Algebra & Number Theory

Volume 7

2013

No. 5

msp

Algebra & Number Theory

msp.org/ant

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

The subscription price for 2013 is US $200/year for the electronic version, and $350/year (+$40, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscribers address should be sent to MSP.

Algebra & Number Theory (ISSN 1944-7833 electronic, 1937-0652 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

ALGEBRA AND NUMBER THEORY 7:5 (2013)
dx.doi.org/10.2140/ant.2013.7.1019

Sharp upper bounds for the Betti numbers of a given Hilbert polynomial

Giulio Caviglia and Satoshi Murai

We show that there exists a saturated graded ideal in a standard graded polynomial ring which has the largest total Betti numbers among all saturated graded ideals for a fixed Hilbert polynomial.

1. Introduction

A classical problem consists in studying the number of minimal generators of ideals in a local or a graded ring in relation to other invariants of the ring and of the ideals themselves. In particular, a great amount of work has been done to establish bounds for the number of generators in terms of certain invariants, for instance, multiplicity, Krull dimension, and Hilbert functions (see [Macaulay 1927; Sally 1978]). An important result was proved in [Elías et al. 1991], where the authors established a sharp upper bound for the number of generators $\nu(I)$ of all perfect ideals I in a regular local ring $(R, \boldsymbol{m}, K)$ (or in a polynomial ring over a field K) in terms of their multiplicity and their height.

In a subsequent paper, Valla [1994] provides, under the same hypotheses, sharp upper bounds for every Betti number $\beta_i^R(I) = \dim_K \operatorname{Tor}_i^R(I, K)$; notice that with this notation $\beta_0^R(I) = \nu(I)$. More surprisingly, Valla proved that among all perfect ideals with a fixed multiplicity and height in a formal power series ring over a field K, there exists one which has the largest possible Betti numbers β_i.

The main result of this paper is an extension of Valla's theorem. We will consider both the local and the graded case, although the result we present for the local case follows directly from the graded case.

We first consider the graded case. We show that for every fixed Hilbert polynomial $p(t)$, there exists a point Y in the Hilbert scheme $\operatorname{Hilb}_{\mathbb{P}^{n-1}}^{p(t)}$ such that $\beta_i(I_Y) \geq \beta_i(I_X)$ for all i and for all $X \in \operatorname{Hilb}_{\mathbb{P}^{n-1}}^{p(t)}$. Equivalently, let $S = K[X_1, \ldots, X_n]$ be a standard graded polynomial ring over a field K. We prove:

The first author was supported by a grant from the Simons Foundation (209661 to G. C.).
The second author was supported by KAKENHI 22740018.
MSC2010: primary 13D02; secondary 13D40.
Keywords: graded Betti numbers, Hilbert polynomials.

Theorem 1.1. *Let $p(t)$ be the Hilbert polynomial of a graded ideal of S. There exists a saturated graded ideal $L \subset S$ with Hilbert polynomial $p(t)$ such that $\beta_i^S(S/L) \geq \beta_i^S(S/I)$ for all i and for all saturated graded ideals $I \subset S$ with Hilbert polynomial $p(t)$.*

Notice that Valla's result corresponds to the special case of the theorem when $p(t)$ is constant.

An important result in the study of upper bounds for Betti numbers is the Bigatti–Hulett–Pardue theorem, which shows that the lex ideal has the largest Betti numbers among all homogeneous ideals in a standard graded polynomial ring for a fixed Hilbert function. By using the Bigatti–Hulett–Pardue theorem, we reduce Theorem 1.1 to a certain combinatorial problem on lex ideals, and prove the theorem by purely combinatorial methods.

We have chosen to not present an explicit formula of the bounds. We are convinced that such a formula, in the general case, would be hard to read and to interpret. Instead, as a part of the proof, we describe the construction of the lex ideal that achieves the bound. Using the Eliahou–Kervaire resolution it is possible to write an explicit formula for the total Betti numbers of every lex ideal in terms of its minimal generators.

In particular, explicit computations of the bounds can be carried out for a given Hilbert polynomial. Thus, it would be possible to describe an explicit formula of the bounds for classes of simple enough Hilbert polynomials. For example, in the special case when the Hilbert polynomials are constant, such a formula was given by Valla [1994].

Theorem 1.1 induces the following upper bounds of Betti numbers of ideals in a regular local ring (see Section 3 for the proof): For a regular local ring $(R, \boldsymbol{m}, K)$ and an ideal $I \subset R$, let $\boldsymbol{p}_{R/I}(t)$ be the Hilbert–Samuel polynomial of R/I with respect to $\boldsymbol{m}$ (see [Bruns and Herzog 1998, §4.6]).

Theorem 1.2. *Let $(R, \boldsymbol{m}, K)$ be a regular local ring of dimension n, and let $\boldsymbol{p}(t)$ be a polynomial such that there is an ideal $J \subset R$ such that $\boldsymbol{p}(t) = \boldsymbol{p}_{R/J}(t)$. There exists an ideal L in $A = K[\![x_1, \dots, x_n]\!]$ with $\boldsymbol{p}_{A/L}(t) = \boldsymbol{p}(t)$ such that $\beta_i^A(A/L) \geq \beta_i^R(R/I)$ for all i and for all ideals $I \subset R$ with $\boldsymbol{p}_{R/I}(t) = \boldsymbol{p}(t)$.*

Unfortunately, the combinatorial part of the proof of Theorem 1.1 is very long and complicated. Moreover, a construction of ideals which achieve the bound is not easy to understand. Thus, it would be desirable to get a simpler proof of the theorem and to get a better understanding for the structure of ideals which attain maximal Betti numbers.

The paper is structured in the following way: In Sections 2 and 3, we reduce a problem of Betti numbers to a problem of combinatorics of lexicographic sets of monomials with a special structure. In Section 4, we introduce key techniques to prove the main result. In particular, we give a new proof of Valla's result. In Section 5, a construction of ideals which attain maximal Betti numbers of Theorem 1.1 will be given. In Section 6, we give a proof of the main combinatorial result about lexicographic sets of monomials, which essentially proves Theorem 1.1. In Section 7, some examples of ideals with maximal Betti numbers are given.

2. Universal lex ideals

In this section, we introduce basic notations which are used in the paper.

Let $S = K[x_1, \dots, x_n]$ be a standard graded polynomial ring over a field K. Let M be a finitely generated graded S-module. The *Hilbert function* $H(M, -) : \mathbb{Z} \to \mathbb{Z}$ of M is the numerical function defined by

$$H(M, k) = \dim_K M_k$$

for all $k \in \mathbb{Z}$, where M_k is the graded component of M of degree k. We denote $P_M(t)$ by the Hilbert polynomial of M. Thus $P_M(t)$ is a polynomial in t satisfying $P_M(k) = H(M, k)$ for $k \gg 0$. The numbers

$$\beta^S_{i,j}(M) = \dim_K \operatorname{Tor}^S_i(M, K)_j$$

are called the *graded Betti numbers* of M, and $\beta^S_i(M) = \sum_{j \in \mathbb{Z}} \beta^S_{i,j}(M)$ are called the (*total*) *Betti numbers* of M.

A set of monomials $W \subset S$ is said to be *lex* if, for all monomials $u \in W$ and $v >_{\text{lex}} u$ of the same degree, one has $v \in W$, where $>_{\text{lex}}$ is the lexicographic order induced by the ordering $x_1 >_{\text{lex}} \cdots >_{\text{lex}} x_n$. A monomial ideal $I \subset S$ is said to be *lex* if the set of monomials in I is lex. By the classical Macaulay's theorem [1927], for any graded ideal $I \subset S$ there exists the unique lex ideal $L \subset S$ with the same Hilbert function as I. Moreover, Bigatti [1993], Hulett [1993], and Pardue [1996] proved that lex ideals have the largest graded Betti numbers among all graded ideals having the same Hilbert function.

For any graded ideal $I \subset S$, let

$$\operatorname{sat} I = (I : \mathfrak{m}^\infty)$$

be the *saturation* of $I \subset S$, where $\mathfrak{m} = (x_1, \dots, x_n)$ is the graded maximal ideal of S. A graded ideal I is said to be *saturated* if $I = \operatorname{sat} I$. It is well-known that I is saturated if and only if $\operatorname{depth}(S/I) > 0$ or $I = S$.

Let $L \subset S$ be a lex ideal. Then $\operatorname{sat} L$ is also a lex ideal. It is natural to ask which lex ideals are saturated. The theory of universal lex ideals gives an answer.

A lex ideal $L \subset S$ is said to be *universal* if $LS[x_{n+1}]$ is also a lex ideal in $S[x_{n+1}]$. The following are fundamental results on universal lex ideals:

Lemma 2.1 [Murai and Hibi 2008]. *Let $L \subset S$ be a lex ideal. The following conditions are equivalent*:

(i) *L is universal.*

(ii) *L is generated by at most n monomials.*

(iii) *$L = S$ or there exist integers $a_1, a_2, \dots, a_t \geq 0$ with $1 \leq t \leq n$ such that*

$$L = (x_1^{a_1+1}, x_1^{a_1}x_2^{a_2+1}, \dots, x_1^{a_1}x_2^{a_2}\cdots x_{t-1}^{a_{t-1}}x_t^{a_t+1}). \tag{1}$$

A relation between universal lex ideals and saturated lex ideals is the following:

Lemma 2.2 [Murai and Hibi 2008]. *Let $L \subsetneq S$ be a lex ideal. Then* $\operatorname{depth}(S/L) > 0$ *if and only if L is generated by at most $n-1$ monomials.*

A lex ideal $I \subset S$ is called a *proper universal lex ideal* if I is generated by at most $n-1$ monomials or $I = S$.

Let $I \subset S$ be a graded ideal. Then there exists the unique lex ideal $L \subset S$ with the same Hilbert function as I. Then $\operatorname{sat} L$ is a proper universal lex ideal with the same Hilbert polynomial as I. This construction $I \to \operatorname{sat} L$ gives a one-to-one correspondence between Hilbert polynomials of graded ideals and proper universal lex ideals:

Proposition 2.3. *For any graded ideal $I \subset S$ there exists the unique proper universal lex ideal $L \subset S$ with the same Hilbert polynomial as I.*

Proof. The existence is obvious. What we must prove is that, if L and L' are proper universal lex ideals with the same Hilbert polynomial then $L = L'$.

Since L and L' have the same Hilbert polynomial, their Hilbert functions coincide in sufficiently large degrees. This fact shows $L_d = L'_d$ for $d \gg 0$. Thus $\operatorname{sat} L = \operatorname{sat} L'$. Since L and L' are saturated, $L = \operatorname{sat} L = \operatorname{sat} L' = L$. □

3. 1-lexicographic ideals, Betti numbers and max sequences

In this section, we reduce a problem of Betti numbers of graded ideals to a problem of combinatorics of lex sets of monomials.

Let $S = K[x_1, \dots, x_n]$ and $\bar{S} = K[x_1, \dots, x_{n-1}]$. For a monomial ideal $I \subset S$, let $\bar{I} = I \cap \bar{S}$. A monomial ideal $I \subset S$ is said to be 1-*lexicographic* if x_n is a nonzero divisor of S/I and $\bar{I}$ is a lex ideal of $\bar{S}$.

Lemma 3.1 [Iyengar and Pardue 1999, Proposition 4]. *For any saturated graded ideal $I \subset S$, there exists a 1-lexicographic ideal $J \subset S$ with the same Hilbert function as I such that $\beta_{i,j}^S(I) \leq \beta_{i,j}^S(J)$ for all i, j.*

Lemma 3.2. *Let $J \subset S$ be a* 1*-lexicographic ideal. Then*:

(i) $\dim_K J_d = \sum_{k=0}^{d} \dim_K \bar{J}_k$ *for all* $d \geq 0$.

(ii) $\beta_i^S(J) = \beta_i^{\bar{S}}(\bar{J})$ *for all* i.

Proof. Condition (ii) is obvious since x_n is regular on S/J. Also, for all $d \geq 0$, we have a decomposition $J_d = \bigoplus_{k=0}^{d} J_k x_n^{d-k}$ as K-vector spaces. This equality proves (i). □

Corollary 3.3. *Let J and J' be* 1*-lexicographic ideals in S. If J and J' have the same Hilbert polynomial then $\bar{J}_d = \bar{J}'_d$ for $d \gg 0$.*

Proof. Lemma 3.2(i) says that $\dim_K J_d - \dim_K J_{d-1} = \dim \bar{J}_d$, so

$$\dim_K \bar{J}_d = \dim_K \bar{J}'_d \quad \text{for } d \gg 0.$$

Then the statement follows since $\bar{J}$ and $\bar{J}'$ are lex. □

Next, we describe all 1-lexicographic ideals in S. By Proposition 2.3, fixing a Hilbert polynomial is equivalent to fixing a proper universal lex ideal U. For a proper universal lex ideal $U \subset S$, let

$$\mathscr{L}(U)$$
$$= \{I \subset \bar{S} : I \text{ is a lex ideal with } I \subset \operatorname{sat} \bar{U} \text{ and } \dim_K(\operatorname{sat} \bar{U})/I = \dim_K(\operatorname{sat} \bar{U})/\bar{U}\}.$$

Note that $\dim_K(\operatorname{sat} J)/J$ is finite for any graded ideal $J \subset S$ since $(\operatorname{sat} J)/J$ is isomorphic to the zeroth local cohomology module $H^0_{\mathfrak{m}}(S/J)$. By using Lemma 3.2, it is easy to see that if $I \in \mathscr{L}(U)$ then IS has the same Hilbert polynomial as U. Actually, the converse is also true.

Lemma 3.4. *Let U be a proper universal lex ideal. If J is a* 1*-lexicographic ideal such that $P_J(t) = P_U(t)$ then $\bar{J} \in \mathscr{L}(U)$.*

Proof. By Corollary 3.3 we have $\bar{U}_d = \bar{J}_d$ for $d \gg 0$, so $\operatorname{sat} \bar{U} = \operatorname{sat} \bar{J}$. Also, since U and J have the same Hilbert polynomial, for $d \gg 0$, one has

$$\dim_K U_d = \sum_{k=0}^{d} \dim_K \bar{U}_k = \sum_{k=0}^{d} \dim_K(\operatorname{sat} \bar{U}_k) - \dim_K(\operatorname{sat} \bar{U}/\bar{U})$$

and

$$\dim_K J_d = \sum_{k=0}^{d} \dim_K \bar{J}_k = \sum_{k=0}^{d} \dim_K(\operatorname{sat} \bar{J}_k) - \dim_K(\operatorname{sat} \bar{J}/\bar{J}).$$

Since $\operatorname{sat} \bar{J} = \operatorname{sat} \bar{U}$, we have $\dim_K(\operatorname{sat} \bar{J}/\bar{J}) = \dim_K(\operatorname{sat} \bar{U}/\bar{U})$ and $\bar{J} \in \mathscr{L}(U)$. □

By Lemmas 3.1 and 3.4, to prove Theorem 1.1, it is enough to find a lex ideal which has the largest Betti numbers among all ideals in $\mathscr{L}(U)$. We consider a more general setting. For any universal lex ideal $U \subset S$ (not necessarily proper) and for any positive integer $c > 0$, define

$$\mathscr{L}(U; c) = \{I \subset U : I \text{ is a lex ideal with } \dim_K U/I = c\}.$$

We consider the Betti numbers of ideals in $\mathscr{L}(U; c)$.

We first discuss Betti numbers of lex ideals. We need the following notation: For any monomial $u \in S$, let $\max u$ be the largest integer ℓ such that x_ℓ divides u, where $\max(1) = 1$. For a set of monomials (or a K-vector space spanned by monomials) M, let

$$m_{\le i}(M) = \#\{u \in M : \max u \le i\}$$

for $i = 1, 2, \dots, n$, where $\#X$ is the cardinality of a finite set X, and

$$m(M) = \big(m_{\le 1}(M), m_{\le 2}(M), \dots, m_{\le n}(M)\big).$$

These numbers are often used to study Betti numbers of lex ideals. The next formula was proved by Bigatti [1993] and Hulett [1993], by using the famous Eliahou–Kervaire resolution [1990].

Lemma 3.5. *Let $I \subset S$ be a lex ideal. Then, for all i, j,*

$$\beta^S_{i,i+j}(I) = \binom{n-1}{i} \dim_K I_j - \sum_{k=1}^{n} \binom{k-1}{i} m_{\le k}(I_{j-1}) - \sum_{k=1}^{n-1} \binom{k-1}{i-1} m_{\le k}(I_j).$$

For vectors $\boldsymbol{a} = (a_1, \dots, a_n)$, $\boldsymbol{b} = (b_1, \dots, b_n) \in \mathbb{Z}^n$, we define

$$\boldsymbol{a} \succeq \boldsymbol{b} \Leftrightarrow a_i \ge b_i \quad \text{for } i = 1, 2, \dots, n.$$

Corollary 3.6. *Let U be a universal lex ideal and $I, J \in \mathscr{L}(U; c)$. Let $\mathcal{M}_I$ (resp. $\mathcal{M}_J$) be the set of all monomials in $U \setminus I$ (resp. $U \setminus J$). If $m(\mathcal{M}_I) \succeq m(\mathcal{M}_J)$ then $\beta^S_i(I) \ge \beta^S_i(J)$ for all i.*

Proof. Observe that $\beta^S_{i,i+j}(I) = \beta^S_{i,i+j}(J) = 0$ for $j \gg 0$. Thus, for $d \gg 0$, we have $\beta^S_i(I) = \sum_{j=0}^{d} \beta^S_{i,i+j}(I)$. Let $I_{\le d} = \bigoplus_{k=0}^{d} I_k$. Then by Lemma 3.5,

$$\beta^S_i(I) = \binom{n-1}{i} \dim_K I_{\le d} - \sum_{k=1}^{n} \binom{k-1}{i} m_{\le k}(I_{\le d-1}) - \sum_{k=1}^{n-1} \binom{k-1}{i-1} m_{\le k}(I_{\le d})$$

and the same formula holds for J. Since, for $d \gg 0$,

$$m(J_{\le d}) = m(U_{\le d}) - m(\mathcal{M}_J) \succeq m(U_{\le d}) - m(\mathcal{M}_I) = m(I_{\le d}),$$

we have $\beta^S_i(I) \ge \beta^S_i(J)$ for all i, as desired. □

Next, we study the structure of $\mathcal{M}_I$. Let

$$U = (x_1^{a_1+1}, x_1^{a_1} x_2^{a_2+1}, \dots, x_1^{a_1} x_2^{a_2} \cdots x_{t-1}^{a_{t-1}} x_t^{a_t+1})$$

be a universal lex ideal, $\delta_i = x_1^{a_1} \cdots x_{i-1}^{a_{i-1}} x_i^{a_i+1}$, and $b_i = a_1 + \cdots + a_i + 1 = \deg \delta_i$. (If $U = S$ then $t = 1$ and $a_1 = -1$.) Let

$$S^{(i)} = K[x_i, \dots, x_n].$$

Then, as K-vector spaces, we have a decomposition

$$U = \delta_1 S^{(1)} \oplus \delta_2 S^{(2)} \oplus \cdots \oplus \delta_t S^{(t)}.$$

Definition 3.7. A set of monomials $N \subset S^{(i)}$ is said to be *revlex* if, for all monomials $u \in N$ and $v <_{\text{lex}} u$ of the same degree, one has $v \in N$. Moreover, N is said to be *super-revlex* (in $S^{(i)}$) if it is revlex and $u \in N$ implies $v \in N$ for any monomial $v \in S^{(i)}$ of degree $\leq \deg u - 1$. A *multicomplex* is a set of monomials $N \subset S^{(i)}$ satisfying that $u \in N$ and $v | u$ imply $v \in N$. Thus a multicomplex is the complement of the set of monomials in a monomial ideal. Note that super-revlex sets are multicomplexes.

Let $I \in \mathscr{L}(U; c)$ and $\mathcal{M}_I$ be the set of monomials in $U \setminus I$. Then we can uniquely write

$$\mathcal{M}_I = \delta_1 M_{\langle 1 \rangle} \uplus \delta_2 M_{\langle 2 \rangle} \uplus \cdots \uplus \delta_t M_{\langle t \rangle},$$

where $M_{\langle i \rangle} \subset S^{(i)}$ and $\uplus$ denotes the disjoint union. The following facts are obvious:

Lemma 3.8. (i) *Each $M_{\langle i \rangle}$ is a revlex multicomplex.*

(ii) *If $\delta_i M_{\langle i \rangle}$ has a monomial of degree d then $\delta_{i+1} M_{\langle i+1 \rangle}$ contains all monomials of degree d in $\delta_{i+1} S^{(i+1)}$ for all d.*

Lemma 3.8(ii) is equivalent to saying that if $M_{\langle i \rangle}$ contains a monomial of degree d then $M_{\langle i+1 \rangle}$ contains all monomials of degree $d - a_{i+1}$ in $S^{(i+1)}$.

We say that a set of monomials

$$M = \delta_1 M_{\langle 1 \rangle} \uplus \delta_2 M_{\langle 2 \rangle} \uplus \cdots \uplus \delta_t M_{\langle t \rangle} \subset U,$$

where $M_{\langle i \rangle} \subset S^{(i)}$, is a *ladder set* if it satisfies conditions (i) and (ii) of Lemma 3.8. The next result is the key result in this paper:

Proposition 3.9. *Let $U \subset S$ be a universal lex ideal. For any integer $c \geq 0$, there exists a ladder set $N \subset U$ with $\#N = c$ such that for any ladder set $M \subset U$ with $\#M = c$ one has*

$$m(N) \succeq m(M).$$

We prove Proposition 3.9 in Section 6. Here, we prove Theorem 1.1 by using Proposition 3.9.

Proof of Theorem 1.1. Let $U \subset S$ be a proper universal lex ideal with $P_U(t) = p(t)$ and $\bar{U} = U \cap \bar{S}$. Let $c = \dim_K(\operatorname{sat} \bar{U} / \bar{U})$. For any lex ideal $I \subset \operatorname{sat} \bar{U}$, let $\mathcal{M}_I$ be the set of monomials in $(\operatorname{sat} \bar{U} \setminus I)$.

Let $N \subset \operatorname{sat} \bar{U}$ be a ladder set of monomials with $\#N = c$ given in Proposition 3.9. Consider the ideal $J \subset \bar{S}$ generated by all monomials in $\operatorname{sat} \bar{U} \setminus N$. Then $J \subset \operatorname{sat} \bar{U}$ and $\mathcal{M}_J = N$. In particular, $J \in \mathscr{L}(U)$.

Let $L = JS$. By construction, $P_L(t) = P_U(t) = p(t)$. We claim that L satisfies the desired conditions. Let $I \subset S$ be a saturated graded ideal with $P_I(t) = p(t)$. By Lemmas 3.1 and 3.4, we may assume that I is a 1-lexicographic ideal with $\bar{I} \in \mathscr{L}(U) = \mathscr{L}(\operatorname{sat} \bar{U}; c)$. Since $\mathcal{M}_{\bar{I}}$ is a ladder set, by the choice of J, $m(\mathcal{M}_J) \succeq m(\mathcal{M}_{\bar{I}})$. Then, by Corollary 3.6,

$$\beta_i^S(L) = \beta_i^{\bar{S}}(J) \geq \beta_i^{\bar{S}}(\bar{I}) = \beta_i^S(I)$$

for all i, as desired. □

Another interesting corollary of Proposition 3.9 is:

Corollary 3.10. *Let $U \subset S$ be a universal lex ideal and $c \geq 0$. There exists a lex ideal $L \subset U$ with $\dim_K U/L = c$ such that, for any graded ideal $I \subset U$ with $\dim_K U/I = c$, one has $\beta_i^S(L) \geq \beta_i^S(I)$ for all i.*

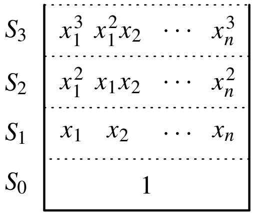
Proof of Theorem 1.2. Let I be an ideal in a regular local ring $(R, \boldsymbol{m}, K)$ such that $\boldsymbol{p}_{R/I}(t) = \boldsymbol{p}(t)$. Then the associated graded ring $\operatorname{gr}_{\boldsymbol{m}}(R/I)$ has the same Hilbert–Samuel polynomial as R/I. Also, we may regard $\operatorname{gr}_{\boldsymbol{m}}(R/I)$ as a quotient of a standard graded polynomial ring $S = K[x_1, \dots, x_n]$ (see [Bruns and Herzog 1998, Proposition 2.2.5]), and it is known that $\beta_i^R(R/I) \leq \beta_i^S(\operatorname{gr}_{\boldsymbol{m}}(R/I))$ for all i (see [Robbiano 1981; Herzog et al. 1986]).

Let $S' = S[x_{n+1}]$. By adjoining a variable to $\operatorname{gr}_{\boldsymbol{m}}(R/I)$ we obtain a graded ring that is isomorphic to S'/J for a saturated graded ideal $J \subset S'$. Then $\boldsymbol{p}_{\operatorname{gr}_{\boldsymbol{m}}(R/I)}(t)$ is equal to the Hilbert polynomial of S'/J and $\beta_i^S(\operatorname{gr}_{\boldsymbol{m}}(R/I)) = \beta_i^{S'}(S'/J)$ for all i. Let $L' \subset S'$ be the saturated ideal with the same Hilbert polynomial as J given in Theorem 1.1. Observe that L' has no generators which are divisible by x_{n+1} by the construction given in the proof of Theorem 1.1.

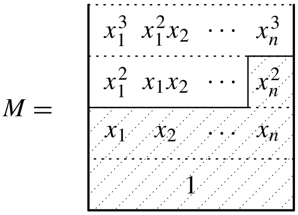
Let $L \subset A = K[\![x_1, \dots, x_n]\!]$ be a monomial ideal having the same generators as L'. We claim that L satisfies the desired conditions. By construction, the Hilbert–Samuel polynomial of A/L is equal to the Hilbert polynomial of S'/L' and $\beta_i^A(A/L) = \beta_i^{S'}(S'/L')$ for all i. Since $\beta_i^R(R/I) \leq \beta_i^{S'}(S'/J) \leq \beta_i^{S'}(S'/L')$ and $\boldsymbol{p}_{R/I}(t) = P_{S'/J}(t) = P_{S'/L'}(t)$, the ideal L satisfies the desired conditions. □

4. Some tools to study max sequence

In this section, we introduce some tools to study $m(-)$. Let $S = K[x_1, \dots, x_n]$ and $\hat{S} = K[x_2, \dots, x_n]$. From now on, we identify vector spaces spanned by monomials (such as polynomial rings and monomial ideals) with the set of monomials in the spaces. First, we introduce pictures, which help to understand the proofs. We associate with the set of monomials in S the following picture:


S_3
x_1^3 $x_1^2x_2$ $\cdots$ x_n^3
S_2
x_1^2 x_1x_2 $\cdots$ x_n^2
S_1
x_1 x_2 $\cdots$ x_n
S_0
1

Each block represents a set of monomials in S of a fixed degree ordered by the lex order. We represent a set of monomials $M \subset S$ by a shaded picture so that the set of monomials in the shade is equal to M. For example, here is a representation of the set $M = \{1, x_1, x_2, \dots, x_n, x_n^2\}$:


$M =$
x_1^3 $x_1^2x_2$ $\cdots$ x_n^3
x_1^2 x_1x_2 $\cdots$ x_n^2
x_1 x_2 $\cdots$ x_n
1

Definition 4.1. We define the *opposite degree lex order* $>_{\text{opdlex}}$ by $u >_{\text{opdlex}} v$ if

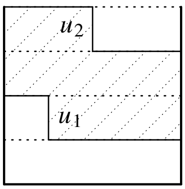
(i) $\deg u < \deg v$ or

(ii) $\deg u = \deg v$ and $u >_{\text{lex}} v$.

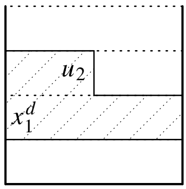
For monomials $u_1 \geq_{\text{opdlex}} u_2$, let

$$[u_1, u_2] = \{v \in S : u_1 \geq_{\text{opdlex}} v \geq_{\text{opdlex}} u_2\}.$$

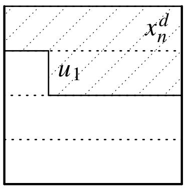
A set of monomials $M \subset S$ is called an *interval* if $M = [u_1, u_2]$ for some monomials $u_1, u_2 \in S$. Moreover, we say that M is a *lower lex set of degree d* if $M = [x_1^d, u_2]$, and that M is an *upper revlex set of degree d* if $M = [u_1, x_n^d]$ (see figure).



Interval



Lower lex set



Upper rev-lex set

A benefit of considering pictures is that we can visualize the map $\rho : S \to \hat{S}$ defined as follows. For any monomial $x_1^k u \in S$ with $u \in \hat{S}$, let

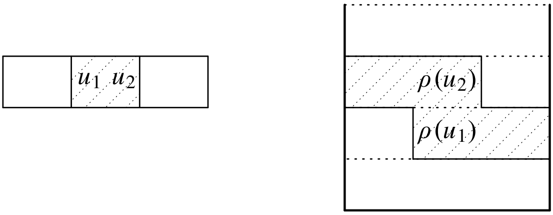
$$\rho(x_1^k u) = u.$$

This induces a bijection

$$\rho : S_d = \bigoplus_{k=0}^{d} x_1^k \hat{S}_{d-k} \longrightarrow \hat{S}_{\le d} = \bigoplus_{k=0}^{d} \hat{S}_k$$

$$x_1^k u \longrightarrow u.$$

It is easy to see that if $[u_1, u_2] \subset S_d$ then $\rho([u_1, u_2]) = [\rho(u_1), \rho(u_2)]$ is an interval in $\hat{S}$:



$[u_1, u_2] \subset S_d$ $\rho([u_1, u_2]) \subset \hat{S}_{\le d}$

In particular:

Lemma 4.2. *Let $M \subset S_d$ be a set of monomials.*

(i) *If M is lex then $\rho(M)$ is a lower lex set of degree* 0 *in $\hat{S}$.*

(ii) *If M is revlex then $\rho(M)$ is an upper revlex set of degree d in $\hat{S}$.*

We define $\max(1) = 1$ in S and $\max(1) = 2$ in $\hat{S}$. For any monomial $u \in S_d$ with $u \neq x_1^d$, one has $\max(u) = \max(\rho(u))$. Hence:

Lemma 4.3. *Let $M \subset S_d$ be a set of monomials. One has $m(M) \succeq m(\rho(M))$. Moreover, if $x_1^d \notin M$ then $m(M) = m(\rho(M))$.*

Lemma 4.4 (Interval Lemma). *Let $[u_1, u_2]$ be an interval in S, $0 \le a \le \deg u_1$, and $b \ge \deg u_2$. Let $L \subset S$ be the lower lex set of degree a and R the upper revlex set of degree b with $\#L = \#R = \#[u_1, u_2]$. Then*

$$m(L) \succeq m\big([u_1, u_2]\big) \succeq m(R).$$

Proof. We use double induction on n and $\#[u_1, u_2]$. The statement is obvious if $n = 1$ or if $\#[u_1, u_2] = 1$. Suppose $n > 1$ and $\#[u_1, u_2] > 1$.

Case 1. We first prove the statement when $[u_1, u_2]$, L, and R are contained in a single component S_d for some degree d. We may assume $L \neq [u_1, u_2]$ and $L \neq R$. Then, since $x_1^d \notin [u_1, u_2]$, $m([u_1, u_2]) = m(\rho([u_1, u_2]))$ and $m(R) = m(\rho(R))$. Since $\rho(L) \subset \hat{S}_{\le d}$ is a lower lex set of degree 0, $\rho([u_1, u_2]) \subset \hat{S}_{\le d}$ is an interval, and $\rho(R) \subset \hat{S}_{\le d}$ is an upper revlex set of degree d in $\hat{S}$. By the induction hypothesis, we have

$$m(L) \succeq m(\rho(L)) \succeq m(\rho([u_1, u_2])) \succeq m(\rho(R)) = m(R).$$

Then the statement follows since $m\big(\rho([u_1, u_2])\big) = m([u_1, u_2])$.

Case 2. Now we prove the statement in general. We first prove the statement for L. We identify S_i with the set of monomials in S of degree i. Suppose $\#[u_1, u_2] > \#S_a$. Then there exist $u_1', u_2' \in S$ such that

$$[u_1, u_2] = [u_1, u_2'] \uplus [u_1', u_2]$$

and $\#[u_1, u_2'] = \#S_a$. Let L' be the lower lex set of degree $a+1$ with $\#L' = \#[u_1', u_2]$. By the induction hypothesis, $m(S_a) \succeq m([u_1, u_2'])$ and $m(L') \succeq m([u_1', u_2])$. Thus

$$m([u_1, u_2]) \preceq m(S_a \uplus L') = m(L).$$

Suppose $\#[u_1, u_2] \le \#S_a$. Then $L \subset S_a$. Let $d = \deg u_1$ and $A \subset S_d$ be the lex set with $\#A = \#[u_1, u_2]$. Then $A = x_1^{d-a} L$. Since $m(A) = m(L)$, what we must prove is:

$$m(A) \succeq m([u_1, u_2]).$$

Since $\#[u_1, u_2] \le \#S_a \le \#S_{d+1}$, we have $\deg u_2 \le d+1$.

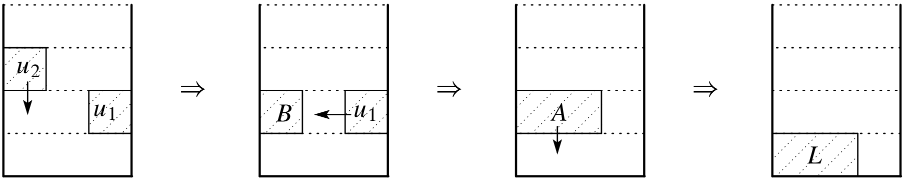
If $\deg u_2 = d$ then $[u_1, u_2] \subset S_d$. Then the desired inequality follows from Case 1. Suppose $\deg u_2 = d+1$. Then

$$[u_1, u_2] = [u_1, x_n^d] \uplus [x_1^{d+1}, u_2].$$

Recall $\#[u_1, u_2] \le \#S_a \le \#S_d$. Let $B \subset S_d$ be the lex set with $\#B = \#[x_1^{d+1}, u_2]$. Then $[x_1^{d+1}, u_2] = x_1 B$. Since $\#B + \#[u_1, x_n^d] = \#[u_1, u_2] \le \#S_d$, $B \cap [u_1, x_n^d] = \varnothing$. Then, by Case 1,

$$m([u_1, u_2]) = m(B) + m\big([u_1, x_n^d]\big) \preceq m(A)$$

(see figure).



$[u_1, u_2]$ $\quad$ $B \uplus [u_1, x_n^d]$ $\quad$ A $\quad$ L

Next, we prove the statement for R. In the same way as in the proof for L, we may assume $\#[u_1, u_2] \le \#S_b$. Let $d = \deg u_2$.

If $\deg u_1 = d$ then $[u_1, u_2] \subset S_d$ and $A = x_1^{b-d}[u_1, u_2]$ is an interval in S_b. Then, by Case 1, we have $m([u_1, u_2]) = m(A) \succeq m(R)$ as desired. Suppose $\deg u_1 < d$. Then

$$[u_1, u_2] = [u_1, x_n^{d-1}] \uplus [x_1^d, u_2].$$

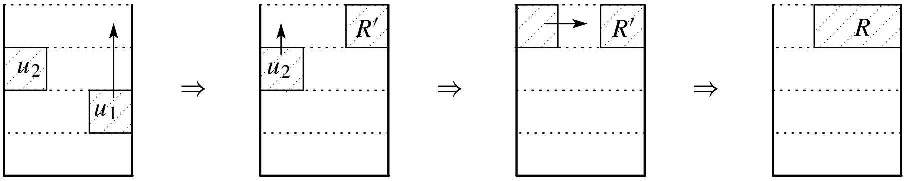
Let R' be the upper revlex set of degree b in S with $\#R' = \#[u_1, x_n^{d-1}]$. Then,

$$m([u_1, u_2]) \succeq m(R') + m\big([x_1^d, u_2]\big) = m(R') + m\big([x_1^b, x_1^{b-d} u_2]\big),$$

where the first inequality follows from the induction hypothesis on the cardinality. Since $R \setminus R' \subset S_b$ is an interval and $[x_1^b, x_1^{b-d} u_2] \subset S_b$ is lex, by Case 1 we have

$$m(R') + m\big([x_1^b, x_1^{b-d} u_2]\big) \succeq m(R') + m(R \setminus R') = m(R),$$

as desired (see figure). □



$[u_1, u_2]$ $R' \uplus [x_1^d, u_2]$ $R' \uplus [x_1^b, x_1^{b-d} u_2]$ R

Recall that a set $M \subset S$ of monomials is said to be super-revlex if it is revlex and $u \in M$ implies $v \in M$ for any monomial $v \in S$ of degree $\le \deg u - 1$.

Corollary 4.5. *Let $R \subset S$ be an upper revlex set of degree d and $M \subset S$ a super-revlex set such that $\#R + \#M \le \#S_{\le d}$. Let $Q \subset S$ be the super-revlex set with $\#Q = \#R + \#M$. Then*

$$m(Q) \succeq m(R) + m(M).$$

Proof. Let $e = \min\{k : x_1^k \notin M\}$ and $F = \{u \in S_e : u \notin M\}$. If $\#F \ge \#R$ then

$$Q = M \uplus (Q \setminus M)$$

and $Q \setminus M \subset F$ is an interval. Thus $m(Q \setminus M) \succeq m(R)$ by the interval lemma.

Suppose $\#F < \#R$. Write

$$R = I \uplus R'$$

such that I is an interval with $\#I = \#F$ and R' is an upper revlex set of degree d. Since F is a lex set, the interval lemma shows

$$m(M) + m(R) = m(M) + m(I) + m(R') \preceq m(F \uplus M) + m(R').$$

Then $F \uplus M$ is a super-revlex set containing x_1^e. By repeating this procedure, we have $m(M) + m(R) \preceq m(Q)$. □

The above corollary proves the next result, which was essentially proved by Elías, Robbiano and Valla [Elías et al. 1991].

Corollary 4.6. *Let $R \subset S$ be a finite revlex set of monomials and $M \subset S$ the super-revlex set with $\#M = \#R$. Then $m(M) \succeq m(R)$.*

Proof. Let $R = \biguplus_{i=0}^{N} R_i$, where R_i is the set of monomials in R of degree i and $N = \max\{i : R_i \neq \varnothing\}$. Let $M_{(\leq j)}$ be the super-revlex set with $\# M_{(\leq j)} = \# \biguplus_{i=0}^{j} R_i$. We claim $m(M_{(\leq j)}) \succeq m\big(\biguplus_{i=0}^{j} R_i\big)$ for all j. This follows inductively from Corollary 4.5 as follows:

$$m\left(\biguplus_{i=0}^{j} R_i\right) = m\left(\biguplus_{i=0}^{j-1} R_i\right) + m(R_j) \preceq m(M_{(\leq j-1)}) + m(R_j) \preceq m(M_{(\leq j)}).$$

(We use the induction hypothesis for the second step and use Corollary 4.5 for the last step.) Then we have $m(M) = m(M_{(\leq N)}) \succeq m\big(\biguplus_{i=0}^{N} R_i\big)$. □

We finish this section by proving the result of Valla, which we mentioned in the introduction.

Corollary 4.7 [Valla 1994]. *Let c be a positive integer and $M \subset S$ the super-revlex set with $\# M = c$. Let $J \subset S$ be the monomial ideal generated by all monomials which are not in M. Then, for any homogeneous ideal $I \subset S$ with* $\dim_K(S/I) = c$, *we have $\beta_i^S(S/J) \geq \beta_i^S(S/I)$ for all i.*

Proof. The proof is similar to that of Corollary 3.6. By the Bigatti–Hulett–Pardue theorem, we may assume that I is lex. Then Lemma 3.5 says, for $d \gg 0$, we have

$$\beta_i^S(I) = \binom{n-1}{i} \dim_K I_{\leq d} - \sum_{k=1}^{n} \binom{k-1}{i} m_{\leq k}(I_{\leq d-1}) - \sum_{k=1}^{n-1} \binom{k-1}{i-1} m_{\leq k}(I_{\leq d})$$

and the same formula holds for J. Let $N \subset S$ be the set of monomials which are not in I. Since N is a revlex set with $\# N = c$, for $d \gg 0$, by Corollary 4.6 we have

$$m(J_{\leq d}) = m(S_{\leq d}) - m(M) \preceq m(S_{\leq d}) - m(N) = m(I_{\leq d}).$$

Hence $\beta_i^S(J) \geq \beta_i^S(I)$ for all i as desired. □

The proof given in this section provides a new short proof of the above result. The most difficult part in the proof is Corollary 4.6. The original proof given in [Elías et al. 1991] is based on computations of binomial coefficients. On the other hand, our proof is based on moves of interval sets of monomials.

5. Construction

In this section, we give a construction of sets of monomials which satisfy the conditions of Proposition 3.9, and study their properties.

Throughout Sections 5 and 6, we fix the following notation: Let $a_1, a_2, \ldots, a_t$ be nonnegative integers, where $t \leq n$, and let $b_i = a_1 + \cdots + a_i + 1$ for $i = 1, 2, \ldots, t$.

Let $F = S\boldsymbol{e}_1 \oplus S\boldsymbol{e}_2 \oplus \cdots \oplus S\boldsymbol{e}_t$ be a free S-module with $\deg \boldsymbol{e}_i = b_i$ for $i = 1, 2, \ldots, t$. We consider the set

$$U = S^{(1)}\boldsymbol{e}_1 \uplus S^{(2)}\boldsymbol{e}_2 \uplus \cdots \uplus S^{(t)}\boldsymbol{e}_t \subset F.$$

Note that we identify each $S^{(k)}$ with the set of monomials in it. For $i = 1, 2, \ldots, t$, let $\delta_i = x_1^{a_1} \cdots x_{i-1}^{a_{i-1}} x_i^{a_i+1}$. Then, by the decomposition given before Definition 3.7, the above set U can be identified with the set of monomials in the universal lex ideal $(\delta_1, \ldots, \delta_t) = \delta_1 S^{(1)} \oplus \cdots \oplus \delta_t S^{(t)}$ via the natural correspondence $u\boldsymbol{e}_i \leftrightarrow \delta_i u$.

We call an element $u\boldsymbol{e}_i \in U$ a monomial in U. For each monomial $u\boldsymbol{e}_i \in U$, we define

$$\max(u\boldsymbol{e}_i) = \begin{cases} i & \text{if } u = 1, \\ \max(u) & \text{otherwise.} \end{cases}$$

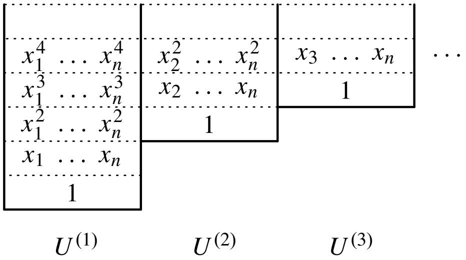
Also, for $M \subset U$, we define $m(M) = (m_{\leq 1}(M), m_{\leq 2}(M), \ldots, m_{\leq n}(M))$ in the same way as in Section 3. We say that a subset $M = M_{\langle 1 \rangle}\boldsymbol{e}_1 \uplus \cdots \uplus M_{\langle t \rangle}\boldsymbol{e}_t \subset U$ is a *ladder set* if $M_{\langle 1 \rangle}, \ldots, M_{\langle t \rangle}$ satisfy the conditions (i) and (ii) of Lemma 3.8. Then, considering $m(-)$ of ladder sets in $U = S^{(1)}\boldsymbol{e}_1 \uplus \cdots \uplus S^{(t)}\boldsymbol{e}_t$ is equivalent to considering $m(-)$ of ladder sets in the universal lex ideal $(\delta_1, \ldots, \delta_t) = \delta_1 S^{(1)} \oplus \cdots \oplus \delta_t S^{(t)}$. In particular, to prove Proposition 3.9, it is enough to consider ladder sets in U.

Let $M \subset U$. We write

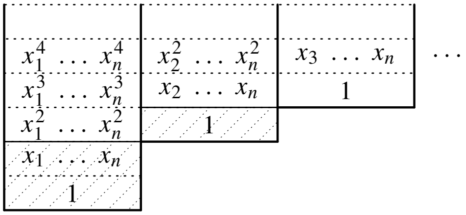
$$U^{(i)} = S^{(i)}\boldsymbol{e}_i,\ M^{(i)} = M \cap U^{(i)},\ U^{(\geq i)} = \biguplus_{k=i}^{t} S^{(k)}\boldsymbol{e}_k, \text{ and } M^{(\geq i)} = M \cap U^{(\geq i)}.$$

Note that $U^{(\geq i)} = \biguplus_{k \geq i} S^{(k)}\boldsymbol{e}_k$ can be identified with the universal lex ideal in $K[x_i, \ldots, x_n]$ generated by $\{(x_i^{b_{i-1}})x_i^{a_i} \cdots x_{k-1}^{a_{k-1}} x_k^{a_k+1} : k = i, i+1, \ldots, t\}$. For a subset $M \subset U$, we write M_k for the set of monomials in M of degree k and $M_{\leq j} = \biguplus_{k=0}^{j} M_k$.

As in Section 4, we use pictures to help to understand the proofs. We identify U with the following picture:



Note that each low represents the set of monomials in U having the same degree. Thus, in the previous figure, $\deg \boldsymbol{e}_2 = \deg \boldsymbol{e}_1 + 2$ and $\deg \boldsymbol{e}_3 = \deg \boldsymbol{e}_2 + 1$. Also, we present a subset $M \subset U$ by a shaded picture. For example, the following figure represents $M = \{1, x_1, x_2, \dots, x_n\}\boldsymbol{e}_1 \uplus \{1\}\boldsymbol{e}_2$:



M

Also, we define the map $\rho : U \to U$ by extending the map given in Section 4 as follows: For $x_i^k u\boldsymbol{e}_i \in U^{(i)}$ with $u \in K[x_{i+1}, \dots, x_n]$, let

$$\rho(x_i^k u\boldsymbol{e}_i) = \begin{cases} u\boldsymbol{e}_{i+1} & \text{if } i \le t-1, \\ 0 & \text{if } i = t. \end{cases}$$

We call the above map $\rho : U \to U$ the *moving map* of U. The moving map induces a bijection from $U_j^{(i)} = \{u\boldsymbol{e}_i \in U^{(i)} : \deg u = j - b_i\}$ to $U_{\le j+a_{i+1}}^{(i+1)} = \{u\boldsymbol{e}_{i+1} \in U^{(i+1)} : \deg u \le j - b_i\}$ for $i = 1, 2, \dots, t-1$.

Lemma 5.1. *For $N \subset U_j^{(i)}$ with $i \le t-1$, one has $m(N) \succeq m(\rho(N))$. Moreover, if $x_i^{j-b_i}\boldsymbol{e}_i \notin N$ then $m(N) = m(\rho(N))$.*

Next, we define ladder sets $M \subset U$ which attain maximal Betti numbers. Recall that a subset $M \subset U$ is called a ladder set if the following conditions hold:

(i) $\{u \in S^{(i)} : u\boldsymbol{e}_i \in M^{(i)}\}$ is a revlex multicomplex for $i = 1, 2, \dots, t$.

(ii) If $M_j^{(i)} \ne \varnothing$ then $M_j^{(i+1)} = U_j^{(i+1)}$ for $i = 1, 2, \dots, t-1$ and for all $j \ge 0$.

To simplify the notation, we say that $N \subset U^{(i)}$ is a super-revlex set (resp. interval, lower lex set or upper revlex set of degree d) if $N' = \{u \in S^{(i)} : u\boldsymbol{e}_i \in N\}$ is super-revlex (resp. interval, lower lex set or upper revlex set of degree $d - b_i$) in $S^{(i)}$. For monomials $u\boldsymbol{e}_i, v\boldsymbol{e}_i \in U$ and for a monomial order $>$ on $S^{(i)}$, we write $u\boldsymbol{e}_i > v\boldsymbol{e}_i$ if $u > v$.

Definition 5.2. A monomial $f = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}\boldsymbol{e}_1 \in U_e^{(1)}$ is said to be *admissible over U* if the following conditions hold:

(i) $\deg \rho^i(f) \le e+1$ or $\rho^i(f) = \boldsymbol{e}_{i+1}$ for $i = 1, 2, \dots, t-2$.

(ii) $\rho^{t-1}(f) = \boldsymbol{e}_t$ or $\rho^{t-1}(f) \ge_{\text{opdlex}} x_t^{e+1-b_t}\boldsymbol{e}_t$.

Note that the second condition in (ii) cannot be satisfied when $e+1-b_t < 0$ and that if $t = 1$ then all monomials in U are admissible. Also, $\rho^{t-1}(f) \ge_{\text{opdlex}} x_t^{e+1-b_t}\boldsymbol{e}_t$ if and only if $\deg \rho^{t-1}(f) \le e$ or $\rho^{t-1}(f) = x_t^{e+1-b_t}\boldsymbol{e}_t$.

We say that $f \in U_e^{(i)}$ is admissible if it is admissible over $U^{(\ge i)}$. Note that $x_i^k\boldsymbol{e}_i \in U^{(i)}$ is admissible for all i and k.

Definition 5.3. Let $>_{\mathrm{dlex}}$ be the degree lex order. Thus for monomials $u, v \in S$, $u >_{\mathrm{dlex}} v$ if $\deg u > \deg v$ or $\deg u = \deg v$ and $u >_{\mathrm{lex}} v$. We extend $>_{\mathrm{dlex}}$ to monomials in U by $u\boldsymbol{e}_i >_{\mathrm{dlex}} v\boldsymbol{e}_j$ if $\delta_i u >_{\mathrm{dlex}} \delta_j v$. Thus, we have $u\boldsymbol{e}_i >_{\mathrm{dlex}} v\boldsymbol{e}_j$ if

(i) $\deg u\boldsymbol{e}_i > \deg v\boldsymbol{e}_j$,

(ii) $\deg u\boldsymbol{e}_i = \deg v\boldsymbol{e}_j$ and $i < j$, or

(iii) $\deg u\boldsymbol{e}_i = \deg v\boldsymbol{e}_j$, $i = j$ and $u >_{\mathrm{dlex}} v$.

Fix an integer $c > 0$. Let

$$f = \max_{>_{\mathrm{dlex}}} \left\{ g \in U^{(1)} : g \text{ is admissible and } \#\{h \in U : h \leq_{\mathrm{dlex}} g\} \leq c \right\}$$

and

$$L_{(c)} = \{h \in U^{(1)} : h \leq_{\mathrm{dlex}} f\}.$$

Let $M = M^{(1)} \uplus \cdots \uplus M^{(t)} \subset U$ be a set of monomials with $\#M = c$. We say that M satisfies the *maximal condition* if $M^{(1)} = L_{(c)}$. Also, we say that M is *extremal* if $M^{(\geq k)} \subset U^{(\geq k)}$ satisfies the maximal condition in $U^{(\geq k)}$ for all k.

Example 5.4. If $t = 1$ then any monomial in $U = S^{(1)}\boldsymbol{e}_1$ is admissible and extremal sets can be identified with super-revlex sets in $S^{(1)}$.

Example 5.5. Suppose $t = 2$. Then $f = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} \boldsymbol{e}_1$, where $f \neq x_1^{\alpha_1}\boldsymbol{e}_1$, is admissible in $U = S^{(1)}\boldsymbol{e}_1 \uplus S^{(2)}\boldsymbol{e}_2$ if $\alpha_1 \geq a_2$ or $f = x_1^{a_2-1} x_2^{\alpha_2}\boldsymbol{e}_1$. In other words, a monomial $f \in S_d^{(1)}\boldsymbol{e}_1$ is admissible if and only if $f \geq_{\mathrm{lex}} x_1^{a_2-1} x_2^{d-a_2+1}\boldsymbol{e}_1$ if $a_2 \leq d$ and $f = x_1^d \boldsymbol{e}_1$ if $a_2 > d$. For example, if $\deg \boldsymbol{e}_1 = 2$ and $\deg \boldsymbol{e}_2 = 4$ then the admissible monomials in $U_5^{(1)} = (S_3^{(1)})\boldsymbol{e}_1$ are

$$x_1^3\boldsymbol{e}_1,\ x_1^2x_2\boldsymbol{e}_1,\ x_1^2x_3\boldsymbol{e}_1,\ \ldots,\ x_1^2x_n\boldsymbol{e}_1,\ x_1x_2^2\boldsymbol{e}_1.$$

Example 5.6. Suppose $t = 3$. The situation is more complicated. A monomial $f = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}\boldsymbol{e}_1 \in U_e^{(1)}$, where $f \neq x_1^{\alpha_1}\boldsymbol{e}_1$ is admissible in U if and only if

- $\alpha_1 \geq a_2 - 1$ and
- $x_3^{\alpha_3} \cdots x_n^{\alpha_n} \geq_{\mathrm{opdlex}} x_3^{e+1-b_3}$ or $x_3^{\alpha_3} \cdots x_n^{\alpha_n} = 1$.

For example, if $\deg \boldsymbol{e}_1 = 2$, $\deg \boldsymbol{e}_2 = 4$, $\deg \boldsymbol{e}_3 = 6$, and $n = 3$ then the set of the admissible monomials in $U_6^{(1)} = (K[x_1, x_2, x_3]_4)\boldsymbol{e}_1$ are

$$\{x_1^4\boldsymbol{e}_1\} \cup \{x_1^3x_2\boldsymbol{e}_1, x_1^3x_3\boldsymbol{e}_1\} \cup \{x_1^2x_2^2\boldsymbol{e}_1, x_1^2x_2x_3\boldsymbol{e}_1\} \cup \{x_1x_2^3\boldsymbol{e}_1, x_1x_2^2x_3\boldsymbol{e}_1\}.$$

Example 5.7. Let $U = x_1^2 S^{(1)} \uplus x_1x_2^3 S^{(2)}$. Suppose $c = \binom{n+2}{2} + 2$. Then

$$\max_{>_{\mathrm{dlex}}} \left\{ f \in U^{(1)} : f \text{ is admissible and } \#\{h \in U : h \leq_{\mathrm{dlex}} f\} \leq c \right\} = x_1^2\boldsymbol{e}_1.$$

Indeed,

$$\#\{h \in U : h \leq_{\mathrm{dlex}} x_1^2\boldsymbol{e}_1\} = \#S_{\leq 2}^{(1)}\boldsymbol{e}_1 \uplus \{1\}\boldsymbol{e}_2 = \binom{n+2}{2} + 1$$

and

$$\#\{h \in U : h \leq_{\text{dlex}} x_1x_2^2\boldsymbol{e}_1\} = \#\big(S^{(1)}_{\leq 3} \setminus \{x_1^3, x_1^2x_2, \dots, x_1^2x_n\}\big)\boldsymbol{e}_1 \uplus S^{(2)}_{\leq 1}\boldsymbol{e}_2$$
$$= \binom{n+3}{3} > c.$$

By Example 5.5, the lex-smallest admissible monomial in $U_5^{(1)}$ is $x_1x_2^2\boldsymbol{e}_1$. Thus the extremal set $L \subset U$ with $\#L = c$ is

$$L = S^{(1)}_{\leq 2}\boldsymbol{e}_1 \uplus \{1, x_n\}\boldsymbol{e}_2.$$

Example 5.8. In general, it is not easy to understand the shape of extremal sets, but in some special cases they are simple.

If $b_1 = b_2 = \cdots = b_t$ then any monomial in U is admissible. Thus any extremal set M in U is of the form

$$M = \{h \in U : h \leq_{\text{dlex}} f\}$$

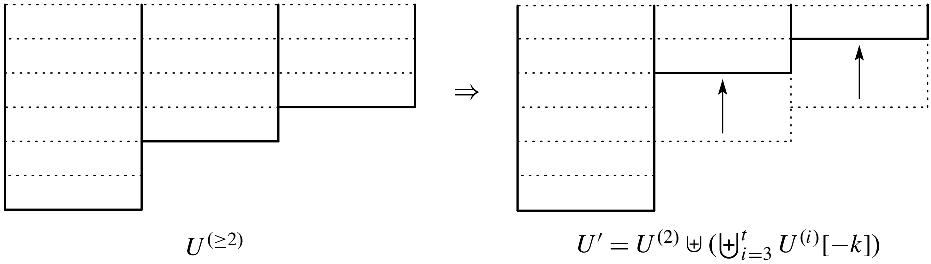
for some $f \in U$.

If $b_2 > e$ then the only admissible monomial in $U_e^{(1)}$ is $x_1^{e-b_1}\boldsymbol{e}_1$. Thus if $b_1 \ll b_2 \ll \cdots \ll b_n$ (for example, if $b_{i+1} - b_i > c$ for all i) then any extremal set M in U with $\#M = c$ is of the form

$$M = S^{(1)}_{\leq d_1}\boldsymbol{e}_1 \uplus S^{(2)}_{\leq d_2}\boldsymbol{e}_2 \uplus \cdots \uplus S^{(t-1)}_{\leq d_{t-1}}\boldsymbol{e}_{t-1} \uplus N,$$

where $N \subset S^{(t)}\boldsymbol{e}_t$ and $\#S^{(i+1)}_{\leq d_{i+1}}\boldsymbol{e}_{i+1} \uplus \cdots \uplus S^{(t-1)}_{\leq d_{t-1}}\boldsymbol{e}_{t-1} \uplus N < \#S^{(i)}_{d_i+1}$ for $i = 1, \dots, t-1$.

In the rest of this section, we study properties of extremal sets. Suppose $t \geq 3$. For an integer $k \geq -a_3$, we write $U^{(i)}[-k] = S^{(i)}\boldsymbol{e}_i'$, where $\boldsymbol{e}_i'$ is a basis element with $\deg \boldsymbol{e}_i = b_i + k$. In the picture, $U^{(i)}[-k]$ is the picture obtained from that of $U^{(i)}$ by moving the blocks k steps above. In particular, for any integer $k \geq -a_3$, $U' = U^{(2)} \uplus \biguplus_{i=3}^t U^{(i)}[-k]$ can be identified with a universal lex ideal in $K[x_2, \dots, x_n]$:



Lemma 5.9. *Suppose $t \geq 3$. Let $f \in U_e^{(1)}$, $d = \deg \rho(f)$, and $k \geq -a_3$ with $e - d + k \geq 0$. Then f is admissible over U if and only if the following conditions hold*:

- $\deg \rho(f) \le e+1$ *or* $\rho(f) = \boldsymbol{e}_2$.
- $x_2^{e-d+k}\rho(f) \in U^{(2)}_{e+k}$ *is admissible in* $U' = U^{(2)} \uplus \biguplus_{i=3}^{t} U^{(i)}[-k]$.

Proof. Let $U' = S^{(2)}\boldsymbol{e}_2 \uplus S^{(3)}\boldsymbol{e}'_3 \uplus \cdots \uplus S^{(t)}\boldsymbol{e}'_t$ with $\deg \boldsymbol{e}'_i = \deg \boldsymbol{e}_i + k$ for $k = 3, \ldots, t$, and let ϕ be the moving map of U'. Let $\rho^i(f) = u_{i+1}\boldsymbol{e}_{i+1}$ for $i = 2, \ldots, t-1$. Then $\phi^i(x_2^{e-d+k}\rho(f)) = u_{i+2}\boldsymbol{e}'_{i+2}$ for $i = 1, 2, \ldots, t-2$. Thus $\deg \rho^i(f) \le e+1$ if and only if $\deg \phi^{i-1}(x_2^{e-d+k}\rho(f)) \le e+1+k$ for $i \ge 2$. Also, $\rho^{t-1}(f) \ge_{\mathrm{opdlex}} x_t^{e+1-b_t}\boldsymbol{e}_t$ if and only if $\phi^{t-2}(x_2^{e+d+k}\rho(f)) \ge_{\mathrm{opdlex}} x_t^{e+1-b_t}\boldsymbol{e}'_t$. Since $\deg x_2^{e-d+k}\rho(f) = e+k$, the above facts prove the statement. □

By the definition of the maximal condition, the next result is straightforward:

Lemma 5.10. *Let $M \subset U$ be an extremal set.*

(i) *If* $\#M \ge \#U_{\le e}$ *then* $M \supset U_{\le e}$.

(ii) *If* $\#M \ge \#U^{(1)}_{\le e-1} \uplus U^{(\ge 2)}_{\le e}$ *then* $M \supset U^{(1)}_{\le e-1} \uplus U^{(\ge 2)}_{\le e}$.

Proof. Since M is extremal, there exists an $f \in U^{(1)}$ such that

$$M^{(1)} = \{h \in U^{(1)} : h \le_{\mathrm{dlex}} f\}.$$

(i) Since $x_1^{e-b_1}\boldsymbol{e}_1$ is admissible and $\{h \in U : h \le_{\mathrm{dlex}} x_1^{e-b_1}\boldsymbol{e}_1\} = U_{\le e}$, $f \ge_{\mathrm{dlex}} x_1^{e-b_1}\boldsymbol{e}_1$. Then $M^{(1)} \supset \{h \in U^{(1)} : h \le_{\mathrm{dlex}} x_1^{e-b_1}\boldsymbol{e}_1\} = U^{(1)}_{\le e}$. Also, since

$$\#M^{(\ge 2)} = \#M - \#M^{(1)} \ge \#\{h \in U : h \le_{\mathrm{dlex}} f\} - \#\{h \in U^{(1)} : h \le_{\mathrm{dlex}} f\} \ge \#U^{(2)}_{\le e},$$

we have $M^{(\ge 2)} \supset U^{(2)}_{\le e}$ by induction on t.

(ii) It is clear that $M \supset U_{\le e-1}$ by (i). If $\deg f \ge e$ then

$$\#M \ge \#\{h \in U : h \le_{\mathrm{dlex}} f\} \ge \#M^{(1)} \uplus U^{(\ge 2)}_{\le e}.$$

Then $\#M^{(\ge 2)} \ge \#U^{(\ge 2)}_{\le e}$ and $M^{(\ge 2)} \supset U^{(\ge 2)}_{\le e}$ by (i) as desired. If $\deg f < e$ then $M^{(1)} = U^{(1)}_{\le e-1}$ and $\#M^{(\ge 2)} \ge \#U^{(\ge 2)}_{\le e}$ by the assumption. Hence $M^{(\ge 2)} \supset U^{(\ge 2)}_{\le e}$ by (i). □

Corollary 5.11. *Extremal sets are ladder sets.*

Proof. If $M \subset U$ is extremal then $M^{(i)}$ is super-revlex for all i by the maximal condition. It is enough to prove that if $M^{(1)}_e \ne \varnothing$ then $M \supset U^{(\ge 2)}_e$. If $M^{(1)}_e \ne \varnothing$ then there exists an admissible monomial $f \in U^{(1)}_e$ such that

$$\#M \ge \#\{h \in U : h \le_{\mathrm{dlex}} f\} \ge \#U^{(1)}_{\le e-1} \uplus U^{(\ge 2)}_{\le e}.$$

Then the statement follows from Lemma 5.10. □

To simplify notation, for $u\boldsymbol{e}_i, v\boldsymbol{e}_i \in U^{(i)}$ with $u \ge_{\mathrm{opdlex}} v$, we write

$$[u\boldsymbol{e}_i, v\boldsymbol{e}_i] = \{w\boldsymbol{e}_i \in U^{(i)} : u \ge_{\mathrm{opdlex}} w \ge_{\mathrm{opdlex}} v\}.$$

Lemma 5.12. *Suppose $t \geq 2$. Let $M \subset U$ be an extremal set.*

(i) *If $a_2 > 0$ then $M_e^{(1)} \neq 0$ if and only if $\#M \geq \#U_{\leq e}^{(1)}$.*

(ii) *If $a_2 = 0$ and $M_e^{(1)} \neq 0$ then $\#M > \#U_{\leq e}^{(1)}$.*

Proof. Let $f \in U_e^{(1)}$ be the lex-smallest admissible monomial in $U_e^{(1)}$ over U.

(i) It suffices to prove that

$$\#\{h \in U : h \leq_{\text{dlex}} f\} = \#U_{\leq e}^{(1)}. \tag{2}$$

If $f = x_1^{e-b_1}\boldsymbol{e}_1$ then $f' = x_1^{e-b_1-1}x_2\boldsymbol{e}_1$ is not admissible. By the definition of admissibility, one has $\deg \rho(f') = \deg x_2\boldsymbol{e}_2 > e+1$ and $b_2 > e$. In this case we have $\{h \in U : h \leq_{\text{dlex}} f\} = U_{\leq e}^{(1)}$.

Suppose $f \neq x_1^{e-b_1}\boldsymbol{e}_1$. We prove (2) by using induction on t. Suppose $t = 2$. Then $f = x_1^{a_2-1}x_2^{e+1-b_2}\boldsymbol{e}_1$, and

$$\{h \in U : h \leq_{\text{dlex}} f\} = U_{\leq e-1}^{(1)} \uplus [f, x_n^{e-b_1}\boldsymbol{e}_1] \uplus U_{\leq e}^{(2)}.$$

Since $\rho([f, x_n^{e-b_1}\boldsymbol{e}_1]) = \biguplus_{j=e+1}^{e+a_2} U_j^{(2)}$, we have

$$\#\{h \in U : h \leq_{\text{dlex}} f\} = \#U_{\leq e-1}^{(1)} + \#U_{\leq e+a_2}^{(2)} = \#U_{\leq e}^{(1)},$$

where we use $\rho(U_e^{(1)}) = U_{\leq e+a_2}^{(2)}$ for the last equality.

Suppose $t \geq 3$. Since $\rho(f) \neq \boldsymbol{e}_2$, we have $\deg \rho(f) = e+1$. Indeed, by Lemma 5.9, $\deg \rho(f) \leq e+1$. On the other hand, since $x_1^{a_2-1}x_2^{e+1-b_2}\boldsymbol{e}_1$ is admissible over U, $f \leq_{\text{lex}} x_1^{a_2-1}x_2^{e+1-b_2}\boldsymbol{e}_1$. Thus $\deg \rho(f) \geq \deg \rho(x_1^{a_2-1}x_2^{e+1-b_2}\boldsymbol{e}_1) = e+1$.

Consider $U' = U^{(2)} \uplus \biguplus_{i=3}^{t} U^{(i)}[-1]$. By Lemma 5.9 (consider the case when $d = e+1$ and $k = 1$), $\rho(f)$ is the lex-smallest admissible monomial in $U_{e+1}^{(2)}$ over U'. Then

$$\begin{aligned}\#[\rho(f), x_n^{e+1-b_2}\boldsymbol{e}_2] \uplus U_{\leq e}^{(\geq 2)} &= \#[\rho(f), x_n^{e+1-b_2}\boldsymbol{e}_2] \uplus U_{\leq e}^{(2)} \uplus U'^{(\geq 3)}_{\leq e+1} \\ &= \#\{h \in U' : h \leq_{\text{dlex}} \rho(f)\} \\ &= \#U_{\leq e+1}^{(2)},\end{aligned} \tag{3}$$

where the last equation follows from the induction hypothesis. On the other hand

$$\{h \in U : h \leq_{\text{dlex}} f\} = [f, x_n^{e-b_1}\boldsymbol{e}_1] \uplus U_{\leq e-1}^{(1)} \uplus U_{\leq e}^{(\geq 2)} \tag{4}$$

and

$$\rho\big([f, x_n^{e-b_1}\boldsymbol{e}_1]\big) = [\rho(f), x_n^{e+1-b_2}\boldsymbol{e}_2] \uplus \biguplus_{j=e+2}^{e+a_2} U_j^{(2)}. \tag{5}$$

Equations (3), (4), and (5) show that

$$\#\{h \in U : h \leq_{\text{dlex}} f\} = \#U_{\leq e-1}^{(1)} \uplus U_{\leq e+a_2}^{(2)} = \#U_{\leq e-1}^{(1)} \uplus U_e^{(1)} = \#U_{\leq e}^{(1)},$$

where the second equality follows since $\rho(U_e^{(1)}) = U_{\leq e+a_2}^{(2)}$.

(ii) It suffices to prove that $\#\{h \in U : h \leq_{\mathrm{dlex}} f\} > \#U^{(1)}_{\leq e}$. Since $a_2 = 0$, $\#U^{(2)}_{\leq e} = \#U^{(1)}_{e}$. Then we have

$$\#\{h \in U : h \leq_{\mathrm{dlex}} f\} > \#U^{(1)}_{\leq e-1} \uplus U^{(2)}_{\leq e} = \#U^{(1)}_{\leq e-1} \uplus U^{(1)}_{e} = U^{(1)}_{\leq e},$$

as desired. □

Corollary 5.13. *Suppose $t \geq 2$. Let $B \subset U^{(1)}_e$ be the revlex set and $N \subset U^{(\geq 2)}$ a ladder set with $\#N \geq \#U^{(\geq 2)}_{\leq e-1}$. Let $Y \subset U$ be the extremal set with*

$$\#Y = \#U^{(1)}_{\leq e-1} \uplus B \uplus N.$$

If $\#B \uplus N < \#U^{(1)}_e$ then

$$Y = U^{(1)}_{\leq e-1} \uplus Y^{(\geq 2)}.$$

Proof. Since $\#Y \geq \#U_{\leq e-1}$, we have $Y \supset U_{\leq e-1}$ by Lemma 5.10. On the other hand, since $\#Y = \#U^{(1)}_{\leq e-1} \uplus B \uplus N < \#U^{(1)}_{\leq e}$ by the assumption, we have $Y^{(1)}_e = \varnothing$ by Lemma 5.12. Hence $Y^{(1)} = U^{(1)}_{\leq e-1}$. □

For monomials $f >_{\mathrm{dlex}} g \in U^{(i)}$, let $[f, g) = [f, g] \setminus \{g\}$.

Lemma 5.14. *Let $f \in U^{(1)}_e$ be the lex-smallest admissible monomial in $U^{(1)}_e$ over U and $g >_{\mathrm{lex}} h \in U^{(1)}_e$ admissible monomials over U such that there are no admissible monomials in $[g, h]$ except for g and h. Then $\#[g, h) \leq \#[f, x_n^{e-b_1}\boldsymbol{e}_1]$.*

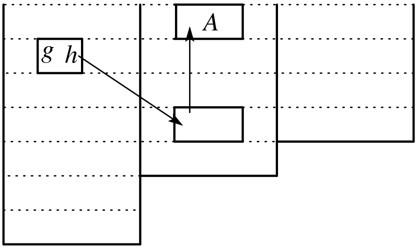
Proof. If $t = 1$ then all monomials are admissible over U. If $t = 2$ then any monomial $w \in U^{(1)}_e$ with $w >_{\mathrm{lex}} f$ is admissible over U. Thus the statement is clear if $t \leq 2$.

Suppose $t \geq 3$. Since $g \neq h$ we have $f \neq x_1^{e-b_1}\boldsymbol{e}_1$. By the definition of admissibility, we have $\deg(\rho(f)) = e$ if $a_2 = 0$ and $\deg(\rho(f)) = e + 1$ if $a_2 > 0$. We consider the case when $a_2 > 0$ (the proof for the case when $a_2 = 0$ is similar).

Consider $U' = U^{(2)} \uplus \biguplus_{i=3}^{t} U^{(i)}[-1]$. Since any monomial $w \in U^{(1)}_e$ such that $\rho(w) = x_2^k\boldsymbol{e}_2$ with $k \leq e + 1 - b_2$ is admissible over U, we have $\rho([g, h)) \subset S_d$ for some $d \leq e + 1$. Let

$$A = x_2^{e+1-d}\rho([g, h)) = \big[x_2^{e+1-d}\rho(g), x_2^{e+1-d}\rho(h)\big) \subset U^{(2)}_{e+1}$$

(see figure).



Let $w \in A$. Then $w = x_2^{e+1-d}\rho(w')$ for some $w' \in [g,h)$. Lemma 5.9 says that w is admissible over U' if and only if w' is admissible over U. Hence A contains no admissible monomial over U' except for $x_2^{e+1-d}\rho(g)$. By Lemma 5.9, $\rho(f) \in U_{e+1}^{(2)}$ is the lex-smallest admissible monomial in $U_{e+1}^{(2)}$ over U'. Then, by the induction hypothesis,

$$\#A \le \#[\rho(f), x_n^{e-b_2}\boldsymbol{e}_2] = \#\rho([f, x_n^{e-b_1}\boldsymbol{e}_1]) \cap U_{e+1}^{(2)} \le \#[f, x_n^{e-b_1}\boldsymbol{e}_1].$$

Then the statement follows since $\#[g,h) = \#\rho([g,h)) = \#A$. □

Lemma 5.15. *Let $M \subset U$ be an extremal set, $e = \min\{k : x_1^{k-b_1}\boldsymbol{e}_1 \notin M\}$, and $H = U_e \setminus M_e$. Let $f \in U_e^{(1)}$ be the lex-smallest admissible monomial in $U_e^{(1)}$ over U. Then*:

(i) $\#U_{\le e} + \#[f, x_n^{e-b_1}\boldsymbol{e}_1] \le \#U_{\le e+1}^{(1)}$.

(ii) $\#M + \#H < \#U_{\le e+1}^{(1)}$.

Proof. We use induction on t. If $t = 1$ then the statements are obvious. Suppose $t > 1$.

(i) If $a_2 > 0$ then by Lemma 5.12

$$\#U_{\le e} + \#[f, x_n^{e-b_1}\boldsymbol{e}_1] = \#\{h \in U : h \le_{\text{dlex}} f\} + \#U_e^{(1)} = \#U_{\le e}^{(1)} + \#U_e^{(1)} < \#U_{\le e+1}^{(1)}$$

as desired. Suppose $a_2 = 0$. Then

$$\rho([f, x_n^{e-b_1}\boldsymbol{e}_1]) = [\rho(f), x_n^{e-b_2}\boldsymbol{e}_1] \subset U_e^{(2)}$$

and $\rho(f)$ is the lex-smallest admissible monomial in $U_e^{(2)}$ over $U^{(\ge 2)}$ by Lemma 5.9. Then by the induction hypothesis,

$$\begin{aligned}\#U_{\le e} + \#[f, x_n^{e-b_1}\boldsymbol{e}_1] &= \#U_{\le e}^{(1)} + \left(\#U_{\le e}^{(\ge 2)} + \#[\rho(f), x_n^{e-b_2}\boldsymbol{e}_2]\right)\\ &\le \#U_{\le e}^{(1)} + \#U_{\le e+1}^{(2)}\\ &= \#U_{\le e+1}^{(1)}\end{aligned}$$

as desired.

(ii) Suppose $M_e^{(2)} \neq U_e^{(2)}$. Then $M_e^{(1)} = \varnothing$. Since $M^{(\ge 2)}$ is extremal over $U^{(\ge 2)}$, by the induction hypothesis,

$$\#M + \#H = \#U_{\le e-1}^{(1)} \uplus M^{(\ge 2)} + \#U_e^{(1)} \uplus H^{(\ge 2)} < \#U_{\le e}^{(1)} + \#U_{\le e+1}^{(2)} \le \#U_{\le e+1}^{(1)},$$

where we use $\#U_{e+1}^{(1)} = \#U_{\le e+1+a_2}^{(2)} \ge \#U_{\le e+1}^{(2)}$ for the last inequality.

Suppose $M_e^{(2)} = U_e^{(2)}$. Let $g = \max_{>_{\text{dlex}}} M^{(1)}$ and let

$$\mu = \min_{>_{\text{dlex}}}\{h \in U_{\le e}^{(1)} : h \text{ is admissible over } U \text{ and } h >_{\text{dlex}} g\}.$$

Then $[\mu, g) \subset U_e^{(1)}$ since $g \geq_{\text{dlex}} x_1^{e-b_1-1}\boldsymbol{e}_1$. Since M is extremal,

$$\#M < \#\{h \in U : h \leq_{\text{dlex}} \mu\}.$$

Since $M^{(1)} = \{h \in U^{(1)} : h \leq_{\text{dlex}} g\}$, $H = [x_1^{e-b_1}\boldsymbol{e}_1, g)$. Thus

$$\begin{aligned}\#M + \#H &< \#\{h \in U : h \leq_{\text{dlex}} \mu\} + \#[x_1^{e-b_1}\boldsymbol{e}_1, g)\\ &= \#U_{\leq e} + \#[\mu, g)\\ &\leq \#U_{\leq e} + \#[f, x_n^{e-b_1}\boldsymbol{e}_1],\end{aligned}$$

where the last inequality follows from Lemma 5.14. Then the desired inequality follows from (i). □

6. Proof of the main theorem

Let $U = S^{(1)}\boldsymbol{e}_1 \uplus S^{(2)}\boldsymbol{e}_2 \uplus \cdots \uplus S^{(t)}\boldsymbol{e}_t$ be as in Section 5. The aim of this section is to prove the next result, which proves Proposition 3.9.

Theorem 6.1. *Let $M \subset U$ be a ladder set and $L \subset U$ the extremal set with $\#L = \#M$. Then $m(L) \succeq m(M)$.*

The proof is by case analysis, and occupies the next three subsections.

In the rest of this section, we fix a ladder set $M \subset U$.

Preliminary of the proof. For two subsets $A, B \subset U$, we define

$$A \gg B \Leftrightarrow \#A = \#B \quad \text{and} \quad m(A) \succeq m(B).$$

Let $X \subset U^{(1)}$ be the super-revlex set with $\#X = \#M^{(1)}$. Then $\{k : M_k^{(1)} \neq \varnothing\} \supset \{k : X_k \neq \varnothing\}$. Thus $X \cup M^{(\geq 2)}$ is also a ladder set in U. Since $X \gg M^{(1)}$ by Corollary 4.6, we have:

Lemma 6.2. *There exists a ladder set $N \subset U$ such that $N^{(1)}$ is super-revlex and $N \gg M$.*

Thus, in the rest of this section we assume that $M^{(1)}$ is super-revlex. Let

$$e = \min\{k + b_1 : x_1^k\boldsymbol{e}_1 \notin M\}$$

and

$$f = \max_{>_{\text{dlex}}} \left\{g \in U_{\leq e}^{(1)} : g \text{ is admissible over } U \text{ and } \#\{h \in U : h \leq_{\text{dlex}} g\} \leq \#M\right\},$$

where $f = 0$ if $\#\{h \in U : h \leq_{\text{dlex}} \boldsymbol{e}_1\} > \#M$. Since $x_1^{e-b_1-1}\boldsymbol{e}_1$ is admissible over U (when $e \neq b_1$), we have $f = x_1^{e-b_1-1}\boldsymbol{e}_1$ or $\deg f = e$. We will prove:

Proposition 6.3. *With the same notation as above, there exists a ladder set N such that $N \gg M$ and*

$$N^{(1)} = \{h \in U^{(1)} : h \leq_{\mathrm{dlex}} f\},$$

where $\{h \in U^{(1)} : h \leq_{\mathrm{dlex}} f\} = \varnothing$ if $f = 0$.

The above proposition proves Theorem 6.1. Indeed, by applying the above proposition repeatedly, one obtains a set N which satisfies the maximal condition and $N \gg M$. Then apply the induction on t. Also, if $t = 1$ then Proposition 6.3 follows from Corollary 4.6. In the rest of this section, we assume that $t > 1$ and that the statement is true when the number of the free basis of U is at most $t-1$. By the above argument, we may assume that Theorem 6.1 is also true when the number of the free basis of U is at most $t-1$.

Lemma 6.4. *There exists a ladder set $N \subset U$ with $N \gg M$ and $\min\{k+b_1 : x_1^k \boldsymbol{e}_1 \notin N^{(1)}\} = e$ satisfying the following conditions*:

(A1) *$N^{(1)}$ is super-revlex and $N^{(\geq 2)}$ is extremal in $U^{(\geq 2)}$.*

(A2) *$\rho(N_e^{(1)}) \cup N^{(2)} \supset U_{\leq e+a_2}^{(2)}$ or $\rho(N_e^{(1)}) \cap N^{(2)} = \varnothing$.*

(A3) *If $t = 2$ and $\rho\big(N_e^{(1)}\big) \cap N^{(2)} = \varnothing$ then $N_e^{(1)} = \varnothing$. If $t \geq 3$ and $\rho\big(N_e^{(1)}\big) \cap N^{(2)} = \varnothing$ then $N_e^{(1)} = \varnothing$ or there exists a $d \geq e$ such that $N^{(2)} = U_{\leq d}^{(2)}$ and $N_{d+1}^{(3)} \neq U_{d+1}^{(3)}$.*

Proof. Let $F = M_e^{(1)}$. Then $M = \big(U_{\leq e-1}^{(1)} \uplus F\big) \uplus M^{(2)} \uplus M^{(\geq 3)}$ since $M^{(1)}$ is super-revlex.

