X) = \mathfrak{N}_G(X)$ *for all affine G-varieties X.*

Proof. 1) Taking the dual cones yields the equivalent condition that

$$\mathsf{con}(\{\varpi_i, \varpi_i - \alpha_i \mid i = 1, \dots, r\}) \subset \mathsf{con}(\Delta^{+}).$$

That is, one has to verify that each $\varpi_i - \alpha_i$ has non-negative coefficients in the expression via the simple roots. Let C denote the Cartan matrix of a simple group G. All the entries of C^{-1} are positive and the rows of C^{-1} provide the expressions of the fundamental weights via the simple roots. Hence it remains to check that the diagonal entries of C^{-1} are $\geqslant 1$. An explicit verification shows that this is true if $G \neq SL_{r+1}$. (The matrices C^{-1} can be found in [21, Table 2].)

2) Suppose that $x \in \mathfrak{N}_G(X)$. Then there exist $g \in G$ and $\tau : \Bbbk^{\times} \to T$ such that $\lim_{t\to 0}\tau(t)g\cdot x = x_0$. Let $y = g\cdot x$. The set of all 1-PS $\nu : \Bbbk^{\times} \to T$ such that $\lim_{t\to 0}\nu(t)\cdot y = x_0$ generates an open cone in $\mathfrak{X}_{\mathbb{R}}$. Therefore, we may assume that τ is a regular 1-PS. Now, in view of the Hilbert–Mumford criterion for G and

Theorem 5.1, it suffices to prove that any regular 1-PS of T is W-conjugate to a U'-admissible one. This follows from part 1), since $\mathsf{con}(\varpi_1, \dots, \varpi_r)$ is a fundamental domain for the W-action on $\mathfrak{X}_{\mathbb{R}}$ and $\mathsf{con}(\varpi_1, \dots, \varpi_r)^o \subset \mathsf{con}(\Delta^+ \setminus \Pi)^o$. □

For $G = SL_{r+1}$, we have $\varpi_1 - \alpha_1, \varpi_r - \alpha_r \notin \mathsf{con}(\Delta^+)$ and therefore, $\mathsf{con}(\varpi_1, \dots, \varpi_r) \not\subset \mathsf{con}(\Delta^+ \setminus \Pi)$. More precisely, $\varpi_1, \varpi_r \not\subset \mathsf{con}(\Delta^+ \setminus \Pi)$. This means that one may expect that, for *some* SL_{r+1}-varieties, there is the strict inclusion $G{\cdot}\mathfrak{N}_{U'}(X) \subsetneqq \mathfrak{N}_G(X)$.

Example 5.4. For $m \geqslant 3$, consider the representation of $G = SL_3$ in the space $V = \mathsf{R}(m\varpi_1)$ of forms of degree m in three variables x, y, z. By Theorem 1.2, $\dim V^{U'} = m+1$. Let U be the subgroup of the unipotent upper-triangular matrices in the basis dual to (x, y, z). The U'-invariants of degree 1 are the coefficients of $x^m, x^{m-1}y, \dots, xy^{m-1}, y^m$. Therefore, $\mathfrak{N}_{U'}(V)$ is contained in the subspace of forms having the linear factor z and all the forms in $SL_3{\cdot}\mathfrak{N}_{U'}(V)$ have a linear factor. On the other hand, the null-form (with respect to SL_3) $x^m + y^{m-1}z$ is irreducible. Hence, $SL_3{\cdot}\mathfrak{N}_{U'}(V) \neq \mathfrak{N}_{SL_3}(V)$.

Remark. In view of Theorem 5.1, it would be much more instructive to have such an example for SL_n, $n \geqslant 4$. However, we are unable to provide it yet.

6. Equidimensional quotients and irreducible representations of simple groups

In this section, we transform the criterion of Theorem 3.9 in a sufficient condition applicable to a wider class of G-varieties. Then we obtain the list of irreducible representations V of simple algebraic groups $G \neq SL_2$ such that $\Bbbk[V]$ is a free $\Bbbk[V]^{U'}$-module.

For any affine irreducible G-variety Z, there is a flat degeneration $\Bbbk[Z] \rightsquigarrow \mathsf{gr}(\Bbbk[Z])$. (Brion attributes this to Domingo Luna in his thesis, see [2, Lemma 1.5]). Here $\mathsf{gr}(\Bbbk[Z])$ is again a finitely generated $\Bbbk$-algebra and a locally-finite G-module, and $\mathsf{gr}Z := \mathsf{Spec}\,(\mathsf{gr}(\Bbbk[Z]))$ is an affine horospherical G-variety. The whole theory of "contractions of actions of reductive groups" is later developed in [15]. (See also [4], [19], [11] for related results and other applications.) The "contraction" $Z \rightsquigarrow \mathsf{gr}Z$ has the property that the algebras $\Bbbk[Z]$ and $\Bbbk[\mathsf{gr}Z] = \mathsf{gr}(\Bbbk[Z])$ are isomorphic as G-modules. But the multiplication in $\Bbbk[\mathsf{gr}Z]$ is simpler than that in $\Bbbk[Z]$; namely, if M and N are two simple G-modules in $\Bbbk[\mathsf{gr}Z]$, then $M{\cdot}N$ (the product in $\Bbbk[\mathsf{gr}Z]$) is again a simple G-module. Furthermore, $\Bbbk[\mathsf{gr}Z]^U \simeq \Bbbk[Z]^U$ and $G{\cdot}((\mathsf{gr}Z)^U) = \mathsf{gr}Z$. This means that if Z is a spherical G-variety, then $\mathsf{gr}Z$ is an $\mathcal{S}$-variety.

Theorem 6.1. *Suppose that G acts on an irreducible affine variety X such that (1) $\Bbbk[X]^U$ is a polynomial algebra and (2) the weights of free generators are fundamental, different and form a sparse set. Then $\Bbbk[X]^{U'}$ is also polynomial, of Krull dimension $2\dim X/\!\!/U$, and the quotient $\pi_{X,U'}: X \to X/\!\!/U'$ is equidimensional.*

Proof. The idea is the same as in the proof of the similar result for U-invariants in [13, Theorem 5.5]. We use the fact that in our situation $\mathrm{gr}X$ is an $\mathcal{S}$-variety whose monoid of dominant weights is generated by a sparse set of fundamental weights.

Let $\varpi_1, \dots, \varpi_m$ be the weights of free generators of $\Bbbk[X]^U$. Set $\Gamma = \sum_{i=1}^m \mathbb{N}\varpi_i$. It follows from the hypotheses on weights that $\Bbbk[X]$ is a multiplicity free G-module, i.e., X is a spherical G-variety [18, Theorem 2]. Therefore, $\Bbbk[X]$ is isomorphic to $\bigoplus_{\lambda\in\Gamma} \mathsf{R}(\lambda)$ as G-module and $\mathrm{gr}X \simeq \mathscr{C}(\Gamma)$.

By [15, §5], there exists a G-variety Y and a function $q \in \Bbbk[Y]^G$ such that $\Bbbk[Y]/(q-a) \simeq \Bbbk[X]$ for all $a \in \Bbbk^\times$, $\Bbbk[Y][q^{-1}] \simeq \Bbbk[X][q, q^{-1}]$, and $\Bbbk[Y]/(q) \simeq \Bbbk[\mathrm{gr}X]$. Recall some details on constructing Y and $\mathrm{gr}X$. Let ϱ be the half-sum of the positive coroots. For $\lambda \in \mathfrak{X}_+$, we set $\mathrm{ht}(\lambda) = (\lambda, \varrho)$. Letting $\Bbbk[X]_{(n)} = \bigoplus_{\lambda:\, \mathrm{ht}(\lambda)\leqslant n} \mathsf{R}(\lambda)$, one obtains an ascending filtration of the algebra $\Bbbk[X]$:

$$\{0\} \subset \Bbbk[X]_{(0)} \subset \Bbbk[X]_{(1)} \subset \cdots \subset \Bbbk[X]_{(n)} \cdots .$$

Each subspace $\Bbbk[X]_{(n)}$ is G-stable and finite-dimensional and $\Bbbk[X]_{(0)} = \Bbbk[X]^G = \Bbbk$. Let q be a formal variable. Then the algebras $\Bbbk[Y]$ and $\mathrm{gr}(\Bbbk[X])$ are defined as follows:

$$\Bbbk[Y] = \bigoplus_{n=0}^{\infty} \Bbbk[X]_{(n)} q^n \subset \Bbbk[X][q] ,$$

$$\mathrm{gr}(\Bbbk[X]) = \bigoplus_{n\geqslant 0} \Bbbk[X]_{(n)}/\Bbbk[X]_{(n-1)} .$$

Let $f_1, \dots, f_m$ be the free generators of $\Bbbk[X]^U$, where $f_i \in \mathsf{R}(\varpi_i)^U$, as usual. They can also be regarded as free generators of $\Bbbk[\mathrm{gr}X]^U$. By Theorem 1.4, $\Bbbk[\mathrm{gr}X]^{U'}$ is freely generated by $f_1, \tilde{f}_1, \dots, f_m, \tilde{f}_m$ and by Theorem 3.9, $\pi_{\mathrm{gr}X,U'}: \mathrm{gr}X \to (\mathrm{gr}X)/\!\!/U'$ is equidimensional. On the other hand, it follows from [14, Theorem 2.4] that $f_1, \tilde{f}_1, \dots, f_m, \tilde{f}_m$ also generate $\Bbbk[X]^{U'}$. Therefore, to conclude that $\Bbbk[X]^{U'}$ is polynomial, it suffices to know that $\dim X/\!\!/U' = \dim(\mathrm{gr}X)/\!\!/U' (= 2m)$. To this end, we exploit the following facts:

a) For an irreducible G-variety X, there always exists a generic stabiliser for the U-action on X [5, Corollaire 1.6], which we denote by $\mathrm{g.s.}(U{:}X)$;

b) If X is affine, then this generic stabiliser depends only on the G-module structure of $\Bbbk[X]$, i.e., on the highest weights of G-modules occurring in $\Bbbk[X]$ [12, Theorem 1.2.9]. Consequently, $\mathrm{g.s.}(U{:}X) = \mathrm{g.s.}(U{:}\mathrm{gr}X)$;

c) the minimal dimension of U'-stablisers in X equals $\dim(U' \cap \mathrm{g.s.}(U{:}X))$ [14, Lemma 2.5]. Therefore it is the same for X and $\mathrm{gr}X$;

d) Since U' is unipotent, we have $\dim X/\!\!/U' = \dim X - \dim U' + \min_{x\in X} \dim U'_x$.

Combining a)-d) yields the desired equality and thereby the assertion that $\Bbbk[X]^{U'}$ is polynomial, of Krull dimension $2m = 2\dim X/\!\!/U$.

Let n_i be the smallest integer such that $\mathsf{R}(\varpi_i) \subset \Bbbk[X]_{(n_i)}$. Using the above description of $\Bbbk[Y]$ and $\Bbbk[\mathrm{gr}X]^{U'}$, one easily obtains that

$$\Bbbk[Y]^U = \Bbbk[q, q^{n_1} f_1, \ldots, q^{n_m} f_m]$$
$$\Bbbk[Y]^{U'} = \Bbbk[q, q^{n_1} f_1, q^{n_1} \tilde{f}_1, \ldots, q^{n_m} f_m, q^{n_m} \tilde{f}_m],$$

i.e., both algebras are polynomial, of Krull dimension $m+1$ and $2m+1$, respectively. By a result of Kraft, the first equality implies that Y has rational singularities (see [2, Theorem 1.6], [15, Theorem 6]). One has the following commutative diagram:

$$\begin{array}{ccccccc}
C(\Gamma) & \simeq & \mathrm{gr}X & \hookrightarrow & Y & \leftarrow & X\times\mathbb{A}^1 \\
 & & \downarrow{\scriptstyle \pi_{\mathrm{gr}X,U'}} & & \downarrow{\scriptstyle \pi_{Y,U'}} & & \\
\mathbb{A}^{2m} & \simeq & (\mathrm{gr}X)/\!\!/U' & \hookrightarrow & Y/\!\!/U' & \simeq & \mathbb{A}^{2m+1} \\
 & & \downarrow & & \downarrow{\scriptstyle q} & & \\
 & & \{0\} & \hookrightarrow & \mathbb{A}^1 & &
\end{array}$$

Consequently,

$$\mathfrak{N}_{U'}(\mathrm{gr}X) = \pi^{-1}_{\mathrm{gr}X,U'}(\pi_{\mathrm{gr}X,U'}(\bar{0})) = \pi^{-1}_{Y,U'}(\pi_{Y,U'}(\bar{0})) = \mathfrak{N}_{U'}(Y),$$

where $\bar{0} \in \mathrm{gr}X \subset Y$ is the unique G-fixed point of $\mathrm{gr}X$. Since $\dim Y = \dim X + 1$, $\dim Y/\!\!/U' = \dim(\mathrm{gr}X)/\!\!/U' + 1$, and $\pi_{\mathrm{gr}X,U'}$ is equidimensional, the morphism $\pi_{Y,U'}$ is equidimensional as well. As Y has rational singularities and hence is Cohen-Macaulay, this implies that $\Bbbk[Y]$ is a flat $\Bbbk[Y]^{U'}$-module. Since $\Bbbk[Y][q^{-1}] \simeq \Bbbk[X][q, q^{-1}]$ and $\Bbbk[Y]^{U'}[q^{-1}] \simeq \Bbbk[X]^{U'}[q, q^{-1}]$, we conclude that $\Bbbk[X]$ is a flat $\Bbbk[X]^{U'}$-module. Thus, $\pi_{X,U'}$ is equidimensional. □

Our next goal is to obtain the list of all irreducible representations V of simple algebraic groups such that $\Bbbk[V]$ is a free $\Bbbk[V]^{U'}$-module. As is well known, $\Bbbk[V]$ is a free $\Bbbk[V]^{U'}$-module if and only if $\Bbbk[V]^{U'}$ is polynomial and $\pi_{V,U'}$ is equidimensional [16, Prop. 17.29]. Therefore, the required representations are contained in [14, Table 1] and our task is to pick from that table the representations having the additional property that $\pi_{V,U'}$ is equidimensional. The numbering of fundamental weights of simple algebraic groups follows [21, Tables].

Theorem 6.2. *Let G be a connected simple algebraic group with $\mathsf{rk}\, G \geqslant 2$ and $\mathsf{R}(\lambda)$ a simple G-module. The following conditions are equivalent:*

(i) $\Bbbk[\mathsf{R}(\lambda)]$ *is a free* $\Bbbk[\mathsf{R}(\lambda)]^{U'}$*-module;*

(ii) *Up to symmetries of the Dynkin diagram of* G*, the pairs* (G, λ) *occur in the following list:* $(\mathbf{A}_r, \varpi_1)$, $(\mathbf{B}_r, \varpi_1)$, $(\mathbf{C}_r, \varpi_1)$, $r \geqslant 2$*;*
$(\mathbf{D}_r, \varpi_1)$, $r \geqslant 3$*;*
$(\mathbf{B}_3, \varpi_3)$, $(\mathbf{B}_4, \varpi_4)$, $(\mathbf{D}_5, \varpi_5)$, $(\mathbf{E}_6, \varpi_1)$, $(\mathbf{G}_2, \varpi_1)$.

Proof. (ii)⇒(i). By [14, Theorem 5.1], all these representations have a polynomial algebra of U'-invariants. Consider $X = \mathfrak{N}_G(\mathsf{R}(\lambda))$, the null-cone with respect to G. The nonzero weights of generators of $\Bbbk[\mathsf{R}(\lambda)]^U$ (and hence the weights of generators of $\Bbbk[X]^U$) given by Brion [3, p. 13] are fundamental and form a sparse set. Consequently, Theorem 6.1 applies to X, and $\pi_{X,U'}$ is equidimensional. Since X is either a G-invariant hypersurface in $\mathsf{R}(\lambda)$ or equal to $\mathsf{R}(\lambda)$, $\pi_{\mathsf{R}(\lambda),U'}$ is also equidimensional.

(i)⇒(ii). We have to prove that, for the other items in [14, Table 1], the quotient is not equidimensional. The list of such "bad" pairs (G, λ) is: $(\mathbf{A}_r, \varpi_2^*)$ with $r \geqslant 4$; $(\mathbf{B}_5, \varpi_5)$, $(\mathbf{D}_6, \varpi_6)$, $(\mathbf{E}_7, \varpi_1)$, $(\mathbf{F}_4, \varpi_1)$. Note that $(\mathbf{A}_3, \varpi_2^*) = (\mathbf{D}_3, \varpi_1)$ and this good pair is included in the list in (ii).
It suffices to check that the free generators of $\Bbbk[\mathsf{R}(\lambda)]^{U'}$ given in that Table do not form a regular sequence. To this end, we point out a certain relation in $\Bbbk[\mathsf{R}(\lambda)]$ using the fact the weights of generators do not form a sparse set (cf. the proof of Theorem 3.4).

The only "bad" serial case is $(\mathbf{A}_r, \varpi_2^*)$ with $r \geqslant 4$. The algebra $\Bbbk[\mathsf{R}(\varpi_2^*)]^U$ has free generators f_{2i} $(1 \leqslant i \leqslant [r/2])$ of degree i and weight ϖ_{2i}, and for r odd, there is also the Pfaffian, which is G-invariant. Then $\Bbbk[\mathsf{R}(\varpi_2^*)]^{U'}$ is freely generated by $f_2, \tilde{f}_2, f_4, \tilde{f}_4, \ldots$ (and the Pfaffian, if r is odd). Using the 4-nodes fragments of the weight posets $\mathcal{P}(\varpi_2)$ and $\mathcal{P}(\varpi_4)$ and notation of the proof of Theorem 3.4, we construct a U-invariant function $f_2q_4 - \tilde{f}_2p_4 + p_2\tilde{f}_4 - q_2f_4$ of degree 3 and weight $\varpi_2 + \varpi_4 - \alpha_2 - \alpha_3 - \alpha_4 = \varpi_1 + \varpi_5$. (Cf. Eq. (3·1).) However, there are no such nonzero U-invariants in $\Bbbk[\mathsf{R}(\varpi_2^*)]$. This yields a relation in $\Bbbk[\mathsf{R}(\varpi_2^*)]$ involving free generators $f_2, \tilde{f}_2, f_4, \tilde{f}_4 \in \Bbbk[\mathsf{R}(\varpi_2^*)]^{U'}$.

In all other cases, we can do the same thing using a pair of generators of $\Bbbk[\mathsf{R}(\lambda)]^U$ corresponding to suitable fundamental weights. The only difference is that one of these two U-invariants is <u>not</u> included in the minimal generating system of $\Bbbk[\mathsf{R}(\lambda)]^{U'}$ and should be expressed via some other U'-invariants. Nevertheless, the resulting relation still shows that the U'-invariants involved do not form a regular sequence.

For instance, consider the pair $(\mathbf{D}_6, \varpi_6)$. Here the free generators of $\Bbbk[\mathsf{R}(\varpi_6)]^U$ have the following degrees and weights: $(1, \varpi_6)$, $(2, \varpi_2)$, $(3, \varpi_6)$, $(4, \varpi_4)$, $(4, \underline{0})$ [3]. The invariants themselves are denoted by $f_6^{(1)}$, f_2, $f_6^{(3)}$, f_4, F, respectively. Starting with the U-invariants f_2 and f_4, we obtain, as a above, a relation of the

form

$$f_2q_4 - \tilde{f}_2p_4 + p_2\tilde{f}_4 - q_2f_4 = 0 \tag{6·1}$$

in $\Bbbk[\mathsf{R}(\varpi_6^*)]$. However, f_4 is not a generator in $\Bbbk[\mathsf{R}(\varpi_6)]^{U'}$. Taking the second U'-invariant in each fundamental G-submodule, we obtain nine functions $f_6^{(1)}$, $\tilde{f}_6^{(1)}$, f_2, $\tilde{f}_2$, $f_6^{(3)}$, $\tilde{f}_6^{(3)}$, f_4, $\tilde{f}_4$, F that generate $\Bbbk[\mathsf{R}(\varpi_6)]^{U'}$. Here $f_4 = f_6^{(1)}\tilde{f}_6^{(3)} - \tilde{f}_6^{(1)}f_6^{(3)}$ and the remaining eight functions freely generate $\Bbbk[\mathsf{R}(\varpi_6)]^{U'}$. Substituting this expression for f_4 in (6·1), we finally obtain the relation

$$f_2q_4 - \tilde{f}_2p_4 + p_2\tilde{f}_4 - q_2\bigl(f_6^{(1)}\tilde{f}_6^{(3)} - \tilde{f}_6^{(1)}f_6^{(3)}\bigr) = 0,$$

which shows that the free generators of $\Bbbk[\mathsf{R}(\varpi_6)]^{U'}$ do not form a regular sequence. □

Some open problems. Let V be a rational G-module.

1^o. *Suppose that $V/\!\!/U$ is an affine space. Is it true that $V/\!\!/U'$ is a complete intersection?*

2^o. *Suppose that $V/\!\!/U'$ is an affine space and G has no simple factors SL_2. Is it true that $V/\!\!/U$ is an affine space?* (In [14], we have proved that $V/\!\!/G$ is an affine space, but this seems to be too modest.)

Direct computations provide an affirmative answer to both questions if G is simple and V is a simple G-module.

Acknowledgements. Part of this work was done while I was visiting MPIM (Bonn). I thank the Institute for the hospitality and inspiring environment. I am grateful to E.B. Vinberg for sending me the preprint [20].

References

[1] А. Берентштейн, А. Зелевинский. Когда кратность веса равна 1? *Функц. анализ и прилож.*, **24**, № 4 (1990), 1–13 (Russian). English translation: A. Berenstein and A. Zelevinsky. When is the multiplicity of a weight equal to 1? *Funct. Anal. Appl.*, **24** (1991), 259–269.

[2] M. Brion. Sur la théorie des invariants, *Publ. Math. Univ. Pierre et Marie Curie*, no. **45** (1981), pp. 1–92.

[3] M. Brion. Invariants d'un sous-groupe unipotent maximal d'un groupe semi-simple, *Ann. Inst. Fourier*, **33** (1983), 1–27.

[4] M. Brion. Quelques propriétés des espaces homogènes sphériques, *Manuscr. Math.*, **55** (1986), 191–198.

[5] M. Brion, D. Luna and Th. Vust. Espaces homogènes sphériques. *Invent. Math.*, **84** (1986), 617–632.

[6] C. Chevalley. "Fondements de la géométrie algébrique", Paris, 1958.

[7] В.И. Данилов. Геометрия торических многообразий, *Успехи Матем. Наук*, т. **XXXIII**, вып. 2 (1978), 85–134 (Russian). English translation: V.I. Danilov. Geometry of toric varieties, *Russian Math. Surveys*, **33**:2, (1978), 97–154.

[8] J.E. HUMPHREYS. "Reflection Groups and Coxeter Groups", Cambridge Univ. Press, 1992.

[9] H. KRAFT. "Geometrische Methoden in der Invariantentheorie", Aspekte der Mathematik **D1**, Braunschweig: Vieweg & Sohn, 1984.

[10] F. GROSSHANS. The variety of points which are not semi-stable, *Illinois J. Math.* **26** (1982), 138–148.

[11] D. PANYUSHEV. On deformation method in Invariant Theory, *Ann. Inst. Fourier*, **47** (1997), 985–1012.

[12] D. PANYUSHEV. Complexity and rank of actions in invariant theory, *J. Math. Sci.* (New York), **95** (1999), 1925–1985.

[13] D. PANYUSHEV. Parabolic subgroups with Abelian unipotent radical as a testing site for Invariant Theory, *Canad. J. Math.*, **51**(3) (1999), 616–635.

[14] D. PANYUSHEV. Actions of the derived group of a maximal unipotent subgroup on G-varieties, *Int. Math. Res. Notices*, **2010**, no. 4 (2010), 674–700.

[15] В.Л. Попов. Стягивания действий редуктивных алгебраических групп, *Матем. Сб.*, т.**130** (1986), 310–334 (Russian). English translation: V.L. POPOV. Contractions of actions of reductive algebraic groups, *Math. USSR-Sbornik*, **58** (1987), 311–335.

[16] G. SCHWARZ. Lifting smooth homotopies of orbit spaces, *Publ. Math. I.H.E.S.*, **51** (1980), 37–135.

[17] J.-P. SERRE. "Local algebra", Springer Monographs in Mathematics, Berlin: Springer-Verlag, 2000.

[18] Э.Б. Винберг, Б.Н. Кимельфельд. Однородные области на флаговых многообразиях и сферические подгруппы полупростых групп Ли, *Функц. анал. и прилож.*, **12**, № 3 (1978), 12–19 (Russian). English translation: E.B. VINBERG and B.N. KIMEL'FELD. Homogeneous domains on flag manifolds and spherical subgroups of semisimple Lie groups, *Funct. Anal. Appl.*, **12** (1978), 168–174.

[19] Э.Б. Винберг. Сложность действий редуктивных групп, *Функц. анал. и прилож.*, **20**, № 1 (1986), 1–13 (Russian). English translation: E.B. VINBERG. Complexity of action of reductive groups, *Funct. Anal. Appl.*, **20** (1986), 1–11.

[20] Э.Б. Винберг, С.Г. Гиндикин. Вырождение орисфер в сферических однородных пространствах (Russian) (= *Degeneration of horospheres in spherical homogeneous spaces*, Preprint dated June 16, 2008; 16 pp.)

[21] Э.Б. Винберг, А.Л. Онищик. "Семинар по группам Ли и алгебраическим группам". Москва: "Наука" 1988 (Russian). English translation: A.L. ONISHCHIK and E.B. VINBERG. "Lie groups and algebraic groups", Berlin: Springer, 1990.

[22] Э.Б. Винберг, В.Л. Попов. Об одном классе аффинных квазиоднородных многообразий, *Изв. АН СССР. Сер. Матем.*, т.**36** (1972), 749–764 (Russian). English translation: E.B. VINBERG and V.L. POPOV. On a class of quasihomogeneous affine varieties, *Math. USSR-Izv.*, **6** (1972), 743–758.

[23] Э.Б. Винберг, В.Л. Попов. "Теория Инвариантов", В кн.: Современные проблемы математики. Фундаментальные направления, т. 55, стр. 137–309. Москва: ВИНИТИ 1989 (Russian). English translation: V.L. POPOV and E.B. VINBERG. "Invariant theory", In: *Algebraic Geometry IV* (Encyclopaedia Math. Sci., vol. 55, pp. 123–284) Berlin Heidelberg New York: Springer 1994.

[24] D. WEHLAU. Equidimensional varieties and associated cones, *J. Algebra*, **159** (1993), 47–53.

Received April 26, 2011. Revised September 11, 2011.

DMITRI I. PANYUSHEV
INSTITUTE FOR INFORMATION TRANSMISSION PROBLEMS OF THE R.A.S.
B. KARETNYI PER. 19
MOSCOW
127994
RUSSIA
panyushev@iitp.ru

PACIFIC JOURNAL OF MATHEMATICS
Vol. 256, No. 2, 2012

INVARIANTS OF TOTALLY REAL LEFSCHETZ FIBRATIONS

NERMİN SALEPCİ

We introduce certain invariants of real Lefschetz fibrations and call these invariants *real Lefschetz chains*. We prove that if the fiber genus is greater than 1, then the real Lefschetz chains are complete invariants of totally real Lefschetz fibrations. If however the fiber genus is 1, real Lefschetz chains are not sufficient to distinguish real Lefschetz fibrations. We show that by adding a certain binary decoration to real Lefschetz chains, we get a complete invariant.

1. Introduction

This note is devoted to a topological study of Lefschetz fibrations equipped with certain $\mathbb{Z}_2$ actions compatible with the fiber structure. The action is generated by an involution, which is called a *real structure*. Intuitively, real structures are topological generalizations of the complex conjugation on complex algebraic varieties defined over the reals. Real Lefschetz fibrations appear, for instance, as blow-ups of pencils of hyperplane sections of complex projective algebraic surfaces defined by real polynomial equations. Regular fibers of real Lefschetz fibrations are compact oriented smooth genus g surfaces, while singular fibers have a single node. The invariant fibers, called the *real fibers*, inherit a real structure from the real structure of the total space. We focus on fibrations whose critical values are all fixed by the action and call such fibrations *totally real*. We also assume that the fixed point set of the base space is oriented. We use the term *directed* to indicate such fibrations.

The main results of this article are exhibited in Sections 6 and 8 in which we treat the cases of fiber genus $g > 1$ and $g = 1$, respectively. In Section 6, we introduce *real Lefschetz chains* and prove that if $g > 1$, real Lefschetz chains are complete invariants of directed genus g totally real Lefschetz fibrations over the disk (Corollary 6.4). The case of $g = 1$ (elliptic fibrations) is considered in Theorem 8.1. We show that directed totally real elliptic Lefschetz fibrations over D^2 are determined uniquely by their decorated real Lefschetz chains. In both cases we study extensions of such fibrations to fibrations over a sphere and obtain complete invariants of directed totally real Lefschetz fibrations over a sphere.

MSC2010: primary 55R15, 55R55; secondary 57M05.
Keywords: Lefschetz fibrations, real structure, monodromy.

It is possible to give a purely combinatorial shape to decorated real Lefschetz chains. We discuss such combinatorial objects, which we call *necklace diagrams*, and their applications in [Salepci 2012]; see [Degtyarev 2011; Degtyarev and Salepci 2011] for other applications of necklace diagrams.

This paper is organized as follows. In Section 2, we settle the definitions and introduce basic notions. Section 3 is devoted to the topological classification of equivariant neighborhoods of real singular fibers. We show that real Lefschetz fibrations around real singular fibers are determined by the pair consisting of the inherited real structure on one of the nearby regular real fibers and the vanishing cycle that is invariant under the action of the real structure. We call such a pair a *real code*.

Real Lefschetz chains are, indeed, sequences of real codes each of which is associated to a neighborhood of a real singular fiber. Obviously, each real Lefschetz fibration with real critical values defines a real Lefschetz chain that is, by definition, invariant of the fibration. The natural question to ask is to what extent real Lefschetz chains determine the fibration.

In Section 4, we compute the fundamental group of the components of the space of real structures on a genus g surface. These computations are applied in Section 5 where we define a *strong boundary fiber sum* (that is, the boundary fiber sum of $\mathbb{C}$-marked real Lefschetz fibrations) and show that if the fiber genus is greater than 1, then the strong boundary fiber sum is well defined. Section 6 is devoted to $\mathbb{C}$-marked genus $g > 1$ fibrations. We show that directed $\mathbb{C}$-marked genus $g > 1$ totally real Lefschetz fibrations are classified by their *strong real Lefschetz chains*. As a corollary, we obtain the result for nonmarked fibrations.

Because of the different geometric nature of the surfaces of genus $g > 1$ and $g = 1$, we apply slightly different techniques to deal with the case of $g = 1$. In Section 7, we define a *boundary fiber sum* of nonmarked real elliptic Lefschetz fibrations. We observe that the boundary fiber sum is not always well defined. This observation leads to a decoration of directed totally real Lefschetz chains. In the last section, we introduce *decorated real Lefschetz chains* and prove that they are complete invariants of real elliptic Lefschetz fibrations. We also study extensions of such fibrations to fibrations over a sphere.

2. Basic definitions

Throughout the paper X will stand for a compact connected oriented smooth 4-manifold and B for a compact connected oriented smooth 2-manifold.

Definition 2.1. A *real structure* c_X on a smooth 4-manifold X is an orientation-preserving involution $c_X^2 = \mathrm{id}$, such that the set $\mathrm{Fix}(c_X)$ of fixed points of c_X is empty or of the middle dimension.

Two real structures c_X and c'_X are considered *equivalent* if there is an orientation-preserving diffeomorphism $\psi : X \to X$ such that $\psi \circ c_X = c'_X \circ \psi$.

A real structure c_B on a smooth 2-manifold B is an orientation-reversing involution $B \to B$. Such structures are similarly considered up to conjugation by orientation-preserving diffeomorphisms of B.

This definition mimics the properties of the standard complex conjugation on complex manifolds. Actually, around a fixed point, every real structure defined as above behaves like complex conjugation.

We will call a manifold together with a real structure a *real manifold* and the fixed point set the *real part*.

Remark 2.2. It is well known that for given g there is a finite number of equivalence classes of *real structures* on a genus g surface Σ_g. These classes can be distinguished by their *types* and the number of real components. Namely, one distinguishes two types of real structures: separating and nonseparating. A real structure is called *separating* if the complement of its real part has two connected components; otherwise we call it *nonseparating* (indeed, in the first case the quotient surface Σ_g/c is orientable while in the second case it is not). The number of real components of a real structure (note that the real part forms the boundary of Σ_g/c), can be at most $g+1$. This estimate is known as *Harnack inequality*. By looking at the possible number of connected components of the real part, one can see that on Σ_g there are $1+[g/2]$ separating real structures and $g+1$ nonseparating ones. A significant property of the case of genus 1 surfaces is that the number of real components, which can be 0, 1 or 2, is enough to distinguish the real structures.

In this article we stick to the following definition of Lefschetz fibrations.

Definition 2.3. A Lefschetz fibration is a surjective smooth map $\pi : X \to B$ such that

- $\pi(\partial X) = \partial B$ and the restriction $\partial X \to \partial B$ of π is a submersion;
- π has only a finite number of critical points (that is, the points where $d\pi$ is degenerate), all the critical points belong to $X \setminus \partial X$, and their images are distinct points of $B \setminus \partial B$; and
- around each of the critical points one can choose orientation-preserving charts $\psi : U \to \mathbb{C}^2$ and $\phi : V \to \mathbb{C}$ so that $\phi \circ \pi \circ \psi^{-1}$ is given by $(z_1, z_2) \to {z_1}^2 + {z_2}^2$.

When we want to specify the genus of the nonsingular fibers, we prefer calling them *genus g Lefschetz fibrations*. In particular, we will use the term *elliptic Lefschetz fibrations* when the genus is equal to one. For each integer g, we will fix a closed oriented surface of genus g, which will serve as a model for the fibers, and denote it by Σ_g. In what follows we will always assume that a Lefschetz fibration is *relatively minimal*; that, is none of its fibers contains a self intersection -1 sphere.

Definition 2.4. A real structure on a Lefschetz fibration $\pi : X \to B$ is a pair of real structures (c_X, c_B) of X and B such that the diagram

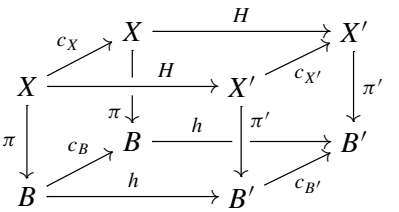
$$\begin{array}{ccc} X & \xrightarrow{c_X} & X \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ B & \xrightarrow{c_B} & B. \end{array}$$

commutes. A Lefschetz fibration equipped with a real structure is called a *real Lefschetz fibration* and is sometimes referred as RLF. When the fiber genus is 1, we call it a *real elliptic Lefschetz fibration* (abbreviated RELF).

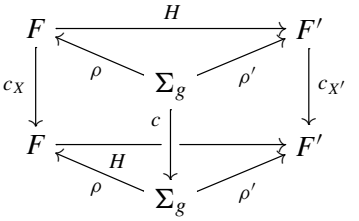
Definition 2.5. An $\mathbb{R}$-*marked* RLF is a triple (π, b, ρ) consisting of a real Lefschetz fibration $\pi : X \to B$, a real regular value b and a diffeomorphism $\rho : \Sigma_g \to F_b$ such that $c_X|_{F_b} \circ \rho = \rho \circ c$, where $c : \Sigma_g \to \Sigma_g$ is a real structure. Note that if $\partial B \neq \varnothing$, then b will be chosen in ∂B.

A $\mathbb{C}$-*marked* RLF is a triple $(\pi, \{m, \overline{m}\}, \{\rho, \bar{\rho}\})$, including a real Lefschetz fibration $\pi : X \to B$, a pair of regular values $m, \overline{m} = c_B(m)$ and a pair of diffeomorphisms $\rho : \Sigma_g \to F_m$ and $\bar{\rho} = c_X|_{F_m} \circ \rho : \Sigma_g \to F_{\overline{m}}$, where F_m and $F_{\overline{m}} = c_X(F_m)$ are the fibers over m and $\overline{m}$, respectively. As in the case of $\mathbb{R}$-marking, if $\partial B \neq \varnothing$, then we choose m in ∂B. When precision is not needed we will denote F_m and $F_{\overline{m}}$ by F and $\overline{F}$, respectively.

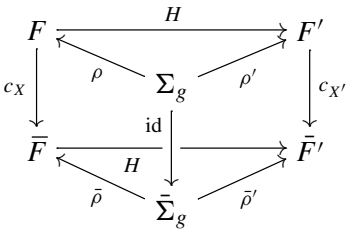
Two real Lefschetz fibrations $\pi : X \to B$ and $\pi' : X' \to B'$ are said to be *isomorphic* if there exist orientation-preserving diffeomorphisms $H : X \to X'$ and $h : B \to B'$ such that this diagram is commutative:



Two $\mathbb{R}$-marked RLFs are called isomorphic if they are isomorphic as RLFs such that $h(b) = b'$ and the following diagram is commutative:



Two $\mathbb{C}$-marked RLFs are called isomorphic if they are isomorphic as RLFs and the following diagram is well defined and commutative:



Definition 2.6. A real Lefschetz fibration $\pi : X \to B$ is called *directed* if the real part of (B, c_B) is oriented. (If c_B is separating, then we consider an orientation on the real part inherited from one of the halves $B \setminus \mathrm{Fix}(c_B)$.)

Two directed RLFs are isomorphic if they are isomorphic as RLFs with the additional condition that the diffeomorphism $h : B \to B$ preserves the chosen orientation on the real part.

Unless otherwise stated, all fibrations we consider are directed.

Remark 2.7. The notion of Lefschetz fibration can be slightly generalized to cover the case of fibrations whose fibers have nonempty boundary. Then, X turns into a manifold with corners and its boundary ∂X becomes naturally divided into two parts, the *vertical boundary* $\partial^v X$ that is the inverse image $\pi^{-1}(\partial B)$, and the *horizontal boundary* $\partial^h X$ that is formed by the boundaries of the fibers. We call such fibrations *Lefschetz fibrations with boundary*.

3. Elementary real Lefschetz fibrations

In this section, we classify real structures on a neighborhood of a real singular fiber of a real Lefschetz fibration. Such a neighborhood can be viewed as a Lefschetz fibration over a disc D^2 with a unique critical value $q = 0 \in D^2$. We call such a fibration an *elementary real Lefschetz fibration*. Without loss of generality, we may assume that the real structure on D^2 is the standard one, conj, induced from $\mathbb{C} \supset D^2$.

Let $\pi : X \to D^2$ be an elementary RLF. By definition, there exist equivariant local charts (U, ϕ_U) and (V, ϕ_V) around the critical point $p \in \pi^{-1}(0)$ and the critical value $0 \in D^2$, respectively, such that U and V are closed discs and $\pi|_U : (U, c_U) \to (V, \mathrm{conj})$ is equivariantly isomorphic (via ϕ_U and ϕ_V) to either of $\xi_\pm : (E_\pm, \mathrm{conj}) \to (D_\epsilon, \mathrm{conj})$, where

$$E_\pm = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1| \le \sqrt{\epsilon},\ |z_1^2 \pm z_2^2| \le \epsilon^2\}$$

and

$$D_\epsilon = \{t \in \mathbb{C} : |t| \le \epsilon^2\} \quad \text{for } 0 < \epsilon < 1,$$

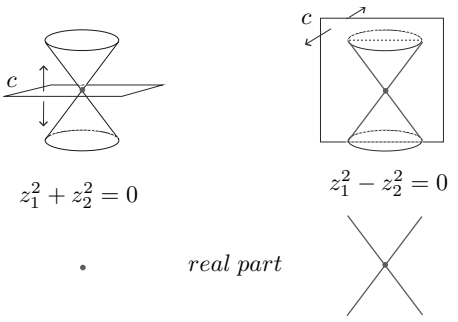


Figure 1. Actions of real structures on the singular fibers of $\xi_\pm$.

with $\xi_\pm(z_1, z_2) = z_1^2 \pm z_2^2$.

The real local models above, $\xi_\pm : E_\pm \to D_\epsilon$, can be seen as two real structures on the neighborhood of a critical point. These two real structures are not equivalent. The difference can be seen already at the level of the singular fibers: In the case of ξ_+, the two branches are imaginary and they are interchanged by the complex conjugation; in the case of ξ_- the two branches are both real (see Figure 1).

To understand the action of the real structures on the regular real fibers of $\xi_\pm$, we can use the branched covering defined by the projection $(z_1, z_2) \to z_1$. Thus:

- In the case of ξ_+, there are two types of real regular fibers; the fibers F_t with $t < 0$ have no real points, their vanishing cycles have invariant representatives (that is, $c(a_t) = a_t$ set-theoretically), and in this case, c acts on the invariant vanishing cycles as an antipodal involution; the fibers F_t with $t > 0$ have a circle as their real part and this circle is an invariant (pointwise fixed) representative of the vanishing cycle.
- In the case of ξ_-, all the real regular fibers are of the same type and the real part of such a fiber consists of two arcs each having its endpoints on the two different boundary components of the fiber; the vanishing cycles have invariant representatives, and c acts on them as a reflection.

Using the ramified covering $(z_1, z_2) \to z_1$, we observe that the horizontal boundary of the fibration $\xi_\pm$ is equivariantly trivial and has a distinguished equivariant trivialization. Moreover, since the complement of U in $\pi^{-1}(V)$ does not contain any critical point, X can be written as union of two RLFs with boundary: One of them, $U \to V$, is isomorphic to $\xi_\pm : E_\pm \to D_\epsilon$, and the other one is isomorphic to the trivial real fiber bundle $R \to D_\epsilon$ whose real fibers are equivariantly diffeomorphic to the complement of an open regular neighborhood of the vanishing cycle $a \subset F_b$. The action of the complex conjugation on the boundary components of the real fibers of $R \to D_\epsilon$ determines the type $\xi_\pm : E_\pm \to D_\epsilon$ of the model glued to $R \to D_\epsilon$: In the case of ξ_+, it switches the boundary components while in the

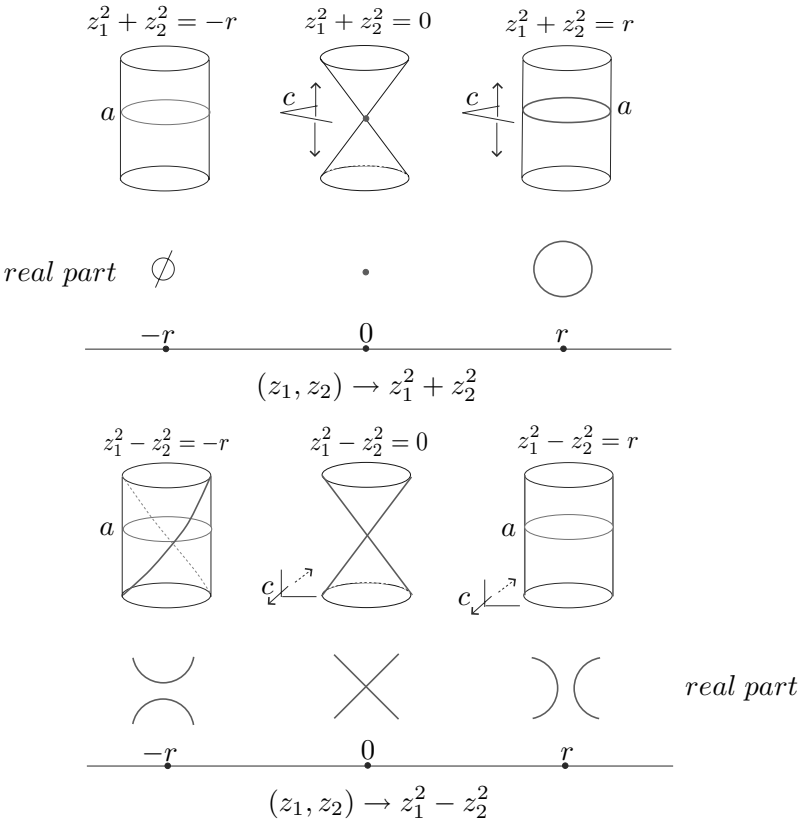


Figure 2. Nearby regular fibers of $\xi_\pm$ and the vanishing cycles.

case of ξ_-, boundary components are preserved (and the complex conjugation acts as a reflection on each of them).

We use the decomposition above to get first a classification of directed $\mathbb{R}$-marked elementary RLF, and then discuss the cases of $\mathbb{C}$-marked and nonmarked fibrations.

Let $\mathscr{A}_g^c$ denote the set of equivariant isotopy classes of noncontractible curves on a real surface (Σ_g, c), let $\mathscr{V}_g^c$ denote the set of equivariant isotopy classes of noncontractible embeddings $\nu : S^1 \times I \to \Sigma_g$ such that $c \circ \nu = \nu$, and let $\mathscr{L}_g^{\mathbb{R},c}$ denote the set of isomorphism classes of directed $\mathbb{R}$-marked elementary real Lefschetz fibrations whose distinguished fiber is identified with (Σ_g, c).

We consider the map $\Omega : \mathscr{V}_g^c \to \mathscr{L}_g^{\mathbb{R},c}$ defined as follows. Let $[\nu]$ be a class in $\mathscr{V}_g^c$ with a representative ν. As $c \circ \nu = \nu$, the closure Σ_g^ν of $\Sigma_g \setminus \nu(S^1 \times I)$ inherits a real structure from (Σ_g, c). Let $R^\nu = \Sigma_g^\nu \times D_\epsilon \to D_\epsilon$ be the trivial real fibration with the real structure $c_{R^\nu} = (c, \text{conj}) : R^\nu \to R^\nu$ and let $E_\pm^\nu \to D_\epsilon$ denote the model $\xi_\pm : E \to D_\epsilon$ whose marked fiber is identified with $\nu(S^1 \times I)$. Depending on the real structure on the horizontal boundary $S^1 \times D_\epsilon \to D_\epsilon$ (where the real structure

on $S^1 \times D_\epsilon$ is taken as $(c_{\partial\Sigma_g^\nu}, \text{conj})$ of $R^\nu \to D_\epsilon$, we choose either of $E_\pm^\nu \to D_\epsilon$. We then glue $R^\nu \to D_\epsilon$ and the suitable model $E_\pm^\nu \to D_\epsilon$ along their horizontal trivial boundaries to get a fibration in $\mathscr{L}_g^{\mathbb{R},c}$.

Lemma 3.1. *$\Omega : \mathscr{V}_g^c \to \mathscr{L}_g^{\mathbb{R},c}$ is well defined.*

Proof. Let $\nu_t : S^1 \times I \to \Sigma_g$ be an isotopy between ν_0 and ν_1. Then, there exists an equivariant ambient isotopy $\Psi_t : \Sigma_g \to \Sigma_g$ such that $\Psi_0 = \text{id}$ and $\nu_t = \Psi_t \circ \nu_0$ with $\Psi_t \circ c = c \circ \Psi_t$ for all t. The diffeomorphism Ψ_1 induces equivariant diffeomorphisms $\Psi_1^R : R^{\nu_0} \to R^{\nu_1}$ and $\Psi_1^E : E_\pm^{\nu_0} \to E_\pm^{\nu_1}$ that respect the fibrations and the gluing; thus, it gives an isomorphism of the images $\Omega([\nu_0])$ and $\Omega([\nu_1])$ as $\mathbb{R}$-marked fibrations. □

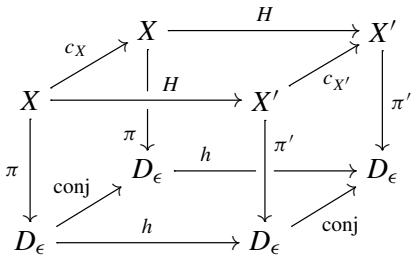
Since $c \circ \nu = \nu$, we have $c(\nu(S^1 \times \{\frac{1}{2}\})) = \nu(S^1 \times \{\frac{1}{2}\}))$. Hence, we can define $\varepsilon : \mathscr{V}_g^c \to \mathscr{A}_g^c$ such that $\varepsilon([\nu]) = [\nu(S^1 \times \{\frac{1}{2}\})]$. This mapping is two-to-one. Since the monodromy does not depend on the orientation of the vanishing cycle, there exists a well-defined mapping $\hat{\Omega}$ such that the following diagram commutes:

$$\begin{array}{ccc} \mathscr{V}_g^c & \xrightarrow{\varepsilon} & \mathscr{A}_g^c \\ {\scriptstyle\Omega}\downarrow & \swarrow_{\hat{\Omega}} & \\ \mathscr{L}_g^{\mathbb{R},c}. & & \end{array}$$

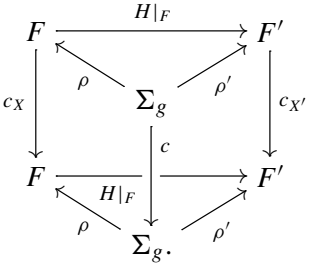
Theorem 3.2. *$\hat{\Omega} : \mathscr{A}_g^c \to \mathscr{L}_g^{\mathbb{R},c}$ is a bijection.*

Proof. As discussed in the beginning of the section, any elementary RLF can be divided equivariantly into two RLFs with boundary: an equivariant neighborhood of the critical point (isomorphic to one of the models, $\xi_\pm$) and the complement of this neighborhood (isomorphic to a trivial real Lefschetz fibration). Such a decomposition defines the equivariant isotopy class of the vanishing cycle. Thus, $\hat{\Omega}$ is surjective.

