Pacific Journal of Mathematics

Volume 263 No. 2 June 2013

PACIFIC JOURNAL OF MATHEMATICS

msp.org/pjm

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

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

The subscription price for 2013 is US $400/year for the electronic version, and $485/year for print and electronic.
Subscriptions, requests for back issues and changes of subscribers address should be sent to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94704-0163, U.S.A. The Pacific Journal of Mathematics is indexed by Mathematical Reviews, Zentralblatt MATH, PASCAL CNRS Index, Referativnyi Zhurnal, Current Mathematical Publications and the Science Citation Index.

The Pacific Journal of Mathematics (ISSN 0030-8730) at the University of California, c/o Department of Mathematics, 798 Evans Hall #3840, Berkeley, CA 94720-3840, is published 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

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

PACIFIC JOURNAL OF MATHEMATICS
Vol. 263, No. 2, 2013
dx.doi.org/10.2140/pjm.2013.263.257

REALIZATIONS OF BC_r-GRADED INTERSECTION MATRIX ALGEBRAS WITH GRADING SUBALGEBRAS OF TYPE B_r, $r \geq 3$

SANDEEP BHARGAVA AND YUN GAO

We study intersection matrix algebras $\mathfrak{im}(A^{[d]})$ that arise from affinizing a Cartan matrix A of type B_r with d arbitrary long roots in the root system Δ_{B_r}, where $r \geq 3$. We show that $\mathfrak{im}(A^{[d]})$ is isomorphic to the universal covering algebra of $\mathfrak{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$, where $\mathfrak{a}$ is an associative algebra with involution η, and C is an $\mathfrak{a}$-module with hermitian form χ. We provide a description of all four of the components $\mathfrak{a}$, η, C, and χ.

1. Introduction

Peter Slodowy [1984; 1986] discovered that matrices like

$$M = \begin{bmatrix} 2 & -1 & 0 & 1 \\ -1 & 2 & -1 & 1 \\ 0 & -2 & 2 & -2 \\ 1 & 1 & -1 & 2 \end{bmatrix}$$

encode the intersection form on the second homology group of Milnor fibers for germs of holomorphic maps with an isolated singularity at the origin. These matrices were like the generalized Cartan matrices of Kac–Moody theory in that they had integer entries, 2's along the diagonal, and M_{ij} was negative if and only if M_{ji} was negative. What was new, however, was the presence of positive entries off the diagonal. Slodowy called such matrices generalized intersection matrices:

Definition 1 [Slodowy 1986]. An $n \times n$ integer-valued matrix M is called a *generalized intersection matrix* ($\mathfrak{gim}$) if the following conditions are satisfied. whenever $1 \leq i, j \leq n$ with $i \neq j$:

$$\begin{gathered} M_{ii} = 2; \\ M_{ij} < 0 \iff M_{ji} < 0; \\ M_{ij} > 0 \iff M_{ji} > 0. \end{gathered}$$

Funding from the National Sciences and Engineering Research Council of Canada is gratefully acknowledged.

MSC2010: primary 17B65, 17B70, 17B05; secondary 17B67, 16W10.

Keywords: Lie algebras, intersection matrix algebras.

Slodowy used these matrices to define a class of Lie algebras that encompassed all the Kac–Moody Lie algebras:

Definition 2 [Slodowy 1986; Berman and Moody 1992]. Given an $n \times n$ generalized intersection matrix $M = (M_{ij})$, define a Lie algebra over $\mathbb{C}$, called a *generalized intersection matrix (*gim*) algebra* and denoted by $\mathfrak{gim}(M)$, with

generators: $e_1, \ldots, e_n, f_1, \ldots, f_n, h_1, \ldots h_n$,

relations:

(R1) For $1 \le i, j \le n$,

$$[h_i, e_j] = M_{ij} e_j,$$
$$[h_i, f_j] = -M_{ij} f_j,$$
$$[e_i, f_i] = h_i.$$

(R2) For $M_{ij} \le 0$,

$$[e_i, f_j] = 0 = [f_i, e_j],$$
$$(\operatorname{ad} e_i)^{-M_{ij}+1} e_j = 0 = (\operatorname{ad} f_i)^{-M_{ij}+1} f_j.$$

(R3) For $M_{ij} > 0$, $i \ne j$,

$$[e_i, e_j] = 0 = [f_i, f_j],$$
$$(\operatorname{ad} e_i)^{M_{ij}+1} f_j = 0 = (\operatorname{ad} f_i)^{M_{ij}+1} e_j.$$

If the M that we begin with is a generalized Cartan matrix, then the $3n$ generators and the first two groups of axioms, (R1) and (R2), provide a presentation of the Kac–Moody Lie algebras [Gabber and Kac 1981; Kac 1990; Carter 2005].

Slodowy [1986] and, later, Berman [1989] showed that the $\mathfrak{gim}$ algebras are also isomorphic to fixed point subalgebras of involutions on larger Kac–Moody algebras. So, in their words, the $\mathfrak{gim}$ algebras lie both "beyond and inside" Kac–Moody algebras.

Further progress came in the 1990s as a byproduct of work on the classification of root-graded Lie algebras [Berman and Moody 1992; Benkart and Zelmanov 1996; Neher 1996], which revealed that some families of intersection matrix ($\mathfrak{im}$) algebras, which are quotient algebras of $\mathfrak{gim}$ algebras, were universal covering algebras of well-understood Lie algebras. For instance the $\mathfrak{im}$ algebras that arise from multiply affinizing a Cartan matrix of type A_r, with $r \ge 3$, are the universal covering algebras of $\mathrm{sl}(\mathfrak{a})$, where $\mathfrak{a}$ is the associative algebra of noncommuting Laurent polynomials in several variables (the number of indeterminates depends on how many times the original Cartan matrix is affinized). A handful of other researchers also began engaging these new algebras. For example, Eswara Rao, Moody, and Yokonuma [Rao et al. 1992] used vertex operator representations to show that $\mathfrak{im}$ algebras were nontrivial. Gao [1996] examined compact forms of $\mathfrak{im}$ algebras arising from conjugations over the complex field. Peng [2002] found relations between $\mathfrak{im}$ algebras and the representations of tilted algebras via

Ringel–Hall algebras. Berman, Jurisich, and Tan [Berman et al. 2001] showed that the presentation of $\mathfrak{gim}$ algebras could be put into a broader framework that incorporated Borcherds algebras.

The chief objective of this paper is to continue advancing our understanding of $\mathfrak{gim}$ and $\mathfrak{im}$ algebras. We construct a generalized intersection matrix $A^{[d]}$ by adjoining d long roots to a base of a root system of type B_r, where $r \geq 3$. This is exactly the analogue of the affinization process in which a single root is adjoined to a Cartan matrix of a finite-dimensional Lie algebra to arrive at a generalized Cartan matrix and, eventually, an affine Kac–Moody algebra. The matrix $A^{[d]}$ is used to define a $\mathfrak{gim}$ algebra $\mathfrak{gim}(A^{[d]})$. Since $\mathfrak{gim}(A^{[d]})$ may possess roots with mixed signs, we quotient out by an ideal $\mathfrak{r}$ that is tailor-made to capture all such roots. The quotient algebra is called the intersection matrix algebra and is denoted by $\mathfrak{im}(A^{[d]})$.

We show that $\mathfrak{im}(A^{[d]})$ is a BC_r-graded Lie algebra, which, in turn, allows us to invoke Allison, Benkart, and Gao's recognition theorem [Allison et al. 2002] and relate $\mathfrak{im}(A^{[d]})$ to an algebraic structure that is better understood. Combining their theorem with the knowledge that $\mathfrak{im}(A^{[d]})$ is centrally closed, we conclude that, up to isomorphism, $\mathfrak{im}(A^{[d]})$ is the universal covering algebra of $\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$. The algebra $\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$ is like the usual matrix model $\mathrm{so}_{2r+1}(\mathbb{C})$ of a finite-dimensional Lie algebra of type B_r, except that we now replace the field $\mathbb{C}$ with an associative algebra $\mathfrak{a}$, which possesses an involution (that is, period two antiautomorphism) η, and we involve a right $\mathfrak{a}$-module C that has a hermitian form $\chi : C \times C \to \mathfrak{a}$. The defining relations of the generalized intersection matrix algebra and, hence, the intersection matrix algebra, in concert with the existence of a central, graded, surjective Lie algebra homomorphism ψ from $\mathfrak{im}(A^{[d]})$ to $\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$ allow us to understand each of $\mathfrak{a}$, η, C, and χ. For example, we get (i) two generators of $\mathfrak{a}$, namely x and x^{-1}, for every long root of the form $\pm(\epsilon_i + \epsilon_{i+1})$, and (ii) four generators of $\mathfrak{a}$, namely y, y^{-1}, z, and z^{-1}, for every other type of long root that we adjoin. We are also able to study the relations among the generators, determine the action of the involution η, and discover that $C = 0$ and $\chi = 0$. Through constructing a surjective Lie algebra homomorphism $\varphi : \mathfrak{gim}(A^{[d]}) \to \mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$ we verify that we indeed have a complete description of the "coordinate algebra" $\mathfrak{a}$.

Our work continues the line of research initiated by Berman, Moody, Benkart and Zelmanov. Berman and Moody [1992] were the first to find realizations of intersection matrix algebras over Lie algebras graded by root systems of types A_r ($r \geq 2$), D_r, E_6, E_7, and E_8. Benkart and Zelmanov [1996] found realizations of intersection matrix algebras over Lie algebras graded by root systems of types A_1, B_r, C_r, F_4, and G_2. In this paper, we find realizations of intersection matrix algebras over Lie algebras graded by root systems of type BC_r with grading subalgebras of type B_r ($r \geq 3$).

2. Multiply affinizing Cartan matrices

In this paper, we focus on generalized intersection matrix algebras that arise from multiply affinizing a Cartan matrix of type B_r, where $r \geq 3$, with long roots in the root system Δ_{B_r}.

Consider a root system of type B_r. Up to isomorphism, Δ_{B_r} may be described as

$$\Delta_{B_r} = \{\pm\epsilon_i \pm \epsilon_j : 1 \leq i \neq j \leq r\} \cup \{\pm\epsilon_i : i = 1, \ldots, r\}.$$

Once we fix an ordering of the simple roots $\alpha_1, \ldots, \alpha_r$ in a base Π, the Cartan matrix A is described by

$$A_{ij} = \frac{2(\alpha_i, \alpha_j)_{\text{Killing}}}{(\alpha_i, \alpha_i)_{\text{Killing}}} \quad \text{for } 1 \leq i, j \leq r.$$

Choose any d long roots in Δ_{B_r}, say $\alpha_{r+1}, \ldots, \alpha_{r+d}$, and consider the $r+d$ by $r+d$ matrix $A^{[d]}$ given by

$$A^{[d]}_{ij} = \frac{2(\alpha_i, \alpha_j)_{\text{Killing}}}{(\alpha_i, \alpha_i)_{\text{Killing}}} \quad \text{for } 1 \leq i, j \leq r+d,$$

with respect to the ordering $(\alpha_1, \ldots, \alpha_r, \alpha_{r+1}, \ldots, \alpha_{r+d})$ of the r roots in the base Π plus the d "adjoined" roots. The axioms of a root system tell us that all the entries of $A^{[d]}$ are integers. Moreover, since the Killing form is symmetric, we have $A^{[d]}_{ji} = 0$ if $A^{[d]}_{ij} = 0$, or if $A^{[d]}_{ij}$ and $A^{[d]}_{ji}$ are nonzero, then they share the same sign. In other words, $A^{[d]}$ is a generalized intersection matrix.

Since the "d-affinized" Cartan matrix $A^{[d]}$ is a generalized intersection matrix, $\mathfrak{gim}(A^{[d]})$ is a generalized intersection matrix algebra.

Note that if we affinize the Cartan matrix A of type B_r with the negative of the highest long root of Δ_{B_r} then the resulting generalized intersection matrix algebra $\mathfrak{gim}(A^{[1]})$ is the affine Kac–Moody Lie algebra of type $B_r^{(1)}$.

3. Intersection matrix algebras

Fix a Cartan matrix A of type B_r ($r \geq 3$) with, say, $\alpha_1, \alpha_2, \ldots, \alpha_r$ being the simple roots in a base of Δ_{B_r} that were used to form A. Let

- $\Omega =$ set of all long roots of the form $\pm(\epsilon_i + \epsilon_{i+1})$ that we adjoin,
- $\Theta =$ set of all remaining long roots that are adjoined,
- $N_\mu =$ the number of copies of the long root μ we have adjoined, and
- $d = \sum_{\mu \in \Omega \cup \Theta} N_\mu$.

Let $A^{[d]}$ be the resulting generalized intersection matrix and $\mathfrak{gim}(A^{[d]})$ the corresponding generalized intersection matrix algebra.

We begin a move towards a quotient algebra of $\mathfrak{gim}(A^{[d]})$ using a slight generalization of the work done by Benkart and Zelmanov [1996]. Let Γ be the integer lattice generated by the Δ, where

$$\Delta = \{\pm\epsilon_i \pm \epsilon_j : 1 \le i \ne j \le r\} \cup \{\pm\epsilon_i, \pm 2\epsilon_i : i = 1, \ldots, r\}$$

is a root system of type BC_r.

We define a Γ-grading on $\mathfrak{gim}(A^{[d]})$ as follows:

$$\deg e_i = \alpha_i = -\deg f_i, \quad \deg h_i = 0$$

for $i = 1, \ldots, r$, and

$$\deg e_{\mu,i} = \mu = -\deg f_{\mu,i}, \quad \deg h_{\mu,i} = 0$$

for $\mu \in \Omega \cup \Theta$ and $i = 1, \ldots, N_\mu$.

Next, we define the *radical* $\mathfrak{r}$ of $\mathfrak{gim}(A^{[d]})$ to be the ideal generated by the root spaces $\mathfrak{gim}(A^{[d]})_\gamma$ where $\gamma \notin \Delta \cup \{0\}$. Since the ideal $\mathfrak{r}$ is homogeneous, the resulting quotient algebra

$$\mathfrak{im}(A^{[d]}) = \mathfrak{gim}(A^{[d]})/\mathfrak{r}$$

is also Γ-graded. Moreover,

$$\mathfrak{im}(A^{[d]})_\gamma = 0 \quad \text{if } \gamma \notin \Delta \cup \{0\}.$$

We call $\mathfrak{im}(A^{[d]})$ the *intersection matrix* ($\mathfrak{im}$) *algebra* corresponding to the generalized intersection matrix algebra $\mathfrak{gim}(A^{[d]})$.

3.1. $\mathfrak{im}(A^{[d]})$ *is* BC_r*-graded.* Allison, Benkart, and Gao gave the following definition of a Lie algebra graded by a root system of type BC.

Definition 3 [Allison et al. 2002]. Let r be a positive integer greater than or equal to 3. A Lie algebra L over $\mathbb{C}$ is *graded by the root system* BC_r or is BC_r*-graded* with a *grading subalgebra of type* B_r if

(i) L contains, as a subalgebra, a finite-dimensional simple Lie algebra $\mathfrak{g}$ whose root system relative to a Cartan subalgebra $\mathfrak{h} = \mathfrak{g}_0$ is Δ_{B_r},

(ii) $L = \bigoplus_{\mu \in \Delta \cup \{0\}} L_\mu$, where $L_\mu = \{x \in L \mid [h, x] = \mu(h)x \text{ for all } h \in \mathfrak{h}\}$ for $\mu \in \Delta \cup \{0\}$, and Δ is the root system of type BC_r, and

(iii) $L_0 = \sum_{\mu \in \Delta} [L_\mu, L_{-\mu}]$.

Proposition 4. *The algebra* $\mathfrak{im}(A^{[d]})$ *is* BC_r*-graded with a grading subalgebra of type* B_r.

Proof. The subalgebra in $\mathfrak{im}(A^{[d]})$ generated by $e_1 + \mathfrak{r}, \ldots, h_r + \mathfrak{r}$, due to the relations on these elements induced by the relations on their preimages in $\mathfrak{gim}(A^{[d]})$, is isomorphic to a finite-dimensional simple Lie algebra $\mathfrak{g}$ of type B_r. We have already shown in Section 3.1 that $\mathfrak{im}(A^{[d]})$ is Γ-graded with $\mathfrak{im}(A^{[d]})_\gamma = 0$ if $\gamma \notin \Delta \cup \{0\}$. That is,

$$\mathfrak{im}(A^{[d]}) = \bigoplus_{\mu \in \Delta \cup \{0\}} \mathfrak{im}(A^{[d]})_\mu.$$

Finally, our initial degree assignments for the generators of $\mathfrak{gim}(A^{[d]})$, the $\mathfrak{gim}$ algebra relations like $h_i = [e_i, f_i]$ and $h_\mu = [e_\mu, f_\mu]$, and the fact that movement into the 0 root space can only occur by bracketing an element from an $\mathfrak{im}(A^{[d]})_\mu$ space with one from the $\mathfrak{im}(A^{[d]})_{-\mu}$ space all combine to lead us to the conclusion that

$$\mathfrak{im}(A^{[d]})_0 = \sum_{\mu \in \Delta} \big[\mathfrak{im}(A^{[d]})_\mu, \mathfrak{im}(A^{[d]})_{-\mu}\big]. \qquad \square$$

3.2. $\mathfrak{im}(A^{[d]})$ *is centrally closed.* Recall that a Lie algebra L is said to be perfect if it equals its derived algebra, that is, $L = [L, L]$. Furthermore, if L is perfect and is its own universal covering then we say that L is centrally closed [Moody and Pianzola 1995].

Proposition 5. *The algebra* $\mathfrak{gim}(A^{[d]})$ *is a perfect Lie algebra.*

Proof. Being a Lie algebra, $\mathfrak{gim}(A^{[d]})$ is closed under the operation of taking brackets; hence $\big[\mathfrak{gim}(A^{[d]}), \mathfrak{gim}(A^{[d]})\big] \subset \mathfrak{gim}(A^{[d]})$. To show the reverse inclusion, it suffices to show that all of the generators of $\mathfrak{gim}(A^{[d]})$ lie in $\big[\mathfrak{gim}(A^{[d]}), \mathfrak{gim}(A^{[d]})\big]$. But this is indeed the case because the generators e_i, f_i, h_i (for $1 \le i \le r$) and the $e_{\mu,i}$, $f_{\mu,i}$, $h_{\mu,i}$, which arise from adjoining the i-th copy of a long root μ, satisfy the relations (R1) of Definition 2. $\square$

Our next theorem is Proposition 1.6 in [Benkart and Zelmanov 1996] adapted to our context.

Theorem 6. *The algebra* $\mathfrak{im}(A^{[d]})$ *is centrally closed.*

Proof. Let $\big(\widetilde{U}, \phi\big)$ be the universal covering algebra of $\mathfrak{im}(A^{[d]})$. Let $\mathfrak{g}$ be the simple finite dimensional subalgebra of type B contained in $\mathfrak{im}(A^{[d]})$ with Cartan subalgebra $\mathfrak{h}$ whose root space decomposition induces a BC-gradation on $\mathfrak{im}(A^{[d]})$. The preimage $\phi^{-1}(\mathfrak{h})$ of $\mathfrak{h}$ contains $\ker\phi$. Since ϕ is a central map, $\ker\phi$ lies in the center of $\widetilde{U}$. So

$$\mathfrak{h}' = \phi^{-1}(\mathfrak{h})/\ker\phi$$

acts on $\widetilde{U}$ via the adjoint action. If $h' \in \mathfrak{h}'$, $\phi(h') = h \in \mathfrak{h}$, and $\mu(t) \in \mathbb{C}[t]$ is the minimal polynomial of $\mathrm{ad}_{U_L}(h)$, then

$$\mu\big(\mathrm{ad}_{\widetilde{U}}(h')\big)\big(\widetilde{U}\big) \subset \ker\phi.$$

So $\mathrm{ad}_{\widetilde{U}}(h')$ satisfies the polynomial $t\mu(t)$. Therefore $\widetilde{U}$ is a sum of root spaces with respect to $\mathrm{ad}_{\widetilde{U}}\,\mathfrak{h}'$, and $\widetilde{U}_\gamma \neq (0)$ if and only if $\gamma \in \Delta \cup \{0\}$. So ϕ induces an isomorphism between the nonzero root spaces of $\widetilde{U}$ and those of $\mathrm{im}(A^{[d]})$. Moreover,

$$\widetilde{U}_0 = \sum_{\gamma\in\Delta} \big[\widetilde{U}_{-\gamma}, \widetilde{U}_\gamma\big] + \ker\phi \quad \text{implies that} \quad \big[\widetilde{U}_0, \widetilde{U}_0\big] \subset \sum_{\gamma\in\Delta} \big[\widetilde{U}_{-\gamma}, \widetilde{U}_\gamma\big].$$

Since $\widetilde{U} = \big[\widetilde{U}, \widetilde{U}\big]$, it follows that

$$\widetilde{U}_0 = \big[\widetilde{U}_0, \widetilde{U}_0\big] + \sum_{\gamma\in\Delta} \big[\widetilde{U}_{-\gamma}, \widetilde{U}_\gamma\big] = \sum_{\gamma\in\Delta} \big[\widetilde{U}_{-\gamma}, \widetilde{U}_\gamma\big].$$

Consequently, ϕ is an isomorphism. □

4. Recognition theorem

The following construction, given in Example 1.23 of [Allison et al. 2002], is a more general version of the classical construction of $\mathrm{so}_{2r+1}(\mathbb{C})$, the simple Lie algebra of type B_r.

Let r be a positive integer, $\mathfrak{a}$ be a unital associative algebra over $\mathbb{C}$ with an involution (that is, period two antiautomorphism) η, C be a right $\mathfrak{a}$-module with a hermitian form $\chi : C \times C \to \mathfrak{a}$, that is a biadditive map $\chi : C \times C \to \mathfrak{a}$ satisfying

$$\chi(c, c'\cdot a) = \chi(c, c')\cdot a, \quad \chi(c\cdot a, c') = \eta(a)\cdot\chi(c, c'), \quad \chi(c, c') = \eta\big(\chi(c', c)\big),$$

for $c, c' \in C$, $a \in \mathfrak{a}$, and G be the $(2r+1)\times(2r+1)$ matrix

$$G = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Also, given any $c \in C$, define $\chi_c \in C^*$ by $\chi_c(c') := \chi(c, c')$, for any $c' \in C$, and given any

$$\underline{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \in C^{2r+1}, \quad \text{define } \chi_{\underline{c}} := \begin{bmatrix} \chi_{c_1} \\ \vdots \\ \chi_{c_n} \end{bmatrix} \in \big(C^*\big)^{2r+1}.$$

Now set

$$\mathfrak{A}(\chi) := \big\{N \in \mathrm{End}_{\mathfrak{a}}(C) : \chi(Nc, c') + \chi(c, Nc') = 0 \text{ for all } c, c' \in C\big\},$$

$$\mathfrak{A} := \left\{\begin{bmatrix} M & \chi_{\underline{c}} \\ \underline{c}^t G & N \end{bmatrix} : M \in \mathrm{M}_{2r+1}(\mathfrak{a}),\ (M^\eta)^t G + GM = 0,\ \underline{c} \in C^{2r+1},\ N \in \mathfrak{A}(\chi)\right\}.$$

It can be checked that $\mathfrak{A}$ is a Lie algebra that contains a simple Lie algebra

$$\mathfrak{g} = \left\{ \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} : M \in \mathrm{M}_{2r+1}(\mathbb{C}),\, M^t G + GM = 0 \right\},$$

of type B_r. If $\mathfrak{h}$ denotes the Cartan subalgebra of diagonal matrices in $\mathfrak{g}$, then the adjoint action of $\mathfrak{h}$ on $\mathfrak{A}$ induces a root space decomposition

$$\mathfrak{A} = \bigoplus_{\mu \in \Delta \cup \{0\}} \mathfrak{A}_\mu, \quad \text{where } \mathfrak{A}_\mu = \left\{ T \in \mathfrak{A} : [h, T] = \mu(h)\, T \text{ for all } h \in \mathfrak{h} \right\}.$$

The following abbreviated notation helps describe these root spaces:

$$\text{for } v = \begin{bmatrix} v_1 \\ \vdots \\ v_{2r+1} \end{bmatrix} \in \mathbb{C}^{2r+1} \quad \text{and} \quad c \in C, \text{ let } vc := \begin{bmatrix} v_1 c \\ \vdots \\ v_{2r+1} c \end{bmatrix} \in C^{2r+1}.$$

Then $C^{2r+1} = \bigoplus_{i=1}^{2r+1} e_i C$, where $e_1, \ldots, e_{2r+1}$ is the standard basis for $\mathbb{C}^{2r+1}$. Letting B denote the set of skew-symmetric elements of $\mathfrak{a}$ relative to the involution η, we have

$$\mathfrak{A}_{\epsilon_i - \epsilon_j} = \left\{ E_{i,j}(a) + E_{2r+2-j,2r+2-i}(-\eta(a)) : a \in \mathfrak{a} \right\}, \qquad 1 \le i \ne j \le r,$$

$$\mathfrak{A}_{\epsilon_i + \epsilon_j} = \left\{ E_{i,2r+2-j}(a) + E_{j,2r+2-i}(-\eta(a)) : a \in \mathfrak{a} \right\}, \qquad 1 \le i, j \le r,$$

$$\mathfrak{A}_{-\epsilon_i - \epsilon_j} = \left\{ E_{2r+2-i,j}(a) + E_{2r+2-j,i}(-\eta(a)) : a \in \mathfrak{a} \right\}, \qquad 1 \le i, j \le r,$$

$$\mathfrak{A}_{\epsilon_i} = \left\{ \begin{bmatrix} 0 & \chi_{e_i c} \\ (e_{2r+2-i} c)^t & 0 \end{bmatrix} : c \in C \right\} + \left\{ E_{i,r+1}(a) + E_{r+1,2r+2-i}(-\eta(a)) : a \in \mathfrak{a} \right\}, \quad 1 \le i \le r,$$

$$\mathfrak{A}_{-\epsilon_i} = \left\{ \begin{bmatrix} 0 & \chi_{e_{2r+2-i} c} \\ (e_i c)^t & 0 \end{bmatrix} : c \in C \right\} + \left\{ E_{r+1,i}(a) + E_{2r+2-i,r+1}(-\eta(a)) : a \in \mathfrak{a} \right\}, \quad 1 \le i \le r,$$

$$\mathfrak{A}_0 = \left\{ \sum_{i=1}^{r} E_{ii}(a) + E_{2r+2-i,2r+2-i}(-\eta(a)) : a \in \mathfrak{a} \right\} + \left\{ \begin{bmatrix} 0 & 0 \\ 0 & N \end{bmatrix} : N \in \mathfrak{A}(\chi) \right\} + \left\{ E_{r+1,r+1}(b) : b \in B \right\} + \left\{ \begin{bmatrix} 0 & \chi_{e_{r+1} c} \\ (e_{r+1} c)^t & 0 \end{bmatrix} : c \in C \right\}.$$

The subalgebra

$$\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi) := \sum_{\mu \in \Delta} \mathfrak{A}_\mu + \sum_{\mu \in \Delta} [\mathfrak{A}_\mu, \mathfrak{A}_{-\mu}]$$

of $\mathfrak{A}$ has the root spaces

$$\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)_0 = \mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi) \cap \mathfrak{A}_0,$$
$$\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)_\mu = \mathfrak{A}_\mu \quad \text{for } \mu \in \Delta.$$

In particular,

$$\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)_0 = \sum_{\mu \in \Delta} \big[\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)_\mu, \mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)_{-\mu}\big].$$

Remark. In [Allison et al. 2002] the notation L is used to refer to the Lie algebra that we are calling $\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$.

To shorten the description of elements in $\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$, we use the following notation: Given any $1 \le k \le r$ and $a \in \mathfrak{a}$, let

$$E^{\boxed{\cdot}}_{k,r+1}(a) := E_{k,r+1}(a) + E_{r+1,2r+2-k}(-\eta(a)),$$
$$E^{\boxed{\cdot}}_{r+1,k}(a) := E_{r+1,k}(a) + E_{2r+2-k,r+1}(-\eta(a)),$$

and for any $1 \le p, q \le r$ and $a \in \mathfrak{a}$, let

$$E^{\blacksquare}_{p,q}(a) := E_{p,q}(a) + E_{2r+2-q,2r+2-p}(-\eta(a)),$$
$$E^{\boxed{\blacksquare}}_{p,2r+2-q}(a) := E_{p,2r+2-q}(a) + E_{q,2r+2-p}(-\eta(a)),$$
$$E^{\boxed{\blacksquare}}_{2r+2-p,q}(a) := E_{2r+2-p,q}(a) + E_{2r+2-q,p}(-\eta(a)).$$

We often also denote the involution η on $\mathfrak{a}$ by $\bar{\cdot}$. So, for example, we would write

$$E^{\boxed{\blacksquare}}_{2r+2-p,q}(a) \text{ (above) as } E_{2r+2-p,q}(a) + E_{2r+2-q,p}(-\bar{a}).$$

Allison, Benkart, and Gao's classification results on BC_r-graded Lie algebras [Allison et al. 2002] say the following in our setting:

Theorem 7 [Allison et al. 2002, Theorem 3.10]. *Let $r \ge 3$. Then L is BC_r-graded with grading subalgebra $\mathfrak{g}$ of type B_r if and only if there exists an associative algebra $\mathfrak{a}$ with involution η, and an $\mathfrak{a}$-module C with a hermitian form χ such that L is centrally isogenous to the BC_r-graded Lie algebra* $\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$.

Since $\mathfrak{im}(A^{[d]})$ is BC_r-graded with a grading subalgebra of type B_r and is centrally closed, we have the following result.

Corollary 8. *The intersection matrix algebra $\mathfrak{im}(A^{[d]})$ is isomorphic to the universal covering algebra of the Lie algebra* $\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$. *In particular, there exists a graded, central, surjective Lie algebra homomorphism*

$$\psi : \mathfrak{im}(A^{[d]}) \to \mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi).$$

5. Arriving at a "minimal" understanding of $\mathfrak{a}$, η, C, and χ

The graded nature of the map $\psi : \mathfrak{im}(A^{[d]}) \to \mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$ along with the relations among the generating elements of $\mathfrak{im}(A^{[d]})$ allow us to study each of components $\mathfrak{a}$, η, C, and χ involved in $\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$.

Since the elements $e_1 + \mathfrak{r}, \ldots, e_r + \mathfrak{r}$, $f_1 + \mathfrak{r}, \ldots, f_r + \mathfrak{r}$, $h_1 + \mathfrak{r}, \ldots, h_r + \mathfrak{r}$ in $\mathfrak{im}(A^{[d]})$ generate a subalgebra isomorphic to a simple Lie algebra of type B_r, and since ψ is a graded homomorphism, we may assume without loss of generality that (after relabeling the $e_i + \mathfrak{r}$, $f_i + \mathfrak{r}$, and $h_i + \mathfrak{r}$ as e_i, f_i, and h_i, respectively)

$$\psi(e_i) = E^{\blacksquare\square}_{i,i+1}(1) \quad \text{for } 1 \le i \le r-1, \qquad \psi(e_r) = E^{\square\,\square}_{r,r+1}(\sqrt{2}),$$

$$\psi(f_i) = E^{\blacksquare\square}_{i+1,i}(1) \quad \text{for } 1 \le i \le r-1, \qquad \psi(f_r) = E^{\square\,\square}_{r+1,r}(\sqrt{2}),$$

$$\psi(h_i) = E^{\blacksquare\square}_{i,i}(1) + E^{\blacksquare\square}_{i+1,i+1}(-1) \text{ for } 1 \le i \le r-1, \quad \psi(h_r) = E^{\blacksquare\blacksquare}_{r,r}(2).$$

Remark. Here we are using the notation established in Section 4. The generators of $\mathfrak{im}(A^{[d]})$ coming from a simple root $\alpha_j \in \Pi$ are denoted by e_j, f_j, and h_j, while the generators coming from an i-th copy of an adjoined root $\alpha \in \Delta_{B_r}$ are denoted by $e_{\alpha,i}$, $f_{\alpha,i}$, and $h_{\alpha,si}$.

5.1. *Understanding the invertibility of some coordinates of* $\mathfrak{a}$.

Proposition 9. (i) *Let* $e_{\epsilon_p-\epsilon_q,i}$, $f_{\epsilon_p-\epsilon_q,i}$, $h_{\epsilon_p-\epsilon_q,i}$ *be the generators of* $\mathfrak{im}(A^{[d]})$ *that result from adjoining the* i*-th copy of a long root* $\epsilon_p - \epsilon_q$ $(1 \le p, q \le r$, $p \ne q)$. *If*

$$\psi(e_{\epsilon_p-\epsilon_q,i}) = E^{\blacksquare\square}_{p,q}(a)$$

for some $a \in \mathfrak{a}$, *then* a *is an invertible element and*

$$\psi(f_{\epsilon_p-\epsilon_q,i}) = E^{\blacksquare\square}_{q,p}(a^{-1}).$$

(ii) *Let* $e_{\epsilon_p+\epsilon_q,i}$, $f_{\epsilon_p+\epsilon_q,i}$, $h_{\epsilon_p+\epsilon_q,i}$ *be the generators of* $\mathfrak{im}(A^{[d]})$ *that result from adjoining the* i*-th copy of a long root* $\epsilon_p + \epsilon_q$ $(1 \le p, q \le r$, $p \ne q)$. *If*

$$\psi(e_{\epsilon_p+\epsilon_q,i}) = E^{\square\blacksquare}_{p,2r+2-q}(b)$$

for some $b \in \mathfrak{a}$, *then* b *is an invertible element and*

$$\psi(f_{\epsilon_p+\epsilon_q,i}) = E^{\blacksquare\square}_{2r+2-q,p}(b^{-1}).$$

(iii) *Let* $e_{-\epsilon_p-\epsilon_q,i}$, $f_{-\epsilon_p-\epsilon_q,i}$, $h_{-\epsilon_p-\epsilon_q,i}$ *be the generators of* $\mathfrak{im}(A^{[d]})$ *that result from adjoining the* i*-th copy of a long root* $-\epsilon_p - \epsilon_q$ $(1 \le p, q \le r$, $p \ne q)$. *If*

$$\psi(e_{-\epsilon_p-\epsilon_q,i}) = E^{\blacksquare\square}_{2r+2-p,q}(c)$$

for some $c \in \mathfrak{a}$, then c is an invertible element and

$$\psi(f_{-\epsilon_p-\epsilon_q,i}) = E^{\square}_{q,2r+2-p}(c^{-1}).$$

Proof. (i) Since ψ is a graded homomorphism,

$$\psi(e_{\epsilon_p-\epsilon_q,i}) = E^{\blacksquare}_{p,q}(a) \quad \text{and} \quad \psi(f_{\epsilon_p-\epsilon_q,i}) = E^{\blacksquare}_{q,p}(a')$$

for some $a, a' \in \mathfrak{a}$. Without loss of generality, assume that $p < q$. Then

$$\begin{aligned}\big[\big[\psi(e_{\epsilon_p-\epsilon_q,i}), \psi(f_{\epsilon_p-\epsilon_q,i})\big], \psi(e_q)\big] &= \begin{cases} \big[\big[E^{\blacksquare}_{p,q}(a), E^{\blacksquare}_{q,p}(a')\big], E^{\blacksquare}_{q,q+1}(1)\big] & \text{if } q < r, \\ \big[\big[E^{\blacksquare}_{p,q}(a), E^{\blacksquare}_{q,p}(a')\big], E^{\square}_{r,r+1}(\sqrt{2})\big] & \text{if } q = r, \end{cases} \\ &= \begin{cases} E^{\blacksquare}_{q,q+1}(-a'a) & \text{if } q < r, \\ E^{\square}_{r,r+1}(-\sqrt{2}\,a'a) & \text{if } q = r. \end{cases}\end{aligned}$$

But since

$$\big[e_{\epsilon_p-\epsilon_q,i}, f_{\epsilon_p-\epsilon_q,i}\big] = h_{\epsilon_p-\epsilon_q,i} = \big[h_{\epsilon_p-\epsilon_q,i}, e_q\big] = \begin{cases} A_{\epsilon_p-\epsilon_q,\epsilon_q-\epsilon_{q+1}}\, e_q & \text{if } q < r, \\ A_{\epsilon_p-\epsilon_q,\epsilon_r}\, e_q & \text{if } q = r \end{cases} = -e_q$$

and ψ is a homomorphism,

$$\big[\big[\psi(e_{\epsilon_p-\epsilon_q,i}), \psi(f_{\epsilon_p-\epsilon_q,i})\big], \psi(e_q)\big] = -\psi(e_q) = \begin{cases} E^{\blacksquare}_{q,q+1}(-1) & \text{if } q < r, \\ E^{\square}_{r,r+1}(-\sqrt{2}) & \text{if } q = r, \end{cases}$$

So whether $q < r$ or $q = r$, we have

$$a'a = 1. \tag{1}$$

We show that aa' also equals 1. Indeed,

$$\begin{aligned}\big[\big[\psi(e_{\epsilon_p-\epsilon_q,i}), \psi(f_{\epsilon_p-\epsilon_q,i})\big], \psi(e_p)\big] &= \big[\big[E^{\blacksquare}_{p,q}(a), E^{\blacksquare}_{q,p}(a')\big], E^{\blacksquare}_{p,p+1}(1)\big] \\ &= \begin{cases} E^{\blacksquare}_{p,p+1}(aa') & \text{if } q - p \geq 2, \\ E^{\blacksquare}_{p,p+1}(aa' + a'a) & \text{if } q = p+1. \end{cases}\end{aligned}$$

But because

$$\begin{aligned}\big[\big[e_{\epsilon_p-\epsilon_q,i}, f_{\epsilon_p-\epsilon_q,i}\big], e_p\big] &= A_{\epsilon_p-\epsilon_q,\epsilon_p-\epsilon_{p+1}} e_p \\ &= (1+\delta_{q,p+1})e_p = \begin{cases} e_p & \text{if } q \geq p+2, \\ 2e_p & \text{if } q = p+1, \end{cases}\end{aligned}$$

we have

$$\big[\big[\psi(e_{\epsilon_p-\epsilon_q,i}),\psi(f_{\epsilon_p-\epsilon_q,i})\big],\psi(e_p)=\begin{cases}E^{\blacksquare\square}_{p,p+1}(1) & \text{if } q\geq p+2,\\ E^{\blacksquare\blacksquare}_{p,p+1}(2) & \text{if } q=p+1.\end{cases}$$

So if $q\geq p+2$, then $aa'=1$. If $q=p+1$, then $aa'+a'a=2$. But, by (1), $a'a=1$. Hence, in either case, $aa'=1$.

(ii) Since ψ is a graded homomorphism,

$$\psi(e_{\epsilon_p+\epsilon_q,i})=E^{\square\square}_{p,2r+2-q}(b)\quad\text{and}\quad\psi(f_{\epsilon_p+\epsilon_q,i})=E^{\square\blacksquare}_{2r+2-q,p}(b')$$

for some $b,b'\in\mathfrak{a}$. Again without loss of generality, we may assume that $p<q$. Then $\big[\big[\psi(e_{\epsilon_p+\epsilon_q,i}),\psi(f_{\epsilon_p+\epsilon_q,i})\big],\psi(e_q)\big]$ equals

$$\begin{aligned}&\begin{cases}\big[\big[E^{\blacksquare\square}_{p,2r+2-q}(b),E^{\blacksquare\blacksquare}_{2r+2-q,p}(b')\big],E^{\blacksquare\blacksquare}_{q,q+1}(1)\big] & \text{if } q<r,\\ \big[\big[E^{\blacksquare\square}_{p,2r+2-q}(b),E^{\blacksquare\blacksquare}_{2r+2-q,p}(b')\big],E^{\square\square}_{r,r+1}(\sqrt{2})\big] & \text{if } q=r,\end{cases}\\ =&\begin{cases}E^{\blacksquare\blacksquare}_{q,q+1}\big(\eta(b)\,\eta(b')\big) & \text{if } q<r,\\ E^{\square\square}_{r,r+1}\big(\sqrt{2}\eta(b)\,\eta(b')\big) & \text{if } q=r.\end{cases}\end{aligned}$$

But it also equals

$$\psi\big(\big[\big[e_{\epsilon_p+\epsilon_q,i},f_{\epsilon_p+\epsilon_q,i}\big],e_q\big]\big)=\psi(e_q)=\begin{cases}E^{\blacksquare\blacksquare}_{q,q+1}(1) & \text{if } q<r,\\ E^{\square\square}_{r,r+1}(\sqrt{2}) & \text{if } q=r,\end{cases}$$

whence $\eta(b)\,\eta(b')=1$. Applying (the antiautomorphism) η to both sides, we get that

$$b'b=1. \tag{2}$$

To show that $bb'=1$, we first compute that

$$\big[\big[\psi(e_{\epsilon_p+\epsilon_q,i}),\psi(f_{\epsilon_p+\epsilon_q,i})\big],\psi(e_p)\big]=\begin{cases}E^{\blacksquare\square}_{p,p+1}(bb') & \text{if } q\geq p+2,\\ E^{\blacksquare\blacksquare}_{p,p+1}(bb'-\eta(b)\,\eta(b')) & \text{if } q=p+1.\end{cases}$$

Since

$$\big[\big[e_{\epsilon_p+\epsilon_q,i},f_{\epsilon_p+\epsilon_q,i}\big],e_p\big]=A_{\epsilon_p+\epsilon_q,\epsilon_p-\epsilon_{p+1}}\,e_p=(1-\delta_{q,p+1})e_p=\begin{cases}e_p & \text{if } q\geq p+2,\\ 0 & \text{if } q=p+1,\end{cases}$$

we also have

$$\big[\big[\psi(e_{\epsilon_p+\epsilon_q,i}),\psi(f_{\epsilon_p+\epsilon_q,i})\big],\psi(e_p)\big]=\begin{cases}E^{\blacksquare\square}_{p,p+1}(1) & \text{if } q-p\geq 2,\\ 0 & \text{if } q=p+1.\end{cases}$$

So if $q \geq p+2$, then $bb' = 1$. If $q = p+1$, then $bb' - \eta(b)\,\eta(b') = 0$, which implies, using (2), that

$$bb' = \eta(b)\,\eta(b') = \eta(b'b) = \eta(1) = 1.$$

In either case, $bb' = 1$.

(iii) The proof follows using similar calculations as above. □

5.2. *Understanding the involution η on $\mathfrak{a}$.*

Proposition 10. (i) *If*

$$\psi(e_{\epsilon_p+\epsilon_{p+1},i}) = E^{\square}_{p,2r+2-(p+1)}(a)$$

for some $1 \leq p \leq r-1$ and $a \in \mathfrak{a}$, then $\eta(a) = a$.

(ii) *If*

$$\psi(e_{-\epsilon_p-\epsilon_{p+1},i}) = E^{\blacksquare}_{2r+2-p,\,p+1}(b)$$

for some $1 \leq p \leq r-1$ and $b \in \mathfrak{a}$, then $\eta(b) = b$.

Proof. We prove (i). The proof of (ii) is similar. Observe that

$$\left[\psi(e_{\epsilon_p+\epsilon_{p+1},i}),\, \psi(e_p)\right] = \left[E^{\square}_{p,2r+2-(p+1)}(a),\, E^{\blacksquare}_{p,p+1}(1)\right] = E^{\square}_{p,2r+2-p}(a).$$

But $A_{\epsilon_p+\epsilon_{p+1},\epsilon_p-\epsilon_{p+1}} = 0$ implies that $(\operatorname{ad} e_{\epsilon_p+\epsilon_{p+1},i})^{-0+1} e_p = \left[e_{\epsilon_p+\epsilon_{p+1},i},\, e_p\right] = 0$, which, in turn, implies that $\left[\psi(e_{\epsilon_p+\epsilon_{p+1},i}),\, \psi(e_p)\right] = 0$. So

$$E^{\square}_{p,2r+2-p}(a) = E_{p,2r+2-p}(a - \eta(a)) = 0$$

and thus

$$\eta(a) = a. \tag{3}$$

□

5.3. *Understanding the relations on generators of $\mathfrak{a}$.*

Proposition 11. *If, as a consequence of adjoining an i-th copy of the long root $\epsilon_p - \epsilon_q$ and a j-th copy of the long root $\epsilon_p + \epsilon_q$, where $1 \leq p, q \leq r$ with $p \neq q$,*

$$\psi(e_{\epsilon_p-\epsilon_q,i}) = E^{\blacksquare}_{p,q}(s) \text{ and } \psi\left(e_{\epsilon_p+\epsilon_q,j}\right) = E^{\square}_{p,2r+2-q}(t),$$

for some $s, t \in \mathfrak{a}$, then:

(a) *If $|p-q| = 1$, the elements s, t, and $\eta(s)$ in $\mathfrak{a}$ satisfy the relation*

$$s \cdot t = t \cdot \eta(s).$$

(b) *If $|p-q| \geq 2$, the elements $s,\ t,\ \eta(s)$, and $\eta(t)$ in $\mathfrak{a}$ satisfy the relation*

$$s \cdot \eta(t) = t \cdot \eta(s).$$

Proof. Observe that

$$\begin{aligned}\left[\psi(e_{\epsilon_p-\epsilon_q,i}), \psi(e_{\epsilon_p+\epsilon_q,j})\right] &= \left[E^{\blacksquare}_{p,q}(s), E^{\square}_{p,2r+2-q}(t)\right] \\ &= E^{\square}_{p,2r+2-p}(-s\cdot\eta(t)) \\ &= E_{p,2r+2-p}(-s\cdot\eta(t)+t\cdot\eta(s)) \\ &= \begin{cases} E_{p,2r+2-p}(-s\cdot t+t\cdot\eta(s)) & \text{if } |p-q|=1, \\ E_{p,2r+2-p}(-s\cdot\eta(t)+t\cdot\eta(s)) & \text{if } |p-q|\geq 2. \end{cases}\end{aligned}$$

(The division into two cases in the last step follows from the use of (3).) But since $A_{\epsilon_p-\epsilon_q,\epsilon_p+\epsilon_q}=0$, the generalized intersection matrix algebra relations tell us that

$$\left(\operatorname{ad} e_{\epsilon_p-\epsilon_q,i}\right)^{-0+1} e_{\epsilon_p+\epsilon_q,j}=0.$$

That is, $\left[e_{\epsilon_p-\epsilon_q,i}, e_{\epsilon_p+\epsilon_q,j}\right]=0$. So we must have $\left[\psi(e_{\epsilon_p-\epsilon_q,i}), \psi(e_{\epsilon_p+\epsilon_q,j})\right]=0$. This implies that

$$\begin{aligned} -s\cdot t+t\cdot\eta(s)&=0 \quad \text{if } |p-q|=1, \\ -s\cdot\eta(t)+t\cdot\eta(s)&=0, \quad \text{if } |p-q|\geq 2. \end{aligned}$$ □

Similarly:

Proposition 12. *If, as a consequence of adjoining an i-th copy of the long root $\epsilon_p-\epsilon_q$ and a j-th copy of the long root $-\epsilon_p-\epsilon_q$, where $1\leq p, q\leq r$ with $p\neq q$,*

$$\psi(e_{\epsilon_p-\epsilon_q,i})=E^{\blacksquare}_{p,q}(s) \quad \textit{and} \quad \psi(e_{-\epsilon_p-\epsilon_q,j})=E^{\square}_{2r+2-p,q}(t),$$

for some $s, t\in\mathfrak{a}$, then:

(a) *If $|p-q|=1$, the elements s, t, and $\eta(s)$ in $\mathfrak{a}$ satisfy the relation*

$$\eta(s)\cdot t=t\cdot s.$$

(b) *If $|p-q|\geq 2$, the elements s, t, $\eta(s)$, and $\eta(t)$ in $\mathfrak{a}$ satisfy the relation*

$$\eta(s)\cdot t=\eta(t)\cdot s.$$

5.4. ***A description of the module C.*** Since ψ is a graded, surjective homomorphism from $\mathfrak{im}(A^{[d]})$ to $\mathrm{so}_{2r+1}(\mathfrak{a},\eta,C,\chi)$ and we are only adjoining long roots, we can examine the image of $\mathfrak{im}(A^{[d]})$ under ψ to help us understand C.

Proposition 13. *The module C is zero.*

Proof. The generators of $\psi\left(\mathfrak{im}(A^{[d]})\right)$ all have the form $\left[\begin{smallmatrix} M & 0 \\ 0 & 0 \end{smallmatrix}\right]$, $M\in \mathrm{M}_{2r+1}(\mathfrak{a})$. Since the matrices of this form in $\mathrm{so}_{2r+1}(\mathfrak{a},\eta,C,\chi)$ form a subalgebra and since ψ is surjective, we have $C=0$. □

6. Achieving a "sufficient" understanding of $\mathfrak{a}$, η, C, and χ

In the previous section we used the homomorphism

$$\psi : \mathrm{im}(A^{[d]}) \to \mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi),$$

given by the recognition theorem of [Allison et al. 2002] to get a sense (i) of what the generators of $\mathfrak{a}$ ought to be, (ii) of what the involution η on $\mathfrak{a}$ ought to be, (iii) of what the relations on the generators of $\mathfrak{a}$ ought to be, and (iv) that $C = 0$ and $\chi = 0$.

In this section, we show that the understanding we have arrived at is complete. We do so as follows:

1. Take the 4-tuple of associative algebra, involution, module, and hermitian form as we presently understand it. That is:

(i) Let $\Omega =$ the set of all long roots of the form $\pm(\epsilon_i + \epsilon_{i+1})$ that we have adjoined,
$\Theta =$ the set of all long roots in Δ_B which we have adjoined but that are not in Ω,

and let

$$X_e = \bigcup_{\omega\in\Omega} \{x_{\omega,1}, \ldots, x_{\omega,N_\omega}\}, \quad X_f = \bigcup_{\omega\in\Omega} \{x_{\omega,1}^{-1}, \ldots, x_{\omega,N_\omega}^{-1}\},$$
$$Y_e = \bigcup_{\theta\in\Theta} \{y_{\theta,1}, \ldots, y_{\theta,N_\theta}\}, \quad Y_f = \bigcup_{\theta\in\Theta} \{y_{\theta,1}^{-1}, \ldots, y_{\theta,N_\theta}^{-1}\},$$
$$Z_e = \bigcup_{\theta\in\Theta} \{z_{\theta,1}, \ldots, z_{\theta,N_\theta}\}, \quad Z_f = \bigcup_{\theta\in\Theta} \{z_{\theta,1}^{-1}, \ldots, z_{\theta,N_\theta}^{-1}\},$$

denote collections of indeterminates indexed by the sets Ω and Θ. Let $\mathfrak{b}$ be the unital associative $\mathbb{C}$-algebra generated by the indeterminates in

$$X_e \cup X_f \cup Y_e \cup Y_f \cup Z_e \cup Z_f,$$

subject to the relations

$$y_{\epsilon_p-\epsilon_q,i} x_{\epsilon_p+\epsilon_q,j} = x_{\epsilon_p+\epsilon_q,j} z_{\epsilon_p-\epsilon_q,i},$$
$$y_{\epsilon_p-\epsilon_q,i} z_{\epsilon_p+\epsilon_q,j} = y_{\epsilon_p+\epsilon_q,j} z_{\epsilon_p-\epsilon_q,i},$$
$$z_{\epsilon_p-\epsilon_q,i} x_{-\epsilon_p-\epsilon_q,k} = x_{-\epsilon_p-\epsilon_q,k} y_{\epsilon_p-\epsilon_q,i},$$
$$z_{\epsilon_p-\epsilon_q,i} y_{-\epsilon_p-\epsilon_q,k} = z_{-\epsilon_p-\epsilon_q,k} y_{\epsilon_p-\epsilon_q,i},$$

where $i = 1, \ldots, N_{\epsilon_p-\epsilon_q}$ for $\epsilon_p - \epsilon_q \in \Theta$, $j = 1, \ldots, N_{\epsilon_p+\epsilon_q}$ for $\epsilon_p + \epsilon_q \in \Omega \cup \Theta$, and $k = 1, \ldots, N_{-\epsilon_p-\epsilon_q}$ for $-\epsilon_p - \epsilon_q \in \Omega \cup \Theta$.

(ii) Define an involution, which we also call η and sometimes denote by $\bar{\cdot}$, on $\mathfrak{b}$ by

$$\eta(x_{\omega,i}) = x_{\omega,i} \quad \text{if } \omega \in \Omega \text{ and } 1 \le i \le N_\omega,$$
$$\eta(y_{\theta,i}) = z_{\theta,i} \quad \text{if } \theta \in \Theta \text{ and } 1 \le i \le N_\theta,$$
$$\eta(z_{\theta,i}) = y_{\theta,i} \quad \text{if } \theta \in \Theta \text{ and } 1 \le i \le N_\theta.$$

(iii) Let $C = 0$ be the trivial $\mathfrak{b}$-module.

(iv) Let $\chi = 0$ be a hermitian form on C.

Remarks. (a) The indeterminates in $X_e \cup X_f \cup \cdots \cup Z_f$ are intended to capture the elements of the form a, a', b, b', c, and c' of $\mathfrak{a}$ that we studied in Section 5, which arose from the images of the map ψ.

(b) In the relations listed above, our use of the indeterminates $x_{\epsilon_p+\epsilon_q,j}$ and $x_{-\epsilon_p-\epsilon_q,j}$ signals that we are working with roots in Ω and, hence, $|p-q| = 1$ in this setting. Likewise, our use of the indeterminates $y_{\epsilon_p+\epsilon_q,j}$, $z_{\epsilon_p+\epsilon_q,j}$, $y_{-\epsilon_p-\epsilon_q,j}$, and $z_{-\epsilon_p-\epsilon_q,j}$ signals that we are working with roots in Θ and p, q such that $|p-q| \geq 2$.

2. Construct a map

$$\varphi : \mathfrak{gim}(A^{[d]}) \to \mathrm{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$$

sending the generators

$$\begin{array}{lll}
e_1, \ldots, e_r, & \bigcup_{\omega\in\Omega} \{e_{\omega,1}, \ldots, e_{\omega,N_\omega}\}, & \bigcup_{\theta\in\Theta} \{e_{\theta,1}, \ldots, e_{\theta,N_\theta}\}, \\
f_1, \ldots, f_r, & \bigcup_{\omega\in\Omega} \{f_{\omega,1}, \ldots, f_{\omega,N_\omega}\}, & \bigcup_{\theta\in\Theta} \{f_{\theta,1}, \ldots, f_{\theta,N_\theta}\}, \\
h_1, \ldots, h_r, & \bigcup_{\omega\in\Omega} \{h_{\omega,1}, \ldots, h_{\omega,N_\omega}\}, & \bigcup_{\theta\in\Theta} \{h_{\theta,1}, \ldots, h_{\theta,N_\theta}\},
\end{array}$$

of $\mathfrak{gim}(A^{[d]})$ to

$$\begin{array}{lll}
\tilde{e}_1, \ldots, \tilde{e}_r, & \bigcup_{\omega\in\Omega} \{\tilde{e}_{\omega,1}, \ldots, \tilde{e}_{\omega,N_\omega}\}, & \bigcup_{\theta\in\Theta} \{\tilde{e}_{\theta,1}, \ldots, \tilde{e}_{\theta,N_\theta}\} \\
\tilde{f}_1, \ldots, \tilde{f}_r, & \bigcup_{\omega\in\Omega} \{\tilde{f}_{\omega,1}, \ldots, \tilde{f}_{\omega,N_\omega}\}, & \bigcup_{\theta\in\Theta} \{\tilde{f}_{\theta,1}, \ldots, \tilde{f}_{\theta,N_\theta}\}, \\
\tilde{h}_1, \ldots, \tilde{h}_r, & \bigcup_{\omega\in\Omega} \{\tilde{h}_{\omega,1}, \ldots, \tilde{h}_{\omega,N_\omega}\}, & \bigcup_{\theta\in\Theta} \{\tilde{h}_{\theta,1}, \ldots, \tilde{h}_{\theta,N_\theta}\},
\end{array}$$

respectively, where

$$\tilde{e}_i := E_{i,i+1}^{\blacksquare\blacksquare}(1), \quad 1 \leq i \leq r-1,$$

$$\tilde{e}_r := E_{r,r+1}^{\square}(\sqrt{2}),$$

$$\tilde{e}_{\omega,i} := \begin{cases} E_{p,2r+2-(p+1)}^{\blacksquare\square}(x_{\omega,i}) & \text{if } \omega = \epsilon_p + \epsilon_{p+1}, \\ E_{2r+2-p,p+1}^{\blacksquare\square}(x_{\omega,i}) & \text{if } \omega = -\epsilon_p - \epsilon_{p+1}, \end{cases} \quad \text{for } \omega \in \Omega \text{ and } 1 \leq i \leq N_\omega,$$

$$\tilde{e}_{\theta,i} := \begin{cases} E_{p,q}^{\blacksquare\blacksquare}(y_{\theta,i}) & \text{if } \theta = \epsilon_p - \epsilon_q, \\ E_{p,2r+2-q}^{\blacksquare\square}(y_{\theta,i}) & \text{if } \theta = \epsilon_p + \epsilon_q, \\ E_{2r+2-p,q}^{\blacksquare\square}(y_{\theta,i}) & \text{if } \theta = -\epsilon_p - \epsilon_q, \end{cases} \quad \text{for } \theta \in \Theta \text{ and } 1 \leq i \leq N_\theta,$$

$$\tilde{f}_i := E^{\blacksquare\blacksquare}_{i+1,i}(1), \quad 1 \le i \le r-1,$$

$$\tilde{f}_r := E^{\square\blacksquare}_{r+1,r}(\sqrt{2}),$$

$$\tilde{f}_{\omega,i} := \begin{cases} E^{\blacksquare\square}_{2r+2-(p+1),p}(x^{-1}_{\omega,i}) & \text{if } \omega = \epsilon_p + \epsilon_{p+1}, \\ E^{\square\blacksquare}_{p+1,2r+2-p}(x^{-1}_{\omega,i}) & \text{if } \omega = -\epsilon_p - \epsilon_{p+1}, \end{cases} \quad \text{for } \omega \in \Omega \text{ and } 1 \le i \le N_\omega,$$

$$\tilde{f}_{\theta,i} := \begin{cases} E^{\blacksquare\blacksquare}_{q,p}(y^{-1}_{\theta,i}) & \text{if } \theta = \epsilon_p - \epsilon_q, \\ E^{\blacksquare\square}_{2r+2-q,p}(y^{-1}_{\theta,i}) & \text{if } \theta = \epsilon_p + \epsilon_q, \\ E^{\square\blacksquare}_{q,2r+2-p}(y^{-1}_{\theta,i}) & \text{if } \theta = -\epsilon_p - \epsilon_q, \end{cases} \quad \text{for } \theta \in \Theta \text{ and } 1 \le i \le N_\theta,$$

$$\tilde{h}_i := E^{\blacksquare\blacksquare}_{i,i}(1) + E^{\blacksquare\blacksquare}_{i+1,i+1}, \quad 1 \le i \le r-1,$$

$$\tilde{h}_r := E^{\blacksquare\blacksquare}_{r,r}(2),$$

$$\tilde{h}_{\omega,i} := \begin{cases} E^{\blacksquare\blacksquare}_{p,p}(1) + E^{\blacksquare\blacksquare}_{p+1,p+1}(1) & \text{if } \omega = \epsilon_p + \epsilon_{p+1} \\ E^{\blacksquare\blacksquare}_{p,p}(-1) + E^{\blacksquare\blacksquare}_{p+1,p+1}(-1) & \text{if } \omega = -\epsilon_p - \epsilon_{p+1} \end{cases} \quad \text{for } \omega \in \Omega,\ 1 \le i \le N_\omega,$$

$$\tilde{h}_{\theta,i} := \begin{cases} E^{\blacksquare\blacksquare}_{p,p}(1) + E^{\blacksquare\blacksquare}_{q,q}(-1) & \text{if } \theta = \epsilon_p - \epsilon_q, \\ E^{\blacksquare\blacksquare}_{p,p}(1) + E^{\blacksquare\blacksquare}_{q,q}(1) & \text{if } \theta = \epsilon_p + \epsilon_q, \\ E^{\blacksquare\blacksquare}_{p,p}(-1) + E^{\blacksquare\blacksquare}_{q,q}(-1) & \text{if } \theta = -\epsilon_p - \epsilon_q, \end{cases} \quad \text{for } \theta \in \Theta \text{ and } 1 \le i \le N_\theta.$$

3. We show that φ is

(a) a Lie algebra homomorphism (Theorem 14),

(b) that is surjective (Proposition 15), and

(c) graded (Proposition 16).

4. We show that the radical $\mathfrak{r}$ of $\mathfrak{gim}(A^{[d]})$ lies in the kernel of this map φ (see just before Proposition 17), hence inducing a surjective, graded, Lie algebra homomorphism

$$\phi : \mathfrak{im}(A^{[d]}) \to \mathfrak{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi).$$

5. Finally, we show that ϕ is a central map and that $\mathfrak{b} \cong \mathfrak{a}$ (Proposition 18).

Theorem 14. *The map $\varphi : \mathfrak{gim}(A^{[d]}) \to \mathfrak{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$ is a Lie algebra homomorphism.*

Proof. We show that the images in $\mathfrak{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$ of the generators of $\mathfrak{gim}(A^{[d]})$, under the map φ, satisfy the relations (R1)–(R3) of Definition 2 with respect to

the same $(r+d)\times(r+d)$ generalized intersection matrix $A^{[d]}$ as used in the construction of the algebra $\mathfrak{gim}(A^{[d]})$.

While working with the various long roots in our proof, we use labels like u or v to denote the indeterminates $x_{\omega,i}$ or $y_{\theta,i}$.

The reason that we can substitute u or v for the actual indeterminates is that the result of taking a bracket like

$$\left[\tilde{e}_{-\epsilon_p-\epsilon_q,i},\tilde{e}_{-\epsilon_k-\epsilon_l,j}\right]=\left[E^{\blacksquare\square}_{2r+2-p,q}(y_{-\epsilon_p-\epsilon_q,i}),E^{\blacksquare\square}_{2r+2-k,l}(y_{-\epsilon_k-\epsilon_l,j})\right]$$

depends primarily on the indices p, q, k, and l rather than on the particular elements of the algebra $\mathfrak{b}$ being housed at these sites.

If we agree on this convention of using substitute variables like u, then we must recognize that

$$\bar{u}=\begin{cases}x_{\omega,i} & \text{if } u=x_{\omega,i},\\ z_{\theta,i} & \text{if } u=y_{\theta,j}.\end{cases}$$

That is, the involution $\bar{\cdot}$ applied to u depends on whether u is substituting for a variable associated to a root in Ω or a root in Θ.

We show the computations for the interactions between the generators corresponding to the long roots $\epsilon_p-\epsilon_q$ and $\epsilon_k-\epsilon_l$. The remaining computations are similar.

Let $1\le p,q,k,l\le r$ with $p\ne q$ and $k\ne l$, $u,v\in\{x_{\omega,i},x_{\omega,j},y_{\theta,i}y_{\theta,j}\}$ and $u^{-1},v^{-1}\in\{x^{-1}_{\omega,i},x^{-1}_{\omega,j},y^{-1}_{\theta,i}y^{-1}_{\theta,j}\}$, where $\omega\in\Omega$, $\theta\in\Theta$, and $1\le i,j\le N_\omega$ or $1\le i,j\le N_\theta$.

Using the definition of $A_{\epsilon_p-\epsilon_q,\epsilon_k-\epsilon_l}$, we see that

$$A_{\epsilon_p-\epsilon_q,\epsilon_k-\epsilon_l}=\delta_{p,k}-\delta_{p,l}-\delta_{q,k}+\delta_{q,l}=\begin{cases}0 & \text{if } p,q\notin\{k,l\},\\ 1 & \text{if } p=k \text{ but } q\ne l,\\ -1 & \text{if } p=l \text{ but } q\ne k,\\ -1 & \text{if } p\ne l \text{ but } q=k,\\ 1 & \text{if } p\ne k \text{ but } q=l,\\ 2 & \text{if } p=k \text{ and } q=l,\\ -2 & \text{if } p=l \text{ and } q=k.\end{cases}$$

A. $\left[\tilde{e}_{\epsilon_p-\epsilon_q,i},\tilde{e}_{\epsilon_k-\epsilon_l,j}\right]=\left[E^{\blacksquare\square}_{p,q}(u),E^{\blacksquare\square}_{k,l}(v)\right]=\delta_{q,k}E^{\blacksquare\square}_{p,l}(uv)+\delta_{l,p}E^{\blacksquare\square}_{k,q}(-vu)$

$$=\begin{cases}E^{\blacksquare\square}_{k,q}(-vu) & \text{if } p=l \text{ but } q\ne k,\\ E^{\blacksquare\square}_{p,l}(uv) & \text{if } p\ne l \text{ but } q=k,\\ E^{\blacksquare\square}_{p,p}(uv)+E^{\blacksquare\square}_{q,q}(-vu) & \text{if } p=l \text{ and } q=k,\\ 0 & \text{otherwise.}\end{cases}$$

- If $p=l$ but $q\ne k$, then $\left[E^{\blacksquare\square}_{p,q}(u),E^{\blacksquare\square}_{k,q}(-vu)\right]=0$ because $q\ne k$ and $q\ne p$.

- If $p \neq l$ but $q = k$, then $\left[E^{\blacksquare\square}_{p,q}(u), E^{\blacksquare\square}_{p,l}(uv)\right] = 0$ because $q \neq p$ and $l \neq p$.
- If $p = l$ and $q = k$, then

$$\begin{aligned}\left[E^{\blacksquare\square}_{p,q}(u),\ E^{\blacksquare\square}_{p,p}(uv) + E^{\blacksquare\square}_{q,q}(-vu)\right] &= E^{\blacksquare\square}_{p,q}(-uvu) + E^{\blacksquare\square}_{p,q}(-uvu)\\ &= E^{\blacksquare\square}_{p,q}(-2uvu).\end{aligned}$$

So

$$\left(\operatorname{ad}\tilde{e}_{\epsilon_p-\epsilon_q,i}\right)^{1+1}\tilde{e}_{\epsilon_k-\epsilon_l,j} = \begin{cases} E^{\blacksquare\square}_{p,q}(-2uvu) & \text{if } p = l \text{ and } q = k,\\ 0 & \text{otherwise.}\end{cases}$$

Since $\left[E^{\blacksquare\square}_{p,q}(u), E^{\blacksquare\square}_{p,q}(-2uvu)\right] = 0$, we get $\left(\operatorname{ad}\tilde{e}_{\epsilon_p-\epsilon_q,i}\right)^{2+1}\tilde{e}_{\epsilon_k-\epsilon_l,j} = 0$.

B. $\left[\tilde{f}_{\epsilon_p-\epsilon_q,i}, \tilde{f}_{\epsilon_k-\epsilon_l,j}\right] = \left[E^{\blacksquare\square}_{q,p}(u^{-1}), E^{\blacksquare\square}_{l,k}(v^{-1})\right]$

$$\begin{aligned}&= \delta_{p,l}E^{\blacksquare\square}_{q,k}(u^{-1}v^{-1}) + \delta_{k,q}E^{\blacksquare\square}_{l,p}(-v^{-1}u^{-1})\\ &= \begin{cases} E^{\blacksquare\square}_{q,k}(u^{-1}v^{-1}) & \text{if } p = l \text{ but } q \neq k,\\ E^{\blacksquare\square}_{l,p}(-v^{-1}u^{-1}) & \text{if } p \neq l \text{ but } q = k,\\ E^{\blacksquare\square}_{p,p}(-v^{-1}u^{-1}) + E^{\blacksquare\square}_{q,q}(u^{-1}v^{-1}) & \text{if } p = l \text{ and } q = k,\\ 0 & \text{otherwise.}\end{cases}\end{aligned}$$

- If $p = l$ but $q \neq k$, then $\left[E^{\blacksquare\square}_{q,p}(u^{-1}), E^{\blacksquare\square}_{q,k}(u^{-1}v^{-1})\right] = 0$ because $p \neq q$ and $k \neq q$.
- If $p \neq l$ but $q = k$, then $\left[E^{\blacksquare\square}_{q,p}(u^{-1}), E^{\blacksquare\square}_{l,p}(-v^{-1}u^{-1})\right] = 0$ because $p \neq l$ and $p \neq q$.
- If $p = l$ and $q = k$, then

$$\begin{aligned}&\left[E^{\blacksquare\square}_{q,p}(u^{-1}), E^{\blacksquare\square}_{p,p}(-v^{-1}u^{-1}) + E^{\blacksquare\square}_{q,q}(u^{-1}v^{-1})\right]\\ &\qquad = E^{\blacksquare\square}_{q,p}(-u^{-1}v^{-1}u^{-1}) + E^{\blacksquare\square}_{q,p}(-u^{-1}v^{-1}u^{-1})\\ &\qquad = E^{\blacksquare\square}_{q,p}\left(-2u^{-1}v^{-1}u^{-1}\right).\end{aligned}$$

So

$$\left(\operatorname{ad}\tilde{f}_{\epsilon_p-\epsilon_q,i}\right)^{1+1}\tilde{f}_{\epsilon_k-\epsilon_l,j} = \begin{cases} E^{\blacksquare\square}_{q,p}(-2u^{-1}v^{-1}u^{-1}) & \text{if } p = l \text{ and } q = k,\\ 0 & \text{otherwise.}\end{cases}$$

Since $\left[E^{\blacksquare\square}_{q,p}(u^{-1}), E^{\blacksquare\square}_{q,p}(-2u^{-1}v^{-1}u^{-1})\right] = 0$, we get that

$$\left(\operatorname{ad}\tilde{f}_{\epsilon_p-\epsilon_q,i}\right)^{2+1}\tilde{f}_{\epsilon_k-\epsilon_l,j} = 0.$$

C. $\left[\tilde{h}_{\epsilon_p-\epsilon_q,i}, \tilde{h}_{\epsilon_k-\epsilon_l,j}\right]$

$$= \left[E^{\blacksquare\square}_{p,p}(1) + E^{\blacksquare\square}_{q,q}(-1),\ E^{\blacksquare\square}_{k,k}(1) + E^{\blacksquare\square}_{l,l}(-1)\right]$$
$$= \delta_{p,k}E^{\blacksquare\square}_{p,p}([1,1]) + \delta_{p,l}E^{\blacksquare\square}_{p,p}([1,-1]) + \delta_{q,k}E^{\blacksquare\square}_{q,q}([-1,1]) + \delta_{q,l}E^{\blacksquare\square}_{q,q}([-1,-1])$$
$$= \delta_{p,k}E^{\blacksquare\square}_{p,p}(0) + \delta_{p,l}E^{\blacksquare\square}_{p,p}(0) + \delta_{q,k}E^{\blacksquare\square}_{q,q}(0) + \delta_{q,l}E^{\blacksquare\square}_{q,q}(0)$$
$$= 0$$

D. $\left[\tilde{e}_{\epsilon_p-\epsilon_q,i}, \tilde{f}_{\epsilon_k-\epsilon_l,j}\right] = \left[E^{\blacksquare\square}_{p,q}(u), E^{\blacksquare\square}_{l,k}(v^{-1})\right]$

$$= \delta_{q,l}E^{\blacksquare\square}_{p,k}\left(u\,v^{-1}\right) + \delta_{k,p}E^{\blacksquare\square}_{l,q}\left(-v^{-1}\,u\right)$$
$$= \begin{cases} E^{\blacksquare\square}_{l,q}(-v^{-1}u) & \text{if } p=k \text{ but } q\neq l,\\ E^{\blacksquare\square}_{p,k}(uv^{-1}) & \text{if } p\neq k \text{ but } q=l,\\ E^{\blacksquare\square}_{p,p}(uv^{-1}) + E^{\blacksquare\square}_{q,q}(-v^{-1}u) & \text{if } p=k \text{ and } q=l,\\ 0 & \text{otherwise.}\end{cases}$$

- If $p=k$ but $q\neq l$, then $\left[E^{\blacksquare\square}_{p,q}(u), E^{\blacksquare\square}_{l,q}(-v^{-1}u)\right]=0$ because $q\neq l$ and $q\neq p$.
- If $p\neq k$ but $q=l$, then $\left[E^{\blacksquare\square}_{p,q}(u), E^{\blacksquare\square}_{p,k}(uv^{-1})\right]=0$ because $q\neq p$ and $k\neq p$.
- If $p=k$ and $q=l$, then

$$\left[E^{\blacksquare\square}_{p,q}(u), E^{\blacksquare\square}_{p,p}(uv^{-1}) + E^{\blacksquare\square}_{q,q}(-v^{-1}u)\right] = E^{\blacksquare\square}_{p,q}(-uv^{-1}u) + E^{\blacksquare\square}_{p,q}(-uv^{-1}u)$$
$$= E^{\blacksquare\square}_{p,q}(-2uv^{-1}u).$$

So

$$\left(\operatorname{ad}\tilde{e}_{\epsilon_p-\epsilon_q,i}\right)^{1+1}\tilde{f}_{\epsilon_k-\epsilon_l,j} = \begin{cases} E^{\blacksquare\square}_{p,q}(-2uv^{-1}u) & \text{if } p=k \text{ and } q=l,\\ 0 & \text{otherwise.}\end{cases}$$

Since $\left[E^{\blacksquare\square}_{p,q}(u), E^{\blacksquare\square}_{p,q}(-2uv^{-1}u)\right]=0$, we get that $\left(\operatorname{ad}\tilde{e}_{\epsilon_p-\epsilon_q,i}\right)^{2+1}\tilde{f}_{\epsilon_k-\epsilon_l,j}=0$.

E. $\left[\tilde{f}_{\epsilon_p-\epsilon_q,i}, \tilde{e}_{\epsilon_k-\epsilon_l,j}\right] = \left[E^{\blacksquare\square}_{q,p}(u^{-1}), E^{\blacksquare\square}_{k,l}(v)\right]$

$$= \delta_{p,k}E^{\blacksquare\square}_{q,l}(u^{-1}v) + \delta_{l,q}E^{\blacksquare\square}_{k,p}(-vu^{-1})$$
$$= \begin{cases} E^{\blacksquare\square}_{q,l}(u^{-1}\,v) & \text{if } p=k \text{ but } q\neq l,\\ E^{\blacksquare\square}_{k,p}(-v\,u^{-1}) & \text{if } p\neq k \text{ but } q=l,\\ E^{\blacksquare\square}_{p,p}(-v\,u^{-1}) + E^{\blacksquare\square}_{q,q}(u^{-1}\,v) & \text{if } p=k \text{ and } q=l,\\ 0 & \text{otherwise.}\end{cases}$$

- If $p=k$ but $q\neq l$, then $\left[E^{\blacksquare\square}_{q,p}(u^{-1}), E^{\blacksquare\square}_{q,l}(u^{-1}v)\right]=0$ because $p\neq q$ and $l\neq q$.

- If $p \neq k$ but $q = l$, then $\big[E^{\blacksquare\blacksquare}_{q,p}(u^{-1}), E^{\blacksquare\blacksquare}_{k,p}(-v\,u^{-1})\big] = 0$ because $p \neq k$ and $p \neq q$.
- If $p = k$ and $q = l$, then

$$\begin{aligned}\big[E^{\blacksquare\blacksquare}_{q,p}(u^{-1}), E^{\blacksquare\blacksquare}_{p,p}(-vu^{-1}) + E^{\blacksquare\blacksquare}_{q,q}(u^{-1}v)\big] &= E^{\blacksquare\blacksquare}_{q,p}(-u^{-1}vu^{-1}) + E^{\blacksquare\blacksquare}_{q,p}(-u^{-1}vu^{-1}) \\ &= E^{\blacksquare\blacksquare}_{q,p}(-2u^{-1}vu^{-1}).\end{aligned}$$

So

$$\big(\operatorname{ad} \tilde{f}_{\epsilon_p-\epsilon_q,i}\big)^{1+1}\tilde{e}_{\epsilon_k-\epsilon_l,j} = \begin{cases} E^{\blacksquare\blacksquare}_{q,p}(-2u^{-1}vu^{-1}) & \text{if } p = k \text{ and } q = l, \\ 0 & \text{otherwise.}\end{cases}$$

Since $\big[E^{\blacksquare\blacksquare}_{q,p}(u^{-1}), E^{\blacksquare\blacksquare}_{q,p}(-2u^{-1}vu^{-1})\big] = 0$, we get

$$(\operatorname{ad} \tilde{f}_{\epsilon_p-\epsilon_q,i})^{2+1}\tilde{e}_{\epsilon_k-\epsilon_l,j} = 0.$$

F. $$\begin{aligned}\big[\tilde{h}_{\epsilon_p-\epsilon_q,i}, \tilde{e}_{\epsilon_k-\epsilon_l,j}\big] &= \big[E^{\blacksquare\blacksquare}_{p,p}(1) + E^{\blacksquare\blacksquare}_{q,q}(-1), E^{\blacksquare\blacksquare}_{k,l}(v)\big] \\ &= \delta_{p,k}E^{\blacksquare\blacksquare}_{p,l}(v) + \delta_{l,p}E^{\blacksquare\blacksquare}_{k,p}(-v) + \delta_{q,k}E^{\blacksquare\blacksquare}_{q,l}(-v) + \delta_{l,q}E^{\blacksquare\blacksquare}_{k,q}(v) \\ &= \delta_{p,k}E^{\blacksquare\blacksquare}_{k,l}(v) - \delta_{p,l}E^{\blacksquare\blacksquare}_{k,l}(v) - \delta_{q,k}E^{\blacksquare\blacksquare}_{k,l}(v) + \delta_{q,l}E^{\blacksquare\blacksquare}_{k,l}(v) \\ &= (\delta_{p,k} - \delta_{p,l} - \delta_{q,k} + \delta_{q,l})E^{\blacksquare\blacksquare}_{k,l}(v) \\ &= A_{\epsilon_p-\epsilon_q,\epsilon_k-\epsilon_l}\tilde{e}_{\epsilon_k-\epsilon_l,j}.\end{aligned}$$

G. $$\begin{aligned}&\big[\tilde{h}_{\epsilon_p-\epsilon_q,i}, \tilde{f}_{\epsilon_k-\epsilon_l,j}\big] \\ &= \big[E^{\blacksquare\blacksquare}_{p,p}(1) + E^{\blacksquare\blacksquare}_{q,q}(-1), E^{\blacksquare\blacksquare}_{l,k}(v^{-1})\big] \\ &= \delta_{p,l}E^{\blacksquare\blacksquare}_{p,k}(v^{-1}) + \delta_{k,p}E^{\blacksquare\blacksquare}_{l,p}(-v^{-1}) + \delta_{q,l}E^{\blacksquare\blacksquare}_{q,k}(-v^{-1}) + \delta_{k,q}E^{\blacksquare\blacksquare}_{l,q}(v^{-1}) \\ &= \delta_{p,l}E^{\blacksquare\blacksquare}_{l,k}(v^{-1}) - \delta_{p,k}E^{\blacksquare\blacksquare}_{l,k}(v^{-1}) - \delta_{q,l}E^{\blacksquare\blacksquare}_{l,k}(v^{-1}) + \delta_{q,k}E^{\blacksquare\blacksquare}_{l,k}(v^{-1}) \\ &= -(\delta_{p,k} - \delta_{p,l} - \delta_{q,k} + \delta_{q,l})E^{\blacksquare\blacksquare}_{l,k}(v^{-1}) \\ &= -A_{\epsilon_p-\epsilon_q,\epsilon_k-\epsilon_l}\tilde{f}_{\epsilon_k-\epsilon_l,j}\end{aligned}$$ □

Proposition 15. *The map $\varphi : \mathfrak{gim}(A^{[d]}) \to \mathfrak{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$ is a surjective Lie algebra homomorphism.*

Proof. Let $\mathfrak{B} = \operatorname{Im}(\varphi) \subseteq L$, where $L = \mathfrak{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$. We show that $\mathfrak{B} = L$ in a sequence of steps.

#1. Let $\mathfrak{g} = \big\{M \in \mathrm{M}_{2r+1}(\mathbb{C}) : M^tG = -GM\big\}$ and

$$\mathfrak{s} = \big\{M \in \mathrm{M}_{2r+1}(\mathbb{C}) : M^tG = GM, \operatorname{tr}(M) = 0\big\}.$$

Let $A = \{a \in \mathfrak{b} : \eta a = a\}$ and $B = \{b \in \mathfrak{b} : \eta b = -b\}$. If $0 \neq a \in A$, then $a\mathfrak{g}$ is an irreducible $\mathfrak{g}$-submodule of $\mathfrak{B}$ with highest weight $\epsilon_1 + \epsilon_2$. If $0 \neq b \in \mathfrak{B}$, then $b\mathfrak{s}$ is an irreducible $\mathfrak{g}$-submodule of $\mathfrak{B}$ with highest weight $2\epsilon_1$. These $\mathfrak{g}$-modules are not isomorphic.

#2. $\mathfrak{B}$ is a subalgebra of L containing $\mathfrak{g}$, so $\mathfrak{B}$ is a $\mathfrak{g}$-submodule of L.

#3. For $1 \leq p, q \leq 2r+1$, $p \neq q$, $p \neq 2r+2-q$, let

$$I_{pq} = \big\{x \in \mathfrak{b} : E_{pq}(x) - E_{2r+2-q,2r+2-p}(\eta x) \in \mathfrak{B}\big\}.$$

Notice that I_{pq} is a subspace of $\mathfrak{B}$.

#4. I_{pq} is invariant under η. Indeed, let $x \in I_{pq}$, in which case

$$X := E_{pq}(x) - E_{2r+2-q,2r+2-p}(\eta x) \in \mathfrak{B}.$$

But $X = X_1 + X_2$, where $X_1 = \frac{1}{2}(x+\eta x)\big(E_{pq}(1) - E_{2r+2-q,2r+2-p}(1)\big) \in (x+\eta x)\mathfrak{g}$ and $X_2 = \frac{1}{2}(x-\eta x)\big(E_{pq}(1) + E_{2r+2-q,2r+2-p}(1)\big) \in (x-\eta x)\mathfrak{s}$. Thus, by #1 and #2, $X_1, X_2 \in \mathfrak{B}$. So $x+\eta x, x-\eta x \in I_{pq}$, which implies that $\eta x \in I_{pq}$.

#5. By #4, $I_{pq} = I_{pq} \cap A + I_{pq} \cap B$. But by #1 and #2, $I_{pq} \cap A$ and $I_{pq} \cap B$ are independent of p, q. So $I := I_{pq}$ is independent of p, q.

#6. We have

$$\big[E_{12}(x) - E_{2r,2r+1}(\eta x),\, E_{23}(y) - E_{2r-1,2r}(\eta y)\big] = E_{13}(xy) - E_{2r-1,2r+1}\big((\eta y)(\eta x)\big).$$

So, by #5, I is a subalgebra of $\mathfrak{B}$, and, by #4, I is invariant under η.

#7. The action of φ on the generators of $\mathfrak{gim}(A^{[d]})$ tells us that I contains the elements $x_{\omega,i}, x_{\omega,i}^{-1}, y_{\theta,i}, y_{\theta,i}^{-1}$. So by #6, $I = \mathfrak{B}$.

#8. By #7, we have $A\mathfrak{g} + B\mathfrak{s} \subseteq \mathfrak{B}$. But since $C = \{0\}$, we have $\sum_{\alpha\in\Delta} L_\alpha \subseteq A\mathfrak{g} + B\mathfrak{s}$. So $\sum_{\alpha\in\Delta} L_\alpha \subseteq \mathfrak{B}$. Hence, since $\mathfrak{B}$ is a subalgebra of L, $\mathfrak{B} = L$. □

Continuing our plan laid out on page 273, we next show that $\varphi : \mathfrak{gim}(A^{[d]}) \to \mathrm{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$ is a graded homomorphism and that it induces a map from $\mathfrak{im}(A^{[d]})$ to $\mathrm{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$.

We saw in Sections 3 and 4, respectively, that $\mathfrak{gim}(A^{[d]})$ and $\mathrm{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$ are both Γ-graded Lie algebras, where

$$\Gamma = \bigoplus_{\mu\in\Delta} \mathbb{Z}\alpha_\mu.$$

The map $\varphi : \mathfrak{gim}(A^{[d]}) \to \mathrm{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$ is engineered so that, for all $\alpha \in \Gamma$,

$$\varphi\big(\mathfrak{gim}(A^{[d]})_\alpha\big) \subset \mathrm{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)_\alpha.$$

That is, the following result holds by design.

Proposition 16. *The map $\varphi : \mathfrak{gim}(A^{[d]}) \to \mathrm{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$ is also a graded homomorphism.*

Moreover, since $\mathrm{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)_\gamma = 0$ for $\gamma \notin \Delta \cup \{0\}$, we get that the radical $\mathfrak{r}$ of $\mathfrak{gim}(A^{[d]})$ lies in the kernel of φ.

Proposition 17. *There exists a surjective, graded Lie algebra homomorphism*

$$\phi : \mathfrak{im}(A^{[d]}) \to \mathrm{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$$

given by $\phi(u + \mathfrak{r}) = \varphi(u)$ *for any* $u + \mathfrak{r} \in \mathfrak{im}(A^{[d]})$, *where* $u \in \mathfrak{gim}(A^{[d]})$.

We now turn to centrality. Let $\mathrm{so}(\mathfrak{a})$ and $\mathrm{so}(\mathfrak{b})$ be shorthand for $\mathrm{so}_{2r+1}(\mathfrak{a}, \eta, C, \chi)$ and $\mathrm{so}_{2r+1}(\mathfrak{b}, \eta, C, \chi)$, respectively.

Since the elements of $\mathfrak{a}$ satisfy the defining relations of $\mathfrak{b}$, by universality, there exists a surjective associative algebra homomorphism $g : \mathfrak{b} \to \mathfrak{a}$. In particular, $g(x_{\omega,i}) = a \in \mathfrak{a}$ if $\psi(e_{\omega,i}) = E_{pq}(a) - E_{2r+2-q,2r+2-p}(\eta a)$, and $g(y_{\theta,i}) = b \in \mathfrak{a}$ if $\psi(e_{\theta,i}) = E_{pq}(b) - E_{2r+2-q,2r+2-p}(\eta b)$. This algebra homomorphism respects the involution and induces a surjective Lie algebra homomorphism

$$\widetilde{g} : \mathrm{so}(\mathfrak{b}) \to \mathrm{so}(\mathfrak{a}) \quad \text{such that } \widetilde{g}\phi = \psi.$$

Hence $\ker\phi \subset \ker\psi \subset \mathfrak{z}\big(\mathfrak{im}(A^{[d]})\big)$, where $\mathfrak{z}\big(\mathfrak{im}(A^{[d]})\big)$ denotes the center of $\mathfrak{im}(A^{[d]})$. Thus $\ker\phi \subset \mathfrak{z}\big(\mathfrak{im}(A^{[d]})\big)$, implying the following result:

Proposition 18. *The map* $\phi : \mathfrak{im}(A^{[d]}) \to \mathrm{so}(\mathfrak{b})$ *is a central homomorphism.*

We also know that $\psi : \mathfrak{im}(A^{[d]}) \to \mathrm{so}(\mathfrak{a})$ is a universal central extension: so there exists a Lie algebra homomorphism $\widetilde{f} : \mathrm{so}(\mathfrak{a}) \to \mathrm{so}(\mathfrak{b})$ such that $\widetilde{f}\psi = \phi$. Since ψ is surjective, the generators of $\mathfrak{a}$ are of the form a, a^{-1}, $\eta(a)$, where a is the element in $\mathfrak{a}$ corresponding to the image $\psi(e_{\lambda,i})$ of the i-th copy of a long root λ in Δ_B which was adjoined.

Since $\widetilde{f}\psi = \phi$, the map $\widetilde{f}$ induces an associative algebra homomorphism $f : \mathfrak{a} \to \mathfrak{b}$ given by

$$f(a) = \begin{cases} x_{\omega,i} & \text{if } a \text{ is the element in } \mathfrak{a} \text{ corresponding to} \\ & \text{the image } \psi(e_{\omega,i}) = E_{pq}(a) - E_{2r+2-q,2r+2-p}(\eta a) \\ & \text{for some } 1 \le p, q \le 2r+1 \text{ with } p \ne q,\ p \ne 2r+2-q \\ & \text{and } \omega \in \Omega,\ 1 \le i \le N_\omega, \\ y_{\theta,i} & \text{if } a \text{ is the element in } \mathfrak{a} \text{ corresponding to} \\ & \text{the image } \psi\left(e_{\theta,i}\right) = E_{pq}(a) - E_{2r+2-q,2r+2-p}(\eta a) \\ & \text{for some } 1 \le p, q \le 2r+1 \text{ with } p \ne q,\ p \ne 2r+2-q \\ & \text{and } \theta \in \Theta,\ 1 \le i \le N_\theta. \end{cases}$$

We define $f(\eta a)$ to be $\eta f(a)$ so that f preserves the involution.

But then $g \circ f = \mathrm{id}_{\mathfrak{a}}$ and $f \circ g = \mathrm{id}_{\mathfrak{b}}$, that is, $\mathfrak{a} \cong \mathfrak{b}$ as associative algebras.

Acknowledgments

We thank Frank Lemire and Michael Lau for their helpful discussions, and the anonymous referees for their useful comments and shortened proofs.

References

[Allison et al. 2002] B. Allison, G. Benkart, and Y. Gao, *Lie algebras graded by the root systems BC_r, $r \geq 2$*, Mem. Amer. Math. Soc. **751**, Amer. Math. Soc., Providence, RI, 2002. MR 2003h:17038 Zbl 0998.17031

[Benkart and Zelmanov 1996] G. Benkart and E. Zelmanov, "Lie algebras graded by finite root systems and intersection matrix algebras", *Invent. Math.* **126**:1 (1996), 1–45. MR 97k:17044 Zbl 0871.17024

[Berman 1989] S. Berman, "On generators and relations for certain involutory subalgebras of Kac–Moody Lie algebras", *Comm. Algebra* **17**:12 (1989), 3165–3185. MR 91a:17030 Zbl 0693.17012

[Berman and Moody 1992] S. Berman and R. V. Moody, "Lie algebras graded by finite root systems and the intersection matrix algebras of Slodowy", *Invent. Math.* **108**:2 (1992), 323–347. MR 93e:17031 Zbl 0778.17018

[Berman et al. 2001] S. Berman, E. Jurisich, and S. Tan, "Beyond Borcherds Lie algebras and inside", *Trans. Amer. Math. Soc.* **353**:3 (2001), 1183–1219. MR 2001f:17048 Zbl 0973.17031

[Carter 2005] R. W. Carter, *Lie algebras of finite and affine type*, Cambridge Studies in Advanced Mathematics **96**, Cambridge University Press, 2005. MR 2006i:17001 Zbl 1110.17001

[Gabber and Kac 1981] O. Gabber and V. G. Kac, "On defining relations of certain infinite-dimensional Lie algebras", *Bull. Amer. Math. Soc.* (*N.S.*) **5**:2 (1981), 185–189. MR 84b:17011 Zbl 0474.17007

[Gao 1996] Y. Gao, "Involutive Lie algebras graded by finite root systems and compact forms of IM algebras", *Math. Z.* **223**:4 (1996), 651–672. MR 98b:17026 Zbl 0884.17012

[Kac 1990] V. G. Kac, *Infinite-dimensional Lie algebras*, 3rd ed., Cambridge University Press, 1990. MR 92k:17038 Zbl 0716.17022

[Moody and Pianzola 1995] R. V. Moody and A. Pianzola, *Lie algebras with triangular decompositions*, Canadian Mathematical Society Series of Monographs and Advanced Texts, Wiley, New York, 1995. MR 96d:17025 Zbl 0874.17026

[Neher 1996] E. Neher, "Lie algebras graded by 3-graded root systems and Jordan pairs covered by grids", *Amer. J. Math.* **118**:2 (1996), 439–491. MR 97a:17029 Zbl 0857.17019

[Peng 2002] L. Peng, "Intersection matrix Lie algebras and Ringel–Hall Lie algebras of tilted algebras", pp. 98–108 in *Representations of algebra, I*, edited by D. Happel and Y. B. Zhang, Beijing Norm. Univ. Press, Beijing, 2002. MR 2005m:16022 Zbl 1086.16517

[Rao et al. 1992] S. E. Rao, R. V. Moody, and T. Yokonuma, "Lie algebras and Weyl groups arising from vertex operator representations", *Nova J. Algebra Geom.* **1**:1 (1992), 15–57. MR 93h:17040 Zbl 0867.17022

[Slodowy 1984] P. Slodowy, *Singularitäten, Kac–Moody Lie-Algebren, assoziierte Gruppen und Verallgemeinerungen*, Habilitationsschrift, Universität Bonn, 1984.

[Slodowy 1986] P. Slodowy, "Beyond Kac–Moody algebras, and inside", pp. 361–371 in *Lie algebras and related topics* (Windsor, Canada, 1984), edited by D. J. Britten et al., CMS Conf. Proc. **5**, Amer. Math. Soc., Providence, RI, 1986. MR 832211 Zbl 0582.17011

Received April 22, 2012. Revised August 28, 2012.

SANDEEP BHARGAVA
MATHEMATICS DEPARTMENT, SCHOOL OF LIBERAL ARTS AND SCIENCES
HUMBER INSTITUTE OF TECHNOLOGY AND ADVANCED LEARNING
205 HUMBER COLLEGE BOULEVARD
TORONTO, ON M9W 5L7
CANADA
sandeep.bhargava@humber.ca

YUN GAO
DEPARTMENT OF MATHEMATICS AND STATISTICS
YORK UNIVERSITY
4700 KEELE STREET
TORONTO, ON M3J 1P3
CANADA
ygao@yorku.ca

PACIFIC JOURNAL OF MATHEMATICS
Vol. 263, No. 2, 2013
dx.doi.org/10.2140/pjm.2013.263.283

STABLE FLAGS, TRIVIALIZATIONS AND REGULAR CONNECTIONS

ELIE COMPOINT AND EDUARDO COREL

We study stalkwise modifications of a holomorphic vector bundle endowed with a meromorphic connection on a compact Riemann surface. We introduce the notion of Birkhoff–Grothendieck trivialization, in the case of the Riemann sphere, and show that its computation corresponds to shortest paths in some local affine Bruhat–Tits building. We use this to compute how the type of a bundle changes under stalk modifications, and give several corresponding algorithmic procedures. We finally deduce from these results some applications to the Riemann–Hilbert problem.

Introduction

The motivation of this article originates in the Riemann–Hilbert problem on a compact Riemann surface, and the present work follows it as a guideline. The results presented herein are however not directly related to this problem. The reader who is exclusively interested in new advances on the Riemann–Hilbert problem will nevertheless find a couple of improvements on already known results. The real interest of the present paper in the eyes of its authors consists in translating this classical problem in a "new" setting (the setting of Bruhat–Tits buildings). In this new context, the Riemann–Hilbert problem reveals new geometric objects (such as *Birkhoff–Grothendieck trivializations*), whose study appears to be interesting by itself, and seems also promising for the original problem.

The Riemann–Hilbert problem (RHP) has a long and distinguished history, not even devoid of suspense, for it has been solved several times, using different tools, in a seemingly complete and positive way. It was finally A. A. Bolibrukh, in a celebrated series of papers at the beginning of the 1990s, who clarified the situation, by rigorously defining (and exhibiting a counterexample to) the strongest version of the RHP, thereby showing that people before him had either committed a mistake, or solved in reality a weaker problem.

MSC2010: primary 51N30, 34M03; secondary 34M50.

Keywords: meromorphic connection, vector bundle, Birkhoff–Grothendieck theorem, Bruhat–Tits building, Riemann–Hilbert problem.

The modern approach to the RHP was initiated by H. Röhrl in the 1950s, who used the theory of vector bundles in a way that has been conserved since. First, one constructs a vector bundle $\mathscr{E}$ outside the singular points, whose cocycle mimics the monodromy. We call this the *topological* RH problem, since the monodromy is so much encoded in the topology of the constructed bundle, that construction of the required connection becomes essentially trivial. The second step consists of extending the bundle (and the connection) to the singular points by means of a *local solution* to the inverse monodromy problem. It has been exposed in great generality in [Deligne 1970] how to extend a holomorphic vector bundle $\mathscr{E}$, defined over the complement of a divisor $\mathscr{D}$ and endowed with a holomorphic connection ∇ having a prescribed monodromy about $\mathscr{D}$, into a logarithmic connection $(\overline{\mathscr{E}}, \overline{\nabla})$ with singularities on the divisor.

In this way, we get all logarithmic extensions of $\mathscr{E}$ with *nonresonant* residue (the *Deligne lattices*). These two steps are sufficient to solve positively the weak Riemann–Hilbert problem (i.e., with regular singularities). Note, however, that in this second level, two different types of problems have been mixed. The connection constructed is essentially unique up to *meromorphic* equivalence whereas the holomorphic vector bundle setting already introduces much finer *holomorphic* equivalence problems. This fact can contribute to explain some of the confusions that have surrounded the precise formulation of the RHP.

The *strong Riemann–Hilbert problem* asks for a logarithmic bundle (with the prescribed monodromy) which is moreover *trivial*. So, to solve the Riemann–Hilbert problem in this way, one must modify the constructed Deligne bundle, over the support of the singular divisor exclusively (to keep the singular set invariant), while conserving its logarithmic character, until a trivial bundle is eventually found. Until Bolibrukh's celebrated counterexample [1990], it was widely acknowledged that this was possible, and it is indeed so in several "generic" instances, although some mistakes in the seemingly general solution by Plemelj had already been pointed out (e.g., in [Treibich Kohn 1983]).

The counterexample found by Bolibrukh to the strong Riemann–Hilbert problem requires the knowledge of *all* the logarithmic extensions of a regular connection, in order to prove that none is trivial. Despite the production of both counterexamples and sufficient conditions for a positive answer, no general necessary and sufficient conditions for the solubility of the strong Riemann–Hilbert problem have been given in terms of the monodromy representation only, except for the remarkable case of an *irreducible* representation [Bolibrukh 1990; Kostov 1992].

As already stated, the strong Riemann–Hilbert problem admits a solution if and only if the stalks of the Deligne bundle over the singular set can be replaced by logarithmic lattices in such a way that the resulting bundle is trivial. To tackle this problem, one should be able to

(a) determine the set of all logarithmic lattices above a given point, and

(b) get a criterion for the triviality of the modified bundle.

Problem (a) can be considered classical, in the sense that it has been repeatedly solved, under different guises, going back as far as [Gantmacher 1959], and including [Levelt 1961; Babbitt and Varadarajan 1983; Bolibrukh 1990; Sabbah 2002]. In this paper, we give a complete description of the logarithmic lattices in terms of flags stabilized under the action of the residue of the connection; in a way, our contribution is to make the solution to problem (a) given in [Sabbah 2002] completely explicit and effective. We also give a partial answer to problem (b). In the case of $\mathbb{P}^1(\mathbb{C})$, the type of a vector bundle gives such a triviality criterion. In our approach, starting with the Deligne bundle $\mathfrak{D}$, we perform a modification of a finite number of stalks, resulting in the bundle $\mathfrak{D}^{\mathrm{mod}}$. The question is then to compute the type of the modified bundle $\mathfrak{D}^{\mathrm{mod}}$. Generalizing a result by Gabber and Sabbah (Proposition 28), we show how to determine the type of $\mathfrak{D}^{\mathrm{mod}}$ from the type of $\mathfrak{D}$. Thus, problem (b) is reduced to computing the type of the Deligne bundle. In a second step, we show that this problem in turn is reduced to the well-known problem of *connection matrices*.

With these problems in mind, we introduce and study in this paper the notion of *Birkhoff–Grothendieck trivialization* of a bundle $\mathscr{E}$, which is a pointwise modification of $\mathscr{E}$ such that

(i) the resulting bundle $\mathscr{F}$ is holomorphically trivial, and

(ii) the relative elementary divisors of the stalks give the *type* of the bundle $\mathscr{E}$.

The paper is organized as follows. In a first section, we define the category in which we will work, and what we precisely mean by "modifying a bundle over one or several points". In a second part, we describe the geometry on the local lattices involved. We describe this geometry in terms of the affine Bruhat–Tits building of SL_n.

The third part contains the main results of the paper. We use the previously introduced setting to give an effective method to compute how the type of an arbitrary bundle E is modified under certain pointwise modifications. This algorithm can also be applied to compute the type of the bundle E. This third section concludes with a generalization of an essential result originally due to Bolibrukh, the *permutation lemma*, for which we provide an interesting geometric interpretation. This result allows us to give a quite complete insight into the structure of Birkhoff–Grothendieck trivializations, which we sum up as follows.

Theorem 1. *Let $\mathbf{\Lambda}^0$ be the set of pointwise modifications over $x \in \mathbb{P}^1(\mathbb{C})$ of a bundle $\mathscr{E}$ that give a trivial bundle. Then $M \in \mathbf{\Lambda}^0$ is a Birkhoff–Grothendieck trivialization of $\mathscr{E}$ if and only if M realizes the minimum of the canonical metric between $\mathscr{E}_x$ and $\mathbf{\Lambda}^0$ in the local affine Bruhat–Tits building of* SL_n *at x.*

The fourth section gives the complete description of the set of logarithmic lattices in terms of flags which are stable under the action of the residue of the connection on the Deligne bundle[1].

In the last part, we use these tools for study of the Riemann–Hilbert problem. After recalling the construction of the classical Röhrl–Deligne bundle, we give a very short proof of Plemelj's theorem on the Riemann–Hilbert solubility. This well-known result becomes an immediate consequence of the geometrical interpretations of the permutation lemma and the set of logarithmic lattices. We describe all trivializations of the Deligne bundle over an arbitrary point, and we give a concise proof of the Bolibrukh–Kostov theorem on the solvability of Riemann–Hilbert in the irreducible case. Finally, we give algorithmically effective procedures that allow to search the space of weak solutions.

1. Holomorphic vector bundles

Let X be a compact Riemann surface and let $\pi : E \to X$ be a holomorphic vector bundle of rank n. The sheaf $\mathscr{E}$ of holomorphic sections of E is a locally free sheaf of $\mathscr{O}_X$-modules of the same rank n, where $\mathscr{O}_X$ denotes as usual the sheaf of holomorphic functions on X. There is a well-known equivalence between these two categories. However, this equivalence does not hold for subobjects of the same rank. Therefore we will privilege the sheaf-theoretical formulation.

Meromorphic connections. Let $\mathscr{D} = \sum_{i=1}^{p} m_i x_i$ be a positive divisor on X. Let $\mathscr{O}_{\mathscr{D}}$ be the sheaf of meromorphic functions on X having pole orders bounded by $\mathscr{D}$, and $|\mathscr{D}| = \{x_1, \dots, x_p\}$ be the *support* of $\mathscr{D}$. For any finite set $S = \{y_1, \dots, y_t\}$, let $(S) = y_1 + \cdots + y_t$.

Let $\nabla : \mathscr{E} \to \mathscr{E} \otimes_{\mathscr{O}_X} \Omega^1_X(\mathscr{D})$ be a meromorphic connection with singular divisor $\mathscr{D}$ on a vector bundle $\mathscr{E}$ of rank n. Sometimes for simplicity we'll just say "connection" for the pair $(\mathscr{E}, \nabla)$. If we assume that $\mathscr{D}$ is the smallest possible (as we will do), then the *Poincaré rank* of ∇ at $x \in X$ is the integer $\mathfrak{p}_x(\nabla) = \max(0, m_x - 1)$. We will omit ∇ whenever possible. If $\mathfrak{p}_x = 0$, the sheaf $\mathscr{E}$ is said to be *logarithmic at* x. Let $\mathscr{S} = |\mathscr{D}|$ be the *singular*, and $\mathscr{S}_{\log} = \{x \in \mathscr{S} \mid \mathfrak{p}_x = 0\}$ the *logarithmic singular sets* of ∇. If $\mathscr{S}_{\log} \neq \varnothing$, then one can define the residue map $\mathrm{Res}\nabla \in \mathrm{End}(\mathscr{E}/\mathscr{E}_{-(\mathscr{S}_{\log})})$. We will specify in parentheses the bundle if necessary.

The meromorphic bundle. Let $\mathscr{M}_X$ be the sheaf of meromorphic functions on X and let $\mathscr{V} = \mathscr{E} \otimes_{\mathscr{O}_X} \mathscr{M}_X$ be the sheaf of meromorphic sections of E. A meromorphic connection ∇ on $\mathscr{E}$ has a canonical extension to $\mathscr{V}$. Since the sheaf $\mathscr{E}$ can be

[1] Independently, P. Boalch [2011] has taken a similar view on local logarithmic lattices, in terms of stable filtrations and Bruhat–Tits buildings, albeit on the more general setting of a complex reductive group. Restricting to GL_n and SL_n enables us to give however more explicit formulæ; see Sections 3.2 and 3.3.

embedded into $\mathcal{V}$, we consider from now on the set

$$\mathcal{H} = \left\{ \mathcal{F} \subset \mathcal{V} \;\middle|\; \mathcal{F} \overset{\text{loc.}}{\simeq} \mathcal{O}_X^n \right\}$$

of *holomorphic vector bundles* of $\mathcal{V}$. Each such bundle $\mathcal{F} \in \mathcal{H}$ is automatically endowed with a meromorphic connection induced by ∇. For simplicity, we won't make any notational difference between all these connections.

We say that $\mathcal{F} \in \mathcal{H}$ is *trivial* if $\mathcal{F} \simeq \mathcal{O}_X^n$, or, equivalently, if $\mathcal{F}$ is generated by its global sections, and *quasitrivial* if there exists a line bundle $\mathcal{L}$ such that $\mathcal{F} \otimes_{\mathcal{O}_X} \mathcal{L}$ is trivial. Let $\mathcal{H}_0 \subset \mathcal{H}$ be the subset of trivial holomorphic bundles in $\mathcal{V}$. The following result is well known (e.g., [Sabbah 2002, p. 134]).

Lemma 2. *Let $\mathcal{F} \in \mathcal{H}_0$ be a trivial holomorphic vector bundle in $(\mathcal{V}, \nabla)$. The space $\mathcal{F}(X) = \Gamma(X, \mathcal{F})$ of global sections is a $\mathbb{C}$-vector space of dimension n. For any logarithmic singularity $s \in \mathcal{S}_{\log}(\mathcal{F})$, the residue $\operatorname{Res}_s^{\mathcal{F}} \nabla$ induces a well-defined endomorphism $\psi_s \in \operatorname{End}_{\mathbb{C}}(\mathcal{F}(X))$.*

Stalks and lattices. For $\mathcal{F} \in \mathcal{H}$, the stalk $\mathcal{F}_x$ at any $x \in X$ is a free $(\mathcal{O}_X)_x$-submodule of rank n (or *lattice*) of $\mathcal{V}_x$, which is a vector space of dimension n over $(\mathcal{M}_X)_x$. Let $\mathcal{O}_x = \widehat{(\mathcal{O}_X)_x}$ be the formal completion of $(\mathcal{O}_X)_x$, and $K_x = \operatorname{Frac}(\mathcal{O}_x)$ its field of fractions. Similarly, the formal completion $\widehat{\mathcal{F}_x} = \mathcal{F}_x \otimes_{(\mathcal{O}_X)_x} \mathcal{O}_x$ is a lattice in the vector space $V_x = \mathcal{V}_x \otimes_{(\mathcal{M}_X)_x} K_x$. This operation is harmless, as $\Lambda \mapsto \widehat{\Lambda}$ is a bijection between the sets of lattices in $\mathcal{V}_x$ and V_x (cf. [Malgrange 1996]). We define an equivalence relation $\sim_x$ on $\mathcal{H}$ as

$$\mathcal{F} \sim_x \widetilde{\mathcal{F}} \quad \text{if and only if } \mathcal{F}_{|X\setminus\{x\}} = \widetilde{\mathcal{F}}_{|X\setminus\{x\}}.$$

For simplicity, we will drop the index x as soon as no ambiguity can arise. Let $\mathbf{\Lambda}_x$ be the set of lattices in V_x. Any coset $[\mathcal{F}]$ of $\mathcal{H}/\sim_x$ can be identified with the set $\mathbf{\Lambda}_x$, by identifying $\mathcal{F}' \in [\mathcal{F}]$ with its formal stalk $\widehat{\mathcal{F}'_x} \in \mathbf{\Lambda}_x$ at x. Since X is compact, two vector bundles $\mathcal{E}$, $\mathcal{F} \in \mathcal{H}$ have equal stalks outside a finite set.

Lemma 3. *Let $\mathcal{E} \in \mathcal{H}$ be a holomorphic vector bundle. For any family of formal lattices $M_x \in \mathbf{\Lambda}_x$ for x in a discrete set $\mathcal{S}$, there exists a unique vector bundle $\mathcal{E}^M \in \mathcal{H}$ such that*

$$(\mathcal{E}^M)_x = \begin{cases} \mathcal{E}_x & \text{if } x \notin \mathcal{S}, \\ M_x \cap \mathcal{V}_x & \text{if } x \in \mathcal{S}. \end{cases}$$

Conversely, for any $\mathcal{F} \in \mathcal{H}$, there exists a discrete set $\mathcal{S}$ and a family $(M_x \in \mathbf{\Lambda}_x)_{x\in\mathcal{S}}$ of lattices such that $\mathcal{F} = \mathcal{E}^M$. If $\mathcal{E}$ is endowed with a meromorphic connection ∇, there is a canonical extension ∇^M of $\nabla_{|X\setminus\mathcal{S}}$ as a meromorphic connection on $\mathcal{E}^M$. In particular, a bundle $\mathcal{E} \in \mathcal{H}$ is completely determined by its coset $[\mathcal{E}] \in \mathcal{H}/\sim_x$ and the lattice $\widehat{\mathcal{E}_x} \in \mathbf{\Lambda}_x$.