Step 1. We first prove that there exits N satisfying (A1). Let X be the extremal set in $U^{(\geq 2)}$ with $\#X = \#M^{(\geq 2)}$. Let

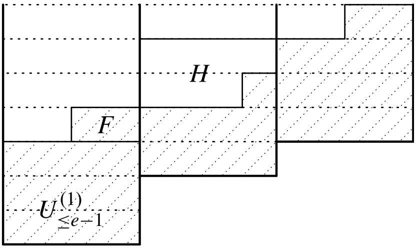
$$N = M^{(1)} \uplus X = U_{\leq e-1}^{(1)} \uplus F \uplus X.$$

Since we assume that Theorem 6.1 is true for $U^{(\geq 2)}$, $N \gg M$. What we must prove is that N is a ladder set. Since $M^{(\geq 2)} \supset U_{\leq e-1}^{(\geq 2)}$, $\#X = \#M^{(\geq 2)} \geq \#U_{\leq e-1}^{(\geq 2)}$. Then Lemma 5.10 says $X \supset U_{\leq e-1}^{(\geq 2)}$, which shows that N is a ladder set if $F = \varnothing$. If $F \neq \varnothing$ then by the definition of ladder sets, $M^{(\geq 2)} \supset U_{\leq e}^{(\geq 2)}$, and $X \supset U_{\leq e}^{(\geq 2)}$ by Lemma 5.10. Hence N is a ladder set.

Step 2. We prove that if M satisfies (A1) but does not satisfy either (A2) or (A3) then there exists an N satisfying (A2) and (A3) such that $N \gg M$ and $\#N^{(1)}$ is strictly smaller than $\#M^{(1)}$. We may assume $\rho(F) \cup M^{(2)} \not\supset U_{\leq e+a_2}^{(2)}$ and $F \neq \varnothing$, otherwise M itself satisfies the desired conditions. Note that $F \neq \varnothing$ implies $M^{(2)} \supset U_{\leq e}^{(2)}$. Let

$$\begin{aligned} a &= \min\{k : M_k^{(2)} \neq U_k^{(2)}\}, \\ b &= \max\{k : k \leq e + a_2,\ \rho(F)_k \neq U_k^{(2)}\}, \\ d &= \max\{k : M_k^{(3)} = U_k^{(3)}\}, \end{aligned}$$

where $d=\infty$ if $n=2$. Let $H=U^{(2)}_{\le d}\setminus M^{(2)}$ (see figure).



M

The set $\rho(F)$ equals $\rho(F)_b \uplus \biguplus_{j=b+1}^{e+a_2} U^{(2)}_j$, since it is an upper revlex set of degree $e+a_2$. Suppose $H=\varnothing$. Then $M^{(2)}=U^{(2)}_{\le d}$. Since $\rho(F)\cup M^{(2)} \not\supset U^{(2)}_{\le e+a_2}$, we have $b>d$ and $\rho(F)\cap M^{(2)}=\varnothing$, which say that M satisfies (A2) and (A3). Suppose $H\neq\varnothing$. Observe that for any super-revlex set L with $U^{(2)}_{\le e}\subset L\subset U^{(2)}_{\le d}$, $M^{(1)}\uplus L\uplus M^{(\ge 3)}$ is a ladder set.

Case 1: Suppose $\#H\ge\#F$. (Note that if $t=2$ then we always have $\#H\ge\#F$.) Then $M^{(2)}$ is super-revlex since we assume that $M^{(\ge 2)}$ is extremal and $\rho(F)$ is an upper revlex set of degree $e+a_2$ with $\#M^{(2)}+\#\rho(F)\le\#U^{(2)}_{\le d}$. Let $R\subset U^{(2)}$ be the super-revlex set in $U^{(2)}$ with $\#R=\#M^{(2)}+\#\rho(F)$. By Corollary 4.5,

$$m(R)\succeq m(M^{(2)})+m(\rho(F))=m(M^{(2)})+m(F). \tag{6}$$

Also, since R is super-revlex, $U^{(2)}_{\le e}\subset R\subset U^{(2)}_{\le d}$. Thus

$$N=U^{(1)}_{\le e-1}\uplus R\uplus M^{(\ge 3)}$$

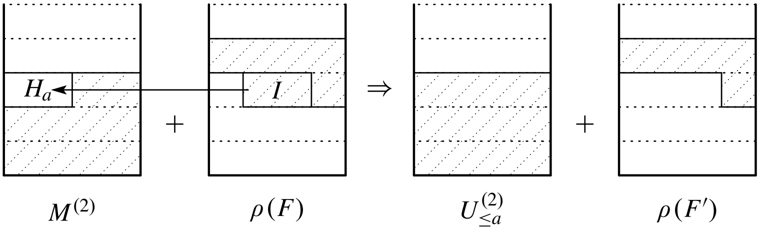
is a ladder set. Then $N^{(1)}_e=\varnothing$ and $N\gg M$ by (6). Hence N satisfies (A2) and (A3).

Case 2: Suppose $\#H<\#F$. Observe that $M^{(2)}\cup\rho(F)$ contains all monomials of degree k in $U^{(2)}$ for $k<a$ and $b<k\le e+a_2$. Since $M\cup\rho(F)\not\supset U^{(2)}_{\le e+a_2}$, we have $a\le b$.

Let $I\subset\rho(F)$ be the interval in $U^{(2)}$ such that $\#I=\#H_a$ and $\rho(F)\setminus I$ is an upper revlex set of degree $e+a_2$, and let $F'\subset F$ be the revlex set with $\rho(F')=\rho(F)\setminus I$. Since H_a is a lower lex set of degree a, the interval lemma gives

$$\begin{aligned} m\big(M^{(2)}\big)+m\big(\rho(F)\big) &\ll m\big(H_a\uplus M^{(2)}\big)+m\big(\rho(F)\setminus I\big)\\ &=m\big(U^{(2)}_{\le a}\big)+m\big(\rho(F')\big).\end{aligned}$$

This is illustrated at the top of the next page.



Suppose $\rho(F') \cup U^{(2)}_{\le a} \supset U^{(2)}_{\le e+a_2}$. Then

$$N = \bigl(U^{(1)}_{\le e-1} \uplus F'\bigr) \uplus U^{(2)}_{\le a} \uplus M^{(\ge 3)}$$

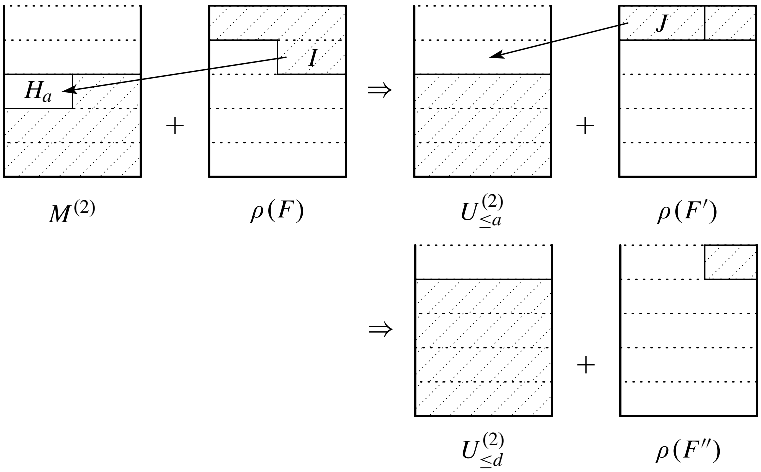
is a ladder set and satisfies $N \gg M$ and conditions (A2) and (A3) since

$$\rho(N^{(1)}_e) \cup N^{(2)} \supset U^{(2)}_{\le e+a_2}.$$

Suppose $\rho(F') \cup U^{(2)}_{\le a} \not\supset U^{(2)}_{\le e+a_2}$. Then $\rho(F') \subset \biguplus_{j=a+1}^{e+a_2} U^{(2)}_j$. Since we assume $\#H < \#F$, $\#F' = \#F - \#H_a > \#(H \setminus H_a)$. Let $J \subset \rho(F')$ be the interval in $U^{(2)}$ such that $\#J = \#(H \setminus H_a)$ and $\rho(F') \setminus J$ is an upper revlex set of degree $e + a_2$, and let $F'' \subset F'$ be the revlex set satisfying $\rho(F'') = \rho(F') \setminus J$. Since $H \setminus H_a = \biguplus_{j=a+1}^{d} U^{(2)}_j$ is a lower lex set of degree $a+1$, the interval lemma yields

$$m\bigl(U^{(2)}_{\le a}\bigr) + m\bigl(\rho(F')\bigr) \preceq m\bigl(M^{(2)} \uplus H\bigr) + m\bigl(\rho(F'')\bigr) = m\bigl(U^{(2)}_{\le d}\bigr) + m\bigl(\rho(F'')\bigr)$$

(see figure).



Then

$$N = \bigl(U^{(1)}_{\le e-1} \uplus F''\bigr) \uplus U^{(2)}_{\le d} \uplus M^{(\ge 3)}$$

is a ladder set and satisfies $N \gg M$ and conditions (A2) and (A3).

Finally, since Step 1 does not change the first component $M^{(1)}$ and Step 2 decreases the first component, by applying Steps 1 and 2 repeatedly, we obtain a set $N \subset U$ satisfying conditions (A1), (A2), and (A3). □

Lemma 6.4 says that to prove Proposition 6.3 we may assume that M satisfies (A1), (A2), and (A3). Thus in the rest of this section we assume that M satisfies these conditions. Also, we may assume $f \neq 0$ since the proposition follows from the induction hypothesis when $f = 0$.

Proof of Proposition 6.3 when $f \neq x_1^{e-b_1-1} e_1$. In this case we have $\deg f = e$. Let

$$f = x_1^{\alpha_1} \cdots x_n^{\alpha_n} \boldsymbol{e}_1$$

and $F = M_e^{(1)}$. Since $x_1^{e-b_1} \boldsymbol{e}_1 \notin F$ by the choice of e, we have $m(F) = m(\rho(F))$. Also, we have

$$M^{(\geq 2)} \supset U_{\leq e}^{(\geq 2)}.$$

Indeed, this is obvious when $F \neq \varnothing$ by the definition of ladder sets. If $F = \varnothing$ then

$$\#M^{(\geq 2)} = \#M - \#U_{\leq e-1}^{(1)} \geq \#\{h \in U : h \leq_{\mathrm{dlex}} f\} - \#U_{\leq e-1}^{(1)} \geq \#U_{\leq e}^{(2)},$$

and since $M^{(\geq 2)}$ is extremal we have $M^{(\geq 2)} \supset U_{\leq e}^{(\geq 2)}$ by Lemma 5.10. Let

$$\epsilon = \deg \rho(f) = \alpha_2 + \cdots + \alpha_n + b_2.$$

Case 1. Suppose $\rho(F) \subset \biguplus_{j=\epsilon}^{e+a_2} U_j^{(2)}$ and $\#F + \#M^{(2)} \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)} \leq \#U_{\leq e+a_2}^{(2)}$. Observe that $M^{(2)} \supset \biguplus_{j=\epsilon}^{e} U_j^{(2)}$. Let P be the super-revlex set with

$$\#P = \#M^{(2)} \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)},$$

and let $Q \subset U^{(2)}$ be the super-revlex set with $\#Q = \#F + \#M^{(2)} \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)}$. Since $\rho(F)$ is an upper revlex set of degree $e + a_2$ and $M^{(2)} \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)}$ is revlex, by Corollaries 4.5 and 4.6, we have

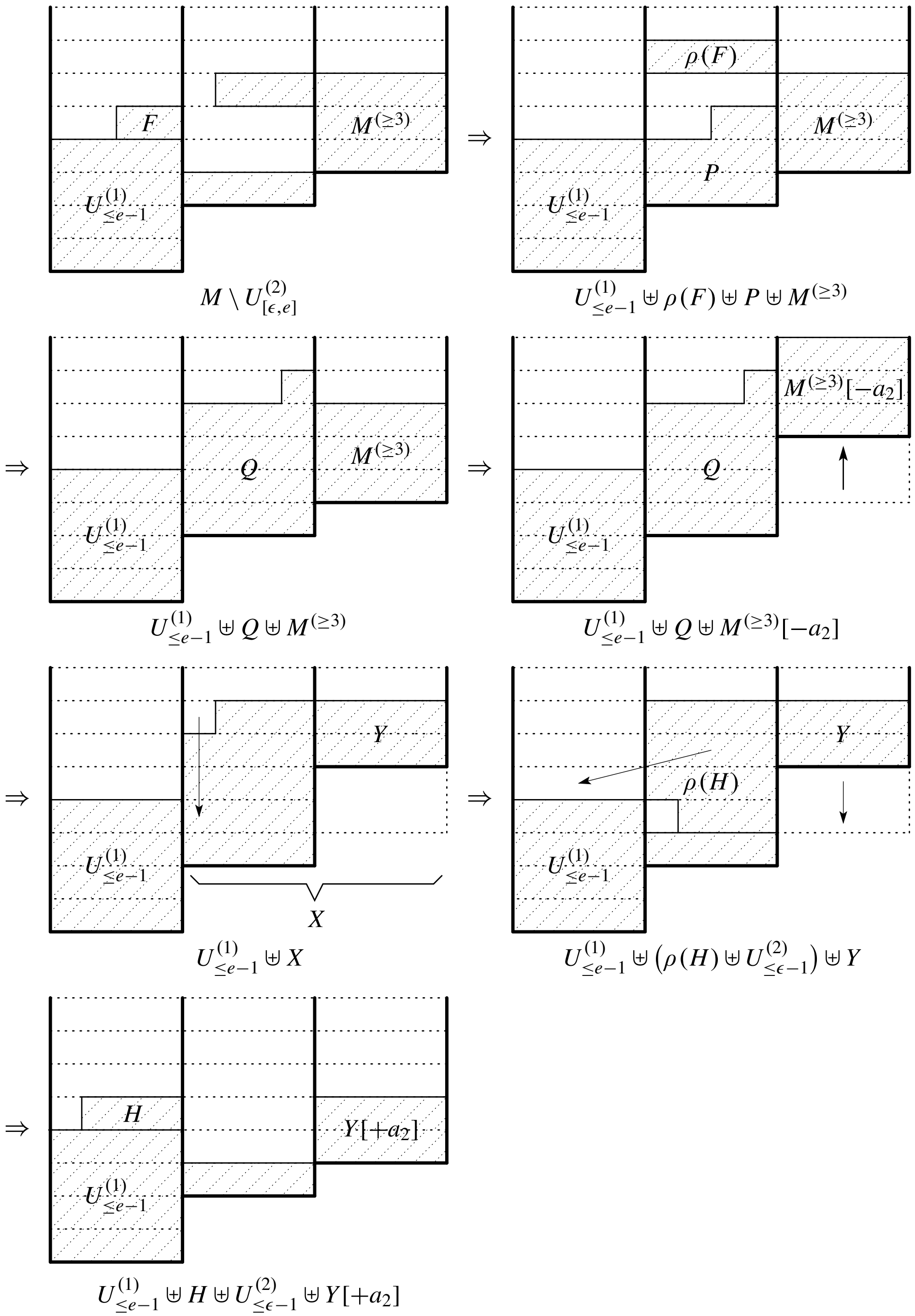
$$m(Q) \succeq m(P) + m\bigl(\rho(F)\bigr) \succeq m\left(M^{(2)} \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)}\right) + m(F) \tag{7}$$

(see the first two steps in Figure 1).

Observe that $Q \subset U_{\leq e+a_2}^{(2)}$ since $\#Q \leq \#U_{\leq e+a_2}^{(2)}$ by the assumption of Case 1. Let $U' = U^{(2)} \uplus \biguplus_{i=3}^{t} U^{(i)}[-a_2]$. Since $M^{(\geq 3)}[-a_2] \supset U_{\leq e}^{(\geq 3)}[-a_2] = U'^{(\geq 3)}_{\leq e+a_2}$,

$$Q \uplus M^{(\geq 3)}[-a_2] \subset U'$$

is a ladder set in U' (see the third step in Figure 1).



$M \setminus U^{(2)}_{[\epsilon,e]}$

$U^{(1)}_{\leq e-1} \uplus \rho(F) \uplus P \uplus M^{(\geq 3)}$

$U^{(1)}_{\leq e-1} \uplus Q \uplus M^{(\geq 3)}$

$U^{(1)}_{\leq e-1} \uplus Q \uplus M^{(\geq 3)}[-a_2]$

$U^{(1)}_{\leq e-1} \uplus X$

$U^{(1)}_{\leq e-1} \uplus \big(\rho(H) \uplus U^{(2)}_{\leq \epsilon-1}\big) \uplus Y$

$U^{(1)}_{\leq e-1} \uplus H \uplus U^{(2)}_{\leq \epsilon-1} \uplus Y[+a_2]$

Figure 1. Some steps in the proof of Proposition 6.3 in the case when $f \neq x_1^{e-b_1-1}\boldsymbol{e}_1$. See bottom of previous page and middle and bottom of page 1047.

Let g be the largest admissible monomial in $U^{(2)}_{\le e+a_2}$ over U' with respect to $>_{\text{dlex}}$ satisfying

$$\#\{h \in U' : h \le_{\text{dlex}} g\} \le \#Q \uplus M[-a_2]^{(\ge 3)}.$$

By the induction hypothesis, there exists $Y \subset U'^{(\ge 3)}$ such that

$$X = \{h \in U^{(2)} : h \le_{\text{dlex}} g\} \uplus Y \subset U'$$

is a ladder set in U' and

$$X \gg Q \uplus M^{(\ge 3)}. \tag{8}$$

Lemma 6.5. *Let $d = e + a_2 - \epsilon$. Then $g \ge_{\text{lex}} x_2^d \rho(f)$.*

Proof. Consider

$$L = \{h \in U : h \le_{\text{dlex}} f\}.$$

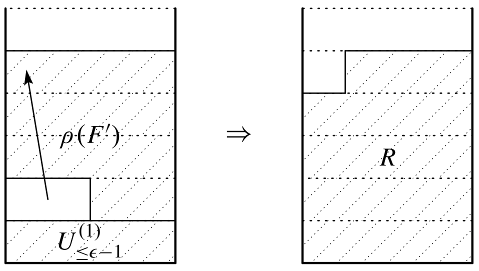
Then $\#M \ge \#L$ and $L^{(\ge 2)} = U^{(\ge 2)}_{\le e}$. Thus $L^{(2)} \setminus \biguplus_{j=\epsilon}^{e} U^{(2)}_j = U^{(2)}_{\le \epsilon - 1}$. Let $F' = L^{(1)}_e = [f, x_n^{e-b_1}\boldsymbol{e}_1]$. Then $\rho(F') = [\rho(f), x_n^{\epsilon - b_2}\boldsymbol{e}_2] \uplus \biguplus_{j=\epsilon+1}^{e+a_2} U^{(2)}_j$. Also, $\rho(F')$ is disjoint from $L^{(2)} \setminus \biguplus_{j=\epsilon}^{e} U^{(2)}_j$ and

$$\begin{aligned} m\Bigg(\rho(F') \uplus \Bigg(L^{(2)} \setminus \biguplus_{j=\epsilon}^{e} U^{(2)}_j\Bigg)\Bigg) &= m\big(U^{(2)}_{\le e+a_2} \setminus \big[x_2^{\epsilon - b_2}\boldsymbol{e}_2, \rho(f)\big)\big) \\ &= m\big(U^{(2)}_{\le e+a_2} \setminus \big[x_2^{e+a_2-b_2}\boldsymbol{e}_2, x_2^d \rho(f)\big)\big). \end{aligned}$$

Let

$$R = U^{(2)}_{\le e+a_2} \setminus \big[x_2^{e+a_2-b_2}\boldsymbol{e}_2, x_2^d\rho(f)\big) = U^{(2)}_{\le e+a_2-1} \uplus \big[x_2^d\rho(f), x_n^{e+a_2-b_2}\boldsymbol{e}_2\big]$$

(see figure).



$U^{(1)}_{\le \epsilon - 1} \uplus \rho(F')$ R

Then $R \uplus L^{(\ge 3)}[-a_2] \subset U'$ is a ladder set in U' and $x_2^d\rho(f)$ is admissible over U' by Lemma 5.9. On the other hand,

$$\#R \uplus L^{(\ge 3)} = \#L - \#U^{(1)}_{\le e-1} - \#\biguplus_{j=\epsilon}^{e} U^{(2)}_j \le \#M - \#U^{(1)}_{\le e-1} - \#\biguplus_{j=\epsilon}^{e} U^{(2)}_j = \#X.$$

Since $x_2^d \rho(f)$ is admissible over U' and since $R \uplus L^{(\geq 3)}[-a_2] = \{h \in U' : h \leq_{\mathrm{dlex}} x_2^d \rho(f)\}$, by the choice of g, we have

$$g \geq_{\mathrm{lex}} x_2^d \rho(f)$$

as desired. □

By Lemma 6.5, g is divisible by x_2^d. Let $H \subset U_e^{(1)}$ be the revlex set such that

$$\rho(H) = \biguplus_{j=\epsilon}^{e+a_2} U_j^{(2)} \setminus x_2^{-d}\bigl[x_2^{e+a_2-b_2}\boldsymbol{e}_2, g\bigr).$$

Then by Lemma 4.3

$$m(H) + m\bigl(U_{\leq \epsilon-1}^{(2)}\bigr) \succeq m\bigl(U_{\leq e+a_2}^{(2)} \setminus [x_2^{e+a_2-b_2}\boldsymbol{e}_2, g)\bigr) = m\bigl(X^{(2)}\bigr). \tag{9}$$

Let

$$N = \bigl(U_{\leq e-1}^{(1)} \uplus H\bigr) \uplus U_{\leq e}^{(2)} \uplus Y[+a_2] \subset U.$$

Since X is a ladder set, $Y \supset U'^{(\geq 3)}_{\leq e+a_2}$ and $Y[+a_2] \supset U_{\leq e}^{(\geq 3)}$. Thus N is a ladder set in U. We claim that N satisfies the desired conditions.

A routine computation shows

$$\#M \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)} = \#U_{\leq e-1}^{(1)} \uplus Q \uplus M^{(\geq 3)} = \#U_{\leq e-1}^{(1)} \uplus X = \#N \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)}$$

(see Figure 1). Thus $\#N = \#M$. Let $\mu = \max_{>_{\mathrm{lex}}} H$. Then $x_2^d \rho(\mu) = g$. We claim that $\mu = f$. Since $g \geq_{\mathrm{lex}} x_2^d \rho(f)$, $\mu \geq_{\mathrm{lex}} f$. Since g is admissible over U', μ is admissible over U by Lemma 5.9 (If $t = 2$ then Lemma 5.9 is not applicable; however, if $t = 2$ then any monomial $h \in U_e^{(1)}$ with $h >_{\mathrm{lex}} f$ is admissible). However, since $\#N = \#M$ and $N \supset \{h \in U : h \leq_{\mathrm{dlex}} \mu\}$, by the choice of f, we have $f = \mu$.

It remains to prove $N \gg M$. This follows from (7), (8), and (9) as follows:

$$\begin{aligned} M \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)} &= \bigl(U_{\leq e-1}^{(1)} \uplus F\bigr) \uplus \left(M^{(2)} \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)}\right) \uplus M^{(\geq 3)} \\ &\ll U_{\leq e-1}^{(1)} \uplus Q \uplus M^{(\geq 3)} \\ &\ll U_{\leq e-1}^{(1)} \uplus X \\ &\ll \bigl(U_{\leq e-1}^{(1)} \uplus H\bigr) \uplus U_{\leq \epsilon-1}^{(2)} \uplus Y[+a_2] = N \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)} \end{aligned}$$

(see Figure 1).

Case 2. Suppose $\rho(F) \subset \biguplus_{j=\epsilon}^{e+a_2} U_j^{(2)}$ and $\#F + \#M^{(2)} \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)} > \#U^{(2)}_{\le e+a_2}$.

Lemma 6.6. *We have* $f = x_1^{\alpha_1} x_2^{\alpha_2} \boldsymbol{e}_1$; *that is*, $\alpha_3 = \cdots = \alpha_n = 0$.

Proof. Suppose $f \neq x_1^{\alpha_1} x_2^{\alpha_2} \boldsymbol{e}_1$. Let $g = x_1^{\alpha_1} x_2^{\alpha_2+\alpha_3+\cdots+\alpha_n} \boldsymbol{e}_1$. Then $g >_{\mathrm{dlex}} f$ is admissible over U by the definition of admissibility. Also,

$$\#M < \#\{h \in U : h \le_{\mathrm{dlex}} g\} = \#\big(U^{(1)}_{\le e-1} \uplus [g, x_n^{e-b_1} \boldsymbol{e}_1]\big) \uplus U^{(2)}_{\le e} \uplus U^{(\ge 3)}_{\le e}.$$

Since $\rho([g, x_n^{e-b_1} \boldsymbol{e}_1]) = \biguplus_{i=\epsilon}^{e+a_2} U_i^{(2)}$ and $M^{(\ge 3)} \supset U^{(\ge 3)}_{\le e}$,

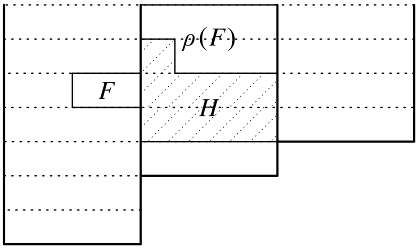
$$\begin{aligned} \#F + \#\Big(M^{(2)} \setminus \biguplus_{j=\epsilon}^{e} U_j^{(2)}\Big) &= \big(\#M - \#U^{(1)}_{\le e-1} - \#M^{(\ge 3)}\big) - \#\biguplus_{j=\epsilon}^{e} U_j^{(2)} \\ &< \#[g, x_n^{e-b_1} \boldsymbol{e}_1] + \#U^{(2)}_{\le e} - \#\biguplus_{j=\epsilon}^{e} U_j^{(2)} = \#U^{(2)}_{\le e+a_2}, \end{aligned}$$

which contradicts the assumption of Case 2. Thus $f = x_1^{\alpha_1} x_2^{\alpha_2} \boldsymbol{e}_1$. □

Lemma 6.6 says that $\rho(f) = x_2^{\epsilon-b_2} \boldsymbol{e}_2$. In particular, $\rho([f, x_n^{e-b_1} \boldsymbol{e}_1]) = \bigcup_{j=\epsilon}^{e+a_2} U_j^{(2)}$. Let

$$H = \biguplus_{j=\epsilon}^{e+a_2} U_j^{(2)} \setminus \rho(F)$$

(see figure).



H

Since $\rho(F)$ is an upper revlex set of degree $e + a_2$, H is a lower lex set of degree ϵ. Also, since $\#F + \#M^{(2)} > \#U^{(2)}_{\le e+a_2}$, $\rho(F) \cup M^{(2)} \supset U^{(\ge 2)}_{\le e+a_2}$ by (A2). Thus $M^{(2)} \supset H$.

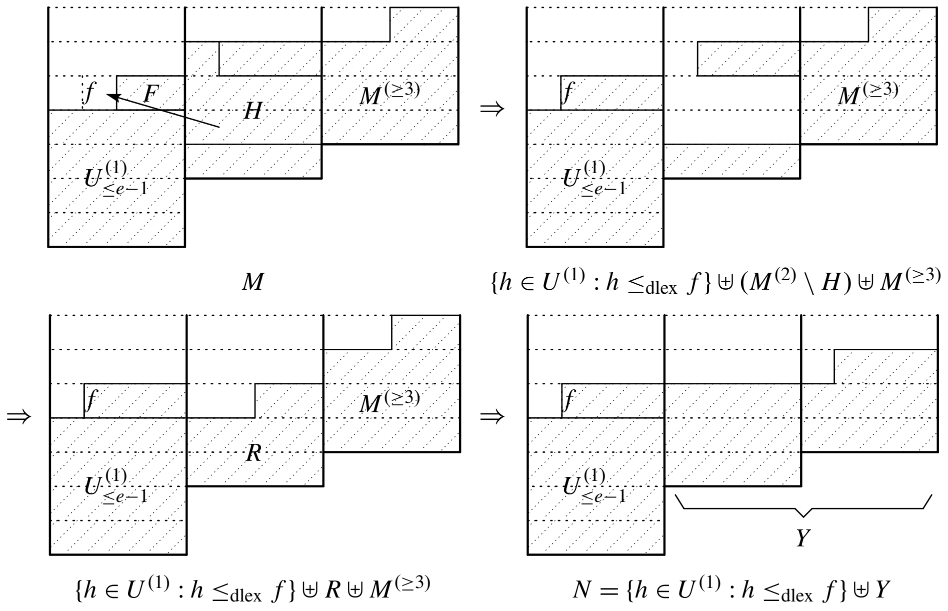
Let R be the super-revlex set in $U^{(2)}$ with $\#R = \#M^{(2)} \setminus H$. Since $M^{(2)} \setminus H$ is revlex, by Corollary 4.6 we have

$$R \gg M^{(2)} \setminus H. \tag{10}$$

Then since $\#R \le \#M^{(2)}$,

$$R \uplus M^{(\ge 3)} \subset U^{(\ge 2)}$$

is a ladder set (see the third picture in Figure 2).



M

$\{h \in U^{(1)} : h \leq_{\text{dlex}} f\} \uplus (M^{(2)} \setminus H) \uplus M^{(\geq 3)}$

$\{h \in U^{(1)} : h \leq_{\text{dlex}} f\} \uplus R \uplus M^{(\geq 3)}$

$N = \{h \in U^{(1)} : h \leq_{\text{dlex}} f\} \uplus Y$

Figure 2. Toward the proof of Case 2.

Let $Y \subset U^{(\geq 2)}$ be the extremal set in $U^{(\geq 2)}$ with $\#Y = \#R \uplus M^{(\geq 3)}$. We claim that

$$N = \{h \in U^{(1)} : h \leq_{\text{dlex}} f\} \uplus Y$$

satisfies the desired conditions. Indeed, we have

$$\begin{aligned} M &= \big(U^{(1)}_{\leq e-1} \uplus F \uplus H\big) \uplus (M^{(2)} \setminus H) \uplus M^{(\geq 3)} \\ &\ll \big(U^{(1)}_{\leq e-1} \uplus [f, x_n^{e-b_1}\boldsymbol{e}_1]\big) \uplus R \uplus M^{(\geq 3)} \\ &\ll \{h \in U^{(1)} : h \leq_{\text{dlex}} f\} \uplus Y = N \end{aligned}$$

(see Figure 2) since $\rho(F) \uplus H = \biguplus_{j=\epsilon}^{e+a_2} U_j^{(2)} = \rho([f, x_n^{e-b_1}\boldsymbol{e}_2])$, by (10).

It remains to prove that N is a ladder set. Since

$$\#Y = \#M - \#\{h \in U^{(1)} : h \leq_{\text{dlex}} f\} \geq \#U^{(\geq 2)}_{\leq e}$$

by the choice of f, we have $Y \supset U^{(\geq 2)}_{\leq e}$ by Lemma 5.10. This fact guarantees that N is a ladder set.

Case 3. Suppose $\rho(F) \not\subset \biguplus_{j=\epsilon}^{e+a_2} U_j^{(2)}$. Then $\rho(F)$ properly contains $\biguplus_{j=\epsilon}^{e+a_2} U_j^{(2)}$ since $\rho(F)$ is an upper revlex set of degree $e+a_2$. In particular, F properly contains $[f, x_n^{e-b_1}\boldsymbol{e}_1]$. We claim:

Lemma 6.7. *We have $f = x_1^{\alpha_1}x_2^{\alpha_2}\boldsymbol{e}_1$ and $\alpha_2 \neq 0$.*

Proof. If $\alpha_k \neq 0$ for some $k \geq 3$ then $x_1^{\alpha_1} x_2^{\alpha_2+\cdots+\alpha_n} \boldsymbol{e}_1 >_{\text{dlex}} f$ is admissible over U. Then by the choice of f, $F \subset [x_1^{\alpha_1} x_2^{\alpha_2+\cdots+\alpha_n} \boldsymbol{e}_1, x_n^{e-b_1} \boldsymbol{e}_1]$ and

$$\rho(F) \subset \rho\big([x_1^{\alpha_1} x_2^{\alpha_2+\cdots+\alpha_n} \boldsymbol{e}_1, x_n^{e-b_1} \boldsymbol{e}_1]\big) = \biguplus_{j=\epsilon}^{e+a_2} U_j^{(2)},$$

a contradiction. Also, if $\alpha_2 = 0$ then $\epsilon = \deg \rho(f) = 0$ which implies

$$\rho(F) \subset \rho(U_e^{(1)}) = U_{\leq e+a_2}^{(2)} = \biguplus_{j=\epsilon}^{e+a_2} U_j^{(2)},$$

a contradiction. □

Recall $\epsilon = \deg \rho(f)$. Thus $\alpha_2 = \epsilon - b_2$ by Lemma 6.7. Let

$$H = \{h \in F : h >_{\text{lex}} f\}$$

and

$$g = \max_{>_{\text{lex}}} H.$$

By the choice of f, H contains no admissible monomials over U. By Lemma 6.7, $\rho(F \setminus H) = \biguplus_{j=\epsilon}^{e+a_2} U_j^{(2)}$. Hence $H \neq \varnothing$ by the assumption of Case 3. Since $x_1^{\alpha_1+1} x_2^{\alpha_2-1} \boldsymbol{e}_1$ is admissible over U,

$$\rho(H) \subset \rho\big([x_1^{\alpha_1+1} x_2^{\alpha_2-1} \boldsymbol{e}_1, x_1^{\alpha_1} x_2^{\alpha_2} \boldsymbol{e}_1)\big) = U_{\epsilon-1}^{(2)}$$

is revlex. Also, $\epsilon - 1 > b_2$ since $U_{b_2}^{(2)} = \{\boldsymbol{e}_2\}$ and $H \neq \varnothing$.

If $t = 2$ then any monomial $h \in U_e^{(1)}$ with $h >_{\text{lex}} f$ is admissible, which implies $H = \varnothing$. Thus we may assume $t \geq 3$.

To prove the statement, it is enough to prove that there exists an extremal set $Z \subset U^{(\geq 3)}$ such that

$$Z \gg H \uplus M^{(\geq 3)}. \tag{11}$$

Indeed, if such a Z exists then $N = (M^{(1)} \setminus H) \uplus M^{(2)} \uplus Z$ satisfies the desired conditions. Recall that $\epsilon \leq e + 1$ by the definition of admissibility.

<u>Subcase 3-1</u>. Suppose $a_3 \geq e - (\epsilon - 1)$.

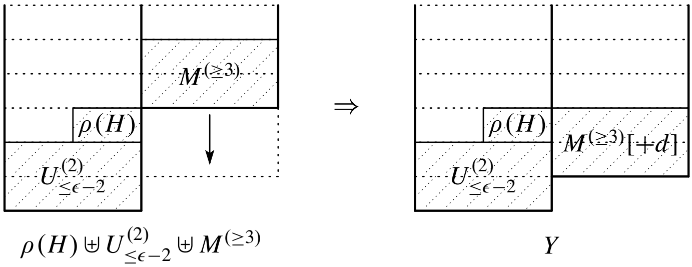
Let $d = e - (\epsilon - 1)$. We consider

$$U' = U^{(2)} \uplus \biguplus_{i=3}^{t} U^{(i)}[+d].$$

This set is well-defined since $a_3 \geq d$. Recall $\rho(H) \subset U_{\epsilon-1}^{(2)}$. Let

$$Y = \rho(H) \uplus U_{\leq \epsilon-2}^{(2)} \uplus M^{(\geq 3)}[+d]$$

(see figure).


$M^{(\geq 3)}$
$\rho(H)$
$U^{(2)}_{\leq \epsilon-2}$
$\Rightarrow$
$\rho(H)$
$M^{(\geq 3)}[+d]$
$U^{(2)}_{\leq \epsilon-2}$
$\rho(H) \uplus U^{(2)}_{\leq \epsilon-2} \uplus M^{(\geq 3)}$
Y

Then Y is a ladder set since $M^{(\geq 3)} \supset U^{(\geq 3)}_{\leq \epsilon-1+d} = U^{(\geq 3)}_{\leq e}$. Also, $U^{(2)}_{\leq \epsilon-2} \neq \varnothing$ since $\epsilon - 1 > b_2$.

Let $\mu \in U^{(2)}_{\leq \epsilon-1}$ be the largest admissible monomial in $U^{(2)}_{\leq \epsilon-1}$ over U' with respect to $>_{\text{dlex}}$ satisfying $\#\{h \in U' : h \leq_{\text{dlex}} \mu\} \leq \#Y$. Then since we assume that Proposition 6.3 is true for U', there exists an extremal set $Z \subset U'^{(\geq 3)}$ such that

$$Y \ll \{h \in U^{(2)} : h \leq_{\text{dlex}} \mu\} \uplus Z.$$

To prove (11), it is enough to prove $\{h \in U^{(2)} : h \leq_{\text{dlex}} \mu\} = U^{(2)}_{\leq \epsilon-2}$; in other words:

Lemma 6.8. $$\mu = x_2^{\epsilon-2-b_2} \boldsymbol{e}_2.$$

Proof. Recall that $U^{(2)}_{\leq \epsilon-2} \neq \varnothing$. It is enough to prove that $\deg \mu \neq \epsilon - 1$. Suppose to the contrary that $\deg \mu = \epsilon - 1$. Let $\mu' \in U^{(1)}_e$ be a monomial such that $\rho(\mu') = \mu$. Then μ' is admissible over U by Lemma 5.9. Also,

$$\#Y - \#U^{(2)}_{\leq \epsilon-2} \geq \#[\mu, x_n^{\epsilon-1-b_2} \boldsymbol{e}_2] + \#U'^{(\geq 3)}_{\leq \epsilon-1} = \#[\mu, x_n^{\epsilon-1-b_2} \boldsymbol{e}_2] + \#U^{(\geq 3)}_{\leq e}.$$

Since $\#M^{(\geq 3)} + \#H = \#Y - \#U^{(2)}_{\leq \epsilon-2}$ and since $\rho\big([\mu', f)\big) = [\mu, x_n^{\epsilon-1-b_2} \boldsymbol{e}_2]$, we have

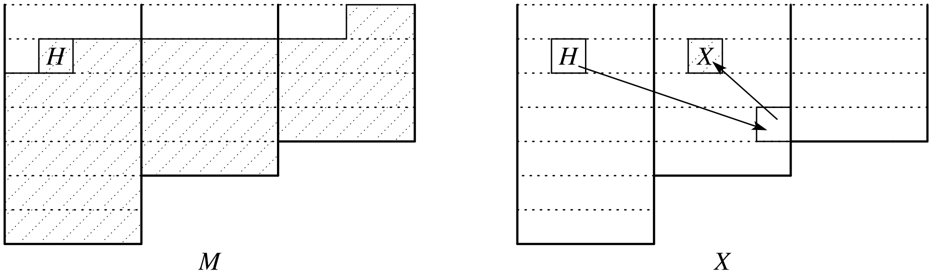
$$\begin{aligned} \#M &= \#(M \setminus H) \uplus M^{(2)} \uplus H \uplus M^{(\geq 3)} \\ &\geq \#(M \setminus H) \uplus U^{(2)}_{\leq e} + \#[\mu, x_n^{\epsilon-1-b_2} \boldsymbol{e}_2] \uplus Z \\ &\geq \#[\mu', f) \uplus (M \setminus H) \uplus U^{(\geq 2)}_{\leq e} = \#\{h \in U : h \leq_{\text{dlex}} \mu'\}, \end{aligned}$$

which contradicts the maximality of f since $\mu' >_{\text{lex}} g >_{\text{lex}} f$ and μ' is admissible over U. □

Subcase 3-2. Suppose $a_3 < e - (\epsilon - 1)$. We consider

$$X = x_2^{e-(\epsilon-1)} \rho(H) \subset U^{(2)}_e,$$

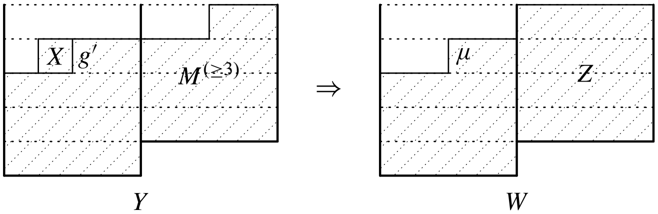
as illustrated at the top of the next page.



Let

$$Y=\left\{h\in U^{(2)}:h\le_{\mathrm{dlex}}x_2^{e-(\epsilon-1)}\rho(g)\right\}\uplus M^{(\geq 3)},$$

as on the left part of the figure:



Further, let

$$g'=\max_{>_{\mathrm{dlex}}}(Y^{(2)}\setminus X).$$

Since $e-(\epsilon-1)>a_3$, $e-(\epsilon-1)\geq 1$. Thus

$$g'=x_2^{e-(\epsilon-1)-1}x_3^{\epsilon-b_2}\boldsymbol{e}_2$$

and

$$Y^{(2)}=X\uplus\{h\in U^{(2)}:h\le_{\mathrm{dlex}}g'\}.$$

Since $a_3<e-(\epsilon-1)$, $\deg\rho(g')=\epsilon+a_3\le e$. Thus g' is admissible over $U^{(\geq 2)}$.

Let μ be the largest admissible monomial in $U^{(2)}_{\le e}$ over $U^{(\geq 2)}$ with respect to $>_{\mathrm{dlex}}$ with $\#\{h\in U^{(\geq 2)}:h\le_{\mathrm{dlex}}\mu\}\le\#Y$. Since Lemma 5.9 says that X contains no admissible monomials over $U^{(\geq 2)}$,

$$\mu\geq_{\mathrm{dlex}}g'\text{ and }\mu\notin X.$$

Since we assume that Proposition 6.3 is true for $U^{(\geq 2)}$, there exists an extremal set $Z\subset U^{(\geq 3)}$ such that

$$W=\{h\in U^{(2)}:h\le_{\mathrm{dlex}}\mu\}\uplus Z$$

is a ladder set and

$$W\gg Y,$$

as shown in the figure immediately above.

Lemma 6.9. $$\mu = g'.$$

Proof. Suppose to the contrary that $\mu \neq g'$. Then $\mu >_{\text{dlex}} g'$ and

$$W = \big[\mu, x_2^{e-(\epsilon-1)}\rho(g)\big) \uplus Y^{(2)} \uplus Z.$$

Then there exists $\mu' \in U_e^{(1)}$ such that

$$x_2^{e-(\epsilon-1)}\rho(\mu') = \mu.$$

By Lemma 5.9, μ' is admissible over U and $\mu' >_{\text{lex}} g >_{\text{lex}} f$. Observe that

$$\#M^{(\geq 3)} + \#H = \#Z \uplus [\mu, g') = \#Z + \#[\mu', f)$$

by the construction of Y and Z. Since $Z \supset U_{\leq e}^{(\geq 3)}$,

$$\begin{aligned}\#M &\geq \#(M^{(1)} \setminus H) \uplus H \uplus U_{\leq e}^{(2)} \uplus M^{(\geq 3)}\\ &= \#(M^{(1)} \setminus H) \uplus U_{\leq e}^{(2)} \uplus Z \uplus [\mu', f)\\ &\geq \#(M^{(1)} \setminus H) \uplus [\mu', f) \uplus U_{\leq e}^{(2)} \uplus U_{\leq e}^{(\geq 3)}\\ &= \#\{h \in U : h \leq_{\text{dlex}} \mu'\}.\end{aligned}$$

Since μ' is admissible over U, this contradicts the maximality of f. □

Now

$$W = \{h \in U^{(2)} : h \leq_{\text{dlex}} g'\} \uplus Z$$

and since $W \gg Y$ and $Y = X \uplus \{h \in U^{(2)} : h \leq_{\text{dlex}} g'\} \uplus M^{(\geq 3)}$, we have

$$m(Z) \succeq m(X \uplus M^{(\geq 3)}) = m(H \uplus M^{(\geq 3)}),$$

which proves (11). This completes the proof of Proposition 6.3 when $f \neq x_1^{e-b_1-1}\boldsymbol{e}_1$.

Proof of Proposition 6.3 when $f = x_1^{e-b_1-1}\boldsymbol{e}_1$. Let $F = M_e^{(1)}$. If $F = \varnothing$ then there is nothing to prove. Thus we may assume $F \neq \varnothing$. Then $M \supset U_{\leq e}^{(\geq 2)}$ since M is a ladder set.

Case 1. Suppose $a_2 = 0$. Then $\deg \boldsymbol{e}_1 = \deg \boldsymbol{e}_2 = b_1$. Since $x_2^{e-b_1}\boldsymbol{e}_1$ is admissible over U, $x_2^{e-b_1}\boldsymbol{e}_1 \notin F$. Indeed, if $x_2^{e-b_1}\boldsymbol{e}_1 \in F$ then $M \supset \{h \in U : h \leq_{\text{dlex}} x_2^{e-b_1}\boldsymbol{e}_1\}$, which contradicts the maximality of f. Thus

$$F \subset [x_2^{e-b_1}\boldsymbol{e}_1, x_n^{e-b_1}\boldsymbol{e}_1]$$

and

$$\rho(F) \subset \rho\big([x_2^{e-b_1}\boldsymbol{e}_1, x_n^{e-b_1}\boldsymbol{e}_1]\big) = U_e^{(2)}.$$

Consider

$$X = \rho(F) \uplus U_{\leq e-1}^{(2)} \uplus M^{(\geq 3)} \subset U^{(\geq 2)}$$

and let $Y \subset U^{(\geq 2)}$ be the extremal set with $\#Y = \#X$. Since X is a ladder set in $U^{(\geq 2)}$, by the induction hypothesis we have

$$Y \gg X.$$

Lemma 6.10. $$Y^{(2)} = U^{(2)}_{\leq e-1}.$$

Proof. Suppose to the contrary that $Y^{(2)} \neq U^{(2)}_{\leq e-1}$. Let $g = \bar{g}\boldsymbol{e}_2$ be the largest admissible monomial in $Y^{(2)}_{\leq e}$ over $U^{(\geq 2)}$ with respect to $>_{\text{dlex}}$. Since $X \supset U^{(\geq 2)}_{\leq e-1}$, we have $Y \supset U^{(2)}_{\leq e-1}$ by Lemma 5.10. Thus $\deg g = e$ and $Y \supset U^{(\geq 3)}_{\leq e}$.

Let $g' = \bar{g}\boldsymbol{e}_1$. Since $g = \bar{g}\boldsymbol{e}_2$ is admissible over $U^{(\geq 2)}$ and since $\rho(g') = g$, g' is admissible over U by Lemma 5.9. Observe that $\#Y = \#X \leq \#F + \#M^{(\geq 2)} - \#U^{(2)}_e$. Then

$$\begin{aligned}
\#M &= \#U^{(1)}_{\leq e-1} \uplus F \uplus M^{(\geq 2)}\\
&\geq \#U^{(1)}_{\leq e-1} + \#U^{(2)}_e + \#Y\\
&\geq \#U^{(1)}_{\leq e-1} + \#U^{(2)}_e + \#\{h \in U^{(\geq 2)} : h \leq_{\text{dlex}} g\}\\
&= \#U^{(1)}_{\leq e-1} + \#U^{(2)}_e + \#U^{(2)}_{\leq e-1} \uplus [g, x_n^{e-b_1}\boldsymbol{e}_2] \uplus U^{(\geq 3)}_{\leq e}\\
&= \#U^{(1)}_{\leq e-1} + \#U^{(\geq 2)}_{\leq e} + \#[g', x_n^{e-b_1}\boldsymbol{e}_1]\\
&= \#\{h \in U : h \leq_{\text{dlex}} g'\},
\end{aligned}$$

which contradicts the maximality of f. Hence $Y^{(2)} = U^{(2)}_{\leq e-1}$. □

Then, since $Y \gg X$, we have

$$Y^{(\geq 3)} \gg F \uplus M^{(\geq 3)}. \tag{12}$$

Let

$$N = U^{(1)}_{\leq e-1} \uplus M^{(2)} \uplus Y^{(\geq 3)}.$$

Then N is a ladder set since $\#Y^{(\geq 3)} \geq \#M^{(\geq 3)}$. Also, $N \gg M$ by (12). Thus N satisfies the desired conditions.

Case 2. Suppose $a_2 > 0$. Since $\deg f \neq e$, by Lemma 5.12 we have

$$\#M < \#U^{(1)}_{\leq e}. \tag{13}$$

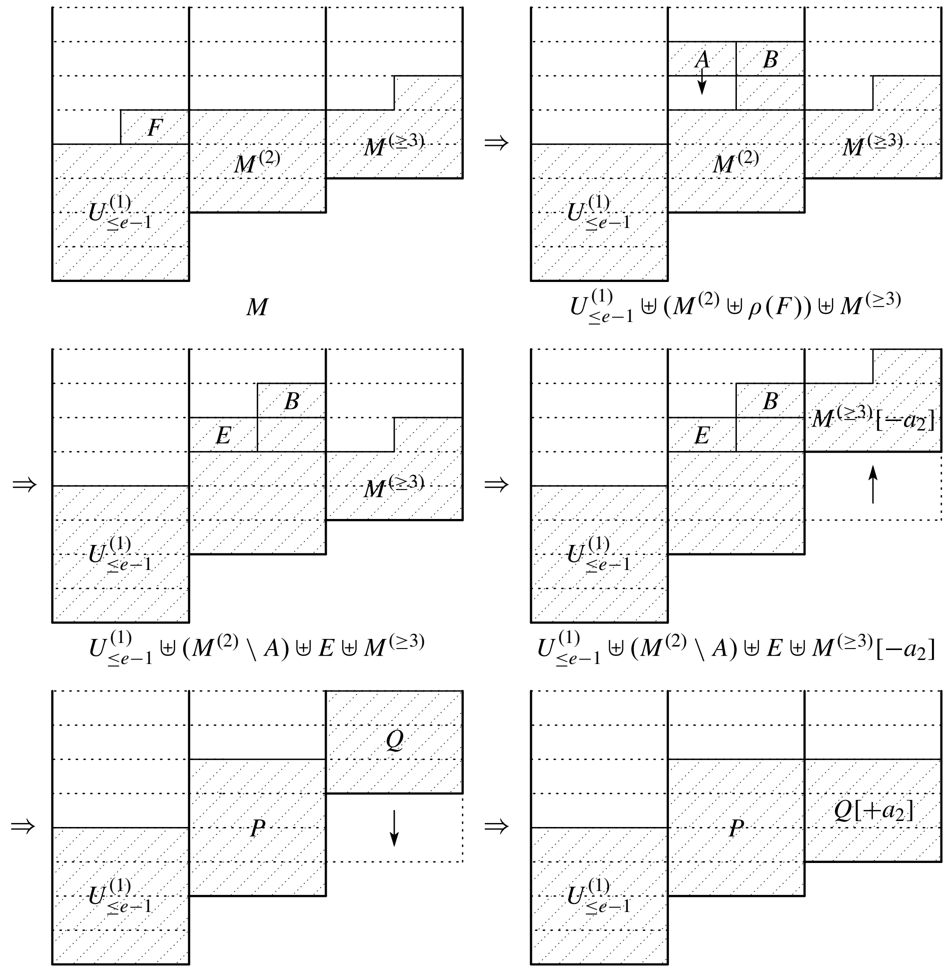
Hence

$$\#F + \#M^{(2)} \leq \#M - \#U^{(1)}_{\leq e-1} < \#U^{(1)}_e \leq \#U^{(2)}_{\leq e+a_2}. \tag{14}$$

Then, by (A2) and (A3), we may assume that $\rho(F) \cap M^{(2)} = \varnothing$, $t \geq 3$, and there exists a $d \geq e$ such that $M^{(2)} = U^{(2)}_{\leq d}$ and $M^{(3)}_{d+1} \neq U^{(3)}_{d+1}$. Let

$$A = \left\{u\boldsymbol{e}_2 \in \rho(F)_{e+a_2} : x_2^{(e+a_2)-(d+1)} \text{ divides } u \text{ and } u/x_2^{(e+a_2)-(d+1)}\boldsymbol{e}_2 \notin \rho(F)_{d+1}\right\},$$

as illustrated in the second picture at the top of the next page.



M

$U^{(1)}_{\leq e-1} \uplus (M^{(2)} \uplus \rho(F)) \uplus M^{(\geq 3)}$

$U^{(1)}_{\leq e-1} \uplus (M^{(2)} \setminus A) \uplus E \uplus M^{(\geq 3)}$

$U^{(1)}_{\leq e-1} \uplus (M^{(2)} \setminus A) \uplus E \uplus M^{(\geq 3)}[-a_2]$

$U_{\leq e-1} \uplus P \uplus Q$

$N = U^{(1)}_{\leq e-1} \uplus P \uplus Q[+a_2]$

Also set

$$E = x_2^{-(e+a_2+d+1)} A \subset U^{(2)}_{d+1} \quad \text{and} \quad B = \rho(F)_{e+a_2} \setminus A \subset U^{(2)}_{e+a_2}.$$

Subcase 2-1. Suppose $\#B + \#M^{(\geq 3)} < \#U^{(2)}_{e+a_2}$. Consider

$$U' = U^{(2)} \uplus \biguplus_{i=3}^{t} U^{(i)}[-a_2].$$

Since $M^{(\geq 3)}[-a_2] \supset U'^{(\geq 3)}_{\leq e+a_2}$, by Corollary 5.13 and by the induction hypothesis, there exists the extremal set $Q \subset U'^{(\geq 3)}$ such that

$$Q \gg B \uplus M^{(\geq 3)}. \tag{15}$$

Let P be the super-revlex set in $U^{(2)}$ with $\#P = \#M^{(2)} + \#\rho(F) \setminus B$. Then since $\rho(F)_{\le e+a_2-1} \uplus E$ is revlex, Corollary 4.6 shows

$$m\big(M^{(2)} \uplus \rho(F) \setminus B\big) = m(M^{(2)}) + m\big(\rho(F)_{\le e+a_2-1} \uplus E\big) \preceq m(P) \qquad (16)$$

(see the second step in the figure on the previous page). We claim that

$$N = U^{(1)}_{\le e-1} \uplus P \uplus Q[+a_2] \subset U$$

satisfies the desired conditions. Indeed, by (15) and (16),

$$m(N) \succeq m\big(U^{(1)}_{\le e-1} \uplus M^{(2)} \uplus (\rho(F) \setminus B) \uplus (B \uplus M^{(\ge 3)})\big) = m(M)$$

(see figure on the previous page).

It remains to prove that N is a ladder set. If $\rho(F) \setminus B = \varnothing$ then $P = M^{(2)}$, and therefore N is a ladder set since $\#Q \ge \#M^{(\ge 3)}$. Suppose $\rho(F) \setminus B \ne \varnothing$. Recall that $\rho(F) \cap M^{(2)} = \varnothing$. Since

$$\#U^{(2)}_{\le e} \le \#M^{(2)} \le \#P = \#\rho(F)_{\le e+a_2-1} \uplus E \uplus M^{(2)} \le \#U^{(2)}_{\le e+a_2-1},$$

we have

$$U^{(2)}_{\le e} \subset P \subset U^{(2)}_{\le e+a_2-1}.$$

Then by Lemma 5.10 what we must prove is that

$$\#Q \ge \#U^{(\ge 3)}_{\le e+a_2-1}.$$

Since $\#S^{(i)}_k = \sum_{j=i}^{n} \#S^{(j)}_{k-1}$ for all $i > 0$ and $k > 0$, we have

$$\#U^{(3)}_k \ge \sum_{j=3}^{t} \#U^{(j)}_{k-1} = \#U^{(\ge 3)}_{k-1} \qquad (17)$$

for all $k > 0$. Since $\rho(F) \setminus B \ne \varnothing$, $\#B = \#\rho(F)_{e+a_2} \setminus A \ge \#U^{(2)}_{e+a_2} - \#U^{(2)}_{d+1}$. Thus

$$\#B \ge \#U^{(2)}_{e+a_2} - \#U^{(2)}_{d+1} = \# \biguplus_{j=d+2}^{e+a_2} U^{(3)}_{j+a_3} \ge \# \biguplus_{j=d+2}^{e+a_2} U^{(3)}_{j} \ge \sum_{j=d+1}^{e+a_2-1} \#U^{(\ge 3)}_j,$$

(we use (17) for the last step) and therefore

$$\#Q = \#M^{(\ge 3)} + \#B \ge \#U^{(\ge 3)}_{\le d} + \sum_{d+1}^{e+a_2-1} U^{(\ge 3)}_j \ge \#U^{(\ge 3)}_{\le e+a_2-1}$$

as desired.

Subcase 2-2. Suppose $\#B + \#M^{(\ge 3)} \ge \#U^{(2)}_{e+a_2}$.

Lemma 6.11. $\rho(F) \not\supset \biguplus_{j=d+2}^{e+a_2} U^{(2)}_j$.

Proof. Suppose to the contrary that $\rho(F) \supset \biguplus_{j=d+2}^{e+a_2} U_j^{(2)}$. Then

$$\#\rho(F) \setminus B = \#\big(\rho(F) \setminus (A \uplus B)\big) \uplus E = \# \biguplus_{j=d+1}^{e+a_2-1} U_j^{(2)}$$

by the choice of E. Then $\#(\rho(F) \setminus B) \uplus M^{(2)} = \#U^{(2)}_{\leq e+a_2-1}$ and

$$\#M = \#U^{(1)}_{\leq e-1} \uplus \rho(F) \uplus M^{(2)} \uplus M^{(\geq 3)} \geq \#U^{(1)}_{\leq e-1} + \#U^{(2)}_{\leq e+a_2-1} + \#U^{(2)}_{e+a_2} = \#U^{(1)}_{\leq e},$$

where we use the assumption $\#B + \#M^{(\geq 3)} \geq \#U^{(2)}_{e+a_2}$ for the second step. However, this contradicts (13). □

The above lemma says that $e + a_2 \geq d + 2$ and $\rho(F)_{d+1} = \varnothing$. Thus B does not contain any monomial $u\boldsymbol{e}_2$ such that u is divisible by $x_2^{(e+a_2)-(d+1)}$. Hence

$$\rho(B) \subset \biguplus_{j=d+2+a_3}^{e+a_2+a_3} U_j^{(3)}. \tag{18}$$

Since $M^{(3)}_{d+1} \neq U^{(3)}_{d+1}$, by Lemma 5.15,

$$\#M^{(\geq 3)} < \#U^{(3)}_{\leq d+2}.$$

Lemma 6.12. $a_3 = 0.$

Proof. If $a_3 > 0$ then

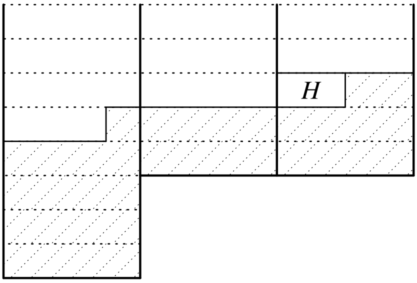
$$\#B + \#M^{(\geq 3)} < \# \biguplus_{j=d+2+a_3}^{e+a_2+a_3} U_j^{(3)} + \#U^{(3)}_{\leq d+2} \leq U^{(3)}_{\leq e+a_2+a_3} = \#U^{(2)}_{e+a_2},$$

which contradicts the assumption of Subcase 2-2. □

Let

$$H = \left\{h \in U^{(\geq 3)}_{d+1} : h \notin M^{(\geq 3)}\right\}$$

(see figure).



M

By Lemma 5.15,

$$\#H + \#M^{(\geq 3)} < \#U^{(3)}_{\leq d+2}.$$

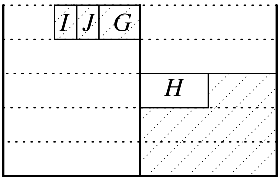
Since $a_3 = 0$, by the assumption of Subcase 2-2,

$$\#B \geq \#U^{(2)}_{e+a_2} - \#M^{(\geq 3)} = \#U^{(3)}_{\leq e+a_2} - \#M^{(\geq 3)} > \#H + \# \biguplus_{j=d+3}^{e+a_2} U^{(3)}_j.$$

Let

$$B = I \uplus J \uplus G,$$

where I is the set of lex-largest $\#H$ monomials in B and G is the revlex set with $\rho(G) = \biguplus_{j=d+3}^{e+a_2} U^{(3)}_j$ (see figure):



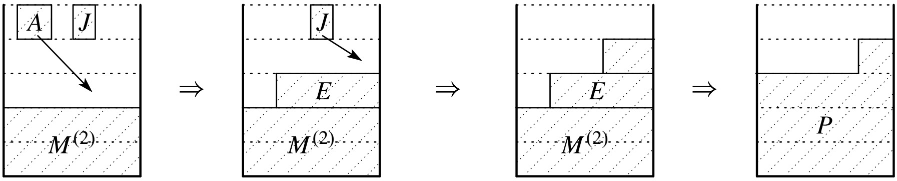
$B \uplus M^{(\geq 3)}$

Since $a_3 = 0$, (18) says $\rho(B) \subset \biguplus_{j=d+2}^{e+a_2} U^{(2)}_j$. Hence $\rho(I) \subset U^{(3)}_{d+2}$. Let $C \subset U^{(3)}_{d+2}$ be the lex set in $U^{(3)}_{d+2}$ with $\#C = \#H$. If we regard $U^{(\geq 3)}$ as a universal lex ideal in $K[x_3, \ldots, x_n]$, then H and C are lex sets in $K[x_3, \ldots, x_n]$ with the same cardinality. Hence $C = x_3 H$. Then, by the interval lemma,

$$m(H) = m(C) \succeq m(\rho(I)) = m(I). \tag{19}$$

Let $P \subset U^{(2)}$ be the super-revlex set with $\#P = \#A + \#J + \#M^{(2)}$. By the choice of G, G is the set of all monomials $u\boldsymbol{e}_2 \in \rho(F)$ such that u is not divisible by $x_2^{e+a_2-(d+2)}$. Also, since B does not contain any monomial $u\boldsymbol{e}_2$ such that u is divisible by $x_2^{e+a_2-(d+1)}$, any monomial in J is divisible by $x_2^{e+a_2-(d+2)}\boldsymbol{e}_2$. Then $x_2^{-(e+a_2)+d+2} J \subset U^{(2)}_{d+2}$ is a revlex set. Since $M^{(2)} \uplus E \uplus (x_2^{-(e+a_2)+(d+2)} J)$ is revlex, we have

$$m(P) \succeq m\big(M^{(2)} \uplus E \uplus x_2^{-(e+a_2)+(d+1)} J\big) = m\big(M^{(2)} \uplus A \uplus J\big). \tag{20}$$



$M^{(2)} \uplus A \uplus J$ $\quad$ $M^{(2)} \uplus E \uplus J$ $\quad$ $M^{(2)} \uplus E \uplus x_2^{-(e+a_2)+d+2} J$ $\quad$ P

Let

$$Q = \rho(F) \setminus (A \uplus B) = \rho(F)_{\le e+a_2-1}.$$

Subcase 2-2-a. Suppose that $\#P + \#Q \le \#U^{(2)}_{\le e+a_2-1}$. Let $R \subset U^{(2)}$ be the super-revlex set with $\#R = \#P + \#Q$. Then since Q is an upper revlex set of degree $e+a_2-1$, by Corollary 4.5 and (20)

$$R \gg P \uplus Q \gg M^{(2)} \uplus A \uplus J \uplus Q. \tag{21}$$

On the other hand, by Lemma 5.15,

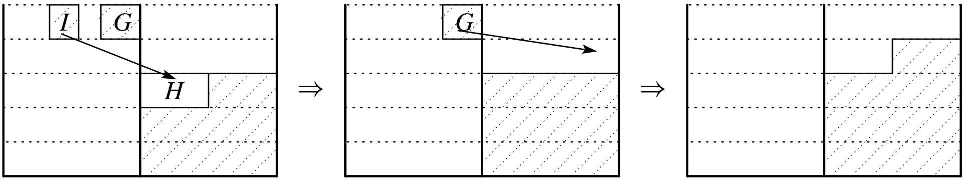
$$\#H + \#M^{(\ge 3)} < \#U^{(3)}_{\le d+2}.$$

Then since $\rho(G) = \biguplus_{j=d+3}^{e+a_2} U^{(3)}_j$,

$$\#I \uplus G \uplus M^{(\ge 3)} = \#G \uplus H \uplus M^{(\ge 3)} < \#U^{(3)}_{\le e+a_2} = \#U^{(2)}_{e+a_2}.$$

Let $U' = U^{(2)} \biguplus_{i=3}^{t} U^{(i)}[-a_2]$. Observe that $M^{(3)}[-a_2] \supset U'^{(\ge 3)}_{\le e+a_2}$. Then Corollary 5.13 and (19) say that there exists an extremal set $Z \subset U^{(\ge 3)}[-a_2]$ such that

$$Z \gg G \uplus H \uplus \big(M^{(\ge 3)}[-a_2]\big) \gg G \uplus I \uplus M^{(\ge 3)}. \tag{22}$$



$I \uplus G \uplus M^{(\ge 3)}$ $\quad$ $G \uplus H \uplus M^{(\ge 3)}$ $\quad$ $Z[+a_2]$

We claim that

$$N = U^{(1)}_{\le e-1} \uplus R \uplus Z[+a_2]$$

satisfies the desired conditions. Indeed, by (21) and (22),

$$\begin{aligned} N &\gg U^{(1)}_{\le e-1} \uplus (M^{(2)} \uplus A \uplus J \uplus Q) \uplus G \uplus I \uplus M^{(\ge 3)} \\ &\gg U^{(1)}_{\le e-1} \uplus F \uplus M^{(2)} \uplus M^{(\ge 3)} = M. \end{aligned}$$

(We use $\rho(F) = A \uplus I \uplus J \uplus G \uplus Q$ and $m(F) = m(\rho(F))$ for the second step.)

It remains to prove that N is a ladder set. Since $U^{(2)}_{\le d} \subset R \subset U^{(2)}_{\le e+a_2-1}$ it is enough to prove that $Z[+a_2] \supset U^{(\ge 3)}_{\le e+a_2-1}$. Since $\rho(G) = \biguplus_{j=d+3}^{e+a_2} U^{(3)}_j$,

$$\#Z = \#(H \uplus M^{(\ge 3)} \uplus G) \ge \#U^{(\ge 3)}_{\le d+1} \uplus \biguplus_{j=d+3}^{e+a_2} U^{(3)}_j \ge \#U^{(\ge 3)}_{\le e+a_2-1}.$$

(We use $\#U_j^{(3)} \ge \#U_{j-1}^{(\ge 3)}$ for the last step.) Then $Z[+a_2] \supset U^{(\ge 3)}_{\le e+a_2-1}$ by Lemma 5.10 as desired.

Subcase 2-2-b. Suppose that $\#P + \#Q > \#U^{(2)}_{\le e+a_2-1}$. Note that

$$\#P + \#Q + \#I + \#G = \#F + \#M^{(2)}.$$

Then $\#M^{(2)} \uplus F > \#U^{(2)}_{\le e+a_2-1}$. Let R be the super-revlex set with $\#R = \#M^{(2)} + \#F$. Then $\#R = \#M^{(2)} + \#F \le \#U^{(2)}_{\le e+a_2}$ by (14). Since $\#R \ge \#P + \#Q > U^{(2)}_{\le e+a_2-1}$, there exists a revlex set $B' \subset U^{(2)}_{e+a_2}$ such that

$$R = U^{(2)}_{\le e+a_2-1} \uplus B'.$$

Also by Corollary 4.5,

$$B' \uplus U^{(2)}_{\le e+a_2-1} = R \gg M^{(2)} \uplus \rho(F). \tag{23}$$

Since $\#F + \#M^{(\ge 2)} < \#U^{(2)}_{\le e+a_2}$, we have $\#B' + \#M^{(\ge 3)} < \#U^{(2)}_{e+a_2}$. Then by Corollary 5.13 there exists the extremal set $Z \subset U^{(\ge 3)}[-a_2]$ such that

$$B' \uplus (M^{(\ge 3)}[-a_2]) \ll Z. \tag{24}$$

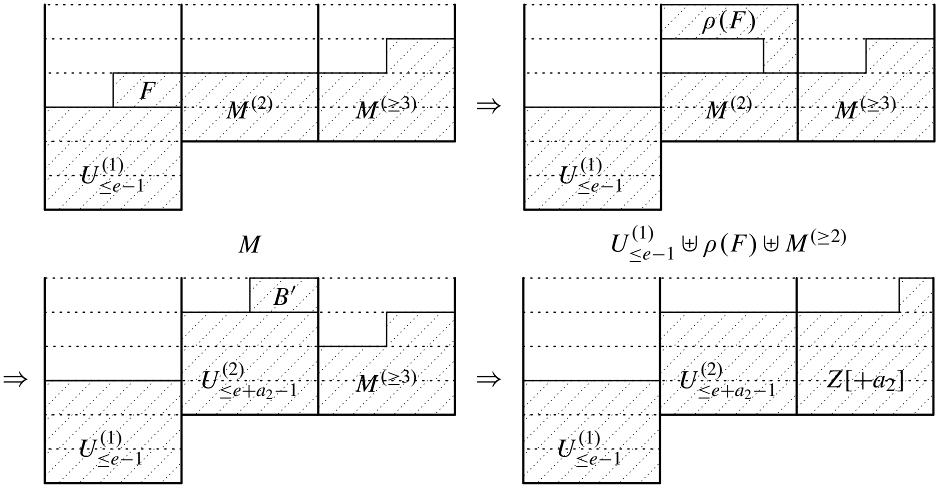
We claim that

$$N = U^{(1)}_{\le e-1} \uplus U^{(2)}_{\le e+a_2-1} \uplus Z[+a_2]$$

satisfies the desired conditions.

By (23) and (24),

$$N \gg U^{(1)}_{\le e-1} \uplus U^{(2)}_{\le e+a_2-1} \uplus B' \uplus M^{(\ge 3)} \gg U^{(1)}_{\le e-1} \uplus F \uplus M^{(2)} \uplus M^{(\ge 3)} = M.$$



M

$U^{(1)}_{\le e-1} \uplus \rho(F) \uplus M^{(\ge 2)}$

$U^{(1)}_{\le e-1} \uplus R \uplus M^{(\ge 3)}$

$N = U^{(1)}_{\le e-1} \uplus U^{(2)}_{\le e+a_2-1} \uplus Z[+a_2]$

It remains to prove that N is a ladder set. What we must prove is:

$$Z[+a_2] \supset U^{(\geq 3)}_{\leq e+a_2-1}.$$

By the assumption of Subcase 2-2-b,

$$\#M^{(2)} + \#F - \#(I \uplus G) = \#Q + \#P > \#U^{(2)}_{\leq e+a_2-1}.$$

Then

$$\#B' = \#M^{(2)} + \#F - \#U^{(2)}_{\leq e+a_2-1} > \#I \uplus G.$$

Then in the same way as the computation of $\#Z$ in Subcase 2-2-a, we have

$$\#Z = \#M^{(\geq 3)} \uplus B' \geq \#M^{(\geq 3)} \uplus (I \uplus G) \geq \#U^{(\geq 3)}_{\leq e+a_2-1}.$$

Then by Lemma 5.10, $Z[+a_2] \supset U^{(\geq 3)}_{\leq e+a_2-1}$ as desired.

7. Examples

In this section, we give some examples of saturated graded ideals which attain maximal Betti numbers for a fixed Hilbert polynomial. Observe that, by the decomposition given before Definition 3.7, the Hilbert polynomial of a proper universal lex ideal $I = (\delta_1, \delta_2, \dots, \delta_t)$ is given by

$$H_I(t) = \binom{t-b_1+n-1}{n-1} + \binom{t-b_2+n-2}{n-2} + \cdots + \binom{t-b_t+n-t}{n-t},$$

where $b_i = \deg \delta_i$ for $i = 1, 2, \dots, t$.

Example 7.1. Let $S = K[x_1, \dots, x_4]$ and $\bar{S} = K[x_1, \dots, x_3]$. Consider the ideal $I = (x_1^3, x_1^2x_2, x_1x_2^2, x_2^3, x_1^2x_3) \subset S$. Then

$$H_I(t) = \tfrac{1}{6}t^3 + t^2 - \tfrac{19}{6}t + 1 = \binom{t+2}{3} + \binom{t-4}{2} + \binom{t-9}{1}$$

and the proper universal lex ideal with the same Hilbert polynomial as I is

$$L = (x_1, x_2^6, x_2^5x_3^5).$$

Let

$$U = \operatorname{sat} \bar{L} = (\bar{L} : x_3^{\infty}) = (x_1, x_2^5) \subset \bar{S}$$

and $c = \dim_K U/\bar{L} = 5$. Then the extremal set $M \subset U$ with $\#M = 5$ is

$$M = x_1\{1, x_1, x_2, x_3\} \uplus x_2^5\{1\}.$$

Then the ideal in S generated by all monomials in $U \setminus M$ is

$$J = x_1(x_1^2, x_1x_2, x_1x_3, x_2^2, x_2x_3, x_3^2) + x_2^5(x_2, x_3) \subset S,$$

and J has the largest total Betti numbers among all saturated graded ideals in S having the same Hilbert polynomial as I.

Example 7.2. Let $S = K[x_1, \dots, x_5]$ and $\bar{S} = K[x_1, \dots, x_4]$. Consider the ideal $I = (x_1, x_2^2, x_2x_3^3, x_2x_3^2x_4^{15})$. Then I is a proper universal lex ideal. Let

$$U = \operatorname{sat} \bar{I} = (\bar{I} : x_4^\infty) = (x_1, x_2^2, x_2x_3^2) \subset \bar{S}$$

and $c = \dim U/\bar{I} = 15$. Then the extremal set $M \subset U$ with $\#M = 15$ is

$$M = x_1\{1, x_1, x_2, x_3, x_4, x_2x_3, x_2x_4, x_3^2, x_3x_4, x_4^2\} \uplus x_2^2\{1, x_2, x_3, x_4\} \uplus x_2x_3^2\{1\}.$$

Then the ideal in S generated by all monomials in $U \setminus M$ is

$$\begin{aligned} J =& x_1(x_1^2, x_1x_2, x_1x_3, x_1x_4, x_2^2, x_2x_3^2, x_2x_3x_4, x_2x_4^2, x_3^3, x_3^2x_4, x_3x_4^2, x_4^3) \\ &+ x_2^2(x_2^2, x_2x_3, x_2x_4, x_3^2, x_3x_4, x_4^2) + x_2x_3^2(x_3, x_4) \end{aligned}$$

and J has the largest total Betti numbers among all saturated graded ideals in S having the same Hilbert polynomial as I.

Finally, we give an explicit formula of the bounds in Theorem 1.1 for one special case. For positive integers a and d, let

$$a = \binom{a_d + d}{d} + \binom{a_{d-1} + d - 1}{d-1} + \cdots + \binom{a_t + t}{t}$$

be the d-th binomial representation of a. Thus $a_d, \dots, a_t$ are integers satisfying $a_d \geq a_{d-1} \geq \cdots \geq a_t \geq 0$ with $t \geq 1$. We define

$$a_{\langle d \rangle} = \binom{a_d - 1 + d}{d} + \binom{a_{d-1} - 1 + d - 1}{d-1} + \cdots + \binom{a_t - 1 + t}{t}.$$

Also, for $k = 0, 1, \dots, n-1$, we inductively define $a_{\langle d,k \rangle}$ by $a_{\langle d,0 \rangle} = a$ and $a_{\langle d,k \rangle} = (a_{\langle d,k-1 \rangle})_{\langle d \rangle}$ for $k \geq 1$, where $0_{\langle d \rangle} = 0$. The following formula is due to Valla [1994, Proposition 5]:

Lemma 7.3. *Let c be a positive integer, $M \subset S$ the super-revlex set with $\#M = c$, and let $J \subset S$ be the ideal generated by all monomials which are not in M. Let e be the unique integer such that $\binom{e-1+n}{n} \leq c < \binom{e+n}{n}$ and let $r = c - \binom{e-1+n}{n}$. Then, for $i \geq 1$, one has*

$$\beta_i^S(S/J) = \binom{e+i-2}{e-1}\binom{e+n-1}{i+e-1} + \sum_{k=1}^{n-1} \binom{k}{i-1} r_{\langle e,n-k \rangle}. \tag{25}$$

The right-hand side of (25) only depends on c, n, and i. Thus we denote it by $B_i(c, n)$.

Let b and c be positive integers. Consider the polynomial

$$p(t)=\binom{t-b+n-1}{n-1}+\cdots+\binom{t-b+2}{2}+\binom{t-b-c+2}{1}. \tag{26}$$

The universal lex ideal having the Hilbert polynomial (26) is

$$L=(x_1^b,\ldots,x_1^{b-1}x_{n-2},x_1^{b-1}x_{n-1}^c).$$

Then $U=\operatorname{sat}\bar{L}=(x_1^{b-1})$ and $\dim_K(\operatorname{sat}\bar{L})/\bar{L}=c$. In this case, an ideal which attains the bound in Theorem 1.1 was considered in Example 5.4. Let $M\subset\bar{S}=K[x_1,\ldots,x_{n-1}]$ be the super-revlex set with $\#M=c$ and let $J\subset S$ be the ideal generated by all monomials in $\bar{S}$ which are not in M. Then the ideal $L=x_1^{b-1}J$ attains the bound. In particular, by Lemma 7.3, we have:

Proposition 7.4. *Let $I\subset S$ be a saturated graded ideal whose Hilbert polynomial is of the form (26). Then $\beta_i^S(S/I)\le B_i(c,n-1)$ for all $i\ge 1$.*

Remark 7.5. When $b=1$, the above proposition is the result of Valla [1994] who considered the case when the Hilbert polynomial of S/I is constant. Indeed, if $P_{S/I}(t)$ is equal to a constant number c then

$$P_I(t)=\binom{t+n-1}{n-1}-c=\binom{t-1+n-1}{n-1}+\cdots+\binom{t-1+2}{2}+\binom{t-1-c+2}{1}.$$

References

[Bigatti 1993] A. M. Bigatti, "Upper bounds for the Betti numbers of a given Hilbert function", *Comm. Algebra* **21**:7 (1993), 2317–2334. MR 94c:13014 Zbl 0817.13007

[Bruns and Herzog 1998] W. Bruns and J. Herzog, *Cohen–Macaulay rings*, 2nd ed., Cambridge Studies in Advanced Mathematics **39**, Cambridge University Press, Cambridge, 1998. MR 95h:13020 Zbl 0909.13005

[Eliahou and Kervaire 1990] S. Eliahou and M. Kervaire, "Minimal resolutions of some monomial ideals", *J. Algebra* **129**:1 (1990), 1–25. MR 91b:13019 Zbl 0701.13006

[Elías et al. 1991] J. Elías, L. Robbiano, and G. Valla, "Number of generators of ideals", *Nagoya Math. J.* **123** (1991), 39–76. MR 92h:13023 Zbl 0714.13016

[Herzog et al. 1986] J. Herzog, M. E. Rossi, and G. Valla, "On the depth of the symmetric algebra", *Trans. Amer. Math. Soc.* **296**:2 (1986), 577–606. MR 87f:13009 Zbl 0604.13008

[Hulett 1993] H. A. Hulett, "Maximum Betti numbers of homogeneous ideals with a given Hilbert function", *Comm. Algebra* **21**:7 (1993), 2335–2350. MR 94c:13015 Zbl 0817.13006

[Iyengar and Pardue 1999] S. Iyengar and K. Pardue, "Maximal minimal resolutions", *J. Reine Angew. Math.* **512** (1999), 27–48. MR 2000d:13023 Zbl 0927.13019

[Macaulay 1927] F. S. Macaulay, "Some properties of enumeration in the theory of modular systems", *Proc. London Math. Soc.* (2) **26**:1 (1927), 531–555. MR 1576950 JFM 53.0104.01

[Murai and Hibi 2008] S. Murai and T. Hibi, "The depth of an ideal with a given Hilbert function", *Proc. Amer. Math. Soc.* **136**:5 (2008), 1533–1538. MR 2009b:13028 Zbl 1148.13006

[Pardue 1996] K. Pardue, "Deformation classes of graded modules and maximal Betti numbers", *Illinois J. Math.* **40**:4 (1996), 564–585. MR 97g:13029 Zbl 0903.13004

[Robbiano 1981] L. Robbiano, "Coni tangenti a singolarità razionali", Contribution 6 in *Curve algebriche: atti del Convegno di Geometria Algebrica* (Florence, 1981), edited by F. Gheradelli, Consiglio Nazionale delle Ricerche, Istituto di Analisi Globale ed Applicazioni, Florence, 1981.