To show that $\hat{\Omega}$ is injective, let us consider the classes $[a], [a'] \in \mathscr{A}_g^c$ such that $\hat{\Omega}([a]) = \hat{\Omega}([a'])$. Let $\pi : X \to D_\epsilon$ and $\pi' : X' \to D_\epsilon$ denote the images of $[a]$ and $[a']$, respectively. Since $\hat{\Omega}$ is well defined, there exist equivariant orientation-preserving diffeomorphisms $H : X \to X'$ and $h : D_\epsilon \to D_\epsilon$ such that we have the commutative diagrams



and



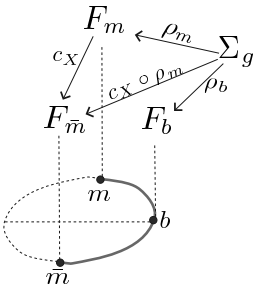


Figure 3. Relation between $\mathbb{R}$-marking and $\mathbb{C}$-marking.

Clearly, $H(\rho(a))$ is equivariantly isotopic to $\rho'(a')$, where a and a' are representatives of $[a]$ and $[a']$, respectively. Moreover, since $H|_F \circ \rho = \rho'$, we have $H(\rho(a)) = \rho'(a)$, so $\rho'(a)$ is equivariant isotopic to $\rho'(a')$.

For $t \in [0, 1]$, let $\psi_t : F' \to F'$ such that $\psi_0 = \mathrm{id}$ and $\psi_1(\rho'(a)) = \rho'(a')$ and that $\psi_t \circ c' = c' \circ \psi_t$. Then, $\Psi_t = \rho'^{-1} \circ \psi_t \circ \rho' : \Sigma_g \to \Sigma_g$ is the required isotopy between a and a'. □

Theorem 3.2 shows that c-equivariant isotopy classes of vanishing cycles on (Σ_g, c) classify directed $\mathbb{R}$-marked elementary RLFs. To obtain a classification for directed $\mathbb{C}$-marked RLFs, we study the difference between two $\mathbb{C}$-markings; see Figure 3.

Let $(\{m, \overline{m}\}, \{\rho_m, c_X \circ \rho_m\})$ be a $\mathbb{C}$-marking on a directed RLF $\pi : X \to D^2$. The complement, $\partial D^2 \setminus \{m, \overline{m}\}$, has two pieces $S_\pm$ (left/right semicircles) distinguished by the direction. By considering a trivialization of the fibration over the piece of S_+ connecting m to the marked real point b (the trivialization over the piece connecting $\overline{m}$ to the real point obtain by the symmetry), we can pull the marking $\rho_m : \Sigma_g \to F_m$ to F_b in order to obtain a marking $\rho_b : \Sigma_g \to F_b$ and a real structure $c = \rho_b^{-1} \circ c_X \circ \rho_b : \Sigma_g \to \Sigma_g$. Any other trivialization results in another marking isotopic to ρ_b and a real structure isotopic to $c : \Sigma_g \to \Sigma_g$. Hence, a directed elementary $\mathbb{C}$-marked RLF defines a vanishing cycle defined up to c-equivariant isotopy, where the real structure c is also considered up to isotopy.

Definition 3.3. A pair (c, a) of a real structure $c : \Sigma_g \to \Sigma_g$ and a noncontractible simple closed curve $a \in \Sigma_g$ is called a *real code* if $c(a) = a$.

Two real codes (c_0, a_0) and (c_1, a_1) are said to be isotopic if there exist a pair (c_t, a_t) of isotopies of real structures and vanishing cycles such that $c_t(a_t) = a_t$ for all $t \in [0, 1]$. Two real codes (c_0, a_0) and (c_1, a_1) are called conjugate if there is an orientation-preserving diffeomorphism $\phi : \Sigma_g \to \Sigma_g$ such that $\phi \circ c_0 = c_1 \circ \phi$ and that $\phi(a_0)$ is isotopic to a_1.

We denote the isotopy class of the real code by $[c, a]$ and the conjugacy class by $\{c, a\}$.

Proposition 3.4. *There is a bijection between the isomorphism classes of directed* $\mathbb{C}$*-marked elementary* RLF*s and the isotopy classes of real codes.*

Proof. We already discussed how to assign a real code to a directed $\mathbb{C}$-marked elementary RLF. It is straightforward to check that this map is well defined and surjective.

To show that it is injective, we consider two isotopy classes $[c_1, a_1]$ and $[c_2, a_2]$ such that $[c_1, a_1] = [c_2, a_2]$. Let

$$(\pi_1 : X_1 \to D^2, \{m_1, \overline{m}_1\}, \{\rho_{m_1}, \bar{\rho}_{m_1}\}) \quad \text{and} \quad (\pi_2 : X_2 \to D^2, \{m_2, \overline{m}_2\}, \{\rho_{m_2}, \bar{\rho}_{m_2}\})$$

be two directed $\mathbb{C}$-marked elementary RLFs associated to the classes $[c_1, a_1]$ and $[c_2, a_2]$, respectively. We need to show that π_1 and π_2 are isomorphic as directed $\mathbb{C}$-marked RLFs.

Note that we can always choose a representative c for both $[c_1]$ and $[c_2]$ such that $[a_1] = [a_2] \in \mathscr{A}_g^c$. Then, by Theorem 3.2, π_1 is isomorphic to π_2 as $\mathbb{R}$-marked RLF*s*. An isomorphism of $\mathbb{R}$-marked RLFs may not preserve the $\mathbb{C}$-markings; however, it can be modified to preserve them.

Up to homotopy one can identify X_2 with a subset $\mathring{X}_2$ of X_1. Since the difference $X_1 \setminus \mathring{X}_2$ has no singular fiber, one can transform the marking $\mathring{m}_2$ of $\mathring{X}_2$ to m_1, preserving the real marking and the trivializations over the corresponding paths, S_+ and $\mathring{S}_+$ (see Figure 4). This way we get an isomorphism of $\mathbb{C}$-marked RLFs preserving the isomorphism class of $\mathbb{R}$-marked RLFs. □

For fibrations without marking we allow $[c, a]$ to change by an equivariant diffeomorphism. Hence, we have the following:

Corollary 3.5. *There is a bijection between the set of conjugacy classes of real codes and the set of classes of directed nonmarked elementary real Lefschetz fibrations.*

Remark 3.6. As the classification of real structures on a genus g surface is known, it is possible to enumerate the conjugacy classes $\{c, a\}$ of real codes. In the case when a is nonseparating, there are 6 classes if $g = 1$; $8g - 3$ classes if $g > 1$ and

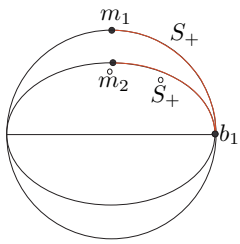


Figure 4. The difference of two $\mathbb{C}$-markings.

is odd; and $8g-4$ classes otherwise. The formulas for separating curves can be found in [Salepci 2007].

Remark 3.7. There is no preferable real fiber over the boundary of the disk if the fibration is not directed. Thus, to an elementary nondirected RLF, we can associate two real codes (c_-, a_-) and (c_+, a_+) extracted from the "left" and "right" real fibers, respectively. It is a fundamental property of the monodromies of real Lefschetz fibrations that the real structures c_- and c_+ are related by the monodromy such that $c_+ \circ c_- = t_{a_-} = t_{a_+}$; see [Salepci 2010].

4. Equivariant diffeomorphisms and the space of real structures

In this section we compute the fundamental group of the space of real structures on a genus g surface. The computations will be essential in next sections.

Let $\mathscr{C}^c(\Sigma_g)$ denote the space of real structures on Σ_g that are isotopic to a fixed real structure c, and let $\mathrm{Diff}_0(\Sigma_g)$ denote the group of orientation-preserving diffeomorphisms of Σ_g that are isotopic to the identity. We consider two subgroups of $\mathrm{Diff}_0(\Sigma_g)$: One, denoted $\mathrm{Diff}_0^c(\Sigma_g)$, consists of those diffeomorphisms that commute with c, and the other, $\mathrm{Diff}_0(\Sigma_g, c)$, is the group of diffeomorphisms that are c-equivariantly isotopic to the identity. The group $\mathrm{Diff}_0(\Sigma_g)$ acts transitively on $\mathscr{C}^c(\Sigma_g)$ by conjugation. The stabilizer of this action is the group $\mathrm{Diff}_0^c(\Sigma_g)$. Hence, $\mathscr{C}^c(\Sigma_g)$ can be identified with the homogeneous space $\mathrm{Diff}_0(\Sigma_g)/\mathrm{Diff}_0^c(\Sigma_g)$.

Lemma 4.1. *The space* $\mathrm{Diff}_0^c(\Sigma_g)$ *is connected for all* $c : \Sigma_g \to \Sigma_g$ *if* $g > 1$, *and for* $c : \Sigma_g \to \Sigma_g$, *which has one real component, if* $g = 1$.

Proof. We will use different techniques for the cases $g > 1$ and $g = 1$.

The case of $g > 1$: We consider the fiber bundle description of conformal structures on Σ_g, introduced in [Earle and Eells 1969]. Let Conf_{Σ_g} denote the space of conformal structures on Σ_g equipped with C^∞-topology. The group $\mathrm{Diff}_0(\Sigma_g)$ acts on Conf_{Σ_g} by composition from the right. This action is proper, continuous, and effective; hence, $\mathrm{Conf}_{\Sigma_g} \to \mathrm{Conf}_{\Sigma_g}/\mathrm{Diff}_0(\Sigma_g)$ is a principal $\mathrm{Diff}_0(\Sigma_g)$-fiber bundle; see [Earle and Eells 1969]. The quotient is the Teichmüller space of Σ_g, denoted $\mathrm{Teich}_{\Sigma_g}$. Note that conformal structures can be seen as equivalence classes of Riemannian metrics with respect to the relation that two Riemannian metrics are equivalent if they differ by a positive function on Σ_g. Let Riem_{Σ_g} denote the space of Riemannian metrics on Σ_g. Then, we have the fibrations

$$\begin{array}{ccc} \{u : \Sigma_g \to \mathbb{R} : u > 0\} & \longrightarrow & \mathrm{Riem}_{\Sigma_g} \\ & & \downarrow{\scriptstyle p_2} \\ \mathrm{Diff}_0(\Sigma_g) & \longrightarrow & \mathrm{Conf}_{\Sigma_g} \\ & & \downarrow{\scriptstyle p_1} \\ & & \mathrm{Teich}_{\Sigma_g}. \end{array}$$

The real structure c has an action $\mathrm{Diff}_0(\Sigma_g)$ by conjugation. This action extends to Conf_{Σ_g} and Riem_{Σ_g} as follows: Fix a section $s : \mathrm{Teich}_{\Sigma_g} \to \mathrm{Conf}_{\Sigma_g}$ of the bundle p_1 and consider a family of diffeomorphisms $\phi_\zeta^s : \mathrm{Diff}_0(\Sigma_g) \to p_1^{-1}(\zeta)$ parametrized by $\mathrm{Teich}_{\Sigma_g}$ such that $\phi_\zeta^s(\mathrm{id}) = s(\zeta)$. Let $s(\zeta) = [\mu_x]$ for some Riemannian metric μ_x on Σ_g. Then, define $\phi_\zeta^s(f(x)) = [\mu_{f(x)}]$ for all $f \in \mathrm{Diff}_0(\Sigma_g)$. The action of the real structure, thus, can be written as $c.[\mu_{f(x)}] = [\mu_{c\circ f\circ c(x)}]$. Clearly the definition does not depend on the choice of the representative of the class $[\mu_{f(x)}]$, so the action extends to Riem_{Σ_g}.

Let $\mathrm{Fix}_{\mathrm{Conf}_{\Sigma_g}}(c)$ denote the set of fixed points of the action of c on Conf_{Σ_g} and let $\mathrm{Fix}_{\mathrm{Riem}_{\Sigma_g}}(c)$ be the set of fixed points on Riem_{Σ_g}. Note that $s(\zeta) = \phi_\zeta^s(\mathrm{id})$ is in $\mathrm{Fix}_{\mathrm{Conf}_{\Sigma_g}}(c)$ for all $\zeta \in \mathrm{Teich}_{\Sigma_g}$. Indeed, each $[\mu_{f(x)}]$ for $f \in \mathrm{Diff}_0^c(\Sigma_1)$ is in $\mathrm{Fix}_{\mathrm{Conf}_{\Sigma_g}}(c)$.

The space $\mathrm{Fix}_{\mathrm{Conf}_{\Sigma_g}}(c)$ is connected. If $\mathrm{Fix}_{\mathrm{Conf}_{\Sigma_g}}(c)$ were disconnected, then the inverse image $\mathrm{Fix}_{\mathrm{Riem}_{\Sigma_g}}(c)$ would also be disconnected in Riem_{Σ_g}. However, it is known that Riem_{Σ_g} is convex; thus, $\mathrm{Fix}_{\mathrm{Riem}_{\Sigma_g}}(c)$ is convex, so it is connected. Therefore, $\mathrm{Fix}_{\mathrm{Conf}_{\Sigma_g}}(c) \cap \mathrm{Diff}_0(\Sigma_g) = \mathrm{Diff}_0^c(\Sigma_g)$ is connected since $\mathrm{Fix}_{\mathrm{Conf}_{\Sigma_g}}(c)$ is a union of sections.

The case of $g = 1$: If c has one real component, then the quotient Σ_1/c is the Möbius band. The space of diffeomorphisms of the Möbius band has two connected components [Hamstrom 1965]: the identity component and the component of the diffeomorphism induced from the reflection of $I \times I$ with respect to $I \times \frac{1}{2}$ (if the Möbius band is obtained from $I \times I$, we identify the points $t \times 0$ with the points $1-t \times 1$ for $t \in I = [0, 1]$). This diffeomorphism is not isotopic to the identity because before identifying the ends it reverses the orientation of $I \times I$, and it lifts to a diffeomorphism of Σ_1 (considered as the obvious quotient of $[-1, 1] \times [-1, 1]$) induced from the central symmetry of $[-1, 1] \times [-1, 1]$. This diffeomorphism is not isotopic to the identity on Σ_1 since it reverses the orientation of the real curve.

Therefore, we have

$$\{f : \Sigma_1/c \to \Sigma_1/c : \hat{f} : \Sigma_1 \to \Sigma_1 \text{ is isotopic to id}\} = \{f : \Sigma_1/c \to \Sigma_1/c : f \cong \mathrm{id}\}.$$

The former is identified by $\mathrm{Diff}_0^c(\Sigma_1)$ and the latter is connected. □

Lemma 4.2. *For any real structure $c : \Sigma_g \to \Sigma_g$,*

$$\pi_1(\mathrm{Diff}_0(\Sigma_g)/\mathrm{Diff}_0(\Sigma_g, c), \mathrm{id}) = \begin{cases} 0 & \text{if } g > 1\,, \\ \mathbb{Z} & \text{if } g = 1. \end{cases}$$

Proof. Note that the subgroup $\mathrm{Diff}_0(\Sigma_g, c)$ acts on $\mathrm{Diff}_0(\Sigma_g)$ by composition from the left. Such an action is free, so $\mathrm{Diff}_0(\Sigma_g) \to \mathrm{Diff}_0(\Sigma_g)/\mathrm{Diff}_0(\Sigma_g, c)$ is a $\mathrm{Diff}_0(\Sigma_g, c)$-fiber bundle. The fibers, $\mathrm{Diff}_0(\Sigma_g, c)$, can be identified with the group $\mathrm{Diff}_0(\Sigma_g/c)$ because the lifting of diffeomorphisms of Σ_g/c can always be assured by means of the orientation double cover of Σ_g/c. (Note that if c is nonseparating,

then Σ_g/c is nonorientable. In this case, $\mathrm{Diff}_0(\Sigma_g/c)$ denotes the space of all diffeomorphisms of Σ_g/c and $\mathrm{Diff}_0(\Sigma_g/c)$ is component of the identity.)

Now, we consider the long exact homotopy sequence of this fibration:

$$\cdots \to \pi_2(\mathrm{Diff}_0(\Sigma_g)) \to \pi_2(\mathrm{Diff}_0(\Sigma_g)/\mathrm{Diff}_0(\Sigma_g, c)) \to \pi_1(\mathrm{Diff}_0(\Sigma_g, c)) \\ \to \pi_1(\mathrm{Diff}_0(\Sigma_g)) \to \pi_1(\mathrm{Diff}_0(\Sigma_g)/\mathrm{Diff}_0(\Sigma_g, c)) \to \pi_0(\mathrm{Diff}_0(\Sigma_g)) \to \cdots.$$

The case of $g > 1$: The space $\mathrm{Diff}_0(\Sigma_g)$ is contractible for $g > 1$ [Earle and Eells 1969], as is $\mathrm{Diff}_0(\Sigma_g/c)$ [Earle and Schatz 1970]. Therefore, from the homotopy long exact sequence of the fibration we get $\pi_1(\mathrm{Diff}_0(\Sigma_g)/\mathrm{Diff}_0(\Sigma_g, c), \mathrm{id}) = 0$.

The case of $g = 1$: It is known that Σ_1 is a deformation retract of $\mathrm{Diff}_0(\Sigma_1)$ [Ivanov 2001], so the space $\mathrm{Diff}_0(\Sigma_1)$ can be considered as a group generated by the rotations that lift to the standard translations on the universal cover.

To understand $\mathrm{Diff}_0(\Sigma_g, c)$, we first consider the case when c has two real components. Note that, in this case, the quotient Σ_1/c is topologically an annulus, so $\pi_1(\mathrm{Diff}_0(\Sigma_1/c), \mathrm{id}) = \mathbb{Z}$; see [Ivanov 2001]. We fix an identification of $\varrho : \mathbb{C}/\mathbb{Z}^2 \to \Sigma_1$ such that the real structure c is the one induced from the standard complex conjugation on $\mathbb{C}$. We consider the family

$$\begin{aligned} R'^1_t :\ \mathbb{C}/\mathbb{Z}^2 &\to \mathbb{C}/\mathbb{Z}^2, & R'^2_t :\ \mathbb{C}/\mathbb{Z}^2 &\to \mathbb{C}/\mathbb{Z}^2, \\ (x+iy)_{\mathbb{Z}^2} &\mapsto (x+t+iy)_{\mathbb{Z}^2}, & (x+iy)_{\mathbb{Z}^2} &\mapsto (x+i(y+t))_{\mathbb{Z}^2} \end{aligned}$$

of diffeomorphisms, where $t \in [0, 1]$ and $(x+iy)_{\mathbb{Z}^2}$ denotes the equivalence class of $x+iy$ in $\mathbb{C}/\mathbb{Z}^2$. Clearly $R'^j_0 = R'^j_1 = \mathrm{id}$, and R'^j_t is isotopic to identity for each $t \in [0, 1]$ and $j = 1, 2$. The homotopy classes of $R^1_t = \varrho \circ R'^1_t \circ \varrho^{-1}$ and $R^2_t = \varrho \circ R'^2_t \circ \varrho^{-1}$ form a basis of $\pi_1(\mathrm{Diff}_0(\Sigma_1), \mathrm{id})$. Moreover, with respect to the identification ϱ, each diffeomorphism R^1_t is in $\mathrm{Diff}_0(\Sigma_1, c)$, so the loop R^1_t is a generator of $\pi_1(\mathrm{Diff}_0(\Sigma_1, c), \mathrm{id})$. Thus, from the homotopy exact sequence we get $\pi_1(\mathrm{Diff}_0(\Sigma_1)/\mathrm{Diff}_0(\Sigma_1, c), \mathrm{id}) = \mathbb{Z}$.

If c has no real component, then the quotient Σ_1/c is a Klein bottle, so the group $\mathrm{Diff}_0(\Sigma_1/c)$ is isomorphic to S^1 and is generated by the rotation that lifts to a translation in the universal cover of the Klein bottle [Hamstrom 1965]. Let us now fix an identification $\varrho : \mathbb{R}^2/\mathbb{Z}^2 \to \Sigma_1$ such that the real structure c is induced from the real structure

$$\mathbb{R}^2/\mathbb{Z}^2 \to \mathbb{R}/\mathbb{Z}^2, \quad (x, y)_{\mathbb{Z}^2} \mapsto (x + \tfrac{1}{2}, -y)_{\mathbb{Z}^2}.$$

The classes of family of diffeomorphisms $R^j_t = \varrho \circ R'^j_t \circ \varrho^{-1}$ for $j = 1, 2$, where

$$\begin{aligned} R'^1_t :\ \mathbb{R}^2/\mathbb{Z}^2 &\to \mathbb{R}^2/\mathbb{Z}^2, & R'^2_t :\ \mathbb{R}^2/\mathbb{Z}^2 &\to \mathbb{R}^2/\mathbb{Z}^2, \\ (x, y)_{\mathbb{Z}^2} &\mapsto (x+t, y)_{\mathbb{Z}^2}, & (x, y)_{\mathbb{Z}^2} &\mapsto (x, y+t)_{\mathbb{Z}^2}, \end{aligned}$$

form a basis of $\pi_1(\mathrm{Diff}_0(\Sigma_1), \mathrm{id})$. With respect to the identification ϱ each diffeomorphism R_t^1 is in $\mathrm{Diff}_0(\Sigma_1, c)$, so R_t^1 is a generator of $\pi_1(\mathrm{Diff}_0(\Sigma_1, c), \mathrm{id}) = \mathbb{Z}$. Therefore, we get $\pi_1(\mathrm{Diff}_0(\Sigma_1)/\mathrm{Diff}_0(\Sigma_1, c), \mathrm{id}) = \mathbb{Z}$.

If c has a unique real component, C, then the restriction $f|_C$ of $f \in \mathrm{Diff}_0(\Sigma_1, c)$ defines a diffeomorphism of C. Such a restriction defines a fibration $\mathrm{Diff}_0(\Sigma_1, c) \to \mathrm{Diff}_0(C)$ whose fibers isomorphic to $\mathrm{Diff}_0(\Sigma_1, C) = \{f \in \mathrm{Diff}_0(\Sigma_1, c) : f|_C = \mathrm{id}\}$. Note that $\mathrm{Diff}_0(\Sigma_1, C) \cong \mathrm{Diff}_0(\overline{\Sigma_1 \setminus C}, \partial)$ where $\overline{\Sigma_1 \setminus C}$ denotes the closure of $\Sigma_1 \setminus C$ and $\mathrm{Diff}_0(\overline{\Sigma_1 \setminus C}, \partial)$ the group diffeomorphisms of $\overline{\Sigma_1 \setminus C}$ that are identity on the boundary.

Topologically $\Sigma_1 \setminus C$ is an annulus, so $\mathrm{Diff}_0(\overline{\Sigma_1 \setminus C}, \partial)$ is contractible; see [Ivanov 2001]. From the homotopy long exact sequence of the fibration

$$\begin{array}{ccc} \mathrm{Diff}_0(\Sigma_1, C) & \longrightarrow & \mathrm{Diff}_0(\Sigma_1, c) \\ & & \big\downarrow \\ & & \mathrm{Diff}_0(C), \end{array}$$

we get $\pi_k(\mathrm{Diff}_0(\Sigma_1, c), \mathrm{id}) \cong \pi_k(\mathrm{Diff}_0(C), \mathrm{id})$ for all k.

Let us now choose an identification $\varrho : \mathbb{C}/\Lambda \to \Sigma_1$, where Λ is the lattice generated by $v_1 = (1/\sqrt{2}, 1/\sqrt{2})$ and $v_2 = (1/\sqrt{2}, -1/\sqrt{2})$. Then, the real structure c can be taken as the one induced from the complex conjugation on $\mathbb{C}$.

We consider $R_i'(t) : \mathbb{C}/\Lambda \to \mathbb{C}/\Lambda$, $t \in [0, 1]$ such that

$$\begin{array}{rcl} R'^1_t : \mathbb{C}/\Lambda & \to & \mathbb{C}/\Lambda, \\ (x+iy)_\Lambda & \mapsto & (x+t+iy)_\Lambda, \end{array} \qquad \begin{array}{rcl} R'^2_t : \mathbb{C}/\Lambda & \to & \mathbb{C}/\Lambda, \\ (x+iy)_\Lambda & \mapsto & (x+i(y+t))_\Lambda. \end{array}$$

Again, the classes of $R_t^j = \varrho \circ R'^j_t \circ \varrho^{-1}$ for $j = 1, 2$ form a basis for $\mathrm{Diff}_0(\Sigma_1)$, while R_t^1 can be taken as a generator for $\pi_1(\mathrm{Diff}_0(\Sigma_1, c), \mathrm{id}) = \mathbb{Z}$. Therefore, $\pi_1(\mathrm{Diff}_0(\Sigma_1)/\mathrm{Diff}_0(\Sigma_1, c), \mathrm{id}) = \mathbb{Z}$. □

Proposition 4.3. *For any real structure $c : \Sigma_g \to \Sigma_g$,*

$$\pi_1(\mathscr{C}^c(\Sigma_g)) = \pi_1(\mathrm{Diff}_0(\Sigma_g)/\mathrm{Diff}_0^c(\Sigma_g), \mathrm{id}) = \begin{cases} 0 & \text{if } g > 1, \\ \mathbb{Z} & \text{if } g = 1. \end{cases}$$

Proof. By Lemma 4.1, $\mathrm{Diff}_0^c(\Sigma_g)$ is connected for all real $c : \Sigma_g \to \Sigma_g$ with $g > 1$ and for the real structure $c : \Sigma_1 \to \Sigma_1$ that has one real component. Hence, in these cases $\mathrm{Diff}_0^c(\Sigma_1) = \mathrm{Diff}_0(\Sigma_1, c)$, so the result follows from Lemma 4.2.

In the case when $c : \Sigma_1 \to \Sigma_1$ has 2 real components, the space $\mathrm{Diff}_0^c(\Sigma_1)$ has two connected components. Note that the diffeomorphism $R_{1/2}^2$ (induced from the translation $(x+iy)_{\mathbb{Z}^2} \to (x+i(y+1/2))_{\mathbb{Z}^2}$ on $\mathbb{C}/\mathbb{Z}^2$) is equivariant; however, it is not equivariantly isotopic to the identity. Hence, $\mathrm{Diff}_0^c(\Sigma_1)$ has two components: the component $\mathrm{Diff}_0(\Sigma_1, c)$ of the identity and the component of the rotation $R_{1/2}^2$. (In what follows, we denote $R_{1/2}^2$ by $R_{1/2}$.)

We identify rotations in $\mathrm{Diff}_0(\Sigma_1) \setminus \mathrm{Diff}_0(\Sigma_1, c)$ with S^1 by letting $R_t^2 \to 2\pi t$. Then rotations in the quotient $\mathrm{Diff}_0(\Sigma_1)/\mathrm{Diff}_0^c(\Sigma_1)$ are identified with $S^1/_{\theta\sim(\theta+\pi)}$, so we have $\pi_1(\mathrm{Diff}_0(\Sigma_1)/\mathrm{Diff}_0^c(\Sigma_1), \mathrm{id}) = \mathbb{Z}$.

The case when $c : \Sigma_1 \to \Sigma_1$ has no real component can be treated similarly using the identification $\varrho : \mathbb{R}^2/\mathbb{Z}^2 \to \Sigma_1$. □

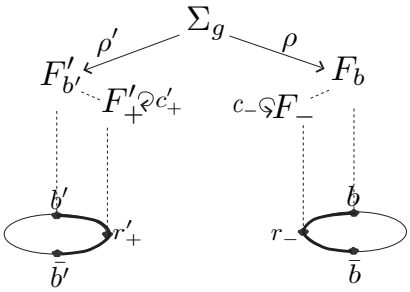
5. Boundary fiber sum of $\mathbb{C}$-marked real Lefschetz fibrations

Let (D^2, conj) be a real disk with oriented real part. We denote by $S^{\pm}$ the upper/lower semicircles of ∂D^2. We consider also left/right semicircles, denoted by $S_{\pm}$, and the quarter circles $S_{\pm}^{\pm} = S^{\pm} \cap S_{\pm}$. (Here directions right/left and up/down are determined by the orientations of D^2 and its real part.) Let $r_{\pm}$ be the real points of $S_{\pm}$, and $c_{\pm}$ the real structures on $F_{\pm} = \pi^{-1}(r_{\pm})$.

Definition 5.1. Let

$$(\pi' : X' \to D^2, \{b', \bar{b}'\}, \{\rho', \bar{\rho}'\}) \quad \text{and} \quad (\pi : X \to D^2, \{b, \bar{b}\}, \{\rho, \bar{\rho}\})$$

be two directed $\mathbb{C}$-marked real Lefschetz fibrations such that the real structures c'_+ on F'_+ and c_- on F_- induce (via the markings) isotopic real structures on Σ_g. Then, we define the *strong boundary fiber sum* (the boundary fiber sum of $\mathbb{C}$-marked RLFs) as follows.



We choose trivializations of $\pi'^{-1}(S_+^+)$ and $\pi^{-1}(S_-^+)$ such that the pull backs of c'_+ and c_- give the same real structure c on Σ_g. The trivialization of $\pi'^{-1}(S_+)$ can be obtained as a union $\Sigma_g \times S_+^+ \cup \Sigma_g \times S_+^-/_{(x,1_+)\sim(c(x),1_-)}$ and similarly $\pi^{-1}(S_-)$ is obtained as $\Sigma_g \times S_-^+ \cup \Sigma_g \times S_-^-/_{(x,-1_+)\sim(c(x),-1_-)}$. The strong boundary fiber sum $X' \natural_{\Sigma_g} X \to D^2 \natural D^2$ is thus obtained by gluing $\pi'^{-1}(S_+)$ to $\pi^{-1}(S_-)$ via the identity map.

Remark 5.2. (1) In fact, the construction described above creates a manifold with corners, but there is a canonical way to smooth the corners; hence, the strong boundary fiber sum is the manifold obtained by smoothing the corners.

(2) By definition, the strong boundary fiber sum is associative but not commutative.

(3) The strong boundary fiber sum of $\mathbb{C}$-marked RLFs is naturally $\mathbb{C}$-marked.

Proposition 5.3. *If $g > 1$, then the strong boundary fiber sum $X' \natural_{\Sigma_g} X \to D^2$ of directed $\mathbb{C}$-marked genus g real Lefschetz fibrations is well defined up to isomorphism of $\mathbb{C}$-marked* RLF*s.*

Proof. The boundary fiber sum does not affect the fibrations outside a small neighborhood of the interval where the gluing is made. Let us choose a neighborhood N that is real and far from the critical set. Obviously, the real structures on the fibers over the real points of N are isotopic. Therefore, each fiber sum defines a path in the space of real structures on Σ_g, and the difference of two strong boundary fiber sums gives a loop in this space. Thus, the result follows from the contractibility (shown in Proposition 4.3) of this loop in the case of $g > 1$. □

6. Strong real Lefschetz chains associated to $\mathbb{C}$-marked real Lefschetz fibrations

Let's consider a directed $\mathbb{C}$-marked totally real Lefschetz fibration $\pi : X \to D^2$. We slice D^2 into smaller discs $D_1, D_2, \ldots, D_n$ (ordered with respect to the orientation of the real part of (D^2, conj)) such that each D_i contains only one critical value and the base point b (which is chosen to be the "north pole" as in Figure 5). Let $r_1, r_2, \ldots, r_n, r_{n+1}$ be the real points of $\bigcup_{i=1}^{n} \partial D_i$ and let c_i be the real structure on Σ_g pulled back from the inherited real structure of F_{r_i}.

As claimed in Remark 3.7, we have $c_{i+1} \circ c_i = t_{a_i}$ for each fibration over D_i, where a_i denotes the corresponding vanishing cycle. As shown in Proposition 3.4, each $\mathbb{C}$-marked real Lefschetz fibration over D_i is determined by the isotopy class $[c_i, a_i]$ of a real code. Hence, the fibration $\pi : X \to D^2$ yields a sequence of real codes $[c_i, a_i]$ satisfying $c_{i+1} \circ c_i = t_{a_i}$. Clearly this sequence is an invariant of π.

Definition 6.1. A sequence $[c_1, a_1], [c_2, a_2], \ldots, [c_n, a_n]$ of isotopy classes of real codes is called a *strong real Lefschetz chain* if we have $c_{i+1} \circ c_i = t_{a_i}$ for all $i = 1, \ldots, n$.

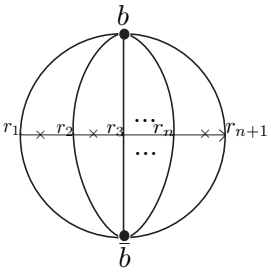


Figure 5. Slicing D^2 into small discs having one critical value.

Theorem 6.2. *If $g > 1$, then there is a one-to-one correspondence between the strong real Lefschetz chains $[c_1, a_1], [c_2, a_2], \ldots, [c_n, a_n]$ and the isomorphism classes of directed $\mathbb{C}$-marked genus g totally real Lefschetz fibrations over D^2.*

Proof. Necessity is clear. As for the converse, we consider the unique class (assured by Proposition 3.4) of directed $\mathbb{C}$-marked elementary real Lefschetz fibration associated to each real code $[c_i, a_i]$. We then glue these elementary fibrations (from left to right respecting the order determined by the chain) using the strong boundary fiber sum. The result, thus, follows from Proposition 5.3. □

Note that if we consider nonmarked fibrations, then the real codes around real singular fibers are defined up to conjugation. Thus, we are motivated to give the following definition and state the immediate corollary of Theorem 6.2.

Definition 6.3. A sequence $\{c_1, a_1\}, \{c_2, a_2\}, \ldots, \{c_n, a_n\}$ of conjugacy classes of real codes is called a *real Lefschetz chain* if $t_{a_i} \circ c_i$ is conjugate to c_{i+1} for all $1 \leq i \leq n$.

Corollary 6.4. *If $g > 1$, then there is a one-to-one correspondence between the real Lefschetz chains $\{c_1, a_1\}, \{c_2, a_2\}, \ldots, \{c_n, a_n\}$ and the isomorphism classes of nonmarked directed genus g totally real Lefschetz fibrations over D^2.*

If the total monodromy of the fibration $\pi : X \to D^2$ is the identity, then we can consider the extension of π to a fibration $\hat{\pi} : \hat{X} \to S^2$. Two such extensions $\hat{\pi} : \hat{X} \to S^2$ and $\check{\pi} : \check{X} \to S^2$ are considered *isomorphic* if there is an equivariant orientation-preserving diffeomorphism $H : \hat{X} \to \check{X}$ such that $\hat{\pi} = \check{\pi} \circ H$.

Proposition 6.5. *Let $\pi : X \to D^2$ be a $\mathbb{C}$-marked genus g totally real Lefschetz fibration whose total monodromy is the identity. If $g > 1$, then π can be extended uniquely up to isomorphism to a totally real Lefschetz fibration over S^2.*

Proof. Once again, the difference of two extensions corresponds to a loop in the space of real structures. Hence, the result follows from Proposition 4.3. □

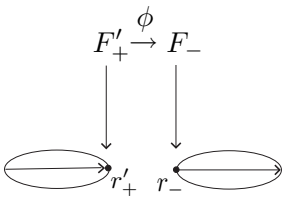
Corollary 6.6. *If $g > 1$, then there is a one-to-one correspondence between the strong real Lefschetz chains $[c_1, a_1], [c_2, a_2], \ldots, [c_n, a_n]$ such that $c_{n+1} \circ c_1 = (t_{a_n} \circ c_n) \circ c_1 = \mathrm{id}$ and the isomorphism classes of directed $\mathbb{C}$-marked genus g totally real Lefschetz fibrations over S^2.*

Remark 6.7. It is known that the components of the space of diffeomorphisms of the torus fixing a point are contractible [Earle and Eells 1969], so Theorem 6.2 can be adapted to $\mathbb{C}$-marked real elliptic Lefschetz fibration admitting a real section (a section compatible with the real structures). See [Salepci 2007, Section 5.4] for details. In the next section, we treat the case of nonmarked elliptic Lefschetz fibrations, which possibly do not admit a real section.

7. Boundary fiber sum of nonmarked real elliptic Lefschetz fibrations

To deal with the case of elliptic fibrations, we introduce the boundary fiber sum for nonmarked fibrations. (Although we concentrate on the case of $g(F) = 1$, the definition applies to any genus.)

Definition 7.1. Let $\pi' : X' \to D^2$ and $\pi : X \to D^2$ be two directed nonmarked RLFs. We consider the real fibers F'_+ and F_- of π' and π over the real points r'_+ and r_-, respectively. Let us assume that the real structure $c'_+ : F'_+ \to F'_+$ is conjugate to $c_- : F_- \to F_-$. That is, there is an orientation-preserving equivariant diffeomorphism $\phi : F'_+ \to F_-$. Then, the *boundary fiber sum* of $X' \natural_{F,\phi} X \to D^2$ is obtained by identifying the fibers F'_+ and F_- via ϕ, as below.



The boundary fiber sum does depend on the choice of ϕ in such a way that the two boundary fiber sums defined by the equivariant diffeomorphisms $\phi, \psi : F'_+ \to F_-$ are isomorphic, if $\psi \circ \phi^{-1} : F_- \to F_-$ can be extended to an equivariant diffeomorphism of $X \to D^2$ (or similarly if $\phi^{-1} \circ \psi : F'_+ \to F'_+$ can be extended to an equivariant diffeomorphism of $X' \to D^2$). The necessary and sufficient condition for $\psi \circ \phi^{-1} : F_- \to F_-$ to extend to an equivariant diffeomorphism of the fibration $X \to D^2$ is that $\psi \circ \phi^{-1}$ takes the unique vanishing cycle a of $X \to D^2$ to a curve equivariantly isotopic to a.

Now note that if $c(a) = a$, then c induces an action on a. Such an action can be the identity, a reflection, or an antipodal involution. It is not hard to show that if $c : \Sigma_1 \to \Sigma_1$ has one real component, then Σ_1 contains a unique c-equivariant isotopy class of noncontractible curves on which c acts as a reflection, a unique class of curves where the action of c is an antipodal involution, and a unique real curve; if c has 2 real components, then Σ_1 contains no c-equivariant isotopy class of curves on which c acts as an antipodal involution, a unique class of curves on which c acts as a reflection, and two classes of real curves (in which case, we call a pair of representatives of different classes *c-twin curves*); if c has no real components, then there exist two c-equivariant isotopy classes where c acts as an antipodal involution (as above, a pair of representatives of different classes are called *c-twin curves*) and no classes of other types. The boundary fiber sum is, therefore, well defined unless the real structure c has no real component or c has two real components one of which is the vanishing cycle a.

Recall that the rotation $R_{1/2}$ (introduced in the proof of Proposition 4.3) switches the c-twin curves. Hence, c-twin curves can be carried to each other via equivariant diffeomorphisms although they are not equivariantly isotopic, so in the case of existence of c-twin curves, there is an ambiguity in the definition of the boundary fiber sum $X' \natural X \to D^2$ (it can be defined in two ways). To resolve the ambiguity, we should specify how we identify the c'_+-twin curves on the fiber F'_+ in X' with the c_--twin curves on the fiber F_- in X. In a certain case, namely, if the real structure c'_+ has two real components and acts on the vanishing cycle a' as a reflection, the problem of switching c-twin curves can be eliminated via the transformation introduced below.

Let $\pi : X \to D^2$ be an elementary directed real elliptic Lefschetz fibration such that the real structure $c_+ : F_+ \to F_+$ acts on the vanishing cycle as a reflection. As a result, one of $c_\pm : F_\pm \to F_\pm$ has 1 real component while the other has 2 real components. Without loss of generality, we can assume that the real structure c_- has 1 real component. Our aim is to construct a transformation T_{sing} of X that does not change the isomorphism class of the fibration $\pi : X \to D^2$ and that is identity over $S_- \subset \partial D^2$ and interchanges the real components of F_+. To construct T_{sing}, we consider the following well known model for elementary elliptic fibrations.

Let $\hat{\Omega} = \{z \mid |\text{Re}(z)| \le \frac{1}{2}, \text{Im}(z) \ge 1\} \cup \infty$. This is the subset bounded by $\text{Im}(z) \ge 1$ of the one point compactification of the standard fundamental domain $\{z \mid |\text{Re}(z)| \le \frac{1}{2}, |z| \ge 1\}$ of the modular action on $\mathbb{C}$; see Figure 6.

We consider the real structure $c_{\hat{\Omega}} : \hat{\Omega} \to \hat{\Omega}$ such that $c_{\hat{\Omega}}(\omega) = \overline{-\omega}$. Let Ω denote the quotient $\hat{\Omega} / \frac{1}{2}+iy \sim -\frac{1}{2}+iy$. The real structure $c_{\hat{\Omega}}$ induces a real structure on Ω. Note that Ω is a topological real disc and can be identified with D^2 so that the real part of D^2 corresponds to the union of the half-lines iy and $\frac{1}{2} + iy$, where $y \ge 1$. For any $\omega \in \Omega$, the fiber over ω is given by $F_\omega = \mathbb{C}/(\mathbb{Z} + \omega\mathbb{Z})$, where the fiber F_∞ has the required nodal-type singularity.

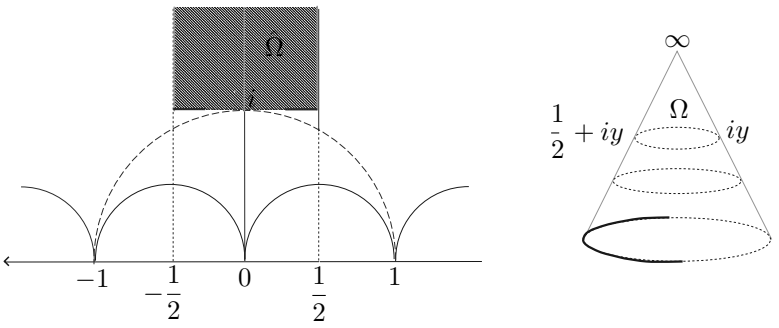


Figure 6. Moduli space of prescribed RELFs.

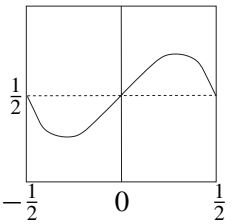


Figure 7. The graph of f.

Let $\pi_\Omega : X_\Omega \to \Omega$ denote the fibration such that $\pi_\Omega^{-1}(\omega) = F_\omega = \mathbb{C}/(\mathbb{Z} + \omega\mathbb{Z})$ for all $\omega \in \Omega$. Then, we consider the translation T'_Ω defined by

$$T'_\Omega : X_\Omega \to X_\Omega, \quad (z)_{\mathbb{Z}+\omega\mathbb{Z}} \in F_\omega \mapsto (z + \tau(w))_{\mathbb{Z}+\omega\mathbb{Z}} \in F_\omega,$$

where $(\cdot)_{\mathbb{Z}+\omega\mathbb{Z}}$ denotes the equivalence class in $\mathbb{C}/(\mathbb{Z} + \omega\mathbb{Z})$.

The map $\tau : \Omega \to \Omega$ is defined by

$$\tau(\omega) = -\tfrac{1}{2} + \left(\tfrac{1}{2} - f(\mathrm{Re}(\omega)) + i\right) \exp(-\mathrm{Im}(\omega) + 1),$$

where $f : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ is a smooth mapping whose graph is as shown in Figure 7 and that satisfies the following properties:

- $f(0) = \frac{1}{2}$ (modulo $\mathbb{Z}$),
- $f(1 - x) = 1 - f(x)$, (which implies $f(\frac{1}{2}) = \frac{1}{2}$) (modulo $\mathbb{Z}$),
- f is linear on $[\frac{1}{4}, \frac{3}{4}]$ (modulo $\mathbb{Z}$).

Note that τ has the following properties. (Equations are considered modulo the relation $-\frac{1}{2} + iy \sim \frac{1}{2} + iy$, with $y \geq 1$.)

- $\tau(\overline{-\omega}) = \overline{-\tau(\omega)}$.
- $\tau(\infty) = \frac{1}{2}$.
- $\tau(\frac{1}{2} + iy) = -\frac{1}{2} + i\exp(-y + 1) = \frac{1}{2} + i\exp(-y + 1)$; in particular, if $y = 1$, then $\tau(\frac{1}{2} + i) = \frac{1}{2} + i$.
- $\tau(iy) = -\frac{1}{2} + i\exp(-y + 1) = \frac{1}{2} + i\exp(-y + 1)$; in particular, if $y = 1$, then $\tau(i) = \frac{1}{2} + i$.

Let $T_{\mathrm{sing}} : X \to X$ denote the transformation induced from $T'_{\mathrm{sing}} : X_\Omega \to X_\Omega$. By definition T_{sing} is equivariant and the identity over $S_- \subset \partial D^2$, and its restriction to F_+ is the rotation $R_{1/2}$. (Figure 8 shows the action of T_{sing}.)

Lemma 7.2. *Let $\pi' : X' \to D^2$ and $\pi : X \to D^2$ be two nonmarked elementary* RELF*s such that both c'_+ and c_- have 2 real components. We assume that the vanishing cycle a of π is real with respect to c_-. Then, the boundary fiber sum $X' \natural_F X \to D^2$ is well defined if c'_+ acts on the vanishing cycle a' as a reflection.*

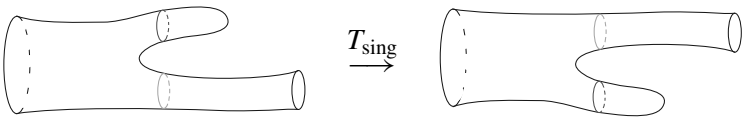


Figure 8. The action of T_{sing} on the real part (in gray).

Proof. The boundary fiber sums $X' \natural_{F,\phi} X \to D^2$ and $X' \natural_{F,\psi} X \to D^2$ are not isomorphic if $\phi \circ \psi^{-1}(a)$ and a are c-twin curves, but in the case when c'_+ acts on the vanishing cycle a' as a reflection, we can apply T_{sing} to X' so that $T_{\text{sing}}(F'_+)$ differs from the fiber F'_+ by the rotation $R_{1/2}$. Hence, $X' \natural_{F,\phi} X \to D^2$ is isomorphic to $T_{\text{sing}}(X') \natural_{F,\phi\circ R_1/2} X \to D^2$, which is isomorphic to $X' \natural_{F,\psi} X \to D^2$. □

8. Real Lefschetz chains associated to nonmarked real elliptic Lefschetz fibrations

We now consider a nonmarked directed totally real elliptic Lefschetz fibration $\pi : X \to D^2$, with $q_1 < q_2 < \cdots < q_n$. Around each critical value q_i we choose a small real disc D_i such that

$$D_i \cap \{q_1, q_2, \ldots, q_n\} = \{q_i\} \quad \text{and} \quad D_i \cap D_{i+1} = \{r_{i+1}\} \subset [q_i, q_{i+1}];$$

see Figure 9. Let c_i be the real structures on the fibers F_{r_i} for $1 \le i \le n$ (where r_1 is the left real point of ∂D^2) and a_i be the corresponding vanishing cycle.

By Corollary 3.5, each directed (nonmarked) fibration over D_i is classified by the conjugacy class $\{c_i, a_i\}$ of the real code. Thus, we can encode the fibration $\pi : X \to D^2$ by the *real Lefschetz chain* $\{c_1, a_1\}, \{c_2, a_2\}, \ldots, \{c_n, a_n\}$.

Clearly, real Lefschetz chains are invariants of directed nonmarked totally real elliptic Lefschetz fibrations over D^2, but they are not sufficient for classifying such fibrations. Additional information is needed, if for some i the real structure c_i has 2 real components and vanishing cycles corresponding to the critical values q_i and

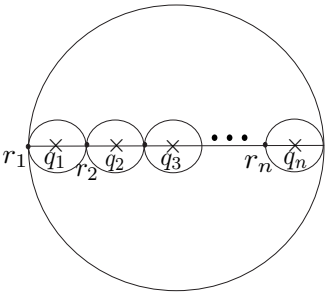


Figure 9. Subdividing D^2 into smaller discs.