The sheaf $\mathscr{V}$ is trivial (as a meromorphic bundle), and the group G of (meromorphic) automorphisms of the space $\Gamma(X,\mathscr{V})$ is isomorphic to $\mathrm{GL}_n(\mathbb{C}(X))$. Let $\mathbf{\Lambda}_x^0 = \mathbf{\Lambda}_x \cap \mathscr{H}_0$ be the set of trivial bundles in the coset $\mathbf{\Lambda}_x$. The subgroup $G_x \subset G$ of automorphisms of $\Gamma(X,\mathscr{V})$ that leave $\mathbf{\Lambda}_x^0$ globally invariant is called the *group of monopole gauge transforms at* x. Each element of G_x sends a trivial sheaf $\mathscr{F}$ to a trivial sheaf $\widetilde{\mathscr{F}}$ such that $\mathscr{F}_{|X\setminus\{x\}} = \widetilde{\mathscr{F}}_{|X\setminus\{x\}}$. An element of G_x modifies at most the stalk $\mathscr{F}_x$.

2. Lattices and the affine building of SL_n

2.1. ***Flags and filtrations.*** Let V be a vector space over a field $\mathbb{K}$ of characteristic 0. Given a flag $\boldsymbol{F}$ of vector spaces $0 = F_0 \subset F_1 \subset \cdots \subset F_s = V$, with *length* $|\boldsymbol{F}| = s$ and *signature* $\sigma(\boldsymbol{F}) = (n_1,\dots,n_s)$ where $n_i = \dim_{\mathbb{K}}(F_i/F_{i-1})$, a map $u \in \mathrm{End}_{\mathbb{K}}(V)$ *stabilizes* the flag $\boldsymbol{F}$ if $u(F_i) \subset F_i$ for all $0 \leqslant i \leqslant |\boldsymbol{F}|$. Let $\mathfrak{Fl}(V)$ denote the set of flags of V, and $\mathfrak{Fl}_u(V)$ the subset of flags that are stabilized by u. A flag $\boldsymbol{F}'$ is *transversal* to $\boldsymbol{F}$ if $|\boldsymbol{F}'| = |\boldsymbol{F}| = s$ and $F_i' \oplus F_{s-i} = V$ for $1 \leqslant i \leqslant s$. In this case, the signature of $\boldsymbol{F}'$ is equal to $\sigma^- = (n_s,\dots,n_1)$. The left action of the permutation group S_n on a sequence $a = (a_1,\dots,a_n)$ is given by $\tau a = (a_{\tau(1)},\dots,a_{\tau(n)})$ for $\tau \in S_n$. Let s_i be the transposition $(i, i+1)$ exchanging i and $i+1$. The set $S = \{s_1,\dots,s_{n-1}\}$ makes (S_n, S) into a Coxeter group of type A_{n-1}.

Let (e) be a basis of V, and let $\sigma = (n_1,\dots,n_s)$ be a signature. Let $\boldsymbol{F}^\sigma(e)$ be the flag with elements $F_i^\sigma(e) = \langle e_1,\dots,e_{\nu_i}\rangle$ where $\nu_i = n_1 + \cdots + n_i$. The basis (e) is said to be *adapted* to $\boldsymbol{F}$ if any element of the flag is spanned by a subfamily of (e), *strictly adapted* if $\boldsymbol{F} = F^\sigma(e)$, and *transversal* to $\boldsymbol{F}$ if $F_i \oplus \langle e_{\nu_i+1},\dots,e_n\rangle = V$ for all i. The *parabolic subgroup* stabilizing a flag $\boldsymbol{F}$ (for the action on a strictly adapted basis) is the subgroup $W_{\boldsymbol{F}} = \langle s_i \mid i \in I\rangle$ of S_n generated by the generators corresponding to the missing dimensions $I = [n]\setminus\{\dim_{\mathbb{K}} F_i \mid 1 \leqslant i \leqslant n\}$. These properties depend in fact only on the $\mathbb{K}$-vector subspaces spanned by the vectors of (e). The *opposite flag* $\boldsymbol{F}^{-(e)}$ is then defined as the unique flag transversal to $\boldsymbol{F}$ for which (e) is adapted. This last notion does not even depend on the *order* in which the vectors of (e) are taken. A flag $\boldsymbol{F}'$ is transversal to $\boldsymbol{F}$ if and only if there is a basis (e) of V strictly adapted to $\boldsymbol{F}$ such that $\boldsymbol{F}' = \boldsymbol{F}^{-(e)}$. A $\mathbb{K}$-*frame* is an unordered set $\Phi = \{L_1,\dots,L_n\}$ of one-dimensional $\mathbb{K}$-vector subspaces of V such that $L_1 + \cdots + L_n = V$. The notions defined in the previous paragraph make sense for a frame (with a fixed order on the lines for some of them). The *relative position* $\rho(\boldsymbol{F},\boldsymbol{F}') \in S_n$ of two flags $\boldsymbol{F}$ and $\boldsymbol{F}'$ is the[2] permutation $\tau \in S_n$ such that

[2]Strictly speaking, ρ is only unique when both flags are complete. Otherwise, we shall a bit imprecisely consider ρ either as its double coset in $W_{\boldsymbol{F}'}\rho W_{\boldsymbol{F}}$ modulo the parabolic subgroups attached to $\boldsymbol{F}$ and $\boldsymbol{F}'$, or to the unique minimal length representative of this coset (or possibly even to *any* such representative).

there exists a basis (e) strictly adapted to $\boldsymbol{F}$ for which $(\tau e) = (e_{\tau(1)}, \dots, e_{\tau(n)})$ is strictly adapted to $\boldsymbol{F}'$.

Similarly, a *flag* $\boldsymbol{S}$ of length s in $[n] = \{1, \dots, n\}$ is an increasing sequence $\boldsymbol{S} : S_0 = \varnothing \subset S_1 \subset \cdots \subset S_s = [n]$ of subsets of $[n]$, whose *signature* is the sequence $\sigma(\boldsymbol{S}) = (n_1, \dots, n_s)$ where $n_i = |S_i| - |S_{i-1}|$. For a given signature $\sigma = (n_1, \dots, n_s)$, the *standard ascending flag* $\boldsymbol{S}^{\nearrow}(\sigma)$ *of signature* σ is the flag composed of initial segments of $[n]$ of lengths $n_1 + \cdots + n_i$ for $1 \leqslant i \leqslant s$.

Given a sequence $D = (d_1, \dots, d_n) \in \mathbb{Z}^n$, let $D^{\nearrow}$ be the sequence of the elements of D arranged in increasing order. Define the *ascendent flag* $\boldsymbol{S}^{\nearrow}(D)$ *of* D as the sequence of subsets of indices corresponding to blocks of equal elements of $D^{\nearrow}$. Let also $(\mathbb{Z}^n)^{\nearrow}$ denote the set of nondecreasing integer sequences. Finally, let

$$D \cong D' \quad \text{if } D^{\nearrow} = (D')^{\nearrow}. \tag{1}$$

We denote with a $\searrow$ symbol all similarly defined *descending* quantities. Note that $D^{\searrow} = w_0(D^{\nearrow})$ where the permutation $w_0 = (n, n-1, \dots, 1)$ is the largest element of S_n in the Bruhat order.

We further define $D + \ell = (d_1 + \ell, \dots, d_n + \ell)$, $D_0 = D - \min D \in \mathbb{N}^n$ and $D^0 = \max D - D \in \mathbb{N}^n$. Let $\operatorname{Tr} D = \sum_{i=1}^n d_i$ and $\Delta D = \max D - \min D$, and finally

$$\|D\| = \sum_{j=1}^{n} \left(d_j - \frac{\operatorname{Tr} D}{n} \right)^2 \quad \text{and} \quad i(D) = \sum_{j=1}^{n} (\max D - d_j) \in \mathbb{N}.$$

We list some useful and obvious properties in the following lemma.

Lemma 4. *For $D \in \mathbb{Z}^n$ and $\ell \in \mathbb{Z}$, we have*

(i) $\Delta(D + \ell) = \Delta D$ *and* $i(D + \ell) = i(D)$,

(ii) $\Delta(-D) = \Delta D$ *and* $i(-D) = i(D^0) = \sum_{j=1}^n (d_j - \min D)$,

(iii) $i(D) + i(D^0) = n \Delta D$.

An $\boldsymbol{F}$*-admissible sequence* is an integer sequence whose ascending flag is equal to the standard ascendent flag of signature $\sigma(\boldsymbol{F}) = (n_1, \dots, n_s)$; in more concrete terms, an integer sequence

$$D = (\underbrace{d_1, \dots, d_1}_{n_1 \text{ times}}, \underbrace{d_2, \dots, d_2}_{n_2 \text{ times}}, \dots, \underbrace{d_s, \dots, d_s}_{n_s \text{ times}}) \quad \text{with } d_1 < \cdots < d_s.$$

Let $\mathbb{Z}^n(\boldsymbol{F})$ be the set of integer $\boldsymbol{F}$-admissible sequences, and let

$$\Xi(V) = \{(\boldsymbol{F}, D) \mid \boldsymbol{F} \in \mathfrak{Fl}(V) \text{ and } D \in \mathbb{Z}^n(\boldsymbol{F})\}$$

be the set of $\boldsymbol{F}$*-filtrations* of V.

2.2. ***Lattices.*** In the remainder of the section, we fix a point $x \in X$ and a coset $\mathbf{\Lambda} \in H/\sim_x$. We drop the index x for simplicity. The field $K = \widehat{(\mathcal{M}_X)_x}$ is local, and endowed with the discrete valuation $v = \mathrm{ord}_x$, whose valuation ring and maximal ideal we denote by $\mathbb{O}$ and $\mathfrak{m}$. Let V be the K-vector space $\mathcal{V}_x \otimes_{(\mathcal{M}_X)_x} K$ of dimension n.

Let $\mathcal{L}(u)$ denote the free $\mathbb{O}$-module spanned by a family (u) of vectors in V. An $\mathbb{O}$-module $M \in V$ is a lattice if there exists a K-basis (e) of V such that $M = \mathcal{L}(e)$. Let

$$v_\Lambda(x) = \max\{k \in \mathbb{Z} \mid x \in \mathfrak{m}^k \Lambda\}$$

be the natural valuation of V induced by Λ. For any lattices $M \subset \Lambda$ in V, we define the *interval* $[M, \Lambda]$ as

$$[M, \Lambda] = \{N \in \mathbf{\Lambda} \mid M \subset N \subset \Lambda\}.$$

Let $\pi_\Lambda : \Lambda \to \bar{\Lambda} = \Lambda/\mathfrak{m}\Lambda \simeq \mathbb{C}^n$ denote the canonical surjection on the quotient module.

Elementary divisors. Let z be a uniformizing parameter of K. For any two lattices Λ and M in V, there exists a unique increasing sequence of integers $d_1 \leqslant \cdots \leqslant d_n$ (the *elementary divisors of M in Λ*) and an $\mathbb{O}$-basis $(e_1, \ldots, e_n)$ of Λ such that $(z^{d_1} e_1, \ldots, z^{d_n} e_n)$ is a basis of M. Such a basis (e) is called a *Smith basis of Λ for M*. We will write them $d_i^\Lambda(M)$ if we want to specify the respective lattices, and we put

$$\mathbf{ED}_\Lambda(M) \;=\; \big(d_1^\Lambda(M), \ldots, d_n^\Lambda(M)\big).$$

Note that $d_1^\Lambda(M) = v_\Lambda(M)$ and $\mathbf{ED}_\Lambda(z^k M) = \mathbf{ED}_\Lambda(M) + k$. Let also

$$M_\Lambda = z^{-v_\Lambda(M)} M \quad \text{and} \quad \mu_\Lambda(M) = \dim_{\mathbb{C}} M_\Lambda/\mathfrak{m}\Lambda.$$

If $P \in \mathrm{GL}_n(K)$ is a basis change from Λ to M, the sequence $\mathbf{ED}_\Lambda(M) = (d_1, \ldots, d_n)$ can be computed in the following manner. For a subset $I \subset [n]$ of cardinality $|I|$, let $S_I = \sum_{i \in I} s_i$, and for $|I| = |J| = k$, let $P_{I,J}$ denote the (I, J)-submatrix of P. The sequence $(d_1, \ldots, d_n)$ satisfies

$$d_k = e_k - e_{k-1} \quad \text{where } e_k = \min_{|I|=|J|=k} \{v(\det P_{I,J})\} \text{ and } e_0 = 0. \tag{2}$$

It is convenient to be a bit more lax in the definition, and allow the elementary divisors to appear in another order. To avoid ambiguities, we will specify that (e) is an *ascending Smith basis* of Λ for M, if the vectors in (e) are ordered according to $\mathbf{ED}_\Lambda(M)^\nearrow$.

We say that a matrix P is *D-parabolic* if $P_{ij} \neq 0 \Rightarrow d_i \leqslant d_j$ for $1 \leqslant i, j \leqslant n$. For any commutative ring R, put $G_D(R)$ for the group of D-parabolic matrices of

$\mathrm{GL}_n(R)$. The subgroup of $\mathrm{GL}_n(\mathbb{C})$ that acts on the set of Smith bases of Λ for M is the subgroup of $\mathcal{G}_D$-*parabolic* matrices

$$\mathcal{G}_D = \{P \in \mathrm{GL}_n(\mathbb{C}) \mid v(P_{ij}) \geqslant d_i - d_j\}.$$

Being $\mathcal{G}_D$-parabolic is stronger than D-parahoric, which only stabilizes the induced D-flag in $\Lambda/\mathfrak{m}\Lambda$, and weaker than D-parabolic, which stabilizes a K-flag of signature $\sigma(D)$. For any sequence $D = (d_1, \ldots, d_n)$, let z^D be the diagonal matrix $\mathrm{diag}(z^{d_1}, \cdots, z^{d_n})$. We will frequently use the following type of diagram:

$$\begin{array}{ccc} \Lambda : (e) & \xrightarrow{z^D} & M : (\varepsilon) \\ \Big\downarrow P & & \Big\downarrow \tilde{P} \\ \Lambda : (e') & \xrightarrow{z^D} & M : (\varepsilon') \end{array}$$

which means that (e) and (e') are two Smith bases of Λ for M. Note that in this case $P \in \mathrm{GL}_n(\mathbb{O})$ is $\mathcal{G}_D$-parabolic. Since $G_D(\mathbb{C}) \subset \mathcal{G}_D$ holds, the obvious factorization $P = P_0 P_I$ of a lattice gauge $P \in \mathrm{GL}_n(\mathbb{O})$ into a constant term $P_0 = P(0) \in \mathrm{GL}_n(\mathbb{C})$ and a term $P_I = I + zU$ with $U \in \mathfrak{gl}_n(\mathbb{O})$ tangent to I satisfies the property

$$P \in \mathcal{G}_D \iff (P \in G_D(\mathbb{C}) \text{ and } P_I \in \mathcal{G}_D).$$

We can therefore usually assume that P is tangent to I. Note that this also holds for a right factorization $P = P_I' P_0$.

Sometimes, we will find it more convenient to consider the elementary divisors with their multiplicities. In this case, we will put $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_s$ for the distinct elementary divisors of M in Λ and let n_i be their respective multiplicities. The set $[n] = \{1, \ldots, n\}$ of indices of ordinary (*simple*) elementary divisors is partitioned into the subsets I_j corresponding to a single value of the elementary divisors:

$$I_j = \{1 \leqslant \ell \leqslant n \mid d_\ell = \boldsymbol{d}_j\} \quad \text{for } 1 \leqslant j \leqslant s.$$

2.3. ***Relative flag of a lattice.*** Any lattice M induces a natural flag in $\overline{\Lambda} = \Lambda/\mathfrak{m}\Lambda$. For any $k \in \mathbb{Z}$, let

$$M_k = (\mathfrak{m}^{-k} M \cap \Lambda) + \mathfrak{m}\Lambda \in [\mathfrak{m}\Lambda, \Lambda].$$

Lemma 5. *Let Λ, M be lattices in V. Let $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_s$ be the distinct elementary divisors of M in Λ. The flag in $\overline{\Lambda}$*

$$\boldsymbol{F}^\Lambda(M) : 0 \subset \pi_\Lambda(M_{\boldsymbol{d}_1}) \subset \cdots \subset \pi_\Lambda(M_{\boldsymbol{d}_s}) \subset \overline{\Lambda}$$

has signature $\sigma(\mathbf{ED}_\Lambda(M))$.

Proof. Let (e) be a basis of elementary divisors of Λ for M, and $I = \{1 \leqslant i \leqslant n \mid d_i \leqslant k\}$. Then M_k admits (u) as basis where $u_i = e_i$ if $i \in I$ and $u_i = ze_i$ if $i \notin I$.

The spaces M_k are thus embedded lattices, all belonging to the interval $[\mathfrak{m}\Lambda, \Lambda]$, so they take at most $n+1$ different values. Their images $\overline{M}_k = \pi_\Lambda(M_k)$ in the quotient space $\overline{\Lambda}$ form a flag $\boldsymbol{F}^\Lambda(M)$, and it is clear that $\overline{M}_{k-1} \subsetneq \overline{M}_k$ if and only if k is an elementary divisor of M. If $\boldsymbol{d}_1, \dots, \boldsymbol{d}_s$ are the distinct elementary divisors of M in Λ, with multiplicities n_i, the subset of indices corresponding to $\boldsymbol{d}_j$ can be written as

$$I_j = [n_1 + \cdots + n_j, n_1 + \cdots + n_{j+1} - 1].$$

The lattices M_k, M_ℓ coincide if and only if there exists i such that $\boldsymbol{d}_i \leqslant k, \ell < \boldsymbol{d}_{i+1}$ (with the conventions $\boldsymbol{d}_0 = -\infty$ and $\boldsymbol{d}_{s+1} = +\infty$). Therefore the flag $\boldsymbol{F}^\Lambda(M)$ has exactly length s, and its signature is equal to the sequence $(n_1, \dots, n_s)$. □

The components of $\boldsymbol{F}^\Lambda(M)$ can be indexed either as $\overline{M}_{\boldsymbol{d}_i}$, by the value of the elementary divisor $\boldsymbol{d}_i$ to which it is attached (if known), or as $\overline{M}_i$, by its index in the flag (here i). In this latter case, we will also use the notation $F_i^\Lambda(M)$. It will hopefully be always clear which convention we are using.

Lemma 6. *Let Λ, M, N be lattices in V. Let $D = \mathbf{ED}_\Lambda(M)^\nearrow$ and $D' = \mathbf{ED}_\Lambda(N)^\nearrow$. If either*

(i) *there exists a common Smith basis for Λ, M and N, or*

(ii) *the flags $\boldsymbol{F}^\Lambda(N)$ and $\boldsymbol{F}^\Lambda(M)$ are transverse,*

then we have

$$\mathbf{ED}_M(N) \cong D' - \sigma D,$$

where $\sigma = \rho(\boldsymbol{F}^\Lambda(N), \boldsymbol{F}^\Lambda(M))$ is the relative position of the induced flags in $\Lambda/\mathfrak{m}\Lambda$.

A similar formula holds for the descending sequences $D^\searrow$, but with $w_0 \sigma w_0$ instead of σ.

Proof. We summarize the setting by means of the following scheme:

$$\begin{array}{ccc} \Lambda : (e) & \xrightarrow{z^{S^\nearrow}} & M : (\varepsilon) \\ \Big\downarrow P & & \Big\downarrow \tilde{P} \\ \Lambda : (e') & \xrightarrow{z^{T^\nearrow}} & M : (\varepsilon') \end{array}$$

If there is an apartment containing Λ, M, N, then one can assume that P is a permutation matrix, namely $\Pi_{\sigma^{-1}}$. In the second case, it is possible to choose $P = \Pi_{w_0}(I + zU)$ with $U \in \mathfrak{gl}(\mathbb{O})$. Let $Q = I + zU$. Then we have $\tilde{P}_{ij} = Q_{ij} z^{t_j - s_{n+1-i}}$. According to formula (2), we have $d_1 = \min_{1 \leqslant i,j \leqslant n}(v(P_{ij}) + t_j - s_{n+1-i})$. The minimum of $t_j - s_{n+1-i}$ is attained for $(i,j) = (1,1)$, and by assumption we have $v(P_{11}) = 0$. Therefore, we get $d_1 = t_1 - s_n$. Let us prove by induction that $d_i = t_i - s_{n+1-i}$ for all i. Assume that $e_i = T_{\{1,\dots,i\}} - S_{\{n+1-i,\dots,n\}}$ for $i \leqslant k-1$.

Formula (2) yields $e_k = \min_{|I|=|J|=k}\{v(\det P_{I,J}) + T_J - (w_0 S)_I\}$. The minimum for $T_J - (w_0 S)_I$ is attained for $I = J = \{1, \dots, k\}$, while by assumption the principal minor $P_{I,I}$ has valuation 0. Hence $e_k = T_{[k]} - (w_0 S)_{[k]}$ and thus $d_k = e_k - e_{k-1} = t_k - s_{n+1-k}$ for all k. □

For a flag $\boldsymbol{F} = (\boldsymbol{F}_i)_{1 \leqslant i \leqslant s}$ of K-vector spaces, and a lattice Λ, let $\boldsymbol{F}_\Lambda = (\boldsymbol{F} \cap \Lambda)/\mathfrak{m}\Lambda$ be the flag $((\boldsymbol{F}_i \cap \Lambda)/\mathfrak{m}\Lambda)_{1 \leqslant i \leqslant s}$ induced in $\Lambda/\mathfrak{m}\Lambda$. The following result is easily established.

Lemma 7. *Let $\boldsymbol{F}$ and $\boldsymbol{F}'$ be two K-flags of V. For any basis (e), adapted for $\boldsymbol{F}$ and $\boldsymbol{F}'$, one has*

$$\rho(\boldsymbol{F}, \boldsymbol{F}') = \rho(\boldsymbol{F}_\Lambda, \boldsymbol{F}'_\Lambda), \quad \text{where } \Lambda = \mathcal{L}(e).$$

2.4. ***The affine building of* SL(V).** For this section, good references are [Garrett 1997; Ronan 1989], and especially [Abramenko and Brown 2008]. The affine building B_n naturally attached to $\mathrm{SL}(V)$ is the following $(n-1)$-dimensional simplicial complex. Two lattices Λ and M are *homothetic* if there exists $\alpha \in K^*$ such that $M = \alpha\Lambda$. Let $[\Lambda]$ be the homothety class of the lattice Λ in V. Two classes L and L' are *adjacent* if and only if there exist representatives Λ of L and M of L' such that $\mathfrak{m}\Lambda \subset M \subset \Lambda$. Consider a graph whose vertices are the homothety classes of lattices in V, and edges connect all adjacent vertices. The affine building B_n is the *flag simplicial complex* associated with this graph, or in other terms, its *clique complex*. A simplex A is a set $\{L_1, \dots, L_k\}$ of mutually adjacent vertices, and the *face relation* $B \leqslant A$ is defined by the inclusion $B \subset A$ of these sets.

Lemma 8. *Let L, L', L'' be vertices in B_n. If L' and L'' are adjacent, then for any representatives $\Lambda \in L$, $M \in L'$ and $N \in L''$, the flags $\boldsymbol{F}^\Lambda(M)$ and $\boldsymbol{F}^\Lambda(N)$ are compatible (in the sense that their components are all pairwise comparable). In particular, a maximal simplex C induces a complete flag in $\Lambda/\mathfrak{m}\Lambda$.*

Proof. One can find representatives $\Lambda \in L$, $M \in L'$ and $N \in L''$, and a basis (e) of Λ such that $M = \mathcal{L}(z^D e)$ and $N = \mathcal{L}(z^{D'} e)$, and that $D' - D \in \{0, 1\}^n$. Indeed, let (u) be a basis of $M/\mathfrak{m}M$ which is adapted to both flags $\boldsymbol{F}^M(\Lambda)$ and $\boldsymbol{F}^M(N)$. There exists a lifting (ε) of (u) in M which is a Smith basis of M for Λ. Since M and N are adjacent, *any* lifting of (u) is a Smith basis for N. Then $(e) = (z^{\mathbf{ED}_M(\Lambda)}\varepsilon)$ is a common Smith basis of Λ for M and N. By definition of adjacency, one can ensure that the representatives satisfy $\mathfrak{m}M \subset N \subset M$; that is, $D' - D \in \{0, 1\}^n$. Let now $I_k = \{1 \leqslant i \leqslant n \mid d_i \leqslant k\}$ and $I'_k = \{1 \leqslant i \leqslant n \mid d'_i \leqslant k\}$. Then we have $I_{k-1} \subset I'_k \subset I_k$ for all i. Since $\boldsymbol{F}^\Lambda(M)$ is spanned by the subfamilies $(\overline{e_i})_{i \in I_k}$ for $k \in \mathbb{Z}$, the claim follows. □

Simplices and chambers. A maximal simplex, or *chamber* in B_n, is an n-chain C of vertices $L_0, \dots, L_{n-1}$ with representatives Λ_i for $0 \leqslant i \leqslant n-1$ satisfying

$$\mathfrak{m}\Lambda_0 \subset \Lambda_1 \subset \cdots \subset \Lambda_{n-1} \subset \Lambda_0.$$

A basis (e) is a *fundamental basis for C at L_i* if

$$L_{(i+j) \bmod n} = [\mathscr{L}(e_1, \dots, e_j, ze_{j+1}, \dots, ze_n)] \quad \text{for } 0 \leqslant j \leqslant n-1.$$

Lemma 9. *The set of chambers which contain a given vertex L is in bijection with the set of complete flags in $\Lambda/\mathfrak{m}\Lambda$ with $\Lambda \in L$. A basis (e) is fundamental for C at L if and only if its image in $\Lambda/\mathfrak{m}\Lambda$ is strictly adapted to the flag $\boldsymbol{F}^{\Lambda}(C)$.*

A (partial) flag $\boldsymbol{F}$ in $\Lambda/\mathfrak{m}\Lambda$ can be lifted (by $\pi_{[\Lambda]}^{-1}$) to a uniquely defined simplex in B_n containing the vertex $L = [\Lambda]$.

Definition 10. The *graph-theoretic distance*, *canonical metric* and *index* on B_n are defined, respectively, by

$$d(L, L') = \Delta\big(\mathbf{ED}_{\Lambda}(\Lambda')\big), \quad \boldsymbol{d}(L, L') = \big\|\mathbf{ED}_{\Lambda}(\Lambda')\big\|, \quad [L : L'] = i\big(\mathbf{ED}_{\Lambda}(\Lambda')^0\big)$$

for any representatives Λ, Λ' of L, L'.

Note that $d(L, L') = -v_{\Lambda}(\Lambda') - v_{\Lambda'}(\Lambda)$ also holds, and that these three maps are indeed symmetric. The $\boldsymbol{d}$ metric makes the geometric realization of B_n into a CAT(0) space [Abramenko and Brown 2008, Theorem 11.16, p. 555].

Apartments. Let $\Phi = \{d_1, \dots, d_n\}$ be a K-frame of V. The set

$$\mathscr{A}_{\Phi} = \{\Lambda = \ell_1 + \cdots + \ell_n \,|\, \ell_i \text{ is a lattice in } d_i\}$$

of lattices spanned over multiples of the vectors in Φ induces a simplicial subcomplex in the affine building B_n called the *apartment* spanned by Φ. For any lattice $\Lambda \in \boldsymbol{\Lambda}$, a Λ-*basis* of the apartment $\mathscr{A}_{\Phi}$ is a collection $(e) = (e_1, \dots, e_n)$ of vectors such that e_i spans d_i and $v_{\Lambda}(e_i) = 0$. Such a family is unique up to permutation and to multiplication of each e_i by a scalar $\lambda_i \in \mathbb{O}^*$. The lattice is an element of the apartment $\mathscr{A}_{\Phi}$ if and only if the family $(e) = (e_1, \dots, e_n)$ is actually a basis of the lattice Λ. Equivalently, and without reference to a basis, this means that

$$\Lambda = \bigoplus_{i=1}^{n} \Lambda \cap d_i.$$

In the general case, the lattice $\Lambda_{\Phi} = \bigoplus_{i=1}^{n} \Lambda \cap d_i$ is the *largest sublattice* of Λ in the apartment $\mathscr{A}_{\Phi}$. The homothety class $L_{\Phi} = [\Lambda_{\Phi}]$ is therefore the closest point projection of $L = [\Lambda]$ on $\mathscr{A}_{\Phi}$, and the map

$$\begin{aligned} \rho_{\Phi} : B_n &\to \mathscr{A}_{\Phi} \\ \Lambda &\mapsto \Lambda_{\Phi} \end{aligned}$$

is the *retract* onto $\mathscr{A}_\Phi$.

Galleries and type. Two chambers C and C' are *adjacent* if they share $n-1$ vertices in common (a so-called *panel*). A *gallery* Γ from C to C' of length $\ell(\Gamma) = \ell$ is a sequence of chambers $\Gamma = (C_0 = C, C_1, \ldots, C_\ell = C')$ such that C_i is adjacent to C_{i+1}. The *gallery distance* is defined as

$$\ell(C, D) = \min\{\ell(\Gamma) \mid \Gamma \text{ is a gallery from } C \text{ to } D\}.$$

The vertices in the building B_n can be labelled by $\mathbb{Z}/n\mathbb{Z}$, which we assume given from now on. Let $S = \{s_0, \ldots, s_{n-1}\}$. The affine Weyl group W of the building then has a presentation

$$W = \langle S \mid s_i^2 = 1 \text{ and } (s_i s_{i+1})^3 = 1\rangle,$$

where the indices are understood modulo n. Every chamber has exactly one vertex labelled i, that we denote $M_i(C)$, for every $i \in \mathbb{Z}/n\mathbb{Z}$. Two chambers are called *i-adjacent* when their common panel precisely does not contain those vertices labelled i. Two i-adjacent chambers C and C' are then said to have *W-distance* $\delta(C, C') = s_i$. If $\Gamma = (C, C_1, \ldots, C_{t-1}, C')$ is any gallery from C to C' lying inside an apartment, the W-distance $\delta(C, C')$ is defined as the product

$$\delta(C, C') = \delta(C, C_1) \cdots \delta(C_{t-1}, C').$$

There is a bijection between minimal galleries between C and C' and minimal length decompositions of the W-distance $\delta(C, C') \in W$ into products of generators in S. Let $\mathbb{Z}_0^n$ be the set of n-tuples of integers summing to 0. A labelling of the vertices by $\mathbb{Z}/n\mathbb{Z}$ induces an isomorphism

$$W \overset{\phi}{\simeq} S_n \rtimes \mathbb{Z}_0^n \quad \text{given by} \quad \begin{cases} \phi(s_i) = ((i, i+1), 0) & \text{for } 1 \leqslant i \leqslant n, \\ \phi(s_0) = ((1, n), (1, 0, \ldots, 0, -1)). \end{cases}$$

Let $\delta = \delta(C, C')$ be the W-distance between two chambers C and C'. Let $L = M_0(C)$ and $L' = M_0(C')$ be the respective unique vertices of type 0 of C and C'. Then $\phi(\delta) = (\sigma, K)$ is the unique couple such that there exists a fundamental basis (e) of C at L for which $(z^K e_\sigma) = (z^{k_1} e_{\sigma(1)}, \ldots, z^{k_n} e_{\sigma(n)})$ is a fundamental basis of C' at L'.

Walls. Let (e) be a basis of V, and let $\mathscr{A}$ be the apartment spanned by (e). In the basis (e), any lattice L in $\mathscr{A}$ can be represented by a unique (up to an integer multiple of $(1, \ldots, 1)$) n-tuple $(x_1(L), \ldots, x_n(L)) \in \mathbb{Z}^n$. The set $\mathscr{H}$ of *walls* of V for W is the set of hyperplanes

$$H_{i,j}^{(k)} = \{x \in \mathbb{R}^n \mid x_i - x_j = k\} \quad \text{for } 1 \leqslant i < j \leqslant n \text{ and } k \in \mathbb{Z}. \tag{3}$$

Define the corresponding half-spaces

$$H_{i,j}^{(k)+} = \{x \in \mathbb{R}^n \mid x_i - x_j \geqslant k\},$$
$$H_{i,j}^{(k)-} = \{x \in \mathbb{R}^n \mid x_i - x_j \leqslant k\}.$$

For a hyperplane $H \in \mathscr{H}$ and a simplex A, let

$$\sigma_H(A) = \begin{cases} + & \text{if } A \subset H^+ \backslash H^-, \\ - & \text{if } A \subset H^- \backslash H^+, \\ 0 & \text{if } A \subset H^+ \cap H^-. \end{cases}$$

A wall H *separates* the simplices A and B if $\sigma_H(A)\sigma_H(B) = -$ for the usual sign product rule. Any simplex A is completely defined by its sign sequence $(\sigma_H(A))_{H\in\mathscr{H}}$. However, there is only a finite number of them which are relevant (i.e., whose defining equations or inequalities are not redundant). If we define the fundamental chamber as $C_0 = (L_0, \ldots, L_{n-1})$, where $L_i = [\mathscr{L}(e_1, \ldots, e_i, ze_{i+1}, \ldots, ze_n)]$, the (essential) walls of C_0 are the hyperplanes $H^{(0)}_{i,(i \bmod n)+1}$ for $1 \leqslant i \leqslant n$, and its sign sequence is

$$\sigma_{H^{(k)}_{i,(i \bmod n)+1}}(C_0) = \begin{cases} + & \text{for } 1 \leqslant i \leqslant n-1 \text{ and } k \leqslant 0 \text{ or } i = n \text{ and } k \geqslant 0, \\ - & \text{for } 1 \leqslant i \leqslant n-1 \text{ and } k \geqslant 0 \text{ or } i = n \text{ and } k \leqslant 0. \end{cases}$$

Definition 11. Let A, B be simplices in B_n. The gallery distance $\ell(A, B)$ is defined as $\min_{A\leqslant C, B\leqslant D} \ell(C, D)$.

Lemma 12. *For two simplices A, B we have*

$$\ell(A, B) = |\{H \in \mathscr{H} \mid \sigma_H(A)\sigma_H(B) = -\}|.$$

If $A = [\Lambda]$ and $B = [M]$ are vertices, we have

$$\ell(A, B) = \sum_{1\leqslant i<j\leqslant n} \max(0, d_j - d_i - 1), \quad \textit{where } (d_1, \ldots, d_n) = \mathbf{ED}_\Lambda(M)^\nearrow.$$

Proof. The first assertion is known (see [Abramenko and Brown 2008, p. 32]). Take a basis (e) of V where A has coordinates a and B coordinates b. The description (3) of walls shows that A and B are separated by $H_{i,j}^{(k)}$ if and only if $a_i - a_j \leqslant k - 1$ and $b_i - b_j \geqslant k + 1$. Taking for (e) a Smith basis of Λ for M gives $a = 0$ and $b = (d_1, \ldots, d_n)$. The result is then a straightforward count. □

2.5. ***Forms.*** For $L \in B_n$, let $\overline{L} = \Lambda/\mathfrak{m}\Lambda \simeq \mathbb{C}^n$ for $\Lambda \in L$. This definition is independent of the choice of Λ as there is a canonical isomorphism between $\Lambda/\mathfrak{m}\Lambda$ and $\mathfrak{m}^k\Lambda/\mathfrak{m}^{k+1}\Lambda$ for any k. For $L' \in B_n$, let N be the unique representative of L' such that $v_\Lambda(N) = 0$, and define

$$\pi_L(L') = (N + \mathfrak{m}\Lambda)/\mathfrak{m}\Lambda \in \overline{L}. \tag{4}$$

The map π_L induces an isomorphism of simplicial complexes between the link $\mathrm{lk}(L)$ of $L = [\Lambda]$ and the set E of chains of linear subspaces of $\overline{L}$. A *form* in a lattice Λ is

a $\mathbb{C}$-vector subspace Y of Λ spanned by an $\mathbb{O}$-basis (e) of Λ. For any $x \in \Lambda$, there is a unique representative x_Y of the coset $x + \mathfrak{m}\Lambda$ in Y. This induces a well-defined isomorphism

$$\phi_Y : Y \xrightarrow{\simeq} \Lambda/\mathfrak{m}\Lambda.$$

We will say that $Y \subset \Lambda$ is a *Smith form* for M if there is a basis of Y which is a Smith basis of Λ for M.

Lemma 13. *Let Λ be a lattice in V and Y be a form in Λ.*

(i) *For any basis (e) of the lattice Λ, there exists a unique $\mathbb{C}$-basis (e_Y) of the form Y whose image in $\Lambda/\mathfrak{m}\Lambda$ coincides with the image of (e). We call (e_Y) the Y-*basis of (e).

(ii) *Given a filtration $(\boldsymbol{F}, D)$ in $\Xi(\Lambda/\mathfrak{m}\Lambda)$, there exists a unique lattice $M = \mathscr{L}_Y(\boldsymbol{F}, D)$ such that $\boldsymbol{F}^\Lambda(M) = \boldsymbol{F}$ and $\mathbf{ED}_\Lambda(M) = D$ that admits a Smith basis in Y.*

(iii) *For any lattice M such that $d(\Lambda, M) = 1$, let $(\bar{e})$ a basis of $\Lambda/\mathfrak{m}\Lambda$ respecting $\pi_\Lambda(M) = M/\mathfrak{m}\Lambda$. Then the Y-basis (e_Y) is a Smith basis for M.*

Proof. The basis (ε) obtained by putting $\varepsilon_i = (e_i)_Y = \phi_Y^{-1}(\pi_\Lambda(e_i))$ obviously satisfies the conditions of (i). For any $\mathbb{C}$-basis (e) of Y which respects the flag $\boldsymbol{F}$, put $M = \bigoplus_{i=1}^n z^{d_i} e_i$. Let $(\tilde{e})$ be another basis of Y and $\tilde{M} = \bigoplus_{i=1}^n z^{d_i}\tilde{e}_i$. The matrix of the change of basis from $(z^D e)$ to $(z^D \tilde{e})$ is equal to $P = z^D C z^{-D}$, where $C \in \mathrm{GL}_n(\mathbb{C})$ is the matrix of the change of basis from (e) to $(\tilde{e})$. By definition of the parabolic subgroup $P_{\boldsymbol{F}}$, one has $z^D C z^{-D} \in \mathrm{GL}_n(\mathbb{O}) \iff C \in P_{\boldsymbol{F}}$; hence $M = \tilde{M}$ if and only if (e) and $(\tilde{e})$ both respect the flag $\boldsymbol{F}$. Note that the gauge from the basis (e) to its Y-basis is always of the form $P = I + zU \in \mathrm{GL}_n(\mathbb{O})$. Let $W = \pi_\Lambda(M)$ and let

$$T = \begin{pmatrix} 0_r & 0 \\ 0 & I_{n-r} \end{pmatrix}$$

be the elementary divisors of $M_\Lambda = z^{-v_\Lambda(M)} M$ with respect to Λ. Assume that (e) satisfies the assumptions of (iii). Then the Y-basis (e_Y) is obtained by a gauge $P = I + zU \in \mathrm{GL}_n(\mathbb{O})$. Putting

$$U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix},$$

we have

$$\begin{array}{ccc} Y : (e_Y) & \xrightarrow{t^T} & M_\Lambda \xrightarrow{z^{v_\Lambda(M)} I} M \\ {\scriptstyle I+tU}\uparrow & & \uparrow{\scriptstyle \tilde{U} \in \mathrm{GL}_n(\mathbb{O})} \\ \Lambda : (e) & \xrightarrow{z^T} & M_\Lambda \end{array}$$

since

$$\widetilde{U} = z^{-T}(I+zU)z^{T} = \begin{pmatrix} I_r + zU_{11} & z^2U_{12} \\ U_{21} & I_{n-r}+zU_{22} \end{pmatrix}.$$

The basis (e_Y) is therefore indeed a Smith basis of Λ for M. □

For any two forms Y and $\widetilde{Y}$ in Λ, the set of gauges between bases of Y and $\widetilde{Y}$ is an element of the double coset $\mathrm{GL}_n(\mathbb{C})\backslash\mathrm{GL}_n(\mathcal{O})/\mathrm{GL}_n(\mathbb{C})$. Let z be a uniformizing parameter. With the convention that $\deg_z P = \infty$ if $P \in \mathrm{GL}_n(\mathcal{O})\backslash\mathfrak{gl}(\mathbb{C}[z])$, the following definition makes sense.

Definition 14. Let Y, $\widetilde{Y}$ be two forms in Λ. Define the *z-distance* as

$$\delta_z(Y,\widetilde{Y}) = \min(\deg_z P, \deg_z P^{-1}) \in \mathbb{N}\cup\{\infty\}$$

for any gauge P from a basis of Y to a basis of $\widetilde{Y}$.

Lemma 15. *If $t = d(\Lambda, M)$, then for any form Y of Λ, and any uniformizing parameter z, there exists a Smith form $\widetilde{Y}$ for M at a z-distance $\delta_z(Y,\widetilde{Y}) \leqslant t-1$.*

Proof. There exists a Smith form Y' of Λ for M. Let $P = P_0 + P_1 z + \cdots \in \mathrm{GL}_n(\mathcal{O})$ be a gauge corresponding to a basis change from Y to Y'. Let $\overline{P} = P_0 + \cdots + P_k z^k$ for some $k \geqslant 0$, and let $\widetilde{Y}$ be the form obtained by this gauge transformation, as explained in the following scheme:

$$\begin{array}{ccccc} \widetilde{Y} & \xleftarrow{\overline{P}} & Y & \xrightarrow{P} & Y' \\ \downarrow{\scriptstyle z^D} & & & & \downarrow{\scriptstyle z^D} \\ M & & \xleftarrow[Q]{} & & M \end{array}$$

We have $Q = z^{-D}P^{-1}\overline{P}z^{D} = (\widetilde{P}_{ij}z^{d_j-d_i})$, where $\widetilde{P} = P^{-1}\overline{P}$. By construction, we have $\widetilde{P} = P^{-1}(P-(P-\overline{P})) = I + z^{k+1}U$ with $U \in \mathfrak{gl}(\mathcal{O})$. As soon as $k \geqslant t-1$, we have $Q \in \mathrm{GL}_n(\mathcal{O})$; hence the form $\widetilde{Y}$ is a Smith form for M. □

2.6. ***Shortest paths and elementary splittings.*** A *shortest path* is a path Γ in B_n such that for any vertices L, $L' \in \Gamma$, the length of the path between L and L' induced by Γ is equal to $d(L,L')$. For a path $\Gamma = (L_0,\ldots,L_t)$ in B_n, and a representative $\Lambda_0 \in L_0$, let the Λ_0*-normalized sequence* of Γ be the unique sequence of lattices $\Lambda_i \in L_i$ such that $v_{\Lambda_0}(\Lambda_i) = 0$.

Lemma 16. *Let Λ, $M \in \boldsymbol{\Lambda}$. The Λ-normalized sequence $(\Lambda, \Lambda_1, \ldots, \Lambda_t = M_\Lambda)$ of a shortest path from $[\Lambda]$ to $[M]$ satisfies $\Lambda \supset \Lambda_1 \supset \cdots \supset \Lambda_t \supset z^t\Lambda$.*

Proof. Let $\Gamma = (\Lambda = \Lambda_0, \Lambda_1, \ldots, \Lambda_{t-1}, M = \Lambda_t)$ be a path of minimal length t, and let us prove that $\Lambda \supset \Lambda_1 \supset \cdots \supset \Lambda_t$ by induction on t. For $t = 1$, this is the very definition of adjacency in B_n. Assuming that the claim is established for any pair of

lattices at distance $\leqslant t-1$, we have $M \subset \Lambda \supset \Lambda_1 \supset \cdots \supset \Lambda_{t-1}$ for the normalized sequence of Γ. Since Λ_{t-1} and M are adjacent, there exists a unique $k \in \mathbb{Z}$ such that $z^k\Lambda_{t-1} \supset M \supset z^{k+1}\Lambda_{t-1}$. We know that $d(\Lambda, \Lambda_{t-1}) = t-1$, and hence we have $\Lambda \supset \Lambda_{t-1} \supset z^{t-1}\Lambda$; therefore we get

$$z^k\Lambda \supset z^k\Lambda_{t-1} \supset M \supset z^{k+1}\Lambda_{t-1} \supset z^{t+k}\Lambda.$$

If $k > 0$, then $v_\Lambda(M) \geqslant k > 0$, which was excluded by assumption. But if $k < 0$, then $d(\Lambda, M) < t$, which is also excluded. Thus we have $k = 0$, and the claim is proved. □

The following result explains how to construct some shortest paths algebraically.

Proposition 17. *Let L, $L' \in B_n$, and let $t = d(L, L')$. For $k \in \mathbb{N}$, let $L_k = [\Lambda' + \mathfrak{m}^k\Lambda]$, and $M_k = [\Lambda \cap \mathfrak{m}^{k-t}\Lambda']$, where $\Lambda \in L$ and $\Lambda' \in L'$ are such that $v_\Lambda(\Lambda') = 0$. Then $d(L_k, L_{k+1}) = 1$ for $0 \leqslant k \leqslant t-1$ and $L_t = L'$. The paths*

$$\Gamma_{\min}(L, L') = (L_0, L_1, \ldots, L_t) \quad \text{and} \quad \Gamma_{\max}(L, L') = (M_0, M_1, \ldots, M_t)$$

are shortest paths from L to L', respectively called the min- *and the* max-shortest path *from L to L'. For any Λ-normalized sequence $(N_0 = \Lambda, N_1, \ldots, N_t = M)$ of a shortest path Γ' from L to L', one has*

$$L_i \subset N_i \subset M_i \quad \text{for } 0 \leqslant i \leqslant t.$$

Proof. The existential part of the lemma is easy to verify by using Smith bases of the representatives $\Lambda \in L$ and $M \in L'$, and is left to the reader. First note that the min- and max-shortest paths are in correspondence under the duality map $\Lambda \mapsto \Lambda^* = \hom_K(\Lambda, \mathbb{O})$, so we only need to prove claims on one type of path. Note that the shortest path interval is symmetric. Indeed, letting $\Gamma(L', L) = (L'_0, L'_1, \ldots, L'_t)$, we have $L_k = [M_\Lambda + z^k\Lambda]$ and $L'_{t-k} = [\Lambda_M + z^{t-k}M]$. By definition we have

$$\Lambda_M + z^{t-k}M = z^{-v_M(\Lambda)}\Lambda + z^{-v_M(\Lambda)-v_\Lambda(M)-k}M = z^{-v_M(\Lambda)-k}\big(z^k\Lambda + z^{-v_\Lambda(M)}M\big).$$

Therefore $L'_{t-k} = L_k$.

Now we turn to the proof of $\Lambda_i = M + \mathfrak{m}^i\Lambda \subset N_i \subset \Lambda \cap \mathfrak{m}^{i-t}M = M_i$ for the shortest path $\Gamma' = (N_0, \ldots, N_t)$, which we will prove by induction on the distance $t = d(L, L')$. For convenience, let any path $([\Lambda] = L_0, L_1, \ldots, L_{t-1}, [M] = L_t)$ be represented by its Λ-normalized sequence $(\Lambda, \Lambda_1, \ldots, \Lambda_{t-1}, \Lambda_t = M)$ of lattices $\Lambda_i \in L_i$ such that $v_\Lambda(\Lambda_i) = 0$. The claim is obvious for $t = 1$. Let us assume it holds for any shortest path between pairs of vertices in B_n whose distance is at most $t-1$. Suppose then that $d(\Lambda, M) = t$. Let $\Lambda = \Lambda_0 \supset \Lambda_1 \supset \cdots \supset \Lambda_{t-1} \supset M$ be the min-shortest path from Λ to M and let $\Lambda \supset \Lambda'_1 \supset \cdots \supset \Lambda'_{t-1} \supset M$ represent the path Γ of minimal length. By assumption, $d(\Lambda, \Lambda'_{t-1}) = t-1$, therefore

we have $\Lambda \supset \Lambda'_{t-1} \supset z^{t-1}\Lambda$, and by definition, we have $\Lambda_{t-1} = M + z^{t-1}\Lambda$. Therefore, we get

$$\begin{aligned}\Lambda_{t-1} \cap \Lambda'_{t-1} &= (M + z^{t-1}\Lambda) \cap \Lambda'_{t-1} \\ &= M + (z^{t-1}\Lambda \cap \Lambda'_{t-1}) \quad \text{(because } M \subset \Lambda'_{t-1}\text{)} \\ &= M + z^{t-1}\Lambda \quad \text{(since } z^{t-1}\Lambda \subset \Lambda'_{t-1}\text{)} \\ &= \Lambda_{t-1}.\end{aligned}$$

Thus $\Lambda_{t-1} \subset \Lambda'_{t-1}$ holds. On the other hand, $M_{t-1} = \Lambda \cap \mathfrak{m}^{-1} M$ is by construction the largest lattice containing M, contained in Λ and adjacent to M. Therefore $\Lambda_{t-1} \subset \Lambda'_{t-1} \subset M_{t-1}$. By the induction assumption, the claim is established. □

If $D = (d_1, \dots, d_n)$ are the elementary divisors of M_Λ in Λ, then the lattices Λ_k on the (normalized) min-shortest path from Λ to M have elementary divisors $D_k = (\min(d_1, k), \dots, \min(d_n, k))$ in Λ. The differences $\Delta_k = D_k - D_{k-1}$ are the intermediate elementary divisors as explained in the following:

$$\Lambda \xrightarrow{z^{\Delta_1}} \Lambda_1 \xrightarrow{z^{\Delta_2}} \Lambda_2 \longrightarrow \cdots \xrightarrow{z^{\Delta_{t-1}}} \Lambda_{t-1} \xrightarrow{z^{\Delta_t}} M_\Lambda \xrightarrow{z^{v_\Lambda(M)}} M. \tag{5}$$

Let for simplicity $I = (1, \dots, 1)$ and $\ell I = (\ell, \dots, \ell)$.

Definition 18. For $D = (d_1, \dots, d_n) \in \mathbb{Z}^n$, let $v = \min D$, let $t = \max D$ and

$$D_i = \big(\min(d_1 - v, i), \dots, \min(d_n - v, i)\big) \quad \text{for } 1 \leqslant i \leqslant t.$$

The *elementary splitting* of D is the unique decomposition $D = \Delta_0 + \Delta_1 + \cdots + \Delta_t$ where $\Delta_0 = vI \in \mathbb{Z}^n$ and $\Delta_k = D_k - D_{k-1} \in \{0, 1\}^n$ for $1 \leqslant k \leqslant t$.

We write $\Delta_i(D)$ if we need to specify the sequence D. The next straightforward lemma is left to the reader.

Lemma 19. *Let* $\mathbf{ED}_\Lambda(M) = \Delta_0 + \Delta_1 + \cdots + \Delta_t$ *be the elementary splitting of* $\mathbf{ED}_\Lambda(M)$ *for* Λ, $M \in \mathbf{\Lambda}$. *Then we have* $\Delta_0 = v_\Lambda(M)I$ *and the following statements hold*:

(i) $\delta_i = \operatorname{Tr} \Delta_i = \mu_M(\Lambda_i)$, *where* $\Lambda_i \in \Gamma_{\min}(\Lambda, M)$ *as in* (5).

(ii) $\Delta_i(\mathbf{ED}_M(\Lambda)) = I - \Delta_{d-i+1}(\mathbf{ED}_\Lambda(M))$ *for* $1 \leqslant i \leqslant t$. *In other terms*,

$$\mathbf{ED}_M(\Lambda) = (-t - v_\Lambda(M))I + (I - \Delta_t) + \cdots + (I - \Delta_1)$$

is the elementary splitting of $\mathbf{ED}_M(\Lambda)$.

(iii) *Let* $A = [\Lambda]$ *and* $B = [M]$. *The gallery distance* $\ell(A, B)$ *is given by*

$$\ell(A, B) = \sum_{i=1}^{t-1} (n - \delta_i)\delta_{i+1}.$$

Note that the sequence $\delta_1, \dots, \delta_t$ is nonincreasing, and that we also have

$$\delta_i = |\{1 \leqslant j \leqslant n \mid d_j \geqslant i + v_\Lambda(M)\}|.$$

3. Birkhoff–Grothendieck trivializations

The central result in the theory of holomorphic vector bundles on $\mathbb{P}^1(\mathbb{C})$ is the Birkhoff–Grothendieck theorem, which states that any such bundle is isomorphic to a direct sum of line bundles. In this section, we take $X = \mathbb{P}^1(\mathbb{C})$, and we investigate what properties of the vector bundle can be retrieved by considering only the Bruhat–Tits building at a point $x \in X$.

3.1. ***The Birkhoff–Grothendieck property.*** According to Lemma 3, a holomorphic vector bundle $\mathscr{E} \in \mathscr{H}$ is completely described by the coset $\mathbf{\Lambda} = [\mathscr{E}] \in \mathscr{H}/\sim_x$ and the lattice $\Lambda = \mathscr{E}_x \in \mathbf{\Lambda}$. Let us take up the notations of Section 2 again. Let V denote the formalized meromorphic stalk $\widehat{\mathscr{V}_x}$ and let B be the corresponding Bruhat–Tits building. Let $B_0 = \{[M] \mid \mathscr{E}^M \text{ is trivial}\}$ be the subset of *trivializing lattices* of B.

Note 20. Strictly speaking, $L \in B_0$ contains exactly one lattice $M \in L$ such that the extension $\mathscr{E}^M$ is trivial. Any other $\tilde{M} \in L$ gives a *quasitrivial* vector bundle $\mathscr{E}^{\tilde{M}}$. There is no real need to make the difference, since one gets a trivial bundle by simply tensoring with a line bundle.

Lemma 21. *$B_0 \neq \varnothing$ for any $x \in X$ and $\mathbf{\Lambda} \in \mathscr{H}/\sim_x$.*

The existence of an analytic trivialization outside of an arbitrary point is a very general result on holomorphic bundles over complex analytic manifolds. This much more restrictive *algebraicity* statement follows easily from the Birkhoff–Grothendieck theorem.

Proof. Let $\mathscr{E} \in \mathbf{\Lambda}$ be a representative of the coset $\mathbf{\Lambda}$. Let $X^* = X \backslash \{x\}$, and Δ a neighborhood of x. The Birkhoff–Grothendieck theorem implies that there exists a sequence of integers $(a_1, \dots, a_n) \in \mathbb{Z}^n$, and linearly independent sections $\sigma_1, \dots, \sigma_n \in \Gamma(X^*, \mathscr{E})$ and $\widetilde{\sigma}_1, \dots, \widetilde{\sigma}_n \in \Gamma(\Delta, \mathscr{E})$ such that $\widetilde{\sigma}_i = z^{-a_i}\sigma_i$ for a local coordinate z at x. The section σ_i admits therefore a *meromorphic* extension at x, that we still denote σ_i. The trivial bundle $\mathscr{F} \in H_0$ spanned by $(\sigma_1 \ \dots, \sigma_n)$ is a trivialization of $\mathscr{E}$ at x as required. □

We say that any such bundle $\mathscr{F}$ is a *Birkhoff–Grothendieck trivialization of $\mathscr{E}$ at x*. We'll often write BG for short. Let $\mathrm{BG}_x(\mathscr{E}) \subset \mathscr{H}_0$ for the set of BG trivializations of $\mathscr{E}$ at x. Note that any $\mathscr{F} \in \mathrm{BG}_x(\mathscr{E})$ corresponds uniquely to a decomposition of $\mathscr{E}$ as a direct sum of line subbundles (a *BG decomposition*)

$$\mathscr{E} = \bigoplus_{i=1}^{n} \mathscr{L}_i. \tag{6}$$

Note 22. A line bundle on P is characterized by its degree. Recall that, for any decomposition (6), if the integers $a_i = \deg \mathscr{L}_i$ satisfy $a_1 \geqslant \cdots \geqslant a_n$, then the sequence $T(\mathscr{E}) = (a_1, \ldots, a_n)$ is unique and called the *type* of $\mathscr{E}$.

In the case where the coset $[\mathscr{E}] = \mathbf{\Lambda} \in \mathscr{H}/\sim_x$ is fixed, we write

$$\mathrm{BG}(\Lambda) = \{M \in \mathbf{\Lambda}^0 \mid \mathscr{E}^M \in \mathrm{BG}_x(\mathscr{E})\},$$

and $T(\Lambda)$ instead of $T(\mathscr{E})$. The global sections $\mathscr{F}(X)$ of a trivial bundle $\mathscr{F} \in [\mathscr{E}]$ induce, by taking stalks at x, a form Y_M in the corresponding lattice $M = \mathscr{F}_x \in \mathbf{\Lambda}$, that we call the *global form* of M. The link between the Birkhoff–Grothendieck theorem and the algebraic structure of the local lattices is then given by the following straightforward characterization.

Lemma 23. *The trivial lattice M is in* $\mathrm{BG}(\Lambda)$ *if and only if there exists a Smith basis (e) of M with respect to Λ that is simultaneously a $\mathbb{C}$-basis of the global form Y_M of M.*

In this case, the basis (e) will be called a *BG basis* for Λ. We state now the following result separately for further reference.

Proposition 24. *Let $\mathscr{E} \in \mathscr{H}$ be a holomorphic vector bundle. For any BG trivialization $\mathscr{F}$ of $\mathscr{E}$ at any $x \in \mathbb{P}^1(\mathbb{C})$, the type $T(\mathscr{E})$ of the bundle $\mathscr{E}$ is equal to the sequence of elementary divisors $\mathbf{ED}_\Lambda(M)$ of the stalk $M = \mathscr{F}_x$ with respect to $\Lambda = \mathscr{E}_x$ (viewed as lattices in $\mathscr{V}_x$).*

The BG trivializations of a bundle $\mathscr{E}$ are as a rule not unique, nor is any trivialization of $\mathscr{E}$ necessarily BG. One of the aims of this section is to prove the following local characterization of BG trivializations. Recall that the Bruhat–Tits building is endowed with three distance maps: the graph-theoretical distance d, the canonical metric $\boldsymbol{d}$ and the gallery distance ℓ (which is actually a pseudodistance on the vertices).

Theorem 25. *Let $\Lambda = \mathscr{E}_x$. Let $\preceq_{\mathrm{lex}}$ be the lexicographic, and $\preceq_{\mathrm{lexrev}}$ the reverse lexicographic orderings on $\mathbb{Z}^n$. Then we have*:

(i) $T(\mathscr{E}) \cong \min_{\preceq_{\mathrm{lex}}}\{\mathbf{ED}_\Lambda(M)^\searrow \mid M \in \mathbf{\Lambda}^0\} = \max_{\preceq_{\mathrm{lexrev}}}\{\mathbf{ED}_\Lambda(M)^\searrow \mid M \in \mathbf{\Lambda}^0\}$.

(ii) *If $M \in \mathbf{\Lambda}^0$ then the following conditions are equivalent*:

- (a) $M \in \mathrm{BG}(\Lambda)$;
- (b) $\mathbf{ED}_\Lambda(M) \cong T(\mathscr{E})$;
- (c) $\boldsymbol{d}(\Lambda, M) = \min_{\tilde{M} \in \mathbf{\Lambda}^0} \boldsymbol{d}(\Lambda, \tilde{M})$;
- (d) $\ell(\Lambda, M) = \min_{\tilde{M} \in \mathbf{\Lambda}^0} \ell(\Lambda, \tilde{M})$.

(iii) *Moreover, if $M \in \mathrm{BG}(\Lambda)$ then $d(\Lambda, M) = \min_{\tilde{M} \in \mathbf{\Lambda}^0} d(\Lambda, \tilde{M})$.*

In other terms, in the set of elementary divisors on $\mathbf{\Lambda}^0$ with respect to $\mathscr{E}_x$, ordered by decreasing values, the minimum by direct lexicographic ordering coincides with the maximum by reverse lexicographic ordering, and this value is exactly the type $T(\mathscr{E})$. Moreover, any trivial lattice having $T(\mathscr{E})$ as elementary divisors is a BG trivialization of $\mathscr{E}$. Finally, the BG trivializations of a lattice Λ are exactly the geodesic projections of Λ onto $\mathbf{\Lambda}^0$ for both metrics $\boldsymbol{d}$ and ℓ. This result will be proved in two steps: Section 3.2 is dedicated to proving the first item and the implications (a) $\Rightarrow$ (b), (c), (d), (iii), and Section 3.3 to the converse implications.

3.1.1. *Monopoles and BG trivializations.* Recall from Section 1 that the group of monopole gauge transforms at x sends a trivial sheaf $\mathscr{F}$ to a trivial sheaf $\widetilde{\mathscr{F}}$ such that $\mathscr{F}_{|X\setminus\{x\}} = \widetilde{\mathscr{F}}_{|X\setminus\{x\}}$. This group is described by the group of unimodular polynomial matrices $\mathrm{GL}_n(\mathbb{C}[T])$, that is, matrices of the form

$$P = P_0 + P_1 T + \cdots + P_k T^k \quad \text{where } \exists \alpha \in \mathbb{C}^*, \det P = \alpha.$$

More precisely, assuming that $x = 0$, a matrix Π is a monopole at x if and only $\Pi = P(1/z)$ with $P \in \mathrm{GL}_n(\mathbb{C}[X])$ and z is the standard coordinate on $\mathbb{P}^1(\mathbb{C})$.

A trivial bundle $\mathscr{F} \in \mathscr{H}_0$ is a trivialization of $\mathscr{E} \in \mathscr{H}$ at x if any basis $(\sigma) = (\sigma_1, \ldots, \sigma_n)$ of global sections of $\mathscr{F}$ spans the stalk $\mathscr{E}_y$ over the local ring $\mathbb{O}_y = (\mathbb{O}_X)_y$ for every $y \neq x$. Any other basis of meromorphic sections $(\widetilde{\sigma})$ spans a trivialization $\widetilde{\mathscr{F}}$ of $\mathscr{E}$ at x if and only if the gauge Π from (σ) to $(\widetilde{\sigma})$ is a monopole at x. In particular, since the group of units of $\mathbb{C}[X]$ is $\mathbb{C}^*$, a line bundle $\mathscr{L}$ admits a unique trivialization $\mathscr{L}^x$ at x.

Lemma 26. *Two BG bases for Λ are related by a $-T(\Lambda)$-parabolic monopole gauge.*

Proof. Consider two BG trivializations M, $\tilde{M}$ of Λ, as in the following diagram, where $D = T(\Lambda)$:

$$\begin{array}{ccc} \Lambda & \xrightarrow{z^D} & Y_M \\ {\scriptstyle P}\downarrow & & \downarrow{\scriptstyle \Pi} \\ \Lambda & \xrightarrow{z^D} & Y_{\tilde{M}} \end{array}$$

Since $v(P_{ij}) \geqslant 0$, the gauge Π satisfies $v(\Pi_{ij}) \geqslant d_j - d_i$. Therefore $\Pi_{ij} = 0$ as soon as $d_j > d_i$, which means that Π is $(-D)$-parabolic. □

According to the previous section, $\mathscr{F}$ is a BG trivialization of $\mathscr{E}$ at x if there exists a basis $(\sigma) = (\sigma_1, \ldots, \sigma_n)$ of global sections of $\mathscr{F}$ and an integer sequence $D = (d_1, \ldots, d_n)$, such that $(e) = (t^{-d_1}\sigma_1, \ldots, t^{-d_n}\sigma_n)$ spans the stalk $\mathscr{E}_x$ over the local ring $\mathbb{O} = (\mathbb{O}_X)_x$, where t is a local coordinate at x. This coordinate t can be arbitrarily chosen, since the local behavior of $\mathscr{E}$ only depends on the local ring $\mathbb{O}$. If we choose as coordinate t a *meromorphic* function on X, then the sections

$(e) = (t^{-d_1}\sigma_1, \ldots, t^{-d_n}\sigma_n)$ form a basis of global (meromorphic) sections of $\mathscr{V}$. The $\mathbb{O}_X$-module $\widetilde{\mathscr{F}}$ spanned by (e) in this case does coincide with $\mathscr{E}$ at x, and differs from it at most on the support of the divisor of the function t.

When $X = \mathbb{P}^1(\mathbb{C})$, we can obviously find a function t with divisor $(t) = x - y$ for any arbitrary point $y \neq x$, that we call a *global coordinate at* x. In this case, the bundle $\widetilde{\mathscr{F}}$ is a BG trivialization of $\mathscr{E}$ at y. It is clearly independent of the global basis (σ) of $\mathscr{F}$, which is defined up to a $(-D)$-parabolic constant matrix $C \in \mathrm{GL}_n(\mathbb{C})$, and of the specific meromorphic function t, which is only defined up to a nonzero constant. We call $t_y(\mathscr{F}) = \widetilde{\mathscr{F}}$ the *transport at* y of the BG trivialization $\mathscr{F}$ of $\mathscr{E}$ at x.

Understood otherwise, this is the description of a nontrivial bundle $\mathscr{E}$ by means of *two trivial* bundles $\mathscr{F}$ and $\widetilde{\mathscr{F}}$ coinciding outside $\{x, y\}$, and glued along the cocycle $g = t^D$, where $(t) = x - y$.

3.1.2. *The Harder–Narasimhan flag.* The Harder–Narasimhan filtration $\mathrm{HN}(\mathscr{E})$ of $\mathscr{E}$ over $\mathbb{P}^1(\mathbb{C})$ can be obtained (see [Sabbah 2002, p. 65]) from a BG decomposition $\mathscr{E} = \bigoplus_{i=1}^n \mathscr{L}_i$ of $\mathscr{E}$ as a direct sum of line bundles $\mathscr{L}_i \simeq \mathbb{O}(a_i)$ of the appropriate degree, by

$$F^k(\mathscr{E}) = \bigoplus_{i\,|\,a_i \geqslant k} \mathscr{L}_i.$$

Locally, the Harder–Narasimhan filtration can be described as follows. Let (e) be a BG basis for $\Lambda = \mathscr{E}_x$. The *Harder–Narasimhan flag* HN_Λ *of* $V = \mathscr{V}_x$ defined by

$$F^k = \bigoplus_{i\,|\,a_i \geqslant k} Ke_i \tag{7}$$

is independent of the BG basis for Λ. For a lattice M, define the *residual HN flag*

$$\boldsymbol{H}_\Lambda^M = (\mathrm{HN}_\Lambda \cap M)/\mathfrak{m}M \quad \text{and, for simplicity,} \quad \boldsymbol{H}^\Lambda = \boldsymbol{H}_\Lambda^\Lambda.$$

Lemma 27. *Let* $\Lambda \in \boldsymbol{\Lambda}$ *and let* HN_Λ *be the HN flag of* V.

(i) *If* $M \in \mathrm{BG}(\Lambda)$, *then* $\boldsymbol{H}_\Lambda^M = \boldsymbol{F}^M(\Lambda)$, *hence* $\boldsymbol{H}^\Lambda$ *is transversal to* $\boldsymbol{F}^\Lambda(M)$.

(ii) *Conversely, for any flag* $\boldsymbol{F}'$ *which is transversal to* $\boldsymbol{H}^\Lambda$, *there exists* $M \in \mathrm{BG}(\Lambda)$ *such that* $\boldsymbol{F}^\Lambda(M) = \boldsymbol{F}'$.

Proof. Let $T = T(\Lambda) = \mathrm{diag}(a_1 I_{n_1}, \ldots, a_s I_{n_s})$ with $a_i > a_{i+1}$ be the type of the lattice Λ, so that

$$\Lambda \xrightarrow{z^T} M$$

sends the basis (e) of Λ into the global basis $(\varepsilon) = z^T(e)$ of Y_M. Let $\nu_i = \sum_{1 \leqslant k \leqslant i} n_i$, and let generically $\bar{x}$ denote the class modulo $\mathfrak{m}$ of a vector x. The i-th component of the flag $\boldsymbol{H}_\Lambda^M$ is spanned by $(\bar{\varepsilon}_1, \ldots, \bar{\varepsilon}_{\nu_i})$, so $\boldsymbol{H}_\Lambda^M = \boldsymbol{F}^M(\Lambda)$ indeed

holds. Conversely, the $(n-i+1)$-th component of $\boldsymbol{F}^{\Lambda}(M)$ is $F_i = \langle \bar{e}_{\nu_i+1}, \ldots, \bar{e}_n \rangle$ while the i-th component H_i of $\boldsymbol{H}^{\Lambda}$ satisfies $H_i = \langle \bar{e}_1, \cdots, \bar{e}_{\nu_i} \rangle$; hence both flags are transversal to each other. Any other BG trivialization $\tilde{M}$ is obtained from (ε) by a monopole gauge transform Π such that $P = z^T \Pi z^{-T} \in \mathrm{GL}_n(\mathbb{O})$. According to Lemma 26, Π is block-upper-triangular with respect to the blocks of equal elements of T, hence so is P. Let $P \in G_T(\mathbb{C})$. If z is a global coordinate, the matrix $z^{-T} P z^T$ is a monopole. The orbit of (ε) under the set of the constant T-parabolic matrices covers the set of all flags in E which are transversal to the image of $\mathrm{HN}(\mathscr{E})_x$ in E. □

For any BG trivialization $\mathscr{F}$ of $\mathscr{E}$ at x, let $Y = \Gamma(X, \mathscr{F})$ be the $\mathbb{C}$-vector space of global sections of $\mathscr{F}$. The Harder–Narasimhan filtration $\mathrm{HN}(\mathscr{E})$ also induces a canonical filtration $\mathrm{HN}_\Lambda(Y)$ of $\mathbb{C}$-vector spaces of Y. To avoid defining new concepts, we will also refer to this filtration as the *Harder–Narasimhan filtration of* Y. Note that it depends solely on the lattice $\Lambda \in \boldsymbol{\Lambda}$.

3.2. ***Modification of the type.*** We wish to answer algebraically the following question: "what does the type of $\mathscr{E}$ become when the stalk $\mathscr{E}_x = \Lambda$ at x is replaced by another lattice $\widetilde{\Lambda}$?" It turns out that the question can be very explicitly answered when the lattice $\widetilde{\Lambda}$ is not too far from Λ, namely at distance 1 in the graph-theoretic distance of the Bruhat–Tits building. The following proposition generalizes a result of Gabber and Sabbah.

Proposition 28. *Let $\mathscr{E} \simeq \bigoplus_{i=1}^n \mathbb{O}(a_i)$ be a holomorphic vector bundle on $X = \mathbb{P}^1(\mathbb{C})$, with $a_1 \geqslant \cdots \geqslant a_n$, and let $x \in X$. Let $\widetilde{\Lambda} \in \boldsymbol{\Lambda}_x$ be a lattice such that $\mathfrak{m}_x \mathscr{E}_x \subset \widetilde{\Lambda} \subset \mathscr{E}_x$. Let $E = \mathscr{E}_x / \mathfrak{m}_x \mathscr{E}_x$ be the local fiber at x, let*

$$\boldsymbol{H} : H_0 = 0 \subset H_1 \subset \cdots \subset H_s = E$$

be the residual HN flag in E, and $W = \widetilde{\Lambda} / \mathfrak{m}_x \mathscr{E}_x$ the image of $\widetilde{\Lambda}$. Assume that the type of $\mathscr{E}$ is written as

$$a = (\underbrace{a_1, \ldots, a_1}_{n_1 \text{ times}}, \underbrace{a_2, \ldots, a_2}_{n_2 \text{ times}}, \ldots, \underbrace{a_s, \ldots, a_s}_{n_s \text{ times}}).$$

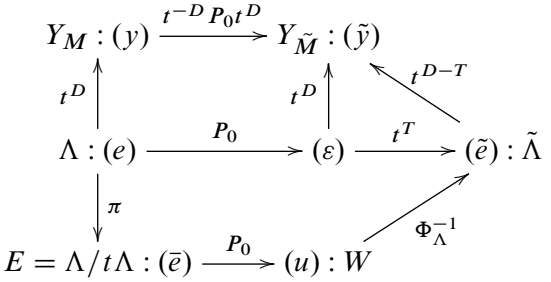
Then the modified bundle $\mathscr{F} = \mathscr{E}^{\widetilde{\Lambda}}$ has type

$$\tilde{a} = (\underbrace{a_1, \ldots, a_1}_{m_1 \text{ times}}, \underbrace{a_1 - 1, \ldots, a_1 - 1}_{n_1 - m_1 \text{ times}}, \ldots, \underbrace{a_s, \ldots, a_s}_{m_s \text{ times}}, \underbrace{a_s - 1, \ldots, a_s - 1}_{n_s - m_s \text{ times}})$$

where $m_i = \dim_{\mathbb{C}} H_i \cap W - \dim_{\mathbb{C}} H_{i-1} \cap W$.

Proof. This is explained in the following scheme. Let $\Lambda = \mathscr{E}_x$, and let t be a local coordinate at x (that we assume without loss of generality to be ∞). Let

$D = \mathrm{diag}(a_1, \ldots, a_n)$ be the elementary divisors of the BG trivialization M in Λ (or, in this case, the type of $\mathscr{E}$).

(8)


Let (e) be a basis of Λ, such that $(\sigma) = (t^D e)$ is a basis of the form Y_M. Under the canonical projection $\pi : \Lambda \to E = \Lambda / t\Lambda$, the HN filtration of Λ descends to a flag of $\mathbb{C}$-vector spaces $\boldsymbol{H} : 0 = H_0 \subset \cdots \subset H_s = E$, and the quotient basis $(\bar{e})$ is a basis respecting this flag. Let $t\Lambda \subset \widetilde{\Lambda} \subset \Lambda$ be the new lattice, and let $W \subset E$ be the subspace it is projected upon by π. Let (u) be a basis respecting both W and the flag $\boldsymbol{H}$, and let P_0 be a change of basis from (e) to (u). Consequently, the matrix P_0 belongs to the parabolic subgroup $P_{\boldsymbol{H}}$ stabilizing the flag $\boldsymbol{H}$; therefore it is block-upper-triangular, with blocks given by the equal elements among the a_i. Define now the basis (ε) of Λ as the image of (e) under the constant gauge P_0. Here is where $d(\Lambda, \widetilde{\Lambda}) \leqslant 1$ is important: the basis (ε) is a Smith basis of $\widetilde{\Lambda}$ (this would be not necessarily true if the lattices were further apart). Let $T = \mathrm{diag}(t_1, \ldots, t_n)$ be the diagonal matrix such that $t_i = 0$ if $\pi(\varepsilon_i) \in W$ and $t_i = 1$ otherwise. Then $(\tilde{e}) = t^T(\varepsilon)$ is a basis of $\widetilde{\Lambda}$. Let now $(\widetilde{\varepsilon}) = t^D(\varepsilon)$ be the basis of $\widetilde{M}$ deduced from (ε). The matrix of the basis change from $\widetilde{\Lambda}$ to $\widetilde{M}$ corresponding to the bases (σ) and $(\widetilde{\varepsilon})$ is equal to $Q = t^{-D} P_0 t^D = (P_0)_{ij} t^{d_j - d_i}$. Now, since $P_0 \in P_{\boldsymbol{H}}$, we have $(P_0)_{ij} = 0$ whenever $d_i - d_j < 0$. Therefore this gauge $Q = \frac{1}{t^k} Q_k + \cdots + Q_0$ is a Laurent polynomial in t with only nonpositive terms, where moreover $Q_0 \in \mathrm{GL}_n(\mathbb{C})$. Since $X = \mathbb{P}^1(\mathbb{C})$, it is possible to choose as local coordinate at ∞ a meromorphic function with divisor $(\infty) - (0)$, namely $t = 1/z$. Accordingly, Q is a polynomial in z, whereas $\det Q = \det P_0 \in \mathbb{C}^*$. Hence $Q \in \mathrm{GL}_n(\mathbb{C}[z])$ is a monopole gauge. Since (σ) was a basis of global meromorphic sections of E, then $(\widetilde{\varepsilon})$ also is. Therefore $\widetilde{M} \in B_0$ is a trivializing lattice. Moreover, $\widetilde{M}$ is a BG trivialization of both $\mathscr{E}$ and $\mathscr{F} = \mathscr{E}^{\widetilde{\Lambda}}$, because the basis $(\widetilde{\varepsilon})$ is a Smith basis for Λ and $\widetilde{\Lambda}$. Therefore, we can explicitly compute the new elementary divisors of $\widetilde{\Lambda}$ in $\widetilde{M}$, which are given by the matrix $D - T$. Summing up, we see that the change of lattice has subtracted 1 from all the elementary divisors corresponding to the vectors of the basis (ε) whose image under π do not fall into the subspace W. We obtain the Harder–Narasimhan filtration of the modified bundle by reordering the type by decreasing values. □

Proposition 28 generalizes the construction given by Sabbah based on an idea of O. Gabber in [Sabbah 2002, Proposition 4.11] (where only the case where W is one-dimensional is tackled). This result is independent of the valuation of $\widetilde{\Lambda}$, and can thus be formulated as follows.

Corollary 29. *Let* $\Lambda, \widetilde{\Lambda} \in \mathbf{\Lambda}$, *and let* $\boldsymbol{H} = (\mathrm{HN}_\Lambda \cap \Lambda)/\mathfrak{m}\Lambda$ *be the residual HN flag and* $\boldsymbol{F} = \boldsymbol{F}^\Lambda(\widetilde{\Lambda}) = (0 \subset W \subset \Lambda/\mathfrak{m}\Lambda)$ *the flag induced by* $\widetilde{\Lambda}$ *in* $\Lambda/\mathfrak{m}\Lambda$. *If* $d(\Lambda, \widetilde{\Lambda}) = 1$, *then we have*

$$T(\widetilde{\Lambda}) \cong T(\Lambda)^{\searrow} - \sigma\big(\mathbf{ED}_\Lambda(\widetilde{\Lambda})^{\nearrow}\big)$$

where $\sigma \in \rho(\boldsymbol{F}, \boldsymbol{H})$ *represents the relative position of* $\boldsymbol{F}$ *and* $\boldsymbol{H}$. *Actually, one can even say, putting* $\widetilde{\boldsymbol{H}} = (\mathrm{HN}_{\widetilde{\Lambda}} \cap \Lambda)/\mathfrak{m}\Lambda$, *that*

$$T(\widetilde{\Lambda})^{\searrow} = \rho(\boldsymbol{H}, \widetilde{\boldsymbol{H}})\, T(\Lambda)^{\searrow} - \rho(\boldsymbol{F}, \widetilde{\boldsymbol{H}})\big(\mathbf{ED}_\Lambda(\widetilde{\Lambda})^{\nearrow}\big).$$

Proof. With the notations of the diagram (8), the basis $(\tilde{y})$ is a common BG basis for Λ and $\widetilde{\Lambda}$. Therefore, the HN flags are spanned in $(\tilde{y})$ over K by the flags of indices $\boldsymbol{S}^{\searrow}(D)$ and $\boldsymbol{S}^{\searrow}(D-T)$, respectively. By applying any representative of the coset $\rho(\boldsymbol{H}, \widetilde{\boldsymbol{H}})$ to $D-T$, one gets $T(\widetilde{\Lambda})^{\searrow}$. Therefore, $T(\widetilde{\Lambda})^{\searrow} = \rho(\boldsymbol{H}, \widetilde{\boldsymbol{H}})\, T(\Lambda)^{\searrow} - \rho(\boldsymbol{H}, \widetilde{\boldsymbol{H}})\rho(\boldsymbol{F}, \boldsymbol{H})(\mathbf{ED}_\Lambda(\widetilde{\Lambda})^{\nearrow})$. Take as representatives of the cosets $\rho(\boldsymbol{H}, \widetilde{\boldsymbol{H}})$ and $\rho(\boldsymbol{F}, \boldsymbol{H})$ their minimal length element (see, e.g., [Abramenko and Brown 2008, Proposition 2.23, p. 83]). Since the quotient basis $(\bar{\varepsilon})$ of $\Lambda/\mathfrak{m}\Lambda$ is adapted to the three flags $\boldsymbol{H}$, $\widetilde{\boldsymbol{H}}$ and $\boldsymbol{F}$, we get $\rho(\boldsymbol{H}, \widetilde{\boldsymbol{H}})\rho(\boldsymbol{F}, \boldsymbol{H}) = \rho(\boldsymbol{F}, \widetilde{\boldsymbol{H}})$ [Abramenko and Brown 2008, Lemma 5.55, p. 236]. □

Corollary 30. *Let* $\Lambda \in \mathbf{\Lambda}$. *Then we have*

(i) $\mathrm{BG}(\Lambda) \neq \varnothing$,

(ii) *for any adjacent lattice* $\widetilde{\Lambda}$, *we have* $\mathrm{BG}(\Lambda) \cap \mathrm{BG}(\widetilde{\Lambda}) \neq \varnothing$,

(iii) *all the elements in a chamber of* B_n *have a common BG trivialization.*

Proof. According to Proposition 28, if a lattice Λ admits a BG trivialization, so does M for any adjacent lattice M. However, according to Proposition 17, two lattices are always connected by a path of adjacent lattices. Since a trivial lattice is its own BG trivialization, the two first results are simultaneously established. The third stems from the fact that any chamber appears as a complete flag in the quotient space of any representative. According to Bruhat's lemma, there is a basis respecting simultaneously two flags, in this case the one corresponding to the chamber and the one induced by the HN filtration. □

3.2.1. *An algorithm to compute a Birkhoff–Grothendieck trivialization.* Let $x \in X$, and let $\mathbf{\Lambda} = [\mathscr{E}]_x$ be the $\sim_x$-equivalence class of $\mathscr{E}$. Let $\Lambda = \mathscr{E}_x$ and $M = \mathscr{F}_x \in \mathbf{\Lambda}$ where $\mathscr{F}$ is an arbitrary trivialization of $\mathscr{E}$ at x. In this local setting, we "see" the

global sections of $\mathscr{F}$ as the *global form* $Y \subset M$. An arbitrary trivialization of $\mathscr{E}$ is not necessarily BG. The following result is easily established.

Lemma 31. *Let $M \in \mathrm{BG}(\Lambda)$. Then we have*

(i) $M \in \mathrm{BG}(z^k\Lambda)$ *for* $k \in \mathbb{Z}$,

(ii) $M \in \mathrm{BG}(\Lambda')$ *for any lattice* Λ' *on the shortest path* $\Gamma_{\min}(\Lambda, M)$.

Proposition 28 allows to construct effectively from an arbitrary trivialization M a BG one, by following shortest paths in the Bruhat–Tits building from M to $\mathscr{E}_x$. The following result shows how to start the construction. We can assume that $v_M(\Lambda) = 0$.

Lemma 32. *Let $M \in \mathbf{\Lambda}^0$ be a trivial lattice in B_n. For any lattice $\Lambda \in \mathbf{\Lambda}$ such that $d(\Lambda, M) = 1$, we have $M \in \mathrm{BG}(\Lambda)$. More precisely, let $Y \subset M$ be the global form of M. For any basis (e) respecting $W = \Lambda/\mathfrak{m}M$, the Y-basis (e_Y) is a Smith basis for Λ.*

Proof. Assume that (e) satisfies the assumptions of the lemma. Then, according to Lemma 13, the Y-basis (e_Y) is a Smith basis of M for Λ. Since it is a basis of the global form of M, the result follows, and in particular, the Harder–Narasimhan filtration of the corresponding bundle is equal to the Y-lifting of the flag $(0 \subset W \subset M/\mathfrak{m}M)$. □

If there existed a Smith basis of M for Λ which spanned simultaneously Y_M, the lattice M would be a BG trivialization of Λ, and the sequence $\mathbf{ED}_\Lambda(M)$ would represent the type of $\mathscr{E}$.

Theorem 33. *Let $\Lambda \in \mathbf{\Lambda} = [\mathscr{E}]$, and let $M \in \mathbf{\Lambda}^0$ be an arbitrary trivial lattice. Let $\mathbf{ED}_\Lambda(M) = \Delta_0 + \cdots + \Delta_t$ be the elementary splitting of $\mathbf{ED}_\Lambda(M)$. There exists a sequence of permutations $w_k \in S_n$ such that the type $T(\Lambda)$ satisfies*

$$T(\Lambda) \cong \Delta_0 + w_1\Delta_1 + w_2\Delta_2 + \cdots + w_t\Delta_t.$$

Proof. We prove the result first on the sequence $D = \mathbf{ED}_M(\Lambda) = -\mathbf{ED}_\Lambda(M)$. Let $\Gamma = (M = M_0, M_1, \ldots, M_t = \Lambda_M)$ be (a normalized representative of) the min-shortest path from $[M]$ to $[\Lambda]$. Let $D = \mathbf{ED}_M(\Lambda_M) = (\boldsymbol{k}_1 I_{n_1}, \ldots, \boldsymbol{k}_s I_{n_s})$ where $0 = \boldsymbol{k}_1 < \cdots < \boldsymbol{k}_s = t$. Consider the elementary splitting of D

$$D = \Delta_1 + \cdots + \Delta_t \quad \text{where } \Delta_i = (0_{n-m_i}, I_{m_i}) \tag{9}$$

for a nonincreasing sequence (m_i). Recall that

$$D_k = \Delta_1 + \cdots + \Delta_k = \mathbf{ED}_M(M_k).$$

Let (e) be a Smith basis of M for Λ, and let $(e^{(k)}) = z^{D_k}(e)$ for $0 \leqslant k \leqslant t$ (with $D_0 = 0$ by convention). The induced basis $(\bar{e}^{(k)})$ of $E_k = M_k/\mathfrak{m}M_k$ respects both flags $\boldsymbol{F}^{M_k}(M)$ and $\boldsymbol{F}^{M_k}(\Lambda)$. With the help of Proposition 28,

we construct a BG trivialization $\tilde{M}_{k+1}$ of the k-th element M_k of the shortest path Γ, which is simultaneously a BG trivialization of $\Lambda_M + \mathfrak{m}M_k = \Lambda_M + \mathfrak{m}(\Lambda_M + \mathfrak{m}^k M) = M_{k+1}$.

Let us describe this in more detail. By assumption, there exists a BG basis (σ) of $\tilde{M}_k$ for M_k. Let

$$T = \mathbf{ED}_{\tilde{M}_k}(M_k)^{\nearrow} \quad \text{and} \quad T' = \mathbf{ED}_{M_k}(\Lambda_M)^{\nearrow}.$$

Let $(y) = z^T(\sigma)$ be the corresponding basis of M_k, and (ε) a Smith basis of M_k for Λ_M. The gauge U from (y) to (ε) can be factored as

$$U = U_0(I + tU') \quad \text{with } U_0 \in \mathrm{GL}_n(\mathbb{C}). \tag{10}$$

Since (σ) is a basis of the global form $Y_{\tilde{M}_k}$, the basis $(\overline{y})$ of $E = M_k/\mathfrak{m}M_k$ is strictly adapted to the HN flag

$$\boldsymbol{H}^{(k)} = (\mathrm{HN}_{M_k} \cap M_k)/\mathfrak{m}M_k.$$

Similarly, $(\overline{\varepsilon})$ is strictly adapted to the flag $\boldsymbol{F}^{(k)} = \boldsymbol{F}^{M_k}(\Lambda_M)$. Let B be the standard Borel subgroup of $\mathrm{GL}_n(\mathbb{C})$. By the Bruhat decomposition, the group $\mathrm{GL}_n(\mathbb{C})$ is a disjoint union of double cosets:

$$\mathrm{GL}_n(\mathbb{C}) = \coprod_{w \in W} BwB$$

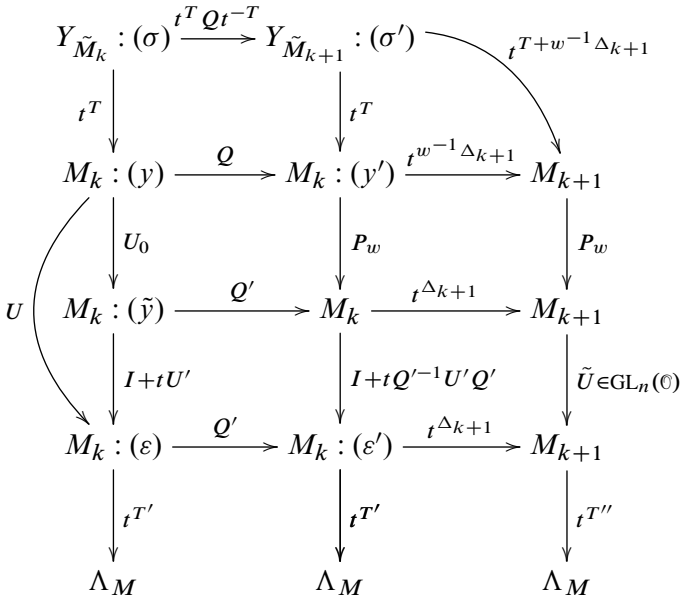
where W is the Weyl group $W = S_n$. The constant term U_0 of the gauge U belongs to only one such cell: let $w \in S_n$ be the label of the corresponding Schubert cell. We have a decomposition $U_0 = QP_wQ'^{-1}$ with $Q, Q' \in B$, where P_w is the matrix representation of the permutation w. Accordingly, the gauge transforms Q and Q' respect, respectively, the flags $\boldsymbol{H} = \boldsymbol{H}^{(k)}$ and $\boldsymbol{F} = \boldsymbol{F}^{(k)}$. In the quotient space E, we have:

$$\begin{array}{ccc} E : (\overline{y}) & \xrightarrow{Q} & E : (\overline{y}') \\ \Big\downarrow{\scriptstyle U_0} & & \Big\downarrow{\scriptstyle P_w} \\ E : (\overline{\varepsilon}) & \xrightarrow{Q'} & E : (\overline{\varepsilon}') \end{array}$$

The gauge U_0 represents geometrically the change of a basis that spans the Harder–Narasimhan flag $\boldsymbol{H}$ to one that spans the flag $\boldsymbol{F}$ induced by Λ; therefore w is a representative of the relative position $\rho(\boldsymbol{H}, \boldsymbol{F})$.

Let $T' = \Delta_{k+1} + \cdots + \Delta_t$ be the elementary splitting of T'. Since $(\overline{\varepsilon})$ respects the flag $\boldsymbol{F}$, it will in particular respect the trace of the first element $M_{k+1} = \Lambda + \mathfrak{m}M_k$ of the shortest path $\Gamma_{\min}(M_k, \Lambda)$; therefore any lifting of $(\overline{\varepsilon})$ will be a Smith basis of M_{k+1} with elementary divisors Δ_{k+1}. Put $T'' = T' - \Delta_{k+1}$. The previous

scheme gets thus lifted to the following complete picture.



As a result, the elementary divisors of M_{k+1} with respect to the common BG trivialization $\tilde{M}_{k+1}$ of M_k and M_{k+1} are not $T+\Delta_{k+1}$ (as with respect to $\tilde{M}_k$), but $T+w^{-1}\Delta_{k+1}$; namely, the elements of Δ_{k+1} have been twisted according to the permutation $w^{-1}=\rho(\boldsymbol{F},\boldsymbol{H})$ indexing the Bruhat cell that contains the matrix $U_0\in \mathrm{GL}_n(\mathbb{C})$. The resulting matrix $\widetilde{T}=T+w^{-1}\Delta_{k+1}$ is not necessarily ordered by increasing values: therefore we cannot ensure that $w_{k+1}=\rho(\boldsymbol{F},\boldsymbol{H})$, since the *ordered* diagonal has the form $\sigma T+\sigma w^{-1}\Delta_{k+1}$. According to Corollary 29, however, we know that we can take

$$\sigma=\rho(\boldsymbol{H}^{M_k},\boldsymbol{H}_{M_k}^{M_{k+1}})=\rho(\mathrm{HN}_{M_k},\mathrm{HN}_{M_{k+1}}).$$

Thus

$$\widetilde{T}^{\nearrow}=\rho(\mathrm{HN}_{M_k},\mathrm{HN}_{M_{k+1}})T+\rho(\boldsymbol{H}^{M_k},\boldsymbol{H}_{M_k}^{M_{k+1}})\rho(\boldsymbol{F}^{(k)},\boldsymbol{H}^{M_k})\Delta_{k+1}.$$

Putting $T_k=\mathbf{ED}_{\tilde{M}_k}(M_k)^{\nearrow}$, we get

$$T_{k+1}=\rho(\mathrm{HN}_{M_k},\mathrm{HN}_{M_{k+1}})T_k+\rho\big(\boldsymbol{F}^{M_k}(\Lambda),\boldsymbol{H}_{M_k}^{M_{k+1}}\big)\Delta_{k+1}. \tag{11}$$

At the end of at most t steps, the lattice $\tilde{M}_t$ is a BG trivialization of Λ_M, and thus of Λ. We have $\widetilde{T}=w_1\Delta_1+w_2\Delta_2+\cdots+w_t\Delta_t$ such that

$$\tilde{M}_t\xrightarrow{z^{\widetilde{T}}}\Lambda_M\xrightarrow{z^{v_M(\Lambda)}}\Lambda.$$

Since $T(\Lambda) = \mathbf{ED}_\Lambda(\tilde{M}_t) = -\tilde{T} - v_M(\Lambda)$, we get

$$\begin{aligned} T(\Lambda) &= -\Delta_0 - w_1\Delta_1 - w_2\Delta_2 - \cdots - w_t\Delta_t \\ &= -\Delta_0 - tI + w_1(I-\Delta_1) + (I - w_2\Delta_2) + \cdots + (I - w_t\Delta_t) \\ &= (-\Delta_0 - tI) + w_t(I-\Delta_t) + w_{t-1}(I-\Delta_{t-1}) + \cdots + w_1(I-\Delta_1). \end{aligned}$$

By Lemma 19, the result is established. Note that a similar relation holds for $\mathbf{ED}_M(\Lambda)$, and that if $T(\Lambda) = \Delta_0 + w_1\Delta_1 + w_2\Delta_2 + \cdots + w_t\Delta_t$, then $T(\Lambda_k) = \Delta_0 + w_{k+1}\Delta_{k+1} + \cdots + w_t\Delta_t$ for the normalized elements Λ_k of the shortest path $\Gamma_{\min}(\Lambda, M)$. □

We can even specify an actual (noneffective) formula for the permutations w_i appearing in $T(\Lambda)^\nearrow = \Delta_0 + w_1\Delta_1 + w_2\Delta_2 + \cdots + w_t\Delta_t$, attached to the shortest path

$$\Lambda \xrightarrow{z^{\Delta_1}} \Lambda_1 \xrightarrow{z^{\Delta_2}} \Lambda_2 \longrightarrow \cdots \xrightarrow{z^{\Delta_{t-1}}} \Lambda_{t-1} \xrightarrow{z^{\Delta_t}} M_\Lambda \xrightarrow{z^{\Delta_0}} M;$$

namely,

$$w_i = \rho(\mathrm{HN}_{\Lambda_1}, \mathrm{HN}_\Lambda) \cdots \rho(\mathrm{HN}_{\Lambda_{i-1}}, \mathrm{HN}_{\Lambda_{i-2}}) \rho\big(\boldsymbol{F}^{\Lambda_i}(\Lambda), \boldsymbol{H}^{\Lambda_{i-1}}_{\Lambda_i}\big).$$

The formula cannot be reduced as in Corollary 29 since there is not necessarily an apartment adapted for three such consecutive flags.

3.2.2. *The abacus.* Theorem 33 has a nice combinatorial interpretation in terms of Young diagrams. For any integer sequence $D = (d_1, \ldots, d_n) \in \mathbb{Z}^n$, let $D = \Delta_0 + \cdots + \Delta_t$ be the elementary splitting of D. For $w = (w_1, w_2, \ldots, w_t) \in (S_n)^t$, let $w(D) = \Delta_0 + w_1\Delta_1 + w_2\Delta_2 + \cdots + w_t\Delta_t$. The *abacus* of D is the set of sequences

$$\mathrm{ab}(D) = \{A \in \mathbb{Z}^n \mid \exists w \in (S_n)^t, A = w(D)\}.$$

The name comes from the following analogy. Let $Y(D)$ be the Young diagram with n rows whose respective lengths are the elements of $D_0 = D - \min D \in \mathbb{N}^n$, assumed to be arranged in increasing order (rows of length 0 are included). Assume from now on that $\min D = 0$. Let $d_i^* = |\{1 \leqslant j \leqslant n \mid d_j \geqslant i\}|$ be the number of boxes in the i-th column of $Y(D)$. The sequence $D^* = (d_1^*, \ldots, d_t^*)$ where $t = \max D$ is the *dual* sequence of D. The elementary splitting of D is given by $\Delta_i = (0_{n-d_i^*}, I_{d_i^*})$. In view of Lemma 19, put

$$\ell(D) = \sum_{i=1}^{t-1} (n - d_i^*) d_{i+1}^*.$$

Any sequence A in the abacus $\mathrm{ab}(D)$ of $D \in \mathbb{Z}^n$ comes from a box diagram obtained from $Y(D)$ by allowing to move some boxes *only vertically* inside the

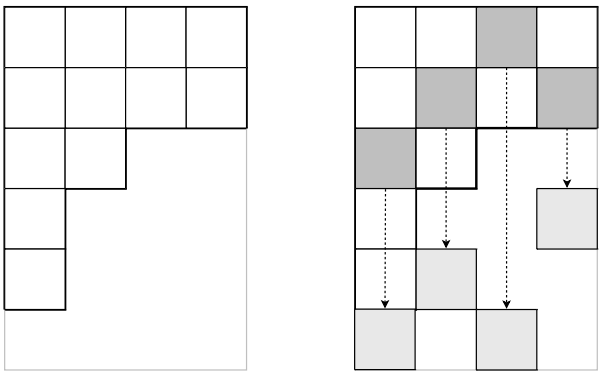


Figure 1. The Young diagram $Y(D)$ of the sequence $D = (0,1,1,2,4,4)$, and an element $A = (2,2,2,1,2,3) \in \mathrm{ab}(D)$, featuring the moved boxes (in shades of gray), so that $A^{\nearrow} = (1,2,2,2,2,3)$. The complement (thin gray line) corresponds to the sequence $D^0 = (0,0,2,3,3,4)$, and the bijection from Lemma 35, to reversing the arrows.

whole corresponding column of length n. The diagram thus obtained, that we call an *abacus diagram*, can have nonadjacent boxes (as shown in Figure 1). For any such diagram, the sequence $(a_1,\dots,a_n)$ of number of boxes contained in each of the n rows is the required element of $\mathrm{ab}(D)$.

Lemma 34. *Let $D = \Delta_1 + \cdots + \Delta_t$ be the elementary decomposition of $D \in (\mathbb{Z}^n)^{\nearrow}$, and let $A = w_1\Delta_1 + \cdots + w_t\Delta_t \in \mathrm{ab}(D)$. There exist $w'_1,\dots,w'_t \in S_n$ such that*

$$w'_1\Delta_1 + \cdots + w'_i\Delta_i = (w_1\Delta_1 + \cdots + w_i\Delta_i)^{\nearrow} \quad \textit{for all } 1 \leqslant i \leqslant d.$$

Proof. We proceed by induction on the number s of columns in $Y(D)$. Let Y be the Young diagram for $D = (d_1,\dots,d_n)$ and $\widetilde{Y}$ the one obtained from Y by erasing the last column, i.e., corresponding to the sequence $\widetilde{T} = (\Delta_1,\dots,\Delta_{t-1})$. Let $\widetilde{D} = (\tilde{d}_1,\dots,\tilde{d}_n)$ be the associated sequence. Then we have $d_i = \tilde{d}_i$ for $1 \leqslant i \leqslant n - d_t^*$ and $d_i = \tilde{d}_i + 1$ for $n - d_t^* + 1 \leqslant i \leqslant n$. An element $A \in \mathrm{ab}(D)$ given by the permutations $w = (w_2,\dots,w_t)$ corresponds uniquely to the pair $(\widetilde{A}, w_t)$ where $\widetilde{A} \in \mathrm{ab}(\widetilde{D})$ is given by the restriction $\widetilde{w} = (w_2,\dots,w_{t-1})$.

The claim is clear for $t = 1$, so assume that it is established for any $\widetilde{D}$ such that $\Delta\widetilde{D} \leqslant t-1$. We have $D = \Delta_1 + \cdots + \Delta_t = \widetilde{D} + \Delta_t$ where $\widetilde{D} = \Delta_1 + \cdots + \Delta_{t-1}$ is the elementary decomposition of $\widetilde{D}$. Let $A \in \mathrm{ab}(D)$ be described by the number a_i of boxes in the i-th row for $1 \leqslant i \leqslant n$, and let $\widetilde{A} = (\tilde{a}_1,\dots,\tilde{a}_n)$ be the restriction of A to the $t-1$ first columns. Let $\mathcal{J} = \{i \,|\, a_i = \tilde{a}_i + 1\}$. Note that $\mathcal{J} = w_t(\{n - d_t^* + 1,\dots,n\})$, and thus $|\mathcal{J}| = d_t^*$. According to the assumption, we can find $w'_1,\dots,w'_{t-1}$ such that $w'_1\Delta_1 + \cdots + w'_i\Delta_i = (w_1\Delta_1 + \cdots + w_i\Delta_i)^{\nearrow}$ for all $1 \leqslant i \leqslant t-1$. In particular,

we can assume that $\tilde{a}_1 \leqslant \cdots \leqslant \tilde{a}_n$. If there exists an index i such that $a_i > a_{i+1}$ holds, then we have necessarily $\tilde{a}_i = \tilde{a}_{i+1} = a_{i+1}$ and $a_i = \tilde{a}_i + 1$, so $i \in \mathcal{J}$ and $i+1 \notin \mathcal{J}$. Exchanging i and $i+1$ will not change the resulting sequence A when reordered, so one can avoid the inversion by putting $w'_t = (i, i+1)w_t$, so that $i \notin \mathcal{J}$ and $i+1 \in \mathcal{J}$ instead. By repeating this procedure for all the inverted indices, we get the claimed result for t. □

For an $n \times t$ rectangular matrix $M = (M_{i,j})$, define the *row sum vector*

$$r(M) = (r_1, \ldots, r_t), \quad r_j = \sum_{i=1}^{n} M_{i,j},$$

and *column sum vector*

$$c(M) = (c_1, \ldots, c_n), \quad c_i = \sum_{j=1}^{t} M_{i,j}.$$

An element $A \in \mathrm{ab}(D)$ in the abacus of D can also be seen as a $(0,1)$ rectangular $n \times t$ matrix $\mathscr{A} = (A_{i,j})$ where $A_{i,j} = 1$ whenever $i \in w_j(\{n - d_j^* + 1, \ldots, n\})$ holds. This matrix $\mathscr{A}$ has row sum vector $r(\mathscr{A}) = D^*$, and column sum vector $c(\mathscr{A}) = A$. Recall that for two sequences $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_t)$ having the same sum, one says that p *dominates* q when $\sum_{i=1}^{k} q_i \leqslant \sum_{i=1}^{k} p_i$ for all integers k, where one completes the missing elements with 0. The following lemma sums up the behavior of the quantities introduced in Section 2.1 under the abacus transformations.

Lemma 35. *Let* $D \in \mathbb{Z}^n$.

(i) *The maps* max, Δ, i, ℓ *and* $\|\cdot\|$ *all attain their maximum over* $\mathrm{ab}(D)$ *at* D. *The map* min *attains its minimum at* D.

(ii) *For any sequence* $A \in \mathrm{ab}(D)$, *the following hold*:
  (a) $\max A = \max D \iff i(A) = i(D)$,
  (b) $\Delta A = \Delta D \iff \max A = \max D$ *and* $\min A = \min D$,
  (c) $\|A\| = \|D\| \iff \ell(A) = \ell(D) \iff A^{\nearrow} = D^{\nearrow}$.

Moreover, the map $A \mapsto \max D - A$ *is a bijection between* $\mathrm{ab}(D_0)$ *and* $\mathrm{ab}(D^0)$.

Proof. We will only prove the claims about $\ell(D)$ and $\|D\|$, since the others are easily derived from their definitions. Although the assertions are similar, the methods of proof will be different. Let us start with ℓ. Assume without loss of generality that $\min D = 0$. According to Lemma 34, one can also assume that $A = A^{\nearrow}$ holds. Define the integers k_i by induction as $k_1 = a_1^* - d_1^*$ and $k_i = a_i^* - d_i^* + k_{i-1}$ for $i \geqslant 2$. Since A can be seen as a $(0,1)$ matrix $\mathscr{A}$ with $r(\mathscr{A}) = D^*$ and $c(\mathscr{A}) = A$, the Gale–Ryser theorem [Krause 1996] implies that $c(\mathscr{A})^* = A^*$ dominates $r(\mathscr{A}) = D^*$.

Therefore

$$\sum_{i=1}^{k} d_i^* \leqslant \sum_{i=1}^{k} a_i^*$$

for all integers k. Hence $k_i \geqslant 0$ for all i. Putting $k_0 = k_t = 0$, we then have $a_i^* = d_i^* - k_{i-1} + k_i$ for $1 \leqslant i \leqslant d$. According to the definition, we have

$$\begin{aligned}
\ell(A) &= \sum_{i=1}^{t-1}(n-a_i^*)a_{i+1}^* = \sum_{i=1}^{t-1}(n-d_i^*+k_{i-1}-k_i)(d_{i+1}^*-k_i+k_{i+1}) \\
&= \sum_{i=1}^{t-1}(n-d_i^*)d_{i+1}^* + (k_{i-1}-k_i)(d_{i+1}^*-k_i+k_{i+1}) - (k_i-k_{i+1})(n-d_i^*) \\
&= \ell(D) - \sum_{i=1}^{t-1}(k_i-k_{i+1})(n-d_i^*+d_{i+1}^*+k_{i+1}-k_i).
\end{aligned}$$

After some algebraic manipulation, we get

$$\begin{aligned}
\ell(A)-\ell(D) &= -k_1\big(n-d_1^*+d_2^*-k_1+k_2-(d_3^*-k_2+k_3)\big) \\
&\quad \cdots - \sum_{i=2}^{t-2} k_i\big(d_{i+1}^*-k_i+k_{i+1}-(d_{i+2}^*-k_{i+1}+k_{i+2})+d_{i-1}^*-d_i^*\big) \\
&\quad \cdots - k_{t-1}(d_t^*-k_{t-1}+k_t+d_{t-2}^*-d_{t-1}^*) \\
&= -k_1(n-d_1^*+a_2^*-a_3^*)\cdots - \sum_{i=2}^{t-2} k_i(a_{i+1}^*-a_{i+2}^*+d_{i-1}^*-d_i^*) \\
&\quad \cdots - k_{t-1}(a_t^*+d_{t-2}^*-d_{t-1}^*).
\end{aligned}$$

Since both sequences (d_i^*) and (a_i^*) are nonincreasing, we get $\ell(A) \leqslant \ell(D)$. Moreover, if $\ell(A) = \ell(D)$ holds, then all these terms must be zero. Let us prove then by induction that $k_i = 0$ for all i. If $k_1 \neq 0$, then one has $n = d_1^*$ and $a_2^* = a_3^*$. Since $a_1^* \geqslant d_1^*$, one must therefore have $a_1^* = n$, so $k_1 = 0$ holds, and hence we get $d_1^* = a_1^*$. Assume now that $k_j = 0$ and $a_j^* = d_j^*$ hold for $j \leqslant i-1$. If $k_i \neq 0$ then one has $a_i^* > d_i^*$ and $d_{i-1}^* = d_i^*$ (and $a_{i+1}^* = a_{i+2}^*$ also). But then $d_{i-1}^* = d_i^* = a_{i-1}^* \geqslant a_i^* \geqslant d_i^*$, so we have $d_{i-1}^* = d_i^* = a_{i-1}^* = a_i^*$. Thus $k_i = a_i^* - d_i^* + k_{i-1} = 0$, and so the result for ℓ is established.

Let us turn now to $\|\cdot\|$. We can assume here that $w_1 = \mathrm{id}$. We proceed by induction on the number t of columns in the Young diagram $Y(D)$. Like in the proof of the previous result, the restricted column sequence $T' = (\Delta_1, \ldots, \Delta_{t-1})$ corresponds to the Young diagram $Y(D')$ of $D' = (d_1', \ldots, d_n')$ with $d_i = d_i'$ for $1 \leqslant i \leqslant n - d_t^*$ and $d_i = d_i' + 1$ for $n - d_t^* + 1 \leqslant i \leqslant n$. By the König–Huygens

identity, we have

$$\begin{aligned}\|D\| &= \sum_{i=1}^{n}\left(d_i - \frac{\operatorname{Tr} D}{n}\right)^2 = \sum_{i=1}^{n} d_i^2 - \frac{(\operatorname{Tr} D)^2}{n}\\ &= \sum_{i=1}^{n-d_t^*} (d_i')^2 + \sum_{i=n-d_t^*+1}^{n} (d_i'+1)^2 - \frac{(\operatorname{Tr} D' + d_t^*)^2}{n}\\ &= \sum_{i=1}^{n}(d_i')^2 - \frac{(\operatorname{Tr} D')^2}{n} + d_t^* + 2\sum_{i=n-d_t^*+1}^{n} d_i' - \frac{2d_t^* \operatorname{Tr} D' + (d_t^*)^2}{n}\\ &= \|D'\| + d_t^* + 2d_t^*(t-1) - \frac{2d_t^* \operatorname{Tr} D' + (d_t^*)^2}{n}.\end{aligned}$$

Let $A = w(D) \in \operatorname{ab}(D)$ with $w = (w_2, \ldots, w_t)$ and $A' = w'(D') \in \operatorname{ab}(D')$ where $w' = (w_2, \ldots, w_{t-1})$.