[Sally 1978] J. D. Sally, *Numbers of generators of ideals in local rings*, Marcel Dekker, New York, 1978. MR 58 #5654 Zbl 0395.13010

[Valla 1994] G. Valla, "On the Betti numbers of perfect ideals", *Compositio Math.* **91**:3 (1994), 305–319. MR 95d:13012 Zbl 0815.14033

Communicated by David Eisenbud
Received 2010-10-12 Revised 2012-12-06 Accepted 2013-01-06

gcavigli@math.purdue.edu *Department of Mathematics, Purdue University,*
West Lafayette, IN 47901, United States

murai@yamaguchi-u.ac.jp *Department of Mathematical Science, Yamaguchi University,*
1677-1 Yoshida, Yamaguchi 753-8512, Japan

ALGEBRA AND NUMBER THEORY 7:5 (2013)
dx.doi.org/10.2140/ant.2013.7.1065

Comparing numerical dimensions

Brian Lehmann

The numerical dimension is a numerical measure of the positivity of a pseudoeffective divisor L. There are several proposed definitions of the numerical dimension due to Nakayama and Boucksom et al. We prove the equality of these notions and give several additional characterizations. We also prove some new properties of the numerical dimension.

1. Introduction

Suppose that X is a smooth complex projective variety and L is an effective divisor. An important principle in birational geometry is that the geometry of L is captured by the asymptotic behavior of the spaces $H^0(X, \mathcal{O}_X(mL))$ as m increases. When L is a big divisor, this asymptotic behavior has close ties to the cohomological and numerical properties of L. These connections have been applied profitably in many situations in birational geometry, most notably in the minimal model program.

However, when L is an effective divisor that is not big, these close relationships no longer hold. In order to understand the interplay between numerical and asymptotic properties, Kawamata [1985] defined the numerical dimension of a nef divisor. Nakayama [2004] and Boucksom et al. [2012] proposed several different extensions of this notion to pseudoeffective divisors. Our goal is to give a unifying framework for the numerical dimension by proving the equality of these definitions and giving other natural descriptions as well. We also describe some new properties of the numerical dimension. The crucial perspectives are the following:

(1) The numerical dimension measures the asymptotic behavior of L when it is perturbed by adding a small ample divisor ϵA.

(2) The numerical dimension measures the largest dimension of a subvariety $W \subset X$ such that L is positive along W. An important subtlety is that one should not simply consider $L|_W$ but should "remove" contributions of the base locus of L.

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship.

MSC2010: 14C20.

Keywords: divisor, numerical dimension.

Since some of the definitions used in the main theorem are rather technical, we simply give references here. We will describe in Section 1A the intuition behind the theorem. The notation $\mathbf{B}_{-}(L)$ denotes the diminished base locus defined in Section 2A, $\mathrm{vol}_{X|W}$ denotes the restricted volume defined in Section 2D, $P_\sigma(-)$ denotes the divisorial Zariski decomposition defined in Section 3, and $\langle - \rangle$ denotes the restricted positive product defined in Section 4.

Theorem 1.1. *Let X be a normal projective variety over $\mathbb{C}$, and let L be a pseudoeffective $\mathbb{R}$-Cartier $\mathbb{R}$-Weil divisor. In the following, A will denote some fixed sufficiently ample $\mathbb{Z}$-divisor, and W will range over all subvarieties of X not contained in $\mathbf{B}_{-}(L) \cup \mathrm{Supp}(L) \cup \mathrm{Sing}(X)$. The following quantities coincide*:

Perturbed growth condition:

(1) $\max\{k \in \mathbb{Z}_{\geq 0} \mid \limsup_{m\to\infty} h^0(X, \mathcal{O}_X(\lfloor mL \rfloor + A))/m^k > 0\}$.

Volume conditions:

(2) $\max\{k \in \mathbb{Z}_{\geq 0} \mid \exists C > 0$ *such that* $Ct^{n-k} < \mathrm{vol}(L + tA)$ *for all* $t > 0\}$.

(3) $\max\{\dim W \mid \lim_{\epsilon\to 0} \mathrm{vol}_{X|W}(L + \epsilon A) > 0\}$.

(4) $\max\{\dim W \mid \inf_{\phi: Y\to X} \mathrm{vol}_{\widetilde{W}}(P_\sigma(\phi^*L)|_{\widetilde{W}}) > 0\}$, *where ϕ varies over all birational maps such that no exceptional center contains W and $\widetilde{W}$ denotes the strict transform of W.*

Positive product conditions:

(5) $\max\{k \in \mathbb{Z}_{\geq 0} \mid \langle L^k \rangle \neq 0\}$.

(6) $\max\{\dim W \mid \langle L^{\dim W} \rangle_{X|W} > 0\}$.

Seshadri-type condition:

(7) $\min\{\dim W \mid \phi^*L - \epsilon E$ *is not pseudoeffective for any* $\epsilon > 0\}$, *where ϕ denotes the blow-up $\phi : Bl_W X \to X$ and E denotes the Cartier divisor on $Bl_W X$ such that $\mathcal{O}_{Bl_W X}(-E) = \phi^{-1}\mathcal{I}_W \cdot \mathcal{O}_{Bl_W X}$.* (By convention, if L is big, we interpret this expression as returning $\dim X$.)

This common quantity is known as the numerical dimension of L and is denoted $\nu(L)$. It only depends on the numerical class of L.

The definitions κ_σ and κ_ν of [Nakayama 2004, pp. 174 and 181] are listed as (1) and (7), respectively; the definition ν of [Boucksom et al. 2012] is listed as (5). When L is numerically effective, this definition agrees with the definition of [Kawamata 1985].

Remark 1.2. The numerical dimension also admits a natural interpretation with respect to separation of jets, reduced volumes, and the other invariants considered in [Ein et al. 2009].

The numerical dimension is natural from the viewpoint of birational geometry. It is established in [Nakayama 2004] that for a pseudoeffective divisor L,

- $0 \leq \nu(L) \leq \dim X$,
- $\nu(L) = \dim X$ if and only if L is big and $\nu(L) = 0$ if and only if $P_\sigma(L) \equiv 0$,
- $\kappa(L) \leq \nu(L)$, and
- if $\phi : Y \to X$ is a surjective morphism, then $\nu(\phi^* L) = \nu(L)$.

We prove two additional basic properties, answering a question of Nakayama:

- We have $\nu(L) = \nu(P_\sigma(L))$.
- Fix some sufficiently ample $\mathbb{Z}$-divisor A. Then there are positive constants C_1 and C_2 such that
$$C_1 m^{\nu(L)} < h^0(X, \mathbb{O}_X(\lfloor mL \rfloor + A)) < C_2 m^{\nu(L)}$$
for every sufficiently large m.

The properties of $\nu(L)$ will be discussed in more depth in Section 6.

1A. ***Intuitive description.*** We now turn to an intuitive description of several of the definitions in Theorem 1.1. Classically, one measures the positivity of a divisor using the rate of growth of sections of $H^0(X, \mathbb{O}_X(mL))$ as m increases. More precisely, the Iitaka dimension is defined as

$$\kappa(L) = \max \left\{ k \in \mathbb{Z}_{\geq 0} \;\middle|\; \limsup_{m \to \infty} \frac{h^0(X, \mathbb{O}_X(\lfloor mL \rfloor))}{m^k} > 0 \right\}.$$

(If $H^0(X, \mathbb{O}_X(\lfloor mL \rfloor)) = 0$ for every m, we set $\kappa(L) = -\infty$.) To obtain a numerical invariant, we must instead consider sections of $mL + A$ for some sufficiently ample divisor A. Thus, definition (1) indicates that $\nu(L)$ can be viewed as a numerical analogue of the Iitaka dimension.

Another way to calculate the positivity of L is to use intersection products. [Kawamata 1985] defined the numerical dimension of a numerically effective divisor L as

$$\nu(L) := \max\{ k \in \mathbb{Z}_{\geq 0} \mid L^k \cdot A^{n-k} \neq 0 \}$$

for some (thus any) ample divisor A. The naïve extension of this definition to pseudoeffective divisors does not work as the diminished base locus of L might contribute positively to this intersection and distort the measurement. The positive product of [Boucksom et al. 2012] gives a precise method of taking intersection products while discounting these contributions. Definition (5) shows that $\nu(L)$ can be defined as in [Kawamata 1985] by replacing the intersection product by the positive product.

A third way to measure the positivity of a divisor is the volume: if $n = \dim X$,

$$\operatorname{vol}(L) := \limsup_{m\to\infty} \frac{h^0(X, \mathcal{O}_X(mL))}{m^n/n!}.$$

Conceptually, we can view the volume as a loose analogue of the top self-intersection of L. While this latter quantity does not usually yield geometric information, the volume is a useful alternative that still shares many of the desirable properties of intersection products. It is shown in [Lazarsfeld and Mustaţă 2009; Boucksom et al. 2009] that vol is a differentiable function on the space of big $\mathbb{R}$-Cartier divisors. Definition (2) demonstrates that $\nu(L)$ controls the derivative of vol near L.

1B. ***Restricted numerical dimension.*** It is useful to study not only numerical invariants on X but also restricted versions that measure positivity along a subvariety V. We will define a restricted numerical dimension of L along a subvariety V of X. Just as in the nonrestricted case, the restricted numerical dimension should measure the maximal dimension of a very general subvariety $W \subset V$ such that the "positive restriction" of L is big along W.

Definition 1.3. Let X be a smooth variety, V a subvariety, and L a pseudoeffective $\mathbb{R}$-divisor such that $V \not\subset \mathbf{B}_-(L)$. Fix an ample divisor A. We define the restricted numerical dimension $\nu_{X|V}(L)$ to be

$$\nu_{X|V}(L) := \max\Bigl\{ \dim W \Bigm| \lim_{\epsilon\to 0} \operatorname{vol}_{X|W}(L+\epsilon A) > 0 \Bigr\},$$

where W ranges over smooth subvarieties of V not contained in $\mathbf{B}_-(L)$. The restricted numerical dimension is an invariant of the numerical class of L.

The restricted numerical dimension satisfies (slightly weaker) analogues of Theorems 1.1 and 6.7. For numerically effective divisors, we obtain nothing new because $\nu_{L|V}(L) = \nu_V(L|_V)$. Nevertheless, the restricted numerical dimension plays an important role in understanding the geometry of a pseudoeffective divisor L.

1C. ***Organization.*** The paper is organized as follows. Section 3 is devoted to the study of the divisorial Zariski decomposition, giving the technical background for the rest of the paper. Sections 4 and 5 prove some basic facts about the invariants of Theorem 1.1. We then turn to the proof of Theorem 1.1 in Section 6. Section 7 is devoted to a discussion of the restricted numerical dimension.

2. Preliminaries

All schemes will lie over the base field $\mathbb{C}$. A variety will always be an irreducible reduced projective scheme. The ambient variety X is assumed to be normal unless otherwise noted. The term "divisor" will always refer to an $\mathbb{R}$-Cartier $\mathbb{R}$-Weil divisor. Let $N^p(X)$ denote the $\mathbb{R}$-vector space of codimension-p cycles quotiented out by those numerically equivalent to 0, and $CD(X)$ will denote the $\mathbb{R}$-vector space of Cartier divisors quotiented out by those that have degree 0 along every irreducible curve.

2A. ***Base loci.*** Let L be a pseudoeffective divisor. The $\mathbb{R}$-stable base locus of L is defined to be

$$\mathbf{B}_{\mathbb{R}}(L) := \bigcap \{\mathrm{Supp}(D) \mid D \geq 0 \text{ and } D \sim_{\mathbb{R}} L\}.$$

When L is not $\mathbb{R}$-linearly equivalent to an effective divisor, we use the convention that $\mathbf{B}_{\mathbb{R}}(L) = X$. The $\mathbb{R}$-stable base locus is always a Zariski-closed subset of X; we do not associate any scheme structure to it.

We obtain a much better behaved invariant by perturbing by an ample divisor. This approach to invariants was first considered in [Nakamaye 2000] and was studied systematically in [Ein et al. 2006].

Definition 2.1. Let L be a pseudoeffective divisor. The augmented base locus of L is

$$\mathbf{B}_{+}(L) := \bigcap_{A \text{ ample}} \mathbf{B}_{\mathbb{R}}(L - A).$$

Note that $\mathbf{B}_{+}(L) \supset \mathbf{B}_{\mathbb{R}}(L)$. [Ein et al. 2006, Corollary 1.6] verifies that the augmented base locus is equal to $\mathbf{B}_{\mathbb{R}}(L - A)$ for any sufficiently small ample divisor A. Thus, $\mathbf{B}_{+}(L)$ is a Zariski-closed subset of X, and it only depends on the numerical class of L.

For the second variant, we add on a small ample divisor.

Definition 2.2. Let L be a pseudoeffective divisor. The diminished base locus of L is

$$\mathbf{B}_{-}(L) := \bigcup_{A \text{ ample}} \mathbf{B}_{\mathbb{R}}(L + A).$$

Remark 2.3. Although Nakayama [2004] uses a different definition, it is equivalent to ours by his Theorem V.1.3.

Proposition 1.15 of [Ein et al. 2006] checks that the diminished base locus only depends on the numerical class of L. Unlike the augmented base locus, the diminished base locus is probably not a Zariski-closed subset (although no examples are known of such pathological behavior). However, it is a countable union of closed subsets by the following theorem:

Theorem 2.4 [Nakayama 2004, Theorem V.1.3]. *Let X be a smooth variety, and let L be a pseudoeffective divisor. There is an ample divisor A such that*

$$\mathbf{B}_{-}(L) = \bigcup_{m} \mathrm{Bs}(\lceil mL \rceil + A),$$

where Bs *denotes the* (*set-theoretic*) *base locus.*

From [Nakayama 2004] we know $\mathbf{B}_-(L)$ is invariant under surjective morphisms.

Proposition 2.5. *Let $\phi : Y \to X$ be a surjective morphism from a normal variety Y onto a normal variety X. Suppose that L is a pseudoeffective divisor on X. Then we have an equality of sets*

$$\phi^{-1}\mathbf{B}_-(L) \cup \phi^{-1}\operatorname{Sing}(X) = \mathbf{B}_-(\phi^* L) \cup \phi^{-1}\operatorname{Sing}(X).$$

Proof. Fix an ample divisor H on Y and an ample divisor A on X. We have

$$\begin{aligned}
\phi^{-1}\mathbf{B}_-(L) &= \phi^{-1}\Bigl(\bigcup_m \mathbf{B}_{\mathbb{R}}\bigl(L+\tfrac{1}{m}A\bigr)\Bigr) && \text{by [Ein et al. 2006, Remark 1.20]}\\
&= \bigcup_m \mathbf{B}_{\mathbb{R}}\bigl(\phi^*\bigl(L+\tfrac{1}{m}A\bigr)\bigr)\\
&\supset \bigcup_m \mathbf{B}_{\mathbb{R}}\bigl(\phi^*\bigl(L+\tfrac{1}{m}A\bigr)+\tfrac{1}{m}H\bigr)\\
&= \mathbf{B}_-(\phi^* L) && \text{by [Ein et al. 2006, Remark 1.20].}
\end{aligned}$$

This proves the inclusion $\supset$. Furthermore, the same argument shows that it suffices to prove the reverse inclusion $\subset$ after replacing Y by any higher birational model.

We next reduce to the case where X and Y are smooth. Let $\psi : \widetilde{X} \to X$ denote a resolution that is an isomorphism away from $\operatorname{Sing}(X)$. Suppose that the closed point $\tilde{x} \notin \mathbf{B}_-(\phi^* L) \cup \phi^{-1}\operatorname{Sing}(X)$. Fix an ample divisor $\widetilde{A}$ on $\widetilde{X}$, and choose an ample divisor A on X so that $\phi^* A - \widetilde{A}$ is an effective divisor E. Since $\tilde{x}$ is not contained in the ψ-exceptional locus, we may also ensure that $\tilde{x} \notin \operatorname{Supp}(E)$. Then

$$\tilde{x} \notin \mathbf{B}_{\mathbb{R}}(\phi^*(L) + \epsilon H + \epsilon E) = \phi^{-1}\mathbf{B}_{\mathbb{R}}(\phi^*(L+\epsilon A))$$

for any $\epsilon > 0$, showing that

$$\psi^{-1}\mathbf{B}_-(L) \cup \psi^{-1}\operatorname{Sing}(X) = \mathbf{B}_-(\psi^* L) \cup \psi^{-1}\operatorname{Sing}(X).$$

As discussed earlier, we may verify the desired equality of sets by replacing Y by a smooth birational model that dominates $\widetilde{X}$. Thus, we have reduced to the case when both X and Y are smooth.

[Nakayama 2004, Lemmas III.2.3 and III.5.15] together show that for a smooth variety Z and a pseudoeffective divisor M on Z, a closed point $z \in Z$ is contained in $\mathbf{B}_-(M)$ if and only if, for every birational map $\psi : W \to Z$ from a smooth variety W and every ψ-exceptional divisor E with $\psi(E) = z$, we have $E \subset \mathbf{B}_-(\psi^* L)$. This immediately implies the inclusion $\subset$ when both X and Y are smooth. □

2B. ***V-pseudoeffective cone and V-big cone.*** The perturbed base loci can be used to describe when a divisor L sits in "general position" with respect to a subvariety V.

Definition 2.6. Suppose that $V \subset X$ is a subvariety. We define the V-pseudoeffective cone $\mathrm{Psef}_V(X)$ to be the cone in $CD(X)$ generated by classes of divisors L with $V \not\subset \mathbf{B}_-(L)$. We define the V-big cone $\mathrm{Big}_V(X)$ to be the cone generated by classes of divisors L with $V \not\subset \mathbf{B}_+(L)$.

It is easy to verify that $\mathrm{Psef}_V(X)$ is closed and $\mathrm{Big}_V(X)$ is its interior. Note also that $L|_V$ is pseudoeffective whenever L has numerical class in $\mathrm{Psef}_V(X)$. The following perspective will sometimes be useful:

Definition 2.7. Suppose that $V \subset X$ is a subvariety. If L is an effective divisor such that $\mathrm{Supp}(L) \not\supset V$, we say $L \geq_V 0$.

The relationship with the earlier criteria is given by a trivial lemma.

Lemma 2.8. *Suppose that $V \subset X$ is a subvariety. If L is a V-big divisor, then $L \sim_{\mathbb{R}} L'$ for some $L' \geq_V 0$.*

2C. *Admissible and V-birational models.* Suppose that X is a normal variety and V is a subvariety. In order to study how V-pseudoeffective divisors behave under birational pullbacks, we need to be careful about how V intersects the exceptional centers of the map. The most general situation is the following:

Definition 2.9. Let X be a normal variety and V a subvariety of X. Suppose that $\phi : Y \to X$ is a birational map and that W is a subvariety of Y such that the induced map $\phi|_W : W \to V$ is generically finite. We say that (Y, W) or $\phi : (Y, W) \to (X, V)$ is an admissible model for (X, V). When both Y and W are smooth, we say that (Y, W) is a smooth admissible model.

The disadvantage of admissible models is that in many circumstances we need to keep track of the degree of $\phi|_W$. Since we want to focus on the birational geometry of V, we will usually restrict ourselves to the following situation:

Definition 2.10. Let X be a normal variety and V a subvariety not contained in $\mathrm{Sing}(X)$. Suppose that $\phi : \widetilde{X} \to X$ is a birational map from a normal variety $\widetilde{X}$ such that V is not contained in any ϕ-exceptional center. Let $\widetilde{V}$ denote the strict transform of V. We say that $(\widetilde{X}, \widetilde{V})$ or $\phi : \widetilde{X} \to X$ is a V-birational model for (X, V). When both $\widetilde{X}$ and $\widetilde{V}$ are smooth, we say that $(\widetilde{X}, \widetilde{V})$ is a smooth V-birational model.

Suppose that V is a subvariety not contained in $\mathrm{Sing}(X)$ and $\phi : (Y, W) \to (X, V)$ is an admissible model. By Proposition 2.5, the pullback of a V-pseudoeffective divisor under ϕ is W-pseudoeffective. If ϕ is a V-birational model, then more is true.

Proposition 2.11. *Let X be a normal variety and V a subvariety not contained in* $\mathrm{Sing}(X)$. *Suppose that $\phi : \widetilde{X} \to X$ is a V-birational model. If L is a V-big divisor, then $\phi^* L$ is a $\widetilde{V}$-big divisor.*

Proof. The V-pseudoeffectiveness of L implies that ϕ^*L is $\widetilde{V}$-pseudoeffective. By openness of the $\widetilde{V}$-big cone, it suffices to check that ϕ^*H is $\widetilde{V}$-big for an ample divisor H on X. Let $\psi : \widetilde{Y} \to \widetilde{X}$ be a smooth model such that ψ is an isomorphism away from $\mathrm{Sing}(\widetilde{X})$. Note that for some sufficiently small ϵ,

$$\begin{aligned}\psi^{-1}\mathbf{B}_+(\phi^*H) &= \psi^{-1}\mathbf{B}_{\mathbb{R}}((1-\epsilon)\phi^*H) \quad \text{by [Ein et al. 2006, Corollary 1.6]}\\ &= \mathbf{B}_{\mathbb{R}}((1-\epsilon)\psi^*\phi^*H)\\ &\subset \mathbf{B}_+(\psi^*\phi^*H).\end{aligned}$$

But clearly $\mathbf{B}_+(\psi^*\phi^*H)$ is contained in the $(\phi \circ \psi)$-exceptional locus. Thus, $\mathbf{B}_+(\phi^*H)$ is contained inside the union of the ϕ-exceptional locus and $\mathrm{Sing}(\widetilde{X})$. In particular, it does not contain $\widetilde{V}$. □

2D. ***Restricted volume.*** Just as the volume measures the asymptotic rate of growth of sections, the restricted volume measures the rate of growth of restrictions of sections to a subvariety V. This notion originated in the work of Hacon–McKernan and Takayama and is systematically developed in [Ein et al. 2009].

Definition 2.12. Suppose that X is a normal variety, V is a d-dimensional subvariety of X, and L is a divisor. We define

$$H^0(X|V, \mathcal{O}_X(\lfloor L \rfloor)) := \mathrm{Im}\big(H^0(X, \mathcal{O}_X(\lfloor mL \rfloor)) \to H^0(V, \mathcal{O}_V(\lfloor mL \rfloor))\big)$$

and $h^0(X|V, \mathcal{O}_X(\lfloor L \rfloor))$ to be the dimension of this space. We then define the restricted volume $\mathrm{vol}_{X|V}(L)$ to be

$$\mathrm{vol}_{X|V}(L) := \limsup_{m\to\infty} \frac{h^0(X|V, \mathcal{O}_X(\lfloor mL \rfloor))}{m^d/d!}.$$

Remark 2.13. Although this definition of $\mathrm{vol}_{X|V}$ is formulated differently from that of [Ein et al. 2009], the two definitions agree (whenever the restricted volume is defined in [Ein et al. 2009]). An elementary argument proves that $\mathrm{vol}_{X|V}$ is homogeneous of degree d so that Definition 2.12 agrees with the definition in [Ein et al. 2009] for $\mathbb{Q}$-divisors. In particular, $\mathrm{vol}_{X|V}$ is a continuous function on the space of V-big $\mathbb{Q}$-divisors. Using this fact, one readily checks that $\mathrm{vol}_{X|V}$ is continuous on the set of V-big $\mathbb{R}$-divisors by perturbing by ample divisors and thus coincides with the definition of [Ein et al. 2009].

As with the other quantities we consider, the restricted volume is a numerical and birational invariant. More precisely, [Ein et al. 2009, Theorem A] shows that if L and L' are numerically equivalent V-big divisors, then $\mathrm{vol}_{X|V}(L) = \mathrm{vol}_{X|V}(L')$. Furthermore, [Ein et al. 2009, Proposition 2.4] proves that the restricted volume remains unchanged upon pulling back to an admissible model.

2E. ***Twisted linear series.*** It was observed by Iitaka that linear series of the form $|\lfloor mL \rfloor + A|$ play an important role in governing the numerical behavior of L. Due to the presence of the auxiliary divisor A, we call these "twisted" linear series. In this section, we recall the work of Nakayama [2004] analyzing the asymptotic behavior of twisted linear series.

Definition 2.14. Let X be a normal variety, L a pseudoeffective $\mathbb{R}$-divisor, and A any divisor. If $H^0(X, \mathbb{O}_X(\lfloor mL + A \rfloor))$ is nonzero for infinitely many values of m, we define

$$\kappa_\sigma(L;\, A) := \max\left\{ k \in \mathbb{Z}_{\geq 0} \;\middle|\; \limsup_{m\to\infty} \frac{h^0(X, \mathbb{O}_X(\lfloor mL + A \rfloor))}{m^k} > 0 \right\}.$$

Otherwise, we define $\kappa_\sigma(L;\, A) = -\infty$. The σ-dimension $\kappa_\sigma(X, L)$ is defined to be

$$\kappa_\sigma(L) := \max_A \{\kappa_\sigma(L;\, A)\}.$$

Note that this maximum will be computed by some sufficiently ample divisor A. Thus, we restrict our attention to the case when A is an ample $\mathbb{Z}$-divisor from now on.

Remark 2.15. As we increase m, the class of the divisor $\lceil mL \rceil - \lfloor mL \rfloor$ is bounded. Thus, if we replace $\lfloor - \rfloor$ by $\lceil - \rceil$ in the definition of $\kappa_\sigma(L)$, the result is unchanged as the difference can be absorbed by the divisor A.

Remark 2.16. Nakayama asks whether $\kappa_\sigma(L)$ coincides with

- $\kappa_\sigma^-(L)$, where we replace the lim sup by a lim inf, and
- $\kappa_\sigma^+(L)$, where we replace > 0 by $< \infty$.

The equality of these three notions is a consequence of Theorem 6.7(7).

Nakayama shows that κ_σ is a birational and numerical invariant. In fact, since κ_σ is one of the many equivalent definitions of the numerical dimension, it satisfies all of the properties of Theorem 6.7. The following key result shows that κ_σ is nonnegative for pseudoeffective divisors:

Proposition 2.17 [Nakayama 2004, Corollary V.1.4]. *Let X be a smooth variety of dimension n. Fix a big basepoint-free divisor B on X. Then a divisor L is pseudoeffective if and only if $h^0(X, \mathbb{O}_X(K_X + (n+2)B + \lceil mL \rceil)) > 0$ for every $m \geq 0$.*

Proof. Nakayama's Corollary V.1.4 is actually a similar statement for B very ample. We explain how to extend the argument to the case when B is big and basepoint-free. The main point is to show that there is an effective divisor $D \equiv (n+1)B + \lceil mL \rceil$ such that $\mathcal{J}(D)$ has an isolated point. There is an effective divisor $E \equiv B + \lceil mL \rceil$. Choose a general point x that does not lie in $\operatorname{Supp}(E) \cup \mathbf{B}_+(B)$. Let $B_1, \ldots, B_{n^2} \in |B|$ be irreducible smooth divisors going through x. Since B is big, by choosing the B_i sufficiently general, we may ensure the intersections of any collection of at most n

of them has the expected dimension. Thus, $D := \sum \frac{1}{n} B_i + E$ has multiplicity n at x and less than 1 in a neighborhood of x. By [Lazarsfeld 2004, Propositions 9.3.2 and 9.5.13], $\mathcal{J}(D)$ has an isolated point. The proof then proceeds as in [Nakayama 2004, Corollary V.1.4]. □

3. Divisorial Zariski decomposition

The divisorial Zariski decomposition is a higher dimension analogue of the classical Zariski decomposition on surfaces. It was introduced by Nakayama [2004] and by Boucksom [2004] in the analytic setting.

Definition 3.1. Let X be a smooth variety, and let L be a pseudoeffective divisor. Fix an ample divisor A on X. For any prime divisor Γ on X, we define

$$\sigma_\Gamma(L) = \lim_{\epsilon \to 0^+} \inf\{\, \mathrm{mult}_\Gamma(L') \mid L' \sim_{\mathbb{R}} L + \epsilon A \text{ and } L' \geq 0\,\}.$$

By Lemma III.1.5 of [Nakayama 2004], this is independent of the choice of A.

Lemma III.1.7 of the same reference says that for any pseudoeffective divisor L there are only finitely many prime divisors Γ with $\sigma_\Gamma(L) > 0$. Thus, we can define the following:

Definition 3.2. Let X be a smooth variety and L a pseudoeffective divisor. Define

$$N_\sigma(L) = \sum \sigma_\Gamma(L)\Gamma \quad \text{and} \quad P_\sigma(L) = L - N_\sigma(L).$$

The decomposition $L = N_\sigma(L) + P_\sigma(L)$ is called the divisorial Zariski decomposition of L.

The following proposition records the basic properties of the divisorial Zariski decomposition. The key point is that $P_\sigma(L)$ captures all of the interesting geometric information about L.

Proposition 3.3 [Nakayama 2004, Lemma III.1.4, Corollary III.1.9, Theorem V.1.3]. *Let X be a smooth variety and L a pseudoeffective divisor. Then*

(1) $N_\sigma(L)$ *depends only on the numerical class of* L,

(2) $N_\sigma(L) \geq 0$ *and* $\kappa(N_\sigma(L)) = 0$,

(3) $\mathrm{Supp}(N_\sigma(L))$ *is precisely the divisorial part of* $\mathbf{B}_-(L)$, *and*

(4) $H^0(X, \mathcal{O}_X(\lfloor m P_\sigma(L) \rfloor)) \to H^0(X, \mathcal{O}_X(\lfloor mL \rfloor))$ *is an isomorphism for all* $m \geq 0$.

Note that $N_\sigma(L) = 0$ if and only if $\mathbf{B}_-(L)$ has no divisorial components. This simple observation leads to a different perspective on the divisorial Zariski decomposition.

Definition 3.4. Let X be a smooth variety. The movable cone $\overline{Mov}^1(X) \subset CD(X)$ is the cone consisting of the classes of all pseudoeffective divisors L such that $\mathbf{B}_-(L)$ has no divisorial components.

The positive part $P_\sigma(L)$ of the divisorial Zariski decomposition can be understood as a "projection" of L onto the movable cone. We will need a slightly modified version of [Nakayama 2004, Proposition III.1.14] that takes into account a subvariety V.

Proposition 3.5. *Let X be smooth, V a subvariety, and L a V-pseudoeffective divisor. If M is a movable divisor, then $L \geq_V M$ if and only if $P_\sigma(L) \geq_V M$. Thus, $L - M$ is V-big or V-pseudoeffective if and only if $P_\sigma(L) - M$ is V-big or V-pseudoeffective, respectively.*

Proof. First suppose that $P_\sigma(L) \geq_V M$. Since L is V-pseudoeffective, no component of $N_\sigma(L)$ contains V. Thus, $L \geq_V M$. Conversely, suppose $L = M + E$ with $E \geq_V 0$. Since M is movable, $N_\sigma(L) \leq E$ by [Nakayama 2004, Proposition III.1.14]. Thus, $E - N_\sigma(L)$ is still effective and does not contain V in its support, showing that $P_\sigma(L) \geq_V M$.

Suppose now that $L - M$ is V-big. Choose an ample divisor A sufficiently small so that $L - M - A$ is V-big. By Lemma 2.8, there is some $D \sim_{\mathbb{R}} L - M - A$ such that $D \geq_V 0$. Applying the first step to $L - D$ shows that $P_\sigma(L) - L + D \equiv P_\sigma - M - A$ is V-pseudoeffective so that $P_\sigma(L) - M$ is V-big. The converse is straightforward. The analogous statement for V-pseudoeffectiveness follows by taking limits. □

3A. *Birational properties.* Although the divisorial Zariski decomposition is not a birational invariant, its birational behavior is relatively nice.

Proposition 3.6 [Nakayama 2004, Theorem III.5.16]. *Let $\phi : Y \to X$ be a birational map of smooth varieties, and let L be a pseudoeffective divisor on X. Then $N_\sigma(\phi^* L) - \phi^* N_\sigma(L)$ is effective and ϕ-exceptional.*

We say L admits a Zariski decomposition if there is a birational map $\phi : Y \to X$ from a smooth variety Y such that $P_\sigma(\phi^* L)$ is numerically effective. An important example due to Nakayama [2004, Section IV.2] shows that Zariski decompositions do not always exist. Nevertheless, there is a sense in which the positive part $P_\sigma(\phi^* L)$ becomes "more numerically effective" as we pass to higher models $\phi : Y \to X$. We will give two versions of this fact. In the first, we consider a V-big divisor L.

Proposition 3.7. *Let X be smooth, V a subvariety, and L a V-big divisor with $L \geq_V 0$. Then there is an effective divisor G so that for any sufficiently large m there is a model $\phi_m : \widetilde{X}_m \to X$ centered in $\mathbf{B}_+(L)$ and a big and numerically effective divisor N_m on $\widetilde{X}_m$ such that, with $\widetilde{V}_m$ denoting the strict transform of V on $\widetilde{X}_m$,*

$$N_m \leq_{\widetilde{V}_m} P_\sigma(\phi_m^* L) \leq_{\widetilde{V}_m} N_m + \tfrac{1}{m} \phi_m^* G.$$

The second version handles V-pseudoeffective divisors L. Although the statement is slightly more technical, the additional flexibility will be useful later on.

Proposition 3.8. *Let X be smooth, and let L be a pseudoeffective divisor. There are birational maps $\phi_m : \widetilde{X}_m \to X$ centered in $\mathbf{B}_-(L)$, an ample $\mathbb{Z}$-divisor A, and an effective divisor G satisfying the following condition. Suppose that V is a subvariety of X not contained in $\mathbf{B}_-(L)$. Then there is some $G_V \sim_{\mathbb{Q}} G$, and for every m, there is an effective divisor $D_m \sim \lceil mL \rceil + A$ and a big and numerically effective divisor M_{m,D_m} such that*

$$M_{m,D_m} \leq_{\widetilde{V}_m} P_\sigma(\phi_m^* D_m) \leq_{\widetilde{V}_m} M_{m,D_m} + \phi_m^* G_V,$$

where $\widetilde{V}_m$ denotes the strict transform of V on $\widetilde{X}_m$. We may furthermore assume that $A+D$ is ample for every D supported on Supp(L) *with coefficients in the set* $[-3, 3]$.

Proposition 3.7 is equivalent to the following comparison between asymptotic multiplier ideals and base loci. It is the analogue for $\mathbb{R}$-divisors of [Lazarsfeld 2004, Theorem 11.2.21]. Note that the theory of asymptotic multiplier ideals for big $\mathbb{R}$-divisors works just as in the case of $\mathbb{Q}$-divisors.

Lemma 3.9. *Let X be smooth, and let L be a big divisor on X. Fix a very ample $\mathbb{Z}$-divisor H on X such that $H+D$ is ample for every divisor D supported on* Supp(L) *with coefficients in the set* $[-3, 3]$. *Suppose that b is a sufficiently large positive integer so that $\lfloor bL \rfloor - (K_X + (n+1)H)$ is numerically equivalent to an effective $\mathbb{Z}$-divisor G. Then for every $m \geq b$, we have*

$$\mathcal{J}(\|mL\|) \otimes \mathcal{O}_X(-G) \subseteq \mathfrak{b}(|\lfloor mL \rfloor|).$$

Proof. The condition on H guarantees that for $m \geq b$, we can write

$$\begin{aligned}\lfloor mL \rfloor - G &\equiv \lfloor mL \rfloor - \lfloor bL \rfloor + K_X + (n+1)H \\ &\equiv ((m-b)L + A) + K_X + nH\end{aligned}$$

for some ample $\mathbb{R}$-divisor A. By applying Nadel vanishing and Castelnuovo–Mumford regularity, we find that

$$\mathcal{O}_X(\lfloor mL \rfloor) \otimes (\mathcal{O}_X(-G) \otimes \mathcal{J}(\|(m-b)L\|))$$

is globally generated for $m \geq b$. Then $\mathcal{J}(\|mL\|) \subset \mathcal{J}(\|(m-b)L\|)$. □

Proof of Proposition 3.7. Fix a very ample $\mathbb{Z}$-divisor H and an integer b as in Lemma 3.9. Thus, for any $m \geq b$, we have

$$\mathcal{J}(\|mL\|) \otimes \mathcal{O}_X(-G) \subseteq \mathfrak{b}(|\lfloor mL \rfloor|).$$

Recall that G can be chosen to be any effective $\mathbb{Z}$-divisor numerically equivalent to $\lfloor bL \rfloor - (K_X + (n+1)H)$. In particular, for b large enough, the base locus of $|G|$ is contained in $\mathbf{B}_+(L)$. Since this set does not contain V, we may ensure that $G \geq_V 0$.

Let $\phi_m : \widetilde{X}_m \to X$ be a resolution of the ideals $\mathfrak{b}(|\lfloor mL \rfloor|)$ and $\mathcal{J}(\|mL\|)$. Note that each ϕ_m is centered in $\mathbf{B}_+(L)$. We write $\phi_m^{-1}\mathfrak{b}(|\lfloor mL \rfloor|) \cdot \mathcal{O}_{Y_m} = \mathcal{O}_{Y_m}(-E_m)$ and $\phi_m^{-1}\mathcal{J}(\|mL\|) \cdot \mathcal{O}_{Y_m} = \mathcal{O}_{Y_m}(-F_m)$. We also define the big and numerically effective divisor $M_m := m\phi_m^* L - E_m - \phi_m^*\{mL\}$.

We know that $F_m + \phi_m^* G \geq E_m$ for all sufficiently large m. Let $M = \sum_{D \subset \mathrm{Supp}(L)} D$ be the sum of the components of $\mathrm{Supp}(L)$. Replacing G by $G + M$ allows us to take into account the fractional part of mL so that

$$F_m + \phi_m^* G \geq E_m + \phi_m^*\{mL\}.$$

Note that still $G \geq_V 0$. Since L is V-big, we know that $F_m \geq_{\widetilde{V}_m} 0$. Thus, the inequality in the equation above is a $\widetilde{V}_m$-inequality. Furthermore, $N_\sigma(m\phi_m^* L) \geq_{\widetilde{V}_m} F_m$ by [Ein et al. 2006, Proposition 2.5]. In all, we get $P_\sigma(m\phi_m^* L) \leq_{\widetilde{V}_m} M_m + \phi_m^* G$. Dividing by m and setting $N_m := M_m/m$ yields $P_\sigma(\phi_m^* L) \leq_{\widetilde{V}_m} N_m + \frac{1}{m}\phi_m^* G$. The inequality $N_m \leq_{\widetilde{V}_m} P_\sigma(\phi_m^* L)$ follows from Proposition 3.5 and the fact that $E_m + \phi_m^*\{mL\} \geq_{\widetilde{V}_m} 0$. □

Proof of Proposition 3.8. Fix very ample divisors H and G. By Theorem 2.4, there is an ample $\mathbb{Z}$-divisor A such that $\mathrm{Bs}(|\lceil mL \rceil + A|) \subset \mathbf{B}_-(L)$ for every positive integer m. We may assume that A is sufficiently ample so that

- $\lceil mL \rceil + A - K_X - (n+1)H$ is numerically equivalent to an effective divisor G_m for every $m > 0$ and
- $A + D$ is ample for every D supported on $\mathrm{Supp}(L)$ with coefficients in $[-3, 3]$.

Choose $D_m \sim \lceil mL \rceil + A$ so that $D_m \geq_V 0$. Note that we can apply Proposition 3.7 to D_m using G_m as our choice of effective divisor (since D_m is an integral divisor, there is no need to set conditions on the ampleness of H along the components of D_m). In particular, for every positive integer m, choose an $\epsilon_m > 0$ such that $G - \epsilon_m G_m$ is ample. Proposition 3.7 constructs a birational map $\phi_m : X_m \to X$ and big and numerically effective divisors M_{m,D_m} such that

$$M_{m,D_m} \leq_{\widetilde{V}_m} P_\sigma(\phi_m^* D_m) \leq_{\widetilde{V}_m} M_{m,D_m} + \epsilon_m \phi_m^* G_m.$$

Since $G - \epsilon_m G_m$ is V-big, we may replace G by some $\mathbb{Q}$-linearly equivalent divisor G_V so that

$$M_{m,D_m} \leq_{\widetilde{V}_m} P_\sigma(\phi_m^* D_m) \leq_{\widetilde{V}_m} M_{m,D_m} + \phi_m^* G_V. \qquad \square$$

4. The restricted positive product

Fujita realized that one can study the asymptotic behavior of sections of a big divisor L by analyzing the ample divisors sitting beneath L on higher birational models. The positive product (developed in [Boucksom 2004; Boucksom et al. 2012]) is a construction that encapsulates this approach to asymptotic behavior.

In this section, we discuss the restricted positive product $\langle L_1 \cdot L_2 \cdot \cdots \cdot L_k \rangle_{X|V}$ of Boucksom, Favre, and Jonsson [Boucksom et al. 2009]. Unlike the usual intersection product $L_1 \cdot L_2 \cdot \cdots \cdot L_k \cdot V$, the restricted positive product throws away the contributions of the base loci of the L_i. The result is a numerical equivalence class of cycles on V that gives a more precise measure of the positivity of the L_i along V.

4A. ***Definition and basic properties.*** We start by reviewing the construction of the restricted positive product in [Boucksom et al. 2009]. Throughout, we will use the intersection product of [Fulton 1984]. We will use the following notation:

Definition 4.1. Let X be a normal variety. Suppose that V is a subvariety of X and that $[L] \in CD(X)$. We will let $[L]|_V$ denote the image under the restriction map $CD^1(X) \to CD^1(V)$.

Note that if L is a divisor such that $\mathrm{Supp}(L) \not\supset V$, then $[L|_V] = [L]|_V$.

Definition 4.2. Let X be a normal variety of dimension n. Suppose that K and K' are two classes in $N^k(X)$. We write $K \succeq K'$ if $K - K'$ is contained in the closure of the cone generated by effective cycles of dimension $n - k$.

Lemma 4.3 [ibid., Proposition 2.3, Definition 4.4]. *Let X be a smooth variety and V a subvariety of X. Suppose that $N_1, \dots, N_k$ and $N'_1, \dots, N'_k$ are numerically effective divisors on X satisfying $N_i \geq_V N'_i$. Then*

$$N_1 \cdot \cdots \cdot N_k \cdot V \succeq N'_1 \cdot \cdots \cdot N'_k \cdot V.$$

Theorem 4.4 [ibid., Lemmas 2.6 and 2.7]. *Let X be a normal variety, V a subvariety not contained in* $\mathrm{Sing}(X)$, *and $L_1, \dots, L_k$ V-big divisors. Consider the classes*

$$\phi_*(N_1 \cdot N_2 \cdot \cdots \cdot N_k \cdot \widetilde{V}) \in N^k(V),$$

where $\phi : (\widetilde{X}, \widetilde{V}) \to (X, V)$ varies over all smooth V-birational models, the N_i are numerically effective, and $E_i := \phi^* L_i - N_i$ is a $\mathbb{Q}$-divisor satisfying $E_i \geq_{\widetilde{V}} 0$. These classes form a directed set under the relation $\preceq$ and admit a unique maximum under this relation.

Remark 4.5. Although [Boucksom et al. 2009] only proves this when V is a prime divisor in X, the proof works without change in this more general situation.

The restricted positive product is defined as the maximum class occurring in the previous theorem.

Definition 4.6. Let X be a normal variety, and let V be a subvariety not contained in $\mathrm{Sing}(X)$. Let $L_1, L_2, \dots, L_k$ be V-big divisors. We define the cycle

$$\langle L_1 \cdot L_2 \cdot \cdots \cdot L_k \rangle_{X|V} \in N^k(V)$$

as the maximum under $\preceq$ of $\phi_*(N_1 \cdot N_2 \cdot \cdots \cdot N_k \cdot \widetilde{V})$, where $\phi : (\widetilde{X}, \widetilde{V}) \to (X, V)$

runs over smooth V-birational models, the N_i are numerically effective and $E_i := \phi^*L_i - N_i$ is a $\mathbb{Q}$-divisor satisfying $E_i \geq_{\widetilde{V}} 0$. In the special case $X = V$, we write $\langle L_1 \cdot L_2 \cdot \cdots \cdot L_k \rangle_X$.

In fact, [ibid., Proposition 2.13] shows that the definition is unchanged if we allow E_i to be a V-pseudoeffective $\mathbb{R}$-divisor. The restricted positive product satisfies a number of important properties.

Proposition 4.7 [ibid., Proposition 4.6]. *As a function on the k-fold product of the V-big cone, the restricted positive product is continuous, symmetric, homogeneous of degree* 1, *and superadditive in each variable in the sense that*

$$\langle (L + L') \cdot L_2 \cdot \cdots \cdot L_k \rangle_{X|V} \succeq \langle L \cdot L_2 \cdot \cdots \cdot L_k \rangle_{X|V} + \langle L' \cdot L_2 \cdot \cdots \cdot L_k \rangle_{X|V}.$$

Since the product is continuous, this allows us to define a limit as we approach the pseudoeffective cone.

Definition 4.8. Let X be a normal variety, V a subvariety not contained in $\mathrm{Sing}(X)$, and $L_1, L_2, \ldots, L_k$ V-pseudoeffective divisors. For each i, fix a sequence of V-big divisors $B_{i,j}$ converging to 0 as j increases. We define the class

$$\langle L_1 \cdot L_2 \cdot \cdots \cdot L_k \rangle_{X|V} = \lim_{j\to\infty} \big\langle (L_1 + B_{1,j}) \cdot (L_2 + B_{2,j}) \cdot \cdots \cdot (L_k + B_{k,j}) \big\rangle_{X|V}.$$

Note that this limit is independent of the choice of the $B_{i,j}$ since by superadditivity any two choices are comparable under $\succeq$.

We will sometimes abuse notation by allowing the restricted positive product to take numerical classes as arguments rather than actual divisors. Since the restricted positive product is compatible under pushforward, we can extend the definition to arbitrarily singular varieties in the following way:

Definition 4.9. Let X be an integral variety, and let $\phi : Y \to X$ be a smooth model. For $[L_1], \ldots, [L_k] \in CD(X)$, we define

$$\langle [L_1] \cdot \cdots \cdot [L_k] \rangle_X := \phi_* \langle \phi^*[L_1] \cdot \cdots \cdot \phi^*[L_k] \rangle_Y.$$

Even though the restricted positive product is continuous along the V-big cone, it is only semicontinuous along the V-pseudoeffective boundary in the sense that if $L_{i,j}$ is a sequence of V-pseudoeffective divisors whose limit is L_i, then

$$\langle L_1 \cdot \cdots \cdot L_k \rangle_{X|V} \succeq \lim_{j\to\infty} \langle L_{1,j} \cdot \cdots \cdot L_{k,j} \rangle_{X|V}.$$

As noted in [Boucksom et al. 2009], it is most natural to consider the restricted positive product as the set of classes $\{\langle \phi^*L_1 \cdot \cdots \cdot \phi^*L_k \rangle_{\widetilde{X}|\widetilde{V}}\}$ on all smooth V-birational models $\phi : \widetilde{X} \to X$ or, in other words, as a class on the Riemann–Zariski space of V. Although we will not develop this principle systematically, this idea appears implicitly as some theorems will only hold upon taking a limit over all sufficiently high birational models.

Since the restricted positive product should be considered as a birational object, the class in $N^k(V)$ may not be closely related to the geometry of L and V. The class $\langle L_1 \cdot \cdots \cdot L_k \rangle_{X|V}$ seems to be most interesting in the following two situations:

Example 4.10. When X is smooth, $\langle L \rangle_X$ is the numerical class of $P_\sigma(L)$. It suffices to check this when L is big. Recall that for any birational map $\phi : Y \to X$ from a smooth variety Y, we have $\phi_* P_\sigma(\phi^* L) = P_\sigma(L)$. Thus, choosing an effective divisor G as in Proposition 3.7, the result of the proposition implies that for any $\epsilon > 0$, we have $\langle L \rangle_X \preceq [P_\sigma(L)] \preceq \langle L + \epsilon G \rangle_X$. Letting $\epsilon \to 0$ demonstrates the equality.

Example 4.11. Consider $\langle L_1 \cdot \cdots \cdot L_d \rangle_{X|V}$, where $d = \dim V$. Since the restricted positive product is compatible under pushforward, $\deg\langle \phi^* L_1 \cdot \cdots \cdot \phi^* L_d \rangle_{\widetilde{X}|\widetilde{V}}$ is independent of the choice of V-birational model $(\widetilde{X}, \widetilde{V})$ by the projection formula. In fact, we have the following:

Proposition 4.12 [Ein et al. 2009, Proposition 2.11, Theorem 2.13]. *Let X be a smooth variety, V a d-dimensional subvariety, and L a V-big divisor. Then* $\deg\langle L^d \rangle_{X|V} = \mathrm{vol}_{X|V}(L)$.

4B. ***Properties of the restricted positive product.*** In this section, we study the properties of the restricted positive product. The main goal of the section is to show that the restricted positive product can be interpreted as the usual intersection product of $P_\sigma(\phi^* L_i)$ if we take a limit over all birational models ϕ. The advantage of this viewpoint is that it gives us a natural interpretation of the restricted positive product along the boundary of the pseudoeffective cone.

We first show that the restricted positive product has a natural compatibility with the divisorial Zariski decomposition.

Proposition 4.13. *Let X be a smooth variety, V a subvariety, and $L_1, \dots, L_k$ V-pseudoeffective divisors. Then*

$$\langle L_1 \cdot \cdots \cdot L_k \rangle_{X|V} = \langle P_\sigma(L_1) \cdot \cdots \cdot P_\sigma(L_k) \rangle_{X|V}.$$

Proof. First suppose that the L_i are V-big. Since any numerically effective divisor is movable, Proposition 3.5 shows that for any of the N_i as in Definition 4.6, we have $P_\sigma(\phi^* L_i) \geq_{\widetilde{V}} N_i$. We also know that $N_\sigma(\phi^* L_i) \geq_{\widetilde{V}} \phi^* N_\sigma(L_i)$ since V is not contained in $\mathbf{B}_-(L_i)$. Combining the two inequalities yields

$$\phi^* P_\sigma(L_i) \geq_{\widetilde{V}} N_i.$$

Thus, the classes $\langle L_1 \cdot \cdots \cdot L_k \rangle_{X|V}$ and $\langle P_\sigma(L_1) \cdot \cdots \cdot P_\sigma(L_k) \rangle_{X|V}$ are computed by taking a maximum over the same sets, showing that they are equal.

Now suppose that the L_i are only V-pseudoeffective. Fix an ample divisor A on X. Note that

$$P_\sigma(L+\epsilon A)-P_\sigma(L)=\epsilon A+(N_\sigma(L)-N_\sigma(L+\epsilon A))$$

is V-big. As ϵ goes to 0, these V-big classes also converge to 0. Thus,

$$\langle P_\sigma(L_1)\cdot\cdots\cdot P_\sigma(L_k)\rangle_{X|V}=\lim_{\epsilon\to 0}\langle P_\sigma(L_1+\epsilon A)\cdot\cdots\cdot P_\sigma(L_k+\epsilon A)\rangle_{X|V}.$$

Applying the V-big case to the right-hand side finishes the proof. □

The following proposition compares the restricted positive product of the L_i along V with the positive product of the restrictions $L_i|_V$. The statement is proved in [Boucksom et al. 2009] only when the L_i are V-big, but the proposition extends to the V-pseudoeffective case by taking limits.

Proposition 4.14 [Boucksom et al. 2009, Remark 4.5]. *Let X be a smooth variety, V a subvariety, and $L_1,\dots,L_k$ V-pseudoeffective divisors. Then*

$$\langle L_1\cdot\cdots\cdot L_k\rangle_{X|V}\preceq\langle[L_1]|_V\cdot\cdots\cdot[L_k]|_V\rangle_V.$$

By combining Propositions 4.13 and 4.14, we obtain

$$\langle L_1\cdot\cdots\cdot L_k\rangle_{X|V}\preceq\phi_*\big\langle[P_\sigma(\phi^*L_1)]|_{\widetilde{V}}\cdot\cdots\cdot[P_\sigma(\phi^*L_k)]|_{\widetilde{V}}\big\rangle_{\widetilde{V}},$$

where $\phi:(\widetilde{X},\widetilde{V})\to(X,V)$ is any V-birational model. The main theorem of this section states that by taking a limit over all birational models, the right-hand side approaches the left.

Theorem 4.15. *Let X be a smooth variety, V a subvariety, and $L_1,\dots,L_k$ V-pseudoeffective divisors. Fix an ample divisor A. Then for any ϵ, there is some V-birational map $\phi:(\widetilde{X},\widetilde{V})\to(X,V)$ such that*

$$\phi_*\big\langle[P_\sigma(\phi^*L_1)]|_{\widetilde{V}}\cdot\cdots\cdot[P_\sigma(\phi^*L_k)]|_{\widetilde{V}}\big\rangle_{\widetilde{V}}\preceq\langle L_1\cdot\cdots\cdot L_k\rangle_{X|V}+\epsilon A^k\cdot V.$$

Proof. First suppose the L_i are V-big. By Lemma 2.8, we may replace the L_i by some $\mathbb{R}$-linearly equivalent divisors to ensure that $L_i\geq_V 0$. Proposition 3.7 then yields an effective divisor G_i such that for any m there is a V-birational model $\phi:\widetilde{X}_m\to X$ with

$$N_{m,i}\leq_{\widetilde{V}}P_\sigma(\phi_m^*L_i)\leq_{\widetilde{V}}N_{m,i}+\tfrac{1}{m}\phi_m^*G_i$$

for some numerically effective divisors $N_{m,i}$. Fix some ample divisor A on X such that $A-L_i$ and $A-G_i$ are ample for every i. By Lemma 4.3, there is some constant C such that

$$\phi_{m*}\big\langle[P_\sigma(\phi_m^*L_1)]|_{\widetilde{V_m}}\cdot\cdots\cdot[P_\sigma(\phi_m^*L_k)]|_{\widetilde{V_m}}\big\rangle_{\widetilde{V_m}}\preceq\phi_*(N_{m,1}\cdot\cdots\cdot N_{m,k}\cdot\widetilde{V})+\frac{C}{m}A^k\cdot V.$$

Now suppose that the L_i are only V-pseudoeffective. We first choose an ample divisor H so that

$$\langle (L_1 + H) \cdot \cdots \cdot (L_k + H)\rangle_{X|V} \preceq \langle L_1 \cdot \cdots \cdot L_k\rangle_{X|V} + \tfrac{\epsilon}{2} A^k \cdot V.$$

Construct a model ϕ by applying the V-big case to the $L_i + H$ and $\epsilon/2$. Since $P_\sigma(\phi^*(L_i + H)) - P_\sigma(\phi^* L)$ is $\widetilde{V}$-pseudoeffective, the conclusion follows. □

Corollary 4.16. *Let X be a smooth variety, and let $L_1, \dots, L_k$ be pseudoeffective divisors. There is a sequence of birational maps $\phi_m : X_m \to X$ centered in $\cup_i \mathbf{B}_-(L_i)$ such that for any subvariety V not contained in $\cup_i \mathbf{B}_-(L_i)$, we have*

$$\langle L_1 \cdot \cdots \cdot L_k\rangle_{X|V} = \lim_{m\to\infty} \phi_{m*}\big\langle [P_\sigma(\phi_m^* L_1)]|_{\widetilde{V}_m} \cdot \cdots \cdot [P_\sigma(\phi_m^* L_k)]|_{\widetilde{V}_m}\big\rangle_{\widetilde{V}_m}.$$

Proof. Fix a sequence of birational maps ϕ_m, an ample divisor A, and an effective divisor G as in Proposition 3.8 for each of the L_i simultaneously. The proposition constructs divisors $D_{m,i} \equiv \lceil mL_i\rceil + A$ and big and numerically effective divisors $M_{m,i,D_{m,i}}$ such that

$$M_{m,i,D_{m,i}} \leq_{\widetilde{V}_m} P_\sigma(\phi_m^* D_{m,i}) \leq_{\widetilde{V}_m} M_{m,i,D_{m,i}} + \phi_m^* G_V.$$

Just as in the previous proposition, we have

$$\begin{aligned}
&\lim_{m\to\infty} \frac{1}{m^k} \phi_{m*}(M_{m,1,D_{m,1}} \cdot \cdots \cdot M_{m,k,D_{m,k}} \cdot \widetilde{V}_m) \\
&\qquad \preceq \lim_{m\to\infty} \frac{1}{m^k} \langle D_{m,1} \cdot \cdots \cdot D_{m,k}\rangle_{X|V} \\
&\qquad \preceq \lim_{m\to\infty} \frac{1}{m^k} \phi_{m*}\big\langle [P_\sigma(\phi_m^* D_{m,1})]|_{\widetilde{V}_m} \cdot \cdots \cdot [P_\sigma(\phi_m^* D_{m,k})]|_{\widetilde{V}_m}\big\rangle_{\widetilde{V}_m} \\
&\qquad \preceq \lim_{m\to\infty} \frac{1}{m^k} \phi_{m*}\big((M_{m,1,D_{m,1}} + \phi_m^* G_V) \cdot \cdots \cdot (M_{m,k,D_{m,k}} + \phi_m^* G_V) \cdot \widetilde{V}_m\big).
\end{aligned}$$

Arguing as in the previous proof, we see that the leftmost and rightmost expressions converge as m increases. Recall that by our choice of A we have $\lceil mL_i\rceil + A - mL_i$ is V-big for every m. Thus,

$$\langle L_1 \cdot \cdots \cdot L_k\rangle_{X|V} = \lim_{m\to\infty} \big\langle \tfrac{1}{m} D_{m,1} \cdot \cdots \cdot \tfrac{1}{m} D_{m,k}\big\rangle_{X|V}$$

so that the sequence converges to the restricted positive product as desired. □

We extract a useful feature of the previous arguments as a definition.

Definition 4.17. Let X be a smooth variety, V a subvariety, and $L_1, \dots, L_k$ V-big divisors. Choose $L_i' \sim_{\mathbb{Q}} L_i$ satisfying $L_i' \geq_V 0$. Suppose that ϕ_m is a countable sequence of maps that satisfy the conclusion of Proposition 3.7 for every L_i' simultaneously. We say that the ϕ_m compute the restricted positive product of the L_i.

Note that for any finite set of subvarieties $V_1, \dots, V_r$, we can choose ϕ_m and N_m to simultaneously compute the restricted positive product for each V_j. The key property of Definition 4.17 is that only countably many maps are needed to compute the restricted positive product.

The restricted positive product reduces to the usual product for numerically effective divisors.

Lemma 4.18. *Let X be a smooth variety, V a subvariety, and $L_1, \dots, L_k$ V-pseudoeffective divisors.*

(1) *Suppose N is a numerically effective divisor. Then*

$$\langle L_1 \cdot L_2 \cdot \dots \cdot L_k \cdot N \rangle_{X|V} = \langle L_1 \cdot L_2 \cdot \dots \cdot L_k \rangle_{X|V} \cdot N|_V.$$

(2) *If H is a very general element of a basepoint-free linear system, then*

$$\langle L_1 \cdot L_2 \cdot \dots \cdot L_k \rangle_{X|V} \cdot H = \langle L_1 \cdot L_2 \cdot \dots \cdot L_k \rangle_{X|V \cap H}.$$

(3) *If $f : X \to Z$ is a morphism and F is a very general fiber, then*

$$\langle L_1 \cdot L_2 \cdot \dots \cdot L_k \rangle_{X|V} \cdot F = \langle L_1 \cdot L_2 \cdot \dots \cdot L_k \rangle_{X|V \cap F}.$$

Proof. For each of these properties, it is enough to check the case when the L_i are V-big.

The first property is shown in [Boucksom et al. 2009, Proposition 4.7]; one simply notes that for an ample divisor A the pullback ϕ^*A is already numerically effective so that one may take ϕ^*A to be the numerically effective divisor in Definition 4.6. By taking limits as A approaches N, we obtain the statement.

To show the second property, consider a countable set of smooth V-birational models $\phi_m : \widetilde{X}_m \to X$ that compute the restricted positive product. Choose H sufficiently general so that it does not contain any ϕ_m-exceptional center. Then the strict transform of $V \cap H$ is a cycle representing the class $\phi_m^* H \cdot \widetilde{V}$. Thus, we can identify the classes

$$\begin{aligned} \phi_{m*}(N_1 \cdot N_2 \cdot \dots \cdot N_k \cdot \widetilde{V}) \cdot H &= \phi_{m*}(N_1 \cdot N_2 \cdot \dots \cdot N_k \cdot \phi_m^* H \cdot \widetilde{V}) \\ &= \phi_{m*}(N_1 \cdot N_2 \cdot \dots \cdot N_k \cdot \widetilde{V \cap H}). \end{aligned}$$

The third property can be proved by a similar argument. One uses the second property inductively by pulling back very ample divisors from Z. □

Corollary 4.19. *Let X be a normal variety, V a subvariety not contained in* $\mathrm{Sing}(X)$, *and $L_1, \dots, L_k$ V-pseudoeffective divisors. Suppose $\phi : (\widetilde{X}, \widetilde{V}) \to (X, V)$ is a smooth V-birational model. If $\langle \phi^* L_1 \cdot \dots \cdot \phi^* L_k \rangle_{\widetilde{X}|\widetilde{V}} \neq 0$, then $\langle L_1 \cdot \dots \cdot L_k \rangle_{X|V} \neq 0$.*

Proof. Let A be an ample divisor on $\widetilde{X}$, and let H be an ample divisor on X such that $\phi^*H \geq A$. Since ϕ is V-birational, we may ensure that $\operatorname{Supp}(\phi^*H - A)$ does not contain $\widetilde{V}$. Setting $d = \dim V$, we have

$$\begin{aligned}\langle L_1 \cdot \cdots \cdot L_k\rangle_{X|V} \cdot H^{d-k} &= \langle \phi^*L_1 \cdot \cdots \cdot \phi^*L_k\rangle_{\widetilde{X}|\widetilde{V}} \cdot \phi^*H^{d-k} \\ &= \langle \phi^*L_1 \cdot \cdots \cdot \phi^*L_k \cdot \phi^*H^{d-k}\rangle_{\widetilde{X}|\widetilde{V}} \\ &\geq \langle \phi^*L_1 \cdot \cdots \cdot \phi^*L_k \cdot A^{d-k}\rangle_{\widetilde{X}|\widetilde{V}} \\ &= \langle \phi^*L_1 \cdot \cdots \cdot \phi^*L_k\rangle_{\widetilde{X}|\widetilde{V}} \cdot A^{d-k} > 0. \qquad \square\end{aligned}$$

We next consider how the restricted positive product behaves when passing to an admissible model.

Proposition 4.20. *Let X be a smooth variety, V a subvariety, and $L_1, \dots, L_k$ V-pseudoeffective divisors. Suppose $f : (Y, W) \to (X, V)$ is an admissible model. Then*

$$f_*\langle f^*L_1 \cdot \cdots \cdot f^*L_k\rangle_{Y|W} = \deg(f|_W)\langle L_1 \cdot \cdots \cdot L_k\rangle_{X|V}.$$

Note that f^*L_i is W-pseudoeffective by Proposition 2.5.

Proof. It suffices to consider the case when the L_i are V-big. By Lemma 2.8, we may suppose that $L_i \geq_V 0$. Let $\phi_m : X_m \to X$ be a sequence of V-birational models that computes $\langle L_1 \cdot \cdots \cdot L_k\rangle_{X|V}$, and let $\psi_m : Y_m \to Y$ be a sequence of W-birational models that computes $\langle f^*L_1 \cdot \cdots \cdot f^*L_k\rangle_{Y|W}$. Since the natural map $\phi_m^{-1} \circ f \circ \psi_m$ is a morphism on the generic point of W, by passing to higher W-birational models, we may assume that Y_m admits a morphism $f_m : Y_m \to X_m$. Note that

$$f_m^*N_{i,m} \leq_{\widetilde{W}_m} P_\sigma(\psi_m^* f_m^* L_i) \leq_{\widetilde{W}_m} f_m^* P_\sigma(\phi_m^* L_i) \leq_{\widetilde{W}_m} f_m^* N_{i,m} + \tfrac{1}{m} f_m^* \phi_m^* G_i.$$

By construction, the pushforwards

$$\phi_{m*} f_{m*}(f_m^* N_{1,m} \cdot \cdots \cdot f_m^* N_{k,m} \cdot \widetilde{W}_m)$$

converge to $\deg(f|_W)\langle L_1 \cdot \cdots \cdot L_k\rangle_{X|V}$. The same is true for the terms on the right-hand side. Thus, $f_*\psi_{m*}\langle P_\sigma(\psi_m^* f^* L_1) \cdot \cdots \cdot P_\sigma(\psi_m^* f^* L_k)\rangle_{Y|\widetilde{W}_m}$ converges to the same thing, and Theorem 4.15 finishes the proof. □

It is worth pointing out that Proposition 4.20 does not contradict the invariance of $\operatorname{vol}_{X|V}(L)$ under passing to admissible models. Even if L is V-big, ϕ^*L will not be W-big when $\deg(f|_W) > 1$, so Proposition 4.12 does not apply to W.

Proposition 4.21. *Let X be a smooth variety, V a subvariety of dimension d, and L a V-pseudoeffective divisor. Suppose that* $\deg(\langle L^d\rangle_{X|V}) > 0$. *Then for a very general intersection of very ample divisors W of dimension d, we also have* $\deg(\langle L^d\rangle_{X|W}) > 0$.

Proof. Fix a sequence of maps $\phi_m : \widetilde{X}_m \to X$ for L as in Corollary 4.16. By choosing very ample divisors $H_1, \dots, H_{n-d}$ very general in their linear systems, we may ensure that no H_i contains any ϕ_m-exceptional center and the intersection $W = H_1 \cap \cdots \cap H_{n-d}$ is smooth of the expected dimension.

For each $i = 1, 2, \dots, n-d$, choose a positive integer c_i so that $\mathcal{I}_V(c_i H_i)$ is generated by global sections, and set $C = \prod_i c_i^{-1}$. Note that for any V-birational model $\phi : (Y, \widetilde{V}) \to (X, V)$, there are $D_i \in |c_i \phi^* H_i|$ such that each D_i has multiplicity at least 1 along $\widetilde{V}$ and $D_1 \cap \cdots \cap D_{n-k}$ has dimension k. In particular for ϕ_m,

$$\begin{aligned} [\widetilde{W}] &= C[\phi_m^* c_1 H_1] \cap [\phi_m^* c_2 H_2] \cap \cdots \cap [\phi_m^* c_{n-d} H_{n-d}] \\ &\succeq C[\widetilde{V}], \end{aligned}$$

where $\widetilde{W}$ and $\widetilde{V}$ denote the strict transforms of W and V on $\widetilde{X}_m$. In particular, for any numerically effective divisor N on $\widetilde{X}_m$, we have $N^d \cdot \widetilde{W} \geq N^d \cdot \widetilde{V}$, and the conclusion follows. □

5. Nakayama constants

Suppose that L is an ample divisor and V is a subvariety in X. Let $\phi : Y \to X$ be a smooth resolution of the ideal $\mathcal{I}_V$, and define the divisor E by the equation $\mathcal{O}_Y(-E) = \phi^{-1}\mathcal{I}_V \cdot \mathcal{O}_Y$. The Seshadri constant

$$\varepsilon(L, V) := \max\{\, \tau \mid \phi^* L - \tau E \text{ is numerically effective} \,\}$$

measures “how ample” L is along the subvariety V. Seshadri constants play an important role in understanding the positivity properties of ample divisors. We will be interested in a related notion that can be defined for an arbitrary pseudoeffective divisor L. It first appears in connection with the numerical dimension in [Nakayama 2004].

Definition 5.1. Let X be a normal variety, $\mathcal{I}$ an ideal sheaf on X, and L a pseudoeffective divisor. Choose a smooth resolution $\phi : Y \to X$ of $\mathcal{I}$, and define E by setting $\mathcal{O}_Y(-E) = \phi^{-1}\mathcal{I} \cdot \mathcal{O}_Y$. We define the Nakayama constant

$$\varsigma(L, \mathcal{I}) := \max\{\, \tau \mid \phi^* L - \tau E \text{ is pseudoeffective} \,\}.$$

Of course, ς is independent of the choice of resolution. When $\mathcal{I}$ is the ideal sheaf of a subvariety V, $\varsigma(L, V)$ will denote the Nakayama constant.

One advantage of $\varsigma(L, V)$ is that it can be positive even when L is pseudoeffective but not big. Thus, the Nakayama constant is a more sensitive measure of positivity than the moving Seshadri constant of [Nakamaye 2003], which always vanishes as we approach the pseudoeffective boundary. It turns out that the Nakayama constant is closely related to the other notions of positivity we have considered.

Remark 5.2. Nakayama [2004] works with a slightly different formulation of this concept. His definition is equivalent to ours; the equivalence is demonstrated in the first paragraph of the proof of Proposition 5.3.

There is a useful criterion for nonvanishing of ς that is closer in spirit to Nakayama's original formulation.

Proposition 5.3. *Let X be a normal variety, $\mathcal{I}$ an ideal sheaf, and L a pseudoeffective divisor. Then $\varsigma(L,\mathcal{I})>0$ if and only if there is an ample divisor A on X so that for any q*

$$h^0(X,\overline{\mathcal{I}^q}\otimes\mathcal{O}_X(\lceil mL\rceil+A))>0$$

for sufficiently large m, where $\overline{\mathcal{I}^q}$ denotes the integral closure of $\mathcal{I}^q$.

Note that we can replace $\lceil - \rceil$ by $\lfloor - \rfloor$ by absorbing the difference into A.

Proof. Let $\phi:Y\to X$ denote a smooth resolution of $\mathcal{I}$ and $\mathcal{O}_Y(-E)=\phi^{-1}\mathcal{I}\cdot\mathcal{O}_Y$ define E. Suppose that $\varsigma(L,\mathcal{I})=0$ so that $m\phi^*L-E$ is not pseudoeffective for any m. Let $p:N^1(Y)\to V$ denote the cokernel of the inclusion $\mathbb{R}[\phi^*L]\to N^1(Y)$. Note that $p(-E)$ is disjoint from $p(\overline{NE}^1(Y))$. Thus, there is a small ample divisor H on Y so that $p(-E+H)$ is still disjoint from $p(\overline{NE}^1(Y))$. In other words, $m\phi^*L-E+H$ is not pseudoeffective for any m.

Let A be any ample divisor on X. Choose q so that $qH-\phi^*A$ is pseudoeffective. Then $m\phi^*L-qE+\phi^*A$ is not pseudoeffective for any m. Thus, for any A there is a q so that

$$h^0(Y,\mathcal{O}_Y(\phi^*(\lfloor mL\rfloor+A)-qE))=0$$

for every m. Since the class of $\lceil mL\rceil-\lfloor mL\rfloor$ is bounded as m varies, by absorbing the difference into A, the condition using $\lceil mL\rceil$ also fails.

Conversely, suppose that $\varsigma(L,\mathcal{I})>0$. Then for any real number $b>0$, $a\phi^*L-bE$ is pseudoeffective for any $a\geq b/\varsigma(L,\mathcal{I})$. By Proposition 2.17 (and Remark 2.15), there is an ample divisor H on Y (independent of b) so that

$$h^0(Y,\mathcal{O}_Y(\lfloor c(a\phi^*L-bE)\rfloor+H))>0$$

for every $c>0$ and every $a\geq b/\varsigma(L,\mathcal{I})$. Choose an ample $\mathbb{Z}$-divisor $A\geq\phi_*H$. Then $\phi^*A\geq\phi^*\phi_*H\geq H$ so that

$$h^0(Y,\mathcal{O}_Y(\phi^*(\lceil acL\rceil+A)-\lfloor bcE\rfloor))>0.$$

Fix an integer q and choose c so that $\lfloor cbE\rfloor\geq qE$. Then for any $m>bc/\varsigma(L,\mathcal{I})$,

$$h^0(X,\overline{\mathcal{I}^q}\otimes\mathcal{O}_X(\lceil mL\rceil+A))>0. \qquad \square$$

If we are only interested in whether $\varsigma(L,\mathcal{I})>0$, we can replace the condition of Proposition 5.3 by several alternatives. We have $\mathcal{I}^q\subset\overline{\mathcal{I}^q}\subset\mathcal{I}^{\langle q\rangle}$, and by the comparison theorems for symbolic powers (for example, [Swanson 2000, Theorem 3.1]), there is some k independent of q so that $\mathcal{I}^{\langle kq\rangle} \subset \mathcal{I}^q$. When X is smooth, we have $\mathcal{I}^q \subset \mathcal{J}(\mathcal{I}^q)$, and by Skoda's theorem, $\mathcal{J}(\mathcal{I}^q) \subset \mathcal{I}^{q-\dim X+1}$ for sufficiently large q. Thus, the nonvanishing of $\varsigma(L, \mathcal{I})$ is equivalent to the statement that for any q

$$h^0(X, *_q \otimes \mathcal{O}_X(\lceil mL\rceil + A)) > 0$$

for sufficiently large m, where $*_q$ can be

- $\mathcal{I}^q$,
- $\mathcal{I}^{\langle q\rangle}$, or
- $\mathcal{J}(\mathcal{I}^q)$ when X is smooth.

Applying the statement for symbolic powers, we immediately get the following:

Proposition 5.4. *Let X be a normal variety, V a subvariety not contained in* $\operatorname{Sing}(X)$, *and L a divisor. If $(\widetilde{X}, \widetilde{V})$ is a smooth V-birational model for (X, V), then $\varsigma(\phi^* L, \widetilde{V}) > 0$ if and only if $\varsigma(L, V) > 0$.*

The following proposition indicates that the Nakayama constant satisfies the usual compatibility relations:

Proposition 5.5. *Let X be a smooth variety, let L be a pseudoeffective divisor, and let $\mathcal{I}$ be an ideal such that no associated prime of $\mathcal{I}$ is centered in $\mathbf{B}_-(L)$. Then*

(1) *$\varsigma(L, \mathcal{I}) = \varsigma(P_\sigma(L), \mathcal{I})$, and*

(2) *if L is big, then $\varsigma(L, \mathcal{I}) = \max_{\phi^* L \geq A} \varsigma(A, \phi^{-1}\mathcal{I} \cdot \mathcal{O}_Y)$, where $\phi : Y \to X$ varies over all birational maps and A is big and numerically effective.*

Proof. (1) It suffices to show the inequality $\leq$. Let $\phi : Y \to X$ denote a smooth resolution of $\mathcal{I}$, and let E denote the divisor satisfying $\mathcal{O}_X(-E) = \phi^{-1}\mathcal{I} \cdot \mathcal{O}_Y$. Suppose that $\phi^* L - \tau E$ is pseudoeffective. Fix an ample A on Y. For any $\epsilon > 0$, we find that $\phi^* L + \epsilon A \sim_{\mathbb{R}} \tau E + F$ for some effective F. Since $\operatorname{Supp}(E)$ is not contained in the diminished base locus of $\phi^* L$, we know that $N_\sigma(\phi^* L + \epsilon A) \leq F$. Subtracting, we find that $P_\sigma(\phi^* L + \epsilon A) - \tau E$ is pseudoeffective. Taking a limit over ϵ and noting that $\phi^* P_\sigma(L) \geq P_\sigma(\phi^* L)$ completes the proof of the inequality.

(2) It suffices to show the inequality $\leq$. We may also replace L by some $\mathbb{Q}$-linearly equivalent divisor so that $L \geq 0$. Fix an effective ample divisor H on X. Proposition 3.7 indicates that there are birational maps ϕ_m and big and numerically effective divisors N_m satisfying $N_m \leq P_\sigma(\phi_m^* L) \leq N_m + \frac{1}{m}\phi_m^* H$. The expression on the right-hand side can be made arbitrarily close to $\varsigma(P_\sigma(L), \phi^{-1}\mathcal{I} \cdot \mathcal{O}_Y)$. By (1), this equals $\varsigma(L, \mathcal{I})$. □

[Nakayama 2004] shows that $\varsigma(L, V)$ is controlled by what happens to a very general subvariety of dimension equal to $\dim V$.

Proposition 5.6 [Nakayama 2004, Lemma V.2.21]. *Let X be a smooth variety of dimension n, and let L be a pseudoeffective divisor. Suppose there is a d-dimensional subvariety V such that $\varsigma(L, V) = 0$. Then there is a very ample divisor H so that any complete intersection W of $(n-d)$ very general elements of $|H|$ satisfies $\varsigma(L, W) = 0$.*

6. The numerical dimension

Our goal in this section is to show that the different definitions of the numerical dimension coincide. We start by giving an example of effective divisors that are numerically equivalent but have different Iitaka dimensions.

Example 6.1. We give an example of a threefold X and effective divisors L and L' so that $L \equiv L'$ but $\kappa(L) \neq \kappa(L')$. Fix an elliptic curve E, and consider $S = E \times E$ with projection maps p_1 and p_2. Let F be a fiber of p_1. Choose a degree-0 divisor T on E that is nontorsion, and define $N = p_2^* T$. We have $\kappa(F) = 1$ and $\kappa(F + N) = -\infty$.

Let X be the $\mathbb{P}^1$-bundle $\mathbb{P}_S(\mathcal{O}_S \oplus \mathcal{O}_S(F + N))$ with the morphism $\pi : X \to S$. Define L to be the section $\mathbb{P}_S(\mathcal{O}_S)$, and define $L' = L - \pi^* N$. Note that L and L' are numerically equivalent. By identifying the pushforwards of $\mathcal{O}_X(mL)$ with symmetric powers of $\mathcal{O}_S \oplus \mathcal{O}_S(F + N)$, we see that $\kappa(L) = 0$. Similarly, since $\mathcal{O}_X(L')$ can be realized as the relative dualizing sheaf of $\mathbb{P}_S(\mathcal{O}_S(-N) \oplus \mathcal{O}_S(F))$, we see that $\kappa(L') \geq \kappa(F) = 1$.

We first prove Theorem 1.1 for smooth varieties X. For convenience, we arrange the definitions in a more suitable order. Definition (1) in the following theorem is the definition of numerical dimension in [Boucksom et al. 2012] while (5) and (6) correspond to $\kappa_\sigma(L)$ and $\kappa_\nu(L)$ (by Remark 5.2) in [Nakayama 2004]. Note that we allow varieties $W \subset \operatorname{Supp}(L)$ at the slight cost of using numerical restrictions in (4).

Theorem 6.2. *Let X be a smooth variety, and let L be a pseudoeffective divisor. Here A will denote some fixed sufficiently ample $\mathbb{Z}$-divisor and W will range over all subvarieties of X not contained in $\mathbf{B}_-(L)$. Then the following quantities coincide*:

(1) $\max\{ k \in \mathbb{Z}_{\geq 0} \mid \langle L^k \rangle_X \neq 0 \}$.

(2) $\max\{ \dim W \mid \langle L^{\dim W} \rangle_{X|W} > 0 \}$.

(3) $\max\{ \dim W \mid \lim_{\epsilon \to 0} \operatorname{vol}_{X|W}(L + \epsilon A) > 0 \}$.