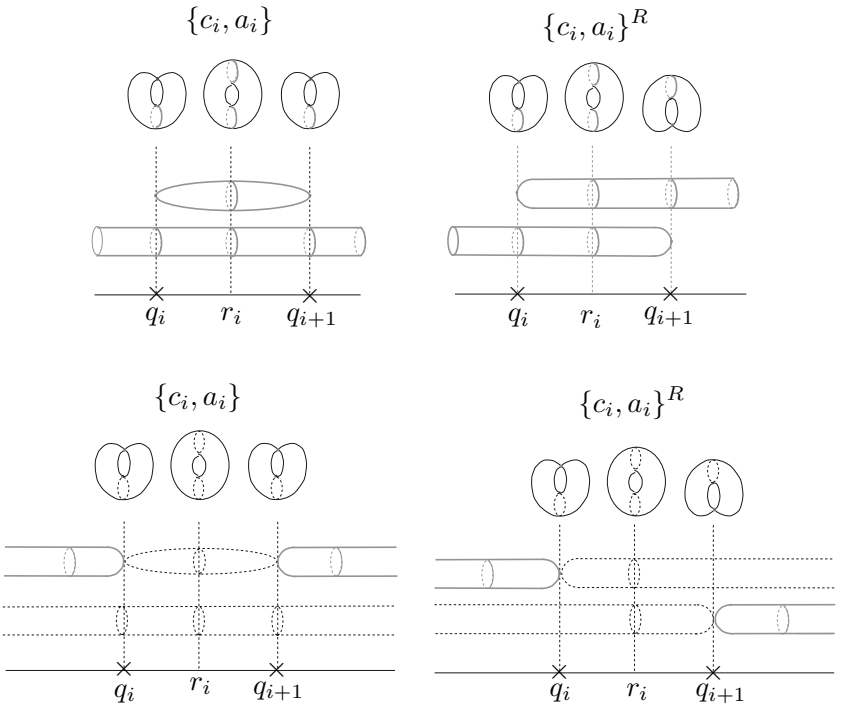


Figure 10. Real parts (in gray) of the fibrations associated to $\{c_i, a_i\}$ and $\{c_i, a_i\}^R$.

q_{i+1} are real or if c_i has no real component. Indeed, in these cases the vanishing cycles corresponding to the critical values q_i and q_{i+1} can be the same curve, or they can be c_i-twin curves. If they are c_i-twin curves, then we mark $\{c_i, a_i\}^R$ the corresponding real code $\{c_i, a_i\}$ by adding R (here R refers to the rotation $R_{1/2}$ that interchanges c-twin curves). The real Lefschetz chain we obtain is called the *decorated real Lefschetz chain*. Figure 10 shows all possible configurations of the real locus associated to $\{c_i, a_i\}$ and $\{c_i, a_i\}^R$.

Theorem 8.1. *There exists a one-to-one correspondence between the decorated real Lefschetz chains and the isomorphism classes of directed nonmarked totally real elliptic Lefschetz fibrations over D^2.*

Proof. Necessity is clear. As for the converse, we consider the unique class of the directed nonmarked elementary RELF (assured by Corollary 3.5) associated to each real code $\{c_i, a_i\}$. Then, we construct the required fibration by gluing elementary fibrations (from left to right) using the boundary fiber sum. As discussed above, the boundary fiber sum is uniquely defined in the case when the real structure on the fiber where the sum is performed has 1 real component or when it has 2 real components and acts on the vanishing cycle of the fibration glued to

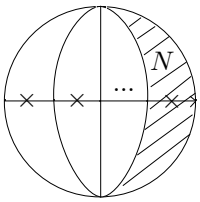


Figure 11. Neighborhood over which T_{sing} is applied.

right as a reflection. In the case when the real structure has 2 real components and acts as reflection on the vanishing cycle corresponding to the rightmost critical value of the already constructed $\pi' : X' \to D^2$, the two possible boundary fiber sums are isomorphic by Lemma 7.2 since in this case we can apply T_{sing} to X' (by considering T_{sing} on a neighborhood N of the last critical value, as shown in Figure 11, and extending it to X' as the identity outside of $\pi'^{-1}(N)$). In all the other cases, the boundary fiber sum is defined uniquely by the decoration. □

If c_1 is conjugate to c_{n+1}, then we can consider an extension of $\pi : X \to D^2$ to a fibration over S^2. As before, in the case when c_{n+1} has no real components or it has 2 real components and both a_1 and a_n are real, a decoration at infinity will be needed.

Proposition 8.2. *Let $\pi : X \to D^2$ be a totally real elliptic Lefschetz fibration associated to a decorated real Lefschetz chain. We assume that the real structures c_1 and c_{n+1} on the fibers over left and respectively right real point of ∂D^2 are conjugate. If c_{n+1} (and thus c_1) has 1 real component or if c_{n+1} (and thus c_1) has 2 real components and either c_{n+1} acts on the vanishing cycle a_n as a reflection, or c_1 acts on the vanishing cycle a_1 as a reflection, then π extends uniquely to a fibration over S^2. Otherwise, there are two extensions distinguished by the decoration at infinity.*

Proof. An extension of $\pi : X \to D^2$ to a fibration over S^2 defines a trivialization $\phi : \Sigma_1 \times S^1 \to \pi^{-1}(\partial D^2)$ over the boundary ∂D^2. Two trivializations ϕ and ϕ' correspond to isomorphic real fibrations if $\phi^{-1} \circ \phi' : \Sigma_1 \times S^1 \to \Sigma_1 \times S^1$ can be extended to an equivariant diffeomorphism of $\Sigma_1 \times D^2$ with respect to the real structure $(c_{n+1}, \text{conj}) : \Sigma_1 \times D^2 \to \Sigma_1 \times D^2$. Let $\Phi_t = (\phi^{-1} \circ \phi')_t : \Sigma_1 \to \Sigma_1$, $t \in S^1$. Since there is no fixed marking, up to change of marking we assume that $\Phi_t \in \text{Diff}_0(\Sigma_1)$.

The real structure splits the boundary into two symmetric pieces, so instead of considering an equivariant map over the entire boundary we consider a diffeomorphism over one the symmetric pieces. Let Φ_t for $t \in [0, 1]$ denote the family of such diffeomorphisms. This family defines a path in $\text{Diff}_0(\Sigma_1)$ whose end points lie in the group $\text{Diff}_0^{c_{n+1}}(\Sigma_1)$; therefore, Φ_t defines a relative loop in

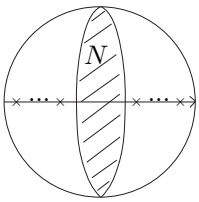


Figure 12. Neighborhood over which T is applied.

$\pi_1(\mathrm{Diff}_0(\Sigma_1), \mathrm{Diff}_0^{c_{n+1}}(\Sigma_1))$, and we are interested in the contractibility of this relative loop.

We consider the exact sequence

$$\cdots \to \pi_1(\mathrm{Diff}_0^{c_{n+1}}) \to \pi_1(\mathrm{Diff}_0) \xrightarrow{f} \pi_1(\mathrm{Diff}_0, \mathrm{Diff}_0^{c_{n+1}}) \xrightarrow{g} \pi_0(\mathrm{Diff}_0^{c_{n+1}}) \xrightarrow{h} \pi_0(\mathrm{Diff}_0) \to \pi_0(\mathrm{Diff}_0, \mathrm{Diff}_0^{c_{n+1}}) \to 0$$

of the pair $(\mathrm{Diff}_0(\Sigma_1), \mathrm{Diff}_0^{c_{n+1}}(\Sigma_1))$.

In the case when c_{n+1} has one real component, $\mathrm{Diff}_0^{c_{n+1}}(\Sigma_1)$ is connected, so the map h is injective; hence f is surjective. Therefore, we can see elements of the group $\pi_1(\mathrm{Diff}_0(\Sigma_1), \mathrm{Diff}_0^{c_{n+1}}(\Sigma_1), \mathrm{id})$ as being in $\pi_1(\mathrm{Diff}_0(\Sigma_1), \mathrm{id})$.

In all the other cases, $\mathrm{Diff}_0^{c_{n+1}}(\Sigma_1)$ has two components. We mark one of the components to make the map h injective when restricted to the marked component. Thus, g becomes the zero map, and so f is surjective over the marked component of $\mathrm{Diff}_0^{c_{n+1}}(\Sigma_1)$. Note that decoration of real Lefschetz chains distinguishes one of the component of $\mathrm{Diff}_0^{c_{n+1}}(\Sigma_1)$; hence, marking one component or other give the two extensions distinguished by the decoration. The distinctive feature of the case when c_{n+1} has 2 real components and acts a_n as a reflection (or c_1 acts on a_1 as a reflection) is that the transformation T_{sing} changes one marking to other, so the marking is not essential.

The proposition, thus, follows from Lemma 8.4 in which we show that any relative loop can be made contractible by means of some transformations T of the fibration $\pi : X \to D^2$. □

Let us first define the transformation T of real elliptic Lefschetz fibrations that is defined over a regular slice N of D^2.

Let $\pi : X \to D^2$ be a directed RELF. We consider a real slice N of D^2 that contains no critical value; see Figure 12.

Let $\xi : I \times I \to N$, where $I = [0, 1]$, be an orientation-preserving diffeomorphism such that first interval corresponds to the real direction on N. The fibration over N has no singular fiber; hence, it is trivializable. Let us consider a trivialization

$\Xi : \Sigma_1 \times I \times I \to \pi^{-1}(N)$ such that the following diagram commutes:

$$\begin{array}{ccc} \Sigma_1 \times I \times I & \xrightarrow{\Xi} & \pi^{-1}(N) \\ \downarrow & & \downarrow \pi \\ I \times I & \xrightarrow{\xi} & N. \end{array}$$

Since N has no critical value, the isotopy type of the real structure on the fibers over the real part of N remains fixed. If the real structure c has 2 real components, then we consider the model $\varrho : \mathbb{C}/\mathbb{Z}^2 \to \Sigma_1$ and set $\bar{\varrho} = (\varrho, \mathrm{id}) : \mathbb{C}/\mathbb{Z}^2 \times I \times I \to \Sigma_1 \times I \times I$ to define T as follows:

$$T' : \mathbb{C}/\mathbb{Z}^2 \times I \times I \to \mathbb{C}/\mathbb{Z}^2 \times I \times I, \quad ((x+iy)_{\mathbb{Z}^2}, t, s) \mapsto ((x+t+iy)_{\mathbb{Z}^2}, t, s).$$

Then, we set $T = \Xi \circ (\bar{\varrho} \circ T' \circ \bar{\varrho}^{-1}) \circ \Xi^{-1}$ on $\pi^{-1}(N)$. Since T is the identity at $t = 0, 1$, we can extend T to X by the identity outside of $\pi^{-1}(N)$.

If c has one real component, we construct T using $\varrho : \mathbb{C}/\Lambda \to \Sigma_1$. Similarly, if c has no real component, then we repeat the same using $\varrho : \mathbb{R}^2/\mathbb{Z}^2 \to \Sigma_1$.

Remark 8.3. First, since the transformation T is defined by a real rotation, T preserves the isomorphism class of the real Lefschetz fibration.

Second, the map T depends only on the isotopy type of $\pi^{-1}(N)$.

Lemma 8.4. *Let $\pi : X \to D^2$ be a totally real elliptic Lefschetz fibration. We assume that there exists at least one vanishing cycle on which corresponding real structure acts as a reflection. Then, there exists a generating set for*

$$\pi_1(\mathrm{Diff}_0(\Sigma_1), \mathrm{id}) = \mathbb{Z} + \mathbb{Z}$$

consisting of transformations $T_\pm$ for some nonsingular slices $N_\pm$.

Proof. Let q_i be the critical value such that the real structure on a nearby regular real fiber acts on the vanishing cycle as a reflection. This assumption assures that the neighboring real fibers have one real component on one side and two real components on the other side of the critical value q_i. Without loss of generality we can assume that the real structure over a fiber over a real point that lies on the left of q_i has two real components. (The other case can be treated similarly.)

We choose an auxiliary $\mathbb{C}$-marking $(\{b, \bar{b}\}, \{\rho : \Sigma_1 \to F_b, \bar{\rho} : \Sigma_1 \to F_{\bar{b}}\})$ and fix an identification $\varrho : S^1 \times S^1 \to \Sigma_1$. Since the real structure has 2 real components, we can assume that the induced real structure on $S^1 \times S^1$ is the reflection $(\alpha, \beta) \to (\alpha, -\beta)$. The real part consists of the curves $C_1 = (\alpha, 0)$ and $C_2 = (\alpha, \pi)$. Moreover, a representative of the vanishing cycle can be chosen as $(0, \beta)$. As $c_+ = t_{a_i} \circ c_-$ on $S^1 \times S^1$, the real part of c_+ is the curve C_3 given homologically by $2\alpha - \beta$; see Figure 13.

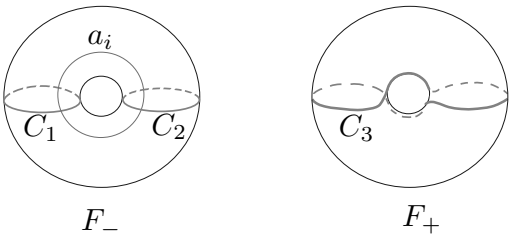


Figure 13. Real fibers (in gray) over the real points neighboring q_i.

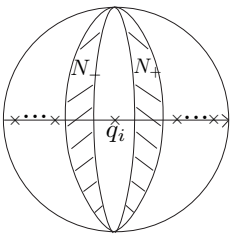


Figure 14. Regular slices $N_\pm$.

We now consider two nonsingular real slices N_- and N_+ of D^2 as shown in Figure 14. Let us suppose that the real fibers over N_- are identified to F_- while real fibers over N_+ are identified to F_+ (where $F_\pm$ are as shown Figure 13). Let C_3' and C_1' be curves on F_b obtained by pulling back $C_3 \subset F_+$ and $C_1 \subset F_-$, respectively. The curves C_3' and C_1' intersect at one point, so we can identify Σ_1 with $C_1' \times C_3'$ so that rotations along C_1' and C_3' generate the group $\mathrm{Diff}_0(\Sigma_1, \mathrm{id})$. Hence, $\{T_+, T_-\}$ generates $\pi_1(\mathrm{Diff}_0(\Sigma_1), \mathrm{id})$. □

Theorem 8.1 applies naturally to directed nonmarked RELFs over D^2 that admit a real section in which a real-case Lefschetz chain does not contain a real code (c_i, a_i) where the real structure has no real component. Besides, in the case when the real structure has 2 real components and the vanishing cycle is real, the decoration is not needed since the existence of a real section determines naturally the gluing. Moreover, the extension to a fibration over S^2 is uniquely defined by the section. Hence we have the following proposition.

Proposition 8.5. *Two directed totally real elliptic Lefschetz fibrations over S^2 admitting a real section and having the same real Lefschetz chain up to cyclic ordering are isomorphic.*

Remark 8.6. Indeed, the proposition holds even for fibrations with a fixed real section. If there are only real critical values, then the real sections are determined in a neighborhood of a real part. Moreover, over the real part one can carry one real section to another using the transformations T and *double* T_{sing}. Indeed, the

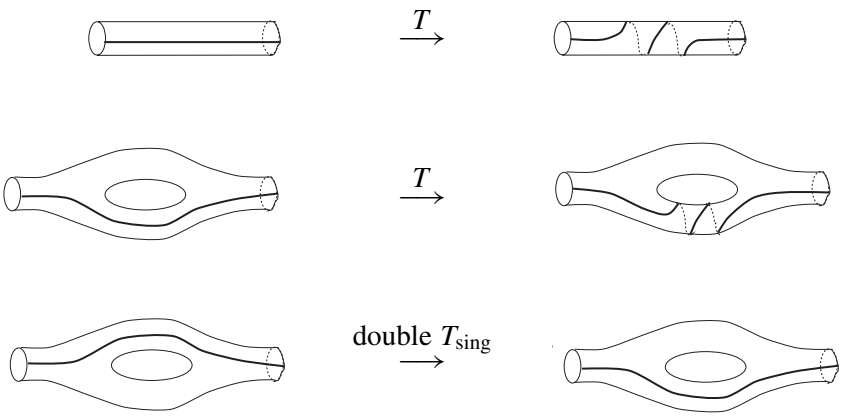


Figure 15. Modification of the real section over the real part.

double T_{sing} is defined for real Lefschetz fibrations with two critical values where the real structure extracted from the real fiber over a real point between the critical values acts on the vanishing cycles as a reflection. The model we use to define the double T_{sing} is as follows. Consider the disc D with two critical values as the double cover of a disc with one critical value branched at a regular real point. Let D_- and D_+ be two corresponding copies of the disk on the branched cover. By pulling back the fibration X_Ω over D, we obtain a model fibration over $D_- \cup D_+$. Thus, we can apply T_{sing} at the same time to fibrations over D_- and D_+. The possible modifications of the section is shown in the Figure 15.

Acknowledgments

This work is extracted from my thesis. I would like to express my gratitude to my supervisors Sergey Finashin and Viatcheslav Kharlamov for sharing their deep insight and knowledge.

References

[Degtyarev 2011] A. Degtyarev, "Hurwitz equivalence of braid monodromies and extremal elliptic surfaces", *Proc. Lond. Math. Soc.* **103**:6 (2011), 1083–1120. MR 2861751 Zbl 1234.14030

[Degtyarev and Salepci 2011] A. Degtyarev and N. Salepci, "Products of pairs of Dehn twists and maximal real Lefschetz fibrations", preprint 2011 - 32, Mathematisches Forschungsinstitut Oberwolfach, 2011, available at http://www.mfo.de/scientific-programme/publications/owp/2011/OWP2011_32.pdf. To appear in *Nagoya Math. J.*

[Earle and Eells 1969] C. J. Earle and J. Eells, "A fibre bundle description of Teichmüller theory", *J. Differential Geom.* **3**:1–2 (1969), 19–43. MR 43 #2737a Zbl 0185.32901

[Earle and Schatz 1970] C. J. Earle and A. Schatz, "Teichmüller theory for surfaces with boundary", *J. Differential Geom.* **4**:2 (1970), 169–185. MR 43 #2737b Zbl 0194.52802

[Hamstrom 1965] M.-E. Hamstrom, "Homotopy properties of the space of homeomorphisms on P^2 and the Klein bottle", *Trans. Amer. Math. Soc.* **120**:1 (1965), 37–45. MR 32 #1707 Zbl 0148.17201

[Ivanov 2001] N. V. Ivanov, "Mapping class groups", pp. 523–633 in *Handbook of geometric topology*, edited by R. J. Daverman and R. B. Sher, Elsevier, Amsterdam, 2001. MR 2003h:57022 Zbl 1002.57001

[Salepci 2007] N. Salepci, *Real Lefschetz fibrations*, thesis, Université Louis Pasteur, Strasbourg, 2007, available at http://tinyurl.com/Salepci-thesis. MR 2780321 Zbl 1216.55004

[Salepci 2010] N. Salepci, "Real elements in the mapping class group of T^2", *Topology Appl.* **157**:16 (2010), 2580–2590. MR 2011g:57022 Zbl 1213.57024

[Salepci 2012] N. Salepci, "Classification of totally real elliptic Lefschetz fibrations via necklace diagrams", *J. Knot Theory Ramifications* **21**:9 (2012), 1250089.

Received April 21, 2011.

NERMİN SALEPCİ
INSTITUT CAMILLE JORDAN
UNIVERSITÉ LYON I
43, BOULEVARD DU 11 NOVEMBRE 1918
69622 VILLEURBANNE CEDEX
FRANCE

salepci@math.univ-lyon1.fr
http://math.univ-lyon1.fr/∼salepci/

PACIFIC JOURNAL OF MATHEMATICS
Vol. 256, No. 2, 2012

STABLE TRACE FORMULAS AND DISCRETE SERIES MULTIPLICITIES

STEVEN SPALLONE

Let G be a reductive algebraic group over $\mathbb{Q}$, and suppose that $\Gamma \subset G(\mathbb{R})$ is an arithmetic subgroup defined by congruence conditions. A basic problem in arithmetic is to determine the multiplicities of discrete series representations in $L^2(\Gamma\backslash G(\mathbb{R}))$, and in general to determine the traces of Hecke operators on these spaces. In this paper we give a conjectural formula for the traces of Hecke operators, in terms of stable distributions. It is based on a stable version of Arthur's formula for L^2-Lefschetz numbers, which is due to Kottwitz. We reduce this formula to the computation of elliptic p-adic orbital integrals and the theory of endoscopic transfer. As evidence for this conjecture, we demonstrate the agreement of the central terms of this formula with the unipotent contributions to the multiplicity coming from Selberg's trace formula of Wakatsuki, in the case $G = \mathrm{GSp}_4$ and $\Gamma = \mathrm{GSp}_4(\mathbb{Z})$.

1. Introduction

Let G be a reductive algebraic group over $\mathbb{Q}$, and Γ an arithmetic subgroup of $G(\mathbb{R})$ defined by congruence conditions. Then $G(\mathbb{R})$ acts on $L^2(\Gamma\backslash G(\mathbb{R}))$ via right translation; let us write R for this representation. A fundamental problem in arithmetic is to understand R. As a first step, we may decompose R as

$$R = R_{\mathrm{disc}} \oplus R_{\mathrm{cont}},$$

where R_{disc} is a direct sum of irreducible representations, and R_{cont} decomposes continuously. The continuous part may be understood inductively through Levi subgroups of G as in [Langlands 1976], leaving us with the study of R_{disc}. Given an irreducible representation π of $G(\mathbb{R})$, write $R_{\mathrm{disc}}(\pi)$ for the π-isotypic subspace of R_{disc}. Then

$$R_{\mathrm{disc}}(\pi) \cong \pi^{\oplus m_{\mathrm{disc}}(\pi)}$$

MSC2010: 11F46, 11F72, 22E55, 32N10.

Keywords: discrete series, Hecke operators, orbital integrals, Shimura varieties, endoscopy, fundamental lemma, stable trace formula.

for some integer $m_{\text{disc}}(\pi)$. (We may also write $m_{\text{disc}}(\pi, \Gamma)$.) A basic problem is to compute these integers.

There is more structure than simply these dimensions, however. Arithmetic provides us with a multitude of Hecke operators h on $L^2(\Gamma\backslash G(\mathbb{R}))$ that commute with R. Write $R_{\text{disc}}(\pi, h)$ for the restriction of h to $R_{\text{disc}}(\pi)$. The general problem is to find a formula for the trace of $R_{\text{disc}}(\pi, h)$.

We focus on discrete series representations π. These are representations that behave like representations of compact or finite groups, in the sense that their associated matrix coefficients are square integrable. Like other smooth representations, they have a theory of characters developed by Harish-Chandra. They separate naturally into finite sets called L-packets. For an irreducible finite-dimensional algebraic representation E of $G(\mathbb{C})$, there is a corresponding L-packet Π_E of discrete series representations, consisting of those with the same infinitesimal and central characters as E.

We follow the tradition of computing tr $R_{\text{disc}}(\pi, h)$ through trace formulas. This method has gone through several incarnations, beginning with Selberg [1956] for GL_2, in which he also investigated the continuous Eisenstein series. A goal was to compute dimensions of spaces of modular forms, and traces of Hecke operators on these spaces. These spaces of modular forms correspond to the spaces $R_{\text{disc}}(\pi)$ we are discussing in this case. His trace formula is an integral, over the quotient of the upper half space X by Γ, of a sum of functions H_γ, one for each element of Γ. Let us write it roughly as

$$\dim_{\mathbb{C}} S(\Gamma) = \int_{\Gamma\backslash X} \sum_{\gamma\in\Gamma} H_\gamma(Z)\,dZ,$$

for some space $S(\Gamma)$ of cusp forms with a suitable Γ-invariance condition.

Here dZ is a $G(\mathbb{R})$-invariant measure on X. When the quotient $\Gamma\backslash X$ is compact, the sum and integral may be interchanged, leading to a simple expression for the dimensions in terms of orbital integrals. The interference of the Eisenstein series precludes this approach in the noncompact quotient case. Here there are several convergence difficulties, which Selberg overcomes by employing a truncation process. Unfortunately the truncation process leads to notoriously complicated expressions, which are far from being in closed form. This study of $R_{\text{disc}}(\pi)$ has been expanded to other reductive groups using what is called the Arthur–Selberg trace formula. See [Arthur 2005].

Generally, a trace formula is an equality of distributions on $G(\mathbb{R})$, or on the adelic group $G(\mathbb{A})$. One distribution is called the geometric side; it is a sum of terms corresponding to conjugacy classes of G. Given a test function f, the formula is essentially made up of combinations $I_M(\gamma, f)$ of weighted integrals of f over the conjugacy classes of elements γ. (Here M is a Levi subgroup of G.)

The other distribution is called the spectral side, involving the Harish-Chandra transforms $\operatorname{tr}\pi(f)$ for various representations π. Here, the operator $\pi(f)$ is given by weighting the representation π by f. The geometric and spectral sides agree, and in applications we can learn much about the latter from the former. Some of the art is in picking test functions to extract information about both sides.

The best general result using the trace formula to study $\operatorname{tr} R_{\mathrm{disc}}(\pi, h)$ seems to be Arthur's [1989]. He produces a formula for

$$\sum_{\pi\in\Pi} \operatorname{tr} R_{\mathrm{disc}}(\pi, h), \tag{1-1}$$

where Π is a given discrete series L-packet for $G(\mathbb{R})$. He uses test functions f which he calls "stable cuspidal". Their Fourier transforms $\pi \mapsto \operatorname{tr}\pi(f)$ are "stable" in that they are constant on L-packets, and "cuspidal" in that, considered as a function defined on tempered representations, they are supported on discrete series. (Tempered representations are those that appear in the Plancherel formula for $G(\mathbb{R})$.) Using his invariant trace formula, Arthur [1988a; 1988b] obtains (1-1) as the spectral side. The geometric side is a combination of orbital integrals for h and values of Arthur's Φ-function, which describes the asymptotic values of discrete series characters averaged over an L-packet.

In particular, he produces a formula for

$$\sum_{\pi\in\Pi} m_{\mathrm{disc}}(\pi), \tag{1-2}$$

for an L-packet Π of (suitably regular) discrete series representations.

In the case of $G = \mathrm{GL}_2$, there is a discrete series representation π_k for each integer $k \geq 1$. In this case $m_{\mathrm{disc}}(\pi_k)$ is the dimension of the space $S_k(\Gamma)$ of Γ-cusp forms of weight k on the upper half plane. Restriction to $\mathrm{SL}_2(\mathbb{R})$ gives two discrete series $\{\pi_k^+, \pi_k^-\}$ in each L-packet. However we may still use Arthur's formula here since $m_{\mathrm{disc}}(\pi_k^+, \Gamma) = m_{\mathrm{disc}}(\pi_k^-, \Gamma)$ for every arithmetic subgroup Γ. (Endoscopy does not play a role.)

For the group $\mathrm{GSp}_4(\mathbb{R})$ there are two discrete series representations in each L-packet: one "holomorphic" and one "large" discrete series. Let π be a holomorphic discrete series, and write π' for the large discrete series representation in the same L-packet as π. The multiplicity $m_{\mathrm{disc}}(\pi, \Gamma)$ is also the dimension of a certain space of vector-valued Siegel cusp forms (see [Wallach 1984]) on the Siegel upper half space, an analogue of the usual cusp forms on the upper half plane. For $\Gamma = \mathrm{Sp}_4(\mathbb{Z})$, the dimensions of these spaces of cusp forms were calculated by Tsushima [1983; 1997] by using the Riemann–Roch–Hirzebruch formula, and later by Wakatsuki [2012] by using the Selberg trace formula and the properties of prehomogeneous

vector spaces. In [$\geq$ 2012], Wakatsuki then evaluated Arthur's formula to compute $m_{\text{disc}}(\pi, \Gamma) + m_{\text{disc}}(\pi', \Gamma)$, thereby deducing a formula for $m_{\text{disc}}(\pi', \Gamma)$.

A natural approach to isolating the individual $m_{\text{disc}}(\pi)$, or generally the individual $\operatorname{tr} R_{\text{disc}}(\pi, h)$, is to apply a trace formula to a matrix coefficient, or more properly, a pseudocoefficient f. This means that f is a test function whose Fourier transform picks out π rather than the entire packet Π containing π; see Definition 6 below. Such a function will not be stable cuspidal, but merely cuspidal. Arthur [1989] (see also [2005]) showed that $I_M(\gamma, f)$ vanishes when f is stable cuspidal and the unipotent part of γ is nontrivial. If we examine the geometric side of Arthur's formula for a pseudocoefficient f, we must evaluate the more complicated terms $I_M(\gamma, f)$ for elements γ with nontrivial unipotent part. At the time of this writing, such calculations have not been made in general; we take another approach.

Distinguishing the individual representations π from others in its L-packet leads to the theory of endoscopy, and stable trace formulas. The grouping of representations π into packets Π on the spectral side mirrors the fusion of conjugacy classes that occurs when one extends the group $G(\mathbb{R})$ to the larger group $G(\mathbb{C})$. If F is a local or global field, then a stable conjugacy class in $G(F)$ is, roughly, the union of classes which become conjugate in $G(\overline{F})$. (See [Langlands 1979] for a precise definition.)

The distribution that takes a test function to its integral over a regular semisimple stable conjugacy class is a basic example of a stable distribution. Indeed, a stable distribution is defined to be a closure of the span of such distributions; see [Langlands 1983; 1979]. A distribution on $G(F)$ is stabilized if it can be written as a sum of stable distributions, the sum being over smaller subgroups H related to G. These groups H are called endoscopic groups for G; they are tethered to G not as subgroups but through their Langlands dual groups. As part of a series of techniques called endoscopy, one writes unstable distributions on G as combinations of stable distributions on the groups H. Part of this process is the theory of transfer, associating suitable test functions f^H on $H(F)$ to test functions f on $G(F)$ that yield a matching of orbital integrals. Indeed this was the drive for [Ngô 2010]. As the name suggests, the theory of endoscopy, while laborious, leads to an intimate understanding of G.

There has been much work in stabilizing Arthur's formula. See for example [Langlands 1983; Arthur 2002; 2001; 2003]. In Kottwitz's preprint [$\geq$ 2012], he defines a stable version of Arthur's Lefschetz formula, which we review below. (See also [Morel 2010].) It is a combination $\mathcal{K}(f) = \sum_H \iota(G, H)\, ST_g(f^H)$ of distributions $f \mapsto f^H \mapsto ST_g(f^H)$ over endoscopic groups H for G. Here the distributions ST_g, defined for each H, are stable. (See Section 5.1 for the definition of the rational numbers $\iota(G, H)$.) Each ST_g is a sum of terms corresponding to

stable conjugacy classes of elliptic elements $\gamma \in H(\mathbb{Q})$. Kottwitz's main result is that $\mathcal{K}$ agrees with Arthur's distribution, at least for functions f that are stable cuspidal at the real place.

As part of the author's thesis [Spallone 2004], the identity terms of $\mathcal{K}$ were evaluated for the group $G = \mathrm{SO}_5$ at a function f that was a pseudocoefficient for a discrete series representation at the real place. Later, Wakatsuki noted that the resulting expressions match up with the terms in his multiplicity formulas for $m_{\mathrm{disc}}(\pi, \Gamma)$ and $m_{\mathrm{disc}}(\pi', \Gamma)$ corresponding to unipotent elements. Moreover, the contribution in [Spallone 2004] from the endoscopic group accounted for the difference in these multiplicity formulas, while the stable part corresponded to the sum. After further investigation, we conjecture simply that Kottwitz's distribution evaluated at a function $f = f_{\pi,\Gamma}$ suitably adapted to π and Γ is equal to $m_{\mathrm{disc}}(\pi, \Gamma)$, under a regularity condition on π. (See Section 5.3 for the precise statement.) Of course this is compatible with Arthur's results in [1989].

In this paper we give some computational evidence for this conjecture. We also reduce the computation of each $ST(f^H_{\pi,\Gamma})$ to evaluating elliptic orbital p-adic integrals for the transfer $f^{\infty H}$ at the finite places. The rest breaks naturally into a problem at the real points and a global volume computation.

The main ingredient at the archimedean place is the Φ-function $\Phi_M(\gamma, \Theta^E)$ of Arthur, which we review. This quantity gives the contribution from the real place to the trace formulas in [Arthur 1989] and [Goresky et al. 1997]. It also plays a prominent role in Kottwitz's formula. This function, originally defined by the asymptotic behavior of a stable character near a singular element γ, was expressed in closed form in many cases by the author in [Spallone 2009].

There are two volume-related constants that enter into any explicit computation of ST_g. The first is $\bar{v}(G)$, which is essentially the volume of an inner form of G over $\mathbb{R}$. It depends on the choice of local measure dg_∞. The second comes about from orbital integrals at the finite adeles, and depends on the choice of local measure dg_f. These integrals may frequently be written in terms of the volumes of open compact subgroups K_f of $G(\mathbb{A}_f)$. In practice, one is left computing expressions such as $\bar{v}(G)^{-1} \operatorname{vol}_{dg_f}(K_f)^{-1}$, which are independent of the choice of local measures. More specifically, we define

$$\chi_{K_f}(G) = \bar{v}(G)^{-1} \operatorname{vol}_{dg_f}(K_f)^{-1} \tau(G) d(G).$$

Here $\tau(G)$ is the Tamagawa number of G and $d(G)$ is the index of the real Weyl group in the complex Weyl group. A main general result of this paper, Theorem 2, interprets $\chi_{K_f}(G)$ via Euler characteristics of arithmetic subgroups. It extends a computation of Harder [1971], which was for semisimple simply connected groups, to the case of reductive groups, under some mild hypotheses on G.

We work out two examples in this paper, one for SL_2 and another for GSp_4. It

is easy to verify our conjecture for $G = \mathrm{SL}_2$ and $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ using the classic dimension formula for cusp forms. In this case endoscopy does not appear. The calculations for GSp_4 are more complex; we content ourselves with working out the central terms of Kottwitz's formula.

If π is a holomorphic discrete series representation of $\mathrm{GSp}_4(\mathbb{R})$, write H_1^π for the central-unipotent terms of the Selberg trace formula, as evaluated in [Wakatsuki $\geq$ 2012] to compute $m_{\mathrm{disc}}(\pi, \Gamma)$. Here $\Gamma = \mathrm{GSp}_4(\mathbb{Z})$. If π is a large discrete series representation, write H_1^π for the central-unipotent terms in [Wakatsuki $\geq$ 2012] contributing to $m_{\mathrm{disc}}(\pi, \Gamma)$. In both cases, write $f = f_{\pi,\Gamma} = f_\infty f^\infty$, with f_∞ a pseudocoefficient for π, and f^∞ the (normalized) characteristic function of the integer adelic points of G. Write $\mathcal{H}(f, \pm 1)$ for the central terms of Kottwitz's formula applied to f.

As evidence for our conjecture, we show this:

Theorem 1. *For each regular discrete series representation π of $G(\mathbb{R})$, we have*

$$\mathcal{H}(f_{\pi,\Gamma}, \pm 1) = H_1^\pi.$$

We believe that the $\mathcal{H}(f_{\pi,\Gamma}, \pm 1)$ terms will generally match up with the difficult central-unipotent terms of the Arthur–Selberg formula, as in this case.

Our conjecture reduces the computation of discrete series multiplicities to the computation of stable elliptic orbital integrals of various transfers f_p^H, written for functions on $G(\mathbb{Q}_p)$. Let us write this as $SO_{\gamma_H}(f_p^H)$. Here f_p are characteristic functions of congruence subgroups of $G(\mathbb{Q}_p)$ related to Γ. Certainly at suitably regular elements, $SO_{\gamma_H}(f_p^H)$ is an unstable combination of orbital integrals of f_p; however there are also contributions from elliptic singular γ_H, notably $\gamma_H = 1$. At present, there are expressions for f_p^H in the parahoric case and of course for $G(\mathbb{Z}_p)$, but less seems to be known for smaller congruence subgroups. On the other hand, there are many formulas for dimensions of Siegel cusp forms and discrete series multiplicities for these cases (for example, [Wakatsuki $\geq$ 2012]). This suggests that one could predict stable singular elliptic orbital integrals $SO_{\gamma_H}(f_p^H)$ for the transfer f_p^H of characteristic functions of congruence subgroups (see for example Klingen, Iwahori and Siegel), by comparing our formulas.

Finally, we refer the casual reader to our survey [Spallone 2011] of the present approach to discrete series multiplicities.

In Section 2, we set up the conventions for this study. We explain how we are setting up the orbital integrals, and indicate our main computational tools. We also review the Langlands correspondence for real groups.

The theory of Arthur's Φ-function is reviewed in Section 4. In Section 5, we review Kottwitz's stable version of Arthur's formula from [Kottwitz $\geq$ 2012]. We also state our conjecture here. The heart of the volume computations in this paper

is in Section 6, where we determine $\chi_K(G)$. As a warm up, we work out the classic case of SL_2, with $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ in Section 7.

The case of $G = \mathrm{GSp}_4$ is considerably more difficult. We must work out several isomorphisms of real tori. These are described in Section 8. The basic structure of G and its Langlands dual $\hat{G}$ is set up in Section 9. In Section 10 we work out the Langlands parameters for discrete series of $G(\mathbb{R})$. There is only one elliptic endoscopic group H for G. We describe H in Section 11. In Section 12, we describe the Langlands parameters for discrete series of $H(\mathbb{R})$ and describe the transfer of discrete series in this case. In Section 13, we describe the Levi subgroups of G and H and compute various constants that occur in Kottwitz's formula for these groups. In Section 14, we compute explicitly Arthur's Φ-function for Levi subgroups of G, and we do this for Levi subgroups of H in Section 15. In Section 16, we write out the terms of Kottwitz's formula corresponding to central elements of G and H, for a general arithmetic subgroup Γ. In Section 17, we specialize to the case of $\Gamma = \mathrm{GSp}_4(\mathbb{Z})$, and in Section 18 we gather our results to demonstrate Theorem 1.

2. Preliminaries and notation

If F is a field, write Γ_F for the absolute Galois group of F. Suppose G is an algebraic group over F. If E is an extension field of F, we write G_E for G viewed as an algebraic group over E (by restriction). If γ is an element of $G(F)$, we denote by G_γ the centralizer of γ in G. By G° we denote the identity component of G (with the Zariski topology). Write G_{der} for the derived group of G. If G is a reductive group, write G_{sc} for the simply connected cover of G_{der}. Let $X^*(G) = \mathrm{Hom}(G_{\bar{F}}, \mathbb{G}_m)$ and $X_*(G) = \mathrm{Hom}(\mathbb{G}_m, G_{\bar{F}})$. These are abelian groups. Write $X^*(G)_{\mathbb{C}}$ and $X_*(G)_{\mathbb{C}}$ for the tensor product of these groups over $\mathbb{Z}$ with $\mathbb{C}$. Similarly with the subscript $\mathbb{R}$. Write A_G for the maximal F-split torus in the center of G.

We denote by $\mathbb{A}$ the ring of adeles over $\mathbb{Q}$. We denote by $\mathbb{A}_f$ the ring of finite adeles over $\mathbb{Q}$, so that $\mathbb{A} = \mathbb{A}_f \times \mathbb{R}$. Write $\mathbb{O}_f$ for the integral points of $\mathbb{A}_f$.

If G is a real Lie group, we write G^+ for the connected component of G (using the classical topology rather than any Zariski topology).

Let G be a connected reductive group over $\mathbb{R}$. A torus T in G is elliptic if T/A_G is anisotropic (as an $\mathbb{R}$-torus). Say that G is cuspidal if it contains a maximal torus T that is elliptic. An element of $G(\mathbb{R})$ is elliptic if it is contained in an elliptic maximal torus of G. Having fixed an elliptic maximal torus T, the absolute Weyl group Ω_G of T in G is the quotient of the normalizer of $T(\mathbb{C})$ in $G(\mathbb{C})$ by $T(\mathbb{C})$. The real Weyl group $\Omega_{G,\mathbb{R}}$ of T in G is the quotient of the normalizer of $T(\mathbb{R})$ in $G(\mathbb{R})$ by $T(\mathbb{R})$. We may drop the subscript G if it is clear from context. Also fix a maximal compact subgroup $K_{\mathbb{R}}$ of $G(\mathbb{R})$.

Write $q(G)$ for half the dimension of $G(\mathbb{R})/K_{\mathbb{R}}Z(\mathbb{R})$. If we write R for the roots of G, with a set of positive roots R^+, then

$$q(G) = \tfrac{1}{2}(|R^+| + \dim(X)),$$

where X is the span of R.

If G is an algebraic group over $\mathbb{Q}$, let $G(\mathbb{Q})^+ = G(\mathbb{R})^+ \cap G(\mathbb{Q})$.

2.1. ***Endoscopy.*** Here we review the theory of based root data and endoscopy in the form we will use in this paper.

The notion of a based root datum is defined in [Springer 1979]. First, a root datum is a quadruple $\Psi = (X, R, X^\vee, R^\vee)$, where

- X and $X^\vee$ are free, finitely generated abelian groups, in duality by a pairing
$$\langle \cdot, \cdot \rangle : X \times X^\vee \to \mathbb{Z};$$
- R and $R^\vee$ are finite subsets of X and $X^\vee$, respectively;
- there is a bijection $\alpha \mapsto \alpha^\vee$ from R onto $R^\vee$;
- we have $\langle \alpha, \alpha^\vee \rangle = 2$ for all $\alpha \in R$;
- $s_\alpha(R) = R$ if s_α is the reflection of X determined by α, and similarly with α replaced by $\alpha^\vee$ and R by $R^\vee$.

A based root datum is a quadruple $\Psi_0 = (X, \Delta, X^\vee, \Delta^\vee)$, where Δ and $\Delta^\vee$ are sets of simple roots of root system R and $R^\vee$ respectively, so that $(X, R, X^\vee, R^\vee)$ is a root datum. The dual of $\Psi_0 = (X, \Delta, X^\vee, \Delta^\vee)$ is given simply by $\Psi_0^\vee = (X^\vee, \Delta^\vee, X, \Delta)$.

Let $\Psi_0 = (X, \Delta, X^\vee, \Delta^\vee)$ and $\Psi_0' = (X', \Delta', X'^\vee, \Delta'^\vee)$ be two based root data. Then an isomorphism between Ψ and Ψ' is an isomorphism of groups $f : X \to X'$ so that f induces a bijection of Δ onto Δ' and so that the transpose of f induces a bijection of $\Delta^\vee$ onto $\Delta'^\vee$.

Let G be a connected reductive group over an algebraically closed field F. Fix a maximal torus T and a Borel subgroup B of G with $T \subseteq B$. We say in this situation that (T, B) is a pair (for G). The choice of pair determines a based root datum

$$\Psi_0(G, T, B) = (X^*(T), \Delta(T, B), X_*(T), \Delta^\vee(T, B))$$

for G. Here $\Delta(T, B)$ is the set of simple B-positive roots of T, and $\Delta^\vee(T, B)$ is the set of simple B-positive coroots of T. If another pair $T' \subseteq B'$ is chosen, the new based root datum obtained is canonically isomorphic to the original via an inner automorphism α of G. We have $\alpha(T') = T$ and $\alpha(B') = B$. Although the inner automorphism α need not be unique, its restriction to an isomorphism $T' \xrightarrow{\sim} T$ is unique.

We may remove the dependence of the based root datum on the choice of pair as follows. Write X^*, Δ, X_*, and $\Delta^\vee$ for the inverse limit over the set of pairs (T, B) of $X^*(T)$, $\Delta(T, B)$, $X_*(T)$ and $\Delta^\vee(T, B)$, respectively. Then we simply define the based root datum of G to be

$$\Psi_0(G) = (X^*, \Delta, X_*, \Delta^\vee).$$

Let G be a connected reductive group over a field F, and $\Psi_0(G)$ a based root datum of $G_{\bar{F}}$. Then Γ_F acts naturally (via isomorphisms) on $\Psi_0(G)$. The action of Γ_F on G is said to be an L-action if it fixes some splitting of G; see [Kottwitz 1984, Section 1.3].

Definition 1. A dual group for G is the following data:

(i) A connected complex reductive group with a based root datum $\Psi_0(\hat{G})$. We write its complex points as $\hat{G}$.

(ii) An L-action of Γ_F on $\hat{G}$.

(iii) A Γ_F-isomorphism from $\Psi_0(\hat{G})$ to the dual of $\Psi_0(G)$.

To specify the isomorphism for (iii) above, one typically fixes pairs (T_0, B_0) of G and $(\hat{S}_0, \hat{B}_0)$ of a dual group $\hat{G}$ and an isomorphism from $\Psi_0(\hat{G}, \hat{S}_0, \hat{B}_0)$ to the dual of $\Psi_0(G, T_0, B_0)$.

In the case that G is a torus T, the dual group $\hat{T}$ is simply given by

$$\hat{T} = X^*(T) \otimes_{\mathbb{Z}} \mathbb{C}^\times, \tag{2-1}$$

with the Γ_F-action induced from $X^*(T)$. There are canonical Γ_F-isomorphisms $X^*(\hat{T}) \xrightarrow{\sim} X_*(T)$ and $X_*(\hat{T}) \xrightarrow{\sim} X^*(T)$.

The formalism for dual groups encodes canonical isomorphisms between tori. If T and T' are tori, and $\varphi : T \to T'$ is a homomorphism, it induces a homomorphism $\hat{T}' \to \hat{T}$ in the evident way.

Suppose that (T, B) is a pair for G and $(\hat{S}, \hat{B})$ is a pair for $\hat{G}$. By (iii) above, one has in particular a fixed isomorphism from $\Psi_0(G, T, B)$ to the dual of $\Psi_0(\hat{G}, \hat{S}, \hat{B})$. In particular this yields an isomorphism from $X^*(T)$ to $X_*(\hat{S})$, which induces an isomorphism

$$\hat{T} \xrightarrow{\sim} \hat{S}. \tag{2-2}$$

Next, let G be a connected reductive group over a field F, which is either local or global.

Definition 2. An endoscopic group for G is a triple (H, s, η) as follows:

- H is a quasisplit connected group, with a fixed dual group $\hat{H}$ as above;
- $s \in Z(\hat{H})$.

- $\eta : \hat{H} \to \hat{G}$ is an embedding.
- The image of η is $(\hat{G})^{\circ}_{\eta(s)}$, the connected component of the centralizer in $\hat{G}$ of $\eta(s)$.
- The $\hat{G}$-conjugacy class of η is fixed by Γ_F.
  Cohomology of Γ_F-modules then yields a boundary map

$$[Z(\hat{H})/Z(\hat{G})]^{\Gamma_F} \to H^1(F, Z(\hat{G})).$$

- The image of s in $Z(\hat{H})/Z(\hat{G})$ is fixed by Γ, and its image under the boundary map above is trivial if F is local and locally trivial if F is global.

An endoscopic group is elliptic if the identity components of $Z(\hat{G})^{\Gamma_F}$ and $Z(\hat{H})^{\Gamma_F}$ agree.

Isomorphism of endoscopic groups is defined in [Kottwitz 1984, Section 7.5]; we do not review it here.

2.2. ***Langlands correspondence.*** Let G be a connected reductive group over $\mathbb{R}$. In this section we review elliptic Langlands parameters for G and the corresponding L-packets for discrete series representations of $G(\mathbb{R})$. Our main references are [Borel 1979] and [Kottwitz 1990]. Write $W_{\mathbb{R}}$ for the Weil group of $\mathbb{R}$, and $W_{\mathbb{C}}$ for the canonical image of $\mathbb{C}^{\times}$ in $W_{\mathbb{R}}$. There is an exact sequence

$$1 \to W_{\mathbb{C}} \to W_{\mathbb{R}} \to \Gamma_{\mathbb{R}} \to 1.$$

The Weil group $W_{\mathbb{R}}$ is generated by $W_{\mathbb{C}}$ and a fixed element τ satisfying $\tau^2 = -1$ and $\tau z \tau^{-1} = \bar{z}$ for $z \in W_{\mathbb{C}}$. The action of $\Gamma_{\mathbb{R}}$ on $\hat{G}$ inflates to an action of $W_{\mathbb{R}}$ on $\hat{G}$, and through this action we form the L-group ${}^L G = \hat{G} \rtimes W_{\mathbb{R}}$.

A Langlands parameter φ for G is an equivalence class of continuous homomorphisms $\varphi : W_{\mathbb{R}} \to {}^L G$ commuting with projection to $\Gamma_{\mathbb{R}}$, satisfying a mild hypothesis on the image; see [Borel 1979]. The equivalence relation is via inner automorphisms from $\hat{G}$. One associates to a Langlands parameter φ an L-packet $\Pi(\varphi)$ of irreducible admissible representations of G.

Suppose that G is cuspidal, so that there is a discrete series representation of $G(\mathbb{R})$. This implies that the longest element w_0 of the Weyl group Ω acts as -1 on $X_*(T)$. If φ is a Langlands parameter, write C_φ for the centralizer of $\varphi(W_{\mathbb{R}})$ in $\hat{G}$ and $\hat{S}$ for the centralizer of $\varphi(W_{\mathbb{C}})$ in $\hat{G}$. Write S_φ for the product $C_\varphi Z(\hat{G})$. We say φ is elliptic if $S_\varphi / Z(\hat{G})$ is finite, and describe the L-packet $\Pi(\varphi)$ in this case.