For $t = 1$, the claim is clear, for $\|w(D)\| = \|D\|$ and $w(D)^{\nearrow} = D$ for any $w \in (S_t)^n$. Assume then that for any diagram $Y' = Y(D')$ with at most $t-1$ columns, we have $\|A'\| \leqslant \|D'\|$ for $A' \in \operatorname{ab}(D')$, and $\|A'\| = \|D'\|$ if and only if $(A')^{\nearrow} = D'$. Let $Y = Y(D)$ have t columns. Let $A \in \operatorname{ab}(D)$ be described by the number a_i of boxes in the i-th row for $1 \leqslant i \leqslant n$, and let $A' = (a_1', \ldots, a_n')$ be the restriction of A to the $t-1$ first columns. Let again $\mathcal{J} = \{i \mid a_i = a_i' + 1\}$. Then one has

$$\begin{aligned}\|A\| &= \sum_{i=1}^{n} a_i^2 - \frac{(\operatorname{Tr} A)^2}{n} = \sum_{i\in\mathcal{J}}(a_i'+1)^2 + \sum_{i\notin\mathcal{J}}^{n}(a_i')^2 - \frac{(\operatorname{Tr} A' + d_t^*)^2}{n}\\ &= \sum_{i=1}^{n}(a_i')^2 - \frac{(\operatorname{Tr} A')^2}{n} + |\mathcal{J}| + 2\sum_{i\in\mathcal{J}}^{n} a_i' - \frac{2d_t^* \operatorname{Tr} A' + (d_t^*)^2}{n}\\ &= \|A'\| + d_t^* - \frac{2d_t^* \operatorname{Tr} A' + (d_t^*)^2}{n} + 2\sum_{i\in\mathcal{J}}^{n} a_i'.\end{aligned}$$

By construction, we have $\operatorname{Tr} A' = \operatorname{Tr} D'$, so, under the induction assumption, we get

$$\|A\| \leqslant \|D'\| + d_t^* - \frac{2d_t^* \operatorname{Tr} D' + (d_t^*)^2}{n} + 2\sum_{i\in\mathcal{J}}^{n} a_i' = \|D\| + 2\left(\sum_{i\in\mathcal{J}} a_i' - d_t^*(t-1)\right).$$

Since $a_i' \leqslant t-1 = \max D'$ holds by construction, and $|\mathcal{J}| = d_t^*$, we get $\|A\| \leqslant \|D\|$. Moreover, $\|A\| = \|D\|$ can only happen when $a_i' = t-1$ for $i \in \mathcal{J}$. In this case,

we get

$$\|A'\| = \|D\| - \left(d_t^* + 2d_t^*(t-1) - \frac{2d_t^* \operatorname{Tr} D' + (d_t^*)^2}{n} \right) = \|D'\|.$$

By the induction assumption, we have $(A')^{\nearrow} = D'$, and by the definition of $\mathcal{J}$, we get $a_i = a'_i + 1 = d = \max D$ for $i \in \mathcal{J}$. Therefore, we get d_t^* elements in A which are equal to d, and hence $A^{\nearrow} = D$. □

3.2.3. *Local criteria for BG trivializations.* In this section, we use the fact that the type $T(\Lambda)$ is an element of $\mathrm{ab}(\mathbf{ED}_\Lambda(M))$ for any trivial $M \in \mathbf{\Lambda}^0$ to derive local criteria satisfied by the BG trivializations. Let $d(\Lambda, \mathbf{\Lambda}^0) = \min_{M \in \mathbf{\Lambda}^0} d(\Lambda, M)$.

Lemma 36. *Let $\Lambda \in \mathbf{\Lambda}$ be a lattice. For any $M \in \mathbf{\Lambda}^0$, we have*

$$d(\Lambda, M) = d(\Lambda, \mathbf{\Lambda}^0) \Longleftrightarrow v_\Lambda(M) = \max_{\tilde{M} \in \mathbf{\Lambda}^0} v_\Lambda(\tilde{M}) \text{ and } v_M(\Lambda) = \max_{\tilde{M} \in \mathbf{\Lambda}^0} v_{\tilde{M}}(\Lambda).$$

If $M \in \mathrm{BG}(\Lambda)$, then $\delta(\Lambda, M) = \min_{\tilde{M} \in \mathbf{\Lambda}^0} \delta(\Lambda, \tilde{M})$ for $\delta \in \{d, \boldsymbol{d}, \ell\}$.

Proof. Let $\mathbf{\Lambda}^0 \ni \tilde{M} \xrightarrow{z^D} \Lambda$ represent the elementary divisors of an arbitrary trivialization $\tilde{M}$ of Λ. The BG algorithm of Section 3.2.1 can be applied to $\tilde{M}$ to obtain $M \in \mathrm{BG}(\Lambda)$, as in the following scheme, where $A = w(D) \in \mathrm{ab}(D)$:

$$\mathbf{\Lambda}^0 \ni \tilde{M} \xleftarrow{z^D} \Lambda \xrightarrow{z^A} M \in \mathrm{BG}(\Lambda) \ .$$

By definition, we have $T(\Lambda) \cong A$. By Lemma 35, for any $\tilde{M} \in \mathbf{\Lambda}^0$, we have $v_\Lambda(\tilde{M}) = \min D \leqslant \min A = v_\Lambda(M)$ and $v_{\tilde{M}}(\Lambda) = -\max D \leqslant -\max A = v_M(\Lambda)$. This holds for any trivial $\tilde{M}$; hence, for any BG trivialization $M \in \mathrm{BG}(\Lambda)$, we get

$$v_\Lambda(M) = \max_{\tilde{M} \in \mathbf{\Lambda}^0} v_\Lambda(\tilde{M}) \quad \text{and} \quad v_M(\Lambda) = \max_{\tilde{M} \in \mathbf{\Lambda}^0} (v_{\tilde{M}}(\Lambda)).$$

On the other hand, we have, for any $\tilde{M} \in \mathbf{\Lambda}^0$,

$$d(\Lambda, \tilde{M}) = -v_\Lambda(\tilde{M}) - v_{\tilde{M}}(\Lambda) \geqslant -v_\Lambda(M) - v_M(\Lambda) = d(\Lambda, M) \text{ for } M \in \mathrm{BG}(\Lambda).$$

The rest in a direct consequence of Lemma 35. □

Proposition 37. *Let $\Lambda \in \mathbf{\Lambda}$. Then*

$$T(\Lambda) = \max_{\preceq_{\mathrm{lex}}} \{\mathbf{ED}_\Lambda(M)^{\nearrow} \mid M \in \mathbf{\Lambda}^0\} = \min_{\preceq_{\mathrm{lexrev}}} \{\mathbf{ED}_\Lambda(M)^{\nearrow} \mid M \in \mathbf{\Lambda}^0\}.$$

Proof. Let $M \in \mathbf{\Lambda}^0$, and let $D = (d_1, \ldots, d_n) = \mathbf{ED}_\Lambda(M)^{\nearrow}$. By Theorem 33, there exists $w = (w_1, \ldots, w_t) \in (S_n)^t$ such that $w(D) \cong T(\Lambda)$, and we can, by Lemma 34, ensure that $w(D) = T(\Lambda)^{\nearrow}$. Representing D as its Young diagram $Y(D)$, the sequence $T(\Lambda)$ is given by an abacus diagram $A \in \mathrm{ab}(D)$ derived from $Y(D)$ by moving boxes vertically (downwards), and such that the number

of boxes t_i in row i weakly increases with the row index (i.e., $t_{i+1} \geqslant t_i$). Then we have $d_1 = v_\Lambda(M) \leqslant t_1 = \max_{\tilde{M} \in \mathbf{\Lambda}^0} v_\Lambda(\tilde{M})$. Suppose that $d_1 = t_1$, and assume that $d_2 > t_2$ holds. Accordingly some boxes from the second row must be lowered. Therefore, we necessarily have $t_1 > d_1$. This contradicts the maximality assumption on $v_\Lambda(M)$. Thus $d_2 \leqslant t_2$, and so $t_2 = \max_{M \in \mathbf{\Lambda}^0, v_\Lambda(M) = t_1} d_2^\Lambda(M)$. Let $i = \max\{j \mid d_k = t_k \text{ for } k \leqslant j\}$ and assume that $d_{i+1} > t_{i+1}$ holds. The same argument holds and shows that $d_{i+1} \leqslant t_{i+1}$. Therefore we have proved that $t_i = \max_{M \in \mathbf{\Lambda}^0, d_k^\Lambda(M) = t_k, 1 \leqslant k \leqslant i-1} d_i^\Lambda(M)$. A similar argument starting from $d_n = -\max v_M(\Lambda)$ proves the second relation. □

This result establishes the first part of Theorem 25. The proof will be complete when we prove that M is indeed a BG trivialization of Λ if and only if $M \in \mathbf{\Lambda}^0$ and $\mathbf{ED}_\Lambda(M) \cong T(\Lambda)$. This will be established in the next section.

3.3. ***The permutation lemma.*** In the previous section, we showed how to construct a lattice M whose global form Y_M contains a Smith basis for a given lattice Λ. The following result allows us to give a fairly complete geometric view of the set of BG trivializations. We recall that a *principal minor* of a matrix $A \in \mathfrak{gl}_n(\mathbb{C})$ is a minor $A_{I,I}$ where $I \subset [n]$ obtained by deleting rows and columns whose indices are not elements of I.

Proposition 38 (permutation lemma). *Let $D = (d_1, \ldots, d_n) \in \mathbb{Z}^n$ be an integer sequence and $P \in \mathrm{GL}_n(\mathbb{C}[\![t]\!])$ a lattice gauge.*

(1) (Bolibrukh) *There exist a permutation $\tau \in S_n$ and a lattice gauge $\tilde{P} \in \mathrm{GL}_n(\mathbb{C}[\![t]\!])$ such that*

$$\Pi = t^{-D} P^{-1} t^{\tau D} \tilde{P} \in \mathrm{GL}_n(\mathbb{C}[t^{-1}]),$$

where $t^D = \mathrm{diag}(t^{d_1}, \ldots, t^{d_n})$ and $\tau D = (d_{\tau(1)}, \ldots, d_{\tau(n)})$.

(2) *There exists moreover a lattice gauge $Q \in \mathrm{GL}_n(\mathbb{C})$ such that $t^D \Pi = Q t^D$.*

(3) *Furthermore, one can choose $\sigma = 1$ in item (1) if and only if all principal minors of $P(0)$ indexed by the elements of the ascending flag $\boldsymbol{D}^\nearrow$ are nonzero.*

We will give a self-contained proof of this result, following for the first item, due to Bolibrukh, the same lines as the proof of this lemma given in [Ilyashenko and Yakovenko 2007]. Item (2) is, up to our knowledge, new, as well as the necessity statement in (3) (sufficiency appears in the cited work). The proof proceeds by induction, using the following simple lemma.

Lemma 39. *Let $d \leqslant n$ and*

$$T = \begin{pmatrix} I_d & 0 \\ 0 & 0_{n-d} \end{pmatrix}.$$

Let $H = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \mathrm{GL}_n(\mathbb{C}[\![t]\!])$ *be a lattice gauge matrix, decomposed as a* 2×2*-block matrix according to the blocks of* T*. If* $\det A(0) \neq 0$*, then there exists a monopole gauge matrix*

$$\Pi = \begin{pmatrix} I_d & t^{-1}\widetilde{\Pi} \\ 0 & I_{n-d} \end{pmatrix}$$

with $\widetilde{\Pi}$ *a constant matrix, such that* $\widetilde{H} = t^{-T} H t^T \Pi$ *is a lattice gauge matrix; that is,* $\widetilde{H} \in \mathrm{GL}_n(\mathbb{C})$.

Proof. For simplicity, put $M_0 = M(0)$ for a holomorphic matrix M. One checks that putting $\widetilde{\Pi} = -A_0^{-1} B_0$, we have

$$\widetilde{H} = t^{-T} H t^T \Pi = \begin{pmatrix} A & \widetilde{B} \\ tC & \widetilde{D} \end{pmatrix},$$

where

$$\widetilde{B} = t^{-1}(B + A\widetilde{\Pi}) \quad \text{and} \quad \widetilde{D} = D + C\widetilde{\Pi}.$$

By construction, the residue of $\widetilde{B}$ is equal to $B_0 - A_0 A_0^{-1} B_0 = 0$, and hence $\widetilde{B}$ is holomorphic; therefore $\widetilde{H}$ also is. To check that $\widetilde{H} \in \mathrm{GL}_n(\mathbb{C})$, it is sufficient to check the invertibility of

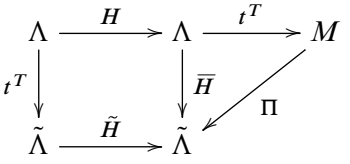
$$\widetilde{H}_0 = \begin{pmatrix} A_0 & \widetilde{B}_0 \\ 0 & D_0 - C_0 A_0^{-1} B_0 \end{pmatrix}.$$

By assumption A_0 is invertible, and it is well-known that the *Schur complement* $D - CA^{-1}B$ is invertible when $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \mathrm{GL}_n(\mathbb{C})$ and A both are so. □

Note that the upper-left block of H appears unchanged in $\widetilde{H}$. Note also that

$$\overline{H} = t^T \Pi = \begin{pmatrix} tI_d & \widetilde{\Pi} \\ 0 & I_{n-d} \end{pmatrix}.$$

Geometrically, we can summarize the construction of Lemma 39 as the following scheme.



We need two technical lemmata before giving the actual proof of the permutation lemma. Let D denote the integer sequence

$$(\underbrace{d_1, \ldots, d_1}_{n_1 \text{ times}}, \ldots, \underbrace{d_s, \ldots, d_s}_{n_s \text{ times}})$$

with $d_i > d_{i+1}$. We say that a matrix H is *strongly D-parabolic* if it has the form

$$H = \begin{pmatrix} t^{d_1} I_{n_1} & \cdots & P_{ij} \\ & t^{d_2} I_{n_2} & \vdots \\ 0 & & t^{d_s} I_{n_s} \end{pmatrix},$$

where P_{ij} is a $n_i \times n_j$ polynomial matrix in t satisfying $\deg P_{ij} < d_i$ and $v(P_{ij}) \geqslant d_j$.

Lemma 40. *Let H be strongly D-parabolic, and let*

$$H' = \begin{pmatrix} tI_m & \widetilde{\Pi} \\ 0 & I_{n-m} \end{pmatrix},$$

where $\widetilde{\Pi}$ is a constant matrix and $m \leqslant n_1$. Then the product HH' is strongly D'-parabolic, where $D' = (\underbrace{d_1+1, \ldots, d_1+1}_{m \text{ times}}, \underbrace{d_1, \ldots, d_1}_{n_1 - m \text{ times}}, \ldots, \underbrace{d_s, \ldots, d_s}_{n_s \text{ times}})$.

Proof. Let $\bar{D} = (\underbrace{d_2, \ldots, d_2}_{n_2 \text{ times}}, \ldots, \underbrace{d_s, \ldots, d_s}_{n_s \text{ times}})$. The matrix H can be written as

$$H = \begin{pmatrix} t^{d_1} I_{n_1} & P \\ 0 & \overline{H} \end{pmatrix},$$

where $\overline{H}$ is strongly $\bar{D}$-parabolic, and $P = (P_2 \ \cdots \ P_s)$ where the blocks P_i satisfy $\deg P_i < d_1$ and $v(P_i) \geqslant d_i$. Then, if $m = n_1$, the product HH' is simply

$$HH' = \begin{pmatrix} t^{d_1+1} I_{n_1} & t^{d_1} \widetilde{\Pi} + P \\ 0 & \overline{H} \end{pmatrix}.$$

Otherwise, we split the matrices in 3×3-blocks, as

$$\begin{aligned} HH' &= \begin{pmatrix} t^m I_m & 0 & P_1 \\ 0 & t^{d_1} I_{n_1 - m} & P_2 \\ 0 & 0 & \overline{H} \end{pmatrix} \begin{pmatrix} tI_m & \widetilde{\Pi}_1 & \widetilde{\Pi}_2 \\ 0 & I_{n_1 - m} & 0 \\ 0 & 0 & I_{n-n_1} \end{pmatrix} \\ &= \begin{pmatrix} t^{d_1+1} I_m & t^{d_1} \widetilde{\Pi}_1 & t^{d_1} \widetilde{\Pi}_1 + P_1 \\ 0 & t^{d_1} I_{n_1 - m} & P_2 \\ 0 & 0 & \overline{H} \end{pmatrix}. \end{aligned}$$

In both cases, we see that the product HH' is strongly D'-parabolic as requested. □

In the following lemma, we prove that the factorization in Proposition 38 exists if and only if the condition on the minors of the constant term holds.

Lemma 41. *Let $P, Q \in \mathrm{GL}_n(\mathbb{C})$, and let*

$$D = (\underbrace{d_1, \ldots, d_1}_{n_1 \text{ times}}, \ldots, \underbrace{d_s, \ldots, d_s}_{n_s \text{ times}})$$

with $d_i > d_{i+1}$. Assume that $\Pi = z^{-D} P z^D Q \in \mathfrak{gl}_n(\mathbb{C}[\![1/z]\!])$. Decomposing a matrix M in blocks $M_{i,j}$ according to the multiplicities $(n_1, \dots, n_s)$ of D, let

$$P_i = \begin{pmatrix} P_{i,i} & \cdots & P_{i,s} \\ \vdots & \ddots & \vdots \\ P_{s,i} & \cdots & P_{s,s} \end{pmatrix} \quad \textit{and} \quad \widetilde{\Pi}_i = \begin{pmatrix} \Pi_{i,1}^{(0)} & \cdots & \Pi_{i,s}^{(0)} \\ \vdots & \ddots & \vdots \\ \Pi_{s,1}^{(0)} & \cdots & \Pi_{s,s}^{(0)} \end{pmatrix}.$$

Then $\widetilde{\Pi}_i$ has maximal rank if and only if $P_i(0), \dots, P_s(0)$ are all invertible.

Proof. We prove this result by induction. We put $M = \sum_{\ell \in \mathbb{Z}} M^{(\ell)} z^\ell$ for a formal series matrix, with $M^{(\ell)} = 0$ when needed. Let $M_{i,(j:k)} = (M_{i,j} \ \cdots \ M_{i,k})$ for $j \leqslant k$, and $M_{i,\bullet} = M_{i,(1:s)}$, and put further

$$\widetilde{M}_j = \begin{pmatrix} M_{j,\bullet}^{(0)} \\ M_{j+1,\bullet}^{(0)} \\ \vdots \\ M_{s,\bullet}^{(0)} \end{pmatrix} \quad \text{and} \quad M_j'^{(\ell)} = \begin{pmatrix} M_{j,\bullet}^{(\ell)} \\ M_{j+1,\bullet}^{(\ell+d_j-d_{j+1})} \\ \vdots \\ M_{s,\bullet}^{(\ell+d_j-d_s)} \end{pmatrix},$$

so that

$$M_j'^{(\ell)} = \begin{pmatrix} M_{j,\bullet}^{(\ell)} \\ M_{j+1}'^{(\ell+d_j-d_{j+1})} \end{pmatrix}.$$

Let $\check{P} = z^{-D} P z^D$. By construction, we have $\check{P} = (\check{P}_{i,j})$ where

$$\check{P}_{i,j} = P_{i,j} z^{d_j - d_i} = \sum_{\ell \geqslant 0} P_{i,j}^{(\ell)} z^{\ell + d_j - d_i} = \sum_{\ell \geqslant d_j - d_i} P_{i,j}^{(\ell + d_i - d_j)} z^\ell.$$

Since $\Pi = \check{P} Q$ holds, we have $\Pi_{i,\bullet} = \sum_{i=1}^s \check{P}_{i,k} Q_{k,\bullet}$, with $Q_{k,\bullet} = \sum_{\ell \geqslant 0} Q_{k,\bullet}^{(\ell)} z^\ell$. After some algebra, we get

$$\Pi_{i,\bullet} = \sum_{t \leqslant 0} \left(\sum_{k=1}^{s} \sum_{\ell = d_k - d_i}^{t} P_{i,k}^{(\ell + d_i - d_k)} Q_{k,\bullet}^{(t-\ell)} \right) z^t \quad \text{since } \Pi \in \mathfrak{gl}_n(\mathbb{C}[\![1/z]\!]). \tag{12}$$

We will establish the main claim (MC) of the lemma by proving simultaneously the following two additional results. If the assumption of the lemma holds, then we have

(A) $\widetilde{Q}_j$ is in the row span of $\widetilde{\Pi}_j$ for $i \leqslant j \leqslant s$,

(B) for $i \leqslant j \leqslant s$ and $\ell \leqslant d_{j-1} - d_j - 1$, there exists a matrix $X_{j\ell}$ such that $Q_j'^{(\ell)} = X_{j\ell} \widetilde{Q}_j$.

Assume first that $i = s$. According to formula (12), we have

$$\Pi_{s,\bullet} = \sum_{t\leqslant 0}\Bigg(\sum_{k=1}^{s}\sum_{\ell=d_k-d_s}^{t} P_{s,k}^{(\ell+d_s-d_k)} Q_{k,\bullet}^{(t-\ell)}\Bigg) z^t.$$

Since $d_k - d_s > 0$ for $k \neq s$, we have that $\Pi_{s,\bullet}$ is a constant, and the formula reduces to $t = 0$ and $k = s$; thus $\Pi_{s,\bullet}^{(0)} = P_{s,s}^{(0)} Q_{s,\bullet}^{(0)}$. Since $\operatorname{rk} Q_{s,\bullet}^{(0)} = n_s$, we get

$$\operatorname{rk} \Pi_{s,\bullet}^{(0)} = n_s \iff \det P_{s,s}^{(0)} \neq 0,$$

which establishes (MC) and (A) for $i = s$. For $0 < t < d_{s-1} - d_s$, formula (12) reduces to

$$\Pi_{s,\bullet}^{(t)} = \sum_{\ell=0}^{t} P_{s,s}^{(\ell)} Q_{s,\bullet}^{(t-\ell)} = 0$$

since $\Pi^{(t)} = 0$ for $t > 0$. We can arrange all these equations into the large matrix equation

$$\begin{pmatrix} P_{s,s}^{(0)} & 0 & \cdots & 0 \\ P_{s,s}^{(1)} & P_{s,s}^{(0)} & & \vdots \\ \vdots & & \ddots & \vdots \\ P_{s,s}^{(d_{s-1}-d_s-1)} & \cdots & \cdots & P_{s,s}^{(0)} \end{pmatrix} \begin{pmatrix} Q_{s,\bullet}^{(0)} \\ Q_{s,\bullet}^{(1)} \\ \vdots \\ Q_{s,\bullet}^{(d_{s-1}-d_s-1)} \end{pmatrix} = \begin{pmatrix} \Pi_{s,\bullet}^{(0)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

It follows easily that $Q_{s,\bullet}^{(\ell)}$ is then a left multiple of $Q_{s,\bullet}^{(0)}$ for $0 \leqslant \ell \leqslant d_{s-1} - d_s - 1$, and also for $\ell < 0$, since $Q_{s,\bullet}^{(\ell)} = 0$, which establishes the claim (B) for $i = s$.

Assume now that (MC), (A) and (B) hold for $i + 1 \leqslant j \leqslant s$ and $\ell \leqslant d_{j-1} - d_j - 1$. Formula (12) gives

$$\begin{aligned} \Pi_{j,\bullet}^{(t)} &= \sum_{k=1}^{s}\sum_{\ell=d_k-d_j}^{t} P_{j,k}^{(\ell+d_j-d_k)} Q_{k,\bullet}^{(t-\ell)} \\ &= \sum_{\ell\geqslant 0}\sum_{k=1}^{s} P_{j,k}^{(\ell)} Q_{k,\bullet}^{(t+d_j-d_k-\ell)}, \end{aligned}$$

with the convention that $Q^{(\ell)} = 0$ when $\ell < 0$. Considering exponents $t = d_i - d_j + m$ for $i \leqslant j \leqslant s$ and $0 \leqslant m \leqslant d_{i-1} - d_i - 1$, we get

$$\begin{aligned} \Pi_{j,\bullet}^{(d_i-d_j+m)} &= \sum_{\ell\geqslant 0}\sum_{k=i}^{s} P_{j,k}^{(\ell)} Q_{k,\bullet}^{(m+d_i-d_k-\ell)} \\ &= \sum_{\ell\geqslant 0} P_{j,(i:s)}^{(\ell)} Q_i^{\prime(m-\ell)}, \end{aligned}$$

since $k < i$ implies $m + d_i - d_k < 0$. Therefore, we get

$$\begin{aligned}
&\Pi_{j,\bullet}^{(d_i-d_j+m)}\\
&=\sum_{\ell\geqslant 0} P_{j,i}^{(\ell)} Q_{i,\bullet}^{(m-\ell)} + P_{j,(i+1:s)}^{(\ell)} Q_{i+1}^{\prime(m+d_i-d_{i+1}-\ell)}\\
&=\sum_{\ell=0}^{m}\Big(P_{j,i}^{(\ell)} Q_{i,\bullet}^{(m-\ell)} + P_{j,(i+1:s)}^{(\ell)} Q_{i+1}^{\prime(m+d_i-d_{i+1}-\ell)}\Big) + \sum_{\ell>m} P_{j,(i+1:s)}^{(\ell)} Q_{i+1}^{\prime(m+d_i-d_{i+1}-\ell)}\\
&=\sum_{\ell=0}^{m} P_{j,(i:s)}^{(\ell)} Q_{i}^{\prime(m-\ell)} + \sum_{\ell>m} P_{j,(i+1:s)}^{(\ell)} Q_{i+1}^{\prime(m+d_i-d_{i+1}-\ell)}.
\end{aligned}$$

According to the induction $Q_{i+1}^{\prime(m+d_i-d_{i+1}-\ell)}$ is a left multiple of $\widetilde{Q}_{i+1}$ for $\ell > m$, so we have

$$\Pi_{j,\bullet}^{(d_i-d_j+m)} = \sum_{\ell=0}^{m}\big(P_{j,i}^{(\ell)} \;\cdots\; P_{j,s}^{(\ell)}\big) Q_{i}^{\prime(m-\ell)} + [\widetilde{Q}_{i+1}] \quad \text{for } i \leqslant j \leqslant s, \tag{13}$$

where we let, for notational simplicity, $A = B + [Q]$ mean "there exists a matrix X such that $A - B = XQ$". Assume $m = 0$ first. The equation for $j = i$ is then

$$\Pi_{i,\bullet}^{(0)} = P_{i,i}^{(0)} Q_{i,\bullet}^{(0)} + \big(P_{i,i+1}^{(0)} \;\cdots\; P_{i,s}^{(0)}\big) Q_{i+1}^{\prime(d_i-d_{i+1})} + [\widetilde{Q}_{i+1}]. \tag{14}$$

The remaining equations,

$$\Pi_{j,\bullet}^{(d_i-d_j)} = P_{j,i}^{(0)} Q_{i,\bullet}^{(0)} + \big(P_{j,i+1}^{(0)} \;\cdots\; P_{j,s}^{(0)}\big) Q_{i+1}^{\prime(d_i-d_{i+1})} + [\widetilde{Q}_{i+1}] = 0, \quad i+1 \leqslant j \leqslant s,$$

can be rewritten as

$$P_{i+1}^{(0)} Q_{i+1}^{\prime(d_i-d_{i+1})} = -\begin{pmatrix} P_{i+1,i}^{(0)} \\ \vdots \\ P_{s,i}^{(0)} \end{pmatrix} Q_{i,\bullet}^{(0)} + [\widetilde{Q}_{i+1}], \quad \text{where } P_{i+1}^{(0)} = \begin{pmatrix} P_{i+1,i+1}^{(0)} & \cdots & P_{i+1,s}^{(0)} \\ \vdots & \ddots & \vdots \\ P_{s,i+1}^{(0)} & \cdots & P_{s,s}^{(0)} \end{pmatrix}$$

is invertible by the induction assumption. Put

$$B = \big(P_{i,i+1}^{(0)} \;\cdots\; P_{i,s}^{(0)}\big) \quad \text{and} \quad C = \begin{pmatrix} P_{i+1,i}^{(0)} \\ \vdots \\ P_{s,i}^{(0)} \end{pmatrix} \quad \text{so that } P_{i}^{(0)} = \begin{pmatrix} P_{i,i}^{(0)} & B \\ C & P_{i+1}^{(0)} \end{pmatrix}.$$

Thus, we get

$$Q_{i+1}^{\prime(d_i-d_{i+1})} = -\big(P_{i+1}^{(0)}\big)^{-1} C Q_{i,\bullet}^{(0)} + X\widetilde{Q}_{i+1} = X'\widetilde{Q}_i;$$

hence $Q_i'^{(0)}$ is a left multiple of $\widetilde{Q}_i$. Moreover, substituting in (14), we get

$$\Pi_{i,\bullet}^{(0)} = \big(P_{i,i}^{(0)} - B\big(P_{i+1}^{(0)}\big)^{-1}C\big)Q_{i,\bullet}^{(0)} + [\widetilde{Q}_{i+1}]. \tag{15}$$

By assumption $\widetilde{Q}_{i+1}$ is in the row space of

$$\begin{pmatrix} \Pi_{i+1,\bullet}^{(0)} \\ \vdots \\ \Pi_{s,\bullet}^{(0)} \end{pmatrix}.$$

Hence $\operatorname{rk}\Pi_{i,\bullet}^{(0)} = n_i$ if and only if $\operatorname{rk}\big(P_{i,i}^{(0)} - B\big(P_{i+1}^{(0)}\big)^{-1}C\big)Q_{i,\bullet}^{(0)} = n_i$; that is, $\det\big(P_{i,i}^{(0)} - B\big(P_{i+1}^{(0)}\big)^{-1}C\big) \neq 0$. This matrix is the Schur complement of

$$P_i^{(0)} = \begin{pmatrix} P_{i,i}^{(0)} & B \\ C & P_{i+1}^{(0)} \end{pmatrix},$$

which is invertible exactly when $P_{i+1}^{(0)}$ is. Therefore (MC) is established in general, and by (15), $\widetilde{Q}_i$ is in the row span of $\widetilde{\Pi}_i$, so claim (A) is proved. Similarly, for a given $0 \leqslant m \leqslant d_{i-1} - d_i - 1$, we can stack the remaining equations, (13), corresponding to $i \leqslant j \leqslant s$,

$$\Pi_{j,\bullet}^{(d_i - d_j + m)} = \sum_{\ell=0}^{m} P_{j,(i:s)}^{(\ell)} Q_i'^{(m-\ell)} + [\widetilde{Q}_{i+1}] = 0,$$

to get the relation

$$\begin{pmatrix} \Pi_{i,\bullet}^{(d_i - d_j + m)} \\ \vdots \\ \Pi_{s,\bullet}^{(d_i - d_j + m)} \end{pmatrix} = \sum_{\ell=0}^{m} P_i^{(\ell)} Q_i'^{(m-\ell)} = \begin{cases} \widetilde{\Pi}_i & \text{if } m = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Putting for notational simplicity $d = d_{i-1} - d_i$, we finally get

$$\begin{pmatrix} P_i^{(0)} & 0 & \cdots & 0 \\ P_i^{(1)} & P_i^{(0)} & & \vdots \\ \vdots & & \ddots & \vdots \\ P_s^{(d-1)} & \cdots & \cdots & P_i^{(0)} \end{pmatrix} \begin{pmatrix} Q_i'^{(0)} \\ \\ Q_i'^{(1)} \\ \vdots \\ Q_i'^{(d-1)} \end{pmatrix} = \begin{pmatrix} \Pi_i' \\ 0 \\ \vdots \\ 0 \end{pmatrix} + [\widetilde{Q}_{i+1}] \quad \text{where } \Pi_i' = \begin{pmatrix} \widetilde{\Pi}_i \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Accordingly, $Q_i'^{(\ell)}$ is a left multiple of $Q_i'^{(0)}$ for $0 \leqslant \ell \leqslant k-1$. Since we have established that $Q_i'^{(0)}$ is a left multiple of $\widetilde{Q}_i$, this ends the proof of claim (B). □

Proof of Proposition 38. Assume for simplicity that $D = \mathrm{diag}(d_1 I_{n_1}, \dots, d_s I_{n_s})$ is written by blocks, and that $d_1 > d_2 > \cdots > d_s$. Then there exist $m = d_1 - d_s$ matrices $T_1, \dots, T_m$ of the type

$$T_i = \begin{pmatrix} I_{b_i} & 0 \\ 0 & 0_{n-b_i} \end{pmatrix},$$

where every b_i is equal to some $n_1 + \cdots + n_{t_i}$ for some decreasing sequence t_i, such that $D = T_1 + \cdots + T_m$. Secondly, assume that all left-upper square blocks of $P(0)$ of sizes b_i are invertible. Letting $P = H_0$, according to Lemma 39, there exists a sequence of monopole matrices

$$\Pi_i = \begin{pmatrix} I_{b_i} & t^{-1}\widetilde{\Pi}_i \\ 0 & I_{n-b_i} \end{pmatrix}$$

with a constant matrix $\widetilde{\Pi}_i$, and a sequence of lattice gauge transforms $H_i \in \mathrm{GL}_n(\mathbb{O})$ such that

$$H_{i+1} = t^{-T_i} H_i t^{T_i} \Pi_i. \tag{16}$$

Let

$$\overline{H}_i = t^{T_i}\Pi_i = \begin{pmatrix} tI_{b_i} & \widetilde{\Pi}_i \\ 0 & I_{n-b_i} \end{pmatrix}.$$

It follows from Lemma 40 that $\overline{H} = \overline{H}_1 \cdots \overline{H}_m$ is strongly D-parabolic. It follows then, as a remarkable consequence, that the diagonal matrix t^D can be both factored from the matrix $\overline{H}$ on the left as $\overline{H} = t^D \Pi$ with a monopole matrix Π, and simultaneously from the right as $\overline{H} = Q t^D$ with a lattice gauge $Q \in \mathrm{GL}_n(\mathbb{O})$. Since $t^D H_{d+1} = H\overline{H}$ holds, we get on the one hand that $t^{-D} P^{-1} t^D H_{m+1} = \Pi \in \mathrm{GL}_n(\mathbb{C}[t^{-1}])$ as required for the first claim of the permutation lemma. However, and this was not stated in [Bolibrukh 1990] or [Ilyashenko and Yakovenko 2007], we also have the relation $t^D \Pi = Q t^D$, which yields the second claim.

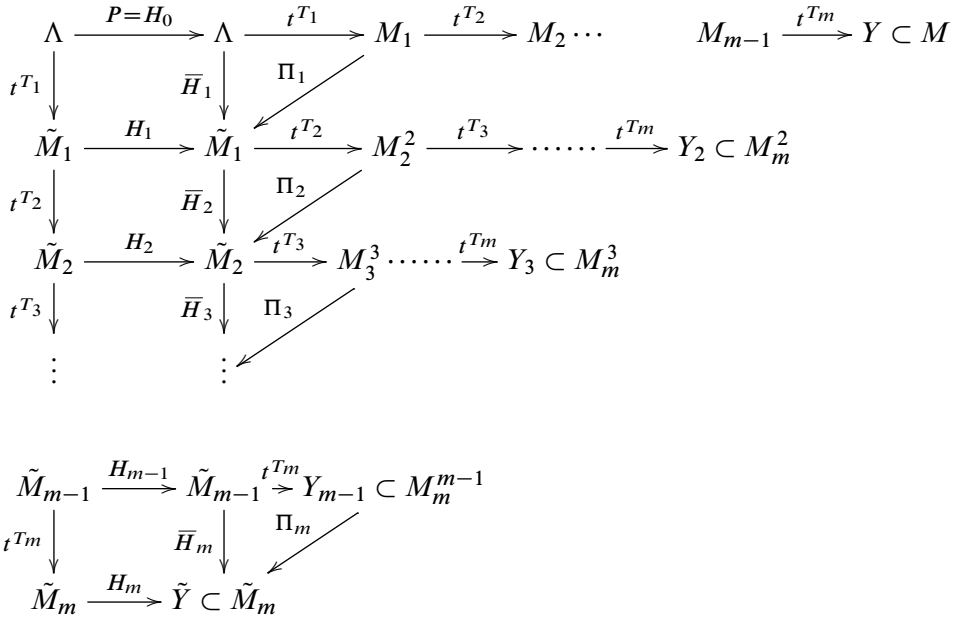
For the third claim, the sufficiency of the minors condition has just been established. For the converse, assume that we have the following diagram, which we have completed with a monopole Π and a lattice gauge Q such that $z^{-D} P z^D Q = \Pi$:

$$\begin{array}{ccccc} \Lambda : (e) & \xrightarrow{z^D} & Y_M \subset M & & \\ \Big\downarrow P & & & \overset{\Pi}{\searrow} & \\ \Lambda : (\varepsilon) & \xrightarrow{z^D} & \tilde{M} & \overset{Q}{\dashrightarrow} & Y_{\tilde{M}} \subset \tilde{M} \end{array}$$

Let $Q = Q_0 + Q_1 z + \cdots \in \mathrm{GL}_n(\mathbb{O})$, and $\Pi = \Pi_0 + \cdots + \Pi_t z^{-t} \in \mathrm{GL}_n(\mathbb{C}[z^{-1}])$. Recall that if Π is a monopole, then one must in particular have $\Pi_0 \in \mathrm{GL}_n(\mathbb{C})$. Then

Lemma 41 implies that such a factorization only exists if the matrix P_0 satisfies the condition on the minors. □

The following scheme sums up this construction.



The first row corresponds to a min-shortest path $\Gamma = (\Lambda, M_1, \ldots, M)$ from Λ to a given BG trivialization M. This path Γ is included in an apartment $\mathcal{B}$, namely the one spanned by a BG basis (e) of Λ corresponding to the trivialization M. By definition, the apartment $\mathcal{B}$ goes through the global form Y of M. The gauge H^{-1} does not map the shortest path Γ onto anything special. However, if we call $\mathcal{A} = H^{-1}(\mathcal{B})$ the image of the apartment spanned by (e), the permutation lemma tells us how to construct a shortest path Γ' in $\mathcal{A}$ whose end point is also a BG trivialization of Λ. Lemma 39 gives the step-by-step modification of the shortest path Γ. Row i of the diagram corresponds indeed to a partial shortest path $\Gamma_i = (\tilde{M}_i, M_i^i, \ldots, M_m^i)$ whose end-point is a BG trivialization of the i-th element $\tilde{M}_i$ of the shortest path $\Gamma' = (\Lambda, \tilde{M}_1, \ldots, \tilde{M}_m)$. Even if the end-point $\tilde{M}_m$ is a BG trivialization of Λ, note that the apartment $\mathcal{A}$ does not contain the global form $\tilde{Y}$ of $\tilde{M}_m$, and that we still need the gauge transform H_m to obtain it.

3.3.1. *Consequences of the permutation lemma.* As stated in [Ilyashenko and Yakovenko 2007], one can assume that $\tau = \mathrm{id}$ if all leading principal minors of P are holomorphically invertible; that is, the corresponding minors of $P(0)$ are nonzero. This condition can always be ensured by a permutation of the columns of P. Actually, as stated in Proposition 38(3), it is sufficient that this condition

holds only for the leading principal minors of orders $n_1, n_1+n_2, \ldots, n_1+\cdots+n_{s-1}$ of $P(0)$. We then say that P *respects the minors condition* with respect to D (or to $\sigma = (n_1, \ldots, n_s)$). Recall the following well-known result.

Lemma 42. *Let $P \in \mathrm{GL}_n(\mathbb{C})$, let (e) be the standard basis of $\mathbb{C}^n$ and (ε) be the column vectors of P. Then P respects the minors condition with respect to a signature σ if and only if (e) is transversal to the flag $\boldsymbol{F}^\sigma(\varepsilon)$.*

Definition 43. Let $\mathscr{A}$ be an apartment induced by the K-frame Φ in V. Consider $\Lambda \in \mathscr{A}$. Let $\boldsymbol{H}$ be the Harder–Narasimhan flag of Λ in $E = \Lambda/\mathfrak{m}\Lambda$. Let $W \subset S_n$ be the parabolic subgroup of S_n associated to $\boldsymbol{H}$, and W' be the set of right cosets $W\backslash S_n$. Let (e) be a basis of Λ in the frame Φ whose image $(\bar{e})$ in E is transversal to $\boldsymbol{H}$. The integer

$$\iota_\Lambda(\mathscr{A}) = |\{\tau \in W' \mid \tau(\bar{e}) \text{ transversal to } \boldsymbol{H}\}|$$

is independent of (e) and is called the *transversality index* of $\mathscr{A}$ with respect to Λ.

Theorem 44. *Let $\mathscr{E}$ be a holomorphic vector bundle over X, and let $\Lambda = \mathscr{E}_x \in \boldsymbol{\Lambda}$ be its stalk at $x \in X$.*

(i) *For any apartment $\mathscr{A}$ in the Bruhat–Tits building B at x such that $[\Lambda] \in \mathscr{A}$, there exists a BG trivialization of Λ in $\mathscr{A}$.*

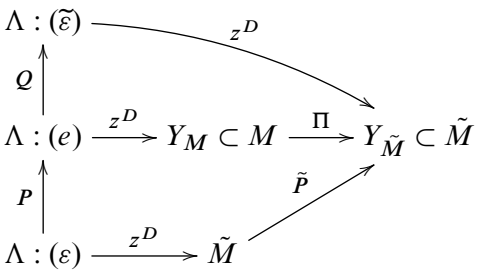
(ii) *More precisely, the number of BG trivializations of Λ in $\mathscr{A}$ is exactly*

$$|\mathscr{A} \cap \mathrm{BG}(\Lambda)| = \iota_\Lambda(\mathscr{A}).$$

Proof. Let (e) be a BG basis of Λ, and $M \in \mathrm{BG}(\Lambda)$. Let (ε) be a basis of the lattice Λ which spans the apartment $\mathscr{A}$. Since $\mathscr{A}$ is invariant under S_n, we can assume that the matrix $P \in \mathrm{GL}_n(\mathbb{O})$ of the basis change from (ε) to (e) has invertible principal leading minors. According to the permutation lemma, there exists a matrix $\tilde{P} \in \mathrm{GL}_n(\mathbb{O})$ such that

$$\Pi = z^{-D} P^{-1} z^D \tilde{P} \in \mathrm{GL}_n(\mathbb{C}[z^{-1}]).$$

The gauge Π sends the basis of global sections $(\sigma) = (z^D e)$ of the BG trivialization of $\mathscr{E}$, given at x by M, into a basis $(\tilde{e})$ of $\tilde{M}$. Since Π is a monopole, the basis $(\tilde{e})$ is also a global basis of sections, but spans another trivializing bundle, namely $\mathscr{F} = \mathscr{E}^{\tilde{M}}$. Therefore the arbitrary apartment $\mathscr{A}$ spanned by (ε) indeed contains a trivial bundle. Now the matrix $\overline{H} = z^D \Pi$ admits a right factorization $\overline{H} = Q z^D$. As a consequence, if we let $(\tilde{\varepsilon})$ be the basis of Λ obtained from (e) by the matrix Q, then $z^D(\tilde{\varepsilon})$ is also a basis of $Y_{\tilde{M}}$. The following scheme sums up the situation.



Therefore the lattice $\tilde{M}$ is also a BG trivialization of Λ.

The monopole gauge Π is block-upper-triangular according to D, and its block matrices Π_{ij} satisfy

$$d_j - d_i \leqslant v(\Pi_{ij}) \leqslant \deg \Pi_{ij} \leqslant 0.$$

This means that Π respects the Harder–Narasimhan filtration HN_Λ of V corresponding to the lattice Λ. Conversely, the lattice gauge Q has a lower D-block-triangular constant term Q_0. Since $z^{-D} P Q z^D \in \mathrm{GL}_n(\mathbb{O})$, the matrix PQ has an upper D-block-triangular constant term $P_0 Q_0$. Therefore, if we put P_D, P_D^- for the pair of opposite D-parabolic standard subgroups of $\mathrm{GL}_n(\mathbb{C})$, the matrix P_0 satisfies $P_0 \in P_D^- P_D$. This means exactly that P satisfies the minors condition with respect to D. If we permute the vectors of (ε) with $\tau \in S_n$ in such a way that $D_{\tau^{-1}} \neq D$, the permutation lemma ensures that $\mathscr{L}(z^D(\varepsilon_\tau)) \neq \tilde{M}$ is again a BG trivialization of Λ. This establishes the second claim of the theorem. □

Corollary 45. *For any apartment $\mathscr{A} \ni \Lambda$, there exists an ordered basis (ε) of $\mathscr{A}$ and a BG basis (e) of Λ such that the gauge P from (e) to (ε) has a lower block-D-triangular unipotent constant term P_0, and that the following picture holds.*

$$\begin{array}{ccc} \Lambda : (\varepsilon) & \xrightarrow{z^D} & M \\ {\scriptstyle P = P_0 + z\tilde{U}} \uparrow & & \uparrow {\scriptstyle P_I = I + zU} \\ \Lambda : (e) & \xrightarrow{z^D} & Y_M \end{array}$$

The number of BG trivializations in $\mathscr{A}$ can hence be computed from a matrix $P_0 \in \mathrm{GL}_n(\mathbb{C})$ with a simple structure:

$$P_0 = \begin{pmatrix} I_{n_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ X_{ij} & \cdots & I_{n_s} \end{pmatrix}$$

as the number of permutations not leaving D invariant whose action on the columns of P_0 preserve the minors condition. The following result completes the proof of Theorem 25.

Corollary 46. *Let* $\Lambda \in \mathbf{\Lambda}$ *and* $T = T(\Lambda)$.

(i) *For any basis* (ε) *of* Λ, *let* $\tilde{M} = \mathscr{L}(z^T(\varepsilon))$. *Then either* $\tilde{M} \in \mathrm{BG}(\Lambda)$ *or* $\tilde{M} \notin \mathbf{\Lambda}^0$.

(ii) *For any flag* $\boldsymbol{F}'$ *in* $\Lambda/\mathfrak{m}\Lambda$ *transversal to* HN_Λ, *and any form* Y *in* Λ, *the lattice* $\tilde{M} = \mathscr{L}_Y(\boldsymbol{F}', T)$ *is a BG trivialization of* Λ.

In particular, if $M \in \mathrm{BG}(\Lambda)$, *then* $\tilde{M} = \mathscr{L}_Y(\boldsymbol{F}^\Lambda(M), T) \in \mathrm{BG}(\Lambda)$ *for any form* Y *in* Λ.

Proof. For (i), if $\tilde{M}$ is trivial, then, by Lemma 41, the gauge from (e) to (ε) satisfies the minors condition; hence $\tilde{M}$ is BG. According to Lemma 27, there exists $M \in \mathrm{BG}(\Lambda)$ such that $\boldsymbol{F}' = \boldsymbol{F}^\Lambda(M)$. Let (e) be a Smith basis of Λ for M. The lattice $\tilde{M}$ is spanned by $z^T(e_Y)$ where (e_Y) is the Y-basis of (e). The gauge P from (e) to (e_Y) has invertible principal minors, since it is tangent to I. Item (ii) follows. □

The permutation lemma is a sort of converse to the Birkhoff–Grothendieck theorem. It can also be seen as a lifting and factorization lemma. For a standard parabolic subgroup $P \subset \mathrm{GL}_n(\mathbb{C})$, let

$$\mathbb{Z}^n_P = \{D \in \mathbb{Z}^n \mid d_i < d_j \Rightarrow A_{ij} = 0 \text{ for all } A \in P\}.$$

Corollary 47. *Let* $A \in \mathrm{GL}_n(\mathcal{O})$. *Let* P, P^- *be a pair of opposed standard parabolic subgroups of* $\mathrm{GL}_n(\mathbb{C})$. *Then we have, for* $w \in S_n$,

$$A_0 \in PwP^- \Leftrightarrow \begin{cases} \forall D \in \mathbb{Z}^n_P,\ \exists \Pi \in \mathrm{GL}_n(\mathbb{C}[z^{-1}]),\ \exists B \in \mathrm{GL}_n(\mathcal{O}) \\ \qquad \text{such that } A = z^D \Pi B z^{-wD}. \end{cases}$$

Moreover, if this holds, then $B \in \mathcal{G}_{-D}$ *and* $\Pi \in \hat{P}$ *hold, where* $\hat{P}$ *stands here for the subgroup of upper-*D*-triangular matrices of* $\mathrm{GL}_n(\mathbb{C}[z^{-1}])$.

This result means that there is a cell decomposition of $\mathrm{GL}_n(\mathcal{O})$ whose cells (defined by the right-hand side of the previous relation) are mapped surjectively (by the canonical surjection $\mathrm{GL}_n(\mathcal{O}) \to \mathrm{GL}_n(\mathbb{C})$) on the Schubert–Bruhat cells of $\mathrm{GL}_n(\mathbb{C})$. Finally, a last consequence of Proposition 38 is that the lattices involved in the Birkhoff–Grothendieck algorithm given in Section 3.2.1 can be taken inside a single apartment.

Corollary 48. *Let* $\Lambda \in \mathbf{\Lambda}$. *For any* $M \in \mathbf{\Lambda}^0$, *let* $\Gamma(M, \Lambda) = (\Lambda_0 = M, \ldots, \Lambda_t = \Lambda)$ *be the* min*-shortest path from* M *to* Λ. *For any apartment* $\mathscr{A} \ni \Lambda, M$, *there exist* $\tilde{M}_1, \ldots, \tilde{M}_t \in \mathscr{A} \cap \mathbf{\Lambda}^0$ *such that* $\tilde{M}_i \in \mathrm{BG}(\Lambda_i)$ *for* $1 \leqslant i \leqslant t$.

Proof. Since $\mathscr{A} \ni \Lambda, M$, we have $\Lambda_i \in \mathscr{A}$ for $1 \leqslant i \leqslant t$. By Theorem 44, any lattice in $\mathscr{A}$ admits a BG trivialization in $\mathscr{A}$. □

4. Local meromorphic connections

Let $\mathrm{Der}_{\mathbb{C}}(K)$ be the K-vector space of dimension one of $\mathbb{C}$-derivations of K and $\Omega = \Omega^1_{\mathbb{C}}(K)$ the dual composed of differentials of K. The valuation v extends naturally to these spaces by the formulæ $v(\vartheta) = v(f)$ and $v(\omega) = v(g)$ if $\vartheta = f\,d/dz$ and $\omega = g\,dz$ for any uniformizing parameter z of K. The space Ω is naturally filtered by the rank-one free $\mathbb{O}$-modules $\Omega(k) = \{\omega \in \Omega \mid v(\omega) \geqslant -k\}$.

Let V be a K-vector space of finite dimension n and let $\Omega(V) = V \otimes_K \Omega^1_{\mathbb{C}}(K)$. We fix a *meromorphic connection* ∇ on V. This is an additive map $\nabla : V \to \Omega(V)$ satisfying the Leibniz rule

$$\nabla(f\,v) = v \otimes df + f \nabla v \quad \text{for all } f \in K \text{ and all } v \in V.$$

For any basis $(e) = (e_1, \ldots, e_n)$ of V, the *matrix* $\mathrm{Mat}(\nabla, (e))$ *of the connection* ∇ *in the basis* (e) is the matrix $A = (A_{ij}) \in \mathrm{M}_n(\Omega)$ such that

$$\nabla e_j = -\sum_{i=1}^{n} e_i \otimes A_{ij} \quad \text{for all } j = 1, \ldots, n.$$

If the matrix $P = \mathrm{Mat}(\mathrm{id}_V, (\varepsilon), (e)) \in \mathrm{GL}_n(K)$ is the basis change from (e) to any other basis (ε), then the matrix of ∇ in (ε) is given by the *gauge transform* of A:

$$A_{[P]} = P^{-1} A P - P^{-1} dP. \tag{17}$$

For any derivation $\tau \in \mathrm{Der}(K/\mathbb{C})$, the contraction of ∇ with τ induces a differential operator ∇_θ on V. The connection ∇ is *regular* whenever the set of *logarithmic* lattices

$$\mathbf{\Lambda}_{\log} = \{\Lambda \in \mathbf{\Lambda} \mid \nabla(\Lambda) \subset \Lambda \otimes_{\mathbb{O}} \Omega(1)\}$$

is nonempty. For any logarithmic lattice $\Lambda \in \mathbf{\Lambda}_{\log}$, the connection ∇ induces a well-defined residue endomorphism $\mathrm{Res}_\Lambda \nabla \in \mathrm{End}_{\mathbb{C}}(\Lambda/\mathfrak{m}\Lambda)$. Note that, since the set $\mathbf{\Lambda}_{\log}$ is closed under homothety and module sums and intersections [Corel 2004, Lemma 2.5], it induces a path-convex subset of the Bruhat–Tits building: if $L, L' \in \mathbf{\Lambda}_{\log}$, then every shortest path between L and L' is a subset of $\mathbf{\Lambda}_{\log}$. This applies in particular to both $\Gamma_{\max}(L, L')$ and $\Gamma_{\min}(L, L')$.

4.1. ***The Deligne lattice.*** As is well known, the choice of a matrix logarithm of the monodromy corresponds to fixing a special lattice in the space V. More precisely, let $V^\nabla \subset V \otimes_K H$ be the $\mathbb{C}$-vector space of horizontal sections on any Picard–Vessiot extension H of K. Let $g = g_s g_u \in \mathrm{End}(V^\nabla)$ be the multiplicative Jordan decomposition of the corresponding local monodromy map. Then the logarithm of the unipotent part g_u is canonically defined (by the Taylor expansion formula for $\log(1 + x)$), but there are several ways to define the logarithm of the semisimple

part g_s. Namely, one must fix a branch of the complex logarithm for every *distinct* eigenvalue of g_s.

A well known result (variously attributed to Deligne, Manin, ...) says that this choice uniquely defines a lattice in V. In Deligne's terms, for any section σ of $\mathbb{C} \to \mathbb{C}/\mathbb{Z}$, there is a unique logarithmic lattice Δ_σ such that the eigenvalues of the residue map $\mathrm{Res}_{\Delta_\sigma} \nabla$ are in the image $\mathrm{Im}\,\sigma$ of σ. As a habit, one usually takes $\mathrm{Re}(\mathrm{Im}\,\sigma) \subset [0, 1[$. In fact, such a habit is not as arbitrary as it seems.

Proposition 49. *Assume that the connection* ∇ *admits an* apparent singularity (*i.e.*, *the monodromy map is trivial*). *Then the matrix* $\mathrm{Mat}(\nabla, (e))$ *is* holomorphic *if and only if the lattice spanned by* (e) *is equal to the Deligne lattice* Δ *attached to* $\mathrm{Re}(\mathrm{Im}\,\sigma) \subset [0, 1[$.

Proof. Since the monodromy map is trivial, its normalized logarithm with respect to Δ is 0. Hence, there is a basis of Δ where the connection has matrix 0. In any other basis (e) of Δ, the connection has matrix $A = P^{-1}dP$ with $P \in \mathrm{GL}_n(\mathbb{O})$, which is holomorphic. Let M be another lattice, and let (e) be a Smith basis of Δ for M. Then the matrix in a basis of M is given by the gauge equation

$$\widetilde{A} = z^{-D} A z^{D} - z^{-D} d(z^{D}) = (A_{ij} z^{k_j - k_i}) - D\frac{dz}{z}.$$

The nonzero diagonal terms of the matrix D of elementary divisors of M give necessarily rise to a pole of order 1 in $\widetilde{A}$. Therefore, Δ is the only lattice where the connection has a holomorphic matrix. □

As a result, we will call Δ *the* Deligne lattice of V.

4.1.1. *Birkhoff forms.* According to a very classical result (see, e.g., [Gantmacher 1959, p. 150]), if

$$\Omega = \mathrm{Mat}(\nabla, (e)) = \sum_{k \geqslant 0} A_k z^k \frac{dz}{z}$$

is the series expansion in z of the matrix of ∇ in a basis (e) of Δ, the gauge $P = \sum_{k \geqslant 0} P_k z^k \in \mathrm{GL}_n(\mathbb{O})$ defined recursively by

$$\begin{cases} P_0 = I, \\ P_k = \Phi^{-1}_{A_0, A_0 - kI}(Q_k), \quad \text{where } Q_k = \sum_{i=1}^{k} A_i P_{k-i}, \end{cases} \tag{18}$$

transforms Ω into $A_0 dz/z$. Here we put $\Phi_{U,V}(X) = XU - VX$. Recall that the map $\Phi_{U,V}$ is an automorphism of $\mathfrak{gl}(\mathbb{C})$ when the spectra of U and V are disjoint. The gauge P thus defined is uniquely determined; moreover, the set of bases where ∇ has matrix $L\frac{dz}{z}$ where $L \in \mathrm{M}_n(\mathbb{C})$ is a constant matrix spans a form Υ_z of Δ, that we call the *Birkhoff form* of the Deligne lattice Δ. The gauge transform P sends in fact the basis (e) to its Υ_z-basis, that we denote here for simplicity by (e_z).

As it results from the proof of Proposition 49, when the singularity is apparent, the Birkhoff form is uniquely defined. Otherwise, however, the form Υ_z depends on the choice of the local coordinate z. Two Birkhoff forms are nevertheless canonically isomorphic.

Lemma 50. *Let z, t be two local coordinates, and let $\alpha \in \mathbb{C}^*$ such that $z = \alpha t$. Let P_z and P_t be the gauge transforms that send (e) to (e_z) and (e_t), respectively. There is a* unique *gauge transform $\widetilde{P}$ that sends (e_z) to (e_t).*

Proof. One has $\frac{dz}{z} = u\frac{dt}{t}$ with $u = 1 + \frac{\theta_t \alpha}{\alpha}$ where $\theta_t = t\frac{d}{dt}$. Put $u = \sum_{i=0}^{\infty} u_i t^i$. Accordingly, the matrix of the connection in (e_z) satisfies

$$\mathrm{Mat}(\nabla, (e_z)) = A_0 \frac{dz}{z} = A_0 \left(\sum_{i=0}^{\infty} u_i t^i \right) \frac{dt}{t}.$$

There exists therefore a uniquely defined gauge transform $\widetilde{P} = \sum_{i=0}^{\infty} \widetilde{P}_i t^i$ that transforms the expression $A_0 dz/z$ into $A_0 dt/t$, as explained in the following scheme.

$$\begin{array}{ccc} \Omega & \xrightarrow{P_z} & A_0 \frac{dz}{z} = \sum_{i=0}^{\infty} u_i A_0 t^i \frac{dt}{t} \\ {\scriptstyle P_t}\downarrow & \swarrow_{\widetilde{P}} & \\ A_0 \frac{dt}{t} & & \end{array}$$

The matrix series $\widetilde{P}$ is determined recursively by the equations (18) applied to the series $\sum_{i=0}^{\infty} A_0 u_i t^i$. The coefficients $\widetilde{P}_i$ are even *polynomials in* A_0, defined by the induction rule

$$\widetilde{P}_0 = I, \quad \widetilde{P}_k = \frac{1}{k} \sum_{i=1}^{k} u_i A_0 \widetilde{P}_{k-i}. \qquad \square$$

4.2. ***Logarithmic lattices and stable flags.*** When two lattices Λ, M are adjacent, all the relevant information on M can be retrieved from the quotient $M/\mathfrak{m}\Lambda$. This is also true in presence of a connection.

Lemma 51. *Let $\Lambda \in \mathbf{\Lambda}_{\log}$ be a logarithmic lattice. For any adjacent lattice $M \in [\mathfrak{m}\Lambda, \Lambda]$, we have $M \in \mathbf{\Lambda}_{\log}$ if and only if $M/\mathfrak{m}\Lambda$ is $\mathrm{Res}_\Lambda \nabla$-stable.*

Proof. In any basis (e) of Λ such that the images of the first $m = \dim W$ vectors span $W = M/\mathfrak{m}\Lambda$, the connection matrix $\Omega = \mathrm{Mat}(\nabla, (e))$ has a residue of the form $\left(\begin{smallmatrix} A & B \\ 0 & C \end{smallmatrix}\right) \in \mathrm{M}_n(\mathbb{C})$, where $A \in \mathrm{M}_m(\mathbb{C})$. Putting $T = \mathrm{diag}(0_m, I_{n-m})$, the basis $(\varepsilon) = z^T(e)$ spans M. It is then straightforward that the matrix $z^{-T}\Omega z^T - T\frac{dz}{z}$ of ∇ in (ε) has a simple pole. $\square$

When the lattices are further apart, this correspondence fails. However, there is also a complete description of the logarithmic lattices as follows. Let Δ be

the Deligne lattice, and let $\delta_\Delta = \mathrm{Res}_\Delta \nabla$ be the residue $\mathbb{C}$-endomorphism on $\mathbf{D} = \Delta/\mathfrak{m}\Delta$. Let Υ be the Birkhoff form of Δ attached to a uniformizing parameter z. Logarithmic lattices can then be characterized as stable flags (as already remarked in [Sabbah 2002, Theorem III.1.1]).

Proposition 52. *The set $\mathbf{\Lambda}_{\log}$ of logarithmic lattices is in bijection with the subset $\Xi_0(\Upsilon)$ of filtrations of Υ which are stable under the residue, namely*

$$\Xi_0(\Upsilon) = \{(\boldsymbol{F}, D) \in \Xi(\Upsilon) \mid \boldsymbol{F} \in \mathfrak{Fl}_{\delta_\Delta}\}.$$

Proof. According to a classical, although not so well known, result (which can be found for instance in [Babbitt and Varadarajan 1983; Bolibrukh 1990]), a lattice $\Lambda \in \mathbf{\Lambda}$ is logarithmic if and only if

(i) there exists a basis (e) of Υ such that $(z^D e)$ is a basis of Λ, with $D = \mathbf{E}\mathbf{D}_\Delta(\Lambda)$,

(ii) $z^{-D} L z^D \in \mathrm{M}_n(\mathbb{O})$, where $L = \mathrm{Mat}(\nabla_{z\frac{d}{dz}}, (e))$.

It results from (ii) that in this case, the matrix L is D-parabolic. Since the flag $\boldsymbol{F}^\Delta(\Lambda)$ induced by Λ on $\mathbf{D} = \Delta/\mathfrak{m}\Delta$ is spanned by the images of the basis (e) in $\mathbf{D}$, it is stable under δ_Δ. Conversely, it is simply a matter of computation to show that any lattice in the Υ-fiber of a δ_Δ-stable flag of $\mathbf{D}$ is logarithmic. □

A difference between our result and Sabbah's is that he only states this result as an equivalence of categories between the set of stable filtrations of $\mathbf{D}$ and the logarithmic lattices, whereas we give the explicit correspondence based on the lifting of $\mathbf{D}$ to a Birkhoff form. Although it would seem that the previous result has little value to effectively determine all logarithmic lattices, it is always possible to determine them in finite terms.

Lemma 53. *Let $M \in \mathbf{\Lambda}_{\log}$ and let $(\boldsymbol{F}, D) = \Pi_\Delta(M)$. Let Y be a form of Δ, and let (e) be a basis of Y respecting the flag $\boldsymbol{F}$. Fix a coordinate z, and let $P = I + P_1 z + \cdots$ be the gauge from (e) to its Υ_z-basis (e_z). Then the Laurent polynomial gauge $Q \in \mathfrak{gl}(\mathbb{C}[z, z^{-1}])$ defined by*

$$Q = (I + \cdots + P_{d-1} z^{d-1}) z^D, \quad \text{where } d = d(\Delta, M),$$

sends the basis (e) of Δ to a basis of M.

Proof. This is an almost direct consequence of Lemma 15. □

Note that the polynomial gauge Q can be explicitly computed from formula (18). On the other hand, one can also explicitly describe the set $\Xi_0(\Upsilon)$. For a linear map $f \in \mathrm{End}(\mathbb{C}^n)$, say that an apartment $\mathscr{A}_\Phi$ is a *diagonalizing apartment* of f if the frame Φ is composed of eigenlines of f.

Lemma 54. *Let $\delta_\Delta = \mathfrak{d} + \mathfrak{n}$ be the additive Jordan decomposition of the residue map $\delta_\Delta = \mathrm{Res}_\Delta \nabla$. The pair $(\boldsymbol{F}, D) \in \Xi(\Upsilon)$ is an element of $\Xi_0(\Upsilon)$ if and only if $\boldsymbol{F}$ admits a complete flag refinement $\hat{\boldsymbol{F}}$ such $\hat{\boldsymbol{F}} \in \mathfrak{Fl}_\mathfrak{n}$ and there is a diagonalizing apartment $\mathscr{A}$ for $\mathfrak{d}$ that respects the flag $\hat{\boldsymbol{F}}$.*

Proof. $\boldsymbol{F}$ is δ-stable if and only if it is stable under both $\mathfrak{d}$ and $\mathfrak{n}$. It is known that $\boldsymbol{F}$ is stable under $\mathfrak{d}$ if and only if every component F_i of $\boldsymbol{F}$ is a direct sum of $\mathfrak{d}$-stable lines, and under $\mathfrak{n}$ if and only if it admits a complete flag refinement $\hat{\boldsymbol{F}} \in \mathfrak{Fl}_\mathfrak{n}$. □

5. The Riemann–Hilbert problem

This problem is by now very well-known, so we will just state the necessary notations and definitions, and refer to the classical paper of Bolibrukh [1990] and to the account he gives of the construction of the Deligne bundle (see [Sabbah 2002; Ilyashenko and Yakovenko 2007] and also [André and Baldassarri 2001] for a purely algebraic construction).

Let $\mathcal{S} = \{s_1, \dots, s_p\}$ be a prescribed set of singular points, $z_0 \notin \mathcal{S}$ be an arbitrary base point, and let χ denote a representation

$$\chi : \pi_1(X \backslash \mathcal{S}, z_0) \to \mathrm{GL}_n(\mathbb{C}). \tag{19}$$

The Riemann–Hilbert problem asks informally for a linear differential system having χ as monodromy representation. In the terms used in this paper, it asks for a regular meromorphic connection ∇ with singular set $\mathcal{S}$ and monodromy χ on a holomorphic vector bundle $\mathscr{E}$. If the bundle is required to be logarithmic with respect to ∇ one speaks of a *weak solution* to RH. In its strongest form, the Riemann–Hilbert problem asks for a differential system $Y' = A(z)Y$ having simple poles on $\mathcal{S}$ as only singularities, and whose monodromy representation is globally conjugate to χ. This amounts to asking for a weak solution $(\mathscr{E}, \nabla)$ which is moreover *trivial*.

5.1. ***The Röhrl–Deligne construction.*** We briefly recall H. Röhrl's construction (as presented, for instance, in [Bolibrukh 1990; Bolibrukh et al. 2006]). Let $\mathfrak{U} = (U_i)_{i \in I}$ be a finite open cover of $X^* = X \backslash \mathcal{S}$ by connected and simply connected open subsets $U_i \subset X^*$ such that their intersection has the same property, and all triple intersections are empty. Consider arbitrary points $z_i \in U_i$ and $z_{ij} \in U_i \cap U_j$, and paths $\gamma_i : z_0 \to z_i$ and $\gamma_{ij} : z_i \to z_{ij}$, so that $\delta_{ij} = \gamma_i \gamma_{ij} \gamma_{ji}^{-1} \gamma_j^{-1}$ is a positively oriented loop around z_i having winding number 1. Then the cocycle $g = (g_{ij})$ defined over $\mathfrak{U}$ by the constant functions $g_{ij} = \chi([\delta_{ij}])$ defines a flat vector bundle $\mathscr{F}$ over X^*. Define the connection ∇ over U_i by the (0) matrix in the basis of sections corresponding to the cocycle g. The ∇-horizontal sections of $\mathscr{F}$ have by construction the prescribed monodromy behavior. This solves what we called the *topological* Riemann–Hilbert problem in our introduction.

Now add a small neighborhood D of each singular point $s \in \mathcal{S}$ to the cover $\mathfrak{U}$, in such a way that $D\backslash\{s\}$ is covered by k pairwise overlapping sectors $\Sigma_1 = D \cap U_{j_1}$, $\ldots$, $\Sigma_k = D \cap U_{j_k}$. On an arbitrarily chosen sector among the Σ_i, say, Σ_1, let $\tilde{g}_{s1} = z^L$ where z is a local coordinate at s and

$$L = \frac{1}{2i\pi} \log \chi(\delta),$$

normalized with eigenvalues having their real part in the interval $[0, 1[$. Since the open subset Σ_1 only intersects Σ_2 and Σ_k, the only necessary cocycle relations to satisfy are $\tilde{g}_{s2} = \tilde{g}_{s1} g_{12}$ and $\tilde{g}_{sk} = \tilde{g}_{s1} g_{1k}$, which we take as *definition* of the cocycle elements $\tilde{g}_{s2}$ and $\tilde{g}_{sk}$. Define in this way the remaining elements of the cocycle $\tilde{g}$ on $D \cap U_{j_i}$. By construction, the result defines a holomorphic vector bundle $\mathfrak{D}$ on the whole of X, and the connection ∇ can be extended as $L\frac{dz}{z}$ in the basis of sections (σ) of $\mathfrak{D}$ over D chosen to construct $\tilde{g}_{s1}$. The pair $(\mathfrak{D}, \nabla)$ is called the *Deligne bundle* of χ. This construction solves simultaneously the meromorphic and the weak Riemann–Hilbert problem.

Note 55. The basis (σ) is, in our terms, a basis of the Birkhoff form attached to the coordinate z at s.

5.2. ***Weak and strong solutions.*** The Riemann–Hilbert problem can be seen as involving three different levels. The topological level is only governed by the (analytic) monodromy around the prescribed singular set. The meromorphic level is essentially based on the solution of the *local* inverse problem. The third one, that we call *holomorphic* is global and asks for the existence of a trivial holomorphic vector bundle. In fact, separating these three aspects is not so easy to do, because the Röhrl–Deligne construction in fact yields a particular *holomorphic* vector bundle $\mathscr{E}$ with a connection ∇ that already respects the holomorphic prescribed behavior.

What makes the strong Riemann–Hilbert problem a difficult one is precisely this third level. The local meromorphic invariants added to the topological solution of the inverse monodromy specify up to *meromorphic equivalence class* the connection ∇ on X. In this respect, the natural category to state this construction is not the category of holomorphic vector bundles with meromorphic connections, but the meromorphic vector bundles, that is, pairs $(\mathscr{V}, \nabla)$ where $\mathscr{V}$ is locally (but in fact globally) isomorphic to $\mathscr{M}_X^n$. This is why we call the second step *meromorphic*. The Riemann–Hilbert problem with the given data solved here corresponds to the *very weak* Riemann–Hilbert problem (as coined in [Sabbah 2002]): any subsheaf $\mathscr{F}$ of locally free $\mathscr{O}_X$-modules contained in the (trivial) meromorphic bundle $\mathscr{V}$ is endowed naturally with the connection ∇, and therefore is a holomorphic vector bundle with a regular connection having the prescribed monodromy. As stated by the next result (and otherwise well known), all solutions to the weak problem are obtained as local modifications of the Deligne bundle.

Proposition 56. *Let $\widetilde{\pi} : \widetilde{E} \to X$ and $\widetilde{\nabla} : \widetilde{\mathscr{E}} \to \widetilde{\mathscr{E}} \otimes_{\mathbb{O}} \Omega$ be a weak solution to the Riemann–Hilbert problem. Then there exist a finite set $S \subset X$, and local lattices M_x for $x \in S$ such that the pair $(\widetilde{\mathscr{E}}, \widetilde{\nabla})$ is holomorphically isomorphic to $(\mathfrak{D}^M, \nabla)$.*

The last step of the strong Riemann–Hilbert problem consists of searching the set of holomorphic vector bundles endowed with the connection ∇ for a bundle which at the same time has the required holomorphic invariants and is holomorphically trivial. A negative answer requires to know all the holomorphic vector bundles with this prescribed logarithmic property. Note that up to this point, the discussion presented in this section holds over an arbitrary compact Riemann surface.

5.2.1. *Plemelj's theorem.* In 1908, the Slovenian mathematician J. Plemelj (see [Plemelj 1964]) proved a first version of the strong Riemann–Hilbert problem, under the assumption that at least one monodromy is diagonalizable. Whereas his first proof used an analytic approach (Fredholm integrals) to construct the actual matrix of solutions, to thence deduce the differential system and prove that it has only simple poles, the general framework of vector bundles recalled so far allows to establish this fact in an amazingly concise way.

Theorem 57 (Plemelj). *If one of the elementary monodromy maps from representation $\chi : \pi_1(X \backslash \mathcal{S}, z_0) \to \mathrm{GL}_n(\mathbb{C})$ is diagonalizable, then the Riemann–Hilbert problem has a strong solution.*

Proof. Let $(\mathfrak{D}, \nabla)$ be the Röhrl–Deligne bundle attached to the representation χ. Let, say $G = \chi(\gamma)$ around $s \in \mathcal{S}$, be diagonalizable. Let Υ be a Birkhoff form at s, and let (e) be a basis of Υ where G is diagonal. According to condition (ii) in Section 4.2, the whole apartment $\mathscr{A}$ spanned by (e) consists of logarithmic lattices, whereas Theorem 44 implies that $\mathscr{A}$ contains a trivializing lattice M. The vector bundle $\mathfrak{D}^M$ is therefore both logarithmic and trivial. □

Note 58. Here we have a solution by modifying the Deligne bundle only at one point. Note that the lattice M corresponds to a BG trivialization of $\mathfrak{D}$ (see Theorem 59 below). Also note that this result also holds replacing $\mathfrak{D}$ with any other weak solution to Riemann–Hilbert.

5.2.2. *Trivializations of weak solutions.* Let $\mathscr{E}$ be a weak solution of the Riemann–Hilbert problem, and let $\mathscr{F}$ be a trivialization of $\mathscr{E}$ at $x \notin \mathcal{S}$. In a global basis of sections (e) of the bundle $\mathscr{F}$, the connection ∇ is expressed by the matrix of global meromorphic 1-forms Ω, which has a simple pole at every $s \in \mathcal{S}$, and an *a priori* uncontrolled pole at x. Assuming for simplicity that $x \notin \mathcal{S}$ is the point at infinity $\infty \in \mathbb{P}^1(\mathbb{C})$, there exist matrices $A_i \in \mathrm{M}_n(\mathbb{C})$ for $1 \leqslant i \leqslant p$ and a matrix

$$B(z) = B_0 + \cdots + B_t z^t$$

such that the connection has the matrix

$$\Omega = \left(\sum_{i=1}^{p} \frac{A_i}{z - s_i} + B(z)\right) dz.$$

The most surprising consequence of the permutation lemma, as we state it, concerns the analytic invariants of the weak solutions to the Riemann–Hilbert problem.

Theorem 59. *Let $\mathscr{E}$ be a weak solution to the Riemann–Hilbert problem for χ. Then, for any $x \notin \mathcal{S}$, there exists a BG trivialization $\mathscr{F}$ of $\mathscr{E}$ at x which is also* logarithmic *at x. Let $Y = \Gamma(X, \mathscr{F})$ and let $\psi_s = \mathrm{Res}_s^{\mathscr{F}} \nabla \in \mathrm{End}(Y)$.*

(1) *The map $\Psi = \sum_{s \in \mathcal{S}} \psi_s = -\mathrm{Res}_x \nabla$ is semisimple, and has integer eigenvalues, which are equal to the type of the bundle $\mathscr{E}$.*

(2) *The image of the Harder–Narasimhan filtration of $\mathscr{E}$ in Y is equal to the flag induced by the eigenspaces of Ψ ordered by increasing values.*

Proof. If $x \notin \mathcal{S}$, the monodromy at x is trivial, and the stalk $\mathscr{E}_x$ of $\mathscr{E}$ coincides with $\mathfrak{D}_x$. The Birkhoff form Υ of D (which is then unique) is equal to the space V^{∇} of horizontal sections at x. All flags in $\mathbf{D} = \mathfrak{D}_x / \mathfrak{m}_x \mathfrak{D}_x$ are stable under $\mathrm{Res}_x^{\mathfrak{D}} \nabla = 0$. According to Corollary 46, the Υ-lifting of the flag induced by any BG trivialization of $\mathscr{E}$ at x is a *logarithmic BG trivialization* of $\mathscr{E}$ at x. In a global basis of sections (e) of $\mathscr{F}$, the connection has the matrix

$$\begin{aligned} A &= \sum_{s \in \mathcal{S} \setminus \{\infty\}} \frac{A_s}{z - s} + \frac{B}{z - x} \quad \left(\text{where } B = -\sum_{s \in \mathcal{S}} A_s \text{ if } x \neq \infty\right) \\ &= \sum_{s \in \mathcal{S}} \frac{A_s}{z - s} \quad \text{if } x = \infty \notin \mathcal{S}, \end{aligned}$$

since ∇ has no other singularities outside $\mathcal{S} \cup \{x\}$. The eigenvalues of $-B = \sum_{s \in \mathcal{S}} A_s$ are therefore equal to the type of $\mathscr{E}$, and the Harder–Narasimhan filtration is defined by the blocks of equal eigenvalues ordered by increasing values. □

As a consequence, we deduce the following new sufficient condition for the solubility of the strong Riemann–Hilbert problem.

Corollary 60. *Let $\mathscr{E} \in \mathcal{H}$ and let $\mathfrak{D}$ be the Deligne lattice of $(\mathcal{V}, \nabla)$. Let $x \in X$, such that $\mathscr{E}_x = \mathfrak{D}_x = \Delta$. Let $\mathbf{D} = \Delta / \mathfrak{m}_x \Delta$. Let $\mathscr{F} \in \mathrm{BG}_x(\mathscr{E})$, and $M = \mathscr{F}_x$. If the flag $\boldsymbol{F}^{\Delta}(M)$ induced in D is stable under $\mathrm{Res}_x^{\mathfrak{D}} \nabla \in \mathrm{End}(\mathbf{D})$, then there exists $\widetilde{\mathscr{F}} \in \mathrm{BG}_x(\mathscr{E})$ which is moreover* logarithmic *at x.*

Proof. Let $\tilde{M}$ be the Υ-lifting of the flag $\boldsymbol{F}^{\Delta}(M)$, where Υ is a Birkhoff form of Δ. According to Proposition 52, the lattice $\tilde{M}$ is logarithmic, and by the permutation lemma, it is a BG trivializing lattice. Therefore, the bundle $\mathscr{E}^{\tilde{M}}$ satisfies the conclusions of the corollary. □

At this point, we would like to sum up our findings about trivial bundles in the following proposition.

Proposition 61. *Let $\mathcal{F} \in \mathcal{H}_0$ be a trivial bundle in $\mathcal{V}$, and let $Y = \mathcal{F}(X)$ be the $\mathbb{C}$-vector space of global sections. Let $x \in X$, and $\mathcal{E} \in H$ such that $\mathcal{F} \sim_x \mathcal{E}$. Let moreover $\Lambda = \mathcal{E}_x$ and $M = \mathcal{F}_x$.*

(i) *Y has a well-defined flag* $\mathrm{HN}_{\mathcal{E}}$ *induced by the Harder–Narasimhan filtration of $\mathcal{E}$.*

(ii) *If $\mathcal{F} \in \mathrm{BG}_x(\mathcal{E})$, then $T(\mathcal{E}) \cong \mathbf{ED}_\Lambda(M)$ and any Smith basis of Y for Λ ordered according to $K^{\searrow}$ is strictly adapted to* $\mathrm{HN}_{\mathcal{E}}$.

(iii) *If $\mathcal{F}$ is additionally* logarithmic *at x, and the stalk $\mathcal{E}_x$ coincides with the Deligne lattice $\mathfrak{D}_x$, then the type $T(\mathcal{E})$ is given by the integer parts of the eigenvalues of the residue* $\mathrm{Res}_x^{\mathcal{F}} \nabla \in \mathrm{End}(Y)$, *that is, of the* exponents *of ∇ on $\mathcal{F}$ at x.*

(iv) *Finally, if $\mathcal{E} \in \mathrm{RH}_\chi$ is moreover a weak solution to Riemann–Hilbert, then*

$$\sum_{x \in X} \mathrm{Res}_x^{\mathcal{F}} \nabla = 0.$$

When $(\mathcal{E}, \mathcal{F})$ satisfy (i) *to* (iv), *we say that $\mathcal{F}$ is a* good RH trivialization of $\mathcal{E}$ *at x.*

Let $\mathcal{F}$ be a good RH trivialization of $\mathcal{E}$ at $x \notin \mathcal{S}$. Let (σ) be any basis of $Y = \Gamma(X, \mathcal{F})$. In (σ), the connection has a matrix of the form (21). The identification of Y to $\mathbb{C}^n$ by means of (σ) endows $\mathbb{C}^n$ with $p+1$ linear maps ψ_s for $s \in \mathcal{S}^* = \mathcal{S} \cup \{x\}$, that we can identify with the matrices $\widetilde{L}_s$ for $s \in \mathcal{S}$ and $-\sum_{s \in \mathcal{S}} \widetilde{L}_s$ for $s = x$. With these notations, we set the following definition.

Definition 62. The space $\mathbb{C}^n$, endowed with the maps ψ_s for $s \in \mathcal{S}^*$, is called a *linear Fuchsian model* of $\mathcal{E}$.

With this notion, we can reduce some questions about vector bundles to linear algebra statements. For instance we can give the following computable version of a criterion due to Gabber for the reducibility of the triviality index originally appearing in [Sabbah 2002, Corollary I.4.14], that we state here only for the case of a logarithmic modification.

Corollary 63. *Let $\mathcal{E} \in \mathrm{RH}_\chi$ be a weak solution, and consider a linear Fuchsian model at $x \notin \mathcal{S}$, given by p matrices A_s for $s \in \mathcal{S}$ such that*

$$\sum_{s \in \mathcal{S}} A_s = \mathrm{diag}(\boldsymbol{t}_1 I_{n_1}, \ldots, \boldsymbol{t}_s I_{n_s})$$

where the integers $\boldsymbol{t}_i$ satisfy $\boldsymbol{t}_i > \boldsymbol{t}_{i+1}$, in such a way that the flag HN *is the flag $0 = F_0 \subset F_1 \subset \cdots \subset F_s = \mathbb{C}^n$ having signature $(n_1, \ldots, n_s)$ in the canonical basis of $\mathbb{C}^n$. There exists a weak solution $\mathcal{E}'$ adjacent to $\mathcal{E}$ at $s \in X$ and such that*

$i(\mathscr{E}') < i(\mathscr{E})$ if and only if there exists an A_s-stable subspace $W \subset \mathbb{C}^n$ such that $W \cap F_1 = (0)$.

Proof. Let $T = \operatorname{diag}(\boldsymbol{t}_1 I_{n_1}, \dots, \boldsymbol{t}_s I_{n_s}) = \operatorname{diag}(t_1, \dots, t_n)$ be the type of $\mathscr{E}$. We have $i(\mathscr{E}) = \sum_{i=1}^{s} n_i(\boldsymbol{t}_1 - \boldsymbol{t}_i)$. According to Proposition 28, any adjacent weak solution $\mathscr{E}'$ is given by an A_s-stable subspace $W \subset \mathbb{C}^n$. For any basis (e) of $\mathbb{C}^n$ respecting the flag HN, the bundle $\mathscr{E}'$ has type $T' = T - K$, where $k_i = 0$ when $e_i \in W$ and $k_i = 1$ otherwise; therefore $i(\mathscr{E}') = \sum_{i=1}^{n}(\max(t_i - k_i) - t_i + k_i)$ where t_i represent the elements of T without multiplicities. Accordingly, we have

$$i(\mathscr{E}) - i(\mathscr{E}') = \sum_{i=1}^{n}(t_1 - k_i - \max(t_i - k_i)).$$

Now, if there exists i such that $t_i = \boldsymbol{t}_1$ and $k_i = 0$, then $\max(t_i - k_i) = \boldsymbol{t}_1$; thus $i(\mathscr{E}) - i(\mathscr{E}') = \sum_{i=1}^{n} -k_i < 0$ (because we exclude the trivial case $W = \mathbb{C}^n$). Otherwise we have $\max(t_i - k_i) = \boldsymbol{t}_1 - 1$, and then $i(\mathscr{E}) - i(\mathscr{E}') = \sum_{i=1}^{n}(1 - k_i) > 0$. Therefore $\mathscr{E}'$ exists if and only if there exists W stable under some A_s such that $W \cap F_1 = 0$. □

Proposition 64. *Let $\mathscr{F}$ be a BG trivialization of $\mathfrak{D}$ at $x \notin \mathcal{S}$. If there exists a flag $\boldsymbol{F}$ in $Y = \mathscr{F}(X)$ which is* transversal *to* $\mathrm{HN}_{\mathfrak{D}}$*, and is moreover* stable *under the action of one of the maps ψ_s for $s \in \mathcal{S}$, then the strong Riemann–Hilbert problem has a solution, which moreover coincides with $\mathfrak{D}$ outside s.*

Proof. Let $\boldsymbol{F}$ be a flag of Y, which is stable under ψ_s. Taking stalks at x of a $\mathbb{C}$-basis of F, we can see the flag $\boldsymbol{F}$ in $\mathbf{D} = \mathfrak{D}_s/\mathfrak{m}_s\mathfrak{D}_s$. According to Lemma 27(ii), there exists a BG trivialization $\mathscr{E}$ of $\mathfrak{D}$ at x, whose image in $D = \mathfrak{D}_s/\mathfrak{m}_s\mathfrak{D}_s$ is $\boldsymbol{F}$. Let (e) be a BG basis of $\mathfrak{D}_s$ with respect to $\mathscr{E}_s$. Consequently, its image in $\mathbf{D}$ respects the flag $\boldsymbol{F}$. Let Υ be a Birkhoff form of $\mathfrak{D}_s$, and let (e_Υ) be the Υ-basis of (e). Since the gauge from (e) to (e_Υ) is tangent to I, the lattice M induced from (e_Υ) by the elementary divisors K of $\mathscr{E}_s$ in Λ is also a trivializing BG lattice for $\mathfrak{D}$ at s. However, the lattice M is also logarithmic, since by construction it induces in $\mathbf{D}$ the ψ_s-stable flag $\boldsymbol{F}$, and moreover sits inside an apartment that contains the Birkhoff form Υ. Hence, the bundle $\mathfrak{D}^M$ is both trivial and logarithmic. □

We have represented the weak solutions to the Riemann–Hilbert problem as points in a product of subvarieties of stable flags.

Theorem 65. *Let $\mathfrak{D}$ be the Deligne bundle, and $\mathscr{F}$ a BG trivialization at an apparent singularity $x \notin \mathcal{S}$. The set of weak solutions to the Riemann–Hilbert problem for χ is parametrized by the set*

$$\mathrm{RH}_\chi = \{(\boldsymbol{F}^s, D_s)_{s\in\mathcal{S}} \mid \boldsymbol{F}^s \in \mathfrak{Fl}_{\psi_s}(Y),\, D_s \in \mathbb{Z}^n(\boldsymbol{F}^s)\},$$

where $Y = \mathscr{F}(X)$ and $\psi_s = \operatorname{Res}_s^{\mathscr{F}} \nabla \in \operatorname{End}_{\mathbb{C}}(Y)$ for $s \in \mathcal{S}$.

5.3. ***The type of the Deligne bundle.*** The strong version of the Riemann–Hilbert problem would directly have a solution if the Deligne bundle were trivial. However, this is not the case, unless all singular points are apparent, since the exponents of ∇ are normalized in such a way that their sum is nonnegative. This means that the type of the Deligne bundle as a rule is not trivial. We have seen several ways to characterize this nontriviality. The *type* characterizes the isomorphism classes of holomorphic vector bundles, so it would seem possible to work with this sole information. However, we are not in the right category to do so, since we consider holomorphic bundles with an embedding in a meromorphic one, denoted by $\mathscr{V}$. This is the reason for which there are *several* trivial bundles in $\mathscr{V}$. From another point of view, it is not possible to determine on the sole basis of the sequence $T = (a_1, \ldots, a_n)$, what the effect of changing the stalk of $\mathfrak{D}$ at x will be. Obviously the geometry of the Harder–Narasimhan filtration will play a decisive role.

5.3.1. *Trivializations of the Deligne bundle.* Let us examine in further detail the case of the Deligne bundle $\mathfrak{D}$. Let us say that δ_i is an elementary generator of the homotopy group $G = \pi_1(X \setminus \mathcal{S}, z_0)$, if δ_i is a closed path based at z_0, having winding number $+1$ around the singularity s_i and 0 around the others. Let $G_i = \chi(\delta_i)$ and $L_i = \frac{1}{2i\pi} \log G_i$, normalized as for the Deligne lattice. Let (σ_i) be a basis of the Birkhoff form Υ_i at s_i described in Note 55, such that the connection has locally as matrix $\Omega_i = L_i \frac{dz}{z}$, on a neighborhood, say D_i of s_i. On the other hand, let D_0 be a neighborhood of z_0, and consider a basis (σ_0) of the local Birkhoff form. According to what precedes, (σ_0) is a basis of local ∇-*horizontal sections* of $\mathfrak{D}$ over D_0. One can moreover choose this basis in such a way that the monodromy of (σ_0) around s_i is exactly given by the matrix G_i.

Assume now for simplicity that $x \notin \mathcal{S}$ is the point at infinity $\infty \in \mathbb{P}^1(\mathbb{C})$, and let $\mathscr{F}$ be a trivialization of $\mathfrak{D}$ at x. In a global basis of sections (e) of the bundle $\mathscr{F}$, there exist matrices $B_i \in \mathrm{M}_n(\mathbb{C})$ and a matrix

$$B(z) = B_0 + \cdots + B_t z^t \quad \text{and} \quad C_i \in \mathrm{GL}_n(\mathbb{C}) \text{ for } 1 \leqslant i \leqslant p$$

such that the connection has the matrix

$$\Omega = \left(\sum_{i=1}^{p} \frac{C_i^{-1} L_i C_i}{z - s_i} + B(z) \right) dz.$$

Note 66. If the bundle $\mathscr{F}$ is moreover logarithmic at ∞ — which can be achieved, for example, by Plemelj's theorem — then $B = 0$ and the residue at infinity, $L_\infty = -\sum_{i=1}^{p} C_i^{-1} L_i C_i$, is semisimple with integer eigenvalues (ssie). At the cost of a (harmless) global conjugation, we can already assume that

$$L_\infty = \mathrm{diag}(b_1 I_{n_1}, \ldots, b_s I_{n_s}) \quad \text{with } b_1 < \cdots < b_s.$$

Note that the sequence $\mathcal{B} = (b_1 I_{n_1}, \dots, b_s I_{n_s})$ coincides with the elementary divisors of the stalk $\mathcal{F}_\infty$ in $\mathfrak{D}_\infty$.

Definition 67. We say that $(C_1, \dots, C_p) \in \mathrm{GL}_n(\mathbb{C})^p$ is a *normalizing p-tuple* for χ if $\sum_{i=1}^p C_i^{-1} L_i C_i$ is ssie for some (and therefore any) normalized logarithms L_i of the generators $\chi(\gamma_i)$ of the monodromy group.

Normalizing p-tuples always exist. Putting t as the coordinate $1/z$ at infinity, the Taylor expansion of ∇ at $x = \infty$ has then the nice expression

$$\Omega = -\sum_{k \geqslant 0} \sum_{i=1}^{p} s_i^k \widetilde{L}_i t^k \frac{dt}{t} \quad \text{with } \widetilde{L}_i = C_i^{-1} L_i C_i. \tag{20}$$

We have thus reduced the computation of the type of the Deligne bundle to the computation of the matrices C_i (the so-called *connection matrices*, because they connect the different local expressions of ∇ on the local Birkhoff forms). It is however well known that the computation of the connection matrices is difficult. Any other trivialization of $\mathfrak{D}$ at infinity is given by a *monopole gauge* (as coined in [Ilyashenko and Yakovenko 2007]), namely a unimodular polynomial matrix $\Pi \in \mathrm{GL}_n(\mathbb{C}[z])$, that is, a matrix satisfying

$$\Pi = P_0 + P_1 z + \cdots + P_k z^k \quad \text{such that } \det \Pi(z) = \mathrm{cst} \in \mathbb{C}^*.$$

Proposition 68. *Given a family of points $s_1, \dots, s_p \in \mathbb{C}$ and invertible matrices $C_1, \dots, C_p \in \mathrm{GL}_n(\mathbb{C})$, there exists a monopole gauge $\Pi \in \mathrm{GL}_n(\mathbb{C}[z])$ such that $\Pi(s_i) = C_i$ for $1 \leqslant i \leqslant p$.*

Proof. The group $\mathrm{GL}_n(R)$ on a ring is generated by transformations $T_{ij}(\lambda) = I + \lambda E_{ij}$ where $\lambda \in R$ and E_{ij} is the (i, j) element of the canonical basis of the vector space $\mathfrak{gl}_n$. At the cost of introducing the trivial transformations $T_{ij}(0) = I$, one can assume that all the matrices C_i can be expressed as a product of the same transformations with different parameters:

$$C_i = T_1(\mu_1^i) \cdots T_s(\mu_s^i) \quad \text{with } \mu_t^i \in \mathbb{C}.$$

Define then $\lambda_k \in \mathbb{C}[z]$ such that $\lambda_k(s_i) = \mu_k^i$ for $1 \leqslant i \leqslant p$. By construction, the product $\widetilde{\Pi} = T_1(\lambda_1) \cdots T_s(\lambda_s) \in \mathrm{SL}_n(\mathbb{C}[z])$ indeed interpolates the matrices C_i at the points s_i. □

As a consequence of this result, one can find a trivialization $\mathscr{E}$ at infinity of the Deligne bundle such that the residues of the connection ∇ are expressed in a basis of $Y = \Gamma(X, \mathscr{E})$ as the actual matrices L_i (and not *conjugated* to them). Although the point at infinity of $\mathscr{E}$ is still an apparent singularity, we have no control on the Poincaré rank of ∇ at ∞.

The results of this section also hold (with the adequate modifications) if the apparent singularity is assumed to be located at $z_0 \notin \mathcal{S} \cup \{\infty\}$. We will refer to the trivialization $\mathscr{E}$ as an *adapted trivialization of* $\mathfrak{D}$ *at* z_0.

5.3.2. *A Deligne–Simpson-type problem.* We know that there exists a family of invertible matrices (C_i) such that $\sum_{i=1}^{p} C_i^{-1} L_i C_i$ is semisimple with integer eigenvalues and that these eigenvalues are equal to the type of the Deligne bundle. This raises two questions:

(1) Does there exist a logarithmic trivialization of $\mathfrak{D}$ for any such family (C_i)?

(2) If there exist several families with this property, how do we recognize those that indeed give the type of the Deligne bundle?

This also raises an interesting computational problem akin to the well-known Deligne–Simpson problem (see, e.g., [Crawley-Boevey 2003]). Let $\mathscr{C}_i$ be the conjugacy class of $L_i = \frac{1}{2i\pi} \log G_i$ under $\mathrm{GL}_n(\mathbb{C})$.

(DS) Determine all conjugates $\widetilde{L}_i \in \mathscr{C}_i$ such that $\sum_{i=1}^{p} \widetilde{L}_i = \mathrm{diag}(b_1, \ldots, b_n) \in \mathbb{Z}^n$.

5.4. ***The Bolibrukh–Kostov theorem.*** The most celebrated recent result on the Riemann–Hilbert problem is the following fact, proved first independently by A. Bolibrukh and V. Kostov.

Theorem 69 (Bolibrukh–Kostov). *The strong Riemann–Hilbert problem is solvable for any* irreducible *monodromy representation* χ.

We give first an algebraic proof of a classical result of Bolibrukh [Anosov and Bolibrukh 1994, Proposition 4.2.1].

Proposition 70. *If the representation* χ *is irreducible, then for any weak solution* $\mathscr{E} \in \mathrm{RH}_\chi$, *the type* $T(\mathscr{E}) = (t_1, \ldots, t_n)$ *of* $\mathscr{E}$ *satisfies* $|t_i - t_j| \leq |i - j|(p-2)$.

Proof. Assume here for simplicity that $x = \infty \notin \mathcal{S}$, and consider again the setting of Section 5.3.1. Let $\mathscr{E}$ be any weak solution to Riemann–Hilbert, and $\mathscr{F}$ be a logarithmic BG trivialization of $\mathscr{E}$ at x. Let $T = (t_1, \ldots, t_n)$ be the type of $\mathscr{E}$. In a basis (e) of global sections of $\mathscr{F}$, there exist constant matrices $\widetilde{L}_a$ for $a \in \mathcal{S}$ such that the connection ∇ has in (e) the matrix

$$\Omega = \sum_{a \in \mathcal{S}} \frac{\widetilde{L}_a}{z-a} dz = -\frac{d\tilde{z}}{\tilde{z}} \sum_{k \geqslant 0} \Omega_k \tilde{z}^k \quad \text{with } \Omega_k = \sum_{a \in \mathcal{S}} a^k \widetilde{L}_a \text{ and } \tilde{z} = \frac{1}{z}. \tag{21}$$

By Proposition 49, the shearing $\tilde{z}^{-T}$ suppresses the singularity at x, since the basis $\tilde{z}^{-T}(e)$ spans the Deligne lattice. As a consequence, $\widetilde{\Omega} = \tilde{z}^T \Omega(\tilde{z}) \tilde{z}^{-T} + T \frac{d\tilde{z}}{\tilde{z}}$ must satisfy $v(\widetilde{\Omega}) \geq 0$. Therefore, the residue matrix $B = -\sum_{a \in \mathcal{S}} \widetilde{L}_a$ of Ω at x is diagonal and equal to $-T$. We can assume further that

$$B = \mathrm{diag}(b_1 I_{n_1}, \ldots, b_s I_{n_s}) \quad \text{with } b_1 = -\boldsymbol{t}_1 < \cdots < b_s = -\boldsymbol{t}_s$$

where $(\boldsymbol{t}_1 I_{n_1}, \dots, \boldsymbol{t}_s I_{n_s})$ represents the type of $\mathscr{E}$ with multiplicities. Partition any matrix M according to the eigenvalue multiplicities of B, as $(M_{\ell,m})$ for $1 \leq \ell, m \leq s$. Then the matrix of the connection can be rewritten by blocks as

$$\widetilde{\Omega}_{\ell,m} = \Omega_{\ell,m} t^{\boldsymbol{t}_\ell - \boldsymbol{t}_m} + T\frac{d\tilde{z}}{\tilde{z}} = \left(-\sum_{j \geq 0} \Omega_{\ell,m}^{(j)} \tilde{z}^{j + \boldsymbol{t}_\ell - \boldsymbol{t}_m} + \delta_{\ell,m} \boldsymbol{t}_\ell I_{n_\ell} \right) \frac{d\tilde{z}}{\tilde{z}}.$$

For each (ℓ, m) block, this series must have strictly positive valuation. The sum $\sum_{a \in \mathscr{S}} \widetilde{L}_a = T$ imposes conditions on all blocks of the residues $\widetilde{L}_a$, while when $\ell > m$ we get the following equations:

$$\Omega_{\ell,m}^{(j)} = \sum_{a \in \mathscr{S}} a^j (\widetilde{L}_a)_{\ell,m} = 0 \text{ for } 0 \leq j \leq \boldsymbol{t}_m - \boldsymbol{t}_\ell \quad \text{when } \ell > m. \tag{22}$$

For a fixed pair (ℓ, m), let $k = \max(0, \boldsymbol{t}_m - \boldsymbol{t}_\ell)$, and let $X_i \in \mathbb{C}^{n_\ell \times n_m}$ be the (ℓ, m)-block of the matrix $\widetilde{L}_{s_i}$, for $1 \leq i \leq p$. For $1 \leq \alpha \leq n_\ell$ and $1 \leq \beta \leq n_m$, let $v_{\alpha,\beta} \in \mathbb{C}^p$ be the vector constructed by taking the coefficient of index (α, β) of X_i, for $1 \leq i \leq p$. Then, the equations (22) can be reformulated as

$$v_{\alpha,\beta} \in \ker M_k(\underline{s}) \quad \text{where } M_k(\underline{s}) = \begin{pmatrix} 1 & \cdots & 1 \\ s_1 & \cdots & s_p \\ \vdots & & \vdots \\ s_1^k & \cdots & s_p^k \end{pmatrix}.$$

The matrix $M_k(\underline{s})$ is an upper-left submatrix of a Vandermonde matrix with coefficients

$$\underline{s} = (s_1, \dots, s_p) \in \mathbb{C}^p \setminus \bigcup_{i \neq j} \{x_i \neq x_j\}.$$

Since all the s_i are distinct, this matrix has always full rank. In particular, as soon as $\boldsymbol{t}_m - \boldsymbol{t}_\ell \geq p - 1$, it has a null kernel, and so all the blocks X_i are zero. Due to the ordering of the $\boldsymbol{t}_i$, we also have $\boldsymbol{t}_{m'} - \boldsymbol{t}_{\ell'} \geqslant p - 1$ for $m' \leqslant m$ and $\ell' \geqslant \ell$; thus all matrices $\widetilde{L}_a$ have a lower-left common zero block. If $m = \ell + 1$, this means that the representation χ is reducible. □

Proof of Theorem 69. Consider the Deligne bundle $\mathfrak{D}$ and an arbitrary singularity $s \in \mathscr{S}$. Put $\Delta = \mathfrak{D}_s$ and let $\boldsymbol{F}$ be a complete flag in $\Delta/\mathfrak{m}\Delta$ which is stable under $\delta = \operatorname{Res}_s^{\mathfrak{D}} \nabla$. Let $D = (0, d, 2d, \dots, (n-1)d)$. By Proposition 52, the lattice Λ obtained by lifting the pair $(\boldsymbol{F}, D)$ in the z-Birkhoff form $\Upsilon \subset \Delta$ is logarithmic. Let (e) be a basis of Υ respecting the flag $\boldsymbol{F}$. Since the residue δ is upper-triangular in (e), and the elements of D are distinct, the elements of the matrix $\Omega = \operatorname{Mat}(\nabla, (z^D e))$ outside the diagonal have valuation at least d. Hence $\Omega = (A + z^d U)\frac{dz}{z}$ where A is diagonal and U holomorphic. By Theorem 44, the

apartment $\mathscr{A}$ spanned by (e) contains a Birkhoff–Grothendieck trivialization M of Λ, as in the following figure:

$$\Delta \xrightarrow{z^D} \Lambda \in \mathbf{\Lambda}_{\log} \xrightarrow{z^{\widetilde{D}}} M \in \mathrm{BG}(\Lambda).$$

According to Proposition 70, the sequence $\widetilde{D}$ satisfies $\Delta\widetilde{D} \leqslant (n-1)(p-2)$, while the Poincaré rank $\mathfrak{p}$ of ∇ on M satisfies $\mathfrak{p} = \min(0, -v(z^{-\widetilde{D}}\Omega z^{\widetilde{D}}))$. Since $v(z^{-\widetilde{D}}\Omega z^{\widetilde{D}}) \geqslant d - \Delta\widetilde{D}$ holds, it is sufficient to impose $d \geqslant (n-1)(p-2)$ to ensure that $\mathfrak{p} = 0$, that is, that the lattice M is logarithmic. Then $\mathfrak{D}^M$ is a strong solution. □

The previous proof holds in fact for any representation χ for which there exists a constant R such that for any weak solution $\mathscr{E}$ of RH for χ, the type of $\mathscr{E}$ satisfies $\Delta T(\mathscr{E}) \leqslant R$. Finally, as a byproduct of the proof of Proposition 70, we have the following result.

Corollary 71. *Let $\widetilde{L}_i \in \mathscr{C}_i$ such that $\sum_{i=1}^p \widetilde{L}_i = \mathrm{diag}(t_1, \ldots, t_n) \in \mathbb{Z}^n$. Then the sequence $T = (t_1, \ldots, t_n)$ represents the type of the Deligne bundle of the monodromy representation χ of the Fuchsian system (with singular locus $\mathscr{S} = \{s_1, \ldots, s_p\} \subset \mathbb{C}$)*

$$Y' = \sum_{i=1}^{p} \frac{\widetilde{L}_i}{z - s_i} Y$$

if and only if $\sum_{i=1}^p s_i^j (\widetilde{L}_i)_{\ell,m} = 0$ holds for $0 \leq j \leq k_m - k_\ell$ and $\ell > m$.

This gives a partial answer to our question (2) from Section 5.3.2. For any explicit solution $L = (L_1, \ldots, L_p)$ to the generalized Deligne–Simpson problem, we can generate nontrivial explicit examples of Deligne bundles (and their Harder–Narasimhan filtrations) corresponding to monodromy representations which are *locally conjugate* to the original one. In the following section, we give the algorithmic procedures that can then be used to determine effectively if the corresponding representation admits or not a strong solution to the Riemann–Hilbert problem. Note that the equations in Corollary 71 show that the set of singular loci $\mathscr{S}$ for which L corresponds to a logarithmic BG trivialization of a Deligne bundle is an algebraic (projective) subvariety of $\mathbb{C}^p \setminus \bigcup_{i \neq j} \{x_i = x_j\}$.

5.5. ***Testing the solubility of the Riemann–Hilbert problem.*** In this section, we apply the results of this paper to the experimental investigation of the solubility of the Riemann–Hilbert problem. We present two ways to search the space of weak solutions, which are completely effective (up to the known problem of connection matrices): one that follows paths of adjacent logarithmic lattices, based on Lemma 51, the other that uses the characterization as stable flags given in

Proposition 52. Note that, if any (not necessarily logarithmic) trivial holomorphic bundle of the meromorphic solution to Riemann–Hilbert is explicitly given, the procedures that we present, coupled with classical Poincaré rank reduction methods, implemented on a computer algebra system, allow to make the actual computations. We however do not know if this bypasses the problem of the connection matrices.

Let $\mathfrak{D}$ be the Deligne bundle of the representation χ. Let $x \notin \mathcal{S}$, and consider a logarithmic BG trivialization $\mathcal{F}$ of $\mathfrak{D}$ at x. Let $Y = \Gamma(X, \mathcal{F})$ and choose a basis (σ) of Y in which the residue matrix at x is equal to the diagonal that represents the type of $\mathfrak{D}$:

$$\mathrm{Mat}(\mathrm{Res}_x^{\mathcal{F}}\nabla, (\sigma)) = -D = \mathrm{diag}(-k_1 I_{n_1}, \dots, -k_s I_{n_s}) \quad \text{where } k_1 > \cdots > k_s.$$

In the basis (σ), the connection has a matrix of the form (21), and the Harder–Narasimhan filtration is expressed as the flag HN_Y of signature $(n_1, \dots, n_s)$ of Y. Let $V = \Gamma(X, \mathcal{V})$ be the $\mathcal{K}$-vector space of meromorphic sections of $\mathcal{V}$, where $\mathcal{K} = \Gamma(X, \mathcal{M}_X)$ is the field of meromorphic functions on X.

For $s \in \mathcal{S}$, let t be a coordinate at x with divisor $(t) = x - s$, and $(\widetilde{\sigma}) = t^{-D}(\sigma)$. Recall that t^{-1} is a coordinate at s. For clarity's sake, we will put $t_x = t$ and $t_s = t^{-1}$ when we are dealing with local sections. Let $\widetilde{\mathcal{F}} = t_s(\mathcal{F})$ be the transport of $\mathcal{F}$ at s and $\widetilde{Y} = \Gamma(X, \widetilde{\mathcal{F}})$. We regard Y and $\widetilde{Y}$ as sub-$\mathbb{C}$-vector spaces of V, spanned respectively by the $\mathcal{K}$-bases (σ) and $(\widetilde{\sigma})$ of V. The relation $(\widetilde{\sigma}) = t^{-D}(\sigma)$ induces a well-defined fixed isomorphism between Y and $\widetilde{Y}$.