(4) $\max\{ \dim W \mid \inf_\phi \operatorname{vol}_{\widetilde{W}}([P_\sigma(\phi^* L)]|_{\widetilde{W}}) > 0 \}$, *where* $\phi : (\widetilde{X}, \widetilde{W}) \to (X, W)$ *ranges over W-birational models.*

(5) $\max\{ k \in \mathbb{Z}_{\geq 0} \,\big|\, \limsup_{m \to \infty} h^0(X, \lfloor mL \rfloor + A)/m^k > 0 \}$.

(6) $\min\{ \dim W \mid \varsigma(L, W) = 0 \}$ (by convention, if L is big we interpret this expression as returning $\dim X$).

(7) $\max\{ k \in \mathbb{Z}_{\geq 0} \mid \exists C > 0$ *such that* $Ct^{n-k} < \operatorname{vol}(L + tA)$ *for all* $t > 0 \}$.

We call this common quantity the numerical dimension of L and denote it $\nu_X(L)$. It only depends on the numerical class of L.

We will prove Theorem 6.2 using a cycle of inequalities. The equivalence of (1)–(4) is an easy consequence of the properties of the positive product, and the inequality (5) $\leq$ (6) was proved in [Nakayama 2004, Proposition V.2.22]. The other inequalities will require more work.

Proof. (1) = (2). Let $H_1, \dots, H_{d-k}$ represent very general elements of a very ample linear system. Since $\langle L^k \rangle_X$ is in the closure of the cone generated by effective cycles, it is nonzero if and only if $\langle L^k \rangle_X \cdot H_1 \cdot \cdots \cdot H_{d-k} > 0$. By Lemma 4.18, this is equivalent to $\langle L^k \rangle_{X|H_1 \cap \cdots \cap H_{d-k}} > 0$. Thus, (1) $\leq$ (2). By Proposition 4.21, the same argument in reverse shows that (2) $\leq$ (1).

(2) = (3). Proposition 4.12 shows that the conditions set on W in (2) and (3) are the same.

(3) = (4). Proposition 4.12 allows us to translate between restricted volume and the restricted positive product in the V-big case. Thus, Theorem 4.15 implies that

$$\operatorname{vol}_{X|W}(L+\epsilon A) = \inf_{\phi:\widetilde{X}\to X} \operatorname{vol}_{\widetilde{W}}\bigl([P_\sigma(\phi^*(L+\epsilon A))]|_{\widetilde{W}}\bigr),$$

where $\phi : (\widetilde{X}, \widetilde{W}) \to (X, W)$ varies over W-birational models. Consider

$$\lim_{\epsilon\to 0} \operatorname{vol}_{X|W}(L+\epsilon A) = \lim_{\epsilon\to 0} \inf_{\phi:\widetilde{X}\to X} \operatorname{vol}_{\widetilde{W}}\bigl([P_\sigma(\phi^*(L+\epsilon A))]|_{\widetilde{W}}\bigr).$$

Note that on any model $\operatorname{vol}_{\widetilde{W}}([P_\sigma(\phi^*(L+\epsilon A))]|_{\widetilde{W}})$ is nondecreasing and continuous as a function of ϵ. Thus, on the right-hand side, we may commute the limit with the infimum.

(4) $\leq$ (5). The first step is to show that there is some ample divisor on W whose pullback lies beneath each restriction $P_\sigma(\phi^*L)|_{\widetilde{W}}$. Using this ample divisor, we find a lower bound for the growth of sections of a certain twisted linear series on W. The last step is to prove a lifting theorem for twisted linear series to conclude that $h^0(\lfloor mL \rfloor + A)$ satisfies the necessary growth conditions.

Lemma 6.3. *Let X be a smooth variety of dimension n, let L be a big divisor, and let N be a general element of a big basepoint-free linear system. Then we have* $\operatorname{vol}(L-N) \geq \operatorname{vol}(L) - n \operatorname{vol}_{X|N}(L)$.

The easiest demonstration appeals to the results of [Boucksom et al. 2009].

Proof. Let $\alpha = \sup_{t\in[0,1]}\{L - tN \text{ is pseudoeffective}\}$. Note $1 \geq \alpha$, and since L is big, $0 < \alpha$. We will prove the stronger result $\operatorname{vol}(L-N) \geq \operatorname{vol}(L) - n\alpha \operatorname{vol}_{X|N}(L)$.

By [Boucksom et al. 2009, Corollary C], the function vol is continuously differentiable on the big cone. More precisely, for $t \in (0, \alpha)$ we have

$$\frac{d}{dt}\operatorname{vol}(L - tN) = -n \operatorname{vol}_{X|N}(L - tN).$$

Note that $\operatorname{vol}_{X|N}(L - tN) \le \operatorname{vol}_{X|N}(L)$ for any $t \ge 0$. Thus, for every $t \in (0, \alpha)$ there is an inequality

$$\frac{d}{dt}\operatorname{vol}(L - tN) \ge -n \operatorname{vol}_{X|N}(L).$$

Integrating both sides over $t \in [0, \alpha]$, we get $\operatorname{vol}(L-\alpha N) \ge \operatorname{vol}(L) - n\alpha \operatorname{vol}_{X|N}(L)$. But if $\alpha \ne 1$, then $\operatorname{vol}(L - \alpha N) = 0 = \operatorname{vol}(L - N)$, finishing the proof. □

Lemma 6.4. *Let W be a smooth variety. Suppose that for every smooth birational model $\phi : \widetilde{W} \to W$ we associate a divisor $L_{\widetilde{W}}$ so that for any birational map $\psi : \widehat{W} \to \widetilde{W}$ we have $\psi^* L_{\widetilde{W}} \ge L_{\widehat{W}}$. Suppose furthermore that*

$$\inf_{\widetilde{W}} \operatorname{vol}(L_{\widetilde{W}}) > 0.$$

There is some ample divisor H on W and constant ϵ such that $\operatorname{vol}(L_{\widetilde{W}} - \phi^* H) > \epsilon$ *for every ϕ.*

Note that $\operatorname{vol}(L_{\widetilde{W}}) \ge \operatorname{vol}(L_{\widehat{W}})$ for every higher model $\widehat{W}$.

Proof. For convenience, set $n = \dim W$ and $\tau = \inf \operatorname{vol}(L_{\widetilde{W}})$. Fix a very ample divisor H on W. It suffices to show that there is some constant k such that for any smooth model $\phi : \widetilde{W} \to W$, there is an $H' \equiv H$ so that

$$\operatorname{vol}(L_{\widetilde{W}} - \tfrac{1}{k}\phi^* H') > \tau/2.$$

Choose a prime very ample divisor $H' \equiv H$ sufficiently general so that $\psi^* H'$ is equal to the strict transform of H'. Note that

$$\operatorname{vol}_{\widetilde{W}|\phi^* H'}(L_{\widetilde{W}}) \le \operatorname{vol}_{\widetilde{W}|\phi^* H'}(\phi^* L_W),$$

and by [Ein et al. 2009, Lemma 2.4], the latter quantity is equal to $\operatorname{vol}_{W|H'}(L_W)$. Choose some constant k so that

$$\tfrac{1}{k}\operatorname{vol}_{W|H'}(L_W) < \frac{\tau}{2n}.$$

(Note that by [Boucksom et al. 2009, Proposition 4.8], k is independent of the choice of H' and thus also independent of the choice of $\widetilde{W}$.) Lemma 6.3 implies

$$\begin{aligned}\operatorname{vol}(kL_{\widetilde{W}} - \phi^* H') &\ge \operatorname{vol}(kL_{\widetilde{W}}) - n \operatorname{vol}_{\widetilde{W}|\phi^* H'}(kL_{\widetilde{W}}) \\ &\ge \operatorname{vol}(kL_{\widetilde{W}}) - n \operatorname{vol}_{\widetilde{W}|\phi^* H'}(k\phi^* L_W).\end{aligned}$$

Rescaling the above expression by k, we find

$$\operatorname{vol}(L_{\widetilde{W}} - \tfrac{1}{k}\phi^* H') \geq \operatorname{vol}(L_{\widetilde{W}}) - \frac{n}{k} \operatorname{vol}_{\widetilde{W}|\phi^* H'}(\phi^* L_W) > \tau/2. \qquad \square$$

In our situation, we find the following:

Corollary 6.5. *Assume that W is a very general intersection of very ample divisors such that* $\inf_\phi \operatorname{vol}_{\widetilde{W}}(P_\sigma(\phi^* L)|_{\widetilde{W}}) > 0$, *where* $\phi : (\widetilde{X}, \widetilde{W}) \to (X, W)$ *varies over all W-birational models. Then there is an ample divisor H on W so that for any W-birational model $\phi : \widetilde{X} \to X$, we have*

$$\operatorname{vol}_{\widetilde{W}}(P_\sigma(\phi^* L)|_{\widetilde{W}} - \phi^* H) > 0.$$

Proof. Consider the set of divisors $P_\sigma(\phi^* L)|_{\widetilde{W}}$. Since $N_\sigma(\phi^* L) \geq_{\widetilde{W}} 0$, they satisfy the comparison condition of Lemma 6.4. By assumption, the infimum condition of Lemma 6.4 also holds. The lemma yields an appropriate ample divisor H on W. □

Our next goal is a lifting theorem for twisted linear series.

Proposition 6.6. *Let X be a smooth variety, and let L be an effective divisor. Suppose that N is a big and numerically effective divisor satisfying* $0 \leq N \leq L$ *such that N has simple normal crossing support. Let $|B|$ be a basepoint-free linear system defining a birational morphism on X. For sufficiently general elements* $B_1, \dots, B_k \in |B|$, *we have an inequality*

$$h^0\big(W, \mathcal{O}_W(K_W + \lceil N|_W \rceil + A|_W)\big) \leq h^0\big(X|W, \mathcal{O}_X(K_X + \lceil L \rceil + B_1 + \cdots + B_k + A)\big),$$

where W is the complete intersection $B_1 \cap \cdots \cap B_k$ and A is any numerically effective $\mathbb{Z}$-divisor on X.

Proof. For convenience, define $W_j := B_1 \cap \cdots \cap B_j$ and $M_i := B_{i+1} + \cdots + B_k$. Note that since the B_i are sufficiently general, we may assume that each W_j is smooth, that $N \geq_{W_j} 0$, and that $N|_{W_j}$ has simple normal crossing support. Note furthermore that B is big and numerically effective so that $M_i|_{W_j}$ is also a big and numerically effective divisor for any i and j.

Kawamata–Viehweg vanishing implies that we have surjections

$$H^0\big(W_i, \mathcal{O}_{W_i}(K_{W_i} + \lceil N|_{W_i} \rceil + (A + M_i)|_{W_i})\big) \to \\ H^0\big(W_{i+1}, \mathcal{O}_{W_{i+1}}(K_{W_{i+1}} + \lceil N|_{W_i} \rceil|_{W_{i+1}} + (A + M_{i+1})|_{W_{i+1}})\big).$$

Furthermore, since $N \geq_{W_i} 0$ for every i, we have $\lceil N|_{W_i} \rceil|_{W_{i+1}} \geq \lceil N|_{W_{i+1}} \rceil$. Thus, by induction we obtain

$$h^0\big(X|W_i, \mathcal{O}_X(\lceil N \rceil + (K_X + A + B_1 + \cdots + B_k))\big) \\ \geq h^0\big(W_i, \mathcal{O}_{W_i}(\lceil N|_{W_i} \rceil + (K_X + A + B_1 + \cdots + B_k)|_{W_i})\big).$$

When $i = k$, we obtain the desired statement. □

We now finish the proof of the inequality (4) $\leq$ (5). Set k to be the value of (4). Fix an ample divisor A on X as in Theorem 2.4 so that for any m there is an $L_m \sim \lceil mL \rceil + A$ such that $L_m \geq 0$.

For each L_m, we can apply Proposition 3.7 to find an effective divisor G_m, a countable sequence of maps $\phi_{i,m}$, and a big and numerically effective divisor $N_{i,m}$ satisfying

$$N_{i,m} \leq P_\sigma(\phi_{i,m}^* L_m) \leq N_{i,m} + \tfrac{1}{i}\phi_{i,m}^* G_m.$$

We may of course assume that each $N_{i,m}$ has simple normal crossing support and each $\phi_{i,n}$ is a composition of blowups along smooth centers.

Note that the set of maps $\phi_{i,m}$ is countable as m and i vary. Fix a very ample linear system $|B|$ on X. We can choose very general elements $B_1, \dots, B_k \in |B|$ so that the $\phi_{i,m}^* B_j$ satisfy the conditions of Proposition 6.6 for each $\widetilde{X}_{i,m}$ and $N_{i,m}$ simultaneously. We may also choose the B_j sufficiently general so that the strict transform of B_j over $\phi_{i,m}$ is the same as the pullback for every i and m. Set $W = B_1 \cap \cdots \cap B_k$. Then each $\phi_{i,m}$ is W-birational and $\widetilde{W}_{i,m,j} = \phi_{i,m}^* B_1 \cap \cdots \cap \phi_{i,m}^* B_j$ is smooth for every j between 1 and k.

Choose an ample divisor H on W as in Corollary 6.5. For each G_m, choose a sufficiently small $\epsilon_m > 0$ so that $H - \epsilon_m G_m|_W$ is pseudoeffective. By choosing $i > 1/\epsilon_m$, we find models $\phi_m : \widetilde{X}_m \to X$ so that

$$N_m \leq_{\widetilde{W}_m} P_\sigma(\phi_m^* L_m) \leq_{\widetilde{W}_m} N_m + \epsilon_m \phi_m^* G_m.$$

Thus,

$$\begin{aligned} N_m|_{\widetilde{W}_m} - (m-1)\phi_m^* H &\geq (P_\sigma(\phi_m^* L_m) - \epsilon_m \phi_m^* G_m)|_{\widetilde{W}_m} - (m-1)\phi_m^* H \\ &\geq (P_\sigma(\phi_m^* L_m) - P_\sigma(\phi_m^* mL))|_{\widetilde{W}_m} \\ &\qquad + m(P_\sigma(\phi_m^* L)|_{\widetilde{W}_m} - \phi_m^* H) + \phi_m^*(H - \epsilon_m G_m|_W). \end{aligned}$$

We analyze this last sum term by term. Since $L_m - mL$ is W-pseudoeffective and $N_\sigma(\phi_m^* L) \geq_{\widetilde{W}_m} 0$, the first term is pseudoeffective by Proposition 3.5. The conclusion of Corollary 6.5 is that the second term is big. The third term is also pseudoeffective by construction. Thus, $D_m := N_m|_{\widetilde{W}_m} - (m-1)\phi_m^* H$ is big.

Fix a very ample divisor M on X. Then

$$\begin{aligned} &h^0\big(\widetilde{W_m}, \mathcal{O}_{\widetilde{W_m}}(K_{\widetilde{W_m}} + (k+2)\phi_m^* M|_W + \lceil N_m|_{\widetilde{W}_m}\rceil)\big) \\ &\qquad \geq h^0\big(\widetilde{W_m}, \mathcal{O}_{\widetilde{W_m}}(K_{\widetilde{W_m}} + (k+2)\phi_m^* M|_W + \lceil D_m \rceil + \lfloor (m-1)\phi_m^* H \rfloor)\big) \\ &\qquad \geq h^0(W, \lfloor (m-1)H \rfloor) \quad \text{by Proposition 2.17} \\ &\qquad \geq Cm^k \end{aligned}$$

for some constant $C > 0$ and for m sufficiently large.

We conclude by applying Proposition 6.6. We have already chosen the divisors $B_1, B_2, \dots, B_k$ sufficiently general so that their pullbacks satisfy the conditions of the theorem. For convenience, define $A' = B_1 + \cdots + B_k$. Proposition 6.6 shows that the dimensions of the spaces of restricted sections

$$h^0\big(\widetilde{X}_m | \widetilde{W}_m, \mathbb{O}_{\widetilde{X}_m}(K_{\widetilde{X}_m} + \phi_m^*(L_m + A' + (k+2)\phi^* M))\big) > Cm^k$$

for some constant $C > 0$ and for sufficiently large m. Since $K_{\widetilde{X}_m/X}$ is ϕ_m-exceptional, these dimensions are equal to

$$\begin{aligned} h^0\big(X|W, \mathbb{O}_X(K_X + L_m + A' + (k+2)\phi^* M)\big) \\ = h^0\big(X|W, \mathbb{O}_X(K_X + \lceil mL \rceil + A + A' + (k+2)\phi^* M)\big). \end{aligned}$$

Thus, $h^0(X, \mathbb{O}_X(K_X + \lceil mL \rceil + A + A' + (k+2)\phi^* M))$ is also bounded below by Cm^k for sufficiently large m.

(5) $\leq$ (6). This is proved in [Nakayama 2004, Proposition V.2.22].

(6) $\leq$ (1). By Proposition 5.6, we may assume that W is a very general intersection of very ample divisors. We need to consider the 0-case separately. Note that (1) is 0 precisely when $P_\sigma(L)$ is numerically trivial. This means that (6) is also 0. Thus, we can prove that (6) $\leq$ (1) by considering the case where (6) is at least 2 and (1) is at least 1.

Suppose for a contradiction that (1) is less than the value of (6). For convenience, we set k to be the value of (1). Let W be a k-dimensional intersection of very general, very ample divisors. Set $\tau = \varsigma(L, W) > 0$, and let $\phi : Y \to X$ be the blowup of W with exceptional divisor E.

Fix a very ample divisor H on Y. We first analyze $\phi^* L + \epsilon H$. Choose models $\psi_i : \widetilde{Y}_i \to Y$ computing positive products $\langle (\phi^* L + \epsilon H)^k \rangle_{Y|E}$ and $\langle (\phi^* L + \epsilon H)^{k+1} \rangle_Y$. Choose big and numerically effective divisors $A_i \leq \psi_i^*(\phi^* L + \epsilon H)$ on $\widetilde{Y}_i$ that compute the product. By Proposition 5.5, $P_\sigma(\psi_i^*(\phi^* L + \epsilon H)) - \tau \psi_i^* E$ is always pseudoeffective, so by choosing ψ_i appropriately, we may also assume $A_i - \frac{\tau}{2} \psi_i^* E$ is pseudoeffective for each A_i. Thus, $A_i - \frac{\tau}{2} \widetilde{E}$ is also pseudoeffective, where $\widetilde{E}$ denotes the strict transform of E on $\widetilde{Y}_i$. Then

$$0 \leq (A_i - \tfrac{\tau}{2} \widetilde{E}) \cdot A_i^k \cdot \psi_i^* H^{d-k-1}.$$

By taking a limit over pushforwards on all such models, we find

$$0 \leq \langle (\phi^* L + \epsilon H)^{k+1} \rangle_Y \cdot H^{d-k-1} - \tfrac{\tau}{2} \langle (\phi^* L + \epsilon H)^k \rangle_{Y|E} \cdot H^{d-k-1}.$$

This is true for all sufficiently small ϵ, so

$$0 \leq \langle \phi^* L^{k+1} \rangle_Y \cdot H^{d-k-1} - \tfrac{\tau}{2} \langle \phi^* L^k \rangle_{Y|E} \cdot H^{d-k-1}.$$

By choosing sufficiently general elements $H_1, \dots, H_{d-k-1} \in |H|$, we may ensure that $E \cap H_1 \cap \cdots \cap H_{d-k-1}$ maps finitely onto W via ϕ. Letting the $A_1, \dots, A_{d-k}$ denote the ample divisors whose intersection is W, we have

$$\begin{aligned}\langle \phi^* L^k \rangle_{Y|E} \cdot H^{d-k-1} &= \langle \phi^* L^k \rangle_{Y|E \cap H_1 \cap \cdots \cap H_{d-k-1}} \\ &= C \langle L^k \rangle_{X|W} \\ &= C \langle L^k \rangle_X \cdot A_1 \cdot \cdots \cdot A_{d-k}\end{aligned}$$

for some positive constant C. By assumption, this latter quantity is positive, so

$$0 < \langle \phi^* L^{k+1} \rangle_Y \cdot H^{d-k-1},$$

contradicting the fact that $\langle L^{k+1} \rangle_X = 0$.

(7) $\leq$ (1). Let k denote the value of (1). Note that

$$\begin{aligned}t^{n-k} \langle (L + tA)^k \rangle \cdot A^{n-k} &= \langle (L + tA)^k \cdot (tA)^{n-k} \rangle \\ &\leq \langle (L + tA)^n \rangle.\end{aligned}$$

The expression in (1) implies that there is some constant C such that $C < \langle (L + tA)^k \rangle \cdot A^{n-k}$ for every $t > 0$. Thus, we obtain $Ct^{n-k} < \mathrm{vol}(L + tA)$ for every $t > 0$.

(1) $\leq$ (7). Let k denote the value of (7). For every constant C, there is some $t > 0$ such that

$$\langle (L + tA)^n \rangle < Ct^{n-k-1}.$$

This implies that

$$t^{n-k-1} \langle (L + tA)^{k+1} \rangle \cdot A^{n-k-1} < Ct^{n-k-1}$$

so that for any C there is some t such that $\langle (L + tA)^{k+1} \rangle \cdot A^{n-k-1} < C$. Note that the left-hand side is increasing in t so that the inequality must hold for arbitrarily small t. Thus, the value of (1) is at most k. □

The numerical dimension satisfies a number of natural properties. All of the following are checked in [Nakayama 2004, Proposition V.2.7] except for (5) and (7):

Theorem 6.7 [Nakayama 2004, Proposition V.2.7]. *Let X be a smooth variety, and let L be a pseudoeffective $\mathbb{R}$-divisor.*

(1) *We have $0 \leq \nu(L) \leq \dim X$ and $\kappa(L) \leq \nu(L)$.*

(2) *We have $\nu(L) = \dim X$ if and only if L is big and $\nu(L) = 0$ if and only if $P_\sigma(L) \equiv 0$.*

(3) *If L' is pseudoeffective, then $\nu(L + L') \geq \nu(L)$.*

(4) *If $f : Y \to X$ is any surjective morphism from a normal variety Y, then $\nu(f^* L) = \nu(L)$.*

(5) *We have* $\nu(L) = \nu(P_\sigma(L))$.

(6) *Suppose that* $f : X \to Z$ *has connected fibers and* F *is a very general fiber of* f. *Then* $\nu(L) \leq \nu(L|_F) + \dim Z$.

(7) *Fix some sufficiently ample* $\mathbb{Z}$*-divisor* A. *Then there are positive constants* C_1 *and* C_2 *so that*

$$C_1 m^{\nu(L)} < h^0(X, \mathbb{O}_X(\lfloor mL \rfloor + A)) < C_2 m^{\nu(L)}$$

for every sufficiently large m.

Proof. Part (5) follows from the invariance of the positive product under passing to P_σ.

Consider the inequality of (7). The leftmost inequality was stated explicitly while demonstrating the implication (4) $\leq$ (5) in the proof of Theorem 6.2. To show the rightmost inequality, let W be a subvariety of dimension $\nu(L)$ with $\varsigma(L, W) = 0$. Proposition 5.3 (and the following discussion) shows that there is a positive integer q with

$$h^0(X, \mathcal{I}_W^q \otimes \mathbb{O}_X(\lceil mL \rceil + A)) = 0$$

for sufficiently large m. Writing W_q for the subscheme defined by the ideal $\mathcal{I}_W^q$, for sufficiently large m there is an injection

$$h^0(X, \mathbb{O}_X(\lceil mL \rceil + A)) \to h^0(W_q, \mathbb{O}_{W_q}(\lceil mL \rceil + A)),$$

and the rate of growth of the latter is bounded by $m^{\dim(W_q)} = m^{\nu(L)}$. □

It is interesting to note that ν is *not* lower semicontinuous as might be expected. This is a consequence of the fact that the restricted positive product is only semicontinuous on the boundary of the V-pseudoeffective cone.

Example 6.8 [Boucksom et al. 2009, Example 3.8]. Let X be any smooth surface with infinitely many -1-curves. Take some compact slice of $\overline{NE}^1(X)$. We can choose a convergent sequence of distinct classes $\{\alpha_i\}$ on this compact slice such that each α_i lies on a ray generated by a different -1-curve. Note that for any irreducible curve C, there is at most one i for which $\alpha_i \cdot C < 0$. Thus, $\beta := \lim_{i\to\infty} \alpha_i$ must be a numerically effective class. A nontrivial numerically effective class β has $\nu(\beta) \geq 1$ but $\nu(\alpha_i) = 0$ for every i. Thus, ν is not lower semicontinuous.

Question 6.9. What properties does ν satisfy along the V-pseudoeffective boundary?

6A. *The numerical dimension for normal varieties.*

Since the numerical dimension is a birational invariant, we can extend the definition to any normal variety X.

Definition 6.10. Let X be a normal variety, and let L be an $\mathbb{R}$-Cartier divisor on X. We define $\nu(L)$ to be $\nu(f^*L)$, where $f : Y \to X$ is any smooth model.

We now complete the proof of Theorem 1.1 by showing that the criteria of Theorem 6.2 can be applied directly to a normal variety. Note that the numbering in the two theorems is different; we will use the numbering of Theorem 1.1.

Proof of Theorem 1.1. We have (1) $= \nu(L)$ since the arguments in the proof of [Nakayama 2004, Proposition V.2.7] show that (1) is a birational invariant even for normal varieties.

We next show that (3) $= \nu(L)$. We first claim there is a complete intersection W of very general, very ample divisors that maximizes (3). Suppose that $V \subset X$ is a k-dimensional subvariety that achieves the maximum value in (3). Choose very ample divisors $A_1, \dots, A_{n-k}$ whose (scheme-theoretic) complete intersection W_0 contains V and also has dimension k. Set $P = \mathbb{P}H^0(X, \mathscr{O}_X(A_1)) \times \cdots \times \mathbb{P}H^0(X, \mathscr{O}_X(A_{n-k}))$.

Let $\mathscr{I}$ be the ideal sheaf on $X \times P$ whose restriction to a fiber of the second projection is the ideal sheaf of the corresponding complete intersection on X. Note that $\mathscr{I}$ is flat over the locus on P representing intersections of the expected dimension. By upper-semicontinuity, we find that for any fixed divisor D we have

$$h^0(X, \mathscr{I}_W(\lfloor D \rfloor)) \leq h^0(X, \mathscr{I}_{W_0}(\lfloor D \rfloor))$$

for a general complete intersection W. Thus,

$$\begin{aligned} h^0(X|W, \mathscr{O}_X(\lfloor D \rfloor)) &\geq h^0(X|W_0, \mathscr{O}_X(\lfloor D \rfloor)) \\ &\geq h^0(X|V, \mathscr{O}_X(\lfloor D \rfloor)) \end{aligned}$$

since the restriction map $\mathscr{O}_X \to \mathscr{O}_V$ factors through restriction to $\mathscr{O}_{W_0}$. In particular, if we fix a countable collection of divisors D_i, then for a very general complete intersection W, we have $\mathrm{vol}_{X|W}(D_i) \geq \mathrm{vol}_{X|V}(D_i)$ for every i. Setting $D_i := L + \frac{1}{i}A$ yields the claim.

Let $\phi : Y \to X$ be a smooth model of X. For any ample divisor A on Y, there is an ample divisor H on X such that $\phi^*H \geq A$. Since W is not contained in any ϕ-exceptional center, we may furthermore ensure that $\mathrm{Supp}(\phi^*H - A)$ does not contain W.

In particular, for any ample divisor A on Y there is some H on X such that

$$\mathrm{vol}_{X|W}(L + \epsilon H) = \mathrm{vol}_{Y|\widetilde{W}}(\phi^*(L + \epsilon H)) \geq \mathrm{vol}_{Y|\widetilde{W}}(\phi^*L + \epsilon A).$$

Similarly, for any ample divisor H on X there is an A on Y with $A - \phi^*H$ ample. Thus, (3) $= \nu(L)$ is proved.

Then (2) $= \nu(L)$ follows from the arguments of the previous two paragraphs, (4) $= \nu(L)$ since (4) remains unchanged upon passing to a smooth V-birational model, both (5) $= \nu(L)$ and (6) $= \nu(L)$ follow from Corollary 4.19, and (7) $= \nu(L)$ by Proposition 5.4. □

7. The restricted numerical dimension

We now turn to the restricted numerical dimension. For a subvariety V, $\nu_{X|V}(L)$ should measure the maximal dimension of a subvariety $W \subset V$ such that the "positive restriction" of L to W is big.

Theorem 7.1. *Let X be a smooth variety, let V be a subvariety of X, and let L be a V-pseudoeffective divisor. In the following, A denotes some fixed sufficiently ample $\mathbb{Z}$-divisor, and W will range over all subvarieties of V not contained in $\mathbf{B}_-(L)$. Then the following quantities coincide*:

(1) $\max\{\, k \in \mathbb{Z}_{\geq 0} \mid \langle L^k \rangle_{X|V} \neq 0 \,\}$,

(2) $\max\{\, \dim W \mid \langle L^{\dim W} \rangle_{X|W} > 0 \,\}$,

(3) $\max\{\, \dim W \mid \lim_{\epsilon \to 0} \operatorname{vol}_{X|W}(L + \epsilon A) > 0 \,\}$, *and*

(4) $\max\{\, \dim W \mid \liminf_{\phi} \operatorname{vol}_{\widetilde{W}}([P_\sigma(\phi^* L)]|_{\widetilde{W}}) > 0 \,\}$, *where* $\phi : (\widetilde{X}, \widetilde{W}) \to (X, W)$ *ranges over W-birational models.*

This common quantity is known as the restricted numerical dimension of L along V and is denoted $\nu_{X|V}(L)$. It only depends on the numerical class of L.

The argument is the same as in the proof of the first four equivalences in Theorem 6.2. One wonders whether the other equalities in Theorem 6.2 can be extended to analogous notions for the restricted numerical dimension. Perhaps the most important is the restricted version of κ_σ.

Definition 7.2. Let X be a smooth variety, let V be a subvariety, and let L be a V-pseudoeffective divisor. Fix any divisor A. If $H^0(X|V, \mathbb{O}_X(\lfloor mL + A \rfloor))$ is nonzero for infinitely many values of m, we define

$$\kappa_\sigma(X|V, L; A) := \max\left\{ k \in \mathbb{Z}_{\geq 0} \;\middle|\; \limsup_{m \to \infty} \frac{h^0(X|V, \mathbb{O}_X(\lfloor mL + A \rfloor))}{m^k} > 0 \right\}.$$

Otherwise, define $\kappa_\sigma(X|V, L; A) := -\infty$. The restricted σ-dimension $\kappa_\sigma(X|V, L)$ is defined to be

$$\kappa_\sigma(X|V, L) := \max_A \{\kappa_\sigma(X|V, L; A)\}.$$

Arguing as in the proof of [Nakayama 2004, Proposition V.2.7], one can check that the restricted σ-dimension is a numerical and birational invariant.

Question 7.3. Let X be a smooth variety, V a subvariety, and L a V-pseudoeffective divisor. Does $\nu_{X|V}(L) = \kappa_\sigma(X|V, L)$?

Since the restricted numerical dimension is invariant under passing to admissible models, we can extend the definition to pairs with singularities.

Definition 7.4. Let X be a normal variety, V a subvariety not contained in $\operatorname{Sing}(X)$, and L a V-pseudoeffective divisor. We define $\nu_{X|V}(L) = \nu_{Y|W}(f^*L)$, where (Y, W) is any smooth V-birational model of (X, V).

7A. *Properties of the restricted numerical dimension.* The restricted numerical dimension satisfies similar properties to the numerical dimension. Since we know less about $\nu_{X|V}$, the statements are slightly weaker.

Theorem 7.5. *Let X be a smooth variety, V a subvariety of X, and L a V-pseudoeffective divisor.*

(1) *We have $\nu_{X|V}(L) \leq \nu(L)$, and if V is normal, then $\nu_{X|V}(L) \leq \nu(L|_V)$.*

(2) *We have $\nu_{X|V}(L) = \nu_{X|V}(P_\sigma(L))$.*

(3) *When L is numerically effective, $\nu_{X|V}(L) = \nu_V(L|_V)$.*

(4) *If L' is also V-pseudoeffective, then $\nu_{X|V}(L + L') \geq \nu_{X|V}(L)$.*

(5) *Suppose that $\nu_{X|V}(L) < \dim V$. If H is a very general, very ample divisor on X, then $\nu_{X|V}(L) = \nu_{X|V \cap H}(L)$.*

(6) *If $\phi : (\widetilde{X}, \widetilde{V}) \to (X, V)$ is an admissible model with $\widetilde{X}$ smooth, then we have $\nu_{\widetilde{X}|\widetilde{V}}(\phi^*L) = \nu_{X|V}(L)$.*

(7) *Let $\phi : Y \to X$ be a smooth birational model, and let W be a subvariety of Y such that $\phi|_W$ maps surjectively onto V. Then $\nu_{Y|W}(\phi^*L) = \nu_{X|V}(L)$.*

Proof. (1) Note that if Z and Z' are subvarieties of X with $Z \subset Z'$, then we have $\operatorname{vol}_{X|Z'}(L) \geq \operatorname{vol}_{X|Z}(L)$ since the restriction map on sections of L from X to Z factors through the restriction map to Z'.

Fix an ample divisor A on X, and let W be an intersection of very general, very ample divisors on X. The two inequalities follow from the two facts that $\operatorname{vol}_{X|W}(L + \epsilon A) \geq \operatorname{vol}_{X|V \cap W}(L + \epsilon A)$ and $\operatorname{vol}_{X|V}(L + \epsilon A) \geq \operatorname{vol}_{X|V \cap W}(L + \epsilon A)$.

(2) This follows from the fact that the restricted positive product is invariant under passing to P_σ as demonstrated in Proposition 4.13.

(3) The restricted volume of an ample divisor can be calculated as an intersection product, so the equality follows from characterization (3) in Theorem 7.1.

(4) Fix an ample divisor A. Then the inequality follows from the other inequality $\operatorname{vol}_{X|W}(L + L' + 2\epsilon A) \geq \operatorname{vol}_{X|W}(L + \epsilon A)$.

(5) Using characterization (1) in Theorem 7.1, we see that if $k < \dim V$, then $\langle L^k \rangle_{X|V} \neq 0$ if and only if $\langle L^k \rangle_{X|V \cap H} = \langle L^k \rangle_{X|V} \cdot H \neq 0$.

(6) This is a consequence of Proposition 4.20 that describes how the restricted positive product is compatible with admissible models.

(7) First suppose that $\dim W > \dim V$; we show $\nu_{Y|W}(\phi^*L) < \dim W$. Every fiber of $\phi|_W$ is covered by curves with $\phi^*L \cdot C = 0$. Since $\mathbf{B}_-(\phi^*L) = \phi^{-1}\mathbf{B}_-(L)$, the

general such curve avoids $\mathbf{B}_-(\phi^*L)$. In particular, for any W-birational model $\psi : \widetilde{Y} \to Y$, the subvariety $\widetilde{W}$ is covered by curves satisfying $P_\sigma(\psi^*\phi^*L) \cdot C = 0$. Thus, $\nu_{Y|W}(\phi^*L) < \dim W$ by characterization (4) in Theorem 7.1.

Fix a very general, very ample divisor H on Y. Then $\nu_{Y|W}(\phi^*L) = \nu_{Y|W\cap H}(\phi^*L)$ by property (5). Proceeding inductively, we reduce to the case $\dim W = \dim V$, which is (6). □

It is important to note we can have $\nu_{X|V}(L) = \dim V$ even when L is not V-big.

Example 7.6. Let X be a smooth variety, V a smooth subvariety, and L a V-big divisor. Let $\phi : (Y, W) \to (X, V)$ be an admissible model such that some ϕ-exceptional center contains V. Then ϕ^*L is W-pseudoeffective but not W-big. Nevertheless, the invariance of $\nu_{X|V}(L)$ under passing to admissible models shows that we still have $\nu_{X|V}(L) = \dim V$.

We next show that the nonvanishing of $\nu(L)$ can be detected by the restricted numerical dimension $\nu_{X|C}(L)$ for a very general curve C.

Proposition 7.7. *Let X be a smooth variety, and let L be a pseudoeffective divisor on X. Then $\nu(L) > 0$ if and only if there is a curve C on X defined as a very general complete intersection of very ample divisors with $\nu_{X|C}(L) > 0$.*

Proof. If $\nu(L) = 0$, then $\nu_{X|C}(L) = 0$ by Theorem 7.5.

Conversely, suppose that C is a very general intersection of very ample divisors. By choosing C appropriately, we may assume that it avoids every component of $\mathbf{B}_-(P_\sigma(L))$. In particular, for any C-birational model $\phi : Y \to X$, we have

$$\operatorname{vol}(P_\sigma(\phi^*L)|_{\widetilde{C}}) = P_\sigma(\phi^*L) \cdot \widetilde{C} = \phi^*P_\sigma(L) \cdot \widetilde{C} = P_\sigma(L) \cdot C.$$

Thus, if $\nu_{X|C}(L) = 0$, then $P_\sigma(L) \cdot C = 0$. But since C is an intersection of ample divisors, this implies that $P_\sigma(L) \equiv 0$ and $\nu(L) = 0$. □

Acknowledgments

I would like to thank my advisor J. M$^{\text{c}}$Kernan for his advice and support. I also thank R. Lazarsfeld for some helpful conversations, Y. Gongyo for pointing out several mistakes in an earlier draft, and the referee for the careful revisions.

References

[Boucksom 2004] S. Boucksom, "Divisorial Zariski decompositions on compact complex manifolds", *Ann. Sci. École Norm. Sup.* (4) **37**:1 (2004), 45–76. MR 2005i:32018 Zbl 1054.32010

[Boucksom et al. 2009] S. Boucksom, C. Favre, and M. Jonsson, "Differentiability of volumes of divisors and a problem of Teissier", *J. Algebraic Geom.* **18**:2 (2009), 279–308. MR 2009m:14005 Zbl 1162.14003

[Boucksom et al. 2012] S. Boucksom, J.-P. Demailly, M. Păun, and T. Peternell, "The pseudo-effective cone of a compact Kähler manifold and varieties of negative Kodaira dimension", preprint, 2012. To appear in print *J. Algebraic Geom.*

[Ein et al. 2006] L. Ein, R. Lazarsfeld, M. Mustaţă, M. Nakamaye, and M. Popa, "Asymptotic invariants of base loci", *Ann. Inst. Fourier* (*Grenoble*) **56**:6 (2006), 1701–1734. MR 2007m:14008 Zbl 1127.14010

[Ein et al. 2009] L. Ein, R. Lazarsfeld, M. Mustaţă, M. Nakamaye, and M. Popa, "Restricted volumes and base loci of linear series", *Amer. J. Math.* **131**:3 (2009), 607–651. MR 2010g:14005 Zbl 1179.14006

[Fulton 1984] W. Fulton, *Intersection theory*, Ergeb. Math. Grenzgeb. (3) **2**, Springer, Berlin, 1984. MR 85k:14004 Zbl 0541.14005

[Kawamata 1985] Y. Kawamata, "Pluricanonical systems on minimal algebraic varieties", *Invent. Math.* **79**:3 (1985), 567–588. MR 87h:14005 Zbl 0593.14010

[Lazarsfeld 2004] R. Lazarsfeld, *Positivity in algebraic geometry, II: Positivity for vector bundles, and multiplier ideals*, Ergeb. Math. Grenzgeb. (3) **49**, Springer, Berlin, 2004. MR 2005k:14001b Zbl 1093.14500

[Lazarsfeld and Mustaţă 2009] R. Lazarsfeld and M. Mustaţă, "Convex bodies associated to linear series", *Ann. Sci. Éc. Norm. Supér.* (4) **42**:5 (2009), 783–835. MR 2011e:14012 Zbl 1182.14004

[Nakamaye 2000] M. Nakamaye, "Stable base loci of linear series", *Math. Ann.* **318**:4 (2000), 837–847. MR 2002a:14008 Zbl 1063.14008

[Nakamaye 2003] M. Nakamaye, "Base loci of linear series are numerically determined", *Trans. Amer. Math. Soc.* **355**:2 (2003), 551–566. MR 2003j:14007 Zbl 1017.14017

[Nakayama 2004] N. Nakayama, *Zariski-decomposition and abundance*, MSJ Memoirs **14**, Mathematical Society of Japan, Tokyo, 2004. MR 2005h:14015 Zbl 1061.14018

[Swanson 2000] I. Swanson, "Linear equivalence of ideal topologies", *Math. Z.* **234**:4 (2000), 755–775. MR 2001f:13037 Zbl 1010.13015

Communicated by Yujiro Kawamata
Received 2011-09-20 Revised 2012-07-09 Accepted 2012-08-03

blehmann@rice.edu *Department of Mathematics, Rice University,*
6100 Main Street, Houston, TX, 77005, United States

ALGEBRA AND NUMBER THEORY 7:5 (2013)
dx.doi.org/10.2140/ant.2013.7.1101

Some consequences of a formula of Mazur and Rubin for arithmetic local constants

Jan Nekovář

We prove a very general case of the parity conjecture for Selmer groups of elliptic curves over totally real fields, as well as slightly less general results for classical modular forms, Hilbert modular forms of parallel weight two and for abelian varieties with real multiplication.

The main results of this article are the following two instances of the parity conjecture for Selmer groups (see [Nekovář 2006, Section 12.1] for a general discussion of this conjecture). Along the way we also prove slightly weaker results for Hilbert modular forms of parallel weight two with trivial character (Theorems 1.4 and 3.5) and for abelian varieties with real multiplication (Theorem 4.3).

Theorem A. *Let E be an elliptic curve over a totally real number field F and let p be a prime number. The p-Selmer rank of E over F*

$$s_p(E/F) := \mathrm{rk}_{\mathbb{Z}}\, E(F) + \mathrm{cork}_{\mathbb{Z}_p}\, Ш(E/F)[p^\infty]$$

(*which is also equal to the dimension* $\dim_{\mathbb{Q}_p} H^1_f(F, V_p(E))$ *of the Bloch–Kato Selmer group* [Bloch and Kato 1990, Definition 5.1] *of the Galois representation* $V_p(E) = T_p(E) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ *over* F) *and the analytic rank of E over F*

$$r_{\mathrm{an}}(E/F) := \mathrm{ord}_{s=1}\, L(E/F, s)$$

satisfy

$$s_p(E/F) \equiv r_{\mathrm{an}}(E/F) \pmod 2$$

in each of the following cases:

(1) *E does not have complex multiplication*,

(2) *E has complex multiplication and* $2 \nmid [F : \mathbb{Q}]$, *and*

(3) *E has complex multiplication by an imaginary quadratic field K' and p splits in $K'/\mathbb{Q}$.*

MSC2010: 11G40.
Keywords: Selmer groups, parity, elliptic curves, modular forms.

Note that potential modularity of E [Wintenberger 2009, Theorem A.1] implies that the L-function $L(E/F, s)$ has a meromorphic continuation to $\mathbb{C}$ and satisfies the expected functional equation [Taylor 2002, proof of Corollary 2.2; Nekovář 2006, 12.11.6]. As a result, the integer $\mathrm{ord}_{s=1} L(E/F, s) \in \mathbb{Z}$ is well defined.

Various special cases of Theorem A (for $F \neq \mathbb{Q}$) were proved in [Nekovář 2006; Kim 2009; Nekovář 2009].

If the p-primary part of $Ш(E/F)$ is finite for some prime number p, then $s_p(E/F) = \mathrm{rk}_{\mathbb{Z}} E(F)$ and the statement of Theorem A is the conjecture of Birch and Swinnerton-Dyer for E over F modulo 2.

Theorem B. *Let $g = \sum_{n=1}^{\infty} a_n q^n \in S_{2r}(\Gamma_0(N))$ for $r \geq 1$ be a normalised $(a_1 = 1)$ newform, and let $L = \mathbb{Q}(a_1, a_2, \dots)$ be the (totally real) number field generated by its coefficients. For any prime $\mathfrak{p}$ of L above a rational prime $p \neq 2$, denote by $V_{\mathfrak{p}}(g)$ the two-dimensional representation of $G_{\mathbb{Q}} = \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ over $L_{\mathfrak{p}}$ attached to g:*

$$\det(1 - X\,\mathrm{Fr}_{\mathrm{geom}}(l) \mid V_{\mathfrak{p}}(g)) = 1 - a_l X + l^{2r-1} X^2, \quad \textit{for all } l \nmid pN.$$

In the case when $r > 1$, assume that the residual representation of $V_{\mathfrak{p}}(g)$ is irreducible. Then

$$\dim_{L_{\mathfrak{p}}} H^1_f(\mathbb{Q}, V_{\mathfrak{p}}(g)(r)) \equiv \mathrm{ord}_{s=r} L(g, s) \pmod 2.$$

If g is (the newform associated to) a twist of a p-ordinary eigenform, Theorem B was proved in [Nekovář 2006, Theorem 12.2.3], even for $p = 2$ and without the assumption on the residual representation.

The proofs of Theorems A and B combine the techniques developed in [Nekovář 2001; 2006; 2007a; 2007b; 2008; 2009] and [Aflalo and Nekovář 2010] — namely, a combination of suitable relative parity results involving two Selmer groups with an Euler system argument [Nekovář 2007a] applied to a nontrivial Euler system [Cornut and Vatsal 2007; Aflalo and Nekovář 2010] — with a formula of Mazur and Rubin [2007, Theorem 1.4]. This formula expresses the difference of the parities of ranks of Selmer groups corresponding to two self-dual Selmer structures on a given finite (self-dual) Galois module as a finite sum of terms depending on purely local data at a finite set of (finite) primes. In a motivic setting, when the two Selmer structures are obtained by propagation from the Bloch–Kato Selmer structures for two self-dual geometric Galois representations that are congruent modulo a prime ideal dividing p, these local terms are expected to mirror the local ε-factors of the corresponding L-functions. Unfortunately, such a relation to ε-factors remains conjectural (in the required generality) even in the fairly simple situation relevant to us, when the two Galois representations come from two congruent Hilbert modular forms of parallel weight (as in Section 3). This means that we do not have at our disposal appropriate relative parity results in the generality we desire. To get around this problem we apply the formula of Mazur and Rubin in two different global situations for which the local data agree. We obtain a "birelative" global result (Theorem 2.2) for the parities of ranks of four different Selmer groups. If we are able to control three of them (in our case, Theorem 1.4 applies to two of them and the auxiliary global situation is chosen in such a way that the third Selmer group is trivial, by an application of another Euler system argument [Kato 2004; Nekovář 2012]), the sought-for parity result for the remaining Selmer group follows. Note that the formula of Mazur and Rubin is used in the proofs of both Theorems 1.1 (on which Theorem 1.4 relies) and 2.2. This program is carried out for Hilbert modular forms in Section 3; the results for abelian varieties with real multiplication are deduced in Section 4. The assumptions on E in Theorem A come from an application of [Nekovář 2012, corollary of Theorem B′].

Notation and conventions

All representations (in particular, characters) of various Galois groups are assumed to be continuous. Given a number field F, a choice of an embedding $\overline{F} \hookrightarrow \overline{F}_v$, for each prime v of F, identifies $G_{F_v} = \mathrm{Gal}(\overline{F}_v/F_v)$ with a subgroup of $G_F = \mathrm{Gal}(\overline{F}/F)$. For each representation V of G_F, we denote by V_v its restriction to G_{F_v}. Denote by S_∞ the set of all archimedean primes of F, and by S_p the set of all primes above a rational prime p of F. For any $R[G]$-module M and a character $\chi : G \to R^\times$ we denote by $M^{(\chi)} = \{m \in M \mid g(m) = \chi(g)m \text{ for all } g \in G\}$ the χ-eigenspace for the action of G on M.

1. A parity result for Hilbert modular forms of parallel weight two

Theorem 1.1 (an abstract cohomological version of the case $\mathfrak{S} = \varnothing$ of [Mazur and Rubin 2007, Theorem 7.1]). *Let F be a number field, and let V be a geometric representation* (*in the sense of Fontaine and Mazur*) *of G_F with coefficients in a finite extension $\mathcal{K}$ of $\mathbb{Q}_p$, where $p \neq 2$. Assume that*

(1) *there exists a nondegenerate skew-symmetric G_F-equivariant bilinear pairing $\langle \cdot , \cdot \rangle : V \times V \to \mathcal{K}(1)$ and*

(2) *after possibly multiplying $\langle \cdot , \cdot \rangle$ by an element of $\mathcal{K}^\times$, there exists a G_F-stable $\mathbb{O}_{\mathcal{K}}$-lattice $T \subset V$ that is self-dual* (*that is, for which the rescaled pairing defines an isomorphism $T \xrightarrow{\sim} T^*(1)$*). (*This is automatic if* $\dim_{\mathcal{K}}(V) = 2$, *for any T.*)

Let K/F be a quadratic extension, and let K' be a cyclic extension of K of p-power order, dihedral over F. Assume that no finite prime of K stable under $\mathrm{Gal}(K/F)$

ramifies in K'/K. Then, for each character $\chi : \mathrm{Gal}(K'/K) \to \mathcal{K}^\times$,

$$\dim_{\mathcal{K}} H^1_f(K', V)^{(\chi^{\pm 1})} - \dim_{\mathcal{K}} H^0(K', V)^{(\chi^{\pm 1})} \equiv \dim_{\mathcal{K}} H^1_f(K, V) - \dim_{\mathcal{K}} H^0(K, V) \pmod 2.$$

Proof. Fix a finite set S of primes of F containing $S_\infty \cup S_p$ such that V is unramified outside S. Fix a uniformiser $t \in \mathcal{O} = \mathcal{O}_{\mathcal{K}}$ and denote by $k = \mathcal{O}/t\mathcal{O}$ the residue field of $\mathcal{K}$. The $\mathcal{K}$-subspaces $H^1_f(F_v, V) \subset H^1(F_v, V)$ for $v \notin S_\infty$ define, by propagation [Mazur and Rubin 2004, Example 1.1.2], a Selmer structure $H^1_f(F_v, X) \subset H^1(F_v, X)$ on each $X = T$, V/T, T/t^nT, $\overline{T} = T/tT$, which is cartesian on $\{T/t^nT\}_{n \le \infty}$ [Mazur and Rubin 2004, Lemma 3.7.1]. The exact sequences

$$0 \to H^0(F, V/T) \otimes_{\mathcal{O}} k \to H^1_f(F, \overline{T}) \to H^1_f(F, V/T)[t] \to 0,$$
$$0 \to H^0(F_v, T) \otimes_{\mathcal{O}} k \to H^0(F_v, \overline{T}) \to H^1_f(F_v, T)[t] \to 0$$

imply that

$$\dim_k H^1_f(F, V/T)[t] - \dim_{\mathcal{K}} H^0(F, V) = \dim_k H^1_f(F, \overline{T}) - \dim_k H^0(F, \overline{T}), \quad (1.1.1)$$

and

$$\dim_k (H^1_f(F_v, \overline{T}) = H^1_f(F_v, T) \otimes_{\mathcal{O}} k) = \dim_k H^0(F_v, \overline{T}) + \dim_{\mathcal{K}} H^1_f(F_v, V) - \dim_{\mathcal{K}} H^0(F_v, V). \quad (1.1.2)$$

So far we have not used the assumptions (1) and (2) of the theorem, but we are going to do it now. The existence of a nondegenerate skew-symmetric bilinear pairing on $H^1_f(F, V/T)/(H^1_f(F, T) \otimes_{\mathcal{O}} \mathcal{K}/\mathcal{O})$ with values in $\mathcal{K}/\mathcal{O}$ constructed in [Flach 1990] (taking into account [Bloch and Kato 1990, Proposition 3.8]) implies that

$$\dim_{\mathcal{K}} H^1_f(F, V) = \mathrm{cork}_{\mathcal{O}}\, (H^1_f(F, T) \otimes_{\mathcal{O}} \mathcal{K}/\mathcal{O}) \equiv \dim_k H^1_f(F, V/T)[t] \pmod 2;$$

we deduce from (1.1.1) that

$$\dim_{\mathcal{K}} H^1_f(F, V) - \dim_{\mathcal{K}} H^0(F, V) \equiv \dim_k H^1_f(F, \overline{T}) - \dim_k H^0(F, \overline{T}) \pmod 2. \quad (1.1.3)$$

The induced representation $\mathrm{Ind}^{\mathrm{Gal}(K'/F)}_{\mathrm{Gal}(K'/K)}(\chi)$ has a natural model $I[\chi]$ (free of rank two) over $\mathcal{O}$, which is equipped with a nondegenerate symmetric G_F-equivariant pairing $I[\chi] \times I[\chi] \to \mathcal{O}$ inducing an isomorphism $I[\chi] \xrightarrow{\sim} I[\chi]^*$. By Shapiro's lemma,

$$\begin{aligned} H^1_f(F, V\otimes I[\chi]) = H^1_f(K, V\otimes\chi) &= (H^1_f(K', V)\otimes\chi)^{\mathrm{Gal}(K'/K)} \\ &= H^1_f(K', V)^{(\chi^{-1})}, \\ H^j(F, V\otimes I[\chi]) = H^j(K, V\otimes\chi) &= H^j(K', V)^{(\chi^{-1})}. \end{aligned}$$

Since $I[\chi] \xrightarrow{\sim} I[\chi^{-1}]$, these groups are respectively isomorphic to $H^1_f(K', V)^{(\chi)}$ and $H^j(K', V)^{(\chi)}$.

The discussion leading to (1.1.1)–(1.1.3) applies to $V\otimes I[\chi]$ and the self-dual lattice $T\otimes_{\mathbb{O}} I[\chi]$. Note there is a canonical identification $\overline{T}\otimes I[\chi] = \overline{T}\otimes I[1]$, where we have denoted by "1" the trivial character of $\mathrm{Gal}(K'/K)$ (this notation, which occurs only in Theorem 1.1 and Lemma 1.2, should not be confused with the Tate twist "(1)"). However, the Selmer structures $H^1_{f,\chi}(F_v, \cdot)$ and $H^1_{f,1}(F_v, \cdot)$ on the G_F-module $\overline{T}\otimes I[\chi] = \overline{T}\otimes I[1]$ obtained by propagation of the subspaces $H^1_f(F_v, V\otimes I[\chi]) \subset H^1(F_v, V\otimes I[\chi])$ and $H^1_f(F_v, V\otimes I[1]) \subset H^1(F_v, V\otimes I[1])$, respectively, are not necessarily the same. The formula [Mazur and Rubin 2007, Theorem 1.4] applies in our case, since both Selmer structures $H^1_{f,\chi}$ and $H^1_{f,1}$ are self-dual, thanks to [Bloch and Kato 1990, Proposition 3.8]; it yields

$$\dim_k H^1_{f,\chi}(F, \overline{T}\otimes I[\chi]) - \dim_k H^1_{f,1}(F, \overline{T}\otimes I[1]) \equiv \sum_{v\in S-S_\infty} \delta_v \pmod 2, \quad (1.1.4)$$

where

$$\delta_v \equiv \dim_k H^1_{f,1}(F, \overline{T}\otimes I[1])/(H^1_{f,1}(F, \overline{T}\otimes I[1]) \cap H^1_{f,\chi}(F, \overline{T}\otimes I[\chi])) \pmod 2.$$

Combining (1.1.4) with (1.1.3) for $T\otimes_{\mathbb{O}} I[\chi]$ and $T\otimes_{\mathbb{O}} I[1]$, we obtain

$$\chi_f(K, V\otimes\chi) - \chi_f(K, V) \equiv \sum_{v\in S-S_\infty} \delta_v \pmod 2, \quad (1.1.5)$$

where we have put

$$\chi_f(K, W) := \dim_{\mathcal{K}} H^1_f(K, W) - \dim_{\mathcal{K}} H^0(K, W). \quad (1.1.6)$$

To conclude the proof, it remains to prove the following lemma.

Lemma 1.2. *Under the assumptions of Theorem 1.1, we have* $\delta_v \equiv 0 \pmod 2$ *for all* $v \in S - S_\infty$.

Proof. If there is a unique prime $w \mid v$ in K, then χ_w (that is, the restriction of χ to G_{K_w}) is unramified by assumption, and therefore trivial [Mazur and Rubin 2007, Lemma 6.5]. It follows that $I[\chi]_v = I[1]_v$; hence $H^1_{f,\chi}(F_v, \overline{T}\otimes I[\chi]) = H^1_{f,1}(F_v, \overline{T}\otimes I[1])$.

The case when v splits as $v\mathcal{O}_K = ww'$ requires a more detailed argument. In this case $K_w = F_v = K_{w'}$, $I[1]_v = 1 \oplus 1$ and $I[\chi]_v = \chi_w \oplus \chi_w^{-1}$. As

$$\delta_v \equiv \dim_k\left(\frac{Y \oplus Y}{(Y \cap Z_+) \oplus (Y \cap Z_-)}\right) \pmod 2,$$

where

$$\begin{aligned} Y &= \mathrm{Im}(H^1_f(F_v, T) \otimes_{\mathcal{O}} k \hookrightarrow H^1(F_v, \overline{T})),\\ Z_\pm &= \mathrm{Im}(H^1_f(F_v, T \otimes \chi_w^{\pm 1}) \otimes_{\mathcal{O}} k \hookrightarrow H^1(F_v, \overline{T} \otimes \chi_w^{\pm 1}) = H^1(F_v, \overline{T})),\end{aligned}$$

we must show that

$$\dim_k(Y \cap Z_+) \equiv \dim_k(Y \cap Z_-) \pmod 2.$$

Firstly, the local duality

$$H^1(F_v, \overline{T}) \times H^1(F_v, \overline{T}) \to H^2(F_v, k(1)) \xrightarrow{\sim} k$$

is a nondegenerate symmetric bilinear pairing under which $Y^\perp = Y$ and $Z_\pm^\perp = Z_\mp$, by [Bloch and Kato 1990, Proposition 3.8]. Secondly, (1.1.2) applied to $T \otimes \chi_w^{\pm 1}$ yields (since $\overline{T} \otimes \chi_w^{\pm 1} = \overline{T}$)

$$\dim_k(Z_\pm) - \dim_k H^0(F_v, \overline{T}) = \dim_{\mathcal{K}} H^1_f(F_v, V \otimes \chi_w^{\pm 1}) - \dim_{\mathcal{K}} H^0(F_v, V \otimes \chi_w^{\pm 1}).$$

If $v \nmid p$, then the right-hand side is equal to zero, but if $v \mid p$, then it is equal, by [Bloch and Kato 1990, Corollary 3.8.4], to

$$\dim_{\mathcal{K}} D_{dR}(V_v \otimes \chi_w^{\pm 1})/Fil^0 = \dim_{\mathcal{K}} D_{dR}(V_v)/Fil^0,$$

which does not depend on the sign ±. In either case,

$$\dim_k(Z_+) = \dim_k(Z_-) = \tfrac{1}{2} \dim_k H^1(F_v, \overline{T}) = \dim_k(Y)$$

and

$$\begin{aligned} \dim_k(Y \cap Z_+) &= \dim_k(Y) + \dim_k(Z_+) - \dim_k(Y + Z_+)\\ &= \dim_k H^1(F_v, \overline{T}) - \dim_k(Y + Z_+)\\ &= \dim_k(Y + Z_+)^\perp = \dim_k(Y^\perp \cap Z_+^\perp) = \dim_k(Y \cap Z_-),\end{aligned}$$

as required. The lemma (and Theorem 1.1) is proved. □

1.3. If V arises as a subquotient of $H^{2r-1}_{et}(X \otimes_F \overline{F}, \mathcal{K})(r)$ for some proper and smooth scheme X over F, then $H^0(L, V) = 0$ for all finite extensions L/F, by Deligne's proof of Weil's conjectures. Theorem 1.1 in this case states that

$$\dim_{\mathcal{K}} H^1_f(K', V)^{(\chi^{\pm 1})} \equiv \dim_{\mathcal{K}} H^1_f(K, V) \pmod 2. \tag{1.3.1}$$

This remark applies, in particular, to $V = V_{\mathfrak{p}}(g)(r)$ as in Theorem B, and to any subrepresentation of $V_p(A) \otimes_{\mathbb{Q}_p} \mathcal{K}$, where A is an abelian variety over F.

Theorem 1.4 (generalisation of [Nekovář 2009, Theorem 1]). *Let $g \in S_2(\mathfrak{n}, 1)$ be a cuspidal Hilbert modular newform of parallel weight two and trivial character over a totally real number field F. Let L be the (totally real) number field generated by its Hecke eigenvalues $\lambda_v(g)$. For any prime $\mathfrak{p}$ of L above a rational prime $p \neq 2$, denote by $V_{\mathfrak{p}}(g)$ the two-dimensional representation of G_F over $L_{\mathfrak{p}}$ attached to g:*

$$\det(1 - X\,\mathrm{Fr}_{\mathrm{geom}}(v) \mid V_{\mathfrak{p}}(g)) = 1 - \lambda_v(g)X + N(v)X^2, \quad \textit{for all } v \nmid p\mathfrak{n}.$$

Assume that at least one of the following three conditions is satisfied:

(a) $2 \nmid [F : \mathbb{Q}]$,

(b) *there exists a nonarchimedean prime of F at which the local component of the automorphic representation $\pi(g)$ of $\mathrm{PGL}_2(\mathbf{A}_F)$ attached to g is a twist of the Steinberg representation, or*

(c) *there exists a nonarchimedean prime v_0 of F at which the local component of $\pi(g)$ is supercuspidal.*

Then

$$\dim_{L_{\mathfrak{p}}} H^1_f(F, V_{\mathfrak{p}}(g)(1)) \equiv r_{\mathrm{an}}(F, g) \pmod 2,$$

where $r_{\mathrm{an}}(F, g) := \mathrm{ord}_{s=1}\ L(g, s)$.

Proof. Assume either (a) or (b). In the case when g corresponds to an elliptic curve defined over F this result was proved in [Nekovář 2009]. The argument there applies in general, with the following modifications: We replace the conductor of E by $\mathfrak{n}$ (the level of g) and use Theorem 1.1 instead of [Mazur and Rubin 2007, Theorem 7.1]. As $V_{\mathfrak{p}}(g)(1)$ arises as a subrepresentation of $V_p(A) \otimes_{\mathbb{Q}_p} L_{\mathfrak{p}}$, where A is the Jacobian of a suitable Shimura curve, (1.3.1) applies in this case.

Now assume (c). Thanks to (a) we can assume that $2 \mid [F : \mathbb{Q}]$. In addition, we can assume, as in [Nekovář 2009, Step 3] (after replacing F by a suitable cyclic extension of odd degree), that there exists a prime $P \mid p$ in F, with $P \neq v_0$. Let K be any totally imaginary quadratic extension of F in which P splits and that satisfies the properties of Lemma 1.5 below (and such that g does not have CM by K). As in [Nekovář 2008, 1.2–1.5] (for $\chi = 1$, $\Sigma = \{P\}$, and $c = 1$), the generalisation of [Cornut and Vatsal 2007, Theorem 4.1] proved in [Aflalo and Nekovář 2010, Theorem 4.3.1] combined with [Nekovář 2007a, Theorem 3.2] implies that there is a finite cyclic subextension K'/K of the ring class field extension $K[P^\infty]/K$ and a character χ of $\mathrm{Gal}(K'/K)$ for which $2 \nmid \dim_{\mathcal{K}} H^1_f(K', V_{\mathfrak{p}}(g)(1))^{(\chi)}$, where

$\mathcal{K} = L_{\mathfrak{p}}(\chi)$. Theorem 1.1 then yields

$$2 \nmid \dim_{L_{\mathfrak{p}}} H^1_f(K, V_{\mathfrak{p}}(g)(1)) \\ = \dim_{L_{\mathfrak{p}}} H^1_f(F, V_{\mathfrak{p}}(g)(1)) + \dim_{L_{\mathfrak{p}}} H^1_f(F, V_{\mathfrak{p}}(g \otimes \alpha)(1)), \qquad (\star)$$

where α is the quadratic character associated to K/F. We can now vary K as in the endgame of [Nekovář 2001]:

If $2 \nmid r_{\mathrm{an}}(F, g)$, then $2 \mid r_{\mathrm{an}}(F, g \otimes \alpha)$ for any α as in Lemma 1.5 below. According to [Waldspurger 1991, Theorem 4] and [Friedberg and Hoffstein 1995, Theorem B.1] there exists such an α satisfying $r_{\mathrm{an}}(F, g \otimes \alpha) = 0$, which implies that $H^1_f(F, V_{\mathfrak{p}}(g \otimes \alpha)(1)) = 0$, by [Nekovář 2012, Theorem B(b)]; thus $2 \nmid \dim_{L_{\mathfrak{p}}} H^1_f(F, V_{\mathfrak{p}}(g)(1))$, by $(\star)$.

If $2 \mid r_{\mathrm{an}}(F, g)$, then $2 \nmid r_{\mathrm{an}}(F, g \otimes \alpha)$ for any α as in Lemma 1.5. The previous argument applies to $g \otimes \alpha$, yielding $2 \nmid \dim_{L_{\mathfrak{p}}} H^1_f(F, V_{\mathfrak{p}}(g \otimes \alpha)(1))$. Applying $(\star)$ again, we obtain $2 \mid \dim_{L_{\mathfrak{p}}} H^1_f(F, V_{\mathfrak{p}}(g)(1))$. □

Lemma 1.5. *Let g be as in Theorem 1.4(c). If $2 \mid [F : \mathbb{Q}]$, then there exists a character $\mu : G_{F_{v_0}} \to \{\pm 1\}$ such that, for any character $\alpha : G_F \to \{\pm 1\}$ satisfying*

$$\alpha_{v_0} = \mu, \qquad \alpha_v = 1 \textit{ for all } v \mid \mathfrak{n} \textit{ with } v \neq v_0, \quad \alpha_v(-1) = -1 \textit{ for all } v \in S_\infty,$$

the corresponding quadratic extension $K = \overline{F}^{\mathrm{Ker}(\alpha)}$ of F is totally imaginary and $2 \nmid r_{\mathrm{an}}(F, g) + r_{\mathrm{an}}(F, g \otimes \alpha)$.

Proof. See [Nekovář 2012, Proposition 2.10.2]. □

2. A relative parity result with a twist

2.1. Assume that V satisfies the assumption (1) of Theorem 1.1. For each nonarchimedean prime v of F we write, as in [Nekovář 2007b, Proposition 2.2.1(1)],

$$\varepsilon_v(V) = \varepsilon_v(V_v) = \varepsilon(WD(V_v), \psi, dx_\psi) \in \{\pm 1\},$$

where ψ is any nontrivial additive character of F_v, where dx_ψ is the corresponding self-dual Haar measure on F_v, and where $WD(V_v)$ is the representation of the Weil–Deligne group of F_v attached to V_v if $v \nmid p$, or to $D_{pst}(V_v)$ if $v \mid p$ (see [Deligne 1973, 8.4; Fontaine 1994; Fontaine and Perrin-Riou 1994, I.1.3.2]).

Theorem 2.2. *Let F and $\mathcal{K}$ be as in Theorem 1.1 (in particular, $p \neq 2$). Let V and V' be geometric representations of G_F with coefficients in $\mathcal{K}$ that satisfy assumptions (1) and (2) of Theorem 1.1. Let $T \subset V$ and $T' \subset V'$ be G_F-stable $\mathbb{O}$-lattices, self-dual with respect to the corresponding pairings $\langle \cdot , \cdot \rangle : T \times T \to \mathbb{O}(1)$ and $\langle \cdot , \cdot \rangle' : T' \times T' \to \mathbb{O}(1)$. Assume that there exists an isomorphism of $k[G_F]$-modules $u : \overline{T}' = T' \otimes_{\mathbb{O}} k \xrightarrow{\sim} \overline{T} = T \otimes_{\mathbb{O}} k$ compatible with the pairings induced by $\langle \cdot , \cdot \rangle$ on $\overline{T}$ and by $\langle \cdot , \cdot \rangle'$ on $\overline{T}'$. Let S be a finite set of primes of F containing $S_\infty \cup S_p$ and* *all primes at which V or V' is ramified. If $\alpha : G_F \to \{\pm 1\}$ is a character such that $\alpha_v = 1$ for all $v \in S - S_\infty$, then* (*using the notation from* (1.1.6)):

$$\chi_f(F, V) - \chi_f(F, V') \equiv \chi_f(F, V \otimes \alpha) - \chi_f(F, V' \otimes \alpha) \pmod 2,$$
$$\varepsilon_v(V)/\varepsilon_v(V') = \varepsilon_v(V \otimes \alpha)/\varepsilon_v(V' \otimes \alpha) \quad \textit{for all } v \notin S_\infty.$$

Proof. As remarked in the course of the proof of Theorem 1.1, the Selmer structure $H^1_f(F_v, \overline{T})$ obtained by propagation of $H^1_f(F_v, V) \subset H^1(F_v, V)$ is self-dual; so is the structure $H^1_{f'}(F_v, \overline{T})$ obtained by propagation of $H^1_f(F_v, V') \subset H^1(F_v, V')$, composed with the isomorphism $H^1(F_v, \overline{T}') \xrightarrow{\sim} H^1(F_v, \overline{T})$ induced by u. Combining [Mazur and Rubin 2007, Theorem 1.4] with (1.1.3) we obtain

$$\begin{aligned}\chi_f(F, V) - \chi_f(F, V') &\equiv \dim_k H^1_f(F, \overline{T}) - \dim_k H^1_{f'}(F, \overline{T}) \\ &\equiv \sum_{v \in S - S_\infty} \delta_v(T_v, T'_v) \pmod 2,\end{aligned} \tag{2.2.1}$$

where

$$\delta_v(T_v, T'_v) \equiv \dim_k H^1_f(F_v, \overline{T})/(H^1_f(F_v, \overline{T}) \cap H^1_{f'}(F, \overline{T})) \pmod 2.$$

Set $S(\alpha) = S \cup \{v \mid \alpha_v \text{ is ramified}\}$. We claim that

$$H^j(F_v, \overline{T} \otimes \alpha) = 0 \quad \text{for all } v \in S(\alpha) - S \text{ and } j = 0, 1, 2. \tag{2.2.2}$$

Indeed, $H^0(F_v, \overline{T} \otimes \alpha) \subset (\overline{T} \otimes \alpha)^{I_v} = 0$ (since $p \neq 2$) and $H^2(F_v, \overline{T} \otimes \alpha) = H^0(F_v, (\overline{T} \otimes \alpha)^*(1))^* = H^0(F_v, \overline{T} \otimes \alpha)^* = 0$, by local duality. Finally, by the local Euler characteristic formula, $H^1(F_v, \overline{T} \otimes \alpha) = 0$.

The pairings $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle'$ and the isomorphism u induce the same data for $T \otimes \alpha$ and $T' \otimes \alpha$. Applying (2.2.1) to these twisted modules, we obtain

$$\begin{aligned}\chi_f(F, V \otimes \alpha) - \chi_f(F, V' \otimes \alpha) &\equiv \sum_{v \in S(\alpha) - S_\infty} \delta_v((T \otimes \alpha)_v, (T' \otimes \alpha)_v) \\ &\equiv \sum_{v \in S - S_\infty} \delta_v((T \otimes \alpha)_v, (T' \otimes \alpha)_v) \\ &\equiv \sum_{v \in S - S_\infty} \delta_v(T_v, T'_v) \\ &\equiv \chi_f(F, V) - \chi_f(F, V') \pmod 2,\end{aligned}$$

where the second congruence follows from (2.2.2) and the third from the fact that $\alpha_v = 1$ for all $v \in S - S_\infty$.

Let us now prove the statement about local ε-constants. For $v \in S - S_\infty$ there is nothing to prove, as $(W \otimes \alpha)_v = W_v$ (here $W = V, V'$); hence $\varepsilon_v(W \otimes \alpha) = \varepsilon_v(W)$. For $v \notin S(\alpha)$ all four ε-constants are equal to 1. Finally, for $v \in S(\alpha) - S$, $\varepsilon_v(W) = 1$ ($W = V, V'$). It follows from (2.2.2) that $(W \otimes \alpha)^{I_v} = 0$, which implies that

$\varepsilon_v(W \otimes \alpha) = \varepsilon_{0,v}(W \otimes \alpha)$. As the local ε_0-constants at primes not dividing p are compatible with congruences modulo p [Deligne 1973, Theorem 6.5], the isomorphism $\overline{T}' \otimes \alpha \xrightarrow{\sim} \overline{T} \otimes \alpha$ implies that $\varepsilon_v(V \otimes \alpha), \varepsilon_v(V' \otimes \alpha) \in \{\pm 1\}$ are congruent modulo p; therefore they are equal to each other. □

2.3. In practice, we are often given a slightly different set of data:

2.3.1 representations V and V' that satisfy the assumption (1) of Theorem 1.1;

2.3.2 a G_F-stable $\mathcal{O}$-lattice $T \subset V$, self-dual with respect to $\langle\,\cdot\,,\cdot\,\rangle : T \times T \to \mathcal{O}(1)$,

2.3.3 for which $\overline{T} = T \otimes_{\mathcal{O}} k$ is an absolutely irreducible representation of G_F, and

2.3.4 a dense set of elements $g \in G_F$ for which $\mathrm{Tr}(g \mid V) \equiv \mathrm{Tr}(g \mid V') \pmod{t\mathcal{O}}$.

The condition 2.3.4 implies that, for any G_F-stable $\mathcal{O}$-lattice $T' \subset V'$, the semisimplification $\overline{T}'^{ss}$ of $\overline{T}'$ is isomorphic to $\overline{T}^{ss}$, which is in turn equal to $\overline{T}$, by condition 2.3.3. It follows that there is an isomorphism $u : \overline{T}' \xrightarrow{\sim} \overline{T}$ of $k[G_F]$-modules, which is unique up to a scalar in $k^\times$ (again by condition 2.3.3). Irreducibility of $\overline{T}'$ implies that any G_F-stable $\mathcal{O}$-lattice in V' is of the form aT' for some $a \in \mathcal{K}^\times$; as a result, T' satisfies the assumption (2) of Theorem 1.1. Finally, the pairings induced on $\overline{T}$ by $\langle\,\cdot\,,\cdot\,\rangle$ (and respectively by $\langle\,\cdot\,,\cdot\,\rangle'$ and u) coincide up to a multiplicative factor $b \in k^\times$ (by condition 2.3.3). After multiplying $\langle\,\cdot\,,\cdot\,\rangle'$ by a suitable element of $\mathcal{O}^\times$, we obtain $b = 1$. In other words, the conditions 2.3.1–2.3.4 give rise to the data required in Theorem 2.2.

3. Two applications of Theorem 2.2 to modular forms

3.1. Let F be a totally real number field. If $g \in S_k(\mathfrak{n}, 1)$ is a cuspidal Hilbert newform over F of level $\mathfrak{n}$, of trivial character and parallel weight k (necessarily even), then its completed L-function coincides, up to a shift, with the L-function of the automorphic representation $\pi(g)$ of $\mathrm{PGL}_2(\mathbf{A}_F)$ associated to g:

$$(L_\infty \cdot L)(g, s) = L(\pi(g), s - (k-1)/2), \qquad L_\infty(g, s) = \Gamma_{\mathbb{C}}(s)^{[F:\mathbb{Q}]}.$$

Since the Γ-factor $L_\infty(g, s)$ has no zero nor pole at the central point $s = k/2$ of the functional equation, the parity of the analytic rank of g over F,

$$r_{\mathrm{an}}(F, g) := \mathrm{ord}_{s=k/2}\, L(g, s),$$

can be read off from the corresponding ε-constant in the functional equation

$$L(\pi(g), s) = \varepsilon(\pi(g), s)\, L(\pi(g), 1-s),$$
$$(-1)^{r_{\mathrm{an}}(F,g)} = \varepsilon\big(\pi(g), \tfrac{1}{2}\big) = \prod_v \varepsilon_v\big(\pi(g)_v, \tfrac{1}{2}\big).$$

If L, $L_{\mathfrak{p}}$, and $V_{\mathfrak{p}}(g)$ are as in Theorem B (with an appropriate modification if $F \neq \mathbb{Q}$; see Theorem 1.4 in the case $k = 2$), then the Galois representation $V = V_{\mathfrak{p}}(g)(k/2)$

satisfies the assumption (1) of Theorem 1.1. The conjectures of Bloch and Kato [1990; Fontaine and Perrin-Riou 1994] predict that

$$\dim_{L_\mathfrak{p}} H^1_f(F, V) = r_{\mathrm{an}}(F, g).$$

We are interested in this conjecture modulo 2:

$$\dim_{L_\mathfrak{p}} H^1_f(F, V) \equiv r_{\mathrm{an}}(F, g) \pmod 2.$$

3.2. Let $g \in S_k(\mathfrak{n}, 1)$ be as in Section 3.1. If F'/F is a quadratic extension and $\alpha : \mathrm{Gal}(F'/F) \xrightarrow{\sim} \{\pm 1\}$ the corresponding quadratic character, then we have

$$H^1_f(F', V) = H^1_f(F, V) \oplus H^1_f(F, V \otimes \alpha) \tag{3.2.1}$$

and

$$\begin{aligned} L(g \otimes F', s) &= L(g, s)L(g \otimes \alpha, s), \\ r_{\mathrm{an}}(F', g) &= r_{\mathrm{an}}(F, g) + r_{\mathrm{an}}(F, g \otimes \alpha), \end{aligned} \tag{3.2.2}$$

where we have denoted, somewhat abusively, by $g' = g \otimes F'$ the base change of g to an automorphic form on $\mathrm{PGL}_2(\mathbf{A}_{F'})$ and by $r_{\mathrm{an}}(F', g)$ the analytic rank $r_{\mathrm{an}}(F', g \otimes F')$ (strictly speaking, it is the automorphic representation of $\mathrm{PGL}_2(\mathbf{A}_{F'})$ attached to g' that is the base change of $\pi(g)$).

3.3. ***Proof of Theorem B.*** The claim for $r = 1$ is a special case of Theorem 1.4(a). If $r > 1$, then it follows from [Ribet 1994, Theorems 2.1 and 2.2, Corollary 3.2] (the author would like to thank F. Diamond for pointing out this reference) and from our assumption about the residual representation of $V_\mathfrak{p}(g)$ that there exists a normalised newform $g_1 \in S_2(N_1, \omega^{2-2r})$ of level N_1 dividing pN whose coefficients lie in a number field $L' \supset L$ and that satisfies, for a suitable prime $\mathfrak{p}' \mid \mathfrak{p}$ of L',

$$\mathrm{Tr}(g \mid V_{\mathfrak{p}'}(g_1)) \equiv \mathrm{Tr}(g \mid V_\mathfrak{p}(g) \otimes_{L_\mathfrak{p}} L'_{\mathfrak{p}'}) \pmod{\mathfrak{p}'} \quad \text{for all } g \in G_\mathbb{Q}.$$

Let $g' \in S_2(N', 1)$ be the newform associated to $g_1 \otimes \omega^{r-1}$ (of level dividing N multiplied by a suitable power of p); set $\mathcal{K} = L'_{\mathfrak{p}'}$, $\mathbb{O} = \mathbb{O}_\mathcal{K}$, $V = V_\mathfrak{p}(g)(r) \otimes_{L_\mathfrak{p}} \mathcal{K}$ and $V' = V_{\mathfrak{p}'}(g')(1) = V_{\mathfrak{p}'}(g_1)(1) \otimes \omega^{r-1}$.