Since φ is elliptic, the centralizer $\hat{S}$ is a maximal torus in $\hat{G}$. Since φ commutes with the projection to $\Gamma_{\mathbb{R}}$, it restricts to a homomorphism

$$W_{\mathbb{C}} \to \hat{S} \times \{1\}.$$

We may view this restriction as a continuous homomorphism $\varphi : \mathbb{C}^\times \to \hat{S}$, which may be written in exponential form

$$\varphi(z) = z^\mu \bar{z}^\nu$$

with μ and ν regular elements of $X_*(\hat{T})_\mathbb{C}$. Write $\hat{B}$ for the unique Borel subgroup of $\hat{G}$ containing $\hat{S}$ so that $\langle \mu, \alpha \rangle$ is positive for every root α of $\hat{S}$ that is positive for $\hat{B}$. We say that φ determines the pair $(\hat{S}, \hat{B})$, at least up to conjugacy in $\hat{G}$.

Let B be a Borel subgroup of $G_\mathbb{C}$ containing T. Then φ and B determine a quasicharacter $\chi_B = \chi(\varphi, B)$, as follows. There is a canonical (up to $\hat{G}$-conjugacy) homomorphism $\eta_B : {}^L T \to {}^L G$ described in [Kottwitz 1990] such that

$$\eta_B(z) = z^\rho \bar{z}^{-\rho} \times z \in \hat{G} \rtimes W_\mathbb{R} \quad \text{for } z \in W_\mathbb{C}.$$

Here $\rho = \rho_G$ is the half sum of the B-positive roots for T. Then a Langlands parameter φ_B for T may be chosen so that $\varphi = \eta_B \circ \varphi_B$. Finally χ_B is the quasicharacter associated to φ_B by the Langlands correspondence for T (as described in [Borel 1979, Section 9.4]).

Write $\mathcal{B}$ for the set of Borels of $G_\mathbb{C}$ containing T. The L-packet associated to φ is indexed by $\Omega_\mathbb{R} \backslash \mathcal{B}$. For $B \in \Omega_\mathbb{R} \backslash \mathcal{B}$, a representation $\pi(\varphi, B)$ in the L-packet is given by the irreducible discrete series representation of $G(\mathbb{R})$ whose character Θ_π is given on regular elements γ of $T(\mathbb{R})$ by

$$(-1)^{q(G)} \sum_{\omega \in \Omega_\mathbb{R}} \chi_{\omega(B)}(\gamma) \cdot \Delta_{\omega(B)}(\gamma)^{-1}.$$

Here Δ_B is the usual discriminant

$$\Delta_B(\gamma) = \prod_{\alpha > 0 \text{ for } B} (1 - \alpha(\gamma)^{-1}).$$

Finally, let

$$\Pi(\varphi) = \{\pi(\varphi, B) \mid B \in \Omega_\mathbb{R} \backslash \mathcal{B}\}.$$

It has order $d(G) = |\Omega / \Omega_\mathbb{R}|$. There is a unique irreducible finite-dimensional algebraic complex representation E of $G(\mathbb{C})$ with the same infinitesimal character and central character as the representations in this L-packet. It has highest weight $\mu - \rho \in X^*(T)$ with respect to B. The isomorphism classes of such E are in one-to-one correspondence with elliptic Langlands parameters φ, and we often write Π_E for $\Pi(\varphi)$.

Definition 3. We say that a discrete series representation $\pi \in \Pi_E$ is regular if the highest weight of E is regular.

2.3. ***Measures and orbital integrals.*** Let G be a locally compact group with Haar measure dg. If f is a continuous function on G, write fdg for the measure on G given by

$$\varphi \mapsto \int_G \varphi(g) f(g) dg,$$

for φ continuous and compactly supported in G. We will refer to the measures obtained in this way simply as "measures". If G is a p-adic, real, or adelic Lie group, we require that f be suitably smooth.

In this paper, we will view orbital integrals and Fourier transforms as distributions defined on measures, rather than on functions. This approach eases their dependence on choices of local measures, choices that do not matter in the end.

For K an open compact subset of G, write e_K for the measure given by fdg, where f is the characteristic function of K divided by $\mathrm{vol}_{dg}(K)$. Note that the measure e_K is independent of the choice of Haar measure dg.

Let G be a reductive group defined over a local field F. Fix a Haar measure dg on $G(F)$. Let fdg be a measure on $G(F)$, and take a semisimple element $\gamma \in G(F)$. Fix a Haar measure dt of $G(F)^\circ_\gamma$. Then we write $O_\gamma(fdg; dt)$ for the usual orbital integral

$$O_\gamma(fdg; dt) = \int_{G_\gamma{}^\circ(F)\backslash G(F)} f(g^{-1}\gamma g)\frac{dg}{dt}.$$

Many cases of finite orbital integrals are easy to compute by the following result, a special case extracted from [Kottwitz 1986, Section 7].

Proposition 1. *Let F be a p-adic field with ring of integers $\mathbb{O}$. Let G be a split connected reductive group defined over $\mathbb{O}$, and let $K = G(\mathbb{O})$. Suppose that $\gamma \in K$ is semisimple, and that $1 - \alpha(\gamma)$ is either 0 or a unit for every root α of G. Let γ' be stably conjugate to γ. Then $O_{\gamma'}(e_K; dt)$ vanishes unless γ' is conjugate to γ, in which case*

$$O_{\gamma'}(e_K; dt) = \mathrm{vol}_{dt}(G_\gamma{}^\circ(F) \cap K)^{-1}.$$

Now let G be a reductive group defined over $\mathbb{Q}$.

Let $f^\infty dg_f$ be a measure on $G(\mathbb{A}_f)$ and take a semisimple element $\gamma \in G(\mathbb{A}_f)$. Fix a Haar measure dt_f of $G^\circ_\gamma(\mathbb{A}_f)$. Write $O_\gamma(f^\infty dg_f; dt_f)$ for the orbital integral

$$O_\gamma(f^\infty dg_f; dt_f) = \int_{G_\gamma{}^\circ(\mathbb{A}_f)\backslash G(\mathbb{A}_f)} f^\infty(g^{-1}\gamma g)\frac{dg_f}{dt_f}.$$

We also have the stable orbital integrals

$$SO_\gamma(f^\infty dg_f; dt_f) = \sum_i e(\gamma_i) O_{\gamma_i}(f^\infty dg_f; dt_{i,f}),$$

the sum being over $\gamma_i \in G(\mathbb{A}_f)$ (up to $G(\mathbb{A}_f)$-conjugacy) whose local components are stably conjugate to γ. The centralizers of γ and a given γ_i are inner forms of each other, and we use corresponding measures dt_f and $dt_{i,f}$. The number $e(\gamma_i)$ is defined as follows: For a reductive group A over a local field, Kottwitz [1983] has defined an invariant $e(A)$. It is equal to 1 if A is quasisplit. For each place v of $\mathbb{Q}$, write $\gamma_{i,v}$ for the vth component of γ_i. Let

$$e(\gamma_{i,v}) = e(G^\circ_{\gamma_{i,v}}(\mathbb{Q}_v)).$$

Finally, let

$$e(\gamma_i) = \prod_v e(\gamma_{i,v}).$$

Definition 4. Let M be a Levi component of a parabolic subgroup P of G, and dm_f a Haar measure on $M(\mathbb{A}_f)$. Given a measure $f^\infty dg_f$, its *M-constant term* is the measure $f_M^\infty dm_f$, where f_M^∞ is defined via

$$f_M^\infty(m) = \delta_{P(\mathbb{A}_f)}^{-1/2}(m) \int_{N(\mathbb{A}_f)} \int_{K_f} f^\infty(k^{-1}nmk) dk_f dn_f.$$

Here we fix the Haar measure dk_f on K_f giving it mass one, and the Haar measure dn_f on $N(\mathbb{A}_f)$ is chosen so that $dg_f = dk_f dn_f dm_f$. The function $\delta_{P(\mathbb{A}_f)}$ is the modulus function on $P(\mathbb{A}_f)$.

It is independent of the choice of parabolic subgroup P.

Proposition 2. *Let G be a split group defined over $\mathbb{Z}$ and let $K_f = G(\mathbb{O}_f)$. Then*

$$(e_{K_f})_M = e_{M(\mathbb{A}_f) \cap K_f}.$$

Proof. Write $e_{K_f} = f^\infty dg_f$. Then it is easy to see that $f_M^\infty(m) = 0$ unless $m \in K_f$. If $m \in K_f$, we compute that

$$f_M^\infty(m) = \frac{\mathrm{vol}_{dk_f}(K_f)\, \mathrm{vol}_{dn_f}(K_f \cap N(\mathbb{A}_f))}{\mathrm{vol}_{dg_f}(K_f)}.$$

The result follows since

$$\mathrm{vol}_{dg_f}(K_f) = \mathrm{vol}_{dm_f}(M(\mathbb{A}_f) \cap K_f)\, \mathrm{vol}_{dn_f}(N(\mathbb{A}_f) \cap K_f)\, \mathrm{vol}_{dk_f}(K_f). \qquad \square$$

2.4. ***Pseudocoefficients.*** We continue with a connected reductive group G over $\mathbb{Q}$, and adopt some terminology from [Arthur 1989]. Fix a maximal compact subgroup $K_\mathbb{R}$ of $G(\mathbb{R})$. We put $K'_\mathbb{R} = K_\mathbb{R} A_G(\mathbb{R})^+$. Given a quasicharacter (smooth homomorphism to $\mathbb{C}^\times$) ξ on $A_G(\mathbb{R})^+$, write $\mathscr{H}_{\mathrm{ac}}(G(\mathbb{R}), \xi)$ for the space of smooth, $K'_\mathbb{R}$-finite functions on $G(\mathbb{R})$ that are compactly supported modulo $A_G(\mathbb{R})^+$, and

transform under $A_G(\mathbb{R})^+$ according to ξ. Write $\Pi(G(\mathbb{R}), \xi)$ for the set of irreducible representations of $G(\mathbb{R})$ whose central character restricted to $A_G(\mathbb{R})^+$ is equal to ξ.

Given a function $f \in \mathscr{H}_{\mathrm{ac}}(G(\mathbb{R}), \xi^{-1})$, a representation $\pi \in \Pi(G(\mathbb{R}), \xi)$, and a Haar measure dg_∞ on $G(\mathbb{R})$, write $\pi(f dg_\infty)$ for the operator on the space of π given by the formula

$$\pi(f dg_\infty) = \int_{G(\mathbb{R})/A_G(\mathbb{R})^+} f(x)\pi(x) dg_\infty.$$

Here we give $A_G(\mathbb{R})^+$ the measure corresponding to Lebesgue measure on $\mathbb{R}^n$, if A_G is n-dimensional. The operator is of trace class.

Write $\Pi_{\mathrm{temp}}(G(\mathbb{R}), \xi)$ (respectively $\Pi_{\mathrm{disc}}(G(\mathbb{R}), \xi)$) for the subset of tempered (respectively discrete series) representations in $\Pi(G(\mathbb{R}), \xi)$.

Definition 5. Suppose that $f \in \mathscr{H}_{\mathrm{ac}}(G(\mathbb{R}), \xi^{-1})$. We say that the measure $f dg_\infty$ is cuspidal if $\operatorname{tr} \pi(f dg_\infty)$, viewed as a function on $\Pi_{\mathrm{temp}}(G(\mathbb{R}), \xi)$, is supported on $\Pi_{\mathrm{disc}}(G(\mathbb{R}), \xi)$.

Write $\tilde{E}$ for the contragredient of the representation E. Arthur [1989] employs functions $f_E \in \mathscr{H}_{\mathrm{ac}}(G(\mathbb{R}), \xi^{-1})$ with $f_E dg_\infty$ cuspidal, whose defining property is that, for all $\pi \in \Pi_{\mathrm{temp}}(G(\mathbb{R}), \xi)$,

$$\operatorname{tr} \pi(f_E dg_\infty) = \begin{cases} (-1)^{q(G)} & \text{if } \pi \in \Pi_{\tilde{E}}, \\ 0 & \text{otherwise.} \end{cases} \tag{2-3}$$

Such measures can be broken down further.

Definition 6. Fix a representation $\pi_0 \in \Pi_{\mathrm{disc}}(G(\mathbb{R}), \xi^{-1})$, and suppose that $f_0 \in \mathscr{H}_{\mathrm{ac}}(G(\mathbb{R}), \xi^{-1})$. Suppose the measure $f_0 dg_\infty$ satisfies, for all $\pi \in \Pi_{\mathrm{temp}}(G(\mathbb{R}), \xi)$,

$$\operatorname{tr} \pi(f_0 dg_\infty) = \begin{cases} (-1)^{q(G)} & \text{if } \pi \cong \tilde{\pi}_0, \\ 0 & \text{otherwise.} \end{cases}$$

It follows from the corollary in [Clozel and Delorme 1984, Section 5.2] that such functions exist. Pick such a function f_0, and put $e_{\pi_0} = f_0 dg_\infty$.

Suppose that for each $\pi \in \Pi_E$ we fix measures e_π as above. Let

$$f_E dg_\infty = \sum_\pi e_\pi,$$

the sum being over $\pi \in \Pi_E$. Then clearly $f_E dg_\infty$ satisfies Arthur's condition (2-3).

We remark that the measure $(-1)^{q(G)} e_\pi$ is called a pseudocoefficient of $\tilde{\pi}$.

3. Transfer

We sketch the important theory of transfer in the form that we will use in this paper.

Suppose that G is a real connected reductive group, and that (H, s, η) is an elliptic endoscopic group for G. Fix an elliptic maximal torus T_H of H, an elliptic maximal torus T of G, and an isomorphism $j : T_H \xrightarrow{\sim} T$ between them. Also fix a Borel subgroup B of $G_{\mathbb{C}}$ containing T and a Borel subgroup B_H of $H_{\mathbb{C}}$ containing T_H.

Suppose that ξ is a quasicharacter on $A_G(\mathbb{R})$, and that $f_\infty \in \mathcal{H}_{\mathrm{ac}}(G(\mathbb{R}), \xi^{-1})$, with $f_\infty dg_\infty$ cuspidal. There is a corresponding quasicharacter ξ_H on $A_H(\mathbb{R})$ described in [Kottwitz $\geq$ 2012, Section 5.5].

There is also a measure $f_\infty^H dh_\infty$ on $H(\mathbb{R})$ with $f_\infty^H \in \mathcal{H}_{\mathrm{ac}}(H(\mathbb{R}), \xi_H^{-1})$, having matching character values. See [Shelstad 1982; Clozel and Delorme 1984; 1990; Langlands and Shelstad 1987]. More specifically, let φ_H be a tempered Langlands parameter for $H_{\mathbb{R}}$, and write $\Pi_H = \Pi(\varphi_H)$ for the corresponding L-packet of discrete series representations of $H(\mathbb{R})$. Transport φ_H via η to a tempered Langlands parameter φ_G for G. The parameters φ_G and φ_H determine pairs $(\hat{S}, \hat{B})$ and $(\hat{S}_H, \hat{B}_H)$ as in Section 2.2.

Then

$$\operatorname{tr} \Pi_H(f_\infty^H dh_\infty) = \sum_{\pi \in \Pi} \Delta_\infty(\varphi_H, \pi) \cdot \operatorname{tr} \pi(f_\infty dg_\infty), \tag{3-1}$$

using Shelstad's transfer factors $\Delta_\infty(\varphi_H, \pi)$. Both sides of (3-1) vanish unless Π_H is a discrete series packet. In particular, $f_\infty^H dh_\infty$ is cuspidal, and it may be characterized by (3-1). (The transfer $f_\infty^H dh_\infty$ is only defined up to the kernel of stable distributions.) We may use this formula to identify it as a combination of pseudocoefficients.

It is a delicate matter to specify the transfer factors. We will use a formula for $\Delta_\infty(\varphi_H, \pi)$ from [Kottwitz 1990], which is itself a reformulates a formula from [Shelstad 1982]. One must carefully specify the duality between G and $\hat{G}$, and between H and $\hat{H}$, because this factor depends on precisely how this is done. It also depends on the isomorphism $j : T_H \xrightarrow{\sim} T$, which must be compatible with correspondences of tori determined by the Langlands parameters, as specified below.

Definition 7. The triple (j, B_T, B_{T_H}) is aligned with φ_H if the following diagram commutes:

$$\begin{array}{ccc} \hat{T} & \longrightarrow & \hat{S} \\ {\scriptstyle \hat{j}}\downarrow & & \uparrow{\scriptstyle \eta} \\ \hat{T}_H & \longrightarrow & \hat{S}_H. \end{array} \tag{3-2}$$

Here the isomorphisms $\hat{T} \to \hat{S}$ and $\hat{T}_H \to \hat{S}_H$ are determined, as in (2-2), by $(B, \hat{B})$ and $(B_H, \hat{B}_H)$, respectively. The map $\hat{j}$ is the map dual to j using the identification (2-1) of the dual tori.

For each $\omega \in \Omega$, there is a character

$$a_\omega : (\hat{T}/Z(\hat{G}))^{\Gamma_{\mathbb{R}}} \to \{\pm 1\}$$

described in [Kottwitz 1990].

If the triple (j, B_T, B_{T_H}) is aligned with φ_H, then we may take as transfer factors

$$\Delta_\infty(\varphi_H, \pi(\varphi, \omega^{-1}(B))) = \langle a_\omega, \hat{j}^{-1}(s)\rangle.$$

Next, let G be a connected reductive algebraic group over $\mathbb{Q}$, and let (H, s, η) be an endoscopic group for G. Given a measure $f^\infty dg_f$ on $G(\mathbb{A}_f)$, there is a measure $f^{\infty H} dh_f$ on $H(\mathbb{A}_f)$ such that for all $\gamma_H \in H(\mathbb{A}_f)$ suitably regular, one has

$$SO_{\gamma_H}(f^{\infty H} dh_f) = \sum_\gamma \Delta^\infty(\gamma_H, \gamma) O_\gamma(f^\infty dg_f).$$

The sum is taken over $G(\mathbb{A}_f)$-conjugacy classes of "images" $\gamma \in G(\mathbb{A}_f)$ of γ_H. We have written $\Delta^\infty(\gamma_H, \gamma)$ for the Langlands–Shelstad transfer factors. One takes matching measures on the centralizers of γ_H and the various γ in forming the quotient measures for the orbital integrals. We have left out many details; please see [Langlands and Shelstad 1987] and [Kottwitz and Shelstad 1999] for definitions, and [Ngô 2010] for the celebrated proof.

4. Arthur's Φ-function

In this section we consider a reductive group G defined over $\mathbb{R}$. Let T be a maximal torus contained in a Borel subgroup B of $G_{\mathbb{C}}$. Let A be the split part of T, let T_c be the maximal compact subtorus of T, and let M be the centralizer of A in G. It is a Levi subgroup of G containing T. Let E be an irreducible finite-dimensional (algebraic) representation of $G(\mathbb{C})$, and consider the L-packet Π_E of discrete series representations π of $G(\mathbb{R})$ that have the same infinitesimal and central characters as E. Write Θ_π for the character of π, and put

$$\Theta^E = (-1)^{q(G)} \sum_{\pi \in \Pi_E} \Theta_\pi.$$

Note that $\Theta^E(\gamma)$ will not extend continuously to all elements $\gamma \in T(\mathbb{R})$, and in particular not to $\gamma = 1$. Define the function D_M^G on T by

$$D_M^G(\gamma) = \det(1 - \mathrm{Ad}(\gamma); \mathrm{Lie}(G)/\mathrm{Lie}(M)).$$

Then a result of Arthur and Shelstad [Arthur 1989] states that the function

$$\gamma \mapsto |D_M^G(\gamma)|^{1/2}\Theta^E(\gamma),$$

defined on the set of regular elements $T_{\mathrm{reg}}(\mathbb{R})$, extends continuously to $T(\mathbb{R})$. We denote this extension by $\Phi_M(\gamma, \Theta^E)$. The following closed expression for $\Phi_M(\gamma, \Theta^E)$ when $\gamma \in T_c$ is given in [Spallone 2009].

Proposition 3. *If $\gamma \in T_c(\mathbb{R})$, then*

$$\Phi_M(\gamma, \Theta^E) = (-1)^{q(L)}|\Omega_L| \sum_{\omega \in \Omega^{LM}} \varepsilon(\omega)\,\mathrm{tr}(\gamma; V^M_{\omega(\lambda_B+\rho_B)-\rho_B}). \tag{4-1}$$

In particular,

(i) *if T is compact, then $M = G$ and $\Phi_G(\gamma, \Theta^E) = \mathrm{tr}(\gamma; E)$;*

(ii) *if T is split, then $M = A$ and $\Phi_A(1, \Theta^E) = (-1)^{q(G)}|\Omega_G|$.*

The notation needs to be explained. Here L is the centralizer of T_c in G. The roots of T in L and M are the real and imaginary roots, respectively, of T in G. Write Ω_L and Ω_M for the respective Weyl groups. Write Ω^{LM} for the set of elements that are simultaneously Kostant representatives for both L and M, relative to B. We write ε for the sign character of Ω_G. Finally by $V^M_{\omega(\lambda_B+\rho_B)-\rho_B}$ we denote the irreducible finite-dimensional representation of $M(\mathbb{C})$ with highest weight $\omega(\lambda_B + \rho_B) - \rho_B$, where λ_B is the B-dominant highest weight of E.

If $z \in G(\mathbb{R})$ is central, it is easy to see that $\Phi_M(\gamma z, \Theta^E) = \lambda_E(z)\Phi_M(\gamma, \Theta^E)$, where λ_E is the central character of E. Thus, for the case of central $\gamma = z$, computing $\Phi_M(z, \Theta^E)$ amounts to computing the dimensions of finite-dimensional representations of $M(\mathbb{C})$ with various highest weights. For this we use the Weyl dimension formula, in the following form.

Proposition 4 (Weyl dimension formula). *Let G be a complex reductive group and T a maximal torus in G, contained in a Borel subgroup B. Write ρ_B for the half-sum of the positive roots for T in G (with respect to B). Let $\lambda_B \in X^*(T)$ be a positive weight. Then there is a unique irreducible representation V_{λ_B} of G with highest weight λ_B. Its dimension is given by*

$$\dim_{\mathbb{C}} V_{\lambda_B} = \prod_{\alpha>0} \frac{\langle \alpha, \lambda_B + \rho_B\rangle}{\langle \alpha, \rho_B\rangle}.$$

Here $\langle \cdot, \cdot \rangle$ is a nondegenerate Ω_G-invariant inner product on $X^*(T)_{\mathbb{R}}$, which is unique up to a scalar.

5. Kottwitz's formula

5.1. ***Various invariants.*** In this section we introduce some invariants involved in Kottwitz's formula.

By $\overline{G}$ we generally denote an inner form of $G_{\mathbb{R}}$ such that $\overline{G}/A_G$ is anisotropic over $\mathbb{R}$.

Definition 8. Let G be a cuspidal reductive group over $\mathbb{R}$, and dg_∞ a Haar measure on $G(\mathbb{R})$. Let

$$\bar{v}(G; dg_\infty) = e(\overline{G}) \operatorname{vol}(\overline{G}(\mathbb{R})/A_G(\mathbb{R})^+).$$

This is a stable version of the constant $v(G)$ that appears in [Arthur 1989]. As before, $e(\overline{G})$ is the sign defined in [Kottwitz 1983]. (Note that $e(\overline{G}) = (-1)^{q(G)}$ when G is quasisplit.) In both cases the Haar measure on $\overline{G}(\mathbb{R})$ is transported from dg_∞ on $G(\mathbb{R})$ in the usual way, and the measure on $A_G(\mathbb{R})^+$ is the standard Lebesgue measure.

Definition 9. Let G be a cuspidal connected reductive group over $\mathbb{Q}$. Then G contains a maximal torus T such that T/A_G is anisotropic over $\mathbb{R}$. Write T_{sc} for the inverse image in G_{sc} of T. Then $k(G)$ is the cardinality of the image of $H^1(\mathbb{R}, T_{\mathrm{sc}}) \to H^1(\mathbb{R}, T)$.

Definition 10. If G is a reductive group over $\mathbb{Q}$, write $\tau(G)$ for the Tamagawa number of G, as defined in [Ono 1966].

By [Kottwitz 1988] or [Kottwitz $\geq$ 2012], the Tamagawa numbers $\tau(G)$ for a reductive group G over $\mathbb{Q}$ may be computed using the formula

$$\tau(G) = |\pi_0(Z(\hat{G})^{\Gamma_{\mathbb{Q}}})| \cdot |\ker^1(\mathbb{Q}, Z(\hat{G}))|^{-1}.$$

Here π_0 denotes the topological connected component.

Definition 11. Let M be a Levi subgroup of G. Then put

$$n_M^G = [N_G(M)(\mathbb{Q}) : M(\mathbb{Q})].$$

Here $N_G(M)$ denotes the normalizer of M in G.

Definition 12. Let $\gamma \in M(\mathbb{Q})$ be semisimple. Then put

$$\bar{\iota}^M(\gamma) = |(M_\gamma/M_\gamma^\circ)(\mathbb{Q})| \quad \text{and} \quad \iota^M(\gamma) = [M_\gamma(\mathbb{Q}) : M_\gamma^\circ(\mathbb{Q})].$$

Let (H, s, η) be an endoscopic triple for G, and write $\mathrm{Out}(H, s, \eta)$ for its outer automorphisms. Put

$$\iota(G, H) = \tau(G)\tau(H)^{-1}|\mathrm{Out}(H, s, \eta)|^{-1}.$$

5.2. ***The formula.*** In this section we give Kottwitz's formula [$\geq$ 2012].

Our G will now be a cuspidal connected reductive group over $\mathbb{Q}$. Let $f^\infty \in C_c^\infty(G(\mathbb{A}_f))$ and $f_\infty \in \mathcal{H}_{\mathrm{ac}}(G(\mathbb{R}), \xi)$ for some ξ. We consider measures $f\,dg$ of the form $f\,dg = f^\infty dg_f \cdot f_\infty dg_\infty \in C_c^\infty(G(\mathbb{A}))$, for some decomposition $dg = dg_f dg_\infty$ of the Tamagawa measure on $G(\mathbb{A}_f)$. Also choose such decompositions for every cuspidal Levi subgroup M of G.

First we define the stable distribution $S\Phi_M$ at the archimedean place:

Definition 13. Let M be a cuspidal Levi subgroup of G. Let $\gamma \in M(\mathbb{Q})$ be elliptic, and pick a Haar measure dt_∞ of $M_\gamma^\circ(\mathbb{R})$. Then $S\Phi_M(\gamma, f_\infty dg_\infty; dt_\infty)$ is defined to be

$$(-1)^{\dim(A_M/A_G)} k(M) k(G)^{-1} \bar{v}(M_\gamma{}^\circ; dt_\infty)^{-1} \sum_\Pi \Phi_M(\gamma^{-1}, \Theta_\Pi) \operatorname{tr} \Pi(f_\infty dg_\infty),$$

the sum being taken over L-packets of discrete series representations.

Here is the basic building block of Kottwitz's formula:

Definition 14. Let M be a cuspidal Levi subgroup of G, and $\gamma \in M(\mathbb{Q})$ an elliptic element. Pick Haar measures dt_f on $M_\gamma^\circ(\mathbb{A}_f)$ and dt_∞ on $M_\gamma^\circ(\mathbb{R})$ whose product is the Tamagawa measure dt on $M_\gamma^\circ(\mathbb{A})$.

We define

$$ST_g(f\,dg, \gamma, M) = (n_M^G)^{-1} \tau(M) \bar{\iota}^M(\gamma)^{-1}\, SO_\gamma(f_M^\infty dm_f; dt_f)\, S\Phi_M(\gamma, f_\infty dg_\infty; dt_\infty).$$

Here $f_M^\infty dm_f$ is the M-constant term of $f^\infty dg_f$. The product

$$SO_\gamma(f_M^\infty dm_f; dt_f) \bar{v}(M; dt_\infty)$$

is independent of the decompositions of dt and dg. We will therefore often write this simply as $SO_\gamma(f_M^\infty dm_f)\bar{v}(M)$, and similarly for other such products.

Kottwitz defines

$$ST_g(f\,dg) = \sum_M \sum_{\gamma \in M} ST_g(f\,dg, \gamma, M).$$

Here M runs over $G(\mathbb{Q})$-conjugacy classes of cuspidal Levi subgroups in G, and the second sum runs over stable $M(\mathbb{Q})$-conjugacy classes of semisimple elements $\gamma \in M(\mathbb{Q})$ that are elliptic in $M(\mathbb{R})$.

For convenience we also define, for $\gamma \in G(\mathbb{Q})$ semisimple,

$$ST_g(f\,dg, \gamma) = \sum_M ST_g(f\,dg, \gamma, M),$$

the sum being taken over cuspidal Levi subgroups of G with semisimple $\gamma \in M(\mathbb{Q})$ that are elliptic in $M(\mathbb{R})$.

Kottwitz's stable version of Arthur's trace formula is given by

$$\mathcal{K}(f\,dg) = \sum_{(H,s,\eta)\in\mathscr{E}_0} \iota(G,H)\,ST_g(f^H dh),$$

where $\mathscr{E}_0$ is the set of (equivalence classes of) elliptic endoscopic groups for G.

We record here the simpler form of $ST_g(f\,dg,\gamma,M)$ when $\gamma = z$ is in the rational points $Z(\mathbb{Q})$ of the center of G. We have

$$ST_g(f\,dg,z,M) = (-1)^{\dim(A_M/A_G)}\frac{k(M)}{k(G)}(n_M^G)^{-1}\tau(M)\,f_M^\infty(z)\bar{v}(M;dm_\infty)^{-1}\Phi_M(z^{-1},\Theta_\Pi).$$

5.3. ***Conjecture.*** Recall the stable cuspidal measure $f_E dg_\infty$ from Section 2.4. Fix any test function $f^\infty dg_f$ and put $f = f^\infty f_E dg$.

Let

$$T_g(f\,dg) = \sum_M (n_M^G)^{-1}\sum_\gamma \iota^M(\gamma)^{-1}\tau(M_\gamma)O_\gamma(f_M^\infty dm_f)\Phi_M(\gamma, f_E dg_\infty).$$

Again, the sum is over cuspidal Levi subgroups M and semisimple $\gamma\in M(\mathbb{Q})$ that are elliptic in $M(\mathbb{R})$. Here as in [Arthur 1989], $\Phi_M(\gamma,\ \cdot\ )$ is the unnormalized form of the distribution I_M defined in [Arthur 1988a].

Now suppose that $\pi\in\Pi_{\mathrm{disc}}(G(\mathbb{R}),\xi)$, and let K_f be an open compact subgroup of $G(\mathbb{A}_f)$. Write

$$L^2(G(\mathbb{Q})\backslash G(\mathbb{A})/K_f,\xi)$$

for the space of functions on this double coset space that transform by $A_G(\mathbb{R})^+$ according to ξ and are square integrable modulo center. Write $R_{\mathrm{disc}}(\pi,K_f)$ for the π-isotypical subspace of $L^2(G(\mathbb{Q})\backslash G(\mathbb{A})/K_f,\xi)$; it is finite-dimensional. If $f^\infty dg_f$ is K_f-biinvariant, then convolution gives an operator $R_{\mathrm{disc}}(\pi,f^\infty dg_f)$ on $R_{\mathrm{disc}}(\pi,K_f)$. According to [Arthur 1989, Corollary 6.2], if the highest weight of E is regular, then

$$\sum_{\pi\in\Pi_E}\operatorname{tr} R_{\mathrm{disc}}(\pi,f^\infty dg_f) = T_g(f\,dg).$$

The main result of [Kottwitz $\geq$ 2012] is that if $f_\infty dg_\infty$ is stable cuspidal, then $T_g(f\,dg) = \mathcal{K}(f\,dg)$. Since we may assume $f_E dg_\infty = \sum_{\pi\in\Pi_E} e_\pi$, the following conjecture is plausible:

Conjecture 1. *Fix a regular discrete series representation π of $G(\mathbb{R})$. As in Section 2.4, let $f_\infty dg_\infty = e_\pi$. Pick a measure $f^\infty dg_f$ with $f^\infty\in C_c(G(\mathbb{A}_f))$, and $dg_f dg_\infty = dg$ the Tamagawa measure on $G(\mathbb{A})$. Put $f = f^\infty f_\infty$. Then*

$$\mathcal{K}(f\,dg) = \operatorname{tr} R_{\mathrm{disc}}(\pi,f^\infty dg_f).$$

In particular, if we choose a compact open subgroup K_f of $G(\mathbb{A}_f)$, and put $f^\infty dg_f = e_{K_f}$, we obtain

$$m_{\text{disc}}(\pi, K_f) = \mathcal{H}(e_\pi e_{K_f}).$$

In this paper we give some evidence for this conjecture. Moreover, we will see that $\mathcal{H}(f dg)$ is given by a closed algebraic expression, which is straightforward to evaluate, so long as one can compute the transfers e_π^H at the real place, and evaluate the semisimple orbital integrals of $f^{\infty H} dh_f$ at the finite adeles.

6. Euler characteristics

We have finished our discussion of Kottwitz's formula, and now solve the arithmetic volume problem mentioned in the introduction. For simplicity we will write K rather than K_f for open compact subgroups of $G(\mathbb{A}_f)$ in this section.

Definition 15. For K a compact open subgroup of $G(\mathbb{A}_f)$, we define

$$\chi_K(G) = \overline{v}(G; dg_\infty)^{-1} \operatorname{vol}_{dg_f}(K)^{-1} \tau(G) d(G)$$

if G is cuspidal. If G is not cuspidal, then $\chi_K(G) = 0$.

Note that if K_0 is another compact open subgroup of $G(\mathbb{A}_f)$, with $K \subseteq K_0$ of finite index, then $\chi_K(G) = [K_0 : K]\chi_{K_0}(G)$. In this section we compute the quantities $\chi_K(G)$ under some mild hypotheses on G.

6.1. ***Statement of theorem.*** Before getting embroiled in details, let us sketch the idea of the computation of $\chi_K(G)$. The computation is considerably easier if K is sufficiently small. In this case, $\chi_K(G)$ is the classical Euler characteristic of a Shimura variety. This in turn may be written in terms of Euler characteristics of an arithmetic subgroup of $G_{\text{ad}}(\mathbb{R})$. For G a semisimple and simply connected Chevalley group, such Euler characteristics were computed in [Harder 1971].

Our work is to reduce to this case. Given a compact open subgroup K_0 of $G(\mathbb{A}_f)$, we will pick a sufficiently small subgroup K of K_0. By the above we know the analogue of $\chi_K(G)$ for G^{sc}. To compute $\chi_{K_0}(G)$ we have two tasks: to change between G and G^{sc}, and to change between K and K_0.

The resulting formula entails several standard definitions:

Definition 16. Write $G(\mathbb{R})_+ \subseteq G(\mathbb{R})$ for the inverse image of $G_{\text{ad}}(\mathbb{R})^+$. Let $G(\mathbb{Q})_+ = G(\mathbb{Q}) \cap G(\mathbb{R})_+$. Write $\nu : G \twoheadrightarrow C$ for the quotient of G by G_{der}. Let $C(\mathbb{R})^\dagger = \nu(Z(\mathbb{R}))$, and $C(\mathbb{Q})^\dagger = C(\mathbb{Q}) \cap C(\mathbb{R})^\dagger$. Write $\rho : G_{\text{sc}} \to G_{\text{der}}$ for the usual covering of G_{der} by G_{sc}. For K a compact open subgroup of $G(\mathbb{A}_f)$, let $K^{\text{der}} = G_{\text{der}}(\mathbb{A}_f) \cap K$, and let K^{sc} be the preimage of K in $G_{\text{sc}}(\mathbb{A}_f)$. Let $\Gamma_K = G(\mathbb{Q})_+ \cap K$, let $\Gamma_K^{\text{der}} = G_{\text{der}}(\mathbb{Q})_+ \cap K$, let $\Gamma_K^{\text{sc}} = K^{\text{sc}} \cap G_{\text{sc}}(\mathbb{Q})_+$, and write Γ_K^{ad} for the image of Γ_K in $G_{\text{ad}}(\mathbb{Q})$.

In this section we avoid certain awkward tori for simplicity, preferring the following kind:

Definition 17. A torus T over $\mathbb{Q}$ is $\mathbb{QR}$-equitropic if the largest $\mathbb{Q}$-anisotropic torus in T is $\mathbb{R}$-anisotropic.

Here are some basic facts about $\mathbb{QR}$-equitropic tori.

Proposition 5. *If T is a $\mathbb{QR}$-equitropic torus, then $T(\mathbb{Q})$ is discrete in $T(\mathbb{A}_f)$. If G is a reductive group, and the connected component Z° of the center of G is $\mathbb{QR}$-equitropic, then its derived quotient C is also $\mathbb{QR}$-equitropic.*

Proof. The first statement follows from [Milne 2005, Theorem 5.26]. The second is straightforward. □

Serre [1971] introduces an Euler characteristic $\chi_{\mathrm{alg}}(\Gamma) \in \mathbb{Q}$ applicable to any group Γ with a finite index subgroup Γ_0 that is torsion-free and has finite cohomological dimension. In particular, it applies to our congruence subgroups $\Gamma = \Gamma_K$. Here are some simple properties of χ_{alg}:

- For an exact sequence of the form
$$1 \to A \to B \to C \to 1,$$
with A, B and C groups as above, we have $\chi_{\mathrm{alg}}(B) = \chi_{\mathrm{alg}}(A) \cdot \chi_{\mathrm{alg}}(C)$.
- If Γ is a finite group, then $\chi_{\mathrm{alg}}(\Gamma) = |\Gamma|^{-1}$.

The theorem of this section relates $\chi_K(G)$ to $\chi_{\mathrm{alg}}(\Gamma_K^{\mathrm{sc}})$. More precisely:

Theorem 2. *Let G be a reductive group over $\mathbb{Q}$. Assume that G_{sc} has no compact factors and that the connected component Z° of the center of G is $\mathbb{QR}$-equitropic. Let $K_0 \subset G(\mathbb{A}_f)$ be a compact open subgroup. Then $\chi_{K_0}(G)$ is equal to*

$$\frac{\begin{array}{c}|\ker(\rho(\mathbb{Q}))|[G_{\mathrm{der}}(\mathbb{A}_f) : G_{\mathrm{der}}(\mathbb{Q})_+ K_0^{\mathrm{der}}] \\ \cdot [\Gamma_{K_0}^{\mathrm{der}} : G_{\mathrm{der}}(\mathbb{Q})_+ \cap \rho(K_0^{\mathrm{sc}})][C(\mathbb{A}_f) : C(\mathbb{Q})^\dagger \nu(K_0)]\end{array}}{[G(\mathbb{R}) : G(\mathbb{R})_+]|\nu(K_0) \cap C(\mathbb{Q})^\dagger|} \chi_{\mathrm{alg}}(\Gamma_{K_0}^{\mathrm{sc}}).$$

Here $\rho(\mathbb{Q})$ denotes the map $\rho(\mathbb{Q}) : G_{\mathrm{sc}}(\mathbb{Q}) \to G(\mathbb{Q})$ on $\mathbb{Q}$-points. The assumption on the absence of compact factors is needed for strong approximation, and is discussed in [Milne 2005].

When G_{sc} is a Chevalley group and $\Gamma_{K_0}^{\mathrm{sc}} = G_{\mathrm{sc}}(\mathbb{Z})$, this reduces the problem to the calculation of Harder [1971]:

Proposition 6. *Let G be a simply connected, semisimple Chevalley group over $\mathbb{Z}$. Write $m_1, \ldots, m_r$ for the exponents of its Weyl group Ω, and put $\Gamma = G(\mathbb{Z})$. We have*

$$\chi_{\mathrm{alg}}(\Gamma) = (-\tfrac{1}{2})^r |\Omega_{\mathbb{R}}|^{-1} \prod_{i=1}^{r} B_{m_i+1}.$$

Here B_n denotes the n-th Bernoulli number. Recall that $\Omega_{\mathbb{R}}$ is the real Weyl group of G.

6.2. ***Shimura varieties.*** To prove Theorem 2, we will use some basic Shimura variety theory, which may be found in [Deligne 1979] or [Milne 2005]. Much of the theory holds only for K sufficiently small. For simplicity, we will say "K is small" rather than "K is a sufficiently small finite index subgroup of K_0".

For convenience, we gather here many simplifying properties of small K, which we will often use without comment. For the rest of this section assume that $Z(G)^\circ$ is $\mathbb{QR}$-equitropic, and that G_{sc} has no compact factors.

Proposition 7. *Let K be small.*

(i) $K \cap Z(\mathbb{Q}) = \{1\}$.

(ii) $\nu(K) \cap C(\mathbb{Q}) = \{1\}$.

(iii) $G(\mathbb{Q}) \cap K G_{\mathrm{der}}(\mathbb{A}_f) \subseteq G_{\mathrm{der}}(\mathbb{Q})$.

(iv) $G_{\mathrm{der}}(\mathbb{A}_f) \cap G(\mathbb{Q})K = G_{\mathrm{der}}(\mathbb{Q})K_{\mathrm{der}}$.

(v) $K \cap G_{\mathrm{der}}(\mathbb{Q}) \subseteq \rho(G_{\mathrm{sc}}(\mathbb{Q}))$.

(vi) $K \cap G(\mathbb{Q}) \subseteq G(\mathbb{Q})^+$.

Proof. The first two items follow because Z° and thus C are $\mathbb{QR}$-equitropic. Item (iii) follows from [Deligne 1979, Corollaire 2.0.12], and the next is a corollary. Items (v) and (vi) follow from [Deligne 1979, Corollaire 2.0.5 and 2.0.14], respectively. □

Recall that we have chosen a maximal compact subgroup $K_{\mathbb{R}}$ of $G(\mathbb{R})$.

Definition 18. Let

$$X = G(\mathbb{R})/K_{\mathbb{R}}^+ Z(\mathbb{R}), \quad \overline{X} = G(\mathbb{R})/K_{\mathbb{R}} Z(\mathbb{R}), \quad S_K = G(\mathbb{Q})\backslash X \times G(\mathbb{A}_f)/K$$

be the double coset space obtained through the action $q(x, g)k = (qx, qgk)$ of $q \in G(\mathbb{Q})$ and $k \in K$.

Similarly, let

$$\overline{S}_K = G(\mathbb{Q})\backslash \overline{X} \times G(\mathbb{A}_f)/K,$$

with the action of $G(\mathbb{Q}) \times K$ defined in the same way.

The component group of S_K is finite and given (see [Deligne 1979, 2.1.3]) by

$$\pi_0(S_K) = G(\mathbb{A}_f)/G(\mathbb{Q})_+ K. \tag{6-1}$$

There is some variation in the literature regarding the use of X versus $\overline{X}$. Deligne [1979] and Milne [2005] implicitly use X (in light of Deligne's [Proposition 1.2.7]). Harder [1971] uses $\overline{X}$. Arthur [1989] uses

$$G(\mathbb{R})/K'_{\mathbb{R}}.$$

(Recall that $K'_{\mathbb{R}} = A_G(\mathbb{R})^+ K_{\mathbb{R}}$.) Since for us Z° is $\mathbb{Q}\mathbb{R}$-equitropic, we have

$$K'_{\mathbb{R}} = Z(\mathbb{R}) K_{\mathbb{R}},$$

and so this quotient is equal to $\overline{X}$.

Since we would like to combine results stated in terms of X with others stated in terms of $\overline{X}$, we must understand the precise relationship between the two. This is the purpose of Proposition 8 below.

Definition 19. Let G be a real group, and Z its center. Write

$$\mathrm{ad} : G(\mathbb{R}) \to G(\mathbb{R})/Z(\mathbb{R}) \tag{6-2}$$

for the quotient map.

Note that $\mathrm{ad}(G(\mathbb{R}))$ has finite index in $G_{\mathrm{ad}}(\mathbb{R})$.

Lemma 1. *For this lemma, let G be a Zariski-connected reductive real group, and $K_{\mathbb{R}}$ a maximal compact subgroup of $G(\mathbb{R})$. Let $L_{\mathbb{R}}$ be a maximal compact subgroup of $G_{\mathrm{ad}}(\mathbb{R})$ containing $\mathrm{ad}(K_{\mathbb{R}})$. Then the following hold*:

(i) *$K_{\mathbb{R}}$ meets all the connected components of $G(\mathbb{R})$.*

(ii) $K_{\mathbb{R}} \cap G(\mathbb{R})^+ = K_{\mathbb{R}}^+$.

(iii) *$\mathrm{ad}(K_{\mathbb{R}})$ is a maximal compact subgroup of $\mathrm{ad}(G(\mathbb{R}))$.*

(iv) $\mathrm{ad}(K_{\mathbb{R}}^+) = L_{\mathbb{R}}^+$.

(v) $K_{\mathbb{R}} Z(\mathbb{R}) \cap G(\mathbb{R})_+ = K_{\mathbb{R}}^+ Z(\mathbb{R})$.

Proof. The first two statements follow from the Cartan decomposition [Satake 1980, Corollary 4.5].

For (iii), suppose that C is a subgroup of $G(\mathbb{R})$ with $\mathrm{ad}(K_{\mathbb{R}}) \subseteq \mathrm{ad}(C)$ and $\mathrm{ad}(C)$ compact. If $\mathrm{ad}(K_{\mathbb{R}}) \neq \mathrm{ad}(C)$, there is an element $a \in CZ(\mathbb{R}) - K_{\mathbb{R}} Z(\mathbb{R})$. By the Cartan decomposition, we may assume that $a = \exp(H)$, with H a semisimple element of $\mathrm{Lie}(G)$, and $\alpha(H)$ real and nonnegative for every root α of G. Since $a \notin Z(\mathbb{R})$, we have $\alpha(H) > 0$ for some root α. Thus $\mathrm{ad}(C)$ is not compact, a contradiction. Thus $\mathrm{ad}(K_{\mathbb{R}}) = \mathrm{ad}(C)$, and statement (iii) follows.

For (iv), note that $L_{\mathbb{R}} \cap \mathrm{ad}(G) = \mathrm{ad}(K_{\mathbb{R}})$, and therefore $L_{\mathbb{R}}/\mathrm{ad}(K_{\mathbb{R}})$ injects into $G_{\mathrm{ad}}(\mathbb{R})/\mathrm{ad}(G(\mathbb{R}))$. It follows that $\mathrm{ad}(K_{\mathbb{R}}^+)$ has finite index in $L_{\mathbb{R}}$. Since it is connected, statement (iv) follows.

For (v), let $g \in K_{\mathbb{R}} Z(\mathbb{R}) \cap G(\mathbb{R})_+$. Then $\mathrm{ad}(g) \in L_{\mathbb{R}} \cap G_{\mathrm{ad}}(\mathbb{R})^+$, so by statement (ii), we see $\mathrm{ad}(g) \in L_{\mathbb{R}}^+ = \mathrm{ad}(K_{\mathbb{R}}^+)$. Thus $g \in K_{\mathbb{R}}^+ Z(\mathbb{R})$. The other inclusion is obvious. □

Proposition 8.

(i) *The natural projection $p_X : X \to \overline{X}$ has fibers of order $[G(\mathbb{R}) : G(\mathbb{R})_+]$.*

(ii) *Let X^+ be a connected component of X. It is stabilized by $G(\mathbb{R})_+$, and the restriction of p_X to X^+ is a $G(\mathbb{R})_+$-isomorphism onto $\overline{X}$.*

(iii) *Let K be small. Then the natural projection $p_S : S_K \to \overline{S}_K$ has fibers of order $[G(\mathbb{R}) : G(\mathbb{R})_+]$.*

Proof. Consider the natural map

$$K_{\mathbb{R}} Z(\mathbb{R})/K_{\mathbb{R}}^+ Z(\mathbb{R}) \to G(\mathbb{R})/G(\mathbb{R})_+. \tag{6-3}$$

It is surjective because $K_{\mathbb{R}}$ meets every connected component of $G(\mathbb{R})$. It is injective because $K_{\mathbb{R}} Z(\mathbb{R}) \cap G(\mathbb{R})_+ \subseteq K_{\mathbb{R}}^+ Z(\mathbb{R})$. It follows that (6-3) is an isomorphism, and the first statement follows.

We now prove the second statement. Note that p_X is both an open and closed map, so that $p_X(X^+)$ is a component of $\overline{X}$. Since $K_{\mathbb{R}}$ meets every connected component of $G(\mathbb{R})$, the set $\overline{X}$ is connected. Therefore $p_X(X^+) = \overline{X}$. By [Milne 2005, Proposition 5.7], there are $[G(\mathbb{R}) : G(\mathbb{R})_+]$ connected components of X, each stabilized by $G(\mathbb{R})_+$. Thus the fiber over a point in $\overline{X}$ is composed of exactly one point from each component of X. So p_X restricted to X^+ is an isomorphism; it is clear that it respects the $G(\mathbb{R})_+$-action.

To prove the third statement, we require K to be sufficiently small, in the following way. Suppose K_* is an open compact subgroup of $G(\mathbb{A}_f)$ satisfying $K_* \cap G(\mathbb{Q}) \subseteq G(\mathbb{Q})^+$. Let $g_1, \ldots, g_r$ be representatives of the finite quotient group $G(\mathbb{Q})K_* \backslash G(\mathbb{A}_f)$. Then we require that

$$K \subseteq \bigcap_{i=1}^{r} g_i^{-1} K_* g_i. \tag{6-4}$$

Now for $x \in X$, let $\mathrm{Fib}(x)$ be the fiber of p_X containing x. If we further fix $g \in G(\mathbb{A}_f)$, let $\mathrm{Fib}(x, g)$ be the fiber of p_S containing (x, g). (Here we understand (x, g) as an element of S_K.) We claim that for all such x and g, the map

$$\mathrm{Fib}(x) \to \mathrm{Fib}(x, g) \tag{6-5}$$

given by $x' \mapsto (x', g)$ is a bijection. This will imply the third statement.