Claim 72. *The trivial bundle $\widetilde{\mathcal{F}}$ is a BG trivialization of $\mathfrak{D}$ at s.*

Claim 73. *The flag $\mathrm{HN}_{\widetilde{Y}}$ is the flag of signature $(n_1, \dots, n_s)$ spanned by $(\widetilde{\sigma})$.*

Claim 74. *The germ (σ_s) of the global basis of Y at s is a local basis of $\mathfrak{D}_s$.*

Indeed, we have the two dual schematic representations, where $(\sigma_x) : \mathcal{E}_x$ means that (σ) is a local basis of $\mathcal{E}$ at x and $(\sigma) : Y$ means that (σ) is a global basis of the form Y:

$$(\widetilde{\sigma}_x) : \mathfrak{D}_x \xrightarrow{t_x^D} (\sigma) : Y \quad \text{and} \quad (\sigma_s) : \mathfrak{D}_s \xrightarrow{t_s^D} (\widetilde{\sigma}) : \widetilde{Y}.$$

5.5.1. *Bolibrukh's first counterexample.* Bolibrukh's first published counterexample is the 3×3 system

$$dX/dz = AX, \tag{23}$$

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & z & 0 \\ 0 & 0 & -z \end{pmatrix} \frac{1}{z^2} + \begin{pmatrix} 0 & 1 & 0 \\ 0 & -\frac{1}{6} & \frac{1}{6} \\ 0 & -\frac{1}{6} & \frac{1}{6} \end{pmatrix} \frac{1}{z+1} + \begin{pmatrix} 0 & 0 & 1 \\ 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \frac{1}{z-1} + \begin{pmatrix} 0 & -1 & -1 \\ 0 & -\frac{1}{3} & \frac{1}{3} \\ 0 & -\frac{1}{3} & \frac{1}{3} \end{pmatrix} \frac{1}{z-\frac{1}{2}}.$$

Let us show what the different notions introduced in the paper are in this case. We consider the system (23) to be the expression of the connection ∇ on a trivial

bundle $\mathscr{E}$, in a basis (σ) of global sections. The singular divisor is $\mathscr{D} = 2\cdot 0 + 1 - 1 + \frac{1}{2}$ and the matrices at 1, -1 and $\frac{1}{2}$ are nilpotent. The point at infinity is not singular, since, putting $t = 1/z$, we have

$$-\frac{1}{t^2} A\left(\frac{1}{t}\right) = \frac{1}{2}\begin{pmatrix} -1 & 0 & -1 \\ -1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} + o(1). \tag{24}$$

Therefore the stalks $\mathscr{E}_x$ for $x \in \mathbb{P}^1(\mathbb{C})\backslash\{0\}$ coincide with the Deligne bundle $\mathfrak{D}$ of ∇: in the terms of Section 1, the bundle $\mathscr{E}$ is a trivialization of $\mathfrak{D}$ at 0. However, the singular point 0 is not an apparent singularity. The gauge $P = \mathrm{diag}(1, z, 1/z)$ brings the system to the form

$$A_{[P]} = \frac{1}{z}\widetilde{A}_0 + \frac{1}{z+1}\widetilde{A}_{-1} + \frac{1}{z-1}\widetilde{A}_1 + \frac{1}{z-\frac{1}{2}}\widetilde{A}_{\frac{1}{2}} + B$$

where

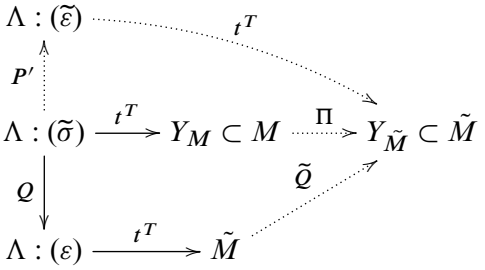
$$\widetilde{A}_0 = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} \widetilde{A}_{-1} = \frac{1}{6}\begin{pmatrix} 0 & -6 & 0 \\ 0 & -1 & -1 \\ 0 & -1 & 1 \end{pmatrix} \widetilde{A}_1 = \frac{1}{2}\begin{pmatrix} 0 & 0 & 2 \\ 0 & -1 & 1 \\ 0 & -1 & 1 \end{pmatrix},$$

$$\widetilde{A}_{\frac{1}{2}} = \frac{1}{12}\begin{pmatrix} 0 & -6 & -24 \\ 0 & -4 & 16 \\ 0 & -1 & 4 \end{pmatrix} \quad \text{and} \quad B = \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

The bundle $\mathscr{F}$ spanned by the global sections $(\sigma') = (\sigma_1, z\sigma_2, \sigma_3/z)$ is trivial by construction, and since all residues over $\mathscr{S} = \{0, 1, -1, 1/2\}$ are nilpotent, $\mathscr{F}$ is a trivialization of $\mathfrak{D}$ at ∞. Moreover, the stalk $\mathfrak{D}_\infty$ is spanned by (σ_∞), so the germ of basis (σ'_∞) is a Smith basis of $\mathfrak{D}_\infty$; hence $\mathscr{F}$ is actually a BG trivialization of the Deligne bundle $\mathfrak{D}$ at $x = \infty$. In general, such a gauge can be found explicitly by combining a Poincaré rank reducing method at all finite singularities (e.g., Gérard–Levelt saturation [1973]) and the BG trivialization algorithm from Section 3.2.1.

Accordingly, the type of the Deligne bundle is $T = T(\mathfrak{D}) = (1, 0, -1)$. However, an apparent singularity of Poincaré rank 2 appears at ∞; hence the BG trivialization $\mathscr{F}$ is not logarithmic at ∞. To get one, we apply the permutation lemma (Proposition 38). We reorder the basis at ∞ as $(\widetilde{\sigma}) = (\sigma_3, \sigma_1, \sigma_2)$ according to the decreasing elements of the type. Putting $\Lambda = \mathfrak{D}_\infty$ and $M = \mathscr{F}_\infty$, for any lattice gauge $Q \in \mathrm{GL}_3(\mathbb{C}[\![t]\!])$, we can find P', $\widetilde{Q} \in \mathrm{GL}_3(\mathbb{C}[\![t]\!])$ and a monopole

$\Pi \in \mathrm{GL}_3(\mathbb{C}[z])$ as in the following diagram:



We will get a BG logarithmic trivialization if (ε) is a basis of the Birkhoff form Υ of $\mathfrak{D}_\infty$. Since ∞ is a regular point, the gauge Q is a holomorphic fundamental matrix of solutions of the system (24). Lemma 15 ensures that we can actually truncate Q at order $\Delta T - 1 = 1$; hence we can take

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} -1 & 0 & -1 \\ -1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} t.$$

The gauge P' is obtained as in the proof of Proposition 38:

$$P' = \begin{pmatrix} 1 & 0 & -\frac{t}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{so that } \Pi = \begin{pmatrix} 1 & 0 & -\frac{1}{2t} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Finally, we get

$$A_{[P' t^T]} = \frac{1}{z} A_0 + \frac{1}{z+1} A_{-1} + \frac{1}{z-1} A_1 + \frac{1}{z-\frac{1}{2}} A_{\frac{1}{2}},$$

$$\text{with} \quad A_0 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ -1 & 0 & 0 \end{pmatrix}, \quad A_{-1} = \frac{1}{24} \begin{pmatrix} 2 & 0 & -1 \\ 0 & 0 & -24 \\ 4 & 0 & -2 \end{pmatrix},$$

$$A_1 = \frac{1}{8} \begin{pmatrix} 2 & 0 & 1 \\ 8 & 0 & -4 \\ -4 & 0 & -2 \end{pmatrix} \quad \text{and} \quad A_{\frac{1}{2}} = \frac{1}{3} \begin{pmatrix} 2 & 0 & -1 \\ -6 & 0 & 0 \\ 4 & 0 & -2 \end{pmatrix}.$$

This is a Fuchsian linear model of (23) at $x = \infty$, and we check that the eigenvalues of

$$A_0 + A_{-1} + A_1 + A_{\frac{1}{2}} = \begin{pmatrix} 1 & 0 & -\frac{1}{4} \\ 0 & 0 & -\frac{1}{2} \\ 0 & 0 & -1 \end{pmatrix}$$

give indeed the type of $\mathfrak{D}$. The residues A_s are all nilpotent with maximal rank; hence there is a unique complete flag $\boldsymbol{F}^{(s)}$ which is stable by A_s. In the canonical basis (e) of $\mathbb{C}^3$, the Harder–Narasimhan filtration corresponds to the coordinate flag

$$\boldsymbol{H} : (0) \subset \langle e_1 \rangle \subset \langle e_1, e_2 \rangle \subset \mathbb{C}^3$$

and we have

$$\begin{aligned} \boldsymbol{F}^{(1)} = \boldsymbol{F}^{(\frac{1}{2})} &: (0) \subset \langle e_2 \rangle \subset \langle e_2, e_1 + 2e_3 \rangle \subset \mathbb{C}^3, \\ \boldsymbol{F}^{(-1)} &: (0) \subset \langle e_2 \rangle \subset \langle e_2, e_1 - 2e_3 \rangle \subset \mathbb{C}^3, \\ \boldsymbol{F}^{(0)} &: (0) \subset \langle e_2 \rangle \subset \langle e_2, e_3 \rangle \subset \mathbb{C}^3. \end{aligned}$$

We see that no stable flag under any A_s is transversal to $\boldsymbol{H}$, which is a necessary condition to be a counterexample (by Proposition 64). However, the condition of Corollary 63 is satisfied (at each $s \in \mathcal{S}$), which means that there is an adjacent weak solution $\mathscr{E}$ with strictly smaller default $i(\mathscr{E}) < i(\mathfrak{D}) = 3$.

5.5.2. *Adjacent lattices.* In this section, we consider a weak solution $\mathscr{E} \in \mathrm{RH}_\chi$. In the following proposition, we describe a procedure which allows to read off at an apparent singularity $x \notin \mathcal{S}$, fixed once and for all, the effect on the weak solution $\mathscr{E}$ of a change of logarithmic adjacent lattice at any singularity $s \in \mathcal{S}$. More precisely, let (σ) be a global basis of a logarithmic BG trivialization of $\mathscr{E}$ at x, and Ω the matrix in Fuchsian form (21) of the connection ∇ in (σ), whose residue at x gives precisely the type of $\mathscr{E}$. Let M be a logarithmic lattice at s that is adjacent to $\mathscr{E}_s$. We determine explicitly a gauge transform Π_M which is a monopole at x, such that $\Omega_{[\Pi_M]}$ has again Fuchsian form (21). From its semisimple residue at x we read directly the type of the modified bundle $\mathscr{E}^M$, equal to the eigenvalues, and the Harder–Narasimhan filtration of $\mathscr{E}^M$, spanned by the eigenspaces ordered by increasing values. This procedure is completely effective once the connection matrices C_s that relate the *local* residue matrices $L_s = \frac{1}{2i\pi} \log G_s$ in the Birkhoff form at s and the *global* residue matrices $\widetilde{L}_s = C_s^{-1} L_s C_s$ in the basis (σ) have been determined.

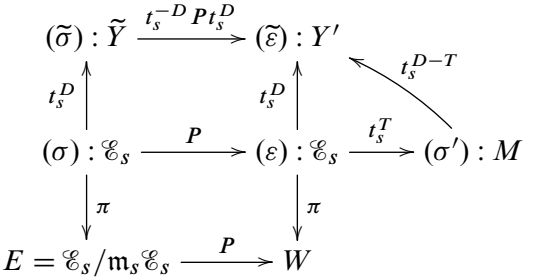
Let M be a lattice at s that is adjacent to $\mathscr{E}_s$. This lattice is uniquely characterized by its image $W = M/\mathfrak{m}_s \mathscr{E}_s$, that can be seen as a sub-$\mathbb{C}$-vector space $W \subset Y$. It is logarithmic if and only if W is stable under the map $\mathrm{Res}_s^{\mathscr{E}} \nabla$.

According to Proposition 28, a BG trivialization of $\mathscr{E}^M$ is obtained from a basis of $\mathscr{E}_s$ that simultaneously respects the space W and the flag HN. Moreover, we can choose (ε) in the $\mathrm{GL}_n(\mathbb{C})$-orbit of (σ).

Claim 75. *There exists a basis (ε) of Y such that $t_s^D(\varepsilon)$ spans a BG trivialization of both $\mathscr{E}$ and $\mathscr{E}^M$ at s.*

Claim 76. *The matrix $P \in \mathrm{GL}_n(\mathbb{C})$ of the basis change from (σ) to (ε) is $(-D)$-parabolic.*

Claim 77. *The gauge $t_s^{-D} P t_s^D = t_x^D P t_x^{-D}$ is a monopole at s and an element of $\mathrm{GL}_n(\mathcal{O}_x)$.*



Claim 78. *The basis (σ') generates M at s and $\mathscr{E}_y$ at $y \neq x$.*

Claim 79. *The trivial bundle $\mathscr{F}'$ spanned by (σ') is a BG trivialization of $\mathscr{E}^M$ at x.*

Claim 80. *The gauge transform from (σ) to (σ') is $P t_s^T = P t_x^{-T}$.*

Claim 81. *The Harder–Narasimhan filtration of $\mathscr{E}^M$ is given by the flag of Y' spanned by (σ') according to $D - T$.*

Indeed, the last arrow on the right implies that at x, we have

$$(\widetilde{\varepsilon}_x) : \mathscr{E}_x = \mathscr{E}_x^M \xrightarrow{t_x^{D-T}} (\sigma') : Y' \quad \text{where } Y' \subset V \text{ is spanned over } \mathbb{C} \text{ by } (\sigma').$$

Therefore the type of $\mathscr{E}^M$ is, as expected, equal to $D - T$.

Proposition 82. *Assume that $\mathcal{S} \subset \mathbb{C}$ and $x = \infty$. Let $\mathscr{E} \in \mathrm{RH}_\chi$ be a weak solution to the Riemann–Hilbert problem. Let the connection ∇ have a matrix Ω of the form* (21) *in a basis (σ) of a logarithmic BG trivialization $\mathscr{F}$ of $\mathscr{E}$ at x. Then, for any $\widetilde{L}_s$-stable subspace W_s of $\mathbb{C}^n$, there exists a computable monopole gauge $\Pi \in \mathrm{GL}_n(\mathbb{C}[z])$, a constant matrix $P_0 \in \mathrm{GL}_n(\mathbb{C})$ and a diagonal matrix T with only* 0, 1 *elements such that $\Omega_{[P_0(z-s)^T \Pi]}$ has again the form* (21) *corresponding to the modification $\mathscr{E}^M$, where M is the lattice of $\mathcal{V}_s$ adjacent to $\mathscr{E}_s$ canonically defined by W_s.*

Proof. We identify $\Gamma(X, \mathscr{F})$ with $\mathbb{C}^n$ by means of the basis (σ). The residue of ∇ at s is then equal to the matrix $L = \widetilde{L}_s$ of formula (21). A logarithmic adjacent lattice M is uniquely defined by an L-stable subspace $W \subset \mathbb{C}^n$. Let (ε) be a basis respecting both W and the Harder–Narasimhan flag $\boldsymbol{H}$, and let $P \in \mathrm{GL}_n(\mathbb{C})$ be the basis change from (σ) to (ε). Assume for simplicity that we have ordered the vectors $\varepsilon_1, \ldots, \varepsilon_n$ in such a way that if $\varepsilon_i \in H_k \cap W$ and $\varepsilon_{i+1} \notin W$ then $\varepsilon_{i+1} \notin H_k$. Let $T = \mathrm{diag}(t_1, \ldots, t_n)$ be the diagonal integer matrix defined by $t_i = 1$ if and only if $\varepsilon_i \notin W$. With the simplifying assumption, the type of $\mathscr{E}^M$

is equal to $D-T$, including the ordering condition, and the Harder–Narasimhan filtration is exactly obtained by putting together the groups of vectors corresponding to equal values of $D-T$. Therefore the basis $(\sigma') = (z-s)^{D-T}(\varepsilon)$ spans a BG trivialization $\mathscr{F}'$ of $\mathscr{E}^M$ at s, and it is simultaneously a global basis of V. The transport $t_x(\mathscr{F}')$ is again a BG trivialization of $\mathscr{E}^M$ at x, but it needs not be logarithmic anymore. Since $\mathscr{E}$ is a weak solution, we have $\mathscr{E}_x = \mathscr{E}^M_x = \mathfrak{D}_x$. Therefore, there exists a lattice gauge transformation $P = I + P_1 t_x + P_2 t_x^2 + \cdots$ which sends the basis (σ') into its Υ-basis (ε'), where Υ is the Birkhoff form at x. The lattice M' spanned by $t_x^{D-T}(\sigma')$ is then necessarily logarithmic, according to Lemma 15. We can effectively determine M' by truncating the gauge P at order $d(M', D) - 1 = k_n - k_1 - 2$, and then applying Gantmacher's classical recursive formulæ (18). Then, the permutation lemma yields a monopole gauge transform Π at x so that the resulting trivialization $\overline{\mathscr{F}}$ is both BG and logarithmic. In this last basis, the connection has again the form (21), where the spectrum of the residue at x gives the type of the modified logarithmic bundle $\mathscr{E}^M$. □

Proposition 56 implies that iterated applications of this procedure describe the set of all weak solutions to the Riemann–Hilbert problem. The strong problem is solvable if and only if one of the bundles $\overline{\mathscr{F}}$ has a 0 residue at x in the orbit under these transformations.

5.5.3. *The general case.* For an arbitrary weak solution $\mathscr{E}$, we must start with the Deligne bundle $\mathfrak{D}$, for we only have the complete description of the local logarithmic lattices from the Deligne lattice. According to the description given in Proposition 52, any logarithmic lattice $N_s \in \mathbf{\Lambda}_s$ is given by an admissible pair $(\boldsymbol{F}, T)$ where $\boldsymbol{F}$ is a $\mathrm{Res}_s^{\mathfrak{D}}\nabla$-stable flag. If we put ourselves in the situation of Section 5.5.2, and consider a logarithmic BG trivialization $\mathscr{F}$ of $\mathfrak{D}$ at x, and identify $Y = \Gamma(X, \mathscr{F})$ to $\mathbb{C}^n$ by means of the basis (σ), then the flag $\boldsymbol{F}$ can be viewed as a flag in $\mathbb{C}^n$ stable under the matrix $\widetilde{L}_s$. As exposed in Theorem 65, the bundle $\mathscr{E}$ is then described by an element $(\boldsymbol{F}^s, D_s)_{s\in\mathcal{S}} \in \mathrm{RH}_\chi$ such that $\boldsymbol{F}^s \in \mathfrak{Fl}_{\psi_s}(Y)$ and $D_s \in \mathbb{Z}^n(\boldsymbol{F}^s)$. In order to actually construct the lattice N_s, one should in principle reach first a Birkhoff form Υ_z in $\Lambda = \mathfrak{D}_s$. Since the Deligne lattice is nonresonant, it is possible to do so by a lattice gauge P_s tangent to the identity, as described in Section 4.1.1. We know from Definition 14 that if we put $d_s = \Delta D_s - 1$, the local gauge P_s can be truncated to t_s-degree d_s, as remarked in the proof of Proposition 82. Assume for simplicity that $x = \infty \notin \mathcal{S}$. Taking $Q_s \in \mathrm{GL}_n(\mathbb{C})$ that brings (σ) to a basis respecting $\boldsymbol{F}^s$, the local gauge can be written as

$$P_s = Q_s\big(I + \cdots + P_{d_s}^{(s)}(z-s)^{d_s}\big).$$

Take a rational interpolation $Z \in \mathrm{GL}_n(\mathbb{C}(z))$ of the local gauges P_s to the prescribed orders d_s, and having only a singularity outside $\mathcal{S}$ at x (see, e.g., [van

Barel et al. 1994]). The global basis of sections $(\widetilde{\sigma})$ whose matrix in (σ) is P spans by construction a trivialization $\widetilde{\mathscr{F}}$ of $\mathscr{E}$ at x. Everything can be seen at x as explained in the following scheme. Putting $\Lambda = \mathfrak{D}_x = \mathscr{E}_x$, we have

$$\begin{array}{ccc}
\Lambda : (\sigma_x) & \xrightarrow{Z} & \widetilde{Y} = \Gamma(X, \widetilde{\mathscr{F}}) \\
\Big\downarrow P & & \Big\downarrow \Pi_1 \\
\Lambda : (\varepsilon) & \xrightarrow{t_x^D} & Y_{\tilde{M}} \\
\Big\downarrow Q & & \Big\downarrow \Pi_2 \\
\Lambda : (\tilde{\varepsilon}) & \xrightarrow{t_x^D} & Y'
\end{array}$$

The effect of having changed the stalks over $\mathscr{S}$ is translated in a purely local fashion by a change in the set $\mathbf{\Lambda}^0$. Indeed, when Λ represented the class of $[\mathfrak{D}]_x$, the germ (σ_x) was a diagonal shift of a global basis. In this scheme however, Λ represents the class $[\mathscr{E}]_x$, and we have now to apply the gauge Z, considered as an element of $\mathrm{GL}_n(\mathbb{C}((t_x)))$, to get a basis of global sections of $\widetilde{Y} = \Gamma(X, \widetilde{\mathscr{F}})$. The monopole Π_1 corresponds to the construction of a BG trivialization, as in Section 3.2.1. The second gauge Π_2, which can be constructed by the enhanced permutation lemma (Proposition 38), brings the system to an optional logarithmic BG trivialization, where the system is again Fuchsian.

References

[Abramenko and Brown 2008] P. Abramenko and K. S. Brown, *Buildings: Theory and applications*, Graduate Texts in Mathematics **248**, Springer, New York, 2008. MR 2009g:20055 Zbl 1214.20033

[André and Baldassarri 2001] Y. André and F. Baldassarri, *De Rham cohomology of differential modules on algebraic varieties*, Progress in Mathematics **189**, Birkhäuser, Basel, 2001. MR 2002h:14031 Zbl 0995.14003

[Anosov and Bolibrukh 1994] D. V. Anosov and A. A. Bolibruch, *The Riemann–Hilbert problem*, Aspects of Mathematics **22**, Friedr. Vieweg & Sohn, Braunschweig, 1994. MR 95d:32024 Zbl 0801.34002

[Babbitt and Varadarajan 1983] D. G. Babbitt and V. S. Varadarajan, "Formal reduction theory of meromorphic differential equations: A group theoretic view", *Pacific J. Math.* **109**:1 (1983), 1–80. MR 86b:34010 Zbl 0533.34010

[van Barel et al. 1994] M. van Barel, B. Beckermann, A. Bultheel, and G. Labahn, "Matrix rational interpolation with poles as interpolation points", pp. 137–148 in *Nonlinear numerical methods and rational approximation, II: Proceedings of the 3rd International Conference* (University of Antwerp, Belgium, 5–11 September 1993), edited by A. Cuyt, Mathematics and its Applications **296**, Kluwer Academic Publishers, Dordrecht, 1994. Zbl 0803.41006

[Boalch 2011] P. P. Boalch, "Riemann–Hilbert for tame complex parahoric connections", *Transform. Groups* **16**:1 (2011), 27–50. MR 2012m:14020 Zbl 1232.34117

[Bolibrukh 1990] A. A. Bolibrukh, "The Riemann–Hilbert problem", *Uspekhi Mat. Nauk* **45**:2(272) (1990), 3–47. In Russian; translated in *Russ. Math. Surv.* **45**:2 (1990), 1–58. MR 92j:14014 Zbl 0706.34005

[Bolibrukh et al. 2006] A. A. Bolibruch, S. Malek, and C. Mitschi, "On the generalized Riemann–Hilbert problem with irregular singularities", *Expo. Math.* **24**:3 (2006), 235–272. MR 2007i:34148 Zbl 1106.34061

[Corel 2004] E. Corel, "On Fuchs' relation for linear differential systems", *Compos. Math.* **140**:5 (2004), 1367–1398. MR 2005e:34270 Zbl 1070.34120

[Crawley-Boevey 2003] W. Crawley-Boevey, "On matrices in prescribed conjugacy classes with no common invariant subspace and sum zero", *Duke Math. J.* **118**:2 (2003), 339–352. MR 2004g:16014 Zbl 1046.15013

[Deligne 1970] P. Deligne, *Équations différentielles à points singuliers réguliers*, Lecture Notes in Mathematics **163**, Springer, Berlin, 1970. MR 54 #5232 Zbl 0244.14004

[Gantmacher 1959] F. R. Gantmacher, *The theory of matrices, II*, Chelsea, New York, 1959. MR 21 #6372c Zbl 0927.15002

[Garrett 1997] P. Garrett, *Buildings and classical groups*, Chapman-Hall, London, 1997. MR 98k: 20081 Zbl 0933.20019

[Gérard and Levelt 1973] R. Gérard and A. H. M. Levelt, "Invariants mesurant l'irrégularité en un point singulier des systèmes d'équations différentielles linéaires", *Ann. Inst. Fourier* (*Grenoble*) **23**:1 (1973), 157–195. MR 49 #10947 Zbl 0243.35016

[Ilyashenko and Yakovenko 2007] Y. Ilyashenko and S. Yakovenko, *Lectures on analytic differential equations*, Graduate Studies in Mathematics **86**, American Mathematical Society, Providence, RI, 2007. MR 2009b:34001 Zbl 1186.34001

[Kostov 1992] V. P. Kostov, "Fuchsian linear systems on $\mathbf{C}\mathrm{P}^1$ and the Riemann–Hilbert problem", *C. R. Acad. Sci. Paris Sér. I Math.* **315**:2 (1992), 143–148. MR 94a:34007 Zbl 0772.34007

[Krause 1996] M. Krause, "A simple proof of the Gale–Ryser theorem", *Amer. Math. Monthly* **103**:4 (1996), 335–337. MR 1383671 Zbl 0855.05032

[Levelt 1961] A. H. M. Levelt, "Hypergeometric functions, II", *Nederl. Akad. Wetensch. Proc. Ser. A 64 = Indag. Math.* **23** (1961), 373–385. MR 25 #1302 Zbl 0124.03602

[Malgrange 1996] B. Malgrange, "Connexions méromorphes, II: Le réseau canonique", *Invent. Math.* **124**:1-3 (1996), 367–387. MR 97h:32060 Zbl 0849.32003

[Plemelj 1964] J. Plemelj, *Problems in the sense of Riemann and Klein*, Interscience Tracts in Pure and Applied Mathematics **16**, Wiley, New York, 1964. MR 30 #5008 Zbl 0124.28203

[Ronan 1989] M. Ronan, *Lectures on buildings*, Perspectives in Mathematics **7**, Academic Press, Boston, MA, 1989. MR 90j:20001 Zbl 0694.51001

[Sabbah 2002] C. Sabbah, *Déformations isomonodromiques et variétés de Frobenius*, EDP Sciences, Paris, 2002. MR 2003m:32013 Zbl 1101.14001

[Treibich Kohn 1983] A. Treibich Kohn, "Un résultat de Plemelj", pp. 307–312 in *Mathematics and physics* (Paris, 1979/1982), edited by L. Boutet de Monvel et al., Progr. Math. **37**, Birkhäuser, Boston, 1983. MR 87c:32016 Zbl 0527.34007

Received December 3, 2010. Revised March 14, 2013.

ELIE COMPOINT

UNIVERSITÉ DES SCIENCES ET TECHNOLOGIES - LILLE 1
CITÉ SCIENTIFIQUE
59655 VILLENEUVE D'ASCQ
FRANCE
elie.compoint@math.univ-lille1.fr

EDUARDO COREL
UNIVERSITÉ D'EVRY-VAL-D'ESSONNE
IBGBI, 23 BOULEVARD DE FRANCE
91037 EVRY
FRANCE
eduardo.corel@genopole.cnrs.fr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 263, No. 2, 2013
dx.doi.org/10.2140/pjm.2013.263.353

ELLIPTIC ALIQUOT CYCLES OF FIXED LENGTH

NATHAN JONES

Silverman and Stange define the notion of an aliquot cycle of length L for a fixed elliptic curve E over $\mathbb{Q}$, and conjecture an order of magnitude for the function which counts such aliquot cycles. In the present note, we combine heuristics of Lang–Trotter with those of Koblitz to refine their conjecture to a precise asymptotic formula by specifying the appropriate constant. We give a criterion for positivity of the conjectural constant, as well as some numerical evidence for our conjecture.

1. Introduction

Let E be an elliptic curve over $\mathbb{Q}$ and fix a positive integer $L \geq 2$. In analogy with the classical notion of an aliquot cycle, Silverman and Stange [2011] define an L-tuple $(p_1, p_2, \ldots, p_L)$ of distinct positive integers to be an *aliquot cycle of length L for E* if each p_i is a prime number of good reduction for E,

$$p_1 = |E(\mathbb{F}_{p_L})| \qquad \text{and} \qquad p_{i+1} = |E(\mathbb{F}_{p_i})| \ \text{ for all } i \in \{1, 2, \ldots, L-1\},$$

which may be more succinctly written as

$$p_{i+1} = |E(\mathbb{F}_{p_i})| \quad \text{for all } i \in \mathbb{Z}/L\mathbb{Z}. \tag{1}$$

When $L = 2$, an aliquot cycle is also referred to as an *amicable pair for E*. As observed in [Silverman and Stange 2011, Remark 1.5], there is an intimate connection between aliquot cycles for E and elliptic divisibility sequences, which relate to generalizations of classical index divisibility questions about Lucas sequences (see also [Gottschlich 2012], which studies some distributional aspects of elliptic divisibility sequences).

It is of interest to know how common such aliquot cycles are, so we presently consider the function which counts aliquot cycles of fixed length for a fixed elliptic curve E over $\mathbb{Q}$. More precisely, define an aliquot cycle $(p_1, p_2, \ldots, p_L)$ to be *normalized* if $p_1 = \min\{p_i : 1 \leq i \leq L\}$, and then write

$$\pi_{E,L}(x) := \left|\left\{p_1 \leq x : \exists \text{ a normalized aliquot cycle } (p_1, p_2, \ldots, p_L) \text{ for } E\right\}\right|.$$

Work partially supported by the National Security Agency under grant H98230-12-1-0210.
MSC2010: 11G05.
Keywords: elliptic curve, aliquot cycle, amicable pair.

The behavior of $\pi_{E,L}(x)$ for large x depends heavily on whether or not E has complex multiplication (CM), as the following conjecture indicates.

Conjecture 1.1 (Silverman–Stange). Let E be an elliptic curve over $\mathbb{Q}$ and $L \geq 2$ a fixed integer, and assume that there are infinitely many primes p such that $|E(\mathbb{F}_p)|$ is prime. Then, as $x \to \infty$, one has

$$\pi_{E,L}(x) \quad \begin{cases} \asymp \dfrac{\sqrt{x}}{(\log x)^L} & \text{if } E \text{ has no CM,} \\ \sim A_E \dfrac{x}{(\log x)^2} & \text{if } E \text{ has CM and } L = 2, \end{cases}$$

where the implied constants in $\asymp$ are both positive and depend only on E and L, and A_E is a positive constant.

Remark 1.2. We may interpret the case $L = 1$ of (1) as describing primes p_1 for which $p_1 = |E(\mathbb{F}_{p_1})|$. Such primes are called *anomalous* primes and have been considered in [Mazur 1972]. The asymptotic count for anomalous primes up to x is a special case of a conjecture of Lang and Trotter [1976].

Silverman and Stange [2011] focus on the intricacies of the CM case, proving that if E has CM, $j_E \neq 0$ and $L \geq 3$, then any normalized aliquot cycle $(p_1, p_2, \ldots, p_L)$ for E must have $p_1 < 5$ (so, in particular, $\pi_{E,L}(x) = O(1)$). The case $j_E = 0$ is apparently more complicated, and no proof is given that $\pi_{E,L}(x) = O(1)$ when $j_E = 0$ and $L > 3$.

In this note, we refine Conjecture 1.1 to an asymptotic formula in the non-CM case. Heuristics will be developed which lead to the following conjecture.

Conjecture 1.3. Let E be an elliptic curve over $\mathbb{Q}$ without complex multiplication and $L \geq 2$ a fixed integer. Then there is a nonnegative real constant $C_{E,L} \geq 0$ (see (5) below) so that, as $x \longrightarrow \infty$,

$$\pi_{E,L}(x) \sim C_{E,L} \int_2^x \frac{1}{2\sqrt{t}(\log t)^L}\, dt.$$

Remark 1.4. It is possible for the constant $C_{E,L}$ to be zero, in which case the limit $\lim_{x\to\infty} \pi_{E,L}(x)$ is provably finite. Thus, in case $C_{E,L} = 0$, let us interpret the above asymptotic to mean that $\lim_{x\to\infty} \pi_{E,L}(x) < \infty$.

Remark 1.5. By integration by parts, one has

$$\int_2^x \frac{1}{2\sqrt{t}(\log t)^L}\, dt \;=\; \frac{\sqrt{x}}{(\log x)^L} + O\bigg(\frac{\sqrt{x}}{(\log x)^{L+1}}\bigg).$$

Thus, Conjecture 1.3 is consistent with Conjecture 1.1. In practice, the error term

$$\left|\pi_{E,L}(x) - C_{E,L} \int_2^x \frac{1}{2\sqrt{t}(\log t)^L}\, dt\right|$$

E	$x = 10^6$	$x = 10^8$	$x = 10^{10}$	$x = 10^{12}$	$x = 10^{13}$
$E_1 : y^2 + y = x^3 - x$	0	1	16	115	332
$E_2 : y^2 = x^3 + 6x - 2$	0	5	32	208	564
$E_3 : y^2 = x^3 - 3x + 4$	0	0	0	0	0

Table 1. Values of $\pi_{E,2}(x)$.

should be smaller than $\left|\pi_{E,L}(x) - C_{E,L}\frac{\sqrt{x}}{(\log x)^L}\right|$, just as in the case of the prime number theorem.

Consider Table 1, which lists the values of $\pi_{E,2}(x)$ for a few non-CM curves E and various magnitudes x. Note that $\pi_{E_2,2}(x)$ is larger than $\pi_{E_1,2}(x)$. This difference is explained by the associated constants appearing in Conjecture 1.3. Indeed, a computation shows that

$$\frac{C_{E_2,2}}{C_{E_1,2}} \approx 1.714.$$

Also note that $\pi_{E_3,2}(10^{13}) = 0$. The additional fact that

$$\left|\left\{p \le 10^{12} : p \text{ is of good reduction for } E_3 \text{ and } |E_3(F_p)| \text{ is prime}\right\}\right| = 715,698,540$$

indicates that there probably are infinitely many primes p for which $|E_3(\mathbb{F}_p)|$ is prime, in which case the above data suggests that E_3 might be a counterexample to Conjecture 1.1. We will later see that $C_{E_3,2} = 0$, and that E_3 is indeed a counterexample, assuming a conjecture of Koblitz on the primality of $|E(\mathbb{F}_p)|$.

Remark 1.6. The heuristics which lead to Conjecture 1.3 are in the style of Koblitz and Lang–Trotter, whose conjectures have been proven "on average over elliptic curves E" (see [Balog et al. 2011; David and Pappalardi 1999]). It might be interesting to see if one could also prove an average version of Conjecture 1.3.

1.1. ***Positivity of $C_{E,L}$ and a directed graph $\mathcal{G}_E$.*** In the interest of characterizing the non-CM elliptic curves which have infinitely many aliquot cycles of length L, we will state a graph-theoretic criterion for positivity of $C_{E,L}$. Recall that a *directed graph* $\mathcal{G}$ is a pair $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{V}(\mathcal{G})$ is an arbitrary set of *vertices* and $\mathcal{E} = \mathcal{E}(\mathcal{G}) \subseteq \mathcal{V} \times \mathcal{V}$ is a subset of *directed edges*. The sequence of vertices $(v_1, v_2, v_3, \ldots, v_n)$ is a *closed walk of length n* if and only if $(v_i, v_{i+1}) \in \mathcal{E}$ for each $i \in \mathbb{Z}/n\mathbb{Z} = \{1, 2, 3, \ldots, n\}$. Note that closed walks may have repeated vertices. For instance, if $(v, v) \in \mathcal{E}$ for some vertex v (i.e., if $\mathcal{G}$ has a *loop* at a vertex v), then $\mathcal{G}$ has closed walks of any length.

We will associate to an elliptic curve E a directed graph $\mathcal{G}_E$. First, consider the n-th division field $\mathbb{Q}(E[n])$ of E, obtained by adjoining to $\mathbb{Q}$ the x and y-coordinates

of the n-torsion $E[n]$ of a given Weierstrass model of E. The extension $\mathbb{Q}(E[n])$ is Galois over $\mathbb{Q}$, and once we fix a basis over $\mathbb{Z}/n\mathbb{Z}$ of $E[n]$, we may view

$$\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q}) \subseteq \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z}). \tag{2}$$

We will now attach to $\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})$ a directed graph $\mathcal{G}_E(n)$. Viewing Galois automorphisms as 2×2 matrices via (2), the vertex set $\mathcal{V}(n)$ of our graph $\mathcal{G}_E(n)$ is

$$\mathcal{V}(n) := \left\{(t, d) \in \mathbb{Z}/n\mathbb{Z} \times (\mathbb{Z}/n\mathbb{Z})^\times : \exists g \in \mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q}) \text{ with } \operatorname{tr} g = t, \det g = d\right\}.$$

We define the set $\mathcal{E}(n) \subseteq \mathcal{V}(n) \times \mathcal{V}(n)$ of directed edges by declaring that $(v_1, v_2) \in \mathcal{E}(n)$ if and only if $d_1 + 1 - t_1 = d_2$, where $v_i = (t_i, d_i) \in \mathcal{V}(n)$.

Let m_E denote the *torsion conductor* of E, which is defined as the smallest positive integer m for which

$$\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q}) = \pi^{-1}\big(\mathrm{Gal}(\mathbb{Q}(E[\gcd(m, n)])/\mathbb{Q})\big) \quad \text{for all } n \in \mathbb{Z}_{>0},$$

where $\pi : \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z}) \to \mathrm{GL}_2(\mathbb{Z}/\gcd(m, n)\mathbb{Z})$ is the canonical projection. (The existence of a torsion conductor m_E for a non-CM elliptic curve E is a celebrated theorem of Serre [1972].) Finally, we define the directed graph $\mathcal{G}_E$ to be the above graph at level m_E:

$$\mathcal{G}_E := \mathcal{G}_E(m_E).$$

The following version of Conjecture 1.3 states a criterion for positivity of $C_{E,L}$ in terms of the directed graph $\mathcal{G}_E$.

Conjecture 1.7. Let E be an elliptic curve over $\mathbb{Q}$ without complex multiplication and $L \geq 2$ a fixed integer. Suppose that the directed graph $\mathcal{G}_E$ has a closed walk of length L. Then there are infinitely many aliquot cycles of length L for E. More precisely, there is a positive constant $C_{E,L} > 0$ so that, as $x \longrightarrow \infty$,

$$\pi_{E,L}(x) \sim C_{E,L} \int_2^x \frac{1}{2\sqrt{t}(\log t)^L}\, dt.$$

Remark 1.8. If $\mathcal{G}_E$ does not have a closed walk of length L, then $C_{E,L} = 0$ and there are at most finitely many aliquot cycles of length L for E (see Proposition 2.6).

In Section 2, we will write down the constant $C_{E,L}$ explicitly as an "almost Euler product" and discuss its positivity in terms of the graph $\mathcal{G}_E$. In Section 3, we will develop the heuristics which lead to Conjecture 1.3. In Section 4, we will provide some numerical evidence for Conjecture 1.3 by examining the order of magnitude of $\pi_{E,L}(x) - C_{E,L} \displaystyle\int_2^x \frac{1}{2\sqrt{t}(\log t)^L}\, dt$ for various elliptic curves E and $L \in \{2, 3\}$.

2. The constant

We now describe in detail the constant $C_{E,L}$. The next lemma allows us to interpret (1) in terms of the Frobenius automorphisms[1] $\mathrm{Frob}_{\mathbb{Q}(E[n])}(p_i) \in \mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})$ attached to the various primes p_i. Recall the trace of Frobenius $a_p(E) \in \mathbb{Z}$, which satisfies the equation

$$|E(\mathbb{F}_p)| = p + 1 - a_p(E)$$

as well as the Hasse bound

$$|a_p(E)| \le 2\sqrt{p}. \tag{3}$$

Lemma 2.1 [Serre 1968, IV-4–IV-5]. *For any positive integer n and any prime p of good reduction for E which does not divide n, p is unramified in $\mathbb{Q}(E[n])$ and, for any Frobenius automorphism $\mathrm{Frob}_{\mathbb{Q}(E[n])}(p) \in \mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})$, we have*

$$\mathrm{tr}(\mathrm{Frob}_{\mathbb{Q}(E[n])}(p)) \equiv a_p(E) \bmod n \quad \textit{and} \quad \det(\mathrm{Frob}_{\mathbb{Q}(E[n])}(p)) \equiv p \bmod n.$$

For any subset $G \subseteq \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$, define

$$G^L_{\text{ali-cycle}} := \big\{(g_1, g_2, \dots, g_L) \in G^L : \forall i \in \mathbb{Z}/L\mathbb{Z}, \det(g_{i+1}) = \det(g_i) + 1 - \mathrm{tr}(g_i)\big\}.$$

Note that, by Lemma 2.1, if $(p_1, p_2, \dots, p_L)$ is an aliquot cycle of length L for E, then

$$\big(\mathrm{Frob}_{\mathbb{Q}(E[n])}(p_1), \mathrm{Frob}_{\mathbb{Q}(E[n])}(p_2), \dots, \mathrm{Frob}_{\mathbb{Q}(E[n])}(p_L)\big) \in \mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^L_{\text{ali-cycle}}. \tag{4}$$

Next, let $\phi(x) := \frac{2}{\pi}\sqrt{1-x^2}$ be the distribution function of Sato–Tate, which (assuming E has no CM) conjecturally[2] satisfies

$$\lim_{x\to\infty} \frac{\big|\big\{p \le x : \frac{a_p(E)}{2\sqrt{p}} \in I \subseteq [-1,1]\big\}\big|}{|\{p \le x\}|} = \int_I \phi(x)\,dx.$$

In other words, ϕ is the density function of $a_p(E)/2\sqrt{p}$, viewed as a random variable. Denote by $\phi_L := \phi * \phi * \cdots * \phi$ the L-fold convolution of ϕ with itself,

[1] The Frobenius automorphism in

$$\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})$$

attached to an unramified rational prime p is only defined up to conjugation in $\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})$. Here and throughout the paper, we understand $\mathrm{Frob}_{\mathbb{Q}(E[n])}(p)$ to be any choice of such a Frobenius automorphism.

[2] Assuming E has nonintegral j-invariant, the Sato–Tate conjecture is now a theorem of L. Clozel, M. Harris, N. Shepherd-Barron, and R. Taylor (see [Taylor 2008] and the references therein).

which (again assuming the Sato–Tate conjecture) is the density function of the random variable

$$\sum_{i=1}^{L} \frac{a_{p_i}(E)}{2\sqrt{p_i}},$$

provided the various terms $a_{p_i}(E)/2\sqrt{p_i}$ are "statistically independent." Since the primes $p_1, p_2, \ldots, p_L$ belonging to an aliquot cycle must be close to one another (i.e., within $\approx L\sqrt{t}$ of one another where $p_1 \approx t$, by the Hasse bound (3)), we are really assuming statistical independence *in short intervals* of the various terms $a_{p_i}(E)/2\sqrt{p_i}$. Finally, for a positive integer k, put

$$n_k := \prod_{p \le k} p^k.$$

In Section 3, we will develop heuristics which predict Conjecture 1.3, with

$$C_{E,L} := \frac{\phi_L(0)}{L} \cdot \lim_{k\to\infty} \frac{n_k^L \left|\mathrm{Gal}(\mathbb{Q}(E[n_k])/\mathbb{Q})^L_{\text{ali-cycle}}\right|}{\left|\mathrm{Gal}(\mathbb{Q}(E[n_k])/\mathbb{Q})^L\right|}. \tag{5}$$

2.1. ***The constant as a product.*** We will presently prove the following proposition, which gives a more explicit expression of $C_{E,L}$ as a convergent Euler product. Recall that m_E denotes the torsion conductor of E, i.e., the smallest positive integer m for which

$$\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q}) = \pi^{-1}\big(\mathrm{Gal}(\mathbb{Q}(E[\gcd(m,n)])/\mathbb{Q})\big) \quad \text{for all } n \in \mathbb{Z}_{>0},$$

where $\pi : \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z}) \to \mathrm{GL}_2(\mathbb{Z}/\gcd(m,n)\mathbb{Z})$ is the canonical projection.

Proposition 2.2. *For a positive integer k, let $n_k := \prod_{p\le k} p^k$. Then one has*

$$\lim_{k\to\infty} \frac{n_k^L \left|\mathrm{Gal}(\mathbb{Q}(E[n_k])/\mathbb{Q})^L_{\text{ali-cycle}}\right|}{\left|\mathrm{Gal}(\mathbb{Q}(E[n_k])/\mathbb{Q})^L\right|} = \frac{m_E^L \left|\mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})^L_{\text{ali-cycle}}\right|}{\left|\mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})^L\right|} \cdot \prod_{l\nmid m_E} \frac{l^L \left|\mathrm{GL}_2(\mathbb{F}_l)^L_{\text{ali-cycle}}\right|}{\left|\mathrm{GL}_2(\mathbb{F}_l)^L\right|}.$$

Furthermore,

$$0 < \frac{l^L \left|\mathrm{GL}_2(\mathbb{F}_l)^L_{\text{ali-cycle}}\right|}{\left|\mathrm{GL}_2(\mathbb{F}_l)^L\right|} = 1 + O_L\left(\frac{1}{l^2}\right), \tag{6}$$

so the infinite product $\displaystyle\prod_{l\nmid m_E} \frac{l^L \left|\mathrm{GL}_2(\mathbb{F}_l)^L_{\text{ali-cycle}}\right|}{\left|\mathrm{GL}_2(\mathbb{F}_l)^L\right|}$ converges absolutely.

The proof of Proposition 2.2 involves the following two lemmas.

Lemma 2.3. *Let n_1 and n_2 be relatively prime positive integers, and pick any subgroups $G_1 \subseteq \mathrm{GL}_2(\mathbb{Z}/n_1\mathbb{Z})$ and $G_2 \subseteq \mathrm{GL}_2(\mathbb{Z}/n_2\mathbb{Z})$. Then, viewing $G_1 \times G_2 \subseteq \mathrm{GL}_2(\mathbb{Z}/n_1n_2\mathbb{Z})$, one has*

$$(G_1 \times G_2)^L_{\text{ali-cycle}} = (G_1)^L_{\text{ali-cycle}} \times (G_2)^L_{\text{ali-cycle}}.$$

Proof. Let $\iota : \mathrm{GL}_2(\mathbb{Z}/n_1\mathbb{Z}) \times \mathrm{GL}_2(\mathbb{Z}/n_2\mathbb{Z}) \to \mathrm{GL}_2(\mathbb{Z}/n_1n_2\mathbb{Z})$ be the isomorphism of the Chinese remainder theorem, and set $G := \iota(G_1 \times G_2)$. For each L-tuple $(g_i)_i \in G^L$, we have

$$\begin{aligned}\det g_{i+1} \equiv \det g_i + 1 - \operatorname{tr} g_i \pmod{n_1n_2} \quad \text{for all } i \in \mathbb{Z}/L\mathbb{Z}\\ \iff \quad \left\{\begin{array}{l}\det g_{i+1} \equiv \det g_i + 1 - \operatorname{tr} g_i \pmod{n_1}\\ \det g_{i+1} \equiv \det g_i + 1 - \operatorname{tr} g_i \pmod{n_2}\end{array}\right\} \text{ for all } i \in \mathbb{Z}/L\mathbb{Z}.\end{aligned}$$

This implies the conclusion of Lemma 2.3. □

Lemma 2.4. *Let n be a positive integer and n' any multiple of n such that, for every prime number l, $l \mid n' \Rightarrow l \mid n$. Let $\pi : \mathrm{GL}_2(\mathbb{Z}/n'\mathbb{Z}) \to \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$ denote the canonical projection and let $G \subseteq \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$ be any subgroup. Then one has*

$$\frac{(n')^L \left|(\pi^{-1}(G))^L_{\text{ali-cycle}}\right|}{|\pi^{-1}(G)^L|} = \frac{n^L \left|G^L_{\text{ali-cycle}}\right|}{|G^L|}. \tag{7}$$

Proof. By induction, it suffices to check the case $n' = ln$, where l is some prime dividing n. In this case, since $|\pi^{-1}(G)| = l^4|G|$, (7) is equivalent to

$$\left|(\pi^{-1}(G))^L_{\text{ali-cycle}}\right| = l^{3L}\left|G^L_{\text{ali-cycle}}\right|, \tag{8}$$

which we now show. Fix an element $g = (g_1, g_2, \ldots, g_L) \in G^L_{\text{ali-cycle}}$, and note that any element $g' \in \pi^{-1}(g)$ has the form

$$g' = (g'_1, g'_2, \ldots, g'_L) = \big(\tilde{g}_1(I + nA_1), \tilde{g}_2(I + nA_2), \ldots, \tilde{g}_L(I + nA_L)\big) \in \pi^{-1}(g),$$

where for each i, $\tilde{g}_i$ is any fixed lift to $\mathrm{GL}_2(\mathbb{Z}/l n\mathbb{Z})$ of g_i, and $A_i \in M_{2\times 2}(\mathbb{F}_l)$ is arbitrary. We will presently determine the exact conditions on the A_i which force $(g'_1, g'_2, \ldots, g'_L) \in (\pi^{-1}(G))^L_{\text{ali-cycle}}$. First, since $(g_1, g_2, \ldots, g_L) \in G^L_{\text{ali-cycle}}$, we must have

$$g_i \pmod{l} \notin \{0, I\} \quad \text{for each } i \in \mathbb{Z}/L\mathbb{Z} \tag{9}$$

and furthermore, the quantity

$$\gamma_i := \frac{\det \tilde{g}_{i+1} - \det \tilde{g}_i - 1 + \operatorname{tr} \tilde{g}_i}{n} \in \mathbb{F}_l$$

is well-defined. One checks that

$$\begin{aligned}\det g'_{i+1} \equiv \det g'_i + 1 - \operatorname{tr} g'_i \bmod ln\\ \iff \quad \gamma_i \equiv -\det g_{i+1} \cdot \operatorname{tr} A_{i+1} + \det g_i \cdot \operatorname{tr} A_i - \operatorname{tr}(g_i A_i) \bmod l.\end{aligned} \tag{10}$$

The condition on the right-hand side is (affine) linear in the coefficients of A_{i+1} and A_i. We consider the linear transformation

$$T : \mathbb{F}_l^{4L} \simeq M_{2\times 2}(\mathbb{F}_l)^L \to \mathbb{F}_l^L,$$

given by

$$(A_i)_{i=1}^L \mapsto \big(-\det g_{i+1} \cdot \operatorname{tr} A_{i+1} + \det g_i \cdot \operatorname{tr} A_i - \operatorname{tr}(g_i A_i)\big)_{i=1}^L.$$

In light of (10), the condition (8) will follow from the surjectivity of the above linear transformation, which we now verify. Writing coordinates as

$$g_i =: \begin{pmatrix} x_i & y_i \\ z_i & w_i \end{pmatrix} \quad \text{and} \quad A_i =: \begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix},$$

we have

$$T((A_i)) = \big((\det g_i - x_i)a_i + (\det g_i - w_i)d_i - y_i c_i - z_i b_i - \det g_{i+1}a_{i+1} - \det g_{i+1}d_{i+1}\big).$$

By (9), at least one of $\det g_i - x_i$, $\det g_i - w_i$, y_i and z_i must be nonzero modulo l, and so

$$T\big(\{0\}\times\cdots\times\{0\}\times M_{2\times 2}(\mathbb{F}_l)\times\{0\}\times\cdots\times\{0\}\big) = \{0\}\times\cdots\times\{0\}\times\mathbb{F}_l\times\{0\}\times\cdots\times\{0\},$$

where the nonzero entries correspond to the same index i. In particular, the linear transformation in question is surjective and we have verified (8), finishing the proof of Lemma 2.4. □

Proof of Proposition 2.2. Choose k large enough so that $m_E \mid n_k$, and write $n_k = n_k^{(1)} \cdot n_k^{(2)}$, where $n_k^{(1)}$ is divisible by primes dividing m_E and $\gcd(m_E, n_k^{(2)}) = 1$. By definition of m_E, we then have

$$\operatorname{Gal}(\mathbb{Q}(E[n_k])/\mathbb{Q}) \simeq \pi^{-1}\big(\operatorname{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})\big) \times \prod_{\substack{l^k \parallel n_k \\ l \nmid m_E}} \mathrm{GL}_2(\mathbb{Z}/l^k\mathbb{Z}),$$

where $\pi : \mathrm{GL}_2(\mathbb{Z}/n_k^{(1)}\mathbb{Z}) \to \mathrm{GL}_2(\mathbb{Z}/m_E\mathbb{Z})$ is the canonical projection. By Lemmas 2.3 and 2.4, we have

$$\frac{n_k^L \big|\operatorname{Gal}(\mathbb{Q}(E[n_k])/\mathbb{Q})^L_{\text{ali-cycle}}\big|}{\big|\operatorname{Gal}(\mathbb{Q}(E[n_k])/\mathbb{Q})^L\big|} = \frac{m_E^L \big|\operatorname{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})^L_{\text{ali-cycle}}\big|}{\big|\operatorname{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})^L\big|} \cdot \prod_{\substack{l \mid n_k \\ l \nmid m_E}} \frac{l^L \big|\operatorname{Gal}(\mathbb{Q}(E[l])/\mathbb{Q})^L_{\text{ali-cycle}}\big|}{\big|\operatorname{Gal}(\mathbb{Q}(E[l])/\mathbb{Q})^L\big|}.$$

Taking the limit as $k \to \infty$, we arrive at the product representation of $C_{E,L}$ stated in Proposition 2.2. We leave the verification of (6) as an exercise. □

2.2. ***Positivity of the constant.*** We will now discuss the positivity of $C_{E,L}$. The following corollary of Proposition 2.2 is immediate.

Corollary 2.5. *One has*

$$C_{E,L} > 0 \iff \mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})^L_{\text{ali-cycle}} \neq \varnothing.$$

We will now prove the following proposition, which allows one to deduce Conjecture 1.7 from Conjecture 1.3.

Proposition 2.6. *For any non-CM elliptic curve E over $\mathbb{Q}$, one has*

$$C_{E,L} > 0 \iff \mathcal{G}_E \text{ has a closed walk of length } L. \tag{11}$$

Furthermore, if $\mathcal{G}_E$ has no closed walks of length L, then there are only finitely many aliquot cycles $(p_1, p_2, \ldots, p_L)$ of length L for E.

Proof. First we prove (11). By Corollary 2.5, we are reduced to showing that

$$\mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})^L_{\text{ali-cycle}} \neq \varnothing \iff \mathcal{G}_E \text{ has a closed walk of length } L. \tag{12}$$

The mapping

$$\mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q}) \to \mathcal{V}(\mathcal{G}_E),$$
$$g \mapsto (\operatorname{tr} g, \det g)$$

induces a mapping $\mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})^L_{\text{ali-cycle}} \longrightarrow \{\text{closed walks of length } L \text{ in } \mathcal{G}_E\}$. Thus, if $\mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})^L_{\text{ali-cycle}} \neq \varnothing$ then $\mathcal{G}_E$ has a closed walk of length L. Conversely, suppose $\mathcal{G}_E$ has a closed walk $(v_1, v_2, v_3, \ldots, v_L)$ of length L. Recall that $\mathcal{V} = \mathbb{Z}/m_E\mathbb{Z} \times (\mathbb{Z}/m_E\mathbb{Z})^\times$ and write $v_i = (t_i, d_i)$. Choosing any element $g_i \in \mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})$ with $\operatorname{tr} g_i = t_i$ and $\det g_i = d_i$, we have then constructed an element $(g_1, g_2, \ldots, g_L) \in \mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})^L_{\text{ali-cycle}}$, so $\mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})^L_{\text{ali-cycle}} \neq \varnothing$. By Corollary 2.5, we conclude the proof of (11).

To see why the nonexistence of closed walks of length L in $\mathcal{G}_E$ implies that $\lim_{x\to\infty} \pi_{E,L}(x) < \infty$, note that, by (12), one has $\mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})^L_{\text{ali-cycle}} = \varnothing$. But then (4) implies that $\lim_{x\to\infty} \pi_{E,L}(x) < \infty$, and the proof of Proposition 2.6 is complete. □

3. Heuristics

We will construct a probabilistic model in the style of [Koblitz 1988] and [Lang and Trotter 1976]. We shall call the L-tuple $(p_1, p_2, \ldots, p_L)$ of distinct prime numbers an *aliquot sequence of length L for E* if it satisfies

$$p_{i+1} = |E(\mathbb{F}_{p_i})| \quad \text{for all } i \in \{1, 2, \ldots L-1\}.$$

Thus, an aliquot cycle of length L is an aliquot sequence of length L which additionally satisfies $p_1 = |E(\mathbb{F}_{p_L})|$. Suppose that $(p_1, p_2, \ldots, p_L)$ is an aliquot

sequence of length L for E. By substituting $p_2 = p_1 + 1 - a_{p_1}(E)$ into the equation $p_3 = p_2 + 1 - a_{p_2}(E)$, one finds that $p_3 = p_1 + 2 - (a_{p_1}(E) + a_{p_2}(E))$, and continuing in this manner one obtains

$$p_1 = |E(\mathbb{F}_{p_L})| \iff \sum_{j=1}^{L} a_{p_j}(E) = L. \tag{13}$$

Thus, a given L-tuple $(p_1, p_2, \dots, p_L)$ of positive integers is an aliquot cycle of length L for E if and only if the following conditions hold:

(1_L) the L-tuple $(p_1, p_2, \dots, p_L)$ is an aliquot sequence of length L for E;

(2_L) one has $\sum_{j=1}^{L} a_{p_j}(E) = L$.

Consider the following condition, which generalizes condition (2_L) above by replacing L with an arbitrary fixed integer r:

$(2'_L)$ one has $\sum_{j=1}^{L} a_{p_j}(E) = r$.

We now develop the heuristic "probability" that a given L-tuple $(p_1, p_2, \dots, p_L)$ of positive integers satisfies (1_L) and $(2'_L)$. First, we must gather some notation. Fix a positive integer n and elements $a, b \in \mathbb{Z}/n\mathbb{Z}$. For any subset $S \subseteq \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$, let

$$S_{\mathcal{N}=a} := \{g \in S : \det(g) + 1 - \mathrm{tr}(g) = a\} = \{g \in S : \det(I - g) = a\},$$
$$S^{\det=b} := \{g \in S : \det(g) = b\}, \quad S^{\det=b}_{\mathcal{N}=a} := S_{\mathcal{N}=a} \cap S^{\det=b}.$$

Finally, for $L \geq 1$ and $G \subseteq \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$, put

$$G^L_{\text{ali-sequence}} := \big\{(g_1, g_2, \dots, g_L) \in G^L : \text{for all } i \in \{1, 2, \dots, L-1\}, \det(g_{i+1}) = \det(g_i) + 1 - \mathrm{tr}(g_i)\big\}.$$

Note that when $L = 1$, the defining conditions become empty and we have $G^{L=1}_{\text{ali-sequence}} = G$. For a general $L \geq 1$, note that any aliquot sequence $(p_1, p_2, \dots, p_L)$ for E will satisfy

$$\big(\mathrm{Frob}_{\mathbb{Q}(E[n])}(p_1), \mathrm{Frob}_{\mathbb{Q}(E[n])}(p_2), \dots, \mathrm{Frob}_{\mathbb{Q}(E[n])}(p_L)\big) \in \mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^L_{\text{ali-sequence}}.$$

Finally, for a fixed integer r, define

$$G^{L,\sum \mathrm{tr}=r}_{\text{ali-sequence}} := \left\{(g_1, g_2, \dots, g_L) \in G^L_{\text{ali-sequence}} : \sum_{i=1}^{L} \mathrm{tr}(g_i) \equiv r \bmod n\right\}.$$

We will presently derive an expression for the probability

$$\mathcal{P}_{(1_L),(2'_L)}(t) := \mathrm{Prob}\big((p_1, p_2, \dots, p_L) \text{ satisfies } (1_L) \text{ and } (2'_L), \text{ given that } p_1 \approx t\big).$$

Putting $\mathcal{P}_{(1_L)}(t)$ for the probability that $(p_1, p_2, \ldots, p_L)$ satisfies (1_L) above, and $\mathcal{P}_{(2'_L)}^{\text{given } (1_L)}(t)$ for the conditional probability that $(p_1, p_2, \ldots, p_L)$ satisfies $(2'_L)$, given that it satisfies (1_L), we have

$$\mathcal{P}_{(1_L),(2'_L)}(t) = \mathcal{P}_{(1_L)}(t) \cdot \mathcal{P}_{(2'_L)}^{\text{given } (1_L)}(t). \tag{14}$$

In Section 3.1 below, we will derive the probability formula

$$\mathcal{P}_{(1_L)}(t) \approx \frac{n^{L-1} \cdot \left|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^L_{\text{ali-sequence}}\right|}{\left|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^L\right|} \cdot \frac{1}{(\log t)^L}. \tag{15}$$

Following this, in Section 3.2, we will derive

$$\mathcal{P}_{(2'_L)}^{\text{given } (1_L)}(t) \approx \phi_L\left(\frac{r}{2\sqrt{t}}\right) \frac{n \cdot \left|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{L,\,\sum \text{tr}\,=\,r}_{\text{ali-sequence}}\right|}{\left|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^L_{\text{ali-sequence}}\right|} \cdot \frac{1}{2\sqrt{t}}. \tag{16}$$

Before deriving (15) and (16), we will now observe that, taken together, they lead to Conjecture 1.3. Indeed, using (14), (15) and (16), one concludes

$$\mathcal{P}_{(1_L),(2'_L)}(t) \approx \phi_L\left(\frac{r}{2\sqrt{t}}\right) \cdot \frac{n^L \left|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{L,\,\sum \text{tr}\,=\,r}_{\text{ali-sequence}}\right|}{\left|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^L\right|} \cdot \frac{1}{2\sqrt{t}(\log t)^L}.$$

Just as with (13), one verifies that, for each $(g_1, g_2, \ldots, g_L) \in \text{GL}_2(\mathbb{Z}/n\mathbb{Z})^L_{\text{ali-sequence}}$, one has

$$\det(g_L) + 1 - \text{tr}(g_L) = \det g_1 \iff \sum_{i=1}^{L} \text{tr}(g_i) \equiv L \bmod n.$$

It follows that $\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^L_{\text{ali-cycle}} = \text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{L,\,\sum \text{tr}\,=\,L}_{\text{ali-sequence}}$. Thus, putting $r = L$, $n = n_k$ and taking the limit as $k \to \infty$, one arrives at

$$\mathcal{P}_{(1_L),(2_L)}(t) \approx \phi_L\left(\frac{L}{2\sqrt{t}}\right) \cdot \lim_{k\to\infty} \frac{n_k^L \left|\text{Gal}(\mathbb{Q}(E[n_k])/\mathbb{Q})^L_{\text{ali-cycle}}\right|}{\left|\text{Gal}(\mathbb{Q}(E[n_k])/\mathbb{Q})^L\right|} \cdot \frac{1}{2\sqrt{t}(\log t)^L}.$$

Thus, using

$$\pi_{E,L}(x) \approx \frac{1}{L} \int_2^x \mathcal{P}_{(1_L),(2_L)}(t)\, dt,$$

one arrives at Conjecture 1.3. The reason for the extra factor of L in the denominator above is that $\pi_{E,L}(x)$ counts *normalized* aliquot cycles, whereas the heuristic probabilities above do not take normalization into account. Also, since L is fixed, one verifies that the estimation $\phi_L(L/(2\sqrt{t})) \approx \phi_L(0)$ does not affect the asymptotic.

3.1. ***The probability that*** $(p_1, p_2, \dots, p_L)$ ***satisfies*** (1_L)**.** We will now derive a refined probability formula which implies (15). Fix a vector $\boldsymbol{a} = (a_2, a_3, \dots, a_L) \in ((\mathbb{Z}/n\mathbb{Z})^\times)^{L-1}$, and consider the probability

$$\begin{aligned}&\mathcal{P}^{\boldsymbol{a}}_{(1_L)}(t)\\ &:= \mathrm{Prob}\big((p_1, p_2, \dots p_L) \text{ satisfies } (1_L) \text{ and for all } i \in \{2, 3, \dots, L\},\ p_i \equiv a_i \bmod n\big)\end{aligned}$$

and (for any subset $G \subseteq \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$) the subset

$$G^{L,\boldsymbol{a}}_{\text{ali-sequence}} := \big\{(g_1, g_2, \dots, g_L) \in G^L_{\text{ali-sequence}} : \text{for all } i \in \{2, 3, \dots, L\}, \det(g_i) = a_i\big\}.$$

In case $L = 1$, the vector $\boldsymbol{a} \in ((\mathbb{Z}/n\mathbb{Z})^\times)^0$ is nonexistent, and as before we interpret the empty condition as $G^{1,\boldsymbol{a}}_{\text{ali-sequence}} = G$. Also note the decomposition

$$G^{L,\boldsymbol{a}}_{\text{ali-sequence}} = G_{\mathcal{N}=a_2} \times G^{\det=a_2}_{\mathcal{N}=a_3} \times G^{\det=a_3}_{\mathcal{N}=a_4} \times \cdots \times G^{\det=a_{L-1}}_{\mathcal{N}=a_L} \times G^{\det=a_L}. \tag{17}$$

Finally, note that if $\boldsymbol{a}_1 \neq \boldsymbol{a}_2$, then $G^{L,\boldsymbol{a}_1}_{\text{ali-sequence}} \cap G^{L,\boldsymbol{a}_2}_{\text{ali-sequence}} = \varnothing$, and so we have a disjoint union

$$G^L_{\text{ali-sequence}} = \bigsqcup_{\boldsymbol{a} \in ((\mathbb{Z}/n\mathbb{Z})^\times)^{L-1}} G^{L,\boldsymbol{a}}_{\text{ali-sequence}}.$$

For similar reasons, we have

$$\mathcal{P}_{(1_L)}(t) = \sum_{\boldsymbol{a} \in ((\mathbb{Z}/n\mathbb{Z})^\times)^{L-1}} \mathcal{P}^{\boldsymbol{a}}_{(1_L)}(t).$$

Thus, (15) will follow from

$$\mathcal{P}^{\boldsymbol{a}}_{(1_L)}(t) \approx \frac{n^{L-1} \cdot \big|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{L,\boldsymbol{a}}_{\text{ali-sequence}}\big|}{\big|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^L\big|} \cdot \frac{1}{(\log t)^L}, \tag{18}$$

which we will now derive by induction on L.

Base case: $L = 1$. Suppose that p_1 is a positive integer of size about t. One may interpret the prime number theorem as the probabilistic statement that

$$\mathcal{P}_{(1_{L=1})}(t) = \mathrm{Prob}(p_1 \text{ is prime}) \approx \frac{1}{\log t},$$

which is base case $L = 1$ of (18).

Induction step. Assume now that (18) holds for some fixed $L \geq 1$, and fix any vector $\boldsymbol{a} = (a_2, a_3, \dots, a_{L+1}) \in ((\mathbb{Z}/n\mathbb{Z})^\times)^L$. Since the statement

$$(p_1, p_2, \dots, p_{L+1}) \text{ satisfies } (1_{L+1}) \text{ and for all } i \in \{2, 3, \dots, L+1\},\ p_i \equiv a_i \bmod n$$

is equivalent to

$$(p_1, p_2, \ldots, p_L) \text{ satisfies } (1_L) \text{ and for all } i \in \{2, 3, \ldots, L\},\ p_i \equiv a_i \bmod n,$$
$$p_{L+1} := p_L + 1 - a_{p_L}(E) \text{ is prime, and } p_{L+1} \equiv a_{L+1} \bmod n,$$

we see that

$$\mathcal{P}_{(1_{L+1})}^{(a_2, a_3, \ldots, a_L, a_{L+1})}(t) = \mathcal{P}_{(1_L)}^{(a_2, a_3, \ldots, a_L)}(t) \cdot \mathcal{P}(t), \tag{19}$$

where $\mathcal{P}(t)$ is the conditional probability that $p_{L+1} := p_L + 1 - a_{p_L}(E)$ is prime, and that $p_{L+1} \equiv a_{L+1} \bmod n$, given that (1_L) holds. To estimate $\mathcal{P}(t)$, let us assume that (1_L) holds. First note that, by the Hasse bound $|a_p(E)| \le 2\sqrt{p}$, one has

$$p_{L+1} = p_1 + L - \sum_{i=1}^{L} a_{p_i}(E) \in \left[p_1 + L - 2L\sqrt{p_{\max}},\ p_1 + L + 2L\sqrt{p_{\max}}\right],$$

where $p_{\max} := \max\{p_i : i = 1, 2, \ldots, L\}$. By induction we have $p_{\max} = t + O_L(\sqrt{t})$, and so $p_{L+1} \approx t$, with an error of $O_L(\sqrt{t})$. Now, if p_{L+1} were a positive integer of size about t selected independently of $(p_1, p_2, \ldots, p_L)$, then

$$\text{Prob}(p_{L+1} \text{ is prime and } p_{L+1} \equiv a_{L+1} \bmod n) \approx \frac{1}{\varphi(n) \log t}, \tag{20}$$

by the prime number theorem in arithmetic progressions. If the positive integer p_{L+1} were chosen randomly and independently of the previous primes, then the probability that $p_{L+1} \equiv a_{L+1} \bmod n$ would be $1/n$. However, p_{L+1} is not chosen independently of $(p_1, p_2, \ldots, p_L)$; it is related to p_L by the formula $p_{L+1} = p_L + 1 - a_{p_L}(E)$. Thus, the congruence $p_{L+1} \equiv a_{L+1} \bmod n$ is really the demand that

$$\text{Frob}_{\mathbb{Q}(E[n])}(p_L) \in \text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})_{\mathcal{N}=a_{L+1}}.$$

Since we assume that (1_L) holds, we know that $\text{Frob}_{\mathbb{Q}(E[n])}(p_L) \in \text{GL}_2(\mathbb{Z}/n\mathbb{Z})^{\det = a_L}$. It is thus natural to multiply (20) by the correction factor

$$\frac{\left|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})_{\mathcal{N}=a_{L+1}}^{\det=a_L}\right| \Big/ \left|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{\det=a_L}\right|}{1/n},$$

obtaining

$$\begin{aligned} \mathcal{P}(t) &\approx \frac{\left|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})_{\mathcal{N}=a_{L+1}}^{\det=a_L}\right| \Big/ \left|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{\det=a_L}\right|}{1/n} \cdot \frac{1}{\varphi(n) \log t} \\ &= \frac{n\left|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})_{\mathcal{N}=a_{L+1}}^{\det=a_L}\right|}{|\text{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})|} \cdot \frac{1}{\log t}. \end{aligned} \tag{21}$$

By (17), we may rewrite (18) as

$$\mathcal{P}^{\boldsymbol{a}}_{(1_L)}(t) \approx n^{L-1} \cdot \frac{\left|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})_{\mathcal{N}=a_2}\right|}{\left|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})\right|} \cdot \left(\prod_{i=2}^{L-1} \frac{\left|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{\det=a_i}_{\mathcal{N}=a_{i+1}}\right|}{\left|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})\right|} \right) \cdot \frac{\left|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{\det=a_L}\right|}{\left|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})\right|} \cdot \frac{1}{(\log t)^L}.$$

Plugging this expression and (21) into (19), and using the fact that

$$\left|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{\det=a_L}\right| = \left|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{\det=a_{L+1}}\right|,$$

one concludes the induction step, completing the derivation of (18), and thus of (15).

Our analysis has motivated the following conjecture, wherein

$$\pi^{\text{ali-sequence}}_{E,L}(x) := \left|\left\{p_1 \le x : \exists \text{ an aliquot sequence } (p_1, p_2, \dots, p_L) \text{ for } E\right\}\right|,$$

$$C^{\text{ali-sequence}}_{E,L} := \lim_{k\to\infty} \frac{n_k^{L-1} \cdot \left|\mathrm{Gal}(\mathbb{Q}(E[n_k])/\mathbb{Q})^L_{\text{ali-sequence}}\right|}{\left|\mathrm{Gal}(\mathbb{Q}(E[n_k])/\mathbb{Q})^L\right|}.$$

Conjecture 3.1. Let E be an elliptic curve over $\mathbb{Q}$ without complex multiplication and $L \ge 2$ a fixed integer. Then, as $x \longrightarrow \infty$, one has

$$\pi^{\text{ali-sequence}}_{E,L}(x) \sim C^{\text{ali-sequence}}_{E,L} \int_2^x \frac{1}{(\log t)^L}\, dt.$$

Similarly to Proposition 2.6, one has

$$C^{\text{ali-sequence}}_{E,L} > 0 \iff \mathcal{G}_E \text{ has a (directed) walk of length } L.$$

3.2. ***The conditional probability that*** $(p_1, p_2, \dots, p_L)$ ***satisfies*** $(2'_L)$**.** We will now derive (16), completing the heuristic derivation of Conjecture 1.3. Suppose that $(p_1, p_2, \dots, p_L)$ is an aliquot sequence of length L for E, i.e., that it satisfies (1_L). What is the conditional probability that $\sum_{i=1}^L a_{p_i}(E) = r$? In the case $L = 1$, condition (1_L) is empty, and our question becomes identical to the Lang–Trotter conjecture for fixed Frobenius trace. In what follows, we will develop a probabilistic model in the same style as theirs.

Fixing a level n, the number $f_n(r, p) \ge 0$ will estimate the probability of the event that $\sum_{i=1}^L a_{p_i}(E) = r$, given that $(p = p_1, p_2, \dots, p_L)$ is an aliquot sequence of length L for E. We will model the situation by assuming that the vector

$$\left(\mathrm{Frob}_{\mathbb{Q}(E[n])}(p_1), \mathrm{Frob}_{\mathbb{Q}(E[n])}(p_2), \dots \mathrm{Frob}_{\mathbb{Q}(E[n])}(p_L)\right) \in \mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^L_{\text{ali-sequence}} \tag{22}$$

is randomly distributed according to counting measure, and we will assume that the various $a_{p_i}(E)/(2\sqrt{p_i})$ are independent at infinity, i.e., that ϕ_L is the distribution

function for their sum. We will also assume independence of the random variables $\sum_{i=1}^{L} a_{p_i}(E)/(2\sqrt{p_i})$ and (22). Finally, in order to simplify our model, we will also regard all of the various primes p_i as having the same size, namely p. These considerations lead us to the following assumptions about the probabilities $f_n(r, p)$:

$$f_n(r, p) = 0 \quad \text{if } |r| > 2L\sqrt{p},$$

$$f_n(r, p) = \phi_L\left(\frac{r}{2\sqrt{p}}\right) \cdot \frac{n\left|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{L,\sum \mathrm{tr}=r}_{\text{ali-sequence}}\right|}{\left|\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})^{L}_{\text{ali-sequence}}\right|} \cdot c_p \quad \text{if } |r| \le 2L\sqrt{p},$$

where c_p is some constant chosen so that $\sum_{r\in\mathbb{Z}} f_n(r, p) = 1$. Then, similarly to [Lang and Trotter 1976, pp. 31–32], one concludes that $c_p \sim \frac{1}{2\sqrt{p}}$, as $p \to \infty$. This leads to (16), completing the derivation of Conjecture 1.3.

4. Examples

We will now give some numerical evidence for Conjecture 1.3.

4.1. ***Elliptic curves with $C_{E,L} > 0$.*** Table 2 and Table 3 display some data for four elliptic curves. In each table, the column labeled "predicted" lists the approximate values of

$$C_{E,L} \int_2^{10^{13}} \frac{dt}{2\sqrt{t}(\log t)^L},$$

"actual" lists the values of $\pi_{E,L}(10^{13})$, and "% error" lists as a percentage the approximate values of

$$\frac{C_{E,L} \displaystyle\int_2^{10^{13}} \frac{dt}{2\sqrt{t}(\log t)^L} - \pi_{E,L}(10^{13})}{C_{E,L} \displaystyle\int_2^{10^{13}} \frac{dt}{2\sqrt{t}(\log t)^L}}.$$

The first and third curves were already considered in [Silverman and Stange 2011], and are included here largely to show the contrast with the second curve. For each of these curves, a detailed list of the aliquot cycles with $p_1 \le 10^{13}$ may be found in an expanded version of this paper [Jones 2012].

E	predicted	actual	% error
$y^2 + y = x^3 - x$	318.98	332	−4.08%
$y^2 = x^3 + 6x - 2$	546.78	564	−2.97%
$y^2 + y = x^3 + x^2$	318.97	328	−2.83%
$y^2 + xy + y = x^3 - x^2$	318.95	331	−3.78%

Table 2. Data on $\pi_{E,2}(10^{13})$ for various E.

E	predicted	actual	% error
$y^2+y=x^3-x$	3.03	3	1.05%
$y^2=x^3+6x-2$	12.59	12	4.66%
$y^2+y=x^3+x^2$	3.04	2	34.10%
$y^2+xy+y=x^3-x^2$	3.02	4	−32.48%

Table 3. Data on $\pi_{E,3}(10^{13})$ for various E.

The four elliptic curves E under consideration satisfy

$$\left[\mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z}) : \mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})\right] \le 2 \tag{23}$$

for each $n \ge 1$ (see [Serre 1972, pp. 309–311; Lang and Trotter 1976, p. 51]). As shown in [Serre 1972, pp. 310–311], this is the smallest index that one can have for general n when the elliptic curve E is defined over $\mathbb{Q}$. We call any elliptic curve E satisfying (23) a *Serre curve*. Serre curves are thus elliptic curves for which $\mathrm{Gal}(\mathbb{Q}(E[n])/\mathbb{Q})$ is "as large as possible for all n," and it has been shown that, when ordered by height, almost all elliptic curves are Serre curves (see [Jones 2010; Radhakrishnan 2008]). One can show that for any Serre curve E, one has $C_{E,L} > 0$. In fact, if we define the constant C_L by

$$C_L := \frac{\phi_L(0)}{L} \cdot \lim_{k\to\infty} \frac{n_k^L \left|\mathrm{GL}_2(\mathbb{Z}/n_k\mathbb{Z})^L_{\text{ali-cycle}}\right|}{\left|\mathrm{GL}_2(\mathbb{Z}/n_k\mathbb{Z})^L\right|} = \frac{\phi_L(0)}{L} \cdot \prod_{l \text{ prime}} \frac{l^L \left|\mathrm{GL}_2(\mathbb{F}_l)^L_{\text{ali-cycle}}\right|}{\left|\mathrm{GL}_2(\mathbb{F}_l)^L\right|},$$

then for any Serre curve E one has that $C_{E,L} = C_L \cdot f_L(\Delta_{sf}(E))$, where $\Delta_{sf}(E)$ denotes the square-free part of the discriminant of any Weierstrass model of E and f_L is a positive function which approaches 1 as $|\Delta_{sf}(E)|$ approaches infinity. For $L = 2$ one has

$$\begin{aligned} C_2 &= \frac{\phi_2(0)}{2} \cdot \prod_{l \text{ prime}} \frac{l^2 \left|\mathrm{GL}_2(\mathbb{F}_l)^2_{\text{ali-cycle}}\right|}{\left|\mathrm{GL}_2(\mathbb{F}_l)^2\right|} = \frac{8}{3\pi^2} \cdot \prod_{l \text{ prime}} \frac{l^2(l^4-2l^3-2l^2+3l+3)}{[(l^2-1)(l-1)]^2} \\ &\approx 0.077088124, \end{aligned}$$

whereas for $L = 3$ one has

$$\begin{aligned} C_3 &= \frac{\phi_3(0)}{3} \prod_{l \text{ prime}} \frac{l^3 \left|\mathrm{GL}_2(\mathbb{F}_l)^3_{\text{ali-cycle}}\right|}{\left|\mathrm{GL}_2(\mathbb{F}_l)^3\right|} \\ &= \frac{\phi_3(0)}{3} \prod_{l \text{ prime}} \frac{l^3\left[l^6-3l^5-3l^4+14l^3+(3+\chi(l))l^2-(19+3\chi(l))l-10-3\chi(l)\right]}{[(l^2-1)(l-1)]^3} \\ &\approx 0.019759298, \end{aligned}$$

E	$C_{E,2}$	$C_{E,3}$	$\Delta_{sf}(E)$
$y^2+y=x^3-x$	≈ 0.077093	≈ 0.019841	37
$y^2=x^3+6x-2$	≈ 0.132151	≈ 0.082365	-3
$y^2+y=x^3+x^2$	≈ 0.077091	≈ 0.019861	-43
$y^2+xy+y=x^3-x^2$	≈ 0.077088	≈ 0.019759	-53

Table 4. Values of $C_{E,2}$, $C_{E,3}$ and $\Delta_{sf}(E)$.

where $\chi(l)=\left(\frac{-3}{l}\right)$ denotes the character of conductor 3. Table 4 gives the values of $C_{E,2}$, $C_{E,3}$ and $\Delta_{sf}(E)$ for each of the four curves under consideration. The reason the second curve has a larger value of $C_{E,L}$ is that $|\Delta_{sf}(E)|$ is smaller for this curve than for the others.

4.2. ***An elliptic curve with $C_{E,L}=0$.*** We will now discuss briefly the elliptic curve

$$E\ :\ y^2=x^3-3x+4 \tag{24}$$

which was mentioned in the introduction, for which $\pi_{E,L}(x)\equiv 0$ and whose associated graph $\mathcal{G}_E$ contains no closed walks at all. We will presently describe the Galois group $\mathrm{Gal}(\mathbb{Q}(E[4])/\mathbb{Q})$, which is an index 4 subgroup of $\mathrm{GL}_2(\mathbb{Z}/4\mathbb{Z})$. First, define the subgroup $H(4)\subseteq \mathrm{GL}_2(\mathbb{Z}/4\mathbb{Z})$ by

$$H(4):=\left\{\begin{pmatrix}1&0\\0&1\end{pmatrix},\begin{pmatrix}0&1\\-1&-1\end{pmatrix},\begin{pmatrix}-1&-1\\1&0\end{pmatrix},\begin{pmatrix}-1&-1\\0&1\end{pmatrix},\begin{pmatrix}1&0\\-1&-1\end{pmatrix},\begin{pmatrix}0&1\\1&0\end{pmatrix}\right\}.$$

We then have

$$\mathrm{Gal}(\mathbb{Q}(E[4])/\mathbb{Q})=H(4)\cdot\left(I+2\left\{\begin{pmatrix}0&0\\0&0\end{pmatrix},\begin{pmatrix}1&1\\0&1\end{pmatrix},\begin{pmatrix}1&0\\1&1\end{pmatrix},\begin{pmatrix}0&1\\1&0\end{pmatrix}\right\}\right). \tag{25}$$

(To see that the right-hand expression defines a subgroup of $\mathrm{GL}_2(\mathbb{Z}/4\mathbb{Z})$, note that

$$\left\{\begin{pmatrix}0&0\\0&0\end{pmatrix},\begin{pmatrix}1&1\\0&1\end{pmatrix},\begin{pmatrix}1&0\\1&1\end{pmatrix},\begin{pmatrix}0&1\\1&0\end{pmatrix}\right\}\subseteq M_{2\times 2}(\mathbb{Z}/2\mathbb{Z})$$

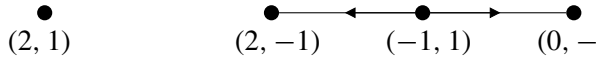
is closed under addition and under $\mathrm{GL}_2(\mathbb{Z}/2\mathbb{Z})$-conjugation.)

Even though $\mathrm{Gal}(\mathbb{Q}(E[2])/\mathbb{Q})=\mathrm{GL}_2(\mathbb{Z}/2\mathbb{Z})$, $\mathrm{Gal}(\mathbb{Q}(E[4])/\mathbb{Q})$ is a proper subgroup of $\mathrm{GL}_2(\mathbb{Z}/4\mathbb{Z})$, and so one has $4\mid m_E$. Furthermore, in this case the restriction map $\mathrm{Gal}(\mathbb{Q}(E[m_E])/\mathbb{Q})\twoheadrightarrow \mathrm{Gal}(\mathbb{Q}(E[4])/\mathbb{Q})$ induces a graph morphism

$$\mathcal{G}_E=\mathcal{G}_E(m_E)\twoheadrightarrow \mathcal{G}_E(4), \tag{26}$$

which is surjective in the sense that it carries the vertex set $\mathcal{V}(m_E)$ onto $\mathcal{V}(4)$ and likewise carries $\mathcal{E}(m_E)$ onto $\mathcal{E}(4)$.

On the other hand, using (25), one finds that the directed graph $\mathcal{G}_E(4)$ is:

(27) 

Infinitely many primes p for which $|E(\mathbb{F}_p)|$ is prime. The non-CM case of a conjecture of Koblitz (see [Koblitz 1988] and also [Zywina 2011]) expresses (in our terminology) that for any non-CM elliptic curve E, the existence of a single directed edge in $\mathcal{G}_E$ implies the existence of infinitely many primes p for which $|E(\mathbb{F}_p)|$ is prime. Taking E to be the elliptic curve given by (24), we see by the surjectivity of (26) together with (27) that $\mathcal{G}_E$ contains at least one directed edge. Thus, assuming Koblitz's conjecture, there are infinitely many primes p for which $|E(\mathbb{F}_p)|$ is prime.

Finitely many aliquot cycles for E. Continuing with the example (24), by the surjectivity of (26) together with (27), we see that $\mathcal{G}_E$ contains no closed walks at all. By Proposition 2.6, there are only finitely many aliquot cycles $(p_1, p_2, \ldots, p_L)$ for E. This particular example may be explained as follows. Whenever $p_2 = |E(\mathbb{F}_{p_1})|$ for some prime p_1, we see from (27) that $(\mathrm{tr}(\mathrm{Frob}_{\mathbb{Q}(E[4])}(p_1)), \det(\mathrm{Frob}_{\mathbb{Q}(E[4])}(p_1))) = (-1, 1)$ (otherwise, $|E(\mathbb{F}_{p_1})|$ would be even). But then

$$\big(\mathrm{tr}(\mathrm{Frob}_{\mathbb{Q}(E[4])}(p_2)), \det(\mathrm{Frob}_{\mathbb{Q}(E[4])}(p_2))\big) \in \{(0, -1), (2, -1)\},$$

in which case $|E(\mathbb{F}_{p_2})|$ must be even. One deduces that E has no aliquot cycles of length $L \geq 2$, and indeed no aliquot *sequences* of length $L \geq 3$.

Remark 4.1. There is a modular curve X of level 4 and genus 0 with $|X(\mathbb{Q})| = \infty$, whose noncuspidal $\mathbb{Q}$-rational points correspond to elliptic curves E' for which $-\Delta_{E'}$ is a perfect square. For almost all such elliptic curves E', one may find an appropriate twist E of E' for which (25) holds, and thus for which $\lim_{x\to\infty} \pi_{E,L}(x) < \infty$ for $L \geq 2$. The elliptic curve (24) is one such example.

Acknowledgments

The author gratefully acknowledges A. C. Cojocaru, who first brought this question to my attention, and also J. Silverman for a stimulating discussion. Also many thanks to A. Sutherland, who provided help with the computations (a description of the software used therein may be found in [Kedlaya and Sutherland 2008]). Finally, thanks to the anonymous referee for a careful reading of the manuscript and several helpful comments.

References

[Balog et al. 2011] A. Balog, A.-C. Cojocaru, and C. David, "Average twin prime conjecture for elliptic curves", *Amer. J. Math.* **133**:5 (2011), 1179–1229. MR 2012j:11118 Zbl 05969056

[David and Pappalardi 1999] C. David and F. Pappalardi, "Average Frobenius distributions of elliptic curves", *Internat. Math. Res. Notices* 4 (1999), 165–183. MR 2000g:11045 Zbl 0934.11033

[Gottschlich 2012] A. Gottschlich, "On positive integers n dividing the nth term of an elliptic divisibility sequence", *New York J. Math.* **18** (2012), 409–420. MR 2928585 Zbl 06098855

[Jones 2010] N. Jones, "Almost all elliptic curves are Serre curves", *Trans. Amer. Math. Soc.* **362**:3 (2010), 1547–1570. MR 2011d:11130 Zbl 1204.11088

[Jones 2012] N. Jones, "Elliptic aliquot cycles of fixed length", preprint, 2012. arXiv 1212.1010

[Kedlaya and Sutherland 2008] K. S. Kedlaya and A. V. Sutherland, "Computing L-series of hyperelliptic curves", pp. 312–326 in *Algorithmic number theory*, edited by A. J. van der Poorten and A. Stein, Lecture Notes in Comput. Sci. **5011**, Springer, Berlin, 2008. MR 2010d:11070 Zbl 1232.11078

[Koblitz 1988] N. Koblitz, "Primality of the number of points on an elliptic curve over a finite field", *Pacific J. Math.* **131**:1 (1988), 157–165. MR 89h:11023 Zbl 0608.10010

[Lang and Trotter 1976] S. Lang and H. Trotter, *Frobenius distributions in* GL_2*-extensions: Distribution of Frobenius automorphisms in* $\mathrm{G}L_2$*-extensions of the rational numbers*, Lecture Notes in Mathematics **504**, Springer, Berlin, 1976. MR 58 #27900 Zbl 0329.12015

[Mazur 1972] B. Mazur, "Rational points of abelian varieties with values in towers of number fields", *Invent. Math.* **18** (1972), 183–266. MR 56 #3020 Zbl 0245.14015

[Radhakrishnan 2008] V. Radhakrishnan, *An asymptotic formula for the number of non–Serre curves in a two-parameter family of elliptic curves*, Ph.D. thesis, University of Colorado at Boulder, 2008, Available at http://search.proquest.com/docview/304629348. MR 2711545

[Serre 1968] J.-P. Serre, *Abelian l-adic representations and elliptic curves*, W. A. Benjamin, New York, 1968. MR 41 #8422 Zbl 0186.25701

[Serre 1972] J.-P. Serre, "Propriétés galoisiennes des points d'ordre fini des courbes elliptiques", *Invent. Math.* **15**:4 (1972), 259–331. In French. MR 52 #8126 Zbl 0235.14012

[Silverman and Stange 2011] J. H. Silverman and K. E. Stange, "Amicable pairs and aliquot cycles for elliptic curves", *Exp. Math.* **20**:3 (2011), 329–357. MR 2012g:11109

[Taylor 2008] R. Taylor, "Automorphy for some l-adic lifts of automorphic mod l Galois representations, II", *Publ. Math. Inst. Hautes Études Sci.* 108 (2008), 183–239. MR 2010j:11085 Zbl 1169.11021

[Zywina 2011] D. Zywina, "A refinement of Koblitz's conjecture", *Int. J. Number Theory* **7**:3 (2011), 739–769. MR 2012e:11107 Zbl 05913798

Received July 16, 2012. Revised December 10, 2012.

NATHAN JONES
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MISSISSIPPI
HUME HALL 305
P.O. BOX 1848
UNIVERSITY, MS 38677
UNITED STATES
ncjones@olemiss.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 263, No. 2, 2013
dx.doi.org/10.2140/pjm.2013.263.373

ASYMPTOTIC L^4 NORM OF POLYNOMIALS DERIVED FROM CHARACTERS

DANIEL J. KATZ

Littlewood investigated polynomials with coefficients in $\{-1, 1\}$ (Littlewood polynomials), to see how small their ratio of norms $\|f\|_4/\|f\|_2$ on the unit circle can become as $\deg f \to \infty$. A small limit is equivalent to slow growth in the mean square autocorrelation of the associated binary sequences of coefficients of the polynomials. The autocorrelation problem for arrays and higher dimensional objects has also been studied; it is the natural generalization to multivariable polynomials. Here we find, for each $n > 1$, a family of n-variable Littlewood polynomials with lower asymptotic $\|f\|_4/\|f\|_2$ than any known hitherto. We discover these through a wide survey, infeasible with previous methods, of polynomials whose coefficients come from finite field characters. This is the first time that the lowest known asymptotic ratio of norms $\|f\|_4/\|f\|_2$ for multivariable polynomials $f(z_1, \ldots, z_n)$ is strictly less than what could be obtained by using products $f_1(z_1) \cdots f_n(z_n)$ of the best known univariate polynomials.

1. Introduction

1A. ***History and main result.*** Littlewood [1966; 1968] pioneered the study of the L^4 norm on the complex unit circle of polynomials whose coefficients lie in $\{-1, 1\}$, and in particular wanted to know how small their ratio of norms $\|f\|_4/\|f\|_2$ can become as $\deg f \to \infty$. He suspected, based on calculations of Swinnerton-Dyer, that this ratio could be made to approach 1 asymptotically, but the smallest limiting ratio he could find was $\sqrt[4]{4/3}$ for the Rudin–Shapiro polynomials [Littlewood 1968]. The L^4 norm is of particular interest since it serves as a lower bound for the L^∞ norm and is easier to calculate than most other L^r norms. Erdős [1957, Problem 22; 1962] had conjectured that $\|f\|_\infty/\|f\|_2$ is bounded away from 1 for nonconstant polynomials with complex coefficients of unit magnitude. This was disproved in [Kahane 1980], but the modified problem where we restrict to coefficients in $\{-1, 1\}$ remains open [Newman and Byrnes 1990], and would be solved if one could prove

The author was supported by funding from an NSERC grant awarded to Jonathan Jedwab.
MSC2010: primary 11C08; secondary 11T24, 42A05, 11B83.
Keywords: L^4 norm, Littlewood polynomial, character polynomial, Fekete polynomial, character sum.

that $\|f\|_4/\|f\|_2$ is bounded away from 1 as $\deg f \to \infty$. Polynomials in one or more variables with coefficients in $\{-1, 1\}$ and small $\|f\|_4/\|f\|_2$ are equivalent to binary sequences and arrays (that is, those that simply list the coefficients of the polynomials) with low mean square aperiodic autocorrelation. Such sequences and arrays are important in the theory of communications[1] [Golay 1977] and statistical physics [Bernasconi 1987]. Accordingly, we define a *Littlewood polynomial* in n variables to have the form

$$f(z_1, \ldots, z_n) = \sum_{j_1=0}^{s_1-1} \cdots \sum_{j_n=0}^{s_n-1} f_{j_1,\ldots,j_n} z_1^{j_1} \cdots z_n^{j_n},$$

with coefficient $f_{j_1,\ldots,j_n}$ in $\{-1, 1\}$, and our L^r norm for $f(z_1, \ldots, z_n)$ is

$$\|f\|_r = \left(\frac{1}{(2\pi)^n} \int_0^{2\pi} \cdots \int_0^{2\pi} \left| f\big(\exp(i\theta_1), \ldots, \exp(i\theta_n)\big) \right|^r d\theta_1 \cdots d\theta_n \right)^{1/r}.$$

Note that $\|f\|_2^2 = s_1 \cdots s_n$ for our Littlewood polynomial.

For univariate Littlewood polynomials, the lowest asymptotic ratio of norms $\|f\|_4/\|f\|_2$ found by Littlewood himself [1968] was $\sqrt[4]{4/3}$ for the Rudin–Shapiro polynomials. Two decades later, this was improved to $\sqrt[4]{7/6}$ by Høholdt and Jensen [1988], using modifications of Fekete polynomials. Over two decades later still, in [Jedwab et al. 2013b], another modification was shown to yield further improvement:

Theorem 1.1 (Jedwab, Katz, Schmidt). *There is a family of univariate Littlewood polynomials that, as* $\deg f \to \infty$, *has* $\|f\|_4/\|f\|_2 \to B_1$, *the largest real root of* $27x^{12} - 498x^8 + 1164x^4 - 722$, *which is less than* $\sqrt[4]{22/19}$.

Prior to this paper, for each n, the lowest known asymptotic $\|f\|_4/\|f\|_2$ for n-variable Littlewood polynomials

$$f(z_1, \ldots, z_n) \quad \text{(in the limit as } \deg_{z_1} f, \ldots, \deg_{z_n} f \to \infty)$$

was simply the n-th power of the lowest known ratio for univariate polynomials, based on the fact that if $f(z_1, \ldots, z_n) = f_1(z_1) \cdots f_n(z_n)$, then $\|f\|_r = \|f_1\|_r \cdots \|f_n\|_r$. For bivariate Littlewood polynomials, Schmidt [2011] obtained an asymptotic $\|f\|_4/\|f\|_2$ of $\sqrt{7/6}$ in this way (via Høholdt and Jensen's univariate polynomials mentioned above), and foresaw the possibility that the asymptotic $\|f\|_4/\|f\|_2$ could be lowered to B_1^2, contingent upon the conjecture that was later established as Theorem 1.1. In this paper, we show that one can do better than this product construction, even when based on the best univariate polynomials now known (those of Theorem 1.1).

[1] In this milieu, results are expressed in terms of the *merit factor*, defined as $\|f\|_2^4/(\|f\|_4^4 - \|f\|_2^4)$.

Theorem 1.2. *For each $n > 1$, there is a family of n-variable Littlewood polynomials $f(z_1, \dots, z_n)$ which, as $\deg_{z_1} f, \dots, \deg_{z_n} f \to \infty$, has $\|f\|_4/\|f\|_2$ tending to a value strictly less than B_1^n.*

The lowest known asymptotic $\|f\|_4/\|f\|_2$ for n-variable Littlewood polynomials is an algebraic number depending on n, and is specified precisely in Section 1C after we define in Section 1B the polynomials that are involved.[2]

1B. ***Character polynomials.*** The polynomials used in [Høholdt and Jensen 1988], [Jedwab et al. 2013b], and the current paper to break previous records for the lowest known asymptotic $\|f\|_4/\|f\|_2$ are all character polynomials, that is, polynomials whose coefficients are given by characters of finite fields. The L^4 norms of character polynomials have already been studied extensively [Høholdt and Jensen 1988; Jensen and Høholdt 1989; Jensen et al. 1991; Bömer and Antweiler 1993; Borwein 2002; Borwein and Choi 2000; 2002; Jedwab 2005; Høholdt 2006; Jedwab and Schmidt 2010; Schmidt 2011], but it took the new methods of this paper to discover and verify the properties of the polynomials of our Theorem 1.2.

The interrelation between the additive and multiplicative structures of finite fields endow character polynomials with their remarkable qualities: the coefficients of an *additive character polynomial* are obtained by applying an additive character of a finite field to its nonzero elements arranged multiplicatively (listed as successive powers of a primitive element), while the coefficients of a *multiplicative character polynomial* are obtained by applying a multiplicative character of a finite field to its elements arranged additively (as $\mathbb{Z}$-linear combinations of the generators, arrayed in a box whose dimensionality equals the number of generators). Thus an additive character polynomial has the form

$$f(z) = \sum_{j \in S} \psi(\alpha(j+t)) z^j, \tag{1}$$

where $\psi : \mathbb{F}_q \to \mathbb{C}$ is a nontrivial additive character, the *support* S is a set of the form $\{0, 1, \dots, s-1\}$, the *translation* t is an element of $\mathbb{Z}$, and the *arrangement* α is a group epimorphism from $\mathbb{Z}$ to $\mathbb{F}_q^*$. A multiplicative character polynomial has the form

$$f(z_1, \dots, z_e) = \sum_{j=(j_1,\dots,j_e) \in S} \chi(\alpha(j+t))\, z_1^{j_1} \cdots z_e^{j_e}, \tag{2}$$

where e is a positive integer, χ is a nontrivial complex-valued multiplicative character of $\mathbb{F}_q = \mathbb{F}_{p^e}$ with p prime, the *support* S is $S_1 \times \cdots \times S_e$ with each S_k a set of

[2] Gulliver and Parker [2005] have also studied $\|f\|_4/\|f\|_2$ for multivariable Littlewood polynomials, but in a very different limit: they let the number of variables tend to infinity while keeping the degree in each variable less than or equal to one.

the form $\{0, 1, \ldots, s_k - 1\}$, while the *translation* $t = (t_1, \ldots, t_e)$ is in $\mathbb{Z}^e$, and the *arrangement* α is a group epimorphism from $\mathbb{Z}^e$ to $\mathbb{F}_q$. We always extend nontrivial multiplicative characters to take 0 to 0.

We now define the Fekete polynomials and their modifications used in [Høholdt and Jensen 1988] and [Jedwab et al. 2013b]. For an odd prime p, the p-th Fekete polynomial is a multiplicative character polynomial using the quadratic character (Legendre symbol) over the prime field $\mathbb{F}_p$, support $S = \{0, 1, \ldots, p-1\}$, translation $t = 0$, and arrangement $\alpha : \mathbb{Z} \to \mathbb{F}_p$ given by reduction modulo p. Fekete polynomials are themselves the subject of many fascinating studies linking number theory and analysis [Fekete and Pólya 1912; Pólya 1919; Montgomery 1980; Baker and Montgomery 1990; Conrey et al. 2000; Borwein et al. 2001; Borwein and Choi 2002].

The polynomials used in [Høholdt and Jensen 1988] to obtain asymptotic $\|f\|_4/\|f\|_2$ of $\sqrt[4]{7/6}$ have the same character, support, and arrangement, but the translations t are chosen such that $t/p \to 1/4$ as $p \to \infty$, and any coefficient of 0 (arising from the extended multiplicative character) is replaced with 1 to obtain Littlewood polynomials. To obtain asymptotic $\|f\|_4/\|f\|_2$ less than $\sqrt[4]{22/19}$, we used in [Jedwab et al. 2013b] a different limit for t/p, and allowed the support $S = \{0, 1, \ldots, s-1\}$ to be of size other than p, and in fact let s/p tend to a number slightly larger than 1 as $p \to \infty$.

The families of character polynomials used here are based on similar asymptotics: we say that a family $\{f_\iota\}_{\iota \in I}$ of additive character polynomials is *size-stable* to mean that if we write $\mathbb{F}_{q_\iota}$ and S_ι for the field and support of f_ι, then $\{q_\iota : \iota \in I\}$ is infinite and $|S_\iota|/(q_\iota - 1)$ tends to a positive real number σ (called the *limiting size*) as $q_\iota \to \infty$. Likewise, we say that a family of e-variable multiplicative character polynomials $\{f_\iota\}_{\iota \in I}$ is *size-stable* to mean that if we write $\mathbb{F}_{q_\iota} = \mathbb{F}_{p_\iota^e}$ and $S_\iota = S_{\iota,1} \times \cdots \times S_{\iota,e}$ for the field and support of f_ι, then the set of primes $\{p_\iota : \iota \in I\}$ is infinite and for each $k \in \{1, \ldots, e\}$, the ratio $|S_{\iota,k}|/p_\iota$ tends to a positive real number σ_k as $q_\iota \to \infty$. We call $\sigma_1, \ldots, \sigma_e$ the *limiting sizes*. And we say that a family of e-variable multiplicative character polynomials $\{f_\iota\}_{\iota \in I}$ is *translation-stable* to mean that if we write $\mathbb{F}_{q_\iota} = \mathbb{F}_{p_\iota^e}$ and $t_\iota = (t_{\iota,1}, \ldots, t_{\iota,e})$ for the field and translation of f_ι, then the set of primes $\{p_\iota : \iota \in I\}$ is infinite and for each $k \in \{1, \ldots, e\}$, the ratio $t_{\iota,k}/p_\iota$ tends to a real number τ_k as $q_\iota \to \infty$. We call $\tau_1, \ldots, \tau_e$ the *limiting translations*.

1C. ***Subsidiary results.*** We discovered the polynomials of Theorem 1.2 via a survey, enabled by the methods presented in this paper, of the asymptotic L^4 norms of both additive and multiplicative character polynomials. Quadratic multiplicative characters behave differently than nonquadratic ones, so we treat them separately: we have *quadratic families* in which every character is quadratic, and *nonquadratic* families in which none is. We then have three theorems: one for additive characters

and two for the different types of multiplicative characters, and we express our limiting norms in terms of the function

$$\Omega(x, y) = \sum_{n\in\mathbb{Z}} \max(0, 1 - |xn - y|)^2, \tag{3}$$

which is defined and continuous on $\{(x, y) \in \mathbb{R}^2 : x \neq 0\}$.

Theorem 1.3. *Let $\{f_\iota\}_{\iota\in I}$ be a size-stable family, with limiting size σ, of additive character polynomials over fields $\{\mathbb{F}_{q_\iota}\}_{\iota\in I}$.*

(i) *As $q_\iota \to \infty$, we have*

$$\frac{\|f_\iota\|_4^4}{\|f_\iota\|_2^4} \to -\tfrac{2}{3}\sigma + 2\,\Omega\Big(\frac{1}{\sigma}, 0\Big).$$

(ii) *This limit is globally minimized if and only if σ is the unique root in $\big(1, 1+\frac{9}{64}\big)$ of $x^3 - 12x + 12$.*

Theorem 1.4. *Let $\{f_\iota\}_{\iota\in I}$ be a size-stable family, with limiting sizes $\sigma_1, \dots, \sigma_e$, of e-variable nonquadratic multiplicative character polynomials over fields $\{\mathbb{F}_{q_\iota}\}_{\iota\in I}$.*

(i) *As $q_\iota \to \infty$, we have*

$$\frac{\|f_\iota\|_4^4}{\|f_\iota\|_2^4} \to -\frac{2^e}{3^e}\prod_{j=1}^{e}\sigma_i + 2\prod_{j=1}^{e}\Omega\Big(\frac{1}{\sigma_j}, 0\Big).$$

(ii) *This limit is globally minimized if and only if $\sigma_1, \dots, \sigma_e$ all equal the unique root in $(1, 1 + 3^{e+1}/2^{2e+4})$ of*

$$x^{3e} - \frac{3^e}{2^{e-3}}(x-1)(3x^2 - 4x + 2)^{e-1}.$$

Theorem 1.5. *Let $\{f_\iota\}_{\iota\in I}$ be a size- and translation-stable family, with limiting sizes $\sigma_1, \dots, \sigma_e$ and limiting translations $\tau_1, \dots, \tau_e$, of e-variable quadratic multiplicative character polynomials over fields $\{\mathbb{F}_{q_\iota}\}_{\iota\in I}$.*

(i) *As $q_\iota \to \infty$, we have*

$$\frac{\|f_\iota\|_4^4}{\|f_\iota\|_2^4} \to -\frac{2^{e+1}}{3^e}\prod_{j=1}^{e}\sigma_i + 2\prod_{j=1}^{e}\Omega\Big(\frac{1}{\sigma_j}, 0\Big) + \prod_{j=1}^{e}\Omega\Big(\frac{1}{\sigma_j}, 1 + \frac{2\tau_j}{\sigma_j}\Big).$$

(ii) *This limit is globally minimized if and only if $\sigma_1, \dots, \sigma_e$ all equal the unique root in $(1, 1 + 3^{e+1}/2^{2e+3})$ of*

$$x^{3e} - \frac{3^e}{2^{e-2}}(x-1)(3x^2 - 4x + 2)^{e-1} - \frac{3^e}{2^{2e}}(2x-1)^{2e-1},$$

and $\tau_j \in \{\frac{1}{4}(1 - 2\sigma_j) + n/2 : n \in \mathbb{Z}\}$ for each $j \in \{1, \dots, e\}$.

These new theorems are much more general than the compositum of all previous results on the limiting ratio of L^4 to L^2 norm for character polynomials [Høholdt and Jensen 1988; Jensen and Høholdt 1989; Jensen et al. 1991; Borwein and Choi 2000; Schmidt 2011; Jedwab et al. 2013a; 2013b], and reveal for the first time the full functional form of the asymptotic ratio of norms as it depends on choice of character, limiting size, and limiting translation, thus enabling us to find multivariable Littlewood polynomials with lower asymptotic $\|f\|_4/\|f\|_2$ than any known hitherto.

For each $e \geq 1$, let A_e (resp., B_e) be the minimum asymptotic ratio of norms for a family of e-variable nonquadratic (resp., quadratic) character polynomials, as described in Theorem 1.4(ii) (for A_e) and Theorem 1.5(ii) (for B_e). Note that A_1 is also the minimum asymptotic ratio of norms achievable by a family of additive character polynomials as described in Theorem 1.3(ii). Rational approximations of $B_1^4, B_2^4, \ldots, B_5^4$ are obtained later in Lemma 7.1, and if desired, a computer may be used to obtain more accurate approximations of values of various A_e and B_e. For each $e \geq 1$, B_e is to date the lowest known asymptotic $\|f\|_4/\|f\|_2$ for a family of e-variable Littlewood[3] polynomials $f(z_1, \ldots, z_e)$ in the limit as $\deg_{z_1} f, \ldots, \deg_{z_e} f \to \infty$. For $e = 1$, this recapitulates Corollary 3.2 of [Jedwab et al. 2013b], while for $e > 1$, the ratio obtained here is strictly lower than any found to date. Until now, the smallest known asymptotic ratio has been whatever can be obtained from the best univariate polynomials and the product construction $\|f(z_1) \cdots f(z_e)\|_r = \|f(z)\|_r^e$, and so we are claiming that $B_e < B_1^e$ for every $e > 1$. This will give our main result, Theorem 1.2, but in fact we prove something more general: one always obtains a lower asymptotic ratio of norms with a single optimal family of quadratic character polynomials than one does using the product construction with two or more families of character polynomials (which could draw coefficients from $\{-1, 1\}$ or a larger set, depending on the characters involved).

Theorem 1.6. *For each $e \geq 1$, let A_e (resp., B_e) be the minimum asymptotic ratio of L^4 to L^2 norm achievable by families of e-variable nonquadratic* (*resp.*, *quadratic*) *multiplicative character polynomials, as described in Theorem 1.4*(*ii*) (*resp.*, *Theorem 1.5*(*ii*)).

Then $B_e < A_e$ for every $e \geq 1$ and $B_{e_1+e_2} < B_{e_1} B_{e_2}$ for every $e_1, e_2 \geq 1$.