The representations V and V' satisfy conditions 2.3.1 and 2.3.4 (note that $\mathbb{Z}_p(r)$ and $\mathbb{Z}_p(1) \otimes \omega^{r-1}$ have the same residual representation $\mathbf{F}_p(r)$). Fix any $G_\mathbb{Q}$-stable $\mathbb{O}$-lattice $T \subset V$. It satisfies condition 2.3.3 (irreducibility implies absolute irreducibility, as the action of the complex conjugation on $\overline{T}$ has two distinct eigenvalues ± 1 contained in $k = \mathbb{O}/t\mathbb{O}$) and, after rescaling the symplectic form $\langle \cdot, \cdot \rangle : V \times V \to \mathcal{K}(1)$, also condition 2.3.2. The discussion in Section 2.3 implies that the assumptions of Theorem 2.2 are satisfied. Using, in addition, Section 1.3, we deduce that

$$\begin{aligned}\dim_{\mathcal{K}} H^1_f(\mathbb{Q}, V) &- \dim_{\mathcal{K}} H^1_f(\mathbb{Q}, V') \\ &\equiv \dim_{\mathcal{K}} H^1_f(\mathbb{Q}, V \otimes \alpha) - \dim_{\mathcal{K}} H^1_f(\mathbb{Q}, V' \otimes \alpha) \pmod 2, \qquad (3.3.1)\end{aligned}$$

whenever $\alpha : G_{\mathbb{Q}} \to \{\pm 1\}$ is a character satisfying

$$\alpha_l = 1 \quad \text{for all } l \mid pN. \qquad (3.3.2)$$

According to Theorem 1.4(a),

$$\begin{aligned}\dim_{\mathcal{K}} H^1_f(\mathbb{Q}, V') &\equiv r_{\mathrm{an}}(\mathbb{Q}, g') \qquad \pmod 2, \\ \dim_{\mathcal{K}} H^1_f(\mathbb{Q}, V' \otimes \alpha) &\equiv r_{\mathrm{an}}(\mathbb{Q}, g' \otimes \alpha) \pmod 2.\end{aligned} \qquad (3.3.3)$$

Combining (3.3.1) and (3.3.3) with Lemma 3.4 below, we obtain

$$\begin{aligned}\dim_{\mathcal{K}} H^1_f(\mathbb{Q}, V) &- r_{\mathrm{an}}(\mathbb{Q}, g) \\ &\equiv \dim_{\mathcal{K}} H^1_f(\mathbb{Q}, V \otimes \alpha) - r_{\mathrm{an}}(\mathbb{Q}, g \otimes \alpha) \pmod 2. \qquad (3.3.4)\end{aligned}$$

It follows from the nonvanishing results of [Waldspurger 1991, Theorem 4; Friedberg and Hoffstein 1995, Theorem B.1] that there exists a character α satisfying (3.3.2) for which $r_{\mathrm{an}}(\mathbb{Q}, g \otimes \alpha) = 0$. A fundamental result of Kato [2004, Theorem 14.2(2)] then implies that $H^1_f(\mathbb{Q}, V \otimes \alpha) = 0$. The congruence (3.3.4) for this particular α becomes

$$\dim_{\mathcal{K}} H^1_f(\mathbb{Q}, V) \equiv r_{\mathrm{an}}(\mathbb{Q}, g) \pmod 2,$$

which proves Theorem B.

Lemma 3.4. *For any character α satisfying (3.3.2) we have*

$$r_{\mathrm{an}}(\mathbb{Q}, g) - r_{\mathrm{an}}(\mathbb{Q}, g') \equiv r_{\mathrm{an}}(\mathbb{Q}, g \otimes \alpha) - r_{\mathrm{an}}(\mathbb{Q}, g' \otimes \alpha) \pmod 2.$$

Proof. To simplify the notation we write $\varepsilon_v(h) = \varepsilon_v\big(\pi(h)_v, \tfrac{1}{2}\big)$ for the corresponding local ε-constants. It is enough to show that, for any prime v of $\mathbb{Q}$,

$$\varepsilon_v(g)/\varepsilon_v(g') = \varepsilon_v(g \otimes \alpha)/\varepsilon_v(g' \otimes \alpha). \qquad (3.4.1)$$

Firstly, $\varepsilon_\infty(h) = \varepsilon_\infty(h \otimes \alpha)$ $(h = g, g')$, since the twist by α does not change the weight. Secondly, if l is a prime number dividing pN, then (3.3.2) implies that $\pi(h \otimes \alpha)_l = \pi(h)_l$ $(h = g, g')$; hence $\varepsilon_l(h \otimes \alpha) = \varepsilon_l(h)$. Finally, if l does not divide pN, then $\pi(g)_l = \pi(\mu, \mu^{-1})$ and $\pi(g')_l = \pi(\mu', \mu'^{-1})$ are unramified principal series representations with trivial central characters; it follows that $\pi(g \otimes \alpha) = \pi(\mu\alpha_l, \mu^{-1}\alpha_l)$, $\pi(g' \otimes \alpha) = \pi(\mu'\alpha_l, \mu'^{-1}\alpha_l)$ and

$$\begin{aligned}\varepsilon_l(g) = \mu(-1) = 1 = \mu'(-1) = \varepsilon_l(g'), \\ \varepsilon_l(g \otimes \alpha) = (\mu\alpha_l)(-1) = \alpha_l(-1) = (\mu'\alpha_l)(-1) = \varepsilon_l(g' \otimes \alpha),\end{aligned}$$

which completes the proof of (3.4.1). □

Theorem 3.5. *Let $g \in S_2(\mathfrak{n}, 1)$, L and $\mathfrak{p} \mid p$ ($p \neq 2$) be as in Theorem 1.4. Assume that $2 \mid [F : \mathbb{Q}]$, that the residual representation $T_{\mathfrak{p}}(g)/\mathfrak{p}T_{\mathfrak{p}}(g)$ (where $T_{\mathfrak{p}}(g) \subset V_{\mathfrak{p}}(g)$ is a G_F-stable $O_{L,\mathfrak{p}}$-lattice) is an irreducible G_F-module and that one of the following two conditions holds*:

1. *g has no complex multiplication and $V_{\mathfrak{p}}(g)$ is not quaternionic (in the sense of Section 3.6 below)*;
2. *g has complex multiplication: g is the theta series attached to an algebraic Hecke character $\mathbf{A}^{\times}_{K(g)} \to L'^{\times}$, where $K(g)$ and L' are totally imaginary quadratic extensions of F and L, respectively, $\mathfrak{p}$ splits in L'/L and $V_{\mathfrak{p}}(g)|_{G_{K(g)}} = \psi_1 \oplus \psi_2$, where $\psi_i : G_{K(g)} \to L^{\times}_{\mathfrak{p}}$ are characters for which $\psi_2(\mathrm{Ker}(\psi_1))$ is infinite.*

Then

$$\dim_{L_{\mathfrak{p}}} H^1_f(F, V_{\mathfrak{p}}(g)(1)) \equiv r_{\mathrm{an}}(F, g) \pmod 2.$$

Proof. As in the proof of Theorem B, the G_F-modules $V_{\mathfrak{p}}(g)(1) \supset T_{\mathfrak{p}}(g)(1)$ satisfy conditions 2.3.1–2.3.3. The level raising machinery [Taylor 1989] together with [Deligne and Serre 1974, Lemme 6.11] imply that there exists a newform $g' \in S_2(\mathfrak{n}', 1)$ of level $\mathfrak{n}'$ satisfying $\mathfrak{q} \mid \mathfrak{n}' \mid \mathfrak{n}\mathfrak{q}$ (for a suitable prime $\mathfrak{q} \nmid \mathfrak{n}$) whose Hecke eigenvalues lie in a number field $L' \supset L$ and satisfy

$$\lambda_v(g') \equiv \lambda_v(g) \pmod{\mathfrak{p}'} \quad \text{for all } v \nmid p\mathfrak{n}\mathfrak{q}$$

for a suitable prime $\mathfrak{p}' \mid \mathfrak{p}$ of L'. It follows from the Čebotarev density theorem that the representations $V = V_{\mathfrak{p}}(g)(1) \otimes_{L_{\mathfrak{p}}} \mathcal{K}$, $T = T_{\mathfrak{p}}(g)(1) \otimes_{O_{L,\mathfrak{p}}} \mathbb{O}_{\mathcal{K}}$ (where $\mathcal{K} = L'_{\mathfrak{p}'}$), and $V' = V_{\mathfrak{p}'}(g')(1)$ satisfy conditions 2.3.1–2.3.4. Applying Theorem 2.2 and taking into account Section 1.3, we obtain, for any character $\alpha : G_F \to \{\pm 1\}$ satisfying

$$\alpha_v = 1 \quad \text{for all } v \mid p\mathfrak{n}\mathfrak{q}, \tag{3.5.1}$$

that

$$\dim_{\mathcal{K}} H^1_f(F, V) - \dim_{\mathcal{K}} H^1_f(F, V') \equiv \dim_{\mathcal{K}} H^1_f(F, V \otimes \alpha) - \dim_{\mathcal{K}} H^1_f(F, V' \otimes \alpha) \pmod 2. \tag{3.5.2}$$

Since $\mathrm{ord}_{\mathfrak{q}}(\mathfrak{n}') = 1$, the local representation $\pi(g')_{\mathfrak{q}}$ is the twist of the Steinberg representation by an unramified character of order one or two. Then Theorem 1.4(b) applies to g' and its quadratic twists:

$$\begin{aligned} \dim_{\mathcal{K}} H^1_f(F, V') &\equiv r_{\mathrm{an}}(F, g') \pmod 2, \\ \dim_{\mathcal{K}} H^1_f(F, V' \otimes \alpha) &\equiv r_{\mathrm{an}}(F, g' \otimes \alpha) \pmod 2. \end{aligned} \tag{3.5.3}$$

The argument used in the proof of Lemma 3.4 applies, yielding

$$r_{\mathrm{an}}(F, g) - r_{\mathrm{an}}(F, g') \equiv r_{\mathrm{an}}(F, g \otimes \alpha) - r_{\mathrm{an}}(F, g' \otimes \alpha) \pmod 2. \tag{3.5.4}$$

Combining (3.5.2)–(3.5.4), we obtain

$$\begin{aligned} &\dim_{\mathcal{K}} H^1_f(F, V) - r_{\mathrm{an}}(F, g) \\ &\qquad \equiv \dim_{\mathcal{K}} H^1_f(F, V \otimes \alpha) - r_{\mathrm{an}}(F, g \otimes \alpha) \pmod 2, \end{aligned} \tag{3.5.5}$$

for any quadratic character α satisfying (3.5.1). As in the proof 3.3, it follows from [Waldspurger 1991, Theorem 4; Friedberg and Hoffstein 1995, Theorem B.1] that there exists α satisfying (3.5.1) such that $r_{\mathrm{an}}(F, g \otimes \alpha) = 0$. A generalisation of [Longo 2006, Theorem C] proved in [Nekovář 2012, Theorem B] implies that $H^1_f(F, V \otimes \alpha) = 0$ (this is where the assumptions (1) and (2) come in, by [Nekovář 2012, B.5.5(2) and B.6.5(2)], respectively). The congruence (3.5.5) for this α yields the desired result. □

3.6. ***(Non)quaternionic representations.*** If g from Theorem 3.5 does not have complex multiplication, recall from [Nekovář 2012, Appendix B.3] that there exists a finite abelian group $\Gamma \subset \mathrm{Aut}(L/\mathbb{Q})$ of exponent at most two and a quaternion algebra D over L^Γ such that, for each finite prime $\mathfrak{p}$ of L, the Lie algebra of the Galois image

$$\mathrm{Im}(G_F \to \mathrm{Aut}_{L_\mathfrak{p}}(V_\mathfrak{p}(g)) \xrightarrow{\sim} GL_2(L_\mathfrak{p}))$$

is equal to

$$\{x \in D_{\mathfrak{p}_\Gamma} \subset M_2(L_\mathfrak{p}) \mid \mathrm{Trd}(x) \in \mathbb{Q}_p\},$$

where $\mathfrak{p}_\Gamma$ is the prime of $L^\Gamma \subset L$ below $\mathfrak{p}$ and $D_{\mathfrak{p}_\Gamma} = D \otimes_{L^\Gamma} (L^\Gamma)_{\mathfrak{p}_\Gamma}$.

As in [Nekovář 2012, B.4.7] we say that $V_\mathfrak{p}(g)$ is *quaternionic* if $D_{\mathfrak{p}_\Gamma}$ is a division algebra (which can happen only for finitely many $\mathfrak{p}$).

According to [Nekovář 2012, B.4.8(1)], if the extension $L_\mathfrak{p}/(L^\Gamma)_{\mathfrak{p}_\Gamma}$ is unramified and the residual representation $T_\mathfrak{p}(g)/\mathfrak{p}T_\mathfrak{p}(g)$ is an irreducible G_F-module, then $V_\mathfrak{p}(g)$ is not quaternionic. In particular, the condition "$V_\mathfrak{p}(g)$ is not quaternionic" can be omitted in Theorem 3.5(1) if $L_\mathfrak{p}/(L^\Gamma)_{\mathfrak{p}_\Gamma}$ is unramified.

4. Parity results for abelian varieties with real multiplication

4.1. Let F and L be totally real number fields, and let A be an abelian variety over F satisfying

$$\dim(A) = [L : \mathbb{Q}], \qquad O_L = \mathrm{End}_F(A). \tag{4.1.1}$$

For each finite prime $\mathfrak{p}$ of L the two-dimensional $L_\mathfrak{p}$-representation $V_\mathfrak{p}(A) := T_p(A) \otimes_{O_L \otimes \mathbb{Z}_p} L_\mathfrak{p}$ of G_F satisfies the assumptions of Theorem 1.1 (with $\mathcal{K} = L_\mathfrak{p}$).

Recall that A is *modular* (over F) if there exists a cuspidal Hilbert modular newform $g \in S_2(\mathfrak{n}, 1)$ whose field of Hecke eigenvalues is equal to $\iota(L) \subset \mathbb{C}$ (for some embedding $\iota : L \hookrightarrow \mathbb{C}$) and that satisfies

$$V_{\mathfrak{p}}(A) \xrightarrow{\sim} V_{\mathfrak{p}}(g)(1)$$

for one (equivalently, for each) finite prime $\mathfrak{p}$ of L. This is, in turn, equivalent to an equality of L-functions,

$$L(\iota A/F, s) = L(g, s)$$

(Euler factor by Euler factor), which implies that

$$L(\sigma\iota A/F, s) = L({}^{\sigma}g, s) \quad \text{for all } \sigma \in \mathrm{Aut}(\mathbb{C}).$$

4.2. The potential automorphy results of [Barnet-Lamb et al. 2010, Theorems 4.5.1 and 5.3.1] imply that every abelian variety A satisfying (4.1.1) is potentially modular in the following sense: For each finite extension M/F there exists a totally real finite extension F'/F that is linearly disjoint from M/F such that $A \otimes_F F'$ is modular over F'.

As in [Nekovář 2006, 12.11.6; 2009, Step 4], a minor improvement (use of Solomon's induction theorem [Curtis and Reiner 1981, Theorem 15.10] instead of the usual Brauer theorem) of an argument of Taylor [2002, proof of Corollary 2.2] implies that there exist intermediate fields $F \subset F_i \subset F'$ and integers n_i with the following properties:

4.2.1 A is modular over each F_i: there exists a Hilbert modular newform g_i of parallel weight 2 over F_i such that $L(\iota A/F_i, s) = L(g_i, s)$ and $V_{\mathfrak{p}}(A)|_{G_{F_i}} \xrightarrow{\sim} V_{\mathfrak{p}}(g_i)(1)$ for each finite prime $\mathfrak{p}$ of L.

4.2.2 $L(\iota A/F, s) = \prod_i L(\iota A/F_i, s)^{n_i} = \prod_i L(g_i, s)^{n_i}$.

4.2.3 $V_{\mathfrak{p}}(A) = \bigoplus_i n_i \operatorname{Ind}_{G_{F_i}}^{G_F}(V_{\mathfrak{p}}(A)|_{G_{F_i}}) = \bigoplus_i n_i \operatorname{Ind}_{G_{F_i}}^{G_F}(V_{\mathfrak{p}}(g_i)(1))$ in the Grothendieck ring of $L_{\mathfrak{p}}[G_F]$-modules.

It follows that, for each $\sigma \in \mathrm{Aut}(\mathbb{C})$, the L-function

$$L(\sigma\iota A/F, s) = \prod_i L({}^{\sigma}g_i, s)^{n_i}$$

has a meromorphic continuation to $\mathbb{C}$ and satisfies the expected functional equation. In particular, the analytic rank

$$r_{\mathrm{an}}(\sigma\iota A/F) := \mathrm{ord}_{s=1} L(\sigma\iota A/F, s) \in \mathbb{Z}$$

is defined. Since the ε-constant in the functional equation of $L({}^{\sigma}g_i, s)$ does not depend on σ, the parity

$$r_{\mathrm{an}}(\tau A/F) \pmod 2 \in \mathbb{Z}/2\mathbb{Z}$$

of the analytic rank $r_{\mathrm{an}}(\tau A/F)$ does not depend on the embedding $\tau : L \hookrightarrow \mathbb{C}$.

Theorem 4.3. *Let A, F and L be as in (4.1.1). Let $\mathfrak{p}$ be a prime of L above a rational prime $p \neq 2$. Assume that at least one of the following conditions holds*:

(a) *A is modular over F and $2 \nmid [F : \mathbb{Q}]$.*

(b) *A does not have potentially good reduction everywhere.*

(c) *A does not have complex multiplication, $A[\mathfrak{p}]$ is an irreducible G_F-module, and the simple algebra $C := \mathrm{End}_{\overline{F}}(A) \otimes \mathbb{Q}$ satisfies $C \otimes_{Z(C)} Z(C)_{\mathfrak{p}_C} \xrightarrow{\sim} M_n(Z(C)_{\mathfrak{p}_C})$, where $\mathfrak{p}_C$ is the prime of $Z(C) \subset L$ below $\mathfrak{p}$ (the latter condition follows from the irreducibility of $A[\mathfrak{p}]$ if $L_{\mathfrak{p}}/Z(C)_{\mathfrak{p}_C}$ is unramified).*

(d) *A has complex multiplication by a totally imaginary quadratic extension L' of L (defined over a totally imaginary quadratic extension $K(A)$ of F), $A[\mathfrak{p}]$ is an irreducible G_F-module, $\mathfrak{p}$ splits in L'/L, and the image of $G_{K(A)}$ in $\mathrm{Aut}_{L'\otimes_L L_{\mathfrak{p}}}(V_{\mathfrak{p}}(A)) = L_{\mathfrak{p}}^{\times} \times L_{\mathfrak{p}}^{\times}$ contains an open subgroup of $\mathbb{Z}_p^{\times} \times \mathbb{Z}_p^{\times}$.*

(e) *$A[\mathfrak{p}]$ is a reducible G_F-module, $L_{\mathfrak{p}}/\mathbb{Q}_p$ is unramified and $p > 2[L_{\mathfrak{p}} : \mathbb{Q}_p] + 1$.*

Then the Selmer rank

$$\dim_{L_{\mathfrak{p}}} H^1_f(F, V_{\mathfrak{p}}(A)) = \mathrm{rk}_{O_L} A(F) + \mathrm{cork}_{O_{L,\mathfrak{p}}} Ш(A/F)[\mathfrak{p}^{\infty}]$$

satisfies

$$\dim_{L_{\mathfrak{p}}} H^1_f(F, V_{\mathfrak{p}}(A)) \equiv r_{\mathrm{an}}(\tau A/F) \pmod 2,$$

for each embedding $\tau : L \hookrightarrow \mathbb{C}$.

Proof. The case (a) follows from Theorem 1.4(a). In the cases (b)–(e) we have, thanks to Section 4.2,

$$\dim_{L_{\mathfrak{p}}} H^1_f(F, V_{\mathfrak{p}}(A)) - r_{\mathrm{an}}(\tau A/F) \equiv \sum_i n_i \big(\dim_{L_{\mathfrak{p}}} H^1_f(F_i, V_{\mathfrak{p}}(g_i)(1)) - r_{\mathrm{an}}(F_i, g_i)\big) \pmod 2,$$

which means that we can replace F by F_i and assume that A is modular over F (taking $M = F(A[\mathfrak{p}])$ in Section 4.2 we ensure that $A[\mathfrak{p}]$ is irreducible as a G_{F_i}-module in cases (c) or (d)). The case (b) then follows from Theorem 1.4(b) and the cases (c) and (d) from Theorem 3.5 (using [Nekovář 2012, B.6.5(2)]). In case (e) we can assume, thanks to Theorem 1.4(c), that $\pi(g)$ is a principal series representation at each finite prime of F, which implies that A acquires locally at each completion of F (hence also globally, by [Artin and Tate 1990, Chapter 10, Theorem 5]) good reduction over a suitable cyclic extension. The result then follows from an $O_{L,\mathfrak{p}}$-equivariant version of the proof of [Coates et al. 2010, Theorem 2.1]. □

4.4. ***Proof of Theorem A.*** As in the proof of Theorem 4.3, potential modularity of E [Wintenberger 2009, Theorem A.1] together with properties 4.2.2 and 4.2.3 imply that we can write $s_p(E/F) - r_{\mathrm{an}}(E/F)$ as an integral linear combination of $s_p(E/F_i) - r_{\mathrm{an}}(E/F_i)$, for suitable totally real extensions F_i/F over which E is modular. It is enough, therefore, to replace F by F_i and consider only the case when E is modular over F (which is automatic if E has complex multiplication).

Assume first that $p = 2$. It follows from [Waldspurger 1991, Theorem 4; Friedberg and Hoffstein 1995, Theorem B.1] that there exists a nontrivial quadratic character $\alpha : G_F \to \{\pm 1\}$ such that $r_{\mathrm{an}}(E \otimes \alpha/F) = 0$. This implies, by [Nekovář 2012, corollary of Theorem B′], that $s_2(E \otimes \alpha/F) = 0$. Let F'/F be the quadratic extension corresponding to α. Since

$$s_2(E/F') \equiv r_{\mathrm{an}}(E/F') \pmod 2$$

by [Dokchitser and Dokchitser 2011, Corollary 4.8], we conclude by the following analogue of (3.2.1) and (3.2.2):

$$s_p(E/F') = s_p(E/F) + s_p(E \otimes \alpha/F), \quad r_{\mathrm{an}}(E/F') = r_{\mathrm{an}}(E/F) + r_{\mathrm{an}}(E \otimes \alpha/F).$$

If $p \neq 2$, we can assume that $2 \mid [F : \mathbb{Q}]$, in view of [Nekovář 2009, Theorem 1(a)]. Theorem 4.3(c),(d) (respectively (e)) then implies the desired result if $E[p]$ is an irreducible G_F-module (respectively when $E[p]$ is reducible and $p > 3$). The remaining case when $p = 3$ and $E[3]$ is a reducible G_F-module is treated in [Dokchitser and Dokchitser 2011, Corollary 5.8].

4.5. Further absolute parity results (it would be too cumbersome to list them all here) follow from a combination of Theorem A with the relative parity results proved in [Mazur and Rubin 2007, Theorems 6.4 and 7.1; 2008, Theorem 1.1; Dokchitser and Dokchitser 2009, Theorems 4.3 and 4.5; 2011, Proposition 6.12; Greenberg 2011, Section 11.8; de La Rochefoucauld 2011, Theorem 2.1].

4.6. Our proof of Theorem A in the case when $E[p]$ is a reducible G_F-module uses Theorem 1.4(c), which relies on several very recent technical advances: [Aflalo and Nekovář 2010; Nekovář 2012] and [Yuan et al. 2008] (used in the proof of [Nekovář 2012, Theorem B(b)]). It would be desirable to have a more direct proof in the reducible case.[1]

4.7. The conclusion of Theorem A also holds in the case when E has complex multiplication (and hence is modular over F), $p \neq 2$ and the conductor of E is not a square, by Theorem 1.4(c) (conductors are preserved under the local Langlands correspondence and the conductor of any principal series representation of $\mathrm{PGL}_2(F_v)$ is a square).

[1] Added in proof: This is done in [Česnavičius 2012].

Acknowledgements

This work owes its origin to my stay at the Centre de Recerca Matemàtica at Universitat Autònoma de Barcelona in December 2009. I am grateful to the organisers of the program "Arithmetic Geometry" for their invitation and to the CRM for its hospitality. I thank the referee for helpful comments.

References

[Aflalo and Nekovář 2010] E. Aflalo and J. Nekovář, "Non-triviality of CM points in ring class field towers", *Israel J. Math.* **175** (2010), 225–284. MR 2011j:11108 Zbl 05789726

[Artin and Tate 1990] E. Artin and J. Tate, *Class field theory*, 2nd ed., Addison-Wesley, Redwood City, CA, 1990. MR 91b:11129 Zbl 0681.12003

[Barnet-Lamb et al. 2010] T. Barnet-Lamb, T. Gee, D. Geraghty, and R. Taylor, "Potential automorphy and change of weight", preprint, 2010. arXiv 1010.2561v1

[Bloch and Kato 1990] S. Bloch and K. Kato, "L–functions and Tamagawa numbers of motives", pp. 333–400 in *The Grothendieck Festschrift, I*, edited by P. Cartier et al., Progr. Math. **86**, Birkhäuser, Boston, MA, 1990. MR 92g:11063 Zbl 0768.14001

[Česnavičius 2012] K. Česnavičius, "The p-parity conjecture for elliptic curves with a p-isogeny", preprint, 2012. arXiv 1207.0431

[Coates et al. 2010] J. Coates, T. Fukaya, K. Kato, and R. Sujatha, "Root numbers, Selmer groups, and non-commutative Iwasawa theory", *J. Algebraic Geom.* **19**:1 (2010), 19–97. MR 2011a:11127 Zbl 1213.11135

[Cornut and Vatsal 2007] C. Cornut and V. Vatsal, "Nontriviality of Rankin–Selberg L–functions and CM points", pp. 121–186 in *L–functions and Galois representations* (Durham, 2004), edited by D. Burns et al., London Math. Soc. Lecture Note Ser. **320**, Cambridge Univ. Press, 2007. MR 2009m:11088 Zbl 1153.11025

[Curtis and Reiner 1981] C. W. Curtis and I. Reiner, *Methods of representation theory, I: With applications to finite groups and orders*, Wiley, New York, 1981. MR 82i:20001 Zbl 0469.20001

[Deligne 1973] P. Deligne, "Les constantes des équations fonctionnelles des fonctions L", pp. 501–597 in *Modular functions of one variable, II* (Antwerp, 1972), edited by P. Deligne and W. Kuyk, Lecture Notes in Math. **349**, Springer, Berlin, 1973. MR 50 #2128 Zbl 0271.14011

[Deligne and Serre 1974] P. Deligne and J.-P. Serre, "Formes modulaires de poids 1", *Ann. Sci. École Norm. Sup.* (4) **7** (1974), 507–530. MR 52 #284 Zbl 0321.10026

[Dokchitser and Dokchitser 2009] T. Dokchitser and V. Dokchitser, "Regulator constants and the parity conjecture", *Invent. Math.* **178**:1 (2009), 23–71. MR 2010j:11089 Zbl 1219.11083

[Dokchitser and Dokchitser 2011] T. Dokchitser and V. Dokchitser, "Root numbers and parity of ranks of elliptic curves", *Crelle's Journal* **658** (2011), 39–64. MR 2012h:11084 Zbl 05962772

[Flach 1990] M. Flach, "A generalisation of the Cassels–Tate pairing", *Crelle's Journal* **412** (1990), 113–127. MR 92b:11037 Zbl 0711.14001

[Fontaine 1994] J.-M. Fontaine, "Représentations l–adiques potentiellement semi-stables", pp. 321–347 in *Périodes p–adiques* (Bures-sur-Yvette, 1988), Astérisque **223**, Société Mathématique de France, Paris, 1994. MR 95k:14031 Zbl 0873.14020

[Fontaine and Perrin-Riou 1994] J.-M. Fontaine and B. Perrin-Riou, "Autour des conjectures de Bloch et Kato: Cohomologie galoisienne et valeurs de fonctions L", pp. 599–706 in *Motives* (Seattle, 1991), vol. 1, edited by U. Jannsen et al., Proc. Sympos. Pure Math. **55**, Amer. Math. Soc., 1994. MR 95j:11046 Zbl 0821.14013

[Friedberg and Hoffstein 1995] S. Friedberg and J. Hoffstein, "Nonvanishing theorems for automorphic L–functions on GL(2)", *Ann. of Math.* **142**:2 (1995), 385–423. MR 96e:11072 Zbl 0847.11026

[Greenberg 2011] R. Greenberg, *Iwasawa theory, projective modules, and modular representations*, Mem. Amer. Math. Soc. **992**, Amer. Math. Soc., 2011. MR 2807791 Zbl 1247.11085

[Kato 2004] K. Kato, "p–adic Hodge theory and values of zeta functions of modular forms", pp. 117–290 in *Cohomologies p–adiques et applications arithmétiques, III*, Astérisque **295**, Société Mathématique de France, Paris, 2004. MR 2006b:11051 Zbl 1142.11336

[Kim 2009] B. D. Kim, "The symmetric structure of the plus/minus Selmer groups of elliptic curves over totally real fields and the parity conjecture", *J. Number Theory* **129**:5 (2009), 1149–1160. MR 2010f:11089 Zbl 1170.11011

[de La Rochefoucauld 2011] T. de La Rochefoucauld, "Invariance of the parity conjecture for p–Selmer groups of elliptic curves in a D_{2p^n}–extension", *Bull. Soc. Math. France* **139**:4 (2011), 571–592. MR 2869306 Zbl 1244.11062 arXiv 1002.0554

[Longo 2006] M. Longo, "On the Birch and Swinnerton–Dyer conjecture for modular elliptic curves over totally real fields", *Ann. Inst. Fourier* (*Grenoble*) **56**:3 (2006), 689–733. MR 2008f:11071 Zbl 1152.11028

[Mazur and Rubin 2004] B. Mazur and K. Rubin, *Kolyvagin systems*, Mem. Amer. Math. Soc. **799**, Amer. Math. Soc., 2004. MR 2005b:11179 Zbl 1055.11041

[Mazur and Rubin 2007] B. Mazur and K. Rubin, "Finding large Selmer rank via an arithmetic theory of local constants", *Ann. of Math.* **166**:2 (2007), 579–612. MR 2009a:11127 Zbl 1219.11084

[Mazur and Rubin 2008] B. Mazur and K. Rubin, "Growth of Selmer rank in nonabelian extensions of number fields", *Duke Math. J.* **143**:3 (2008), 437–461. MR 2009g:11070 Zbl 1151.11023

[Nekovář 2001] J. Nekovář, "On the parity of ranks of Selmer groups, II", *C. R. Acad. Sci. Paris Sér. I Math.* **332**:2 (2001), 99–104. MR 2002e:11060 Zbl 1090.11037

[Nekovář 2006] J. Nekovář, *Selmer complexes*, Astérisque **310**, Société Mathématique de France, Paris, 2006. MR 2009c:11176 Zbl 1211.11120

[Nekovář 2007a] J. Nekovář, "The Euler system method for CM points on Shimura curves", pp. 471–547 in *L–functions and Galois representations* (Durham, 2004), edited by D. Burns et al., London Math. Soc. Lecture Note Ser. **320**, Cambridge University Press, 2007. MR 2010a:11110 Zbl 1152.11023

[Nekovář 2007b] J. Nekovář, "On the parity of ranks of Selmer groups, III", *Doc. Math.* **12** (2007), 243–274. Corrected in *Doc. Math.* **14** (2009), 191–194. MR 2009k:11109 Zbl 1201.11067

[Nekovář 2008] J. Nekovář, "Growth of Selmer groups of Hilbert modular forms over ring class fields", *Ann. Sci. Éc. Norm. Supér.* (4) **41**:6 (2008), 1003–1022. MR 2010g:11084 Zbl 1236.11047

[Nekovář 2009] J. Nekovář, "On the parity of ranks of Selmer groups, IV", *Compos. Math.* **145**:6 (2009), 1351–1359. MR 2010j:11106 Zbl 1221.11150

[Nekovář 2012] J. Nekovář, "Level raising and anticyclotomic Selmer groups for Hilbert modular forms of weight two", *Canad. J. Math.* **64**:3 (2012), 588–668. MR 2962318 Zbl 06042742

[Ribet 1994] K. A. Ribet, "Report on mod l representations of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$", pp. 639–676 in *Motives* (Seattle, 1991), vol. 2, edited by U. Jannsen et al., Proc. Sympos. Pure Math. **55**, Amer. Math. Soc., 1994. MR 95d:11056 Zbl 0822.11034

[Taylor 1989] R. Taylor, "On Galois representations associated to Hilbert modular forms", *Invent. Math.* **98**:2 (1989), 265–280. MR 90m:11176 Zbl 0705.11031

[Taylor 2002] R. Taylor, "Remarks on a conjecture of Fontaine and Mazur", *J. Inst. Math. Jussieu* **1**:1 (2002), 125–143. MR 2004c:11082 Zbl 1047.11051

[Waldspurger 1991] J.-L. Waldspurger, "Correspondances de Shimura et quaternions", *Forum Math.* **3**:3 (1991), 219–307. MR 92g:11054 Zbl 0724.11026

[Wintenberger 2009] J.-P. Wintenberger, "Potential modularity of elliptic curves over totally real fields", 2009, available at http://www-irma.u-strasbg.fr/∼wintenb/potmodcomp.pdf. Appendix to [Nekovář 2009].

[Yuan et al. 2008] X. Yuan, S.-w. Zhang, and W. Zhang, "Heights of CM points, I: Gross–Zagier formula", preprint, 2008, available at http://www.math.columbia.edu/∼szhang/papers/HCMI.pdf.

Communicated by Brian Conrad
Received 2011-10-03 Revised 2012-05-18 Accepted 2012-05-18

nekovar@math.jussieu.fr *Institut de Mathématiques de Jussieu, Université Pierre et Marie Curie (Paris 6), Théorie des Nombres, Case 247, 4, place Jussieu, F-75252 Paris cedex 05, France*

ALGEBRA AND NUMBER THEORY 7:5 (2013)
dx.doi.org/10.2140/ant.2013.7.1121

Quantized mixed tensor space and Schur–Weyl duality

Richard Dipper, Stephen Doty and Friederike Stoll

Let R be a commutative ring with 1 and q an invertible element of R. The (specialized) quantum group $\mathbf{U} = U_q(\mathfrak{gl}_n)$ over R of the general linear group acts on mixed tensor space $V^{\otimes r} \otimes V^{*\otimes s}$, where V denotes the natural $\mathbf{U}$-module R^n, r and s are nonnegative integers and V^* is the dual $\mathbf{U}$-module to V. The image of $\mathbf{U}$ in $\operatorname{End}_R(V^{\otimes r} \otimes V^{*\otimes s})$ is called the rational q-Schur algebra $S_q(n; r, s)$. We construct a bideterminant basis of $S_q(n; r, s)$. There is an action of a q-deformation $\mathfrak{B}^n_{r,s}(q)$ of the walled Brauer algebra on mixed tensor space centralizing the action of $\mathbf{U}$. We show that $\operatorname{End}_{\mathfrak{B}^n_{r,s}(q)}(V^{\otimes r} \otimes V^{*\otimes s}) = S_q(n; r, s)$. By a previous result, the image of $\mathfrak{B}^n_{r,s}(q)$ in $\operatorname{End}_R(V^{\otimes r} \otimes V^{*\otimes s})$ is $\operatorname{End}_{\mathbf{U}}(V^{\otimes r} \otimes V^{*\otimes s})$. Thus, a mixed tensor space as $(\mathbf{U}, \mathfrak{B}^n_{r,s}(q))$-bimodule satisfies Schur–Weyl duality.

Introduction

Schur–Weyl duality plays an important role in representation theory since it relates the representations of the general linear group with the representations of the symmetric group. The classical Schur–Weyl duality, due to Schur [1927], states that the actions of the general linear group $G = \mathrm{GL}_n(\mathbb{C})$ and the symmetric group $\mathfrak{S}_m$ on the tensor space $V^{\otimes m}$ with $V = \mathbb{C}^n$ satisfy the bicentralizer property, that is, $\operatorname{End}_{\mathfrak{S}_m}(V^{\otimes m})$ is generated by the action of G and correspondingly, $\operatorname{End}_G(V^{\otimes m})$ is generated by the action of $\mathfrak{S}_m$. This duality has been generalized to subgroups of G (e.g., orthogonal, symplectic groups, and Levi subgroups) and corresponding algebras related with the group algebra of the symmetric group (e.g., Brauer algebras and Ariki–Koike algebras) as well as deformations of these algebras. In general, the phrase "Schur–Weyl duality" has come to indicate such a bicentralizer property for two algebras acting on some module.

One such generalization is the mixed tensor space $V^{\otimes r} \otimes V^{*\otimes s}$, where V is the natural and V^* its dual $\mathbb{C}G$-module. The centralizer algebra is known to be the walled Brauer algebra $\mathfrak{B}^n_{r,s}$, and it was shown by Benkart, Chakrabarti, Halverson, Leduc, Lee and Stroomer [Benkart et al. 1994] that mixed tensor space under the

MSC2010: primary 33D80; secondary 16D20, 16S30, 17B37, 20C08.
Keywords: Schur–Weyl duality, walled Brauer algebra, mixed tensor space, rational q-Schur algebra.

action of $\mathbb{C}G$ and $\mathfrak{B}^n_{r,s}$ satisfies Schur–Weyl duality; see also [Koike 1989; Turaev 1989]. In [Kosuda and Murakami 1993] the authors introduced a one-parameter deformation $\mathfrak{B}^n_{r,s}(q)$ of the walled Brauer algebra and proved Schur–Weyl duality in the generic case (i.e., over $\mathbb{C}(q)$), where $\mathbb{C}G$ is replaced by the generic quantum group $U_{\mathbb{C}(q)}(\mathfrak{gl}_n)$.

In this paper, we generalize the results of [Benkart et al. 1994; Kosuda and Murakami 1993] to a very general setting. Let R be a commutative ring with 1 and $q \in R$ be invertible. Let $\mathbf{U}$ be (a specialized version of) the quantum group over R, which replaces the general linear group in the quantized case. Let $\mathfrak{B}^n_{r,s}(q)$ be the q-deformation of the walled Brauer algebra defined in [Leduc 1994]. Here we use a specialized version of Leduc's multiparameter version that acts on mixed tensor space $V^{\otimes r} \otimes V^{*\otimes s}$, where $V = R^n$ is the natural $\mathbf{U}$-module.

In [Dipper et al. 2012], one side of Schur–Weyl duality was shown in this situation, namely that the image of $\mathfrak{B}^n_{r,s}(q)$ in $\mathrm{End}_R(V^{\otimes r} \otimes V^{*\otimes s})$ is the centralizing algebra of the action of $\mathbf{U}$ on mixed tensor space.

In this paper, which is a revised version of a preprint that has circulated since 2008, the other side of Schur–Weyl duality will be proven, namely that the image of $\mathbf{U}$ in $\mathrm{End}_R(V^{\otimes r} \otimes V^{*\otimes s})$ is the endomorphism algebra of mixed tensor space under the action of $\mathfrak{B}^n_{r,s}(q)$. We call this image the *rational q-Schur algebra* and denote it $S_q(n; r, s)$. It is a q-analogue of the rational Schur algebra introduced and studied in [Dipper and Doty 2008]. In case $q = 1$, we obtain a similar statement (which is also new) for the rational Schur algebra with respect to the hyperalgebra over R of $\mathfrak{gl}_n$. In the meantime, this result was shown in [Tange 2012] in the special case $q = 1$ by different methods. One may also wish to consult [Brundan and Stroppel 2011], which enlarges the landscape on walled Brauer algebras considerably.

For technical reasons, it will be useful to turn things around and instead define $S_q(n; r, s)$ to be $\mathrm{End}_{\mathfrak{B}^n_{r,s}(q)}(V^{\otimes r} \otimes V^{*\otimes s})$. Since we show at the end that this coincides with the image of $\mathbf{U}$ in $\mathrm{End}_R(V^{\otimes r} \otimes V^{*\otimes s})$, there is no harm in this abuse of notation. In our proof, we will show that $\mathrm{End}_{\mathfrak{B}^n_{r,s}(q)}(V^{\otimes r} \otimes V^{*\otimes s}) = S_q(n; r, s)$ is free as R-module of rank independent of the choice of R and q. We shall accomplish this by constructing an R-basis of $S_q(n; r, s)$ that is dual to a certain bideterminant basis of the dual coalgebra $A_q(n; r, s)$ of $S_q(n; r, s)$.

As a guide for the reader, we briefly outline the main ideas behind the proof. There is a natural embedding of mixed tensor space $V^{\otimes r} \otimes V^{*\otimes s}$ into ordinary tensor space $V^{\otimes r+(n-1)s}$. This embedding κ is not $\mathbf{U}$-linear but is $\mathbf{U}'$-linear, where $\mathbf{U}'$ is the subalgebra of $\mathbf{U}$ corresponding to the special linear Lie algebra. We will see that replacing $\mathbf{U}$ by $\mathbf{U}'$ is not significant. For $u \in \mathbf{U}'$, the restriction of the action of u on $V^{\otimes r+(n-1)s}$ to $V^{\otimes r} \otimes V^{*\otimes s} \leq V^{\otimes r+(n-1)s}$ commutes with the action of $\mathfrak{B}^n_{r,s}(q)$ on $V^{\otimes r} \otimes V^{*\otimes s}$ and hence lies in $S_q(n; r, s)$. Thus, κ induces an algebra homomorphism π from the ordinary q-Schur algebra $S_q(n, r + (n-1)s)$, which is the image of $\mathbf{U}'$ in $\mathrm{End}_R(V^{\otimes r+(n-1)s})$ into $S_q(n; r, s)$. This homomorphism was motivated by a similar homomorphism in [Dipper and Doty 2008].

Let $\rho_{\mathrm{ord}} : \mathbf{U}' \to S_q(n, r + (n-1)s)$ be the representation of $\mathbf{U}'$ on $V^{\otimes r+(n-1)s}$ and $\rho_{\mathrm{mxd}} : \mathbf{U}' \to S_q(n; r, s)$ the representation of $\mathbf{U}'$ on mixed tensor space. Then $\rho_{\mathrm{mxd}} = \pi \circ \rho_{\mathrm{ord}}$ by construction. By classical quantized Schur–Weyl duality, ρ_{ord} is surjective, so ρ_{mxd} is surjective (i.e., $\rho_{\mathrm{mxd}}(\mathbf{U}') = S_q(n; r, s)$) if π is surjective. We show that π possesses an R-linear right inverse, thus proving the surjectivity of π.

At this point, we switch over to coefficient spaces. It is well known that the dual coalgebra $A_q(n, r + (n-1)s) = S_q(n, r + (n-1)s)^*$ is the coefficient space of $\mathbf{U}'$ acting on ordinary tensor space $V^{\otimes r+(n-1)s}$. There is no problem here with dualization since the classical q-Schur algebra $S_q(n, r + (n-1)s)$ is known to be free as R-module of fixed rank independent of the choice of R and q. Moreover, $A_q(n, r + (n-1)s)$ possesses a bideterminant basis [Huang and Zhang 1993]. The endomorphism algebra $S_q(n; r, s) = \mathrm{End}_{\mathfrak{B}^n_{r,s}(q)}(V^{\otimes r} \otimes V^{*\otimes s})$ may be described by a system of linear equations in the endomorphism algebra $\mathrm{End}_R(V^{\otimes r} \otimes V^{*\otimes s})$, which is free as R-module. Using these equations, we apply a general argument (Lemma 2.3) to construct a factor coalgebra $A_q(n; r, s)$ of the R-coalgebra $\mathrm{End}_R(V^{\otimes r} \otimes V^{*\otimes s})$ such that $A_q(n; r, s)^*$ is isomorphic to the R-algebra $S_q(n; r, s)$. In Section 5, we exhibit a map $\iota : A_q(n; r, s) \to A_q(n, r + (n-1)s)$ and show explicitly that $\iota^* = \pi : S_q(n, r + (n-1)s) \to S_q(n; r, s)$. In Section 6, we show that $A_q(n; r, s)$ and hence $S_q(n; r, s)$ are free as R-module by constructing a (rational) bideterminant basis. From this it is not hard to find an (R-linear) left inverse of the map ι whose dual map is then the required right inverse of $\iota^* = \pi$, proving that $S_q(n; r, s)$ is the image of $\mathbf{U}'$ (and hence $\mathbf{U}$) acting on mixed tensor space.

1. Preliminaries

Let n be a given positive integer. In this section, we introduce the quantized enveloping algebra of the general linear Lie algebra $\mathfrak{gl}_n$ over a commutative ring R with parameter q and summarize some well known results; see for example [Hong and Kang 2002; Jantzen 1996; Lusztig 1990]. We will start by recalling the definition of the quantized enveloping algebra over $\mathbb{Q}(q)$, where q is an indeterminate.

Let $P^\vee$ be the free $\mathbb{Z}$-module with basis $h_1, \dots, h_n$, and let $\varepsilon_1, \dots, \varepsilon_n \in P^{\vee *}$ be the corresponding dual basis: ε_i is given by $\varepsilon_i(h_j) := \delta_{i,j}$ for $j = 1, \dots, n$, where δ is the usual Kronecker symbol. For $i = 1, \dots, n-1$, let $\alpha_i \in P^{\vee *}$ be defined by $\alpha_i := \varepsilon_i - \varepsilon_{i+1}$.

Definition 1.1. The quantum general linear algebra $U_q(\mathfrak{gl}_n)$ is the associative $\mathbb{Q}(q)$-algebra with 1 generated by the elements e_i, f_i ($i = 1, \dots, n-1$) and q^h ($h \in P^\vee$) with the defining relations

$$q^0 = 1, \qquad q^h q^{h'} = q^{h+h'}, \qquad q^h e_i q^{-h} = q^{\alpha_i(h)} e_i, \qquad q^h f_i q^{-h} = q^{-\alpha_i(h)} f_i,$$

$$e_i f_j - f_j e_i = \delta_{i,j}\frac{K_i - K_i^{-1}}{q-q^{-1}}, \quad \text{where } K_i := q^{h_i - h_{i+1}},$$
$$e_i^2 e_j - (q+q^{-1})e_i e_j e_i + e_j e_i^2 = 0 \quad \text{for } |i-j| = 1,$$
$$f_i^2 f_j - (q+q^{-1})f_i f_j f_i + f_j f_i^2 = 0 \quad \text{for } |i-j| = 1,$$
$$e_i e_j = e_j e_i \quad \text{and} \quad f_i f_j = f_j f_i \quad \text{for } |i-j| > 1.$$

We note that the subalgebra generated by the K_i, e_i and f_i $(i = 1, \dots, n-1)$ is isomorphic with $U_q(\mathfrak{sl}_n)$. Also, $U_q(\mathfrak{gl}_n)$ is a Hopf algebra with comultiplication Δ, counit ε the unique algebra homomorphisms and antipode S the unique invertible antihomomorphism of algebras, defined on generators by

$$\Delta(q^h) = q^h \otimes q^h,$$
$$\Delta(e_i) = e_i \otimes K_i^{-1} + 1 \otimes e_i, \qquad \Delta(f_i) = f_i \otimes 1 + K_i \otimes f_i,$$
$$\varepsilon(q^h) = 1, \qquad \varepsilon(e_i) = \varepsilon(f_i) = 0,$$
$$S(q^h) = q^{-h}, \qquad S(e_i) = -e_i K_i, \qquad S(f_i) = -K_i^{-1} f_i.$$

Let $V_{\mathbb{Q}(q)}$ be a free $\mathbb{Q}(q)$-vector space with basis $\{v_1, \dots, v_n\}$. We make $V_{\mathbb{Q}(q)}$ a $U_q(\mathfrak{gl}_n)$-module via

$$q^h v_j = q^{\varepsilon_j(h)} v_j \quad \text{for } h \in P^\vee \text{ and } j = 1, \dots, n,$$
$$e_i v_j = \begin{cases} v_i & \text{if } j = i+1, \\ 0 & \text{otherwise,} \end{cases} \qquad f_i v_j = \begin{cases} v_{i+1} & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

We call $V_{\mathbb{Q}(q)}$ the *vector representation* of $U_q(\mathfrak{gl}_n)$. This is also a $U_q(\mathfrak{sl}_n)$-module by restriction of the action.

Let $[l]_q$ in $\mathbb{Z}[q, q^{-1}]$ (or in R) be defined by

$$[l]_q := \sum_{i=0}^{l-1} q^{2i-l+1}$$

and set $[l]_q! := [l]_q[l-1]_q \cdots [1]_q$. Define the divided powers $e_i^{(l)} := e_i^l/[l]_q!$ and $f_i^{(l)} := f_i^l/[l]_q!$. Let $\mathbf{U}_{\mathbb{Z}[q,q^{-1}]}$ (resp. $\mathbf{U}'_{\mathbb{Z}[q,q^{-1}]}$) be the $\mathbb{Z}[q, q^{-1}]$-subalgebra of $U_q(\mathfrak{gl}_n)$ generated by the q^h (resp. the K_i) and the $e_i^{(l)}$ and $f_i^{(l)}$ for $l \geq 0$. Then $\mathbf{U}_{\mathbb{Z}[q,q^{-1}]}$ is a Hopf algebra, and we have

$$\Delta(e_i^{(l)}) = \sum_{k=0}^{l} q^{k(l-k)} e_i^{(l-k)} \otimes K_i^{k-l} e_i^{(k)}, \quad \Delta(f_i^{(l)}) = \sum_{k=0}^{l} q^{-k(l-k)} f_i^{(l-k)} K_i^k \otimes f_i^{(k)},$$
$$S(e_i^{(l)}) = (-1)^l q^{l(l-1)} e_i^{(l)} K_i^l, \qquad S(f_i^{(l)}) = (-1)^l q^{-l(l-1)} K_i^{-l} f_i^{(l)},$$
$$\varepsilon(e_i^{(l)}) = \varepsilon(f_i^{(l)}) = 0.$$

Furthermore, the $\mathbb{Z}[q,q^{-1}]$-lattice $V_{\mathbb{Z}[q,q^{-1}]}$ in $V_{\mathbb{Q}(q)}$ generated by the v_i is invariant under the action of $\mathbf{U}_{\mathbb{Z}[q,q^{-1}]}$ and of $\mathbf{U}'_{\mathbb{Z}[q,q^{-1}]}$. Now, make the transition from $\mathbb{Z}[q,q^{-1}]$ to an arbitrary commutative ring R with 1. Let $q \in R$ be invertible, and consider R as a $\mathbb{Z}[q,q^{-1}]$-module via specializing $q \in \mathbb{Z}[q,q^{-1}] \mapsto q \in R$.

Let $\mathbf{U}_R := R \otimes_{\mathbb{Z}[q,q^{-1}]} \mathbf{U}_{\mathbb{Z}[q,q^{-1}]}$ and $\mathbf{U}'_R := R \otimes_{\mathbb{Z}[q,q^{-1}]} \mathbf{U}'_{\mathbb{Z}[q,q^{-1}]}$. Then $\mathbf{U}_R$ inherits a Hopf algebra structure from $\mathbf{U}_{\mathbb{Z}[q,q^{-1}]}$, and $V_R := R \otimes_{\mathbb{Z}[q,q^{-1}]} V_{\mathbb{Z}[q,q^{-1}]}$ is a $\mathbf{U}_R$-module and by restriction also a $\mathbf{U}'_R$-module.

If no ambiguity arises, we will henceforth omit the index R and write $\mathbf{U}$, $\mathbf{U}'$ and V instead of $\mathbf{U}_R$, $\mathbf{U}'_R$ and V_R. Furthermore, we will write $e_i^{(l)}$ as shorthand for $1 \otimes e_i^{(l)} \in \mathbf{U}_R$, similarly for the $f_i^{(l)}$, K_i for $1 \otimes K_i$ and q^h for $1 \otimes q^h$.

Suppose W, W_1 and W_2 are $\mathbf{U}$-modules; then one can define $\mathbf{U}$-module structures on $W_1 \otimes W_2 = W_1 \otimes_R W_2$ and $W^* = \operatorname{Hom}_R(W, R)$ using the comultiplication and the antipode by setting $x(w_1 \otimes w_2) = \Delta(x)(w_1 \otimes w_2)$ and $(xf)(w) = f(S(x)w)$.

Definition 1.2. Let r and s be nonnegative integers. The $\mathbf{U}$-module $V^{\otimes r} \otimes V^{*\otimes s}$ is called *mixed tensor space*.

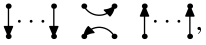
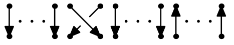
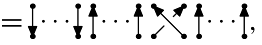
Let $I(n,r)$ be the set of r-tuples with entries in $\{1,\dots,n\}$, and let $I(n,s)$ be defined similarly. The elements of $I(n,r)$ (and $I(n,s)$) are called *multi-indices*. Note that the symmetric groups $\mathfrak{S}_r$ and $\mathfrak{S}_s$ act on $I(n,r)$ and $I(n,s)$ respectively from the right by place permutation, that is, if $\boldsymbol{i} = (i_1, i_2, \dots)$ is a multi-index and s_j is a Coxeter generator, then let $\mathbf{i}.s_j := (i_1, \dots, i_{j-1}, i_{j+1}, i_j, i_{j+2}, \dots)$. Then a basis of the mixed tensor space $V^{\otimes r} \otimes V^{*\otimes s}$ can be indexed by $I(n,r) \times I(n,s)$. For $\boldsymbol{i} = (i_1, \dots, i_r) \in I(n,r)$ and $\boldsymbol{j} = (j_1, \dots, j_s) \in I(n,s)$, let

$$v_{\boldsymbol{i}|\mathbf{j}} := v_{i_1} \otimes \cdots \otimes v_{i_r} \otimes v_{j_1}^* \otimes \cdots \otimes v_{j_s}^* \in V^{\otimes r} \otimes V^{*\otimes s},$$

where $\{v_1^*, \dots, v_n^*\}$ is the basis of V^* dual to $\{v_1, \dots, v_n\}$. Then $\{\, v_{\boldsymbol{i}|\boldsymbol{j}} : \boldsymbol{i} \in I(n,r), \boldsymbol{j} \in I(n,s) \,\}$ is a basis of $V^{\otimes r} \otimes V^{*\otimes s}$.

We have another algebra acting on $V^{\otimes r} \otimes V^{*\otimes s}$, namely the quantized walled Brauer algebra $\mathfrak{B}_{r,s}^n(q)$ introduced in [Dipper et al. 2012]. This algebra is defined as a diagram algebra in terms of Kauffman's tangles. A presentation by generators and relations can be found in [Dipper et al. 2012]. Note that this algebra and its action coincide with Leduc's algebra [1994] (see the remarks in [Dipper et al. 2012]).

Here, all we need is the action of generators given in the following diagrams. The Brauer algebra $\mathfrak{B}_{r,s}^n(q)$ is generated by the elements

$E =$ , $S_i =$  and $\hat{S}_j =$ ,

where the nonpropagating edges in E connect vertices in columns r and $r+1$ while the crossings in S_i and $\hat{S}_j$ connect vertices in columns i and $i+1$ and columns $r+j$ and $r+j+1$, respectively. If $v_{\boldsymbol{i}|\boldsymbol{j}} = v \otimes v_{i_r} \otimes v_{j_1}^* \otimes v'$, then the action of the generators on $V^{\otimes r} \otimes V^{*\otimes s}$ is given by

$$v_{\boldsymbol{i}|\boldsymbol{j}} E = \delta_{i_r, j_1} \sum_{s=1}^{n} q^{2i_r - n - 1} v \otimes v_s \otimes v_s^* \otimes v',$$

$$v_{\boldsymbol{i}|\boldsymbol{j}} S_i = \begin{cases} q^{-1} v_{\boldsymbol{i}|\boldsymbol{j}} & \text{if } i_i = i_{i+1}, \\ v_{\boldsymbol{i}.s_i|\boldsymbol{j}} & \text{if } i_i < i_{i+1}, \\ v_{\boldsymbol{i}.s_i|\boldsymbol{j}} + (q^{-1} - q) v_{\boldsymbol{i}|\boldsymbol{j}} & \text{if } i_i > i_{i+1}, \end{cases}$$

$$v_{\boldsymbol{i}|\boldsymbol{j}} \hat{S}_j = \begin{cases} q^{-1} v_{\boldsymbol{i}|\boldsymbol{j}} & \text{if } j_j = j_{j+1}, \\ v_{\boldsymbol{i}|\boldsymbol{j}.s_j} & \text{if } j_j > j_{j+1}, \\ v_{\boldsymbol{i}|\boldsymbol{j}.s_j} + (q^{-1} - q) v_{\boldsymbol{i}|\boldsymbol{j}} & \text{if } j_j < j_{j+1}. \end{cases}$$

The action of $\mathfrak{B}^n_{r,s}(q)$ on $V^{\otimes r} \otimes V^{*\otimes s}$ commutes with the action of $\mathbf{U}$.

Theorem 1.3 [Dipper et al. 2012]. *Let $\sigma : \mathfrak{B}^n_{r,s}(q) \to \operatorname{End}_{\mathbf{U}}(V^{\otimes r} \otimes V^{*\otimes s})$ be the representation of the quantized walled Brauer algebra on the mixed tensor space. Then σ is surjective, that is,*

$$\operatorname{End}_{\mathbf{U}}(V^{\otimes r} \otimes V^{*\otimes s}) \cong \mathfrak{B}^n_{r,s}(q)/_{\operatorname{ann}_{\mathfrak{B}^n_{r,s}(q)}(V^{\otimes r} \otimes V^{*\otimes s})}.$$

The main result of this paper is the other half of the preceding theorem.

Theorem 1.4. *Let $\rho_{\mathrm{mxd}} : \mathbf{U} \to \operatorname{End}_{\mathfrak{B}^n_{r,s}(q)}(V^{\otimes r} \otimes V^{*\otimes s})$ be the representation of the quantum group. Then ρ_{mxd} is surjective, that is,*

$$\operatorname{End}_{\mathfrak{B}^n_{r,s}(q)}(V^{\otimes r} \otimes V^{*\otimes s}) \cong \mathbf{U}/_{\operatorname{ann}_{\mathbf{U}}(V^{\otimes r} \otimes V^{*\otimes s})}.$$

Theorems 1.3 and 1.4 together state that the mixed tensor space is a $(\mathbf{U}, \mathfrak{B}^n_{r,s}(q))$-bimodule with the double centralizer property. In the literature, this is also called *Schur–Weyl Duality*. Theorem 1.4 will be proved at the end of this paper.

For $s = 0$, this is well known; $\mathfrak{B}^n_{m,0}(q)$ is the Hecke algebra $\mathcal{H}_m$, and $V^{\otimes m}$ is the (ordinary) tensor space.

Definition 1.5. If m is a positive integer, let $\mathcal{H}_m$ be the associative R-algebra with 1 generated by elements $T_1, \dots, T_{m-1}$ with respect to the relations

$$(T_i + q)(T_i - q^{-1}) = 0 \quad \text{for } i = 1, \dots, m-1,$$

$$T_i T_{i+1} T_i = T_{i+1} T_i T_{i+1} \quad \text{for } i = 1, \dots, m-2,$$

$$T_i T_j = T_j T_i \quad \text{for } |i - j| \geq 2.$$

If $w \in \mathfrak{S}_m$ is an element of the symmetric group on m letters and $w = s_{i_1} s_{i_2} \dots s_{i_l}$ is a reduced expression as a product of Coxeter generators, let $T_w := T_{i_1} T_{i_2} \dots T_{i_l}$. Then the set $\{ T_w : w \in \mathfrak{S}_m \}$ is a basis of $\mathcal{H}_m$.

Note that $\mathcal{H}_m$ acts on $V^{\otimes m}$ since $\mathcal{H}_m \cong \mathfrak{B}^n_{m,0}(q)$, the isomorphism given by $T_i \mapsto S_i$.

Theorem 1.6 [Dipper and James 1989; Green 1996]. *Let $\rho_{\mathrm{ord}} : \mathbf{U} \to \mathrm{End}_R(V^{\otimes m})$ be the representation of* $\mathbf{U}$ *on $V^{\otimes m}$. Then* $\operatorname{im} \rho_{\mathrm{ord}} = \mathrm{End}_{\mathscr{H}_m}(V^{\otimes m})$. *This algebra is called the q-Schur algebra and denoted $S_q(n, m)$.*

We will refer to $V^{\otimes m}$ as ordinary tensor space.

2. Mixed tensor space as a submodule

Recall that $\mathbf{U}'$ is the subalgebra of $\mathbf{U}$ corresponding to the Lie algebra $\mathfrak{sl}_n$.

Theorem 2.1. *If m is a nonnegative integer, let $\rho_{\mathrm{ord}} : \mathbf{U} \to \mathrm{End}_R(V^{\otimes m})$ be the representation of* $\mathbf{U}$ *on $V^{\otimes m}$. Then*

$$\rho_{\mathrm{ord}}(\mathbf{U}) = \rho_{\mathrm{ord}}(\mathbf{U}').$$

Proof. Define the *weight* of $\boldsymbol{i} \in I(n, m)$ to be $\mathrm{wt}(\boldsymbol{i}) := \lambda = (\lambda_1, \dots, \lambda_n)$ such that λ_i is the number of entries in $\boldsymbol{i}$ that are equal to i. If $\lambda = (\lambda_1, \dots, \lambda_n)$ is a composition of m into n parts, i.e., $\lambda_1 + \cdots + \lambda_n = m$, let $V_\lambda^{\otimes m}$ be the R-submodule of $V^{\otimes m}$ generated by all $v_{\mathbf{i}}$ with $\mathrm{wt}(\boldsymbol{i}) = \lambda$. Then $V^{\otimes m}$ is the direct sum of all $V_\lambda^{\otimes m}$, where λ runs through the set of compositions of m into n parts. Let φ_λ be the projection onto $V_\lambda^{\otimes m}$. By [Green 1996], the restriction of $\rho_{\mathrm{ord}} : \mathbf{U} \to S_q(n, m)$ to any subalgebra $\mathbf{U}' \subseteq \mathbf{U}$ is surjective if the subalgebra $\mathbf{U}'$ contains the divided powers $e_i^{(l)}$ and $f_i^{(l)}$ and preimages of the projections φ_λ.

Therefore, we define a partial order on the set of compositions of m into n parts by $\lambda \preceq \mu$ if and only if

$$(\lambda_1 - \lambda_2, \lambda_2 - \lambda_3, \dots, \lambda_{n-1} - \lambda_n) \leq (\mu_1 - \mu_2, \mu_2 - \mu_3, \dots, \mu_{n-1} - \mu_n)$$

in the lexicographical order. It suffices to show that for each composition λ, there exists an element $u \in \mathbf{U}'$ such that $uv_{\boldsymbol{i}} = 0$ whenever $\mathrm{wt}(\boldsymbol{i}) \prec \lambda$ (i.e., $\mathrm{wt}(\boldsymbol{i}) \preceq \lambda$ and $\mathrm{wt}(\boldsymbol{i}) \neq \lambda$) and $uv_{\boldsymbol{i}} = v_{\boldsymbol{i}}$ whenever $\mathrm{wt}(\boldsymbol{i}) = \lambda$. In Theorem 4.5 of [Lusztig 1990], it is shown that certain elements

$$\begin{bmatrix} K_i; c \\ t \end{bmatrix} := \prod_{s=1}^{t} \frac{K_i q^{c-s+1} - K_i^{-1} q^{-c+s-1}}{q^s - q^{-s}}$$

are elements of $\mathbf{U}'$ for $i = 1, \dots, n-1$, $c \in \mathbb{Z}$ and $t \in \mathbb{N}$. Let

$$u := \prod_{i=1}^{n-1} \begin{bmatrix} K_i; m+1 \\ \lambda_i - \lambda_{i+1} + m + 1 \end{bmatrix},$$

which is an element of $\mathbf{U}'$ since $\lambda_i - \lambda_{i+1} + m + 1 > 0$. Then u has the desired properties. □

The next lemma is motivated by [Dipper and Doty 2008, §6.3].

Lemma 2.2. *There is a well defined* $\mathbf{U}'$*-monomorphism* $\kappa : V^* \to V^{\otimes n-1}$ *given by*

$$\begin{aligned} v_i^* \mapsto (-q)^i & \sum_{w \in \mathfrak{S}_{n-1}} (-q)^{l(w)} v_{(12\ldots\hat{\imath}\ldots n).w} \\ = (-q)^i & \sum_{w \in \mathfrak{S}_{n-1}} (-q)^{l(w)} v_{(12\ldots\hat{\imath}\ldots n)} T_w = (-q)^i v_{(12\ldots\hat{\imath}\ldots n)} \sum_{w \in \mathfrak{S}_{n-1}} (-q)^{l(w)} T_w, \end{aligned}$$

where $\hat{\imath}$ *means leaving out* i*.*

Proof. Clearly κ is a monomorphism of R-modules, and $K_i v_j^* = q^{\delta_{i+1,j}-\delta_{i,j}} v_j^*$ and $K_i v_{(1\ldots\hat{\jmath}\ldots n)} = q^{1-\delta_{i,j}} q^{\delta_{i+1,j}-1} v_{(1\ldots\hat{\jmath}\ldots n)}$ by definition. Thus, κ commutes with K_i. Now $e_i v_j^* = -\delta_{i,j} q^{-1} v_{j+1}^*$. If $j \neq i, i+1$, then

$$\begin{aligned} e_i \kappa(v_j^*) &= (-q)^j e_i \sum_w (-q)^{l(w)} v_{(1\ldots i i+1 \ldots\hat{\jmath}\ldots n)} T_w \\ &= -(-q)^j \sum_w (-q)^{l(w)} v_{(1\ldots i i \ldots\hat{\jmath}\ldots n)} T_w = 0 = \kappa(e_i v_j^*). \end{aligned}$$

For $j = i$ (resp. $i+1$), we get

$$\begin{aligned} e_i \kappa(v_{i+1}^*) &= (-q)^{i+1} \sum_w (-q)^{l(w)} (e_i v_{(1\ldots\widehat{i+1}\ldots n)}) T_w = 0, \\ e_i \kappa(v_i^*) &= (-q)^i \sum_w (-q)^{l(w)} (e_i v_{(1\ldots\hat{\imath} i+1\ldots n)}) T_w \\ &= (-q)^i \sum_w (-q)^{l(w)} v_{(1\ldots i\,\widehat{i+1}\ldots n)} T_w = -q^{-1} \kappa(v_{i+1}^*). \end{aligned}$$

Furthermore, for $l \geq 2$ we clearly have $e_i^{(l)} v_j^* = 0$ and $e_i^{(l)} \kappa(v_j^*) = 0$. The argument for f_i works similarly. □

Lemma 2.2 enables us to consider the mixed tensor space $V^{\otimes r} \otimes V^{*\otimes s}$ as a $\mathbf{U}'$-submodule $T^{r,s}$ of $V^{\otimes r+(n-1)s}$ via an embedding that we will also denote κ. Thus, $\mathfrak{B}_{r,s}^n(q)$ acts on $T^{r,s}$.

If we restrict the action of an element of $\mathbf{U}'$ on $V^{\otimes r+(n-1)s}$ or equivalently of the q-Schur algebra $S_q(n, r+(n-1)s)$ to $T^{r,s}$, then we get an element of $\operatorname{End}_R(T^{r,s})$. Since the actions of $\mathbf{U}'$ and $\mathfrak{B}_{r,s}^n(q)$ commute, this is also an element of $\operatorname{End}_{\mathfrak{B}_{r,s}^n(q)}(T^{r,s})$. Let $S_q(n; r, s) := \operatorname{End}_{\mathfrak{B}_{r,s}^n(q)}(V^{\otimes r} \otimes V^{*\otimes s})$; thus, we have an algebra homomorphism $\pi : S_q(n, r+(n-1)s) \to S_q(n; r, s)$ by restriction of the action to $T^{r,s} \cong V^{\otimes r} \otimes V^{*\otimes s}$. Our aim is to show that π is surjective, for then each element of $\operatorname{End}_{\mathfrak{B}_{r,s}^n(q)}(V^{\otimes r} \otimes V^{*\otimes s})$ is given by the action of an element of $\mathbf{U}'$.

Lemma 2.3. *Let* M *be a free* R*-module with basis* $\mathcal{B} = \{b_1, \ldots, b_l\}$ *and* U *a submodule of* M *given by a set of linear equations on the coefficients with respect to the basis* $\mathcal{B}$*, i.e.,* $a_{ij} \in R$ *such that* $U = \left\{ \sum c_i b_i \in M : \sum_j a_{ij} c_j = 0 \text{ for all } i \right\}$

exist. Let $\{b_1^*, \dots, b_l^*\}$ *be the basis of* $M^* = \operatorname{Hom}_R(M, R)$ *dual to* $\mathcal{B}$, *and let* X *be the submodule generated by all* $\sum_j a_{ij} b_j^*$. *Then* $U \cong (M^*/X)^*$.

Proof. We have that $(M^*/X)^*$ is isomorphic to the submodule of M^{**} given by linear forms on M^* that vanish on X. Via the natural isomorphism $M^{**} \cong M$, this is isomorphic to the set of elements of M that are annihilated by X. An element $m = \sum_k c_k b_k$ is annihilated by X if and only if $0 = \sum_{j,k} a_{ij} b_j^*(c_k b_k) = \sum_k a_{ik} c_k$ for all i, and this is true if and only if $m \in U$. □

Note an element $\tilde{\varphi} \in (M^*/X)^*$ corresponds to the element $\varphi = \sum_i \tilde{\varphi}(b_i^* + X) b_i$ of U. In our case, $S_q(n, m)$ and $S_q(n; r, s)$ are R-submodules of R-free algebras, namely $\operatorname{End}_R(V^{\otimes m})$ and $\operatorname{End}_R(V^{\otimes r} \otimes V^{*\otimes s})$ respectively, given by a set of linear equations, which we will determine more precisely in Sections 3 and 4.

Definition 2.4. Let $M := \operatorname{End}_R(V^{\otimes m})$ and $U := S_q(n, m)$. Then U is defined as the algebra of endomorphisms commuting with a certain set of endomorphisms and thus is given by a system of linear equations on the coefficients. Let $A_q(n, m) := M^*/X$ as in Lemma 2.3. Similarly, let $A_q(n; r, s) := M^*/X$ with $M := \operatorname{End}_R(V^{\otimes r} \otimes V^{*\otimes s})$ and $U := S_q(n; r, s)$.

By Lemma 2.3, $A_q(n, m)^* = S_q(n, m)$ and $A_q(n; r, s)^* = S_q(n; r, s)$. We will proceed as follows. We will take $m = r + (n-1)s$ and define an R-homomorphism $\iota : A_q(n; r, s) \to A_q(n, r+(n-1)s)$ so that $\iota^* = \pi : S_q(n, r+(n-1)s) \to S_q(n; r, s)$. Then we will define an R-homomorphism $\phi : A_q(n, r+(n-1)s) \to A_q(n; r, s)$ such that $\phi \circ \iota = \mathrm{id}_{A_q(n;r,s)}$ by giving suitable bases for $A_q(n, r+(n-1)s)$ and $A_q(n; r, s)$. Dualizing this equation, we get $\pi \circ \phi^* = \iota^* \circ \phi^* = \mathrm{id}_{S_q(n;r,s)}$, and this shows that π is surjective. Actually, $A_q(n, r+(n-1)s)$ and $A_q(n; r, s)$ are coalgebras, and ι is a morphism of coalgebras, but we do not need this for our results.

3. $A_q(n, m)$

The description of $A_q(n, m)$ is well known; see, e.g., [Dipper and Donkin 1991]. Let $A_q(n)$ be the free R-algebra on generators x_{ij} $(1 \le i, j \le n)$ subject to the relations

$$\begin{aligned}
x_{ik}x_{jk} &= q x_{jk}x_{ik} && \text{if } i < j,\\
x_{ki}x_{kj} &= q x_{kj}x_{ki} && \text{if } i < j,\\
x_{ij}x_{kl} &= x_{kl}x_{ij} && \text{if } i < k \text{ and } j > l,\\
x_{ij}x_{kl} &= x_{kl}x_{ij} + (q - q^{-1}) x_{il}x_{kj} && \text{if } i < k \text{ and } j < l.
\end{aligned}$$

Note that these relations define the commutative algebra in n^2 commuting indeterminates x_{ij} in case $q = 1$. The free algebra on the generators x_{ij} is obviously graded (with all generators in degree 1), and since the relations are homogeneous, this induces a grading on $A_q(n)$. Then we have the following lemma:

Lemma 3.1 [Dipper and Donkin 1991]. *$A_q(n,m)$ is the R-submodule of $A_q(n)$ of elements of homogeneous degree m.*

Proof. Since our relations of the Hecke algebra differ from those in [Dipper and Donkin 1991] ($(T_i - q)(T_i + 1) = 0$ is replaced by $(T_i + q)(T_i - q^{-1}) = 0$) and thus $A_q(n,m)$ differs as well, we include a proof here.

Suppose φ is an endomorphism of $V^{\otimes m}$ commuting with the action of a generator S_i. For convenience, we assume that $m = 2$ and $S = S_1$. Then φ can be written as a linear combination of the basis elements $E_{(ij),(kl)}$ mapping $v_k \otimes v_l$ to $v_i \otimes v_j$ and all other basis elements to 0. For the coefficient of $E_{(ij),(kl)}$, we write $c_{ik}c_{jl}$ so that $\varphi = \sum_{i,j,k,l} c_{ik}c_{jl}E_{(ij),(kl)}$. On the one hand, we have

$$\begin{aligned}
&S(\varphi(v_k \otimes v_l))\\
&= S\biggl(\sum_{i,j} c_{ik}c_{jl}v_i \otimes v_j\biggr)\\
&= \sum_{i<j} c_{ik}c_{jl}v_j \otimes v_i + q^{-1}\sum_{i} c_{ik}c_{il}v_i \otimes v_i + \sum_{i>j} c_{ik}c_{jl}(v_j \otimes v_i + (q^{-1} - q)v_i \otimes v_j)\\
&= \sum_{i\neq j} c_{ik}c_{jl}v_j \otimes v_i + q^{-1}\sum_{i} c_{ik}c_{il}v_i \otimes v_i + (q^{-1} - q)\sum_{i<j} c_{jk}c_{il}v_j \otimes v_i.
\end{aligned}$$

Now, suppose that $k > l$. Then

$$\begin{aligned}
\varphi(S(v_k \otimes v_l)) &= \varphi(v_l \otimes v_k + (q^{-1} - q)v_k \otimes v_l)\\
&= \sum_{i,j}(c_{jl}c_{ik} + (q^{-1} - q)c_{jk}c_{il})v_j \otimes v_i.
\end{aligned}$$

Similar formulas hold for $k = l$ and $k < l$. Comparing coefficients leads to the relations given above. □

$A_q(n,m)$ has a basis consisting of monomials, but it will turn out to be more convenient for our purposes to work with a basis of standard bideterminants; see [Huang and Zhang 1993]. In that reference, the supersymmetric quantum letterplace algebra for $L^- = P^- = \{1, \dots, n\}$ and $L^+ = P^+ = \varnothing$ is isomorphic to $A_{q^{-1}}(n) \cong A_q(n)^{\mathrm{opp}}$, and we will adjust the results to our situation.

A *partition* λ of m is a sequence $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k)$ of nonnegative integers such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ and $\sum_{i=1}^{k} \lambda_i = m$. Denote the set of partitions of m by $\Lambda^+(m)$. The *Young diagram* $[\lambda]$ of a partition λ is $\{\, (i, j) \in \mathbb{N} \times \mathbb{N} : 1 \leq i \leq k,\ 1 \leq j \leq \lambda_i \,\}$. It can be represented by an array of boxes: λ_1 boxes in the first row, λ_2 boxes in the second row, etc.

A λ*-tableau* $\mathfrak{t}$ is a map $f : [\lambda] \to \{1, \dots, n\}$. A tableau can be represented by writing the entry $f(i, j)$ into the (i, j)th box. A tableau $\mathfrak{t}$ is called *standard* if the entries in each row are strictly increasing from left to right and the entries in each column are nondecreasing downward. In the literature, this property is also called semistandard, and the role of rows and columns may be interchanged. Note that if $\mathfrak{t}$ is a standard λ-tableau, then $\lambda_1 \leq n$. A pair $[\mathfrak{t}, \mathfrak{t}']$ of λ-tableaux is called a *bitableau*. It is standard if both $\mathfrak{t}$ and $\mathfrak{t}'$ are standard λ-tableaux.

Note that the next definition differs from the definition in [Huang and Zhang 1993] by a sign.

Definition 3.2. Let $i_1, \dots, i_k, j_1, \dots, j_k \in \{1, \dots, n\}$. For $i_1 < i_2 < \dots < i_k$, let the *right quantum minor* be defined by

$$(i_1 i_2 \dots i_k | j_1 j_2 \dots j_k)_r := \sum_{w \in \mathfrak{S}_k} (-q)^{l(w)} x_{i_{w1} j_1} x_{i_{w2} j_2} \dots x_{i_{wk} j_k}.$$

For arbitrary $i_1, \dots, i_k$, the right quantum minor is then defined by the rule

$$(i_1 \dots i_l i_{l+1} \dots i_k | j_1 j_2 \dots j_k)_r := -q^{-1} (i_1 \dots i_{l-1} i_{l+1} i_l i_{l+2} \dots i_k | j_1 j_2 \dots j_k)_r$$

for $i_l > i_{l+1}$. Similarly, let the *left quantum minor* be defined by

$$(i_1 \dots i_k | j_1 \dots j_k)_l := \sum_{w \in \mathfrak{S}_k} (-q)^{l(w)} x_{i_1, j_{w1}} x_{i_2 j_{w2}} \dots x_{i_k j_{wk}} \text{ if } j_1 < \dots < j_k,$$

$$(i_1 \dots i_k | j_1 \dots j_k)_l := -q^{-1} (i_1 \dots i_k | j_1 \dots j_{l+1} j_l \dots j_k)_l \text{ if } j_l > j_{l+1}.$$

Finally, let the *quantum determinant* be defined by

$$\det_q := (12 \dots n | 12 \dots n)_r = (12 \dots n | 12 \dots n)_l.$$

If $[\mathfrak{t}, \mathfrak{t}']$ is a bitableau and $\mathfrak{t}_1, \mathfrak{t}_2, \dots, \mathfrak{t}_k$ (resp. $\mathfrak{t}'_1, \mathfrak{t}'_2, \dots, \mathfrak{t}'_k$) are the rows of $\mathfrak{t}$ (resp. $\mathfrak{t}'$), then let

$$(\mathfrak{t} | \mathfrak{t}') := (\mathfrak{t}_k | \mathfrak{t}'_k)_r \dots (\mathfrak{t}_2 | \mathfrak{t}'_2)_r (\mathfrak{t}_1 | \mathfrak{t}'_1)_r.$$

Then $(\mathfrak{t} | \mathfrak{t}')$ is called a *bideterminant*.

Remark 3.3. We note the following properties of quantum minors:

(1) $(i_1 \dots i_k | j_1 \dots j_k)_r = -q (i_1 \dots i_k | j_1 \dots j_{l+1} j_l \dots j_k)_r$ for $j_l > j_{l+1}$,
$(i_1 \dots i_k | j_1 \dots j_k)_l = -q (i_1 \dots i_{l+1} i_l \dots i_k | j_1 \dots j_k)_l$ for $i_l > i_{l+1}$.

(2) If $i_1 < i_2 < \dots < i_k$ and $j_1 < j_2 < \dots < j_k$, then right and left quantum minors coincide, and we simply write $(i_1 \dots i_k | j_1 \dots j_k)$. This notation thus indicates that the sequences of numbers are increasing. In general, right and left quantum minors differ by a power of $-q$.

(3) If two i_ls or j_ls coincide, then the quantum minors vanish.

(4) The quantum determinant $\det_q$ is an element of the center of $A_q(n)$.

Definition 3.4. Let the *content* of a monomial $x_{i_1 j_1} \dots x_{i_m j_m}$ be defined as the tuple $(\alpha, \beta) = ((\alpha_1, \dots, \alpha_n), (\beta_1, \dots, \beta_n))$, where α_i is the number of indices i_t such that $i_t = i$ and β_j is the number of indices j_t such that $j_t = j$. Note that $\sum \alpha_i = \sum \beta_j = m$ for each monomial of homogeneous degree m. For such a tuple (α, β), let $P(\alpha, \beta)$ be the subspace of $A_q(n, m)$ generated by the monomials of content (α, β). Furthermore, let the *content* of a bitableau $[\mathfrak{t}, \mathfrak{t}']$ be defined similarly as the tuple (α, β) such that α_i is the number of entries in $\mathfrak{t}$ equal to i and β_j is the number of entries in $\mathfrak{t}'$ equal to j.