For surjectivity of (6-5), pick $(x', g') \in \mathrm{Fib}(x, g)$. Then there are $q \in G(\mathbb{Q})$ and $k \in G(\mathbb{A}_f)$ such that $q p_X(x') = p_X(x)$ and $q g' k = g$. Let $x'' = q x'$. Then $x'' \in \mathrm{Fib}(x)$ and $(x'', g) = (x', g')$.

For injectivity of (6-5), suppose that $(x_1, g) = (x_2, g)$ in S_K with $x_1, x_2 \in \mathrm{Fib}(x)$. Then in particular, there is an element $q \in G(\mathbb{Q})$ and $k \in K$ such that $qgk = g$ and $q x_1 = x_2$. Write $g = q_0 k_0 g_i$ with $q_0 \in G(\mathbb{Q})$ and $k_0 \in K_*$. Then we have

$$q(q_0 k_0 g_i)k = q_0 k_0 g_i,$$

which we rewrite as

$$q_0^{-1} q q_0 = k_0 g_i k^{-1} g_i^{-1} k_0^{-1}.$$

Using this and (6-4) we see that $q_0^{-1} q q_0 \in G(\mathbb{Q}) \cap K_* \subseteq G(\mathbb{Q})^+$. Since $G(\mathbb{Q})^+$ is normal in $G(\mathbb{Q})$, in fact $q \in G(\mathbb{Q})^+$.

Meanwhile, pick $\xi_1, \xi_2 \in G(\mathbb{R})$ representing x_1 and x_2, respectively. Since $x_1, x_2 \in \mathrm{Fib}(x)$ we have $\xi_1^{-1}\xi_2 \in K_{\mathbb{R}} Z(\mathbb{R})$. Write $\xi_2 = \xi_1 k z$, with $k \in K_{\mathbb{R}}$ and $z \in Z(\mathbb{R})$. Since $q x_1 = x_2$, we have $\xi_2^{-1} q \xi_1 \in K_{\mathbb{R}}^+ Z(\mathbb{R})$, and thus $z^{-1} k^{-1} \xi_1^{-1} q \xi_1 \in K_{\mathbb{R}}^+ Z(\mathbb{R})$. Using the fact that q is in the normal subgroup $G(\mathbb{R})_+$ of $G(\mathbb{R})$, it follows that $k \in G(\mathbb{R})_+ \cap K_{\mathbb{R}} \subseteq K_{\mathbb{R}}^+ Z(\mathbb{R})$. Thus $x_1 = x_2$, as desired. □

Proposition 9 (Harder; see [Harder 1971; Serre 1971]). *If G is semisimple and K is small, then $\chi_{\mathrm{top}}(\Gamma_K \backslash \overline{X}) = \chi_{\mathrm{alg}}(\Gamma_K)$.*

Proposition 10 [Arthur 1989; Goresky et al. 1997]. *If K is small, then we have $\chi_K(G) = \chi_{\mathrm{top}}(\overline{S}_K)$.*

6.3. ***Computations.*** The next three lemmas will allow us to convert our computation for K_0 to a computation for K.

Lemma 2. *If K is small, then*

$$|C(\mathbb{Q})^\dagger \backslash C(\mathbb{A}_f)/\nu(K)| = [\nu(K_0) : \nu(K)] |\nu(K_0) \cap C(\mathbb{Q})^\dagger|^{-1} |C(\mathbb{Q})^\dagger \backslash C(\mathbb{A}_f)/\nu(K_0)|.$$

Proof. This follows from the exactness of the sequence

$$1 \to \nu(K_0) \cap C(\mathbb{Q})^\dagger \to \nu(K_0)/\nu(K) \to C(\mathbb{Q})^\dagger \backslash C(\mathbb{A}_f)/\nu(K) \to C(\mathbb{Q})^\dagger \backslash C(\mathbb{A}_f)/\nu(K_0) \to 1. \quad \square$$

Lemma 3. *If $K \subseteq K_0$ is small, then*

$$[\Gamma_{K_0}^{\mathrm{ad}} : \Gamma_K^{\mathrm{ad}}] = \frac{[\Gamma_{K_0} : \rho(\Gamma_{K_0}^{\mathrm{sc}})][K_0 : K]}{|K_0 \cap Z(\mathbb{Q})|[\nu(K_0) : \nu(K)][K_0^{\mathrm{der}} : K^{\mathrm{der}} \rho(K_0^{\mathrm{sc}})]}. \tag{6-6}$$

In the proof we refer to conditions of Proposition 7.

Proof. Consider the map $\Gamma_{K_0}^{\mathrm{der}} / \Gamma_K^{\mathrm{der}} \to \Gamma_{K_0}^{\mathrm{ad}} / \Gamma_K^{\mathrm{ad}}$.

The kernel of this map sits in the middle of the exact sequence

$$1 \to \Gamma_{K_0}^{\mathrm{der}} \cap Z(\mathbb{Q}) \to (\Gamma_K Z(\mathbb{Q}) \cap \Gamma_{K_0}^{\mathrm{der}}) / \Gamma_K^{\mathrm{der}} \to (\Gamma_K Z(\mathbb{Q}) \cap \Gamma_{K_0}^{\mathrm{der}}) / \Gamma_K^{\mathrm{der}} (\Gamma_{K_0}^{\mathrm{der}} \cap Z(\mathbb{Q})) \to 1,$$

using condition (i). This last quotient is trivial, because actually $\Gamma_K = \Gamma_K^{\mathrm{der}}$ by condition (iii).

We have established the exactness of the sequence

$$1 \to \Gamma_{K_0}^{\mathrm{der}} \cap Z(\mathbb{Q}) \to \Gamma_{K_0}^{\mathrm{der}} / \Gamma_K^{\mathrm{der}} \to \Gamma_{K_0}^{\mathrm{ad}} / \Gamma_K^{\mathrm{ad}} \to \Gamma_{K_0} Z(\mathbb{Q}) / \Gamma_{K_0}^{\mathrm{der}} Z(\mathbb{Q}) \to 1.$$

The last quotient is isomorphic to $\Gamma_{K_0}/(Z(\mathbb{Q}) \cap K_0)\Gamma_{K_0}^{\mathrm{der}}$, which itself sits inside the exact sequence

$$1 \to K_0 \cap Z(\mathbb{Q}) / \Gamma_{K_0}^{\mathrm{der}} \cap Z(\mathbb{Q}) \to \Gamma_{K_0} / \Gamma_{K_0}^{\mathrm{der}} \to \Gamma_{K_0}/(Z(\mathbb{Q}) \cap K_0)\Gamma_{K_0}^{\mathrm{der}} \to 1.$$

The quantity $|\Gamma_{K_0}^{\mathrm{der}} \cap Z(\mathbb{Q})|$ cancels, and it follows that

$$[\Gamma_{K_0}^{\mathrm{ad}} : \Gamma_K^{\mathrm{ad}}] = \frac{[\Gamma_{K_0}^{\mathrm{der}} : \Gamma_K^{\mathrm{der}}] \cdot [\Gamma_{K_0} : \Gamma_{K_0}^{\mathrm{der}}]}{|K_0 \cap Z(\mathbb{Q})|}. \tag{6-7}$$

By condition (v) we have

$$1 \to \rho(\Gamma_{K_0}^{\mathrm{sc}})/\rho(\Gamma_K^{\mathrm{sc}}) \to \Gamma_{K_0}^{\mathrm{der}} / \Gamma_K^{\mathrm{der}} \to \Gamma_{K_0}^{\mathrm{der}} / \rho(\Gamma_{K_0}^{\mathrm{sc}}) \to 1.$$

Strong approximation tells us that $G_{\mathrm{sc}}(\mathbb{Q})$ is dense in $G_{\mathrm{sc}}(\mathbb{A}_f)$. Therefore we have isomorphisms

$$\rho(\Gamma_{K_0}^{\mathrm{sc}})/\rho(\Gamma_K^{\mathrm{sc}}) \xrightarrow{\sim} \Gamma_{K_0}^{\mathrm{sc}} / \Gamma_K^{\mathrm{sc}} \xrightarrow{\sim} K_0^{\mathrm{sc}}/K^{\mathrm{sc}} \xrightarrow{\sim} \rho(K_0^{\mathrm{sc}})/\rho(K^{\mathrm{sc}}).$$

Combining this with the exact sequences

$$1 \to K_0^{\mathrm{der}}/K^{\mathrm{der}} \to K_0/K \to \nu(K_0)/\nu(K) \to 1$$

and

$$1 \to \rho(K_0^{\mathrm{sc}})/\rho(K^{\mathrm{sc}}) \to K_0^{\mathrm{der}}/K^{\mathrm{der}} \to K_0^{\mathrm{der}}/K^{\mathrm{der}}\rho(K_0^{\mathrm{sc}}) \to 1, \tag{6-8}$$

we obtain

$$[\Gamma_{K_0}^{\mathrm{der}} : \Gamma_K^{\mathrm{der}}] = \frac{[\Gamma_{K_0}^{\mathrm{der}} : \rho(\Gamma_{K_0}^{\mathrm{sc}})][K_0 : K]}{[K_0^{\mathrm{der}} : K^{\mathrm{der}}\rho(K_0^{\mathrm{sc}})][\nu(K_0) : \nu(K)]}.$$

Plugging this into (6-7) gives the lemma. □

Corollary 1. *Suppose that $K \subseteq K_0$ is small, and $g \in G(\mathbb{A}_f)$ with $gKg^{-1} \subseteq K_0$ also small. Then*

$$[\Gamma_{K_0}^{\mathrm{ad}} : \Gamma_{gKg^{-1}}^{\mathrm{ad}}] = [\Gamma_{K_0}^{\mathrm{ad}} : \Gamma_K^{\mathrm{ad}}].$$

Proof. We show that the expression (6-6) does not change when K is replaced with gKg^{-1}. Clearly $\nu(K) = \nu(gKg^{-1})$. Since

$$[K_0 : K] = \mathrm{vol}_{dg_f}(K_0)/\mathrm{vol}_{dg_f}(K),$$

we have $[K_0 : gKg^{-1}] = [K_0 : K]$. Finally, we claim that

$$[K_0^{\mathrm{der}} : (gKg^{-1})^{\mathrm{der}}\rho(K_0^{\mathrm{sc}})] = [K_0^{\mathrm{der}} : K^{\mathrm{der}}\rho(K_0^{\mathrm{sc}})].$$

From the exact sequence (6-8), it is enough to show that $[K_0^{\mathrm{der}} : (gKg^{-1})^{\mathrm{der}}] = [K_0^{\mathrm{der}} : K^{\mathrm{der}}]$ and $[\rho(K_0^{\mathrm{sc}}) : \rho((gKg^{-1})^{\mathrm{sc}})] = [\rho(K_0^{\mathrm{sc}}) : \rho(K^{\mathrm{sc}})]$. These hold because $(gKg^{-1})^{\mathrm{der}} = gK^{\mathrm{der}}g^{-1}$ and $\rho((gKg^{-1})^{\mathrm{sc}}) = g\rho(K^{\mathrm{sc}})g^{-1}$. □

Lemma 4. *If G is semisimple and K is small, then*

$$|\pi_0(S_K)| = [K_0 : K\rho(K_0^{\mathrm{sc}})][\Gamma_{K_0} : G(\mathbb{Q})_+ \cap \rho(K_0^{\mathrm{sc}})]|\pi_0(S_{K_0})|.$$

Proof. The kernel of the projection $\pi_0(S_K) \twoheadrightarrow \pi_0(S_{K_0})$ is isomorphic to

$$K_0/(KG(\mathbb{Q})_+ \cap K_0).$$

By [Deligne 1979, Section 2.1.3], we have $\rho(G_{\mathrm{sc}}(\mathbb{A}_f)) \subseteq KG(\mathbb{Q})_+$. Using the exact sequence

$$1 \to (K_0 \cap KG(\mathbb{Q})_+)/K\rho(K_0^{\mathrm{sc}}) \to K_0/K\rho(K_0^{\mathrm{sc}}) \to K_0/(KG(\mathbb{Q})_+ \cap K_0) \to 1,$$

we are reduced to computing the order of

$$(K_0 \cap KG(\mathbb{Q})_+)/K\rho(K_0^{\mathrm{sc}}) \xrightarrow{\sim} \Gamma_{K_0}/(K\rho(K_0^{\mathrm{sc}}) \cap G(\mathbb{Q})_+).$$

This group sits in the sequence

$$\begin{aligned} 1 \to (G(\mathbb{Q})_+ \cap K\rho(K_0^{\mathrm{sc}}))/(G(\mathbb{Q})_+ \cap \rho(K_0^{\mathrm{sc}})) \\ \to \Gamma_{K_0}/(G(\mathbb{Q})_+ \cap \rho(K_0^{\mathrm{sc}})) \to \Gamma_{K_0}/(K\rho(K_0^{\mathrm{sc}}) \cap G(\mathbb{Q})_+) \to 1. \end{aligned}$$

We claim the kernel is trivial. Note that $K\rho(K_0^{\mathrm{sc}}) \subseteq K\rho(G_{\mathrm{sc}}(\mathbb{Q})K^{\mathrm{sc}})$ by strong approximation. So

$$\begin{aligned} G(\mathbb{Q})_+ \cap K\rho(K_0^{\mathrm{sc}}) &\subseteq G(\mathbb{Q})_+ \cap K\rho(G_{\mathrm{sc}}(\mathbb{Q})) \\ &= G(\mathbb{Q})_+ \cap (K \cap G(\mathbb{Q}))\rho(G_{\mathrm{sc}}(\mathbb{Q})). \end{aligned}$$

Since $K \cap G(\mathbb{Q}) \subseteq \rho(G_{\mathrm{sc}}(\mathbb{Q}))$ by Proposition 7(v), we have $G(\mathbb{Q})_+ \cap K\rho(K_0^{\mathrm{sc}}) \subseteq G(\mathbb{Q})_+ \cap \rho(K_0^{\mathrm{sc}})$. This proves the claim, and the lemma follows. □

In the course of proving the theorem, we will pass to the adjoint group to apply Harder's theorem (Proposition 9), but lift to G_{sc} to apply Harder's calculation (Proposition 6). We must record the difference between Serre's Euler characteristic at G_{ad} and G_{sc}.

Lemma 5. *We have*

$$\chi_{\mathrm{alg}}(\Gamma_{K_0}^{\mathrm{ad}}) = \frac{|\ker(\rho(\mathbb{Q}))||K_0 \cap Z(\mathbb{Q})|}{[\Gamma_{K_0}^{\mathrm{der}} : \rho(\Gamma_{K_0}^{\mathrm{sc}})][\Gamma_{K_0} : \Gamma_{K_0}^{\mathrm{der}}]} \chi_{\mathrm{alg}}(\Gamma_{K_0}^{\mathrm{sc}}).$$

Proof. This follows from the properties of χ_{alg} mentioned earlier. □

Proof of Theorem 2. Pick a set $g_1, \ldots, g_r$ of representatives of $\pi_0(S_{K_0})$, viewed as a quotient of $G(\mathbb{A}_f)$ as in (6-1).

Let K be small subgroup of finite index in K_0. Possibly by intersecting finitely many conjugates of K, we may assume that

- K is normal in K_0 and
- $g_i K g_i^{-1}$ is a small subgroup of K_0 for all i.

By Proposition 10, $\chi_K(G) = \chi_{\text{top}}(\overline{S}_K)$. By Proposition 8, this is equal to $[G(\mathbb{R}) : G(\mathbb{R})_+]^{-1}\chi_{\text{top}}(S_K)$. Write Γ_g for $\Gamma^{\text{ad}}_{gKg^{-1}}$. By [Deligne 1979, 2.1.2], the components of S_K are each isomorphic to $\Gamma_g\backslash X^+$, where X^+ is a component of X. Here g runs over $\pi_0(S_K)$.

By Proposition 8, the topological spaces $\Gamma_g\backslash X^+$ and $\Gamma_g\backslash\overline{X}$ are isomorphic. Therefore we have $\chi_{\text{top}}(\Gamma_g\backslash X^+) = \chi_{\text{top}}(\Gamma_g\backslash\overline{X})$.

Applying Proposition 9 to G_{ad}, this is equal to $\chi_{\text{alg}}(\Gamma_g)$. Therefore

$$\chi_K(G) = [G(\mathbb{R}) : G(\mathbb{R})_+]^{-1} \sum_{g\in\pi_0(S_K)} \chi_{\text{alg}}(\Gamma_g).$$

Every element in $\pi_0(S_K)$ may be written as the product of an element of $\pi_0(S_{K_0})$ with an element of K_0. Since K is normal in K_0, the groups Γ_{gk_0} and Γ_g are equal for $k_0 \in K_0$. It follows that

$$\chi_K(G) = \frac{|\pi_0(S_K)|}{[G(\mathbb{R}) : G(\mathbb{R})_+]|\pi_0(S_{K_0})|} \sum_{i=1}^{r} \chi_{\text{alg}}(\Gamma_{g_i}).$$

By Corollary 1 we have

$$\chi_{\text{alg}}(\Gamma_{g_i}) = [\Gamma^{\text{ad}}_{K_0} : \Gamma_{g_i}]\chi_{\text{alg}}(\Gamma^{\text{ad}}_{K_0}) = [\Gamma^{\text{ad}}_{K_0} : \Gamma^{\text{ad}}_K]\chi_{\text{alg}}(\Gamma^{\text{ad}}_{K_0}).$$

This gives

$$\chi_K(G) = [G(\mathbb{R}) : G(\mathbb{R})_+]^{-1}[\Gamma^{\text{ad}}_{K_0} : \Gamma^{\text{ad}}_K]|\pi_0(S_K)|\chi_{\text{alg}}(\Gamma^{\text{ad}}_{K_0}).$$

The component group $\pi_0(S_K)$ fits into the exact sequence

$$1 \to G_{\text{der}}(\mathbb{A}_f)/(G_{\text{der}}(\mathbb{A}_f) \cap G(\mathbb{Q})_+K) \to \pi_0(S_K) \to C(\mathbb{Q})^\dagger\backslash C(\mathbb{A}_f)/\nu(K) \to 1$$

This gives

$$\chi_K(G) = [G(\mathbb{R}) : G(\mathbb{R})_+]^{-1}|\pi_0(S_{K^{\text{der}}})||C(\mathbb{Q})^\dagger\backslash C(\mathbb{A}_f)/\nu(K)|[\Gamma^{\text{ad}}_{K_0} : \Gamma^{\text{ad}}_K]\chi_{\text{alg}}(\Gamma^{\text{ad}}_{K_0}).$$

where here $\pi_0(S_{K^{\text{der}}}) = G_{\text{der}}(\mathbb{A}_f)/G_{\text{der}}(\mathbb{Q})_+K^{\text{der}}$.

Using $\chi_{K_0}(G) = [K_0 : K]^{-1}\chi_K(G)$ together with Lemma 2 gives

$$\chi_{K_0}(G) = \frac{|\pi_0(S_{K^{\text{der}}})|[\nu(K_0) : \nu(K)]|C(\mathbb{Q})^\dagger\backslash C(\mathbb{A}_f)/\nu(K_0)|[\Gamma^{\text{ad}}_{K_0} : \Gamma^{\text{ad}}_K]}{[G(\mathbb{R}) : G(\mathbb{R})_+]|\nu(K_0) \cap C(\mathbb{Q})^\dagger|[K_0 : K]}\chi_{\text{alg}}(\Gamma^{\text{ad}}_{K_0}).$$

By Lemmas 3 and 5,

$$\chi_{K_0}(G) = \frac{|\ker(\rho(\mathbb{Q}))||\pi_0(S_{K^{\mathrm{der}}})||C(\mathbb{Q})^\dagger\backslash C(\mathbb{A}_f)/\nu(K_0)|}{[G(\mathbb{R}) : G(\mathbb{R})_+]|\nu(K_0)\cap C(\mathbb{Q})^\dagger|[K_0^{\mathrm{der}} : K^{\mathrm{der}}\rho(K_0^{\mathrm{sc}})]}\chi_{\mathrm{alg}}(\Gamma_{K_0}^{\mathrm{sc}}).$$

The theorem then follows from Lemma 4. □

6.4. ***Examples.*** We now use Theorem 2 and Proposition 6 to explicitly compute some cases of $\chi_{K_0}(G)$. Recall that we write $\mathbb{O}_f$ for the integer points of $\mathbb{A}_f$.

Corollary 2. *If T is a torus and $K_0 \subset T(\mathbb{A}_f)$ is a compact open subgroup, then*

$$\chi_{K_0}(T) = |T(\mathbb{Q})\backslash T(\mathbb{A}_f)/K_0| \cdot |K_0 \cap T(\mathbb{Q})|^{-1}.$$

Let $T = \mathbb{G}_m$, and $K_0 = T(\mathbb{O}_f)$. Then $\chi_{K_0}(T) = 1/2$.

Let T be the norm-one subgroup of an imaginary quadratic extension E of $\mathbb{Q}$. Let $K_0 = T(\mathbb{O}_f)$. Write $\mathbb{O}(E)$ for the integer points of the adeles $\mathbb{A}_E$ over E. Then $T(\mathbb{Q})\backslash T(\mathbb{A}_f)/K_0$ injects into $E^\times\backslash\mathbb{A}_{E,f}^\times/\mathbb{O}(E)^\times$, which is in bijection with the class group. If the class number of E is trivial, it follows that $\chi_{K_0}(T) = |T(\mathbb{Z})|^{-1}$.

Corollary 3. *If G is semisimple and simply connected, then*

$$\chi_{K_0}(G) = [G(\mathbb{R}) : G(\mathbb{R})_+]^{-1}\chi_{\mathrm{alg}}(\Gamma_{K_0}).$$

Let $G = \mathrm{SL}_2$ and $K_0 = G(\mathbb{O}_f)$. Then

$$\chi_{K_0}(G) = \chi_{\mathrm{alg}}(\mathrm{SL}_2(\mathbb{Z})) = -\tfrac{1}{2}B_2 = -2^{-2}3^{-1}.$$

Let $G = \mathrm{Sp}_4$ and $K_0 = G(\mathbb{O}_f)$. Then

$$\chi_{K_0}(G) = \chi_{\mathrm{alg}}(\mathrm{Sp}_4(\mathbb{Z})) = -\tfrac{1}{8}B_2B_4 = -2^{-5}3^{-2}5^{-1}.$$

When the derived group is simply connected the calculation is not much harder.

Corollary 4. *If G_{der} is simply connected, then*

$$\chi_{K_0}(G) = \frac{|C(\mathbb{Q})^\dagger\backslash C(\mathbb{A}_f)/\nu(K_0)|}{[G(\mathbb{R}) : G(\mathbb{R})_+]|\nu(K_0)\cap C(\mathbb{Q})^\dagger|}\chi_{\mathrm{alg}}(\Gamma_{K_0}^{\mathrm{der}}).$$

Let $G = \mathrm{GL}_2$ and $K_0 = G(\mathbb{O}_f)$. Then $\chi_{K_0}(G) = \frac{1}{2}\chi_{\mathrm{alg}}(\mathrm{SL}_2(\mathbb{Z})) = -2^{-3}3^{-1}$.

Let $G = \mathrm{GSp}_4$ and $K_0 = G(\mathbb{O}_f)$. Then $\chi_{K_0}(G) = \frac{1}{2}\chi_{\mathrm{alg}}(\mathrm{Sp}_4(\mathbb{Z})) = -2^{-6}3^{-2}5^{-1}$.

Lemma 6. *If all the points of* $\ker\rho$ *are $\mathbb{Q}$-rational, then*

$$[\Gamma_{K_0}^{\mathrm{der}} : G_{\mathrm{der}}(\mathbb{Q})_+ \cap \rho(K_0^{\mathrm{sc}})] = 1.$$

Proof. By [Deligne 1979, Section 2.0.3], we have an injection

$$G_{\mathrm{der}}(\mathbb{Q})/\rho(G_{\mathrm{sc}}(\mathbb{Q})) \hookrightarrow H^1(\mathrm{im}(\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})), (\ker\rho)(\overline{\mathbb{Q}})),$$

using the cohomology group defined in that paper. We also have an injection

$$\Gamma^{\mathrm{der}}_{K_0}/(G_{\mathrm{der}}(\mathbb{Q})_+ \cap \rho(K_0^{\mathrm{sc}})) \hookrightarrow G_{\mathrm{der}}(\mathbb{Q})/\rho(G_{\mathrm{sc}}(\mathbb{Q})).$$

Since all the points of $\ker\rho$ are $\mathbb{Q}$-rational, all these groups are trivial. □

Let $G = \mathrm{PGL}_2$ and $K_0 = G(\mathbb{O}_f)$. The only nontrivial factors in the formula are $[G(\mathbb{R}) : G(\mathbb{R})_+] = 2$, $|\ker\rho(\mathbb{Q})| = 2$, and $\chi_{\mathrm{alg}}(\mathrm{SL}_2(\mathbb{Z})) = -2^{-2}3^{-1}$. Thus $\chi_{K_0}(G) = -2^{-2}3^{-1}$.

7. The case of SL_2

Let $G = \mathrm{SL}_2$, defined over $\mathbb{Q}$. Let A be the subgroup of diagonal matrices in G, and let T be the maximal elliptic torus of G given by matrices

$$\gamma_{a,b} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}, \tag{7-1}$$

with $a^2 + b^2 = 1$.

The characters and cocharacters of T are both isomorphic to $\mathbb{Z}$. We identify $\mathbb{Z} \xrightarrow{\sim} X^*(T)$ via $n \mapsto \chi_n$, where $\chi_n(\gamma_{a,b}) = (a+bi)^n$. We specify $\mathbb{Z} \xrightarrow{\sim} X_*(T)$ by identifying n with the cocharacter taking α to $\mathrm{diag}(\alpha, \alpha^{-1})$. The roots of T in G are then $\{\pm 2\}$, and the coroots of T in G are $\{\pm 1\}$. The Weyl group Ω of these systems has order 2 and the compact Weyl group $\Omega_{\mathbb{R}}$ is trivial. Thus each L-packet of discrete series has order 2. The group dual to G is $\hat{G} = \mathrm{PGL}_2(\mathbb{C})$ in the usual way.

Pick an element $\xi \in G(\mathbb{C})$ such that

$$\mathrm{Ad}(\xi)\begin{pmatrix} a & -b \\ b & a \end{pmatrix} = \begin{pmatrix} a+ib & \\ & a-ib \end{pmatrix},$$

and put $B_T = \mathrm{Ad}(\xi^{-1})B_A$. Then B_T is a Borel subgroup of $G(\mathbb{C})$ containing T.

Consider the Langlands parameter $\varphi_G : W_{\mathbb{R}} \to \hat{G}$ given by $\varphi_G(\tau) = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) \times 1$, and

$$\varphi_G(z) = \mathrm{diag}(z^n, \bar{z}^n) \times z = z^{\mu}\bar{z}^{\nu} \times z,$$

where μ corresponds to $n \in X_*(\hat{T}) \xrightarrow{\sim} X^*(T)$ and ν corresponds to $-n$. The corresponding representation E of $G(\mathbb{C})$ has highest weight $\lambda_B = n-1 \in X^*(T)$. It is the $(n-1)$-st symmetric power of the standard representation. Its central character is $\lambda_E(z) = z^{n-1}$, where $z = \pm 1$.

We put $\pi_G = \pi(\varphi_G, B_T)$, in the notation from Section 2.2. Write π'_G for the other discrete series representation in Π_E. Thus the L-packet determined by φ_G is

$$\Pi_E = \{\pi_G, \pi'_G\}.$$

We will put $f_\infty dg_\infty = e_{\pi_G}$ as in Section 2.4.

7.1. ***Main term.*** First we consider the terms $ST_g(f\,dg, \pm 1)$.

We have $S\Phi_G(1, e_{\pi_G}) = -n\bar{v}(G; dg_\infty)^{-1}$, and so

$$ST_g(f\,dg, \pm 1, G) = (\pm 1)^n n\bar{v}(G; dg_\infty)^{-1} f^\infty(\pm 1).$$

We have $S\Phi_A(1, e_{\pi_G}) = -\bar{v}(G; dg_\infty)^{-1}$, and so

$$ST_g(f\,dg, \pm 1, A) = (\pm 1)^n \tfrac{1}{2}\bar{v}(G; dg_\infty)^{-1} f_A^\infty(\pm 1).$$

If γ is a regular semisimple element of $G(\mathbb{C})$ with eigenvalues α and α^{-1}, then according to the Weyl character formula,

$$\mathrm{tr}(\gamma; E) = \frac{\alpha^n - \alpha^{-n}}{\alpha - \alpha^{-1}}.$$

Define $t_4(n) = \mathrm{tr}(\mathrm{diag}(i, -i); E)$, where i is a fourth root of unity. Then $t_4(n) = 0$ if n is even, and $t_4(n) = (-1)^{(n-1)/2}$ if n is odd.

Similarly, define $t_3(n) = \mathrm{tr}(\mathrm{diag}(\zeta, \zeta^2); E)$, where ζ is a third root of unity. Then $t_3(n) = [0, 1, -1; 3]_n$, meaning that

$$t_3(n) = \begin{cases} 0 & \text{if } n \equiv 0, \\ 1 & \text{if } n \equiv 1, \\ -1 & \text{if } n \equiv 2. \end{cases}$$

Here the congruence is modulo 3.

There are three stable conjugacy classes of elliptic $\gamma \in G(\mathbb{Q})$, which we represent by

$$\gamma_3 = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}, \quad \gamma_4 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \gamma_6 = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}.$$

Note that $-\gamma_4 \sim \gamma_4$, $\gamma_6^2 = \gamma_3$, and $-\gamma_3 \sim \gamma_6$.

Write T_3 for the elliptic torus consisting of elements

$$\begin{pmatrix} a & a-b \\ b-a & b \end{pmatrix}, \quad \text{with } a^2 - ab + b^2 = 1.$$

We have $S\Phi_G(\gamma_3, e_{\pi_G}) = -\bar{v}(T_3)^{-1} t_3(n)$, and so

$$ST_g(f\,dg, \gamma_3, G) = -\bar{v}(T_3)^{-1}\, SO_{\gamma_3}(f^\infty dg_f) t_3(n).$$

We have $S\Phi_G(\gamma_4, e_{\pi_G}) = -\bar{v}(T)^{-1} t_4(n)$, and so

$$ST_g(f\,dg, \gamma_4, G) = -\bar{v}(T)^{-1}\, SO_{\gamma_4}(f^\infty dg_f) t_4(n).$$

Finally $S\Phi_G(\gamma_6, e_{\pi_G}) = -\bar{v}(T_3) t_3(n)(-1)^{n-1}$, and so

$$ST_g(f\,dg, \gamma_6, G) = -\bar{v}(T_3)^{-1}\, SO_{-\gamma_3}(f^\infty dg_f) t_3(n)(-1)^{n-1}.$$

Thus, $ST_g(f\,dg)$ is equal to the sum

$$-n\overline{v}(G;dg_\infty)^{-1}f^\infty(1)+n\overline{v}(G;dg_\infty)^{-1}f^\infty(-1)(-1)^n-\tfrac{1}{2}\overline{v}(A;da_\infty)^{-1}f_A^\infty(1)$$
$$+\tfrac{1}{2}\overline{v}(A;da_\infty)^{-1}f_A^\infty(-1)(-1)^n-\overline{v}(T_3)^{-1}SO_{\gamma_3}(f^\infty dg_f)t_3(n)$$
$$-\overline{v}(T)^{-1}SO_{\gamma_4}(f^\infty dg_f)t_4(n)+\overline{v}(T_3)^{-1}SO_{-\gamma_3}(f^\infty dg_f)t_3(n)(-1)^n.$$

7.2. *Endoscopic terms.*

Definition 20. Let E be an imaginary quadratic extension of $\mathbb{Q}$. Write H_E for the kernel of the norm map $\mathrm{Res}^E_{\mathbb{Q}}\,\mathbb{G}_m\to\mathbb{G}_m$.

The H_E comprise the (proper) elliptic endoscopic groups for $G=\mathrm{SL}_2$. For each $H=H_E$ we have $\tau(H)=2$ and $|\mathrm{Out}(H,s,\eta)|=1$; see [Kottwitz 1984, Section 7]. Therefore $\iota(G,H)=\frac{1}{2}$. The character identities of Shelstad [1982] give $e^H_{\pi_G}=e_{\chi_n}+e_{\chi_n^{-1}}$.

Write $f^H dh=f^{\infty H}dh_f e^H_{\pi_G}$, where $f^{\infty H}dh_f$ is the transfer of $f^\infty dg_f$. Choose dh_∞ so that $dh_f dh_\infty$ is the Tamagawa measure on H. Then we obtain

$$ST_g(f^H dh)=2\overline{v}(H;dh_\infty)\sum_{\gamma_H}f^{\infty,H}(\gamma_H)\,\mathrm{Tr}^E_{\mathbb{Q}}(\gamma_H^n),$$

the sum being taken over $\gamma_H\in H(\mathbb{Q})$.

Remark. Consider the local transfer, where $f_p dg_p$ is a spherical (that is, invariant under $G(\mathbb{Z}_p)$) measure on $G(\mathbb{Q}_p)$. Then if H ramifies over p, a representation π_p in one of the L-packets transferring from H will also be ramified. This means that $\mathrm{tr}\,\pi_p(f_p dg_p)=0$. So we take $f_p^H=0$ in this case. Thus

$$\mathcal{H}(f\,dg)=ST_g(f\,dg);$$

there is no (proper) endoscopic contribution. This is compatible with the fact that m_{disc} is constant on L-packets in this case.

7.3. *Case of* $\Gamma=\mathrm{SL}_2(\mathbb{Z})$.

We take $K_f=K_0$ to be the integral points of $G(\mathbb{A}_f)$. Also let $K_A=K_0\cap A(\mathbb{A}_f)$ and $K_T=K_0\cap T(\mathbb{A}_f)$. Each of these breaks into a product of local groups $K_{0,p}$, etc.

We put $f^\infty dg_f=e_{K_0}$. Note that $f^\infty(g)=f^\infty(-g)$ for all $g\in G(\mathbb{A}_f)$ and $f_A^\infty(a)=f_A^\infty(-a)$ for all $a\in A(\mathbb{A}_f)$. Therefore, if n is even, then $ST_g(f\,dg)=0$. So assume henceforth that n is odd. Then our expression is equal to

$$-2n\overline{v}(G;dg_\infty)^{-1}f^\infty(1)-\overline{v}(A;da_\infty)^{-1}f_A^\infty(1)$$
$$-2\overline{v}(T_3)^{-1}SO_{\gamma_3}(f^\infty dg_f)t_3(n)+\overline{v}(T)^{-1}SO_{\gamma_4}(f^\infty dg_f)(-1)^{(n+1)/2}.$$

We have

$$\begin{aligned} -2n\bar{v}(G; dg_\infty)^{-1} f^\infty(1) &= -2n\bar{v}(G; dg_\infty)^{-1} \operatorname{vol}_{dg_f}(K_0)^{-1} \\ &= -2n\tau(G)^{-1} d(G)^{-1} \chi_{K_0}(G) = \tfrac{1}{12} n, \\ -\bar{v}(A; da_\infty)^{-1} f_A^\infty(1) &= -\bar{v}(A; da_\infty)^{-1} \operatorname{vol}_{da_f}(K_A)^{-1} \\ &= -\tau(A)^{-1} d(A)^{-1} \chi_{K_A}(A) = -\tfrac{1}{2}. \end{aligned}$$

Now we consider $SO_{\gamma_4}(f^\infty dg_f; dt_f)$. We have $1 - \alpha(\gamma_4) = 2$ for the positive root α of G. Therefore by Proposition 1, the local orbital integrals are equal to $\operatorname{vol}_{dt_p}(K_{T,2})^{-1}$ for $p \neq 2$. At $p = 2$, one has two stable conjugacy classes γ_4 and γ_4' in the conjugacy class of γ_4, where $\gamma_4' = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$.

It follows that

$$SO_{\gamma_4}(f^\infty dg_f; dt_f) = \big(O_{\gamma_4}(e_{K_2}; dt_2) + O_{\gamma_4'}(e_{K_2}; dt_2)\big) \prod_{p \neq 2} \operatorname{vol}_{dt_p}(T(\mathbb{Q}_p) \cap K_p)^{-1}.$$

To compute the local integral at $p = 2$, we reduce to a GL_2-computation by the following lemma. Its proof is straightforward.

Lemma 7. *Let F be a p-adic local field with ring of integers $\mathbb{O}$. Put $G = \mathrm{SL}_2$, $\tilde{G} = \mathrm{GL}_2$, and Z for the center of $\tilde{G}$. Pick Haar measures dg on $G(F)$, $d\tilde{g}$ on $\tilde{G}(F)$, and dz on $Z(F)$. Let $f \in C_c(Z(F)\backslash\tilde{G}(F))$. Then*

$$\frac{\operatorname{vol}_{dz}(Z(\mathbb{O}))}{\operatorname{vol}_{d\tilde{g}}(\tilde{G}(\mathbb{O}))} \int_{Z(F)\backslash\tilde{G}(F)} f(g) \frac{d\tilde{g}}{dz} = \operatorname{vol}_{dg}(G(\mathbb{O}))^{-1} |\mathbb{O}^\times / \mathbb{O}^{\times 2}|^{-1} \sum_\alpha \int_{G(F)} f(t_\alpha g) dg.$$

Here α runs over the square classes in $F^\times$, and $t_\alpha = \operatorname{diag}(\alpha, 1)$.

Proposition 11. *We have*

$$O_{\gamma_4}(e_{K_2}; dt_2) + O_{\gamma_4'}(e_{K_2}; dt_2) = 2 \operatorname{vol}_{dt_2}(K_{T,2})^{-1}.$$

Proof. Write $\tilde{f}_2$ for the characteristic function of $\mathrm{GL}_2(\mathbb{Z}_2) Z(\mathbb{Q}_2)$. By the lemma,

$$\int_{Z(\mathbb{Q}_2)\backslash \mathrm{GL}_2(\mathbb{Q}_2)} \tilde{f}_2(g^{-1}\gamma_4 g) \frac{d\tilde{g}}{dz} = \operatorname{vol}_{dt_2}(K_{T,2}) |\mathbb{Z}_2^\times / \mathbb{Z}_2^{\times 2}|^{-1} \sum_\alpha O_{\mathrm{Ad}(t_\alpha)(\gamma_4)}(e_{K_0}; dt_2).$$

Here we are normalizing $d\tilde{g}$ and dz so that $\operatorname{vol}_{dz}(Z(\mathbb{Z}_2)) = \operatorname{vol}_{d\tilde{g}}(\mathrm{GL}_2(\mathbb{Z}_2)) = 1$.

In fact, $\mathrm{Ad}(t_\alpha)(\gamma_4)$ is conjugate in $G(\mathbb{Q}_2)$ to γ_4 if and only if α is a norm from $\mathbb{Q}_2(\sqrt{-1})$, and in the contrary case, it is conjugate to γ_4'. It follows that

$$\int_{Z(\mathbb{Q}_2)\backslash \mathrm{GL}_2(\mathbb{Q}_2)} \tilde{f}_2(g^{-1}\gamma_4 g) \frac{d\tilde{g}}{dz} = \big(O_{\gamma_4}(e_{K_2}; dt_2) + O_{\gamma_4'}(e_{K_2}; dt_2)\big) \operatorname{vol}_{dt_2}(K_{T,2}).$$

By an elliptic orbital integral computation in [Kottwitz 2005], the left hand side is equal to 2. □

We conclude that

$$SO_{\gamma_4}(f^\infty dg_f; dt_f) = 2\,\mathrm{vol}_{dt_f}(T(\mathbb{A}_f) \cap K_0)^{-1},$$

and so

$$\begin{aligned} -\bar{v}(T)^{-1}\, SO_{\gamma_4}(f^\infty dg_f)t_4(n) &= -2\bar{v}(T)^{-1}\,\mathrm{vol}_{dt_f}(T(\mathbb{A}_f) \cap K_0)^{-1} t_4(n) \\ &= -2\tau(T)^{-1}\chi_{K_T}(T)t_4(n) = 2^{-2}(-1)^{(n+1)/2}. \end{aligned}$$

Similarly, we find that

$$SO_{\gamma_3}(f^\infty dg_f) = 2\,\mathrm{vol}_{dt_{3,f}}(T_3(\mathbb{A}_f) \cap K_0)^{-1},$$

and so

$$-2\bar{v}(T_3)^{-1}\, SO_{\gamma_3}(f^\infty dg_f)t_3(n) = -3^{-1}t_3(n).$$

We conclude that in this case,

$$ST_g(f\,dg) = \tfrac{1}{12}n - \tfrac{1}{2} + \tfrac{1}{4}(-1)^{(n+1)/2} - \tfrac{1}{3}t_3(n).$$

Note that for $n > 1$ this agrees precisely with the discrete series multiplicities. For $n = 1$, this expression is equal to -1, but of course in this case π is not regular.

8. Real tori

We have finished our discussion of SL_2. Starting with this section, we begin to work out the example of GSp_4. Various isomorphisms of tori must be written carefully, so we begin by explicitly working out their parametrizations.

8.1. ***The real tori $\mathbb{G}_m$, S, and T_1.*** We identify the group of characters of $\mathbb{G}_m$ with $\mathbb{Z}$ in the usual way, via $(a \mapsto a^n) \leftrightarrow n$.

Let $A_0 = \mathbb{G}_m \times \mathbb{G}_m$, viewed as a maximal torus in GL_2 in the usual way. Via the identification above we obtain $X^*(A_0) \cong \mathbb{Z}^2$ and $X_*(A_0) \cong \mathbb{Z}^2$.

Let $S = \mathrm{Res}^{\mathbb{C}}_{\mathbb{R}}\mathbb{G}_m$. Recall that $\mathrm{Res}^{\mathbb{C}}_{\mathbb{R}}\mathbb{G}_m$ denotes the algebraic group over $\mathbb{R}$ whose $\mathscr{A}$-points are $(\mathscr{A}\otimes\mathbb{C})^\times$ for an $\mathbb{R}$-algebra $\mathscr{A}$. By choosing the basis $\{1, i\}$ of $\mathbb{C}$ over $\mathbb{R}$, we have an injection $(\mathscr{A}\times\mathbb{C})^\times \to \mathrm{GL}(\mathscr{A}\otimes\mathbb{C}) \cong \mathrm{GL}_2(\mathscr{A})$. Thus we have an embedding $\iota_S : S \to \mathrm{GL}_2$ as an elliptic maximal torus.

There is a ring isomorphism $\varphi : \mathbb{C}\otimes\mathbb{C} \xrightarrow{\sim} \mathbb{C}\times\mathbb{C}$ such that $\varphi(z_1\otimes z_2) = (z_1 z_2, z_1\bar{z}_2)$, which restricts to an isomorphism $\varphi : S(\mathbb{C}) \xrightarrow{\sim} \mathbb{G}_m(\mathbb{C}) \times \mathbb{G}_m(\mathbb{C})$. This isomorphism is also actualized by conjugation within $\mathrm{GL}_2(\mathbb{C})$. Fix $x \in \mathrm{GL}_2(\mathbb{C})$ so that

$$\mathrm{Ad}(x)\begin{pmatrix} a & -b \\ b & a \end{pmatrix} = \begin{pmatrix} a+ib & \\ & a-ib \end{pmatrix};$$

then $\mathrm{Ad}(x) : S(\mathbb{C}) \xrightarrow{\sim} A_0(\mathbb{C})$ is identical to φ, viewing these two tori under the embeddings above.

We fix the isomorphism from $\mathbb{Z}^2$ to $X^*(S)$ that sends $(1, 0)$ and $(0, 1)$ to the character φ composed with projection to the first and, respectively, second component of $\mathbb{G}_m \times \mathbb{G}_m$. Similarly we fix the isomorphism from $\mathbb{Z}^2$ to $X_*(S)$ that sends $(1, 0)$ and $(0, 1)$ to the cocharacters $a \mapsto \varphi^{-1}(a, 1)$ and $a \mapsto \varphi^{-1}(1, a)$, respectively.

Write $\hat{S}$ for the Langlands dual torus to S. It is isomorphic to $\mathbb{C}^\times \times \mathbb{C}^\times$ as a group, with $\Gamma_{\mathbb{R}}$-action defined by $\sigma(\alpha, \beta) = (\beta, \alpha)$. We fix the isomorphism $X^*(S) \xrightarrow{\sim} X_*(\hat{S})$ given by $(a, b) \mapsto (z \mapsto (z^a, z^b))$.

We have an inclusion $\iota_S : \mathbb{G}_m \to S$ given on $\mathscr{A}$-points by $a \mapsto a \otimes 1$. Write σ_S for the automorphism of S given by $1 \otimes \sigma$ on $\mathscr{A}$-points. Note that the fixed point set of σ_S is precisely the image of ι_S.

Write $\mathrm{Nm} : S \to \mathbb{G}_m$ for the norm map given by $s \mapsto s \cdot \sigma_S(s)$. Note that the product $s \cdot \sigma_S(s)$ is in $\iota_S(\mathbb{G}_m)$, which we identify here with $\mathbb{G}_m$. One computes that the norm map induces the map $n \mapsto (n, n)$ from $X^*(\mathbb{G}_m)$ to $X^*(S)$ with the identifications above.

Write T_1 for the kernel of this norm map. Its group of characters fits into the exact sequence

$$0 \to X^*(\mathbb{G}_m) \to X^*(S) \to X^*(T_1) \to 0.$$

We identify $X^*(T_1)$ with $\mathbb{Z}$ so that the restriction map $X^*(S) \to X^*(T_1)$ is given by $(a, b) \mapsto a - b$. The corresponding map $\hat{S} \to \hat{T}$ is given by $(\alpha, \beta) \mapsto \alpha\beta^{-1}$.

8.2. *The kernel and cokernel tori.*

Definition 21. We define A_{ker} to be the kernel of the map from ${\mathbb{G}_m}^4 \to \mathbb{G}_m$ given by $(a, b, c, d) \mapsto (ab)/(cd)$. We define A_{cok} to be the cokernel of the map from $\mathbb{G}_m$ to ${\mathbb{G}_m}^4$ given by $x \mapsto (x, x, x^{-1}, x^{-1})$. Write T_{ker} for the kernel of the map

$$S \times S \to \mathbb{G}_m, \quad (\alpha, \beta) \mapsto \mathrm{Nm}(\alpha/\beta),$$

and T_{cok} for the cokernel of the map

$$\mathbb{G}_m \to S \times S, \quad x \mapsto (\iota_S(x), \iota_S(x^{-1})).$$

Identifying $X_*(\mathbb{G}_m)$ and $X^*(\mathbb{G}_m)$ with $\mathbb{Z}$ as before, we obtain exact sequences

$$0 \to X_*(A_{\mathrm{ker}}) \to \mathbb{Z}^4 \to \mathbb{Z} \to 0,$$
$$0 \to \mathbb{Z} \to \mathbb{Z}^4 \to X^*(A_{\mathrm{ker}}) \to 0,$$
$$0 \to \mathbb{Z} \to \mathbb{Z}^4 \to X_*(A_{\mathrm{cok}}) \to 0,$$
$$0 \to X^*(A_{\mathrm{cok}}) \to \mathbb{Z}^4 \to \mathbb{Z} \to 0.$$

Here the maps from $\mathbb{Z} \to \mathbb{Z}^4$ are both $n \mapsto (n, n, -n, -n)$, and the maps from $\mathbb{Z}^4 \to \mathbb{Z}$ are both $(n_1, n_2, n_3, n_4) \mapsto n_1 + n_2 - n_3 - n_4$.

Thus we obtain isomorphisms

$$g_{\mathrm{kc}} : X^*(A_{\mathrm{ker}}) \xrightarrow{\sim} X_*(A_{\mathrm{cok}}) \quad \text{and} \quad g_{\mathrm{ck}} : X^*(A_{\mathrm{cok}}) \xrightarrow{\sim} X_*(A_{\mathrm{ker}}),$$

obtained from the exact sequences defining A_{ker} and A_{cok}. In this way we view $A_{\mathrm{cok}}(\mathbb{C})$ and $A_{\mathrm{ker}}(\mathbb{C})$ as the dual tori $\hat{A}_{\mathrm{ker}}$ and $\hat{A}_{\mathrm{cok}}$, respectively.

The isomorphism $\varphi \times \varphi : S(\mathbb{C}) \times S(\mathbb{C}) \xrightarrow{\sim} (\mathbb{C}^\times)^4$ gives isomorphisms $\Phi_{\mathrm{ker}} : T_{\mathrm{ker}}(\mathbb{C}) \xrightarrow{\sim} A_{\mathrm{ker}}(\mathbb{C})$ and $\Phi_{\mathrm{cok}} : T_{\mathrm{cok}}(\mathbb{C}) \xrightarrow{\sim} A_{\mathrm{cok}}(\mathbb{C})$.