1D. ***Organization of this paper.*** To prove Theorems 1.3–1.5, we first establish a general theorem for obtaining the L^4 norm of a polynomial from its Fourier interpolation in Section 3, after setting down notational conventions in Section 2. Our general theorem reduces the problem of computing L^4 norms of character

[3]Quadratic character polynomials are not always Littlewood because the extended quadratic character χ has $\chi(0) = 0$, so we replace each coefficient of 0 thus produced with a 1, and Corollary A.3 shows that this has no effect on the asymptotic ratio of norms.

polynomials to a pair of calculations (one for additive and one for multiplicative characters) involving Gauss sums, which are presented in Section 4. We use these in Section 5 to prove Theorems 1.3(i), 1.4(i), and 1.5(i). These respectively imply Theorems 1.3(ii), 1.4(ii), and 1.5(ii), but showing this demands delicate arguments which are sketched in Section 6. We prove Theorem 1.6 in Section 7. Some technical results used in Sections 5 and 6 are collected and proved in the Appendix.

2. Notation and conventions

For the rest of this paper p is a prime, and $q = p^e$ with e a positive integer. For any group Γ, we use $\widehat{\Gamma}$ to denote the group of characters from Γ into $\mathbb{C}$: thus $\widehat{\mathbb{F}_q}$ is the group of additive characters from $\mathbb{F}_q$ to $\mathbb{C}$ and $\widehat{\mathbb{F}_q^*}$ the group of multiplicative characters from $\mathbb{F}_q^*$ to $\mathbb{C}$. We extend any nontrivial $\chi \in \widehat{\mathbb{F}_q^*}$ so that $\chi(0) = 0$.

To write the multiplicative character polynomial (2) compactly, we use the convention that if $j = (j_1, \ldots, j_e) \in \mathbb{Z}^e$, the notation z^j is a shorthand for $z_1^{j_1} \cdots z_e^{j_e}$. To make it easier to speak about supports of character polynomials (1) and (2), we call a finite set of consecutive integers a *segment*, and a finite Cartesian product of segments a *box*. If S is a subset of $\mathbb{Z}^n$ and $t \in \mathbb{Z}^n$, then $S + t$ is the translated subset $\{s + t : s \in S\}$.

3. L^4 norms via the Fourier transform

If Γ is a finite abelian group and $\{F_g\}_{g\in\Gamma}$ is a family of complex numbers, then for any $\eta \in \widehat{\Gamma}$, we have the Fourier transform

$$\hat{F}_\eta = \sum_{g\in\Gamma} F_g \eta(g),$$

with inverse

$$F_g = \frac{1}{|\Gamma|} \sum_{\eta\in\widehat{\Gamma}} \hat{F}_\eta \overline{\eta(g)}.$$

We express the L^4 norm in terms of the Fourier interpolation.

Theorem 3.1. *Let Γ be a finite abelian group, $\{F_g\}_{g\in\Gamma}$ a family of complex numbers, n a positive integer, and $\pi \in \mathrm{Hom}(\mathbb{Z}^n, \Gamma)$. For any $\eta \in \widehat{\Gamma}$, let $\eta' \in \widehat{\mathbb{Z}^n}$ be $\eta \circ \pi$. If U is a finite subset of $\mathbb{Z}^n$ and $F(z) = \sum_{u\in U} F_{\pi(u)} z^u \in \mathbb{C}[z_1, \ldots, z_n]$, then*

$$\|F\|_4^4 = \frac{1}{|\Gamma|^5} \sum_{\substack{a,b,c,d\in U \\ a+b=c+d}} \sum_{\kappa,\lambda,\mu,\nu\in\widehat{\Gamma}} \overline{\kappa'(a)\lambda'(b)}\mu'(c)\nu'(d) H(\kappa, \lambda, \mu, \nu),$$

where

$$H(\kappa, \lambda, \mu, \nu) = \sum_{\xi\in\widehat{\Gamma}} \hat{F}_{\xi\kappa} \hat{F}_{\xi\lambda} \overline{\hat{F}_{\xi\mu} \hat{F}_{\xi\nu}}.$$

Proof. By the definition of the L^4 norm, we have

$$\|F\|_4^4 = \sum_{\substack{a,b,c,d\in U\\ a+b=c+d}} F_{\pi(a)}F_{\pi(b)}\overline{F_{\pi(c)}F_{\pi(d)}},$$

and thus, using the inverse Fourier transform,

$$\|F\|_4^4 = \frac{1}{|\Gamma|^4}\sum_{\substack{a,b,c,d\in U\\ a+b=c+d}}\sum_{\kappa,\lambda,\mu,\nu\in\widehat{\Gamma}} \hat{F}_\kappa\hat{F}_\lambda\overline{\hat{F}_\mu\hat{F}_\nu\kappa'(a)\lambda'(b)}\mu'(c)\nu'(d).$$

Since we are summing κ over all $\widehat{\Gamma}$, we can replace κ by $\xi\kappa$ for any given $\xi\in\widehat{\Gamma}$, and also do likewise with λ, μ, ν to obtain

$$\|F\|_4^4 = \frac{1}{|\Gamma|^4}\sum_{\substack{a,b,c,d\in U\\ a+b=c+d}}\sum_{\kappa,\lambda,\mu,\nu\in\widehat{\Gamma}} \hat{F}_{\xi\kappa}\hat{F}_{\xi\lambda}\overline{\hat{F}_{\xi\mu}\hat{F}_{\xi\nu}\kappa'(a)\lambda'(b)}\mu'(c)\nu'(d)$$

where we have omitted mention of the resulting factor of $\overline{\xi'(a)\xi'(b)}\xi'(c)\xi'(d)$, which equals 1 in view of the constraint in the first summation. Now sum ξ over $\widehat{\Gamma}$ and divide by $|\Gamma| = |\widehat{\Gamma}|$ to finish. □

We apply this general theorem to additive and multiplicative character polynomials in two corollaries below. Such polynomials have Gauss sums as their Fourier coefficients, so for any $\psi\in\widehat{\mathbb{F}}_q$ and $\chi\in\widehat{\mathbb{F}_q^*}$, we define the *Gauss sum* associated with ψ and χ to be

$$G(\psi,\chi) = \sum_{a\in\mathbb{F}_q^*}\psi(a)\chi(a).$$

Corollary 3.2. *If $f(z)$ is an additive character polynomial with character $\psi\in\widehat{\mathbb{F}}_q$, support S, translation t, and arrangement α, then*

$$\|f\|_4^4 = \frac{1}{(q-1)^5}\sum_{\substack{a,b,c,d\in S+t\\ a+b=c+d}}\sum_{\kappa,\lambda,\mu,\nu\in\widehat{\mathbb{F}_q^*}} \overline{\kappa'(a)\lambda'(b)}\mu'(c)\nu'(d)H(\kappa,\lambda,\mu,\nu)$$

where for any $\eta\in\widehat{\mathbb{F}_q^*}$, we let $\eta'\in\widehat{\mathbb{Z}}$ be $\eta\circ\alpha$, and

$$H(\kappa,\lambda,\mu,\nu) = \sum_{\xi\in\widehat{\mathbb{F}_q^*}} G(\psi,\xi\kappa)G(\psi,\xi\lambda)\overline{G(\psi,\xi\mu)G(\psi,\xi\nu)}.$$

Proof. Our additive character polynomial $f(z) = \sum_{s\in S}\psi(\alpha(s+t))z^s$ has the same L^r norms as $F(z) = z^t f(z) = \sum_{u\in S+t}\psi(\alpha(u))z^u$, so take $\Gamma = \mathbb{F}_q^*$, $F_g = \psi(g)$, $n=1$, $\pi=\alpha$, and $U = S+t$ in Theorem 3.1, and note that for $\eta\in\widehat{\mathbb{F}_q^*}$ we have $\hat{F}_\eta = G(\psi,\eta)$. □

Corollary 3.3. *If $f(z)$ is a multiplicative character polynomial with character $\chi \in \widehat{\mathbb{F}_q^*}$, support S, translation t, and arrangement α, then*

$$\|f\|_4^4 = \frac{1}{q^5} \sum_{\substack{a,b,c,d \in S+t \\ a+b=c+d}} \sum_{\kappa,\lambda,\mu,\nu \in \widehat{\mathbb{F}_q}} \overline{\kappa'(a)\lambda'(b)}\mu'(c)\nu'(d)H(\kappa, \lambda, \mu, \nu)$$

where for any $\eta \in \widehat{\mathbb{F}_q}$, we let $\eta' \in \widehat{\mathbb{Z}^e}$ be $\eta \circ \alpha$, and

$$H(\kappa, \lambda, \mu, \nu) = \sum_{\xi \in \widehat{\mathbb{F}_q}} G(\xi\kappa, \chi)G(\xi\lambda, \chi)\overline{G(\xi\mu, \chi)G(\xi\nu, \chi)}.$$

Proof. Our multiplicative character polynomial $f(z) = \sum_{s \in S} \chi(\alpha(s+t))z^s$ has the same L^r norms as $F(z) = z^t f(z) = \sum_{u \in S+t} \chi(\alpha(u))z^u$, so take $\Gamma = \mathbb{F}_q$, $F_g = \chi(g)$, $n = e$, $\pi = \alpha$, and $U = S + t$ in Theorem 3.1, and note that for $\eta \in \widehat{\mathbb{F}_q}$ we have $\hat{F}_\eta = G(\eta, \chi)$. □

The key to L^4 norms is then the evaluation of the sums $H(\kappa, \lambda, \mu, \nu)$ in the above two corollaries, which we take up in the next section.

4. Two propositions on summations of Gauss sums

Here we estimate the values of the summations H that appear in Corollaries 3.2 and 3.3. We begin with some basic facts about Gauss sums, which are proved in Theorems 5.11 and 5.12 of [Lidl and Niederreiter 1997].

Lemma 4.1. *If $\psi \in \widehat{\mathbb{F}_q}$ and $\chi \in \widehat{\mathbb{F}_q^*}$, then*

(i) $G(\psi, \chi) = q - 1$ *if both characters are trivial,*

(ii) $G(\psi, \chi) = 0$ *if ψ is trivial and χ is not,*

(iii) $G(\psi, \chi) = -1$ *if χ is trivial and ψ is not,*

(iv) $|G(\psi, \chi)| = \sqrt{q}$ *if both characters are nontrivial, and*

(v) $\sum_{a \in \mathbb{F}_q^*} \psi(ba)\chi(a) = \overline{\chi}(b)G(\psi, \chi)$ *for any $b \in \mathbb{F}_q^*$.*

We first estimate the summation H appearing in Corollary 3.2.

Proposition 4.2. *Let ψ be a nontrivial character in $\widehat{\mathbb{F}_q}$, and $\kappa, \lambda, \mu, \nu \in \widehat{\mathbb{F}_q^*}$. If*

$$H = \sum_{\xi \in \widehat{\mathbb{F}_q^*}} G(\psi, \xi\kappa)G(\psi, \xi\lambda)\overline{G(\psi, \xi\mu)G(\psi, \xi\nu)},$$

$$M = \begin{cases} (q-1)^3 & \text{if } \{\kappa, \lambda\} = \{\mu, \nu\}, \\ 0 & \text{otherwise}, \end{cases}$$

then $|H - M| \le (q-1)q\sqrt{q}$.

Proof. First we consider the case where $\{\kappa, \lambda\} = \{\mu, \nu\}$, wherein

$$H = \sum_{\xi \in \widehat{\mathbb{F}_q^*}} |G(\psi, \xi\kappa)|^2 |G(\psi, \xi\lambda)|^2.$$

One can work out from parts (iii) and (iv) of Lemma 4.1 that $H = (q-2)q^2 + 1$ if $\kappa = \lambda = \mu = \nu$ and $H = (q-3)q^2 + 2q$ otherwise. Thus $H - M = (q-1)(q-2)$ or $1-q$.

Now we consider the case where $\{\kappa, \lambda\} \neq \{\mu, \nu\}$, wherein

$$\begin{aligned} H &= \sum_{\xi \in \widehat{\mathbb{F}_q^*}} \sum_{w,x,y,z \in \mathbb{F}_q^*} \psi(w+x-y-z)\xi(wxy^{-1}z^{-1})\kappa(w)\lambda(x)\overline{\mu(y)\nu(z)} \\ &= (q-1) \sum_{\substack{w,x,y,z \in \mathbb{F}_q^* \\ wx = yz}} \psi(w+x-y-z)\kappa(w)\lambda(x)\overline{\mu(y)\nu(z)}. \end{aligned}$$

Now reparametrize the sum with $w = uy$ and $z = ux$ to obtain

$$H = (q-1) \sum_{u,x,y \in \mathbb{F}_q^*} \overline{\psi}((u-1)x)\psi((u-1)y)\kappa\overline{\nu}(u)\lambda\overline{\nu}(x)\kappa\overline{\mu}(y),$$

and since $\{\kappa, \lambda\} \neq \{\mu, \nu\}$, we can restrict to $u \neq 1$ without changing the value of the summation. Then Lemma 4.1(v) tells us that when we sum over x and y, we obtain

$$H = (q-1)G(\psi, \kappa\overline{\mu})G(\overline{\psi}, \lambda\overline{\nu}) \sum_{u \neq 0,1} \overline{\kappa\lambda}\mu\nu(u-1)\kappa\overline{\nu}(u).$$

Now $\overline{\kappa\lambda}\mu\nu$ and $\kappa\overline{\nu}$ cannot both be the trivial character since $\{\kappa, \lambda\} \neq \{\mu, \nu\}$. If $\overline{\kappa\lambda}\mu\nu$ is trivial, then the sum over u is -1; if $\kappa\overline{\nu}$ is trivial, the sum is $-\overline{\kappa\lambda}\mu\nu(-1)$; otherwise, let ω be a generator of $\widehat{\mathbb{F}_q^*}$ and we can write the sum over u as

$$\sum_{u \neq 0,1} \omega((u-1)^a u^b)$$

for some nonzero $a, b \in \mathbb{Z}/(q-1)\mathbb{Z}$, and use the Weil bound [Weil 1948; Lidl and Niederreiter 1997, Theorem 5.41] to see that this sum is bounded in magnitude by $\sqrt{q}$. We can use this fact, along with Lemma 4.1(iii), (iv), to see that $|H| \leq (q-1)q\sqrt{q}$. □

Similarly, we estimate the summation H appearing in Corollary 3.3.

Proposition 4.3. *Let χ be a nontrivial character in $\widehat{\mathbb{F}_q^*}$, and $\kappa, \lambda, \mu, \nu \in \widehat{\mathbb{F}_q}$. If*

$$H = \sum_{\xi \in \widehat{\mathbb{F}_q}} G(\xi\kappa, \chi)G(\xi\lambda, \chi)\overline{G(\xi\mu, \chi)G(\xi\nu, \chi)}$$

and

$$M = \begin{cases} q^3 & \text{if } \{\kappa, \lambda\} = \{\mu, \nu\}, \\ q^3 & \text{if } \kappa = \lambda,\ \mu = \nu, \text{ and } \chi \text{ is the quadratic character}, \\ 0 & \text{otherwise}, \end{cases}$$

then $|H - M| \le 3q^2\sqrt{q}$.

Proof. Let ϵ be the canonical additive character over $\mathbb{F}_q$. Then for any $\eta \in \widehat{\mathbb{F}}_q$, there is a unique $y \in \mathbb{F}_q$ such that $\eta(z) = \epsilon(yz)$ for all $z \in \mathbb{F}_q$. Let a, b, c, d be chosen so that $\kappa(z) = \epsilon(az)$, $\lambda(z) = \epsilon(bz)$, $\mu(z) = \epsilon(cz)$, and $\nu(z) = \epsilon(dz)$ for all $z \in \mathbb{F}_q$. Furthermore, we shall parametrize the sum of ξ over $\widehat{\mathbb{F}}_q$ in the definition of H by a sum over $x \in \mathbb{F}_q$, and replace $\xi(z)$ with $\epsilon(xz)$ wherever it occurs. Thus, in view Lemma 4.1(v) and (ii), we have

$$H = |G(\epsilon, \chi)|^4 \sum_{x \in \mathbb{F}_q} \overline{\chi}((x+a)(x+b))\chi((x+c)(x+d)),$$

and $|G(\epsilon, \chi)| = \sqrt{q}$ by Lemma 4.1(iv). Let m be the order of χ. Then

$$H = q^2 \sum_{x \in \mathbb{F}_q} \chi((x+a)^{m-1}(x+b)^{m-1}(x+c)(x+d)).$$

The magnitude of the Weil sum over x is bounded by $3\sqrt{q}$ unless the polynomial $(x+a)^{m-1}(x+b)^{m-1}(x+c)(x+d)$ is an m-th power in $\mathbb{F}_q[x]$. (See [Weil 1948; Lidl and Niederreiter 1997, Theorem 5.41].) It is an m-th power only if $\{a, b\} = \{c, d\}$ or if $m = 2$, $a = b$, and $c = d$, in which cases the Weil sum is either $q - 1$ (if $a = b = c = d$) or $q - 2$ (if there are two distinct roots). □

5. Asymptotic L^4 norm

We prove Theorems 1.3(i), 1.4(i), and 1.5(i) in this section, by using the propositions from the previous section with Corollaries 3.2 and 3.3.

Proof of Theorems 1.4(i) *and* 1.5(i). Let χ be a nontrivial character in $\widehat{\mathbb{F}_q^*}$, let α be an epimorphism from $\mathbb{Z}^e$ to $\mathbb{F}_q$, let $t \in \mathbb{Z}^e$, and let $S = S_1 \times \cdots \times S_e$ be a box where each S_j is a nonempty segment of the form $\{0, 1, \ldots, s_j - 1\}$. Recall from Section 2 our notational convention that z^s is shorthand for $z_1^{s_1} \cdots z_e^{s_e}$ when $s = (s_1, \ldots, s_e) \in \mathbb{Z}^e$. Let $f(z)$ be the multiplicative character polynomial $\sum_{s \in S} \chi(\alpha(s + t))z^s$. We shall calculate $\|f\|_4^4$ first, and then investigate what happens asymptotically to this quantity in the limits considered in Theorems 1.4(i) and 1.5(i). By Corollary 3.3, we have

$$\|f\|_4^4 = \frac{1}{q^5} \sum_{\substack{a,b,c,d \in U \\ a+b=c+d}} \sum_{\kappa,\lambda,\mu,\nu \in \widehat{\mathbb{F}}_q} \overline{\kappa'(a)\lambda'(b)}\mu'(c)\nu'(d)H(\kappa, \lambda, \mu, \nu), \tag{4}$$

where we let $U = S + t$, and for any $\eta \in \widehat{\mathbb{F}}_q$, we let $\eta' = \eta \circ \alpha$, and

$$H(\kappa, \lambda, \mu, \nu) = \sum_{\xi \in \widehat{\mathbb{F}}_q} G(\xi\kappa, \chi) G(\xi\lambda, \chi) \overline{G(\xi\mu, \chi) G(\xi\nu, \chi)}.$$

By Proposition 4.3, we can write $H(\kappa, \lambda, \mu, \nu) = M(\kappa, \lambda, \mu, \nu) + N(\kappa, \lambda, \mu, \nu)$ with

$$M(\kappa, \lambda, \mu, \nu) = \begin{cases} q^3 & \text{if } \{\kappa, \lambda\} = \{\mu, \nu\}, \\ q^3 & \text{if } \kappa = \lambda,\ \mu = \nu, \text{ and } \chi \text{ is the quadratic character}, \\ 0 & \text{otherwise}, \end{cases}$$

and

$$|N(\kappa, \lambda, \mu, \nu)| \leq 3q^2\sqrt{q}, \tag{5}$$

for all $\kappa, \lambda, \mu, \nu \in \widehat{\mathbb{F}}_q^*$.

If χ is nonquadratic, when we write out separately the contributions from M and N to (4), we get $\|f\|_4^4 = A + B - D + E$, where

$$A = \frac{1}{q^2} \sum_{\substack{a,b,c,d \in U \\ a+b=c+d}} \sum_{\kappa, \lambda \in \widehat{\mathbb{F}}_q} \kappa'(c-a)\lambda'(d-b),$$

$$B = \frac{1}{q^2} \sum_{\substack{a,b,c,d \in U \\ a+b=c+d}} \sum_{\kappa, \lambda \in \widehat{\mathbb{F}}_q} \kappa'(d-a)\lambda'(c-b),$$

$$D = \frac{1}{q^2} \sum_{\substack{a,b,c,d \in U \\ a+b=c+d}} \sum_{\kappa \in \widehat{\mathbb{F}}_q} 1,$$

$$E = \frac{1}{q^5} \sum_{\kappa, \lambda, \mu, \nu \in \widehat{\mathbb{F}}_q} N(\kappa, \lambda, \mu, \nu) \sum_{\substack{a,b,c,d \in U \\ a+b=c+d}} \overline{\kappa'(a)\lambda'(b)} \mu'(c)\nu'(d).$$

Here A accounts for the value of M when $(\kappa, \lambda) = (\mu, \nu)$, and B accounts for the value of M when $(\kappa, \lambda) = (\nu, \mu)$, while D corrects for the double counting by A and B of the case $\kappa = \lambda = \mu = \nu$.

Note that $A = B$, and that A counts the number of $(a, b, c, d) \in U^4$ with $c - a = b - d \in \ker\alpha$. If we write $a = (a_1, \ldots, a_e)$, $b = (b_1, \ldots, b_e)$, $c = (c_1, \ldots, c_e)$, and $d = (d_1, \ldots, d_e)$, then $c - a \in \ker\alpha$ is equivalent to $c_1 - a_1 \equiv \cdots \equiv c_e - a_e \equiv 0 \pmod{p}$, because α is an epimorphism from $\mathbb{Z}^e$ to $\mathbb{F}_q = \mathbb{F}_{p^e}$ and so factors as $\alpha = \gamma \circ \beta$, with $\beta\colon \mathbb{Z}^e \to (\mathbb{Z}/p\mathbb{Z})^e$ coordinate-wise reduction modulo p and $\gamma\colon (\mathbb{Z}/p\mathbb{Z})^e \to \mathbb{F}_q$ a group isomorphism. Now $U = U_1 \times \cdots \times U_e$ with each $U_j = \{t_j, t_j + 1, \ldots, t_j + |S_j| - 1\}$, so for each $n \in \mathbb{Z}$, there are $\max(0, |S_j| - p|n|)$ ways for $c_j - a_j$ to equal pn and the same number of ways for $b_j - d_j$ to equal pn.

So

$$A = B = \prod_{j=1}^{e} \sum_{n_j \in \mathbb{Z}} \max(0, |S_j| - p|n_j|)^2.$$

On the other hand, qD counts the number of $(a, b, c, d) \in U^4$ with $c - a = b - d$, so by the same argument we just used (with modulus 1 instead of p),

$$qD = \prod_{j=1}^{e} \sum_{n_j \in \mathbb{Z}} \max(0, |S_j| - |n_j|)^2,$$

from which we can compute

$$D = \prod_{j=1}^{e} \left(\frac{2|S_j|^3 + |S_j|}{3p} \right).$$

Now we bound E via two bounds: our bound (5) on N, and a technical result, Lemma A.1 in the Appendix, bounding the inner sum of E. We satisfy the condition on α demanded by this lemma, since $\alpha = \gamma \circ \beta$ with $\beta \colon \mathbb{Z}^e \to (\mathbb{Z}/p\mathbb{Z})^e$ coordinate-wise reduction modulo p and $\gamma \colon (\mathbb{Z}/p\mathbb{Z})^e \to \mathbb{F}_q$ a group isomorphism. With these two bounds, we obtain

$$|E| \leq 3 \cdot 64^e q\sqrt{q} \prod_{j=1}^{e} \max\left(1, \frac{|S_j|}{p}\right)^3 \prod_{j=1}^{e} (1 + \log p)^3. \tag{6}$$

Now we divide $\|f\|_4^4 = A + B - D + E$ by $\|f\|_2^4$ and consider the limit where each $|S_j|/p \to \sigma_j$ as $q \to \infty$, that is, consider what happens in a size-stable family of polynomials. Another technical result, Lemma A.2, shows that we can replace the denominator $\|f\|_2^4$ with $|S|^2$ without changing the limit. Then recall the definition (3) of Ω, and note that $A/|S|^2$ and $B/|S|^2$ tend to $\prod_{j=1}^{e} \Omega(1/\sigma_j, 0)$, that $D/|S|^2$ tends to $\left(\frac{2}{3}\right)^e \prod_{j=1}^{e} \sigma_j$, and that $|E|/|S|^2$ tends to 0 in this limit.

If χ is quadratic, the proof is done in the same manner, except that there is now a contribution from M in the case where $\kappa = \lambda$ and $\mu = \nu$, and so we get $\|f\|_4^4 = A + B + C - 2D + E$, where A, B, D, and E are as defined above, and

$$C = \frac{1}{q^2} \sum_{\substack{a,b,c,d \in U \\ a+b=c+d}} \sum_{\kappa,\mu \in \widehat{\mathbb{F}_q}} \kappa'(-a-b)\mu'(c+d).$$

Note that we subtract D twice now because A, B, and C count the case where $\kappa = \lambda = \mu = \nu$ three times. C counts the number of $(a, b, c, d) \in U^4$ with $a + b = c + d \in \ker \alpha$. Following the method we used to determine A, write $a = (a_1, \ldots, a_e)$, $b = (b_1, \ldots, b_e)$, $c = (c_1, \ldots, c_e)$, and $d = (d_1, \ldots, d_e)$, and note that $a + b \in \ker \alpha$ is equivalent to $a_1 + b_1 \equiv \cdots \equiv a_e + b_e \equiv 0 \pmod{p}$.

Since $U = U_1 \times \cdots \times U_e$ with each $U_j = \{t_j, t_j+1, \ldots, t_j+|S_j|-1\}$, there are $\max(0, |S_j| - |np - (2t_j + |S_j| - 1)|)$ ways to obtain $a_j + b_j = np$ with $(a_j, b_j) \in U_j^2$, and the same number of ways to obtain $c_j + d_j = np$ with $(c_j, d_j) \in U_j^2$, so

$$C = \prod_{j=1}^{e} \sum_{n_j \in \mathbb{Z}} \max\bigl(0, |S_j| - |pn_j - |S_j| - 2t_j + 1|\bigr)^2,$$

and if we have both size- and translation-stability, then $|S_j|/p \to \sigma_j$ and $t_j/p \to \tau_j$ as $q \to \infty$, so that $C/|S|^2 \to \prod_{j=1}^{e} \Omega\left(1/\sigma_j, 1 + 2\tau_j/\sigma_j\right)$. □

Proof of Theorem 1.3(i). The proof is the same, *mutatis mutandis*, as for the $e = 1$ case of Theorem 1.4(i), with the roles of $\mathbb{F}_q$ and $\mathbb{F}_q^*$ exchanged. Corollary 3.2 and Proposition 4.2 replace Corollary 3.3 and Proposition 4.3, and Lemma A.2 becomes unnecessary as $\|f\|_2^2$ for an additive character polynomial f is always precisely equal to the cardinality of the support of f. These, and other attendant minor changes resulting from the exchange of $\mathbb{F}_q$ and $\mathbb{F}_q^*$, cause (5) to become $|N(\kappa, \lambda, \mu, \nu)| \le (q-1)q\sqrt{q}$, and (6) to become

$$|E| \le 64q\sqrt{q} \max\bigl(1, |S|/(q-1)\bigr)^3 \bigl(1 + \log(q-1)\bigr)^3,$$

and any other printed instance of p or q should be replaced with $q - 1$. □

6. Minimizing the asymptotic ratio of L^4 to L^2 norm

Here we prove Theorems 1.3(ii), 1.4(ii), and 1.5(ii) by finding the limiting sizes and (for quadratic multiplicative character polynomials) the limiting translations that globally minimize the ratio of the L^4 to L^2 norm.

Proof of Theorem 1.4(ii). In view of Theorem 1.4(i), we are trying to minimize the limiting ratio of norms, given by the function

$$K(x_1, \ldots, x_e) = -\frac{2^e}{3^e} \prod_{j=1}^{e} x_i + 2 \prod_{j=1}^{e} \Phi(x_j),$$

for $x_1, \ldots, x_e$ positive real numbers (the limiting sizes), where for positive x, we define

$$\Phi(x) = \Omega\Bigl(\frac{1}{x}, 0\Bigr) = \sum_{n \in \mathbb{Z}} \max\Bigl(0, 1 - \frac{|n|}{x}\Bigr)^2, \tag{7}$$

which is differentiable for $x \neq 0$ and is C^∞ for $x \notin \mathbb{Z}$.

Step 1. We can assume that each $x_j > 1$ because otherwise the partial derivative of K with respect to x_j would be negative.

Step 2. We can assume that $(x_1, \dots, x_e) \in (1, 3)^e$: indeed, Lemma A.6 shows that $K(x_1, \dots, x_e) \geq \Phi(x_1) \cdots \Phi(x_e)$, and note that $\Phi(x)$ is increasing for $x > 1$, that $\Phi(1) = 1$, $\Phi(3) = \frac{19}{9} > 2$, and $K(1, \dots, 1) = 2 - \left(\frac{2}{3}\right)^e < 2$. This proves that a global minimum exists and lies in $(1, 3)^e$: the closure of $(1, 3)^e$ is compact and K is continuous thereupon.

Step 3. Suppose $(\sigma_1, \dots, \sigma_e)$ to be global minimizer of K. Then the partial derivatives of K must vanish there, whence for each $k \in \{1, \dots, e\}$, we have $2u(\sigma_k) \prod_{j=1}^{e} U(\sigma_j) = 1$, where $u(x) = x\Phi'(x)/\Phi(x)$ and $U(x) = 3\Phi(x)/(2x)$ for $x > 0$.

Step 4. Then one can show that $U(x)$ is strictly decreasing on $[1, 3]$, with $U(3) = \frac{19}{18}$. Thus we must have $u(\sigma_k) \leq \frac{1}{2} \cdot \left(\frac{18}{19}\right)^e < \frac{1}{2}$ for all $k \in \{1, \dots, e\}$. Then examination of $u(x)$ shows that $u(x)$ strictly increases from 0 to $\frac{1}{2}$ for $x \in [1, 2 - \sqrt{2/3}]$, and then $u(x) > \frac{1}{2}$ for $x \in (2 - \sqrt{2/3}, 3)$. This then forces $\sigma_1 = \cdots = \sigma_e < 2 - \sqrt{2/3} < \frac{6}{5}$.

Step 5. Now $U(\sigma_1) > U\left(\frac{6}{5}\right) > \frac{9}{7}$, so this forces $u(\sigma_1) < \frac{1}{2} \cdot \left(\frac{7}{9}\right)^e \leq \frac{7}{18}$, which in turn forces $\sigma_1 < \frac{8}{7}$. Then $U(\sigma_1) > U\left(\frac{8}{7}\right) > \frac{4}{3}$, so this forces $u(\sigma_1) < \frac{1}{2} \cdot \left(\frac{3}{4}\right)^e$. Since $u(x) \geq 8(x-1)/3$ for $x \in \left[1, \frac{8}{7}\right]$, this forces $\sigma_1 < 1 + 3^{e+1}/2^{2e+4}$.

Step 6. Now our problem is reduced to the single-variable minimization of

$$\Theta(x) = K(x, \dots, x) = -\left(\frac{2x}{3}\right)^e + 2\Phi(x)^e$$

on the interval $(1, 1 + 3^{e+1}/2^{2e+4})$. It is not hard to see that $d\Theta/dx$ vanishes if and only if $x^{3e} - (3^e/2^{e-3})(x-1)(3x^2-4x+2)^{e-1}$ vanishes. Meanwhile $d^2\Theta/dx^2 > 0$ on our interval: by computing its value and then dropping a nonnegative term, we can see that $d^2\Theta/dx^2$ is at least

$$-e(e-1)\frac{2^e x^{e-2}}{3^e} + 8e\frac{3-2x}{x^4}\Phi(x)^{e-1} \geq -e(e-1)\frac{19^e}{24^e} + 2e > 0.$$

This proves that there is a unique minimum: the unique root a_e of

$$x^{3e} - \frac{3^e}{2^{e-3}}(x-1)(3x^2-4x+2)^{e-1}$$

lying in $(1, 1 + 3^{e+1}/2^{2e+4})$. □

Proof of Theorem 1.3(ii). This is accomplished exactly as the $e = 1$ case of the proof of Theorem 1.4(ii) above, save that Lemma A.5 replaces Lemma A.6. □

Proof of Theorem 1.5(ii). In view of Theorem 1.5 (i), we are trying to minimize the limiting ratio of norms, given by the function

$$-\frac{2^{e+1}}{3^e}\prod_{j=1}^{e} x_j + 2\prod_{j=1}^{e}\Omega\left(\frac{1}{x_j}, 0\right) + \prod_{j=1}^{e}\Omega\left(\frac{1}{x_j}, 1 + \frac{2y_j}{x_j}\right) \tag{8}$$

for $x_1, \dots, x_e$ positive real numbers (the limiting sizes) and $y_1, \dots, y_e$ arbitrary real numbers (the limiting translations).

Step 1. We invoke Lemma A.4(i) to see that we can confine our search to $x_1, \dots, x_e$ greater than or equal to $\frac{1}{2}$. For as long as $x_j \le \frac{1}{2}$, the lemma shows that we can always arrange for y_j to be such that $\Omega(x_j^{-1}, 1 + 2x_j^{-1} y_j) = 0$, and we note that $\Omega(x_j^{-1}, 0) = 1$ for all $x_j \in \left(0, \frac{1}{2}\right]$. Thus we can increase x_j to $\frac{1}{2}$ to lower (8) through the term $-(2^{e+1}/3^e) \prod_{j=1}^{e} x_j$ while keeping the other terms constant.

Step 2. Now we invoke Lemma A.4(ii) to see that for fixed $x_1, \dots, x_e$, we minimize the last term of (8) if and only if we arrange that

$$y_j \in \left\{ \frac{1-2x_j}{4} + \frac{m}{2} : m \in \mathbb{Z} \right\}$$

for each $j \in \{1, \dots, e\}$. The problem is thus reduced to the minimization of

$$\Lambda(x_1, \dots, x_e) = -\frac{2^{e+1}}{3^e} \prod_{j=1}^{e} x_j + 2 \prod_{j=1}^{e} \Phi(x_j) + \prod_{j=1}^{e} \Psi(x_j)$$

for $x_1, \dots, x_e$ positive real numbers, where $\Phi(x)$ is as defined in (7), and

$$\Psi(x) = \sum_{n \in \mathbb{Z}} \max\left(0, 1 - \frac{|2n+1|}{2x}\right)^2$$

for $x > 0$. Note that Ψ is differentiable for $x \neq 0$ and is C^∞ for $x \notin \mathbb{Z} + \frac{1}{2}$.

Step 3. We can assume that each $x_j > \frac{1}{2}$ because otherwise the partial derivative of Λ with respect to x_j would be negative.

Step 4. We can assume that $x_1, \dots, x_e \in (\frac{1}{2}, 3)$: indeed, Lemma A.6 shows that $\Lambda(x_1, \dots, x_e) \ge \Phi(x_1) \cdots \Phi(x_e)$, and note that $\Phi(x)$ is nondecreasing for $x > \frac{1}{2}$, that $\Phi\left(\frac{1}{2}\right) = 1$, $\Phi(3) = \frac{19}{9} > 2$, and $\Lambda(1, \dots, 1) = 2 - 2\left(\frac{2}{3}\right)^e + \left(\frac{1}{2}\right)^e < 2$. This proves that a global minimum exists and lies in $\left(\frac{1}{2}, 3\right)^e$: the closure of $\left(\frac{1}{2}, 3\right)^e$ is compact and Λ is continuous thereupon.

Step 5. Suppose that $(\sigma_1, \dots, \sigma_e)$ is a global minimizer of Λ. Then the partial derivatives of Λ must vanish there, whence

$$u(\sigma_k) \prod_{j=1}^{e} U(\sigma_j) + \tfrac{1}{2} v(\sigma_k) \prod_{j=1}^{e} V(\sigma_j) = 1, \tag{9}$$

where $u(x) = \dfrac{x\Phi'(x)}{\Phi(x)}$, $U(x) = \dfrac{3\Phi(x)}{2x}$, $v(x) = \dfrac{x\Psi'(x)}{\Psi(x)}$, and $V(x) = \dfrac{3\Psi(x)}{2x}$ for $x > \frac{1}{2}$.

Step 6. We can assume $\sigma_1, \dots, \sigma_e \in (1, 3)$: see (9) and note that $u(x) = 0$ for $x \in \left(\frac{1}{2}, 1\right]$, $\frac{1}{2} v(x) V(x) < 1$ for $x \in \left(\frac{1}{2}, 1\right]$, and $V(x) < 1$ for $x \in \left(\frac{1}{2}, 3\right)$.

Step 7. It is not difficult to show that $U(x)$ strictly decreases and $V(x)$ strictly increases on $[1, 3]$ with $U(3) = \frac{19}{18}$ and $V(1) = \frac{3}{4}$, and that $0 \le u(x) < 1 \le v(x)$ for $x \in [1, 3]$. Thus (9) shows that we must have $u(\sigma_k) < \left(\frac{18}{19}\right)^e (1 - 3^e/2^{2e+1})$ for all k. This forces $u(\sigma_k) < \frac{7}{10}$ for all k, and examination of the function u shows that $u(x) \ge \frac{7}{10}$ for $x \in \left[\frac{5}{2}, 3\right]$, and so we must have $\sigma_k < \frac{5}{2}$ for all k.

Now one can repeat the argument on the interval $\left[1, \frac{5}{2}\right]$ to show that every $\sigma_k < 2$, then repeat it again on $[1, 2]$ to show $\sigma_k < \frac{5}{4}$. Further repetitions give $\sigma_k < \frac{6}{5}$, $\sigma_k < \frac{13}{11}$, and $\sigma_k < \frac{7}{6}$. Since $U(x) > \frac{4}{3}$ while $v(x)$, $V(x) \ge 0$ for $x \in \left(1, \frac{7}{6}\right)$, we have $u(\sigma_k) < \left(\frac{3}{4}\right)^e$ for all k, and since $u(x) \ge 8(x-1)/3$ for $x \in \left[1, \frac{7}{6}\right]$, this means that $\sigma_k < 1 + 3^{e+1}/2^{2e+3}$ for all k.

Step 8. So we have $1 < \sigma_k < \min\left(\frac{7}{6}, 1 + 3^{e+1}/2^{2e+3}\right)$ for all k. Consider the products in (9): since each $\sigma_j \in \left(1, \frac{7}{6}\right)$, we have

$$\left(\tfrac{4}{3}\right)^e < \prod_{j=1}^{e} U(\sigma_j) < \left(\tfrac{3}{2}\right)^e, \quad \text{while } \left(\tfrac{3}{4}\right)^e < \prod_{j=1}^{e} V(\sigma_j) < \left(\tfrac{7}{8}\right)^e.$$

We now claim that for a given $A \in \left[\left(\frac{4}{3}\right)^e, \left(\frac{3}{2}\right)^e\right]$ and $B \in \left[\left(\frac{3}{4}\right)^e, \left(\frac{7}{8}\right)^e\right]$, there is at most one solution $x \in \left(1, \frac{7}{6}\right)$ to $Au(x) + \frac{1}{2}Bv(x) = 1$, which will force $\sigma_1 = \cdots = \sigma_e$. For if we set $w(x) = Au(x) + \frac{1}{2}Bv(x)$, we can show that $w'(x) > 0$ for $x \in \left(1, \frac{7}{6}\right)$: on this interval, we have $u(x) = 4(x-1)/(3x^2 - 4x + 2)$ $v(x) = 2/(2x-1)$, and it is not difficult to show that $u'(x) > \frac{3}{2}$ and $v'(x) > -4$, so that

$$w'(x) > \tfrac{3}{2}A - 2B \ge \left(\tfrac{3}{2}\right)\left(\tfrac{4}{3}\right)^e - 2\left(\tfrac{7}{8}\right)^e > 0.$$

Step 9. Now our problem is reduced to the single-variable minimization of

$$T(x) = \Lambda(x, \ldots, x) = -2(2x/3)^e + 2\Phi(x)^e + \Psi(x)^e$$

for $x \in (1, 1 + 3^{e+1}/2^{2e+3})$. It is not hard to see that dT/dx vanishes if and only if

$$x^{3e} - \frac{3^e}{2^{e-2}}(x-1)(3x^2 - 4x + 2)^{e-1} - \frac{3^e}{2^{2e}}(2x-1)^{2e-1}$$

does. Meanwhile we claim that the second derivative of T is strictly positive on our interval: by dropping some nonnegative terms we see that

$$\frac{d^2T}{dx^2}(x) \ge -e(e-1)\frac{2^{e+1}x^{e-2}}{3^e} + e\frac{8(3-2x)}{x^4}\Phi(x)^{e-1} + e\frac{3-4x}{x^4}\Psi(x)^{e-1}.$$

Thus for $e = 1$, the second derivative is at least $(27 - 20x)/x^4$, which is strictly positive on our interval $\left(1, 1 + \frac{9}{32}\right)$. For $e = 2$, we can use the fact that

$$0 \le \Psi(x) \le 1 \le \Phi(x)$$

on our interval $\left(1, 1+\frac{27}{128}\right)$ to show that the second derivative is at least

$$-\frac{16}{9}+\frac{54-40x}{x^4},$$

which is strictly positive on our interval. Finally, if $e \geq 3$, we have

$$1+\frac{3^{e+1}}{2^{2e+3}} < \frac{7}{6},$$

and so on our interval $(1, 1+3^{e+1}/2^{2e+3})$ we have $8(3-2x)/x^4 > \frac{20}{7}$, $\Phi(x) \geq 1$, $(3-4x)/x^4 \geq -1$, and $0 \leq \Psi(x) \leq 1$, so that

$$\frac{d^2T}{dx^2}(x) \geq -\tfrac{8}{9}\left(\tfrac{7}{9}\right)^{e-2}e(e-1)+\tfrac{13}{7}e > 0.$$

This proves that there is a unique global minimum for this single-variable problem: the unique root b_e of

$$x^{3e}-\frac{3^e}{2^{e-2}}(x-1)(3x^2-4x+2)^{e-1}-\frac{3^e}{2^{2e}}(2x-1)^{2e-1}$$

lying in $(1, 1+3^{e+1}/2^{2e+3})$. Thus we have found that global minima are obtained precisely when $\sigma_1 = \cdots = \sigma_e = b_e$ and $\tau_j \in \{\frac{1}{4}(1-2b_e)+m/2 : m \in \mathbb{Z}\}$ for each $j \in \{1, \ldots, e\}$. □

7. Proof of Theorem 1.6

For A_e and B_e as defined in Theorem 1.6, we first show that $B_e < A_e$ for each $e \geq 1$. Given the minimizing conditions described in Theorems 1.4(ii) and 1.5(ii), it suffices to show that

$$-\frac{2^{e+1}}{3^e}x^e+2\Omega\Big(\frac{1}{x},0\Big)^e+\Omega\Big(\frac{1}{x},\frac{1}{2x}\Big)^e < -\frac{2^e}{3^e}x^e+2\Omega\Big(\frac{1}{x},0\Big)^e.$$

for $x \in \left[1, \frac{3}{2}\right]$. This follows if $\Omega(1/x, 1/2x) < 2x/3$, or using the definition (3) of Ω, if $4x^3-12x^2+12x-3 > 0$ for $x \in \left[1, \frac{3}{2}\right]$, which is routine to show.

Now we show that $B_{e_1+e_2} < B_{e_1}B_{e_2}$ for any $e_1, e_2 \geq 1$. We use a technical Lemma 7.1 below, which provides bounds on the B_e. It shows that if $e_1 \geq 5$ or $e_2 \geq 5$, then $B_{e_1}B_{e_2} > \sqrt[4]{2} > B_{e_1+e_2}$. So we may confine ourselves to the case where $1 \leq e_1 \leq e_2 \leq 4$. If we define $B_0 = 1$, then the bounds in Lemma 7.1 also show us that

$$\frac{B_1}{B_0} > \frac{B_2}{B_1} > \frac{B_3}{B_2} > \frac{B_4}{B_3} > \frac{B_5}{B_4}.$$

Thus we note that $B_{e_1}B_{e_2} > B_{e_1-1}B_{e_2+1}$. If $e_1+e_2 > 5$, we can repeat this argument to show that $B_{e_1}B_{e_2}$ is greater than $B_{e_1+e_2-5}B_5$, which we have already shown to exceed $B_{e_1+e_2}$. On the other hand, if $e_1+e_2 \leq 5$, repetition of the same argument produces $B_{e_1}B_{e_2} > B_0B_{e_1+e_2} = B_{e_1+e_2}$.

Lemma 7.1. *For each $e \geq 1$, let B_e be the minimum asymptotic ratio of L^4 to L^2 norm achievable by a family of e-variable quadratic multiplicative character polynomials as described in Theorem 1.5(ii). Then*

(i) $\sqrt[4]{103/89} < B_1 < \sqrt[4]{22/19}$,

(ii) $\sqrt[4]{86/65} < B_2 < \sqrt[4]{75/56}$,

(iii) $\sqrt[4]{142/95} < B_3 < \sqrt[4]{116/77}$,

(iv) $\sqrt[4]{100/61} < B_4 < \sqrt[4]{107/65}$,

(v) $\sqrt[4]{7/4} < B_5 < \sqrt[4]{128/73}$, *and*

(vi) $\sqrt[4]{7/4} < B_e < \sqrt[4]{2}$ *for all* $e \geq 6$.

Proof. By Section 6 of the proof of Theorem 1.5(ii) (see page 389), for each $e \geq 1$ the quantity B_e^4 is the minimum of the function

$$T(x) = -\frac{2^{e+1}}{3^e}x^e + 2\Omega\Big(\frac{1}{x}, 0\Big)^e + \Omega\Big(\frac{1}{x}, \frac{1}{2x}\Big)^e$$

on the interval $(1, 1+3^{e+1}/2^{2e+3})$, upon which the second derivative of T is shown to be positive. Thus if we can find x_1, x_2, x_3 in this interval with $x_1 < x_2 < x_3$ and $T(x_2) < \min(T(x_1), T(x_3))$, then we will have shown that the minimizing value of x lies in the interval (x_1, x_3). Then we can use $B_e^4 \leq T(x_2)$ for our upper bound and, by the monotonicity of $\Omega(1/x, 0)$ and $\Omega(1/x, 1/(2x))$, we can use

$$B_e^4 > -\frac{2^{e+1}}{3^e}x_3^e + 2\Omega\Big(\frac{1}{x_1}, 0\Big)^e + \Omega\Big(\frac{1}{x_1}, \frac{1}{2x_1}\Big)^e$$

as a lower bound. We use this technique to prove bounds in (i)–(v). The calculations done by hand are tedious, so here we simply state choices of the triple (x_1, x_2, x_3) that establish stricter bounds than the ones we claim above:

for B_1 use $\big(\frac{55}{52}, \frac{128}{121}, \frac{73}{69}\big)$, for B_2 use $\big(\frac{18}{17}, \frac{17}{16}, \frac{16}{15}\big)$, for B_3 use $\big(\frac{21}{20}, \frac{20}{19}, \frac{19}{18}\big)$,
for B_4 use $\big(\frac{26}{25}, \frac{25}{24}, \frac{24}{23}\big)$, and for B_5 use $\big(\frac{36}{35}, \frac{35}{34}, \frac{34}{33}\big)$.

Henceforth assume that $e \geq 6$, and let b_e be the unique value in the interval $(1, 1+3^{e+1}/2^{2e+3})$ such that $T(b_e) = B_e^4$. Now note that $\Omega(1/x, 0) \geq 1$ and $\Omega(1/x, 1/(2x)) \geq \frac{1}{2}$ for $x \geq 1$, and that $2b_e/3 < (2^{15}+3^7)/(3\cdot 2^{14}) < 1$, so that

$$B_6^4 > -2\Big(\frac{2^{15}+3^7}{3\cdot 2^{14}}\Big)^6 + 2 + \frac{1}{2^6} > \frac{7}{4},$$

and if $e \geq 7$, then

$$B_e^4 > -2\Big(\frac{2^{15}+3^7}{3\cdot 2^{14}}\Big)^7 + 2 > \frac{7}{4}.$$

This proves our lower bound on B_e when $e \geq 6$.

To prove our upper bound on B_e when $e \geq 6$, write $b_e = 1 + c_e$ with $0 < c_e < 3^{e+1}/2^{2e+3}$, and use the definition (3) of Ω and the fact that $1/(1+c_e) > 1 - c_e$ to estimate $B_e = T(b_e)$ as

$$\begin{aligned} B_e^4 &< -\frac{2^{e+1}}{3^e} + 2(1+2c_e^2)^e + \frac{1}{2^e}(1+c_e)^{2e} \\ &\leq -\frac{2^{e+1}}{3^e} + 2 + 4ec_e^2 + 2^{e+3}c_e^4 + \frac{1}{2^e}(1+c_e)^{2e}, \end{aligned}$$

where the second inequality follows from a crude approximation with the binomial expansion. Now note that

$$2^{e+3}c_e^4 < \frac{3^{4e+4}}{2^{7e+9}} < \frac{2^e}{6\cdot 3^e} \quad \text{and} \quad 4ec_e^2 < e\frac{3^{2e+2}}{2^{4e+4}} < \frac{5\cdot 2^e}{4\cdot 3^e},$$

so that

$$B_e^4 < 2 - \frac{7}{12}\cdot\frac{2^e}{3^e} + \frac{1}{2^e}(1+c_e)^{2e},$$

which will imply $B_e^4 < 2$ if we can show that $(3(1+c_e)^2/4)^e \leq \frac{7}{12}$. Given that $c_e \leq 3^7/2^{15} < \frac{1}{10}$, the quantity being raised to the e-th power on the left hand side is less than 1, so it suffices to show that $(3(\frac{11}{10})^2/4)^6 \leq \frac{7}{12}$. □

Appendix A: Proofs of auxiliary results

Here we collect, for the sake of completeness, technical results used in our proofs. The first is a bound on a character sum used in the proofs of Theorems 1.3(i), 1.4(i), and 1.5(i) in Section 5.

Lemma A.1. *Let Γ be a finite abelian group, n a positive integer, and*

$$\pi_1, \ldots, \pi_n \in \mathrm{Hom}(\mathbb{Z}, \Gamma)$$

such that $\operatorname{im}\pi_1 + \cdots + \operatorname{im}\pi_n$ is the internal direct sum of $\operatorname{im}\pi_1, \ldots, \operatorname{im}\pi_n$ in Γ. Let $\pi \in \mathrm{Hom}(\mathbb{Z}^n, \Gamma)$ with $\pi(u_1, \ldots, u_n) = \pi_1(u_1) + \cdots + \pi_n(u_n)$ for all $(u_1, \ldots, u_n) \in \mathbb{Z}^n$, and let $U = U_1 \times \cdots \times U_n$ be a box in $\mathbb{Z}^n$. Then

$$T = \sum_{\kappa,\lambda,\mu,\nu\in\widehat{\Gamma}} \left| \sum_{\substack{a,b,c,d\in U \\ a+b=c+d}} \overline{\kappa(\pi(a))\lambda(\pi(b))}\mu(\pi(c))\nu(\pi(d)) \right|$$

is no greater than

$$64^n|\Gamma|^4 \prod_{j=1}^n \max\left(1, \frac{|U_j|}{|\operatorname{im}\pi_j|}\right)^3 \prod_{j=1}^n (1 + \log|\operatorname{im}\pi_j|)^3.$$

Proof. Write $K = \bigoplus_{j=1}^n \operatorname{im}\pi_j$, so that $\widehat{K} = \prod_{j=1}^n \widehat{\operatorname{im}\pi_j}$. Each character of K extends to $[\Gamma : K]$ characters of Γ, and for any $\eta \in \widehat{\operatorname{im}\pi_j}$, let $\eta' \in \widehat{\mathbb{Z}}$ be $\eta \circ \pi_j$,

so that

$$T = [\Gamma : K]^4 \prod_{j=1}^{n} \sum_{\kappa_j,\lambda_j,\mu_j,\nu_j \in \widehat{\operatorname{im}\pi_j}} \left| \sum_{\substack{a_j,b_j,c_j,d_j \in U_j \\ a_j+b_j=c_j+d_j}} \overline{\kappa'_j(a_j)\lambda'_j(b_j)}\mu'_j(c_j)\nu'_j(d_j) \right|,$$

and so it suffices to prove the bound when $n = 1$ and π is surjective. In this case Γ is a finite cyclic group, which we identify with $\mathbb{Z}/m\mathbb{Z}$ by identifying $\pi(1)$ with $1 \in \mathbb{Z}/m\mathbb{Z}$. Then we set $\epsilon_a = \exp(2\pi i a/m)$ for every $a \in \mathbb{Z}/m\mathbb{Z}$, and note that $\widehat{\Gamma}$ is the set of maps $a \mapsto \epsilon_{xa}$ with $x \in \mathbb{Z}/m\mathbb{Z}$. Thus,

$$T = \sum_{w,x,y,z \in \mathbb{Z}/m\mathbb{Z}} \left| \sum_{\substack{a,b,c,d \in U \\ a+b=c+d}} \epsilon_{-wa-xb+yc+zd} \right|.$$

U is a set of consecutive integers in $\mathbb{Z}$, and note that translation of U does not influence the magnitude of the inner sum in T, so without loss of generality, we assume that $U = \{0, 1, \ldots, |U| - 1\}$. Then reparametrize the outer sum in T with $x' = w - x$, $y' = y - w$, $z' = z - w$, and w to obtain

$$T = m \sum_{x',y',z' \in \mathbb{Z}/m\mathbb{Z}} \left| \sum_{\substack{a,b,c,d \in U \\ a+b=c+d}} \epsilon_{-x'b+y'c+z'd} \right|,$$

which is not more than $64m \max(m, |U|)^3(1 + \log m)^3$ by [Jedwab et al. 2013b, Lemma 2.2]. □

The next result is used in the proofs of Section 5 to understand the asymptotic behavior of the L^2 norm for multiplicative character polynomials.

Lemma A.2. *Let $\{f_\iota\}_{\iota \in I}$ be a size-stable family of e-variable multiplicative character polynomials with $\mathbb{F}_{q_\iota}$, S_ι, t_ι, and α_ι the field, support, translation, and arrangement, respectively, of f_ι for each $\iota \in I$. Then there is a Q and an N such that for all $\iota \in I$ with $q_\iota \geq Q$, we have $|S_\iota \cap (\ker \alpha_\iota - t_\iota)| \leq N$. Thus $|S_\iota \cap \ker(\alpha_\iota - t_\iota)|/|S_\iota| \to 0$ and $\|f_\iota\|_2^2/|S_\iota| \to 1$ as $q_\iota \to \infty$.*

Proof. Suppose that the limiting sizes for our size-stable family $\{f_\iota\}_{\iota \in I}$ of multiplicative character polynomials are $\sigma_1, \ldots, \sigma_e$. For each $\iota \in I$, let χ_ι be the character of f_ι, so that $f_\iota(z) = \sum_{s \in S_\iota} \chi_\iota(\alpha_\iota(s + t_\iota))z^s$, and let $p_\iota = {q_\iota}^{1/e}$, which is the characteristic of the field $\mathbb{F}_{q_\iota}$ of f_ι. Since α_ι is an epimorphism, its restriction to each $p_\iota \times \cdots \times p_\iota$ cubical box in $\mathbb{Z}^e$ is a bijection to $\mathbb{F}_{q_\iota}$, and by the definition of size-stability, there is some Q such that for every $q_\iota \geq Q$, the support $S_\iota = S_{\iota,1} \times \cdots \times S_{\iota,e}$ can be covered with $N = \prod_{j=1}^{e} \left(\lfloor \sigma_j \rfloor + 1\right)$ such cubes, each of which contains one point of $\ker \alpha_\iota - t_\iota$, so $|S_\iota \cap (\ker \alpha_\iota - t_\iota)| \leq N$. Since the family is size-stable, $|S_\iota| \to \infty$ as $q_\iota \to \infty$. The squared L^2 norm of a polynomial is the sum of the squared magnitudes of its coefficients, and $\chi_\iota(\alpha_\iota(s + t_\iota)) = 0$ for $s \in S_\iota \cap (\ker \alpha_\iota - t_\iota)$ while

$|\chi_\iota(\alpha_\iota(s+t_\iota))| = 1$ for all other $s \in S_\iota$. Thus $\|f_\iota\|_2^2 = |S_\iota \smallsetminus (S_\iota \cap (\ker \alpha_\iota - t_\iota))|$, and so $\|f_\iota\|_2^2/|S_\iota| \to 1$ as $q_\iota \to \infty$. □

Recall from footnote 3 of Section 1C that we sometimes wish to obtain a Littlewood polynomial from a quadratic character polynomial $f(z) = \sum_{s\in S} \chi(\alpha(s+t))z^s$, but f may have some coefficients equal to 0 because an extended nontrivial multiplicative character χ has $\chi(0) = 0$. More generally, if χ is a nontrivial multiplicative character (not necessarily quadratic), we may wish to obtain from f a polynomial with coefficients of unit magnitude. So we replace the zero coefficient for each z^s such that $s \in S \cap (\ker \alpha - t)$ with a coefficient of unit magnitude. We may choose each replacement coefficient independently of the others, and any polynomial g resulting from such replacements is called a *unimodularization* of f. The following corollary to Lemma A.2 shows that unimodularizing all the polynomials in a size-stable family of multiplicative character polynomials does not affect asymptotic ratio $\|f\|_4/\|f\|_2$.

Corollary A.3. *Let $\{f_\iota\}_{\iota\in I}$ be a size-stable family of multiplicative character polynomials over fields $\{\mathbb{F}_{q_\iota}\}_{\iota\in I}$, and let g_ι be a unimodularization of f_ι for each $\iota \in I$. If r is a real number with $r \geq 2$ or if $r = \infty$, then $\|f_\iota\|_r/\|g_\iota\|_r \to 1$ as $q_\iota \to \infty$.*

Proof. If $u \in \mathbb{C}$ with $|u| = 1$ and if $s = (s_1, \ldots, s_e) \in \mathbb{Z}^e$, then L^r norm of $uz^s = uz_1^{s_1} \cdots z_e^{s_e}$ is 1. By Lemma A.2 there is an N and a Q such that whenever $q_\iota \geq Q$, the two polynomials f_ι and g_ι differ by the sum of N or fewer such monomials, and so by the L^r triangle inequality, the difference between $\|f_\iota\|_r$ and $\|g_\iota\|_r$ cannot be greater in magnitude than N. Now $\|g_\iota\|_r \geq \|g_\iota\|_2 = \sqrt{|S_\iota|}$ by monotonicity of L^r norms and the fact that the squared L^2 norm of a polynomial is the sum of the squared magnitudes of its coefficients, and $|S_\iota| \to \infty$ as $q_\iota \to \infty$ for a size-stable family. □

The next result is used in Section 6 to find the limiting translations that globally minimize the asymptotic ratio of L^4 to L^2 norm for quadratic character polynomials.

Lemma A.4. *Let x be a fixed nonzero value in $\mathbb{R}$ and let y vary over $\mathbb{R}$.*

(i) *If $|x| \geq 2$, the function $\Omega(x, y)$, considered as a function of y, achieves a global minimum value of 0 for $y \in \bigcup_{m\in\mathbb{Z}}[m|x|+1, (m+1)|x|-1]$ and for no other value of y.*

(ii) *If $0 < |x| \leq 2$, the function $\Omega(x, y)$, considered as a function of y, achieves a global minimum value of*

$$\Omega\Big(x, \frac{x}{2}\Big) = \sum_{n\in\mathbb{Z}} \max\big(0, 1 - \big|\big(n+\tfrac{1}{2}\big)x\big|\big)^2$$

for $y \in \big\{x\big(m+\frac{1}{2}\big) : m \in \mathbb{Z}\big\}$ and for no other value of y.

Proof. For part (i), note that all the terms of $\Omega(x, y)$ are nonnegative. Since $\Omega(x, y) = \Omega(-x, y)$, we may assume without loss of generality that $x \geq 2$, and then the term $\max(0, 1-|xn-y|)^2$ is nonvanishing if and only if $(y-1)/x < n < (y+1)/x$. Thus we can obtain a global minimum value of 0 if we can arrange that no integer lie in the interval $((y-1)/x, (y+1)/x)$. If m is the greatest integer lying below this interval (so that $y \geq mx+1$), then for the next integer $m+1$ to lie above the interval, it is necessary and sufficient that $y \leq (m+1)x - 1$.

For part (ii), it is clear from the definition (3) of Ω that $\Omega(-x, y) = \Omega(x, y)$, $\Omega(x, -y) = \Omega(x, y)$, and $\Omega(x, y) = \Omega(x, y+x)$. So without loss of generality we may restrict our attention to the case where $0 < x \leq 2$ and $0 \leq y \leq x/2$. In this case

$$\Omega(x, y) = \sum_{\lceil (y-1)/x \rceil \leq n \leq 0} (y - 1 - nx)^2 + \sum_{0 < n \leq \lfloor (y+1)/x \rfloor} (y + 1 - nx)^2,$$

and we reparametrize the sums to obtain

$$\Omega(x, y) = \sum_{0 \leq n \leq \lfloor (1-y)/x \rfloor} (y - 1 + nx)^2 + \sum_{0 \leq n \leq \lfloor (1+y-x)/x \rfloor} (y - x + 1 - nx)^2,$$

and calculate

$$\begin{aligned}\frac{\partial}{\partial y}\Omega(x, y) &= \sum_{0 \leq n \leq \lfloor (1-y)/x \rfloor} 2(y - 1 + nx) + \sum_{0 \leq n \leq \lfloor (1+y-x)/x \rfloor} 2(y - x + 1 - nx) \\ &= 2\left\lfloor \frac{y+1}{x} \right\rfloor (2y - x) + \sum_{\lfloor (1+y-x)/x \rfloor < n \leq \lfloor (1-y)/x \rfloor} 2(y - 1 + nx),\end{aligned}$$

because $(1-y)/x$ is greater than or equal to $(1+y-x)/x$, and note for the remainder of this proof that their difference is at most 1. Since $0 \leq y \leq x/2 \leq 1$, we can see that both terms in the last expression for our partial derivative are nonpositive, with the summation over n strictly negative if $y < x - 1$, and the other term is strictly negative if $x - 1 \leq y < x/2$. Thus our partial derivative is strictly negative for $0 \leq y < x/2$, and so for our ranges of x and y values, the unique minimum is obtained when $y = x/2$. □

The last two results are used in Section 6 to show that a large limiting size will make the ratio of L^4 to L^2 norm large.

Lemma A.5. *If $\{f_\iota\}_{\iota \in I}$ is a size-stable family, with limiting size σ, of additive character polynomials over fields $\{\mathbb{F}_{q_\iota}\}_{\iota \in I}$, then*

$$\liminf_{q_\iota \to \infty} \frac{\|f_\iota\|_4^4}{\|f_\iota\|_2^4} \geq \Omega\Big(\frac{1}{\sigma}, 0\Big).$$

Proof. For each $\iota \in I$, let f_ι have character ψ_ι, support S_ι, translation t_i, and arrangement α_ι, so that $f_\iota(z) = \sum_{s \in S_\iota} \psi_\iota(\alpha_\iota(s + t_i))z^s$. When we confine the values

of z to the complex unit circle, we have

$$\overline{f_\iota(z)} = \sum_{s\in S_\iota} \overline{\psi_\iota(\alpha_\iota(s+t_i))} z^{-s}.$$

We can consider $f_\iota(z)$ and $\overline{f_\iota(z)}$ formally as elements of $\mathbb{C}[z, z^{-1}]$, and view $\|f_\iota\|_4^4$ as the sum of the squared magnitudes of the coefficients of $f_\iota(z)\overline{f_\iota(z)}$. The coefficient of z^s in $f_\iota(z)\overline{f_\iota(z)}$ is

$$\sum_{\substack{u,v\in S_\iota \\ u-v=s}} \psi_\iota(\alpha_\iota(u+t_i))\overline{\psi_\iota(\alpha_\iota(v+t_i))}.$$

Since α_ι is an epimorphism from $\mathbb{Z}$ to $\mathbb{F}_{q_\iota}^*$, we see that $\psi_\iota(\alpha_\iota(u+t_i)) = \psi_\iota(\alpha_\iota(v+t_i))$ whenever $u \equiv v \pmod{q_\iota - 1}$. Thus if $s \equiv 0 \pmod{q_\iota - 1}$, the coefficient of z^s in $f_\iota(z)\overline{f_\iota(z)}$ is equal to $|S_\iota \cap (s + S_\iota)|$. Since $\|f_\iota\|_4^4$ is the sum of the squared magnitudes of the coefficients of $f_\iota(z)\overline{f_\iota(z)}$ while $\|f_\iota\|_2^2$ is the coefficient of z^0 of the same, we have

$$\frac{\|f_\iota\|_4^4}{\|f_\iota\|_2^4} \geq \frac{\sum_{n\in\mathbb{Z}} |S_\iota \cap (n(q_\iota - 1) + S_\iota)|^2}{|S_\iota|^2},$$

and then we note that $|S_\iota \cap (n(q_\iota - 1) + S_\iota)| = \max(0, |S_\iota| - |n(q_\iota - 1)|)$ and apply the size-stability limit $|S_\iota|/(q_\iota - 1) \to \sigma$ as $q_\iota \to \infty$. □

Lemma A.6. *If $\{f_\iota\}_{\iota\in I}$ is a size-stable family, with limiting sizes $\sigma_1, \ldots, \sigma_e$, of e-variable multiplicative character polynomials over fields $\{\mathbb{F}_{q_\iota}\}_{\iota\in I}$, then*

$$\liminf_{q_\iota\to\infty} \frac{\|f_\iota\|_4^4}{\|f_\iota\|_2^4} \geq \prod_{j=1}^{e} \Omega\Big(\frac{1}{\sigma_j}, 0\Big).$$

Proof. For each $\iota \in I$, let f_ι have character χ_ι, support $S_\iota \subseteq \mathbb{Z}^e$, translation $t_i \in \mathbb{Z}^e$, and arrangement α_ι, so that $f_\iota(z_1, \ldots, z_e) = \sum_{s\in S_\iota} \chi_\iota(\alpha_\iota(s+t_i))z^s$, where we write z^s for $z_1^{s_1}\cdots z_e^{s_e}$ when $s = (s_1, \ldots, s_e)$. Our proof runs the same as that of the previous lemma for additive character polynomials once we replace ψ_ι with χ_ι, but we must take care of the fact that $\chi_\iota(\alpha_\iota(s + t_i)) = 0$ when $s \in -t_i + \ker\alpha_\iota$; otherwise, the coefficients are of unit magnitude. And of course the polynomials are in e variables and the coefficients have periodicity p in each direction. Thus if we define $V_\iota = S_\iota \setminus (-t_i + \ker\alpha_\iota)$ we have

$$\frac{\|f_\iota\|_4^4}{\|f_\iota\|_2^4} \geq \frac{\sum_{n\in\mathbb{Z}^e} |V_\iota \cap (np_\iota + V_\iota)|^2}{|V_\iota|^2},$$

but Lemma A.2 can be used to show that the ratio $|V_\iota \cap (np_\iota + V_\iota)|/|V_\iota|$ has the same limit as $|S_\iota \cap (np_\iota + S_\iota)|/|S_\iota|$ as $q_\iota \to \infty$. □

Acknowledgements

The author thanks Jonathan Jedwab and Kai-Uwe Schmidt for helpful suggestions on the presentation of these results.

References

[Baker and Montgomery 1990] R. C. Baker and H. L. Montgomery, "Oscillations of quadratic L-functions", pp. 23–40 in *Analytic number theory* (Allerton Park, IL, 1989), edited by B. C. Berndt et al., Progr. Math. **85**, Birkhäuser, Boston, MA, 1990. MR 91k:11071 Zbl 0718.11039

[Bernasconi 1987] J. Bernasconi, "Low autocorrelation binary sequences: Statistical mechanics and configuration space analysis", *J. Phys. France* **48**:4 (1987), 559–567.

[Bömer and Antweiler 1993] L. Bömer and M. Antweiler, "Optimizing the aperiodic merit factor of binary arrays", *Signal Process.* **30**:1 (1993), 1–13. Zbl 0769.93022

[Borwein 2002] P. Borwein, *Computational excursions in analysis and number theory*, CMS Books in Mathematics **10**, Springer, New York, 2002. MR 2003m:11045 Zbl 1020.12001

[Borwein and Choi 2000] P. Borwein and K.-K. S. Choi, "Merit factors of character polynomials", *J. London Math. Soc.* (2) **61**:3 (2000), 706–720. MR 2002d:11103 Zbl 1011.11059

[Borwein and Choi 2002] P. Borwein and K.-K. S. Choi, "Explicit merit factor formulae for Fekete and Turyn polynomials", *Trans. Amer. Math. Soc.* **354**:1 (2002), 219–234. MR 2002i:11065 Zbl 1010.11017

[Borwein et al. 2001] P. Borwein, K.-K. S. Choi, and S. Yazdani, "An extremal property of Fekete polynomials", *Proc. Amer. Math. Soc.* **129**:1 (2001), 19–27. MR 2001j:11061 Zbl 0987.11010

[Conrey et al. 2000] B. Conrey, A. Granville, B. Poonen, and K. Soundararajan, "Zeros of Fekete polynomials", *Ann. Inst. Fourier* (*Grenoble*) **50**:3 (2000), 865–889. MR 2001h:11108 Zbl 1007.11053

[Erdős 1957] P. Erdős, "Some unsolved problems", *Michigan Math. J.* **4** (1957), 291–300. MR 20 #5157 Zbl 0081.00102

[Erdős 1962] P. Erdős, "An inequality for the maximum of trigonometric polynomials", *Ann. Polon. Math.* **12** (1962), 151–154. MR 25 #5330 Zbl 0106.27702

[Fekete and Pólya 1912] M. Fekete and G. Pólya, "Über ein Problem von Laguerre", *Rend. Circ. Mat. Palermo* **34**:1 (1912), 89–120.

[Golay 1977] M. J. E. Golay, "The merit factor of long low autocorrelation binary sequences", *IEEE Trans. Inform. Theory* **23**:1 (1977), 43–51.

[Gulliver and Parker 2005] T. Gulliver and M. Parker, "The multivariate merit factor of a Boolean function", in *Proceedings of the IEEE ITSOC Information Theory Workshop 2005 on Coding and Complexity* (Rotorua, New Zealand, 2005), IEEE, New York, 2005.

[Høholdt 2006] T. Høholdt, "The merit factor problem for binary sequences", pp. 51–59 in *Applied algebra, algebraic algorithms and error-correcting codes*, edited by M. Fossorier et al., Lecture Notes in Comput. Sci. **3857**, Springer, Berlin, 2006. MR 2007c:94080 Zbl 1125.94008

[Høholdt and Jensen 1988] T. Høholdt and H. E. Jensen, "Determination of the merit factor of Legendre sequences", *IEEE Trans. Inform. Theory* **34**:1 (1988), 161–164. Zbl 0652.40006

[Jedwab 2005] J. Jedwab, "A survey of the merit factor problem for binary sequences", pp. 19–21 in *Sequences and their applications* (SETA 2004), edited by T. Helleseth et al., Lecture Notes in Computer Science **3486**, Springer, Berlin, 2005.

[Jedwab and Schmidt 2010] J. Jedwab and K.-U. Schmidt, "Appended m-sequences with merit factor greater than 3.34", pp. 204–216 in *Sequences and their applications — SETA 2010*, edited by C. Carlet and A. Pott, Lecture Notes in Comput. Sci. **6338**, Springer, Berlin, 2010. MR 2012m:94245 Zbl pre05781409

[Jedwab et al. 2013a] J. Jedwab, D. J. Katz, and K.-U. Schmidt, "Advances in the merit factor problem for binary sequences", *J. Combin. Theory Ser. A* **120**:4 (2013), 882–906. MR 3022619 Zbl 06154122

[Jedwab et al. 2013b] J. Jedwab, D. J. Katz, and K.-U. Schmidt, "Littlewood polynomials with small L^4 norm", *Adv. Math.* **241** (2013), 127–136. MR 3053707

[Jensen and Høholdt 1989] H. E. Jensen and T. Høholdt, "Binary sequences with good correlation properties", pp. 306–320 in *Applied algebra, algebraic algorithms and error-correcting codes* (Menorca, 1987), edited by L. Huguet and A. Poli, Lecture Notes in Comput. Sci. **356**, Springer, Berlin, 1989. MR 90h:94009 Zbl 0675.94010

[Jensen et al. 1991] J. M. Jensen, H. E. Jensen, and T. Høholdt, "The merit factor of binary sequences related to difference sets", *IEEE Trans. Inform. Theory* **37**:3, part 1 (1991), 617–626. MR 92j:94009 Zbl 0731.94011

[Kahane 1980] J.-P. Kahane, "Sur les polynômes à coefficients unimodulaires", *Bull. London Math. Soc.* **12**:5 (1980), 321–342. MR 82a:30003 Zbl 0443.30005

[Lidl and Niederreiter 1997] R. Lidl and H. Niederreiter, *Finite fields*, 2nd ed., Encyclopedia of Mathematics and its Applications **20**, Cambridge University Press, 1997. MR 97i:11115 Zbl 0866.11069

[Littlewood 1966] J. E. Littlewood, "On polynomials $\sum^n \pm z^m$, $\sum^n e^{\alpha_m i} z^m$, $z = e^{\theta_i}$", *J. London Math. Soc.* **41** (1966), 367–376. MR 33 #4237 Zbl 0142.32603

[Littlewood 1968] J. E. Littlewood, *Some problems in real and complex analysis*, D. C. Heath and Co. Raytheon Education Co., Lexington, MA, 1968. MR 39 #5777 Zbl 0185.11502

[Montgomery 1980] H. L. Montgomery, "An exponential polynomial formed with the Legendre symbol", *Acta Arith.* **37** (1980), 375–380. MR 82a:10041 Zbl 0369.10024

[Newman and Byrnes 1990] D. J. Newman and J. S. Byrnes, "The L^4 norm of a polynomial with coefficients ± 1", *Amer. Math. Monthly* **97**:1 (1990), 42–45. MR 91d:30006

[Pólya 1919] G. Pólya, "Verschiedene Bemerkungen zur Zahlentheorie", *Jahresber. Dtsch. Math.-Ver.* **28** (1919), 31–40. JFM 47.0882.06

[Schmidt 2011] K.-U. Schmidt, "The merit factor of binary arrays derived from the quadratic character", *Adv. Math. Commun.* **5**:4 (2011), 589–607. MR 2012k:94119 Zbl 1238.94025

[Weil 1948] A. Weil, "On some exponential sums", *Proc. Nat. Acad. Sci. U. S. A.* **34** (1948), 204–207. MR 10,234e Zbl 0032.26102

Received June 13, 2012.

DANIEL J. KATZ
DEPARTMENT OF MATHEMATICS
SIMON FRASER UNIVERSITY
8888 UNIVERSITY DRIVE
BURNABY BC V5A 1S6
CANADA

dkatz@sfu.ca

PACIFIC JOURNAL OF MATHEMATICS
Vol. 263, No. 2, 2013
dx.doi.org/10.2140/pjm.2013.263.399

DEGREE-THREE SPIN HURWITZ NUMBERS

JUNHO LEE

Gunningham (2012) calculated all spin Hurwitz numbers in terms of combinatorics of the Sergeev algebra. Here we use a spin curve degeneration to obtain a recursion formula for degree-three spin Hurwitz numbers.

Let D be a complex curve of genus h and N be a theta characteristic on D, that is, $N^2 = K_D$. The pair (D, N) is called a *spin curve* of genus h with parity $p \equiv h^0(N) \pmod 2$. For $i = 1, \dots, k$, let $m^i = (m^i_1, \dots, m^i_{\ell_i})$ be an odd partition of $d > 0$, namely, all components m^i_j are odd. Fix k points $q^1, \dots, q^k$ in D and consider degree-d maps $f : C \to D$ from possibly disconnected domains C of Euler characteristic χ that are ramified only over the fixed points q^i with ramification data m^i. Observe that the Riemann–Hurwitz formula shows

$$2d(1-h) - \chi + \sum_{i=1}^{k} (\ell(m^i) - d) = 0, \tag{0-1}$$

where $\ell(m^i) = \ell_i$ is the length of m^i. By the Hurwitz formula, the twisted line bundle

$$L_f = f^*N \otimes \mathbb{O}\left(\sum_{i,j} \tfrac{1}{2}(m^i_j - 1)x^i_j \right) \tag{0-2}$$

is a theta characteristic on C where $f^{-1}(q^i) = \{x^i_j\}_{1 \le j \le \ell_i}$ and f has multiplicity m^i_j at x^i_j. We define the parity $p(f)$ of a map f by

$$p(f) \equiv h^0(L_f) \pmod 2. \tag{0-3}$$

Given odd partitions $m^1, \dots, m^k$ of d, the spin Hurwitz number of genus h and parity p is defined as a (weighted) sum of (ramified) covers f satisfying (0-1) with sign determined by the parity $p(f)$:

$$H^{h,p}_{m^1,\dots,m^k} = \sum_f \frac{(-1)^{p(f)}}{|\mathrm{Aut}(f)|} \tag{0-4}$$

MSC2010: 14N35, 53D45.
Keywords: local Gromov–Witten invariants, spin curve, Schiffer variation.

Eskin, Okounkov, and Pandharipande [Eskin et al. 2008] calculated the genus $h = 1$ and odd parity spin Hurwitz numbers in terms of characters of the Sergeev group. Gunningham [2012] calculated all spin Hurwitz numbers in terms of combinatorics of the Sergeev algebra.

The trivial partition (1^d) of d is a partition whose components are all 1. If $m^k = (1^d)$, f has no ramification points over the fixed point q^k and hence we have

$$H^{h,p}_{m^1,\dots,m^{k-1},(1^d)} = H^{h,p}_{m^1,\dots,m^{k-1}}. \tag{0-5}$$

When all partitions $m^i = (1^d)$, denote the spin Hurwitz numbers (0-4) by $H^{h,p}_d$. These are dimension-zero local GW invariants $GT^{\mathrm{loc},h,p}_d$ of spin curve (D, N) that give all dimension-zero GW invariants of Kähler surfaces with a smooth canonical divisor; see [Kiem and Li 2007; 2011; Lee and Parker 2007; Maulik and Pandharipande 2008]. For notational simplicity, we set $H^{h,p}_{(3)^0} = H^{h,p}_3$ and for $k \geq 1$ write

$$H^{h,p}_{(3)^k}$$

for the spin Hurwitz numbers $H^{h,p}_{(3),\dots,(3)}$ with the same k partitions (3). Since there are two odd partitions (1^3) and (3) of $d = 3$, by (0-5) it suffices to compute $H^{h,p}_{(3)^k}$ for $k \geq 0$. The aim of this paper is to use a spin curve degeneration to obtain the following recursion formula.

Theorem 0.1. *If $h = h_1 + h_2$ and $p \equiv p_1 + p_2 \pmod 2$, then, for $k_1 + k_2 = k$,*

$$H^{h,p}_{(3)^k} = 3! H^{h_1,p_1}_{(3)^{k_1}} \cdot H^{h_2,p_2}_{(3)^{k_2}} + 3 H^{h_1,p_1}_{(3)^{k_1+1}} \cdot H^{h_2,p_2}_{(3)^{k_2+1}}. \tag{0-6}$$

One can use Theorem 0.1 and the result of [Eskin et al. 2008] to explicitly compute the spin Hurwitz numbers of degree $d = 3$. In Proposition 7.1, we show that

$$H^{h,\pm}_{(3)^k} = 3^{2h-2}[(-1)^k 2^{k+h-1} \pm 1], \tag{0-7}$$

where $+$ and $-$ denote the even and odd parities. When the degree d is 1 or 2, the dimension-zero local GW invariants are given by the formulas

$$GT^{\mathrm{loc},h,\pm}_1 = \pm 1 \quad \text{and} \quad GT^{\mathrm{loc},h,\pm}_2 = \pm 2^{h-1};$$

see Lemma 2.6 of [Lee 2013]. Since $GT^{\mathrm{loc},h,p}_d = H^{h,p}_d$ as mentioned above, formula (0-7) shows

$$GT^{\mathrm{loc},h,\pm}_3 = 3^{2h-2}(2^{h-1} \pm 1).$$

This calculation is, in fact, the main motivation for the paper.

In Section 1, we express the degree-d spin Hurwitz numbers (0-4) in terms of relative GW moduli spaces. We can then apply a degeneration method for a family of curves $\mathcal{D} \to \Delta$ where the central fiber D_0 is a nodal curve and the general fiber

D_λ ($\lambda \neq 0$) is a smooth curve. Section 2 describes the relative moduli space $\mathcal{M}_0$ of maps f into the nodal curve D_0. In Section 3, we show that the union over $\lambda \in \Delta$ of relative moduli spaces $\mathcal{M}_\lambda$ of maps into D_λ consists of connected components $\mathcal{Z}_{m,f} \to \Delta$ containing $f \in \mathcal{M}_0$. Here m is the ramification data of f over nodes of D_0 such that $d - \ell(m)$ is even.

The (ordinary) Hurwitz numbers are sums of (ramified) maps modulo automorphism without sign. One can easily obtain a recursion formula for Hurwitz numbers by counting maps in the general fiber of $\mathcal{Z}_{m,f} \to \Delta$. For spin Hurwitz numbers, one needs to calculate parities of maps induced from a fixed spin structure on the family of curves $\mathcal{D}$.

The novelty of our approach is to apply a Schiffer variation for the parity calculation. The space $\mathcal{Z}_{m,f}$ is, in general, not smooth. In Section 4, we construct a smooth model for $\mathcal{Z}_{m,f}$ by Schiffer variation. In Section 5, we use the smooth model to twist the pullback of the spin structure on $\mathcal{D}$. When the degree d equals 3, the partition m is odd, either (1^3) or (3). In this case, a suitable twisting immediately yields a required parity calculation. We prove Theorem 0.1 in Section 6 and formula (0-7) in Section 7.

For higher degree $d \geq 4$, the partition m may not be odd! A new parity calculation is needed. In [Lee and Parker 2012], we generalized the recursion formula (0-6) for higher-degree spin Hurwitz numbers by employing additional geometric analysis arguments for parity calculations.

1. Dimension zero relative GW moduli spaces

In this section, we express the spin Hurwitz numbers (0-4) in terms of dimension-zero relative GW moduli spaces. We follow the definitions of [Ionel and Parker 2003] for the relative GW theory.

Let D be a smooth curve of genus h and let $V = \{q^1, \ldots, q^k\}$ be a fixed set of points on D. Given partitions $m^1, \ldots, m^k$ of d, a degree-d holomorphic map $f : C \to D$ from a possibly disconnected curve C is called *V-regular* with contact vectors $m^1, \ldots, m^k$ if $f^{-1}(V)$ consists of $\sum \ell(m^i)$ contact marked points x^i_j ($1 \leq j \leq \ell(m^i)$) with $f(x^i_j) = q^i$ such that f has ramification index (or multiplicity) m^i_j at x^i_j. Two V-regular maps $(f, C; \{x^i_j\})$ and $(\tilde{f}, \widetilde{C}; \{\tilde{x}^i_j\})$ are equivalent if they are isomorphic, that is, there is a biholomorphism $\sigma : C \to \widetilde{C}$ with $\tilde{f} \circ \sigma = f$ and $\sigma(x^i_j) = \tilde{x}^i_j$ for all i, j. The relative moduli space

$$\mathcal{M}^V_{\chi, m^1, \ldots, m^k}(D, d) \tag{1-1}$$

consists of equivalence classes of V-regular maps $(f, C; \{x^i_j\})$ with the Euler characteristic $\chi(C) = \chi$ and with contact vectors $m^1, \ldots, m^k$. Since no confusion can

arise, we regard a point in the space (1-1) as a V-regular map $(f, C; \{x_j^i\})$. For simplicity, we often write a V-regular map $(f, C; \{x_j^i\})$ simply as f.

The (formal) complex dimension of the space (1-1) is given by the left side of the Riemann–Hurwitz formula (0-1):

$$2d(1-h) - \chi - \sum_{i=1}^{k} (d - \ell(m^i)). \tag{1-2}$$

Suppose this dimension is zero. Then, for each V-regular map $(f, C; \{x_j^i\})$ in (1-1), forgetting the contact marked points x_j^i gives a (ramified) cover f that is ramified only over fixed points q^i and satisfies (0-1). The automorphism group $\mathrm{Aut}(f)$ of a (ramified) cover f consists of automorphisms $\sigma \in \mathrm{Aut}(C)$ with $f \circ \sigma = f$. The automorphism group $\mathrm{Aut}(f, V)$ of a V-regular map $(f, C; \{x_j^i\})$ consists of automorphisms $\sigma \in \mathrm{Aut}(f)$ with $\sigma(x_j^i) = x_j^i$ for all i, j.

For a partition m of d, let $\mathrm{Aut}(m)$ be the subgroup of symmetric group $S_{\ell(m)}$ permuting equal parts of the partition m.

Lemma 1.1. *Let $m^1, \ldots, m^k$ be as above and suppose the dimension* (1-2) *is zero.*

(a) *If $m^i = (1^d)$ for some $1 \le i \le k$, $\mathrm{Aut}(f, V)$ is trivial for all f in* (1-1).

(b) *If $m^1, \ldots, m^k$ are all odd partitions,*

$$H^{h,p}_{m^1,\ldots,m^k} = \frac{1}{\prod_{i=1}^k |\mathrm{Aut}(m^i)|} \sum \frac{(-1)^{p(f)}}{|\mathrm{Aut}(f, V)|}$$

where the sum is over all f in (1-1) *and $p(f)$ is the parity* (0-3).

Proof. Let $(f, C; \{x_j^i\})$ be a V-regular map in (1-1) and $\sigma \in \mathrm{Aut}(f, V)$. If $m^i = (1^d)$, the set of branch points B of f is a subset of $V \setminus \{q^i\}$ and the restriction of σ to $C \setminus f^{-1}(B)$ is a covering transformation that fixes contact marked points $x_1^i, \ldots, x_d^i$. Noting $f^{-1}(B)$ is finite, we conclude that σ is an identity map on C. This proves (a).

As mentioned above, forgetting contact marked points x_j^i gives a (ramified) cover f satisfying (0-1). Conversely, given a (ramified) cover f satisfying (0-1), one can mark a point over q^i with ramification index m_j^i as a contact marked point x_j^i. Such marking gives V-regular maps $(f, C; \{x_j^i\})$ in $\prod_{i=1}^k |\mathrm{Aut}(m^i)|$ ways. Observe that $(f, C; \{x_j^i\})$ and $(f, C; \{\sigma(x_j^i)\})$ are isomorphic for each $\sigma \in \mathrm{Aut}(f)$ and that $\mathrm{Aut}(f, V)$ is a normal subgroup of $\mathrm{Aut}(f)$. Consequently, the quotient group $G = \mathrm{Aut}(f)/\mathrm{Aut}(f, V)$ acts freely on the set of V-regular maps $(f, C; \{x_j^i\})$ obtained by the (ramified) cover f. Its orbits give $\prod_{i=1}^k |\mathrm{Aut}(m^i)|/|G|$ points (that is, equivalence classes of V-regular maps) in the space (1-1), each of which has the same automorphism group $\mathrm{Aut}(f, V)$. Now (b) follows from counting maps with the parity of map modulo automorphisms. □

2. Maps into a nodal curve

Let $D_0 = D_1 \cup E \cup D_2$ be a connected nodal curve of (arithmetic) genus h with two nodes p^1 and p^2 such that, for $i = 1, 2$, $E = \mathbb{P}^1$ meets D_i at node p^i and D_i has genus h_i with $h_1 + h_2 = h$. In this section, we consider maps into D_0 that are relevant to our subsequent discussion.

Below, we fix d, h, χ, and odd partitions $m^1, \dots, m^k$ of d so that the Riemann–Hurwitz formula (0-1) holds, or equivalently, the dimension formula (1-2) is zero. For each partition m of d, consider the product space

$$\mathcal{P}_m = \mathcal{M}^{V_1}_{\chi_1,(1^d),m^1,\dots,m^{k_1},m}(D_1, d) \times \mathcal{M}^{V_0}_{\chi_0,m,(1^d),m}(E, d) \times \mathcal{M}^{V_2}_{\chi_2,m,m^{k_1+1},\dots,m^k,(1^d)}(D_2, d)$$

where

$$V_1 = \{q^{k+1}, q^1, \dots, q^{k_1}, p^1\}, \quad V_0 = \{p^1, q^{k+2}, p^2\}, \quad V_2 = \{p^2, q^{k_1+1}, \dots, q^k, q^{k+3}\}$$

and

$$\chi_1 + \chi_0 + \chi_2 - 4\ell(m) = \chi. \tag{2-1}$$

For simplicity, let $\mathcal{M}^1_m$, $\mathcal{M}^0_m$, and $\mathcal{M}^2_m$ denote the first, second, and third factors of $\mathcal{P}_m$.

Lemma 2.1. *If $\mathcal{P}_m \neq \varnothing$, the spaces $\mathcal{M}^1_m$, $\mathcal{M}^0_m$, and $\mathcal{M}^2_m$ have dimension zero. Consequently, $\chi_0 = 2\ell(m)$ and $d - \ell(m)$ is even.*

Proof. Each $\mathcal{M}^i_m$ $(0 \le i \le 2)$ has nonnegative dimension by the Riemann–Hurwitz formula. The formula (2-1) and our assumption that the dimension (1-2) is zero thus imply that each $\mathcal{M}^i_m$ has dimension zero. The dimension formulas for $\mathcal{M}^0_m$ and $\mathcal{M}^i_m$ $(i = 1, 2)$ then show that $\chi_0 = 2\ell(m)$ and $d - \ell(m)$ is even because $d - \ell(m^i) = \sum(m^i_j - 1)$ is even for all $1 \le i \le k$. □

Let $|A|$ denote the cardinality of a set A.

Lemma 2.2.
$$|\mathcal{M}^0_m| = \frac{d!\,|\mathrm{Aut}(m)|}{\prod m_j}.$$

Proof. Let $f \in \mathcal{M}^0_m$. Since $\chi_0 = 2\ell(m)$, the domain of f is a disjoint union of smooth rational curves E_j for $1 \le j \le \ell(m)$, and each restriction $f_j = f|_{E_j}$ has exactly one contact marked point over p^i $(i = 1, 2)$ with multiplicity m_j, so f_j has degree m_j.

Consequently, forgetting contact marked points of maps in $\mathcal{M}^0_m$ gives exactly one map (as a cover) with automorphism group of order $|\mathrm{Aut}(m)| \prod m_j$. Here the factor $|\mathrm{Aut}(m)|$ appears because we can relabel maps f_j in $|\mathrm{Aut}(m)|$ ways and the factor $\prod m_j$ appears because each restriction map f_j (as a cover) has an automorphism group of order m_j. We then argue as in the proof of Lemma 1.1. □

For each $(f_1, f_0, f_2) \in \mathcal{P}_m$, by identifying contact marked points over $p^i \in D^i \cap E$ $(i = 1, 2)$, one can glue the domains of f_i and f_0 to obtain a map $f : C \to D_0$ with $\chi(C) = \chi$. For notational convenience, we often write the glued map f as $f = (f_1, f_0, f_2)$. Denote by

$$\mathcal{M}_{m,0} \tag{2-2}$$

the space of such glued maps $f = (f_1, f_0, f_2)$. Contact marked points are labeled, but nodal points of C are not labeled. Thus, we have the following.

Lemma 2.3. *$\mathcal{P}_m$ is a cover of $\mathcal{M}_{m,0}$ of degree $|\mathrm{Aut}(m)|^2$.*

3. Limiting and gluing

Following [Ionel and Parker 2004], this section describes limiting and gluing arguments under a degeneration of target curves. Let $D_0 = D_1 \cup E \cup D_2$ be the nodal curve with fixed points $q^1, \ldots, q^{k+3}$ as in Section 2. In Section 4, we construct a family of curves together with $k+3$ sections:

$$\begin{array}{ccc} & & \mathcal{D} \\ & Q^i \nearrow & \downarrow \rho \\ & & \Delta \end{array} \tag{3-1}$$

Here the total space $\mathcal{D}$ is a smooth complex surface, $\Delta \subset \mathbb{C}$ is a disk with parameter λ, the central fiber is D_0, the general fiber D_λ $(\lambda \neq 0)$ is a smooth curve of genus h, and $Q^i(0) = q^i$ for $1 \le i \le k+3$. By Gromov's convergence theorem, a sequence of holomorphic maps into D_λ with $\lambda \to 0$ has a map into D_0 as a limit. For notational simplicity, for $\lambda \neq 0$ we set

$$\mathcal{M}_\lambda = \mathcal{M}^{V_\lambda}_{\chi, m^1, \ldots, m^{k+3}}(D_\lambda, d), \quad \text{where } V_\lambda = \{Q^1(\lambda), \ldots, Q^{k+3}(\lambda)\}, \tag{3-2}$$

and denote the set of limits of sequences of maps in $\mathcal{M}_\lambda$ as $\lambda \to 0$ by

$$\lim_{\lambda \to 0} \mathcal{M}_\lambda. \tag{3-3}$$

Lemma 3.1 shows that limit maps in (3-3) lie in the union of spaces (2-2), namely,

$$\lim_{\lambda \to 0} \mathcal{M}_\lambda \subset \bigcup_m \mathcal{M}_{m,0} \tag{3-4}$$

where the union is over all partitions m of d with $d - \ell(m)$ even.

Conversely, by the gluing theorem of [Ionel and Parker 2004], the domain of each map in $\mathcal{M}_{m,0}$ can be smoothed to produce maps in $\mathcal{M}_\lambda$ for small $|\lambda|$. Shrinking Δ if necessary, for $\lambda \in \Delta$, one can assign to each $f_\lambda \in \mathcal{M}_\lambda$ a partition m of d by (3-4). Let $\mathcal{M}_{m,\lambda}$ be the set of all pairs (f_λ, m). For each $f \in \mathcal{M}_{m,0}$, let

$$\mathcal{Z}_{m,f} \to \Delta \tag{3-5}$$

be the connected component of $\bigcup_{\lambda\in\Delta}\mathcal{M}_{m,\lambda}\to\Delta$ that contains f, and let

$$\mathfrak{L}_{m,f,\lambda} \tag{3-6}$$

denote the fiber of (3-5) over $\lambda\in\Delta$. It follows that, for $\lambda\neq 0$,

$$\mathcal{M}_\lambda = \bigsqcup_{f\in\mathcal{M}_{m,0}} \mathfrak{L}_{m,f,\lambda}. \tag{3-7}$$

For $f=(f_1, f_0, f_2)\in\mathcal{M}_{m,0}$ where $m=(m_1,\dots,m_\ell)$, let y_j^i be the node mapped to p^i at which f_i and f_0 have multiplicity m_j. The gluing theorem shows that one can smooth each node y_j^i in m_j ways to produce $\big(\prod m_j\big)^2$ maps in $\mathfrak{L}_{m,f,\lambda}$, so

$$|\mathfrak{L}_{m,f,\lambda}| = \big(\textstyle\prod m_j\big)^2 \quad (\lambda\neq 0). \tag{3-8}$$

In order to prove (3-4), we use the following fact on stable maps. An irreducible component of a stable holomorphic map f is a ghost component if its image is a point. Write the domain of f as $C^g\cup C$ where C^g is a connected curve whose irreducible components are all ghost components. Then the stability of f implies that

$$\chi(C^g)-\ell^g-n\le -1 \tag{3-9}$$

where $\ell^g=|C^g\cap C|$ and n is the number of marked points on C^g.

Lemma 3.1. *Let $\mathcal{M}_r$ and $\mathcal{M}_{m,0}$ be as above. Then we have*

$$\lim_{\lambda\to 0}\mathcal{M}_\lambda \subset \bigcup_m \mathcal{M}_{m,0}$$

where the union is over all partitions m of d with $d-\ell(m)$ even.

Proof. Let f be a limit map in (3-3). The domain C of f can be written as

$$C = C_1\cup C_0\cup C_2\cup\Big(\bigcup_{i=1}^{k+3} C_i^g\Big)\cup C^g\cup\widetilde{C}^g \tag{3-10}$$

where C_0 maps to E, C_1 and C_2 map to D_1 and D_2, C_i^g is the union of all ghost components over q^i, where $i=1,\dots,k+3$, C^g is the union of all ghost components over points in $D_0\setminus(V_1\cup V_0\cup V_2)$, and $\widetilde{C}^g$ is the union of all ghost components over $\{p^1,p^2\}$. Let $f_j=f|_{C_j}$ for $j=0,1,2$. Observe that f_j is V_j-regular because C_j has no ghost components. Let $\widehat{m}^i$ be a contact vector over q^i, $\widetilde{m}^1$ and $\widetilde{m}^2$ be contact vectors of f_1 and f_2 over p^1 and p^2, and $\widetilde{m}^{0;1}$ and $\widetilde{m}^{0;2}$ be contact vectors of f_0 over p^1 and p^2. The Riemann–Hurwitz formulas for f_0, f_1, and f_2 give

$$\sum_{j=0}^{2}\chi(C_j)\le 2d(1-h)+\sum_{i=1}^{k+3}(\ell(\widehat{m}^i)-d)+\sum_{i=1}^{2}(\ell(\widetilde{m}^i)+\ell(\widetilde{m}^{0;i})). \tag{3-11}$$

For $i = 1, \dots, k+3$, let $\ell_i = |C_1 \cup C_0 \cup C_2 \cap C_i^g|$ and let n_i be the number of marked points on C_i^g. Since all marked points are limits of marked points, we have

$$\ell(\widehat{m}^i) = \ell(m^i) - n_i + \ell_i. \tag{3-12}$$

For $j = 0, 1, 2$, let $\tilde{\ell}_j = |C_j \cap \widetilde{C}^g|$. Counting the number of nodes mapped to p^1 and p^2 shows

$$\sum_{i=1}^{2} (\ell(\widetilde{m}^i) - \tilde{\ell}_i) = \sum_{i=1}^{2} |C_i \cap C_0| = \sum_{i=1}^{2} \ell(\widetilde{m}^{0;i}) - \tilde{\ell}_0. \tag{3-13}$$

Let $\ell^g = |C_1 \cup C_0 \cup C_2 \cap C^g|$. Since $\chi(C) = \chi$, by (3-10) and (3-13) we have

$$\begin{aligned} \chi = \sum_{j=0}^{2} \chi(C_j) + \sum_{i=1}^{k+3} (\chi(C_i^g) - 2\ell_i) + \chi(C^g) - 2\ell^g \\ + \chi(\widetilde{C}^g) - \tilde{\ell} - \sum_{i=1}^{2} (\ell(\widetilde{m}^i) + \ell(\widetilde{m}^{0;i})), \end{aligned} \tag{3-14}$$

where $\tilde{\ell} = \tilde{\ell}_0 + \tilde{\ell}_1 + \tilde{\ell}_2$. By our assumption that formula (0-1) holds, it follows from (3-11), (3-12), and (3-14) that

$$\chi \le \chi + \sum_{i=1}^{k+3} (\chi(C_i^g) - \ell_i - n_i) + \chi(C^g) - 2\ell^g + \chi(\widetilde{C}^g) - \tilde{\ell}. \tag{3-15}$$

Noting that C^g and $\widetilde{C}^g$ have no marked points, by (3-9) and (3-15), we conclude that the domain C of f has no ghost components. Consequently,

- f_j is V_j-regular for $j = 0, 1, 2$,
- $\widetilde{m}^i = \widetilde{m}^{0;i}$ for $i = 1, 2$ (see Lemma 3.3 of [Ionel and Parker 2004]) and $\widehat{m}^i = m^i$ for $i = 1, \dots, k+3$.

In particular, the equality in (3-11) holds; otherwise we have a strict inequality in (3-15). So, we have $\chi(C_0) = \ell(\widetilde{m}^1) + \ell(\widetilde{m}^2)$. But $\chi(C_0) \le 2\min\{\ell(\widetilde{m}^1), \ell(\widetilde{m}^2)\}$. It follows that

- C_0 has $\ell(\widetilde{m}^1) = \ell(\widetilde{m}^2)$ connected components E_j with $\chi(E_j) = 2$ for all j,
- $\widetilde{m}_j^1 = \deg(f_0|_{E_j}) = \widetilde{m}_j^2$ for all j, that is, $\widetilde{m}^1 = \widetilde{m}^2$.

It follows that the Euler characteristics of C_0, C_1, and C_2 satisfy (2-1) by (3-14). Therefore, $f \in \mathcal{M}_{m,0}$ for $m = \widetilde{m}^1 = \widetilde{m}^2$ and $d - \ell(m)$ is even by Lemma 2.1. □

4. Smooth model by Schiffer variation

A *Schiffer variation* of a nodal curve (compare [Arbarello et al. 2011, p. 184]) is obtained by gluing deformations $uv = \lambda$ near nodes with the trivial deformation

away from nodes. In this section, we use the method of Schiffer variation to construct a smooth model for the space $\mathscr{Z}_{m,f}$ in (3-5), which has several branches intersecting at f unless m is trivial.

In this section, we fix an odd partition $m = (n^\ell)$, that is, $m = (m_1, \ldots, m_\ell)$ with

$$m_1 = \cdots = m_\ell = n, \quad \text{where } n = d/\ell \text{ is odd.} \tag{4-1}$$

Let $f = (f_1, f_0, f_2)$ be a map in $\mathcal{M}_{m,0}$ in (2-2). As described in Section 2, the central fiber of $\rho : \mathcal{D} \to \Delta$ is the nodal curve $D_0 = D_1 \cup E \cup D_2$ with two nodes $p^1 \in D_1 \cap E$ and $p^2 \in D_2 \cap E$ where $E = \mathbb{P}^1$. The domain of f is a nodal curve

$$C = C_1 \cup C_0 \cup C_2, \quad \text{where } C_0 = \bigcup_{j=1}^{\ell} E_\ell,$$

with 2ℓ nodes, such that, for $i = 1, 2$ and $j = 1, \ldots, \ell$,

- $f^{-1}(p^i)$ consists of the ℓ nodes $y^i_j \in C_i \cap E_j$,
- C_i is smooth and $f|_{C_i} = f_i$ has ramification index $m_j = n$ at the node y^i_j,
- $E_j = \mathbb{P}^1$ and $f|_{E_j} = f_0|_{E_j} : E_j \to E$ has ramification index $m_j = n$ at the node y^i_j.

The following is the main result of this section.

Proposition 4.1. *Let f be as above. Then, for each vector $\zeta = (\zeta^1_1, \zeta^2_1, \ldots, \zeta^1_\ell, \zeta^2_\ell)$, where ζ^i_j is an n^{th} root of unity, there are a family of curves $\varphi_\zeta : \mathscr{C}_\zeta \to \Delta$, with smooth total space $\mathscr{C}_\zeta$, over a disk Δ (with parameter s) and a holomorphic map $\mathscr{F}_\zeta : \mathscr{C}_\zeta \to \mathcal{D}$ satisfying*:

(a) *the central fiber $C_{\zeta,0} = C$ and the restriction map $\mathscr{F}_\zeta|_C = f$;*

(b) *the general fiber $C_{\zeta,s}$ ($s \neq 0$) is smooth and, for $\lambda = s^n \neq 0$,*

$$\bigcup_\zeta \{f_{\zeta,s}\} = \mathscr{Z}_{m,f,\lambda}, \tag{4-2}$$

where the union is over all ζ, $f_{\zeta,s} = \mathscr{F}_\zeta|_{C_{\zeta,s}}$ and $\mathscr{Z}_{m,f,\lambda}$ is the space (3-6).

Proof. The proof consists of four steps.

Step 1. We first show how to construct the family of curves $\rho : \mathcal{D} \to \Delta$ with $k+3$ sections. For $i = 1, 2$, a neighborhood of the node $p^i \in D_i \cap E$ can be regarded as the union $U^i \cup V^i$ of the two disks

$$U^i = \{u^i \in \mathbb{C} : |u^i| < 1\} \subset D_i \quad \text{and} \quad V^i = \{v^i \in \mathbb{C} : |v^i| < 1\} \subset E$$

with their origins identified. We may assume that the fixed points $q^1, \ldots, q^{k+3}$ in D_0 described in (2-1) lie outside these sets. Consider the regions

$$A^i = \{(u^i, v^i, \lambda) \in U^i \times V^i \times \Delta : u^i v^i = \lambda\},$$
$$B = \bigcup_{i=1}^{2} G^i \cup \Big[\Big(D_0 \setminus \bigcup_{i=1}^{2}(U^i \cup V^i)\Big) \times \Delta\Big],$$

where

$$G^i = \{(u^i, \lambda) \in U^i \times \Delta : |u^i| > \sqrt{|\lambda|}\} \cup \{(v^i, \lambda) \in V^i \times \Delta : |v^i| > \sqrt{|\lambda|}\}.$$

We obtain a smooth complex surface $\mathcal{D}$ by gluing A^1, A^2, and B_0 using the maps

$$G^i \to A^i \quad \text{defined by } (u^i, \lambda) \to \Big(u^i, \frac{\lambda}{u^i}, \lambda\Big) \text{ and } (v^i, \lambda) \to \Big(\frac{\lambda}{v^i}, v^i, \lambda\Big). \tag{4-3}$$

Let $\rho : \mathcal{D} \to \Delta$ be the projection to the last factor and define $k+3$ sections Q^i of ρ by

$$Q^i(\lambda) = (q^i, \lambda).$$

Step 2. We can similarly construct a family of curves over a 2ℓ-dimensional polydisk:

$$\varphi_{2\ell} : \mathcal{X} \to \Delta_{2\ell} = \{t = (t_1^1, t_1^2, \ldots, t_\ell^1, t_\ell^2) \in \mathbb{C}^{2\ell} : |t_j^i| < 1\}. \tag{4-4}$$

For each node $y_j^i \in C_i \cap E_j$, choose a neighborhood obtained from two disks

$$U_j^i = \{u_j^i \in \mathbb{C} : |u_j^i| < 1\} \subset C_i \quad \text{and} \quad V_j^i = \{v_j^i \in \mathbb{C} : |v_j^i| < 1\} \subset E_j$$

by identifying the origins. Consider the regions

$$A_j^i = \{(u_j^i, u_j^i, t) \in U_j^i \times V_j^i \times \Delta_{2\ell} : u_j^i v_j^i = t_j^i\},$$
$$B_{2\ell} = \bigcup_{i,j} G_j^i \cup \Big[\Big(C \setminus \bigcup_{i,j}(U_j^i \cup V_j^i)\Big) \times \Delta_{2\ell}\Big],$$

where

$$G_j^i = \{(u_j^i, t) \in U_j^i \times \Delta_{2\ell} : |u_j^i| > \sqrt{|t_j^i|}\} \cup \{(v_j^i, t) \in V_j^i \times \Delta_{2\ell} : |v_j^i| > \sqrt{|t_j^i|}\}.$$

We can then obtain a smooth complex manifold $\mathcal{X}$ of dimension $2\ell + 1$ by gluing $\bigcup A_j^i$ and $B_{2\ell}$ with the maps

$$G_j^i \to A_j^i \quad \text{defined by } (u_j^i, t) \to \Big(u_j^i, \frac{t_j^i}{u_j^i}, t\Big) \text{ and } (v_j^i, t) \to \Big(\frac{t_j^i}{v_j^i}, v_j^i, t\Big). \tag{4-5}$$

Let $\varphi_{2\ell} : \mathcal{X} \to \Delta$ be the projection to the factor t.