Theorem 3.5 [Huang and Zhang 1993]. *The bideterminants* $(\mathfrak{t}|\mathfrak{t}')$ *of the standard* λ*-tableaux with* λ *a partition of* m *form a basis of* $A_q(n, m)$ *such that the bideterminants of standard* λ*-tableaux of content* (α, β) *form a basis of* $P(\alpha, \beta)$.

The proof in [Huang and Zhang 1993] works over a field, but the arguments are valid if the field is replaced by a commutative ring with 1. The reversed order of the minors is due to the opposite algebra. Note that for $i_1 < i_2 < \cdots < i_k$ and $j_1 < j_2 < \cdots < j_k$, we have

$$q^{k(k-1)/2}(i_1 i_2 \dots i_k | j_1 j_2 \dots j_k)_r = \sum_{w \in \mathfrak{S}_k} (-q)^{-l(w)} x_{i_{wk} j_1} x_{i_{w(k-1)} j_2} \dots x_{i_{w1} j_k},$$

which is a quantum minor of $A_{q^{-1}}(n)^{\mathrm{opp}}$.

Lemma 3.6 (Laplace's expansion [Huang and Zhang 1993]).

(1) *For* $j_1 < j_2 < \cdots < j_l < j_{l+1} < \cdots < j_k$, *we have*

$$\begin{aligned}(i_1 i_2 \dots i_k | j_1 j_2 \dots j_k)_l \\ = \sum_w (-q)^{l(w)} (i_1 \dots i_l | j_{w1} \dots j_{wl})_l (i_{l+1} \dots i_k | j_{w(l+1)} \dots j_{wk})_l,\end{aligned}$$

where the summation is over all $w \in \mathfrak{S}_k$ *such that* $w1 < w2 < \cdots < wl$ *and* $w(l+1) < w(l+2) < \cdots < wk$.

(2) *For* $i_1 < i_2 < \cdots < i_k$, *we have*

$$\begin{aligned}(i_1 i_2 \dots i_k | j_1 j_2 \dots j_k)_r \\ = \sum_w (-q)^{l(w)} (i_{w1} \dots i_{wl} | j_1 \dots j_l)_r (i_{w(l+1)} \dots i_{wk} | j_{l+1} \dots j_k)_r,\end{aligned}$$

the summation again over all $w \in \mathfrak{S}_k$, *such that* $w1 < w2 < \cdots < wl$ *and* $w(l+1) < w(l+2) < \cdots < wk$.

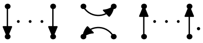
4. $A_q(n; r, s)$

A basis of $\mathrm{End}_R(V^{\otimes r} \otimes V^{*\otimes s})$ is given by matrix units $E_{\boldsymbol{i}|\boldsymbol{j}\ \boldsymbol{k}|\boldsymbol{l}}$ such that $E_{\boldsymbol{i}|\boldsymbol{j}\ \boldsymbol{k}|\boldsymbol{l}} v_{\mathbf{s}|\boldsymbol{t}} = \delta_{\boldsymbol{k}|\boldsymbol{l},\boldsymbol{s}|\mathbf{t}} v_{\boldsymbol{i}|\boldsymbol{j}}$. Suppose $\varphi := \sum_{\mathbf{i},\boldsymbol{j},\boldsymbol{k},\boldsymbol{l}} c_{i|j\ \mathbf{k}|l} E_{\boldsymbol{i}|\boldsymbol{j}\ \boldsymbol{k}|\boldsymbol{l}} \in \mathrm{End}_R(V^{\otimes r} \otimes V^{*\otimes s})$ commutes with the action of $\mathfrak{B}_{r,s}^n(q)$ or equivalently with a set of generators of $\mathfrak{B}_{r,s}^n(q)$. Since coefficient spaces are multiplicative, we can write

$$c_{i_1k_1}c_{i_2k_2}\cdots c_{i_rk_r}c^*_{j_1l_1}c^*_{j_2l_2}\cdots c^*_{j_sl_s}$$

for the coefficient $c_{\boldsymbol{i}|\boldsymbol{j}\;\boldsymbol{k}|\boldsymbol{l}}$. It is easy to see from the description of $A_q(n,m)$ that φ commutes with the generators without nonpropagating edges if and only if the c_{ij} satisfy the relations of $A_q(n)$ and the c^*_{ij} satisfy the relations of $A_{q^{-1}}(n)\cong A_q(n)^{\mathrm{opp}}$.

Now suppose that φ in addition commutes with the action of the generator

$e=$ .

We assume $\varphi=\sum_{i,j,k,l=1}^n c_{ik}c^*_{jl}E_{i|j\;k|l}$ and that $r=s=1$ (the general case being similar). Let $v=v_i\otimes v_j^*$ be a basis element of $V\otimes V^*$. We have (the indices in the sums always run from 1 to n)

$$\varphi(v)e=\sum_{s,t}c_{si}c^*_{tj}(v_s\otimes v_t^*)e=\sum_{s,k}q^{2s-n-1}c_{si}c^*_{sj}(v_k\otimes v_k^*),$$

$$\varphi(ve)=\delta_{ij}q^{2i-n-1}\sum_{k}\varphi(v_k\otimes v_k^*)=\delta_{ij}q^{2i-n-1}\sum_{k,s,t}c_{sk}c^*_{tk}v_s\otimes v_t^*.$$

Comparing coefficients, we get the following conditions:

$$\sum_{k=1}^n c_{ik}c^*_{jk}=0\quad\text{for } i\neq j,$$

$$\sum_{k=1}^n q^{2k}c_{ki}c^*_{kj}=0\quad\text{for } i\neq j,$$

$$\sum_{k=1}^n q^{2k-2i}c_{ki}c^*_{ki}=\sum_{k=1}^n c_{jk}c^*_{jk}.$$

This, combined with Lemma 2.3, shows the following:

Lemma 4.1. *We have*

$$A_q(n;r,s)\cong(F(n,r)\otimes_R F_*(n,s))/Y,$$

where $F(n,r)$ (resp. $F_*(n,s)$) is the R-submodule of the free algebra on generators x_{ij} (resp. x^*_{ij}) generated by monomials of degree r (resp. s) and Y is the R-submodule of $F(n,r)\otimes_R F_*(n,s)$ generated by elements of the form $h_1h_2h_3$, where h_2 is one of the elements

$$x_{ik}x_{jk}-qx_{jk}x_{ik}\qquad \textit{for } i<j, \tag{4.1.1}$$

$$x_{ki}x_{kj}-qx_{kj}x_{ki}\qquad \textit{for } i<j, \tag{4.1.2}$$

$$x_{ij}x_{kl}-x_{kl}x_{ij}\qquad \textit{for } i<k,\ j>l, \tag{4.1.3}$$

$$x_{ij}x_{kl} - x_{kl}x_{ij} - (q-q^{-1})x_{il}x_{kj} \quad \textit{for } i<k,\ j<l, \tag{4.1.4}$$

$$x^*_{ik}x^*_{jk} - q^{-1}x^*_{jk}x^*_{ik} \quad \textit{for } i<j, \tag{4.1.5}$$

$$x^*_{ki}x^*_{kj} - q^{-1}x^*_{kj}x^*_{ki} \quad \textit{for } i<j, \tag{4.1.6}$$

$$x^*_{ij}x^*_{kl} - x^*_{kl}x^*_{ij} \quad \textit{for } i<k,\ j>l, \tag{4.1.7}$$

$$x^*_{ij}x^*_{kl} - x^*_{kl}x^*_{ij} + (q-q^{-1})x^*_{il}x^*_{kj} \quad \textit{for } i<k,\ j<l, \tag{4.1.8}$$

$$\sum_{k=1}^{n} x_{ik}x^*_{jk} \quad \textit{for } i\neq j, \tag{4.1.9}$$

$$\sum_{k=1}^{n} q^{2k}x_{ki}x^*_{kj} \quad \textit{for } i\neq j, \tag{4.1.10}$$

$$\sum_{k=1}^{n} q^{2k-2i}x_{ki}x^*_{ki} - \sum_{k=1}^{n} x_{jk}x^*_{jk} \tag{4.1.11}$$

and h_1 and h_3 are monomials of appropriate degree.

Remark 4.2. The map given by $x_{ik} \mapsto q^{2k-2i}x_{ki}$ and $x^*_{ik} \mapsto x^*_{ki}$ induces an R-linear automorphism of $A_q(n; r, s)$.

Bideterminants can also be formed using the variables x^*_{ij}. In this case, let

$$(\mathfrak{t}|\mathfrak{t}')^* := (\mathfrak{t}_1|\mathfrak{t}'_1)^*_r(\mathfrak{t}_2|\mathfrak{t}'_2)^*_r \cdots (\mathfrak{t}_k|\mathfrak{t}'_k)^*_r,$$

where the quantum minors $(i_1 \dots i_k | j_1 \dots j_k)^*_{r/l}$ are defined as above with q replaced by q^{-1}.

5. The map $\iota : A_q(n; r, s) \to A_q(n, r+(n-1)s)$

For any $1 \le i, j \le n$, let $\iota(x_{ij}) := x_{ij}$ and

$$\iota(x^*_{ij}) := (-q)^{j-i}(12\dots\hat{\imath}\dots n|12\dots\hat{\jmath}\dots n) \in A_q(n, n-1);$$

then there is a unique R-linear map

$$\iota : F(n, r) \otimes_R F_*(n, s) \to A_q(n, r+(n-1)s)$$

such that $\iota(x_{i_1j_1}\cdots x_{i_rj_r}x^*_{k_1l_1}\cdots x^*_{k_sl_s}) = \iota(x_{i_1j_1})\cdots\iota(x_{i_rj_r})\iota(x^*_{k_1l_1})\cdots\iota(x^*_{k_sl_s})$.

Lemma 5.1. *The kernel of ι contains Y, and thus, ι induces an R-linear map*

$$A_q(n; r, s) \to A_q(n, r+(n-1)s),$$

which we will then also denote ι.

Proof. We have to show that the generators of Y lie in the kernel of ι. Generators of Y involving the elements (4.1.1)–(4.1.4) are obviously in the kernel of ι. Theorem 7.3 of [Goodearl 2006] shows that generators involving elements (4.1.5)–(4.1.8) are also in the kernel. Laplace's expansion shows that

$$\begin{aligned}\iota\Bigl(\sum_{k=1}^{n} x_{ik}x^*_{jk}\Bigr) &= \sum_{k=1}^{n}(-q)^{(k-1)-(j-1)}x_{ik}\cdot(1\dots\hat{\jmath}\dots n|1\dots\hat{k}\dots n)_l\\ &= (-q)^{1-j}(i1\dots\hat{\jmath}\dots n|1\dots n)_l = \delta_{i,j}\cdot\det{}_q,\\ \iota\Bigl(\sum_{k=1}^{n} q^{2k-2i}x_{ki}x^*_{kj}\Bigr) &= q^{-2i+j+1}\sum_{k=1}^{n}(-q)^{k-1}x_{ki}\cdot(1\dots\hat{k}\dots n|1\dots\hat{\jmath}\dots n)_r\\ &= (-q)^{j-2i+1}(1\dots n|i1\dots\hat{\jmath}\dots n)_r = \delta_{i,j}\cdot\det{}_q;\end{aligned}$$

thus, the generators involving the elements (4.1.9)–(4.1.11) are in the kernel of ι. □

Now, we have maps

$$\iota^*: A_q(n, r+(n-1)s)^* \to A_q(n; r, s)^* \quad\text{and}\quad \pi: S_q(n, r+(n-1)s) \to S_q(n; r, s).$$

By definition, $A_q(n, r+(n-1)s)^* \cong S_q(n, r+(n-1)s)$ and $A_q(n; r, s)^* \cong S_q(n; r, s)$.

Lemma 5.2. *Under the identifications above, we have* $\iota^* = \pi$.

Proof. We will write

$$\begin{aligned}x_{i_1\dots i_l\ j_1\dots j_l} &= x_{i_1,j_1}\cdots x_{i_l,j_l},\\ x_{i_l\dots i_1|l_1\dots l_m\ j_l\dots j_1|k_1\dots k_m} &= x_{i_l,j_l}\cdots x_{i_1,j_1}x^*_{l_1,k_1}\cdots x^*_{l_m,k_m}.\end{aligned}$$

Suppose that $\tilde{\varphi} \in A_q(n, r+(n-1)s)^*$. Then

$$\varphi = \sum_{\boldsymbol{i},\boldsymbol{j}\in \mathrm{I}(n,r+(n-1)s)} \tilde{\varphi}(x_{\mathbf{ij}})E_{\boldsymbol{ij}}$$

is the corresponding element of $S_q(n, r+(n-1)s)$. Since $\iota^*(\tilde{\varphi}) = \tilde{\varphi}\circ\iota$, we have

$$\iota^*(\varphi) = \sum_{\boldsymbol{i},\boldsymbol{j},\boldsymbol{k},\boldsymbol{l}} \tilde{\varphi}\circ\iota(x_{\boldsymbol{i}|\boldsymbol{j}\,\boldsymbol{k}|\boldsymbol{l}})E_{\boldsymbol{i}|\boldsymbol{j}\,\boldsymbol{k}|\boldsymbol{l}}.$$

In other words, the coefficient of $E_{\boldsymbol{i}|\boldsymbol{j}\,\boldsymbol{k}|\boldsymbol{l}}$ in $\iota^*(\varphi)$ can be computed by substituting each $x_{\mathbf{st}}$ in $\iota(x_{\boldsymbol{i}|\boldsymbol{j}\,\boldsymbol{k}|\boldsymbol{l}})$ by $\tilde{\varphi}(x_{\mathbf{st}})$. On the other hand, to compute the coefficient of $E_{\boldsymbol{i}|\boldsymbol{j}\,\boldsymbol{k}|\boldsymbol{l}}$ in $\pi(\varphi)$, one has to consider the action of φ on a basis element $v = \kappa(v_{\boldsymbol{k}|\boldsymbol{l}})$ of $T^{r,s}$. For a multi-index $\boldsymbol{l} \in \mathrm{I}(n, s)$, let $\boldsymbol{l}^* \in \mathrm{I}(n, (n-1)s)$ be defined by

$$\boldsymbol{l}^* := (1\dots\widehat{l_1}\dots n1\dots\widehat{l_2}\dots n\dots1\dots\widehat{l_s}\dots n).$$

Then

$$v = \kappa(v_{\boldsymbol{k}|\boldsymbol{l}}) = (-q)^{l_1+l_2+\cdots+l_s} \sum_{w \in \mathfrak{S}_{n-1}^{\times s}} (-q)^{l(w)} v_{\boldsymbol{k}} \otimes (v_{\boldsymbol{l}^*} T_w),$$

and thus, we have

$$\begin{aligned}\varphi(v) &= (-q)^{\sum l_k} \sum_{\boldsymbol{s},\boldsymbol{t},w} (-q)^{l(w)} \tilde{\varphi}(x_{\boldsymbol{st}}) E_{\boldsymbol{st}}(v_{\boldsymbol{k}} \otimes (v_{\boldsymbol{l}^*} T_w)) \\ &= \sum_{\boldsymbol{s},w} (-q)^{l(w)+\sum l_k} \tilde{\varphi}(x_{\boldsymbol{s}\ \boldsymbol{k}\boldsymbol{l}^*_{\cdot} w}) v_{\boldsymbol{s}}.\end{aligned}$$

Since φ leaves $T^{r,s}$ invariant, $\varphi(v)$ is a linear combination of the basis elements $\kappa(v_{\boldsymbol{i}|\boldsymbol{j}})$ of $T^{r,s}$. Distinct $\kappa(v_{\boldsymbol{i}|\boldsymbol{j}})$ involve distinct basis vectors of $V^{\otimes r+(n-1)s}$. Thus, if

$$\varphi(v) = \sum_{\boldsymbol{i}|\boldsymbol{j}} \lambda_{\boldsymbol{i}|\boldsymbol{j}} \kappa(v_{\boldsymbol{i}|\boldsymbol{j}}) = \sum_{\boldsymbol{i}|\boldsymbol{j},w} \lambda_{\boldsymbol{i}|\boldsymbol{j}} (-q)^{l(w)+j_1+\cdots+j_s} v_{\boldsymbol{i}\boldsymbol{j}^*.w},$$

then $(-q)^{\sum j_k} \lambda_{\boldsymbol{i}|\boldsymbol{j}}$ is equal to the coefficient of $v_{\boldsymbol{i}\boldsymbol{j}^*}$ when $\varphi(v)$ is written as a linear combination of basis vectors of $V^{\otimes r+(n-1)s}$. The coefficient of $v_{\boldsymbol{i}\boldsymbol{j}^*}$ in $\varphi(v)$ is, by the formula above,

$$(-q)^{\sum l_k} \sum_{w} (-q)^{l(w)} \tilde{\varphi}(x_{\boldsymbol{i}\boldsymbol{j}^*\ \boldsymbol{k}\boldsymbol{l}^*_{\cdot} w}).$$

Thus,

$$\lambda_{\boldsymbol{i}|\boldsymbol{j}} = (-q)^{\sum l_k - j_k} \sum_{w} (-q)^{l(w)} \tilde{\varphi}(x_{\boldsymbol{i}\boldsymbol{j}^*\ \boldsymbol{k}\boldsymbol{l}^*_{\cdot} w}) = \tilde{\varphi} \circ \iota(x_{\boldsymbol{i}|\boldsymbol{j}\ \boldsymbol{k}|\boldsymbol{l}}).$$

But $\lambda_{\boldsymbol{i}|\boldsymbol{j}}$ is also the coefficient of $E_{\boldsymbol{i}|\boldsymbol{j}\ \boldsymbol{k}|\boldsymbol{l}}$ in $\pi(\varphi)$, which shows the result. □

Theorem 5.3 (Jacobi's ratio theorem). *Suppose $n \geq l \geq 0$ and $i_1 < i_2 < \cdots < i_l$ and $j_1 < j_2 < \cdots < j_l$. Let $i'_1 < i'_2 < \cdots < i'_{n-l}$ and $j'_1 < j'_2 < \cdots < j'_{n-l}$ be the unique numbers such that $\{1, \ldots, n\} = \{i_1, \ldots, i_l, i'_1, \ldots, i'_{n-l}\} = \{j_1, \ldots, j_l, j'_1, \ldots, j'_{n-l}\}$. Then*

$$\iota((i_1 \ldots i_l | j_1 \ldots j_l)^*) = (-q)^{\sum_{t=1}^{l}(j_t - i_t)} \det{}_q^{l-1}(i'_1 \ldots i'_{n-l} | j'_1 \ldots j'_{n-l}).$$

Proof. We argue by induction on l. Note that for $l = 0$, $\det_q^{l-1} = \det_q^{-1}$ is not an element of $A_q(n)$. However, $(i'_1 \ldots i'_{n-l} | j'_1 \ldots j'_{n-l})$ turns out to be $\det_q$; thus, the right-hand side of the formula is $\det_q^{-1} \det_q = 1 = \iota(1)$. In this sense, the formula is valid for $l = 0$.

For $l = 1$, the theorem is true by the definition of $\iota(x^*_{ij})$. Now assume the theorem is true for $l - 1$. Apply Laplace's expansion and use induction to get

$$\iota((i_1 \ldots i_l | j_1 \ldots j_l)^*) = \iota\left(\sum_{k=1}^{l} (-q)^{-(k-1)} x^*_{i_k j_1} (i_1 \ldots \widehat{i_k} \ldots i_l | j_2 \ldots\ldots j_l)^*\right)$$

$$= \sum_{k=1}^{l} (-q)^{1-k}(-q)^{j_1-i_k}(1\dots\widehat{i_k}\dots n|1\dots\widehat{j_1}\dots n)\cdot(-q)^{\sum_{t\neq 1} j_t-\sum_{t\neq k} i_t}\det_q^{l-2}$$
$$\cdot(1\dots\widehat{i_1}\dots\widehat{i_2}\dots\dots\widehat{i_{k-1}}\dots\widehat{i_{k+1}}\dots\dots\widehat{i_l}\dots n|1\dots\widehat{j_2}\dots\widehat{j_3}\dots\dots\widehat{j_l}\dots n).$$

We claim that this is equal to

$$(-q)^{\sum_{t=1}^{l}(j_t-i_t)}\det_q^{l-2}\sum_{w}(-q)^{l(w)+1-n}(w1w2\dots w(n-1)|1\dots\widehat{j_1}\dots n)$$
$$\cdot(wn\,1\dots\widehat{i_1}\dots\dots\widehat{i_l}\dots n|1\dots\widehat{j_2}\dots\dots\widehat{j_l}\dots n)_l, \quad (5.3.1)$$

where the summation is over all $w\in\mathfrak{S}_n$ such that $w1<w2<\cdots<w(n-1)$. If wn is not one of the i_ks, then the summand in (5.3.1) vanishes since wn appears twice in the row on the left side of the second minor. Thus, the summation is over all w as above with $wn=i_k$ for some k. Note that $l(w)=n-i_k$ and

$$(i_k1\dots\widehat{i_1}\dots\dots\widehat{i_l}\dots n|\mathfrak{t})_l=(-q)^{i_k-k}(1\dots\widehat{i_1}\dots\widehat{i_{k-1}}\dots\widehat{i_{k+1}}\dots\widehat{i_l}\dots n|\mathfrak{t});$$

the claim follows. Again apply Laplace's expansion to the second minor in (5.3.1) to get

$$(wn\,1\dots\widehat{i_1}\dots\dots\widehat{i_l}\dots n|1\dots\widehat{j_2}\dots\dots\widehat{j_l}\dots n)_l$$
$$=\sum_{v}(-q)^{l(v)}x_{wn\,v1}(1\dots\widehat{i_1}\dots\dots\widehat{i_l}\dots n|v2v3\dots\widehat{vj_2}\dots\dots\widehat{vj_l}\dots vn),$$

the summation being over all $v\in\mathfrak{S}_{\{1,\dots,\widehat{j_2},\dots,\widehat{j_l},\dots,n\}}$ with $v2<v3<\cdots<vn$. After substituting this term in (5.3.1), one can again apply Laplace's expansion to get that (5.3.1) is equal to

$$(-q)^{\sum(j_t-i_t)}\det_q^{l-2}\sum_{v}(-q)^{l(v)+1-n}(12\dots n|1\dots\widehat{j_1}\dots n\,v1)_r$$
$$\cdot(1\dots\widehat{i_1}\dots\dots\widehat{i_l}\dots n|v2v3\dots\widehat{vj_2}\dots\dots\widehat{vj_l}\dots vn). \quad (5.3.2)$$

The only summand in (5.3.2) that does not vanish is the term for $v1=j_1$ with $l(v)=j_1-1$. Thus, (5.3.2) is equal to

$$(-q)^{\sum(j_t-i_t)}\det_q^{l-2}(-q)^{j_1-n}(12\dots n|1\dots\widehat{j_1}\dots nj_1)_r\cdot(i'_1\dots i'_{n-l}|j'_1\dots j'_{n-l})$$
$$=(-q)^{\sum_{t=1}^{l}(j_t-i_t)}\det_q^{l-1}(i'_1\dots i'_{n-l}|j'_1\dots j'_{n-l}). \quad \square$$

6. A basis for $A_q(n;r,s)$

Theorem 5.3 enables us to construct elements of $A_q(n;r,s)$ that are mapped to standard bideterminants under ι. First, we will introduce the notion of rational tableaux although we will slightly differ from the definition of rational tableaux in [Stembridge 1987]. Recall that $\Lambda^+(k)$ is the set of partitions of k.

Definition 6.1. Fix $0 \le k \le \min(r, s)$. Let $\rho \in \Lambda^+(r-k)$ and $\sigma \in \Lambda^+(s-k)$ with $\rho_1 + \sigma_1 \le n$. A *rational* (ρ, σ)*-tableau* is a pair $(\mathfrak{r}, \mathfrak{s})$ with $\mathfrak{r}$ a ρ-tableau and $\mathfrak{s}$ a σ-tableau.

Let $\mathrm{first}_i(\mathfrak{r}, \mathfrak{s})$ be the number of entries of the first row of $\mathfrak{r}$, which are at most i, plus the number of entries of the first row of $\mathfrak{s}$, which are at most i. A rational tableau is called *standard* if $\mathfrak{r}$ and $\mathfrak{s}$ are standard tableaux and the following condition holds:

$$\mathrm{first}_i(\mathfrak{r}, \mathfrak{s}) \le i \quad \text{for all } i = 1, \dots, n. \tag{6.1.1}$$

A pair $[(\mathfrak{r}, \mathfrak{s}), (\mathfrak{r}', \mathfrak{s}')]$ of rational (ρ, σ)-tableaux is called a *rational bitableau*, and it is called a standard rational bitableau if both $(\mathfrak{r}, \mathfrak{s})$ and $(\mathfrak{r}', \mathfrak{s}')$ are standard rational tableaux.

Remark 6.2. In [Stembridge 1987], condition (6.1.1) is already part of the definition of rational tableaux. The condition $\rho_1 + \sigma_1 \le n$ is equivalent to condition (6.1.1) for $i = n$. The reason for the difference will be apparent in the next lemma's proof.

Lemma 6.3. *There is a bijection between the set consisting of all standard rational* (ρ, σ)*-tableaux for* $\rho \in \Lambda^+(r-k)$ *and* $\sigma \in \Lambda^+(s-k)$ *as* k *runs from* 0 *to* $\min(r, s)$ *and the set of all standard* λ*-tableaux for* $\lambda \in \Lambda^+(r+(n-1)s)$ *so* $\sum_{i=1}^{s} \lambda_i \ge (n-1)s$.

Proof. Given a rational (ρ, σ)-tableau $(\mathfrak{r}, \mathfrak{s})$, we construct a λ-tableau $\mathfrak{t}$ as follows. Draw a rectangular diagram with s rows and n columns. Rotate the tableau $\mathfrak{s}$ by 180 degrees, and place it in the bottom right corner of the rectangle. Place the tableau $\mathfrak{r}$ on the left side below the rectangle. Fill the empty boxes of the rectangle with numbers such that in each row the entries that do not appear in $\mathfrak{t}$ appear in the empty boxes in increasing order. Let $\mathfrak{t}$ be the tableau consisting of the formerly empty boxes and the boxes of $\mathfrak{r}$. We illustrate this procedure with an example. Let $n = 5$, $r = 4$, $s = 5$ and $k = 1$, and let

$$(\mathfrak{r}, \mathfrak{s}) = \left(\begin{array}{|c|c|} \hline 1 & 3 \\ \hline 2 & \\ \hline \end{array}, \begin{array}{|c|c|} \hline 3 & 4 \\ \hline 3 & 5 \\ \hline \end{array} \right).$$

Then

$$(\mathfrak{r}, \mathfrak{s}) \rightsquigarrow \begin{array}{|c|c|c|c|c|} \hline & & & & \\ \hline & & & & \\ \hline & & & & \\ \hline & & & 5 & 3 \\ \hline & & & 4 & 3 \\ \hline 1 & 3 & & & \\ \hline 2 & & & & \\ \hline \end{array} \rightsquigarrow \begin{array}{|c|c|c|c|c|} \hline 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 2 & 4 & 5 & 3 \\ \hline 1 & 2 & 5 & 4 & 3 \\ \hline 1 & 3 & & & \\ \hline 2 & & & & \\ \hline \end{array} \rightsquigarrow \mathfrak{t} = \begin{array}{|c|c|c|c|c|} \hline 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 2 & 4 & & \\ \hline 1 & 2 & 5 & & \\ \hline 1 & 3 & & & \\ \hline 2 & & & & \\ \hline \end{array}$$

It is now easy to give an inverse. Just draw the rectangle into the tableau $\mathfrak{t}$, fill the empty boxes of the rectangle in a similar way as before, and rotate these back to obtain $\mathfrak{s}$. Note $\mathfrak{r}$ is the part of the tableau $\mathfrak{t}$ that lies outside the rectangle. We have to show that these bijections provide standard tableaux of the right shape.

Suppose $(\mathfrak{r}, \mathfrak{s})$ is a rational (ρ, σ)-tableau, so $\mathfrak{t}$ is a λ-tableau with $\lambda_i = n - \sigma_{s+1-i}$ for $i \leq s$ and $\lambda_i = \rho_{i-s}$ for $i > s$. So $\lambda_i \geq \lambda_{i+1}$ for $i < s$ is equivalent to $\sigma_{s+1-i} \leq \sigma_{s-i}$, and for $i > s$ it is equivalent to $\rho_{i-s} \geq \rho_{i+1-s}$. Now $\rho_1 + \sigma_1 = \lambda_{s+1} - (\lambda_s - n)$. This shows that λ is a partition if and only if ρ and σ are partitions with $\rho_1 + \sigma_1 \leq n$. We still have to show that $(\mathfrak{r}, \mathfrak{s})$ is standard if and only if $\mathfrak{t}$ is standard.

By definition, all standard tableaux have increasing rows. A tableau has nondecreasing columns if and only if for all $i = 1, \dots, n$ and all rows (except for the last row) the number of entries at most i in this row is greater than or equal to the number of entries at most i in the next row. Now it follows from the construction that $\mathfrak{t}$ has nondecreasing columns inside the rectangle if and only if $\mathfrak{s}$ has nondecreasing columns, $\mathfrak{t}$ has nondecreasing columns outside the rectangle if and only if $\mathfrak{r}$ has nondecreasing columns and the columns in $\mathfrak{t}$ do not decrease from row s to row $s+1$ if and only if condition (6.1.1) holds. □

Definition 6.4. Let $\mathfrak{det}_q^{(k)} \in A_q(n; k, k)$ with $k \geq 1$ be recursively defined by $\mathfrak{det}_q^{(1)} := \sum_{l=1}^{n} x_{1l} x_{1l}^*$ and $\mathfrak{det}_q^{(k)} := \sum_{l=1}^{n} x_{1l} \mathfrak{det}_q^{(k-1)} x_{1l}^*$ for $k > 1$.

Let a (rational) bideterminant $((\mathfrak{r}, \mathfrak{s}) | (\mathfrak{r}', \mathfrak{s}')) \in A_q(n; r, s)$ be defined by

$$((\mathfrak{r}, \mathfrak{s}) | (\mathfrak{r}', \mathfrak{s}')) := (\mathfrak{r} | \mathfrak{r}') \, \mathfrak{det}_q^{(k)} \, (\mathfrak{s} | \mathfrak{s}')^*$$

whenever $[(\mathfrak{r}, \mathfrak{s}), (\mathfrak{r}', \mathfrak{s}')]$ is a rational (ρ, σ)-bitableau such that $\rho \in \Lambda^+(r-k)$ and $\sigma \in \Lambda^+(s-k)$ for some $k = 0, 1, \dots, \min(r, s)$.

Note that the proof of Lemma 5.1 and Remark 3.3(4) show that $\iota(\mathfrak{det}_q^{(k)}) = \det_q^k$. Furthermore, if ρ_1 or $\sigma_1 > n$, then the bideterminant of a (ρ, σ)-bitableau vanishes. As a direct consequence of Theorem 5.3, we get the following:

Lemma 6.5. *Let $(\mathfrak{r}, \mathfrak{s})$ and $(\mathfrak{r}', \mathfrak{s}')$ be two standard rational tableaux, and let $\mathfrak{t}$ and $\mathfrak{t}'$ be the (standard) tableaux obtained from the correspondence of Lemma 6.3. Then*

$$\iota((\mathfrak{r}, \mathfrak{s}) | (\mathfrak{r}', \mathfrak{s}')) = (-q)^{c(\mathfrak{t}, \mathfrak{t}')} (\mathfrak{t} | \mathfrak{t}')$$

for some integer $c(\mathfrak{t}, \mathfrak{t}')$. In particular, the bideterminants of standard rational bitableaux are linearly independent.

Proof. This follows directly from Theorem 5.3, the construction of the bijection and $\iota(\mathfrak{det}_q^{(k)}) = \det_q^k$. The second statement follows from the fact that the $(\mathfrak{t} | \mathfrak{t}')$s are linearly independent. □

Lemma 6.6. *We have*

$$\sum_{l=1}^{n} x_{il} \mathfrak{det}_q^{(k)} x_{jl}^* = 0 \quad \textit{for } i \neq j, \tag{6.6.1}$$

$$\sum_{l=1}^{n} q^{2l} x_{li} \mathfrak{det}_q^{(k)} x_{lj}^* = 0 \quad \textit{for } i \neq j, \tag{6.6.2}$$

and

$$\sum_{l=1}^{n} q^{2l-2i} x_{li} \mathfrak{Det}_q^{(k)} x_{li}^* = \sum_{l=1}^{n} x_{jl} \mathfrak{Det}_q^{(k)} x_{jl}^*. \tag{6.6.3}$$

Proof. Without loss of generality, we may assume $k=1$. Suppose that $i,\ j \neq 1$. Then

$$\begin{aligned}\sum_{l=1}^{n} x_{il} \mathfrak{Det}_q^{(1)} x_{jl}^* = \sum_{k,l=1}^{n} x_{ik} x_{1l} x_{1l}^* x_{jk}^* &= \sum_{k<l} x_{1l} x_{ik} x_{jk}^* x_{1l}^* + q^{-2} \sum_{k} x_{1k} x_{ik} x_{jk}^* x_{1k}^* \\ &\quad + \sum_{k>l} \big(x_{1l} x_{ik} x_{jk}^* x_{1l}^* + (q^{-1}-q)(x_{1k} x_{il} x_{1l}^* x_{jk}^* + x_{1l} x_{ik} x_{1k}^* x_{jl}^*) \big) \\ &= \sum_{k,l} x_{1l} x_{ik} x_{jk}^* x_{1l}^* + (q^{-2}-1) \sum_{k} q x_{1k} x_{ik} x_{1k}^* x_{jk}^* \\ &\quad + (q^{-1}-q) \sum_{k>l} (x_{1k} x_{il} x_{1l}^* x_{jk}^* + x_{1l} x_{ik} x_{1k}^* x_{jl}^*) \\ &= \delta_{ij} \mathfrak{Det}_q^{(2)} + (q^{-1}-q) \sum_{k,l} x_{1k} x_{il} x_{1l}^* x_{jk}^* = \delta_{ij} \mathfrak{Det}_q^{(2)}.\end{aligned}$$

For $j \neq 1$, we have

$$\begin{aligned}\sum_{l=1}^{n} x_{1l} \mathfrak{Det}_q^{(1)} x_{jl}^* = \sum_{k,l=1}^{n} x_{1k} x_{1l} x_{1l}^* x_{jk}^* &= \sum_{k<l} q x_{1l} x_{1k} x_{jk}^* x_{1l}^* + q^{-1} \sum_{k} x_{1k} x_{1k} x_{jk}^* x_{1k}^* \\ &\quad + \sum_{k>l} \big(q^{-1} x_{1l} x_{1k} x_{jk}^* x_{1l}^* + (q^{-1}-q) x_{1k} x_{1l} x_{jl}^* x_{1k}^* \big) \\ &= \sum_{k,l} q^{-1} x_{1l} x_{1k} x_{jk}^* x_{1l}^* = 0.\end{aligned}$$

Similarly, one can show that

$$\sum_{l=1}^{n} x_{il} \mathfrak{Det}_q^{(1)} x_{1l}^* = 0 \quad \text{for } i \neq 1,$$

$$\sum_{l=1}^{n} q^{2l-2i} x_{li} \mathfrak{Det}_q^{(1)} x_{lj}^* = \delta_{ij} \sum_{l=1}^{n} q^{2l-2} x_{l1} \mathfrak{Det}_q^{(1)} x_{l1}^* \quad \text{for } i,\ j \neq 1,$$

$$\sum_{l=1}^{n} q^{2l-2} x_{l1} \mathfrak{Det}_q^{(1)} x_{lj}^* = 0 \quad \text{for } j \neq 1,$$

$$\sum_{l=1}^{n} q^{2l-2i} x_{li} \mathfrak{Det}_q^{(1)} x_{l1}^* = 0 \quad \text{for } i \neq 1.$$

Finally,

$$\begin{aligned}
&\sum_{l=1}^{n} q^{2l-2} x_{l1}\mathfrak{det}_q^{(1)} x_{l1}^* \\
&\qquad = \sum_{l,k} q^{2l-2} x_{l1}x_{1k}x_{1k}^*x_{l1}^* = \sum_{l,k\neq 1} q^{2l-2} x_{1k}x_{l1}x_{l1}^*x_{1k}^* \\
&\qquad\qquad + \sum_{l\neq 1} q^{2l-4} x_{11}x_{l1}x_{l1}^*x_{11}^* + \sum_{k\neq 1} q^{2} x_{1k}x_{11}x_{11}^*x_{1k}^* + x_{11}x_{11}x_{11}^*x_{11}^* \\
&\qquad = \mathfrak{det}_q^{(2)} + \sum_{l\neq 1} q^{2l-4}(1-q^2) x_{11}x_{l1}x_{l1}^*x_{11}^* + \sum_{k\neq 1} (q^2-1) x_{1k}x_{11}x_{11}^*x_{1k}^* \\
&\qquad = \mathfrak{det}_q^{(2)} + (1-q^2)\Bigg(\sum_{l\neq 1} q^{2l-4} x_{11}x_{l1}x_{l1}^*x_{11}^* - q^{-2}\sum_{k\neq 1} x_{11}x_{1k}x_{1k}^*x_{11}^*\Bigg) \\
&\qquad = \mathfrak{det}_q^{(2)}. \qquad\qquad \square
\end{aligned}$$

Lemma 6.7. *Suppose* $\boldsymbol{r} = (r_1, \dots, r_k), \boldsymbol{s} = (s_1, \dots, s_k) \in I(n, k)$ *are fixed. Let* $j \in \{1, \dots, n\}$ *and* $k \geq 1$. *Then we have, modulo* $\mathfrak{det}_q^{(1)}$,

$$\begin{aligned}
&\sum_{j<j_1<j_2<\cdots<j_k} (\boldsymbol{r}|j_k \dots j_2 j_1)_r (\boldsymbol{s}|j_1 j_2 \dots j_k)_r^* \\
&\qquad\qquad \equiv (-1)^k q^{2\sum_{i=0}^{k-1} i} \sum_{j_1<j_2<\cdots<j_k\leq j} (\boldsymbol{r}|j_k \dots j_2 j_1)_r (\boldsymbol{s}|j_1 j_2 \dots j_k)_r^*.
\end{aligned}$$

Proof. The only difference between $(\boldsymbol{s}|j_1 j_2 \dots j_k)_r^*$ and $(\boldsymbol{s}|j_1 j_2 \dots j_k)_l^*$ is on a power of $-q$ not depending on $j_1, j_2, \dots, j_k$. Thus, we can show the lemma with $(\cdot\,,\cdot)_r^*$ replaced by $(\cdot\,,\cdot)_l^*$. Similarly, we can assume that $r_1 < r_2 < \cdots < r_k$ and $s_1 > s_2 > \cdots > s_k$. Note that, modulo $\mathfrak{det}_q^{(1)}$, we have the relations $\sum_{k=1}^{n} x_{ik}x_{jk}^* \equiv 0$. It follows that the lemma is true for $k = 1$. Assume that the lemma holds for $k - 1$. If M is an ordered set, let $M^{k,<}$ be the set of k-tuples in M with increasing entries. For a subset $M \subset \{1, \dots, n\}$, we have

$$\begin{aligned}
&\sum_{\boldsymbol{j}\in M^{k,<}} (\boldsymbol{r}|j_k \dots j_2 j_1)_r (\boldsymbol{s}|j_1 j_2 \dots j_k)_l^* \\
&\qquad\qquad = \sum_{\boldsymbol{j}\in M^{k,<}, w} (-q)^{-l(w)} (\boldsymbol{r}|j_k \dots j_2 j_1)_r x_{s_1 j_{w1}}^* \cdots x_{s_k j_{wk}}^* \\
&\qquad\qquad = \sum_{\boldsymbol{j}\in M^{k,<}, w} (\boldsymbol{r}|j_{wk} \dots j_{w1})_r x_{s_1 j_{w1}}^* \cdots x_{s_k j_{wk}}^* \\
&\qquad\qquad = \sum_{\boldsymbol{j}\in M^{k}} (\boldsymbol{r}|j_k \dots j_1)_r x_{s_1 j_1}^* \cdots x_{s_k j_k}^*.
\end{aligned}$$

Applying Laplace's expansion, we can write a quantum minor $(\boldsymbol{r}|\boldsymbol{j}_1\boldsymbol{j}_2)_r$ as a linear combination of products of quantum minors, say

$$(\boldsymbol{r}|\boldsymbol{j}_1\boldsymbol{j}_2)_r = \sum_l c_l(\boldsymbol{r}'_l|\boldsymbol{j}_1)_r(\boldsymbol{r}''_l|\boldsymbol{j}_2)_r.$$

Then with $\epsilon_k := (-1)^k q^{2\sum_{i=0}^{k-1} i}$, $\boldsymbol{j} = (j_1, \dots, j_k)$, $\boldsymbol{j}' = (j_1, \dots, j_{k-1})$, $C = \{1 \dots j\}$ and $D = \{j+1 \dots n\}$, we have

$$\begin{aligned}
\sum_{\boldsymbol{j}\in D^{k,<}} (\boldsymbol{r}|j_k \dots j_2 j_1)_r(\boldsymbol{s}|j_1 j_2 \dots j_k)^*_l &= \sum_{\boldsymbol{j}\in D^k} (\boldsymbol{r}|j_k \dots j_1)_r x^*_{s_1 j_1} \cdots x^*_{s_k j_k} \\
&= \sum_{\boldsymbol{j}\in D^k, l} c_l(\boldsymbol{r}'_l|j_k)_r(\boldsymbol{r}''_l|j_{k-1} \dots j_1)_r x^*_{s_1 j_1} \cdots x^*_{s_{k-1} j_{k-1}} x^*_{s_k j_k} \\
&\equiv \epsilon_{k-1} \sum_{\boldsymbol{j}'\in C^{k-1}, l,\ j_k > j} c_l(\boldsymbol{r}'_l|j_k)_r(\boldsymbol{r}''_l|j_{k-1} \dots j_1)_r x^*_{s_1 j_1} \cdots x^*_{s_{k-1} j_{k-1}} x^*_{s_k j_k} \\
&= \epsilon_{k-1} \sum_{\boldsymbol{j}'\in C^{k-1}, j_k > j} (\boldsymbol{r}|j_k j_{k-1} \dots j_1)_r x^*_{s_1 j_1} \cdots x^*_{s_{k-1} j_{k-1}} x^*_{s_k j_k} \\
&= \epsilon_{k-1} \sum_{\boldsymbol{j}'\in C^{k-1}, j_k > j} (-q)^{k-1}(\boldsymbol{r}|j_{k-1} \dots j_1 j_k)_r x^*_{s_k j_k} x^*_{s_1 j_1} \cdots x^*_{s_{k-1} j_{k-1}} \\
&= \epsilon_{k-1} \sum_{\boldsymbol{j}'\in C^{k-1}, l,\ j_k > j} (-q)^{k-1} c_l(\boldsymbol{r}'_l|j_{k-1} \dots j_1)_r x_{\boldsymbol{r}''_l j_k} x^*_{s_k j_k} x^*_{s_1 j_1} \cdots x^*_{s_{k-1} j_{k-1}} \\
&\equiv -\epsilon_{k-1} \sum_{\boldsymbol{j}\in C^k, l} (-q)^{k-1} c_l(\boldsymbol{r}'_l|j_{k-1} \dots j_1)_r x_{\boldsymbol{r}''_l j_k} x^*_{s_k j_k} x^*_{s_1 j_1} \cdots x^*_{s_{k-1} j_{k-1}} \\
&= -\epsilon_{k-1} \sum_{\boldsymbol{j}\in C^k} (-q)^{k-1}(\boldsymbol{r}|j_{k-1} \dots j_1 j_k)_r x^*_{s_k j_k} x^*_{s_1 j_1} \cdots x^*_{s_{k-1} j_{k-1}} \\
&= -\epsilon_{k-1} \sum_{\boldsymbol{j}\in C^{k,<}} (-q)^{k-1}(\boldsymbol{r}|j_k \dots j_1)_r(s_k s_1 \dots s_{k-1}|j_1 \dots j_k)^*_l \\
&= -\epsilon_{k-1} \sum_{\boldsymbol{j}\in C^{k,<}} (-q)^{2(k-1)}(\boldsymbol{r}|j_k \dots j_1)_r(s_1 \dots s_k|j_1 \dots j_k)^*_l \\
&= \epsilon_k \sum_{\boldsymbol{j}\in C^{k,<}} (\boldsymbol{r}|j_k \dots j_2 j_1)_r(\boldsymbol{s}|j_1 j_2 \dots j_k)^*_l. \qquad \square
\end{aligned}$$

Lemma 6.8. *Let $\boldsymbol{r}'$ and $\boldsymbol{s}'$ be strictly increasing multi-indices considered as tableaux with one row. Let i be the maximal entry appearing, and suppose that i is minimal such that i violates condition (6.1.1). Let I be the set of entries appearing in both $\boldsymbol{r}'$ and $\boldsymbol{s}'$; then we have $i \in I$. Let $L_1 := \{k_1, \dots, k_{l_1}\}$ be the set of entries of $\boldsymbol{r}'$ not appearing in $\boldsymbol{s}'$, let $L_2 := \{k'_1, \dots, k'_{l_2}\}$ be the set of entries of $\boldsymbol{s}'$ not appearing in $\boldsymbol{r}'$, and let $i_1 < i_2 < \dots < i_k = i$ be the entries of I.*

Let $D := \{i_1, \dots, i_k, i_k + 1, i_k + 2, \dots, n\}$ *and* $C := \{1, \dots, n\} \setminus (D \cup L_1 \cup L_2)$. *Furthermore, for* $j_1, \dots, j_t \in \{1, \dots, n\}$, *let*

$$m(j_1, \dots, j_t) := \bigl|\bigl\{ (l, c) \in \{1, \dots, t\} \times C : j_l < c \bigr\}\bigr|.$$

Let $\boldsymbol{k} := (k_1, \dots, k_{l_1})$ *and* $\boldsymbol{k}' := (k'_1, \dots, k'_{l_2})$, *and let* $\boldsymbol{r}$ *and* $\boldsymbol{s}$ *be multi-indices of the same length as* $\boldsymbol{r}'$ (*resp.* $\boldsymbol{s}'$); *then we have*

$$\sum_{\boldsymbol{j} \in D^{k,<}} q^{2m(\boldsymbol{j})} (\boldsymbol{r} | \boldsymbol{k} j_k \dots j_1)_r (\boldsymbol{s} | j_1 \dots j_k \boldsymbol{k}')^*_r \equiv 0 \bmod \mathfrak{Det}_q^{(1)}.$$

Proof. Note that $i \in I$ and $i = 2k + l_1 + l_2 - 1$; otherwise, $i - 1$ would violate (6.1.1). Therefore, $|C| = k - 1$. Let $c_{\max}$ be the maximal element of C, $\tilde{C} = \{1, \dots, c_{\max}\}$, $\tilde{D} = \{c_{\max} + 1, c_{\max} + 2, \dots, n\} \subset D \cup L_1 \cup L_2$, $D_- = \{ d \in D : d < c_{\max} \}$ and $D_+ = \{ d \in D : d > c_{\max} \}$. With $\tilde{\boldsymbol{j}} = (j_1, \dots, j_l)$ and $\widehat{\boldsymbol{j}} = (j_{l+1}, \dots, j_k)$, we have

$$\begin{aligned}\sum_{\boldsymbol{j} \in D^{k,<}} & q^{2m(\boldsymbol{j})} (\boldsymbol{r} | \boldsymbol{k} j_k \dots j_1)_r (\boldsymbol{s} | j_1 \dots j_k \boldsymbol{k}')^*_r \\ &= \sum_{l=0}^{k} \sum_{\tilde{\boldsymbol{j}} \in D_-^{l,<}} q^{2m(\tilde{\boldsymbol{j}})} \sum_{\widehat{\boldsymbol{j}} \in D_+^{k-l,<}} (\boldsymbol{r} | \boldsymbol{k} j_k \dots j_1)_r (\boldsymbol{s} | j_1 \dots j_k \boldsymbol{k}')^*_r. \qquad (6.8.1)\end{aligned}$$

Without loss of generality, we may assume that the entries in $\boldsymbol{s}$ are increasing. We apply Laplace's expansion and Lemma 6.7 to get for fixed l and $\tilde{\boldsymbol{j}}$

$$\begin{aligned}\sum_{\widehat{\boldsymbol{j}} \in D_+^{k-l,<}} (\boldsymbol{r} | \boldsymbol{k} j_k \dots j_1)_r (\boldsymbol{s} | j_1 \dots j_k \boldsymbol{k}')^*_r &= \sum_{\widehat{\boldsymbol{j}} \in \tilde{D}^{k-l,<}} (\boldsymbol{r} | \boldsymbol{k} j_k \dots j_1)_r (\boldsymbol{s} | j_1 \dots j_k \boldsymbol{k}')^*_r \\ &= q^{2l(k-l)} \sum_{\widehat{\boldsymbol{j}} \in \tilde{D}^{k-l,<}} (\boldsymbol{r} | \boldsymbol{k} j_l \dots j_1 j_k \dots j_{l+1})_r (\boldsymbol{s} | j_{l+1} \dots j_k j_1 \dots j_l \boldsymbol{k}')^*_r \\ &\equiv \epsilon_{k-l} q^{2l(k-l)} \sum_{\widehat{\boldsymbol{j}} \in \tilde{C}^{k-l,<}} (\boldsymbol{r} | \boldsymbol{k} j_l \dots j_1 j_k \dots j_{l+1})_r (\boldsymbol{s} | j_{l+1} \dots j_k j_1 \dots j_l \boldsymbol{k}')^*_r \\ &= \epsilon_{k-l} q^{2l(k-l)} \sum_{\widehat{\boldsymbol{j}} \in (C \cup D_-)^{k-l,<}} (\boldsymbol{r} | \boldsymbol{k} j_l \dots j_1 j_k \dots j_{l+1})_r (\boldsymbol{s} | j_{l+1} \dots j_k j_1 \dots j_l \boldsymbol{k}')^*_r.\end{aligned}$$

This expression can be substituted into (6.8.1). Each nonzero summand belongs to a disjoint union $S_1 \dot{\cup} S_2 = S \subset C \cup D_-$ such that $|S| = k$, $S_1 = \{j_1, \dots, j_l\}$ and $S_2 = \{j_{l+1}, \dots, j_k\}$. We will show that the summands belonging to some fixed set S cancel out.

Therefore, we claim that for each subset $S \subset C \cup D_-$ with k elements, there exists some $d \in D \cap S$ such that $m(d) = |\{ s \in S : s > d \}|$. Suppose not. Since $|C| = k - 1$, S contains at least one element of D. Let $s_1 < s_2 < \dots < s_m$ be the elements of $D \cap S$. We show by downward induction that $m(s_l) > |\{ s \in S : s > s_l \}|$ for $1 \le l \le m$; $m(s_m)$ is the cardinality of $\{s_m + 1, \dots, c_{\max}\} \cap C$. Since all $s \in S$ with $s > s_m$ are elements of C, we have $\{s_m+1,\dots,c_{\max}\}\cap S\subset\{s_m+1,\dots,c_{\max}\}\cap C$, and thus, $m(s_m)\geq|\{\,s\in S:s>s_m\,\}|$. By assumption, we have $>$ instead of $\geq$. Now suppose $m(s_l)>|\{\,s\in S:s>s_l\,\}|$, so $\{\,s\in S:s_{l-1}<s\leq s_l\,\}=\{\,s\in S\cap C:s_{l-1}<s<s_l\,\}\cup\{s_l\}$; thus, S contains at most $m(s_{l-1})-m(s_l)$ elements between s_{l-1} and s_l, so at most $m(s_{l-1})-m(s_l)+1+m(s_l)-1=m(s_{l-1})$ elements are greater than s_{l-1}. By assumption, we have $m(s_{l-1})>|\{\,s\in S:s>s_{l-1}\,\}|$. We have shown that S contains less than $m(s_1)$ elements greater than s_1; thus, S contains less than $|C|+1=k$ elements, which is a contradiction. This shows the claim.

Let $S\subset C\cup D_-$ be fixed subset of cardinality k. By the previous consideration, there is an element $d\in D\cap S$ with $m(d)=|\{\,s\in S:s>d\,\}|$. We claim that the summand for S_1 and S_2 with $d\in S_1$ cancels the summand for $S_1\setminus\{d\}$ and $S_2\cup\{d\}$. Note that

$$\begin{aligned}(\boldsymbol{r}|\boldsymbol{k}\,j_l\dots\widehat{d}\dots j_1\,j_k\dots d\dots j_{l+1})_r(\boldsymbol{s}|\,j_{l+1}\dots d\dots j_k\,j_1\dots\widehat{d}\dots j_l\boldsymbol{k}')^*_r\\ =q^{2|\{\,s\in S:s>d\,\}|-2(l-1)}(\boldsymbol{r}|\boldsymbol{k}\,j_l\dots j_1\,j_k\dots j_{l+1})_r(\boldsymbol{s}|\,j_{l+1}\dots j_k\,j_1\dots j_l\boldsymbol{k}')^*_r.\end{aligned}$$

Comparing coefficients, we see that both summands cancel. □

Theorem 6.9 (Rational Straightening Algorithm). *The set of bideterminants of standard rational bitableaux forms an R-basis of $A_q(n;r,s)$.*

Proof. We have to show that the bideterminants of standard rational bitableaux generate $A_q(n;r,s)$. Clearly, the bideterminants $((\mathfrak{t},\mathfrak{s})|(\mathfrak{t}',\mathfrak{s}'))$ with $\mathfrak{t}$, $\mathfrak{t}'$, $\mathfrak{s}$ and $\mathfrak{s}'$ standard tableaux generate $A_q(n;r,s)$. Let $\operatorname{cont}(\mathfrak{t})$ (resp. $\operatorname{cont}(\mathfrak{s})$) be the content of $\mathfrak{t}$ (resp. $\mathfrak{s}$) defined in Definition 3.4.

Let $\mathfrak{t}$, $\mathfrak{t}'$, $\mathfrak{s}$ and $\mathfrak{s}'$ be standard tableaux, and suppose that the rational bitableau $[(\mathfrak{t},\mathfrak{s}),(\mathfrak{t}',\mathfrak{s}')]$ is not standard. It suffices to show the bideterminant $((\mathfrak{t},\mathfrak{s})|(\mathfrak{t}',\mathfrak{s}'))$ is a linear combination of bideterminants $((\widehat{\mathfrak{t}},\widehat{\mathfrak{s}})|(\widehat{\mathfrak{t}}',\widehat{\mathfrak{s}}'))$ such that $\widehat{\mathfrak{t}}$ has fewer boxes than $\mathfrak{t}$ or $\operatorname{cont}(\mathfrak{t})>\operatorname{cont}(\widehat{\mathfrak{t}})$ or $\operatorname{cont}(\mathfrak{s})>\operatorname{cont}(\widehat{\mathfrak{s}})$ in the lexicographical order. Without loss of generality, we make the following assumptions:

- In the nonstandard rational bitableau $[(\mathfrak{t},\mathfrak{s}),(\mathfrak{t}',\mathfrak{s}')]$, the rational tableau $(\mathfrak{t}',\mathfrak{s}')$ is nonstandard. Note that the automorphism of Remark 4.2 maps a bideterminant $((\mathfrak{t},\mathfrak{s})|(\mathfrak{t}',\mathfrak{s}'))$ to the bideterminant $((\mathfrak{t}',\mathfrak{s}')|(\mathfrak{t},\mathfrak{s}))$.
- Suppose that $(\mathfrak{t},\mathfrak{s})$ and $(\mathfrak{t}',\mathfrak{s}')$ are (ρ,σ)-tableaux. In view of Lemma 6.6, we can assume that $\rho\in\Lambda^+(r)$ and $\sigma\in\Lambda^+(s)$.
- The tableaux $\mathfrak{t}$, $\mathfrak{t}'$, $\mathfrak{s}$ and $\mathfrak{s}'$ have only one row (each bideterminant has a factor of this type), and we can use Theorem 3.5 to write nonstandard bideterminants as a linear combination of standard ones of the same content.
- Let i be minimal such that condition (6.1.1) of Definition 6.1 is violated for i. Applying Laplace's expansion, we may assume that there is no greater entry than i in $\mathfrak{t}'$ and in $\mathfrak{s}'$.

Note that all elements of $A_q(n; r, s)$ having a factor $\mathfrak{Det}_q^{(1)}$ can be written as a linear combination of bideterminants of rational (ρ, σ)-bitableaux with $\rho \in \Lambda^+(r-k)$, $k > 0$. Thus, it suffices to show that $((\mathfrak{r}, \mathfrak{s})|(\mathfrak{r}', \mathfrak{s}'))$ is, modulo $\mathfrak{Det}_q^{(1)}$, a linear combination of bideterminants of "lower content". The summand of highest content in Lemma 6.8 is that one for $\boldsymbol{j} = (i_1, i_2, \dots, i_k)$, and this summand is a scalar multiple (a power of $-q$, which is invertible) of $((\mathfrak{r}, \mathfrak{s})|(\mathfrak{r}', \mathfrak{s}'))$. □

The following is an immediate consequence of the preceding theorem and Lemma 6.3.

Corollary 6.10. *There exists an R-linear map $\phi : A_q(n, r+(n-1)s) \to A_q(n; r, s)$ given on a basis by $\phi(\mathfrak{t}|\mathfrak{t}') := (-q)^{-c(\mathfrak{t},\mathfrak{t}')}((\mathfrak{r}, \mathfrak{s})|(\mathfrak{r}', \mathfrak{s}'))$ if the shape λ of $\mathfrak{t}$ satisfies $\sum_{i=1}^{s} \lambda_i \geq (n-1)s$, where $(\mathfrak{r}, \mathfrak{s})$ and $(\mathfrak{r}', \mathfrak{s}')$ are the rational tableaux respectively corresponding to $\mathfrak{t}$ and $\mathfrak{t}'$ under the correspondence of Lemma 6.3, and $\phi(\mathfrak{t}|\mathfrak{t}') := 0$ otherwise. We have*

$$\phi \circ \iota = \mathrm{id}_{A_q(n;r,s)},$$

and thus, $\pi = \iota^*$ is surjective.

As noted in Section 2, we now have the main result.

Theorem 6.11 (Schur–Weyl duality for mixed tensor space, II). *We have*

$$S_q(n; r, s) = \mathrm{End}_{\mathfrak{B}_{r,s}(q)}(V^{\otimes r} \otimes V^{*\otimes s}) = \rho_{\mathrm{mxd}}(\mathbf{U}) = \rho_{\mathrm{mxd}}(\mathbf{U}'),$$

and $S_q(n; r, s)$ is R-free with a basis indexed by standard rational bitableau.

Proof. The first assertion follows from the surjectivity of π; the second assertion is obtained by dualizing the basis of $A_q(n; r, s)$. □

References

[Benkart et al. 1994] G. Benkart, M. Chakrabarti, T. Halverson, R. Leduc, C. Lee, and J. Stroomer, "Tensor product representations of general linear groups and their connections with Brauer algebras", *J. Algebra* **166**:3 (1994), 529–567. MR 95d:20071 Zbl 0815.20028

[Brundan and Stroppel 2011] J. Brundan and C. Stroppel, "Gradings on walled Brauer algebras and Khovanov's arc algebras", preprint, 2011. arXiv 1107.0999

[Dipper and Donkin 1991] R. Dipper and S. Donkin, "Quantum GL_n", *Proc. London Math. Soc.* (3) **63**:1 (1991), 165–211. MR 92g:16055 Zbl 0734.20018

[Dipper and Doty 2008] R. Dipper and S. Doty, "The rational Schur algebra", *Represent. Theory* **12** (2008), 58–82. MR 2009e:20097 Zbl 1185.20052

[Dipper and James 1989] R. Dipper and G. James, "The q-Schur algebra", *Proc. London Math. Soc.* (3) **59**:1 (1989), 23–50. MR 90g:16026 Zbl 0711.20007

[Dipper et al. 2012] R. Dipper, S. Doty, and F. Stoll, "The quantized walled Brauer algebra and mixed tensor space", preprint, 2012. arXiv 0806.0264

[Goodearl 2006] K. R. Goodearl, "Commutation relations for arbitrary quantum minors", *Pacific J. Math.* **228**:1 (2006), 63–102. MR 2007j:17019 Zbl 1125.16034

[Green 1996] R. M. Green, "q-Schur algebras as quotients of quantized enveloping algebras", *J. Algebra* **185**:3 (1996), 660–687. MR 97k:17016 Zbl 0862.17007

[Hong and Kang 2002] J. Hong and S.-J. Kang, *Introduction to quantum groups and crystal bases*, Graduate Studies in Mathematics **42**, American Mathematical Society, Providence, RI, 2002. MR 2002m:17012 Zbl 1134.17007

[Huang and Zhang 1993] R. Q. Huang and J. J. Zhang, "Standard basis theorem for quantum linear groups", *Adv. Math.* **102**:2 (1993), 202–229. MR 94j:16067 Zbl 0793.05143

[Jantzen 1996] J. C. Jantzen, *Lectures on quantum groups*, Graduate Studies in Mathematics **6**, American Mathematical Society, Providence, RI, 1996. MR 96m:17029 Zbl 0842.17012

[Koike 1989] K. Koike, "On the decomposition of tensor products of the representations of the classical groups: by means of the universal characters", *Adv. Math.* **74**:1 (1989), 57–86. MR 90j:22014 Zbl 0681.20030

[Kosuda and Murakami 1993] M. Kosuda and J. Murakami, "Centralizer algebras of the mixed tensor representations of quantum group $U_q(\mathrm{gl}(n, \mathbf{C}))$", *Osaka J. Math.* **30**:3 (1993), 475–507. MR 94k:17025 Zbl 0806.17012

[Leduc 1994] R. E. Leduc, *A two-parameter version of the centralizer algebra of the mixed tensor representations of the general linear group and quantum general linear group*, Ph.D. thesis, University of Wisconsin–Madison, 1994. MR 2691209

[Lusztig 1990] G. Lusztig, "Finite-dimensional Hopf algebras arising from quantized universal enveloping algebra", *J. Amer. Math. Soc.* **3**:1 (1990), 257–296. MR 91e:17009 Zbl 0695.16006

[Schur 1927] I. Schur, "Über die rationalen Darstellungen der allgemeinen linearen Gruppe", *Sitzungsber. Akad. Berlin* (1927), 58–75. Reprinted as pp. 68–85 in *Gesammelte Abhandlungen*, III, Springer, Berlin, 1973. MR 57 #2858c JFM 53.0108.05

[Stembridge 1987] J. R. Stembridge, "Rational tableaux and the tensor algebra of gl_n", *J. Combin. Theory Ser. A* **46**:1 (1987), 79–120. MR 89a:05012 Zbl 0626.20030

[Tange 2012] R. Tange, "A bideterminant basis for a reductive monoid", *J. Pure Appl. Algebra* **216**:5 (2012), 1207–1221. MR 2012j:20138 Zbl 1251.05179

[Turaev 1989] V. G. Turaev, "Operator invariants of tangles, and R-matrices", *Izv. Akad. Nauk SSSR Ser. Mat.* **53**:5 (1989), 1073–1107, 1135. In Russian; translated in *Math. USSR, Izv.* **35**:2 (1990), 411–444. MR 91e:17011 Zbl 0707.57003

Communicated by Georgia Benkart
Received 2011-11-11 Revised 2012-04-12 Accepted 2012-06-20

rdipper@mathematik.uni-stuttgart.de
Institut für Algebra und Zahlentheorie, Universität Stuttgart,
Pfaffenwaldring 57, 70569 Stuttgart, Germany

doty@math.luc.edu
Department of Mathematics and Statistics,
Loyola University Chicago, 1023 West Sheridan Road,
Chicago, IL 60660, United States

stoll@mathematik.uni-stuttgart.de
Institut für Algebra und Zahlentheorie, Universität Stuttgart,
Pfaffenwaldring 57, 70569 Stuttgart, Germany

ALGEBRA AND NUMBER THEORY 7:5 (2013)
dx.doi.org/10.2140/ant.2013.7.1147

Weakly commensurable S-arithmetic subgroups in almost simple algebraic groups of types B and C

Skip Garibaldi and Andrei Rapinchuk

To Kevin McCrimmon on the occasion of his retirement

Let G_1 and G_2 be absolutely almost simple algebraic groups of types B_ℓ and C_ℓ, respectively, defined over a number field K. We determine when G_1 and G_2 have the same isomorphism or isogeny classes of maximal K-tori. This leads to the necessary and sufficient conditions for two Zariski-dense S-arithmetic subgroups of G_1 and G_2 to be weakly commensurable.

1. Introduction and the statement of main results 1147
2. Steinberg's theorem for algebras with involution 1150
3. Maximal tori in real groups of types B and C 1157
4. Local-global principles for embedding étale algebras with involution 1161
5. Function field analogue of Theorem 1.4 1163
6. Proof of Theorem 1.4 1164
7. Weakly commensurable subgroups and proof of Theorem 1.2 1168
8. Proofs of Proposition 1.3 and Theorem 1.5 1171
9. Alternative proofs via Galois cohomology 1172
Acknowledgements 1177
References 1177

1. Introduction and the statement of main results

This paper has two interrelated goals: first, to complete the investigation of weak commensurability of S-arithmetic subgroups of almost simple algebraic groups begun in [Prasad and Rapinchuk 2009], and second, to contribute to the classical

Garibaldi's research was partially supported by NSA grant H98230-11-1-0178 and the Charles T. Winship Fund. Rapinchuk's research was partially supported by NSF grant DMS-0965758 and the Humboldt Foundation.

MSC2010: primary 20G15; secondary 11E57, 14L35, 20G30.

Keywords: isogenous maximal tori, weak commensurability, isomorphic maximal tori.

problem of characterizing almost simple algebraic groups having the same isomorphism or the same isogeny classes of maximal tori over the field of definition.

Let G_1 and G_2 be two semisimple algebraic groups over a field F of characteristic zero, and let $\Gamma_i \subset G_i(F)$ be a (finitely generated) Zariski-dense subgroup for $i = 1, 2$. We recall in Section 7 below the notion of *weak commensurability* of Γ_1 and Γ_2 introduced in [Prasad and Rapinchuk 2009]. (This notion was inspired by some problems dealing with isospectral and length-commensurable locally symmetric spaces, and we state some geometric consequences of our main results in (7-1) and (7-2).) We further recall that the mere existence of Zariski-dense weakly commensurable subgroups implies that G_1 and G_2 either have the same Killing–Cartan type, or one of them is of type B_ℓ and the other is of type C_ℓ. Moreover, cumulatively the results of [Prasad and Rapinchuk 2009; 2010; Garibaldi 2012] give, by and large, a complete picture of weak commensurability for S-arithmetic subgroups of almost simple algebraic groups having the *same* type.

On the other hand, weak commensurability of S-arithmetic subgroups in the case where G_1 is of type B_ℓ and G_2 is of type C_ℓ has not been investigated so far — it was only pointed out in [Prasad and Rapinchuk 2009] that S-arithmetic subgroups corresponding to the split forms of such groups are indeed weakly commensurable; see also Remark 2.6 below. Our first theorem provides a complete characterization of the situations where S-arithmetic subgroups in the groups of types B and C are weakly commensurable. In its formulation we will employ the description, introduced [ibid., §1], of S-arithmetic subgroups of $G(F)$, where G is an absolutely almost simple algebraic group over a field F of characteristic zero, in terms of triples $(\mathscr{G}, K, S)$ consisting of a number field $K \subset F$, a finite subset S of places of K, and an F/K-form $\mathscr{G}$ of the adjoint group $\overline{G}$ — we briefly recall this description in Section 6.

The following definition will enable us to streamline the statements of our results.

Definition 1.1. Let $\mathscr{G}_1$ and $\mathscr{G}_2$ be absolutely almost simple algebraic groups of types B_ℓ and C_ℓ with $\ell \geqslant 2$, respectively, over a number field K. We say that $\mathscr{G}_1$ and $\mathscr{G}_2$ are *twins* (over K) if for each place v of K, both groups are simultaneously either split or anisotropic over the completion K_v.

Theorem 1.2. *Let G_1 and G_2 be absolutely almost simple algebraic groups over a field F of characteristic zero having Killing–Cartan types B_ℓ and C_ℓ $(\ell \geqslant 3)$, respectively, and let Γ_i be a Zariski-dense $(\mathscr{G}_i, K, S)$-arithmetic subgroup of $G_i(F)$ for $i = 1, 2$. Then Γ_1 and Γ_2 are weakly commensurable if and only if the groups $\mathscr{G}_1$ and $\mathscr{G}_2$ are twins.*

If Zariski-dense $(\mathscr{G}_1, K_1, S_1)$- and $(\mathscr{G}_2, K_2, S_2)$-arithmetic subgroups are weakly commensurable then necessarily $K_1 = K_2$ and $S_1 = S_2$ by [Prasad and Rapinchuk

2009, Theorem 3], so Theorem 1.2 in fact treats the most general situation. Furthermore, for $\ell = 2$ we have $\mathsf{B}_2 = \mathsf{C}_2$, so G_1 and G_2 have the same type; then Γ_1 and Γ_2 are weakly commensurable if and only if $\mathscr{G}_1 \simeq \mathscr{G}_2$ over K by [ibid., Theorem 4]. This shows that the assumption $\ell \geqslant 3$ in Theorem 1.2 is essential — the excluded case of $\ell = 2$ is treated in Theorem 1.5 below.

Turning to the second problem, that of characterizing almost simple algebraic groups having the same (isomorphic classes of) maximal tori, we would like to point out that, as we will see shortly, one gets more satisfactory results if instead of talking about *isomorphic* groups one talks about *isogenous* ones. We recall that algebraic K-groups H_1 and H_2 are called isogenous if there exists a K-group H with central K-isogenies $\pi_i \colon H \to H_i$, $i = 1, 2$. For semisimple K-groups G_1 and G_2, this amounts to the fact that the universal covers $\widetilde{G}_1$ and $\widetilde{G}_2$ are K-isomorphic, and for K-tori T_1 and T_2 this simply means that there exists a K-isogeny $T_1 \to T_2$. Furthermore, we say that two semisimple K-groups G_1 and G_2 have the same isogeny classes of maximal K-tori if every maximal K-torus T_1 of G_1 is K-isogenous to some maximal K-torus T_2 of G_2, and vice versa. Unsurprisingly, K-isogenous groups have the same isogeny classes of maximal tori. Using the results from [Prasad and Rapinchuk 2009; Garibaldi 2012], we prove the following partial converse for almost simple groups over number fields.

Proposition 1.3. *Let G_1 and G_2 be absolutely almost simple algebraic groups over a number field K. Assume that G_1 and G_2 have the same isogeny classes of maximal K-tori. Then at least one of the following holds*:

(1) *G_1 and G_2 are K-isogenous.*

(2) *G_1 and G_2 are of the same Killing–Cartan type*, *which is one of the following*: A_ℓ *for* $\ell > 1$, $\mathsf{D}_{2\ell+1}$ *for* $\ell > 1$, *or* E_6.

(3) *One of the groups is of type* B_ℓ *and the other of type* C_ℓ *for some* $\ell \geqslant 3$.

We will prove the proposition in Section 8. As Theorem 1.5 below shows, it is possible for two isogenous, but not isomorphic, groups to have the same isomorphism classes of maximal K-tori, so the conclusion in (1) cannot be strengthened even if we assume that G_1 and G_2 have the same maximal tori. On the other hand, for each of the types listed in (2) one can construct nonisomorphic simply connected, and hence nonisogenous, groups of this type having the same tori [Prasad and Rapinchuk 2009, §9], so these types are genuine exceptions. In this paper, we will sharpen case (3). Specifically, we prove the following in Section 6.

Theorem 1.4. *Let G_1 and G_2 be absolutely almost simple algebraic groups over a number field K of types* B_ℓ *and* C_ℓ, *respectively*, *for some* $\ell \geqslant 3$.

(1) *The groups G_1 and G_2 have the same isogeny classes of maximal K-tori if and only if they are twins.*

(2) *The groups G_1 and G_2 have the same isomorphism classes of maximal K-tori if and only if they are twins, G_1 is adjoint, and G_2 is simply connected.*

We note that one can give examples of groups G_1 and G_2 of types B_ℓ and C_ℓ, respectively, over the field $\mathbb{R}$ of real numbers, that are neither split nor anisotropic but nevertheless have the same isomorphism classes of maximal $\mathbb{R}$-tori; see Example 3.6. This shows Theorem 1.4, unlike many statements about algebraic groups over number fields, is *not* a global version of the corresponding theorem over local fields. What is crucial for the proof of Theorem 1.4 (and also Theorem 1.2) is that if the real groups G_1 and G_2 are neither split nor anisotropic with G_1 adjoint and G_2 simply connected then they cannot have the same maximal $\mathbb{R}$-tori; see Corollary 3.4.

***The special case* $\mathsf{B}_2 = \mathsf{C}_2$.** Theorem 1.4 completely settles the question of when the groups of types B_ℓ and C_ℓ have isogenous tori for $\ell \geqslant 3$. The case where $\ell = 2$ is special because the root systems B_2 and C_2 are the same.

Let G_1 and G_2 be groups of type $\mathsf{B}_2 = \mathsf{C}_2$. They have the same isogeny classes of maximal tori if and only if they are isogenous by Lemma 8.1 below or [Prasad and Rapinchuk 2009, Theorem 7.5(2)]. In particular, when G_1 and G_2 are both adjoint or both simply connected, they have the same isogeny classes of maximal tori if and only if $G_1 \simeq G_2$ if and only if they have the same maximal tori. It remains only to give a condition for G_1 and G_2 to have the same maximal tori when one is adjoint and the other is simply connected, which we now do.

Theorem 1.5. *Let q_1 and q_2 be 5-dimensional quadratic forms over a number field K. The groups $G_1 = \mathrm{SO}(q_1)$ and $G_2 = \mathrm{Spin}(q_2)$ have the same isomorphism classes of maximal K-tori if and only if*

(1) *q_1 is similar to q_2, and*

(2) *q_1 and q_2 are either both split or both anisotropic at every completion of K.*

Notation. For a number field K, we let V^K denote the set of all places, and let V_∞^K and V_f^K denote the subsets of archimedean and nonarchimedean places. Given a reductive algebraic group G defined over a field K, for any field extension L/K we let $\mathrm{rk}_L\, G$ denote the L-rank of G, that is, the dimension of a maximal L-split torus.

We write $r\langle a\rangle$ for the symmetric bilinear form $(x, y) \mapsto a \sum_{i=1}^r x_i y_i$ on K^r, and adopt similar notation for quadratic forms and hermitian forms.

In Section 6, we systematically use the following: For G_1 and G_2 absolutely almost simple groups of types B_ℓ and C_ℓ, respectively, we put $G_1^\natural$ for the adjoint group of G_1 ("SO"), and $G_2^\natural$ for the simply connected cover of G_2 ("Sp").

2. Steinberg's theorem for algebras with involution

Our proofs of Theorems 1.2 and 1.4 rely on the well-known fact that groups of classical types can be realized as special unitary groups associated with simple algebras with involutions, so their maximal tori correspond to certain commutative étale subalgebras invariant under the involution. This description enables us to apply the local-global principles for the existence of an embedding of an étale algebra with an involutory automorphism into a simple algebra with an involution [Prasad and Rapinchuk 2010]. To ensure the existence of local embeddings, we will use an analogue for algebras with involution of the theorem, due to Steinberg [1965], asserting that if G_0 is a quasisplit simply connected almost simple algebraic group over a field K and G is an inner form of G_0 over K, then any maximal K-torus T of G admits a K-defined embedding into G_0. The required analogue roughly states that if (A, τ) is an algebra with involution such that the corresponding group is quasisplit then any commutative étale algebra with involution (E, σ) that can potentially embed in (A, τ) does embed. It can be deduced from the original Steinberg's theorem along the lines of [Gille 2004, Proposition 3.2(b)], but in fact one can give a simple direct argument. To our knowledge, this has not been recorded in the literature. Further, the argument for type B_n (in Proposition 2.5) extends with minor modifications to other types. So, despite the fact that we will only use this statement for algebras corresponding to groups of type B_n and C_n, we will give the argument for all classical types. We begin by briefly recalling the types of algebras with involution arising in this context, indicating in each case the étale subalgebras that give maximal tori.

Description of tori in terms of étale algebras. Let A be a central simple algebra of dimension n^2 over a field L of characteristic other than 2, and let τ be an involution of A. Set $K = L^{\tau}$. We recall that τ is said to be of the *first* or *second* kind if the restriction $\tau|_L$ is trivial or nontrivial, respectively. Furthermore, if τ is an involution of the first kind, then it is either *symplectic* (that is, $\dim_K A^{\tau} = n(n-1)/2$) or *orthogonal* (that is, $\dim_K A^{\tau} = n(n+1)/2$).

We also recall the well-known correspondence between involutions on $A = M_n(L)$ and nondegenerate hermitian or skew-hermitian forms on L^n [Knus et al. 1998]: Given such a form f, there exists a unique involution τ_f such that

$$f(ax, y) = f(x, \tau_f(a)y)$$

for all $x, y \in L^n$ and all $a \in A$; then the pair $(M_n(L), \tau_f)$ will be denoted by A_f. Moreover, f is symmetric or skew-symmetric if and only if τ_f is orthogonal or symplectic, respectively. Conversely, for any involution τ there exists a form f on L^n of appropriate type such that $\tau = \tau_f$, and any two such forms are proportional. (For involutions of the second kind one can pick the corresponding form to be either hermitian or skew-hermitian as desired.)

Type ${}^2\mathsf{A}_{\ell}$. Let (A, τ) be a central simple L-algebra of dimension n^2 with an involution τ of the second kind. Then $G = \mathrm{SU}(A, \tau)$ is an absolutely almost simple simply connected K-group of type ${}^2\mathsf{A}_\ell$ with $\ell = n-1$, and conversely any such group corresponds to an algebra with involution (A, τ) of this kind. Any τ-invariant étale commutative subalgebra $E \subset A$ gives a maximal K-torus

$$T = \mathrm{R}_{E/K}(\mathrm{GL}_1) \cap G = \mathrm{SU}(E, \tau|_E)$$

of G, and all maximal K-tori are obtained this way; see, for example, [Prasad and Rapinchuk 2010, Proposition 2.3]. The group G is quasisplit if and only if $A = M_n(L)$ and $\tau = \tau_h$, where h is a nondegenerate hermitian form on L^n of Witt index $[n/2]$.

Type B_ℓ $(\ell \geqslant 2)$. Let $A = M_n(K)$ with $n = 2\ell + 1$, and let τ be an orthogonal involution of A. Then $\tau = \tau_f$ for some nondegenerate symmetric bilinear form f on K^n, and $G = \mathrm{SU}(A, \tau) = \mathrm{SO}(f)$ is an adjoint group of type B_ℓ, and every such group is obtained this way. Furthermore, maximal K-tori T of G bijectively correspond to maximal commutative étale τ-invariant subalgebras E of A (of dimension n) such that $\dim_K E^\tau = \ell + 1$ under the correspondence given by $T = \mathrm{R}_{E/K}(\mathrm{GL}_1) \cap G = \mathrm{SU}(E, \tau|_E)$. Furthermore, any such algebra admits a decomposition

$$(E, \tau) = (E', \tau') \times (K, \mathrm{id}_K), \tag{2-1}$$

where $E' \subset E$ is a τ-invariant subalgebra of dimension 2ℓ. Finally, the group G is quasisplit (in fact, split) if and only if f has Witt index ℓ.

Type C_ℓ $(\ell \geqslant 2)$. Let A be a central simple K-algebra of dimension n^2 with $n = 2\ell$, and let τ be a symplectic involution of A. Then $G = \mathrm{SU}(A, \tau)$ is an absolutely almost simple simply connected group of type C_ℓ, and all such groups are obtained this way. Maximal K-tori of G correspond to maximal commutative étale τ-invariant subalgebras $E \subset A$ (of dimension n) such that $\dim_K E^\tau = \ell$ in the fashion described above. The group G is quasisplit (in fact, split) if and only if $A = M_n(K)$. Then $\tau = \tau_f$, where f is a nondegenerate skew-symmetric form on K^n; there is only one equivalence class of such forms, so in this case $G \simeq \mathrm{Sp}_n$.