Consider the map from $S \times S$ to $S \times S$ given by $(a, b) \mapsto (ab, a\sigma_S(b))$. This fits together with the previous maps to form an exact sequence

$$1 \to \mathbb{G}_m \to S \times S \to S \times S \to \mathbb{G}_m \to 1,$$

and yields an isomorphism $\Psi_T : T_{\mathrm{cok}} \xrightarrow{\sim} T_{\mathrm{ker}}$.

Consider the map from ${\mathbb{G}_m}^4$ to ${\mathbb{G}_m}^4$ given by $(a, b, c, d) \mapsto (ac, bd, ad, bc)$. This fits together with the previous maps to form an exact sequence

$$1 \to \mathbb{G}_m \to {\mathbb{G}_m}^4 \to {\mathbb{G}_m}^4 \to \mathbb{G}_m \to 1$$

and yields an isomorphism $\Psi_A : A_{\mathrm{cok}} \xrightarrow{\sim} A_{\mathrm{ker}}$. On $\mathbb{C}$-points we have

$$\Phi_{\mathrm{ker}} \circ \Psi_T(\mathbb{C}) = \Psi_A(\mathbb{C}) \circ \Phi_{\mathrm{cok}}. \tag{8-1}$$

9. Structure of $\mathrm{GSp}_4(F)$

9.1. ***The general symplectic group.*** Let F be a field of characteristic 0. Put

$$J = \begin{pmatrix} & & & 1 \\ & & -1 & \\ & 1 & & \\ -1 & & & \end{pmatrix}.$$

Take G to be the algebraic group $\mathrm{GSp}_4 = \{g \in \mathrm{GL}_4 \mid gJg^t = \mu J, \text{ some } \mu = \mu(g) \in \mathbb{G}_m\}$. It is closely related to the group $G' = \mathrm{Sp}_4 = \{g \in \mathrm{GSp}_4 \mid \mu(g) = 1\}$. Write A for the subgroup of diagonal matrices in G, and Z for the subgroup of scalar matrices in G.

We fix the isomorphism $\iota_A : A_{\mathrm{ker}} \xrightarrow{\sim} A$ given by

$$(a, b, c, d) \mapsto \mathrm{diag}(a, c, d, b). \tag{9-1}$$

Let B_A be the Borel subgroup of upper triangular matrices in G.

9.2. ***Root data.*** Although A and A_{ker} are isomorphic tori, we prefer to parametrize their character and cocharacter groups differently, since the isomorphism ι_A permutes the order of the components.

So we express $X^*(A) = \mathrm{Hom}(A, \mathbb{G}_m)$ as the cokernel of the map

$$i : \mathbb{Z} \to \mathbb{Z}^4, \tag{9-2}$$

given by $i(n) = (n, -n, -n, n)$.

We write $e_1, \ldots, e_4$ for the images in $X^*(A)$ of $(1, 0, 0, 0), \ldots, (0, 0, 0, 1)$. Thus $e_1 + e_4 = e_2 + e_3$. The basis Δ_G of simple roots corresponding to B_A is $\{e_1 - e_2, e_2 - e_3\}$, with corresponding positive roots $\{e_1 - e_2, e_1 - e_4, e_2 - e_3, e_1 - e_3\}$. The half-sum of the positive roots is then $\rho_B = \frac{1}{2}(4e_1 - e_2 - 3e_3) \in X^*(A)$.

Definition 22. Write Ω for the Weyl group of A in G. Write w_0, w_1, w_2 for the elements of Ω that conjugate $\mathrm{diag}(a, b, c, d) \in A$ to

$$\mathrm{diag}(d, c, b, a), \quad \mathrm{diag}(a, c, b, d), \quad \mathrm{diag}(b, a, d, c),$$

respectively.

Ω has order 8 and is generated by w_0, w_1, and w_2.

Express $X_*(A)$ as the kernel of the map

$$p : \mathbb{Z}^4 \to \mathbb{Z}, \quad (a, b, c, d) \mapsto a - b - c + d. \tag{9-3}$$

Let $\vartheta_1 = (1, 0, 0, -1)$ and $\vartheta_2 = (0, 1, -1, 0) \in X_*(A)$. Then the coroots of A in G are given by $R^\vee = R^\vee(A, G) = \{\pm\vartheta_1 \pm \vartheta_2, \pm\vartheta_1, \pm\vartheta_2\}$. The basis $\Delta_G^\vee$ of simple coroots dual to Δ_G is $\{\vartheta_1 - \vartheta_2, \vartheta_2\}$. Then $(X^*(A), \Delta_G, X_*(A), \Delta_G^\vee)$ is a based root datum for G.

9.3. ***The dual group $\hat{G}$.*** We will take $\hat{G}$ to be $\mathrm{GSp}_4(\mathbb{C})$, with trivial L-action, and the same based root data as already discussed for G. The isomorphism

$$X^*(A) \xrightarrow{(\iota_A)^*} X^*(A_{\mathrm{ker}}) \xrightarrow{(\Psi_A)^*} X^*(A_{\mathrm{cok}}) \xrightarrow{g_{ck}} X_*(A_{\mathrm{ker}}) \xrightarrow{(\iota_A)_*} X_*(A) \tag{9-4}$$

(and its inverse) furnish the required isomorphism of based root data. Let us write this out more explicitly. Note that $(\iota_A)_*$ and $(\iota_A)^*$ are given by

$$(\iota_A)_*(a, b, c, d) = (a, c, d, b) \quad \text{and} \quad (\iota_A)^*(a, b, c, d) = (a, d, b, c).$$

The isomorphism in (9-4) is induced from the linear transformation $\Sigma : \mathbb{Z}^4 \to \mathbb{Z}^4$ represented by the matrix

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix},$$

which gives the exact sequence $0 \to \mathbb{Z} \xrightarrow{i} \mathbb{Z}^4 \xrightarrow{\Sigma} \mathbb{Z}^4 \xrightarrow{p} \mathbb{Z} \to 0$, and thus an isomorphism

$$X^*(A) \xrightarrow{\Sigma}{\sim} X_*(A). \tag{9-5}$$

This agrees with the isomorphism used in [Roberts and Schmidt 2007, Section 2.3].

We have $\Sigma(e_1 - e_2) = \vartheta_2$ and $\Sigma(e_2 - e_3) = \vartheta_1 - \vartheta_2$. Thus the based root datum above is self-dual. Note that $\Sigma(\rho) = \frac{3}{2}\vartheta_1 + \frac{1}{2}\vartheta_2$. Write $\hat{A}$ for $A(\mathbb{C})$; it is the torus dual to A via the isomorphism in (9-5).

10. Discrete series for $\mathbf{GSp_4}(\mathbb{R})$

10.1. ***The maximal elliptic torus T of G.*** Consider the map $\mathrm{GL}_2 \times \mathrm{GL}_2 \to \mathrm{GL}_4$ given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \times \begin{pmatrix} e & f \\ g & h \end{pmatrix} \mapsto \begin{pmatrix} a & & & b \\ & e & f & \\ & g & h & \\ c & & & d \end{pmatrix}.$$

The composition of this with the natural inclusion $S \times S \to \mathrm{GL}_2 \times \mathrm{GL}_2$ gives an embedding of $S \times S$ into GL_4. This restricts to an embedding of T_{ker} into G, whose image is an elliptic maximal torus T of G. Thus we have $\iota_T : T_{\mathrm{ker}} \xrightarrow{\sim} T$.

$T(\mathbb{R})$ is the subgroup of matrices of the form

$$\gamma_{r,\theta_1,\theta_2} = \begin{pmatrix} r\cos(\theta_1) & & & -r\sin(\theta_1) \\ & r\cos(\theta_2) & -r\sin(\theta_2) & \\ & r\sin(\theta_2) & r\cos(\theta_2) & \\ r\sin(\theta_1) & & & r\cos(\theta_1) \end{pmatrix} \tag{10-1}$$

for $r > 0$ and angles θ_1, θ_2.

Pick an element $\xi \in G(\mathbb{C})$ so that

$$\mathrm{Ad}(\xi) \begin{pmatrix} a & & & -b \\ & c & -d & \\ & d & c & \\ b & & & a \end{pmatrix} = \begin{pmatrix} a+ib & & & \\ & c+id & & \\ & & c-id & \\ & & & a-ib \end{pmatrix},$$

and put $B_T = \mathrm{Ad}(\xi^{-1})B_A$. Then B_T is a Borel subgroup of $G_{\mathbb{C}}$ containing T, and $\mathrm{Ad}(\xi) : T(\mathbb{C}) \xrightarrow{\sim} A(\mathbb{C})$ is the canonical isomorphism associated to the pairs (T, B_T) and (A, B_A). The definitions have been set up so that

$$\iota_A \circ \Phi_{\mathrm{ker}} = \mathrm{Ad}(\xi) \circ \iota_T.$$

We identify $A(\mathbb{C})$ as the torus dual $\hat{T}$ to T via the isomorphisms

$$X^*(T) \xrightarrow{(\iota_T)^*} X^*(T_{\mathrm{ker}}) \xrightarrow{\Phi^*_{\mathrm{ker}}} X^*(A_{\mathrm{ker}}) \xrightarrow{(\Psi_A)^*} X^*(A_{\mathrm{cok}}) \xrightarrow{g_{ck}} X_*(A_{\mathrm{ker}}) \xrightarrow{(\iota_A)_*} X_*(A). \tag{10-2}$$

10.2. ***Real Weyl group.*** We use $\mathrm{Ad}(\xi)$ to identify Ω with the Weyl group of $T(\mathbb{C})$ in $G(\mathbb{C})$. Recall that $\Omega_{\mathbb{R}}$ denotes the Weyl group of $T(\mathbb{R})$ in $G(\mathbb{R})$. By [Warner 1972, Proposition 1.4.2.1], we have

$$\Omega_{\mathbb{R}} = N_{K_{\mathbb{R}}}(T(\mathbb{R}))/(T(\mathbb{R}) \cap K_{\mathbb{R}}).$$

When discussing maximal compact subgroups of $\mathrm{GSp}_4(\mathbb{R})$, it is convenient to use a different realization of these symplectic groups. Following [Pitale and Schmidt 2009], take for J the symplectic matrix

$$\begin{pmatrix} & & 1 & \\ & & & 1 \\ -1 & & & \\ & -1 & & \end{pmatrix}.$$

Take for $K_{\mathbb{R}}$ the standard maximal compact subgroup of $\mathrm{GSp}_4(\mathbb{R})$ (the intersection of $G(\mathbb{R})$ with the orthogonal group), and $SK_{\mathbb{R}}$ the intersection of $K_{\mathbb{R}}$ with $\mathrm{Sp}_4(\mathbb{R})$. One finds that $SK_{\mathbb{R}}$ is isomorphic to the compact unitary group $U_2(\mathbb{R})$, and yields the Weyl group element w_2. The element $\mathrm{diag}(1, 1, -1, -1) \in N_{G(\mathbb{R})}(T(\mathbb{R})) \cap K_{\mathbb{R}}$ gives $w_0 \in \Omega_{\mathbb{R}}$, and these two elements generate $\Omega_{\mathbb{R}}$. This subgroup has index 2 in Ω, and does not contain the element w_1.

10.3. ***Admissible embeddings.*** Consider the admissible embedding $\eta_B : {}^L T \to {}^L G$. Write $\theta(z) = z/|z|$ for $z \in \mathbb{C}^\times$. We have ${}^L T = \hat{T} \rtimes W_{\mathbb{R}}$, with τ acting as the longest Weyl group element on $\hat{T}$.

Writing ${}^L T = \hat{T} \times W_{\mathbb{R}}$, we put

$$\begin{aligned} &\eta_B(1 \times z) = \mathrm{diag}(\theta(z)^3, \theta(z), \theta(z)^{-1}, \theta(z)^{-3}) \times z && \text{for } z \in \mathbb{C}^\times \cong W_{\mathbb{C}}, \\ &\eta_B(\hat{t} \times 1) = \hat{t} \times 1 && \text{for } \hat{t} \in \hat{T}, \\ &\eta_B(1 \times \tau) = J \times \tau. \end{aligned}$$

10.4. ***Elliptic Langlands parameters.*** Let a, b be odd integers with $a > b > 0$. Let t be an even integer. Put

$$\mu = \tfrac{1}{2}[(t, t, t, t) + (a, b, -b, -a)] \quad \text{and} \quad \nu = \tfrac{1}{2}[(t, t, t, t) + (-a, -b, b, a)],$$

viewed in $X_*(\hat{T})_{\mathbb{C}}$. Then we may define a Langlands parameter $\varphi_G : W_{\mathbb{R}} \to {}^L G$ by

$$\varphi_G(z) = z^{\mu} \bar{z}^{\nu} \times z = |z|^t \,\mathrm{diag}(\theta(z)^a, \theta(z)^b, \theta(z)^{-b}, \theta(z)^{-a}) \times z,$$

and $\varphi_G(\tau) = J \times \tau$.

Note that the centralizer of $\varphi_G(W_{\mathbb{C}})$ in $\hat{G}$ is simply $\hat{A}$, and that $\langle \mu, \alpha \rangle$ is positive for every root of A that is positive for $B_A(\mathbb{C})$. Thus φ_G determines the pair $(\hat{A}, \hat{B}_A)$, where $\hat{B}_A$ is simply $B_A(\mathbb{C})$.

Define a Langlands parameter $\varphi_B : W_{\mathbb{R}} \to {}^L T$ by

$$\varphi_B(z) = |z|^t \operatorname{diag}(\theta(z)^{a-3}, \theta(z)^{b-1}, \theta(z)^{1-b}, \theta(z)^{3-a}) \times z,$$

and $\varphi_B(\tau) = 1 \times \tau$. Then $\varphi_G = \eta_B \circ \varphi_B$.

Let $\pi_G = \pi(\varphi_G, B_T)$ and $\pi'_G = \pi(\varphi_G, w_1(B_T))$, with notation from Section 2.2. The L-packet determined by φ_G is $\Pi = \{\pi_G, \pi'_G\}$. Here π_G is called a holomorphic discrete series representation, and π'_G is called a large discrete series representation.

The highest weight for the associated representation E of $G(\mathbb{C})$ is

$$\lambda_B = \tfrac{1}{2}(a+b-4, t-b+1, t-a+3, 0) \in X^*(A).$$

From this we may read off the central character $\lambda_E(zI) = z^t$ for $zI \in A_G(\mathbb{C})$.

11. The elliptic endoscopic group H

11.1. ***Root data.*** Let H be the cokernel of the map $\mathbb{G}_m \to \mathrm{GL}_2 \times \mathrm{GL}_2$ given by $t \mapsto tI \times t^{-1}I$. Write A^H for the diagonal matrices in H, and B_H for the pairs of upper triangular matrices in H. Fix $\iota_{A^H} : A_{\mathrm{cok}} \xrightarrow{\sim} A^H$ given by

$$(a, b, c, d) \mapsto \operatorname{diag}(a, b) \times \operatorname{diag}(d, c).$$

Write T_H for the image of $S \times S$ in H. It is an elliptic maximal torus in H. Fix $\iota_{T_H} : T_{\mathrm{cok}} \xrightarrow{\sim} T_H$ obtained from the map $S \times S \to \mathrm{GL}_2 \times \mathrm{GL}_2$, $\alpha \mapsto (\iota_S(\alpha), \iota_S(\alpha))$. Put $B_{T_H} = \mathrm{Ad}(x \times x)^{-1} B_H$, a Borel subgroup of $H_{\mathbb{C}}$ containing T_H. Then $\mathrm{Ad}(x \times x)$ is the canonical isomorphism $T_H(\mathbb{C}) \xrightarrow{\sim} A^H(\mathbb{C})$ associated to the pairs (T_H, B_{T_H}) and (A^H, B_H). We view $X^*(T_H)$ as the kernel of the map $p : \mathbb{Z}^2 \times \mathbb{Z}^2 \to \mathbb{Z}$ given by $(a, b) \times (c, d) \mapsto a + b - c - d$. We have a basis of roots Δ_H given by

$$\Delta_H = \{(1, -1) \times (0, 0), (0, 0) \times (1, -1)\}, \tag{11-1}$$

and $\rho_H = \frac{1}{2}(1, -1) \times \frac{1}{2}(1, -1)$.

Furthermore, $X_*(T_H)$ is the cokernel of the map $\iota : \mathbb{Z} \to \mathbb{Z}^2 \times \mathbb{Z}^2$ given by $a \mapsto (a, a) \times (-a, -a)$. We have a basis of coroots $\Delta_H^\vee$ given by

$$\Delta_H^\vee = \{(1, -1) \times (0, 0), (0, 0) \times (1, -1)\}, \tag{11-2}$$

viewed in the quotient $X_*(T_H)$.

11.2. ***Dual group $\hat{H}$.*** Let $\hat{H} = \{(g, h) \in \mathrm{GL}_2(\mathbb{C}) \times \mathrm{GL}_2(\mathbb{C}) \mid \det(g) = \det(h)\}$. We have an inclusion $A_{\mathrm{ker}}(\mathbb{C}) \to \hat{H}$ given by

$$(a, b, c, d) \mapsto \mathrm{diag}(a, b) \times \mathrm{diag}(d, c).$$

Write $\hat{A}^H \subset \hat{H}$ for the image. We thus have an isomorphism $\iota_{\hat{A}^H} : A_{\mathrm{ker}}(\mathbb{C}) \xrightarrow{\sim} \hat{A}^H$.

Also write $\hat{B}_H$ for the subgroup of upper triangular matrices in $\hat{H}$. This Borel subgroup determines a based root datum for $\hat{H}$.

Giving $\hat{H}$ the trivial L-action, we view it as a dual group to H via the isomorphisms

$$X^*(A^H) \xrightarrow{(\iota_{A^H})^*} X^*(A_{\mathrm{cok}}) \xrightarrow{g_{\mathrm{ck}}} X_*(A_{\mathrm{ker}}) \xrightarrow{(\iota_{\hat{A}^H})_*} X_*(\hat{A}^H),$$
$$X^*(\hat{A}^H) \xrightarrow{(\iota_{\hat{A}^H})^*} X^*(A_{\mathrm{ker}}) \xrightarrow{g_{\mathrm{kc}}} X_*(A_{\mathrm{cok}}) \xrightarrow{(\iota_{A^H})_*} X_*(A^H).$$

We identify $\hat{A}^H$ as the torus $\hat{T}_H$ dual to T_H via the isomorphisms

$$X^*(T_H) \xrightarrow{(\iota_{T_H})^*} X^*(T_{\mathrm{cok}}) \xrightarrow{\Phi^*_{\mathrm{cok}}} X^*(A_{\mathrm{cok}}) \xrightarrow{g_{ck}} X_*(A_{\mathrm{ker}}) \xrightarrow{(\iota_{\hat{A}^H})^*} X_*(\hat{A}^H). \tag{11-3}$$

Let $\eta : {}^L H \to {}^L G$ be given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \times \begin{pmatrix} e & f \\ g & h \end{pmatrix} \times w \mapsto \begin{pmatrix} a & & & b \\ & e & f & \\ & g & h & \\ c & & & d \end{pmatrix} \times w. \tag{11-4}$$

Let $s = \mathrm{diag}(1, 1) \times \mathrm{diag}(-1, -1) \in \hat{H}$.

The image $\eta(\hat{H})$ is the connected centralizer in $\hat{G}$ of $\eta(s)$. Thus, (H, s, η) is an elliptic endoscopic triple for G. In fact it is the only one, up to isomorphism.

Moreover note that η restricted to $\hat{A}^H$ is given by

$$\eta|_{\hat{A}^H} = \iota_A \circ (\iota_{\hat{A}^H})^{-1}. \tag{11-5}$$

(Recall that $\hat{A} = A(\mathbb{C})$.)

12. Transfer for $H(\mathbb{R})$

The goal of this section is Proposition 12, in which we identify $e^H_{\pi_G}$ and $e^H_{\pi'_G}$. This is part of the global transfer $f^H dh$ that is to be entered into ST_g for the endoscopic group H. We will recognize it using the character theory of transfer reviewed in Section 3.

12.1. ***Parametrization of discrete series.*** First we must set up the Langlands parameters for discrete series representations of $H(\mathbb{R})$, and describe how they transfer to L-packets in $G(\mathbb{R})$. Recall that we have fixed three integers a, b, t, with a, b odd,

t even, and $a > b > 0$. Define the Langlands parameter $\varphi_H : W_{\mathbb{R}} \to {}^L H = \hat{H} \times W_{\mathbb{R}}$ by

$$\varphi_H(z) = |z|^t \operatorname{diag}(\theta(z)^a, \theta(z)^{-a}) \times |z|^t \operatorname{diag}(\theta(z)^b, \theta(z)^{-b}) \times z$$

for $z \in W_{\mathbb{C}}$, and

$$\varphi_H(\tau) = \left(\begin{smallmatrix} & 1 \\ -1 & \end{smallmatrix}\right) \times \left(\begin{smallmatrix} & -1 \\ 1 & \end{smallmatrix}\right) \times \tau.$$

Then φ_H determines the pair $(\hat{A}_H, \hat{B}_H)$. The L-packet is a singleton $\{\pi_H\}$. The corresponding representation E_H of $H(\mathbb{C})$ has highest weight

$$\lambda_H = \tfrac{1}{2}(t + a - 1, t - a + 1) \times \tfrac{1}{2}(t + b - 1, t - b + 1)$$

and central character $\lambda_{E_H}(z_1, z_2) = (z_1 z_2)^t$. Most importantly, we have $\varphi_G = \eta \circ \varphi_H$.

There is another Langlands parameter φ'_H given by

$$\varphi'_H(z) = |z|^t \operatorname{diag}(\theta(z)^b, \theta(z)^{-b}) \times |z|^t \operatorname{diag}(\theta(z)^a, \theta(z)^{-a}) \times z,$$

and by $\varphi'_H(\tau) = \varphi_H(\tau)$ as above.

Again the L-packet is a singleton $\{\pi'_H\}$. The corresponding representation E'_H has highest weight

$$\lambda'_H = \tfrac{1}{2}(t + b - 1, t - b + 1) \times \tfrac{1}{2}(t + a - 1, t - a + 1),$$

and central character $\lambda_{E'_H} = \lambda_{E_H}$ above.

Let $\varphi'_G = \eta \circ \varphi'_H$. Then $\varphi'_G = \operatorname{Int}(w_2) \circ \varphi_G$, so it is equivalent to φ_G. In particular, both L-packets $\{\pi_H\}$ and $\{\pi'_H\}$ transfer to $\Pi = \{\pi_G, \pi'_G\}$.

12.2. ***Alignment.*** Recall the definition of alignment from Section 3.

Lemma 8. *Define $j : T_H \xrightarrow{\sim} T$ by $j = \iota_T \circ \Psi_T \circ (\iota_{T_H})^{-1}$. Then (j, B_T, B_{T_H}) is aligned with φ_H, and $(j, w_1 B_T, B_{T_H})$ is aligned with φ'_H.*

Proof. Since the parameter φ_G gives the pair $(\hat{A}, \hat{B})$, the parameter φ'_G gives the pair $(\hat{A}, w_1 \hat{B})$, and because φ_H and φ'_H both give $(\hat{A}, \hat{B}_H)$, the horizontal maps in (3-2) are identities. The map $\hat{j} : \hat{T} \to \hat{T}_H$ may be computed by composing the isomorphism $X_*(\hat{T}) \xrightarrow{\sim} X^*(T)$ in (10-2) with the induced map $j^* : X^*(T) \xrightarrow{\sim} X^*(T_H)$ and finally with the inverse of the isomorphism $X_*(\hat{T}_H) \xrightarrow{\sim} X^*(T_H)$ in (11-3). Using equations (8-1) and (11-5), one finds that $\hat{j} = \iota_{\hat{A}_H} \circ (\iota_A)^{-1} = \eta^{-1}$, as desired. □

12.3. ***Transfer for $H_{\mathbb{R}}$.***

Proposition 12. *Let $\pi_G = \pi(\varphi_G, B_T)$ and $\pi'_G = \pi(\varphi_G, \omega^{-1}(B_T))$ as described in Section 10.4. Then (using notation from Section 2.4) we may take $e^H_{\pi_G} = e_{\pi_H} + e_{\pi'_H}$, where π_H and π'_H are the discrete series representation determined by φ_H and φ'_H, respectively, as above. Furthermore, we may take $e^H_{\pi'_G} = -e^H_{\pi_G}$.*

Proof. By Lemma 8, we may use

$$\Delta_\infty(\varphi_H, \pi(\varphi_G, \omega^{-1}(B_T))) = \langle a_\omega, \hat{j}^{-1}(s)\rangle,$$
$$\Delta_\infty(\varphi'_H, \pi(\varphi_G, \omega^{-1}(w_1 B_T))) = \langle a_{w_1\omega}, \hat{j}^{-1}(s)\rangle$$

for $\omega \in \Omega$. In both cases, this is given by

$$\langle a_\omega, s\rangle = \begin{cases} 1 & \text{if } \omega \in \Omega_{\mathbb{R}}, \\ -1 & \text{if } \omega \notin \Omega_{\mathbb{R}}. \end{cases}$$

Note that $\langle a_{w_1\omega}, \hat{j}^{-1}(s)\rangle = -\langle a_\omega, \hat{j}^{-1}(s)\rangle$. Therefore the characterization (3-1) becomes, for a general measure $f_\infty dg_\infty$ at the real place,

$$\begin{aligned}\Theta_{\pi_H}(f_\infty^H dh_\infty) &= \sum_{\pi \in \Pi(\varphi_G)} \Delta_\infty(\varphi_H, \pi)\Theta_\pi(f_\infty dg_\infty) \\ &= \Theta_{\pi_G}(f_\infty dg_\infty) - \Theta_{\pi'_G}(f_\infty dg_\infty)\end{aligned}$$

and similarly

$$\Theta_{\pi'_H}(f_\infty^H dh_\infty) = \Theta_{\pi_G}(f_\infty dg_\infty) - \Theta_{\pi'_G}(f_\infty dg_\infty).$$

In our case, we obtain

$$\Theta_{\pi_H}(e^H_{\pi_G}) = \Theta_{\pi'_H}(e^H_{\pi_G}) = (-1)^{q(G)} \quad \text{and} \quad \Theta_{\pi_H}(e^H_{\pi'_G}) = \Theta_{\pi'_H}(e^H_{\pi'_G}) = -(-1)^{q(G)}.$$

The proposition follows. □

13. Levi subgroups

13.1. ***Levi subgroups.*** We give the standard Levi subgroups of G, which are those of the parabolic subgroups containing B_A. We have the group A, the group G itself, and the following two Levi subgroups:

$$M_1 = \{\mathrm{diag}(g, \lambda g) \mid g \in \mathrm{GL}_2,\ \lambda \in \mathbb{G}_m\},$$
$$M_2 = \{\mathrm{diag}(a, g, b) \mid g \in \mathrm{GL}_2,\ a, b \in \mathbb{G}_m,\ \det(g) = ab\}.$$

Note that both M_1 and M_2 are isomorphic to $\mathbb{G}_m \times \mathrm{GL}_2$.

The group H also has four Levi subgroups, namely A^H, the group H itself, the image M_1^H of $\mathrm{GL}_2 \times A_0$ in H, and the image M_2^H of $A_0 \times \mathrm{GL}_2$ in H. Note that both M_1^H and M_2^H are isomorphic to $\mathrm{GL}_2 \times \mathbb{G}_m$.

13.2. ***Miscellaneous constants.*** We now compute the invariants from Section 5.1 for the Levi subgroups of G and H.

First, we compute the various $k(M)$. When M is the split torus A its derived group is trivial and so $k(A) = 1$. For $i = 1, 2$, the Levi subgroup M_i is isomorphic

to $\mathrm{GL}_2 \times \mathbb{G}_m$, and the torus is isomorphic to $S \times \mathbb{G}_m$. Since S and $\mathbb{G}_m$ have trivial first cohomology, again $k(M_1) = 1$.

Lemma 9. *We have* $k(G) = 2$.

Write T as before for the elliptic torus of G.

Proof. Recall that T_1 is the kernel of Nm and $H^1(\mathbb{R}, T_1)$ has order 2.

Recall that the torus T is isomorphic to the kernel of the map

$$S \times S \to \mathbb{G}_m, \quad (\alpha, \beta) \mapsto \mathrm{Nm}(\alpha/\beta).$$

Projection to the first (or second) component followed by Nm gives an exact sequence

$$1 \to T_1 \times T_1 \to T \to \mathbb{G}_m \to 1. \tag{13-1}$$

We have that $G_{\mathrm{sc}} = G_{\mathrm{der}}$ and the inclusion $T_{\mathrm{sc}} = G_{\mathrm{der}} \cap T \subset T$ may be identified with the map $T_1 \times T_1 \to T$ in the sequence above. In particular, $H^1(\mathbb{R}, T_{\mathrm{sc}})$ has order 4.

Taking the cohomology of (13-1) gives the exact sequence

$$1 \to \mathbb{R}^\times/\mathbb{R}^{\times 2} \to H^1(\mathbb{R}, T_{\mathrm{sc}}) \to H^1(\mathbb{R}, T) \to 1,$$

from which we conclude that $H^1(\mathbb{R}, T_{\mathrm{sc}}) \to H^1(\mathbb{R}, T)$ is surjective and $H^1(\mathbb{R}, T)$ has order 2. □

One must also compute $k(M_H)$ for Levi subgroups M_H of H. The intermediate Levi subgroups are again isomorphic to $\mathrm{GL}(2) \times \mathbb{G}_m$, and for A_H the derived group is trivial. So $k(M_H) = 1$ for each of these.

Lemma 10. *We have* $k(H) = 1$.

Proof. We have $T = P(S \times S)$, $H_{\mathrm{sc}} = \mathrm{SL}_2 \times \mathrm{SL}_2$, and $T_{\mathrm{sc}} = T_1 \times T_1$. The map $T_{\mathrm{sc}} \to T$ factors through $T_1 \times T_1 \to S \times S$. As above we conclude that $k(H) = 1$. □

Secondly, we compute the Tamagawa numbers. Recall that

$$\tau(G) = |\pi_0(Z(\hat{G})^{\Gamma_{\mathbb{Q}}})| \cdot |\ker^1(\mathbb{Q}, Z(\hat{G}))|^{-1}.$$

Proposition 13. *We have* $\tau(M) = 1$ *for all Levi subgroups of* G *and for all proper Levi subgroups of* H, *and* $\tau(H) = 2$.

Proof. For each of these groups, $Z(\hat{M})$ is either the group $\mathbb{C}^\times$ with trivial $\Gamma_{\mathbb{Q}}$-action, or a product of such groups. By the Chebotarev density theorem, the homomorphism

$$\mathrm{Hom}(\Gamma_{\mathbb{Q}}, \mathbb{C}^\times) \to \prod_v \mathrm{Hom}(\Gamma_{\mathbb{Q}_v}, \mathbb{C}^\times)$$

is injective. So $|\ker^1(\mathbb{Q}, Z(\hat{G}))|$ is trivial for our examples. Computing the component group of each $Z(\hat{M})$ is straightforward. □

The quantities n_M^G are easy to compute using $N_G(M) \subseteq N_G(Z(M))$. If M is a maximal torus, n_M^G is of course the order of the Weyl group. For the intermediate cases, one finds that $n_{M_i}^G = n_{M_i^H}^H = 2$.

If $\gamma = 1$, then $\bar{\iota}^M(\gamma) = 1$ for each M, since each M is connected. Note that for Levi subgroups M of G, all proper Levi subgroups M of H, and all semisimple elements γ in G or H, we have $\bar{\iota}^M(\gamma) = 1$ since in all these cases the derived groups are simply connected.

Finally, we compute $\iota(G, H)$, which we recall is given by

$$\iota(G, H) = \tau(G)\tau(H)^{-1}|\mathrm{Out}(H, s, \eta)|^{-1}.$$

One may compute the order of $\mathrm{Out}(H, s, \eta)$ through [Kottwitz 1984, Section 7.6], which shows that this set is in bijection with $\bigwedge(\eta(s), \rho)$, in the notation of that paper. This last set is represented by $\{1, g\}$, where

$$g = \begin{pmatrix} & 1 & & \\ 1 & & & \\ & & & 1 \\ & & 1 & \end{pmatrix}.$$

The conclusion is that $\iota(G, H) = \frac{1}{4}$.

14. Computing $S\Phi_M$ for Levi subgroups of G

Recall from Proposition 3 the formula

$$\Phi_M(\gamma, \Theta^E) = (-1)^{q(L)}|\Omega_L| \sum_{\omega \in \Omega^{LM}} \varepsilon(\omega)\,\mathrm{tr}(\gamma; V^M_{\omega(\lambda_B+\rho_B)-\rho_B}) \quad \text{for } \gamma \in T_e(\mathbb{R}).$$

In this section, the maximal torus will be conjugate to A, and the character group will be identified with $X^*(A)$. We specify an inner product we use on $X^*(A)_{\mathbb{R}}$ for the Weyl dimension formula (Proposition 4).

Definition 23. The usual dot product gives an inner product $(\cdot, \cdot)$ on $X_*(A)_{\mathbb{R}}$, viewing it as a hypersurface in $\mathbb{R}^4$.

Consider the isomorphism

$$\mathrm{pr} : X^*(A)_{\mathbb{R}} \xrightarrow{\sim} X_*(A)_{\mathbb{R}}$$

given by

$$\mathrm{pr}(a, b, c, d) = (a, b, c, d) - \tfrac{1}{4}(a + d - b - c)(1, -1, -1, 1),$$

and let $\langle \lambda, \mu \rangle = (\mathrm{pr}(\lambda), \mathrm{pr}(\mu))$.

For instance,

$$\mathrm{pr}(\lambda_B) = \tfrac{1}{4}(a+b+t-4, a-b+t-2, -a+b+t+2, -a-b+t+4).$$

It will also be necessary to compute Ω^{LM} for each example. Recall that this is the set of $w \in \Omega$ such that $w^{-1}\alpha > 0$ for positive roots α that are either real or imaginary.

14.1. ***The term Φ_G.*** By (4-1) we have $\Phi_G(\gamma, \Theta^E) = \mathrm{tr}(\gamma; E)$. Using the Weyl dimension formula, we compute

$$S\Phi_G(1, e_{\pi_G}) = -\tfrac{1}{24}ab(a+b)(a-b)\bar{v}(G)^{-1}.$$

14.2. ***The term $S\Phi_{M_1}$.*** Consider the torus T_{M_1} given by

$$\begin{pmatrix} a & b & & \\ -b & a & & \\ & & \lambda a & \lambda b \\ & & -\lambda b & \lambda a \end{pmatrix},$$

with $a^2 + b^2 \neq 0$ and $\lambda \neq 0$. This is an elliptic torus in M_1.

There is one positive real root $e_1 - e_3$ and one positive imaginary root $\alpha_{M_1} = e_1 - e_2$. We have $\Omega^{LM} = \{1, w_1\}$, $q(L) = 1$, and $|\Omega_L| = 2$. This gives

$$\Phi_{M_1}(1, \Theta^E) = (-2)\big(\dim_{\mathbb{C}} V_{\lambda_B}^{M_1} - \dim_{\mathbb{C}} V_{\lambda'_B}^{M_1}\big),$$

where $\lambda'_B = \frac{1}{2}(a+b-4, t-a+1, t-b+3, 0) \in X^*(T)$.

Note that $\langle \alpha_{M_1}, \lambda_B \rangle = \frac{1}{2}(b-1)$. The Weyl dimension formula yields

$$\dim_{\mathbb{C}} V_{\lambda_B}^{M_1} = b \quad \text{and} \quad \dim_{\mathbb{C}} V_{\lambda'_B}^{M_1} = a.$$

Thus

$$S\Phi_{M_1}(1, e_{\pi_G}) = -(b-a)\bar{v}(M_1)^{-1}.$$

14.3. ***The term $S\Phi_{M_2}$.*** Consider the torus T_{M_2} given by

$$\begin{pmatrix} s & & & \\ & a & -b & \\ & b & a & \\ & & & t \end{pmatrix},$$

with $st = a^2 + b^2 \neq 0$. This is an elliptic torus in M_2.

We may conjugate this in $G(\mathbb{C})$ to matrices of the form

$$\gamma = \mathrm{diag}(s, a+ib, a-ib, t)$$

in $A(\mathbb{C})$. Composing the roots of A with this composition, we determine the positive imaginary root $\alpha_{M_2} = e_2 - e_3$. We have $\Omega^{LM} = \{1, w_2\}$.

This gives

$$\Phi_{M_2}(1, \Theta^E) = (-2)\big(\dim_{\mathbb{C}} V_{\lambda_B}^{M_2} - \dim_{\mathbb{C}} V_{\lambda_B''}^{M_2}\big),$$

where $\lambda_B'' = \frac{1}{2}(t-b-1, a+b-2, 0, t-a+3) \in X^*(T)$. Note that

$$\mathrm{pr}(\lambda_B'') = \tfrac{1}{4}(t+a-b-4, t+a+b-2, t-a-b+2, t-a+b+4).$$

The Weyl dimension formula yields

$$\dim_{\mathbb{C}} V_{\lambda_B}^{M_2} = \tfrac{1}{2}(a-b) \quad \text{and} \quad \dim_{\mathbb{C}} V_{\lambda_B''}^{M_2} = \tfrac{1}{2}(a+b),$$

and so

$$S\Phi_{M_2}(1, e_{\pi_G}) = b \cdot \overline{v}(M_2)^{-1}.$$

14.4. ***The term $S\Phi_A$.*** By (4-1), we have $\Phi_A(1, \Theta^E) = (-1)^{q(G)}|\Omega_G| = -8$, and so

$$S\Phi_A(1, e_{\pi_G}) = 4\overline{v}(A)^{-1}.$$

15. Computing $S\Phi_{M_H}$ for Levi subgroups of H

Since $e_{\pi_G}^H = e_{\pi_H} + e_{\pi_H'}$, we have

$$S\Phi_{M_H}(1, e_{\pi_G}^H) = (-1)^{q(G)}(-1)^{\dim(A_{M_H}/A_H)}\overline{v}(M_H)^{-1}\big(\Phi_{M_H}(1, \Theta_{\pi_H}) + \Phi_{M_H}(1, \Theta_{\pi_H'})\big).$$

15.1. ***The term $S\Phi_H(1, e_{\pi_G}^H)$.*** In this case H has the elliptic torus T_H.

From (4-1), we obtain $\Phi_H(1, \Theta_{\pi_H}) = \dim_{\mathbb{C}} E_H$. To apply the dimension formula, we compute for instance $\langle \alpha_1, \lambda_H \rangle = a-1$, $\langle \alpha_2, \lambda_H \rangle = b-1$, and $\langle \alpha_i, \rho_H \rangle = 1$.

We find that

$$\Phi_H(1, \Theta^{E_H}) = \Phi_H(1, \Theta^{E_H'}) = ab.$$

Therefore

$$S\Phi_H(1, e_{\pi_G}^H) = -2\overline{v}(H)^{-1}ab.$$

15.2. ***The term $S\Phi_{A^H}(1, e_{\pi_G}^H)$.*** From (4-1), we obtain

$$\Phi_{A^H}(1, \Theta^{E_H}) = \Phi_{A^H}(1, \Theta^{E_H'}) = 4.$$

Therefore

$$S\Phi_{A^H}(1, e_{\pi_G}^H) = -8\overline{v}(A^H)^{-1}.$$

15.3. ***The terms $S\Phi_{M_H}(1, e^H_{\pi_G})$ for the intermediate Levi subgroups.*** For both $M = M^1_H$ and $M = M^2_H$, we have $\Omega_G = \Omega_L \Omega_M$, and so formula (4-1) becomes simply $\Phi_{M_H}(1, \Theta^{E_H}) = (-2)\dim_{\mathbb{C}} V_{\lambda_H}^{M_H}$ for both of these Levi subgroups.

We obtain

$$\Phi_{M^1_H}(1, \Theta^{E_H}) = \Phi_{M^2_H}(1, \Theta^{E'_H}) = -2a$$

and

$$\Phi_{M^2_H}(1, \Theta^{E_H}) = \Phi_{M^1_H}(1, \Theta^{E'_H}) = -2b.$$

Therefore

$$S\Phi_{M^1_H}(1, e^H_{\pi_G}) = S\Phi_{M^2_H}(1, e^H_{\pi_G}) = -2\overline{v}(M^1_H)^{-1}(a+b).$$

16. Final form: γ central

Recall that $G = \mathrm{GSp}_4$. For the convenience of the reader, we recall the setup.

Let a and b be odd integers with $a > b > 0$, and t an even integer. Consider the Langlands parameter $\varphi_G : W_{\mathbb{R}} \to {}^L G$ given by

$$\varphi_G(z) = |z|^t \operatorname{diag}(\theta(z)^a, \theta(z)^b, \theta(z)^{-b}, \theta(z)^{-a}) \times z \quad \text{and} \quad \varphi_G(\tau) = J \times \tau.$$

Let π_G be the discrete series representation $\pi(\varphi_G, B_T)$ of $G(\mathbb{R})$ as in Section 2.2. Write π'_G for the other representation in $\Pi(\varphi_G)$.

Put $f_\infty dg_\infty = e_{\pi_G}$ as in Section 2.4 for π_G and any measure $f^\infty dg_f$ on $G(\mathbb{A}_f)$. Let $f\,dg = e_{\pi_G} f^\infty dg_f$, a measure on $G(\mathbb{A})$. By the theory of endoscopic transfer there is a matching measure $f^H dh$ on $H(\mathbb{A})$, where H is the elliptic endoscopic group $P(\mathrm{GL}_2 \times \mathrm{GL}_2)$ discussed above.

If $z \in A_G(\mathbb{Q})$, then $\sum_M ST_g(f\,dg, z, M)$ is given by the product of $\lambda_E(z) = z^t$ with

$$-\tfrac{1}{24}ab(a+b)(a-b)\overline{v}(G)^{-1} f^\infty(z) + \tfrac{1}{2}(a-b)\overline{v}(M_1)^{-1} f^\infty_{M_1}(z) + \tfrac{1}{2}b\overline{v}(M_2)^{-1} f^\infty_{M_2}(z) + \tfrac{1}{2}\overline{v}(A)^{-1} f^\infty_A(z).$$

If $z = (z_1, z_2) \in A_H(\mathbb{Q})$, then $\sum_{M_H} ST_g(f^H dh, z, M_H)$ is given by the product of $\lambda_{E_H}(z) = (z_1 z_2)^t$ with

$$-4ab\overline{v}(H)^{-1} f^{H,\infty}(z) - 2(a+b)\overline{v}(M^1_H)^{-1} f^\infty_{M_2}(z) - 2\overline{v}(A^H)^{-1} f^\infty_{A^H}(z).$$

17. The case $\Gamma = \mathbf{Sp_4}(\mathbb{Z})$

Let $f^\infty dg_f = e_{K_0}$, where $K_0 = G(\mathbb{O}_f)$. Here dg_f is an arbitrary Haar measure on $G(\mathbb{A}_f)$, so that $dg = dg_f dg_\infty$ is the Tamagawa measure on $G(\mathbb{A})$.

17.1. ***Central terms in G.*** Note that $f_M^\infty(z) = 0$ for all $z \in Z(\mathbb{Q})$ unless $z = \pm 1$, and that $f_M^\infty(1) = f_M^\infty(-1)$ for all Levi subgroups M.

First we compute $ST_g(fdg, \pm 1, G)$. We have

$$\begin{aligned} -\frac{1}{2^3 3} ab(a+b)(a-b)\bar{v}(G)^{-1} f^\infty(\pm 1) & \\ = -\frac{1}{2^3 3} ab(a+b)(a-b)\tau(G)^{-1} d(G)^{-1} \chi_{K_0}(G) & \\ = 2^{-10} 3^{-3} 5^{-1} ab(a+b)(a-b). & \end{aligned}$$

Next we treat the $\pm 1 \in M_i$ terms, for the intermediate Levi subgroups. We have

$$\begin{aligned} ST_g(fdg, \pm 1, M_1) &= \tfrac{1}{2}(a-b)\bar{v}(M_1)^{-1} f_{M_1}^\infty(\pm 1) = -2^{-5} 3^{-1}(a-b), \\ ST_g(fdg, \pm 1, M_2) &= \tfrac{1}{2} b\bar{v}(M_2)^{-1} f_{M_2}^\infty(\pm 1) = -2^{-5} 3^{-1} b. \end{aligned}$$

Next we treat the $\pm 1 \in A$ terms. We have $f_A(1) = \mathrm{vol}_{da_f}(K \cap A(\mathbb{A}_f))^{-1}$, which is 1. Moreover we take Lebesgue measure on $A(\mathbb{R})$ so that $\bar{v}(A) = 8$. It follows that

$$ST_g(fdg, \pm 1, A) = \tfrac{1}{2}\bar{v}(A)^{-1} f_A^\infty(\pm 1) = 2^{-4}.$$

Doubling these terms to account for both central elements, we compute

$$\begin{aligned} \sum_{z,M} ST_g(fdg, z, M) & \\ = 2^{-9} 3^{-3} 5^{-1} ab(a+b)(a-b) - 2^{-4} 3^{-1}(a-b) - 2^{-4} 3^{-1} b + 2^{-3}. & \end{aligned} \tag{17-1}$$

17.2. ***Central terms in H.*** By the fundamental lemma ([Hales 1997; Weissauer 2009] for GSp_4, and of course [Ngô 2010] in general), we may write $(e_{K_0})^H = e_{K_H}$, where $K_H = H(\mathbb{C}_f)$. Thus $(f^\infty)_M^H(z) = 0$ for all $z \in H(\mathbb{Q})$ unless $z = (1, \pm 1)$, and

$$f_M^{H\infty}(1,1) = f_M^{H\infty}(1,-1)$$

for all Levi subgroups $M = M_H$ of H.

The only nontrivial factors in the formula of Theorem 2 are $|\ker \rho(\mathbb{Q})| = 2$, $[H(\mathbb{R}) : H(\mathbb{R})_+] = 4$, and $\chi_{\mathrm{alg}}(H^{\mathrm{sc}}(\mathbb{Z}))$. Note that $H^{\mathrm{sc}} = \mathrm{SL}_2 \times \mathrm{SL}_2$.

Therefore

$$\chi_{K_H}(H) = 2^{-1} \chi_{\mathrm{alg}}(\mathrm{SL}_2(\mathbb{Z}))^2 = 2^{-5} 3^{-2}.$$

We conclude that

$$ST_g(f^H dh, (1, \pm 1), H) = -4ab\bar{v}(H)^{-1} \mathrm{vol}(K_H)^{-1} = -2^{-4} 3^{-2} ab.$$

Next we find that

$$\sum_{i=1}^{2} ST_g(f^H dh, (1, \pm 1), M_i^H) = -2(a+b)\bar{v}(M_1^H)^{-1} \operatorname{vol}(K_M)^{-1}$$
$$= 2^{-3}3^{-1}(a+b).$$

Finally, we have

$$ST_g(f^H dh, (1, \pm 1), A^H) = -2\bar{v}(A)^{-1} \operatorname{vol}(K_A)^{-1} = -2^{-2}.$$

Multiplying by $\iota(G, H) = 4^{-1}$ and then doubling to account for both central elements, we compute

$$\iota(G, H) \sum_{z, M_H} ST_g(f^H dh, z, M_H) = -2^{-5}3^{-2}ab + 2^{-4}3^{-1}(a+b) - 2^{-3}. \tag{17-2}$$

18. Comparison

As mentioned in the introduction, Wakatsuki [$\geq$ 2012; 2012] has used the Selberg trace formula and Arthur's L^2-Lefschetz number formula to compute the discrete series multiplicities $m_{\text{disc}}(\pi, \Gamma)$ for π both holomorphic and large discrete series representations for $\mathrm{Sp}_4(\mathbb{R})$, and for many cases of arithmetic subgroups Γ. We will compare our formula to his when Γ is the full modular group. (Note that if π is a discrete series representation of $\mathrm{GSp}_4(\mathbb{R})$ with trivial central character, and π_1 is its restriction to $\mathrm{Sp}_4(\mathbb{R})$, then $m_{\text{disc}}(\pi, \Gamma) = m_{\text{disc}}(\pi_1, \Gamma_1)$, where $\Gamma_1 = \mathrm{Sp}_4(\mathbb{Z})$.) Since he is using the Selberg trace formula, his formula breaks into contributions from each conjugacy class in Γ. In particular, he identifies the central-unipotent contributions H_1^{Hol} and H_1^{Large} to $m_{\text{disc}}(\pi_G)$ and $m_{\text{disc}}(\pi'_G)$, respectively. Namely,

$$H_1^{\text{Hol}} = 2^{-9}3^{-3}5^{-1}ab(a-b)(a+b) - 2^{-5}3^{-2}ab + 2^{-4}3^{-1}b,$$
$$H_1^{\text{Large}} = 2^{-9}3^{-3}5^{-1}ab(a-b)(a+b) + 2^{-5}3^{-2}ab - 2^{-3}3^{-1}b + 2^{-2}.$$

(To translate from his notation to ours, use $j = b - 1$ and $k = \frac{1}{2}(a-b) + 2$.)