Step 3. Since f_i and $f_0|_{E_j}$ have ramification index $m_j = n$ at y_j^i, we may assume (after coordinates change) that on U_j^i and V_j^i the map f can be written as

$$U_j^i \to U^i \text{ by } u_j^i \to (u_j^i)^n \quad \text{and} \quad V_j^i \to V^i \text{ by } v_j^i \to (v_j^i)^n. \tag{4-6}$$

For each i, j, define a map

$$G^i_j \to G^i \quad \text{by } (u^i_j, t) \to ((u^i_j)^n, (t^i_j)^n) \text{ and } (u^i_j, t) \to ((v^i_j)^n, (t^i_j)^n). \tag{4-7}$$

On the other hand, for each i, j, we have a map

$$A^i_j \to A^i \quad \text{defined by } (u^i_j, v^i_j, t) \to ((u^i_j)^n, (v^i_j)^n, (t^i_j)^n). \tag{4-8}$$

These two maps (4-7) and (4-8) are glued together under the maps (4-3) and (4-5). The glued map extends to a holomorphic map $f_t : \mathscr{X}_t \to D_\lambda$ if and only if

$$(t^1_1)^n = (t^2_1)^n = \cdots = (t^1_\ell)^n = (t^2_\ell)^n = \lambda. \tag{4-9}$$

There are $n^{2\ell}$ solutions t of (4-9) and the extension map f_t is given by

$$(x, t) \to (f(x), \lambda) \quad \text{on } \mathscr{X}_t - \bigcup A^i_j.$$

Step 4. For each vector $\zeta = (\zeta^1_1, \zeta^2_1, \ldots, \zeta^1_\ell, \zeta^2_\ell)$, where each ζ^i_j is an n^{th} root of unity, define

$$\delta_\zeta : \Delta \to \Delta_{2\ell} \text{ by } s \to (\zeta^1_1 s, \zeta^2_1 s, \zeta^1_2 s, \zeta^2_2 s, \ldots, \zeta^1_\ell s, \zeta^2_\ell s).$$

The pullback $\delta^*_\zeta \mathscr{X}$ gives a family of curves:

$$\begin{array}{ccc} \mathscr{C}_\zeta = \delta^*_\zeta \mathscr{X} & \longrightarrow & \mathscr{X} \\ {\scriptstyle \varphi_\zeta}\downarrow & & \downarrow{\scriptstyle \varphi_{2\ell}} \\ \Delta & \xrightarrow{\delta_\zeta} & \Delta_{2\ell} \end{array} \tag{4-10}$$

The central fiber is $C_{\zeta,0} = C$ and the general fiber $C_{\zeta,s}$ ($s \neq 0$) is smooth. A neighborhood of the node y^i_j of C in $\mathscr{C}_\zeta$ can be viewed as

$$\hat{A}^i_j = \{(u^i_j, v^i_j, s) \in \mathbb{C}^3 : |u^i_j| < 1, |v^i_j| < 1, u^i_j v^i_j = \zeta^i_j s\}. \tag{4-11}$$

It follows that the total space $\mathscr{C}_\zeta$ is a complex smooth surface. Noting $\delta_\zeta(s)$ is a solution of (4-9) for $\lambda = s^n$, we obtain a holomorphic map $\mathscr{F}_\zeta : \mathscr{C}_\zeta \to \mathscr{D}$ given by

$$\begin{aligned} (u^i_j, v^i_j, s) &\to ((u^i_j)^n, (v^i_j)^n, s^n) \quad \text{on } \hat{A}^i_j, \\ (x, s) &\to (f(x), s^n) \qquad\qquad \text{on } \mathscr{C}_\zeta - \bigcup \hat{A}^i_j. \end{aligned} \tag{4-12}$$

Since the restriction $\mathscr{F}_\zeta|_C = f$ by (4-6) and (4-12), it remains to show (4-2). By our choice of fixed points q^i on D_0, each contact marked point x^i_j of f lies in $\mathscr{C}_\zeta - \bigcup \hat{A}^i_j$. Thus, by (4-12), the pullback $\mathscr{F}^*_\zeta Q^i$ of the section Q^i of ρ gives a section X^i_j of φ_ζ given by $X^i_j(s) = (x^i_j, s)$. After marking the points $X^i_j(s)$ in $C_{\zeta,s}$, the restriction map

$$f_{\zeta,s} = \mathscr{F}_\zeta|_{C_{\zeta,s}} : C_{\zeta,s} \to D_\lambda, \quad \text{where } \lambda = s^n \neq 0,$$

has contact marked points $X^i_j(s)$ over $Q^i(\lambda)$ with multiplicity m^i_j. This means that $f_{\zeta,s}$ lies in the space $\mathcal{M}_\lambda$ in (3-2) for $\lambda = s^n$. Therefore, noting that $f_{\zeta,s} \to f$ as $s \to 0$ and that $|\mathcal{Z}_{m,f,\lambda}| = n^{2\ell}$ by (3-8), we conclude (4-2). □

5. Spin structure and parity

The aim of this section is to use a spin structure on a family of nodal curves [Cornalba 1989] to show the parity calculation in Proposition 5.4. Twisting a bundle as in (5-6) is a key idea for parity calculation.

We first introduce a spin structure on a family of nodal curves that is relevant to our discussion. We refer to [Cornalba 1989] for the definition of spin structure and more details. The relative dualizing sheaf ω_ρ of the family of curves $\rho : \mathcal{D} \to \Delta$ in (3-1) is the canonical bundle $K_{\mathcal{D}}$ on the total space $\mathcal{D}$, since $\mathcal{D}$ is smooth and K_Δ is trivial. For each $\lambda \neq 0$, the restriction $K_{\mathcal{D}}|_{D_\lambda}$ is the canonical bundle K_{D_λ} on D_λ, and the restriction $K_{\mathcal{D}}|_{D_0}$ is the dualizing sheaf ω_{D_0} of the nodal curve $D_0 = D_1 \cup E \cup D_2$. As described in Section 4, D_0 is locally given by $u^i v^i = 0$ near each node p^i in $D_i \cap E$ for $i = 1, 2$. Then the local generators of ω_{D_0} are du^i/u^i and dv^i/v^i with a relation $du^i/u^i + dv^i/v^i = 0$; see [Harris and Morrison 1998, p. 82]. This implies the restriction $\omega_{D_0}|_{D_i} = K_{D_i} \otimes \mathbb{O}(p^i)$. On the other hand, $1/u^i$ is a local defining function for the divisor $-E$ on $\mathcal{D}$ near p^i. By restricting $1/u^i$ to D_i, one can see that $\mathbb{O}(-E)|_{D_i} = \mathbb{O}(-p^i)$. Consequently, for $i = 1, 2$,

$$K_{\mathcal{D}}|_{D_i} \otimes \mathbb{O}(-E)|_{D_i} = \omega_{D_0}|_{D_i} \otimes \mathbb{O}(-p^i) = K_{D_i}. \tag{5-1}$$

From Cornalba's construction [1989, p. 570], there are a line bundle $\mathcal{N} \to \mathcal{D}$ and a homomorphism $\Phi : \mathcal{N}^2 \to \omega_\rho = K_{\mathcal{D}}$ satisfying the following.

- Φ vanishes identically on the exceptional component E and $\mathcal{N}|_E = \mathbb{O}_E(1)$.
- Since $\Phi|_E \equiv 0$, there is an induced homomorphism $\hat{\Phi} : \mathcal{N}^2 \to K_{\mathcal{D}} \otimes \mathbb{O}(-E)$ such that Φ is the composition of $\hat{\Phi}$ with tensoring with η:

$$\Phi : \mathcal{N}^2 \xrightarrow{\hat{\Phi}} K_{\mathcal{D}} \otimes \mathbb{O}(-E) \xrightarrow{\otimes \eta} K_{\mathcal{D}}, \tag{5-2}$$

  where η is a section of $\mathbb{O}(E)$ with zero divisor E. Then, for $i = 1, 2$, the restriction

$$\hat{\Phi}|_{D_i} : (\mathcal{N}|_{D_i})^2 \to K_{\mathcal{D}}|_{D_i} \otimes \mathbb{O}(-E)|_{D_i} = K_{D_i}$$

  is an isomorphism so that the restriction $N_i = \mathcal{N}|_{D_i}$ is a theta characteristic on D_i.
- For each $\lambda \neq 0$, the restriction $\Phi|_{D_\lambda} : (\mathcal{N}|_{D_\lambda})^2 \to K_{D_\lambda}$ is an isomorphism so that the restriction $N_\lambda = \mathcal{N}|_{D_\lambda}$ is a theta characteristic on D_λ.

The pair $(\mathcal{N}, \Phi)$ is a spin structure on $\rho : \mathcal{D} \to \Delta$ and the restriction $\mathcal{N}|_{D_0}$ is a theta characteristic on the nodal curve D_0.

Remark 5.1. Atiyah [1971] and Mumford [1971] showed that the parity of a theta characteristic on a smooth curve is a deformation invariant. Cornalba [1989, Page 580] used the homomorphism Φ to extend Mumford's proof to the case of spin structure on a family of nodal curves. Thus, if p_1, p_2, and p are the parities of N_1, N_2, and N_λ ($\lambda \neq 0$), we have

$$p \equiv p_1 + p_2 \pmod 2.$$

Let $\varphi_\zeta : \mathscr{C}_\zeta \to \Delta$ be the family of curves in Proposition 4.1. Recall that the central fiber of φ_ζ is $C = C_1 \cup C_0 \cup C_2$, where $C_0 = \bigsqcup_j E_j$ is a disjoint union of ℓ exceptional components E_j and $C_i \cap E_j = \{y_j^i\}$ for $i = 1, 2$ and $1 \le j \le \ell$. Similarly as for (5-1), by restricting local defining functions, we have

$$\text{(5-3)}\quad \mathbb{O}(\pm C_0)|_{C_i} = \mathbb{O}\Big(\pm \sum_j y_j^i\Big) \quad (i = 1, 2) \quad \text{and} \quad \mathbb{O}(\pm C_0)|_{C_{\zeta,s}} = \mathbb{O} \quad (s \neq 0).$$

Since any fiber of φ_ζ is a principal divisor on $\mathscr{C}_\zeta$, $\mathbb{O}(C) = \mathbb{O}$ and hence $\mathbb{O}(C_0) = \mathbb{O}(-C_1 - C_2)$. We also have

$$\text{(5-4)}\quad \mathbb{O}(\pm C_0)|_{E_j} = \mathbb{O}(\mp(C_1 + C_2))|_{E_j} = \mathbb{O}(\mp(y_j^1 + y_j^2)) = \mathbb{O}(\mp 2)(1 \le j \le \ell).$$

Let $f = (f_1, f_0, f_2)$ and $\mathscr{F}_\zeta : \mathscr{C}_\zeta \to \mathscr{D}$ be the maps in Proposition 4.1. The ramification divisor $R_{\mathscr{F}_\zeta}$ of $\mathscr{F}_\zeta$ has local defining functions given by the Jacobian of $\mathscr{F}_\zeta$, so (4-12) shows

$$\text{(5-5)}\qquad R_{\mathscr{F}_\zeta} = \mathbb{O}(X_\zeta + (n-1)C) = \mathbb{O}(X_\zeta),$$

where $X_\zeta = \sum_{i,j}(m_j^i - 1)X_j^i$ and X_j^i is the section of φ_ζ defined in (4-12). Note that

(i) the ramification divisor of $f_i = \mathscr{F}_\zeta|_{C_i}$ ($i = 1, 2$) is $R_{f_i} = X_\zeta|_{C_i} + \sum_j (n-1)y_j^i$;

(ii) the ramification divisor of $f_{\zeta,s} = \mathscr{F}_\zeta|_{C_{\zeta,s}}$ ($s \neq 0$) is $R_{f_{\zeta,s}} = X_\zeta|_{C_{\zeta,s}}$.

Now, noting n is odd, we twist the pullback bundle $\mathscr{F}_\zeta^*\mathcal{N}$ by setting

$$\text{(5-6)}\qquad \mathscr{L}_\zeta = \mathscr{F}_\zeta^*\mathcal{N} \otimes \mathbb{O}\Big(\tfrac{1}{2}X_\zeta + \frac{(n-1)}{2}C_0\Big).$$

The lemma below shows that the twisted line $\mathscr{L}_\zeta$ restricts to a theta characteristic on each fiber of φ_ζ, including the central fiber C.

Lemma 5.2. *Let $\mathscr{L}_\zeta$ be as above. Then*:

(a) $\mathscr{L}_\zeta|_{E_j} = \mathbb{O}(1)$ *for* $1 \le j \le \ell$.

(b) $\mathscr{L}_\zeta|_{C_1} = L_{f_1}$, $\mathscr{L}_\zeta|_{C_2} = L_{f_2}$ *and* $\mathscr{L}_\zeta|_{C_{\zeta,s}} = L_{f_{\zeta,s}}$ *for* $s \neq 0$, *where* L_{f_1}, L_{f_2}, $L_{f_{\zeta,s}}$ *are the theta characteristics on* C_1, C_2, $C_{\zeta,s}$ *defined by* (0-2).

Proof. Part (a) follows from (5-4) and the fact that each restriction map $\mathcal{F}_\zeta|_{E_j}$ has degree n. Part (b) follows from (5-3), (i), and (ii). □

Observe that the relative dualizing sheaf ω_{φ_ζ} is the canonical bundle $K_{\mathscr{C}_\zeta}$ since $\mathscr{C}_\zeta$ is smooth. The Hurwitz formula and (5-5) thus imply that

$$\omega_{\varphi_\zeta} = K_{\mathscr{C}_\zeta} = \mathcal{F}_\zeta^* K_{\mathscr{D}} \otimes \mathbb{O}(X_\zeta). \tag{5-7}$$

Define a homomorphism

$$\begin{aligned}\hat{\Psi}_\zeta : \mathcal{L}_\zeta^2 = \mathcal{F}_\zeta^* \mathcal{N}^2 \otimes \mathbb{O}(X_\zeta + (n-1)C_0) \\ \to \mathcal{F}_\zeta^*(K_{\mathscr{D}} \otimes \mathbb{O}(-E)) \otimes \mathbb{O}(X_\zeta + (n-1)C_0)\end{aligned} \tag{5-8}$$

by $\hat{\Psi}_\zeta = \mathcal{F}_\zeta^* \hat{\Phi} \otimes \mathrm{Id}$, where $\hat{\Phi}$ is the induced homomorphism in (5-2). Noting that $\mathbb{O}(C) = \mathbb{O}$ and $\mathbb{O}(D_0) = \mathbb{O}$, by (4-12), we have

$$\mathcal{F}_\zeta^* \mathbb{O}(-E) = \mathcal{F}_\zeta^* \mathbb{O}(D_1 + D_2) = \mathbb{O}(n(C_1 + C_2)) = \mathbb{O}(-nC_0).$$

Together with (5-7), this implies that the right side of (5-8) is $K_{\mathscr{C}_\zeta} \otimes \mathbb{O}(-C_0)$. Now define a homomorphism $\Psi_\zeta : \mathcal{L}_\zeta^2 \to K_{\mathscr{C}_\zeta}$ to be the composition

$$\Psi_\zeta : \mathcal{L}_\zeta^2 \xrightarrow{\hat{\Psi}_\zeta} K_{\mathscr{C}_\zeta} \otimes \mathbb{O}(-C_0) \xrightarrow{\otimes \xi} K_{\mathscr{C}_\zeta}, \tag{5-9}$$

where ξ is a section of $\mathbb{O}(C_0)$ with zero divisor C_0.

Lemma 5.3. *$(\mathcal{L}_\zeta, \Psi_\zeta)$ is a spin structure on $\varphi_\zeta : \mathscr{C}_\zeta \to \Delta$.*

Proof. First, $\mathcal{L}_\zeta|_E = \mathbb{O}(1)$ by Lemma 5.2(a) and Ψ_ζ vanishes identically on each exceptional component E_j, since $\xi = 0$ on $C_0 = \bigsqcup_j E_j$. Second, since $\hat{\Phi}|_{D_i}$ is an isomorphism, (5-3) and (i) show that, for $i = 1, 2$, the restriction

$$\hat{\Psi}|_{C_i} = f_i^*(\hat{\Phi}|_{D_i}) \otimes \mathrm{Id} : (\mathcal{L}_\zeta|_{C_i})^2 = f_i^* N_i^2 \otimes \mathbb{O}(R_{f_i}) \to f_i^* K_{D_i} \otimes \mathbb{O}(R_{f_i}) = K_{C_i}$$

is an isomorphism. Lastly, let $\lambda = s^n \neq 0$. Since $\Phi|_{D_\lambda}$ is an isomorphism, so is $\hat{\Phi}|_{D_\lambda}$. Thus, by (5-3), (ii), and the facts $K_{\mathscr{D}}|_{D_\lambda} = K_{D_\lambda}$ and $\mathbb{O}(-E)|_{D_\lambda} = \mathbb{O}$, the restriction

$$\hat{\Psi}_\zeta|_{C_{\zeta,s}} = f_{\zeta,s}^* \hat{\Phi}|_{D_\lambda} \otimes \mathrm{Id} : (\mathcal{L}_\zeta|_{C_{\zeta,s}})^2 = f_{\zeta,s}^* N_\lambda^2 \otimes \mathbb{O}(R_{f_{\zeta,s}}) \to f_{\zeta,s}^* K_{D_\lambda} \otimes \mathbb{O}(R_{f_{\zeta,s}}) = K_{C_{\zeta,s}}$$

is an isomorphism. This implies that the restriction

$$\Psi_\zeta|_{C_{\zeta,s}} : (\mathcal{L}_\zeta|_{C_{\zeta,s}})^2 \to K_{C_\zeta}|_{C_{\zeta,s}} = K_{C_\zeta, s}$$

is also an isomorphism. Therefore, we conclude that $(\mathcal{L}_\zeta, \Psi_\zeta)$ is a spin structure on φ_ζ. □

The following is a key fact for the proof of Theorem 0.1.

Proposition 5.4. *Let $f = (f_1, f_0, f_2)$ and $f_{\zeta,s}$ be maps in Proposition 4.1. Then, for all $s \neq 0$,*

$$p(f_{\zeta,s}) \equiv p(f_1) + p(f_2) \pmod 2. \tag{5-10}$$

Proof. Since $(\mathscr{L}_\zeta, \Psi_\zeta)$ is a spin structure on φ_ζ, Cornalba's proof, mentioned in Remark 5.1, shows that, for all $s \neq 0$,

$$h^0(\mathscr{L}_\zeta|_{C_{\zeta,s}}) \equiv h^0(\mathscr{L}_\zeta|_{C_1}) + h^0(\mathscr{L}_\zeta|_{C_2}) \pmod 2.$$

This and Lemma 5.2(b) prove (5-10). □

6. Proof of Theorem 0.1

Proof. Fix a spin structure $(\mathcal{N}, \Phi)$ on $\rho : \mathscr{D} \to \Delta$ given in Section 5. Consider the space $\mathcal{M}_{m,0}$ in (2-2) where m is a partition of $d = 3$. In this case, by Lemma 2.1, either $m = (1^3)$ or $m = (3)$. Note that both of them satisfy (4-1). Fix $\lambda \neq 0$ and let $f = (f_1, f_0, f_2)$ be a map in $\mathcal{M}_{m,0}$. Then (4-2) and (5-10) show that, for all $f_\mu \in \mathscr{Z}_{m,f,\lambda}$,

$$p(f_\mu) \equiv p(f_1) + p(f_2) \pmod 2. \tag{6-1}$$

Lemma 1.1 and (3-7) show that

$$\begin{aligned} H^{h,p}_{(3)^k} &= H^{h,p}_{(3)^k,(1^3)^3} \\ &= \frac{1}{(3!)^3}\Bigg(\sum_{f\in\mathcal{M}_{(1^3),0}}\ \sum_{f_\mu\in\mathscr{Z}_{(1^3),f,\lambda}} (-1)^{p(f_\mu)} + \sum_{f\in\mathcal{M}_{(3),0}}\ \sum_{f_\mu\in\mathscr{Z}_{(3),f,\lambda}} (-1)^{p(f_\mu)}\Bigg). \end{aligned} \tag{6-2}$$

By (3-8) and (6-1), (6-2) becomes

$$H^{h,p}_{(3)^k} = \sum_{f=(f_1,f_0,f_2)\in\mathcal{M}_{(1^3),0}} \frac{(-1)^{p(f_1)+p(f_2)}}{(3!)^3} + \sum_{f=(f_1,f_0,f_2)\in\mathcal{M}_{(3),0}} \frac{3^2(-1)^{p(f_1)+p(f_2)}}{(3!)^3}. \tag{6-3}$$

It then follows from Lemma 2.3 and (6-3) that

$$\begin{aligned} H^{h,p}_{(3)^k} &= \sum_{(f_1,f_0,f_2)\in\mathscr{P}_{(1^3)}} \frac{(-1)^{p(f_1)+p(f_2)}}{(3!)^5} + \sum_{(f_1,f_0,f_2)\in\mathscr{P}_{(3)}} \frac{3^2(-1)^{p(f_1)+p(f_2)}}{(3!)^3} \\ &= \frac{1}{(3!)^3}\sum_{f_1\in\mathcal{M}^1_{(1^3)}}(-1)^{p(f_1)}\sum_{f_2\in\mathcal{M}^2_{(1^3)}}(-1)^{p(f_2)} + \frac{3}{(3!)^2}\sum_{f_1\in\mathcal{M}^1_{(3)}}(-1)^{p(f_1)}\sum_{f_2\in\mathcal{M}^2_{(3)}}(-1)^{p(f_2)} \\ &= 3!H^{h_1,p_1}_{(3)^{k_1}}\cdot H^{h_2,p_2}_{(3)^{k_2}} + 3H^{h_1,p_1}_{(3)^{k_1+1}}\cdot H^{h_2,p_2}_{(3)^{k_2+1}}; \end{aligned}$$

the second equality follows from Lemma 2.2 and the last from Lemma 1.1. □

7. Calculation

Proposition 7.1. $H^{h,\pm}_{(3)^k} = 3^{2h-2}[(-1)^k 2^{k+h-1} \pm 1]$.

Proof. The proof consists of four steps.

Step 1. We first show the following facts which we use in the computation below.

Lemma 7.2. (a) $H^{0,+}_{(3)^0} = H^{0,+}_3 = \frac{1}{3!}$, (b) $H^{0,+}_{(3)^3} = -\frac{1}{3}$, (c) $H^{1,+}_{(3)^0} = H^{1,+}_3 = 2$.

Proof. Consider the dimension-zero space $\mathcal{M}^V_\chi(\mathbb{P}^1, 3)$ where $V = \varnothing$. The Euler characteristic $\chi = 6$ by (0-1), and hence the space contains only one map $f : C \to \mathbb{P}^1$ where C is a disjoint union of three rational curves and $|\mathrm{Aut}(f)| = 3!$. This shows (a). Let (f, C) be a map in the dimension-zero space $\mathcal{M}^V_{\chi,(3),(3),(3)}(\mathbb{P}^1, 3)$. Then C is a connected curve of genus one and the theta characteristic L_f on C defined by (0-2) is

$$L_f = \mathbb{O}(-2x_1 + x_2 + x_3) = \mathbb{O}(x_1 - 2x_2 + x_3) = \mathbb{O}(x_1 + x_2 - 2x_3),$$

where x_1, x_2, and x_3 are ramification points of f. This implies $L_f^3 = \mathbb{O}$, and hence $L_f = \mathbb{O}$ because $L_f^2 = L_f^3 = \mathbb{O}$. We have $p(f) = 1$. Therefore,

$$H^{0,+}_{(3)^3} = -H^0_{(3)^3} = -\tfrac{1}{3},$$

where $H^0_{(3)^3}$ denotes the (ordinary) Hurwitz number, which is calculated by using the character formula; see [Okounkov and Pandharipande 2006, (0.10)]. By Proposition 9.2 of [Lee and Parker 2007], the spin Hurwitz numbers $H^{h,p}_d$ are the dimension-zero local invariants of spin curve that count maps from possibly disconnected domains. Let $GW^{h,p}_d$ denote the dimension-zero local invariants of spin curve that count maps from connected domains. Then $H^{h,p}_d$ and $GW^{h,p}_d$ are related as follows:

$$1 + \sum_{d>0} H^{h,p}_d t^d = \exp\Big(\sum_{d>0} GW^{h,p}_d t^d \Big).$$

Now (c) follows from $GW^{1,+}_1 = 1$, $GW^{1,+}_2 = 1/2$, and $GW^{1,+}_3 = 4/3$; see Section 10 of [Lee and Parker 2007]. □

Step 2. In this step, we compute $H^{1,-}_{(3)^k}$. For a spin curve of genus one with trivial theta characteristic. It follows from formula (3.12) of [Eskin et al. 2008] that

$$H^{1,-}_{(3)^k} = 2^{-k}[(\boldsymbol{f}_{(3)}(21))^k - (\boldsymbol{f}_{(3)}(3))^k]. \tag{7-1}$$

Here the *central character* $\boldsymbol{f}_{(3)}$ can be written as

$$\boldsymbol{f}_{(3)} = \tfrac{1}{3}\boldsymbol{p}_3 + a_2\boldsymbol{p}_1^2 + a_1\boldsymbol{p}_1 + a_0$$

for some $a_i \in \mathbb{Q}$ $(0 \le i \le 2)$, and the *supersymmetric functions* $\boldsymbol{p}_1$ and $\boldsymbol{p}_3$ are defined

by

$$\boldsymbol{p}_1(m) = d - \tfrac{1}{24} \quad \text{and} \quad \boldsymbol{p}_3(m) = \sum_j m_j^3 - \tfrac{1}{240},$$

where $m = (m_1, \ldots, m_\ell)$ is a partition of d. For $k = 0, 1$, (7-1) shows

$$H^{1,-}_{(3)^0} = 0 \quad \text{and} \quad H^{1,-}_{(3)} = -3. \tag{7-2}$$

Lemma 7.2(b), (7-2), and formula (0-6) give $H^{1,-}_{(3)^2} = 3H^{1,-}_{(3)} \cdot H^{0,+}_{(3)^3} = 3$. Together with (7-1) and (7-2), this yields $\boldsymbol{f}_{(3)}(21) = -4$ and $\boldsymbol{f}_{(3)}(3) = 2$. From this and (7-1) we have, for $k \geq 0$,

$$H^{1,-}_{(3)^k} = (-1)^k 2^k - 1. \tag{7-3}$$

Step 3. In this step, we compute $H^{h,+}_{(3)^k}$ for $h = 0, 1$. For $k \geq 1$, (7-2) and formula (0-6) give $H^{1,-}_{(3)^{k-1}} = 3H^{1,-}_{(3)} \cdot H^{0,+}_{(3)^k} = -3^2 H^{0,+}_{(3)^k}$. Combining this with Lemma 7.2(a) we obtain, for $k \geq 0$,

$$H^{0,+}_{(3)^k} = -\frac{1}{3^2}((-1)^{k-1}2^{k-1} - 1). \tag{7-4}$$

Lemma 7.2(c), (7-3), (7-4), and formula (0-6) show

$$\begin{aligned}
H^{2,+}_{(3)^0} &= 3!H^{1,-}_{(3)^0} \cdot H^{1,-}_{(3)^0} + 3H^{1,-}_{(3)} \cdot H^{1,-}_{(3)} = 27,\\
H^{2,+}_{(3)} &= 3!H^{1,-}_{(3)^0} \cdot H^{1,-}_{(3)} + 3H^{1,-}_{(3)} \cdot H^{1,-}_{(3)^2} = -27,\\
H^{2,+}_{(3)^0} &= 3!H^{1,+}_{(3)^0} \cdot H^{1,+}_{(3)^0} + 3H^{1,+}_{(3)} \cdot H^{1,+}_{(3)} = 24 + 3H^{1,+}_{(3)} \cdot H^{1,+}_{(3)},\\
H^{2,+}_{(3)} &= 3!H^{1,+}_{(3)^0} \cdot H^{1,+}_{(3)} + 3H^{1,+}_{(3)} \cdot H^{1,+}_{(3)^2} = 12H^{1,+}_{(3)} + 3H^{1,+}_{(3)} \cdot H^{1,+}_{(3)^2},\\
H^{1,+}_{(3)^2} &= 3!H^{1,+}_{(3)^0} \cdot H^{0,+}_{(3)^2} + 3H^{1,+}_{(3)} \cdot H^{0,+}_{(3)^3} = 4 - H^{1,+}_{(3)}.
\end{aligned}$$

It follows that $H^{1,+}_{(3)} = -1$. Hence, Lemma 7.2(c), (7-4), and formula (0-6) give

$$H^{1,+}_{(3)^k} = 3!H^{1,+}_{(3)^0} \cdot H^{0,+}_{(3)^k} + 3H^{1,+}_{(3)} \cdot H^{0,+}_{(3)^{k+1}} = (-1)^k 2^k + 1. \tag{7-5}$$

Step 4. It remains to compute $H^{h,p}_{(3)^k}$ for $h \geq 2$. The formula (0-6) gives

$$H^{h,p}_{(3)^k} = 3!H^{h-1,p}_{(3)^0} \cdot H^{1,+}_{(3)^k} + 3H^{h-1,p}_{(3)} \cdot H^{1,+}_{(3)^{k+1}}.$$

From this, we can deduce that, for $h \geq 2$,

$$\begin{aligned}
\begin{pmatrix} H^{h,p}_{(3)^k} \\ H^{h,p}_{(3)^{k+1}} \end{pmatrix} &= \begin{pmatrix} 3!H^{1,+}_{(3)^k} & 3H^{1,+}_{(3)^{k+1}} \\ 3!H^{1,+}_{(3)^{k+1}} & 3H^{1,+}_{(3)^{k+2}} \end{pmatrix} \begin{pmatrix} H^{h-1,p}_{(3)^0} \\ H^{h-1,p}_{(3)} \end{pmatrix} \\
&= \begin{pmatrix} 3!H^{1,+}_{(3)^k} & 3H^{1,+}_{(3)^{k+1}} \\ 3!H^{1,+}_{(3)^{k+1}} & 3H^{1,+}_{(3)^{k+2}} \end{pmatrix} \begin{pmatrix} 3!H^{1,+}_{(3)^0} & 3H^{1,+}_{(3)} \\ 3!H^{1,+}_{(3)} & 3H^{1,+}_{(3)^2} \end{pmatrix}^{h-2} \begin{pmatrix} H^{1,p}_{(3)^0} \\ H^{1,p}_{(3)} \end{pmatrix}.
\end{aligned} \tag{7-6}$$

Therefore, (7-3), (7-5), and (7-6) complete the proof. □

References

[Arbarello et al. 2011] E. Arbarello, M. Cornalba, and P. A. Griffiths, *Geometry of algebraic curves*, vol. II, Grundlehren der Mathematischen Wissenschaften **268**, Springer, Heidelberg, 2011. MR 2012e:14059 Zbl 1235.14002

[Atiyah 1971] M. F. Atiyah, "Riemann surfaces and spin structures", *Ann. Sci. École Norm. Sup.* (4) **4** (1971), 47–62. MR 44 #3350 Zbl 0212.56402

[Cornalba 1989] M. Cornalba, "Moduli of curves and theta-characteristics", pp. 560–589 in *Lectures on Riemann surfaces* (Trieste, 1987), edited by M. Cornalba et al., World Scientific, Teaneck, NJ, 1989. MR 91m:14037 Zbl 0800.14011

[Eskin et al. 2008] A. Eskin, A. Okounkov, and R. Pandharipande, "The theta characteristic of a branched covering", *Adv. Math.* **217**:3 (2008), 873–888. MR 2008k:14065 Zbl 1157.14014

[Gunningham 2012] S. Gunningham, "Spin Hurwitz numbers and topological quantum field theory", preprint, 2012. arXiv 1201.1273

[Harris and Morrison 1998] J. Harris and I. Morrison, *Moduli of curves*, Graduate Texts in Mathematics **187**, Springer, New York, 1998. MR 99g:14031 Zbl 0913.14005

[Ionel and Parker 2003] E.-N. Ionel and T. H. Parker, "Relative Gromov–Witten invariants", *Ann. of Math.* (2) **157**:1 (2003), 45–96. MR 2004a:53112 Zbl 1039.53101

[Ionel and Parker 2004] E.-N. Ionel and T. H. Parker, "The symplectic sum formula for Gromov–Witten invariants", *Ann. of Math.* (2) **159**:3 (2004), 935–1025. MR 2006b:53110 Zbl 1075.53092

[Kiem and Li 2007] Y.-H. Kiem and J. Li, "Gromov–Witten invariants of varieties with holomorphic 2-forms", preprint, 2007. arXiv 0707.2986

[Kiem and Li 2011] Y.-H. Kiem and J. Li, "Low degree GW invariants of spin surfaces", *Pure Appl. Math. Q.* **7**:4 (2011), 1449–1475. MR 2918169 Zbl 06107784

[Lee 2013] J. Lee, "Sum formulas for local Gromov–Witten invariants of spin curves", *Trans. Amer. Math. Soc.* **365**:1 (2013), 459–490. MR 2984064

[Lee and Parker 2007] J. Lee and T. H. Parker, "A structure theorem for the Gromov–Witten invariants of Kähler surfaces", *J. Differential Geom.* **77**:3 (2007), 483–513. MR 2010b:53159 Zbl 1130.53059

[Lee and Parker 2012] J. Lee and T. H. Parker, "Recursion formulas for spin Hurwitz numbers", preprint, 2012. arXiv 1212.1825

[Maulik and Pandharipande 2008] D. Maulik and R. Pandharipande, "New calculations in Gromov–Witten theory", *Pure Appl. Math. Q.* **4**:2 (2008), 469–500. MR 2009d:14073 Zbl 1156.14042

[Mumford 1971] D. Mumford, "Theta characteristics of an algebraic curve", *Ann. Sci. École Norm. Sup.* (4) **4** (1971), 181–192. MR 45 #1918 Zbl 0216.05904

[Okounkov and Pandharipande 2006] A. Okounkov and R. Pandharipande, "Gromov–Witten theory, Hurwitz theory, and completed cycles", *Ann. of Math.* (2) **163**:2 (2006), 517–560. MR 2007b:14123 Zbl 1105.14076

Received May 7, 2012. Revised September 1, 2012.

JUNHO LEE
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CENTRAL FLORIDA
ORLANDO, FL 32816
UNITED STATES
junlee@mail.ucf.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 263, No. 2, 2013
dx.doi.org/10.2140/pjm.2013.263.419

$(\mathbb{Z}_2)^3$-COLORINGS AND RIGHT-ANGLED HYPERBOLIC 3-MANIFOLDS

YOULIN LI AND JIMING MA

For a compact 3-manifold N with connected nonempty boundary, let Γ be an admissible trivalent graph in ∂N that decomposes ∂N into a set of disks. As an extension of small covers, from a $(\mathbb{Z}_2)^3$-coloring λ on $\partial N - \Gamma$, one can get a closed 3-manifold M_λ that admits a locally standard $(\mathbb{Z}_2)^3$-action.

Suppose N is irreducible and atoroidal: say, a handlebody. We give a combinatorial necessary and sufficient condition for a $(\mathbb{Z}_2)^3$-colorable pair (N, Γ) to admit a right-angled hyperbolic structure, which naturally induces a hyperbolic structure on M_λ.

1. Introduction

In this note, we study polyhedral hyperbolic 3-manifolds admitting $(\mathbb{Z}_2)^3$-colorings on their connected boundaries, which correspond to closed hyperbolic 3-manifolds admitting locally standard $(\mathbb{Z}_2)^3$-actions.

$(\mathbb{Z}_2)^3$-colorings and locally standard $(\mathbb{Z}_2)^3$-actions. Small covers, or Coxeter orbifolds, were studied in [Davis and Januszkiewicz 1991]. They are a class of manifolds which admit locally standard $(\mathbb{Z}_2)^n$-actions, such that the orbit spaces are n-dimensional simple polyhedra. The algebraic and topological properties of a small cover are closely related to the combinatorics of the orbit polyhedron and the coloring on its boundary. For example, the (mod 2) Betti number β_i of a small cover M agrees with h_i, where $h = (h_0, h_1, \ldots, h_n)$ is the h-vector of the polyhedron.

Those manifolds admitting locally standard $(\mathbb{Z}_2)^n$-actions form a wider class than small covers. In this paper, we focus on the 3-dimensional case.

The first author was partially supported by NSFC 11001171. The second author was partially supported by NSFC 10901038 and Shanghai NSF 10ZR1403600.
MSC2010: primary 57M50, 57M60; secondary 52B70.
Keywords: $(\mathbb{Z}_2)^3$-action, hyperbolic structure with polyhedral boundary, 3-manifold.

A standard representation of the $(\mathbb{Z}_2)^3$-action on $\mathbb{R}^3$ is the natural action defined by

$$e_1 : (x_1, x_2, x_3) \mapsto (-x_1, x_2, x_3), \tag{1-1}$$

$$e_2 : (x_1, x_2, x_3) \mapsto (x_1, -x_2, x_3), \tag{1-2}$$

$$e_3 : (x_1, x_2, x_3) \mapsto (x_1, x_2, -x_3). \tag{1-3}$$

The actions e_1, e_2 and e_3 generate the group $(\mathbb{Z}_2)^3$. This action fixes the origin of $\mathbb{R}^3$ such that its orbit space is the positive cone

$$\mathbb{R}^3_{\geq 0} = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_i \geq 0\}.$$

Definition 1.1. An effective $(\mathbb{Z}_2)^3$-action on a 3-dimensional closed manifold M is said to be *locally standard* if it locally looks like the standard representation of $(\mathbb{Z}_2)^3$-action on $\mathbb{R}^3$. More precisely, if for each point x in M, there is a $(\mathbb{Z}_2)^3$-invariant neighborhood U_x of x such that U_x is equivariantly homeomorphic to an invariant open subset of the standard $(\mathbb{Z}_2)^3$-action on $\mathbb{R}^3$.

The orbit space of a locally standard $(\mathbb{Z}_2)^3$-action on a 3-dimensional closed manifold M is a compact manifold N with corners. In other words, it is a 3-dimensional compact manifold N with a graph Γ on ∂N. The graph Γ on ∂N induces a cell decomposition on ∂N. The vertices of Γ are the image of fixed points of the $(\mathbb{Z}_2)^3$-action, the (open) edges of Γ are the image of fixed points of subgroups $(\mathbb{Z}_2)^2 < (\mathbb{Z}_2)^3$ and (open) components of $\partial N - \Gamma$ are the image of fixed points of subgroups $\mathbb{Z}_2 < (\mathbb{Z}_2)^3$.

Definition 1.2. Let N be a 3-dimensional manifold with nonempty boundary, and Γ a trivalent graph in ∂N that gives a cell decomposition of ∂N. A $(\mathbb{Z}_2)^3$-*coloring* is a map $\lambda : \partial N - \Gamma \to (\mathbb{Z}_2)^3 - 0$ such that $\lambda(f_1)$, $\lambda(f_2)$ and $\lambda(f_3)$ generate $(\mathbb{Z}_2)^3$ for each triple of faces f_1, f_2 and f_3 sharing a common vertex.

Associated to a locally standard $(\mathbb{Z}_2)^3$-action on M, there is a canonical $(\mathbb{Z}_2)^3$-coloring λ on $\partial N - \Gamma$ which colors each face $f \in \partial N - \Gamma$ by the element $e \in (\mathbb{Z}_2)^3$ that fixes f. For an i-dimensional cell f in the cell decomposition, $i = 0, 1, 2$, we have a group $G_f \cong (\mathbb{Z}_2)^{3-i}$ which is generated by the colorings in the faces which are adjacent to f. The locally standard $(\mathbb{Z}_2)^3$-action on M induces a principal $(\mathbb{Z}_2)^3$-bundle over N.

Conversely, by Lemma 3.1 of [Lü and Masuda 2009], from a $(\mathbb{Z}_2)^3$-coloring λ on $(\partial N, \Gamma)$ and a principal $(\mathbb{Z}_2)^3$-bundle over N, we can get a unique closed 3-manifold M. In particular, from a $(\mathbb{Z}_2)^3$-coloring λ and the trivial principal $(\mathbb{Z}_2)^3$-bundle over N, we get a 3-manifold M_λ which depends only on the coloring λ. By this we take eight copies of N, $N \times \{\alpha\}$ for each $\alpha \in (\mathbb{Z}_2)^3$, and construct a quotient

space M_λ under the following gluing rule:

$$(1\text{-}4)\quad (x,\alpha_1)\sim(y,\alpha_2)\Leftrightarrow\begin{cases}x=y,\ \alpha_1=\alpha_2, & \text{if } x \text{ lies in the interior of } N,\\ x=y,\ \alpha_1\alpha_2^{-1}\in G_f, & \text{if } x \text{ lies in a cell } f.\end{cases}$$

Then it is easy to see M_λ is a closed 3-manifold. In this paper, we only consider closed 3-manifolds associated to $(\mathbb{Z}_2)^3$-colorings and trivial principal $(\mathbb{Z}_2)^3$-bundles over N.

A simple example is that if we consider a coloring of the four faces of a tetrahedron by $e_1, e_2, e_3, e_1+e_2+e_3$, respectively, then from the above construction, we get the closed orientable 3-manifold RP^3. A tetrahedron admits a unique right-angled spherical structure, and the spherical structures on eight copies of the tetrahedron, when glued together, give rise to the unique spherical structure on RP^3. This point of view is applied in this paper.

There are many works on manifolds with locally standard $(\mathbb{Z}_2)^3$-actions. For example, 3-dimensional small covers are studied in [Lü 2009; Lü and Yu 2011]. Six operations on small covers were defined in [Lü and Yu 2011], which topologically behave well, such that every 3-dimensional small cover is obtained from the two simple small covers RP^3 and $S^1\times RP^2$ by a sequence of these operations. It should be noted that the operations in [Lü and Yu 2011] give many disks in a simple convex polygon P, which intersects the 1-skeleton of P in at most four points but which is not vertex-linking or edge-linking. So the preimage of these disks are essential spheres or essential tori in the small cover M in general, and hence M does not admit a geometric structure [Scott 1983].

Polyhedral hyperbolic 3-manifolds. Andreev [1971] (see also [Roeder et al. 2007]) gives a complete characterization of compact hyperbolic polyhedra in dimension 3 with nonobtuse angles. The boundary of a compact hyperbolic polyhedron inherits a natural cell decomposition. The 1-skeleton of the cell decomposition is a graph Γ on the boundary of the 3-ball, and a dihedral angle is also given on each edge of Γ from the hyperbolic structure. Andreev's theorem is given in terms of a set of conditions on the dihedral angles. Besides its beauty, Andreev's theorem is also essential in the proof of Thurston's geometrization theorem for Haken 3-manifolds. The natural question is, given a cell decomposition of the boundary of the 3-ball, and a weight $\alpha_e\in(0,\pi)$ attached to each edge e of the cell decomposition, whether there is a compact hyperbolic polyhedron in $\mathbb{H}^3$ realizing this cell decomposition whose dihedral angles coincide with the attached weights. This question is still open now.

A clever approach for working with compact hyperbolic polyhedra having arbitrary dihedral angles is to express necessary and sufficient conditions for existence of a given polyhedron in terms of its polar dual in the de Sitter space; see [Hodgson and Rivin 1993]. For a generalization of Andreev's result to ideal and hyper-ideal hyperbolic polyhedra, see [Rivin 1996; Bao and Bonahon 2002]. Hyperbolic

structures on topologically more complicated 3-manifold N with boundary are also studied. See [Schlenker 2002; 2003; 2005; 2006, Fillastre and Izmestiev 2009; 2011; Guéritaud 2009].

Suppose N is a compact 3-manifold with connected and nonempty boundary. In this note, we consider the right-angled hyperbolic structures on N with compressible boundary.

Given a graph Γ in ∂N, we call it *admissible* if the lift $\tilde{\Gamma}$ of Γ in the universal cover of N, say $\tilde{N}$, gives a cell decomposition of $\partial\tilde{N}$ such that each of its 2-cells has a closure homeomorphic to a disk, and each pair of such two disks shares at most one edge in $\tilde{\Gamma}$. A *right-angled hyperbolic realization* of (N, Γ) is a complete compact hyperbolic manifold N^* with right-angled polyhedral boundary (i.e., modeled on the orthogonal intersection of two half-spaces with totally geodesic boundaries in H^3 and having finite volume), endowed with a homeomorphism to N that sends the nonsmooth points of N^* precisely to the points of Γ. The nonsmooth points of N^* will be called the *singular locus* of this structure. From the homeomorphism between N^* and N, Γ is also called the *singular locus* for this hyperbolic realization. We will call such a structure on N^* a *hyperbolic structure with right-angled polyhedral boundary* on N. Hence these kinds of hyperbolic structures on N look locally like compact convex right-angled hyperbolic polyhedra in H^3.

Similar to all results above, it is interesting to give a kind of characterization of hyperbolic structures with right-angled polyhedral boundary on N. Since all the dihedral angles are right-angled, an easy argument shows that the graph Γ defined above must be trivalent.

It is well known that most 3-manifolds are hyperbolic 3-manifolds [Thurston 1982]. So it is interesting to consider locally standard $(\mathbb{Z}_2)^3$-actions on closed hyperbolic 3-manifolds. It is natural to ask which closed hyperbolic 3-manifold admits a locally standard $(\mathbb{Z}_2)^3$-action. The orbit space of a locally standard $(\mathbb{Z}_2)^3$-action is a compact manifold N with a coloring λ in ∂N. If (N, Γ) admits a right-angled hyperbolic structure, then it is easy to see that M is hyperbolic. A pair (N, Γ) admits a unique right-angled hyperbolic structure. However, it may admit many different colorings. Each coloring, together with a principle $(\mathbb{Z}_2)^3$-bundle over N, gives a manifold with a locally standard $(\mathbb{Z}_2)^3$-action. So these give many different hyperbolic manifolds of the same volume. [Inoue 2008] gives a very clear description of right-angled hyperbolic polyhedra from this point of view.

The most interesting case is that N is a handlebody, or the simplest one, a 3-ball.

Main result. Suppose Γ is an admissible graph in ∂N. For a vertex v of Γ, we take a small closed regular neighborhood B of v in N, then B intersects $N - \operatorname{int} B$ in a disk D_v. We call D_v a *vertex-linking disk*. It intersects Γ in three points. The preimage of a vertex-linking disk in M_λ is a sphere which bounds a 3-ball in M_λ for

a $(\mathbb{Z}_2)^3$-coloring λ. For an edge e of Γ, we also take a small closed neighborhood B of e in N; then B intersects $N - \operatorname{int} B$ in a disk D_e. We call D_e an *edge-linking disk*. It intersects Γ in four points. The preimage of an edge-linking disk in M_λ is a torus (or a Klein bottle) which bounds a solid torus (or a solid Klein bottle) in M_λ. We say a properly embedded disk D in the 3-manifold N intersects Γ *efficiently* if ∂D and Γ are in general position and there is no bigon in $\partial N - (\partial D \cup \Gamma)$. In this note we always assume a disk D intersects with Γ efficiently.

Our main result is the following:

Theorem 1.3. *Let N be an irreducible, atoroidal and compact 3-manifold with connected nonempty boundary, and Γ be an admissible trivalent graph in ∂N which gives a cell decomposition of ∂N, such that $(\partial N, \Gamma)$ admits a $(\mathbb{Z}_2)^3$-coloring. Then (N, Γ) realizes a right-angled hyperbolic structure if and only if every properly embedded disk D in N has $|D \cap \Gamma| \geq 5$, except when D is a vertex-linking disk or an edge-linking disk. Moreover, the realization is unique up to isometry.*

Remark 1.4. In practice, much attention has been paid to the right-angled hyperbolic structures on handlebodies. They are irreducible, atoroidal and compact 3-manifolds with connected nonempty boundaries. So Theorem 1.3 can be applied to the handlebody case.

Remark 1.5. There are two canonical ways to study polyhedral hyperbolic structure on 3-manifold M: Alexandrov's method and the variational method; see, for example, [Fillastre and Izmestiev 2011]. Our approach in this note uses the doubling trick. A $(\mathbb{Z}_2)^3$-coloring helps us find a closed 3-manifold on which we can apply the geometrization theorem.

2. Preliminaries

If (N, Γ) admits a right-angled hyperbolic structure, then Γ is admissible, and each of its 2-dimensional faces is a right-angled hyperbolic n-polygon. So $n \geq 5$.

Definition 2.1. Let Γ^* be the dual graph of Γ in ∂N. A *k-circuit* is a simple closed curve C in Γ^* consisting of k successive edges of Γ^* which is contractible in ∂N. A circuit is *elementary* if it bounds a disk D in ∂N and there is exactly one vertex V of Γ that lies in D. A k-circuit is *prismatic* if the endpoints of all the edges of Γ which intersect C are distinct.

Obviously, there is a one-to-one correspondence between edges of Γ and those of Γ^*.

Lemma 2.2. *Suppose Γ is admissible. If C is a 3-circuit which is not prismatic, then C is isotopic to the boundary of a vertex-linking disk. If Γ contains no prismatic 3-circuit, and C is a 4-circuit which is not prismatic, then C is isotopic to the boundary of an edge-linking disk.*

Proof. The proof is similar to the proofs of Lemmas 1.2 and 1.3 of [Roeder et al. 2007], by which we have that, if C is a nonprismatic 3-circuit, then it is an elementary circuit. So it is isotopic to the boundary of a vertex-linking disk. If Γ contains no prismatic 3-circuit, then every nonprismatic 4-circuit C separates off exactly two vertices of Γ from the remaining vertices of Γ, which in turn implies that C is isotopic to the boundary of an edge-linking disk. Actually, the authors of [Roeder et al. 2007] proved this for any graph in S^2. Since Γ is admissible, their arguments can be extended verbatim in the general case. □

We give a proposition on the orientability of a 3-manifold M_λ with a locally standard $(\mathbb{Z}_2)^3$-action and trivial principal $(\mathbb{Z}_2)^3$-bundle.

Proposition 2.3. *Suppose N is a compact orientable connected 3-manifold with connected boundary. Then, for a $(\mathbb{Z}_2)^3$-coloring λ on $(\partial N, \Gamma)$, M_λ is orientable if and only if there is a basis $\{e_1, e_2, e_3\}$ of $(\mathbb{Z}_2)^3$, such that the image of λ is contained in $\{e_1, e_2, e_3, e_1 + e_2 + e_3\}$.*

Proof. For small covers, this proposition has been proved in Theorem 1.7 of [Nakayama and Nishimura 2005]. Recall that M_λ is determined by the coloring λ and the trivial principal $(\mathbb{Z}_2)^3$-bundle over N. So M_λ is obtained by gluing eight copies of (N, λ), and M_λ is orientable if and only if $H_3(M_\lambda, \mathbb{Z}) = \mathbb{Z}$. To calculate $H_3(M_\lambda, \mathbb{Z})$, we only need to consider the 3-cells and 2-cells in a cellular decomposition of M_λ, which is induced by a cellular decomposition of (N, Γ). Note that ∂N is connected, so the arguments of the proof of Theorem 1.7 of [Nakayama and Nishimura 2005] hold in our case word-by-word. □

3. Proof of Theorem 1.3

Proof of the necessity part of Theorem 1.3. Suppose (N, Γ) realizes a right-angled hyperbolic structure. If $D \subset N$ is a properly embedded disk which intersects Γ efficiently, and is not vertex-linking or edge-linking, then by Gauss–Bonnet theorem, we have $|D \cap \Gamma| \geq 5$. So the necessity part of Theorem 1.3 follows. □

Proof of the sufficiency part of Theorem 1.3 in the case that M_λ is orientable. Recall that a closed orientable 3-manifold M is *irreducible* if every embedded 2-sphere S in M bounds a 3-ball; otherwise M is *reducible*. An embedded 2-sphere S which does not bound a 3-ball in M is called *essential*. A closed irreducible orientable 3-manifold M is *atoroidal* if every embedded torus T in M bounds a solid torus; otherwise M is *toroidal*. An embedded torus T which does not bound a solid torus is *essential* in M. See [Hempel 1976] or [Jaco 1980].

We need the equivariant sphere theorem of Meeks, Simon, and Yau [Meeks et al. 1982], but the reformulation by Dunwoody [1985] is more convenient for us.

Theorem 3.1. *Let G be a finite group that acts on a closed orientable 3-manifold M by homeomorphisms. Suppose M has a G-equivariant triangulation. If there exists an essential 2-sphere S in M, then there exists an essential 2-sphere S_1 in M which is in general position with respect to the triangulation, such that $g(S_1) = S_1$ or $g(S_1) \cap S_1 = \varnothing$ for every $g \in G$.*

We also need the following equivariant torus theorem; see [Freedman et al. 1983; Jaco and Shalen 1979; Johannson 1979].

Theorem 3.2. *Let G be a finite group which acts on a closed 3-manifold M by homeomorphisms. Suppose M is irreducible, orientable and contains an essential torus. Then either M is Seifert-fibered, or M contains a G-equivariant essential torus.*

First, we show the following lemmas.

Lemma 3.3. *M_λ is irreducible.*

Proof. We give a triangulation $\mathcal{T}$ of N, such that the graph Γ is contained in the 1-skeleton of $\mathcal{T}$. So the triangulation $\mathcal{T}$ induces a triangulation of M_λ.

If M_λ is reducible, then by the equivariant sphere theorem, there is a $(\mathbb{Z}_2)^3$-equivariant sphere S which is essential in M_λ. We denote $S \cap N \times \{1\}$ by A, which is a compact surface with nonempty boundary if $A \neq \varnothing$. We may assume $A \neq \varnothing$, otherwise we can use the $(\mathbb{Z}_2)^3$-action to find another $(\mathbb{Z}_2)^3$-invariant sphere S' which has nonempty intersection with $N \times \{1\}$. Since A is obtained from S by the $(\mathbb{Z}_2)^3$-action and S is connected, A is connected.

Since S is in general position with respect the triangulation of M_λ, A is in general position with respect to the triangulation of ∂N, in particular, with respect to Γ. So there is a cell decomposition of ∂A: for each face f of $\partial N - \Gamma$, $f \cap A$ is an edge in ∂A. Moreover, the coloring on $\partial N - \Gamma$ now induces a coloring on ∂A, which we denote by λ_A, and S is obtained from copies of A by the gluing rule from λ_A.

By Definition 1.2, the colorings on any two adjacent edges of ∂A are different. So we have a subgroup G of $(\mathbb{Z}_2)^3$ which has index 1 or 2 in $(\mathbb{Z}_2)^3$, such that for any $g \in G$ we have $g(S) = S$, and for any $h \in (\mathbb{Z}_2)^3 - G$ we have $h(S) \cap S = \varnothing$.

In other words, S is obtained by gluing 4 or 8 copies of A, and the edges in ∂A contribute a 4-valence graph in S. So $\chi(S) = m(\chi(A) - E/2 + E/4) = 2$, where E is the number of edges in ∂A, and $m = 4$ or 8. If $m = 4$, then $E = 2$ and $\chi(A) = 1$. So A is a disk with ∂A consisting of 2 edges. This is impossible by the assumption in Theorem 1.3. If $m = 8$, then $E = 3$ and $\chi(A) = 1$. So A is a disk with ∂A consists of 3 edges. Moreover, we have that $\partial A \cap f$ is connected for each face f. Suppose otherwise; i.e., suppose that $\partial A \cap f$ consists of at least two arcs. We have an edge e of Γ which intersects A such that the two sides of e both are in the face f. Then when we lift ∂N to the universal cover $\tilde{N}$. The closure of the lifting $\tilde{f}$ of the face f

is not a disk, contradicting the assumption that Γ is admissible. So ∂A is a 3-circuit. Thus, by the assumption and Lemma 2.2, A is a vertex-linking disk in N. The preimage of a vertex-linking disk in M_λ is a sphere which bounds a 3-ball in M_λ. This contradicts the assumption that S is essential in M_λ. So M_λ is irreducible. □

Lemma 3.4. *If M_λ is a toroidal Seifert manifold, then there is an essential torus in M_λ which is $(\mathbb{Z}_2)^3$-equivariant.*

Proof. Suppose e_1, e_2 and e_3 are three orientation-reversing involutions which generate the $(\mathbb{Z}_2)^3$-action. Since M_λ admits orientation-reversing involutions, according to Theorems 8.2 and 8.5 of [Neumann and Raymond 1978], M_λ is Seifert-fibered with Euler number 0; i.e., M_λ contains horizontal incompressible surfaces which are transversal to each fiber. In other words, M_λ is a surface bundle over S^1 with horizontal incompressible surfaces as surface fibers. We already proved in Lemma 3.3 that M_λ is irreducible, so the Euler characteristic of the base orbifold of M_λ is negative or zero. Thus M_λ admits the geometries $H^2 \times \mathbb{R}$ or E^3. We refer the readers to [Scott 1983] for the details about these two geometries.

For each $i = 1, 2, 3$, $\mathrm{Fix}(e_i)$ contains no nonorientable closed surfaces since the nonorientable closed surfaces are one-sided in M_λ. According to [Meeks and Scott 1986], for each $i = 1, 2, 3$, e_i is isotopic to an isometry. So $\mathrm{Fix}(e_i)$ consists of some totally geodesic, and hence incompressible, closed surfaces in M_λ.

If M_λ admits the $H^2 \times \mathbb{R}$ geometry, then it has unique Seifert fibration structure. So each homeomorphism sends regular fibers to regular fibers. Then, among $\mathrm{Fix}(e_1)$, $\mathrm{Fix}(e_2)$ and $\mathrm{Fix}(e_3)$, at least two of them, say $\mathrm{Fix}(e_1)$ and $\mathrm{Fix}(e_2)$, consist of vertical essential tori, and e_3 keeps each regular fiber invariant and reverses its orientation. By the definition of $(\mathbb{Z}_2)^3$-action, $\mathrm{Fix}(e_1) \cap \mathrm{Fix}(e_2) \neq \varnothing$, and $\mathrm{Fix}(e_1)$ intersects $\mathrm{Fix}(e_2)$ transversely. Choose a torus component T in $\mathrm{Fix}(e_1)$ which intersects $\mathrm{Fix}(e_2)$ nontrivially. For any point $p \in T$, we have $e_1e_2(p) = e_2e_1(p) = e_2(p)$. So $e_2(T) \subset \mathrm{Fix}(e_1)$. Thus we have that either $e_2(T) = T$ or $e_2(T) \cap T = \varnothing$. By the assumption $T \cap \mathrm{Fix}(e_2) \neq \varnothing$, we have $e_2(T) = T$. Moreover, $e_3(T) = T$. Hence T is invariant by e_1, e_2 and e_3, and hence is invariant by each element of the group $(\mathbb{Z}_2)^3$. So it is an essential torus which is $(\mathbb{Z}_2)^3$-equivariant.

If M_λ admits the E^3 geometry, then according to Theorems 8.2 and 8.5 of [Neumann and Raymond 1978], it is either the 3-torus T^3 or the Seifert manifold with invariant $\{0; (2, 1), (2, -1), (2, 1), (2, -1)\}$. For the former case, we choose a Seifert fibration structure which is fibred by all circles isotopic to the circles in $\mathrm{Fix}(e_1) \cap \mathrm{Fix}(e_2)$. Then we can apply the same argument as above to obtain a $(\mathbb{Z}_2)^3$-equivariant essential torus. For the latter case, we fix the Seifert fibration structure given before, and then all horizontal incompressible surfaces in M_λ are isotopic essential tori. This is because M_λ has a unique structure of surface bundles over S^1, since its first Betti number is 1; see [Thurston 1986]. So we can still

assume that both $\mathrm{Fix}(e_1)$ and $\mathrm{Fix}(e_2)$ consist of vertical essential tori, and e_3 keeps each regular fiber invariant and reverses its orientation. The same argument in the previous paragraph still applies, and the same conclusion still holds. □

Lemma 3.5. *M_λ is atoroidal.*

Proof. By Theorem 3.2 and Lemma 3.4, if M_λ is toroidal, then there is a $(\mathbb{Z}_2)^3$-equivariant essential torus $T \subset M_\lambda$.

Similar to the sphere case in Lemma 3.3, we also give a triangulation $\mathcal{T}$ of N. We denote $T \cap N \times \{1\}$ by A, which is nonempty, and is a compact connected surface with nonempty boundary. Also, there is a cell decomposition of ∂A induced from the triangulation of N.

Similar to the argument in the sphere case in Lemma 3.3, we have $\chi(T) = m(\chi(A) - E/2 + E/4) = 0$, where E is the number of edges in ∂A, and m is an integer. So $E = 4$ and $\chi(A) = 1$. Thus A is a disk with ∂A consists of 4 edges. Moreover, $\partial A \cap f$ is connected for each face f. Suppose otherwise; i.e., suppose that $\partial A \cap f$ consists of at least two arcs. When we lift ∂N to the universal cover $\tilde{N}$, two of the four edges forming ∂A belong to the same face. If these two edges are adjacent in ∂A, then by the same argument as in the proof of Lemma 3.3, we obtain a contradiction. If these two edges are not adjacent in ∂A, then the lift of these two arcs in the universal cover are identified. So in the universal cover, there are two disks which share two distinct edges, contradicting the assumption that Γ is admissible. Thus ∂A is a 4-circuit. Therefore, by the assumption and Lemma 2.2, A is an edge-linking disk. The preimage of an edge-linking disk in M_λ is a torus (or a Klein bottle) which bounds a solid torus (or a solid Klein bottle, which is impossible since we assume M_λ is orientable in this subsection), so it is not essential. This contradicts the assumption that T is essential. So M_λ is atoroidal. □

Lemma 3.6. *M_λ is not a Seifert manifold.*

Proof. Suppose M_λ is an orientable Seifert manifold with orientable base orbifold, and M_λ is neither a lens space nor S^3. Here the lens spaces don't include S^3 or $S^2 \times S^1$. By Theorem 8.2 of [Neumann and Raymond 1978] and its proof, if there is an orientation-reversing involution on M_λ, then the Seifert invariant of M_λ is $\{g; (a_1, b_1), (a_1, -b_1), (a_2, b_2), (a_2, -b_2), \ldots, (a_t, b_t), (a_t, -b_t)\}$, where g is the genus of the base orbifold. Since M_λ is atoroidal, we have $g = 0$ and $t = 1$, and hence M_λ is a lens space. This is a contradiction.

Suppose M_λ is an orientable Seifert manifold with nonorientable base orbifold, and M_λ is not a lens space. By Theorem 8.5 of [Neumann and Raymond 1978] and its proof, if there is an orientation-reversing involution on M_λ, then the Seifert invariant of M_λ is $\{k; (a_1, b_1), (a_1, -b_1), (a_2, b_2), (a_2, -b_2), \ldots, (a_t, b_t), (a_t, -b_t)\}$, where k is the genus of the nonorientable base orbifold. If $t \geq 1$, then M_λ cannot be

atoroidal. If $t = 0$, then M_λ is either reducible or toroidal. In both cases, we arrive at contradictions.

Suppose M_λ is a lens space. By the main result in [Kwun 1970], among all lens spaces, only RP^3 admits orientation-reversing involutions. Moreover, RP^3 admits exactly one orientation-reversing involution up to isotopies, and the set of fixed points of this involution is an RP^2, which has Euler characteristic 1. However, according to the definition of locally standard $(\mathbb{Z}_2)^3$-action, for any nontrivial element $e \in (\mathbb{Z}_2)^3$, its fixed point set $\mathrm{Fix}(e)$ is a union of k-polygons ($k \geq 5$ by our assumption), and each vertex in $\mathrm{Fix}(e)$ is adjacent to 4 edges. Let v be the number of vertices in $\mathrm{Fix}(e)$. Then the number of edges in $\mathrm{Fix}(e)$ is $2v$, and the number of faces of $\mathrm{Fix}(e)$ is less than or equal to $4v/5$. So the Euler characteristic of $\mathrm{Fix}(e)$ is negative — a contradiction.

Suppose M_λ is the 3-sphere S^3. From the fact that the orientation-preserving mapping class group of S^3 is trivial, we know S^3 admits exactly one orientation-reversing involution up to isotopy, and the set of fixed points of this involution is an S^2, which has Euler characteristic 2. Then similar to the argument in the previous paragraph, we get a contradiction. So the lemma follows. □

By Lemmas 3.3, 3.5 and 3.6, M_λ is a closed, irreducible, and atoroidal manifold which is not Seifert-fibered. So by Perelman's proof of Thurston's geometrization theorem (see [Cao and Zhu 2006; Bessières et al. 2010; Kleiner and Lott 2008; Morgan and Tian 2007]), M_λ is a hyperbolic 3-manifold. By [Dinkelbach and Leeb 2009], every smooth action of a finite group on a hyperbolic 3-manifold is conjugate to an isometric action. Since each $e \in (\mathbb{Z}_2)^3$ is conjugate to an isometric involution, its fixed point set is a totally geodesic surface in M_λ. Since $(\mathbb{Z}_2)^3$ is an Abelian group, by elementary arguments for the isometric group of hyperbolic 3-space H^3, all these totally geodesic surfaces intersect orthogonally. So the hyperbolic structure on M_λ induces a hyperbolic structure on (N, Γ). Conversely, each right-angled hyperbolic structure on (N, Γ) induces a hyperbolic structure on M_λ. By Mostow's rigidity theorem [1973], there is only one hyperbolic structure on M_λ. So the right-angled realization of (N, Γ) is unique. This ends the proof of Theorem 1.3 in the case that M_λ is orientable. □

Proof of the sufficiency part of Theorem 1.3 in the case that M_λ is nonorientable. Let $\pi : \tilde{M}_\lambda \to M_\lambda$ be the orientable double cover of M_λ, and τ be the covering transformation of $\tilde{M}_\lambda$. Note that τ is orientation-reversing. By the lifting theorem, for each i, e_i lifts to an action, say $\tilde{e}_i$, on $\tilde{M}_\lambda$ such that $\tilde{e}_i(x_0) = x_0$, where $x_0 \in \tilde{M}_\lambda$ projects to a vertex of Γ in ∂N.

We show that $\tilde{e}_i$ and $\tilde{e}_j$ commute, for $1 \leq i, j \leq 3$. It is easy to verify that $\tilde{e}_i\tilde{e}_j$ is the lift of e_ie_j, and $\tilde{e}_j\tilde{e}_i$ is the lift of e_je_i. Since $e_ie_j = e_je_i$, and $\tilde{e}_i\tilde{e}_j(x_0) = \tilde{e}_j\tilde{e}_i(x_0)$, by the unique lifting property, $\tilde{e}_i\tilde{e}_j = \tilde{e}_j\tilde{e}_i$. We also show that τ and $\tilde{e}_i$ commute,

for $1 \leq i \leq 3$. It is easy to verify that both $\tau\tilde{e}_i$ and $\tilde{e}_i\tau$ are lifts of e_i. So either $\tau\tilde{e}_i = \tilde{e}_i\tau$ or $\tau\tau\tilde{e}_i = \tilde{e}_i\tau$. The latter is $\tilde{e}_i = \tilde{e}_i\tau$ in fact, which is impossible. So $\tau\tilde{e}_i = \tilde{e}_i\tau$. Therefore we have an action of $(\mathbb{Z}_2)^4$ on $\tilde{M}_\lambda$.

If $\tilde{M}_\lambda$ is a toroidal Seifert manifold, then by Lemma 3.4, there is an essential vertical torus T in $\tilde{M}_\lambda$ which is fixed by $\tilde{e}_1$, and is invariant by $\tilde{e}_i$, for $i = 2, 3$. For any point $p \in T$, we have $\tilde{e}_1\tau(p) = \tau\tilde{e}_1(p) = \tau(p)$. So $\tau(T) \subset \mathrm{Fix}(\tilde{e}_1)$. Hence either $\tau(T) = T$ or $\tau(T) \cap T = \varnothing$. It is straightforward to verify that T is $(\mathbb{Z}_2)^4$-equivariant.

Therefore, similar to the previous subsection, we can prove that $\tilde{M}_\lambda$ is irreducible and atoroidal. Moreover, if $\tilde{M}_\lambda$ is an atoroidal Seifert manifold, then it must be S^3 or RP^3. The action of τ on $\tilde{M}_\lambda$ has no fixed points. However, as stated in the previous subsection, any orientation-reversing involution on S^3 or RP^3 must have fixed points. We arrive at a contradiction.

So $\tilde{M}_\lambda$ is hyperbolic. Similar to the arguments in the previous subsection, (N, Γ) admits a unique right-angled hyperbolic structure. □

4. Examples

In this section we give three examples.

Example 4.1. The simplest way to construct a handlebody which admits right-angled hyperbolic structure is from the Löbell polyhedron $L(n)$ for $n \geq 5$ (see, for example, [Inoue 2008]). A Löbell polyhedron $L(n)$ admits a right-angled hyperbolic structure. Gluing two opposite n-gon faces of $L(n)$, we get a solid torus admitting right-angled hyperbolic structures, and whose boundary consists of $2n$ octagons.

For instance, from $L(5)$, which is a dodecahedron, we can get three solid tori, according to the twisting angle of gluing. All these solid tori satisfy Theorem 1.3. It is easy to see that they admit $(\mathbb{Z}_2)^3$-colorings, but don't admit one which satisfies the orientability criterion in Proposition 2.3.

This kind of right-angled hyperbolic solid tori are "simple", by which we mean we can obtain a right-angled hyperbolic polyhedron by cutting along a totally geodesic right-angled n-polygon P from the solid tori, where P intersects the boundaries of the solid tori orthogonally.

Example 4.2. A hexagonal tessellation of $\mathbb{R}^2$ with a coloring is shown in Figure 1. We assume that the diameter of a hexagon is 1. We take a $\mathbb{Z}^2$-action on $\mathbb{R}^2$, such that its fundamental domain is a rectangle R whose vertical edges have length 4.5, and whose horizontal edges have length $3\sqrt{3}$. So there are six hexagons in each horizontal layer and each vertical layer. Gluing the boundaries of R, we get a torus T.

We can show that any solid torus bounded by T with coloring shown in Figure 1 satisfies the orientability criterion in Proposition 2.3 as well as the assumption of Theorem 1.3. So it admits a right-angled hyperbolic structure.

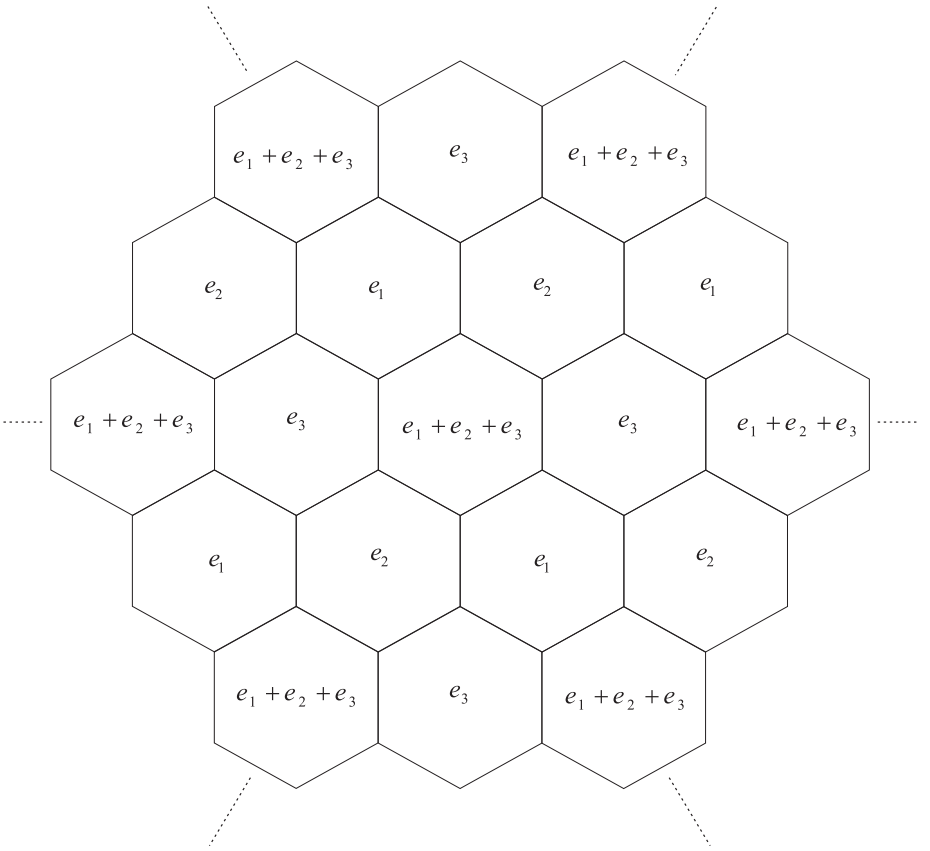


Figure 1

We fix a homeomorphism from T to the boundary of a solid torus J, so it is natural to ask whether the pair (J, Γ) admits a right-angled hyperbolic structure.

It is easy to see that any essential simple closed curve C in this T intersects Γ in at least five points, and any curve C which bounds a disk D in T intersects Γ in at least five points, unless that D is a vertex-linking disk or an edge-linking disk. So for any solid torus J which is bounded by T, (J, Γ) realizes a right-angled hyperbolic structure.

If the boundary of the unique essential disk in the solid torus J is the image of a horizontal line, then the hyperbolic solid torus can be decomposed into three copies of the Löbell polyhedron $L(6)$ along three totally geodesic right-angled hexagons in the solid torus. The same claim holds if the boundary of the unique essential disk in the solid torus is the image of the straight lines which have angles $\pi/3$ or $2\pi/3$ with the horizontal lines.

Except in these three cases, the right-angled hyperbolic structure cannot be obtained by gluing two faces of a right-angled hyperbolic polyhedron by an isometry,

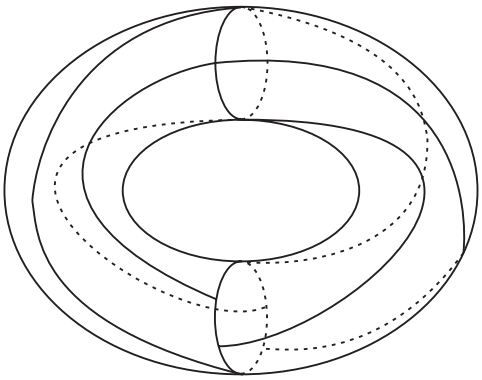


Figure 2

so it is not "simple". Suppose otherwise; then the totally geodesic right-angled k-polygon P which decomposes J is in general position with Γ, and so some faces of ∂J must be decomposed into a set of right-angled hyperbolic n-polygons by ∂P. Note that $n \geq 5$, so if ∂P enters a face f of $\partial J - \Gamma$, then it exits f from the opposite edge of f from where it enters. It is easy to see that ∂P is the image of the lines in $\mathbb{R}^2$ which have angles 0, $\pi/3$ or $2\pi/3$ with the horizontal lines.

Example 4.3. The graph Γ decomposes the torus illustrated in Figure 2 [Chen 2009] into three hexagons, say f_1, f_2 and f_3. We color f_i by $e_i \in (\mathbb{Z}_2)^3$ for $i = 1, 2, 3$. There are two sets of disks in Theorem 1.3. The first one consists of boundary parallel disks. The second one consists of essential disks, i.e., not boundary parallel.

For any embedding of (T^2, Γ) of Figure 2 into a solid torus J, the boundary parallel disks satisfy the assumption of Theorem 1.3. So if we embed (T^2, Γ) into a solid torus J by a map f so that the unique essential disk D (up to isotopy) intersects Γ in at least 5 points, then by Theorem 1.3, we get a right-angled hyperbolic structure on $(J, f(\Gamma))$. Note that for a fixed embedding of $\Gamma \to T^2$, there are only finitely many isotopy classes of simple closed curves which intersect Γ in at most 4 points.

In general, if the pair $(\partial N, \Gamma)$ admits a $(\mathbb{Z}_2)^3$-coloring and ∂N has genus at least one, then it may admit many colorings for a re-embedding of Γ into ∂N. This in turn induces many closed 3-manifolds from locally standard $(\mathbb{Z}_2)^3$-actions.

Acknowledgements

The second author would like to thank Michel Boileau and Jean-Marc Schlenker for discussions on these topics. Both authors would like to thank Zhi Lü for helpful conversations. We also would like to thank the referee for many useful comments.

References

[Andreev 1971] E. M. Andreev, "On convex polyhedra of finite volume in Lobačevski space", *Math. USSR Sbornik* **12**:3 (1971), 225–259. Zbl 0252.52005

[Bao and Bonahon 2002] X. Bao and F. Bonahon, "Hyperideal polyhedra in hyperbolic 3-space", *Bull. Soc. Math. France* **130**:3 (2002), 457–491. MR 2003k:52007 Zbl 1033.52009

[Bessières et al. 2010] L. Bessières, G. Besson, S. Maillot, M. Boileau, and J. Porti, *Geometrisation of 3-manifolds*, EMS Tracts in Mathematics **13**, European Mathematical Society, Zürich, 2010. MR 2012d:57027 Zbl 1244.57003

[Cao and Zhu 2006] H.-D. Cao and X.-P. Zhu, "A complete proof of the Poincaré and geometrization conjectures—application of the Hamilton–Perelman theory of the Ricci flow", *Asian J. Math.* **10**:2 (2006), 165–492. MR 2008d:53090 Zbl 1200.53057

[Chen 2009] B. Chen, *A study of 2-torus topology*, Ph.D. thesis, Fudan University, 2009. In Chinese.

[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

[Dinkelbach and Leeb 2009] J. Dinkelbach and B. Leeb, "Equivariant Ricci flow with surgery and applications to finite group actions on geometric 3-manifolds", *Geom. Topol.* **13**:2 (2009), 1129–1173. MR 2011b:53158 Zbl 1181.57023

[Dunwoody 1985] M. J. Dunwoody, "An equivariant sphere theorem", *Bull. London Math. Soc.* **17**:5 (1985), 437–448. MR 87f:57008 Zbl 0592.57005

[Fillastre and Izmestiev 2009] F. Fillastre and I. Izmestiev, "Hyperbolic cusps with convex polyhedral boundary", *Geom. Topol.* **13**:1 (2009), 457–492. MR 2009i:57039 Zbl 1179.57026

[Fillastre and Izmestiev 2011] F. Fillastre and I. Izmestiev, "Gauss images of hyperbolic cusps with convex polyhedral boundary", *Trans. Amer. Math. Soc.* **363**:10 (2011), 5481–5536. MR 2012b:57034 Zbl 1238.57018

[Freedman et al. 1983] M. Freedman, J. Hass, and P. Scott, "Least area incompressible surfaces in 3-manifolds", *Invent. Math.* **71**:3 (1983), 609–642. MR 85e:57012 Zbl 0482.53045

[Guéritaud 2009] F. Guéritaud, "Deforming ideal solid tori", preprint, 2009. arXiv 0911.3067

[Hempel 1976] J. Hempel, 3*-Manifolds*, Ann. of Math. Studies **86**, Princeton University Press, 1976. MR 54 #3702 Zbl 0345.57001

[Hodgson and Rivin 1993] C. D. Hodgson and I. Rivin, "A characterization of compact convex polyhedra in hyperbolic 3-space", *Invent. Math.* **111**:1 (1993), 77–111. MR 93j:52015 Zbl 0784.52013

[Inoue 2008] T. Inoue, "Organizing volumes of right-angled hyperbolic polyhedra", *Algebr. Geom. Topol.* **8**:3 (2008), 1523–1565. MR 2009k:57025 Zbl 1146.52005

[Jaco 1980] W. Jaco, *Lectures on three-manifold topology*, CBMS Regional Conference Series in Mathematics **43**, American Mathematical Society, Providence, R.I., 1980. MR 81k:57009 Zbl 0433.57001

[Jaco and Shalen 1979] W. H. Jaco and P. B. Shalen, *Seifert fibered spaces in* 3*-manifolds*, Mem. Amer. Math. Soc. **220**, Amer. Math. Soc., Providence, RI, 1979. MR 81c:57010 Zbl 0415.57005

[Johannson 1979] K. Johannson, *Homotopy equivalences of* 3*-manifolds with boundaries*, Lecture Notes in Mathematics **761**, Springer, Berlin, 1979. MR 82c:57005 Zbl 0412.57007

[Kleiner and Lott 2008] B. Kleiner and J. Lott, "Notes on Perelman's papers", *Geom. Topol.* **12**:5 (2008), 2587–2855. MR 2010h:53098 Zbl 1204.53033

[Kwun 1970] K. W. Kwun, "Scarcity of orientation-reversing PL involutions of lens spaces", *Michigan Math. J.* **17** (1970), 355–358. MR 43 #5535 Zbl 0191.54901

[Lü 2009] Z. Lü, "2-torus manifolds, cobordism and small covers", *Pacific J. Math.* **241**:2 (2009), 285–308. MR 2010k:55008 Zbl 1181.57036

[Lü and Masuda 2009] Z. Lü and M. Masuda, "Equivariant classification of 2-torus manifolds", *Colloq. Math.* **115**:2 (2009), 171–188. MR 2010k:52023 Zbl 1165.57023

[Lü and Yu 2011] Z. Lü and L. Yu, "Topological types of 3-dimensional small covers", *Forum Math.* **23**:2 (2011), 245–284. MR 2012h:57038 Zbl 1222.52015

[Meeks and Scott 1986] W. H. Meeks, III and P. Scott, "Finite group actions on 3-manifolds", *Invent. Math.* **86**:2 (1986), 287–346. MR 88b:57039 Zbl 0626.57006

[Meeks et al. 1982] W. Meeks, III, L. Simon, and S. T. Yau, "Embedded minimal surfaces, exotic spheres, and manifolds with positive Ricci curvature", *Ann. of Math.* (2) **116**:3 (1982), 621–659. MR 84f:53053 Zbl 0521.53007

[Morgan and Tian 2007] J. Morgan and G. Tian, *Ricci flow and the Poincaré conjecture*, Clay Mathematics Monographs **3**, American Mathematical Society, Providence, RI, 2007. MR 2008d:57020 Zbl 1179.57045

[Mostow 1973] G. D. Mostow, *Strong rigidity of locally symmetric spaces*, Annals of Mathematics Studies **78**, Princeton University Press, 1973. MR 52 #5874 Zbl 0265.53039

[Nakayama and Nishimura 2005] H. Nakayama and Y. Nishimura, "The orientability of small covers and coloring simple polytopes", *Osaka J. Math.* **42**:1 (2005), 243–256. MR 2006a:57023 Zbl 1065.05041

[Neumann and Raymond 1978] W. D. Neumann and F. Raymond, "Seifert manifolds, plumbing, μ-invariant and orientation reversing maps", pp. 163–196 in *Algebraic and geometric topology* (Santa Barbara, CA, 1977), edited by K. C. Millett, Lecture Notes in Math. **664**, Springer, Berlin, 1978. MR 80e:57008 Zbl 0401.57018

[Rivin 1996] I. Rivin, "A characterization of ideal polyhedra in hyperbolic 3-space", *Ann. of Math.* (2) **143**:1 (1996), 51–70. MR 96i:52008 Zbl 0874.52006

[Roeder et al. 2007] R. K. W. Roeder, J. H. Hubbard, and W. D. Dunbar, "Andreev's theorem on hyperbolic polyhedra", *Ann. Inst. Fourier* (*Grenoble*) **57**:3 (2007), 825–882. MR 2008e:51011 Zbl 1127.51012

[Schlenker 2002] J.-M. Schlenker, "Hyperbolic manifolds with polyhedral boundary", preprint, 2002. arXiv 0111136

[Schlenker 2003] J.-M. Schlenker, "Hyperideal polyhedra in hyperbolic manifolds", preprint, 2003. arXiv 0212355

[Schlenker 2005] J.-M. Schlenker, "Hyperideal circle patterns", *Math. Res. Lett.* **12**:1 (2005), 85–102. MR 2005k:52043 Zbl 1067.52016

[Schlenker 2006] J.-M. Schlenker, "Hyperbolic manifolds with convex boundary", *Invent. Math.* **163**:1 (2006), 109–169. MR 2006m:57023 Zbl 1091.53019

[Scott 1983] P. Scott, "The geometries of 3-manifolds", *Bull. London Math. Soc.* **15**:5 (1983), 401–487. MR 84m:57009 Zbl 0561.57001

[Thurston 1982] W. P. Thurston, "Three-dimensional manifolds, Kleinian groups and hyperbolic geometry", *Bull. Amer. Math. Soc.* (*N.S.*) **6**:3 (1982), 357–381. MR 83h:57019 Zbl 0496.57005

[Thurston 1986] W. P. Thurston, "A norm for the homology of 3-manifolds", *Mem. Amer. Math. Soc.* **59**:339 (1986), i–vi and 99–130. MR 88h:57014

Received March 31, 2012. Revised November 11, 2012.

YOULIN LI
DEPARTMENT OF MATHEMATICS
SHANGHAI JIAOTONG UNIVERSITY
200240 SHANGHAI
CHINA
liyoulin@sjtu.edu.cn

JIMING MA
SCHOOL OF MATHEMATICAL SCIENCES
FUDAN UNIVERSITY
200433 SHANGHAI
CHINA
majiming@fudan.edu.cn

PACIFIC JOURNAL OF MATHEMATICS
Vol. 263, No. 2, 2013
dx.doi.org/10.2140/pjm.2013.263.435

REAL CLOSED SEPARATION THEOREMS AND APPLICATIONS TO GROUP ALGEBRAS

TIM NETZER AND ANDREAS THOM

Dedicated to Konrad Schmüdgen on the occasion of his 65th birthday

In this paper we prove a strong Hahn–Banach theorem: separation of disjoint convex sets by linear forms is possible without any further conditions if the target field $\mathbb{R}$ is replaced by a more general real closed extension field. From this we deduce a general Positivstellensatz for $*$-algebras, involving representations over real closed fields. We investigate the class of group algebras in more detail. We show that the cone of sums of squares in the augmentation ideal has an interior point if and only if the first cohomology vanishes. For groups with Kazhdan's property (T), the result can be strengthened to interior points in the ℓ^1-metric. We finally reprove some strong Positivstellensätze by Helton and Schmüdgen, using our separation method.

1. Introduction

In this article we combine techniques from real algebraic geometry, convex geometry, and the unitary representation theory of discrete groups to address various problems that arise in the emerging field of noncommutative real algebraic geometry [Schmüdgen 2009]. Classical results — like Artin's solution of Hilbert's 17th problem — strive for a characterization of natural notions of positivity in terms of algebraic certificates. For example, Artin proved that every polynomial in n variables that is positive at every point on $\mathbb{R}^n$ must be a sum of squares of rational functions. Much later, Schmüdgen [1991] proved that a strictly positive polynomial on a compact semialgebraic set must be a sum of squares of polynomials plus defining inequalities. More recently, similar questions were asked in a noncommutative context. Typically, the setup involves a $*$-algebra A and a family of representations $\mathcal{F}$. The question is now: Is every self-adjoint element of A that is positive (semi)definite in every representation in $\mathcal{F}$ necessarily of the form $\sum_i a_i^* a_i$ for some $a_i \in A$? It turned out — similar to the more classical commutative case — that the cone $\Sigma^2 A = \left\{\sum_i a_i^* a_i \mid a_i \in A\right\} \subset A$ is an interesting object of study in itself. Natural

MSC2010: 46H15.
Keywords: real closed fields, group rings, Kazhdan's property (T), sums of squares.

questions are: Is $\Sigma^2(A) \cap (-\Sigma^2 A) = \{0\}$? Is $\Sigma^2 A$ closed in a natural topology? Does it contain interior points, in the finest locally convex topology, say?

A point q is called an algebraic interior point of a cone C if the cone intersects each line through q in an open interval around q. A point is an algebraic interior point if and only if it is an interior point in the finest locally convex topology (see [Cimprič et al. 2011, Proposition 1.3], for example). The question for interior points of cones has the following motivation. If a cone C has an (algebraic) interior point q, then for every point $a \in C^{\vee\vee}$ from the double dual, one has $a + \epsilon q \in C$ for all $\epsilon > 0$ (see [loc. cit.] for a proof of this well known fact). Using the standard Gelfand–Naimark–Segal construction, this yields the following Positivstellensatz for unital $*$-algebras:

Theorem. *Assume that q is an interior point of the cone $\Sigma^2 A$. If $a = a^* \in A$ is positive semidefinite in each $*$-representation of A, then $a + \epsilon q \in \Sigma^2 A$ for all $\epsilon > 0$.*

Our first main result is a different Positivstellensatz (Theorem 3.12): we prove that each element from a real reduced unital $*$-algebra that is positive in every generalized representation is necessarily in $\Sigma^2 A$. The notion of a *generalized representation* involves an extension of the standard real and complex numbers to more general real- and algebraically closed fields.

A natural and vast class of examples of $*$-algebras is given by complex group algebras $\mathbb{C}[\Gamma]$ of discrete countable groups. We study the cones $\Sigma^2\mathbb{C}[\Gamma]$ and $\Sigma^2\omega(\Gamma)$ in more detail, where $\omega(\Gamma) \subset \mathbb{C}[\Gamma]$ denotes the augmentation ideal; see Section 4. The situation for $\omega(\Gamma)$ is much more complicated, as the study is closely related to questions about first cohomology with unitary coefficients.We prove that $\Sigma^2\omega(\Gamma)$ has an interior point if and only if $H_1(\Gamma, \mathbb{C}) = 0$. The cone $\Sigma^2\omega(\Gamma)$ has an interior point in the ℓ^1-metric if Γ has Kazhdan's property (T), and the converse holds if $H_2(\Gamma, \mathbb{C}) = 0$ (see Section 5). In Section 6, we analyze the situation for free groups more closely and reprove theorems of Schmüdgen and Helton.

Along the way we prove some new and powerful separation theorems in Sections 2 and 3. The Hahn–Banach separation theorems for convex sets only apply if additional conditions on the involved sets are imposed; sets have to be closed or have to have nonempty interior, etc. We can remove *all* additional assumptions at the expense of enlarging the target $\mathbb{R}$ to some real closed extension of the real numbers; see Theorem 2.1.

2. A real closed separation theorem for convex sets

Throughout, we will work with various real closed fields $\mathbf{R}$ and *always* assume that $\mathbb{R} \subset \mathbf{R}$. The following is a first general separation theorem for convex cones.

Theorem 2.1. *Let V be an $\mathbb{R}$-vector space, $C \subset V$ a convex cone and $x \notin C$. Then there exist a real closed field $\mathbf{R}$ containing $\mathbb{R}$ and an $\mathbb{R}$-linear functional $\varphi\colon V \to R$*

such that

$$\varphi(x) < 0 \quad \textit{and} \quad \varphi(y) \geq 0 \quad \textit{for all} \quad y \in C.$$

We can even ensure $\varphi(y) > 0$ *for all* $y \in C \setminus (C \cap -C)$. *Also,* $\mathbf{R}$ *depends only on* V, *not on* x *or* C.

Proof. Let us first assume that V is finite-dimensional. We construct a complete flag of subspaces $V = H_1 \supset H_2 \supset \cdots \supset H_n = C \cap -C$, starting with $V = H_1$, in the following way. By the standard separation theorem for convex sets (see for example Theorem 2.9 in [Barvinok 2002]), we choose a nontrivial $\mathbb{R}$-linear functional $\varphi_i \colon H_i \to \mathbb{R}$ such that $\varphi_i(y) \geq 0$ for all $y \in C \cap H_i$ and $\varphi_i(x) \leq 0$ (if $x \in H_i$). We then define $H_{i+1} := H_i \cap \{\varphi_i = 0\}$ and iterate the process. We finally extend each φ_i in any way to V. Now let $\mathbf{R}$ be a proper real closed extension field of $\mathbb{R}$. Choose positive elements

$$1 = \epsilon_1 > \epsilon_2 > \cdots > \epsilon_{n-1} > 0$$

from $\mathbf{R}$ such that $k \cdot \epsilon_i < \epsilon_{i-1}$ for all $k \in \mathbb{R}$. For example, ϵ_2 can be any infinitesimal element with respect to $\mathbb{R}$, which exists since $\mathbf{R}$ is a proper extension of $\mathbb{R}$; the following ϵ_i can be taken as powers of ϵ_2.

Then define

$$\varphi := \epsilon_1\varphi_1 + \cdots + \epsilon_{n-1}\varphi_{n-1}.$$

One checks that φ has the desired properties. This proves the claim in the case of finite dimension.

In the general case, consider the set $\mathcal{S}$ of all finite-dimensional subspaces H of V. For each $H \in \mathcal{S}$, choose $\varphi_H \colon H \to \mathbf{R}$, separating x from $C \cap H$ as desired (if $x \in H$). Extend φ_H in any way to V. Now let ω be an ultrafilter on $\mathcal{S}$, containing the sets $\{H \in \mathcal{S} \mid y \in H\}$ for all $y \in V$. Consider the linear functional $\varphi \colon V \to \mathbf{R}^\omega$, $\varphi(v) = (\varphi_H(v))_{H \in \mathcal{S}}$. Here $\mathbf{R}^\omega$ denotes the ultrapower of $\mathbf{R}$ with respect to ω. One checks that φ separates x from C as desired, by the theorem of Łos (see for example Theorem 2.2.9 in [Prestel and Delzell 2001]). □

Remark 2.2. In the usual way, one can now also deduce that any two convex disjoint sets in a vector space can be separated as above with a real-closed valued affine functional.

It turns out that we can also extend functionals quite often if we allow for an extension of the real closed field.

Theorem 2.3. *Let* V *be an* $\mathbb{R}$*-vector space,* $C \subseteq V$ *a convex cone, and* $H \subseteq V$ *a subspace. Assume* $(C + H) \cap -(C + H) = H$. *Then for any real closed extension field* $\mathbf{R}$ *of* $\mathbb{R}$ *and any* $\mathbb{R}$*-linear functional* $\varphi \colon H \to \mathbf{R}$ *with* $\varphi \geq 0$ *on* $C \cap H$, *there is a real closed extension field* $\mathbf{R}'$ *of* $\mathbf{R}$ *and an* $\mathbb{R}$*-linear functional* $\overline{\varphi} \colon V \to \mathbf{R}'$ *with* $\overline{\varphi} \geq 0$ *on* C *and* $\overline{\varphi} = \varphi$ *on* H. *We can even ensure* $\overline{\varphi}(y) > 0$ *for all* $y \in C \setminus H$.

Proof. We apply Theorem 2.1 to the convex cone $C + H$ in V and obtain a real closed field $\widetilde{\mathbf{R}}$ and an $\mathbb{R}$-linear functional $\psi : V \to \widetilde{\mathbf{R}}$ with $\psi = 0$ on H and $\psi(y) > 0$ for $a \in C \setminus H$. By amalgamation of real closed fields, we can assume without loss of generality that $\mathbf{R} = \widetilde{\mathbf{R}}$. Finally, let $\mathbf{R}'$ be a real closed extension field of $\mathbf{R}$ that contains an element $\delta > \mathbf{R}$. Extend φ to an $\mathbf{R}$-valued functional on V and set

$$\overline{\varphi} := \varphi + \delta \cdot \psi.$$

It is clear that $\overline{\varphi}$ coincides with φ on H and also that $\overline{\varphi}(y) > 0$ for all $y \in C \setminus H$. □

We will improve upon the separation results in the case of certain $*$-algebras in the next section.

3. Completely positive separation

Throughout this section, let A be a $\mathbb{C}$-algebra with involution $*$, not necessarily unital. We consider the cone of sums of hermitian squares

$$\Sigma^2 A = \left\{ \sum_{i=1}^{n} a_i^* a_i \;\middle|\; n \in \mathbb{N}, a_i \in A \right\}$$

contained in the real vector subspace of hermitian elements

$$A^h = \{a \in A \mid a^* = a\}.$$

If $b \in A^h \setminus \Sigma^2 A$, we find an $\mathbb{R}$-linear functional $\varphi\colon A^h \to \mathbf{R}$, into some real closed extension field $\mathbf{R}$ of $\mathbb{R}$, such that $\varphi(b) < 0$, $\varphi(a^*a) \geq 0$ for all $a \in A$, by Theorem 2.1. We can extend φ uniquely to a $\mathbb{C}$-linear functional $\varphi\colon A \to \mathbf{R}[i]$ fulfilling $\varphi(a^*) = \overline{\varphi(a)}$. We will denote the algebraically closed field $\mathbf{R}[i]$ by $\mathbf{C}$ from now on.

The condition $\varphi(a^*a) \geq 0$ for all $a \in A$ is called *positivity* of φ. We would now like positive and real-closed valued functionals to fulfill the Cauchy–Schwarz inequality

$$|\varphi(a^*b)|^2 \leq \varphi(a^*a)\varphi(b^*b)$$

for all $a, b \in A$. However, this is not true in general, as the next example shows.

Example 3.1. Let $A = \mathbb{C}[t]$ be the univariate polynomial ring; $*$ is coefficientwise conjugation. The cone $\Sigma^2 A$ equals the cone of nonnegative real polynomials. Consider the functional

$$\varphi\colon \mathbb{R}[t] \to \mathbf{R}, \quad p \mapsto p(0) + \epsilon p''(0),$$

where $\epsilon \in \mathbf{R}$ is positive and infinitesimal with respect to $\mathbb{R}$. One checks that φ is positive, but for $a = 1 + t^2$ and $b = 1$, we have

$$|\varphi(a^*b)|^2 = 1 + 4\epsilon + 4\epsilon^2 > 1 + 4\epsilon = \varphi(a^*a)\varphi(b^*b).$$

Example 3.2. The last example can be modified to even fulfill $\varphi(a^*a) > 0$ if $a \neq 0$. Indeed, let $1 = \epsilon_0 > \epsilon_1 > \epsilon_2 > \cdots > 0$ be a sequence of elements from $\mathbf{R}$, such that $\mathbb{R} \cdot \epsilon_i < \epsilon_{i-1}$ for all i. Then the linear mapping $p \mapsto \sum_{i=0}^{\infty} \epsilon_i \cdot p^{(2i)}(0)$ is well-defined and strictly positive in the desired sense. If we further assume $\mathbb{R} \cdot \epsilon_2 < \epsilon_1^2$, then the same argument as in Example 3.1 shows that the Cauchy–Schwarz inequality is not fulfilled.

Definition 3.3. A $\mathbb{C}$-linear functional $\varphi\colon A \to \mathbf{C}$ with $\varphi(a^*) = \overline{\varphi(a)}$ is called *completely positive* if for all $m \in \mathbb{N}$, the componentwise defined function

$$\varphi^{(m)}\colon \mathrm{M}_m(A) \to \mathrm{M}_m(\mathbf{C})$$

maps sums of hermitian squares to positive semidefinite matrices.

Remark 3.4. It is easily seen that a positive $\mathbb{C}$-linear functional $\varphi\colon A \to \mathbb{C}$ with $\varphi(a^*) = \overline{\varphi(a)}$ is always completely positive.

Example 3.5. The functionals from Examples 3.1 and 3.2 are positive, but not completely positive. Indeed, with $a = 1 + t^2$ and

$$M = \begin{pmatrix} 1 & a \\ 0 & 0 \end{pmatrix},$$

we find that

$$\varphi^{(2)}(M^*M) = \begin{pmatrix} 1 & \varphi(a) \\ \varphi(a^*) & \varphi(a^*a) \end{pmatrix}$$

is not positive semidefinite, since its determinant is negative in $\mathbf{R}$.

Lemma 3.6. *A $\mathbb{C}$-linear functional $\varphi\colon A \to \mathbf{C}$ with $\varphi(a^*) = \overline{\varphi(a)}$ is completely positive if and only if the $\mathbf{C}$-linear extension*

$$\mathrm{id} \otimes \varphi\colon \mathbf{C} \otimes_{\mathbb{C}} A \to \mathbf{C}$$

is positive.

Proof. The condition that $\mathrm{id} \otimes \varphi$ is positive is

$$\begin{aligned} 0 &\leq (\mathrm{id} \otimes \varphi)\left(\left(\sum_{j=1}^{m} z_j \otimes a_j\right)^* \left(\sum_{j=1}^{m} z_j \otimes a_j\right)\right) \\ &= (\mathrm{id} \otimes \varphi)\left(\sum_{j,k} \overline{z}_j z_k \otimes a_j^* a_k\right) \\ &= \sum_{j,k} \overline{z}_j z_k \cdot \varphi(a_j^* a_k) \end{aligned}$$

for all $m \in \mathbb{N}$, $z_j \in \mathbf{C}$, $a_j \in A$. But this just means that the matrix $(a_j^* a_k)_{j,k}$ is mapped to a positive semidefinite matrix under $\varphi^{(m)}$. Since every sum of hermitian squares in $\mathrm{M}_m(A)$ is a finite sum of such rank-one squares, this proves the claim. □

Corollary 3.7. *If $\varphi\colon A \to \mathbf{C}$ is completely positive, then it fulfills the Cauchy–Schwarz inequality*

$$|\varphi(a^*b)|^2 \leq \varphi(a^*a)\varphi(b^*b)$$

for all $a, b \in A$.

Proof. Either consider the positive and $\mathbf{C}$-linear extension $\mathrm{id} \otimes \varphi$ to $\mathbf{C} \otimes_{\mathbb{C}} A$ and use the standard proof for the inequality, or apply $\varphi^{(2)}$ to the sum of hermitian squares

$$\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}^* \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} a^*a & a^*b \\ b^*a & b^*b \end{pmatrix},$$

and use that the obtained matrix is positive semidefinite. □

Remark 3.8. We see from the last proof that in fact only the 2-positivity of φ is needed for the Cauchy–Schwarz inequality.

Corollary 3.9. *Let A be a $\mathbb{C}$-algebra with involution and $\mathbf{R}$ a real closed field that contains $\mathbb{R}$. Let $\varphi\colon A \to \mathbf{C}$ be a completely positive $\mathbb{C}$-linear functional that satisfies $\varphi(a^*) = \overline{\varphi(a)}$ for all $a \in A$. The gauge $\|a\|_\varphi := \varphi(a^*a)^{1/2}$ satisfies*

$$\|\lambda \cdot a\|_\varphi = |\lambda| \cdot \|a\|_\varphi$$

and

$$\|a + b\|_\varphi \leq \cdot \|a\|_\varphi + \|b\|_\varphi.$$

Proof. The first assertion is obvious. Let's compute

$$\begin{aligned} \|a+b\|_\varphi^2 &= \varphi((a+b)^*(a+b)) \\ &= \varphi(a^*a) + \varphi(a^*b) + \varphi(b^*a) + \varphi(b^*b) \\ &\leq \|a\|_\varphi^2 + \|b\|_\varphi^2 + 2|\varphi(a^*b)| \\ &\leq \|a\|_\varphi^2 + \|b\|_\varphi^2 + 2\|a\|_\varphi\|b\|_\varphi \\ &= (\|a\|_\varphi + \|b\|_\varphi)^2. \end{aligned}$$

This proves the claim. □

It turns out that separation from the cone of sums of hermitian squares can often be done with a completely positive functional.

Definition 3.10. Let A be a $\mathbb{C}$-algebra with involution, not necessarily unital. Then A is called *real reduced* if $\sum_i a_i^* a_i = 0$ implies $a_i = 0$ for all i and $a_i \in A$.

Theorem 3.11. *Let A be a $\mathbb{C}$-algebra with involution that is real reduced. Let $b \in A^h \setminus \Sigma^2 A$. Then there is real closed extension field $\mathbf{R}$ of $\mathbb{R}$ and a completely positive $\mathbb{C}$-linear functional $\varphi\colon A \to \mathbf{C}$ with $\varphi(a^*) = \overline{\varphi(a)}$ such that*

$$\varphi(b) < 0 \quad \textit{and} \quad \varphi(a^*a) > 0 \quad \textit{for} \quad a \in A \setminus \{0\}.$$

Proof. For any finite-dimensional subspace H of A, denote by

$$\Sigma^2 H = \left\{ \sum_i a_i^* a_i \,\middle|\, a_i \in H \right\}$$

the set of sums of hermitian squares of elements from H. It is well known that $\Sigma^2 H$ is a closed convex cone in a finite-dimensional subspace of A^h. This follows from the fact that A is real reduced, using for example the approach from [Powers and Scheiderer 2001, Lemma 2.7]. It also follows that $\Sigma^2 H$ is salient, that is, it fulfills

$$\Sigma^2 H \cap -\Sigma^2 H = \{0\}.$$

So for each such H, there is an $\mathbb{R}$-linear functional $\varphi_H \colon A^h \to \mathbb{R}$ with

$$\varphi_H(b) < 0 \quad \text{and} \quad \varphi_H(a^* a) > 0 \quad \text{for all} \quad a \in H \setminus \{0\}.$$

Let $\mathcal{S}$ be the set of all finite-dimensional subspaces H of A, equipped with an ultrafilter ω containing all sets $\{H \in \mathcal{S} \mid c \in H\}$ for $c \in A$. Define

$$\varphi \colon A^h \to \mathbb{R}^\omega, \quad \varphi(a) := (\varphi_H(a))_{H \in \mathcal{S}}.$$

Then φ does the separation as desired. We consider the $\mathbb{C}$-linear extension of φ to A, and finally show that it is completely positive. The $\mathbb{C}$-linear extension of φ_H to A indeed maps a matrix $(a_i^* a_j)_{i,j} \in \mathrm{M}_n(A)$ to a positive semidefinite hermitian matrix, at least if all $a_i \in H$, as is easily checked (compare to Remark 3.4). Since we can check positivity of the matrix $(\varphi(a_i^* a_j))_{i,j} \in \mathrm{M}_n(\mathbb{R}^\omega[i])$ componentwisely in $\mathrm{M}_n(\mathbb{R}[i])$, by the theorem of Łos, this finishes the proof. □

Throughout, we will take the freedom to consider $*$-representations of A on vector spaces that carry a sesquilinear $\mathbf{C}$-valued inner product, where $\mathbf{C} = \mathbf{R}[i]$ for some real closed field $\mathbf{R} \supset \mathbb{R}$. We call these representations *generalized representations*. For every completely positive functional $\varphi \colon A \to \mathbf{C}$, we can perform the usual GNS construction to construct such a representation (see the proof of Theorem 6.1 below for more technical details). The usual concepts of self-adjointness and positive semidefiniteness of operators on such a vector space can be defined without any problems. The first consequence is the following Positivstellensatz (compare to the standard Positivstellensatz from the introduction):

Theorem 3.12. *Let A be a real reduced $*$-algebra, and $a \in A^h$. Then a is positive semidefinite in every generalized representation if and only if $a \in \Sigma^2 A$.*

Proof. If $a \notin \Sigma^2 A$, then there exists a completely positive map $\varphi \colon A \to \mathbf{C}$ such that $\varphi(a) < 0$. Clearly, a will not be positive semidefinite in the generalized GNS representation associated with φ. □

Examples for real reduced $*$-algebras are group algebras $\mathbb{C}[\Gamma]$. In Section 6, we will see that for particular groups, the study of generalized representations of $\mathbb{C}[\Gamma]$ can be reduced to the study of usual (finite-dimensional) unitary representations, using Tarski's transfer principle.

4. Sums of squares in the group algebra

Let Γ be a group and let $\mathbb{C}[\Gamma]$ denote the complex group algebra. A typical element in $\mathbb{C}[\Gamma]$ is denoted by $a = \sum_g a_g g$, where only finitely many of the $a_g \in \mathbb{C}$ are not zero. In $\mathbb{C}[\Gamma]$ we identify $\mathbb{C}$ with $\mathbb{C} \cdot e$, where e denotes the neutral element of Γ. The group algebra comes equipped with an involution $\left(\sum_g a_g g\right)^* := \sum_g \bar{a}_g g^{-1}$ and a trace $\tau : \mathbb{C}[\Gamma] \to \mathbb{C}$ that is given by the formula $\tau\left(\sum_g a_g g\right) = a_e$. The faithfulness of the trace shows that $\mathbb{C}[\Gamma]$ is real reduced. Let $\Sigma^2\mathbb{C}[\Gamma]$ denote the set of sums of hermitian squares in $\mathbb{C}[\Gamma]$. The following appears for example as Example 3 in [Cimprič 2009]:

Remark 4.1. $\|a\|_1^2 - a^*a \in \Sigma^2\mathbb{C}[\Gamma]$ for all $a \in \mathbb{C}[\Gamma]$, where $\|a\|_1 = \sum_g |a_g|$.

Remark 4.2. From the identity

$$2\|a\|_1 \cdot (\|a\|_1 - a) = (\|a\|_1 - a)^* (\|a\|_1 - a) + (\|a\|_1^2 - a^*a)$$

for $a \in \mathbb{C}[\Gamma]^h$, we see that 1 is an algebraic interior point of the cone $\Sigma^2\mathbb{C}[\Gamma]$ in the real vector space $\mathbb{C}[\Gamma]^h$. That means $1 + \epsilon a \in \Sigma^2\mathbb{C}[\Gamma]$ for all $a \in \mathbb{C}[\Gamma]^h$ and sufficiently small $\epsilon > 0$. In fact, the ϵ does only depend on $\|a\|_1$ here.

Remark 4.3. As explained in the introduction, for any element $a \in \mathbb{C}[\Gamma]^h$ that is positive semidefinite in each (usual) $*$-representation of $\mathbb{C}[\Gamma]$, one thus has $a + \epsilon \in \Sigma^2\mathbb{C}[\Gamma]$, for all $\epsilon > 0$.

Remark 4.4. Since $\mathbb{C}[\Gamma]$ is real reduced and unital, the result of Theorem 3.12 holds here as well. So if a is positive semidefinite in each generalized representation, then $a \in \Sigma^2\mathbb{C}[\Gamma]$.

We now consider the augmentation homomorphism $\varepsilon : \mathbb{C}[\Gamma] \to \mathbb{C}$, which is defined by $\varepsilon\left(\sum_g a_g g\right) = \sum_g a_g$. The *augmentation ideal* is

$$\omega(\Gamma) := \ker(\varepsilon) = \left\{ a \in \mathbb{C}[\Gamma] \;\middle|\; \sum_g a_g = 0 \right\}.$$

We set $c(g) := g - 1$ and note that $\{c(g) \mid g \in \Gamma \setminus \{e\}\}$ is a basis of $\omega(\Gamma)$. The multiplication satisfies

$$c(g)c(h) = c(gh) - c(g) - c(h).$$

We denote by $\omega^2(\Gamma)$ the square of $\omega(\Gamma)$, that is, $\omega^2(\Gamma) = \operatorname{span}_{\mathbb{C}}\{ab \mid a, b \in \omega(\Gamma)\}$. Inside $\omega(\Gamma)$, we study the cone of sums of hermitian squares

$$\Sigma^2\omega(\Gamma) := \Big\{\sum a_i^* a_i \,\Big|\, a_i \in \omega(\Gamma)\Big\}.$$

We are interested in interior points of this cone. Note that $\Sigma^2\omega(\Gamma) \subset \omega^2(\Gamma)$ and $\omega(\Gamma)/\omega^2(\Gamma) = \mathbb{C} \otimes_{\mathbb{Z}} \Gamma_{ab}$, where $\Gamma_{ab} = \Gamma/[\Gamma, \Gamma]$ and $[\Gamma, \Gamma]$ denotes the subgroup of Γ generated by commutators. Hence, if Γ has nontorsion abelianization, then $\Sigma^2\omega(\Gamma)$ is contained in a proper subspace of $\omega(\Gamma)$. However, we will show below that $\Sigma^2\omega(\Gamma)$ always has an interior point in $\omega^2(\Gamma)^h$.