Type ${}^{1,2}\mathsf{D}_\ell$ $(\ell \geqslant 4)$. Let A be a central simple K-algebra of dimension n^2, where $n = 2\ell$, and let τ be an orthogonal involution of A. Then $G = \mathrm{SU}(A, \tau)$ is an almost absolutely simple K-group of type ${}^{1,2}\mathsf{D}_\ell$ that is neither simply connected nor adjoint, and any K-group of this type is K-isogenous to such a group. Maximal K-tori of G correspond to maximal commutative étale τ-invariant subalgebras $E \subset A$ (of dimension n^2) such that $\dim_K E^\tau = \ell$. The group G is quasisplit if and only if $A = M_n(K)$ and $\tau = \tau_f$, where f is a symmetric bilinear form on K^n of Witt index $\ell - 1$ or ℓ.

Summary. Thus, if A is a central simple L-algebra of dimension n^2 (and $L = K$ for all types except ${}^2\mathsf{A}_\ell$) then maximal K-tori of the algebraic K-group $G = \mathrm{SU}(A, \tau)$

correspond in the manner described above to maximal abelian étale τ-invariant subalgebras $E \subset A$ with $\dim_L E = n$ such that for $\sigma = \tau|_E$ we have

$$\dim_K E^\sigma = \begin{cases} n & \text{if } \sigma|_L \neq \mathrm{id}_L, \\ [(n+1)/2] & \text{if } \sigma|_L = \mathrm{id}_L. \end{cases} \tag{2-2}$$

(The condition is automatically satisfied if $\sigma|_L \neq \mathrm{id}_L$.)

Now, let (E, σ) be an n-dimensional commutative étale L-algebra with an involution satisfying (2-2). Then the question of whether the K-torus $T = \mathrm{SU}(E, \sigma)^\circ$ can be embedded into $G = \mathrm{SU}(A, \tau)$, where A is a central simple L-algebra of dimension n^2 with an involution τ such that $\sigma|_L = \tau|_L$, translates into the question of whether there is an embedding $(E, \sigma) \hookrightarrow (A, \tau)$ of L-algebras with involution, which we will now investigate in the cases of interest to us. We note that if G is quasisplit, then $A = M_n(L)$ in all cases. In this case, the universal way to construct an embedding $(E, \sigma) \hookrightarrow (M_n(L), \tau)$ is described in the following well-known statement.

Proposition 2.1. *Let (E, σ) be an n-dimensional commutative étale L-algebra with an involution σ.*

(i) *For any $b \in E^\times$, the map $\phi_b \colon E \times E \to K$ given by $\phi_b(x, y) = \mathrm{tr}_{E/L}(x \cdot b \cdot \sigma(y))$ is a nondegenerate sesquilinear form, which is hermitian or skew-hermitian if and only if b is such.*

(ii) *Let $b \in E^\times$ be hermitian or skew-hermitian, and let τ_{ϕ_b} be the involution on $A := \mathrm{End}_L(E) \simeq M_n(L)$ corresponding to ϕ_b; then the regular representation of E gives an embedding $(E, \sigma) \hookrightarrow (A, \tau_{\phi_b}) = A_{\phi_b}$ of algebras with involution.*

(iii) *Let τ be an involution on $A = M_n(L)$, and let f be a hermitian or skew-hermitian form on L^n such that $\tau_f = \tau$. Then the following conditions are equivalent*:

(a) *There exists $b \in E^\times$ of the same type as f such that ϕ_b is equivalent to f.*

(b) *There exists a form h on $E \simeq L^n$ that is equivalent to f and that satisfies*

$$h(ax, y) = h(x, \sigma(a)y) \quad \textit{for all } a, x, y \in E. \tag{2-3}$$

(c) *There exists an embedding $(E, \sigma) \hookrightarrow (A, \tau)$ as L-algebras with involutions.*

Sketch of proof. The nondegeneracy of ϕ_b in (i) follows from the fact that the L-bilinear form on E given by $(x, y) \mapsto \mathrm{tr}_{E/L}(xy)$ is nondegenerate as E/L is étale; other assertions in (i) and (ii) are immediate consequences of the definitions. The implications (a) $\Longrightarrow$ (b) $\Longrightarrow$ (c) in (iii) are obvious, and the equivalence of (a) and (c) (which we will not need) is established in [Prasad and Rapinchuk 2010, Proposition 7.1]. □

We also note that in fact any nondegenerate hermitian/skew-hermitian form h on E satisfying (2-3) is of the form ϕ_b for some $b \in E^\times$ of the respective type. Indeed, since the form ϕ_1 is nondegenerate, we can write h in the form $h(x, y) = \mathrm{tr}_{E/L}(x \cdot g(\sigma(y)))$ for some K-linear automorphism g of E. Then (2-3) implies that g is E-linear, and therefore is of the form $g(x) = bx$ for some $b \in E^\times$, which will necessarily be of appropriate type.

Example 2.2 (involutions of the first kind). According to [Prasad and Rapinchuk 2010, Proposition 2.2], if $L = K$ and (E, σ) is a K-algebra with involution of dimension $n = 2\ell$ satisfying (2-2), then $(E, \sigma) \simeq (F[\delta]/(\delta^2 - d), \theta)$, where $F = E^\sigma$, $d \in F^\times$, and $\theta(\delta) = -\delta$.

For invertible $b \in E^\sigma$ and $x_i, y_i \in F$, we have

$$\phi_b(x_1 + y_1\delta, x_2 + y_2\delta) = \mathrm{tr}_{E/K}(bx_1x_2 - bdy_1y_2) = \mathrm{tr}_{F/K}(2b(x_1x_2 - dy_1y_2)),$$

so ϕ_b is the transfer from F to K of the symmetric bilinear form $\langle 2b, -2bd\rangle$. Clearly, if E is $F \times F$, then ϕ_b is hyperbolic.

The example gives the entries in the ϕ_b column of Table 1.

Proposition 2.3 (type C). *Let (E, σ) be an étale K-algebra of dimension $n = 2\ell$ with involution satisfying (2-2). Then for every symplectic involution τ on $M_n(K)$, there is a K-embedding $(E, \sigma) \hookrightarrow (M_n(K), \tau)$.*

Proof. It follows from the structure of (E, σ) in the example that there exists a skew-symmetric *invertible* $b \in E$ (one can take, for example, the element corresponding to δ); then by Proposition 2.1(i), the form ϕ_b is nondegenerate and skew-symmetric. On the other hand, since τ is symplectic, we have $\tau = \tau_f$ for some nondegenerate skew-symmetric form f on K^n. As any two such forms are equivalent, our assertion follows from Proposition 2.1(iii). □

To handle the algebras corresponding to types B and D, we need the following.

Lemma 2.4. *Let (E, σ) be a commutative étale K-algebra with involution of dimension $n = 2\ell$ satisfying (2-2). Then there exists a nondegenerate symmetric bilinear form h on E that satisfies (2-3) and has Witt index $\geqslant \ell - 1$.*

Proof. If K is finite then one can take, for example, $h = \phi_1$, so we can assume in the rest of the argument that K is infinite. It follows from the description of E that $(E \otimes_K \overline{K}, \sigma \otimes \mathrm{id}_K) \simeq (M, \mu)$ for $\overline{K}$ an algebraic closure of K, where $M = \prod_{i=1}^{\ell}(\overline{K} \times \overline{K})$ and μ acts on each copy of $\overline{K} \times \overline{K}$ by switching components. Viewing M as an affine n-space, we consider the K-defined subvariety $M_- := \{x \in M \mid \mu(x) = -x\}$. Clearly, M_- is a K-defined vector space, so the K-points $E_- := M_- \cap E$ are Zariski-dense in M_-. On the other hand, let $U \subset M$ be the Zariski-open subvariety of elements with pairwise distinct components; then any

$x \in U$ generates M as a $\overline{K}$-algebra. Furthermore, it is easy to see that $U \cap M_- \neq \varnothing$, so $U \cap E_- \neq \varnothing$.

Fix $e \in U \cap E_-$; then $1, e, \dots, e^{n-1}$ form a K-basis of E. For $x \in E$ we define $c_i(x)$ for $i = 0, \dots, n-1$ so that $x = \sum_{i=0}^{n-1} c_i(x)e^i$. Set

$$h(x, y) := c_{n-2}(x\sigma(y)).$$

Clearly, h is symmetric bilinear and satisfies (2-3). Let us show that h is nondegenerate. If $x = \sum_{i=0}^{n-1} c_i(x)e^i$ is in the radical of h, then so is $\sigma(x)$, and therefore also $x_+ := \sum_{i=0}^{\ell-1} c_{2i}(x)e^{2i}$ and $x_- := \sum_{i=0}^{\ell-1} c_{2i+1}(x)e^{2i+1}$. From $h(x_+, 1) = 0$, $h(x_+, e^2) = 0$, etc., we successively obtain that $c_{n-2}(x) = 0$, $c_{n-4}(x) = 0$, etc., that is, $x_+ = 0$. Furthermore, we have $0 = h(x_-, e^{-1}) = -c_{n-1}(x)$. Then from $h(x_-, e) = 0$, $h(x_-, e^3) = 0$, etc., we successively obtain $c_{n-3}(x) = 0$, $c_{n-5}(x) = 0$, etc. Thus, $x_- = 0$; hence $x = 0$, as required. It remains to observe that the subspace spanned by $1, e, \dots, e^{\ell-2}$ is totally isotropic with respect to h. □

Remark. In an earlier version of this paper, we constructed h in Lemma 2.4 in the form $h = \phi_b$ using some matrix computations. The current proof, which minimizes computations, was inspired by [Bhargava and Gross 2011, §5].

Proposition 2.5 (type B). *Let (E, σ) be an étale K-algebra of dimension $n = 2\ell+1$ with involution satisfying (2-2). If τ is an orthogonal involution on $A = M_n(K)$ such that $\tau = \tau_f$, where f is a nondegenerate symmetric bilinear form on K^n of Witt index ℓ, then there exists an embedding $(E, \sigma) \hookrightarrow (A, \tau)$ of K-algebras with involution.*

Proof. Pick a decomposition (2-1), and then use Lemma 2.4 to find a form h' on E' with the properties described therein. We can write $h' = h'_1 \perp h'_2$, where h'_1 is a direct sum of $\ell - 1$ hyperbolic planes and h'_2 is a binary form. Choose a 1-dimensional form h'' so that $h'_2 \perp h''$ is isotropic, and consider $h = h' \perp h''$ on $E = E' \times K$. Then h is a nondegenerate symmetric bilinear form on E satisfying (2-3) and having Witt index ℓ. So, h is equivalent to f; hence (E, σ) embeds in (A, τ) by Proposition 2.1(iii). □

Remark 2.6. Let now G_1 be the K-split adjoint group $\mathrm{SO}_{2\ell+1}$ of type B_ℓ and G_2 be the K-split simply connected group $\mathrm{Sp}_{2\ell}$ of type C_ℓ, where $\ell \geqslant 2$. It was observed in [Prasad and Rapinchuk 2009, Example 6.7] that G_1 and G_2 have the same isomorphism classes of maximal K-tori over any field K of characteristic not 2. This was derived from the fact that G_1 and G_2 have isomorphic Weyl groups using the results of [Gille 2004; Raghunathan 2004]. Now, we are in a position to give a much simpler explanation of this phenomenon. Indeed, $G_1 = \mathrm{SU}(A_1, \tau_1)$, where $A_1 = M_{2\ell+1}(K)$ and τ_1 is an orthogonal involution on A_1 corresponding to a nondegenerate symmetric bilinear form on $K^{2\ell+1}$ of Witt index ℓ, and $G_2 = \mathrm{SU}(A_2, \tau_2)$, where $A_2 = M_{2\ell}(K)$ and τ_2 is a symplectic involution on A_2

corresponding to a nondegenerate skew-symmetric form on $K^{2\ell}$. Any maximal K-torus T_2 of G_2 is of the form $\mathrm{SU}(E_2, \sigma_2)$, where E_2 is a 2ℓ-dimensional commutative τ_2-invariant subalgebra of A_2, and $\sigma_2 = \tau_2|_{E_2}$, with (E_2, σ_2) satisfying (2-2). Set $(E_1, \sigma_1) = (E_2, \sigma_2) \times (K, \mathrm{id}_K)$. According to Proposition 2.5, there exists an embedding $(E_1, \sigma_1) \hookrightarrow (A_1, \tau_1)$, which gives rise to a K-isomorphism between T_2 and the maximal K-torus $T_1 = \mathrm{SU}(E_1, \sigma_1)$ of G_1. This, combined with the symmetric argument based on Proposition 2.3, yields the required fact. Then, repeating the argument given in [Prasad and Rapinchuk 2009, Example 6.7], we conclude that if K is a number field then for any finite subset $S \subset V^K$ containing V^K_∞, the S-arithmetic subgroups of G_1 and G_2 are weakly commensurable.

Turning now to type D_ℓ, we first observe that if (E, σ) is a K-algebra with involution of dimension $n = 2\ell$ satisfying (2-2) then the determinant — viewed as an element of $K^\times/K^{\times 2}$ — of the symmetric bilinear form ϕ_b for invertible $b \in E^\sigma$ does not depend on b [Brusamarello et al. 2003, Corollary 4.2] and will be denoted $d(E, \sigma)$. Now, if τ is an involution on $A = M_n(K)$ that corresponds to a symmetric bilinear form f on K^n having determinant $d(f)$, then it follows from Proposition 2.1(iii) that an embedding $(E, \sigma) \hookrightarrow (A, \tau)$ can exist only if $d(E, \sigma) = d(f)$ in $K^\times/K^{\times 2}$.

Proposition 2.7. *Let (E, σ) be an étale K-algebra of dimension $n = 2\ell$ with involution satisfying (2-2). If τ is an orthogonal involution on $A = M_n(K)$ such that $\tau = \tau_f$, where f is a nondegenerate symmetric bilinear form on K^n of Witt index at least $\ell - 1$ such that $d(E, \sigma) = d(f)$ (in $K^\times/K^{\times 2}$), then there exists an embedding $(E, \sigma) \hookrightarrow (A, \tau)$ of K-algebras with involution.*

Proof. Let h be the symmetric bilinear form on E constructed in Lemma 2.4. As we observed after Proposition 2.1, h is actually of the form $h = \phi_b$ for some invertible $b \in E^\sigma$, so $d(h) = d(E, \sigma)$. We can write $h = h_1 \perp h_2$, where h_1 is a direct sum of $\ell - 1$ hyperbolic planes and h_2 is a binary form. Similarly, $f = f_1 \perp f_2$, where f_1 is a direct sum of $\ell - 1$ hyperbolic planes and f_2 is binary. Then $d(E, \sigma) = d(f)$ implies that $d(h_2) = d(f_2)$, so h_2 and f_2 are similar. Thus, a suitable multiple of h is equivalent to f, and our claim follows from Proposition 2.1(iii). □

Finally, we will treat algebras corresponding to the groups of type ${}^2\mathsf{A}_\ell$. Here L will be a quadratic extension of K and all involutions will restrict to the nontrivial automorphism of L/K.

Proposition 2.8 (type A). *Let (E, σ) be an étale n-dimensional L-algebra with involution. If τ is a unitary involution on $A = M_n(L)$ such that $\tau = \tau_f$, where f is a hermitian form on L^n having Witt index $m := [n/2]$, then there exists an embedding $(E, \sigma) \hookrightarrow (A, \tau)$ of L-algebras with involution.*

Proof. It is enough to construct a nondegenerate hermitian form on E that satisfies (2-3) and has Witt index m. If K is finite, one can take, for example, $h = \phi_1$, so we can assume that K is infinite. Set $F = E^{\sigma}$ so that $E = F \otimes_K L$. Since K is infinite, arguing as in the proof of Lemma 2.4, one can find $e \in F$ so that $F = K[e]$. Then any $x \in E$ admits a unique presentation of the form $x = \sum_{i=0}^{n-1} e^i \otimes c_i(x)$ with $c_i(x) \in L$. Define

$$h(x, y) := c_{n-1}(x\sigma(y)).$$

It is easy to see h is a hermitian form satisfying (2-3); let us show that it is nondegenerate. If x is in the radical of h, then from $h(x, 1) = 0$, $h(x, e) = 0$, etc., we successively obtain that $c_{n-1}(x) = 0$, $c_{n-2}(x) = 0$, etc. Thus, $x = 0$, proving the nondegeneracy of h. Since $2(m-1) < n-1$, the subspace spanned by $1, e, \dots, e^{m-1}$ is totally isotropic; hence the Witt index of h is m, as required. □

3. Maximal tori in real groups of types B and C

This section is devoted to determining the isomorphism classes of maximal tori in certain linear algebraic groups, primarily of types B and C, over the real numbers. Recall that every torus T over $\mathbb{R}$ is $\mathbb{R}$-isomorphic to the product

$$(\mathrm{GL}_1)^{\alpha} \times (\mathrm{R}^{(1)}_{\mathbb{C}/\mathbb{R}}(\mathrm{GL}_1))^{\beta} \times (\mathrm{R}_{\mathbb{C}/\mathbb{R}}(\mathrm{GL}_1))^{\gamma} \tag{3-1}$$

for uniquely determined nonnegative integers α, β, γ [Voskresenskiĭ 1998, p. 64], and then the group $T(\mathbb{R})$ is topologically isomorphic to $(\mathbb{R}^{\times})^{\alpha} \times (S^1)^{\beta} \times (\mathbb{C}^{\times})^{\gamma}$, where S^1 is the group of complex numbers of modulus 1. The fact that T is isomorphic to a maximal $\mathbb{R}$-torus of a given reductive $\mathbb{R}$-group G typically imposes serious restrictions on the numbers α, β and γ. To illustrate this, we first consider the following easy example.

Example 3.1. *Every maximal $\mathbb{R}$-torus in $G = \mathrm{GL}_{n,\mathbb{H}}$, where $\mathbb{H}$ is the algebra of Hamiltonian quaternions, is isomorphic to* $(\mathrm{R}_{\mathbb{C}/\mathbb{R}}(\mathrm{GL}_1))^n$. Indeed, every maximal $\mathbb{R}$-torus in G is of the form $\mathrm{R}_{E/\mathbb{R}}(\mathrm{GL}_1)$, where E is a maximal commutative $2n$-dimensional étale subalgebra of $A = M_n(\mathbb{H})$. Any commutative $2n$-dimensional étale $\mathbb{R}$-algebra E is isomorphic to $\mathbb{R}^{\alpha} \times \mathbb{C}^{\gamma}$ with $\alpha + 2\gamma = 2n$. But in order for E to have an $\mathbb{R}$-embedding in A, we must have $\alpha = 0$ and then $\gamma = n$ [Prasad and Rapinchuk 2010, 2.6], so our claim follows.

We now recall the standard notation for some classical real algebraic groups. We let $\mathrm{SO}(r, n-r)$ denote the special orthogonal group of the n-dimensional quadratic form $q = r\langle 1\rangle \perp (n-r)\langle -1\rangle$. Similarly, we let $\mathrm{Sp}(r, n-r)$ denote the special unitary group of the n-dimensional hermitian form $h = r\langle 1\rangle \perp (n-r)\langle -1\rangle$ over $\mathbb{H}$ with the standard involution. Every adjoint $\mathbb{R}$-group of type B_{ℓ} is isomorphic to some $\mathrm{SO}(r, n-r)$ for $n = 2\ell+1$ and some $0 \leqslant r \leqslant n$, and every nonsplit simply connected $\mathbb{R}$-group of type C_ℓ is isomorphic to $\mathrm{Sp}(r, \ell-r)$ some $0 \leqslant r \leqslant \ell$.

Lemma 3.2 (adjoint B_ℓ over $\mathbb{R}$). *The maximal $\mathbb{R}$-tori in $G = \mathrm{SO}(r, n-r)$, where $n = 2\ell+1$, are of the form* (3-1) *with $\alpha+\beta+2\gamma = \ell$ and $\alpha+2\gamma \leqslant s := \min(r, n-r)$.*

Proof. Let τ be the involution on $A = M_n(K)$ that corresponds to the symmetric bilinear form f associated with the quadratic form $q = r\langle 1\rangle \perp (n-r)\langle -1\rangle$ so that $G = \mathrm{SU}(A, \tau)$. Let T be a maximal $\mathbb{R}$-torus of G written in the form (3-1). Since the rank of G is ℓ, we immediately obtain $\dim T = \alpha+\beta+2\gamma = \ell$. Furthermore, we have $T = \mathrm{SU}(E, \sigma)$, where $E \subset A$ is a τ-invariant maximal commutative étale subalgebra, $\sigma = \tau|_E$, and (2-2) holds. There are exactly 4 isomorphism classes of indecomposable étale $\mathbb{R}$-algebras with involution, which are listed in Table 1. Using this information, we can write

$$(E, \sigma) = \mathbb{R}^{\delta_1} \times (\mathbb{R}\times\mathbb{R})^{\delta_2} \times \mathbb{C}^{\delta_3} \times (\mathbb{C}\times\mathbb{C})^{\delta_4},$$

where the involutions on factors are as in the table. Comparing this with the structure of T, we obtain $\delta_2 = \alpha$, $\delta_3 = \beta$, and $\delta_4 = \gamma$. According to Proposition 2.1(iii), there exists $b \in E^\sigma$ such that ϕ_b is equivalent to f. But the Witt index of f is s (which equals the $\mathbb{R}$-rank of G), and the Witt index of ϕ_b is $\geqslant \delta_2 + 2\delta_4$. Thus, $\alpha + 2\gamma \leqslant s$. (We note that $\mathrm{rk}_{\mathbb{R}}\, T = \alpha+\gamma$, immediately yielding the restriction $\alpha+\gamma \leqslant s$. So, the restriction we have actually obtained is stronger than one can a priori expect.)

Conversely, suppose α, β, γ satisfy the two constraints, and assume that $r > n-r$ (otherwise we can replace the quadratic form q defining G with $-q$); in particular, $r > \ell$. Consider the étale $\mathbb{R}$-algebra

$$(E, \sigma) = \mathbb{R} \times (\mathbb{R}\times\mathbb{R})^{\alpha} \times \mathbb{C}^{\beta} \times (\mathbb{C}\times\mathbb{C})^{\gamma} =: (E_1, \sigma_1) \times \cdots \times (E_4, \sigma_4)$$

of dimension $1 + 2\alpha + 2\beta + 4\gamma = 2\ell+1 = n$, where the involutions on the factors $\mathbb{R}$, $\mathbb{R}\times\mathbb{R}, \ldots$ are as described in Table 1. (Clearly, E satisfies (2-2).) Let us show that there exists $b = (b_1, \ldots, b_4) \in E^\sigma$ such that ϕ_b is equivalent to f. Set $b_2 = ((1,1), \ldots, (1,1))$ and $b_4 = ((1,1), \ldots, (1,1))$. Then the quadratic form associated with the bilinear form $(\phi_{2,4})_{(b_2,b_4)}$ on $E_2 \times E_4$ is equivalent to $(\alpha+2\gamma)(\langle 1\rangle \perp \langle -1\rangle)$. Since $t := (n-r) - (\alpha+2\gamma) \geqslant 0$, we can choose $b_1 = \pm 1$ and $b_3 = (\pm 1, \ldots, \pm 1)$ so that the quadratic form associated with $(\phi_{1,3})_{(b_1,b_3)}$ is equivalent to $(2\beta+1-t)\langle 1\rangle \perp t\langle -1\rangle$. Then $b = (b_1, \ldots, b_4)$ is as required. By Proposition 2.1(iii), there exists an embedding $(E, \sigma) \hookrightarrow (A, \tau)$, and therefore an $\mathbb{R}$-defined embedding $\mathrm{SU}(E, \sigma) \hookrightarrow \mathrm{SU}(A, \tau) = G$. Finally, it follows from our construction and Table 1 that $T = \mathrm{SU}(E, \sigma)$ is a torus having the required structure. □

Lemma 3.3 (simply connected C_ℓ over $\mathbb{R}$). *The maximal $\mathbb{R}$-tori in the group $G = \mathrm{Sp}(r, \ell-r)$ are of the form* (3-1) *with $\alpha = 0$, $\beta+2\gamma = \ell$ and $\gamma \leqslant s := \min(r, \ell-r)$.*

E	σ	ϕ_b for $b \in E^\sigma$	$\mathrm{SU}(E,\sigma)$
$\mathbb{R}$	Id	$\langle b \rangle$	$\{1\}$
$\mathbb{R}\times\mathbb{R}$	switch	$\langle 1, -1\rangle$	GL_1
$\mathbb{C}$	conjugation	$\langle b, b\rangle$	$\mathrm{R}^{(1)}_{\mathbb{C}/\mathbb{R}}(\mathrm{GL}_1)$
$\mathbb{C}\times\mathbb{C}$	switch	$\langle 1, -1\rangle \oplus \langle 1, -1\rangle$	$\mathrm{R}_{\mathbb{C}/\mathbb{R}}(\mathrm{GL}_1)$

Table 1. Isomorphism classes of indecomposable étale $\mathbb{R}$-algebras with involution and their associated symmetric bilinear forms and unitary groups.

Proof. Let τ be the involution on $A = M_\ell(\mathbb{H})$ that gives rise to the hermitian form $f = r\langle 1\rangle \perp (\ell - r)\langle -1\rangle$, so that $G = \mathrm{SU}(A,\tau)$. Every maximal $\mathbb{R}$-torus T of G is of the form $T = \mathrm{SU}(E,\sigma)$ for some (2ℓ)-dimensional étale τ-invariant subalgebra E of A, where $\sigma = \tau|_E$ and condition (2-2) holds. As in Example 3.1, $E \simeq \mathbb{C}^\ell$ as $\mathbb{R}$-algebras, and therefore $(E,\sigma) = \mathbb{C}^{\delta_1} \times (\mathbb{C}\times\mathbb{C})^{\delta_2}$, where the involutions on $\mathbb{C}$ and $\mathbb{C}\times\mathbb{C}$ are as in Table 1. Then in (3-1) for $T = \mathrm{SU}(E,\sigma)$ we have $\alpha = 0$, $\beta = \delta_1$ and $\gamma = \delta_2$. By dimension count, we get $\beta + 2\gamma = \ell$. Furthermore, $\gamma = \mathrm{rk}_{\mathbb{R}}\, T \leqslant \mathrm{rk}_{\mathbb{R}}\, G = s$.

Conversely, suppose that T has parameters α, β and γ satisfying our constraints. Consider $(E,\sigma) = \mathbb{C}^\beta \times (\mathbb{C}\times\mathbb{C})^\gamma$ with the involutions as above, and assume (as we may) that $\ell - r \leqslant r$. Note that

$$(z, w) \mapsto \begin{pmatrix} z & 0 \\ 0 & \overline{w} \end{pmatrix}$$

defines an embedding of algebras with involutions $\mathbb{C}\times\mathbb{C} \hookrightarrow (M_2(\mathbb{H}), \theta)$, where $\theta(x) = J^{-1}\bar{x}^t J$ with $J = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, where $\bar{x}$ is obtained by applying quaternionic conjugation to all entries. Consider the involution $\hat{\theta}$ on A given by $\hat{\theta}(x) = \hat{J}^{-1}\bar{x}^t\hat{J}$, where

$$\hat{J} = \mathrm{diag}(\underbrace{1, \dots, 1}_{r-\gamma}, \underbrace{-1, \dots, -1}_{\beta-(r-\gamma)}, \underbrace{J, \dots, J}_{\gamma}).$$

Then it follows from our construction that there exists an embedding $(E,\sigma) \hookrightarrow (A,\theta)$. Noting that $(A,\tau) \simeq (A,\theta)$, we obtain an embedding $(E,\sigma) \hookrightarrow (A,\tau)$. So, there exists an $\mathbb{R}$-embedding $\mathrm{SU}(E,\sigma) \hookrightarrow \mathrm{SU}(A,\tau) = G$, and it remains to observe that $T = \mathrm{SU}(E,\sigma)$ is a torus having the required structure. □

Alternatively, the results of Lemmas 3.2 and 3.3 can be deduced from the more general classification of maximal $\mathbb{R}$-tori in simple real algebraic groups obtained in [Đoković and Thǎńg 1994]. For the reader's convenience we have included the direct proofs above, written in the same language as the rest of the paper.

Corollary 3.4. *Let G_1 be an adjoint real group of type B_ℓ, and let G_2 be a simply connected real group of type C_ℓ. The groups G_1 and G_2 have the same isomorphism classes of maximal $\mathbb{R}$-tori if and only if G_1 and G_2 are either both split or both anisotropic.*

Proof. Since every $\mathbb{R}$-anisotropic torus T is of the form $(\mathrm{R}^{(1)}_{\mathbb{C}/\mathbb{R}}(\mathrm{GL}_1))^{\dim T}$, there is nothing to prove if both groups are anisotropic. If both groups are split, our claim follows from Remark 2.6. Clearly, G_1 and G_2 cannot have the same maximal tori if one of the groups is anisotropic and the other is isotropic. So, it remains to consider the case, where both groups are isotropic but not split. Then G_1 contains the torus with $\alpha = 1,\ \beta = \ell - 1$, and $\gamma = 0$ by Lemma 3.2, but G_2 does not by Lemma 3.3. □

Remark 3.5. Our argument shows that if G_1 is isotropic and G_2 is not split, then G_1 has a maximal $\mathbb{R}$-torus that is not isomorphic to any $\mathbb{R}$-torus of G_2. Moreover, by Lemma 3.2, a maximal $\mathbb{R}$-torus T_1 of G_1 that contains a maximal $\mathbb{R}$-split torus has parameters $\alpha = s,\ \beta = \ell - s$ and $\gamma = 0$, and hence does not allow an $\mathbb{R}$-embedding into G_2. In particular, if $G_1 = \mathrm{SO}(n-1, 1)$ and G_2 is not split then every isotropic maximal $\mathbb{R}$-torus of G_1 is not isomorphic to a subtorus of G_2.

Example 3.6 (absolute rank 3). As an empirical illustration of the landscape over $\mathbb{R}$, we divide the 14 real groups of types B_3 and C_3 into equivalence classes under the relation "have isomorphic collections of maximal tori". For forms of SO_7 or Sp_6, the maximal tori are described by Lemmas 3.2 and 3.3. Also, the four anisotropic (compact) forms obviously make up one equivalence class. For the other groups one can use a computer program such as the Atlas software [Adams and du Cloux 2009] to find the maximal tori. In summary, the groups $\mathrm{SO}(1, 6)$, $\mathrm{SO}(2, 5)$, and $\mathrm{Spin}(2, 5)$ are each their own equivalence class, and we find the following nonsingleton equivalence classes:

$$\{\text{4 anisotropic forms}\}, \quad \{\mathrm{Sp}_6, \mathrm{SO}(4, 3)\}, \quad \{\mathrm{PSp}_6, \mathrm{Spin}(4, 3)\},$$
$$\text{and} \quad \{\mathrm{Sp}(1, 2), \mathrm{PSp}(1, 2), \mathrm{Spin}(1, 6)\}.$$

In particular, $\mathrm{Spin}(1, 6)$ and $\mathrm{PSp}(1, 2)$ have the same isomorphism classes of maximal tori and yet are neither both split nor both anisotropic. This situation is dual to the one considered and eliminated in Corollary 3.4 (adjoint B_ℓ and simply connected C_ℓ).

For completeness, we mention the (much easier) analogue of Corollary 3.4 for nonarchimedean local fields.

Lemma 3.7. *Let G_1 and G_2 be absolutely almost simple groups of type B_ℓ and C_ℓ, respectively, with $\ell \geqslant 3$, over K a nonarchimedean local field of characteristic not* 2. *The following are equivalent*:

(1) *The groups G_1 and G_2 have the same isogeny classes of maximal K-tori.*

(2) $\mathrm{rk}_K\, G_1 = \mathrm{rk}_K\, G_2$.

(3) *G_1 and G_2 are split.*

Proof. (1) obviously implies (2). Suppose (2) and that G_2 is not split. Then

$$[\ell/2] = \mathrm{rk}_K\, G_2 = \mathrm{rk}_K\, G_1 \geqslant \ell - 1,$$

but this is impossible because $\ell \geqslant 3$, hence (3).

To prove (3) implies (1), we may assume that G_1 is split adjoint and G_2 is split simply connected. Combining Propositions 2.3 and 2.5 with (2-1) gives that G_1 and G_2 have the same isogeny classes of maximal tori. □

4. Local-global principles for embedding étale algebras with involution

The last ingredient we need to develop before proving Theorem 1.4 in Section 6 is a result guaranteeing in our situation the validity of the local-global principle for the existence of an embedding of an étale algebra with involution into a simple algebra with involution. This issue was analyzed in [Prasad and Rapinchuk 2010]: although the local-global principle may fail (see [ibid., Example 7.5]), it can be shown to hold under rather general conditions. For our purposes we need the following case.

Let (E, σ) be an étale algebra with involution over a number field K of dimension $n = 2m$ and satisfying (2-2). Then $E = F[x]/(x^2 - d)$, where $F = E^\sigma$ is an m-dimensional étale K-algebra and $d \in F^\times$, with the involution defined by $x \mapsto -x$ as in Example 2.2. We write $F = \prod_{j=1}^{r} F_j$, where F_j is a field extension of K, and suppose that in terms of this decomposition $d = (d_1, \dots, d_r)$. Let τ be an orthogonal involution on $A = M_n(K)$.

Proposition 4.1 [Prasad and Rapinchuk 2010, Theorem 7.3]. *Assume that for every $v \in V^K$ there exists a K_v-embedding*

$$\iota_v\colon (E \otimes_K K_v, \sigma \otimes \mathrm{id}_{K_v}) \hookrightarrow (A \otimes_K K_v, \tau \otimes \mathrm{id}_{K_v}).$$

If it holds that

$$\begin{aligned}&\textit{for every finite subset } V \subset V^K, \textit{ there exists } v_0 \in V^K \setminus V \textit{ such that}\\ &\textit{for } j = 1, \dots, r, \textit{ if } d_j \notin F_j^{\times 2}, \textit{ then } d_j \notin (F_j \otimes_K K_{v_0})^{\times 2},\end{aligned} \tag{$\diamond$}$$

then there exists an embedding $\iota\colon (E, \sigma) \hookrightarrow (A, \tau)$. Furthermore, ($\diamond$) automatically holds if F is a field.

We will now derive from the proposition the following statement, in which n can be odd or even.

Lemma 4.2. *Let K be a number field, let (E, σ) be an n-dimensional étale algebra with involution satisfying (2-2), and let τ be an orthogonal involution on $A = M_n(K)$. Assume that for every $v \in V^K$ there is an embedding*

$$\iota_v \colon (E \otimes_K K_v, \sigma \otimes \mathrm{id}_{K_v}) \hookrightarrow (A \otimes_K K_v, \tau \otimes \mathrm{id}_{K_v}).$$

Then in each of the situations

(1) *$n \leqslant 5$ or*

(2) *there is a real $v \in V^K$ such that $(E \otimes_K K_v, \sigma \otimes \mathrm{id}_{K_v})$ is isomorphic to $(\mathbb{C}, {}^{-})^m$ or $(\mathbb{C}, {}^{-})^m \times (\mathbb{R}, \mathrm{id}_{\mathbb{R}})$ depending on whether $n = 2m$ or $n = 2m + 1$,*

there exists an embedding $\iota \colon (E, \sigma) \hookrightarrow (A, \tau)$.

Proof. First, we will reduce the argument to the case of even n, that is, when E satisfies one of the following conditions:

(1′) $n = 2$ or 4, or

(2′) there is a real $v \in V^K$ such that $(E \otimes_K K_v, \sigma \otimes \mathrm{id}_{K_v})$ is isomorphic to $(\mathbb{C}, {}^{-})^m$.

Indeed, let $n = 2m + 1$ and suppose E satisfies condition (1) or (2) of the lemma. Then by [Prasad and Rapinchuk 2010, Proposition 7.2], $(E, \sigma) = (E', \sigma') \times (K, \mathrm{id}_K)$ and there exists an orthogonal involution τ' on $A' = M_{n-1}(K)$ such that for every $v \in V^K$ there is an embedding

$$\iota'_v \colon (E' \otimes_K K_v, \sigma' \otimes \mathrm{id}_{K_v}) \hookrightarrow (A' \otimes_K K_v, \tau' \otimes \mathrm{id}_{K_v}),$$

and the existence of an embedding $\iota' \colon (E', \sigma') \hookrightarrow (A', \tau')$ is equivalent to the existence of an embedding $\iota \colon (E, \sigma) \hookrightarrow (A, \tau)$. Clearly, E' satisfies the respective condition (1′) or (2′). So, if we assume that the lemma has already been established for E', then the existence of ι follows.

Now, suppose that $\dim_K E = 2m$ and E satisfies (2-2). Write $E = F[x]/(x^2 - d)$, where $F = E^\sigma = \prod_{j=1}^r F_j$ and $d = (d_1, \ldots, d_r)$ with $d_j \in F_j^\times$. Assume that there exist K-embeddings $\varphi_j \colon F_j \hookrightarrow \overline{K}$ such that if

$$M = \varphi_1(F_1) \cdots \varphi_r(F_r) \quad \text{and} \quad N = M\big(\sqrt{\varphi_1(d_1)}, \ldots, \sqrt{\varphi_r(d_r)}\big),$$

then there is $\lambda \in \mathrm{Gal}(N/M)$ with the property

$$\lambda\big(\sqrt{\varphi_j(d_j)}\big) = -\sqrt{\varphi_j(d_j)} \quad \text{whenever } d_j \notin F_j^\times \text{ for } j = 1, \ldots, r. \tag{4-1}$$

Let P be the normal closure of N over K, and let $\mu \in k\,\mathrm{Gal}(P/K)$ be such that $\mu|_N = \lambda$. By Chebotarev's density theorem [Cassels and Fröhlich 2010, Chapter 7, 2.4], for any finite $V \subset V^K$, there exists a nonarchimedean $v_0 \in V^K \setminus V$ that is unramified in P and for which the Frobenius automorphism $\mathrm{Fr}(w_0|v_0)$ is μ for a suitable extension $w_0|v_0$. Then it follows from (4-1) that $d_j \notin (F_{j\,w_0})^{\times 2}$ for any j such that $d_j \notin F_j^{\times 2}$, and therefore condition $(\diamond)$ holds.

Let now (E, σ) be an étale algebra with involution satisfying (1′) or (2′) for which embeddings ι_v exist for all $v \in V^K$. In order to derive the existence of ι from Proposition 4.1, we need to check $(\diamond)$, for which it is enough to find an automorphism λ as in the previous paragraph. Suppose that (1′) holds. Then $F = E^\sigma$ has dimension 1 or 2, respectively. Since we don't need to consider the case where F is a field (see Proposition 4.1), the only remaining case is where $F = K \times K$. Clearly, $K(\sqrt{d_1}, \sqrt{d_2})$ always has an automorphism λ such that $\lambda(\sqrt{d_j}) = -\sqrt{d_j}$ if $d_j \notin K^{\times 2}$, as required. Finally, suppose that (2′) holds. Then $F \otimes_K K_v \simeq \mathbb{R}^m$, and $d = (\delta_1, \ldots, \delta_m)$ in $\mathbb{R}^m$ with $\delta_i < 0$ for all i. Then for any embeddings $\varphi_j \colon F_j \hookrightarrow \mathbb{C}$ we have $\varphi_j(F_j) \subset \mathbb{R}$ and the restriction λ of complex conjugation satisfies $\lambda(\sqrt{d_j}) = -\sqrt{d_j}$ for all j, concluding the argument. □

Remark. Example 7.5 in [Prasad and Rapinchuk 2010] shows that there exists (E, σ) with E of dimension 6 for which the local-global principle for embeddings fails, so in terms of dimension the condition (1) in Lemma 4.2 is sharp.

For convenience of further reference, we will also quote the local-global principle for embeddings in the case of symplectic involutions.

Lemma 4.3 [Prasad and Rapinchuk 2010, Theorem 5.1]. *Let A be a central simple K-algebra of dimension n^2 with a symplectic involution τ (then, of course, n is necessarily even), and let (E, σ) be an n-dimensional étale K-algebra with involution satisfying (2-2). If for every $v \in V^K$ there exists an embedding*

$$\iota_v \colon (E \otimes_K K_v, \sigma \otimes \mathrm{id}_{K_v}) \hookrightarrow (A \otimes_K K_v, \tau \otimes \mathrm{id}_{K_v}),$$

then there exists an embedding $(E, \sigma) \hookrightarrow (A, \tau)$.

5. Function field analogue of Theorem 1.4

We recall the following immediate consequence of the rationality of the variety of maximal tori (see [Harder 1968; Platonov and Rapinchuk 1994, Corollary 7.3]), which will be used repeatedly: *Let G be a reductive algebraic group over a number field K; then given any $v \in V^K$ and any maximal K_v-torus $T^{(v)}$ of G there exists a maximal K-torus T of G that is conjugate to $T^{(v)}$ by an element of $G(K_v)$.* In particular, for any $v \in V^K$ there exists a maximal K-torus T of G such that $\mathrm{rk}_{K_v} T = \mathrm{rk}_{K_v} G$. It follows that if G_1 and G_2 are reductive K-groups having the same isogeny classes of maximal K-tori, then

$$\mathrm{rk}_{K_v} G_1 = \mathrm{rk}_{K_v} G_2 \quad \text{for all } v \in V^K. \tag{5-1}$$

The remark made in the previous paragraph remains valid for global function fields, which can be used to give the following analogue of Theorem 1.4: Suppose G_1 and G_2 are absolutely almost simple algebraic groups of types B_ℓ and C_ℓ $(\ell \geqslant 3)$

over a global field K of characteristic greater than 2. *The groups G_1 and G_2 have the same isogeny classes of maximal K-tori if and only if they are split.* Indeed, if the two groups have the same isogeny classes of maximal K-tori, then both groups are K_v-split for every v (by (5-1) and Lemma 3.7); hence both groups are K-split (by the Hasse principle). The converse holds by Remark 2.6.

6. Proof of Theorem 1.4

Throughout this section G_1 and G_2 will denote absolutely almost simple algebraic groups of types B_ℓ and C_ℓ for some $\ell \geqslant 3$ defined over a number field K. In Definition 1.1 we defined what it means for G_1 and G_2 to be *twins*. We now observe that since G_1 and G_2 cannot be K_v-anisotropic for $v \in V_f^K$, they are twins if and only if both of the following conditions hold:

$$\operatorname{rk}_{K_v} G_1 = \operatorname{rk}_{K_v} G_2 = \ell \quad \text{for all } v \in V_f^K, \tag{6-1}$$

$$\operatorname{rk}_{K_v} G_1 = \operatorname{rk}_{K_v} G_2 = 0 \text{ or } \ell \quad \text{for all } v \in V_\infty^K. \tag{6-2}$$

We also note that if G_1 and G_2 are twins over K then they remain twins over any finite extension L/K. If K has r real places, then (by the Hasse principle) there are exactly $4 \cdot 2^r$ pairs of K-groups G_1, G_2 that are twins, equivalently, 2^r pairs if one only counts the groups G_1 and G_2 up to isogeny.

Now, let G_1 and G_2 be as above, with G_1 adjoint and G_2 simply connected. Then $G_i = \mathrm{SU}(A_i, \tau_i)$ for $i = 1, 2$, where $A_1 = M_{n_1}(K)$, $n_1 = 2\ell + 1$ and the involution τ_1 is orthogonal, and A_2 is a central simple K-algebra of dimension n_2^2 with $n_2 = 2\ell$ and the involution τ_2 is symplectic. Any maximal K-torus T_i of G_i is of the form $\mathrm{SU}(E_i, \sigma_i)$, where $E_i \subset A_i$ is an n_i-dimensional étale τ_i-invariant K-subalgebra and $\sigma_i = \tau_i|_{E_i}$ so that (2-2) holds. For $i = 1$, we can always write $(E_1, \sigma_1) = (E_1', \sigma_1') \times (K, \mathrm{id}_K)$. For $i = 2$, we set $(E_2^+, \sigma_2^+) = (E_2, \sigma_2) \times (K, \mathrm{id}_K)$.

Proposition 6.1. *Let (A_1, τ_1) and (A_2, τ_2) be algebras with involution as above, and assume that $G_1 = \mathrm{SU}(A_1, \tau_1)$ and $G_2 = \mathrm{SU}(A_2, \tau_2)$ are twins. If (E_1, σ_1) is isomorphic to an n_1-dimensional étale subalgebra of (A_1, τ_1) satisfying (2-2), then (E_1', σ_1') is isomorphic to a subalgebra of (A_2, τ_2). Conversely, if (E_2, σ_2) is isomorphic to an n_2-dimensional étale subalgebra of (A_2, τ_2) satisfying (2-2) then (E_2^+, σ_2^+) is isomorphic to a subalgebra of (A_1, τ_1). Thus, the correspondences*

$$(E_1, \sigma_1) \mapsto (E_1', \sigma_1') \quad \textit{and} \quad (E_2, \sigma_2) \mapsto (E_2^+, \sigma_2^+)$$

implement mutually inverse bijections between the sets of isomorphism classes of n_1- and n_2-dimensional étale subalgebras of (A_1, τ_1) and (A_2, τ_2) that are invariant under the respective involutions and satisfy (2-2).

Proof. If we have $\operatorname{rk}_{K_v} G_1 = \operatorname{rk}_{K_v} G_2 = \ell$ for all $v \in V_\infty^K$ then the groups G_1 and G_2 are K-split by (6-1) and the Hasse principle. Then τ_1 corresponds to a nondegenerate symmetric bilinear form of Witt index ℓ, and $A_2 = M_{n_2}(K)$ with τ_2 corresponding to a nondegenerate skew-symmetric form. In this case, our claim immediately follows from Propositions 2.3 and 2.5, as in Remark 2.6. So, we may assume that there is a real $v_0 \in V_\infty^K$ such that $\mathrm{rk}_{K_{v_0}} G_1 = \mathrm{rk}_{K_{v_0}} G_2 = 0$. Observe that given *any* real $v \in V_\infty^K$ satisfying $\mathrm{rk}_{K_v} G_1 = \mathrm{rk}_{K_v} G_2 = 0$, the data in Table 1 shows that for any n_1-dimensional τ_1-invariant étale subalgebra $E_1 \subset A_1$ satisfying (2-2) and $\sigma_1 = \tau_1|_{E_1}$, we have

$$(E_1 \otimes_K K_v, \sigma_1 \otimes \mathrm{id}_{K_v}) \simeq (\mathbb{C}, \bar{\ })^\ell \times (\mathbb{R}, \mathrm{id}_{\mathbb{R}}), \tag{6-3}$$

and for any n_2-dimensional τ_2-invariant étale subalgebra $E_2 \subset A_2$ satisfying (2-2) and $\sigma_2 = \tau_2|_{E_2}$ we have

$$(E_2 \otimes_K K_v, \sigma_2 \otimes \mathrm{id}_{K_v}) \simeq (\mathbb{C}, \bar{\ })^\ell. \tag{6-4}$$

Let (E_1, σ_1) be as in the statement of the proposition. We first show that for any $v \in V^K$ there is an embedding $\iota_v \colon (E_1' \otimes_K K_v, \sigma_1' \otimes \mathrm{id}_{K_v}) \hookrightarrow (A_2 \otimes_K K_v, \tau_2 \otimes \mathrm{id}_{K_v})$. If $\mathrm{rk}_{K_v} G_1 = \mathrm{rk}_{K_v} G_2 = \ell$, this follows from Proposition 2.3. Otherwise, v is real, and $\mathrm{rk}_{K_v} G_1 = \mathrm{rk}_{K_v} G_2 = 0$, so we see from (6-3) that $(E_1' \otimes_K K_v, \sigma_1' \otimes \mathrm{id}_{K_v}) \simeq (\mathbb{C}, \bar{\ })^\ell$. Then the existence of ι_v follows from the argument given in the proof of Lemma 3.3. Now, applying Lemma 4.3 we obtain the existence of an embedding

$$\iota \colon (E_1', \sigma_1') \hookrightarrow (A_2, \tau_2),$$

as required.

Conversely, let (E_2, σ_2) be as in the proposition. Then arguing as above (using Proposition 2.5 and the proof of Lemma 3.2) we obtain the existence of local embeddings $\iota_v \colon (E_2^+ \otimes_K K_v, \sigma_2^+ \otimes \mathrm{id}_{K_v}) \hookrightarrow (A_1 \otimes_K K_v, \tau_1 \otimes \mathrm{id}_{K_v})$ for all $v \in V^K$. It follows from (6-4) that

$$(E_2^+ \otimes_K K_{v_0}, \sigma_2^+ \otimes \mathrm{id}_{K_{v_0}}) \simeq (\mathbb{C}, \bar{\ })^\ell \times (\mathbb{R}, \mathrm{id}_{\mathbb{R}}).$$

This enables us to use Lemma 4.2 which yields the existence of an embedding $(E_2^+, \sigma_2^+) \hookrightarrow (A_1, \tau_1)$, completing the argument. □

The following consequence of the proposition proves the "if" component in both parts, (1) and (2), of Theorem 1.4.

Corollary 6.2. *Let G_1 and G_2 be absolutely almost simple algebraic groups of types B_ℓ and C_ℓ, respectively, that are twins.*

(i) *G_1 and G_2 have the same isogeny classes of maximal K-tori.*

(ii) *If G_1 is adjoint and G_2 is simply connected then G_1 and G_2 have the same isomorphism classes of maximal K-tori.*

Proof. Statement (ii) easily follows from the proposition, and (i) is an immediate consequence of (ii). □

Remark 6.3. The assumption that $\ell \geqslant 3$ was never used in Proposition 6.1 and Corollary 6.2. So, these statements remain valid also for $\ell = 2$, which will be helpful in Section 8.

We now turn to the proof of the "only if" direction in both parts of Theorem 1.4, where the assumption $\ell \geqslant 3$ becomes essential and will be kept throughout the rest of the section. This direction requires a bit more work and involves the notion of *generic tori*. To recall the relevant definitions, we let G denote a semisimple algebraic K-group, and fix a maximal K-torus T of G. Furthermore, we let $\Phi(G, T)$ denote the corresponding root system, and let K_T denote the minimal splitting field of T over K. The natural action of $\mathrm{Gal}(K_T/K)$ on the group of characters $X(T)$ gives rise to an injective group homomorphism

$$\theta_T\colon \mathrm{Gal}(K_T/K) \to \mathrm{Aut}(\Phi(G, T)).$$

We say that T is *generic* (over K) if $\theta_T(\mathrm{Gal}(K_T/K))$ contains the Weyl group $W(G, T)$. As the following statement shows, generic tori with prescribed local properties always exist.

Proposition 6.4 [Prasad and Rapinchuk 2009, Corollary 3.2]. *Let G be an absolutely almost simple algebraic K-group, and let $V \subset V^K$ be a finite subset. Suppose that for each $v \in V$ we are given a maximal K_v-torus $T^{(v)}$ of G. Then there exists a maximal K-torus T of G which is generic over K and which is conjugate to $T^{(v)}$ by an element of $G(K_v)$ for all $v \in V$.*

We now return to the situation where G_1 and G_2 are absolutely almost simple K-groups of types B_ℓ and C_ℓ ($\ell \geqslant 3$), respectively. We let $G_1^\natural$ denote the adjoint group of G_1, and $G_2^\natural$ the simply connected cover of G_2. Furthermore, given a maximal K-torus T_i of G_i, we let $T_i^\natural$ denote the image of T_i in $G_i^\natural$ if $i = 1$ and the preimage of T_i in $G_i^\natural$ if $i = 2$.

Proposition 6.5. *Let T_i be a generic maximal K-torus of G_i, where $i = 1, 2$. If there exists a K-isogeny $\pi : T_i \to T_{3-i}$ onto a maximal K-torus of G_{3-i}, then there exists a K-isomorphism $T_i^\natural \simeq T_{3-i}^\natural$.*

The proof below is an adaptation of [Prasad and Rapinchuk 2009, Lemma 4.3 and Remark 4.4].

Proof. We have $K_{T_1} = K_{T_2} =: L$, and let $\mathscr{G} = \mathrm{Gal}(L/K)$. Then θ_{T_j} is an isomorphism of $\mathscr{G}$ on $W_j = W(G_j, T_j)$ for $j = 1, 2$. The isogeny π induces a $\mathscr{G}$-equivariant homomorphism of character groups $\pi^*\colon X(T_{3-i}) \to X(T_i)$. Let $X_j^\natural = X(T_j^\natural)$; we need to prove that there is a $\mathscr{G}$-equivariant *isomorphism* $\psi\colon X_{3-i}^\natural \to X_i^\natural$. (We recall that $X_1^\natural$ is the subgroup of $X(T_1)$ generated by all the roots in $\Phi_1 = \Phi(G_1, T_1)$, and $X_2^\natural$ is generated by the weights of the root system $\Phi_2 = \Phi(G_2, T_2)$.)

To avoid cumbersome notation, we will assume that $i = 1$. (This does not restrict generality as along with π there is always a K-isogeny $\pi' \colon T_{3-i} \to T_i$.) Consider

$$\phi = \pi^* \otimes \mathrm{id}_{\mathbb{R}} \colon V_2 = X(T_2) \otimes_{\mathbb{Z}} \mathbb{R} \to X(T_1) \otimes_{\mathbb{Z}} \mathbb{R} = V_2$$

and $\mu \colon W_2 \to W_1$ defined by $\mu = \theta_{T_1} \circ \theta_{T_2}^{-1}$. Then the fact that π^* is $\mathscr{G}$-equivariant implies that

$$\phi(w \cdot v) = \mu(w) \cdot \phi(v) \quad \text{for all } v \in V_2,\ w \in W_2. \tag{6-5}$$

On the other hand, it follows from the explicit description of the root systems as in [Bourbaki 2002] that there exists a linear isomorphism $\phi_0 \colon V_2 \to V_1$ and a group isomorphism $\mu_0 \colon W_2 \to W_1$ such that

$$\phi_0(w \cdot v) = \mu_0(w) \cdot \phi_0(v) \quad \text{for all } v \in V_2, w \in W_2, \tag{6-6}$$

ϕ_0 takes the short roots of Φ_2 to the long roots of Φ_1, and $(1/2)\phi_0$ takes the long roots of Φ_2 to the short roots of Φ_1, consequently $\phi_0(X_2^\natural) = X_1^\natural$. (We identify W_j with the Weyl group of the root system Φ_j.)

We claim that there exists a nonzero $\lambda \in \mathbb{R}$ and $z \in W_1$ such that

$$\phi(v) = \lambda \cdot z \cdot \phi_0(v) \quad \text{and} \quad \mu(w) = z \cdot \mu_0(w) \cdot z^{-1} \quad \text{for all } v \in V_2, w \in W_2.$$

Indeed, it was shown in [Prasad and Rapinchuk 2009, Lemma 4.3] (using that $\ell \geqslant 3$) that a suitable multiple $\phi' = \lambda^{-1} \cdot \phi$ takes the short roots of Φ_2 to the long roots of Φ_2, and $(1/2)\phi_0$ takes the long roots of Φ_2 to the short roots of Φ_1. Then $z := \phi' \circ \phi_0^{-1}$ is an automorphism of Φ_1 and hence can be identified with an element of W_1. This gives the formula for ϕ, and then the formula for μ follows from (6-5) and (6-6).

Put $\psi := \lambda^{-1} \cdot \phi$. Then $\psi(X_2^\natural) = z(\phi_0(X_2^\natural)) = X_1^\natural$, and ψ is $\mathscr{G}$-equivariant, as required. □

Corollary 6.6. *Let T_i be a generic maximal K-torus of G_i. If there exists $v \in V$ such that $T_i^\natural$ does not allow a K_v-defined embedding into $G_{3-i}^\natural$, then T_i is not K-isogenous to any maximal K-torus T_{3-i} of G_{3-i}. Thus, if G_1 and G_2 have the same isogeny classes of maximal K-tori, then $G_1^\natural$ and $G_2^\natural$ have the same isomorphism classes of maximal K_v-tori for all $v \in V$.*

Proof. The first assertion immediately follows from the proposition. To derive the second assertion from the first, we observe that given $v \in V$ and a maximal K_v-torus $\mathscr{T}_i$ of $G_i^\natural$ that does not allow a K_v-embedding into $G_{3-i}^\natural$, we can find a maximal K-torus T_i of G_i such that $T_i^\natural$ is conjugate to $\mathscr{T}_i$ by an element $G_i^\natural(K_v)$. □

Proof of Theorem 1.4, "only if". Assume G_1 and G_2 have the same isogeny classes of maximal K-tori. Then by Corollary 6.6, $G_1^\natural$ and $G_2^\natural$ have the same isomorphism classes of maximal K_v-tori for all v. It follows that G_1 and G_2 are twins (by Corollary 3.4 for v real and Lemma 3.7 for v finite), completing the proof of part (1) of Theorem 1.4.

Now suppose that G_1 and G_2 have the same *isomorphism* classes of maximal K-tori, in particular, there is a K-isomorphism $\pi : T_1 \to T_2$ between two generic K-tori. Then as in the proof of Proposition 6.5, π^* induces $\phi \colon V_2 \to V_1$, which necessarily satisfies $\phi(X(T_2)) = X(T_1)$ and $\phi(X(T_2^\natural)) = X(T_1^\natural)$. Since $X(T_1^\natural) \subseteq X(T_1)$ and $X(T_2^\natural) \supseteq X(T_2)$, this is possible only if both inclusions are in fact equalities, that is, $G_1 = G_1^\natural$ and $G_2 = G_2^\natural$. This completes the proof of part (2) of Theorem 1.4. □

7. Weakly commensurable subgroups and proof of Theorem 1.2

We begin by recalling the notion of weak commensurability of Zariski-dense subgroups introduced in [Prasad and Rapinchuk 2009]. Let G_1 and G_2 be semisimple algebraic groups over a field F of characteristic zero, and let $\Gamma_i \subset G_i(F)$ be a Zariski-dense subgroup for $i = 1, 2$. Semisimple elements $\gamma_i \in \Gamma_i$ are *weakly commensurable* if there exist maximal F-tori T_i of G_i such that $\gamma_i \in T_i(F)$ and for some characters $\chi_i \in X(T_i)$ we have $\chi_1(\gamma_1) = \chi_2(\gamma_2) \neq 1$. Furthermore, the subgroups Γ_1 and Γ_2 are *weakly commensurable* if every semisimple element $\gamma_1 \in \Gamma_1$ of infinite order is weakly commensurable to some $\gamma_2 \in \Gamma_2$ of infinite order, and vice versa.

The focus in [ibid.] was on analyzing when two Zariski-dense S-arithmetic subgroups in absolutely almost simple algebraic groups are weakly commensurable. This analysis was based on a description of such S-arithmetic groups in terms of triples, which we will now briefly recall. Let G be a (connected) absolutely almost simple algebraic group defined over a field F of characteristic zero, $\overline{G}$ be its adjoint group, and $\pi : G \to \overline{G}$ be the natural isogeny. Suppose we are given the following data:

- a number field K with a fixed embedding $K \hookrightarrow F$,
- a finite set S of valuations of K containing all archimedean valuations, and
- an F/K-form $\mathscr{G}$ of $\overline{G}$ (that is, a K-defined algebraic group such that there exists an F-defined isomorphism of algebraic groups ${}_F\mathscr{G} \simeq \overline{G}$, where ${}_F\mathscr{G}$ is the group obtained from $\mathscr{G}$ by the extension of scalars F/K).

(It is assumed in addition that S does not contain any nonarchimedean valuations v such that $\mathscr{G}$ is K_v-anisotropic.) We then have an embedding $\iota \colon \mathscr{G}(K) \hookrightarrow \overline{G}(F)$ and a natural S-arithmetic subgroup $\mathscr{G}(\mathbb{O}_K(S))$, where $\mathbb{O}_K(S)$ is the ring of S-integers in K, defined in terms of a fixed K-embedding $\mathscr{G} \hookrightarrow \mathrm{GL}_n$, that is, $\mathscr{G}(\mathbb{O}_K(S)) =$

$\mathscr{G}(K) \cap \mathrm{GL}_n(\mathbb{O}_K(S))$. A subgroup Γ of $G(F)$ such that $\pi(\Gamma)$ is commensurable with $\iota(\mathscr{G}(\mathbb{O}_K(S)))$ is called $(\mathscr{G}, K, S)$-*arithmetic*. (It should be pointed out that we do not fix an F-defined isomorphism ${}_F\mathscr{G} \simeq \overline{G}$ in this definition, and by varying it we obtain a class of subgroups invariant under F-defined automorphisms of G in the obvious sense.)

It was shown in [ibid.] that if G_i is absolutely almost simple and Γ_i is Zariski-dense and $(\mathscr{G}_i, K_i, S_i)$-arithmetic for $i = 1, 2$, then the weak commensurability of Γ_1 and Γ_2 implies that $K_1 = K_2 =: K$ and $S_1 = S_2 =: S$, and additionally either G_1 and G_2 are of the same type or one of them is of type B_ℓ and the other is of type C_ℓ for some $\ell \geqslant 3$. That paper also contains many precise conditions for two S-arithmetic subgroups to be weakly commensurable in the case where G_1 and G_2 are of the *same* type. The goal of this section is to prove Theorem 1.2, which provides such conditions when one of the groups is of type B_ℓ and the other of type C_ℓ ($\ell \geqslant 3$). In conjunction with the previous results, this completes the investigation of weak commensurability of S-arithmetic subgroups in absolutely almost simple groups over number fields.

Proof of Theorem 1.2. Let G_1 and G_2 be absolutely almost simple algebraic groups of types B_ℓ and C_ℓ ($\ell \geqslant 3$), respectively, defined over a number field K, and let Γ_i be a Zariski-dense $(\mathscr{G}_i, K, S)$-arithmetic subgroup of G_i.

Suppose that $\mathscr{G}_1$ and $\mathscr{G}_2$ are twins. Then by Theorem 1.4, they have the same isogeny classes of maximal K-tori. This *automatically* implies that Γ_1 and Γ_2 are weakly commensurable. To see this, we basically need to repeat the argument given in [Prasad and Rapinchuk 2009, Example 6.5], which we also give here for the reader's convenience. First, we may assume without any loss of generality that G_1 and G_2 are adjoint (see [ibid., Lemma 2.4]); hence $\Gamma_i \subset \mathscr{G}_i(K)$. Let $\gamma_1 \in \Gamma_1$ be a semisimple element of infinite order, and let T_1 be a maximal K-torus of $\mathscr{G}_1$ that contains γ_1. Then there exists a K-isogeny $\varphi \colon T_1 \to T_2$ onto a maximal K-torus T_2 of $\mathscr{G}_2$. The subgroup $\varphi(T_1(K) \cap \Gamma_1)$ is an S-arithmetic subgroup of $T_2(K)$, so there exists $n > 0$ such that $\gamma_2 := \varphi(\gamma_1)^n \in \Gamma_2$. Let $\chi_1 \in \varphi^*(X(T_2))$ be a character such that $\chi_1(\gamma_1)$ is not a root of unity, and let $\chi_2 \in X(T_2)$ be such that $\varphi^*(\chi_2) = \chi_1$. Then

$$(n\chi_1)(\gamma_1) = \chi_1(\gamma_1)^n = \chi_2(\gamma_2) \neq 1,$$

which implies that Γ_1 and Γ_2 are weakly commensurable.

Conversely, suppose that Γ_1 and Γ_2 are weakly commensurable. According to [ibid., Theorem 6.2], this in particular implies that

$$\mathrm{rk}_{K_v} \mathscr{G}_1 = \mathrm{rk}_{K_v} \mathscr{G}_2 \quad \text{for all } v \in V^K.$$

As we have seen in Lemma 3.7, for $v \in V_f^K$ and the groups under consideration, the equality of ranks implies that both groups are actually K_v-split, verifying condition (6-1). Assume that condition (6-2) fails for a real $v_0 \in V_\infty^K$. Then by Corollary 3.4, there is an $i \in \{1, 2\}$ and a maximal K_{v_0}-torus $\mathcal{T}_i$ of $\mathcal{G}_i^\natural$ that does not allow a K_{v_0}-embedding into $\mathcal{G}_{3-i}^\natural$; obviously $\mathcal{T}_i$ is K_{v_0}-isotropic. Let $T_i^{(v_0)}$ be a maximal K_{v_0}-torus of $\mathcal{G}_i$ such that $(T_i^\natural)^{(v_0)} = \mathcal{T}_i$. Furthermore, for $v \in S \setminus \{v_0\}$ we let $T_i^{(v)}$ denote a maximal K_v-torus of $\mathcal{G}_i$ such that $\operatorname{rk}_{K_v} T_i^{(v)} = \operatorname{rk}_{K_v} \mathcal{G}_i$. Using Proposition 6.4, we can find a maximal K-torus T_i of $\mathcal{G}_i$ that is generic and that is conjugate to $T^{(v)}$ by an element of $\mathcal{G}_i(K_v)$ for all $v \in S \cup \{v_0\}$. Then clearly $\operatorname{rk}_S T_i := \sum_{v \in S} \operatorname{rk}_{K_v} T_i > 0$ as $\operatorname{rk}_S \mathcal{G}_i > 0$. By Dirichlet's theorem [Platonov and Rapinchuk 1994, Theorem 5.12], the group of S-integral points $T_i(\mathbb{O}_K(S))$ has the structure $H \times \mathbb{Z}^d$, where $d = \operatorname{rk}_S T_i - \operatorname{rk}_K T_i$. Since T_i is obviously K-anisotropic, we conclude that there exists $\gamma_i \in T_i(K) \cap \Gamma_i$ of infinite order (as in the previous paragraph, we are assuming that G_1 and G_2 are adjoint, and hence $\Gamma_j \subset \mathcal{G}_j(K)$ for $j = 1, 2$). Then γ_i is weakly commensurable to some semisimple $\gamma_{3-i} \in \Gamma_{3-i}$ of infinite order. Let T_{3-i} be a maximal K-torus of $\mathcal{G}_{3-i}$ containing γ_{3-i}. By the isogeny theorem [Prasad and Rapinchuk 2009, Theorem 4.2], the tori T_i and T_{3-i} are K-isogenous. Using Proposition 6.5, we conclude that $T_i^\natural$ and $T_{3-i}^\natural$ are K-isomorphic. This implies that over K_{v_0}, the torus $\mathcal{T}_i \simeq T_i^\natural$ has an embedding into $\mathcal{G}_{3-i}$, a contradiction, proving (6-2) and completing the proof of Theorem 1.2. □

As we already mentioned, the notion of weak commensurability was introduced to tackle some differential-geometric problems dealing with length-commensurable and isospectral locally symmetric spaces, and we conclude this section with a sample of geometric consequences — established in [Prasad and Rapinchuk 2013] — of the results of the current paper. For a Riemannian manifold M, we let $L(M)$ denote the weak length spectrum of M, that is, the collection of lengths of all closed geodesics in M. Two Riemannian manifolds M_1 and M_2 are called *length-commensurable* if $\mathbb{Q} \cdot L(M_1) = \mathbb{Q} \cdot L(M_2)$.

Let M_1 be an arithmetic quotient of the real hyperbolic space $\mathbb{H}^p$ ($p \geqslant 5$), and M_2 be an arithmetic quotient of the quaternionic hyperbolic space $\mathbb{H}^q_{\mathbf{H}}$ ($q \geqslant 2$). Then M_1 and M_2 are not length-commensurable. (7-1)

Theorem 1.2 is used to handle the case $p = 2n$ and $q = n - 1$ for $n \geqslant 3$; for other p and q, the claim follows from [Prasad and Rapinchuk 2009, Theorem 8.15].

Now, let $\mathfrak{X}_1$ be the symmetric space of the real Lie group $\mathcal{G}_1 = \mathrm{SO}(n+1, n)$, and let $\mathfrak{X}_2$ be the symmetric space of the real Lie group $\mathcal{G}_2 = \mathrm{Sp}_{2n}$, where $n \geqslant 3$.

Let M_i be the quotient of $\mathfrak{X}_i$ by a $(\mathcal{G}_i, K)$-arithmetic subgroup of $\mathcal{G}_i$ for $i = 1, 2$. If $\mathcal{G}_1$ and $\mathcal{G}_2$ are twins, then (7-2)

$$\mathbb{Q} \cdot L(M_2) = \lambda \cdot \mathbb{Q} \cdot L(M_1), \quad \textit{where } \lambda = \sqrt{\frac{2n+2}{2n-1}}.$$

(We refer to [Prasad and Rapinchuk 2009, §1] for the notion of arithmeticity and the explanation of other terms used here.) We finally note that even though one can make M_1 and M_2 length-commensurable by scaling the metric on one of them, this will never make them isospectral [Yeung 2011].

8. Proofs of Proposition 1.3 and Theorem 1.5

Proof of Proposition 1.3. We can assume that G_1 and G_2 are connected absolutely almost simple *adjoint* K-groups having the same isogeny classes of maximal K-tori. Assume that provisions (2) and (3) of the proposition do not hold; let us show that (1) must hold. First, by [Prasad and Rapinchuk 2009, Theorem 7.5], G_1 and G_2 have the same Killing–Cartan type. Furthermore, if L_i is the minimal Galois extension of K over which G_i becomes an inner form then $L_1 = L_2$; in other words, G_1 and G_2 are inner twists of the *same* quasisplit K-group. So, the required assertion is a consequence of the following lemma. □

Lemma 8.1. *Let G_1 and G_2 be connected absolutely almost simple* adjoint *K-groups of the same Killing–Cartan type, which is different from A_ℓ ($\ell > 1$), $\mathsf{D}_{2\ell+1}$ ($\ell > 1$) or E_6. Assume that G_1 and G_2 are inner twists of the same quasisplit K-group (which holds automatically if G_1 and G_2 are not of type D). If G_1 and G_2 have the same isogeny classes of maximal K-tori then $G_1 \simeq G_2$.*

Proof. First, suppose that the groups are not of type D. As we have seen in Section 5, the fact that G_1 and G_2 have the same isogeny classes of maximal K-tori implies that $\mathrm{rk}_{K_v} G_1 = \mathrm{rk}_{K_v} G_2$ for all $v \in V^K$. For groups of one of the types under consideration, this implies that $G_1 \simeq G_2$ over K_v for all $v \in V^K$ and then our assertion follows from the Hasse principle for Galois cohomology of adjoint groups; see [Prasad and Rapinchuk 2009, §6] for details of the argument.

Now, suppose the groups are of type $\mathsf{D}_{2\ell}$ for some $\ell \geqslant 2$. There exists a maximal K-torus T_1 of G_1 that is generic and such that $\mathrm{rk}_{K_v} T_1 = \mathrm{rk}_{K_v} G_1$ at every place v where at least one of G_1 or G_2 is not quasisplit. (The set of such v is finite; see [Platonov and Rapinchuk 1994, Theorem 6.7].) By hypothesis, T_1 is isogenous to a maximal K-torus T_2 of G_2, which is necessarily also generic. Following [Prasad and Rapinchuk 2009, Lemma 4.3 and Remark 4.4], one finds a K-isomorphism $T_1 \to T_2$ that extends to a $\overline{K}$-isomorphism $G_1 \to G_2$. Then our assertion follows from Theorem 20 in [Garibaldi 2012]. □

Proof of Theorem 1.5. The "if" direction is actually contained in Corollary 6.2 — see Remark 6.3. For the "only if" direction, we first observe that if G_1 and G_2 have the same isomorphism classes of maximal K-tori then by Lemma 8.1 the groups $\mathrm{SO}(q_1)$ and $\mathrm{SO}(q_2)$ are isomorphic; hence the forms q_1 and q_2 are similar, yielding assertion (1). Thus, we can assume that $G_1 = \mathrm{SO}(q)$ and $G_2 = \mathrm{Spin}(q)$ for a single quadratic form q.

To prove assertion (2), it is enough to show that if $v \in V^K$ is such that the Witt index of q over K_v is 1 then there exists a 2-dimensional K_v-torus T_1 that has a K_v-embedding into G_1 but does not allow a K_v-embedding into G_2. For this we pick a quadratic extension L/K_v and set

$$T_1 = \mathrm{GL}_1 \times \mathrm{R}^{(1)}_{L/K_v}(\mathrm{GL}_1).$$

We can write $q = q' \perp q''$, where q' is a hyperbolic plane. Then $\mathrm{SO}(q') = \mathrm{GL}_1$ and $\mathrm{SO}(q'') = \mathrm{PSL}_{1,D}$, where D is a quaternion division algebra over K_v. Since L embeds in D, the torus $\mathrm{R}^{(1)}_{L/K_v}(\mathrm{GL}_1)$ embeds in $SL_{1,D}$ and then also in $\mathrm{PSL}_{1,D}$. It follows that T_1 embeds in $G_1 = \mathrm{SO}(q)$. On the other hand, let $T_2 \subset G_2$ be a maximal K_v-torus that splits over L. We can identify G_2 with $\mathrm{SU}(A, \tau)$, where $A = M_2(D)$ with D a quaternion division algebra over K and τ is a symplectic involution on A. Let E_2 be the K_v-subalgebra of A generated by $T_2(K_v)$. Then $E_2 \otimes_{K_v} L \simeq L^4$. As in Section 3, we conclude that $(E_2, \tau|_{E_2})$ is isomorphic to $(L, \sigma) \times (L, \sigma)$, where σ is the nontrivial automorphism of L, or to $(L \times L, \lambda)$, where λ is the switch involution. Then $T_2 = \mathrm{SU}(E_2, \tau|_{E_2})$ is isomorphic, respectively, to $\mathrm{R}^{(1)}_{L/K_v}(\mathrm{GL}_1)^2$ or $\mathrm{R}_{L/K_v}(\mathrm{GL}_1)$. Neither such torus can be isomorphic to T_1. □

9. Alternative proofs via Galois cohomology

Although the main body of the paper demonstrates the effectiveness (and in fact the ubiquity) of the technique of étale algebras in dealing with maximal tori of classical groups, it is worth pointing out that some parts of the argument can also be given in the language of Galois cohomology of algebraic groups. In this section, we will illustrate such an exchange by giving a cohomological proof of the "if" direction of Theorem 1.4(2), that is, of Corollary 6.2(ii).

Our main tool is Proposition 9.1, for which we need some notation. Let G be a connected semisimple algebraic group over a number field K. Fix a maximal K-torus T of G, and let $N = N_G(T)$ and $W = N/T$ denote, respectively, its normalizer and the corresponding Weyl group. For any field extension P/K, we let $\theta_P \colon H^1(P, N) \to H^1(P, W)$ denote the map induced by the natural K-morphism $N \to W$, and let

$$\mathscr{C}(P) := \mathrm{Ker}\big(H^1(P, N) \to H^1(P, G)\big).$$

The elements of $\mathscr{C}(P)$ are in one-to-one correspondence with the $G(P)$-conjugacy classes of maximal P-tori in G; see for example [Prasad and Rapinchuk 2009, Lemma 9.1] where this correspondence is described explicitly. There is an obvious K-defined map $W \to \mathrm{Aut}\, T$, so for any $\xi \in H^1(K, W)$ one can consider the corresponding twisted K-torus ${}_{\xi}T$.

Proposition 9.1. *Assume that there exists a subset $V_0 \subset V_\infty^K$ such that G is K_v-anisotropic for all $v \in V_0$ and is K_v-split for all $v \in V^K \setminus V_0$. Then the sequence*

$$\mathscr{C}(K) \xrightarrow{\theta_K} H^1(K, W) \xrightarrow{\prod \rho_v} \prod_{v \in V_0} H^1(K_v, W) \tag{9-1}$$

is exact.

Here ρ_v denotes the natural restriction map $H^1(K, W) \to H^1(K_v, W)$.

Proof. If V_0 is empty then it follows from the Hasse principle for adjoint groups [Platonov and Rapinchuk 1994, Theorem 6.22] that G is K-split. In this case it was shown by Gille [2004] and Raghunathan [2004] (or earlier by Kottwitz [1982]) that $\theta_K(\mathscr{C}(K)) = H^1(K, W)$, and our claim follows. So, we will assume in the rest of the argument that V_0 is not empty.

We first prove that $\rho_v \theta_K = 0$ for all $v \in V_0$. Given $\xi \in \mathscr{C}(K)$, one can pick $g \in G(\overline{K})$ such that $n(\sigma) := g^{-1}\sigma(g)$ belongs to $N(\overline{K})$ for all $\sigma \in \mathrm{Gal}(\overline{K}/K)$, and the cocycle $\sigma \mapsto n(\sigma)$ represents ξ. Then the maximal torus $T' = gTg^{-1}$ is defined over K. Now, let $v \in V_0$. According to our definitions, G is anisotropic over $K_v = \mathbb{R}$, so it follows from the conjugacy of maximal tori in compact Lie groups that T and T' are conjugate by an element of $G(K_v)$. Then the one-to-one correspondence between the elements of $\mathscr{C}(K_v)$ and the $G(K_v)$-conjugacy classes of maximal K_v-tori in G (or a simple direct computation) implies that the image of ξ under the restriction map $\mathscr{C}(K) \to \mathscr{C}(K_v)$ is trivial, and hence the image of $\theta_K(\xi)$ under the restriction map $H^1(K, W) \to H^1(K_v, W)$ is trivial as well.

Now suppose that G is simply connected; we verify that every $\xi \in \bigcap_{v \in V_0} \ker \rho_v$ is in the image of θ_K. Pick $v \in V_0$. Since ξ lies in the kernel of $H^1(K, W) \to H^1(K_v, W)$, the twisted torus ${}_\xi T$ is K_v-isomorphic to T, hence K_v anisotropic (as G is K_v-anisotropic). Thus,

$$\mathrm{Ker}\bigl(H^2(K, {}_\xi T) \to \textstyle\prod_{v \in V^K} H^2(K_v, {}_\xi T)\bigr) = 0$$

by [Prasad and Rapinchuk 2009, Proposition 6.12]. Invoking [ibid., Theorem 9.2], we see that to prove the inclusion $\xi \in \theta_K(\mathscr{C}(K))$, it is enough to show that $\rho_v(\xi) \in \theta_{K_v}(\mathscr{C}(K_v))$ for all $v \in V^K$. If $v \in V_0$ then by construction $\rho_v(\xi)$ is trivial, and there is nothing to prove. Otherwise, the group G is K_v-split, so by the result of Gille, Kottwitz and Raghunathan we have $\theta_{K_v}(\mathscr{C}(K_v)) = H^1(K_v, W)$, and the inclusion $\rho_v(\xi) \in \theta_{K_v}(\mathscr{C}(K_v))$ is obvious. Since ξ was arbitrary, we have proved that $\bigcap \ker \rho_v$ is contained in the image of θ_K.

In case G is not simply connected, we fix a K-defined universal cover $\pi : \widetilde{G} \to G$ of G and use the tilde to denote the objects associated with $\widetilde{G}$. Then π yields a

K-isomorphism of $\widetilde{W}$ and W and we have a commutative diagram

$$\begin{tikzcd}
\widetilde{\mathscr{C}}(K) \arrow[r, "\widetilde{\theta}_K"] \arrow[d] & H^1(K, \widetilde{W}) \arrow[r, "\prod \widetilde{\rho}_v"] \arrow[d, equal] & \prod_{v \in V_0} H^1(K_v, \widetilde{W}) \arrow[d, equal] \\
\mathscr{C}(K) \arrow[r, "\theta_K"] & H^1(K, W) \arrow[r, "\prod \rho_v"] & \prod_{v \in V_0} H^1(K_v, W).
\end{tikzcd}$$

The top row is exact by the previous paragraph; hence $\bigcap \ker \rho_v$ is contained in the image of θ_K. □

We now begin to work our way towards the proof of Theorem 1.4(2) and Corollary 6.2(ii). Let G_1 be adjoint of type B_ℓ and let G_2 be simply connected of type C_ℓ for some $\ell \geqslant 2$. We will use a subscript $i \in \{1, 2\}$ to denote the objects associated with G_i. In particular, we let T_i denote a maximal torus of G_i, and let $N_i = N_{G_i}(T_i)$ and $W_i = N_i/T_i$ be its normalizer and the Weyl group. Then W_i naturally acts on T_i by conjugation. We say that the morphisms of algebraic groups $\varphi \colon T_1 \to T_2$ and $\psi \colon W_1 \to W_2$ are *compatible* if

$$\varphi(w \cdot t) = \psi(w) \cdot \varphi(t) \quad \text{for all } t \in T_1, w \in W_1.$$

Lemma 9.2. *One can pick maximal K-tori T_i of G_i for $i = 1, 2$ so that there exist compatible K-defined isomorphisms $\varphi \colon T_1 \to T_2$ and $\psi \colon W_1 \to W_2$.*

Proof. Imitating the argument given in [Platonov and Rapinchuk 1994, Proposition 6.16], it is easy to see that there exists a quadratic extension L/K that splits *both* G_1 and G_2. Indeed, let V_i be the (finite) set of places $v \in V^K$ such that G_i does not split over K_v, and let $V = V_1 \cup V_2$. Pick a quadratic extension L/K so that the local degree $[L_w : K_v] = 2$ for all $v \in V$ and $w|v$. We claim that L is as required. By the Hasse principle, it is enough to show that both G_1 and G_2 split over L_w for any $w \in V^L$. For a given w, we let $v \in V^K$ be the place that lies below w. If $v \notin V$ then by our construction G_1 and G_2 split already over K_v, and there is nothing to prove. If $v \in V$ then $[L_w : K_v] = 2$, and then the proof of [ibid., Proposition 6.16] gives that G_1 and G_2 split over L_w, as required.