Comparing these formulas to our formulas above, we observe

$$H_1^{\text{Hol}} = \sum_M ST_g(f dg, \pm 1, M) + \iota(G, H) \sum_{M_H} ST_g(f^H dh, (1, \pm 1), M_H)$$

when $f dg = e_{\pi_G} e_{K_0}$ and

$$H_1^{\text{Large}} = \sum_M ST_g(f dg, \pm 1, M) + \iota(G, H) \sum_{M_H} ST_g(f^H dh, (1, \pm 1), M_H).$$

when $f dg = e_{\pi'_G} e_{K_0}$.

This proves Theorem 1. □

Acknowledgments

This paper is founded on my thesis under the direction of Robert Kottwitz. I would like to thank him for his continual help with this project. I am also indebted to Satoshi Wakatsuki for predicting Theorem 1, and for much useful correspondence. I would also like to thank Ralf Schmidt for helpful conversations.

References

[Arthur 1988a] J. Arthur, "The invariant trace formula, I: Local theory", *J. Amer. Math. Soc.* **1**:2 (1988), 323–383. MR 89e:22029 Zbl 0682.10021

[Arthur 1988b] J. Arthur, "The invariant trace formula, II: Global theory", *J. Amer. Math. Soc.* **1**:3 (1988), 501–554. MR 89j:22039 Zbl 0667.10019

[Arthur 1989] J. Arthur, "The L^2-Lefschetz numbers of Hecke operators", *Invent. Math.* **97**:2 (1989), 257–290. MR 91i:22024 Zbl 0692.22004

[Arthur 2001] J. Arthur, "A stable trace formula, II: Global descent", *Invent. Math.* **143**:1 (2001), 157–220. MR 2002m:11043 Zbl 0978.11025

[Arthur 2002] J. Arthur, "A stable trace formula, I: General expansions", *J. Inst. Math. Jussieu* **1**:2 (2002), 175–277. MR 2004b:11066 Zbl 1040.11038

[Arthur 2003] J. Arthur, "A stable trace formula, III: Proof of the main theorems", *Ann. of Math.* (2) **158**:3 (2003), 769–873. MR 2004m:11079 Zbl 1051.11027

[Arthur 2005] J. Arthur, "An introduction to the trace formula", pp. 1–263 in *Harmonic analysis, the trace formula, and Shimura varieties* (Toronto, 2003), edited by J. Arthur et al., Clay Math. Proc. **4**, Amer. Math. Soc., Providence, RI, 2005. MR 2007d:11058 Zbl 1152.11021

[Borel 1979] A. Borel, "Automorphic L-functions", pp. 27–61 in *Automorphic forms, representations and L-functions* (Corvallis, OR, 1977), vol. 2, edited by A. Borel and W. Casselman, Proc. Sympos. Pure Math. **33**, Amer. Math. Soc., Providence, RI, 1979. MR 81m:10056 Zbl 0412.10017

[Clozel and Delorme 1984] L. Clozel and P. Delorme, "Le théorème de Paley–Wiener invariant pour les groupes de Lie réductifs", *Invent. Math.* **77**:3 (1984), 427–453. MR 86b:22015 Zbl 0584.22005

[Clozel and Delorme 1990] L. Clozel and P. Delorme, "Le théorème de Paley–Wiener invariant pour les groupes de Lie réductifs, II", *Ann. Sci. École Norm. Sup.* (4) **23**:2 (1990), 193–228. MR 91g:22013 Zbl 0724.22012

[Deligne 1979] P. Deligne, "Variétés de Shimura: interprétation modulaire, et techniques de construction de modèles canoniques", pp. 247–289 in *Automorphic forms, representations and L-functions* (Corvallis, OR, 1977), vol. 2, edited by A. Borel and W. Casselman, Proc. Sympos. Pure Math. **33**, Amer. Math. Soc., Providence, RI, 1979. MR 81i:10032 Zbl 0437.14012

[Goresky et al. 1997] M. Goresky, R. E. Kottwitz, and R. MacPherson, "Discrete series characters and the Lefschetz formula for Hecke operators", *Duke Math. J.* **89**:3 (1997), 477–554. MR 99e:11064a Zbl 0888.22011

[Hales 1997] T. C. Hales, "The fundamental lemma for Sp(4)", *Proc. Amer. Math. Soc.* **125**:1 (1997), 301–308. MR 97c:22020 Zbl 0876.22022

[Harder 1971] G. Harder, "A Gauss–Bonnet formula for discrete arithmetically defined groups", *Ann. Sci. École Norm. Sup.* (4) **4**:3 (1971), 409–455. MR 46 #8255 Zbl 0232.20088

[Kottwitz 1983] R. E. Kottwitz, "Sign changes in harmonic analysis on reductive groups", *Trans. Amer. Math. Soc.* **278**:1 (1983), 289–297. MR 84i:22012 Zbl 0538.22010

[Kottwitz 1984] R. E. Kottwitz, "Stable trace formula: Cuspidal tempered terms", *Duke Math. J.* **51**:3 (1984), 611–650. MR 85m:11080 Zbl 0576.22020

[Kottwitz 1986] R. E. Kottwitz, "Stable trace formula: Elliptic singular terms", *Math. Ann.* **275**:3 (1986), 365–399. MR 88d:22027 Zbl 0577.10028

[Kottwitz 1988] R. E. Kottwitz, "Tamagawa numbers", *Ann. of Math.* (2) **127**:3 (1988), 629–646. MR 90e:11075 Zbl 0678.22012

[Kottwitz 1990] R. E. Kottwitz, "Shimura varieties and λ-adic representations", pp. 161–209 in *Automorphic forms, Shimura varieties, and L-functions* (Ann Arbor, MI, 1988), vol. 1, edited by L. Clozel and J. S. Milne, Perspect. Math. **10**, Academic Press, Boston, 1990. MR 92b:11038 Zbl 0743.14019

[Kottwitz 2005] R. E. Kottwitz, "Harmonic analysis on reductive p-adic groups and Lie algebras", pp. 393–522 in *Harmonic analysis, the trace formula, and Shimura varieties* (Toronto, 2003), edited by J. Arthur et al., Clay Math. Proc. **4**, Amer. Math. Soc., Providence, RI, 2005. MR 2006m:22016 Zbl 1106.22013

[Kottwitz $\geq$ 2012] R. E. Kottwitz, "Stable version of Arthur's formula", preprint.

[Kottwitz and Shelstad 1999] R. E. Kottwitz and D. Shelstad, *Foundations of twisted endoscopy*, Astérisque **255**, Sociéte Mathématique de France, Paris, 1999. MR 2000k:22024 Zbl 0958.22013

[Langlands 1976] R. P. Langlands, *On the functional equations satisfied by Eisenstein series*, Lecture Notes in Mathematics **544**, Springer, Berlin, 1976. MR 58 #28319 Zbl 0332.10018

[Langlands 1979] R. P. Langlands, "Stable conjugacy: definitions and lemmas", *Canad. J. Math.* **31**:4 (1979), 700–725. MR 82j:10054 Zbl 0421.12013

[Langlands 1983] R. P. Langlands, *Les débuts d'une formule des traces stable*, Publications Mathématiques **13**, Université de Paris VII, 1983. MR 85d:11058 Zbl 0532.22017

[Langlands and Shelstad 1987] R. P. Langlands and D. Shelstad, "On the definition of transfer factors", *Math. Ann.* **278**:1-4 (1987), 219–271. MR 89c:11172 Zbl 0644.22005

[Milne 2005] J. S. Milne, "Introduction to Shimura varieties", pp. 265–378 in *Harmonic analysis, the trace formula, and Shimura varieties* (Toronto, 2003), edited by J. Arthur et al., Clay Math. Proc. **4**, Amer. Math. Soc., Providence, RI, 2005. MR 2006m:11087 Zbl 1148.14011

[Morel 2010] S. Morel, *On the cohomology of certain noncompact Shimura varieties*, Annals of Mathematics Studies **173**, Princeton University Press, 2010. MR 2011b:11073 Zbl 1233.11069

[Ngô 2010] B. C. Ngô, "Le lemme fondamental pour les algèbres de Lie", *Publ. Math. Inst. Hautes Études Sci.* **111**:1 (2010), 1–169. MR 2011h:22011 Zbl 1200.22011

[Ono 1966] T. Ono, "On Tamagawa numbers", pp. 122–132 in *Algebraic groups and discontinuous subgroups* (Boulder, CO, 1965), edited by A. Borel and G. D. Mostow, Proc. Sympos. Pure Math. **9**, Amer. Math. Soc., Providence, RI, 1966. MR 35 #191 Zbl 0223.20050

[Pitale and Schmidt 2009] A. Pitale and R. Schmidt, "Bessel models for lowest weight representations of GSp(4, $\mathbb{R}$)", *Int. Math. Res. Not.* **2009**:7 (2009), 1159–1212. MR 2010h:11084 Zbl 05553685

[Roberts and Schmidt 2007] B. Roberts and R. Schmidt, *Local newforms for* GSp(4), Lecture Notes in Mathematics **1918**, Springer, Berlin, 2007. MR 2008g:11080 Zbl 1126.11027

[Satake 1980] I. Satake, *Algebraic structures of symmetric domains*, Kanô Memorial Lectures **4**, Iwanami Shoten, Tokyo, 1980. MR 82i:32003 Zbl 0483.32017

[Selberg 1956] A. Selberg, "Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series", *J. Indian Math. Soc.* (*N.S.*) **20** (1956), 47–87. MR 19,531g Zbl 0072.08201

[Serre 1971] J.-P. Serre, "Cohomologie des groupes discrets", pp. 77–169 in *Prospects in mathematics* (Princeton, 1970), edited by F. Hirzebruch et al., Ann. of Math. Studies **70**, Princeton University Press, 1971. MR 52 #5876 Zbl 0235.22020

[Shelstad 1982] D. Shelstad, "L-indistinguishability for real groups", *Math. Ann.* **259**:3 (1982), 385–430. MR 84c:22017 Zbl 0506.22014

[Spallone 2004] S. T. Spallone, *Arthur's trace formula for* SO(5) *and individual discrete series matrix coefficients*, thesis, Univ. Chicago, 2004, available at http://www.math.ou.edu/∼sspallone/Papers/Dissertation.pdf.

[Spallone 2009] S. T. Spallone, "Stable discrete series characters at singular elements", *Canad. J. Math.* **61**:6 (2009), 1375–1382. MR 2011c:22025 Zbl 1180.22019

[Spallone 2011] S. T. Spallone, "On the trace formula approach to discrete series multiplicities", pp. 1–13 in *Automorphic forms, trace formulas and zeta functions* (Kyoto, 2011), edited by Y. Gon and T. Moriyama, RIMS Kokyuroku **1767**, Res. Inst. Math. Sci., Kyoto, 2011.

[Springer 1979] T. A. Springer, "Reductive groups", pp. 3–27 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 80h:20062 Zbl 0416.20034

[Tsushima 1983] R. Tsushima, "An explicit dimension formula for the spaces of generalized automorphic forms with respect to $\mathrm{Sp}(2, \mathbb{Z})$", *Proc. Japan Acad. Ser. A Math. Sci.* **59**:4 (1983), 139–142. MR 85a:11011 Zbl 0513.10025

[Tsushima 1997] R. Tsushima, "Dimension formula for the spaces of Siegel cusp forms and a certain exponential sum", *Mem. Inst. Sci. Tech. Meiji Univ.* **36** (1997), 1–56.

[Wakatsuki 2012] S. Wakatsuki, "Dimension formulas for spaces of vector-valued Siegel cusp forms of degree two", *J. Number Theory* **132**:1 (2012), 200–253. MR 2843308 Zbl 05966274

[Wakatsuki ≥ 2012] S. Wakatsuki, "Multiplicity formulas for discrete series representations in $L^2(\Gamma\backslash \mathrm{Sp}(2, \mathbb{R}))$", preprint.

[Wallach 1984] N. R. Wallach, "On the constant term of a square integrable automorphic form", pp. 227–237 in *Operator algebras and group representations* (Neptun, 1980), vol. 2, edited by G. Arsene, Monogr. Stud. Math. **18**, Pitman, Boston, 1984. MR 86i:22029 Zbl 0554.22004

[Warner 1972] G. Warner, *Harmonic analysis on semi-simple Lie groups, I*, Grundlehren Math. Wiss. **188**, Springer, New York, 1972. MR 58 #16979 Zbl 0265.22020

[Weissauer 2009] R. Weissauer, "A special case of the fundamental lemma, I–IV", pp. 211–320 in *Endoscopy for* GSp(4) *and the cohomology of Siegel modular threefolds*, Lecture Notes in Mathematics **1968**, Springer, Berlin, 2009.

Received April 26, 2011. Revised December 5, 2011.

STEVEN SPALLONE
SCHOOL OF MATHEMATICS
TATA INSTITUTE OF FUNDAMENTAL RESEARCH
HOMI BHABHA RD
COLABA
MUMBAI 400005
INDIA

sspallone@gmail.com

PACIFIC JOURNAL OF MATHEMATICS
Vol. 256, No. 2, 2012

SMALL COVERS AND THE HALPERIN–CARLSSON CONJECTURE

LI YU

We prove that the Halperin–Carlsson conjecture holds for any free $(\mathbb{Z}_2)^m$-action on a compact manifold whose orbit space is a small cover.

1. Introduction

For any prime p, let $\mathbb{Z}_p$ denote the quotient group $\mathbb{Z}/p\mathbb{Z}$, and S^1 the circle group.

The Halperin–Carlsson Conjecture. *If $G = (\mathbb{Z}_p)^m$ or $(S^1)^m$ can act freely on a finite CW-complex X, then, respectively,*

$$\sum_{i=0}^{\infty} \dim_{\mathbb{Z}_p} H^i(X, \mathbb{Z}_p) \geq 2^m \quad \text{or} \quad \sum_{i=0}^{\infty} \dim_{\mathbb{Q}} H^i(X, \mathbb{Q}) \geq 2^m.$$

This was proposed by Halperin [1985] for the torus case and by Carlsson [1986] for the $\mathbb{Z}_p$-torus case. It is also called the *toral rank conjecture* in some papers.

At first this conjecture mainly took the form of whether a free $(\mathbb{Z}_p)^m$-action on a product of spheres $S^{n_1} \times \cdots \times S^{n_k}$ implies $m \leq k$. Many authors have studied this intriguing conjecture in its various aspects [Conner 1957; Carlsson 1982; Adem 1987; Adem and Browder 1988; Adem and Benson 1998; Hanke 2009]. For a survey of results on the topic, see [Adem 2004; Allday and Puppe 1993]. The general case is still open for any prime p.

For general finite CW-complexes, the conjecture was proved in [Puppe 2009] for $m \leq 3$ in the torus and $\mathbb{Z}_2$-torus cases and $m \leq 2$ in the odd $\mathbb{Z}_p$-torus case. Also we have the following result, achieved independently, which confirmed the Halperin–Carlsson conjecture for some special $\mathbb{Z}_2$-torus actions on real moment-angle complexes:

This work was partially supported by the Japan Society for the Promotion of Science (JSPS grant no. P10018) and the Natural Science Foundation of China (grant no. 11001120). This work was also funded by the PAPD (priority academic program development) of Jiangsu higher education institutions.

MSC2010: 57R22, 57R91, 57S17, 57S25.

Keywords: free torus action, Halperin–Carlsson conjecture, small cover, moment-angle manifold, glue-back construction.

Theorem 1.1 [Cao and Lü 2009; Ustinovskii 2011]. *Let K^{n-1} be an $(n-1)$-dimensional simplicial complex on the vertex set $[d]$. Then the real moment-angle complex $\mathbb{R}\mathcal{Z}_{K^{n-1}}$ over K^{n-1} must satisfy $\sum_i \dim_{\mathbb{Z}_2} H^i(\mathbb{R}\mathcal{Z}_{K^{n-1}}, \mathbb{Z}_2) \geq 2^{d-n}$. In particular, if P^n is an n-dimensional simple convex polytope with d facets, then the real moment-angle manifold $\mathbb{R}\mathcal{Z}_{P^n}$ must satisfy $\sum_i \dim_{\mathbb{Z}_2} H^i(\mathbb{R}\mathcal{Z}_{P^n}, \mathbb{Z}_2) \geq 2^{d-n}$.*

Remark 1.2. Stronger results were obtained in [Cao and Lü 2009] and [Ustinovskii 2011]; for example, Theorem 1.1 holds even if the $\mathbb{Z}_2$-coefficients are replaced by rational coefficients.

Remark 1.3. There is a purely algebraic analogue of the Halperin–Carlsson conjecture, which was proposed in [Carlsson 1986] in the context of commutative algebras. Some related results were obtained in [Carlsson 1987].

Here we only study the conjecture for $G = (\mathbb{Z}_2)^m$ and X a closed manifold. We use the following conventions:

- we treat $(\mathbb{Z}_2)^m$ as an additive group;
- all manifolds and submanifolds are smooth;
- we do not distinguish between an embedded submanifold and its image.

Suppose that $(\mathbb{Z}_2)^m$ acts freely and smoothly on a closed n-manifold M^n. Let $Q^n = M^n/(\mathbb{Z}_2)^m$ be the orbit space. Then Q^n is a closed n-manifold too. Let $\pi : M^n \to Q^n$ be the orbit map. We can think of M^n either as a principal $(\mathbb{Z}_2)^m$-bundle over Q^n or as a regular covering over Q^n whose deck transformation group is $(\mathbb{Z}_2)^m$. In algebraic topology, we have a standard way to recover M^n from Q^n, using the universal covering space of Q^n and the monodromy of the covering [Hatcher 2002]. However, it is not so easy for us to visualize the total space of the covering with this approach. In [Yu 2012], a new way of constructing principal $(\mathbb{Z}_2)^m$-bundles over closed manifolds is introduced, which allows us to visualize this kind of object more easily.

Indeed, it is shown in [Yu 2012] that $\pi : M^n \to Q^n$ determines a $(\mathbb{Z}_2)^m$-coloring λ_π on a nice manifold with corners V^n (called a $\mathbb{Z}_2$-*core* of Q^n), and up to equivariant homeomorphism, we can recover M^n by a standard *glue-back construction* from V^n and λ_π. Using this new language, we prove the following theorem, which supports the Halperin–Carlsson conjecture.

Theorem 1.4. *Suppose that $(\mathbb{Z}_2)^m$ acts freely on a closed n-manifold M^n whose orbit space is homeomorphic to a small cover; then*

$$\sum_i \dim_{\mathbb{Z}_2} H^i(M^n, \mathbb{Z}_2) \geq 2^m. \tag{1}$$

Recall that an n-dimensional small cover is a closed n-manifold with a locally

standard $(\mathbb{Z}_2)^n$-action whose orbit space can be identified with an n-dimensional simple convex polytope [Davis and Januszkiewicz 1991].

Given an arbitrary n-dimensional simple convex polytope P^n, there may not exist any small cover over P^n. But we can always define a closed manifold $\mathbb{R}\mathcal{Z}_{P^n}$ associated to P^n called a *real moment-angle manifold* [Davis and Januszkiewicz 1991, Construction 4.1]. Let $\mathcal{F}(P^n) = \{F_1, \ldots, F_r\}$ be the set of facets of P^n, and let $\{e_1, \ldots, e_r\}$ be a basis of $(\mathbb{Z}_2)^r$. For $1 \le i \le r$, we define a function $\lambda^* : \mathcal{F}(P^n) \to (\mathbb{Z}_2)^r$ by

$$\lambda^*(F_i) = e_i. \tag{2}$$

For any proper face f of P^n, let G_f denote the subgroup of $(\mathbb{Z}_2)^r$ generated by the set $\{\lambda^*(F_i) \mid f \subset F_i\}$. The real moment-angle manifold $\mathbb{R}\mathcal{Z}_{P^n}$ of P^n is defined to be the quotient space

$$\mathbb{R}\mathcal{Z}_{P^n} := P^n \times (\mathbb{Z}_2)^r / \sim, \tag{3}$$

where $(p, g) \sim (p', g')$ if and only if $p = p'$ and $g^{-1}g' \in G_{f(p)}$, with $f(p)$ being the unique face of P^n that contains p in its relative interior. Let $[(p, g)]$ denote the equivalence class of (p, g) in $\mathbb{R}\mathcal{Z}_{P^n}$. There is a *canonical action* of $(\mathbb{Z}_2)^r$ on $\mathbb{R}\mathcal{Z}_{P^n}$ by

$$g' \cdot [(p, g)] = [(p, g' + g)],$$

for all $p \in P^n$ and $g, g' \in (\mathbb{Z}_2)^r$. This $(\mathbb{Z}_2)^r$-action on $\mathbb{R}\mathcal{Z}_{P^n}$ is not free. But a subgroup $N \subset (\mathbb{Z}_2)^r$ might act freely on $\mathbb{R}\mathcal{Z}_{P^n}$ through the canonical action. In that case, the quotient space $\mathbb{R}\mathcal{Z}_{P^n}/N$ is called a *partial quotient* of $\mathbb{R}\mathcal{Z}_{P^n}$ [Buchstaber and Panov 2002, Section 7.5]. Also, if there is another subgroup $\widetilde{N}$ of $(\mathbb{Z}_2)^r$ with $\widetilde{N} \supset N$, and $\widetilde{N}$ also acts freely on $\mathbb{R}\mathcal{Z}_{P^n}$ through the canonical action, we get an induced free action of $\widetilde{N}/N$ on $\mathbb{R}\mathcal{Z}_{P^n}/N$ whose orbit space is $\mathbb{R}\mathcal{Z}_{P^n}/\widetilde{N}$. By abuse of terminology, we also call this $(\widetilde{N}/N)$-action on $\mathbb{R}\mathcal{Z}_{P^n}/N$ *canonical*.

It is known that any small cover over P^n (if it exists) is a partial quotient of $\mathbb{R}\mathcal{Z}_{P^n}$ by a rank $(r-n)$ subgroup of $(\mathbb{Z}_2)^r$ [Buchstaber and Panov 2002, Section 7.5].

Proposition 1.5. *Suppose that Q^n is a small cover over a simple convex polytope P^n of dimension n, and that M^n is a principal $(\mathbb{Z}_2)^m$-bundle over Q^n. If M^n is connected, then there exists a subgroup N of $(\mathbb{Z}_2)^r$, where r is the number of facets of P^n, such that M^n is equivalent to the partial quotient $\mathbb{R}\mathcal{Z}_{P^n}/N$ as principal $(\mathbb{Z}_2)^m$-bundles over Q^n.*

Recall that two principal $(\mathbb{Z}_2)^m$-bundles M_1^n and M_2^n over a space Q^n are called *equivalent* if there are a homeomorphism $f : M_1^n \to M_2^n$ and a group automorphism $\sigma : (\mathbb{Z}_2)^m \to (\mathbb{Z}_2)^m$ such that

- $f(g \cdot x) = \sigma(g) \cdot f(x)$ for all $g \in (\mathbb{Z}_2)^m$ and $x \in M_1^n$, and
- f induces the identity map on the orbit space.

Under these conditions, we also say that the *free* $(\mathbb{Z}_2)^m$*-actions* on M_1^n and M_2^n *are equivalent*.

This paper is organized as follows. In Section 2, we review how to construct principal $(\mathbb{Z}_2)^m$-bundles over a manifold from the classical theory of fiber bundles and from the glue-back construction introduced in [Yu 2012]. We compare these two constructions, using them to prove several lemmas on principal $(\mathbb{Z}_2)^m$-bundles, and then give a proof of Proposition 1.5. In Section 3, we prove Theorem 1.4.

2. Glue-back construction

Suppose $(\mathbb{Z}_2)^m$ acts freely and smoothly on an n-dimensional closed manifold M^n. Then the orbit space $Q^n = M^n/(\mathbb{Z}_2)^m$ is naturally a closed manifold. In this section, we assume that Q^n is connected and that $H^1(Q^n, \mathbb{Z}_2) \neq 0$. Indeed, if Q^n is not connected, we can just apply our discussion to each connected component of Q^n. And if $H^1(Q^n, \mathbb{Z}_2) = 0$, then M^n must be homeomorphic to $Q^n \times (\mathbb{Z}_2)^m$.

Let $\pi : M^n \to Q^n$ be the orbit map of the free $(\mathbb{Z}_2)^m$-action. If we think of M^n as a principal $(\mathbb{Z}_2)^m$-bundle over Q^n, then it determines an element

$$\Lambda_\pi \in \mathrm{Hom}\big(H_1(Q^n, \mathbb{Z}_2), (\mathbb{Z}_2)^m\big) \cong H^1(Q^n, (\mathbb{Z}_2)^m). \tag{4}$$

If we think of M^n as a regular covering space over Q^n, its *monodromy* is a group homomorphism $\mathscr{H}_\pi : \pi_1(Q^n, q_0) \to (\mathbb{Z}_2)^m$, where q_0 is a base point of Q^n. Then $\mathscr{H}_\pi$ factors through Λ_π via the canonical group homomorphism

$$\pi_1(Q^n, q_0) \to H_1(Q^n, \mathbb{Z}) \to H_1(Q^n, \mathbb{Z}_2). \tag{5}$$

Conversely, given any element $\Lambda \in \mathrm{Hom}(H_1(Q^n, \mathbb{Z}_2), (\mathbb{Z}_2)^m)$, we can obtain a principal $(\mathbb{Z}_2)^m$-bundle $X(Q^n, \Lambda)$ over Q^n as follows. We compose Λ with the group homeomorphism in (5) and obtain a group homomorphism

$$\Phi_\Lambda : \pi_1(Q^n, q_0) \to (\mathbb{Z}_2)^m. \tag{6}$$

Then we define a left action of $\pi_1(Q^n, q_0)$ on $(\mathbb{Z}_2)^m$ by

$$\gamma \cdot g = \Phi_\Lambda(\gamma) + g, \tag{7}$$

for all $\gamma \in \pi_1(Q^n, q_0)$ and $g \in (\mathbb{Z}_2)^m$. Also, suppose $p : \widetilde{Q}^n \to Q^n$ is a universal covering of Q^n, and let $\pi_1(Q^n, q_0)$ act freely on $\widetilde{Q}^n$ from the right. Then we can define a free action of $\pi_1(Q^n, q_0)$ on $\widetilde{Q}^n \times (\mathbb{Z}_2)^m$ thus: for any $\gamma \in \pi_1(Q^n, q_0)$ and $(x, g) \in \widetilde{Q}^n \times (\mathbb{Z}_2)^m$,

$$\gamma \cdot (x, g) := (x \cdot \gamma^{-1}, \gamma \cdot g) = (x \cdot \gamma^{-1}, \Phi_\Lambda(\gamma) + g). \tag{8}$$

Let $X(Q^n, \Lambda)$ be the quotient space of this $\pi_1(Q^n, q_0)$ action on $\widetilde{Q}^n \times (\mathbb{Z}_2)^m$, and let $\Theta_\Lambda : \widetilde{Q}^n \times (\mathbb{Z}_2)^m \to X(Q^n, \Lambda)$ be the corresponding quotient map. So for all

$\gamma \in \pi_1(Q^n, q_0)$ and $(x, g) \in \widetilde{Q}^n \times (\mathbb{Z}_2)^m$, we have

$$\Theta_\Lambda(x \cdot \gamma, g) = \Theta_\Lambda(x, \gamma \cdot g).$$

Now, for any $(x, g) \in \widetilde{Q}^n \times (\mathbb{Z}_2)^m$, we define a map

$$\pi_\Lambda : X(Q^n, \Lambda) \to Q^n, \quad \Theta_\Lambda(x, g) \mapsto p(x). \tag{9}$$

Clearly $\pi_\Lambda : X(Q^n, \Lambda) \to Q^n$ is a principal $(\mathbb{Z}_2)^m$-bundle with a *canonical free* $(\mathbb{Z}_2)^m$*-action* defined by

$$g' \cdot \Theta_\Lambda(x, g) := \Theta_\Lambda(x, g + g'), \tag{10}$$

for all $x \in \widetilde{Q}^n$ and $g, g' \in (\mathbb{Z}_2)^m$. Therefore the monodromy of $X(Q^n, \Lambda)$ is given by Φ_Λ. We call $X(Q^n, \Lambda)$ the bundle associated to $p : \widetilde{Q}^n \to Q^n$ (thought of as a principal $\pi_1(Q^n)$-bundle) and the $\pi_1(Q)$-action (7) on $(\mathbb{Z}_2)^m$. In the theory of fiber bundles, we may also write

$$X(Q^n, \Lambda) = \widetilde{Q}^n \times_\Lambda (\mathbb{Z}_2)^m.$$

Also, any subgroup H of $(\mathbb{Z}_2)^m$ acts freely on $X(Q^n, \Lambda)$ via (10). Then the quotient space $X(Q^n, \Lambda)/H$ is naturally equipped with a free $(\mathbb{Z}_2)^m/H$-action. We call $X(Q^n, \Lambda)/H$ with this free $(\mathbb{Z}_2)^m/H$-action a *partial quotient* of $X(Q^n, \Lambda)$.

For a principal $(\mathbb{Z}_2)^m$-bundle $\pi : M^n \to Q^n$, it is easy to verify that $X(Q^n, \Lambda_\pi)$ is equivalent to M^n as principal $(\mathbb{Z}_2)^m$-bundles over Q^n.

But this way of constructing M^n from Q^n and Λ_π is not so convenient for the proof of Theorem 1.4, so we use another way of constructing principal $(\mathbb{Z}_2)^m$-bundles over Q^n, introduced in [Yu 2012]. First, we construct a manifold with corners from Q^n that can carry the information of any element of $H^1(Q^n, (\mathbb{Z}_2)^m)$. This is done as follows [Yu 2012].

By a standard argument of intersection theory in differential topology, we can show that there exists a collection of $(n-1)$-dimensional compact embedded submanifolds $\Sigma_1, \ldots, \Sigma_k$ in Q^n such that their homology classes $\{[\Sigma_1], \ldots, [\Sigma_k]\}$ form a basis of $H_{n-1}(Q^n, \mathbb{Z}_2) \cong H^1(Q^n, \mathbb{Z}_2) \neq 0$. Also, we can put $\Sigma_1, \ldots, \Sigma_k$ in *general position* in Q^n, which means that

- $\Sigma_1, \ldots, \Sigma_k$ intersect transversely with each other, and
- if $\Sigma_{i_1} \cap \cdots \cap \Sigma_{i_s}$ is not empty, then it is an embedded submanifold of Q^n of codimension s.

Then we cut Q^n open along $\Sigma_1, \ldots, \Sigma_k$; that is, we choose a small tubular neighborhood $N(\Sigma_i)$ of each Σ_i and remove the interior of each $N(\Sigma_i)$ from Q^n. Then we get a nice manifold with corners $V^n = Q^n - \bigcup_i \operatorname{int}(N(\Sigma_i))$, which is called a $\mathbb{Z}_2$*-core* of Q^n from cutting Q^n open along $\Sigma_1, \ldots, \Sigma_k$ (see Figure 1 for an example). A manifold with corners is called *nice* if each codimension-l face

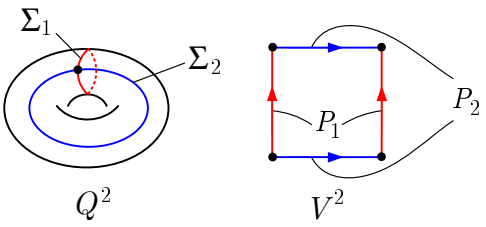


Figure 1. A $\mathbb{Z}_2$-core of a torus.

of the manifold belongs to exactly l facets [Jänich 1968; Davis 1983]. Here, the niceness of V^n follows from $\Sigma_1, \ldots, \Sigma_k$ being in general position in Q^n. The boundary of $N(\Sigma_i)$ is called the *cut section* of Σ_i in Q^n, and $\{\Sigma_1, \ldots, \Sigma_k\}$ is called a $\mathbb{Z}_2$*-cut system* of Q^n. We can choose each Σ_i to be connected.

The projection $\eta_i : \partial N(\Sigma_i) \to \Sigma_i$ is a double cover, either trivial or nontrivial. Let $\bar{\tau}_i$ be the generator of the deck transformation of η_i. Then $\bar{\tau}_i$ is a free involution on $\partial N(\Sigma_i)$; that is, $\bar{\tau}_i$ is a homeomorphism with no fixed point, and $\bar{\tau}_i^2 = id$. By applying some local deformations to these $\bar{\tau}_i$ if necessary [Yu 2012], we can construct an *involutive panel structure* on ∂V^n, which means that the boundary of V^n is the union of some compact subsets $P_1, \ldots, P_k$ (called *panels*) that satisfy the following conditions:

(a) each panel P_i is a disjoint union of facets of V^n, and each facet is contained in exactly one panel;

(b) there exists a free involution τ_i on each P_i that sends a face $f \subset P_i$ to a face $f' \subset P_i$ (it is possible that $f' = f$);

(c) for all $i \neq j$, we have $\tau_i(P_i \cap P_j) \subset P_i \cap P_j$ and $\tau_i \circ \tau_j = \tau_j \circ \tau_i : P_i \cap P_j \to P_i \cap P_j$.

The P_i above consists of those facets of V^n that lie in the cut section of Σ_i, and $\tau_i : P_i \to P_i$ is the restriction of the modified $\bar{\tau}_i$ to P_i (see [Yu 2012] for the details of these constructions).

Remark 2.1. A more general notion of involutive panel structure is introduced in [Yu 2012], where the involution τ_i in (b) is not required to be free. This general notion is used in [Yu 2012] to unify the construction of all locally standard $(\mathbb{Z}_2)^m$-actions on closed manifolds from the orbit spaces.

Let $\mathcal{P}(V^n) = \{P_1, \ldots, P_k\}$ denote the set of all panels in V^n. Any map $\lambda : \mathcal{P}(V^n) \to (\mathbb{Z}_2)^m$ is called a $(\mathbb{Z}_2)^m$*-coloring on* V^n, and any element in $(\mathbb{Z}_2)^m$ is called a *color*.

Now, let us see how to recover a principal $(\mathbb{Z}_2)^m$-bundle $\pi : M^n \to Q^n$ from a $\mathbb{Z}_2$-core V^n of Q^n and the element $\Lambda_\pi \in \mathrm{Hom}(H_1(Q^n, \mathbb{Z}_2), (\mathbb{Z}_2)^m)$. By the Poincaré

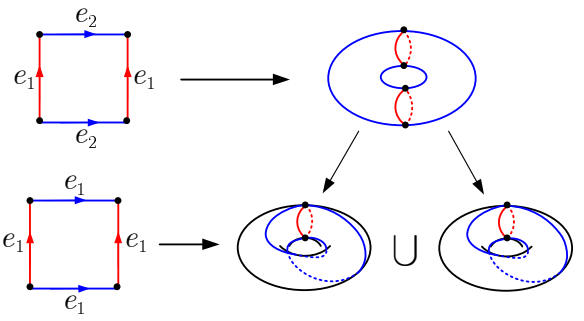


Figure 2

duality for closed manifolds, there is a group isomorphism

$$\kappa : H_{n-1}(Q^n, \mathbb{Z}_2) \to H_1(Q^n, \mathbb{Z}_2).$$

So we can assign an element of $(\mathbb{Z}_2)^m$ to each panel P_i of V^n by

$$\lambda_\pi(P_i) = \Lambda_\pi(\kappa([\Sigma_i])) \in (\mathbb{Z}_2)^m. \tag{11}$$

We call λ_π the *associated* $(\mathbb{Z}_2)^m$*-coloring* of $\pi : M^n \to Q^n$ on V^n.

Generally, for any $(\mathbb{Z}_2)^m$-coloring λ on V^n, we can glue 2^m copies of V^n by

$$M(V^n, \{P_i, \tau_i\}, \lambda) := V^n \times (\mathbb{Z}_2)^m / \sim, \tag{12}$$

where $(x, g) \sim (x', g')$ whenever $x' = \tau_i(x)$ for some P_i and $g' = g + \lambda(P_i) \in (\mathbb{Z}_2)^m$.

If x is in the relative interior of $P_{i_1} \cap \cdots \cap P_{i_s}$, then $(x, g) \sim (x', g')$ if and only if $(x', g') = \big(\tau_{i_s}^{\varepsilon_s} \circ \cdots \circ \tau_{i_1}^{\varepsilon_1}(x), g + \varepsilon_1\lambda(P_1) + \cdots + \varepsilon_s\lambda(P_s)\big)$, where $\varepsilon_j = 0$ or 1 for any $1 \le j \le s$ and $\tau_{i_j}^0 := id$.

$M(V^n, \{P_i, \tau_i\}, \lambda)$ is called the *glue-back construction* from (V^n, λ). Also, we use $M(V^n, \lambda)$ to denote $M(V^n, \{P_i, \tau_i\}, \lambda)$ in contexts where there is no ambiguity about the involutive panel structure on V^n.

Example 2.2. Figure 2 shows two principal $(\mathbb{Z}_2)^2$-bundles over a torus T^2 via glue-back constructions from two different $(\mathbb{Z}_2)^2$-colorings on a $\mathbb{Z}_2$-core of T^2. The $\{e_1, e_2\}$ in the picture is a linear basis of $(\mathbb{Z}_2)^2$. The first $(\mathbb{Z}_2)^2$-coloring gives a torus, and the second one gives a disjoint union of two tori. Also, we can define a double covering map (as defined later in (16)) from the torus on the top to either one of the tori below it.

Example 2.3. Figure 3 shows a $\mathbb{Z}_2$-core of the Klein bottle with three different $\mathbb{Z}_2$-colorings, where $\mathbb{Z}_2 = \langle a \rangle$. So from the glue-back construction, we get three inequivalent double coverings of the Klein bottle. From left to right in Figure 3, the first $\mathbb{Z}_2$-coloring gives a torus, while the second and the third both give a Klein bottle.

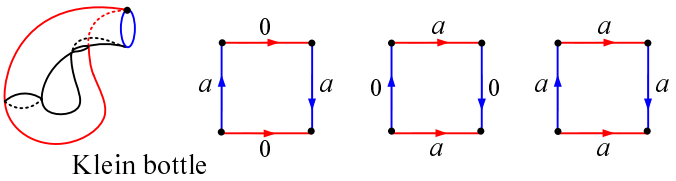


Figure 3

Let $\theta_\lambda : V^n \times (\mathbb{Z}_2)^m \to M(V^n, \lambda)$ be the quotient map defined in (12). It is shown in [Yu 2012] that $M(V^n, \lambda)$ is a closed manifold with a smooth free $(\mathbb{Z}_2)^m$-action defined by

$$g' \cdot \theta_\lambda(x, g) := \theta_\lambda(x, g + g'), \tag{13}$$

for all $x \in V^n$ and $g, g' \in (\mathbb{Z}_2)^m$. The orbit space of $M(V^n, \lambda)$ under this free $(\mathbb{Z}_2)^m$-action is homeomorphic to Q^n. We say that (13) defines the *natural* $(\mathbb{Z}_2)^m$-*action* on $M(V^n, \lambda)$. Here, we always associate this natural free $(\mathbb{Z}_2)^m$-action to $M(V^n, \lambda)$. Any subgroup $H \subset (\mathbb{Z}_2)^m$ also acts freely on $M(V^n, \lambda)$ through the natural action. The induced action of $(\mathbb{Z}_2)^m/H$ on $M(V^n, \lambda)/H$ is also free, and its orbit space is homeomorphic to $M(V^n, \lambda)/(\mathbb{Z}_2)^m = Q^n$. By abuse of terminology, we also call this $(\mathbb{Z}_2)^m/H$-action on $M(V^n, \lambda)/H$ *natural* and call $M(V^n, \lambda)/H$ with the natural $(\mathbb{Z}_2)^m/H$-action a *partial quotient* of $M(V^n, \lambda)$.

We have defined "partial quotient" in three different contexts : $\mathbb{RZ}_{P^n}$, $X(Q^n, \Lambda)$ and $M(V^n, \lambda)$. The common property of these notions is that each of them denotes the quotient space of some free $\mathbb{Z}_2$-torus action on a space.

Theorem 2.4 [Yu 2012, Theorem 3.5]. *Let $\pi : M^n \to Q^n$ be a principal $(\mathbb{Z}_2)^m$-bundle, and let λ_π be the associated $(\mathbb{Z}_2)^m$-coloring on V^n. Then $M(V^n, \lambda_\pi)$ and M^n are equivalent principal $(\mathbb{Z}_2)^m$-bundles over Q^n.*

For any integer $m \geq 1$, define

$$\begin{aligned} \mathrm{Col}_m(V^n) &:= \text{the set of all } (\mathbb{Z}_2)^m\text{-colorings on } V^n \\ &= \{\lambda \mid \lambda : \mathcal{P}(V^n) \to (\mathbb{Z}_2)^m\}, \\ L_\lambda &:= \text{the subgroup of } (\mathbb{Z}_2)^m \text{ generated by } \{\lambda(P_1), \ldots, \lambda(P_k)\}, \\ \mathrm{rank}(\lambda) &:= \dim_{\mathbb{Z}_2} L_\lambda, \text{ for all } \lambda \in \mathrm{Col}_m(V^n). \end{aligned}$$

For any $g \in (\mathbb{Z}_2)^m$, it is clear from (13) that L_λ acts freely on $\theta_\lambda(V^n \times (g + L_\lambda))$, whose orbit space is Q^n.

Theorem 2.5 [Yu 2012, Theorem 3.7]. *For any $(\mathbb{Z}_2)^m$-coloring λ on V^n, $M(V^n, \lambda)$ has $2^{m-\mathrm{rank}(\lambda)}$ connected components that are pairwise homeomorphic, and $L_\lambda \cong (\mathbb{Z}_2)^{\mathrm{rank}(\lambda)}$ acts freely on each connected component of $M(V^n, \lambda)$ whose orbit space*

is Q^n. Each connected component of $M(V^n, \lambda)$ is equivalent to $\theta_\lambda(V^n \times L_\lambda)$ as principal $(\mathbb{Z}_2)^{\mathrm{rank}(\lambda)}$-bundles over Q^n.

An element $\lambda \in \mathrm{Col}_m(V^n)$ is called *maximally independent* if $\mathrm{rank}(\lambda) = k = \dim_{\mathbb{Z}_2} H_{n-1}(Q^n, \mathbb{Z}_2)$. If $\lambda \in \mathrm{Col}_m(V^n)$ is maximally independent, then $m \geq k$.

Obviously, the relation in (11) defines a one-to-one correspondence between the elements of $\mathrm{Col}_m(V^n)$ and $\mathrm{Hom}(H_1(Q^n, \mathbb{Z}_2), (\mathbb{Z}_2)^m) \cong H^1(Q^n, (\mathbb{Z}_2)^m)$. Suppose $\Lambda \in \mathrm{Hom}(H_1(Q^n, \mathbb{Z}_2), (\mathbb{Z}_2)^m)$ is the element corresponding to $\lambda \in \mathrm{Col}_m(V^n)$; then L_λ is nothing but the image $\mathrm{Im}(\Lambda) \subset (\mathbb{Z}_2)^m$ of Λ, and λ is maximally independent if and only if Λ is injective. We define

$$\mathrm{rank}(\Lambda) := \dim_{\mathbb{Z}_2}(\mathrm{Im}(\Lambda)) = \dim_{\mathbb{Z}_2}(L_\lambda) = \mathrm{rank}(\lambda).$$

It is clear that $X(Q^n, \Lambda)$ and $M(V^n, \lambda)$ are equivalent principal $(\mathbb{Z}_2)^m$-bundles over Q^n, and so are $\Theta_\Lambda(\widetilde{Q}^n \times \mathrm{Im}(\Lambda))$ and $\theta_\lambda(V^n \times L_\lambda)$. The canonical free $(\mathbb{Z}_2)^m$-action on $X(Q^n, \Lambda)$ defined by (10) corresponds exactly to the natural $(\mathbb{Z}_2)^m$-action on $M(V^n, \lambda)$ defined by (13). So for any subgroup H of $(\mathbb{Z}_2)^m$, the partial quotients $X(Q^n, \Lambda)/H$ and $M(V^n, \lambda)/H$ are equivalent. Then we can write Theorem 2.5 in terms of $X(Q^n, \Lambda)$ as follows.

Theorem 2.5*. *For any $\Lambda \in \mathrm{Hom}(H_1(Q^n, \mathbb{Z}_2), (\mathbb{Z}_2)^m)$, $X(Q^n, \Lambda)$ has $2^{m-\mathrm{rank}(\Lambda)}$ connected components that are pairwise homeomorphic, and $\mathrm{Im}(\Lambda) \cong (\mathbb{Z}_2)^{\mathrm{rank}(\Lambda)}$ acts freely on each connected component of $X(Q^n, \Lambda)$ whose orbit space is Q^n. Each connected component of $X(Q^n, \Lambda)$ is equivalent to $\Theta_\Lambda(\widetilde{Q}^n \times \mathrm{Im}(\Lambda))$ as principal $(\mathbb{Z}_2)^{\mathrm{rank}(\Lambda)}$-bundles over Q^n.*

We prove several lemmas on principal $(\mathbb{Z}_2)^m$-bundles over a closed manifold. The statements of these lemmas are written in the language of glue-back construction. But we use $X(Q^n, \Lambda)$ and $M(V^n, \lambda)$ alternatively in the proofs of these lemmas, depending on what is convenient.

Lemma 2.6. *For any $m \geq \dim_{\mathbb{Z}_2} H_{n-1}(Q^n, \mathbb{Z}_2)$, if $\lambda_1, \lambda_2 \in \mathrm{Col}_m(V^n)$ are both maximally independent, then $M(V^n, \lambda_1)$ must be equivalent to $M(V^n, \lambda_2)$ as principal $(\mathbb{Z}_2)^m$-bundles over Q^n.*

Proof. Let Λ_1 and Λ_2 be the elements of $\mathrm{Hom}(H_1(Q^n, \mathbb{Z}_2), (\mathbb{Z}_2)^m)$ corresponding to λ_1 and λ_2. Then by our assumption, Λ_1 and Λ_2 are both injective. So there exists a group automorphism σ of $(\mathbb{Z}_2)^m$ such that $\sigma \circ \Lambda_1 = \Lambda_2$. Then we can define a homeomorphism $\phi : \widetilde{Q}^n \times (\mathbb{Z}_2)^m \to \widetilde{Q}^n \times (\mathbb{Z}_2)^m$, for $x \in \widetilde{Q}^n$ and $g \in (\mathbb{Z}_2)^m$, by

$$\phi(x, g) = (x, \sigma(g)).$$

Obviously, $\Theta_{\Lambda_1}(x, g) = \Theta_{\Lambda_1}(x', g')$ if and only if $\Theta_{\Lambda_2}(\phi(x, g)) = \Theta_{\Lambda_2}(\phi(x', g'))$. So ϕ induces an equivalence between the two principal $(\mathbb{Z}_2)^m$-bundles $X(Q^n, \Lambda_1)$ and $X(Q^n, \Lambda_2)$. So $M(V^n, \lambda_1)$ is equivalent to $M(V^n, \lambda_2)$. □

Lemma 2.7. *Suppose M_1 and M_2 are two principal $(\mathbb{Z}_2)^k$-bundles over Q^n, where $k = \dim_{\mathbb{Z}_2} H_{n-1}(Q^n, \mathbb{Z}_2)$. If M_1 and M_2 are both connected, then M_1 must be equivalent to M_2 as principal $(\mathbb{Z}_2)^k$-bundles over Q^n.*

Proof. Using this notation, for some $\lambda_i \in \mathrm{Col}_k(V^n)$, $i = 1, 2$, Theorem 2.4 gives

$$M_i \cong M(V^n, \lambda_i).$$

Also, because M_1 and M_2 are both connected, Theorem 2.5 implies that $\mathrm{rank}(\lambda_1) = \mathrm{rank}(\lambda_2) = k$; that is, λ_1 and λ_2 are both maximally independent. So by Lemma 2.6, $M(V^n, \lambda_1)$ and $M(V^n, \lambda_2)$ are equivalent principal $(\mathbb{Z}_2)^k$-bundles over Q^n. □

We study some relations between $M(V^n, \lambda)$ for different $\lambda \in \mathrm{Col}_m(V^n)$. For conciseness, for any topological space B and field $\mathbb{F}$, we define

$$\mathrm{hrk}(B, \mathbb{F}) := \sum_{i=0}^{\infty} \dim_{\mathbb{F}} H^i(B, \mathbb{F}).$$

Lemma 2.8. *For any double covering $\xi : \widetilde{B} \to B$ and any $i \geq 0$,*

$$\dim_{\mathbb{Z}_2} H^i(\widetilde{B}, \mathbb{Z}_2) \leq 2 \cdot \dim_{\mathbb{Z}_2} H^i(B, \mathbb{Z}_2).$$

So $\mathrm{hrk}(\widetilde{B}, \mathbb{Z}_2) \leq 2 \cdot \mathrm{hrk}(B, \mathbb{Z}_2)$.