Lemma 4.5. *For any group Γ, we have $\Sigma^2\omega(\Gamma) = \Sigma^2\mathbb{C}[\Gamma] \cap \omega(\Gamma)$.*

Proof. The inclusion $\Sigma^2\omega(\Gamma) \subset \Sigma^2\mathbb{C}[\Gamma] \cap \omega(\Gamma)$ is obvious. If $\sum_i a_i^* a_i \in \omega(\Gamma)$ with $a_i \in \mathbb{C}[\Gamma]$, then $\sum_i |\varepsilon(a_i)|^2 = 0$, and hence $\varepsilon(a_i) = 0$ for all i. This proves the converse inclusion. □

Remark 4.6. It turns out that we can always extend positive functionals φ on $\omega(\Gamma)$ to positive functionals $\overline{\varphi}$ on $\mathbb{C}[\Gamma]$, at least if we allow for an extension of the real closed field. Indeed, observe that

$$\big(\Sigma^2\mathbb{C}[\Gamma] + \omega(\Gamma)^h\big) \cap -\big(\Sigma^2\mathbb{C}[\Gamma] + \omega(\Gamma)^h\big) = \omega(\Gamma)^h,$$

which follows immediately from an application of the augmentation homomorphism ε. We can thus apply Theorem 2.3.

Lemma 4.7. *Let $\mathbf{R}$ be a real closed extension field of $\mathbb{R}$, and $\varphi\colon \omega(\Gamma) \to \mathbf{C}$ a completely positive $\mathbb{C}$-linear functional with $\varphi(a^*) = \overline{\varphi(a)}$ for all $a \in \omega(\Gamma)$. Then for all $s, h \in \Gamma$,*

$$\big|\varphi(c(s)^* c(h))\big| \le \frac{1}{\sqrt{2}} \cdot \big(\varphi(c(s)^* c(s)) + \varphi(c(h)^* c(h))\big),$$
$$\varphi(c(sh)^* c(sh)) \le 2 \cdot \big(\varphi(c(s)^* c(s)) + \varphi(c(h)^* c(h))\big).$$

Proof. The first inequality is an application of the Cauchy–Schwarz inequality (which is fulfilled by completely positive functionals) and the inequality $\lambda\mu \le (\lambda + \mu)^2/2$. For the second inequality, first apply the triangle identity from Corollary 3.9 to the equation

$$c(sh) = c(s) + (c(s)c(h) + c(h)),$$

together with the well-known inequality $(a+b)^2 \le 2(a^2 + b^2)$. Then use the easily verified identity $\|c(s)c(h) + c(h)\|^2 = \|c(h)\|^2$. □

Since $1 \notin \omega(\Gamma)$, we need to find a different candidate for an interior point of $\Sigma^2\omega(\Gamma)$.

Definition 4.8. Let $S \subset \Gamma$ be a finite symmetric set, that is, $S^{-1} = S$. We define the Laplace operator on S to be

$$\Delta(S) := |S| - \sum_{s \in S} s.$$

Remark 4.9. Note that for every finite symmetric set $S \subset \Gamma$,

$$\Delta(S) = \tfrac{1}{2} \cdot \sum_{s \in S} c(s)^* c(s) \in \Sigma^2 \omega(\Gamma).$$

Proposition 4.10. *Let Γ be a group generated by a finite symmetric set S. Then for any $b \in \omega^2(\Gamma)$, there exists a constant $C(b) \in \mathbb{R}$ such that for any real closed extension field $\mathbf{R}$ of $\mathbb{R}$ and any completely positive $\mathbb{C}$-linear functional $\varphi \colon \omega(\Gamma) \to \mathbf{C}$ with $\varphi(a^*) = \overline{\varphi(a)}$ for all $a \in \omega(\Gamma)$, one has*

$$|\varphi(b)| \leq C(b) \cdot \varphi(\Delta(S)).$$

Proof. Every element $b \in \omega^2(\Gamma)$ is a finite linear combination of $c(g)^* c(h)$ for $g, h \in \Gamma \setminus \{e\}$. This implies that $|\varphi(b)|$ is bounded by a constant times

$$\max\bigl\{\varphi(c(s)^* c(s)) \mid s \in S\bigr\},$$

using Lemma 4.7 several times. However,

$$\max\bigl\{\varphi(c(s)^* c(s)) \mid s \in S\bigr\} \leq 2 \cdot \varphi(\Delta(S))$$

follows from Remark 4.9. This proves the claim. □

Theorem 4.11. *Let Γ be a group with finite generating symmetric set S. Then for every $b \in \omega^2(\Gamma)^h$, there is a constant $C(b) \in \mathbb{R}$ such that*

$$C(b) \cdot \Delta(S) \pm b \in \Sigma^2 \omega(\Gamma).$$

In particular, $\Delta(S)$ is an inner point of the cone $\Sigma^2 \omega(\Gamma)$ in $\omega^2(\Gamma)^h$. If

$$H_1(\Gamma, \mathbb{C}) = 0,$$

it is an inner point in $\omega(\Gamma)^h$.

Proof. In view of Proposition 4.10, we find that

$$C(b) \cdot \Delta(S) \pm b$$

is nonnegative under each completely positive real-closed valued $\mathbb{C}$-linear functional φ on $\omega(\Gamma)$. In view of Theorem 3.12, this means that $C(b) \cdot \Delta(S) \pm b$ is a sum of hermitian squares in $\omega(\Gamma)$. Note that we use that $\omega(\Gamma)$ is real reduced. Finally note that $H_1(\Gamma, \mathbb{C}) = 0$ just means that $\omega^2(\Gamma) = \omega(\Gamma)$. □

5. Groups with Kazhdan's property (T)

We want to show that the constant $C(b)$ from Theorem 4.11 can be chosen as a fixed multiple of $\|b\|_1$, in case the group Γ has Kazhdan's property.

Definition 5.1. Let Γ be a group and $\pi : \Gamma \to U(H)$ a unitary representation on a Hilbert space H.

(1) A 1-cocycle with respect to the unitary representation π is a map $\delta\colon \Gamma \to H$ such that for all $g, h \in \Gamma$, we have $\delta(gh) = \pi(g)\delta(h) + \delta(g)$.

(2) A 1-cocycle $\delta\colon \Gamma \to H$ is called inner if $\delta(g) = \pi(g)\xi - \xi$ for some vector $\xi \in H$.

Definition 5.2. A group has Kazhdan's property (T) if every 1-cocycle with respect to every unitary representation is inner.

We will use several results on Kazhdan groups, which can be found, for example, in [Bekka et al. 2008]. It is well known that groups with Kazhdan's property (T) admit a finite generating set S, and that

$$\omega^2(\Gamma) = \omega(\Gamma)$$

holds. It is also known that for a fixed finite symmetric and generating set S in a Kazhdan group Γ, there is some $\epsilon > 0$ such that for any unitary representation $\pi : \Gamma \to U(H)$ without nonzero fixed vectors, one has

$$\langle \Delta(S)\xi, \xi \rangle \geq \epsilon \cdot \|\xi\|^2 \quad \text{for all} \quad \xi \in H.$$

Such ϵ is called a *Kazhdan constant* for S.

Let's revisit the standard GNS representation in the context of $\omega(\Gamma)$. Let

$$\varphi\colon \omega(\Gamma) \to \mathbb{C}$$

be a positive linear functional with $\varphi(a^*) = \overline{\varphi(a)}$. We associate to φ a Hilbert space as follows. We define on $\omega(\Gamma)$ a positive semidefinite sesquilinear form

$$\langle a, b \rangle_\varphi := \varphi(b^* a)$$

and set $\|a\|_\varphi := \langle a, a \rangle_\varphi^{1/2}$. Let $N(\varphi) := \{a \in \omega(\Gamma) \mid \|a\|_\varphi = 0\}$ and define $L^2(\omega(\Gamma), \varphi)$ to be the metric completion of $\omega(\Gamma)/N(\varphi)$ with respect to $\| \cdot \|_\varphi$. We denote the image of $c(g)$ in $L^2(\omega(\Gamma), \varphi)$ by $\delta(g)$ and denote by $\delta(\Gamma)$ their complex linear span, which is dense by definition of $L^2(\omega(\Gamma), \varphi))$. It is standard that the left-multiplication of $\omega(\Gamma)$ on itself extends to a homomorphism $\pi^\varphi\colon \omega(\Gamma) \to D(\delta(\Gamma))$, where $D(\delta(\Gamma))$ denotes the algebra of densely defined linear operators mapping $\delta(\Gamma)$ into itself. Indeed, if $a \in N(\varphi)$ and $b \in \omega(\Gamma)$, then $ba \in N(\varphi)$ since

$$\varphi((ba)^* ba) = \varphi(a^* b^* ba) \leq \|b\|_1^2 \cdot \varphi(a^* a),$$

by Remark 4.1. Note that

$$\pi^{\varphi}(c(g))\delta(h) = \delta(gh) - \delta(g) - \delta(h).$$

Now we define a unitary representation π_φ of Γ on $L^2(\omega(\Gamma), \varphi)$ by the rule

$$\pi_\varphi(g) := \pi^{\varphi}(c(g)) + 1_{L^2(\omega(\Gamma),\varphi)}.$$

Lemma 5.3. *If $\omega^2(\Gamma) = \omega(\Gamma)$, then the representation π_φ has no fixed vectors.*

Proof. Assume $\eta \in L^2(\omega(\Gamma), \varphi)$ is a fixed vector. By definition of π_φ, this means $\pi^{\varphi}(c(g))\eta = 0$ for all $g \in \Gamma$. Hence,

$$0 = \left\langle \pi^{\varphi}(c(g^{-1}))\eta, \delta(h)\right\rangle_{\varphi} = \langle \eta, \delta(gh) - \delta(g) - \delta(h)\rangle_{\varphi}.$$

Since $c(g)c(h) = c(gh) - c(g) - c(h)$, the vectors $\delta(gh) - \delta(g) - \delta(h)$ span the image of $\omega^2(\Gamma)$ in $\delta(\Gamma)$, and hence $\delta(\Gamma)$, since $\omega^2(\Gamma) = \omega(\Gamma)$. □

Note that the map $g \mapsto \delta(g)$ satisfies

$$\delta(gh) = \pi_\varphi(g)\delta(h) + \delta(g),$$

and hence defines a 1-cocycle with respect to the representation π_φ. If Γ is a Kazhdan group, then there exists $\Omega \in L^2(\omega(\Gamma), \varphi)$ such that

$$\delta(g) = \pi_\varphi(g)\Omega - \Omega.$$

Proposition 5.4. *Let Γ be a Kazhdan group with finite symmetric generating set S and Kazhdan constant $\epsilon > 0$. Then for every nonzero $b \in \omega(\Gamma)^h$, every real closed extension field $\mathbf{R}$ of $\mathbb{R}$, and every positive nontrivial $\mathbb{C}$-linear functional $\varphi\colon \omega(\Gamma) \to \mathbf{C}$ with $\varphi(a^*) = \overline{\varphi(a)}$, one has*

$$\epsilon \cdot \varphi(b) < 2\|b\|_1 \cdot \varphi(\Delta(S)).$$

Proof. Let us first assume $\mathbf{R} = \mathbb{R}$ and $\mathbf{C} = \mathbb{C}$. We do the GNS construction as just described, and get some $\Omega \in L^2(\omega(\Gamma), \varphi)$ with $\delta(g) = \pi_\varphi(g)\Omega - \Omega$, for all $g \in \Gamma$. We set

$$\overline{\varphi}\colon \mathbb{C}[\Gamma] \to \mathbb{C}, \quad \overline{\varphi}(a) = \langle \pi_\varphi(a)\Omega, \Omega\rangle$$

and compute

$$\bar{\varphi}(c(h)^* c(g)) = \left\langle \pi_\varphi(c(g))\Omega, \pi_\varphi(c(h))\Omega\right\rangle_{\varphi} = \langle \delta(g), \delta(h)\rangle_{\varphi} = \varphi(c(h)^* c(g)).$$

This shows that $\bar{\varphi}$ and φ agree on $\omega^2(\Gamma)$, and hence on $\omega(\Gamma)$. If we now do the standard GNS construction with respect to $\overline{\varphi}$, we see that there is a natural Γ-equivariant identification of $L^2(\mathbb{C}[\Gamma], \bar{\varphi})$ and $L^2(\omega(\Gamma), \varphi)$. Since the representation π_φ has no fixed vectors, we get

$$\varphi(\Delta(S)) = \overline{\varphi}(\Delta(S)) = \langle \Delta(S)1, 1\rangle_{\overline{\varphi}} \geq \epsilon \cdot \overline{\varphi}(1).$$

Since $\overline{\varphi}$ is positive and nontrivial, it follows from Remark 4.2 that $\overline{\varphi}(1) > 0$. So finally, again using Remark 4.2, we find

$$\epsilon \cdot \varphi(b) = \epsilon \cdot \overline{\varphi}(b) \leq \epsilon \cdot \|b\|_1 \cdot \overline{\varphi}(1) < 2\|b\|_1 \cdot \varphi(\Delta(S)),$$

the desired result.

Now let $\mathbf{R}$ be arbitrary, and let $\varphi\colon \omega(\Gamma) \to \mathbf{C}$ be positive and nontrivial. From Theorem 4.11 it follows that $\varphi(\Delta(S)) > 0$. So we can assume without loss of generality that $\varphi(\Delta(S)) = 1$. Again from Theorem 4.11, we see that φ now only takes values in $\mathcal{O}[i]$, where $\mathcal{O}$ is the convex hull of $\mathbb{R}$ in $\mathbf{R}$. It is well known that $\mathcal{O}$ is a valuation ring in $\mathbf{R}$ with maximal ideal $\mathfrak{m}$, and that $\mathcal{O}/\mathfrak{m} = \mathbb{R}$ (see for example [Prestel and Delzell 2001], especially the appendix on valued fields). The residue map $\mathcal{O} \to \mathcal{O}/\mathfrak{m}$ maps nonnegative elements to nonnegative elements. So if we compose φ with the residue map on $\mathcal{O}[i]$, we get a positive linear functional to $\mathbb{C}$. Since we know that the desired strict inequality holds now, it was already valid for φ. □

Theorem 5.5. *Let Γ be a group with finite generating symmetric set S. Consider the statements*:

(1) *Γ has Kazhdan's property (T).*

(2) *$\Delta(S)$ is an algebraic interior point of $\Sigma^2\omega(\Gamma)$ in the ℓ^1-metric of $\omega(\Gamma)^h$. More precisely, there exists a constant $\epsilon > 0$ such that for every $b \in \omega(\Gamma)^h$ with $\|b\|_1 = 1$, we have*

$$\Delta(S) + \epsilon \cdot b \in \Sigma^2\omega(\Gamma).$$

The following implications hold: (1) *implies* (2), *and* (2) *implies* (1) *under the additional assumption $H_2(\Gamma, \mathbb{C}) = 0$.*

Proof. The implication (1) ⇒ (2) is a direct consequence of Theorem 3.11 and Proposition 5.4. Let us now prove (2) ⇒ (1) under the additional assumption $H_2(\Gamma, \mathbb{C}) = 0$. We first prove two lemmas.

Lemma 5.6. *Let Γ be a group. There is an exact sequence as follows*:

$$0 \to H_2(\Gamma, \mathbb{C}) \to \omega(\Gamma) \otimes_{\mathbb{C}[\Gamma]} \omega(\Gamma) \to \omega(\Gamma) \to H_1(\Gamma, \mathbb{C}) \to 0.$$

Proof. It is well known that

$$\omega(\Gamma)/\omega^2(\Gamma) = \Gamma_{ab} \otimes_{\mathbb{Z}} \mathbb{C} = H_1(\Gamma, \mathbb{C}).$$

This shows exactness at $\omega(\Gamma)$, and it remains to show exactness at $\omega(\Gamma) \otimes_{\mathbb{C}[\Gamma]} \omega(\Gamma)$. Since $0 \to \omega(\Gamma) \to \mathbb{C}[\Gamma] \to \mathbb{C} \to 0$ is exact, we have

$$H_2(\Gamma, \mathbb{C}) = H_1(\Gamma, \omega(\Gamma)) = \operatorname{Tor}_1^{\mathbb{C}[\Gamma]}(\mathbb{C}, \omega(\Gamma)).$$

Again, we get an exact sequence

$$\mathrm{Tor}_1^{\mathbb{C}[\Gamma]}(\mathbb{C}[\Gamma], \omega(\Gamma)) \to \mathrm{Tor}_1^{\mathbb{C}[\Gamma]}(\mathbb{C}, \omega(\Gamma)) \to \omega(\Gamma) \otimes_{\mathbb{C}[\Gamma]} \omega(\Gamma) \to \mathbb{C}[\Gamma] \otimes_{\mathbb{C}[\Gamma]} \omega(\Gamma).$$

This finishes the proof, since $\mathrm{Tor}_1^{\mathbb{C}[\Gamma]}(\mathbb{C}[\Gamma], \omega(\Gamma)) = 0$. □

If $\Delta(S)$ is an algebraic interior point of $\Sigma^2\omega(\Gamma)$ in $\omega(\Gamma)^h$, then $\omega(\Gamma) = \omega^2(\Gamma)$, that is, $H_1(\Gamma, \mathbb{C}) = 0$. Hence $H_2(\Gamma, \mathbb{C}) = 0$ ensures that the natural map

$$\omega(\Gamma) \otimes_{\mathbb{C}[\Gamma]} \omega(\Gamma) \to \omega(\Gamma)$$

is an isomorphism. This is what we are going to use.

Lemma 5.7. *Let $\pi : \Gamma \to U(H)$ be a unitary representation and let $\delta : \Gamma \to H$ be a 1-cocycle with respect to H. Then*

$$\varphi(c(h)^* \otimes c(g)) := \langle \delta(g), \delta(h) \rangle$$

yields a well-defined positive linear functional on $\omega(\Gamma) = \omega(\Gamma) \otimes_{\mathbb{C}[\Gamma]} \omega(\Gamma)$.

Proof. It is clear that $(c(h)^*, c(g)) \mapsto \langle \delta(g), \delta(h) \rangle$ defines a bilinear map on $\omega(\Gamma)$, that is, a linear map $\varphi' : \omega(\Gamma) \otimes_{\mathbb{C}} \omega(\Gamma) \to \mathbb{C}$. We show that this map passes to $\omega(\Gamma) \otimes_{\mathbb{C}[\Gamma]} \omega(\Gamma)$. Let $g, h, k \in \Gamma$; then

$$\begin{aligned}
\varphi'(c(h)^* k \otimes c(g)) &= \varphi'(c(h^{-1})k \otimes c(g)) \\
&= \varphi'\big((c(h^{-1}k) - c(k)) \otimes c(g)\big) \\
&= \varphi'(c(k^{-1}h)^* \otimes c(g)) - \varphi'(c(k^{-1})^* \otimes c(g)) \\
&= \langle \delta(g), \delta(k^{-1}h) \rangle - \langle \delta(g), \delta(k^{-1}) \rangle \\
&= \langle \delta(g), \pi(k^{-1})\delta(h) \rangle \\
&= \langle \pi(k)\delta(g), \delta(h) \rangle \\
&= \langle \delta(kg), \delta(h) \rangle - \langle \delta(k), \delta(h) \rangle \\
&= \varphi'(c(h)^* \otimes c(kg)) - \varphi'(c(h)^* \otimes c(k)) \\
&= \varphi'(c(h^*) \otimes kc(g)).
\end{aligned}$$

We can now understand φ' as a linear map on $\omega(\Gamma)$ via the above isomorphism to $\omega(\Gamma) \otimes_{\mathbb{C}[\Gamma]} \omega(\Gamma)$. Since a^*a corresponds to $a^* \otimes a$, one easily checks that φ is positive on $\omega(\Gamma)$. □

We continue with the proof of Theorem 5.5. Condition (2) in Theorem 5.5 and Lemma 5.7 imply that any 1-cocycle with respect to any unitary representation is bounded. This is well known to imply Kazhdan's property (T) for Γ. □

Remark 5.8. It is not clear whether the condition $H_2(\Gamma, \mathbb{C}) = 0$ is necessary.

There is an analogue of the implication (1) $\Rightarrow$ (2) in Theorem 5.5 in

$$\ell^1\Gamma := \left\{ \sum_{g\in\Gamma} a_g g \;\Big|\; \sum_{g\in\Gamma} |a_g| < \infty \right\}.$$

We set $\omega^1\Gamma := \overline{\omega(\Gamma)}^{\|\cdot\|_1}$ and define

$$\Sigma^{2,1}\omega(\Gamma) := \left\{ \sum_{i=1}^{\infty} a_i^* a_i \;\Big|\; a_i \in \omega[\Gamma],\ \sum_{i=1}^{\infty} \|a_i\|_1^2 < \infty \right\}$$

and

$$\Sigma^{2,1}\ell^1(\Gamma) := \left\{ \sum_{i=1}^{\infty} a_i^* a_i \;\Big|\; a_i \in \ell^1[\Gamma],\ \sum_{i=1}^{\infty} \|a_i\|_1^2 < \infty \right\}.$$

We note that $\|a\|_1 - a \in \Sigma^{2,1}\ell^1(\Gamma)$ for every hermitian element $a \in \ell^1\Gamma$. Indeed,

$$\|a\|_1 - a = \sum_{g\in G} 2|a_g| - a_g g - \bar{a}_g g^{-1} = \sum_{g\in G} \Big(|a_g|^{1/2} - \frac{a_g}{|a_g|^{1/2}} g\Big)^* \Big(|a_g|^{1/2} - \frac{a_g}{|a_g|^{1/2}} g\Big).$$

Hence, for $\varphi\colon \ell^1\Gamma \to \mathbf{C}$, $\mathbb{C}$-linear and positive on $\Sigma^{2,1}\ell^1\Gamma$, $|\varphi(a)| \le \|a\|_1$ for all $a \in \ell^1\Gamma$. A priori, there is no reason to assume that $\Sigma^{2,1}\ell^1\Gamma$ or $\Sigma^{2,1}\omega^1\Gamma$ are closed or have nontrivial interior. Nevertheless, our result shows:

Corollary 5.9. *Let Γ be a Kazhdan group with finite generating symmetric set S and Kazhdan constant ϵ. Then for every $b \in \omega^1(\Gamma)^h$ with $\|b\|_1 = 1$, we have*

$$\Delta(S) + \varepsilon \cdot b \in \Sigma^{2,1}\omega(\Gamma).$$

6. Group algebras of free groups

In this section, let $\Gamma = F_n$ be the free group on n generators $g_1, \ldots, g_n$ or $\Gamma = F_\infty$. Schmüdgen (private communication, 2011) has proven that an element from the group algebra $\mathbb{C}[\Gamma]$ that is nonnegative under each finite-dimensional $*$-representation is a sum of squares. We demonstrate how his result can be reproved with our real closed separation approach. The main idea of our proof is the same as in Schmüdgen's work. However, instead of a partial GNS construction, we use a full GNS construction, but over a general real closed field. We then reduce to the standard real numbers by Tarski's transfer principle.

Theorem 6.1 (Schmüdgen). *Let $\Gamma = F_n$ be the free group on n generators. If $b \in \mathbb{C}[\Gamma]^h$ is mapped to a positive semidefinite matrix under each finite-dimensional $*$-representation of $\mathbb{C}[\Gamma]$, then $b \in \Sigma^2\mathbb{C}[\Gamma]$.*

Proof. Assume that $b \notin \Sigma^2\mathbb{C}[\Gamma]$. By Theorem 3.11, there is a real closed extension field $\mathbf{R}$ of $\mathbb{R}$ and a completely positive $\mathbb{C}$-linear functional $\varphi\colon \mathbb{C}[\Gamma] \to \mathbf{C}$ with $\varphi(a^*) = \overline{\varphi(a)}$, such that $\varphi(b) < 0$. By Lemma 3.6, the canonical $\mathbf{C}$-linear extension

of φ to $A = \mathbf{C} \otimes_{\mathbb{C}} \mathbb{C}[\Gamma]$ is still positive, and we denote it again by φ. We apply the usual GNS construction to A. We note that

$$N = \{a \in A \mid \varphi(a^* a) = 0\}$$

is a $*$-subspace of the $\mathbf{C}$-vector space A, which follows from the Cauchy–Schwarz inequality, as shown in Corollary 3.9. We denote the quotient space A/N by H, and note that

$$\langle a + N, c + N \rangle_\varphi := \varphi(c^* a)$$

is a well-defined and positive definite $\mathbf{C}$-valued sesquilinear form on H. We also note that left-multiplication from $\mathbb{C}[\Gamma]$ on A is well-defined on H, as explained in Section 5. So we have a $\mathbb{C}$-linear $*$-representation

$$\pi : \mathbb{C}[\Gamma] \to \mathcal{L}(H)$$

with $\langle \pi(b)\xi, \xi \rangle_\varphi < 0$, where $\xi = 1 + N$.

Now let H' be a finite-dimensional $*$-subspace of H, containing the residue classes of all words in the generators g_i of length at most d, where d is the maximal word length in b. We can choose an orthonormal basis $v_1, \ldots v_m$ of H', using the usual Gram–Schmidt procedure over $\mathbf{C}$. So there is an orthogonal projection map $p\colon H \to H'$, defined as

$$p\colon h \mapsto \sum_{i=1}^{m} \langle h, v_i \rangle v_i.$$

Define

$$M_i := p \circ \pi(g_i) \in \mathcal{L}(H').$$

It is easy to see that all M_i are contractions; thus the linear operators $\sqrt{1 - M_i^* M_i}$ and $\sqrt{1 - M_i M_i^*}$ exist on H'. Using Choi's matrix trick [1980, Theorem 7], we define

$$U_i := \begin{pmatrix} M_i & \sqrt{I - M_i M_i^*} \\ \sqrt{I - M_i^* M_i} & -M_i^* \end{pmatrix} \in \mathcal{L}(H' \oplus H').$$

The U_i are checked to be unitary operators, and thus yield a $\mathbb{C}$-linear $*$-representation $\widetilde{\pi}$ of $\mathbb{C}[\Gamma]$ on $H' \oplus H'$. Since the residue classes of all words occurring in b belong to H', and by the definition of the U_i, we find

$$\langle \widetilde{\pi}(b)\xi', \xi' \rangle_{H' \oplus H'} = \langle \pi(b)\xi, \xi \rangle_\varphi < 0,$$

where $\xi' = (\xi, 0)$. Now finally, since $H' \oplus H'$ is finite-dimensional, the existence of such a representation over $\mathbf{C}$ implies the existence over $\mathbb{C}$, by Tarski's transfer principle. This finishes the proof. □

Remark 6.2. The proof becomes even simpler when considering the $*$-algebra of polynomials in noncommuting variables $\mathbb{C}\langle y_1, \ldots, y_n, z_1, \ldots, z_n \rangle$ with $y_i^* = z_i$, or $\mathbb{C}\langle z_1, \ldots, z_n \rangle$ with $z_i^* = z_i$, instead of the group algebra of a free group. The reason

is that one is not forced to make the matrices M_i unitary (only hermitian in the second case). So Theorem 6.1 also holds for these polynomial algebras. This was first proven by Helton [2002].

It is an interesting problem to study the class of groups Γ for which positivity of $a \in \mathbb{C}[\Gamma]$ in every finite-dimensional unitary representation implies that $a \in \Sigma^2\mathbb{C}[\Gamma]$. It is clear that in order for an analogous argument to work, Γ has to be residually finite-dimensional in a very strong sense. Residual finite-dimensionality of a group means that every unitary representation on a Hilbert space can be approximated in the Fell topology by finite-dimensional representations; see [Brown and Ozawa 2008] for details. If — more generally — every generalized unitary representation of Γ on a Hilbert space can be approximated on finitely many vectors by generalized finite-dimensional unitary representations, then everything works. With additional work, this can be carried out for virtually free groups (Schmüdgen, private communication).

Deep results of Scheiderer [2006] imply that the conclusion holds for $\mathbb{Z}^2$. By a classical result [Rudin 1963], however, the group $\mathbb{Z}^3$ does not satisfy the desired conclusion, and the same holds for every group containing $\mathbb{Z}^3$. This is also implied by seminal work of Scheiderer [2000, Theorem 6.2], who showed that the existence of positive elements that are not sums of squares under general assumptions in dimension ≥ 3.

This shows that the theory of generalized unitary representations is fundamentally different and new pathologies occur.

An intriguing and possibly manageable case is that of surface groups. Lubotzky and Shalom [2004] showed that surface groups are residually finite-dimensional. It is quite possible that their methods extend and lead to a resolution of the case of surface groups.

Conjecture 6.3. Let Γ be a surface group. Every element $a \in \mathbb{C}[\Gamma]^h$ that is positive semidefinite in every finite-dimensional unitary representation lies in $\Sigma^2\mathbb{C}[\Gamma]$.

Similar questions can be studied if one allows the unitary representations to be infinite-dimensional. Again, the only known obstruction is $\mathbb{Z}^3 \subset \Gamma$.

References

[Barvinok 2002] A. Barvinok, *A course in convexity*, Graduate Studies in Mathematics **54**, American Mathematical Society, Providence, RI, 2002. MR 2003j:52001 Zbl 1014.52001

[Bekka et al. 2008] B. Bekka, P. de la Harpe, and A. Valette, *Kazhdan's property (T)*, New Mathematical Monographs **11**, Cambridge University Press, Cambridge, 2008. MR 2009i:22001 Zbl 1146.22009

[Brown and Ozawa 2008] N. P. Brown and N. Ozawa, *C^*-algebras and finite-dimensional approximations*, Graduate Studies in Mathematics **88**, American Mathematical Society, Providence, RI, 2008. MR 2009h:46101 Zbl 1160.46001

[Choi 1980] M. D. Choi, "The full C^*-algebra of the free group on two generators", *Pacific J. Math.* **87**:1 (1980), 41–48. MR 82b:46069 Zbl 0463.46047

[Cimprič 2009] J. Cimprič, "A representation theorem for Archimedean quadratic modules on $*$-rings", *Canad. Math. Bull.* **52**:1 (2009), 39–52. MR 2010c:46154 Zbl 1173.16024

[Cimprič et al. 2011] J. Cimprič, M. Marshall, and T. Netzer, "Closures of quadratic modules", *Israel J. Math.* **183** (2011), 445–474. MR 2012e:13047 Zbl 1235.13021

[Helton 2002] J. W. Helton, "'Positive' noncommutative polynomials are sums of squares", *Ann. of Math.* (2) **156**:2 (2002), 675–694. MR 2003k:12002 Zbl 1033.12001

[Lubotzky and Shalom 2004] A. Lubotzky and Y. Shalom, "Finite representations in the unitary dual and Ramanujan groups", pp. 173–189 in *Discrete geometric analysis* (Sendai, 2002), edited by M. Kotani et al., Contemp. Math. **347**, Amer. Math. Soc., Providence, RI, 2004. MR 2005e:22011 Zbl 1080.22006

[Powers and Scheiderer 2001] V. Powers and C. Scheiderer, "The moment problem for non-compact semialgebraic sets", *Adv. Geom.* **1**:1 (2001), 71–88. MR 2002c:14086 Zbl 0984.44012

[Prestel and Delzell 2001] A. Prestel and C. N. Delzell, *Positive polynomials: from Hilbert's 17th problem to real algebra*, Springer, Berlin, 2001. MR 2002k:13044 Zbl 0987.13016

[Rudin 1963] W. Rudin, "The extension problem for positive-definite functions", *Illinois J. Math.* **7** (1963), 532–539. MR 27 #1779 Zbl 0114.31003

[Scheiderer 2000] C. Scheiderer, "Sums of squares of regular functions on real algebraic varieties", *Trans. Amer. Math. Soc.* **352**:3 (2000), 1039–1069. MR 2000j:14090 Zbl 0941.14024

[Scheiderer 2006] C. Scheiderer, "Sums of squares on real algebraic surfaces", *Manuscripta Math.* **119**:4 (2006), 395–410. MR 2006m:14079 Zbl 1120.14047

[Schmüdgen 1991] K. Schmüdgen, "The K-moment problem for compact semi-algebraic sets", *Math. Ann.* **289**:2 (1991), 203–206. MR 92b:44011 Zbl 0744.44008

[Schmüdgen 2009] K. Schmüdgen, "Noncommutative real algebraic geometry: some basic concepts and first ideas", pp. 325–350 in *Emerging applications of algebraic geometry*, edited by M. Putinar and S. Sullivant, IMA Vol. Math. Appl. **149**, Springer, New York, 2009. MR 2010m:13036 Zbl 1158.13313

Received January 24, 2012. Revised August 8, 2012.

TIM NETZER
MATHEMATISCHES INSTITUT
UNIVERSITY OF LEIPZIG
PF 10 09 20
D-04009 LEIPZIG
GERMANY
netzer@math.uni-leipzig.de

ANDREAS THOM
MATHEMATISCHES INSTITUT
UNIVERSITY OF LEIPZIG
PF 10 09 20
D-04009 LEIPZIG
GERMANY
thom@math.uni-leipzig.de

PACIFIC JOURNAL OF MATHEMATICS
Vol. 263, No. 2, 2013
dx.doi.org/10.2140/pjm.2013.263.453

UNIQUENESS THEOREM FOR ORDINARY DIFFERENTIAL EQUATIONS WITH HÖLDER CONTINUITY

YIFEI PAN, MEI WANG AND YU YAN

We study ordinary differential equations of the type $u^{(n)}(t) = f(u(t))$, with initial conditions $u(0) = u'(0) = \cdots = u^{(m-1)}(0) = 0$ and $u^{(m)}(0) \neq 0$, where $m \geq n$; no additional assumption is made on f. We establish some uniqueness results and show that f is always Hölder continuous.

1. Introduction

The question of finding criteria for the uniqueness of solutions has been a constant theme in the study of ordinary differential equations for a very long time, and a wealth of results have been established. The one most quoted in textbooks is perhaps the Lipschitz uniqueness theorem, which states that in the equation $y^{(n)}(x) = f(x, y, y', \ldots, y^{(n-1)})$, if the function $f(x, z_1, z_2, \ldots, z_n)$ is Lipschitz continuous with respect to $z_1, z_2, \ldots, z_n$, then the initial value problem has a unique local solution. Generally speaking, to ensure the uniqueness of solutions to an ODE, we need to assume some condition on the function f besides continuity, the Lipschitz condition being one example. Most of the research in this topic has been devoted to finding the appropriate condition, and there are many nice results, such as the classical theorems by Peano, Osgood, Montel and Tonelli, and Nagumo. An extensive and systematic treatment of the available results is provided in [Agarwal and Lakshmikantham 1993].

In this paper, we approach the uniqueness problem from a different perspective and relate it to the unique continuation problem. We study autonomous ODEs of the type $u^{(n)}(t) = f(u(t))$, where $u \in C^\infty([0, 1])$ and no additional assumption is made on the function f.

If we assume the initial conditions $u(0) = u'(0) = \cdots = u^{(n-1)}(0) = 0$, the solution is not unique. The following is a trivial example.

Example 1. $u(t) = t^3$ satisfies $u''(t) = 6u^{1/3}$ and $u(0) = u'(0) = 0$. Another solution to this initial value problem is $u \equiv 0$.

The research of Yu Yan was partially supported by the Scholar in Residence program at Indiana University–Purdue University Fort Wayne.
MSC2010: 34A12.
Keywords: uniqueness of solutions, ordinary differential equations.

It is no surprise that uniqueness fails in this example, because the function $f(u) = 6u^{1/3}$ has fairly strong singularity at 0. From another perspective, this example shows that if a solution and its derivatives up to order $n-1$ all vanish at 0, it is not guaranteed to be the zero function. On the other hand, even if all its derivatives vanish at 0, the solution still may not be identically 0.

Example 2. The function

$$u(t) = \begin{cases} e^{-1/t} & 0 < t \le 1, \\ 0 & t = 0 \end{cases}$$

is in $C^\infty([0, 1])$, and

$$u^{(k)}(0) = 0 \quad \text{for all} \quad k \in \mathbb{N}. \tag{1}$$

Let

$$f(s) = \begin{cases} s(\ln s)^2 & s > 0, \\ 0 & s = 0. \end{cases}$$

Then $u(t)$ satisfies the equation

$$u' = f(u).$$

However, this equation has another solution, $u \equiv 0$, which also satisfies (1).

This function $u(t)$ is also a classical example in the study of the unique continuation problem, which asks when we can conclude that a function is locally identically zero if its derivatives all vanish at a point. Here is one result in this line:

Theorem 1.1 [Pan and Wang 2008]. *Let* $g(x) \in C^\infty([a, b])$, $0 \in [a, b]$, *and*

$$|g^{(n)}(x)| \le C \sum_{k=0}^{n-1} \frac{|g^{(k)}(x)|}{|x|^{n-k}}, \qquad x \in [a, b] \tag{2}$$

for some constant C *and some* $n \ge 1$. *Then*

$$g^{(k)}(0) = 0 \quad \textit{for all} \quad k \ge 0$$

implies

$$g \equiv 0 \quad \textit{on} \quad [a, b].$$

The order of singularity of $|x|$ at 0 in (2), that is, $n-k$, is sharp, as one example in [Pan and Wang 2008] shows. This theorem is crucial to the proof of our main theorem below.

The previous two examples suggest that to guarantee uniqueness near 0, the solution needs to vanish to sufficiently high order, but not to the infinite order. So we assume that it satisfies the initial conditions $u(0) = u'(0) = \cdots = u^{(m-1)}(0) = 0$ and $u^{(m)}(0) = a \ne 0$, where $m \ge n$. That is, the order of the lowest nonvanishing derivative of u at 0 is no less than the order of the equation. From the equation

it is not difficult to see that f is differentiable away from 0; however, it is not differentiable at 0, as shown by Example 1 with $m = 3$.

Due to the lack of information about the regularity of f, the available uniqueness theory no longer applies to this type of equation. We will show that because u has sufficiently high vanishing order at 0, such solutions are unique near 0. Specifically, we have the following result.

Theorem 1.2. *Let $u(t) \in C^\infty([0, 1))$ be a solution of the differential equation*

$$u^{(n)}(t) = f(u(t)), \tag{3}$$

where $n \geq 1$ and f is a function. Assume that u satisfies

$$u(0) = u'(0) = \cdots = u^{(m-1)}(0) = 0 \quad \text{and} \quad u^{(m)}(0) = a \neq 0, \tag{4}$$

with $m \geq n$. Then such a solution $u(t)$ is unique for t near 0.

The proof of Theorem 1.2 is carried out in two steps. First, we show the following result concerning the derivatives of u at 0.

Lemma 1.3. *Let $u(t)$ be a solution that satisfies Equations* (3) *and* (4). *The derivative of u at 0 of any order equal to or higher than m, that is, $u^{(k)}(0)$ for any $k \geq m$, depends only on m, n, and the behavior of the function f near 0.*

In the second step, suppose there are two solutions u and v, both satisfying (3) and (4); then by Lemma 1.3, the function $u(t) - v(t)$ and all its derivatives vanish at 0. Making use of Theorem 1.1, we can show that $u - v \equiv 0$.

Typically, for an n-th order ODE, we need only n initial conditions. Theorem 1.1 shows that in some sense, the lack of information about f can be compensated by assuming additional derivative information at the initial point.

Interestingly, it turns out that the solution is unique as long as the vanishing order is no less than the order of the equation, but the actual vanishing order and the value of the lowest nonzero derivative are not essential.

Theorem 1.4. *Suppose u_1 and u_2 are two solutions of* (3) *that satisfy*

$$u_1(0) = u_1'(0) = \cdots = u_1^{(m-1)}(0) = 0, \qquad u_1^{(m)}(0) = a \neq 0 \tag{5}$$

and

$$u_2(0) = u_2'(0) = \cdots = u_2^{(l-1)}(0) = 0, \qquad u_2^{(l)}(0) = b \neq 0, \tag{6}$$

where $m, l \geq n$. Then $m = l$, $a = b$, and $u_1 \equiv u_2$ for small t.

The proof of Lemma 1.3 will be given in Section 2, and the proofs of Theorems 1.2 and 1.4 will be given in Section 3.

Naturally, we would like to ask if the result still holds if one of the solutions in Theorem 1.4 has vanishing order lower than n, the order of the equation.

Conjecture. *Suppose* (3) *has a solution* $u(t)$ *that satisfies* (4) *with* $m \geq n$. *Then it cannot possess another solution* $v(t)$ *that satisfies initial conditions*

$$v(0) = v'(0) = \cdots = v^{(l-1)}(0) = 0 \quad \textit{and} \quad v^{(l)}(0) = b \neq 0,$$

where $l < n$.

We can show that this conjecture is true if $m = n + 1$ or l and m are relatively prime. However, there are some difficulties in the general case and we have not been able to prove the full conjecture.

Although in Theorems 1.2 and 1.4 we do not need to make any assumptions about the function f, we can actually obtain interesting information about it. Suppose there is a function $u(t)$ that satisfies Condition (4); then as shown in Section 2, locally t can be expressed as a function of u, and therefore we can express $u^{(n)}(t)$ locally as a function f of u, so $u^{(n)}(t) = f(u)$. The next theorem shows that the function f is Hölder continuous in an interval $[0, \delta]$ for small $\delta > 0$.

Theorem 1.5. *Suppose a function* $u(t)$ *satisfies Condition* (4); *then Equation* (3) *holds for some function* f, *where* $m \geq n$ *and there is a constant* $\delta > 0$ *such that* f *is uniformly Hölder continuous in the interval* $[0, \delta]$.

This theorem is proved after Theorem 1.4 in Section 3.

A summary of Theorems 1.2 and 1.5 is that any smooth function of finite order vanishing at 0 is a unique local solution of a differential equation in the form of Equation (3), where f is differentiable in the interior and uniformly Hölder continuous up to the boundary.

In Theorems 1.2 and 1.4, the high-order vanishing condition (4) allows us to obtain uniqueness results without any extra assumption on f. This phenomenon is only found in autonomous equations like (3). For general ODEs of the form

$$\frac{d^n u}{dt^n} = f\Big(t, u, \frac{du}{dt}, \ldots, \frac{d^{n-1}u}{dt^{n-1}}\Big),$$

such results cannot be expected because there is more than one expression for f. For example:

Example 3. Let $u(t) = t^4$. It satisfies the initial conditions

$$u(0) = u'(0) = u''(0) = u^{(3)}(0) = 0 \quad \text{and} \quad u^{(4)}(0) = 24.$$

Its derivatives are

$$u'(t) = 4t^3 \quad \text{and} \quad u''(t) = 12t^2.$$

We can express u'' as a function f of u and u' in different ways, such as

$$u''(t) = \frac{192u^2}{(u')^2}, \qquad u''(t) = \frac{3(u')^2}{4u}, \qquad u''(t) = 12\big(\tfrac{1}{4}uu'\big)^{2/7},$$

or

$$u''(t) = 12u^{1/4}\left(\tfrac{1}{4}u'\right)^{1/3}.$$

In the first two equations f is not continuous at the origin, while in the last two equations f is Hölder continuous at the origin. This simple example shows that the function f can be expressed in various ways and that in order to study the uniqueness, we need to impose very specific assumptions on f.

Our work was motivated by that of Li and Nirenberg [2006], who studied a similar second-order PDE: $\Delta u = f(u)$, where $u = u(t, x) \in C^\infty(\mathbb{R}^{k+1})$ has a nonvanishing partial derivative at 0 that can be expressed in the form $u(t, x) = at^m + O(t^{m+1})$, $a \neq 0$, $t \in \mathbb{R}$, and $x \in \mathbb{R}^k$. They showed that if two solutions u and v satisfy $u \geq v$, then $u \equiv v$. Theorem 1.2 can be viewed as an improvement of their result in the one-dimensional case to arbitrary order and without the comparison condition $u \geq v$.

2. The proof of Lemma 1.3

Without loss of generality, we can assume that $a > 0$.

First, we show that a, the m-th derivative of u at 0, only depends on m, n, and the function f.

Define

$$\tilde{x} = \left(\frac{u}{a}\right)^{1/m}. \tag{7}$$

Then

$$\tilde{x} = (t^m + O(t^{m+1}))^{1/m} = t(1 + O(t)).$$

This implies that

$$\frac{\tilde{x}}{t} \to 1 \quad \text{as} \quad t \to 0. \tag{8}$$

We can also write $u = a\tilde{x}^m$. Taking the derivative with respect to t, we get

$$\begin{aligned} \frac{du}{dt} &= am\tilde{x}^{m-1}\frac{d\tilde{x}}{dt}, \\ amt^{m-1} + O(t^m) &= am\tilde{x}^{m-1}\frac{d\tilde{x}}{dt}, \\ \frac{t^{m-1}}{\tilde{x}^{m-1}} + O\left(\frac{t^m}{\tilde{x}^{m-1}}\right) &= \frac{d\tilde{x}}{dt}. \end{aligned}$$

In the second equation above and in the analysis that follows, we formally differentiate the Taylor expansion with the big-O notation. A detailed discussion of this differentiation is provided in the Appendix.

In light of (8), it follows that

$$\frac{d\tilde{x}}{dt}|_{t=0} = 1. \tag{9}$$

By the inverse function theorem, t can be expressed as a function of $\tilde{x}$:

$$t = \tilde{x} + O(\tilde{x}^2).$$

Then

$$t^{m-n} = (\tilde{x} + O(\tilde{x}^2))^{m-n} = \tilde{x}^{m-n}(1 + O(\tilde{x})).$$

Similarly, $t^{m-n+1} = \tilde{x}^{m-n+1}(1 + O(\tilde{x}))$. Thus

$$\begin{aligned} f(u) &= u^{(n)} \\ &= am(m-1)\dots(m-n+1)t^{m-n} + O(t^{m-n+1}) \\ &= am(m-1)\dots(m-n+1)\tilde{x}^{m-n} + O(\tilde{x}^{m-n+1}) \\ &= am(m-1)\dots(m-n+1)\left(\frac{u}{a}\right)^{(m-n)/m} + O(u^{(m-n+1)/m}) \qquad \text{by (7)} \\ &= a^{n/m}m(m-1)\dots(m-n+1)u^{(m-n)/m} + O(u^{(m-n+1)/m}). \end{aligned} \tag{10}$$

Therefore,

$$a^{n/m} = \lim_{u\to 0} \frac{f(u)}{m(m-1)\dots(m-n+1)u^{(m-n)/m}}. \tag{11}$$

This shows that a is completely determined by m, n, and the behavior of the function f near 0.

Next, we show that the $(m+1)$-th derivative of u at 0 also only depends on m, n, and f.

Write u as

$$u(t) = at^m + a_{m+1}t^{m+1} + O(t^{m+2}). \tag{12}$$

We will show that a_{m+1} only depends on m, n, and the behavior of f at 0.

Express t as

$$t = \tilde{x} + b_2\tilde{x}^2 + O(\tilde{x}^3). \tag{13}$$

We would like to obtain an expression for b_2 in terms of the derivatives of u at 0. To do this, we take the derivative with respect to t on both sides of

$$\frac{d\tilde{x}}{dt} \cdot \frac{dt}{d\tilde{x}} = 1.$$

By the product rule and chain rule, we have

$$\frac{d}{dt}\Big(\frac{d\tilde{x}}{dt}\Big)\cdot\frac{dt}{d\tilde{x}}+\frac{d\tilde{x}}{dt}\cdot\frac{d}{dt}\Big(\frac{dt}{d\tilde{x}}\Big)=0, \tag{14}$$

$$\frac{d^2\tilde{x}}{dt^2}\cdot\frac{dt}{d\tilde{x}}+\frac{d\tilde{x}}{dt}\cdot\Big(\frac{d^2t}{d\tilde{x}^2}\cdot\frac{d\tilde{x}}{dt}\Big)=0,$$

$$\frac{d^2\tilde{x}}{dt^2}\cdot\frac{dt}{d\tilde{x}}+\Big(\frac{d\tilde{x}}{dt}\Big)^2\cdot\frac{d^2t}{d\tilde{x}^2}=0.$$

We would like to evaluate (14) at $t=0$.

From

$$\begin{aligned}\tilde{x}&=\Big(\frac{u}{a}\Big)^{1/m}=\Big(t^m+\frac{a_{m+1}}{a}t^{m+1}+O(t^{m+2})\Big)^{1/m}\\&=t\Big(1+\frac{a_{m+1}}{a}t+O(t^2)\Big)^{1/m}\\&=t\Big(1+\frac{1}{m}\Big(\frac{a_{m+1}}{a}t+O(t^2)\Big)+\frac{1}{2}\cdot\frac{1}{m}\Big(\frac{1}{m}-1\Big)\Big(\frac{a_{m+1}}{a}t+O(t^2)\Big)^2+O(t^3)\Big)\\&=t+\frac{a_{m+1}}{ma}t^2+O(t^3),\end{aligned}$$

we know that

$$\frac{d^2\tilde{x}}{dt^2}|_{t=0}=\frac{2a_{m+1}}{ma}. \tag{15}$$

From (13) we know that

$$\frac{dt}{d\tilde{x}}|_{t=0}=1\quad\text{and}\quad\frac{d^2t}{d\tilde{x}^2}|_{t=0}=2b_2.$$

Thus if we evaluate (14) at $t=0$, we get

$$\frac{2a_{m+1}}{ma}\cdot 1+1\cdot 2b_2=0,$$

and therefore

$$b_2=-\frac{a_{m+1}}{ma}. \tag{16}$$

Now, from (13) we have

$$\begin{aligned}t^{m-n}&=(\tilde{x}+b_2\tilde{x}^2+O(\tilde{x}^3))^{m-n}\\&=\tilde{x}^{m-n}[1+b_2\tilde{x}+O(\tilde{x}^2)]^{m-n}\\&=\tilde{x}^{m-n}\big[1+(m-n)(b_2\tilde{x}+O(\tilde{x}^2))+O(\tilde{x}^2)\big]\\&=\tilde{x}^{m-n}+(m-n)b_2\tilde{x}^{m-n+1}+O(\tilde{x}^{m-n+2}).\end{aligned}$$

Similarly,

$$t^{m-n+1} = \tilde{x}^{m-n+1} + (m-n+1)b_2\tilde{x}^{m-n+2} + O(\tilde{x}^{m-n+3}),$$
$$t^{m-n+2} = O(\tilde{x}^{m-n+2}).$$

Then from (12) and the above expressions for the powers of t, we have

$$\begin{aligned}
u^{(n)} &= am(m-1)\dots(m-n+1)t^{m-n} \\
&\quad + a_{m+1}(m+1)m\dots(m-n+2)t^{m-n+1} + O(t^{m-n+2}) \\
&= am(m-1)\dots(m-n+1)\big(\tilde{x}^{m-n} + (m-n)b_2\tilde{x}^{m-n+1} + O(\tilde{x}^{m-n+2})\big) \\
&\quad + a_{m+1}(m+1)m\dots(m-n+2) \\
&\qquad \times \big(\tilde{x}^{m-n+1} + (m-n+1)b_2\tilde{x}^{m-n+2} + O(\tilde{x}^{m-n+3})\big) \\
&\quad + O(\tilde{x}^{m-n+2}) \\
&= am(m-1)\dots(m-n+1)\tilde{x}^{m-n} \\
&\quad + \big(am(m-1)\dots(m-n)b_2 + a_{m+1}(m+1)m\dots(m-n+2)\big)\tilde{x}^{m-n+1} \\
&\quad + O(\tilde{x}^{m-n+2}).
\end{aligned}$$

Thus, in view of $f(u) = u^{(n)}$ and (7), we have

$$\begin{aligned}
f(u) &= am(m-1)\dots(m-n+1)\tilde{x}^{m-n} \\
&\quad + \big(am(m-1)\dots(m-n)b_2 + a_{m+1}(m+1)m\dots(m-n+2)\big)\tilde{x}^{m-n+1} \\
&\quad + O(\tilde{x}^{m-n+2}) \\
&= am(m-1)\dots(m-n+1)\Big(\frac{u}{a}\Big)^{\frac{m-n}{m}} \\
&\quad + \big(am(m-1)\dots(m-n)b_2 + a_{m+1}(m+1)m\dots(m-n+2)\big)\Big(\frac{u}{a}\Big)^{\frac{m-n+1}{m}} \\
&\quad + O\Big(\Big(\frac{u}{a}\Big)^{\frac{m-n+2}{m}}\Big) \\
&= a^{\frac{n}{m}}m(m-1)\dots(m-n+1)u^{\frac{m-n}{m}} \\
&\quad + \big(a^{\frac{n-1}{m}}m(m-1)\dots(m-n)b_2 + a^{\frac{n-m-1}{m}}a_{m+1}(m+1)m\dots(m-n+2)\big) \\
&\qquad\qquad \times u^{\frac{m-n+1}{m}} \\
&\quad + O(u^{\frac{m-n+2}{m}}).
\end{aligned}$$

This means that

$$\begin{aligned}
a^{\frac{n-1}{m}}m(m-1)\dots(m-n)b_2 + a^{\frac{n-m-1}{m}}a_{m+1}(m+1)m\dots(m-n+2) \\
= \lim_{u\to 0}\frac{f(u) - a^{n/m}m(m-1)\dots(m-n+1)u^{\frac{m-n}{m}}}{u^{\frac{m-n+1}{m}}}.
\end{aligned}$$

By (16), this can be written as

$$a^{\frac{n-1}{m}} m(m-1)\dots(m-n)\Big(-\frac{a_{m+1}}{ma}\Big) + a^{\frac{n-m-1}{m}} a_{m+1}(m+1)m\dots(m-n+2)$$
$$= \lim_{u\to 0} \frac{f(u) - a^{\frac{n}{m}} m(m-1)\dots(m-n+1)u^{\frac{m-n}{m}}}{u^{\frac{m-n+1}{m}}}.$$

After collecting similar terms, we get

$$a_{m+1} a^{\frac{n-m-1}{m}} \big((m+1)m\dots(m-n+2) - (m-1)(m-2)\dots(m-n)\big)$$
$$= \lim_{u\to 0} \frac{f(u) - a^{\frac{n}{m}} m(m-1)\dots(m-n+1)u^{\frac{m-n}{m}}}{u^{\frac{m-n+1}{m}}}.$$

Consequently,

$$a_{m+1} = \frac{a^{\frac{m-n+1}{m}} \left(\lim_{u\to 0} \dfrac{f(u) - a^{\frac{n}{m}} m(m-1)\dots(m-n+1)u^{\frac{m-n}{m}}}{u^{\frac{m-n+1}{m}}} \right)}{(m+1)m\dots(m-n+2) - (m-1)(m-2)\dots(m-n)}. \tag{17}$$

Since we have proved that a only depends on m, n, and f, Equation (17) shows that a_{m+1} is also completely determined by m, n, and the behavior of f near 0. By (16), this also shows that b_2 depends only on m, n, and f.

Now we will use mathematical induction to show that all the derivatives of u at 0 of order higher than m are completely determined by m, n, and f.

Express u and t as

$$u(t) = at^m + a_{m+1}t^{m+1} + \dots + a_{m+k}t^{m+k} + a_{m+k+1}t^{m+k+1} + O(t^{m+k+2}) \tag{18}$$

and

$$t = \tilde{x} + b_2\tilde{x}^2 + \dots + b_{k+1}\tilde{x}^{k+1} + b_{k+2}\tilde{x}^{k+2} + O(\tilde{x}^{k+3}). \tag{19}$$

Suppose that for $k \geq 1$, $a, a_{m+1}, \dots, a_{m+k}, b_2, \dots, b_{k+1}$ are all determined only by m, n, and f; we will show that a_{m+k+1} and b_{k+2} also are determined only by m, n, and f.

We start by obtaining an expression for b_{k+2} in terms of $a_{m+1}, \dots, a_{m+k}$ and a_{m+k+1}.

Taking the derivative with respect to t on both sides of (14), we obtain

$$0 = \frac{d^3\tilde{x}}{dt^3}\cdot\frac{dt}{d\tilde{x}} + \frac{d^2\tilde{x}}{dt^2}\cdot\Big(\frac{d^2t}{d\tilde{x}^2}\cdot\frac{d\tilde{x}}{dt}\Big) + 2\cdot\frac{d\tilde{x}}{dt}\cdot\frac{d^2\tilde{x}}{dt^2}\cdot\frac{d^2t}{d\tilde{x}^2} + \Big(\frac{d\tilde{x}}{dt}\Big)^2\cdot\Big(\frac{d^3t}{d\tilde{x}^3}\cdot\frac{d\tilde{x}}{dt}\Big)$$

and

$$0=\frac{d^3\tilde{x}}{dt^3}\cdot\frac{dt}{d\tilde{x}}+3\cdot\frac{d\tilde{x}}{dt}\cdot\frac{d^2\tilde{x}}{dt^2}\cdot\frac{d^2t}{d\tilde{x}^2}+\Big(\frac{d\tilde{x}}{dt}\Big)^3\cdot\frac{d^3t}{d\tilde{x}^3}.$$

Taking the derivative of both sides of these equations, we get

$$\begin{aligned}0=&\left(\frac{d^4\tilde{x}}{dt^4}\cdot\frac{dt}{d\tilde{x}}+\frac{d^3\tilde{x}}{dt^3}\cdot\frac{d^2t}{d\tilde{x}^2}\cdot\frac{d\tilde{x}}{dt}\right)\\&+3\left(\frac{d^2\tilde{x}}{dt^2}\cdot\frac{d^2\tilde{x}}{dt^2}\cdot\frac{d^2t}{d\tilde{x}^2}+\frac{d\tilde{x}}{dt}\cdot\frac{d^3\tilde{x}}{dt^3}\cdot\frac{d^2t}{d\tilde{x}^2}+\frac{d\tilde{x}}{dt}\cdot\frac{d^2\tilde{x}}{dt^2}\cdot\frac{d^3t}{d\tilde{x}^3}\cdot\frac{d\tilde{x}}{dt}\right)\\&+\left(3\Big(\frac{d\tilde{x}}{dt}\Big)^2\cdot\frac{d^2\tilde{x}}{dt^2}\cdot\frac{d^3t}{d\tilde{x}^3}+\Big(\frac{d\tilde{x}}{dt}\Big)^3\cdot\frac{d^4t}{d\tilde{x}^4}\cdot\frac{d\tilde{x}}{dt}\right),\\0=&\frac{d^4\tilde{x}}{dt^4}\cdot\frac{dt}{d\tilde{x}}+4\cdot\frac{d^3\tilde{x}}{dt^3}\cdot\frac{d\tilde{x}}{dt}\cdot\frac{d^2t}{d\tilde{x}^2}+3\cdot\frac{d^2\tilde{x}}{dt^2}\cdot\frac{d^2\tilde{x}}{dt^2}\cdot\frac{d^2t}{d\tilde{x}^2}\\&+6\Big(\frac{d\tilde{x}}{dt}\Big)^2\cdot\frac{d^2\tilde{x}}{dt^2}\cdot\frac{d^3t}{d\tilde{x}^3}+\Big(\frac{d\tilde{x}}{dt}\Big)^4\cdot\frac{d^4t}{d\tilde{x}^4}.\end{aligned}\tag{20}$$

If we take the derivative k times with respect to t and collect the similar terms after each differentiation as shown above, eventually we arrive at an expression of the form

$$\begin{aligned}0=&\frac{d^{k+2}\tilde{x}}{dt^{k+2}}\cdot\frac{dt}{d\tilde{x}}\\&+\left(\text{terms involving }\frac{d^{k+1}\tilde{x}}{dt^{k+1}},\frac{d^k\tilde{x}}{dt^k},\dots,\frac{d\tilde{x}}{dt},\frac{dt}{d\tilde{x}},\frac{d^2t}{d\tilde{x}^2},\dots,\frac{d^{k+1}t}{d\tilde{x}^{k+1}}\right)\\&+\Big(\frac{d\tilde{x}}{dt}\Big)^{k+2}\cdot\frac{d^{k+2}t}{d\tilde{x}^{k+2}}.\end{aligned}\tag{21}$$

From (19) we know that

$$\begin{aligned}\frac{dt}{d\tilde{x}}\Big|_{t=0}&=1,\\\frac{d^2t}{d\tilde{x}^2}\Big|_{t=0}&=2b_2,\\&\vdots\\\frac{d^{k+1}t}{d\tilde{x}^{k+1}}\Big|_{t=0}&=(k+1)!\,b_{k+1},\\\frac{d^{k+2}t}{d\tilde{x}^{k+2}}\Big|_{t=0}&=(k+2)!\,b_{k+2}.\end{aligned}\tag{22}$$

Then we look at $\frac{d\tilde{x}}{dt}\big|_{t=0}, \ldots, \frac{d^{k+1}\tilde{x}}{dt^{k+1}}\big|_{t=0}$, and $\frac{d^{k+2}\tilde{x}}{dt^{k+2}}\big|_{t=0}$. By (7) and (18),

$$\begin{aligned}
\tilde{x} &= \left(\frac{at^m + a_{m+1}t^{m+1} + \cdots + a_{m+k}t^{m+k} + a_{m+k+1}t^{m+k+1} + O(t^{m+k+2})}{a}\right)^{1/m} \\
&= t\left(1 + \frac{a_{m+1}}{a}t + \frac{a_{m+2}}{a}t^2 + \cdots + \frac{a_{m+k}}{a}t^k + \frac{a_{m+k+1}}{a}t^{k+1} + O\left(t^{k+2}\right)\right)^{1/m} \\
&= t\left\{1 + \frac{1}{m}\left[\frac{a_{m+1}}{a}t + \frac{a_{m+2}}{a}t^2 + \cdots + \frac{a_{m+k}}{a}t^k + \frac{a_{m+k+1}}{a}t^{k+1} + O(t^{k+2})\right]\right. \\
&\quad + \tfrac{1}{2}\cdot\frac{1}{m}\left(\frac{1}{m} - 1\right) \\
&\qquad \left[\frac{a_{m+1}}{a}t + \frac{a_{m+2}}{a}t^2 + \cdots + \frac{a_{m+k}}{a}t^k + \frac{a_{m+k+1}}{a}t^{k+1} + O(t^{k+2})\right]^2 \\
&\quad + \cdots \\
&\quad + \frac{1}{(k+1)!}\cdot\frac{1}{m}\cdot\left(\frac{1}{m} - 1\right)\cdots\left(\frac{1}{m} - k\right) \\
&\qquad \left[\frac{a_{m+1}}{a}t + \frac{a_{m+2}}{a}t^2 + \cdots + \frac{a_{m+k}}{a}t^k + \frac{a_{m+k+1}}{a}t^{k+1} + O(t^{k+2})\right]^{k+1} \\
&\quad \left. + O(t^{k+2})\right\}
\end{aligned}$$

After collecting similar terms, we can write

$$\tilde{x} = t + \lambda_2 t^2 + \lambda_3 t^3 + \cdots + \lambda_{k+1}t^{k+1} + \left(\frac{a_{m+k+1}}{ma} + \lambda_{k+2}\right)t^{k+2} + O(t^{k+3}), \tag{23}$$

where

- λ_2 is a constant involving m, a, and a_{m+1},
- λ_3 is a constant involving m, a, a_{m+1} and a_{m+2},

  $\vdots$

- λ_{k+1} is a constant involving $m, a, a_{m+1}, \ldots, a_{m+k-1}, a_{m+k}$, and
- λ_{k+2} is a constant involving $m, a, a_{m+1}, \ldots, a_{m+k-1}, a_{m+k}$.

By the inductive hypothesis, $\lambda_2, \lambda_3, \ldots, \lambda_{k+1}, \lambda_{k+2}$ are all constants that only depend on m, n, and the function f.

From (23), we obtain

$$
\begin{aligned}
\frac{d\tilde{x}}{dt}\Big|_{t=0} &= 1,\\
\frac{d^2\tilde{x}}{dt^2}\Big|_{t=0} &= 2\lambda_2,\\
&\vdots\\
\frac{d^{k+1}\tilde{x}}{dt^{k+1}}\Big|_{t=0} &= (k+1)!\,\lambda_{k+1},\\
\frac{d^{k+2}\tilde{x}}{dt^{k+2}}\Big|_{t=0} &= (k+2)!\,\Big(\frac{a_{m+k+1}}{ma}+\lambda_{k+2}\Big).
\end{aligned}
\tag{24}
$$

Now we evaluate (21) at $t=0$ and make use of (22) and (24):

$$
\begin{aligned}
0 = (k+2)!\Big(\frac{a_{m+k+1}}{ma}+\lambda_{k+2}\Big)\cdot 1\\
+\Big(\text{terms involving } b_2,\dots,b_{k+1},\lambda_2,\dots,\lambda_{k+1}\Big)+1\cdot(k+2)!\,b_{k+2}.
\end{aligned}
$$

Thus we obtain

$$
b_{k+2} = -\frac{a_{m+k+1}}{ma}+Q, \tag{25}
$$

where Q is a constant depending on $b_2,\dots,b_{k+1},\lambda_2,\dots,\lambda_{k+1},\lambda_{k+2}$, and hence Q is completely determined by m, n, and f.

Next we will analyze a_{m+k+1}. From (19) we have

$$
\begin{aligned}
t^{m-n} &= \tilde{x}^{m-n}\big(1+b_2\tilde{x}+\cdots+b_{k+1}\tilde{x}^k+b_{k+2}\tilde{x}^{k+1}+O(\tilde{x}^{k+2})\big)^{m-n}\\
&= \tilde{x}^{m-n}\Big\{1+(m-n)\big(b_2\tilde{x}+\cdots+b_{k+1}\tilde{x}^k+b_{k+2}\tilde{x}^{k+1}+O(\tilde{x}^{k+2})\big)\\
&\quad+\frac{(m-n)(m-n-1)}{2}\big(b_2\tilde{x}+\cdots+b_{k+1}\tilde{x}^k+b_{k+2}\tilde{x}^{k+1}+O(\tilde{x}^{k+2})\big)^2+\cdots\\
&\quad+\frac{(m-n)(m-n-1)\dots(m-n-k)}{(k+1)!}\\
&\qquad\big(b_2\tilde{x}+\cdots+b_{k+1}\tilde{x}^k+b_{k+2}\tilde{x}^{k+1}+O(\tilde{x}^{k+2})\big)^{k+1}\\
&\quad+O(\tilde{x}^{k+2})\Big\}.
\end{aligned}
$$

After collecting similar terms, we can express t^{m-n} as

$$
\begin{aligned}
t^{m-n} = \tilde{x}^{m-n}\big\{1+c_{1,m-n}\tilde{x}+c_{2,m-n}\tilde{x}^2+\cdots+c_{k,m-n}\tilde{x}^k\\
+((m-n)b_{k+2}+c_{k+1,m-n})\tilde{x}^{k+1}+O(\tilde{x}^{k+2})\big\},
\end{aligned}
$$

where $c_{1,m-n}$ is a constant depending on m and b_2; $c_{2,m-n}$ is a constant depending on m, b_2, and b_3; ...; $c_{k,m-n}$ is a constant depending on m, $b_2, \ldots, b_{k+1}$; $c_{k+1,m-n}$ is a constant depending on $m, b_2, \ldots, b_{k+1}$.

By the inductive hypothesis, $c_{1,m-n}, c_{2,m-n}, \ldots, c_{k,m-n}$ and $c_{k+1,m-n}$ are all determined only by m, n, and f. Thus we have

$$t^{m-n} = \tilde{x}^{m-n} + c_{1,m-n}\tilde{x}^{m-n+1} + c_{2,m-n}\tilde{x}^{m-n+2} + \cdots + c_{k,m-n}\tilde{x}^{m-n+k} + \big((m-n)b_{k+2} + c_{k+1,m-n}\big)\tilde{x}^{m-n+k+1} + O(\tilde{x}^{m-n+k+2}), \tag{26}$$

where $c_{1,m-n}, c_{2,m-n}, \ldots, c_{k,m-n}$ and $c_{k+1,m-n}$ are constants depending on m, n, and f.

By the same type of analysis we obtain similar expressions for the other powers of t:

$$t^{m-n+1} = \tilde{x}^{m-n+1} + c_{1,m-n+1}\tilde{x}^{m-n+2} + c_{2,m-n+1}\tilde{x}^{m-n+3} + \cdots + c_{k,m-n+1}\tilde{x}^{m-n+k+1} + \big((m-n+1)b_{k+2} + c_{k+1,m-n+1}\big)\tilde{x}^{m-n+k+2} + O(\tilde{x}^{m-n+k+3}), \tag{27}$$

where $c_{1,m-n+1}, c_{2,m-n+1}, \ldots, c_{k,m-n+1}$, and $c_{k+1,m-n+1}$ are constants depending on m, n, and f.

$$t^{m-n+2} = \tilde{x}^{m-n+2} + c_{1,m-n+2}\tilde{x}^{m-n+3} + c_{2,m-n+2}\tilde{x}^{m-n+4} + \cdots + c_{k,m-n+2}\tilde{x}^{m-n+k+2} + \big((m-n+2)b_{k+2} + c_{k+1,m-n+2}\big)\tilde{x}^{m-n+k+3} + O(\tilde{x}^{m-n+k+4}), \tag{28}$$

where $c_{1,m-n+2}, c_{2,m-n+2}, \ldots, c_{k,m-n+2}$ and $c_{k+1,m-n+2}$ are constants depending on m, n, and f. Proceeding inductively,

$$t^{m-n+k} = \tilde{x}^{m-n+k} + c_{1,m-n+k}\tilde{x}^{m-n+k+1} + c_{2,m-n+k}\tilde{x}^{m-n+k+2} + \cdots + c_{k,m-n+k}\tilde{x}^{m-n+2k} + \big((m-n+k)b_{k+2} + c_{k+1,m-n+k}\big)\tilde{x}^{m-n+2k+1} + O(\tilde{x}^{m-n+2k+2}), \tag{29}$$

where $c_{1,m-n+k}, c_{2,m-n+k}, \ldots, c_{k,m-n+k}$ and $c_{k+1,m-n+k}$ are constants depending on m, n, and f.

$$t^{m-n+k+1} = \tilde{x}^{m-n+k+1} + c_{1,m-n+k+1}\tilde{x}^{m-n+k+2} + c_{2,m-n+k+1}\tilde{x}^{m-n+k+3} + \cdots + c_{k,m-n+k+1}\tilde{x}^{m-n+2k+1} + \big((m-n+k+1)b_{k+2} + c_{k+1,m-n+k+1}\big)\tilde{x}^{m-n+2k+2} + O(\tilde{x}^{m-n+2k+3}), \tag{30}$$

where $c_{1,m-n+k+1}, \ldots, c_{k,m-n+k+1}$ and $c_{k+1,m-n+k+1}$ are constants depending on m, n, and f.

$$t^{m-n+k+2} = O(\tilde{x}^{m-n+k+2}). \tag{31}$$

From (18) we obtain

$$\begin{aligned}u^{(n)} &= am(m-1)\dots(m-n+1)t^{m-n} + a_{m+1}(m+1)m\dots(m-n+2)t^{m-n+1}\\ &\quad + \dots + a_{m+k}(m+k)(m+k-1)\dots(m-n+k+1)t^{m-n+k}\\ &\quad + a_{m+k+1}(m+k+1)(m+k)\dots(m-n+k+2)t^{m-n+k+1} + O(t^{m-n+k+2}).\end{aligned}$$

Then by (26) to (31), we can write

$$\begin{aligned}u^{(n)} &= am(m-1)\dots(m-n+1)\Big\{\tilde{x}^{m-n} + c_{1,m-n}\tilde{x}^{m-n+1} + c_{2,m-n}\tilde{x}^{m-n+2} + \dots\\ &\quad + c_{k,m-n}\tilde{x}^{m-n+k} + \big((m-n)b_{k+2} + c_{k+1,m-n}\big)\tilde{x}^{m-n+k+1} + O(\tilde{x}^{m-n+k+2})\Big\}\\ &\quad + a_{m+1}(m+1)m\dots(m-n+2)\Big\{\tilde{x}^{m-n+1} + c_{1,m-n+1}\tilde{x}^{m-n+2}\\ &\quad + c_{2,m-n+1}\tilde{x}^{m-n+3} + \dots + c_{k,m-n+1}\tilde{x}^{m-n+k+1}\\ &\quad + \big((m-n+1)b_{k+2} + c_{k+1,m-n+1}\big)\tilde{x}^{m-n+k+2} + O(\tilde{x}^{m-n+k+3})\Big\} + \dots\\ &\quad + a_{m+k}(m+k)(m+k-1)\dots(m+k-n+1)\Big\{\tilde{x}^{m-n+k}\\ &\quad + c_{1,m-n+k}\tilde{x}^{m-n+k+1} + c_{2,m-n+k}\tilde{x}^{m-n+k+2} + \dots\\ &\quad + c_{k,m-n+k}\tilde{x}^{m-n+2k} + \big((m-n+k)b_{k+2} + c_{k+1,m-n+k}\big)\tilde{x}^{m-n+2k+1}\\ &\quad + O(\tilde{x}^{m-n+2k+2})\Big\}\\ &\quad + a_{m+k+1}(m+k+1)(m+k)\dots(m+k-n+2)\Big\{\tilde{x}^{m-n+k+1}\\ &\quad + c_{1,m-n+k+1}\tilde{x}^{m-n+k+2} + c_{2,m-n+k+1}\tilde{x}^{m-n+k+3} + \dots\\ &\quad + c_{k,m-n+k+1}\tilde{x}^{m-n+2k+1}\\ &\quad + \big((m-n+k+1)b_{k+2} + c_{k+1,m-n+k+1}\big)\tilde{x}^{m-n+2k+2} + O(\tilde{x}^{m-n+2k+3})\Big\}\\ &\quad + O(\tilde{x}^{m-n+k+2})\\ &= am(m-1)\cdots(m-n+1)\tilde{x}^{m-n} + C(m, a, a_{m+1}, c_{1,m-n})\tilde{x}^{m-n+1}\\ &\quad + C\big(m, a, a_{m+1}, a_{m+2}, c_{1,m-n+1}, c_{2,m-n}\big)\tilde{x}^{m-n+2} + \dots\\ &\quad + C\big(m, a, a_{m+1}, \dots, a_{m+k}, c_{k,m-n}, c_{k-1,m-n+1}, \dots, c_{1,m-n+k-1}\big)\tilde{x}^{m-n+k}\\ &\quad + \Big(am(m-1)\dots(m-n)b_{k+2}\\ &\quad + a_{m+k+1}(m+k+1)(m+k)\dots(m+k-n+2)\\ &\quad + C\big(m, a, a_{m+1}, \dots, a_{m+k}, c_{k+1,m-n}, c_{k,m-n+1}, \dots, c_{1,m-n+k}\big)\Big)\tilde{x}^{m-n+k+1}\\ &\quad + O(\tilde{x}^{m-n+k+2}).\end{aligned}$$

Here $C(m, a, a_{m+1}, c_{1,m-n})$ is a constant depending on m, a, a_{m+1}, and $c_{1,m-n}$; we denote it as p_{m-n+1} to simplify notations. Since a, a_{m+1}, and $c_{1,m-n}$ only depend on m, n, and f, we know that p_{m-n+1} only depends on m, n, and f.

Similarly, the other constants $C\big(m, a, a_{m+1}, a_{m+2}, c_{1,m-n+1}, c_{2,m-n}\big), \ldots$, and $C\big(m, a, a_{m+1}, \ldots, a_{m+k}, c_{k+1,m-n}, c_{k,m-n+1}, \ldots, c_{1,m-n+k}\big)$ all depend on m, n, and f only, and can be denoted simply as $p_{m-n+2}, \ldots, p_{m-n+k}$, and $p_{m-n+k+1}$. Thus we can rewrite the above equation as

$$
\begin{aligned}
(32)\quad u^{(n)} &= am(m-1)\ldots(m-n+1)\tilde{x}^{m-n} + p_{m-n+1}\tilde{x}^{m-n+1} \\
&\quad + p_{m-n+2}\tilde{x}^{m-n+2} + \cdots + p_{m-n+k}\tilde{x}^{m-n+k} \\
&\quad + \big(am(m-1)\ldots(m-n)b_{k+2} \\
&\qquad + a_{m+k+1}(m+k+1)(m+k)\ldots(m+k-n+2) + p_{m-n+k+1}\big)\tilde{x}^{m-n+k+1} \\
&\quad + O(\tilde{x}^{m-n+k+2}).
\end{aligned}
$$

Now because of $u^{(n)} = f(u)$ and definition (7) , we have

$$
\begin{aligned}
f(u) &= am(m-1)\ldots(m-n+1)\Big(\frac{u}{a}\Big)^{(m-n/m)} + p_{m-n+1}\Big(\frac{u}{a}\Big)^{(m-n+1/m)} \\
&\quad + p_{m-n+2}\Big(\frac{u}{a}\Big)^{(m-n+2)/m} + \cdots + p_{m-n+k}\Big(\frac{u}{a}\Big)^{(m-n+k)/m} \\
&\quad + \big(am(m-1)\ldots(m-n)b_{k+2} \\
&\qquad + a_{m+k+1}(m+k+1)(m+k)\cdots(m+k-n+2) \\
&\qquad\qquad + p_{m-n+k+1}\big)\Big(\frac{u}{a}\Big)^{(m-n+k+1)/m} \\
&\quad + O(u^{(m-n+k+2)/m}).
\end{aligned}
$$

Due to (25), we can rewrite this equation as

$$
\begin{aligned}
(33)\quad f(u) &= am(m-1)\ldots(m-n+1)\Big(\frac{u}{a}\Big)^{(m-n)/m} + p_{m-n+1}\Big(\frac{u}{a}\Big)^{(m-n+1)/m} \\
&\quad + p_{m-n+2}\Big(\frac{u}{a}\Big)^{(m-n+2)/m} + \cdots + p_{m-n+k}\Big(\frac{u}{a}\Big)^{(m-n+k)/m} \\
&\quad + \big\{\big((m+k+1)(m+k)\ldots(m+k-n+2) \\
&\qquad - (m-1)(m-2)\ldots(m-n)\big)a_{m+k+1} \\
&\qquad + am(m-1)\ldots(m-n)Q + p_{m-n+k+1}\big\}\Big(\frac{u}{a}\Big)^{(m-n+k+1)/m} \\
&\quad + O(u^{(m-n+k+2)/m}).
\end{aligned}
$$

From (33) we get

$$\begin{aligned}&\big((m+k+1)(m+k)\dots(m+k-n+2)-(m-1)(m-2)\dots(m-n)\big)a_{m+k+1}\\&\quad+am(m-1)\dots(m-n)Q+p_{m-n+k+1}\\&\quad=\lim_{u\to 0}\frac{f(u)-am\dots(m-n+1)\left(\frac{u}{a}\right)^{\frac{m-n}{m}}-p_{m-n+1}\left(\frac{u}{a}\right)^{\frac{m-n+1}{m}}-\dots-p_{m-n+k}\left(\frac{u}{a}\right)^{\frac{m-n+k}{m}}}{\left(\frac{u}{a}\right)^{\frac{m-n+k+1}{m}}}.\end{aligned}$$

Note that $(m+k+1)(m+k)\dots(m+k-n+2)-(m-1)(m-2)\dots(m-n)\neq 0$; then since the constants $Q, a, p_{m-n+1}, \dots, p_{m-n+k+1}$ all depend only on m, n, and f, we know that a_{m+k+1} only depends on m, n, and f. Consequently, b_{k+2} also only depends on m, n, and f because of (25).

Therefore, by mathematical induction, all derivatives of f at 0 are determined completely by m, n, and f. This completes the proof of Lemma 1.3.

3. The proofs of Theorems 1.2, 1.4 and 1.5

Proof of Theorem 1.2. Suppose there are two solutions $u(t)$ and $v(t)$, both satisfying Equations (3) and (4). By Lemma 1.3, at $t=0$, u and v have the same derivative of any order. Let $w=u-v$; then

$$w^{(k)}(0)=0 \quad \text{for any integer} \quad k\geq 0.$$

In order to apply Theorem 1.1, we need to show that w satisfies Condition (2).

$$w^{(n)}(t)=u^{(n)}(t)-v^{(n)}(t)=f(u(t))-f(v(t)).$$

Without loss of generality, we assume $a>0$. By Equation (10), we can write

$$f(u)=\left[a^{n/m}m(m-1)\dots(m-n+1)u^{(m-n)/m}+\alpha(u)\right] \tag{34}$$

and

$$f(v)=\left[a^{n/m}m(m-1)\dots(m-n+1)v^{(m-n)/m}+\alpha(v)\right],$$

where α is a function with the order

$$\alpha(s)=O(s^{(m-n+1)/m}).$$

So we can write

$$w^{(n)}(t)=a^{n/m}m(m-1)\dots(m-n+1)(u^{(m-n)/m}-v^{(m-n)/m})+(\alpha(u)-\alpha(v)). \tag{35}$$

If $m=n$, then

$$a^{n/m}m(m-1)\dots(m-n+1)(u^{(m-n)/m}-v^{(m-n)/m})=0.$$

If $m > n$, by the mean value theorem,

$$|u^{(m-n)/m} - v^{(m-n)/m}| \le \frac{m-n}{m}\zeta^{-n/m}|u-v|, \tag{36}$$

where $\zeta(t)$ is between $u(t)$ and $v(t)$. Since $u(t) = at^m + O(t^{m+1})$ and $v(t) = at^m + O(t^{m+1})$, we know that $\zeta(t) = at^m + O(t^{m+1})$, which implies

$$\zeta^{n/m} = a^{n/m}t^n(1 + O(t)) \ge Ct^n$$

for some constant $C > 0$ when t is sufficiently small. Thus

$$\zeta^{-n/m}|u-v| \le C^{-1}\frac{|u-v|}{t^n},$$

and by (36), we know that

$$a^{n/m}m(m-1)\dots(m-n+1)\left|u^{(m-n)/m} - v^{(m-n)/m}\right| \le C\frac{|u-v|}{t^n} \tag{37}$$

for another constant $C > 0$.

Next, we estimate $|\alpha(u) - \alpha(v)|$.

From (3), we know that f is differentiable with respect to t, since $u^{(n)}(t)$ is differentiable with respect to t. Condition (4) shows that $(du/dt)(t) \ne 0$ when $t > 0$ is sufficiently small. Then by the inverse function theorem, t is differentiable with respect to u. Thus, when u is small and positive, f is differentiable with respect to u and

$$\frac{df}{du} = \frac{df}{dt}\cdot\frac{dt}{du}.$$

Then by (34), since f is differentiable on a small interval $(0, \delta)$, α is also differentiable on a small interval $(0, \delta)$. By the mean value theorem,

$$\alpha(u) - \alpha(v) = \alpha'(\eta)(u-v),$$

where $\eta(t)$ is between $u(t)$ and $v(t)$. Because $u(t) = at^m + O(t^{m+1})$ and $v(t) = at^m + O(t^{m+1})$, we know that $\eta(t) = at^m + O(t^{m+1}) \ge Ct^m$ when t is small, and thus

$$\eta^{-(n-1)/m} = O(t^{-(n-1)}).$$

From $\alpha(s) = O(s^{(m-n+1)/m})$ we get $\alpha'(s) = O(s^{-(n-1)/m})$. Therefore

$$\alpha'(\eta) = O(\eta^{-(n-1)/m}) = O(t^{-(n-1)}).$$

Thus for some $C > 0$,

$$\begin{aligned} |\alpha(u) - \alpha(v)| &\le Ct^{-(n-1)}|u-v| \\ &\le Ct^{-n}|u-v| \quad \text{since} \quad 0 < t < 1. \end{aligned} \tag{38}$$

Combining Equations (35), (37), and (38), we conclude

$$|w^{(n)}(t)| \le C\frac{|u(t)-v(t)|}{t^n} = C\frac{|w(t)|}{t^n}.$$

Finally, extend the domain of $w(t)$ to $[-1, 1]$ by defining $w(t) = w(-t)$ when $-1 \le t < 0$. Then $w \in C^\infty([-1, 1])$ and it satisfies Condition (2). By Theorem 1.1, $w \equiv 0$, which means $u \equiv v$.

This completes the proof of Theorem 1.2. □

Proof of Theorem 1.4. Without loss of generality we assume $a > 0$. We apply the same analysis as in the proof of Lemma 1.3 to u_1 and u_2, respectively. Similar to (11), we have

$$\begin{aligned} a^{n/m} &= \lim_{u_1\to 0} \frac{f(u_1)}{m(m-1)\dots(m-n+1)u_1^{(m-n)/m}} \\ &= \lim_{s\to 0} \frac{f(s)}{m(m-1)\dots(m-n+1)s^{(m-n)/m}} \end{aligned}$$

and

$$b^{n/l} = \lim_{u_2\to 0} \frac{f(u_2)}{l(l-1)\dots(l-n+1)u_2^{(l-n)/l}} = \lim_{s\to 0} \frac{f(s)}{l(l-1)\dots(l-n+1)s^{(l-n)/l}}.$$

Suppose $m \ne l$; without loss of generality we assume $m < l$. Dividing the two equations, we get

$$\begin{aligned} a^{n/m}b^{-n/l} &= \frac{l(l-1)\dots(l-n+1)}{m(m-1)\dots(m-n+1)} \lim_{s\to 0} s^{(l-n)/l-(m-n)/m} \\ &= \frac{l(l-1)\dots(l-n+1)}{m(m-1)\dots(m-n+1)} \lim_{s\to 0} s^{n/m-n/l}. \end{aligned}$$

Since $m < l$, $\lim_{s\to 0} s^{n/m-n/l} = 0$. However, $a^{n/m}b^{-n/l} \ne 0$. This is a contradiction.

Therefore $m = l$, and consequently $a = b$. Then by Theorem 1.2, we know that $u_1 \equiv u_2$ for small t. □

Proof of Theorem 1.5. The proof of Lemma 1.3 shows that near 0, t is a function of u, and therefore $u^{(n)}(t)$ can be expressed as a function f of u. Thus (3) holds when $t > 0$ is small. From Condition (4), we define $f(0) = 0$.

By the first two equations in (10) and the discussions in the Appendix, we know that there is a function h that is C^1 on the closed interval $[0, \epsilon]$ for some $\epsilon > 0$, such that

$$f(u) = am(m-1)\dots(m-n+1)\tilde{x}^{m-n} + h(\tilde{x})\tilde{x}^{m-n}.$$

By definition (7), we have

$$f(u) = am(m-1)\dots(m-n+1)\Big(\frac{u}{a}\Big)^{(m-n)/m} + h\Big(\Big(\frac{u}{a}\Big)^{1/m}\Big)\Big(\frac{u}{a}\Big)^{(m-n)/m}. \tag{39}$$

Since $0 \le (m-n)/m < 1$ and $0 < 1/m < 1$, it is well known that $u^{(m-n)/m}$ and $u^{1/m}$ are Hölder continuous on the closed interval $[0, 1]$ with Hölder coefficients $(m-n)/m$ and $1/m$, respectively. This implies that the first term in (39) is Hölder continuous on $[0, 1]$.

Since h is C^1 on $[0, \epsilon]$, it is also Hölder continuous on $[0, \epsilon]$. Then since the composition of two Hölder continuous functions is Hölder continuous, we know that $h((u/a)^{1/m})$ is Hölder continuous with respect to u on a closed interval $[0, \delta]$ with $\delta > 0$. Next, because the product of two Hölder continuous functions is also Hölder continuous, we know that $h((u/a)^{1/m}) \cdot (u/a)^{(m-n)/m}$ is Hölder continuous. Thus the second term in (39) is Hölder continuous on $[0, \delta]$.

Therefore, f is Hölder continuous on $[0, \delta]$ and the theorem is proved. □

Appendix: Differentiation of the Taylor expansion

We will discuss the regularity of the remainder term in the Taylor expansion of a function that is used in the proof of Theorem 1.5 and the differentiation of the Taylor expansion that is frequently used in the proof of Lemma 1.3.

In general, consider a function $g(x) \in C^{k+1}([a, b])$; by the Taylor theorem, we can write

$$g(x) = g(a) + g'(a)(x-a) + \frac{g''(a)}{2!}(x-a)^2 + \cdots + \frac{g^{(k)}(a)}{k!}(x-a)^k + h(x)(x-a)^k, \tag{40}$$

where $\lim_{x \to a} h(x) = 0$. An explicit expression for $h(x)$ is

$$h(x) = \frac{g^{(k+1)}(\xi)}{(k+1)!}(x-a), \tag{41}$$

where $a < \xi < x$.

From (40) we know that $h(x)$ is C^1 on $(a, b]$. Next we show that it is actually C^1 up to the boundary, on $[a, b]$.

Taking the derivative on both sides of (40), we get

$$g'(x) = g'(a) + g''(a)(x-a) + \cdots + \frac{g^{(k)}(a)}{(k-1)!}(x-a)^{k-1} + h'(x)(x-a)^k + kh(x)(x-a)^{k-1}. \tag{42}$$

Define $h(a) = 0$, so h is continuous on $[a, b]$. Write

$$P(x) = g(a) + g'(a)(x-a) + \frac{g''(a)}{2!}(x-a)^2 + \cdots + \frac{g^{(k)}(a)}{k!}(x-a)^k;$$

then $h(x) = \dfrac{g(x) - P(x)}{(x-a)^k}$. By the definition of limits,

$$h'(a) = \lim_{x\to a}\frac{h(x)-h(a)}{x-a} = \lim_{x\to a}\frac{g(x)-P(x)}{(x-a)^{k+1}} \tag{43}$$

$$\vdots$$

$$= \frac{g^{(k+1)}(a)-P^{(k+1)}(a)}{(k+1)!} \quad \text{applying l'Hospital's rule } k+1 \text{ times}$$

$$= \frac{g^{(k+1)}(a)}{(k+1)!},$$

where we have used the fact that $P^{(k+1)}(a) = 0$.

When $x > a$,

$$h'(x) = \frac{d}{dx}\Big(\frac{g(x)-P(x)}{(x-a)^k}\Big)$$

$$= \frac{(g'(x)-P'(x))(x-a)^k - (g(x)-P(x))k(x-a)^{k-1}}{(x-a)^{2k}}$$

$$= \frac{g'(x)-P'(x)}{(x-a)^k} - \frac{k\,(g(x)-P(x))}{(x-a)^{k+1}}.$$

By repeatedly applying l'Hospital's rule, we know that

$$\lim_{x\to a}\frac{g'(x)-P'(x)}{(x-a)^k} = \frac{g^{(k+1)}(a)-P^{(k+1)}(a)}{k!} = \frac{g^{(k+1)}(a)}{k!}$$

and

$$\lim_{x\to a}\frac{k(g(x)-P(x))}{(x-a)^{k+1}} = \frac{k\left(g^{(k+1)}(a)-P^{(k+1)}(a)\right)}{(k+1)!} = \frac{kg^{(k+1)}(a)}{(k+1)!}.$$

Therefore

$$\lim_{x\to a} h'(x) = \frac{g^{(k+1)}(a)}{k!} - \frac{kg^{(k+1)}(a)}{(k+1)!} = \frac{g^{(k+1)}(a)}{(k+1)!}. \tag{44}$$

Equations (43) and (44) show that $h(x)$ is C^1 on the closed interval $[a, b]$.

Furthermore, we know that for any $x \in [a, b]$, $|h'(x)| \le C$ for some constant C_1, and thus

$$|h'(x)(x-a)^k| \le C_1|x-a|^k.$$

Since $g(x) \in C^{k+1}([a, b])$, from (41) we know that $|h(x)| \le C_2|x-a|$ for some constant C_2, and thus

$$|kh(x)(x-a)^{k-1}| \le kC_2|x-a|^k.$$

Therefore, (42) can be written as

$$g'(x) = g'(a) + g''(a)(x-a) + \cdots + \frac{g^{(k)}(a)}{(k-1)!}(x-a)^{k-1} + O(x-a)^k. \tag{45}$$

Since the first, second, ..., and $(k-1)$-th derivatives of $g'(x)$ at a are $g''(a)$, $g^{(3)}(a)$, ..., and $g^{(k)}(a)$, respectively, Equation (45) is the Taylor expansion of $g'(x)$ at a to order $k-1$.

Usually we write (40) as

$$g(x) = g(a) + g'(a)(x-a) + \frac{g''(a)}{2!}(x-a)^2 + \cdots + \frac{g^{(k)}(a)}{k!}(x-a)^k + O((x-a)^{k+1}). \tag{46}$$

This shows that we can formally differentiate (46) to get (45).

References

[Agarwal and Lakshmikantham 1993] R. P. Agarwal and V. Lakshmikantham, *Uniqueness and nonuniqueness criteria for ordinary differential equations*, Series in Real Analysis **6**, World Scientific Publishing, River Edge, NJ, 1993. MR 96e:34002 Zbl 0785.34003

[Li and Nirenberg 2006] Y. Y. Li and L. Nirenberg, "A geometric problem and the Hopf lemma, II", *Chinese Ann. Math. Ser. B* **27**:2 (2006), 193–218. MR 2007i:53006 Zbl 1149.53302

[Pan and Wang 2008] Y. Pan and M. Wang, "When is a function not flat?", *J. Math. Anal. Appl.* **340**:1 (2008), 536–542. MR 2008m:26048 Zbl 1131.26016

Received January 25, 2012. Revised February 20, 2012.

YIFEI PAN
DEPARTMENT OF MATHEMATICAL SCIENCES
INDIANA UNIVERSITY–PURDUE UNIVERSITY FORT WAYNE
FORT WAYNE, IN 46805
UNITED STATES

and

SCHOOL OF MATHEMATICS AND INFORMATICS
JIANGXI NORMAL UNIVERSITY
NANCHANG
CHINA
pan@ipfw.edu

MEI WANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, IL 60637
UNITED STATES
meiwang@galton.uchicago.edu

YU YAN
DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
HUNTINGTON UNIVERSITY
HUNTINGTON, IN 46750
UNITED STATES
yyan@huntington.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 263, No. 2, 2013
dx.doi.org/10.2140/pjm.2013.263.475

AN ANALOGUE TO THE WITT IDENTITY

G. A. T. F. DA COSTA AND G. A. ZIMMERMANN

We solve combinatorial and algebraic problems associated with a multivariate identity first considered by Sherman, which he called an analog to the Witt identity. We extend previous results obtained for the univariate case.

1. Introduction

S. Sherman [1962] considered the formal identity in the indeterminates $z_1, \dots, z_n$,

$$\prod_{m_1,\dots,m_R\geq 0} (1+z_1^{m_1}\cdots z_R^{m_R})^{N_+}(1-z_1^{m_1}\cdots z_R^{m_R})^{N_-} = \prod_{j=1}^{R}(1+z_j)^2, \tag{1-1}$$

where N_+ and N_- are the number of distinct classes of equivalence of nonperiodic closed paths with positive and negative signs, respectively, which traverse without backtracking m_i times edge i, $i = 1, \dots, R$, of a graph G_R with $R > 1$ edges forming loops counterclockwise oriented and hooked to a single vertex, $\sum m_i \geq 1$.

Sherman [1962] refers to (1-1) as *an analog to the Witt identity*. The reason will become clear soon. The *Sherman identity*, as we call it, is a special nontrivial case of another identity called the *Feynman identity*, first conjectured by Richard Feynman. This identity relates the Euler polynomial of a graph to a formal product over the classes of equivalence of closed nonperiodic paths with no backtracking in the graph, and it is an important ingredient in a combinatorial formulation of the Ising model in two dimensions, much studied in physics. The Feynman identity was proved for planar and toroidal graphs by Sherman [1960], and in great generality by M. Loebl [2004] and D. Cimasoni [2010].

Sherman compared (1-1) with the multivariate Witt identity [Witt 1937]:

$$\prod_{m_1,\dots,m_R\geq 0} (1-z_1^{m_1}\cdots z_R^{m_R})^{\mathcal{M}(m_1,\dots,m_R)} = 1-\sum_{i=1}^{R} z_i, \tag{1-2}$$

$$\mathcal{M}(m_1,\dots,m_R) = \sum_{g\mid m_1,\dots,m_R} \frac{\mu(g)}{g}\frac{(N/g)!}{(N/g)(m_1/g)!\cdots(m_R/g)!} \tag{1-3}$$

MSC2010: primary 05C30; secondary 05C25, 05C38.
Keywords: Sherman identity, paths counting, (generalized) Witt formula, free Lie algebras.

where $N = m_1 + \cdots + m_R > 0$, μ is the Möbius function defined by the rules

(a) $\mu(+1) = +1$,

(b) $\mu(g) = 0$, for $g = p_1^{e_1} \cdots p_q^{e_q}$, $p_1, \ldots, p_q$ primes, and any $e_i > 1$,

(c) $\mu(p_1 \cdots p_q) = (-1)^q$.

The summation runs over all the common divisors of $m_1, \ldots, m_R$.

Originally, the Witt identity appeared associated with Lie algebras. In this context the formula gives the dimensions of the homogeneous subspaces of a finitely generated free Lie algebra L. If $L(m_1, \ldots, m_R)$ is the subspace of L generated by all homogeneous elements of multidegree $(m_1, \ldots, m_R)$, then $\dim L = \mathcal{M}$. However, formula (1-3) has many applications in combinatorics as well [Moree 2005]. Especially relevant is that $\mathcal{M}$ can be interpreted as the number of equivalence classes of closed nonperiodic paths which traverse counterclockwise the edges of G_R, the same graph associated to the Sherman identity (1-1). This property is stated in [Sherman 1962] without a proof, but this combinatorial interpretation of the Witt formula can be reinterpreted as a coloring problem of a necklace with N beads with colors chosen out of a set of R colors such that the colored beads form a nonperiodic configuration. In other words, $\mathcal{M}(m_1, \ldots, m_R)$ is the number of nonperiodic colored necklaces composed of m_i occurrences of the color i, $i = 1, \ldots, R$.

Sherman [1962] called attention to this association of identities (1-1) and (1-2) to paths in the same graph, which motivated him to consider the problem of finding a relation between (1-1) and Lie algebras. To interpret (1-1) in algebraic terms means to relate the exponents $N_\pm$ to some Lie algebraic data.

An investigation of Sherman's problem was initiated in [da Costa 1997; da Costa and Variane 2005] and a solution obtained for the univariate case of identity (1-1). In the present paper we solve the problem in the multivariate formal case, which requires important improvements. The counting method developed in [da Costa 1997; da Costa and Variane 2005] is based on a sign formula for a path given in terms of data encoded in the word representation for the path. It played a crucial role in getting formulas for $N_\pm$ in the univariate case. However, the counting method based on this sign formula is complicated. In the present paper we make improvements in the counting method in order to apply it to the multivariate case without depending too much on the sign formula. The formula is used here only to prove a simple lemma.

S-J. Kang and M-H. Kim [1999] derived dimension formulas for the homogeneous spaces of general free graded Lie algebras. We use some of their results to solve Sherman's problem. At the same time our results give a combinatorial realization for some of theirs in terms of paths in a graph.

The paper is organized as follows. In Section 2, we recall the word representation of a path and some basic definitions. We prove a basic lemma about the distribution

of signs in the set of words of a given length. In Section 3, we compute formulas for the number of equivalence classes of closed nonperiodic paths of given length. The first of these generalizes Witt's formula in the sense that it counts paths that traverse the edges of the graph without backtracking. The other formulas give the exponents in Sherman's identity (1-1). We also interpret these formulas in terms of a coloring problem. Sherman's problem, that is, to give an algebraic meaning to the exponents in (1-1) is solved in Section 4.

2. Preliminaries

A path in G_R is an ordered sequence of the edges which does not necessarily respect their orientation. A path is closed and subjected to the constraint that it never goes immediately backwards over a previous edge.

Given $G_r \subseteq G_R$, denote by $i_1, \ldots, i_r$ an enumeration of the edges of G_r in increasing order. A closed path of length $N \geq r$ in G_r is best represented by a word of the form

$$D_{j_1}^{e_{j_1}} D_{j_2}^{e_{j_2}} \cdots D_{j_l}^{e_{j_l}} \tag{2-1}$$

where $l = r, r+1, \ldots, N$, $j_k \in \{i_1, \ldots, i_r\}$, $j_k \neq j_{k+1}$, $j_l \neq j_1$, and

$$\sum_{k=1}^{l} |e_{j_k}| = N.$$

All edges of G_r are traversed by a path such that each i_k appears at least once in the sequence $S_l = (j_1, j_2, \ldots, j_l)$. The order in which the symbols $D_j^{e_j}$ appear in the word indicates the edges traversed by p and in which order. If the sign of e_j is positive, the path traverses edge j exactly e_j times following the edge's orientation; if negative, it traverses the edge $|e_j|$ times in the opposite direction.

A word is called *periodic* if it equals

$$(D_{j_1}^{e_{j_1}} D_{j_2}^{e_{j_2}} \cdots D_{j_\alpha}^{e_{j_\alpha}})^g$$

for some $g > 1$. The number g is called the *period* of the word if the word in parentheses is nonperiodic.

Permuting the symbols $D_j^{e_j}$ in (2-1) cyclically, one gets l words that represent the same closed path. (For example, $D_1^{-2} D_2^{+1} D_1^{+1} D_2^{+3}$ is a cyclic permutation of $D_2^{+1} D_1^{+1} D_2^{-3} D_1^{-2}$.) Words obtained from one another by a cyclic permutation are taken to be equivalent for this reason. Although the word (2-1) and its inverse

$$D_{j_l}^{-e_{j_l}} \cdots D_{j_1}^{-e_{j_1}}$$

also represent the same path, they are not taken as equivalent here. This is the reason for the exponent 2 on the right side of (1-1), also present in [Sherman 1962].

In Section 3 we consider signed paths. The sign of a path is given by the formula

$$\text{sign}(p) = (-1)^{1+n(p)}, \tag{2-2}$$

where $n(p)$ is the number of integral revolutions of the tangent vector of p. From this definition it follows that if $p = (h)^g$ is a periodic path with odd period g, then $\text{sign}(p) = \text{sign}(h)$. If g is even, $\text{sign}(p) = -1$. The sign of a path can be computed from its word representation (2-1) using the formula [da Costa and Variane 2005]

$$(-1)^{N+l+T+s+1}, \tag{2-3}$$

where T is the number of subsequences in the decomposition of S_l into subsequences (see [da Costa and Variane 2005] for the definition and an example of a decomposition) and s is the number of negative exponents in (2-1). It follows from the previous sign formulas that periodic words with even period have negative sign.

The following lemma is important in the proof of several results in Section 3. It was assumed in [da Costa 1997; da Costa and Variane 2005] without a proof.

Lemma 2.1. *Given $G_r \subseteq G_R$, consider all paths that traverse each edge of G_r at least once (no backtracking allowed) and the set of all representative words (periodic or not, cyclic permutations and inversions included) of fixed length $N \geq r > 1$. Then half of the words have positive sign and the other half have negative sign.*

Proof. It suffices to consider the subset of words associated to a fixed sequence $S_l = (j_1, j_2, \ldots, j_l)$. For this sequence the numbers N, l, and T are fixed. The words with these numbers have signs which depend only on $s \in \{0, 1, 2, \ldots, l\}$. For $N + l + T$ even, the sign of a word is $(-1)^{s+1}$. If $l = 2k$, then, for each odd value of s, there are

$$\binom{2k}{s}$$

words with positive sign. Summing over the odd values of s, we get the total number of 2^{2k-1} words with positive sign. Summing over the even values of s, we get the same number of words with negative sign. If $l = 2k + 1$, a similar counting gives 2^{2k} words with positive (negative) signs. The case $N + l + T$ odd is analogous. □

3. Counting paths in G_r

Fix a subgraph $G_r \subseteq G_R$. Given distinct edges $i_1, \ldots, i_r$ in G_r and positive integers $m_{i_1}, \ldots, m_{i_r}$, with $m_{i_1} + \cdots + m_{i_r} = N > r$, let $\theta_\pm(m_{i_1}, \ldots, m_{i_r})$ be the number of equivalence classes of closed nonperiodic paths of length N with $\pm$ signs that traverse each edge i_j exactly m_{i_j} times, for $j = 1, \ldots, r$, with no backtracking, and traverse the edges in $G_R \backslash G_r$ zero times. In this section we derive formulas for $\theta := \theta_+ + \theta_-$ and $\theta_\pm$. Notice that $\theta_\pm$ is just another name for the exponents $N_\pm$ in (1-1) showing only the nonzero entries in $N_\pm$.

Firstly, we compute θ. In the case $r = 1$, a path with $m_i > 1$ is periodic. The nonperiodic ones are two, the path with length $N = 1$ and its inversion so that $\theta(m_i) = 0$ if $m_i > 1$ and $\theta(m_i) = 2$ if $m_i = 1$. In other cases, θ is given as follows.

Theorem 3.1. *For $r = 2$, define*

$$\mathscr{F}\left(\frac{m_{i_1}}{g}, \frac{m_{i_2}}{g}\right) = \sum_{a=1}^{M/g} \frac{2^{2a}}{a} \binom{m_{i_1}/g - 1}{a-1} \binom{m_{i_2}/g - 1}{a-1}, \tag{3-1}$$

where $M = \min\{m_{i_1}, m_{i_2}\}$ and, if $r \geq 3$,

$$\mathscr{F}\left(\frac{m_{i_1}}{g}, \ldots, \frac{m_{i_r}}{g}\right) = \sum_{a=r}^{N/g} \frac{2^a}{a} \sum_{\{S_a\}} \prod_{c=1}^{r} \binom{m_{i_c}/g - 1}{t_{i_c} - 1} \tag{3-2}$$

where $\{S_a\}$ is the set of sequences $(j_1, \ldots, j_a)$ such that $j_k \in \{i_1, \ldots, i_r\}$ and $j_k \neq j_{k+1}$, $j_a \neq j_1$. Number t_{i_c} counts how many times edge i_c occurs in a sequence S_a. Use is made of the convention that the combination symbol in (3-2) *is zero whenever $t_{i_c} > m_{i_c}/g$. Then*

$$\theta(m_{i_1}, \ldots, m_{i_r}) = \sum_{g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathscr{F}\left(\frac{m_{i_1}}{g}, \ldots, \frac{m_{i_r}}{g}\right). \tag{3-3}$$

The summation is over all the common divisors g of $m_{i_1}, \ldots, m_{i_r}$, and $\mu(g)$ is the Möbius function.

Proof. The number $\mathscr{H}(l, m_{i_1}, \ldots, m_{i_r})$ of words that have the same values of $m_{i_1}, \ldots, m_{i_r}$ and $l \in \{r, r+1, \ldots, N\}$ is given by

$$\mathscr{H}(l, m_{i_1}, \ldots, m_{i_r}) = 2^l \sum_{\{S_l\}} \prod_{c=1}^{r} \binom{m_{i_c} - 1}{n_{i_c} - 1}.$$

Let's explain this formula a bit. The number n_{i_c} counts the number of occurrences of edge i_c in a sequence $S_l = (j_1, \ldots, j_l)$. The combination symbol counts the number of unrestricted partitions of m_{i_c} into n_{i_c} nonzero positive parts [Andrews 1976]; thus the product times 2^l (there are 2^l ways of assigning $\pm$ signs to the exponents in (2-1)) gives the total number of words representing paths traversing each edge i_j of $G_r \subseteq G_R$ exactly m_{i_j} times in all possible ways. Then we sum over all sequences S_l with the convention that a combination symbol equals zero when $m < n$.

In the set of $\mathscr{H}(l, m_{i_1}, \ldots, m_{i_r})$ words, there is the subset of nonperiodic words plus their cyclic permutations and inversions, and the subset of periodic words, if any, whose periods are the common divisors of $l, m_{i_1}, \ldots, m_{i_r}$ plus their cyclic permutations and inversions. Denote by $\overline{\mathscr{H}(l, m_{i_1}, \ldots, m_{i_r})}$ the number of elements

in the former set. The words with period g are of the form

$$(D_{k_1}^{e_{k_1}} D_{k_2}^{e_{k_2}} \cdots D_{k_\alpha}^{e_{k_\alpha}})^g$$

where $\alpha = l/g$ and $D_{k_1}^{e_{k_1}} D_{k_2}^{e_{k_2}} \cdots D_{k_\alpha}^{e_{k_\alpha}}$ is nonperiodic so that the number of periodic words with period g plus their cyclic permutations and inversions is given by $\overline{\mathcal{H}(l/g, m_{i_1}/g, \ldots, m_{i_r}/g)}$. Therefore,

$$\mathcal{H}(l, m_{i_1}, \ldots, m_{i_r}) = \sum_{g|l,k,m_{i_1},\ldots,m_{i_r}} \overline{\mathcal{H}\left(\frac{l}{g}, \frac{m_{i_1}}{g}, \ldots, \frac{m_{i_r}}{g}\right)}.$$

The summation is over all the common divisors g of $l, m_{i_1}, \ldots, m_{i_r}$.