Proof. The Gysin sequence of $\xi : \widetilde{B} \to B$, in $\mathbb{Z}_2$-coefficient, reads:

$$\cdots \to H^{i-1}(B, \mathbb{Z}_2) \xrightarrow{\phi_{i-1}} H^i(B, \mathbb{Z}_2) \xrightarrow{\xi^*} H^i(\widetilde{B}, \mathbb{Z}_2) \to H^i(B, \mathbb{Z}_2) \xrightarrow{\phi_i} \ldots,$$

where $\phi_i(\gamma) = \gamma \cup e$ for all $\gamma \in H^i(B, \mathbb{Z}_2)$ and $e \in H^1(B, \mathbb{Z}_2)$ is the first Stiefel–Whitney class (mod 2 Euler class) of $\widetilde{B}$. Then by the exactness of the Gysin sequence, we have

$$\begin{aligned}
\dim_{\mathbb{Z}_2} H^i(\widetilde{B}, \mathbb{Z}_2) &= \dim_{\mathbb{Z}_2} H^i(B, \mathbb{Z}_2) - \dim_{\mathbb{Z}_2} \mathrm{Im}(\phi_{i-1}) + \dim_{\mathbb{Z}_2} \ker(\phi_i) \\
&= 2 \cdot \dim_{\mathbb{Z}_2} H^i(B, \mathbb{Z}_2) - \dim_{\mathbb{Z}_2} \mathrm{Im}(\phi_{i-1}) - \dim_{\mathbb{Z}_2} \mathrm{Im}(\phi_i) \\
&\leq 2 \cdot \dim_{\mathbb{Z}_2} H^i(B, \mathbb{Z}_2).
\end{aligned}$$

□

Remark 2.9. In Lemma 2.8, if we replace the $\mathbb{Z}_2$-coefficients by $\mathbb{Z}_p$ (p is an odd prime) or $\mathbb{Q}$ (rational) coefficients, the conclusion in the lemma might fail. For example, let $B = \mathbb{R}P^2 \vee \mathbb{R}P^2$ be a one-point union of two $\mathbb{R}P^2$'s, and let $\widetilde{B}$ be the union of two spheres that intersect at two points (see Figure 4). It is clear that $\widetilde{B}$ is a double covering of B. But for any field $\mathbb{F} = \mathbb{Z}_p$ or $\mathbb{Q}$, we have $\mathrm{hrk}(B, \mathbb{F}) = 1$, while $\mathrm{hrk}(\widetilde{B}, \mathbb{F}) = 4$.

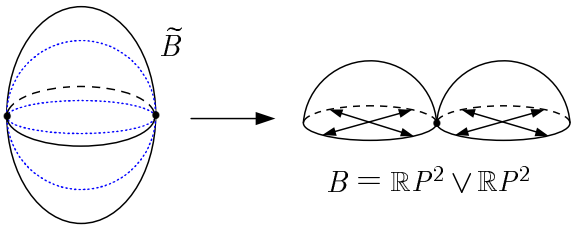


Figure 4

Lemma 2.10. *Suppose that* $\lambda_{\max} \in \mathrm{Col}_k(V^n)$ *is a maximally independent* $(\mathbb{Z}_2)^k$*-coloring on* V^n*, where* $k = \dim_{\mathbb{Z}_2} H_{n-1}(Q^n, \mathbb{Z}_2)$*. Then, for any* $\lambda \in \mathrm{Col}_k(V^n)$*,*

$$\mathrm{hrk}(M(V^n, \lambda), \mathbb{Z}_2) \geq \mathrm{hrk}(M(V^n, \lambda_{\max}), \mathbb{Z}_2).$$

Proof. Suppose Λ is the element of $\mathrm{Hom}(H_1(Q^n, \mathbb{Z}_2), (\mathbb{Z}_2)^k)$ corresponding to λ. Let $\{\alpha_1, \ldots, \alpha_k\}$ be a $\mathbb{Z}_2$-linear basis of $H_1(Q^n, \mathbb{Z}_2)$. Without loss of generality, we assume that $\{\Lambda(\alpha_1), \ldots, \Lambda(\alpha_s)\}$ is a $\mathbb{Z}_2$-linear basis of $\mathrm{Im}(\Lambda) \subset (\mathbb{Z}_2)^k$. Then we can choose $\omega_1, \ldots, \omega_{k-s} \in (\mathbb{Z}_2)^k$ such that $(\mathbb{Z}_2)^k = \mathrm{Im}(\Lambda) \oplus \langle\omega_1\rangle \oplus \cdots \oplus \langle\omega_{k-s}\rangle$.

We define a sequence of elements $\Lambda_0, \ldots, \Lambda_{k-s} \in \mathrm{Hom}(H_1(Q^n, \mathbb{Z}_2), (\mathbb{Z}_2)^k)$ thus: for any $0 \leq j \leq k-s$,

$$\Lambda_j(\alpha_i) := \begin{cases} \Lambda(\alpha_i) & \text{if } 1 \leq i \leq s \text{ or } s+j < i \leq k; \\ \omega_{i-s} & \text{if } s+1 \leq i \leq s+j. \end{cases} \tag{14}$$

Clearly $\Lambda_0 = \Lambda$ and $\mathrm{Im}(\Lambda) = \mathrm{Im}(\Lambda_0) \subset \mathrm{Im}(\Lambda_1) \subset \cdots \subset \mathrm{Im}(\Lambda_{k-s}) = (\mathbb{Z}_2)^k$, and for $1 \leq j \leq k-s$,

$$\mathrm{rank}(\Lambda_j) = \mathrm{rank}(\Lambda_{j-1}) + 1.$$

Let λ_j be the elements of $\mathrm{Col}_k(V^n)$ corresponding to Λ_j, with $0 \leq j \leq k-s$. Then λ_{k-s} is maximally independent. So by Lemma 2.6, we have

$$\mathrm{hrk}(M(V^n, \lambda_{\max}), \mathbb{Z}_2) = \mathrm{hrk}(M(V^n, \lambda_{k-s}), \mathbb{Z}_2) = \mathrm{hrk}(X(Q^n, \Lambda_{k-s}), \mathbb{Z}_2). \tag{15}$$

To prove the lemma, it suffices to show that for all $1 \leq j \leq k-s$,

$$\mathrm{hrk}(X(Q^n, \Lambda_{j-1}), \mathbb{Z}_2) \geq \mathrm{hrk}(X(Q^n, \Lambda_j), \mathbb{Z}_2).$$

Notice that $\mathrm{Im}(\Lambda_j) = \mathrm{Im}(\Lambda_{j-1}) \oplus \langle\omega_j\rangle \subset (\mathbb{Z}_2)^k$, and the only difference between Λ_{j-1} and Λ_j is that $\Lambda_{j-1}(\alpha_{s+j}) = \Lambda(\alpha_{s+j})$ while $\Lambda_j(\alpha_{s+j}) = \omega_j$. Let

$$K_j = \Theta_{\Lambda_j}(\widetilde{Q}^n \times \mathrm{Im}(\Lambda_j))$$

for all $1 \leq j \leq k-s$, where $p : \widetilde{Q}^n \to Q^n$ is a universal covering of Q^n.

We define a free involution t_j on K_j: for any $(x, g) \in \widetilde{Q}^n \times \mathrm{Im}(\Lambda_j)$,

$$t_j(\Theta_{\Lambda_j}(x, g)) = \Theta_{\Lambda_j}(x, g + \Lambda(\alpha_{s+j}) + \omega_j). \tag{16}$$

Let K_j/t_j be the quotient space of K_j under t_j, and let $\overline{\Theta_{\Lambda_j}(x,g)} \in K_j/t_j$ denote the equivalence class of $\Theta_{\Lambda_j}(x,g)$. So K_j is a double covering of K_j/t_j.

By (9), the bundle map $\pi_{\Lambda_j} : X(Q^n, \Lambda_j) \to Q^n$ restricted to K_j gives a bundle map $\pi_{\Lambda_j} : K_j \to Q^n$ that sends any $\Theta_{\Lambda_j}(x,g)$ to $p(x)$, and the monodromy of π_{Λ_j} is $\Phi_{\Lambda_j} : \pi_1(Q^n, q_0) \to \mathrm{Im}(\Lambda_j) \subset (\mathbb{Z}_2)^k$; see (6). So π_{Λ_j} induces a map

$$\overline{\pi}_{\Lambda_j} : K_j/t_j \to Q^n, \quad \overline{\Theta_{\Lambda_j}(x,g)} \mapsto p(x).$$

By the definition (16) of t_j, we can easily see that $\overline{\pi}_{\Lambda_j}$ is a fiber bundle whose fiber is $\mathrm{Im}(\Lambda_j)$ modulo the relation $\sim$, where for all $g \in \mathrm{Im}(\Lambda_j)$,

$$g \sim g + \Lambda(\alpha_{s+j}) + \omega_j,$$

or equivalently, $\omega_j \sim \Lambda(\alpha_{s+j})$. Now by (14), $\mathrm{Im}(\Lambda_j)/\sim$ can be identified with $\mathrm{Im}(\Lambda_{j-1})$, so the fiber of $\overline{\pi}_{\Lambda_j} : K_j/t_j \to Q^n$ is isomorphic to $\mathrm{Im}(\Lambda_{j-1})$. Let

$$\varrho : \mathrm{Im}(\Lambda_j) \to \mathrm{Im}(\Lambda_{j-1}) = \mathrm{Im}(\Lambda_j)/\sim .$$

So the monodromy of $\overline{\pi}_{\Lambda_j}$ is $\varrho \circ \Phi_{\Lambda_j} : \pi_1(Q^n, q_0) \to \mathrm{Im}(\Lambda_{j-1})$. Also, it is easy to check that $\varrho \circ \Phi_{\Lambda_j}$ coincides with the monodromy $\Phi_{\Lambda_{j-1}}$ of the bundle $\pi_{\Lambda_{j-1}} : K_{j-1} \to Q^n$. Therefore, the two bundles K_j/t_j and K_{j-1} over Q^n are actually equivalent. So by Lemma 2.8,

$$\mathrm{hrk}(K_j, \mathbb{Z}_2) \le 2 \cdot \mathrm{hrk}(K_j/t_j, \mathbb{Z}_2) = 2 \cdot \mathrm{hrk}(K_{j-1}, \mathbb{Z}_2).$$

Also, because by Theorem 2.5*, $X(Q^n, \Lambda_j)$ consists of $2^{k-\mathrm{rank}(\Lambda_j)}$ copies of K_j for each $0 \le j \le k-s$ and $\mathrm{rank}(\Lambda_j) = \mathrm{rank}(\Lambda_{j-1}) + 1$, we get

$$\begin{aligned}\mathrm{hrk}(X(Q^n, \Lambda_{j-1}), \mathbb{Z}_2) &= 2^{k-\mathrm{rank}(\Lambda_{j-1})} \mathrm{hrk}(K_{j-1}, \mathbb{Z}_2) \\ &\ge 2^{k-\mathrm{rank}(\Lambda_j)} \mathrm{hrk}(K_j, \mathbb{Z}_2) = \mathrm{hrk}(X(Q^n, \Lambda_{j+1}), \mathbb{Z}_2).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{hrk}(M(V^n, \lambda), \mathbb{Z}_2) = \mathrm{hrk}(X(Q^n, \Lambda_0), \mathbb{Z}_2) &\ge \mathrm{hrk}(X(Q^n, \Lambda_{k-s}), \mathbb{Z}_2) \\ &= \mathrm{hrk}(M(V^n, \lambda_{\max}), \mathbb{Z}_2),\end{aligned}$$

where we use (15) for the final equality. □

Lemma 2.11. *Let M^n be a connected principal $(\mathbb{Z}_2)^s$-bundle over Q^n. Then there exist a maximally independent coloring $\widetilde{\lambda} \in \mathrm{Col}_k(V^n)$, where*

$$k = \dim_{\mathbb{Z}_2} H_{n-1}(Q^n, \mathbb{Z}_2),$$

and a free $(\mathbb{Z}_2)^{k-s}$-action on $M(V^n, \widetilde{\lambda})$ whose orbit space is homeomorphic to M^n. Also, M^n is equivalent to a partial quotient $M(V^n, \widetilde{\lambda})/H$ for some subgroup H of $(\mathbb{Z}_2)^k$ with rank $k-s$.

Proof. We use a similar argument to the proof of Lemma 2.10. Because M^n is connected, Theorem 2.5 implies that $s \leq k$ and that there is an element $\lambda \in \mathrm{Col}_k(V^n)$ such that $\mathrm{rank}(\lambda) = s$ and M^n is homeomorphic to $\theta_\lambda(V^n \times L_\lambda) \subset M(V^n, \lambda)$.

As in the proof of Lemma 2.10, let Λ be the element of $\mathrm{Hom}(H_1(Q^n, \mathbb{Z}_2), (\mathbb{Z}_2)^k)$ corresponding to λ, and let $\{\alpha_1, \ldots, \alpha_k\}$ be a $\mathbb{Z}_2$-linear basis of $H_1(Q^n, \mathbb{Z}_2)$ such that $\{\Lambda(\alpha_1), \ldots, \Lambda(\alpha_s)\}$ is a $\mathbb{Z}_2$-linear basis of $\mathrm{Im}(\Lambda) \subset (\mathbb{Z}_2)^k$. Suppose also that $(\mathbb{Z}_2)^k = \mathrm{Im}(\Lambda) \oplus \langle\omega_1\rangle \oplus \cdots \oplus \langle\omega_{k-s}\rangle$, and define the same sequence of elements $\Lambda = \Lambda_0, \Lambda_1, \ldots, \Lambda_{k-s} \in \mathrm{Hom}(H_1(Q^n, \mathbb{Z}_2), (\mathbb{Z}_2)^k)$ as in (14) and corresponding elements $\lambda_0, \lambda_1, \ldots, \lambda_{k-s} \in \mathrm{Col}_k(V^n)$. So λ_{k-s} is maximally independent.

Let $\widehat{H} = \langle\omega_1\rangle \oplus \cdots \oplus \langle\omega_{k-s}\rangle \subset (\mathbb{Z}_2)^k$. Then $\widehat{H} \cong (\mathbb{Z}_2)^{k-s}$, and there exists a free action $\star$ of $\widehat{H}$ on $X(Q^n, \Lambda_{k-s}) \cong M(V^n, \lambda_{k-s})$ defined by

$$\omega_j \star \Theta_{\Lambda_{j-s}}(x, g) := \Theta_{\Lambda_{j-s}}(x, g + \Lambda(\alpha_{s+j}) + \omega_j),$$

for $1 \leq j \leq k-s$. As in the proof of Lemma 2.10, we can show that the orbit space of the action of $\widehat{H}$ is homeomorphic to $\Theta_\Lambda(\widetilde{Q}^n \times \mathrm{Im}(\Lambda)) \cong \theta_\lambda(V^n \times L_\lambda) \cong M^n$.

The action of $\widehat{H}$ on $X(Q^n, \Lambda_{k-s})$ can be identified with the canonical action (10) of $H = \langle\Lambda(\alpha_{s+1}) + \omega_1\rangle \oplus \cdots \oplus \langle\Lambda(\alpha_k) + \omega_{k-s}\rangle$ on $X(Q^n, \Lambda_{k-s})$ via a group isomorphism $\sigma : \widehat{H} \to H$, where for $1 \leq j \leq k-s$,

$$\sigma(\omega_j) = \Lambda(\alpha_{s+j}) + \omega_j.$$

Here σ is an isomorphism because $(\mathbb{Z}_2)^k = \mathrm{Im}(\Lambda) \oplus \langle\omega_1\rangle \oplus \cdots \oplus \langle\omega_{k-s}\rangle$. So M^n is equivalent to the partial quotient $X(Q^n, \Lambda_{k-s})/H \cong M(V^n, \lambda_{k-s})/H$ as principal $(\mathbb{Z}_2)^s$-bundles over Q^n. This completes the lemma. □

Proof of Proposition 1.5. Suppose the polytope P^n has $k+n$ facets. Then

$$H_{n-1}(Q^n, \mathbb{Z}_2) \cong (\mathbb{Z}_2)^k.$$

So by Lemma 2.11, there exist a maximally independent coloring $\widetilde{\lambda} \in \mathrm{Col}_k(V^n)$ and a subgroup $H \subset (\mathbb{Z}_2)^k$ such that M^n is equivalent to the partial quotient $M(V^n, \widetilde{\lambda})/H$ as principal $(\mathbb{Z}_2)^m$-bundles over Q^n. Both $M(V^n, \widetilde{\lambda})$ and the real moment-angle manifold $\mathbb{R}\mathcal{Z}_{P^n}$ are principal $(\mathbb{Z}_2)^k$-bundles over Q^n, and they are both connected. So by Lemma 2.7, $\mathbb{R}\mathcal{Z}_{P^n}$ is equivalent to $M(V^n, \widetilde{\lambda})$.

Let $\widetilde{N} \subset (\mathbb{Z}_2)^{k+n}$ be a subgroup of rank k such that Q^n is homeomorphic to the partial quotient $\mathbb{R}\mathcal{Z}_{P^n}/\widetilde{N}$ (such a subgroup $\widetilde{N}$ is not unique). The equivalence between $M(V^n, \widetilde{\lambda})$ and $\mathbb{R}\mathcal{Z}_{P^n}$ determines a group isomorphism $\sigma : (\mathbb{Z}_2)^k \to \widetilde{N}$. Then $M(V^n, \widetilde{\lambda})/H$ is equivalent to the partial quotient $\mathbb{R}\mathcal{Z}_{P^n}/N$ of $\mathbb{R}\mathcal{Z}_{P^n}$, where $N = \sigma(H) \subset \widetilde{N} \subset (\mathbb{Z}_2)^{k+n}$. This proves our proposition. □

3. Proof of Theorem 1.4

We adapt the following lemma for our proof of Theorem 1.4.

Lemma 3.1 [Ustinovskii 2011]. *Let (X, A) be a pair of CW-complexes such that A has a collar neighborhood $U(A)$ in X, that is,*

$$(U(A), A) \cong (A \times [0, 1), A \times 0).$$

Take a homeomorphism $\varphi : A \to A$ that can be extended to a homeomorphism $\widetilde{\varphi} : X \to X$. Let $Y = X_1 \cup_\varphi X_2$ be the space obtained by gluing two copies of X along A via the map φ. Then for any field $\mathbb{F}$, we have $\operatorname{hrk}(Y, \mathbb{F}) \geq \operatorname{hrk}(A, \mathbb{F})$.

Proof. The argument is almost the same as in [Ustinovskii 2011]. Let $U_1(A)$ and $U_2(A)$ be the collar neighborhoods of A in X_1 and X_2. Consider an open cover $Y = W_1 \cup W_2$, where $W_1 = X_1 \cup U_2(A)$ and $W_2 = X_2 \cup U_1(A)$. Then the Mayer–Vietoris sequence of cohomology groups for this open cover reads (we omit the coefficients $\mathbb{F}$):

$$\begin{aligned}\cdots \to H^{j-1}(W_1 \cap W_2) &\xrightarrow{\delta^*_{(j)}} H^j(Y) \\ &\xrightarrow{g^*_{(j)}} H^j(W_1) \oplus H^j(W_2) \xrightarrow{p^*_{(j)}} H^j(W_1 \cap W_2) \to \cdots.\end{aligned}$$

Here the map $p^*_{(j)}$ equals $i_1^* \oplus -i_2^*$, where i_1 and i_2 are inclusions of $W_1 \cap W_2$ into W_1 and W_2. Because W_1 and W_2 are both homotopy equivalent to X and $W_1 \cap W_2 = U_1(A) \cup U_2(A) \cong A \times (-1, 1)$, we get another, equivalent, long exact sequence

$$\cdots \to H^{j-1}(A) \xrightarrow{\widehat{\delta}^*_{(j)}} H^j(Y) \xrightarrow{\widehat{g}^*_{(j)}} H^j(X_1) \oplus H^j(X_2) \xrightarrow{\widehat{p}^*_{(j)}} H^j(A) \to \cdots.$$

Now $\widehat{p}^*_{(j)} = \iota_1^* \oplus -(\iota_2 \circ \varphi)^*$, where ι_1 and ι_2 are inclusions of A into X_1 and X_2. For any $\gamma \in H^j(X_1)$, it is easy to see that $(\gamma, (\widetilde{\varphi}^{-1})^*\gamma)$ is in $\ker(\widehat{p}^*_{(j)})$. Thus $\dim \ker(\widehat{p}^*_{(j)}) \geq \dim H^j(X)$, and so $\dim \operatorname{Im}(\widehat{p}^*_{(j)}) \leq \dim H^j(X)$. Then

$$\begin{aligned}\dim H^j(Y) &= \dim \ker(\widehat{g}^*_{(j)}) + \dim \operatorname{Im}(\widehat{g}^*_{(j)}) = \dim \operatorname{Im}(\widehat{\delta}^*_{(j)}) + \dim \ker(\widehat{p}^*_{(j)}) \\ &\geq \dim H^{j-1}(A) - \dim \operatorname{Im}(\widehat{p}^*_{(j-1)}) + \dim H^j(X) \\ &\geq \dim H^{j-1}(A) - \dim H^{j-1}(X) + \dim H^j(X).\end{aligned}$$

Summing up these inequalities over all indices j, we get

$$\begin{aligned}\operatorname{hrk}(Y, \mathbb{F}) = \sum_j \dim H^j(Y) &\geq \sum_j \dim H^{j-1}(A) - \dim H^{j-1}(X) + \dim H^j(X) \\ &= \sum_j \dim H^{j-1}(A) = \operatorname{hrk}(A, \mathbb{F}).\end{aligned}$$

□

Remark 3.2. In Lemma 3.1, the assumption that $\varphi : A \to A$ can be extended to a homeomorphism $\widetilde{\varphi} : X \to X$ is essential; otherwise the claim may not hold. For example, let X be a solid torus and $A \cong T^2$ the boundary of X. Let $\varphi : A \to A$ be the

homeomorphism interchanging the meridian and longitude of T^2. If we glue two copies of X along their boundaries via φ, we get a 3-sphere S^3. But $\operatorname{hrk}(S^3, \mathbb{Z}_2) = 2$, while $\operatorname{hrk}(A, \mathbb{Z}_2) = 4$. The reason why the conclusion of Lemma 3.1 does not hold in this example is that φ cannot be extended to a homeomorphism on the whole X.

We introduce an auxiliary notion that plays an important role in our proof of Theorem 1.4. Suppose V^n is a $\mathbb{Z}_2$-core of a closed manifold Q^n and the involutive panel structure on V^n is $\{P_i, \tau_i\}$. For any panel P_j of V^n, we define the space

$$M_{\setminus P_j}(V^n, \lambda) := V^n \times (\mathbb{Z}_2)^m / \sim_{P_j}, \tag{17}$$

where $(x, g) \sim_{P_j} (x', g')$ whenever $x' = \tau_i(x)$ for some $P_i \neq P_j$ and

$$g' = g + \lambda(P_i) \in (\mathbb{Z}_2)^m.$$

In other words, $M_{\setminus P_j}(V^n, \lambda)$ is the quotient space of $V^n \times (\mathbb{Z}_2)^m$ under the rule in (12), except that we leave the interior of those facets in $P_j \times (\mathbb{Z}_2)^m$ open. We call $M_{\setminus P_j}(V^n, \lambda)$ a *partial glue-back* from (V^n, λ). Let the corresponding quotient map be

$$\theta_\lambda^{\setminus P_j} : V^n \times (\mathbb{Z}_2)^m \to M_{\setminus P_j}(V^n, \lambda). \tag{18}$$

Then the boundary of $M_{\setminus P_j}(V^n, \lambda)$ can be written as $\theta_\lambda^{\setminus P_j}(P_j \times (\mathbb{Z}_2)^m)$.

Proof of Theorem 1.4. The proof is by induction on the dimension of M^n. When $n = 1$, the only small cover is a circle. Because a principal $(\mathbb{Z}_2)^m$-bundle over a circle must be a disjoint union of 2^m or 2^{m-1} circles, the theorem holds. Now we assume the theorem holds for manifolds with dimension less than n.

Suppose that P^n is an n-dimensional simple convex polytope with $k+n$ facets $F_1, \ldots, F_{k+n}$, with $k \geq 1$, and that $\pi_\mu : Q^n \to P^n$ is a small cover over P^n with the characteristic function μ. For any face $f = F_{i_1} \cap \cdots \cap F_{i_l}$ of P^n, let G_f^μ be the rank-l subgroup of $(\mathbb{Z}_2)^n$ generated by $\mu(F_1), \ldots, \mu(F_l)$. Then by definition,

$$\begin{gathered} Q^n = P^n \times (\mathbb{Z}_2)^n / \sim, \quad \text{with} \\ (p, w) \sim (p', w') \iff p = p' \text{ and } w - w' \in G_{f(p)}^\mu, \end{gathered} \tag{19}$$

where $f(p)$ is the unique face of P^n that contains p in its relative interior. It was shown in [Davis and Januszkiewicz 1991] that the $\mathbb{Z}_2$-Betti numbers of Q^n can be computed from the h-vector of P^n. In particular, $H_{n-1}(Q^n, \mathbb{Z}_2) \cong (\mathbb{Z}_2)^k$.

We choose an arbitrary vertex v_0 of P^n. By reindexing the facets of P^n, we can assume that $F_1, \ldots, F_k$ are all the facets of P^n that are not incident to v_0. Then according to [Davis and Januszkiewicz 1991], the homology classes of the embedded submanifolds $\pi_\mu^{-1}(F_1), \ldots, \pi_\mu^{-1}(F_k)$ (called *facial submanifolds* of Q^n) form a $\mathbb{Z}_2$-linear basis of $H_{n-1}(Q^n, \mathbb{Z}_2)$. Cutting Q^n open along $\pi_\mu^{-1}(F_1), \ldots, \pi_\mu^{-1}(F_k)$ gives us a $\mathbb{Z}_2$-core of Q^n, denoted by V^n. We can think of V^n as a partial gluing of the 2^n

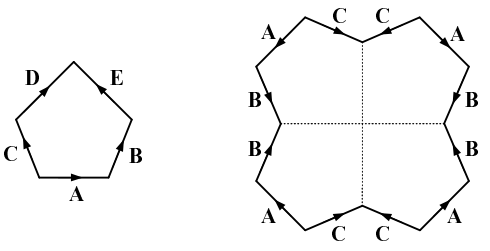


Figure 5. A $\mathbb{Z}_2$-core of a small cover in dimension 2.

copies of P^n according to the rule in (19), except that we leave the facets $F_1, \ldots, F_k$ in each copy of P^n open (see Figure 5 for an example). Let $\zeta : P^n \times (\mathbb{Z}_2)^n \to V^n$ denote the quotient map and let $P_1, \ldots, P_k$ be the panels of V^n corresponding to $\pi_\mu^{-1}(F_1), \ldots, \pi_\mu^{-1}(F_k)$. Then each P_i consists of 2^n copies of F_i, and for all $p \in F_i$ and $w \in (\mathbb{Z}_2)^n$, the involutive panel structure $\{\tau_i : P_i \to P_i\}_{1 \le i \le k}$ on V^n can be written

$$\tau_i(\zeta(p, w)) = \zeta(p, w + \mu(F_i)). \tag{20}$$

Obviously, each τ_i extends to an automorphism $\widetilde{\tau}_i$ of V^n given by the same form: for all $p \in P^n$ and $w \in (\mathbb{Z}_2)^n$,

$$\widetilde{\tau}_i(\zeta(p, w)) = \zeta(p, w + \mu(F_i)). \tag{21}$$

These $\widetilde{\tau}_i$ commute with each other; that is, $\widetilde{\tau}_i \circ \widetilde{\tau}_j = \widetilde{\tau}_j \circ \widetilde{\tau}_i$, for $1 \le i, j \le k$. So each $\widetilde{\tau}_i$ preserves any panel P_j of V^n.

To prove Theorem 1.4, it suffices to show that $\operatorname{hrk}(M(V^n, \lambda), \mathbb{Z}_2) \ge 2^m$ for any $\lambda \in \operatorname{Col}_m(V^n)$, because of Theorem 2.4.

We assume $m = k$. Let λ_0 be a maximally independent $(\mathbb{Z}_2)^k$-coloring of V^n; that is, $\operatorname{rank}(\lambda_0) = k$. By Lemma 2.10, $\operatorname{hrk}(M(V^n, \lambda), \mathbb{Z}_2) \ge \operatorname{hrk}(M(V^n, \lambda_0), \mathbb{Z}_2)$ for all $\lambda \in \operatorname{Col}_k(V^n)$. So it suffices to prove that

$$\operatorname{hrk}(M(V^n, \lambda_0), \mathbb{Z}_2) \ge 2^k. \tag{22}$$

Inequality (22) follows from Theorem 1.1 and Lemma 2.7 (see Remark 3.3 below). But here we give another proof of (22), which only uses Lemma 3.1. This proof takes advantage of the interior symmetries of small covers (see (20) and (21)), and is more natural from the viewpoint of the glue-back construction.

Because λ_0 is maximally independent, by Lemma 2.6, we can assume $\lambda_0(P_i) = e_i$ for $1 \le i \le k$, where $\{e_1, \ldots, e_k\}$ is a linear basis of $(\mathbb{Z}_2)^k$. Let $\theta_{\lambda_0} : V^n \times (\mathbb{Z}_2)^k \to M(V^n, \lambda_0)$ be the quotient map defined by (12).

Now take an arbitrary panel of V^n, say P_1, and let $M_{\setminus P_1}(V^n, \lambda_0)$ be a partial glue-back from (V^n, λ_0) defined by (17). Let $\theta_{\lambda_0}^{\setminus P_1} : V^n \times (\mathbb{Z}_2)^k \to M_{\setminus P_1}(V^n, \lambda_0)$

be the corresponding quotient map. Suppose H is the subgroup of $(\mathbb{Z}_2)^k$ generated by $\{e_2, \dots, e_k\}$. Then we define

$$Y_1 = \theta_{\lambda_0}^{\setminus P_1}(V^n \times H), \quad Y_2 = \theta_{\lambda_0}^{\setminus P_1}(V^n \times (e_1 + H)), \tag{23}$$

$$A_1 = \theta_{\lambda_0}^{\setminus P_1}(P_1 \times H), \quad A_2 = \theta_{\lambda_0}^{\setminus P_1}(P_1 \times (e_1 + H)). \tag{24}$$

Obviously, $A_1 = \partial Y_1$ and $A_2 = \partial Y_2$, and there is a homeomorphism $\Pi : Y_1 \to Y_2$ with $\Pi(A_1) = A_2$. Indeed, for all $x \in V^n$ and $h \in H$, Π is given by

$$\Pi(\theta_{\lambda_0}^{\setminus P_1}(x, h)) = \theta_{\lambda_0}^{\setminus P_1}(x, h + e_1).$$

It is easy to see that $M(V^n, \lambda_0)$ is the gluing of Y_1 and Y_2 along their boundary by a homeomorphism $\varphi : A_1 \to A_2$ defined by

$$\varphi(\theta_{\lambda_0}^{\setminus P_1}(x_1, h)) = \theta_{\lambda_0}^{\setminus P_1}(\tau_1(x_1), h + e_1),$$

for all $x_1 \in P_1$ and $h \in H$.

Also, because $\tau_1 : P_1 \to P_1$ extends to a homeomorphism $\widetilde{\tau}_1 : V^n \to V^n$ (see (20) and (21)), we can extend φ to a homeomorphism $\widetilde{\varphi} : Y_1 \to Y_2$ by

$$\widetilde{\varphi}(\theta_{\lambda_0}^{\setminus P_1}(x, h)) = \theta_{\lambda_0}^{\setminus P_1}(\widetilde{\tau}_1(x), h + e_1),$$

for all $x \in V^n$ and $h \in H$. We know $\widetilde{\varphi}$ is well-defined because $\widetilde{\tau}_1$ commutes with each τ_i on P_i (see (12) and (21)).

Identifying (Y_1, A_1) with (Y_2, A_2) via Π, we get a decomposition of $M(V^n, \lambda_0)$ that satisfies all the conditions in Lemma 3.1. So Lemma 3.1 implies that

$$\operatorname{hrk}(M(V^n, \lambda_0), \mathbb{Z}_2) \geq \operatorname{hrk}(A_1, \mathbb{Z}_2). \tag{25}$$

Also, let $q : Y_1 \cup Y_2 \to M(V^n, \lambda_0)$ be the quotient map and let

$$\xi_{\lambda_0} : M(V^n, \lambda_0) \to Q^n$$

be the orbit map of the natural $(\mathbb{Z}_2)^k$-action on $M(V^n, \lambda_0)$ (see (13)). It is easy to see that

$$A_1 \cong q(A_1) = \xi_{\lambda_0}^{-1}(\pi_\mu^{-1}(F_1)).$$

Because $\xi_{\lambda_0}^{-1}(\pi_\mu^{-1}(F_1))$ is a principal $(\mathbb{Z}_2)^k$-bundle over $\pi_\mu^{-1}(F_1)$ and $\pi_\mu^{-1}(F_1)$ is a small cover over F_1 of dimension $n-1$, we have $\operatorname{hrk}\big(\xi_{\lambda_0}^{-1}(\pi_\mu^{-1}(F_1)), \mathbb{Z}_2\big) \geq 2^k$, by the induction hypothesis. Then $\operatorname{hrk}(A_1, \mathbb{Z}_2) \geq 2^k$ also. So the case $m = k$ is confirmed, because by (25), $\operatorname{hrk}(M(V^n, \lambda_0), \mathbb{Z}_2) \geq 2^k$.

Now we assume $m < k$. Let $\iota : (\mathbb{Z}_2)^m \hookrightarrow (\mathbb{Z}_2)^k$ be the standard inclusion, and define $\widehat{\lambda} := \iota \circ \lambda$. We consider $\widehat{\lambda}$ as a $(\mathbb{Z}_2)^k$-coloring on V^n. So by the above argument, $\operatorname{hrk}(M(V^n, \widehat{\lambda}), \mathbb{Z}_2) \geq 2^k$. By Theorem 2.5, $M(V^n, \widehat{\lambda})$ consists of 2^{k-m} copies of $M(V^n, \lambda)$, so $\operatorname{hrk}(M(V^n, \lambda), \mathbb{Z}_2) \geq 2^m$.

Finally, we assume $m > k$. Because $\operatorname{rank}(\lambda) \le k$, with a proper change of basis, we can assume $L_\lambda \subset (\mathbb{Z}_2)^k \subset (\mathbb{Z}_2)^m$. Let $\varrho : (\mathbb{Z}_2)^m \to (\mathbb{Z}_2)^k$ be the standard projection. Define $\bar{\lambda} := \varrho \circ \lambda$. Similarly, we consider $\bar{\lambda}$ as a $(\mathbb{Z}_2)^k$-coloring on V^n, and so we have $\operatorname{hrk}(M(V^n, \bar{\lambda}), \mathbb{Z}_2) \ge 2^k$. By Theorem 2.5, $M(V^n, \lambda)$ consists of 2^{m-k} copies of $M(V^n, \bar{\lambda})$, so $\operatorname{hrk}(M(V^n, \lambda), \mathbb{Z}_2) \ge 2^m$.

So for any $m \ge 1$ and $\lambda \in \operatorname{Col}_m(V^n)$, we always have $\operatorname{hrk}(M(V^n, \lambda), \mathbb{Z}_2) \ge 2^m$. The induction is complete. □

Remark 3.3. Both $M(V^n, \lambda_0)$ and $\mathbb{R}\mathcal{Z}_{P^n}$ are connected principal $(\mathbb{Z}_2)^k$-bundles over Q^n. Then by Lemma 2.7, $M(V^n, \lambda_0)$ is homeomorphic to $\mathbb{R}\mathcal{Z}_{P^n}$. So the conclusion of Theorem 1.1 also tells us that $\operatorname{hrk}(M(V^n, \lambda_0), \mathbb{Z}_2) \ge 2^k$.

A crucial step in this proof is that when $\lambda_0 \in \operatorname{Col}_k(V^n)$ is maximally independent, we can always get the type of decomposition of $M(V^n, \lambda_0)$ as in Lemma 3.1, which allows us to use the induction hypothesis. However, for an arbitrary $\lambda \in \operatorname{Col}_k(V^n)$, this type of decomposition of $M(V^n, \lambda)$ may not exist (at least not obviously).

For example, in the lower picture in Figure 2, we have a principal $(\mathbb{Z}_2)^2$-bundle $\pi : M^2 \to T^2$, where M^2 is a disjoint union of two tori. The union of the two meridians in M^2 is the inverse image of a meridian in T^2 under π. If we cut M^2 open along these two meridians, we get two circular cylinders. But M^2 is not obtained by gluing these two cylinders together, because the colors of the $(\mathbb{Z}_2)^2$-coloring on the two panels are not linearly independent. So the construction in (23) for this case fails to give us the type of decomposition of M^2 as in Lemma 3.1.

So when $\lambda \in \operatorname{Col}_k(V^n)$ is not maximally independent, we may not be able to directly apply the induction hypothesis to $M(V^n, \lambda)$ as we do to $M(V^n, \lambda_0)$ above. But these cases are settled by Lemma 2.10.

Acknowledgement

The author thanks B. Hanke for Remark 3.2.

References

[Adem 1987] A. Adem, "$\mathbf{Z}/p\mathbf{Z}$ actions on $(S^n)^k$", *Trans. Amer. Math. Soc.* **300**:2 (1987), 791–809. MR 88b:57037 Zbl 0623.57025

[Adem 2004] A. Adem, "Constructing and deconstructing group actions", pp. 1–8 in *Homotopy theory: relations with algebraic geometry, group cohomology, and algebraic K-theory* (Evanston, IL, 2002), edited by P. Goerss and S. Priddy, Contemp. Math. **346**, American Mathematical Society, Providence, RI, 2004. MR 2005d:57048 Zbl 1101.57017 arXiv math.AT/0212280

[Adem and Benson 1998] A. Adem and D. J. Benson, "Elementary abelian groups acting on products of spheres", *Math. Z.* **228**:4 (1998), 705–712. MR 99k:57033 Zbl 0913.57020

[Adem and Browder 1988] A. Adem and W. Browder, "The free rank of symmetry of $(S^n)^k$", *Invent. Math.* **92**:2 (1988), 431–440. MR 89e:57034 Zbl 0644.57022

[Allday and Puppe 1993] C. Allday and V. Puppe, *Cohomological methods in transformation groups*, Cambridge Stud. in Adv. Math. **32**, Cambridge University Press, Cambridge, 1993. MR 94g:55009 Zbl 0799.55001

[Buchstaber and Panov 2002] V. M. Buchstaber and T. E. Panov, *Torus actions and their applications in topology and combinatorics*, University Lecture Ser. **24**, American Mathematical Society, Providence, RI, 2002. MR 2003e:57039 Zbl 1012.52021

[Cao and Lü 2009] X. Cao and Z. Lü, "Möbius transform, moment-angle complex and Halperin–Carlsson conjecture", preprint, 2009. arXiv 0908.3174

[Carlsson 1982] G. Carlsson, "On the rank of abelian groups acting freely on $(S^n)^k$", *Invent. Math.* **69**:3 (1982), 393–400. MR 84e:57033 Zbl 0517.57020

[Carlsson 1986] G. Carlsson, "Free $(\mathbf{Z}/2)^k$-actions and a problem in commutative algebra", pp. 79–83 in *Transformation groups* (Poznań, 1985), edited by S. Jackowski and K. Pawalowski, Lecture Notes in Math. **1217**, Springer, Berlin, 1986. MR 88g:57042 Zbl 0614.57023

[Carlsson 1987] G. Carlsson, "Free $(\mathbf{Z}/2)^3$-actions on finite complexes", pp. 332–344 in *Algebraic topology and algebraic K-theory* (Princeton, NJ, 1983), edited by W. Browder, Ann. of Math. Stud. **113**, Princeton University Press, 1987. MR 89g:57054 Zbl 0701.55003

[Conner 1957] P. E. Conner, "On the action of a finite group on $S^n \times S^n$", *Ann. of Math.* (2) **66**:3 (1957), 586–588. MR 20 #2725 Zbl 0079.38904

[Davis 1983] M. W. Davis, "Groups generated by reflections and aspherical manifolds not covered by Euclidean space", *Ann. of Math.* (2) **117**:2 (1983), 293–324. MR 86d:57025 Zbl 0531.57041

[Davis and Januszkiewicz 1991] M. W. Davis and T. Januszkiewicz, "Convex polytopes, Coxeter orbifolds and torus actions", *Duke Math. J.* **62**:2 (1991), 417–451. MR 92i:52012 Zbl 0733.52006

[Halperin 1985] S. Halperin, "Rational homotopy and torus actions", pp. 293–306 in *Aspects of topology*, edited by I. M. James and E. H. Kronheimer, London Math. Soc. Lecture Note Ser. **93**, Cambridge University Press, Cambridge, 1985. MR 87d:55001 Zbl 0562.57015

[Hanke 2009] B. Hanke, "The stable free rank of symmetry of products of spheres", *Invent. Math.* **178**:2 (2009), 265–298. MR 2010k:57070 Zbl 1177.57028

[Hatcher 2002] A. Hatcher, *Algebraic topology*, Cambridge University Press, Cambridge, 2002. MR 2002k:55001 Zbl 1044.55001

[Jänich 1968] K. Jänich, "On the classification of $O(n)$-manifolds", *Math. Ann.* **176**:1 (1968), 53–76. MR 37 #2261 Zbl 0153.53801

[Puppe 2009] V. Puppe, "Multiplicative aspects of the Halperin–Carlsson conjecture", *Georgian Math. J.* **16**:2 (2009), 369–379. MR 2010j:55008 Zbl 1190.57023 arXiv 0811.3517

[Ustinovskii 2011] Y. M. Ustinovskii, "Гилотеза о тороческом ранге для момент-угол комплексов", *Mat. Zametki* **90**:2 (2011), 300–305. Translated as "Toral rank conjecture for moment-angle complexes" in *Math. Notes* **90**:2 (2011), 279–283. Zbl 06014007

[Yu 2012] L. Yu, "On the constructions of free and locally standard $\mathbb{Z}_2$-torus actions on manifolds", *Osaka J. Math.* **49** (2012). To appear. arXiv 1001.0289

Received March 1, 2011. Revised November 15, 2011.

LI YU
DEPARTMENT OF MATHEMATICS AND IMS
NANJING UNIVERSITY
NANJING, 210093
CHINA

yuli@nju.edu.cn

PACIFIC JOURNAL OF MATHEMATICS
Vol. 256, No. 2, 2012

ACKNOWLEDGEMENT

The editors gratefully acknowledge the valuable advice of the referees who helped them select and better the papers appearing in 2011 in the *Pacific Journal of Mathematics* (reports dated December 16, 2010 through December 31, 2011).

Tetsuya Abe, Ernst Albrecht, Gergley Ambrus, Ben Andrews, Chengming Bai, Erik P. van den Ban, Jonathan Beck, Mikhail Belolipetsky, Alex James Bene, Bruno Benedetti, Michel van den Bergh, Bruce C. Berndt, Daniel Bertrand, Hakima Bessaih, Collin Bleak, Christian Bonatti, Cédric Bonnafé, Steven Boyer, Robert F. Brown, Jonathan Brundan, Ryan Budney, Daniel Bump, Xiaodong Cao, Jianguo Cao, Claudio Carmeli, Sagun Chanillo, Wenxiong Chen, Bang-Yen Chen, Boyong Chen, Shun Jen Cheng, Luke Cherveny, Sangbum Cho, Stephen K. K. Choi, Dan Coman, Caterina Consani, Stéphanie Cupit-Foutou, Mimi Dai, Marcos Dajczer, Benoît Daniel, Donatella Danielli, Bernard Deconinck, Manuel del Pino, Pierre Deligne, Martin Deraux, Qing Ding, Chongying Dong, Slaviša Đorđević, Cristina Draper, Klaus Ecker, Ed Effros, Richard S. Elman, Hélène Esnault, Mario Eudave-Muñoz, Hao Fang, Anna Fino, Joel Foisy, Gerald Folland, Dmitry Fuchs, David Futer, Jorge García Melián, Christof Geiß, Eknath Ghate, Dragos Ghioca, Olga Gil-Medrano, Ernesto Girondo, Frederick M. Goodman, Robert Greene, Wolter Groenevelt, Nicolas Guay, Yuxia Guo, Christopher D. Hacon, Mark Hamilton, Emily Hamilton, Chenxu He, James J. Hebda, Nigel Higson, Toshiaki Hishida, Naihong Hu, Wen Huang, Johannes Huebschmann, Chris Hughes, Stephen Humphries, Jenn-Fang Hwang, Paltin Ionescu, Keiji Izuchi, Michael Jablonski, Louis Jeanjean, Shanyu Ji, David L. Johnson, Robert Kerr, Apoorva Khare, Juha Kinunnen, Maciej Klimek, Mariusz Koras, Dieter Kotschick, Shintaro Kuroki, Joshua Lansky, Jorge Lauret, Claude LeBrun, Young Joo Lee, Yankı Lekili, David Shea Letscher, Congming Li, Haisheng Li, Junfang Li, Song-Ying Li, Fang Li, Charles Livingston, Jiang-Hua Lu, Garving K. Luli, Tao Luo, Feng Luo, Xiaonan Ma, Li Ma, Kirill Mackenzie, Robert Marsh, Wacław Marzantowicz, David McReynolds, Alexander Mednykh, William H. Meeks, III, Alexander Merkurjev, Michal Misiurewicz, Xiaohuan Mo, Christopher Mooney, Dijana Mosić, Vicente Muñoz, Fiona Murnaghan, Tomoki Nakanishi, Sergey Neshveyev, Alexandru Oancea, Masato Okado, Viktor Ostrik, Yelin Ou, Biao Ou, Shengliang Pan, Angela Pasquale, Julien Paupert, Pan Peng, Oscar Perdomo, Carlo Petronio, Raphaël Ponge, Wladimir de Azevedo Pribitkin, Volker Puppe, Mihai Putinar, Sundararaman

Ramanan, Kulumani M. Rangaswamy, Jean Renault, Manuel Ritoré, Paul C. Roberts, Magdalena Rodriguez, David E. Rohrlich, Antonio Ros, Wolfgang Ruppert, Joshua Sabloff, Kentaro Saji, Isabel M. C. Salavessa, Ralf Schmidt, Markus Scholz, Natasa Sesum, Zhongmin Shen, Vladimir Shpilrain, Carlos Simpson, Srinivasa Varadhan, Shaun Stevens, András I. Stipsicz, Jeffrey Streets, Yucai Su, Toshiyuki Sugawa, Junecue Suh, András Szenes, Tadie, Maneesh Thakur, Giuseppe Tinaglia, Peter Tingley, David Shea Vela-Vick, Dietmar Vogt, Luc Vrancken, Adrian Wadsworth, Emmanuel Wagner, Konrad Waldorf, Alden Walker, Changyou Wang, Changping Wang, Qiaoling Wang, Bing Wang, Jiaping Wang, Walter Wei, Guofang Wei, Ben Weinkove, Jean-Yves Welschinger, Henry Wente, Jerzy Weyman, Michael Wiemeler, Bert Wiest, Kenneth S. Williams, Philip Pit Wang Wong, Nancy Court Wrinkle, Siye Wu, William Wylie, Changyu Xia, Yuanlong Xin, Hao Xu, Feng Xu, Aiichi Yamasaki, Paul Yang, Sai Kee Yeung, Fyodor Zak, Xi Zhang, Xiaoyi Zhang, En-Tao Zhao, Detang Zhou, Nguyen Tien Zung.

CONTENTS

Volume 256, no. 1 and no. 2

Guidelines for Authors

Authors may submit manuscripts at msp.berkeley.edu/pjm/about/journal/submissions.html and choose an editor at that time. Exceptionally, a paper may be submitted in hard copy to one of the editors; authors should keep a copy.

By submitting a manuscript you assert that it is original and is not under consideration for publication elsewhere. Instructions on manuscript preparation are provided below. For further information, visit the web address above or write to pacific@math.berkeley.edu or to Pacific Journal of Mathematics, University of California, Los Angeles, CA 90095–1555. Correspondence by email is requested for convenience and speed.

Manuscripts must be in English, French or German. A brief abstract of about 150 words or less in English must be included. The abstract should be self-contained and not make any reference to the bibliography. Also required are keywords and subject classification for the article, and, for each author, postal address, affiliation (if appropriate) and email address if available. A home-page URL is optional.

Authors are encouraged to use LaTeX, but papers in other varieties of TeX, and exceptionally in other formats, are acceptable. At submission time only a PDF file is required; follow the instructions at the web address above. Carefully preserve all relevant files, such as LaTeX sources and individual files for each figure; you will be asked to submit them upon acceptance of the paper.

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used but tags will be converted to the house format (see a current issue for examples).

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple or Mathematica. Many drawing tools such as Adobe Illustrator and Aldus FreeHand can produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, the authors should check with production by sending an email to pacific@math.berkeley.edu.

Each figure should be captioned and numbered, so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript will all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables, which should be used sparingly.

Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Page proofs will be made available to authors (or to the designated corresponding author) at a website in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

PACIFIC JOURNAL OF MATHEMATICS

Volume 256 No. 2 April 2012

0030-8730(201204)256:2;1-8