Applying the Möbius inversion formula [Apostol 1976], it follows that

$$\overline{\mathcal{H}(l, m_{i_1}, \ldots, m_{i_r})} = \sum_{g|(l,m_{i_1},\ldots,m_{i_r})} \mu(g)\mathcal{H}\left(\frac{l}{g}, \frac{m_{i_1}}{g}, \ldots, \frac{m_{i_r}}{g}\right). \tag{3-4}$$

where μ is the Möbius function. To eliminate cyclic permutations divide (3-4) by l. Summing over all possible values of l one gets

$$\theta(m_{i_1}, \ldots, m_{i_r}) = \sum_{l=r}^{N} \frac{\overline{\mathcal{H}(l, m_{i_1}, \ldots, m_{i_r})}}{l}. \tag{3-5}$$

Upon substitution of (3-4) into (3-5) one gets, for the case $r \geq 3$,

$$\theta(m_{i_1}, \ldots, m_{i_r}) = \sum_{l=r}^{N} \frac{1}{l} \sum_{g|(l,m_{i_1},\ldots,m_{i_r})} \mu(g) 2^{l/g} \sum_{\{S_{l/g}\}} \prod_{c=1}^{r} \binom{m_{i_c}/g - 1}{n_{i_c}/g - 1}.$$

Proceed now as follows. For a given common divisor g of $m_{i_1}, \ldots, m_{i_r}$, sum over all values of l which are multiples of g. Then sum over all possible divisors of $m_{i_1}, \ldots, m_{i_r}$. Write $l = ag$, and $n = tg$. If $r \geq 3$, one has $r/g \leq a \leq N/g$, but, unless $g = 1$, it is not admissible to have $a < r$, because all r edges of the graph should be traversed. For this reason, $r \leq a \leq N/g$. Result (3-2) follows. If $r = 2$, l is even and, for each l, only sequences of the form $(i_1, i_2, \ldots, i_1, i_2)$ with $n_{i_1} = n_{i_2} = l/2$ are possible. Put $l = 2a$, $a = 1, 2, \ldots, M = \min\{m_1, m_2\}$ to get (3-1). □

Example 1. From (3-1), we have

$$\mathcal{F}(1, 1) = \mathcal{F}(1, 2) = \mathcal{F}(2, 1) = \mathcal{F}(1, 3) = \mathcal{F}(3, 1) = 4,$$

$$\mathcal{F}(2, 2) = 12,$$

$$\mathcal{F}(1, 4) = \mathcal{F}(4, 1) = \mathcal{F}(1, 5) = \mathcal{F}(5, 1) = 4,$$

$$\mathcal{F}(2, 3) = \mathcal{F}(3, 2) = 20,$$

$$\mathcal{F}(2, 4) = \mathcal{F}(4, 2) = 28,$$

$$\mathcal{F}(3, 3) = 172/3.$$

From (3-3),

$$\theta(1,1)=\theta(1,2)=\theta(2,1)=\theta(1,3)$$
$$=\theta(3,1)=\theta(1,4)=\theta(4,1)=\theta(1,5)=\theta(5,1)=4,$$
$$\theta(2,2)=10,\quad \theta(2,3)=\theta(3,2)=20,\quad \theta(3,3)=56.$$

Example 2. From (3-2),

$$\mathcal{F}(1,1,1)=16,$$
$$\mathcal{F}(1,1,2)=\mathcal{F}(1,2,1)=\mathcal{F}(2,1,1)=32,$$
$$\mathcal{F}(1,2,2)=\mathcal{F}(2,1,2)=\mathcal{F}(2,2,1)=112,$$
$$\mathcal{F}(1,1,3)=\mathcal{F}(1,3,1)=\mathcal{F}(3,1,1)=48,$$
$$\mathcal{F}(1,1,4)=\mathcal{F}(1,4,1)=\mathcal{F}(4,1,1)=64,$$
$$\mathcal{F}(1,2,3)=\mathcal{F}(3,1,2)=\mathcal{F}(2,3,1)=\mathcal{F}(3,2,1)=\mathcal{F}(1,3,2)=\mathcal{F}(2,1,3)=256,$$
$$\mathcal{F}(2,2,2)=1056.$$

From (3-3),

$$\theta(1,1,1)=16,$$
$$\theta(1,1,2)=\theta(2,1,1)=\theta(1,2,1)=32,$$
$$\theta(1,2,2)=\theta(2,1,2)=\theta(2,2,1)=112,$$
$$\theta(1,1,3)=\theta(3,1,1)=\theta(1,3,1)=48,$$
$$\theta(1,1,4)=\theta(4,1,1)=\theta(1,4,1)=64,$$
$$\theta(1,2,3)=\theta(3,1,2)=\theta(2,3,1)=\theta(3,2,1)=\theta(1,3,2)=\theta(2,1,3)=256,$$
$$\theta(2,2,2)=1048.$$

Remark. (a) Notice that θ, and likewise the Witt formula, is given in terms of the Möbius function. However, formula (3-3) counts closed nonperiodic paths traversing the edges of G_R in all directions (without backtracking) and, in that sense, generalizes the Witt formula. Also, our formula has an algebraic meaning of a dimension. See Section 4.

(b) If $m_{i_1},\dots,m_{i_r}$ are coprime, $\mathcal{F}=\theta$. Otherwise, $\mathcal{F}$ can be rational. For instance, $\mathcal{F}(3,3)=172/3$. But $\mathcal{F}':=N\mathcal{F}$, $N=m_{i_1}+\cdots+m_{i_r}$, is always a positive integer which counts the number of words of length N. For example, in the case $N=4$, $m_1=m_2=2$, $\mathcal{F}'=48$. The words are

$$D_1^{\pm2}D_2^{\pm2},\quad D_1^{-1}D_2^{+1}D_1^{+1}D_2^{+1},\quad D_1^{+1}D_2^{-1}D_1^{+1}D_2^{+1},\quad D_1^{-1}D_2^{-1}D_1^{+1}D_2^{+1},$$
$$D_1^{-1}D_2^{+1}D_1^{+1}D_2^{-1},\quad D_1^{-1}D_2^{-1}D_1^{-1}D_2^{+1},\quad D_1^{-1}D_2^{-1}D_1^{+1}D_2^{-1},$$

plus four cyclic permutations for each of them, and the four periodic words $(D_1^{\pm 1} D_2^{\pm 1})^2$ plus two cyclic permutations for each.

In terms of $\mathcal{F}'$,

$$\theta(m_{i_1}, \ldots, m_{i_r}) = \frac{1}{N} \sum_{g | m_{i_1}, \ldots, m_{i_r}} \mu(g) \mathcal{F}'\left(\frac{m_{i_1}}{g}, \ldots, \frac{m_{i_r}}{g}\right).$$

Although the Möbius function is negative for some divisors g, the right hand side is nevertheless always a positive number because $\mathcal{F}'(m_{i_1}/g, \ldots, m_{i_r}/g)$ counts words in a subset of the words counted by $\mathcal{F}'(m_{i_1}, \ldots, m_{i_r})$.

(c) Given a circular necklace with N beads, consider the problem of counting inequivalent nonperiodic colorings of these beads with $2r$ colors $\{c_i, \bar{c}_i\}$, $i = 1, \ldots, r$, with m_i occurrences of the index i, $N = \sum m_i$, with the restriction that no two colors c_i and $\bar{c}_i$ (same index) occur adjacent in a coloring. Now, consider an oriented graph with r loops hooked to a single vertex. Each loop edge corresponds to a color c_i. A nonperiodic closed nonbacktracking path of length N in the graph corresponds to a coloring, and a color $\bar{c}_i$ corresponds to an edge being traversed in the opposite orientation. The presence of a single vertex in the graph reflects the fact that adjacent to a bead with, say, color c_i, any other with distinct index may follow. The number of inequivalent colorings is given by θ.

As a basic test of our counting ideas, we prove Sherman's statement [1962] relating the Witt formula to paths in G_R.

Proposition 3.2. *Relative to graph G_R, formula* (1-2) *gives the number $\mathcal{M}$ of equivalence classes of closed nonperiodic paths of length $N > 0$ that traverse each edge i counterclockwise $m_i \geq 0$ times ($i = 1, 2, \ldots, R$), where $m_1 + \cdots + m_R = N$.*

Proof. Denote by $m_{i_1}, \ldots, m_{i_r}$, $r \leq R$, the nonzero entries in $\mathcal{M}(m_1, \ldots, m_R)$, which we call $\mathcal{M}_r(m_{i_1}, \ldots, m_{i_r})$. Words representing counterclockwise paths have positive exponents so that the factors 2^{2a} and 2^a in formulas (3-1) and (3-2) are not needed. Hence

$$\mathcal{M}_r(m_{i_1}, \ldots, m_{i_r}) = \sum_{g | m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathcal{F}_c\left(\frac{m_{i_1}}{g}, \ldots, \frac{m_{i_r}}{g}\right) \tag{3-6}$$

where

$$\mathcal{F}_c\left(\frac{m_{i_1}}{g}, \frac{m_{i_2}}{g}\right) = \sum_{a=1}^{M/g} \frac{1}{a} \binom{m_{i_1}/g - 1}{a - 1} \binom{m_{i_2}/g - 1}{a - 1} \quad \text{if } r = 2, \tag{3-7}$$

with $M = \min\{m_{i_1}, m_{i_2}\}$, and

$$\mathscr{F}_c\left(\frac{m_{i_1}}{g},\ldots,\frac{m_{i_r}}{g}\right)=\sum_{a=r}^{N/g}\frac{1}{a}\sum_{\{S_a\}}\prod_{c=1}^{r}\binom{m_{i_c}/g-1}{t_{i_c}-1}\quad\text{if } r\geq 3. \tag{3-8}$$

In the case $r=2$ suppose $m_{i_1}\leq m_{i_2}$. Using formula (A-3) from the Appendix (with $l=2$), it follows that

$$\begin{aligned}\sum_{a=1}^{m_{i_1}/g}\frac{1}{a}\binom{m_{i_1}/g-1}{a-1}\binom{m_{i_2}/g-1}{a-1}&=\frac{g}{m_{i_2}}\binom{m_{i_1}/g+m_{i_2}/g-1}{m_{i_1}/g}\\&=\frac{(N/g)!}{(N/g)(m_{i_1}/g)!(m_{i_2}/g)!}.\end{aligned}$$

Similarly if $m_{i_2}\leq m_{i_1}$. In the case $r\geq 3$ define

$$I=\sum_{\substack{m_i>0\\ m_{i_1}+\cdots+m_{i_r}=N}}\mathscr{F}_c\left(\frac{m_{i_1}}{g},\ldots,\frac{m_{i_r}}{g}\right). \tag{3-9}$$

Upon substituting (3-8) into (3-9) and exchanging the summation symbols, we get

$$I=\sum_{a=r}^{N/g}\frac{1}{a}\sum_{\{S_a\}}\sum_{\substack{m_i>0\\ m_{i_1}+\cdots+m_{i_r}=N}}\prod_{c=1}^{r}\binom{\frac{m_{i_c}}{g}-1}{t_{i_c}-1}.$$

Applying Lemma A.2,

$$I=\sum_{a=r}^{N/g}\frac{1}{a}\sum_{\{S_a\}}\binom{N/g-1}{a-1}=\sum_{a=r}^{N/g}\frac{1}{a}\binom{N/g-1}{a-1}rw_r(a)$$

where

$$rw(a)=\sum_{j=1}^{r}(-1)^{r+j}\binom{r}{j}(j-1)^a+(-1)^{a+r}$$

is the number of sequences in $\{S_a\}$ [da Costa and Variane 2005]. Using that

$$\sum_{a=r}^{N/g}\frac{1}{a}\binom{N/g-1}{a-1}(j-1)^a=\frac{g}{N}(j^{N/g}-1)$$

and

$$\sum_{a=r}^{N/g}\frac{1}{a}\binom{N/g-1}{a-1}(-1)^{a+r}=(-1)^{r+1}\frac{g}{N},$$

we get

$$I=\frac{g}{N}\sum_{j=1}^{r}(-1)^{r+j}\binom{r}{j}j^{N/g}. \tag{3-10}$$

The Stirling numbers $S(N/g, r)$ of the second kind are given by [Chen and Koh 1992]

$$S\Big(\frac{N}{g}, r\Big) = \frac{1}{r!}\sum_{k=0}^{r}(-1)^k\binom{r}{k}(r-k)^{N/g} = \frac{1}{r!}\sum_{j=0}^{r}(-1)^{r+j}\binom{r}{j}j^{N/g} \tag{3-11}$$

so that

$$I = r!\frac{g}{N}S\Big(\frac{N}{g}, r\Big). \tag{3-12}$$

Stirling numbers have the property that

$$\sum_{\substack{m_i>0\\ m_{i_1}+\cdots+m_{i_r}=N}} \frac{(N/g)!}{(m_{i_1}/g)!\cdots(m_{i_r}/g)!} = r!S\Big(\frac{N}{g}, r\Big). \tag{3-13}$$

Comparing relations (3-12), (3-13), and (3-9),

$$\mathcal{F}_c\left(\frac{m_{i_1}}{g}, \ldots, \frac{m_{i_r}}{g}\right) = \frac{g}{N}\frac{(N/g)!}{(m_1/g)!\cdots(m_r/g)!}. \tag{3-14}$$

Upon substitution of (3-14) into (3-6), the result follows.

In the following we compute formulas for θ_+ and θ_-.

Theorem 3.3. *Suppose any of the following conditions is satisfied*:

(a) $N = m_{i_1} + \cdots + m_{i_r} < 2r$.

(b) $m_{i_1}, \ldots, m_{i_r}$ *are coprime.*

(c) $m_{i_1}, \ldots, m_{i_r}$ *are neither all odd nor all even.*

(d) $m_{i_1}, \ldots, m_{i_r}$ *are all odd.*

Then

$$\theta_-(m_{i_1}, \ldots, m_{i_r}) = \theta_+(m_{i_1}, \ldots, m_{i_r}). \tag{3-15}$$

Proof. The proof is similar to that of [da Costa 1997, Theorem 1] and uses Lemma 2.1. □

Theorem 3.4. *The number* $\theta_+(m_{i_1}, \ldots, m_{i_r})$ *is given by*

$$\theta_+(m_{i_1}, \ldots, m_{i_r}) = \sum_{\text{odd } g\,|\,m_{i_1},\ldots,m_{i_r}} \frac{\mu(g)}{g}\mathcal{G}\left(\frac{m_{i_1}}{g}, \ldots, \frac{m_{i_r}}{g}\right) \tag{3-16}$$

where the summation is over all the common odd divisors of $m_{i_1}, \ldots, m_{i_r}$, *and* $\mathcal{G} = \mathcal{F}/2$ *with* $\mathcal{F}$ *as in* (3-1) *and* (3-2). *If* $m_{i_1}, \ldots, m_{i_r}$ *are all even numbers, then*

$$\theta_-(m_{i_1}, \ldots, m_{i_r}) = \theta_+(m_{i_1}, \ldots, m_{i_r}) - \theta_+\left(\frac{m_{i_1}}{2}, \ldots, \frac{m_{i_r}}{2}\right). \tag{3-17}$$

Proof. First, suppose that all common divisors of $m_{i_1}, \ldots, m_{i_r}$ are odd numbers. In this case,

$$\theta(m_{i_1}, \ldots, m_{i_r}) = \sum_{\text{odd } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathscr{F}_\Phi\left(\frac{m_{i_1}}{g}, \ldots, \frac{m_{i_r}}{g}\right).$$

Since $\theta = \theta_+ + \theta_-$ and $\theta_+ = \theta_-$ (Theorem 3.3) it follows that $\theta = 2\theta_+$, hence

$$\theta_+ = \tfrac{1}{2} \sum_{\text{odd } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathscr{F}_\Phi\left(\frac{m_{i_1}}{g}, \ldots, \frac{m_{i_r}}{g}\right). \tag{3-18}$$

If the numbers $m_{i_1}, \ldots, m_{i_r}$ are all even, θ_+ is again given by (3-18) because, in this case, the m_i's have common divisors which are even numbers, but since periodic words with even period have negative sign, only the odd divisors are relevant to getting θ_+. The reason one should have the factor $\frac{1}{2}$ is that, by Lemma 2.1, when one considers the set of all possible words representing paths of a given length which traverse the edges of G_r $m_{i_1}, \ldots, m_{i_r}$ times, half of them have positive sign and the other half have negative sign. To account for the positive half, one needs the factor $\frac{1}{2}$. Let's now compute θ_- in the even case. Write

$$\begin{aligned}
\theta &= \sum_{\text{odd } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathscr{F}_\Phi + \sum_{\text{even } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathscr{F}_\Phi \\
&= \tfrac{1}{2} \sum_{\text{odd } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathscr{F}_\Phi + \tfrac{1}{2} \sum_{\text{odd } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathscr{F}_\Phi + \sum_{\text{even } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathscr{F}_\Phi \\
&= 2\theta_+ + \sum_{\text{even } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathscr{F}_\Phi.
\end{aligned}$$

Using that $\theta = \theta_+ + \theta_-$, we obtain

$$\theta_- = \theta_+ + \sum_{\text{even } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathscr{F}_\Phi.$$

Now the relevant even divisors are $\{2n\}$ where n are the odd common divisors of $\{m_i\}$. For the other possible divisors, if any, use that $\mu(2^j n) = 0$, $j \geq 2$. Using the equality $\mu(2n) = -\mu(n)$, we see that the summation over the even divisors is equal to $-\theta_+\big(m_{i_1}/2, \ldots, m_{i_r}/2\big)$, proving the result. □

Remark. Like θ, the numbers $\theta_\pm$ can be interpreted as the number of inequivalent nonperiodic colorings of a circular necklace with N beads. However, now these colorings are classified as positive or negative according to formula (2-3). It is positive (negative) if the number $N + l + T + s$ is odd (even). In this case, s is the number $\bar{c}$ of colors present in a coloring. Interpret T in terms of the color indices.

Definition. Let $s_1, \ldots, s_r$ be arbitrary positive integers. Let the number $\mathcal{P}$ be defined as follows. If $s_1, \ldots, s_r$ are all even numbers,

$$\mathcal{P}(s_1, \ldots, s_r) = \sum_{\text{even } g \mid s_1, \ldots, s_r} \frac{\mu(g)}{g} \mathcal{G}\left(\frac{s_1}{g}, \ldots, \frac{s_r}{g}\right), \tag{3-19}$$

otherwise, $\mathcal{P}(s_1, \ldots, s_r) = 0$. Also, define

$$\mathcal{H}(s_1, \ldots, s_r) = \begin{cases} \mathcal{G}(s_1, \ldots, s_r) & \text{if } s_1, \ldots, s_r \text{ not all even,} \\ \mathcal{G}(s_1, \ldots, s_r) - \displaystyle\sum_{k \mid s_1, \ldots, s_r} \frac{1}{k} \mathcal{P}\left(\frac{s_1}{k}, \ldots, \frac{s_r}{k}\right) & \text{otherwise.} \end{cases} \tag{3-20}$$

Lemma 3.5. $$\mathcal{P} = \sum_{g \mid s_1, \ldots, s_r} \frac{\mu(g)}{g} (\mathcal{G} - \mathcal{H}).$$

Proof. From the definition, $\mathcal{G} = \mathcal{H}$ if $s_1, \ldots, s_r$ are not all even. Otherwise,

$$\mathcal{G} - \mathcal{H} = \sum_{g \mid s_1, \ldots, s_r} \frac{1}{g} \mathcal{P}\left(\frac{s_1}{g}, \ldots, \frac{s_r}{g}\right). \tag{3-21}$$

Now apply Lemma A.1 to get the result. □

Theorem 3.6. $$\theta_+(m_{i_1}, \ldots, m_{i_r}) = \sum_{g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathcal{H}\left(\frac{m_{i_1}}{g}, \ldots, \frac{m_{i_r}}{g}\right).$$

Proof. When $m_{i_1}, \ldots, m_{i_r}$ are not all even, their odd divisors are the only possible common divisors. In this case, $\mathcal{P} = 0$ and

$$\theta_+ = \sum_{\text{odd } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathcal{H}, \tag{3-22}$$

with $\mathcal{H} = \mathcal{G}$. If $m_{i_1}, \ldots, m_{i_r}$ are all even, the sum over odd divisors of $m_{i_1}, \ldots, m_{i_r}$ can be expressed as

$$\begin{aligned} \theta_+ &= \sum_{\text{odd } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathcal{G} \\ &= \sum_{g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathcal{G} - \sum_{\text{even } g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathcal{G} = \sum_{g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathcal{G} - \mathcal{P} \\ &= \sum_{g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathcal{G} - \sum_{g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} (\mathcal{G} - \mathcal{H}) = \sum_{g \mid m_{i_1}, \ldots, m_{i_r}} \frac{\mu(g)}{g} \mathcal{H}. \end{aligned}$$ □

Example 3.

$$\theta_\pm(1,1)=\theta_\pm(1,2)=\theta_\pm(2,1)=\theta_\pm(1,3)=\theta_\pm(3,1)=\theta_\pm(1,4)=\theta_\pm(4,1)$$
$$=\theta_\pm(1,5)=\theta_\pm(5,1)=2,$$
$$\theta_+(2,2)=6, \quad \theta_-(2,2)=4, \quad \theta_\pm(2,3)=\theta_\pm(3,2)=10, \quad \theta_+(2,4)=14,$$
$$\theta_-(2,4)=12, \quad \theta_+(4,2)=14, \quad \theta_-(4,2)=12, \quad \theta_\pm(3,3)=28.$$

Example 4.

$$\theta_\pm(1,1,1)=8,$$
$$\theta_\pm(1,1,2)=\theta_\pm(2,1,1)=\theta_\pm(1,2,1)=16,$$
$$\theta_\pm(1,2,2)=\theta_\pm(2,1,2)=\theta_\pm(2,2,1)=56,$$
$$\theta_\pm(1,1,3)=\theta_\pm(3,1,1)=\theta_\pm(1,3,1)=24,$$
$$\theta_\pm(1,1,4)=\theta_\pm(4,1,1)=\theta_\pm(1,4,1)=32,$$
$$\theta_\pm(1,2,3)=\theta_\pm(3,1,2)=\theta_\pm(2,3,1)=\theta_\pm(3,2,1)=\theta_\pm(1,3,2)=\theta_\pm(2,1,3)=128,$$
$$\theta_+(2,2,2)=524,$$
$$\theta_-(2,2,2)=516.$$

4. Sherman identity and Lie algebras

In this section we relate our previous results with Lie algebras and solve Sherman's problem. The solution is provided by the following proposition.

Proposition 4.1 [Kang and Kim 1999]. *Let* $V=\bigoplus_{(k_1,\dots,k_r)\in\mathbb{Z}^r_{>0}} V_{(k_1,\dots,k_r)}$ *be a* $\mathbb{Z}^r_{>0}$*-graded vector space over* $\mathbb{C}$ *with* $\dim V_{(k_1,\dots,k_r)}=d(k_1,\dots,k_r)<\infty$, *for all* $(k_1,\dots,k_r)\in\mathbb{Z}^r_{>0}$, *and let*

$$L=\bigoplus_{(k_1,\dots,k_r)\in\mathbb{Z}^r_{>0}} L_{(k_1,\dots,k_r)}$$

be the free Lie algebra generated by V. *Then the dimensions of the subspaces* $L_{(k_1,\dots,k_r)}$ *are given by*

$$\dim L_{(k_1,\dots,k_r)}=\sum_{g\mid(k_1,\dots,k_r)}\frac{\mu(g)}{g}\,{}^{\circ}\!\mathcal{W}\left(\frac{k_1}{g},\dots,\frac{k_r}{g}\right) \tag{4-1}$$

where summation is over all common divisors g *of* $k_1,\dots,k_r$ *and* ${}^{\circ}\!\mathcal{W}$ *is given by*

$${}^{\circ}\!\mathcal{W}(k_1,\dots,k_r)=\sum_{s\in T(k_1,\dots,k_r)}\frac{(|s|-1)!}{s!}\prod_{i_1,\dots,i_r=1}^{\infty} d(i_1,\dots,i_r)^{s_{i_1,\dots,i_r}}. \tag{4-2}$$

The exponents $s_{i_1,\ldots,i_r}$ *are the components of* $s \in T$,

$$T(k_1,\ldots,k_r) = \{s = (s_{i_1,\ldots,i_r}) | s_{i_1,\ldots,i_r} \in \mathscr{L}_{\geq 0}, \sum_{i_1,\ldots,i_r=1}^{\infty} s_{i_1,\ldots,i_r}(i_1,\ldots,i_r) = (k_1,\ldots,k_r)\}, \tag{4-3}$$

and

$$|s| = \sum_{i_1,\ldots,i_r=1}^{\infty} s_{i_1,\ldots,i_r}, \quad s! = \prod_{i_1,\ldots,i_r=1}^{\infty} s_{i_1,\ldots,i_r}!. \tag{4-4}$$

Moreover, the numbers $\dim L_{(k_1,\ldots,k_r)}$ *satisfy*

$$\prod_{k_1,\ldots,k_r=1}^{\infty} (1 - z_1^{k_1}\cdots z_r^{k_r})^{\dim L_{(k_1,\ldots,k_r)}} = 1 - f(z_1,\ldots,z_r) \tag{4-5}$$

where

$$f(z_1,\ldots,z_r) := \sum_{k_1,\ldots,k_r=1}^{\infty} d(k_1,\ldots,k_r) z_1^{k_1}\cdots z_r^{k_r}. \tag{4-6}$$

This function is associated with the generating function of the $\mathscr{W}$*'s,*

$$g(z_1,\ldots,z_r) := \sum_{k_1,\ldots,k_r=1}^{\infty} \mathscr{W}(k_1,\ldots,k_r) z_1^{k_1}\cdots z_r^{k_r}, \tag{4-7}$$

by the relation

$$e^{-g} = 1 - f. \tag{4-8}$$

□

Identity (4-5) is a consequence of the famous *Poincaré–Birkhoff–Witt theorem* for the free Lie algebra. Computation of the formal logarithm of the left-hand side of (4-5) and its expansion gives that the infinite product equals the exponential in (4-8). Raise both members of (4-5) to the power -1, compute the formal logarithm of both members, and expand them. Identification of the coefficients of the same order, definition (4-2), and application of the Möbius inversion give (4-1). See [Kang and Kim 1999] for details. In [da Costa and Variane 2005], (4-1) is called the *generalized Witt formula*, $\mathscr{W}$ is called the *Witt partition function*, and (4-5) the *generalized Witt identity*.

Formulas (3-3) and (3-20) have exactly the form of (4-1) with corresponding *Witt partition functions* given by $\mathscr{F}$ and $\mathscr{H}$, respectively, so we interpret θ and θ_+ as giving the dimensions of the homogeneous spaces of graded Lie algebras. In each case, the algebra is generated by a graded vector space whose dimensions can be computed recursively from (4-2) as a function of the Witt partition function.

However, a general formula can be obtained from (4-8) using (4-6) as the formal Taylor expansion of $1-e^{-g}$. This gives

$$d(k_1,\dots,k_r)=\frac{1}{k_1!\cdots k_r!}\frac{\partial^{|k|}}{\partial z_1^{k_1}\cdots\partial z_r^{k_r}}(1-e^{-g})|_{z_1=\cdots=z_r=0} \tag{4-9}$$

with

$$g(z_1,\dots,z_r):=\sum_{k_1,\dots,k_r=1}^{\infty}\mathcal{W}(k_1,\dots,k_r)z^{k_1}\cdots z^{k_r} \tag{4-10}$$

and $\mathcal{W}=\mathcal{F}$, $\mathcal{H}$ given by (3-1), (3-2), and (3-20). Furthermore, dim $L_{(k_1,\dots,k_r)}=\theta$, θ_+ given by (3-3), 3.6 satisfy the *generalized Witt identity* (4-5) with the corresponding dimensions given by (4-9). In fact, an explicit formula for (4-9) can be derived:

Theorem 4.2. *The numbers $d(k_1,\dots,k_r)$ are given by the formula*

$$d(k_1,\dots,k_r)=\sum_{\lambda=1}^{|k|}(-1)^{\lambda+1}\sum_{p(\lambda,k)}\prod_{i=1}^{q}\frac{[\mathcal{W}(l_{i1},\dots,l_{ir})]^{a_i}}{a_i!}, \tag{4-11}$$

where $|k|=k_1+\cdots+k_r$, $q=-1+\prod_{i=1}^{r}(k_i+1)$, $p_{\lambda,k}$ is the set of all $a_i\in\{0,1,2,\dots\}$ such that $\sum_{i=1}^{q}a_i=\lambda$, $\sum_{i=1}^{q}a_il_{ij}=k_j$, and the vectors $l_i=(l_{i1},\dots,l_{ir})$, l_{ij} satisfying $0\le l_{ij}\le k_j$, $\forall j=1,\dots,r$, $\forall i=1,\dots,q$ and $\sum_{j=1}^{r}l_{ij}>0$. Set $\mathcal{W}(l_i)=0$ if $l_{ij}=0$ for some j; otherwise, $\mathcal{W}$ is the Witt partition function.

Proof. A generalization of Faà di Bruno's relation [Constantine and Savits 1996; Savits 2006] gives a formula for the $|k|$-th derivative of the exponential of a function $g(z_1,\dots,z_r)$. From this formula and (4-9), (4-11) follows. □

Example 5. We compute $d(2,2)$ explicitly. In this case, $k_1=k_2=2$, $|k|=4$, $q=8$. The possible vectors $l\le(2,2)$ are $l_1=(0,1)$, $l_2=(1,0)$, $l_3=(1,1)$, $l_4=(0,2)$, $l_5=(2,0)$, $l_6=(2,1)$, $l_7=(1,2)$, and $l_8=(2,2)$. Next we give the values of $a_1,\dots,a_8\ge 0$ satisfying

$$\sum_{i=1}^{8}a_i=\lambda,\quad \sum_{i=1}^{8}a_il_i=(2,2).$$

Define the vector $a=(a_1,\dots,a_8)$. The possible a's for each λ are

for $\lambda=1$, $a=(0,\dots,0,1)$;

for $\lambda=2$, $a=(0,1,0,0,0,0,1,0)$, $(0,0,2,0,0,0,0,0)$,
$(0,0,0,1,1,0,0,0)$, $(1,0,0,0,0,1,0,0)$;

for $\lambda=3$, $a=(0,2,0,1,0,0,0,0)$, $(2,0,0,0,1,0,0,0)$, $(1,1,1,0,0,0,0,0)$;

for $\lambda=4$, $a=(2,2,0,0,0,0,0,0)$.

We get

$$d(2,2) = \mathcal{W}(2,2) - \tfrac{1}{2}\mathcal{W}(1,1)^2.$$

The dimensions up to $d(3,3)$ are

$$\begin{aligned}
N=2,\quad & d(1,1) = \mathcal{W}(1,1),\\
N=3,\quad & d(1,2) = \mathcal{W}(1,2), \quad d(2,1) = \mathcal{W}(2,1),\\
N=4,\quad & d(1,3) = \mathcal{W}(1,3), \quad d(3,1) = \mathcal{W}(3,1),\\
& d(2,2) = \mathcal{W}(2,2) - \tfrac{1}{2}\mathcal{W}(1,1)^2,\\
N=5,\quad & d(1,4) = \mathcal{W}(1,4), \quad d(4,1) = \mathcal{W}(4,1),\\
& d(2,3) = \mathcal{W}(2,3) - \mathcal{W}(1,1)\mathcal{W}(1,2),\\
& d(3,2) = \mathcal{W}(3,2) - \mathcal{W}(1,1)\mathcal{W}(2,1),\\
N=6,\quad & d(1,5) = \mathcal{W}(1,5), \quad d(5,1) = \mathcal{W}(5,1),\\
& d(2,4) = \mathcal{W}(2,4) - \mathcal{W}(1,1)\mathcal{W}(1,3) - \tfrac{1}{2}\mathcal{W}(1,2)^2,\\
& d(4,2) = \mathcal{W}(4,2) - \mathcal{W}(1,1)\mathcal{W}(3,1) - \tfrac{1}{2}\mathcal{W}(2,1)^2\\
& d(3,3) = \mathcal{W}(3,3) - \mathcal{W}(1,1)\mathcal{W}(2,2) - \mathcal{W}(1,2)\mathcal{W}(2,1) + \tfrac{1}{6}\mathcal{W}(1,1)^3.
\end{aligned}$$

For $r = 3$, the dimensions up to $d(2,2,2)$ are

$$\begin{aligned}
N=3,\quad & d(1,1,1) = \mathcal{W}(1,1,1),\\
N=4,\quad & d(1,1,2) = \mathcal{W}(1,1,2), \quad d(1,2,1) = \mathcal{W}(1,2,1), \quad d(2,1,1) = \mathcal{W}(2,1,1),\\
N=5,\quad & d(1,2,2) = \mathcal{W}(1,2,2), \quad d(2,1,2) = \mathcal{W}(2,1,2), \quad d(2,2,1) = \mathcal{W}(2,2,1),\\
& d(1,1,3) = \mathcal{W}(1,1,3), \quad d(1,3,1) = \mathcal{W}(1,3,1), \quad d(3,1,1) = \mathcal{W}(3,1,1),\\
N=6,\quad & d(1,1,4) = \mathcal{W}(1,1,4), \quad d(1,4,1) = \mathcal{W}(1,4,1), \quad d(4,1,1) = \mathcal{W}(4,1,1),\\
& d(1,2,3) = \mathcal{W}(1,2,3), \quad d(3,1,2) = \mathcal{W}(3,1,2), \quad d(2,3,1) = \mathcal{W}(2,3,1),\\
& d(3,2,1) = \mathcal{W}(3,2,1), \quad d(1,3,2) = \mathcal{W}(1,3,2), \quad d(2,1,3) = \mathcal{W}(2,1,3),\\
& d(2,2,2) = \mathcal{W}(2,2,2) - \tfrac{1}{2}\mathcal{W}^2(1,1,1).
\end{aligned}$$

Example 6. Relative to θ with $\mathcal{W} = \mathcal{G}$ and applying data from previous examples, for the case $r = 2$, we find the dimensions

$$\begin{aligned}
d(1,1) = d(1,2) = d(2,1) = d(1,3) = d(3,1) = d(1,4) = d(4,1) = d(2,3)\\
= d(3,2) = d(1,5) = d(5,1) = d(2,2) = d(2,4) = d(4,2) = d(3,3) = 4.
\end{aligned}$$

In the case $r = 3$, the dimensions are

$$d(1,1,1) = 8,$$
$$d(1,1,2) = d(2,1,1) = d(1,2,1) = 16, \quad d(1,2,2) = d(2,1,2) = d(2,2,1) = 56,$$
$$d(1,1,3) = d(3,1,1) = d(1,3,1) = 24, \quad d(1,1,4) = d(4,1,1) = d(1,4,1) = 32,$$
$$d(1,2,3) = d(3,1,2) = d(2,3,1) = d(3,2,1) = d(1,3,2) = d(2,1,3) = 128,$$
$$d(2,2,2) = 496.$$

Example 7. Relative to θ_+ with $\mathcal{W} = \mathcal{H}$, we find for the case $r = 2$

$$d(1,1) = d(1,2) = d(2,1) = d(1,3) = d(3,1) = d(1,4) = d(4,1) = d(2,3)$$
$$= d(3,2) = d(1,5) = d(5,1) = 2,$$
$$d(2,2) = 5, \quad d(2,4) = d(4,2) = 9, \quad d(3,3) = 28,$$

and for $r = 3$

$$d(1,1,1) = 8,$$
$$d(1,1,2) = d(2,1,1) = d(1,2,1) = 16, \quad d(1,2,2) = d(2,1,2) = d(2,2,1) = 56,$$
$$d(1,1,3) = d(3,1,1) = d(1,3,1) = 24, \quad d(1,1,4) = d(4,1,1) = d(1,4,1) = 32,$$
$$d(1,2,3) = d(3,1,2) = d(2,3,1) = d(3,2,1) = d(1,3,2) = d(2,1,3) = 128,$$
$$d(2,2,2) = 504.$$

Remark. In spite of the negative terms in the formulas for the dimensions, they give positive results. To understand why, consider, for example, the case

$$d(2,2) = \mathcal{W}(2,2) - \tfrac{1}{2}\mathcal{W}(1,1)^2$$

with $\mathcal{W}(a,b) = \mathcal{F}' = (a+b)\mathcal{F}$. So $d(2,2)$ is four times the result in example 6. In the set of words counted by $\mathcal{F}'(2,2) = 48$ there is a subset whose elements are words that are obtained by gluing together the words in the set counted by $\mathcal{W}(1,1) = 8$. The gluing produces an overcounting which is corrected by the one half factor. So $d(2,2)$ is positive. The same argument can be used to get positivity for the other formulas.

Theorem 4.3. *For each $G_r \subseteq G_R$, we have*

$$\prod_{m_{i_1},\ldots,m_{i_r}=1}^{\infty} (1 + z_{i_1}^{m_{i_1}} \cdots z_{i_r}^{m_{i_r}})^{\theta_+} = e^{-g(z_{i_1}^2,\ldots,z_{i_r}^2)+g(z_{i_1},\ldots,z_{i_r})}, \tag{4-12}$$

and

$$\prod_{m_{i_1},\ldots,m_{i_r}=1}^{\infty} (1 - z_{i_1}^{m_{i_1}} \cdots z_{i_r}^{m_{i_r}})^{\theta_-} = e^{+g(z_{i_1}^2,\ldots,z_{i_r}^2)-g(z_{i_1},\ldots,z_{i_r})}. \tag{4-13}$$

Proof. To prove (4-12), multiply and divide its left-hand side by

$$\prod_{m_{i_1},\dots,m_{i_r}=1}^{\infty}(1-z_{i_1}^{m_{i_1}}\cdots z_{i_r}^{m_{i_r}})^{\theta_+}$$

and use (4-8). To get (4-13), write

$$\prod_{m_{i_1},\dots,m_{i_r}=1}^{\infty}(1-z_{i_1}^{m_{i_1}}\cdots z_{i_r}^{m_{i_r}})^{\theta_-}=\prod_{N=r}^{\infty}\prod_{\substack{m_i>0\\ m_{i_1}+\cdots+m_{i_r}=N}}(1-z_{i_1}^{m_{i_1}}\cdots z_{i_r}^{m_{i_r}})^{\theta_-}.$$

Decompose the product over N into three products, namely, one over all $N<2r$, one over all even $N\geq 2r$, and another over all odd $N>2r$. Then apply Theorems 3.3 and 3.4 and formula (3-17). □

Theorem 4.4. $$\prod_{m_{i_1},\dots,m_{i_r}=1}^{\infty}(1+z_{i_1}^{m_{i_1}}\cdots z_{i_r}^{m_{i_r}})^{\theta_+}(1-z_{i_1}^{m_{i_1}}\cdots z_{i_r}^{m_{i_r}})^{\theta_-}=1.$$

Proof. Multiply (4-12) and (4-13). □

The left side of (1-1) equals

$$\prod_{j=1}^{R}(1+z_j)^2\prod_{r=2}^{R}\prod_{G_r}\prod_{m_{i_1},\dots,m_{i_r}>0}(1+z_{i_1}^{m_{i_1}}\cdots z_{i_r}^{m_{i_r}})^{\theta_+}(1-z_{i_1}^{m_{i_1}}\cdots z_{i_r}^{m_{i_r}})^{\theta_-}.$$

The Sherman identity now follows from Theorem 4.4.

Appendix

Lemma A.1. *If*

$$g(n_1,\dots,n_k)=\sum_{d|n_1,\dots,n_k}\frac{\mu(d)}{d}f\left(\frac{n_1}{d},\dots,\frac{n_k}{d}\right), \tag{A-1}$$

then

$$f(n_1,\dots,n_k)=\sum_{d|n_1,\dots,n_k}\frac{1}{d}g\left(\frac{n_1}{d},\dots,\frac{n_1}{d}\right). \tag{A-2}$$

Proof. Set $G(n_1,\dots,n_k):=(n_1+\cdots+n_k)g(n_1,\dots,n_k)$ and

$$F\left(\frac{n_1}{d},\dots,\frac{n_k}{d}\right):=\left(\frac{n_1}{d}+\cdots+\frac{n_k}{d}\right)f\left(\frac{n_1}{d},\dots,\frac{n_k}{d}\right).$$

Then (A-1) can be expressed in the form

$$G(n_1,\dots,n_k)=\sum_{d|n_1,\dots,n_k}\mu(d)F\left(\frac{n_1}{d},\dots,\frac{n_1}{d}\right).$$

Möbius inversion gives

$$F(n_1,\dots,n_k)=\sum_{d\mid n_1,\dots,n_k} G\left(\frac{n_1}{d},\dots,\frac{n_k}{d}\right).$$

Therefore,

$$(n_1+\cdots+n_k)\,f(n_1,\dots,n_k)=\sum_{d\mid n_1,\dots,n_k}\left(\frac{n_1}{d}+\cdots+\frac{n_k}{d}\right)g\left(\frac{n_1}{d},\dots,\frac{n_k}{d}\right). \quad \square$$

The converse is also true.

Lemma A.2. *Let $N\geq\alpha=n_1+\cdots+n_l$, $n_1,\dots,n_l$, $n_i>0$, be a partition of α. Then*

$$\sum_{\sum_{i=1}^{l}k_i=N}\prod_{i=1}^{l}\binom{k_i-1}{n_i-1}=\binom{N-1}{\alpha-1} \tag{A-3}$$

with the convention that a bracket in the left side is zero whenever $k_i<n_i$.

Proof. Using

$$\frac{q^\alpha}{(1-q)^\alpha}=\sum_{N=\alpha}^{\infty}\binom{N-1}{\alpha-1}q^N,$$

it follows that

$$\frac{q^\alpha}{(1-q)^\alpha}=\prod_{i=1}^{l}\frac{q^{n_i}}{(1-q)^{n_i}}=\prod_{i=1}^{l}\sum_{k_i=n_i}^{\infty}\binom{k_i-1}{n_i-1}q^{k_i}=\sum_{N=\alpha}^{\infty}\sum_{\substack{k_i\geq n_i\\ \sum_{i=1}^{l}k_i=N}}\prod_{i=1}^{l}\binom{k_i-1}{n_i-1}q^N.$$

Comparison with the previous expression and the convention gives the result. $\square$

Acknowledgements

We thank Professor Peter Moree (Max Planck Institute for Mathematics, Bonn) and Professor Thomas Ward (University of East Anglia, UK) for email correspondence regarding the positivity of the Möbius inversion formula.

References

[Andrews 1976] G. E. Andrews, *The theory of partitions*, Encyclopedia of Mathematics and its Applications **2**, Addison-Wesley, Reading, MA, 1976. MR 58 #27738 Zbl 0371.10001

[Apostol 1976] T. M. Apostol, *Introduction to analytic number theory*, Springer, New York, 1976. MR 55 #7892 Zbl 0335.10001

[Chen and Koh 1992] C. C. Chen and K. M. Koh, *Principles and techniques in combinatorics*, World Scientific, River Edge, NJ, 1992. MR 93h:05001 Zbl 0786.05002

[Cimasoni 2010] D. Cimasoni, "A generalized Kac–Ward formula", *J. Stat. Mech.* **2010** (2010), Article ID #P07023.

[Constantine and Savits 1996] G. M. Constantine and T. H. Savits, "A multivariate Faà di Bruno formula with applications", *Trans. Amer. Math. Soc.* **348**:2 (1996), 503–520. MR 96g:05008 Zbl 0846.05003

[da Costa 1997] G. A. T. F. da Costa, "Feynman identity: a special case, I", *J. Math. Phys.* **38**:2 (1997), 1014–1034. MR 98b:82012 Zbl 0869.05039

[da Costa and Variane 2005] G. A. T. F. da Costa and J. Variane, Jr., "Feynman identity: a special case revisited", *Lett. Math. Phys.* **73**:3 (2005), 221–235. MR 2007a:82012 Zbl 1101.82007

[Kang and Kim 1999] S.-J. Kang and M.-H. Kim, "Dimension formula for graded Lie algebras and its applications", *Trans. Amer. Math. Soc.* **351**:11 (1999), 4281–4336. MR 2000b:17009 Zbl 0926.17002

[Loebl 2004] M. Loebl, "A discrete non-Pfaffian approach to the Ising problem", pp. 145–154 in *Graphs, morphisms and statistical physics* (Piscataway, NJ, 2001), edited by J. Nešetřil and P. Winkler, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. **63**, Amer. Math. Soc., Providence, RI, 2004. MR 2005k:05066 Zbl 1069.82003

[Moree 2005] P. Moree, "The formal series Witt transform", *Discrete Math.* **295**:1-3 (2005), 143–160. MR 2006b:05015 Zbl 1064.05025

[Savits 2006] T. H. Savits, "Some statistical applications of Faa di Bruno", *J. Multivariate Anal.* **97**:10 (2006), 2131–2140. MR 2008g:62144 Zbl 1101.62041

[Sherman 1960] S. Sherman, "Combinatorial aspects of the Ising model for ferromagnetism, I: A conjecture of Feynman on paths and graphs", *J. Math. Phys.* **1** (1960), 202–217. MR 22 #10273 Zbl 0123.45501

[Sherman 1962] S. Sherman, "Combinatorial aspects of the Ising model for ferromagnetism, II: An analogue to the Witt identity", *Bull. Amer. Math. Soc.* **68** (1962), 225–229. MR 25 #2003

[Witt 1937] E. Witt, "Treue Darstellung Liescher Ring", *J. Reine Angew. Math.* **177** (1937), 152–160. JFM 63.0089.02

Received May 5, 2012.

G. A. T. F. DA COSTA
DEPARTAMENTO DE MATEMÁTICA
UNIVERSIDADE FEDERAL DE SANTA CATARINA
88040-900 FLORIANÓPOLIS SC
BRAZIL

gatcosta@mtm.ufsc.br

G. A. ZIMMERMANN
DEPARTAMENTO DE MATEMÁTICA
UNIVERSIDADE FEDERAL DE SANTA CATARINA
88040-900 FLORIANÓPOLIS SC
BRAZIL

graciele@ifsc.edu.br

PACIFIC JOURNAL OF MATHEMATICS
Vol. 263, No. 2, 2013
dx.doi.org/10.2140/pjm.2013.263.495

ON THE CLASSIFICATION OF STABLE SOLUTIONS TO BIHARMONIC PROBLEMS IN LARGE DIMENSIONS

JUNCHENG WEI, XINGWANG XU AND WEN YANG

We give a new bound on the exponent for nonexistence of stable solutions to the biharmonic problem $\Delta^2 u = u^p$ in $\mathbb{R}^n$, where $u > 0$, $p > 1$, and $n \geq 20$.

1. Introduction

Of concern is the biharmonic equation

$$\Delta^2 u = u^p, \quad u > 0 \quad \text{in } \mathbb{R}^n \tag{1-1}$$

where $n \geq 5$ and $p > 1$. Set

$$\Lambda_u(\varphi) := \int_{\mathbb{R}^n} |\Delta\varphi|^2 dx - p \int_{\mathbb{R}^n} u^{p-1}\varphi^2 dx \quad \text{for all } \varphi \in H^2(\mathbb{R}^n). \tag{1-2}$$

The Morse index $\mathrm{ind}(u)$ of a classical solution to (1-1) is defined as the maximal dimension of all subspaces of $H^2(\mathbb{R}^n)$ such that $\Lambda_u(\varphi) < 0$ in $H^2(\mathbb{R}^n) \setminus \{0\}$. We say u is a stable solution to (1-1) if $\Lambda_u(\varphi) \geq 0$ for any test function $\varphi \in H^2(\mathbb{R}^n)$; that is, if the Morse index is zero.

In the first part of the paper, we obtain the following classification result on stable solutions of (1-1).

Theorem 1.1. *Let $n \geq 20$ and $1 < p < 1 + \dfrac{8p^*}{n-4}$. Then* (1-1) *has no stable solutions.*

Here p^* stands for the smallest real root greater than $\dfrac{n-4}{n-8}$ of the algebraic equation

$$\begin{aligned}&512(2-n)x^6+4(n^3-60n^2+670n-1344)x^5-2(13n^3-424n^2+3064n-5408)x^4\\&+2(27n^3-572n^2+3264n-5440)x^3-(49n^3-772n^2+3776n-5888)x^2\\&+4(5n^3-66n^2+288n-416)x-3(n^3-12n^2+48n-64)=0.\end{aligned}$$

The first author was supported from an earmarked grant ("On Elliptic Equations with Negative Exponents") from RGC of Hong Kong.
MSC2010: primary 35B20; secondary 35J60.
Keywords: stable solutions, biharmonic superlinear equations.

Some remarks are in order. Let us recall that for the second-order problem

$$\Delta u + u^p = 0 \quad u > 0 \text{ in } \mathbb{R}^n, \ \ p > 1, \tag{1-3}$$

Farina gave a complete classification of all finite Morse index solutions. The main result of [Farina 2007] is that no stable solution exists to (1-3) if either $n \le 10$, $p > 1$ or $n \ge 11$, $p < p_{JL}$. Here p_{JL} denotes the Joseph–Lundgren exponent [Gui et al. 1992]. On the other hand, a stable radial solution exists for $p \ge p_{JL}$.

For the fourth-order case, the nonexistence of positive solutions to (1-1) is shown if $p < \frac{n+4}{n-4}$, and all entire solutions are classified if $p = \frac{n+4}{n-4}$. See [Lin 1998; Wei and Xu 1999]. When $p > \frac{n+4}{n-4}$, radially symmetric solutions to (1-1) are completely classified in [Ferrero et al. 2009; Gazzola and Grunau 2006; Guo and Wei 2010]. The radial solutions are shown to be stable if and only if $p \ge p'_{JL}$ and $n \ge 13$, where p'_{JL} stands for the corresponding Joseph–Lundgren exponent (see [Ferrero et al. 2009; Gazzola and Grunau 2006]). In the general nonradial case, Wei and Ye [Wei and Ye 2010] showed the nonexistence of stable or finite Morse index solutions when either $n \le 8$, $p > 1$ or $n \ge 9$, $p \le \frac{n}{n-8}$. In dimensions $n \ge 9$, a perturbation argument is used to show the nonexistence of stable solutions for $p < \frac{n}{n-8} + \varepsilon_n$ for some $\varepsilon_n > 0$. However, no explicit value of ε_n was given. The proof of Wei and Ye [2010] follows an earlier idea of Cowan, Esposito and Ghoussoub [2010] in which a similar problem in a bounded domain was studied. Theorem 1.1 gives an explicit value on ε_n for $n \ge 20$.

In the second-order case, the proof of Farina uses basically the Moser iterations: namely multiply (1-3) by the power of u, like u^q, $q > 1$. Moser iteration works because of the following simple identity

$$\int_{\mathbb{R}^n} u^q(-\Delta u) = \frac{4q}{(q+1)^2} \int_{\mathbb{R}^n} |\nabla u^{\frac{q+1}{2}}|^2, \forall u \in C_0^1(\mathbb{R}^n).$$

In the fourth-order case, such equality does not hold, and in fact we have

$$\int_{\mathbb{R}^n} u^q(\Delta^2 u) = \frac{4q}{(q+1)^2} \int_{\mathbb{R}^n} |\Delta u^{\frac{q+1}{2}}|^2 - q(q-1)^2 \int_{\mathbb{R}^n} u^{q-3}|\nabla u|^4, \forall u \in C_0^2(\mathbb{R}^n).$$

The additional term $\int_{\mathbb{R}^n} u^{q-3}|\nabla u|^4$ makes the Moser iteration argument difficult to use. Wei and Ye [2010] used instead the new test function $-\Delta u$ and showed that $\int_{\mathbb{R}^2} |\Delta u|^2$ is bounded. Thus the exponent $\frac{n}{n-8}$ is obtained. In this paper, we use the Moser iteration for the fourth-order problem and give a control on the term $\int_{\mathbb{R}^n} u^{q-3}|\nabla u|^4$ (Lemma 2.3). As a result, we obtain a better exponent $\frac{n}{n-8} + \varepsilon_n$ where ε_n is explicitly given. As far as we know, this seems to be the first result for Moser iteration for a fourth-order problem.

In the second part of this paper, we show that the same idea can be used to establish the regularity of extremal solutions to

$$\text{(1-4)} \qquad \begin{cases} \Delta^2 u = \lambda(u+1)^p, \ \lambda > 0 & \text{in } \Omega, \\ u > 0 & \text{in } \Omega, \\ u = \Delta u = 0 & \text{on } \partial\Omega, \end{cases}$$

where Ω is a smooth and bounded convex domain in $\mathbb{R}^n$.

For problem (1-4), it is known [Berchio and Gazzola 2005] that for $p > \frac{n+4}{n-4}$ there exists a critical value $\lambda^* > 0$ depending on $p > 1$ and Ω such that

- If $\lambda \in (0, \lambda^*)$, (1-4) has a minimal and classical solution which is stable;
- If $\lambda = \lambda^*$, a unique weak solution, called the extremal solution u^* exists for (1-4);
- No weak solution of (1-4) exists whenever $\lambda > \lambda^*$.

The regularity of the extremal solution of problem (1-4) at $\lambda = \lambda_*$ has been studied in [Cowan et al. 2010; Wei and Ye 2010], where it was shown that the extremal solution is bounded provided $n \le 8$ or $p < \frac{n}{n-8} + \varepsilon_n$, $n \ge 9$ (ε_n very small). Here, we also give a explicit bound for the exponent p in large dimensions and our second result is the following.

Theorem 1.2. *The extremal solution u^* of* (1-4) *when $\lambda = \lambda^*$ is bounded provided that $n \ge 20$ and $1 < p < 1 + \frac{8p^*}{n-4}$, where p^* is defined as above.*

As $n \to +\infty$, the value ε_n is asymptotically $8\sqrt{8/3}/(n-8)^{3/2}$ and thus the upper bound for p has the expansion

$$\text{(1-5)} \qquad 1 + \frac{8}{n-8} + \frac{8\sqrt{8/3}}{(n-8)^{3/2}} + O\left(\frac{1}{(n-8)^2}\right).$$

On the other hand, for radial solutions, the Joseph–Lundgren exponent [Gui et al. 1992] has the following asymptotic expansion

$$\text{(1-6)} \qquad 1 + \frac{8}{n-8} + \frac{16}{(n-8)^{3/2}} + O\left(\frac{1}{(n-8)^2}\right).$$

In this paper, we have only considered fourth-order problems with power-like nonlinearity. Other kinds of nonlinearity, such as exponential and negative powers, also appear in many applications; see [Cowan et al. 2010]. However, our technique here yields no improvements of results of that reference in the case of exponential and negative nonlinearities.

This paper is organized as follows. We prove Theorem 1.1 and Theorem 1.2 respectively in Section 2 and Section 3. Some technical inequalities are given in the Appendix.

2. Proof of Theorem 1.1

Lemma 2.1. *For any $\varphi \in C_0^4(\mathbb{R}^n)$ with $\varphi \geq 0$, any $\gamma > 1$ and $\varepsilon > 0$ an arbitrary small number, we have*

$$\int_{\mathbb{R}^n} (\Delta(u^\gamma \varphi^\gamma))^2 \leq \int_{\mathbb{R}^n} ((\Delta u^\gamma \varphi^\gamma)^2 + \varepsilon |\nabla u|^4 \varphi^{2\gamma} u^{2\gamma-4} + C u^{2\gamma} \|\nabla^4(\varphi^{2\gamma})\|), \tag{2-1}$$

$$\int_{\mathbb{R}^n} (\Delta(u^\gamma \varphi^\gamma))^2 \geq \int_{\mathbb{R}^n} ((\Delta u^\gamma \varphi^\gamma)^2 - \varepsilon |\nabla u|^4 \varphi^{2\gamma} u^{2\gamma-4} - C u^{2\gamma} \|\nabla^4(\varphi^{2\gamma})\|), \tag{2-2}$$

$$\int_{\mathbb{R}^n} ((u^\gamma)_{ij})^2 \varphi^{2\gamma} \leq \int_{\mathbb{R}^n} ((u^\gamma \varphi^\gamma)_{ij})^2 + \varepsilon \int_{\mathbb{R}^n} |\nabla u|^4 u^{2\gamma-4} \varphi^{2\gamma} + C \int_{\mathbb{R}^n} u^{2\gamma} \|\nabla^4(\varphi^{2\gamma})\|, \tag{2-3}$$

where C is a positive number that only depends on γ and ε, and $\|\nabla^4(\varphi^{2\gamma})\|$ is defined by

$$\|\nabla^4(\varphi^{2\gamma})\|^2 = \varphi^{-2\gamma} |\nabla \varphi^\gamma|^4 + |\varphi^\gamma (\Delta^2 \varphi^\gamma)| + |\nabla^2 \varphi^\gamma|^2.$$

In the following, unless said otherwise, the constant C always denotes a positive number which may change term by term but only depends on γ, ε.

Proof. Since φ is compactly supported, we can use integration by parts without considering the boundary terms. First, by direct calculation, we get

$$\begin{aligned}(\Delta(u^\gamma \varphi^\gamma))^2 = {} & [(\Delta u^\gamma)\varphi^\gamma]^2 + 4\nabla u^\gamma \nabla \varphi^\gamma \Delta \varphi^\gamma u^\gamma + 4\nabla u^\gamma \nabla \varphi^\gamma \Delta u^\gamma \varphi^\gamma \\ & + 4(\nabla u^\gamma \nabla \varphi^\gamma)^2 + 2\Delta u^\gamma u^\gamma \Delta \varphi^\gamma \varphi^\gamma + u^{2\gamma} (\Delta \varphi^\gamma)^2.\end{aligned} \tag{2-4}$$

We now need to deal with the third and fifth terms on the right side of this equality, up to the integration of both sides.

For the third term, we have

$$\begin{aligned}\int_{\mathbb{R}^n} \Delta u^\gamma \nabla u^\gamma \nabla \varphi^\gamma \varphi^\gamma = {} & -\int_{\mathbb{R}^n} (u^\gamma)_i (u^\gamma)_{ij} (\varphi^\gamma)_j \varphi^\gamma \\ & - \int_{\mathbb{R}^n} (u^\gamma)_i (u^\gamma)_j (\varphi^\gamma)_{ij} \varphi^\gamma - \int_{\mathbb{R}^n} (u^\gamma)_i (u^\gamma)_j (\varphi^\gamma)_j (\varphi^\gamma)_i,\end{aligned}$$

where $f_i = \partial f / \partial x_i$ and $f_{ij} = \partial^2 f / \partial x_j \partial x_i$. (Here and in the sequel, we use the Einstein summation convention, so for example $\partial_i (u_i u_j \varphi_j) = \sum_{1 \leq i,j \leq n} \partial_i (u_i u_j \varphi_j)$.) The first term on the right side of the previous equation can be estimated as

$$\begin{aligned}2\int_{\mathbb{R}^n} (u^\gamma)_i (u^\gamma)_{ij} (\varphi^\gamma)_j \varphi^\gamma = {} & \int_{\mathbb{R}^n} \partial_j ((u^\gamma)_i (u^\gamma)_i (\varphi^\gamma)_j \varphi^\gamma) - \int_{\mathbb{R}^n} ((u^\gamma)_i)^2 (\varphi^\gamma)_{jj} \varphi^\gamma \\ & - \int_{\mathbb{R}^n} ((u^\gamma)_i)^2 (\varphi^\gamma)_j (\varphi^\gamma)_j.\end{aligned}$$

Combining these two equalities, we get

$$\begin{aligned}2\int_{\mathbb{R}^n}\Delta u^\gamma\nabla u^\gamma\nabla\varphi^\gamma\varphi^\gamma =& -\int_{\mathbb{R}^n}\partial_j\big((u^\gamma)_i(u^\gamma)_i(\varphi^\gamma)_j\varphi^\gamma\big)\\ &-\int_{\mathbb{R}^n}2(u^\gamma)_i(u^\gamma)_j(\varphi^\gamma)_{ij}\varphi^\gamma-\int_{\mathbb{R}^n}2(u^\gamma)_i(u^\gamma)_j(\varphi^\gamma)_j(\varphi^\gamma)_i\\ &+\int_{\mathbb{R}^n}((u^\gamma)_i)^2(\varphi^\gamma)_{jj}\varphi^\gamma+\int_{\mathbb{R}^n}((u^\gamma)_i)^2(\varphi^\gamma)_j(\varphi^\gamma)_j.\end{aligned}$$

Rewriting this equality we have

$$\begin{aligned}4\int_{\mathbb{R}^n}\Delta u^\gamma\nabla u^\gamma\nabla\varphi^\gamma\varphi^\gamma =& 2\int_{\mathbb{R}^n}|\nabla u^\gamma|^2\Delta\varphi^\gamma\varphi^\gamma+2\int_{\mathbb{R}^n}|\nabla u^\gamma|^2|\nabla\varphi^\gamma|^2\\ &-4\int_{\mathbb{R}^n}(u^\gamma)_i(u^\gamma)_j(\varphi^\gamma)_{ij}\varphi^\gamma-4\int_{\mathbb{R}^n}\langle\nabla u^\gamma,\nabla\varphi^\gamma\rangle^2.\end{aligned}\tag{2-5}$$

For the fifth term on the right side of (2-4) we have

$$\begin{aligned}\int_{\mathbb{R}^n}\Delta u^\gamma u^\gamma\Delta\varphi^\gamma\varphi^\gamma =& -\int_{\mathbb{R}^n}u^\gamma\langle\nabla u^\gamma,\nabla(\Delta\varphi^\gamma)\rangle\varphi^\gamma\\ &-\int_{\mathbb{R}^n}\langle\nabla u^\gamma,\nabla\varphi^\gamma\rangle u^\gamma\Delta\varphi^\gamma-\int_{\mathbb{R}^n}|\nabla u^\gamma|^2\Delta\varphi^\gamma\varphi^\gamma.\end{aligned}\tag{2-6}$$

Combining (2-4), (2-5) and (2-6), one obtains

$$\begin{aligned}&\int_{\mathbb{R}^n}(\Delta(u^\gamma\varphi^\gamma))^2-\int_{\mathbb{R}^n}(\Delta u^\gamma)^2\varphi^{2\gamma}\\ &\quad=2\int_{\mathbb{R}^n}|\nabla u^\gamma|^2|\nabla\varphi^\gamma|^2-4\int_{\mathbb{R}^n}\varphi^\gamma(\nabla^2\varphi^\gamma(\nabla u^\gamma,\nabla u^\gamma))\\ &\qquad+\int_{\mathbb{R}^n}u^{2\gamma}\varphi^\gamma\Delta^2(\varphi^\gamma)-2\int_{\mathbb{R}^n}u^{2\gamma}(\Delta\varphi^\gamma)^2.\end{aligned}\tag{2-7}$$

Now by the Young equality, for any $\varepsilon>0$, there exists a constant $C=C(\gamma,\varepsilon)$ such that

$$|\nabla u^\gamma|^2|\nabla\varphi^\gamma|^2\le\frac{\varepsilon}{4}|\nabla u^\gamma|^4u^{-2\gamma}\varphi^{2\gamma}+C|\nabla\varphi^\gamma|^4u^{2\gamma}\varphi^{-2\gamma}$$

and

$$|\varphi^\gamma(\nabla^2\varphi^\gamma(\nabla u^\gamma,\nabla u^\gamma))|\le\frac{\varepsilon}{8}|\nabla u^\gamma|^4u^{-2\gamma}\varphi^{2\gamma}+Cu^{2\gamma}|\nabla^2\varphi^\gamma|^2.$$

Thus by (2-7), together with the two estimates above, one gets

$$\left|\int_{\mathbb{R}^n}(\Delta(u^\gamma\varphi^\gamma))^2-\int_{\mathbb{R}^n}(\Delta u^\gamma)^2\varphi^{2\gamma}\right|\le\varepsilon\int_{\mathbb{R}^n}|\nabla u^\gamma|^4u^{-2\gamma}\varphi^{2\gamma}+6C\int_{\mathbb{R}^n}u^{2\gamma}\|\nabla^4\varphi^\gamma\|^2.$$

The estimates (2-1) and (2-2) follow from this easily.

Next we observe that $|\nabla^2 u^\gamma|^2\varphi^{2\gamma} = \left[\frac{1}{2}\Delta|\nabla u^\gamma|^2 - \langle\nabla u^\gamma, \nabla\Delta u^\gamma\rangle\right]\varphi^{2\gamma}$. Thus up to the integration by parts, with the help of (2-5) and the estimates we just proved, the estimate (2-3) also follows by noticing the identity $\int_{\mathbb{R}^n}(\Delta(u^\gamma\varphi^\gamma))^2 = \int_{\mathbb{R}^n}|\nabla^2(u^\gamma\varphi^\gamma)|^2$. The proof of Lemma 2.1 is thus completed. □

Let us return to the equation

$$\Delta^2 u = u^p, \quad u > 0 \text{ in } \mathbb{R}^n. \tag{2-8}$$

Fix $q = 2\gamma - 1 > 0$ and $\gamma > 1$. Let $\varphi \in C_0^\infty(\mathbb{R}^n)$. Multiplying (2-8) by $u^q\varphi^{2\gamma}$ and integration by parts, we obtain

$$\int_{\mathbb{R}^n}\Delta u\Delta(u^q\varphi^{2\gamma}) = \int_{\mathbb{R}^n} u^{p+q}\varphi^{2\gamma}. \tag{2-9}$$

For the left side of (2-9), we have:

Lemma 2.2. *For any $\varphi \in C_0^\infty(\mathbb{R}^n)$ with $\varphi \geq 0$, for any $\varepsilon > 0$ and γ with q defined above, there exists a positive constant C depends on γ, ε such that*

$$\begin{aligned}\int_{\mathbb{R}^n}\frac{\gamma^2}{q}\Delta u\Delta(u^q\varphi^{2\gamma}) \geq \int_{\mathbb{R}^n}(\Delta u^\gamma\varphi^\gamma)^2 - \int_{\mathbb{R}^n} Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\| \\ - \int_{\mathbb{R}^n}(\gamma^2(\gamma-1)^2+\varepsilon)u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma}.\end{aligned} \tag{2-10}$$

Proof. First, by direct computations, we obtain

$$\begin{aligned}\Delta u\Delta(u^{2\gamma-1}\varphi^{2\gamma}) = \Delta u\big((2\gamma-1)u^{2\gamma-2}\Delta u\varphi^{2\gamma} + 2(2\gamma-1)u^{2\gamma-2}\nabla u\nabla(\varphi^{2\gamma}) \\ + (2\gamma-1)(2\gamma-2)u^{2\gamma-3}|\nabla u|^2\varphi^{2\gamma} + u^{2\gamma-1}\Delta\varphi^{2\gamma}\big),\end{aligned}$$

$$\begin{aligned}(\Delta u^\gamma\varphi^\gamma)^2 = \gamma^2u^{2\gamma-2}(\Delta u)^2\varphi^{2\gamma} + \gamma^2(\gamma-1)^2u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} \\ + 2(\gamma-1)\gamma^2u^{2\gamma-3}|\nabla u|^2\Delta u\varphi^{2\gamma}.\end{aligned}$$

Combining these two identities, we get

$$\begin{aligned}\frac{\gamma^2}{q}\Delta u\Delta(u^q\varphi^{2\gamma}) = (\Delta u^\gamma\varphi^\gamma)^2 + 2\gamma^2u^{2\gamma-2}\Delta u\nabla u\nabla\varphi^{2\gamma} + \frac{\gamma^2}{q}u^{2\gamma-1}\Delta u\Delta\varphi^{2\gamma} \\ - \gamma^2(\gamma-1)^2u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma}.\end{aligned} \tag{2-11}$$

For the term $u^{2\gamma-2}\Delta u\nabla u\nabla\varphi^{2\gamma}$, we have

$$\begin{aligned}u^{2\gamma-2}\Delta u\nabla u\nabla\varphi^{2\gamma} = \partial_i(u^{2\gamma-2}u_iu_j(\varphi^{2\gamma})_j) - (2\gamma-2)u^{2\gamma-3}(u_i)^2u_j(\varphi^{2\gamma})_j \\ - u^{2\gamma-2}u_iu_{ij}(\varphi^{2\gamma})_j - u^{2\gamma-2}u_iu_j(\varphi^{2\gamma})_{ij}.\end{aligned}$$

We can regroup the term $u^{2\gamma-2}u_iu_{ij}(\varphi^{2\gamma})_j$ as

$$2u^{2\gamma-2}u_iu_{ij}(\varphi^{2\gamma})_j = \partial_j(u^{2\gamma-2}(u_i)^2(\varphi^{2\gamma})_j) - (2\gamma-2)u^{2\gamma-3}u_j(u_i)^2(\varphi^{2\gamma})_j - u^{2\gamma-2}(u_i)^2(\varphi^{2\gamma})_{jj}.$$

Therefore we get

$$\begin{aligned}(2\text{-}12)\quad 2u^{2\gamma-2}\Delta u\nabla u\nabla\varphi^{2\gamma} &= 2\partial_i(u^{2\gamma-2}u_iu_j(\varphi^{2\gamma})_j) - \partial_j(u^{2\gamma-2}(u_i)^2(\varphi^{2\gamma})_j)\\ &\quad - (2\gamma-2)u^{2\gamma-3}(u_i)^2u_j(\varphi^{2\gamma})_j + u^{2\gamma-2}(u_i)^2(\varphi^{2\gamma})_{jj}\\ &\quad - 2u^{2\gamma-2}u_iu_j(\varphi^{2\gamma})_{ij}.\end{aligned}$$

For the last three terms on the right side of (2-12), applying Young's inequality, we get

$$\begin{aligned}|u^{2\gamma-3}(u_i)^2u_j(\varphi^{2\gamma})_j| &\le \frac{\varepsilon}{6\gamma^2(\gamma-1)}u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} + Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|,\\ |u^{2\gamma-2}(u_i)^2(\varphi^{2\gamma})_{jj}| &\le \frac{\varepsilon}{6\gamma^2}u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} + Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|,\\ |u^{2\gamma-2}u_iu_j(\varphi^{2\gamma})_{ij}| &\le \frac{\varepsilon}{6\gamma^2}u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} + Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|.\end{aligned}$$

These three inequalities and (2-12) imply

$$(2\text{-}13)\quad \int_{\mathbb{R}^n} 2\gamma^2u^{2\gamma-2}\Delta u\nabla u\nabla\varphi^{2\gamma} \ge -\frac{\varepsilon}{2}\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} - C\int_{\mathbb{R}^n} u^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|.$$

Similarly we get

$$(2\text{-}14)\quad \int_{\mathbb{R}^n} \frac{\gamma^2}{q}u^{2\gamma-1}\Delta u\Delta\varphi^{2\gamma} \ge -\frac{\varepsilon}{2}\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} - C\int_{\mathbb{R}^n} u^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|.$$

Inequality (2-10) follows from (2-11), (2-13) and (2-14). □

As a result of (2-1) and (2-10), we have

$$\begin{aligned}(2\text{-}15)\quad \int_{\mathbb{R}^n} \frac{\gamma^2}{q}\Delta u\Delta(u^q\varphi^{2\gamma}) &\ge \int_{\mathbb{R}^n}(\Delta(u^\gamma\varphi^\gamma))^2 - \int_{\mathbb{R}^n} Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|\\ &\quad - \int_{\mathbb{R}^n}(\gamma^2(\gamma-1)^2+\varepsilon)u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma}.\end{aligned}$$

Next we estimate the most difficult term, $\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma}$, in (2-15). This is the key step in proving Theorem 1.1.

Lemma 2.3. *If u is the classical solution to the biharmonic equation* (2-8), *and φ is defined as above, then for any sufficiently small $\varepsilon > 0$, we have the following*

inequality

$$\left(\tfrac{1}{2}-\varepsilon\right)\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} \le \frac{2}{\gamma^2}\int_{\mathbb{R}^n}(\Delta(u^\gamma\varphi^\gamma))^2 + \int_{\mathbb{R}^n} Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\| - \int_{\mathbb{R}^n}\frac{4}{(4\gamma-3+p)(p+1)}u^{2\gamma+p-1}\varphi^{2\gamma}. \tag{2-16}$$

Proof. It is easy to see that

$$\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} = \frac{1}{\gamma^4}\int_{\mathbb{R}^n} u^{-2\gamma}|\nabla u^\gamma|^4\varphi^{2\gamma}, \tag{2-17}$$

and

$$\begin{aligned}\int_{\mathbb{R}^n} u^{-2\gamma}|\nabla u^\gamma|^4\varphi^{2\gamma} &= \int_{\mathbb{R}^n} u^{-2\gamma}|\nabla u^\gamma|^2\nabla u^\gamma\nabla u^\gamma\varphi^{2\gamma}\\ &= \int_{\mathbb{R}^n} -\nabla u^{-\gamma}|\nabla u^\gamma|^2\nabla u^\gamma\varphi^{2\gamma}\\ &= \int_{\mathbb{R}^n} u^{-\gamma}|\nabla u^\gamma|^2\Delta u^\gamma\varphi^{2\gamma} + \int_{\mathbb{R}^n} u^{-\gamma}\nabla(|\nabla u^\gamma|^2)\nabla u^\gamma\varphi^{2\gamma}\\ &\quad + \int_{\mathbb{R}^n} u^{-\gamma}|\nabla u^\gamma|^2\nabla u^\gamma\nabla\varphi^{2\gamma},\end{aligned} \tag{2-18}$$

where in the last step we used integration by parts. For the first term in the last part of this equality, we have

$$\int_{\mathbb{R}^n} u^{-\gamma}|\nabla u^\gamma|^2\Delta u^\gamma\varphi^{2\gamma} = \gamma^3\int_{\mathbb{R}^n}((\gamma-1)u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} + u^{2\gamma-3}|\nabla u|^2\Delta u\varphi^{2\gamma}).$$

Substituting this into (2-18) and combining with (2-17), we obtain

$$\begin{aligned}\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} &= \int_{\mathbb{R}^n}\frac{1}{\gamma^3}u^{-\gamma}\nabla(|\nabla u^\gamma|^2)\nabla u^\gamma\varphi^{2\gamma}\\ &\quad + \int_{\mathbb{R}^n} u^{2\gamma-3}(|\nabla u|^2)\Delta u\varphi^{2\gamma} + \int_{\mathbb{R}^n}\frac{1}{\gamma^3}u^{-\gamma}(|\nabla u^\gamma|^2)\nabla u^\gamma\nabla\varphi^{2\gamma}.\end{aligned} \tag{2-19}$$

The first term on the right side of (2-19) can be estimated as

$$\begin{aligned}u^{-\gamma}\nabla(|\nabla u^\gamma|^2)\nabla u^\gamma &= 2u^{-\gamma}((u^\gamma)_{ij}(u^\gamma)_i(u^\gamma)_j)\\ &\le 2\gamma(u^\gamma)_{ij}(u^\gamma)_{ij} + \frac{u^{-2\gamma}}{2\gamma}(u^\gamma)_i(u^\gamma)_j(u^\gamma)_i(u^\gamma)_j\\ &= 2\gamma|\nabla^2u^\gamma|^2 + \frac{u^{-2\gamma}}{2\gamma}|\nabla u^\gamma|^4.\end{aligned} \tag{2-20}$$

As a consequence, we have

$$\begin{aligned}(2\text{-}21)\quad &\int_{\mathbb{R}^n}\frac{1}{\gamma^3}u^{-\gamma}\nabla(|\nabla u^\gamma|^2)\nabla u^\gamma\varphi^{2\gamma}\\ &\le\int_{\mathbb{R}^n}\frac{2}{\gamma^2}|\nabla^2u^\gamma|^2\varphi^{2\gamma}+\int_{\mathbb{R}^n}\frac{1}{2\gamma^4}u^{-2\gamma}|\nabla u^\gamma|^4\varphi^{2\gamma}\\ &\le\int_{\mathbb{R}^n}\frac{2}{\gamma^2}|\nabla^2(u^\gamma\varphi^\gamma)|^2+\int_{\mathbb{R}^n}Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|+\int_{\mathbb{R}^n}\frac{1+4\gamma^2\varepsilon}{2\gamma^4}u^{-2\gamma}|\nabla u^\gamma|^4\varphi^{2\gamma}\\ &=\int_{\mathbb{R}^n}\frac{2}{\gamma^2}(\Delta(u^\gamma\varphi^\gamma))^2+\int_{\mathbb{R}^n}Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|+\int_{\mathbb{R}^n}\frac{1+4\gamma^2\varepsilon}{2\gamma^4}u^{-2\gamma}|\nabla u^\gamma|^4\varphi^{2\gamma},\end{aligned}$$

where we used (2-3) in the last step.

For the second term on the right side of (2-19), applying estimate (2.3) from [Wei and Ye 2010], that is, $(\Delta u)^2\ge\frac{2}{p+1}u^{p+1}$, and the fact that $\Delta u<0$ from Theorem 3.1 in [Wei and Xu 1999] or Theorem 2.1 in [Xu 2000], we have

$$\begin{aligned}(2\text{-}22)\quad \int_{\mathbb{R}^n}u^{2\gamma-3}(|\nabla u|^2)\Delta u\varphi^{2\gamma}&\le-\int_{\mathbb{R}^n}\sqrt{\frac{2}{p+1}}u^{2\gamma-3+\frac{p+1}{2}}(|\nabla u|^2)\varphi^{2\gamma}\\ &=\int_{\mathbb{R}^n}\frac{\sqrt{\frac{2}{p+1}}}{2\gamma-2+\frac{p+1}{2}}u^{2\gamma-2+\frac{p+1}{2}}\Delta u\varphi^{2\gamma}\\ &\quad+\int_{\mathbb{R}^n}\frac{\sqrt{\frac{2}{p+1}}}{2\gamma-2+\frac{p+1}{2}}u^{2\gamma-2+\frac{p+1}{2}}\nabla u\nabla\varphi^{2\gamma}.\end{aligned}$$

Using the inequality $-\Delta u\ge\sqrt{\frac{2}{p+1}}u^{\frac{p+1}{2}}$, we get

$$(2\text{-}23)\quad \int_{\mathbb{R}^n}\frac{\sqrt{\frac{2}{p+1}}}{2\gamma-2+\frac{2}{p+1}}u^{2\gamma-2+\frac{p+1}{2}}\Delta u\varphi^{2\gamma}\le-\int_{\mathbb{R}^n}\frac{\frac{2}{p+1}}{2\gamma-2+\frac{p+1}{2}}u^{2\gamma+p-1}\varphi^{2\gamma}.$$

On the other hand, for the second term on the right side of (2-22), we have

$$\begin{aligned}(2\text{-}24)\quad \int_{\mathbb{R}^n}u^{2\gamma-2+\frac{p+1}{2}}\nabla u\nabla\varphi^{2\gamma}&=-\int_{\mathbb{R}^n}\frac{1}{L}u^{2\gamma-1+\frac{p+1}{2}}\Delta\varphi^{2\gamma}\\ &=-\int_{\{x\,|\,\Delta\varphi^{2\gamma}>0\}}\frac{1}{L}u^{2\gamma-1+\frac{p+1}{2}}\Delta\varphi^{2\gamma}\\ &\quad-\int_{\{x\,|\,\Delta\varphi^{2\gamma}\le0\}}\frac{1}{L}u^{2\gamma-1+\frac{p+1}{2}}\Delta\varphi^{2\gamma},\end{aligned}$$

where the first equality follows from integration by parts and $L=2\gamma-1+\frac{p+1}{2}$. As for the first term on the last part of (2-24), using the inequality

$$\Delta u\le-\sqrt{\frac{2}{p+1}}u^{\frac{p+1}{2}}<0,$$

we have

$$\frac{\sqrt{\frac{p+1}{2}}}{L}\int_{\{x|\Delta\varphi^{2\gamma}>0\}} u^{2\gamma-1}\Delta u\Delta\varphi^{2\gamma} \le -\int_{\{x|\Delta\varphi^{2\gamma}>0\}} \frac{1}{L}u^{2\gamma-1+\frac{p+1}{2}}\Delta\varphi^{2\gamma}. \tag{2-25}$$

Similarly to the proof of Lemma 2.1, it is easy to get

$$\left|\int_{\{x|\Delta\varphi^{2\gamma}>0\}} \frac{\sqrt{\frac{p+1}{2}}}{L}u^{2\gamma-1}\Delta u\Delta\varphi^{2\gamma}\right| \le \varepsilon\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} + \int_{\mathbb{R}^n} Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|.$$

From this and (2-25), we have

$$\left|\int_{\{x|\Delta\varphi^{2\gamma}>0\}} \frac{1}{L}u^{2\gamma-1+\frac{p+1}{2}}\Delta\varphi^{2\gamma}\right| \le \varepsilon\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} + \int_{\mathbb{R}^n} Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|.$$

Similarly, we also obtain

$$\left|\int_{\{x|\Delta\varphi^{2\gamma}\le 0\}} \frac{1}{L}u^{2\gamma-1+\frac{p+1}{2}}\Delta\varphi^{2\gamma}\right| \le \varepsilon\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} + \int_{\mathbb{R}^n} Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|.$$

From the last two inequalities and (2-24), we have

$$\left|\int_{\mathbb{R}^n} u^{2\gamma-2+\frac{p+1}{2}}\nabla u\nabla\varphi^{2\gamma}\right| \le \varepsilon\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} + \int_{\mathbb{R}^n} Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|. \tag{2-26}$$

Combining (2-22), (2-23) and (2-26), we get the inequality

$$\begin{aligned}\int_{\mathbb{R}^n} u^{2\gamma-3}|\nabla u|^2\Delta u\varphi^{2\gamma} &\le \varepsilon\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} + \int_{\mathbb{R}^n} Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\| \\ &\quad - \int_{\mathbb{R}^n} \frac{4}{(4\gamma-3+p)(p+1)}u^{2\gamma+p-1}\varphi^{2\gamma}.\end{aligned} \tag{2-27}$$

Finally, we apply Young's inequality to the third term on the right side of (2-19), and get

$$\begin{aligned}\int_{\mathbb{R}^n} \frac{1}{\gamma^3}u^{-\gamma}(|\nabla u^\gamma|^2)\nabla u^\gamma\nabla\varphi^{2\gamma} &= \int_{\mathbb{R}^n} u^{2\gamma-3}|\nabla u|^2\nabla u\nabla(\varphi^{2\gamma}) \\ &\le \varepsilon\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} + \int_{\mathbb{R}^n} Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|.\end{aligned} \tag{2-28}$$

By (2-19), (2-21), (2-27) and (2-28), we finally obtain

$$\begin{aligned}\left(\tfrac{1}{2}-\varepsilon\right)\int_{\mathbb{R}^n} u^{2\gamma-4}|\nabla u|^4\varphi^{2\gamma} &\le \frac{2}{\gamma^2}\int_{\mathbb{R}^n}(\Delta(u^\gamma\varphi^\gamma))^2 + \int_{\mathbb{R}^n} Cu^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\| \\ &\quad - \int_{\mathbb{R}^n} \frac{4}{(4\gamma-3+p)(p+1)}u^{2\gamma+p-1}\varphi^{2\gamma}. \quad \square\end{aligned}$$

By (2-9), (2-15) and (2-16), since the number ε is arbitrary small in those three places, we have, for $\delta > 0$ sufficiently small,

$$\begin{aligned}(2\text{-}29)\quad &\int_{\mathbb{R}^n}\big(1-4(\gamma-1)^2-\delta\big)(\Delta(u^\gamma\varphi^\gamma))^2\\ &-\int_{\mathbb{R}^n}\left(\frac{\gamma^2}{2\gamma-1}-\frac{8\gamma^2(\gamma-1)^2}{(4\gamma-3+p)(p+1)}\right)u^{p+2\gamma-1}\varphi^{2\gamma}\le\int_{\mathbb{R}^n}C_\delta u^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|,\end{aligned}$$

where C_δ is a positive constant that depends on δ only. Here, we need to require $1-4(\gamma-1)^2>0$, since we have assumed that $\gamma>1$ in Lemma 2.1. So γ is required be in $\big(1,\frac{3}{2}\big)$. If we can choose δ small enough to make $1-4(\gamma-1)^2-\delta$ positive, by the stability property of function u, we obtain

$$(2\text{-}30)\qquad \int_{\mathbb{R}^n}(E-p\delta)u^{p+q}\varphi^{2\gamma}\le\int_{\mathbb{R}^n}C_\delta u^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|,$$

where E is defined to be

$$(2\text{-}31)\qquad E=p(1-4(\gamma-1)^2)-\frac{\gamma^2}{q}+\frac{8\gamma^2(\gamma-1)^2}{(4\gamma-3+p)(p+1)}.$$

Now we take $\varphi=\eta^m$ with m sufficiently large, and choose η a cut-off function satisfying $0\le\eta\le1$, $\eta=1$ for $|x|<R$ and $\eta=0$ for $|x|>2R$. By Young's inequality again, we have

$$\begin{aligned}(2\text{-}32)\quad \int_{\mathbb{R}^n}u^{2\gamma}\|\nabla^4(\varphi^{2\gamma})\|&\le C_\delta R^{-4}\int_{\mathbb{R}^n}u^{2\gamma}\eta^{2\gamma m-4}\\ &\le C_{\delta,\varepsilon}R^{-\frac{4}{1-\theta}}\int_{\mathbb{R}^n}u^2\eta^{2\gamma m-\frac{4}{1-\theta}}+\varepsilon C_\delta\int_{\mathbb{R}^n}u^{2\gamma+p-1}\eta^{2\gamma m},\end{aligned}$$

where $C_{\delta,\varepsilon}$ is a positive constant depends on δ and ε, and θ is a number such that $2(1-\theta)+(2\gamma+p-1)\theta=2\gamma$, so that $0<\theta<1$ for $2<2\gamma<2\gamma+p-1$. By (2-30) and (2-32), we get

$$(2\text{-}33)\qquad (E-p\delta-\varepsilon C_\delta)\int_{\mathbb{R}^n}u^{p+2\gamma-1}\eta^{2\gamma m}\le C_{\delta,\varepsilon}R^{-\frac{4}{1-\theta}}\int_{\mathbb{R}^n}u^2\eta^{2\gamma m-\frac{4}{1-\theta}}.$$

Since θ is strictly less than 1 and will be fixed for given γ, p, we can choose m sufficiently large to make $2\gamma m-\frac{4}{1-\theta}>0$. On the other hand, if $E>0$, we can find small δ and then small ε, such that $E-p\delta-\varepsilon C_\delta>0$. Therefore, by the definition of function η and (2-33), we obtain

$$(2\text{-}34)\qquad (E-p\delta-\varepsilon C_\delta)\int_{B_R}u^{p+2\gamma-1}\le C_{\delta,\varepsilon}R^{-\frac{4}{1-\theta}}\int_{B_{2R}}u^2.$$

By (2.10) of [Wei and Ye 2010], we have $\displaystyle\int_{B_{2R}}u^2\le CR^{n-\frac{8}{p-1}}$, as a result, the

left side of (2-34) is less equal than $C_{\delta,\varepsilon}R^{n-\frac{8}{p-1}-\frac{4}{1-\theta}}$, which tends to 0 as R tends to ∞, provided the power $n-\frac{8}{p-1}-\frac{4}{1-\theta}$ is negative, which is equivalent to $(p+2\gamma-1)>(p-1)\frac{n}{4}$ according to the definition of θ. So, if $(p+2\gamma-1)>(p-1)\frac{n}{4}$ and $E-p\delta-C_\delta\varepsilon>0$, we have $u\equiv 0$.

Thus, we have proved the nonexistence of stable solution to (2-8) if p satisfies the condition $(p+2\gamma-1)>(p-1)\frac{n}{4}$ and $E>0$ (for δ, ε are arbitrary small). By Lemma A.2 in the Appendix, the power p can be in the interval $(\frac{n}{n-8}, 1+\frac{8p^*}{n-4})$. Combining with Theorem 1.1 of [Wei and Ye 2010], we have proved Theorem 1.1, that is, for any $1<p<1+\frac{8p^*}{n-4}$, $n\geq 20$, (2-8) has no stable solution.

3. Proof of Theorem 1.2

In proving Theorem 1.2, it is enough to consider stable solutions u_λ to (1-4), since $u^*=\lim_{\lambda\to\lambda^*}u_\lambda$. Now we give a uniform bound for the stable solutions to (1-4) when $0<d<\lambda<\lambda^*$, where d is a fixed positive constant from $(0,\lambda^*)$.

First, we need to analyze the solution near the boundary. Specifically, we need the regularity of the stable solutions of the equation

$$\text{(3-1)}\qquad \begin{cases} \Delta^2 u=\lambda(u+1)^p,\ \lambda>0 & \text{in } \Omega,\\ u>0 & \text{in } \Omega,\\ u=\Delta u=0 & \text{on } \partial\Omega. \end{cases}$$

near the boundary (as well as their derivatives; see remark after the next theorem).

Theorem 3.1. *Let Ω be a bounded, smooth, and convex domain. There exists a constant C (independent of λ, u) and small positive number ε, such that for stable solutions u to* (3-1) *we have*

$$\text{(3-2)}\qquad u(x)<C \quad \textit{for all } x\in\Omega_\varepsilon:=\{z\in\Omega:\ d(z,\partial\Omega)<\varepsilon\}.$$

Proof. This result is well known. See [Guo and Wei 2009]. For the sake of completeness, we include a proof here. By Lemma 3.5 of [Cowan et al. 2010], we see that there exists a constant C independent of λ, u, such that

$$\text{(3-3)}\qquad \int_\Omega (1+u)^p\,dx\leq C.$$

We write (3-1) as

$$\begin{cases} \Delta u+v=0, & \text{in } \Omega,\\ \Delta v+\lambda(1+u)^p=0, & \text{in } \Omega,\\ u=v=0, & \text{in } \partial\Omega. \end{cases}$$

If we set $f_1(u,v)=v$, $f_2(u,v)=\lambda(u+1)^p$, we see that $\partial f_1/\partial v=1>0$ and $\partial f_2/\partial u=\lambda p(u+1)^{p-1}>0$. Therefore, the convexity of Ω, Lemma 5.1 of [Troy 1981], and the moving plane method near $\partial\Omega$ (as in the appendix of [Guo and

Webb 2002]) imply that there exist $t_0 > 0$ and α which depends only on the domain Ω, such that $u(x - t\nu)$ and $v(x - t\nu)$ are nondecreasing for $t \in [0, t_0]$, $\quad \nu \in \mathbb{R}^n$ satisfying $|\nu| = 1$ and $(\nu, n(x)) \geq \alpha$ and $x \in \partial\Omega$. Therefore, we can find $\rho, \varepsilon > 0$ such that for any $x \in \Omega_\varepsilon := \{z \in \Omega : \ d(z, \partial\Omega) < \varepsilon\}$ there exists a fixed-sized cone Γ_x (with x as its vertex) with

- $\text{meas}(\Gamma_x) \geq \rho$,
- $\Gamma_x \subset \{z \in \Omega : \ d(z, \partial\Omega) < 2\varepsilon\}$, and
- $u(y) \geq u(x)$ for any $y \in \Gamma_x$.

Then, for any $x \in \Omega_\varepsilon$, we have

$$(1 + u(x))^p \leq \frac{1}{\text{meas}(\Gamma_x)} \int_{\Gamma_x} (1 + u)^p \leq \frac{1}{\rho} \int_\Omega (1 + u)^p \leq C.$$

This implies that $(1 + u(x))^p \leq C$, therefore $u(x) \leq C$. □

Remark. By classical elliptic regularity theory, $u(x)$ and its derivatives up to fourth order are bounded on the boundary by a constant independent of u. See [Wei 1996] for more details.

We now turn to the proof of Theorem 1.2 proper, using the ideas of Section 2. Multiplying (1-4) by $(u + 1)^q$ and integrating by parts, we have

$$\text{(3-4)} \quad \int_\Omega \lambda(u+1)^{p+q} = \int_\Omega \Delta^2 u(u+1)^q = \int_{\partial\Omega} \frac{\partial(\Delta u)}{\partial n} + \int_\Omega \Delta(u+1)\Delta(u+1)^q.$$

Setting $v = u + 1$, by direct calculation, we get

$$\int_\Omega (\Delta v^\gamma)^2 = \int_\Omega \gamma^2 v^{2\gamma-2}(\Delta v)^2 + \int_\Omega \gamma^2(\gamma-1)^2 v^{2\gamma-4}|\nabla v|^4 + 2\int_\Omega \gamma^2(\gamma-1)v^{2\gamma-3}\Delta v|\nabla v|^2,$$

$$\int_\Omega \Delta v \Delta v^q = \int_\Omega q(\Delta v)^2 v^{q-1} + \int_\Omega q(q-1)|\nabla v|^2 \Delta v v^{q-2}.$$

From these two equalities and (3-4) we obtain

$$\text{(3-5)} \quad \int_\Omega \left(\frac{q}{\gamma^2}(\Delta v^\gamma)^2 - q(\gamma-1)^2|\nabla v|^4 v^{2\gamma-4} \right) + \int_{\partial\Omega} \frac{\partial(\Delta v)}{\partial n} = \int_\Omega \lambda v^{p+q}.$$

For the second term in (3-5), we have

$$\begin{aligned}(3\text{-}6)\quad &\int_\Omega |\nabla v|^4 v^{2\gamma-4}\\ &= \frac{1}{\gamma^4}\int_\Omega v^{-2\gamma}|\nabla v^\gamma|^4 = \frac{1}{\gamma^4}\int_\Omega |\nabla v^\gamma|^2 \nabla v^\gamma(-\nabla v^{-\gamma})\\ &= \frac{1}{\gamma^4}\int_\Omega \left(-\nabla\frac{|\nabla v^\gamma|^2\nabla v^\gamma}{v^\gamma} + \frac{\nabla(|\nabla v^\gamma|^2)\nabla v^\gamma}{v^\gamma} + \frac{|\nabla v^\gamma|^2\Delta v^\gamma}{v^\gamma}\right)\\ &= \frac{1}{\gamma^4}\int_\Omega v^{-\gamma}\nabla(|\nabla v^\gamma|^2)\nabla v^\gamma + |\nabla v^\gamma|^2\Delta v^\gamma - \frac{1}{\gamma}\int_{\partial\Omega} v^{2\gamma-3}|\nabla v|^2\frac{\partial v}{\partial n}.\end{aligned}$$

A simple calculation yields

$$(3\text{-}7)\quad \frac{1}{\gamma^4}\int_\Omega v^{-\gamma}|\nabla v^\gamma|^2\Delta v^\gamma = \frac{\gamma-1}{\gamma}\int_\Omega v^{2\gamma-4}|\nabla v|^4 + \frac{1}{\gamma}\int_\Omega v^{2\gamma-3}|\nabla v|^2\Delta v.$$

Substituting (3-7) into (3-6), we get

$$\begin{aligned}(3\text{-}8)\quad &\int_\Omega |\nabla v|^4 v^{2\gamma-4}\\ &\qquad = \int_\Omega v^{2\gamma-3}|\nabla v|^2\Delta v + \frac{1}{\gamma^3}\int_\Omega v^{-\gamma}\nabla(|\nabla v^\gamma|^2)\nabla v^\gamma - \int_{\partial\Omega}|\nabla v|^2\frac{\partial v}{\partial n}.\end{aligned}$$

We now estimate the second term on the right side of (3-8). From the proof of Lemma 2.3, together with the identity $\frac{1}{2}\Delta|\nabla v^\gamma|^2 = |\nabla^2 v^\gamma|^2 + \langle\nabla\Delta v^\gamma, \nabla v^\gamma\rangle$, we have

$$\begin{aligned}(3\text{-}9)\quad \frac{1}{\gamma^3}\int_\Omega v^{-\gamma}\nabla(|\nabla v^\gamma|^2)\nabla v^\gamma &\le \frac{1}{2}\int_\Omega |\nabla v|^4 v^{2\gamma-4} + \frac{2}{\gamma^2}\int_\Omega (\Delta v^\gamma)^2\\ &\quad + \frac{1}{\gamma^2}\int_{\partial\Omega}\frac{\partial|\nabla v^\gamma|^2}{\partial n} - \frac{2}{\gamma^2}\int_{\partial\Omega}(\Delta v^\gamma)\frac{\partial v^\gamma}{\partial n}.\end{aligned}$$

By (3-8) and (3-9), thanks to the convexity of the domain Ω, we get

$$\begin{aligned}(3\text{-}10)\quad &\frac{1}{2}\int_\Omega |\nabla v|^4 v^{2\gamma-4}\\ &\qquad \le \int_\Omega v^{2\gamma-3}|\nabla v|^2\Delta v + \frac{2}{\gamma^2}\int_\Omega (\Delta v^\gamma)^2 - (2\gamma-1)\int_{\partial\Omega}|\nabla v|^2\frac{\partial v}{\partial n}.\end{aligned}$$

For the first term on the right side of (3-10), since $v = u+1$, we have $\Delta v = \Delta u < 0$ by maximal principle, and the inequality

$$(3\text{-}11)\qquad \Delta v < -\sqrt{\frac{2\lambda}{p+1}}\,v^{\frac{p+1}{2}} < 0,$$

by Lemma 3.2 of [Cowan et al. 2010]. Thus

$$\int_\Omega v^{2\gamma-3}|\nabla v|^2\Delta v \le \int_\Omega -\sqrt{\frac{2\lambda}{p+1}}\,v^{2\gamma-3+\frac{p+1}{2}}|\nabla v|^2.$$

Moreover, we have

$$\int_\Omega -\sqrt{\frac{2\lambda}{p+1}}v^{2\gamma-3+\frac{p+1}{2}}|\nabla v|^2 = -\int_\Omega \frac{\sqrt{\frac{2\lambda}{p+1}}}{2\gamma-2+\frac{p+1}{2}}\nabla(v^{2\gamma-2+\frac{p+1}{2}}\nabla v) + \int_\Omega \frac{\sqrt{\frac{2\lambda}{p+1}}}{2\gamma-2+\frac{p+1}{2}}v^{2\gamma-2+\frac{p+1}{2}}\Delta v.$$

For the second term on the right, using (3-11) again, we have

$$\int_\Omega \frac{\sqrt{\frac{2\lambda}{p+1}}}{2\gamma-2+\frac{p+1}{2}}v^{2\gamma-2+\frac{p+1}{2}}\Delta v \le -\int_\Omega \frac{\frac{2\lambda}{p+1}}{2\gamma-2+\frac{p+1}{2}}v^{2\gamma+p-1}.$$

Hence, we obtain

$$\int_\Omega v^{2\gamma-3}|\nabla v|^2\Delta v \le -\int_{\partial\Omega}\frac{\sqrt{\frac{2\lambda}{p+1}}}{2\gamma-2+\frac{p+1}{2}}\frac{\partial v}{\partial n} - \int_\Omega \frac{\frac{2\lambda}{p+1}}{2\gamma-2+\frac{p+1}{2}}v^{2\gamma+p-1}, \tag{3-12}$$

where we used $v|_{\partial\Omega} = u+1|_{\partial\Omega} = 1$, for the boundary term in (3-4), (3-10) and (3-12). By the remark after Theorem 3.1, we find that there exists a constant C (the constant C appeared now and later in this section is independent of u), such that

$$\int_{\partial\Omega}\left(|\nabla u|^2\left|\frac{\partial u}{\partial n}\right| + \left|\frac{\partial(\Delta u)}{\partial n}\right| + \left|\frac{\partial u}{\partial n}\right|\right) \le C. \tag{3-13}$$

Combining (3-5), (3-10), (3-12) and (3-13), we get

$$\left(1-4(\gamma-1)^2\right)\int_\Omega(\Delta(u+1)^\gamma)^2 + \left(\frac{8\lambda\gamma^2(\gamma-1)^2}{(4\gamma+p-3)(p+1)} - \frac{\lambda\gamma^2}{q}\right)\int_\Omega(u+1)^{p+q} \le C.$$

If $1-4(\gamma-1)^2 > 0$ and

$$p(1-4(\gamma-1)^2) + \frac{8\gamma^2(\gamma-1)^2}{(4\gamma+p-3)(p+1)} - \frac{\gamma^2}{q} > 0 \tag{3-14}$$

and u is a stable solution to (1-4), we have

$$\left(p(1-4(\gamma-1)^2) + \frac{8\gamma^2(\gamma-1)^2}{(4\gamma+p-3)(p+1)} - \frac{\gamma^2}{2\gamma-1}\right)\int_\Omega(u+1)^{p+q} \le \frac{C}{\lambda}.$$

This leads to $u+1 \in L^{p+q}$.

If $p+q > (p-1)n/4$, then classical regularity theory implies that $u \in L^\infty(\Omega)$.

Therefore we have established the bound of extremal solutions of (1-4) if (3-14) is satisfied and

$$p < \frac{8\gamma+n-4}{n-4}.$$

By Lemma A.2 and Theorem 3.8 of [Wei and Ye 2010], we have proved that the extremal solution u^*, the unique solution of (1-4) (where $\lambda = \lambda^*$), is bounded provided that one of these conditions hold:

(1) If $n \le 8$, then $p > 1$.

(2) If $9 \le n \le 19$, there exists $\varepsilon_n > 0$ such that for any $1 < p < \frac{n}{n-8} + \varepsilon_n$.

(3) If $n \ge 20$, then $1 < p < 1 + \frac{8p^*}{n-4}$, where p^* was defined immediately after Theorem 1.1.

Appendix

In this appendix, we study the inequalities

$$p(1-4(\gamma-1)^2) - \frac{\gamma^2}{2\gamma-1} + \frac{8\gamma^2(\gamma-1)^2}{(4\gamma-3+p)(p+1)} > 0 \tag{A-1}$$

and

$$p < \frac{8\gamma+n-4}{n-4}. \tag{A-2}$$

In order to get a better range for the power p from (A-1) and (A-2), we must study the following equation obtained by letting $p = \frac{8\gamma+n-4}{n-4}$ in (A-1):

$$\frac{8\gamma+n-4}{n-4}\big(1-4(\gamma-1)^2\big) - \frac{\gamma^2}{2\gamma-1} + \frac{8\gamma^2(\gamma-1)^2}{\big(4\gamma-3+\frac{8\gamma+n-4}{n-4}\big)\big(\frac{8\gamma+n-4}{n-4}+1\big)} = 0. \tag{A-3}$$

We need only consider the behavior of (A-3) for $\gamma \in \big(1, \frac{3}{2}\big)$. Through tedious computations, we see that the equation at the bottom of page 495 is the simplified form of (A-3). As a consequence, they have same roots in $\big(1, \frac{3}{2}\big)$.

We denote the left side of (A-3) by $h(\gamma)$. Notice that if $\gamma = \frac{n-4}{n-8}$, then $p = \frac{n}{n-8}$ and $\gamma - 1 = \frac{4}{n-8}$. Hence

$$h\Big(\frac{n-4}{n-8}\Big) = \frac{8}{n-8}(n^4 - 18n^3 - 56n^2 + 384n - 512).$$

In fact, if $n = 20$, then $h\big(\frac{4}{3}\big) = 512 > 0$. On the other hand, it is also easy to see that $h(\frac{3}{2}) < 0$, while it is obvious that $\big(4\gamma - 3 + \frac{8\gamma+n-4}{n-4}\big)\big(\frac{8\gamma+n-4}{n-4} + 1\big) > 0$ and $(2\gamma - 1) > 0$ when $\gamma \in \big(\frac{n-4}{n-8}, \frac{3}{2}\big)$. Therefore, by continuity, (A-3) possesses a root in $(\frac{n-4}{n-8}, \frac{3}{2})$. We denote the smallest root of (A-3) greater than $\frac{n-4}{n-8}$ by p^*. Once we pick out a γ from the interval $(\frac{n-4}{n-8}, p^*)$, $h(\gamma)$ is of course positive. By continuity, we can find a small positive number δ such that the inequality

$$p(1-4(\gamma-1)^2) - \frac{\gamma^2}{2\gamma-1} + \frac{8\gamma^2(\gamma-1)^2}{(4\gamma-3+p)(p+1)} > 0$$

holds when $p \in \left(\frac{8\gamma+n-4}{n-4} - \delta, \frac{8\gamma+n-4}{n-4}\right)$. So, we conclude that when γ runs in the whole interval $\left(\frac{n-4}{n-8}, p^*\right)$, the power p can be in the whole interval $\left(\frac{n}{n-8}, 1 + \frac{8p^*}{n-4}\right)$. We summarize the result as follows:

Lemma A.2. *When $n \geq 20$, the range of p satisfying* (A-1) *and* (A-2) *equals $\left(\frac{n}{n-8}, 1 + \frac{8p^*}{n-4}\right)$, and this interval is not empty.*

References

[Berchio and Gazzola 2005] E. Berchio and F. Gazzola, "Some remarks on biharmonic elliptic problems with positive, increasing and convex nonlinearities", *Electron. J. Differential Equations* **2005** (2005), Paper No. 34, 1–20. MR 2006e:35067 Zbl 1129.35349

[Cowan et al. 2010] C. Cowan, P. Esposito, and N. Ghoussoub, "Regularity of extremal solutions in fourth order nonlinear eigenvalue problems on general domains", *Discrete Contin. Dyn. Syst.* **28**:3 (2010), 1033–1050. MR 2011e:35085 Zbl 1196.35152

[Farina 2007] A. Farina, "On the classification of solutions of the Lane–Emden equation on unbounded domains of $\mathbb{R}^N$", *J. Math. Pures Appl.* (9) **87**:5 (2007), 537–561. MR 2008c:35070 Zbl 1143.35041

[Ferrero et al. 2009] A. Ferrero, H.-C. Grunau, and P. Karageorgis, "Supercritical biharmonic equations with power-type nonlinearity", *Ann. Mat. Pura Appl.* (4) **188**:1 (2009), 171–185. MR 2010g:35087 Zbl 1179.35125

[Gazzola and Grunau 2006] F. Gazzola and H.-C. Grunau, "Radial entire solutions for supercritical biharmonic equations", *Math. Ann.* **334**:4 (2006), 905–936. MR 2007b:35114 Zbl 1152.35034

[Gui et al. 1992] C. Gui, W.-M. Ni, and X. Wang, "On the stability and instability of positive steady states of a semilinear heat equation in $\mathbb{R}^n$", *Comm. Pure Appl. Math.* **45**:9 (1992), 1153–1181. MR 93h:35095 Zbl 0811.35048

[Guo and Webb 2002] Z. Guo and J. R. L. Webb, "Large and small solutions of a class of quasilinear elliptic eigenvalue problems", *J. Differential Equations* **180**:1 (2002), 1–50. MR 2002k:35246 Zbl 1014.35030

[Guo and Wei 2009] Z. Guo and J. Wei, "On a fourth order nonlinear elliptic equation with negative exponent", *SIAM J. Math. Anal.* **40**:5 (2009), 2034–2054. MR 2010b:35125 Zbl 1175.35144

[Guo and Wei 2010] Z. Guo and J. Wei, "Qualitative properties of entire radial solutions for a biharmonic equation with supercritical nonlinearity", *Proc. Amer. Math. Soc.* **138**:11 (2010), 3957–3964. MR 2012a:35054 Zbl 1203.35105

[Lin 1998] C.-S. Lin, "A classification of solutions of a conformally invariant fourth order equation in $\mathbb{R}^n$", *Comment. Math. Helv.* **73**:2 (1998), 206–231. MR 99c:35062 Zbl 0933.35057

[Troy 1981] W. C. Troy, "Symmetry properties in systems of semilinear elliptic equations", *J. Differential Equations* **42**:3 (1981), 400–413. MR 83b:35051 Zbl 0486.35032

[Wei 1996] J. Wei, "Asymptotic behavior of a nonlinear fourth order eigenvalue problem", *Comm. Partial Differential Equations* **21**:9-10 (1996), 1451–1467. MR 97h:35066 Zbl 0872.35013

[Wei and Xu 1999] J. Wei and X. Xu, "Classification of solutions of higher order conformally invariant equations", *Math. Ann.* **313**:2 (1999), 207–228. MR 2000a:58093 Zbl 0940.35082

[Wei and Ye 2010] J. Wei and D. Ye, "Liouville theorems for finite Morse index solutions of biharmonic problem", preprint, 2010, http://www.math.cuhk.edu.hk/∼wei/4thstable29-10-10.pdf. To appear in *Math. Ann.*

[Xu 2000] X. Xu, "Uniqueness theorem for the entire positive solutions of biharmonic equations in $\mathbb{R}^n$", *Proc. Roy. Soc. Edinburgh Sect. A* **130**:3 (2000), 651–670. MR 2001f:35143 Zbl 0961.35037

Received November 1, 2011. Revised April 24, 2012.

JUNCHENG WEI
DEPARTMENT OF MATHEMATICS
THE CHINESE UNIVERSITY OF HONG KONG
SHATIN, NT
HONG KONG
wei@math.cuhk.edu.hk

XINGWANG XU
DEPARTMENT OF MATHEMATICS
NATIONAL UNIVERSITY OF SINGAPORE
BLOCK S17 (SOC1)
10 LOWER KENT RIDGE ROAD
SINGAPORE 119076
SINGAPORE
matxuxw@nus.edu.sg

WEN YANG
DEPARTMENT OF MATHEMTICS
CHINESE UNIVERSITY OF HONG KONG
SHATIN, NT
HONG KONG
wyang@math.cuhk.edu.hk

CONTENTS

Volume 263, 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 263 No. 2 June 2013

0030-8730(201306)263:2;1-5