Now, let $\sigma \in \operatorname{Gal}(L/K)$ be a generator. According to [ibid., Lemma 6.17], for each $i \in \{1, 2\}$, there exists an L-defined Borel subgroup B_i of G_i such that $T_i := B_i \cap B_i^\sigma$ is a maximal K-torus of G_i that splits over L. Considering the action of σ on the root system $\Phi(G_i, T_i)$, we see that it takes the system of positive roots corresponding to B_i into the system of negative roots. For groups of types B_ℓ and C_ℓ, this implies that σ acts on the character group $X(T_i)$ as multiplication by (-1). It easily follows from the description of the corresponding root systems (see [Bourbaki 2002]) that there exist compatible (in the obvious sense) isomorphisms $\varphi^* \colon X(T_2) \to X(T_1)$ (of abelian groups) and $\psi \colon W_1 \to W_2$ (of abstract groups considered as subgroups of $GL(X(T_1))$ and $GL(X(T_2))$). Then φ^* gives rise to an isomorphism $\varphi\colon T_1 \to T_2$ of algebraic groups that is compatible (as defined above) with ψ (which can be considered as a morphism of algebraic groups). It remains to observe that since σ acts on $X(T_1)$ and $X(T_2)$ as multiplication by (-1), both φ and ψ are K-defined (in fact, σ acts on W_1 and W_2 trivially). □

Remark. If both groups G_1 and G_2 are K-split then one can, of course, take for T_1 and T_2 their maximal K-split tori.

For the rest of the paper, we fix compatible K-defined isomorphisms

$$\varphi^0\colon T_1^0 \to T_2^0 \quad \text{and} \quad \psi^0\colon W_1^0 \to W_2^0.$$

(Thus, we henceforth slightly change the notation used in Lemma 9.2.) Given *arbitrary* maximal K-tori T_i of G_i for $i = 1, 2$, we pick elements $g_i \in G(\overline{K})$ so that $T_i = g_i T_i^0 g_i^{-1}$, and then for any $\sigma \in \mathrm{Gal}(\overline{K}/K)$, the element $n_i(\sigma) := g_i^{-1}\sigma(g_i)$ belongs to $N_i^0(\overline{K})$. Let $\varphi = \varphi(g_1, g_2)$ be the morphism $T_1 \to T_2$ defined by

$$\varphi(t) = g_2\varphi^0(g_1^{-1} t g_1) g_2^{-1},$$

and let $\nu_i^0\colon N_i^0 \to W_i^0$ denote the canonical morphism.

Lemma 9.3. *If*

$$\psi^0(\nu_1^0(n_1(\sigma))) = \nu_2^0(n_2(\sigma)) \quad \textit{for all } \sigma \in \mathrm{Gal}(\overline{K}/K) \tag{9-2}$$

then $\varphi = \varphi(g_1, g_2)$ is defined over K.

Proof. We need to show that φ commutes with every $\sigma \in \mathrm{Gal}(\overline{K}/K)$. Since φ^0 is defined over K, for any $t \in T_1(\overline{K})$, we have

$$\begin{aligned}\sigma(\varphi(t)) &= \sigma(g_2)\varphi^0(\sigma(g_1)^{-1}\sigma(t)\sigma(g_1))\sigma(g_2)^{-1}\\ &= g_2 n_2(\sigma)\varphi^0(n_1(\sigma)^{-1} g_1^{-1}\sigma(t) g_1 n_1(\sigma)) n_2(\sigma)^{-1} g_2^{-1}\\ &= g_2\big((\nu_2^0(n_2(\sigma)))\cdot\varphi^0((\nu_1^0(n_1(\sigma)))\cdot(g_1^{-1}\sigma(t)g_1))\big)g_2^{-1}.\end{aligned}$$

Since φ^0 is compatible with ψ^0, condition (9-2) implies that the latter reduces to

$$g_2\varphi^0(g_1^{-1}\sigma(t)g_1)g_2^{-1} = \varphi(\sigma(t)).$$

It follows that $\sigma(\varphi(t)) = \varphi(\sigma(t))$, that is, φ commutes with σ, as required. □

Pursuant to the notation above, for an extension P/K and $i = 1, 2$, we set

$$\mathscr{C}_i(P) = \mathrm{Ker}\big(H^1(P, N_i^0) \to H^1(P, G_i)\big),$$

and let $\theta_{iP}\colon H^1(P, N_i^0) \to H^1(P, W_i^0)$ denote the canonical map (induced by ν_i). The isomorphism $H^1(K, W_1^0) \to H^1(K, W_2^0)$ induced by ψ^0 will still be denoted by ψ^0.

Lemma 9.4. *Assume that*

$$\psi^0(\mathscr{C}_1(K)) = \mathscr{C}_2(K). \tag{9-3}$$

Then for $i = 1$ or 2, given any maximal K-torus T_i of G_i and an element $g_i \in G_i(\overline{K})$ such that $T_i = g_i T_i^0 g_i^{-1}$, there exists $g_{3-i} \in G_{3-i}(\overline{K})$ such that the maximal torus $T_{3-i} := g_{3-i} T_{3-i}^0 g_{3-i}^{-1}$ and the isomorphism $\varphi(g_1, g_2)\colon T_1 \to T_2$ are K-defined. Thus, in this case G_1 and G_2 have the same isomorphism classes of maximal K-tori.

Proof. To keep our notation simple, we will give an argument for $i = 1$ (the argument in the case $i = 2$ is totally symmetric). As above, we set $n_1(\sigma) = g_1^{-1}\sigma(g_1) \in N_1^0(\overline{K})$ for $\sigma \in \mathrm{Gal}(\overline{K}/K)$, observing that these elements define a cohomology class $n_1 \in \mathscr{C}_1(K)$. Then (9-3) implies that there exists $h_2 \in G_2(\overline{K})$ such that for the cohomology class $m_2 \in \mathscr{C}_2(K)$ defined by the elements $m_2(\sigma) = h_2^{-1}\sigma(h_2) \in N_2^0(\overline{K})$, we have $\psi^0(\theta_{1K}(n_1)) = \theta_{2K}(m_2)$ in $H^1(K, W_2)$. Then there exists $w_2 \in W_2(\overline{K})$ such that

$$\psi^0(\nu_1^0(n_1(\sigma))) = w_2^{-1}\nu_2^0(m_2(\sigma))\sigma(w_2) \quad \text{for all } \sigma \in \mathrm{Gal}(\overline{K}/K). \tag{9-4}$$

Picking $z_2 \in N_2^0(\overline{K})$ so that $\nu_2^0(z_2) = w_2$, and setting

$$g_2 = h_2 z_2 \quad \text{and} \quad n_2(\sigma) = g_2^{-1}\sigma(g_2) \in N_2^0(\overline{K}) \quad \text{for } \sigma \in \mathrm{Gal}(\overline{K}/K),$$

we obtain from (9-4) that (9-2) holds. Then g_2 is as required. Indeed, the fact that $n_2(\sigma) \in N_2^0(\overline{K})$ implies that $T_2 = g_2 T_2^0 g_2^{-1}$ is defined over K, and Lemma 9.3 yields that the morphism $\varphi(g_1, g_2)\colon T_1 \to T_2$ is also defined over K. □

Proof of Corollary 6.2(ii). Suppose that G_1 and G_2 are twins, and let V_0 be the set of all archimedean places $v \in V^K$ such that G_1 and G_2 are both K_v-anisotropic. Then for any $v \in V^K \setminus V_0$, both G_1 and G_2 are K_v-split. Then according to Proposition 9.1 we have

$$\theta_{iK}(\mathscr{C}_i(K)) = \ker\bigl(H^1(K, W_i^0) \to \textstyle\prod_{v \in V_0} H^1(K_v, W_i^0)\bigr)$$

for $i = 1, 2$, and as $\psi_0 : W_1^0 \to W_2^0$ is an isomorphism, condition (9-3) holds, and the claim follows from Lemma 9.4. □

Remark. It follows from the explicit description of the root systems of types B_ℓ and C_ℓ that the isomorphism φ in Lemma 9.2 can be chosen so that for $t \in T_1(\overline{K})$ there exist $\lambda_1, \dots, \lambda_\ell \in \overline{K}^\times$ such that the values of the roots $\alpha \in \Phi(G_1, T_1)$ on t are

$$\lambda_i^{\pm 1}, \quad i = 1, \dots, \ell, \quad \text{and} \quad \lambda_i^{\pm 1} \cdot \lambda_j^{\pm 1}, \quad i, j = 1, \dots, \ell, i \neq j,$$

and the values of the roots $\alpha \in \Phi(G_2, T_2)$ on $\phi(t)$ are

$$\lambda_i^{\pm 2}, \quad i = 1, \dots, \ell, \quad \text{and} \quad \lambda_i^{\pm 1} \cdot \lambda_j^{\pm 1}, \quad i, j = 1, \dots, \ell, i \neq j.$$

Then any identification of the form $\varphi(g_1, g_2)$ also has this property, which was used in [Prasad and Rapinchuk 2013].

Alternatively, suppose that G_i for $i = 1, 2$ is realized as $\mathrm{SU}(A_i, \tau_i)$ as described in the beginning of Section 6. Let E_1 be a $(\tau_1 \otimes \mathrm{id}_{\overline{K}})$-invariant maximal commutative étale $\overline{K}$-subalgebra of $A_1 \otimes_K \overline{K}$ satisfying (2-2), and let $\sigma_1 = \tau_1|_{E_1}$. Then in the notation of Section 6, the algebra (E_1', σ_1') admits a $\overline{K}$-embedding embedding into $(A_2 \otimes_K \overline{K}, \tau_2 \otimes \mathrm{id}_{\overline{K}})$, and we let (E_2, σ_2) the image of this embedding. It is easy to see that if we let T_i denote the maximal torus of G_i defined by (E_i, σ_i) then the isomorphism $T_1 \simeq T_2$ coming from the isomorphism of algebras $(E_1', \sigma_1') \simeq (E_2, \sigma_2)$ is the same as the isomorphism coming from the description of the root systems (see the proof of Lemma 9.2); in particular, it is compatible with the natural isomorphism of the Weyl groups. So, the assertion of Lemma 9.2 means that given *any* K-algebras with involutions (A_1, τ_1) and (A_2, τ_2) as above, there exists a τ_1-invariant maximal commutative étale K-subalgebra E_1 of A_1 that satisfies (2-2) and is such that for $\sigma_1 = \tau_1|_{E_1}$, the algebra (E_1', σ_1') admits an embedding into (A_2, σ_2). Moreover, by Corollary 6.2(ii), if the corresponding groups G_1 and G_2 are twins then the correspondence $(E_1, \sigma_1) \mapsto (E_1', \sigma_1')$ gives a bijection between the sets of isomorphism classes of maximal commutative étale K-subalgebras of (A_1, τ_1) and (A_2, τ_2) that are invariant under the respective involutions and satisfy (2-2). Thus, we recover Proposition 6.1.

Acknowledgements

We thank the referee for helpful comments. During the preparation of the final version of this paper, Rapinchuck was visiting the Mathematics Department of the University of Michigan as a Gehring Professor; the hospitality and generous support of this institution are gratefully acknowledged.

References

[Adams and du Cloux 2009] J. Adams and F. du Cloux, "Algorithms for representation theory of real reductive groups", *J. Inst. Math. Jussieu* **8**:2 (2009), 209–259. MR 2010e:22006 Zbl 1221.22017

[Bhargava and Gross 2011] M. Bhargava and B. Gross, "Arithmetic invariant theory", preprint, 2011. arXiv 1206.4774

[Bourbaki 2002] N. Bourbaki, *Lie groups and Lie algebras: Chapters 4–6*, Springer, Berlin, 2002. MR 2003a:17001 Zbl 0983.17001

[Brusamarello et al. 2003] R. Brusamarello, P. Chuard-Koulmann, and J. Morales, "Orthogonal groups containing a given maximal torus", *J. Algebra* **266**:1 (2003), 87–101. MR 2004k:11050 Zbl 1079.11023

[Cassels and Fröhlich 2010] J. W. S. Cassels and A. Fröhlich (editors), *Algebraic number theory*, 2nd ed., London Math. Soc., London, 2010. MR 88h:11073 Zbl 0645.12001

[Đoković and Thǎńg 1994] D. Ž. Đoković and N. Q. Thǎńg, "Conjugacy classes of maximal tori in simple real algebraic groups and applications", *Canad. J. Math.* **46**:4 (1994), 699–717. MR 95j:20041 Zbl 0835.20060

[Garibaldi 2012] S. Garibaldi, "Outer automorphisms of algebraic groups and determining groups by their maximal tori", *Michigan Math. J.* **61**:2 (2012), 227–237. MR 2944477

[Gille 2004] P. Gille, "Type des tores maximaux des groupes semi-simples", *J. Ramanujan Math. Soc.* **19**:3 (2004), 213–230. MR 2006a:20087 Zbl 1193.20057

[Harder 1968] G. Harder, "Eine Bemerkung zum schwachen Approximationssatz", *Arch. Math.* (*Basel*) **19** (1968), 465–471. MR 39 #2767 Zbl 0205.25104

[Knus et al. 1998] M.-A. Knus, A. Merkurjev, M. Rost, and J.-P. Tignol, *The book of involutions*, American Mathematical Society Colloquium Publications **44**, American Mathematical Society, Providence, RI, 1998. MR 2000a:16031 Zbl 0955.16001

[Kottwitz 1982] R. E. Kottwitz, "Rational conjugacy classes in reductive groups", *Duke Math. J.* **49**:4 (1982), 785–806. MR 84k:20020 Zbl 0506.20017

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

[Prasad and Rapinchuk 2009] G. Prasad and A. S. Rapinchuk, "Weakly commensurable arithmetic groups and isospectral locally symmetric spaces", *Publ. Math. Inst. Hautes Études Sci.* 109 (2009), 113–184. MR 2010e:20074 Zbl 1176.22011

[Prasad and Rapinchuk 2010] G. Prasad and A. S. Rapinchuk, "Local-global principles for embedding of fields with involution into simple algebras with involution", *Comment. Math. Helv.* **85**:3 (2010), 583–645. MR 2011i:11053 Zbl 1223.11047

[Prasad and Rapinchuk 2013] G. Prasad and A. S. Rapinchuk, "On the fields generated by the lengths of closed geodesics in locally symmetric spaces", to appear in *Geom. Dedic.*, 2013. arXiv 1110.0141

[Raghunathan 2004] M. S. Raghunathan, "Tori in quasi-split-groups", *J. Ramanujan Math. Soc.* **19**:4 (2004), 281–287. MR 2005m:20114 Zbl 1080.20042

[Steinberg 1965] R. Steinberg, "Regular elements of semisimple algebraic groups", *Inst. Hautes Études Sci. Publ. Math.* 25 (1965), 49–80. MR 31 #4788 Zbl 0136.30002

[Voskresenskiĭ 1998] V. E. Voskresenskiĭ, *Algebraic groups and their birational invariants*, Translations of Mathematical Monographs **179**, American Mathematical Society, Providence, RI, 1998. MR 99g:20090

[Yeung 2011] S.-K. Yeung, "Isospectral problem of locally symmetric spaces", *Int. Math. Res. Not.* **2011**:12 (2011), 2810–2824. MR 2012k:58056 Zbl 1221.22015

Communicated by Brian Conrad
Received 2012-01-20 Revised 2012-04-29 Accepted 2012-06-07

skip@member.ams.org
Institute for Pure and Applied Mathematics, 460 Portola Plaza,
Box 957121, Los Angeles, CA 90095-7121, United States
http://www.mathcs.emory.edu/~skip/

asr3x@virginia.edu
Department of Mathematics, University of Virginia,
Charlottesville, VA 22904, United States
http://www.math.virginia.edu/Faculty/Rapinchuk/

ALGEBRA AND NUMBER THEORY 7:5 (2013)
dx.doi.org/10.2140/ant.2013.7.1179

Minimisation and reduction of 5-coverings of elliptic curves

Tom Fisher

We consider models for genus-1 curves of degree 5, which arise in explicit 5-descent on elliptic curves. We prove a theorem on the existence of minimal models with the same invariants as the minimal model of the Jacobian elliptic curve and give an algorithm for computing such models. Finally we describe how to reduce genus-1 models of degree 5 defined over $\mathbb{Q}$.

Introduction

Let E be an elliptic curve defined over a number field K. An *n-covering* of E is a pair (C, π), where C is a smooth curve of genus 1 and $\pi : C \to E$ is a morphism, both defined over K, with the property that $\pi = [n] \circ \psi$ for some isomorphism $\psi : C \to E$ defined over $\overline{K}$. An n-descent on E computes the everywhere locally soluble n-coverings of E. For such n-coverings, we have $\psi^*(n.0_E) \sim D$ for some K-rational divisor D on C. The complete linear system $|D|$ defines a morphism $C \to \mathbb{P}^{n-1}$. Thus, in the cases $n = 2, 3, 4$, we may represent C by a binary quartic, ternary cubic, or pair of quadrics in four variables. In the case $n = 5$, we obtain curves $C \subset \mathbb{P}^4$ of degree 5 that are defined by the 4×4 Pfaffians of a 5×5 alternating matrix of linear forms.

The question naturally arises as to how we can choose coordinates on $\mathbb{P}^{n-1}$ so that the equations for C have small coefficients. In the cases $n = 2, 3, 4$, this was answered in [Cremona et al. 2010] using the combination of two techniques called *minimisation* and *reduction*. In this paper, we extend to the case $n = 5$. We establish results on minimisation over an arbitrary local field (immediately implying results over any number field of class number 1), whereas those for reduction are specific to the case $K = \mathbb{Q}$. Implementations of our algorithms in the case $K = \mathbb{Q}$ are available in Magma [Bosma et al. 1997].

MSC2010: primary 11G05; secondary 11G07, 14H52, 14H25.
Keywords: elliptic curves, genus-1 curves, minimisation, reduction, descent.

1. Genus-1 models

A *genus*-1 *model* (of degree 5) is a 5×5 alternating matrix of linear forms in variables $x_1, \dots, x_5$. We write $X_5(R)$ for the space of all genus-1 models with coefficients in a ring R. Models Φ and Φ' are *R-equivalent* if $\Phi' = [A, B]\Phi$ for some $A, B \in \mathrm{GL}_5(R)$. Here the action of A is via $\Phi \mapsto A\Phi A^T$, and the action of B is via $(\Phi_{ij}(x_1, \dots, x_5)) \mapsto (\Phi_{ij}(x'_1, \dots, x'_5))$, where $x'_j = \sum_{i=1}^5 B_{ij}x_i$. The *determinant* of the transformation $g = [A, B]$ is $\det g = (\det A)^2 \det B$.

We write $\mathrm{Pf}(\Phi)$ for the row vector $(p_1, \dots, p_5)$, where p_i is $(-1)^{i-1}$ times the Pfaffian of the 4×4 submatrix obtained by deleting the ith row and column of Φ. This choice of signs is made so that $\mathrm{Pf}(\Phi)\Phi = 0$. For $A \in \mathrm{GL}_5(R)$, we note that $\mathrm{Pf}(A\Phi A^T) = \mathrm{Pf}(\Phi) \operatorname{adj} A$.

A genus-1 model $\Phi \in X_5(K)$ over a field K is *nonsingular* if the subscheme $\mathscr{C}_\Phi = \{\operatorname{rank} \Phi \le 2\} \subset \mathbb{P}^4$ defined by the 4×4 Pfaffians of Φ is a smooth curve of genus 1. We write $K[X_5]$ for the polynomial ring in the fifty coefficients of a genus-1 model. A polynomial $F \in K[X_5]$ is an *invariant* of *weight* k if $F \circ g = (\det g)^k F$ for all $g = [A, B]$ with $A, B \in \mathrm{GL}_5(\overline{K})$. Taking A and B to be scalar matrices shows that an invariant of weight k is a homogeneous polynomial of degree $5k$.

Theorem 1.1. *Let $c_4, c_6, \Delta \in \mathbb{Z}[X_5]$ be the invariants of weights* 4, 6 *and* 12 *satisfying $c_4^3 - c_6^2 = 1728\Delta$ and scaled as specified in [Fisher 2008].*

(i) *A model $\Phi \in X_5(K)$ is nonsingular if and only if $\Delta(\Phi) \neq 0$.*

(ii) *There exist $a_1, a_2, a_3, a_4, a_6 \in \mathbb{Z}[X_5]$ and $b_2, b_4, b_6 \in \mathbb{Z}[X_5]$ satisfying*

$$\begin{aligned} &b_2 = a_1^2 + 4a_2, \quad b_4 = a_1a_3 + 2a_4, \quad b_6 = a_3^2 + 4a_6, \\ &c_4 = b_2^2 - 24b_4 \quad \textit{and} \quad c_6 = -b_2^3 + 36b_2b_4 - 216b_6. \end{aligned} \tag{1}$$

(iii) *If $\Phi \in X_5(K)$ is nonsingular, then $\mathscr{C}_\Phi$ has Jacobian elliptic curve*

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6,$$

where $a_i = a_i(\Phi)$.

For the proof of Theorem 1.1(ii), we use the following lemma:

Lemma 1.2. *Let $c_4, c_6, \Delta \in R = \mathbb{Z}[x_1, \dots, x_N]$ be primitive polynomials satisfying $c_4^3 - c_6^2 = 1728\Delta$. If there exists $a_1 \in R$ satisfying $a_1^2c_4 + c_6 \equiv 0 \pmod 4$, then there exist $a_2, a_3, a_4, a_6, b_2, b_4, b_6 \in R$ satisfying* (1).

Proof. By unique factorisation in $\mathbb{F}_3[x_1, \dots, x_N]$ and the Chinese remainder theorem, there exists some $b_2 \in R$ such that $c_4 \equiv b_2^2 \pmod 3$, $c_6 \equiv -b_2^3 \pmod 3$ and $b_2 \equiv a_1^2 \pmod 4$. Then $b_2c_4 + c_6 \equiv 0 \pmod{12}$, and $c_4^3 \equiv c_6^2 \equiv b_2^2c_4^2 \pmod{24}$. Since c_4 is primitive, it follows that $c_4 \equiv b_2^2 \pmod{24}$. Next, putting $x = b_2$ in an identity of Kraus [1989],

$$(x^2-c_4)^3=(x^3-3xc_4-2c_6)(x^3+2c_6)+3(xc_4+c_6)^2+c_6^2-c_4^3,$$

we deduce $b_2^3-3b_2c_4-2c_6\equiv 0 \pmod{432}$. We put $b_4=(b_2^2-c_4)/24$ and $b_6=(b_2^3-3b_2c_4-2c_6)/432$. Then $0\equiv c_4^3-c_6^2\equiv 16b_2^2(b_2b_6-b_4^2) \pmod{64}$, and so $b_2b_6\equiv b_4^2 \pmod 4$. By unique factorisation in $\mathbb{F}_2[x_1,\dots,x_N]$, there exists $a_3\in R$ with $b_4\equiv a_1a_3 \pmod 2$. Then $b_4^2\equiv a_1^2a_3^2 \pmod 4$, and $b_6\equiv a_3^2 \pmod 4$. We put $a_2=(b_2-a_1^2)/4$, $a_4=(b_4-a_1a_3)/2$ and $a_6=(b_6-a_3^2)/4$. □

Proof of Theorem 1.1. (i) This is [Fisher 2008, Theorem 4.4(ii)].

(ii) By Lemma 1.2, it suffices to construct $a_1\in\mathbb{Z}[X_5]$ with $a_1^2c_4+c_6\equiv 0 \pmod 4$. In [Fisher 2008, Section 10], we constructed an invariant $a_1\in\mathbb{F}_2[X_5]$ of weight 1 and showed that together with Δ it generates the ring of invariants in characteristic 2. In particular, $c_4\equiv a_1^4 \pmod 2$, and $c_6\equiv a_1^6 \pmod 2$. So if we lift a_1 to $\mathbb{Z}[X_5]$, then $a_1^2c_4+c_6=2f$ for some $f\in\mathbb{Z}[X_5]$. Since a_1 is an invariant mod 2, a_1^2 is an invariant mod 4 and f is an invariant mod 2. Therefore, $f\equiv\lambda a_1^6 \pmod 2$ for some $\lambda\in\{0,1\}$. Hence, $a_1^2c_4\pm c_6\equiv 0 \pmod 4$. Specialising to one of the Weierstrass models in [Fisher 2008, Section 6] shows that the sign is $+$.

(iii) It is shown in [Fisher 2008, Theorem 4.4(iii)] that if K is a perfect field with characteristic not 2 or 3, then $\mathscr{C}_\Phi$ has Jacobian $y^2=x^3-27c_4(\Phi)x-54c_6(\Phi)$. The proof is now identical to that of [Cremona et al. 2010, Theorem 2.10]. This generalises a result of Artin, Rodriguez-Villegas and Tate [Artin et al. 2005] in the case $n=3$. □

2. Minimisation theorems

Let K be a discrete valuation field with ring of integers $\mathcal{O}_K$ and normalised valuation $v:K^\times\to\mathbb{Z}$. We assume throughout that the residue field k is perfect. A genus-1 model $\Phi\in X_5(K)$ is *integral* if it has coefficients in $\mathcal{O}_K$. If Φ is nonsingular and integral, then by Theorem 1.1 and the standard formulae for transforming Weierstrass equations, we have $v(\Delta(\Phi))=v(\Delta_E)+12\,\ell(\Phi)$, where Δ_E is the minimal discriminant of $E=\operatorname{Jac}(\mathscr{C}_\Phi)$ and $\ell(\Phi)$ is a nonnegative integer we call the *level*. We say that Φ is *minimal* if $v(\Delta(\Phi))$, or equivalently the level, is minimal among all integral models K-equivalent to Φ. Notice that if $\Phi'=g\Phi$ for some $g=[A,B]$ with $A,B\in\mathrm{GL}_5(K)$, then $\ell(\Phi')=\ell(\Phi)+v(\det g)$.

Theorem 2.1. *Let $\Phi\in X_5(K)$ be nonsingular.*

(i) (*Weak minimisation theorem*) *If $\mathscr{C}_\Phi(K)\neq\varnothing$, then Φ is K-equivalent to an integral model of level* 0.

(ii) (*Strong minimisation theorem*) *If $\mathscr{C}_\Phi(L)\neq\varnothing$, where L is an unramified extension of K, then Φ is K-equivalent to an integral model of level* 0.

In this section, we prove the weak minimisation theorem. In Section 3, we describe an explicit algorithm for minimising. Inspection of this algorithm shows that the minimal level is unchanged by an unramified extension. Theorem 2.1(ii) then follows from Theorem 2.1(i). In Section 7, we prove a converse to the strong minimisation theorem thereby showing this result is best possible.

We refer to [Cremona et al. 2010, Section 2] for notation and results analogous to those in Section 1 for genus-1 models of degree 4, i.e., quadric intersections. Let E be an elliptic curve over K and D a K-rational divisor on E of degree $n = 4$ or 5. The complete linear system $|D|$ defines an embedding $E \subset \mathbb{P}^{n-1}$. The image is defined by a genus-1 model $\Phi \in X_n(K)$, and this model is uniquely determined, up to K-equivalence, by the pair $(E, [D])$. Moreover, every nonsingular model $\Phi \in X_n(K)$ with $\mathscr{C}_\Phi(K) \neq \varnothing$ arises in this way. Therefore, Theorem 2.1(i) is an immediate consequence of the following:

Theorem 2.2. *Let E/K be an elliptic curve with integral Weierstrass equation*

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6, \tag{2}$$

and let $D \in \operatorname{Div}_K(E)$ be a divisor on E of degree $n = 4$ or 5. Then $(E, [D])$ can be represented by an integral genus-1 model with the same discriminant as (2).

The case $n = 4$ is proved in [Cremona et al. 2010, Theorem 3.8]. To deduce the case $n = 5$ from the case $n = 4$, we use the following lemma:

Lemma 2.3. *Let $D \in \operatorname{Div}_K(E)$ be a divisor of degree 4, and let $P \in E(K)$. Let ℓ_i, α_i and β_i for $i = 1, 2, 3$ be linear forms in $x_1, \dots, x_4$ over K. The following statements are equivalent:*

(i) *The pair $(E, [D])$ is represented by the quadric intersection*

$$\ell_1\alpha_1 + \ell_2\alpha_2 + \ell_3\alpha_3 = 0 \quad \textit{and} \quad \ell_1\beta_1 + \ell_2\beta_2 + \ell_3\beta_3 = 0, \tag{3}$$

and P is the point defined by $\ell_1 = \ell_2 = \ell_3 = 0$.

(ii) *The pair $(E, [D + P])$ is represented by the genus-1 model of degree 5*

$$\begin{pmatrix} 0 & \gamma & \alpha_1 & \alpha_2 & \alpha_3 \\ & 0 & \beta_1 & \beta_2 & \beta_3 \\ & & 0 & \ell_3 & -\ell_2 \\ & - & & 0 & \ell_1 \\ & & & & 0 \end{pmatrix}, \tag{4}$$

where $\gamma = x_5$ and P is the point $(x_1 : \cdots : x_5) = (0 : \cdots : 0 : 1)$.

Proof. An isomorphism $\psi : C_4 \to C_5$ between the curves C_4 and C_5 defined by (3) and (4) is given by

$$\psi : (x_1 : x_2 : x_3 : x_4) \mapsto (x_1\ell_i : x_2\ell_i : x_3\ell_i : x_4\ell_i : \alpha_j\beta_k - \alpha_k\beta_j)$$

(where i, j and k are any cyclic permutation of 1, 2 and 3) with inverse

$$\psi^{-1} : (x_1 : x_2 : x_3 : x_4 : x_5) \mapsto (x_1 : x_2 : x_3 : x_4).$$

The points $\{\ell_1 = \ell_2 = \ell_3 = 0\} \in C_4(K)$ and $(0 : \cdots : 0 : 1) \in C_5(K)$ are identified by this isomorphism. To prove the equivalence of (i) and (ii), we note that if $C_4 \subset \mathbb{P}^3$ meets some plane in the divisor $D = P_1 + P_2 + P_3 + P_4$, then the points $\psi(P_i)$ and $(0 : \cdots : 0 : 1)$ are a hyperplane section for $C_5 \subset \mathbb{P}^4$. □

Lemma 2.4. *The genus-*1 *models* (3) *and* (4) *have the same invariants.*

Proof. Let Φ be the matrix (4), and write $P = \mathrm{Pf}(\Phi) = (p_1, \ldots, p_5)$. Then (3) and (4) define curves $C_4 = \{p_1 = p_2 = 0\} \subset \mathbb{P}^3$ and $C_5 = \{\mathrm{rank}\,\Phi \leq 2\} \subset \mathbb{P}^4$. According to [Fisher 2008, Section 5.4], there are invariant differentials ω_4 on C_4 and ω_5 on C_5 given by

$$\omega_n = \frac{x_1^2 d(x_2/x_1)}{\Omega_n(x_1, \ldots, x_n)},$$

where

$$\Omega_4 = \frac{\partial p_1}{\partial x_3}\frac{\partial p_2}{\partial x_4} - \frac{\partial p_1}{\partial x_4}\frac{\partial p_2}{\partial x_3} \quad \text{and} \quad \Omega_5 = \frac{\partial P}{\partial x_3}\frac{\partial \Phi}{\partial x_5}\frac{\partial P^T}{\partial x_4}.$$

In the expression for Ω_5, we have written the partial derivative of a matrix as a shorthand for the matrix of partial derivatives. Since the only entries of Φ to involve x_5 are in the top left 2×2 submatrix, it is clear that $\Omega_4 = \pm\Omega_5$. Hence, the isomorphism $\psi : C_4 \to C_5$ identifies the invariant differentials ω_4 and ω_5 (up to sign). It follows by [Fisher 2008, Proposition 5.23] that (3) and (4) have the same invariants c_4, c_6 and Δ. □

Proof of Theorem 2.2. Let $D \in \mathrm{Div}_K(E)$ be a divisor of degree 4, and let $P \in E(K)$. We show that if the theorem holds for D, then it holds for $D + P$. Suppose $(E, [D])$ is represented by an integral quadric intersection with discriminant Δ. Since $\mathcal{O}_K$ is a principal ideal domain, $\mathrm{SL}_4(\mathcal{O}_K)$ acts transitively on $\mathbb{P}^3(K)$. So we may assume P is the point $(x_1 : x_2 : x_3 : x_4) = (0 : 0 : 0 : 1)$. Our model is now of the form (3) with $\ell_i = x_i$ for $i = 1, 2, 3$. We may choose the linear forms α_i and β_i to have coefficients in $\mathcal{O}_K$. Then the genus-1 model (4) is an integral model of discriminant Δ representing the pair $(E, [D + P])$. □

3. Minimisation algorithms

For $\Phi \in X_5(\mathcal{O}_K)$, we write $\phi \in X_5(k)$ for its reduction mod π. The *singular locus* $\mathrm{Sing}\,\mathscr{C}_\phi$ is the set of points $P \in \mathscr{C}_\phi$ with tangent space of dimension greater than 1. (We make this definition regardless of whether $\mathscr{C}_\phi$ is a curve. In particular, all points on components of dimension at least 2 are singular.) For example, if ϕ takes the form (4) with $\gamma = x_5$ and ℓ_i, α_i and β_i linear forms in $x_1, \ldots, x_4$, then

$P = (0 : \cdots : 0 : 1)$ is singular if and only if ℓ_1, ℓ_2 and ℓ_3 are linearly dependent. An integral genus-1 model $\Phi \in X_5(\mathcal{O}_K)$ is *saturated* if its 4×4 Pfaffians $p_1, \ldots, p_5$ are linearly independent mod π. We write I_m for the $m \times m$ identity matrix.

Our algorithm for minimising genus-1 models of degree 5 generalises the algorithm for models of degree 3 in [Cremona et al. 2010, Section 4B].

Theorem 3.1. *Let $\Phi \in X_5(\mathcal{O}_K)$ be saturated and of positive level.*

(i) *The singular locus $\operatorname{Sing} \mathscr{C}_\phi$ does not span $\mathbb{P}^4$.*

(ii) *Let $B \in \mathrm{GL}_5(\mathcal{O}_K)$ represent a change of coordinates on $\mathbb{P}^4$ mapping the linear span of the singular locus in (i) to $\{x_{m+1} = \cdots = x_5 = 0\}$. Then there exist $A \in \mathrm{GL}_5(K)$ and $\mu \in K^\times$ such that $[A, \mu \operatorname{Diag}(I_m, \pi I_{5-m})B]\Phi$ is an integral model of the same or smaller level.*

(iii) *If Φ is nonminimal, then repeating the procedure in (ii) either gives a nonsaturated model or decreases the level after finitely many iterations.*

As it stands, Theorem 3.1 does not give an algorithm for minimising since we must show how to find A and μ in (ii) and show how to decrease the level of a nonsaturated model. We do this in Theorem 3.2 below. Theorem 3.1 is proved in Sections 4 and 5. In Section 6, we bound the number of iterations required in (iii).

Theorem 3.2. *Let $\Phi \in X_5(\mathcal{O}_K)$ be nonsingular. Let ℓ_0 be the minimum of the levels of all integral models that are K-equivalent to Φ via a transformation of the form $[A, \mu I_5]$, where $A \in \mathrm{GL}_5(K)$ and $\mu \in K^\times$.*

(i) *We may compute an integral model of the form $[A, \mu I_5]\Phi$ with level ℓ_0 as follows*:

Step 1. *Write $\mathrm{Pf}(\Phi) = (p_1, \ldots, p_5)$. Compute $A = (a_{ij}) \in \mathrm{GL}_5(K)$ and quadrics $q_1, \ldots, q_5 \in \mathcal{O}_K[x_1, \ldots, x_5]$ such that $p_j = \sum_{i=1}^5 a_{ij} q_i$ and $q_1, \ldots, q_5$ are linearly independent modulo π. Then replace Φ by $[A, \mu I_5]\Phi$, where $\mu \in K^\times$ is chosen so that Φ has coefficients in $\mathcal{O}_K$ not all in $\pi\mathcal{O}_K$.*

Step 2. *Replace Φ by $[A, I_5]\Phi$, where $A \in \mathrm{GL}_5(\mathcal{O}_K)$ is chosen so that the first two rows of Φ are divisible by π^e with $e \geq 0$ as large as possible. Then divide the first row and column by π^e.*

(ii) *If the model computed in Step 1 is nonsaturated, then we may compute an integral model of level smaller than ℓ_0 by modifying Step 2 so that we divide the first two rows and columns by π^e and then make a transformation of the form $[I_5, B]$ to preserve integrality.*

Proof. With the notation of Step 1, we have

$$\mathrm{Pf}(A\Phi A^T) = \mathrm{Pf}(\Phi) \operatorname{adj} A = (q_1, \ldots, q_5) A \operatorname{adj} A = (\det A)(q_1, \ldots, q_5).$$

So after Step 1, we have $\mathrm{Pf}(\Phi) = (\lambda q_1, \dots, \lambda q_5)$, where $\lambda = \mu^2 \det A \in \mathcal{O}_K$. We split into the cases $v(\lambda) = 0$ and $v(\lambda) \geq 1$. First we need two lemmas.

Lemma 3.3. *Let $\Phi, \Phi' \in X_5(\mathcal{O}_K)$ be nonsingular models with $\Phi' = [A, \mu I_5]\Phi$ for some $A \in \mathrm{GL}_5(K)$ and $\mu \in K^\times$.*

(i) *If Φ is saturated, then $\ell(\Phi') \geq \ell(\Phi)$ with equality if and only if Φ and Φ' are $\mathcal{O}_K$-equivalent.*

(ii) *If Φ and Φ' are of the form output by Step 1, then they are $\mathcal{O}_K$-equivalent.*

Proof. We have $\mathrm{Pf}(\Phi') = \mathrm{Pf}(\Phi)M$, where $M = \mu^2 \operatorname{adj} A$.

(i) Since Φ is saturated, M has entries in $\mathcal{O}_K$. Hence, $\ell(\Phi') - \ell(\Phi) = \frac{1}{2} v(\det M) \geq 0$ with equality if and only if $M \in \mathrm{GL}_5(\mathcal{O}_K)$. If $M \in \mathrm{GL}_5(\mathcal{O}_K)$, then replacing $[A, \mu I_5]$ by $[\lambda A, \lambda^{-2}\mu I_5]$ for suitable $\lambda \in K^\times$, we may assume $A \in \mathrm{GL}_5(\mathcal{O}_K)$. Since Φ and Φ' have the same level, they must therefore be $\mathcal{O}_K$-equivalent.

(ii) Since $\mathrm{Pf}(\Phi)$ and $\mathrm{Pf}(\Phi')$ are scalar multiples of bases for the same $\mathcal{O}_K$-module, some scalar multiple of M belongs to $\mathrm{GL}_5(\mathcal{O}_K)$. So after replacing $[A, \mu I_5]$ by $[\lambda A, \lambda^{-2}\mu I_5]$ for suitable $\lambda \in K^\times$, we may assume $A \in \mathrm{GL}_5(\mathcal{O}_K)$. Since Φ and Φ' are primitive, they must therefore be $\mathcal{O}_K$-equivalent. □

Lemma 3.4. *Let $\phi \in X_5(k)$ be a genus-*1 *model all of whose* 4×4 *Pfaffians are identically zero. Then ϕ is k-equivalent to either*

$$\begin{pmatrix} 0 & \ell_2 & \ell_3 & \ell_4 & \ell_5 \\ & 0 & 0 & 0 & 0 \\ & & 0 & 0 & 0 \\ & - & & 0 & 0 \\ & & & & 0 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0 & x_1 & x_2 & 0 & 0 \\ & 0 & x_3 & 0 & 0 \\ & & 0 & 0 & 0 \\ & - & & 0 & 0 \\ & & & & 0 \end{pmatrix},$$

where $\ell_2, \dots, \ell_5$ are linear forms. □

We now complete the proof of Theorem 3.2. Let $e = v(\lambda)$. If $e = 0$, then Φ is saturated and we are done by Lemma 3.3(i). So suppose $e \geq 1$. In Step 1, the matrix A has entries in $\mathcal{O}_K$. So $v(\mu) \leq 0$, and the level is increased by

$$2v(\det A) + 5v(\mu) \leq 2v(\mu^2 \det A) = 2e.$$

Lemma 3.3(ii) shows that when we apply Step 1 to both Φ and the model implicit in the definition of ℓ_0, we obtain models that are $\mathcal{O}_K$-equivalent. So it will suffice to show that Step 2 reduces the level by $2e$, whereas the modified version in (ii) reduces the level by more than $2e$.

Since $\mathrm{Pf}(\Phi) = (\lambda q_1, \dots, \lambda q_5)$, we have $(q_1, \dots, q_5)\Phi = 0$. The reduction of Φ takes one of the forms specified in Lemma 3.4. In the first case, we have $q_1 \ell_j \equiv 0 \pmod{\pi}$ for $j = 2, \dots, 5$. This contradicts the choices of $q_1, \dots, q_5$ and μ in Step 1. So we must be in the second case. Replacing Φ by an $\mathcal{O}_K$-equivalent model, we may assume it takes the form (4) with $\ell_i = x_i$ for $i = 1, 2, 3$ and α_1, α_2, α_3, β_1, β_2, β_3 and γ linear forms that vanish mod π. By row and column operations, we may assume $\alpha_2 \in \langle x_2, \dots, x_5\rangle$ and $\alpha_3 \in \langle x_3, \dots, x_5\rangle$. Then since $\pi^e \mid (x_1\alpha_1 + x_2\alpha_2 + x_3\alpha_3)$, we have $\pi^e \mid \alpha_1, \alpha_2, \alpha_3$. Likewise, we may assume $\pi^e \mid \beta_1, \beta_2, \beta_3$. The remaining Pfaffians show that $\pi^e \mid \gamma$. Step 2 and its modified version in (ii) now reduce the level by $2e$ and $3e$, respectively. □

Corollary 3.5. *For the proof of Theorem 3.1, we are free to replace Φ by an $\mathcal{O}_K$-equivalent model and to replace K by an unramified field extension.*

Proof. Let $\Phi_1, \Phi_2 \in X_5(\mathcal{O}_K)$ be $\mathcal{O}_K$-equivalent models and $\Phi_1', \Phi_2' \in X_5(\mathcal{O}_K)$ the models returned by Theorem 3.1(ii). Lemma 3.3(i) and [Cremona et al. 2010, Lemma 4.1] together show that if Φ_1' is saturated and $\ell(\Phi_1') = \ell(\Phi_2')$, then Φ_1' and Φ_2' are $\mathcal{O}_K$-equivalent. Therefore, the number of iterations required in Theorem 3.1(iii) depends only on the $\mathcal{O}_K$-equivalence class of Φ.

For the final statement, we note that the performance of the algorithms in Theorems 3.1 and 3.2 is unchanged by an unramified field extension. □

Replacing K by its strict Henselisation, we may assume in the next three sections that K is Henselian and its residue field k is algebraically closed.

4. The singular locus

In this section and the next, we prove Theorem 3.1.

Lemma 4.1. *Let $\phi \in X_5(k)$ be a genus-1 model. Suppose $\Gamma \subset \mathcal{C}_\phi$ is either a line or a (nonsingular) conic. Then either $\Gamma \subset \operatorname{Sing}\mathcal{C}_\phi$ or*

$$\#(\Gamma \cap \operatorname{Sing}\mathcal{C}_\phi) = \begin{cases} 1 & \text{if } c_4(\phi) = c_6(\phi) = 0, \\ 2 & \text{otherwise.} \end{cases}$$

Proof. (i) If $\mathcal{C}_\phi$ contains the line $\Gamma = \{x_3 = x_4 = x_5 = 0\}$ but not every point on Γ is singular, then (unless $\mathcal{C}_\phi$ is a cone, which is an easy special case with $c_4(\phi) = c_6(\phi) = 0$) we may suppose ϕ is

$$\begin{pmatrix} 0 & x_1 & x_2 & * & * \\ & 0 & * & \alpha & \beta \\ & & 0 & \gamma & \delta \\ & - & & 0 & x_5 \\ & & & & 0 \end{pmatrix},$$

where α, β, γ, δ and the entries $*$ are linear forms in x_3, x_4, x_5. By row and column operations (and substitutions for x_1 and x_2), we may suppose α, β, γ and δ do not involve x_5. We write $\alpha = \alpha_3x_3 + \alpha_4x_4, \dots, \delta = \delta_3x_3 + \delta_4x_4$ and put

$$q(s,t) = \det\left(\begin{pmatrix} \gamma_3 & \gamma_4 \\ \delta_3 & \delta_4 \end{pmatrix} s - \begin{pmatrix} \alpha_3 & \alpha_4 \\ \beta_3 & \beta_4 \end{pmatrix} t\right).$$

By the Jacobian criterion, we have

$$\Gamma \cap \operatorname{Sing} \mathscr{C}_\phi = \{\, (s : t : 0 : 0 : 0) \mid q(s,t) = 0 \,\}.$$

A calculation using Lemma 2.4 shows that $c_4(\phi) = \Delta(q)^2$ and $c_6(\phi) = -\Delta(q)^3$, where $\Delta(q)$ is the discriminant of the binary quadratic form q.

(ii) Suppose $\mathscr{C}_\phi$ contains the conic $\Gamma = \{f(x_1,x_2,x_3) = x_4 = x_5 = 0\}$ but not every point on Γ is singular. Let $\operatorname{Pf}(\phi) = (p_1,\dots,p_5)$. Replacing ϕ by an equivalent model, we may suppose $p_i(x_1,x_2,x_3,0,0) = 0$ for $i = 1,2,3,4$ and $p_5(x_1,x_2,x_3,0,0) = f$. Since $\operatorname{Pf}(\phi)\phi = 0$ and Γ is not contained in any component of $\mathscr{C}_\phi$ of higher dimension, we may further suppose the last column of ϕ has entries $x_4, x_5, 0, 0, 0$. The monomials appearing in the invariants c_4 and c_6 are limited by the fact they are invariant under all pairs of diagonal matrices. These restrictions show that $c_4(\phi)$ and $c_6(\phi)$ are unchanged if we set $x_4 = x_5 = 0$ in all entries of ϕ outside the last row and column. Writing $f = \sum_{i \le j} a_{ij} x_i x_j$ and $\phi_{34} = \sum b_i x_i$, we put

$$\delta = \begin{vmatrix} 2a_{11} & a_{12} & a_{13} & b_1 \\ a_{12} & 2a_{22} & a_{23} & b_2 \\ a_{13} & a_{23} & 2a_{33} & b_3 \\ b_1 & b_2 & b_3 & 0 \end{vmatrix}.$$

A calculation using Lemma 2.4 shows that $c_4(\phi) = \delta^2$ and $c_6(\phi) = -\delta^3$. By a change of coordinates, we may suppose $f = x_1x_3 - x_2^2$. Then δ is the discriminant of the binary quadratic form $q(s,t) = \phi_{34}(s^2, st, t^2, 0, 0)$, and by the Jacobian criterion,

$$\Gamma \cap \operatorname{Sing} \mathscr{C}_\phi = \{\, (s^2 : st : t^2 : 0 : 0) \mid q(s,t) = 0 \,\}. \qquad \square$$

Lemma 4.2. *Let $\phi \in X_5(k)$ be a genus-1 model. Suppose the 4×4 Pfaffians $p_1,\dots,p_5$ are linearly independent and $c_4(\phi) = c_6(\phi) = 0$. Then either $\operatorname{Sing} \mathscr{C}_\phi$ is a linear subspace of $\mathbb{P}^4$ or ϕ is equivalent to a model of the form*

$$\begin{pmatrix} 0 & \xi & \alpha & \beta & \eta \\ & 0 & \gamma & \delta & x_5 \\ & & 0 & x_5 & 0 \\ & - & & 0 & 0 \\ & & & & 0 \end{pmatrix}, \tag{5}$$

where $\xi, \eta, \alpha, \beta, \gamma$ and δ are linear forms in $x_1,\dots,x_5$.

Proof. If $P_1, P_2 \in \operatorname{Sing} \mathscr{C}_\phi$ are distinct and the line ℓ between them is contained in $\mathscr{C}_\phi$, then by Lemma 4.1, we have $\ell \subset \operatorname{Sing} \mathscr{C}_\phi$. So either $\operatorname{Sing} \mathscr{C}_\phi$ is a linear subspace of $\mathbb{P}^4$ or there exist $P_1, P_2 \in \operatorname{Sing} \mathscr{C}_\phi$ joined by a line not contained in $\mathscr{C}_\phi$. We move these points to $(1 : 0 : \cdots : 0)$ and $(0 : 1 : \cdots : 0)$. Writing $\phi = \sum x_i M_i$, the matrices M_1 and M_2 have rank 2, but their sum has rank 4. Therefore, ϕ is equivalent to a model with $\phi_{12} = x_1$, $\phi_{34} = x_2$ and all other ϕ_{ij} (for $i < j$) linear forms in x_3, x_4, x_5. Since P_1 and P_2 are singular, ϕ_{35} and ϕ_{45} are linearly dependent and ϕ_{15} and ϕ_{25} are linearly dependent. So the space of linear forms spanned by the entries of the last column has dimension at most 2. In fact, it has dimension exactly 2 since $p_1, \dots, p_5$ are linearly independent.

Replacing ϕ by an equivalent model, we may assume it has last column with entries $x_4, x_5, 0, 0, 0$. The transformation used here does not move P_1 and P_2 but may change the matrices M_1 and M_2. Let $\Gamma = \{x_4 = x_5 = p_5 = 0\} \subset \mathscr{C}_\phi$. Then P_1 and P_2 are contained in Γ, but the line between them is not. It follows that Γ is either a nonsingular conic or a pair of concurrent lines. In either case, Lemma 4.1 shows that $\Gamma \subset \operatorname{Sing} \mathscr{C}_\phi$. By the Jacobian criterion, it follows that $\phi_{34} \in \langle x_4, x_5 \rangle$. However, ϕ_{34} is nonzero since $p_1, \dots, p_5$ are linearly independent. Therefore, ϕ is equivalent to a model of the form (5). □

Lemma 4.3. *Let $\Phi \in X_5(\mathcal{O}_K)$ be a saturated nonsingular model with reduction ϕ of the form (5). Suppose $\operatorname{Sing} \mathscr{C}_\phi$ has linear span $\{x_{m+1} = \cdots = x_5 = 0\}$.*

(i) *There exist $A \in \mathrm{GL}_5(K)$ and $\mu \in K^\times$ such that $[A, \mu \operatorname{Diag}(I_m, \pi I_{5-m})]\Phi$ is an integral model of the same or smaller level.*

(ii) *Suppose that either $\delta = 0$ and $\Phi_{45} \equiv 0 \pmod{\pi^2}$ or $\Phi_{35} \equiv \Phi_{45} \equiv 0 \pmod{\pi^2}$. Then there is a transformation as in (i) that decreases the level.*

Proof. Computing the 4×4 Pfaffians of (5), we find

$$\mathscr{C}_\phi = \{\eta = x_5 = \alpha\delta - \beta\gamma = 0\} \cup \{\gamma = \delta = x_5 = 0\}. \tag{6}$$

First suppose γ, δ and x_5 are linearly dependent. By an $\mathcal{O}_K$-equivalence, we may assume $\delta = 0$. Then $\{\gamma = x_5 = 0\} \subset \operatorname{Sing} \mathscr{C}_\phi \subset \{x_5 = 0\}$. Therefore, $m = 3$ or 4. The required transformations are as follows:

	$m = 3$	$m = 4$
(i)	$A = \operatorname{Diag}(\pi, 1, 1, 1, 1),\ \mu = \pi^{-1}$	$A = \operatorname{Diag}(\pi, \pi, 1, 1, 1),\ \mu = \pi^{-1}$
(ii)	$A = \operatorname{Diag}(\pi, 1, 1, 1, 1),\ \mu = \pi^{-1}$	$A = \operatorname{Diag}(\pi, 1, 1, \pi^{-1}, \pi^{-1}),\ \mu = 1$

Now suppose γ, δ and x_5 are linearly independent. Since Φ is saturated, η and x_5 are linearly independent. A calculation shows $\operatorname{Sing} \mathscr{C}_\phi$ is the first of the two components in (6). Therefore, $m = 2$ or 3. If $m = 2$, then we may assume β, γ, δ, η and ϕ_{25} are linear forms in x_3, x_4 and x_5. The required transformations are as follows:

	$m = 2$	$m = 3$
(i)	$A = \operatorname{Diag}(\pi, 1, 1, 1, 1),\ \mu = \pi^{-1}$	$A = \operatorname{Diag}(1, 1, 1, 1, \pi^{-1}),\ \mu = 1$
(ii)	$A = \operatorname{Diag}(1, 1, 1, \pi^{-1}, \pi^{-1}),\ \mu = 1$	$A = \operatorname{Diag}(\pi, \pi, 1, 1, \pi^{-1}),\ \mu = \pi^{-1}$

□

We now prove the first two parts of Theorem 3.1. Let $\Phi \in X_5(\mathcal{O}_K)$ be saturated and of positive level. Lemma 4.2 shows that either $\operatorname{Sing} \mathscr{C}_\phi$ is a linear subspace or

$\mathscr{C}_\phi$ is contained in a hyperplane. Since $\mathscr{C}_\phi$ is defined by five linearly independent quadrics, it cannot be all of $\mathbb{P}^4$. This proves Theorem 3.1(i).

The proof of Theorem 3.1(ii) in the case ϕ takes the form (5) was already given in Lemma 4.3(i). So by Lemma 4.2, we may assume $\operatorname{Sing}\mathscr{C}_\phi = \{x_{m+1} = \cdots = x_5 = 0\}$. We apply Lemma 3.4 to the reduction mod π of $[I_5, \operatorname{Diag}(I_m, \pi I_{5-m})]\Phi$. In the second case of that lemma, we have $m \geq 3$. We take $A = \operatorname{Diag}(1, 1, 1, 1, \pi^{-1})$ and $\mu = 1$. Otherwise, we are in the first case. If $m \geq 2$, then we take $\mu = \pi^{-1}$ and $A = \operatorname{Diag}(\pi, 1, 1, 1, 1)$. It remains to treat the case $m = 1$; in other words, the case $\operatorname{Sing}\mathscr{C}_\phi$ is a point.

By [Fisher 2008, Lemma 5.8], every component of $\mathscr{C}_\phi$ has dimension at least 1. So if $\operatorname{Sing}\mathscr{C}_\phi$ is just a point, then there are also smooth points on $\mathscr{C}_\phi$. Since K is Henselian, it follows that $\mathscr{C}_\Phi(K) \neq \varnothing$, and so by Theorem 2.1(i), Φ is nonminimal. With this extra hypothesis, we show in the next section that the singular point on $\mathscr{C}_\phi$ is nonregular (as a point on the $\mathcal{O}_K$-scheme $\mathscr{C}_\Phi$).

We may suppose $\phi_{12} = x_1$ and all other ϕ_{ij} (for $i < j$) are linear forms in $x_2, \ldots, x_5$. Since $P = (1 : 0 : \cdots : 0)$ is singular, ϕ_{34}, ϕ_{35} and ϕ_{45} are linearly dependent. So replacing Φ by an $\mathcal{O}_K$-equivalent model, we may assume $\phi_{45} = 0$. In the presence of the stronger condition that P is nonregular, we may further arrange that the coefficient of x_1 in Φ_{45} is divisible by π^2. Taking $A = \operatorname{Diag}(1, 1, 1, \pi^{-1}, \pi^{-1})$ and $\mu = 1$ now preserves the level.

5. Weights and slopes

In this section, we complete the proof of Theorem 3.1.

Definition 5.1. (i) The set of *weights* is

$$\mathscr{W} = \left\{ (r, s) \in \mathbb{Z}^5 \times \mathbb{Z}^5 \;\middle|\; \begin{array}{c} r_1 \leq r_2 \leq \cdots \leq r_5,\ s_1 \leq s_2 \leq \cdots \leq s_5, \\ 2\sum_{i=1}^5 r_i = 1 + \sum_{i=1}^5 s_i \end{array} \right\}.$$

(ii) A *weight for* $\Phi \in X_5(\mathcal{O}_K)$ is $(r, s) \in \mathscr{W}$ such that the model

$$[\operatorname{Diag}(\pi^{-r_1}, \ldots, \pi^{-r_5}), \operatorname{Diag}(\pi^{s_1}, \ldots, \pi^{s_5})]\Phi \tag{7}$$

has coefficients in $\mathcal{O}_K$.

(iii) Let $w = (r, s)$ and $w' = (r', s')$ be weights. Then w *dominates* w' if

$$\max(r_i + r_j - s_k, 0) \geq \max(r'_i + r'_j - s'_k, 0)$$

for all $1 \leq i < j \leq 5$ and $1 \leq k \leq 5$.

Let $\mathbf{1} = (1, 1, \ldots, 1)$. Then $\lambda \in \mathbb{Z}$ acts on $\mathscr{W}$ as $(r, s) \mapsto (r + \lambda\mathbf{1}, s + 2\lambda\mathbf{1})$. Since weights in the same $\mathbb{Z}$-orbit determine the same transformation (7), we may regard such weights as equivalent.

Lemma 5.2. *Let $\Phi \in X_5(\mathcal{O}_K)$ be an integral genus-1 model.*

(i) *If Φ is nonminimal, then it is $\mathcal{O}_K$-equivalent to a model with a weight.*

(ii) *If Φ has weight w and w dominates w', then Φ has weight w'.*

Proof. (i) By hypothesis, there exist $A, B \in \mathrm{GL}_5(K)$ with $[A, B]\Phi$ integral and $2v(\det A) + v(\det B) = -1$. We put A and B in Smith normal form.

(ii) Let $\Phi = (\Phi_{ij})$ with $\Phi_{ij} = \sum_k a_{ijk}x_k$. Then Φ has weight (r, s) if and only if $v(a_{ijk}) \geq \max(r_i + r_j - s_k, 0)$ for all $1 \leq i < j \leq 5$ and $1 \leq k \leq 5$. □

Lemma 5.3. *Let $\Phi \in X_5(\mathcal{O}_K)$ have weight $(r, s) \in \mathcal{W}$ with either $r_1 + r_4 > s_1$ or $r_2 + r_3 > s_1$. Then $P = (1 : 0 : \cdots : 0) \in \mathcal{C}_\phi$ is a singular point. Moreover, if $s_1 < s_3$, then P is nonregular (as a point on the $\mathcal{O}_K$-scheme $\mathcal{C}_\Phi$).*

Proof. We write $\phi = \sum x_i M_i$. If $r_1 + r_4 > s_1$, then the only nonzero entries of M_1 are in the top left 3×3 submatrix. If $r_2 + r_3 > s_1$, then the only nonzero entries of M_1 are in the first row and column. In both cases, $\operatorname{rank} M_1 \leq 2$, and so $P \in \mathcal{C}_\phi$. If $M_1 = 0$, then P is singular (and nonregular). So we may assume $M_1 \neq 0$. We are free to multiply rows of Φ by units in $\mathcal{O}_K$ and to subtract $\mathcal{O}_K$-multiples of later rows from earlier rows (it being understood that we also make the corresponding column operations). In particular, these operations do not upset our hypothesis that Φ has weight (r, s). Let E_{ij} be the 5×5 matrix with a 1 in the (i, j)-place and 0s elsewhere. By row and column operations, we reduce to the case $M_1 = E_{ij} - E_{ji}$, where $(i, j) \in \{(1, 2), (1, 3), (1, 4), (1, 5), (2, 3)\}$. Let $a < b < c$ be chosen such that $\{i, j, a, b, c\} = \{1, \ldots, 5\}$. Since $r_i + r_j \leq s_1 \leq s_2$, it follows by the definition of $\mathcal{W}$ that

$$s_3 + s_4 + s_5 < (r_a + r_b) + (r_a + r_c) + (r_b + r_c).$$

Therefore, at least one of the following three inequalities holds:

$$\begin{aligned} s_3 < r_a + r_b &\implies \phi_{ab}, \phi_{ac}, \phi_{bc} \in \langle x_4, x_5 \rangle, \\ s_4 < r_a + r_c &\implies \phi_{ac}, \phi_{bc} \in \langle x_5 \rangle, \\ s_5 < r_b + r_c &\implies \phi_{bc} = 0. \end{aligned}$$

Since the tangent space at P is $\{\phi_{ab} = \phi_{ac} = \phi_{bc} = 0\}$, it follows that $P \in \mathcal{C}_\phi$ is a singular point.

If $s_1 < s_3$, then the same argument shows there is some $\mathcal{O}_K$-linear combination of Φ_{ab}, Φ_{ac} and Φ_{bc} (with not all coefficients in $\pi\mathcal{O}_K$) that not only vanishes mod π but whose coefficient of x_1 vanishes mod π^2. Hence, P is nonregular. □

Lemma 5.4. *Let $(r, s) \in \mathcal{W}$ be a weight with $r_1 + r_4 \le s_1$ and $r_2 + r_3 \le s_1$. Then (r, s) dominates one of the weights $w_1, \dots, w_7$ in the following table*:

	r_1	r_2	r_3	r_4	r_5	s_1	s_2	s_3	s_4	s_5
w_1	0	0	0	0	1	0	0	0	0	1
w_2	0	0	1	1	1	1	1	1	1	1
w_3	0	0	1	1	2	1	1	1	2	2
w_4	0	1	1	2	2	2	2	2	2	3
w_5	0	1	1	2	3	2	2	2	3	4
w_6	0	1	1	2	3	2	2	3	3	3
w_7	0	1	2	3	4	3	3	4	4	5

Proof. We checked the lemma by writing a computer program using the simplex algorithm. See the proof of Lemma 6.1 for details. □

Definition 5.5. The *slope* of $\Phi \in X_5(\mathbb{O}_K)$ is the least possible value of $v(\det B)$ for $B \in \mathrm{GL}_5(K)$ a matrix with entries in $\mathbb{O}_K$ for which there exist $A \in \mathrm{GL}_5(K)$ and $\mu \in K^\times$ such that $[A, \mu B]\Phi$ is an integral model of smaller level.

We now complete the proof of Theorem 3.1. Since $\Phi \in X_5(\mathbb{O}_K)$ is nonminimal, it has a slope σ, say. Lemma 3.3(i) shows that if $\sigma = 0$, then Φ is nonsaturated. So we may assume $\sigma > 0$. By Lemma 5.2 (and Corollary 3.5), we may replace Φ by an $\mathbb{O}_K$-equivalent model with a weight, say (r, s). Moreover, we may assume the weight realises the slope, i.e., $\sigma = \sum_{i=1}^{5}(s_i - s_1)$.

Suppose that either $r_1 + r_4 > s_1$ or $r_2 + r_3 > s_1$. Since $\sigma > 0$, there exists $1 \le m \le 4$ such that $s_1 = \cdots = s_m < s_{m+1}$. Lemma 5.3 shows (by first making unimodular transformations involving only $x_1, \dots, x_m$) that

$$\{x_{m+1} = \cdots = x_5 = 0\} \subset \operatorname{Sing} \mathscr{C}_\phi. \tag{8}$$

Moreover, if $m = 1$, then the point we have constructed is nonregular. (This is needed to complete the proof of Theorem 3.1(ii) at the end of Section 4.)

Regardless of whether we have equality in (8), it follows that if the level is preserved, then the slope is decreased. So after finitely many iterations, Φ is either nonsaturated or has weight (r, s) with $r_1 + r_4 \le s_1$ and $r_2 + r_3 \le s_1$. In this last case, Lemmas 5.2 and 5.4 show that Φ has weight w for some $w \in \{w_1, \dots, w_7\}$. If $w \in \{w_1, w_2, w_6\}$, then Φ is nonsaturated. If $w \in \{w_5, w_7\}$, then Φ is $\mathbb{O}_K$-equivalent to a model with weight w_3. (This is achieved by a unimodular transformation involving only the second and third rows and columns, respectively a unimodular transformation involving only x_3 and x_4.) Finally, if $w \in \{w_3, w_4\}$, then Φ is $\mathbb{O}_K$-equivalent to a model of the form considered in Lemma 4.3(ii).

6. The number of iterations

We have shown that if we start with a nonminimal model, then iterating the procedure in Theorem 3.1(ii) eventually gives a nonsaturated model or decreases the level. In this section, we show that the maximum number of iterations required is 5. (In our Magma implementation, we count the use of Theorem 3.2 to decrease the level of a nonsaturated model as a further iteration. With this convention, the maximum number of iterations is 6.)

Lemma 6.1. *Let* $(r, s) \in \mathcal{W}$ *be a weight. Then* (r, s) *dominates one of the weights* $w_1, \dots, w_{29}$ *in the following table.* (*The weights in* Lemma 5.4 *appear with new numberings. We have marked these weights in bold.*)

	r_1	r_2	r_3	r_4	r_5	s_1	s_2	s_3	s_4	s_5	λ_ν
w_1	0	0	0	0	0	−1	0	0	0	0	1
$\boldsymbol{w_2}$	0	0	0	0	1	0	0	0	0	1	1
$\boldsymbol{w_3}$	0	0	1	1	1	1	1	1	1	1	1
w_4	0	1	1	1	1	1	1	1	2	2	1
w_5	0	0	0	1	1	0	0	1	1	1	3
w_6	0	0	0	1	1	0	0	0	1	2	3
w_7	0	0	1	1	1	0	0	1	2	2	3
w_8	0	0	1	1	1	0	1	1	1	2	3
$\boldsymbol{w_9}$	0	1	1	2	2	2	2	2	2	3	3
$\boldsymbol{w_{10}}$	0	0	1	1	2	1	1	1	2	2	4
w_{11}	0	0	1	1	2	0	0	2	2	3	5
w_{12}	0	0	1	1	2	0	1	2	2	2	8
w_{13}	0	0	1	1	2	0	1	1	2	3	8
w_{14}	0	1	1	1	2	1	2	2	2	2	4
w_{15}	0	1	1	1	2	1	1	2	2	3	4
w_{16}	0	1	1	2	2	1	2	2	3	3	7
w_{17}	0	1	1	2	2	1	2	2	2	4	6
w_{18}	0	1	1	2	2	1	1	2	3	4	7
$\boldsymbol{w_{19}}$	0	1	1	2	3	2	2	3	3	3	6
$\boldsymbol{w_{20}}$	0	1	1	2	3	2	2	2	3	4	7
w_{21}	0	1	1	2	3	1	2	3	3	4	13
w_{22}	0	1	1	2	3	1	2	2	3	5	12
w_{23}	0	1	2	2	3	2	3	3	3	4	9
w_{24}	0	1	2	2	3	2	2	3	4	4	9
w_{25}	0	1	2	2	3	1	3	3	4	4	10
w_{26}	0	1	2	2	3	1	2	3	4	5	15
$\boldsymbol{w_{27}}$	0	1	2	3	4	3	3	4	4	5	12
w_{28}	0	1	2	3	4	2	3	4	5	5	20
w_{29}	0	1	2	3	4	1	3	4	5	6	22

Proof. We define a *standard inequality* to be an inequality of the form $r_i + r_j \leq s_k + m$, where $1 \leq i < j \leq 5$, $1 \leq k \leq 5$ and m is a nonnegative integer. The condition that $(r, s) \in \mathcal{W}$ does not dominate w_ν is equivalent to a list of λ_ν standard inequalities, at least one of which must hold, where λ_ν is as given in the table. For example, $(r, s) \not\geq w_1$ if and only if $r_1 + r_2 \leq s_1$, whereas $(r, s) \not\geq w_5$ if and only if $r_1 + r_4 \leq s_2$ or $r_4 + r_5 \leq s_2 + 1$ or $r_4 + r_5 \leq s_5$. (We have used the conditions $r_1 \leq \cdots \leq r_5$ and $s_1 \leq \cdots \leq s_5$ to remove redundant inequalities.)

We wrote a program using the simplex algorithm to maximise $\sum(2r_i - s_i)$ for $(r, s) \in \mathbb{R}^{10}$ subject to $0 \leq r_1 \leq \cdots \leq r_5$, $0 \leq s_1 \leq \cdots \leq s_5$ and a list of standard inequalities. Our program starts with the basic feasible solution $(r, s) = (0, 0)$. If there is a finite maximum and it is less than 1, then (by definition of $\mathcal{W}$) there are no weights satisfying these inequalities. If the maximum is 1, then we add the constraint $\sum(2r_i - s_i) = 1$. We then use the simplex algorithm to maximise each of the functions $r_i + r_j - s_k$ in turn. In the case of a finite maximum α, we obtain an additional standard inequality $r_i + r_j \leq s_k + \max(\lfloor \alpha \rfloor, 0)$. Then running our original program on the enlarged set of standard inequalities, we may still be able to show that $\sum(2r_i - s_i) < 1$.

After processing the inequalities coming from $w_1, \dots, w_\nu$ for $\nu = 1, \dots, 29$, the number of cases remaining were

$$1, 1, 1, 1, 3, 5, 8, 13, 16, 30, 31, 49, 58, 47, 60, \\ 64, 58, 53, 45, 36, 39, 34, 25, 15, 14, 10, 3, 1, 0.$$

The final 0 indicates that no cases remain, and this proves the lemma. The proof of Lemma 5.4 is similar but easier. □

If $\Phi \in X_5(\mathcal{O}_K)$ is nonminimal, then by Lemmas 5.2 and 6.1 it has slope at most 14. This already shows that the algorithm in Theorem 3.1(iii) takes at most fourteen iterations. The next lemma improves this bound to seven iterations.

Lemma 6.2. *If the procedure in Theorem 3.1(ii) returns a saturated model with the same level, then the slope goes down by at least* 2.

Proof. We revisit the proof of Theorem 3.1(iii) at the end of Section 5. If the slope goes down by only 1, then Sing $\mathcal{C}_\phi$ spans a hyperplane. If Sing $\mathcal{C}_\phi$ is a hyperplane, then the proof of Theorem 3.1(ii) at the end of Section 4 shows that the level is decreased. Otherwise, by Lemma 4.2 we may assume ϕ takes the form (5). We then follow the proof of Lemma 4.3(i) with $m = 4$. After applying the transformation suggested there, the second row of ϕ has at most one nonzero entry. This implies that Φ is nonsaturated. □

The next lemma will be used to show that only five iterations are required.

Lemma 6.3. *Let $\Phi \in X_5(\mathcal{O}_K)$ be nonminimal and of slope greater than* 10. *Then replacing Φ by an $\mathcal{O}_K$-equivalent model, we may assume it has weight w_{29} and the coefficient of x_k in Φ_{ij} is a unit for*

$$(i, j, k) \in \\ \{(1, 2, 1), (1, 4, 2), (1, 5, 3), (2, 3, 2), (2, 4, 3), (2, 5, 4), (3, 4, 4), (3, 5, 5)\}.$$

Proof. By Lemma 5.2, we know that Φ is $\mathcal{O}_K$-equivalent to a model with one of the twenty-nine weights listed in Lemma 6.1. For all but one of these weights (r, s), we have $\sum_{i=1}^{5}(s_i - s_1) \leq 10$. The remaining case is w_{29}. If one of the coefficients listed is not a unit, then Φ has weight w_ν for some $\nu \in \{1, 5, 13, 26, 16, 21, 8, 12\}$. □

We write $[j, \dots, 5]$ for a linear combination of $x_j, \dots, x_5$ and underline in cases where we know the coefficient is nonzero. If the slope is at most 10, then at most five iterations are needed. Thus, Lemma 6.3 shows that we can reduce to the case where $\Phi \in X_5(\mathcal{O}_K)$ has reduction $\phi \in X_5(k)$ of the form

$$\begin{pmatrix} 0 & [\underline{1},2,3,4,5] & [2,3,4,5] & [\underline{2},3,4,5] & [\underline{3},4,5] \\ & 0 & [\underline{2},3,4,5] & [\underline{3},4,5] & [\underline{4},5] \\ & & 0 & [\underline{4},5] & [\underline{5}] \\ & & & 0 & 0 \\ & & & & 0 \end{pmatrix}.$$

Let $\mathrm{Pf}(\phi) = (p_1, \dots, p_5)$. By considering the partial derivatives of p_1, p_2 and p_4 with respect to x_1, x_2 and x_3, we see that if $P = (x_1 : \dots : x_5) \in \operatorname{Sing} \mathscr{C}_\phi$, then $x_5 = 0$. Then since $P \in \mathscr{C}_\phi$, we have $x_4 = x_3 = x_2 = 0$. So $(1 : 0 : \dots : 0)$ is the unique singular point.

Our algorithm applies the transformation

$$[\mathrm{Diag}(1,1,1,\pi^{-1},\pi^{-1}), \mathrm{Diag}(1,\pi,\pi,\pi,\pi)].$$

The result is a model Φ with weight $w_{26} = (0,1,2,2,3;1,2,3,4,5)$ whose reduction ϕ takes the form

$$\begin{pmatrix} 0 & [\underline{1}] & 0 & [\underline{2},3,4,5] & [\underline{3},4,5] \\ & 0 & 0 & [\underline{3},4,5] & [\underline{4},5] \\ & & 0 & [\underline{4},5] & [\underline{5}] \\ & & & 0 & [5] \\ & & & & 0 \end{pmatrix}.$$

A calculation similar to that above shows that $\operatorname{Sing} \mathscr{C}_\phi = \{x_3 = x_4 = x_5 = 0\}$.

Our algorithm applies the transformation

$$[\mathrm{Diag}(\pi,1,1,1,1), \mathrm{Diag}(\pi^{-1},\pi^{-1},1,1,1)].$$

The result is a model Φ with weight $w_{13} = (0,0,1,1,2;0,1,1,2,3)$ whose reduction ϕ takes the form

$$\begin{pmatrix} 0 & [\underline{1}] & 0 & [\underline{2}] & 0 \\ & 0 & [\underline{2}] & [2,\underline{3},4,5] & [\underline{4},5] \\ & & 0 & [\underline{4},5] & [\underline{5}] \\ & & & 0 & [5] \\ & & & & 0 \end{pmatrix}.$$

A calculation similar to that above shows that $\operatorname{Sing} \mathscr{C}_\phi = \{x_2 = x_4 = x_5 = 0\}$.

The next transformation $[\mathrm{Diag}(1,\pi,1,1,1), \mathrm{Diag}(\pi^{-1},1,\pi^{-1},1,1)]$ gives a model with weight $w_{15} = (0,1,1,1,2;1,1,2,2,3)$. So after three iterations, the slope is at most 4. It follows by Lemma 6.2 that at most five iterations are required.

Example 6.4. The simplest example of a genus-1 model satisfying the conditions of Lemma 6.3 is

$$\Phi = \begin{pmatrix} 0 & x_1 & 0 & x_2 & x_3 \\ & 0 & x_2 & x_3 & x_4 \\ & & 0 & x_4 & x_5 \\ & - & & 0 & 0 \\ & & & & 0 \end{pmatrix}.$$

We find that $\mathscr{C}_\Phi$ is a rational curve with a cusp parametrised by

$$(s:t) \mapsto (-s^5 : s^3t^2 : s^2t^3 : st^4 : t^5).$$

In this case, our algorithm takes the maximum of exactly five iterations to give a nonsaturated model. (The first three iterations are already described above.) Although the model in this example is singular, there are π-adically close nonsingular models that are treated in the same way by our algorithm.

7. Insoluble models

In this section, we prove a result converse to the strong minimisation theorem. This is analogous to the results for models of degrees $n = 2, 3, 4$ proved in [Cremona et al. 2010, Section 5]. As in Section 2, we work over a discrete valuation field K. We write K^{sh} for the strict Henselisation of K. (If K is a p-adic field, then this is the maximal unramified extension.)

Theorem 7.1. *If $\Phi \in X_5(K)$ is nonsingular and $\mathscr{C}_\Phi(K^{\mathrm{sh}}) = \emptyset$, then the minimal level is at least* 1 *and is equal to* 1 *if* $\mathrm{char}(k) \neq 5$.

As in Section 6, we write $[j, \dots, 5]$ for a linear combination of $x_j, \dots, x_5$ and underline in cases where we require the coefficient is nonzero.

Definition 7.2. A genus-1 model $\Phi \in X_5(\mathscr{O}_K)$ is *critical* if it has reduction mod π of the form

$$\begin{pmatrix} 0 & [\underline{1},2,3,4,5] & [\underline{2},3,4,5] & [\underline{3},4,5] & [\underline{4},5] \\ & 0 & [\underline{3},4,5] & [\underline{4},5] & [\underline{5}] \\ & & 0 & [\underline{5}] & 0 \\ & & & 0 & 0 \\ & & & & 0 \end{pmatrix}$$

and $\pi^{-1}\Phi_{35}$ and $\pi^{-1}\Phi_{45}$ have reductions mod π of the form $[\underline{1},2,3,4,5]$ and $[\underline{2},3,4,5]$.

We show in the next three lemmas that critical models are insoluble, minimal and of positive level. We then take $K = K^{\mathrm{sh}}$ and show that every insoluble model $\Phi \in X_5(K)$ is K-equivalent to a critical model.

Lemma 7.3. *Critical models are insoluble over K.*

Proof. Suppose $(x_1, \dots, x_5) \in K^5$ is a nonzero solution with $\min\{v(x_i)\} = 0$. By considering the 4×4 Pfaffians, we successively deduce $\pi \mid x_5, \pi \mid x_4, \dots, \pi \mid x_1$. In particular, $\min\{v(x_i)\} > 0$. This is the required contradiction. □

Since the definition of a critical model is unchanged by an unramified field extension, it follows immediately that critical models are insoluble over K^{sh}.

Lemma 7.4. *Critical models are minimal.*

Proof. It is easy to see that critical models are saturated. Moreover, every point on $\mathscr{C}_\phi = \{x_3 = x_4 = x_5 = 0\}$ is singular. Our algorithm (see Theorem 3.1) makes the transformation $[\mathrm{Diag}(\pi, 1, 1, 1, 1), \pi^{-1} \mathrm{Diag}(1, 1, \pi, \pi, \pi)]$. This gives an integral model of the same level that is $\mathbb{O}_K$-equivalent (by a pair of cyclic permutation matrices) to a critical model.

If Φ were nonminimal, then our algorithm would succeed in reducing the level. But on the contrary, when given a critical model our algorithm endlessly cycles between five $\mathbb{O}_K$-equivalence classes. □

The next lemma describes the possible levels of a critical model. To treat the cases $\mathrm{char}(k) = 2, 3$, we need to work with the a-invariants defined in Section 1. Although these are not $\mathrm{SL}_5 \times \mathrm{SL}_5$-invariant, if we make our choices of a_1, b_2 and a_3 so as not to introduce any new monomials when we lift to characteristic 0, then they will be invariant under all pairs of diagonal matrices. It follows by the proof of Lemma 1.2 that $a_1, \dots, a_6$ are isobaric, i.e.,

$$a_i \circ [\mathrm{Diag}(\lambda_1, \dots, \lambda_5), \mathrm{Diag}(\mu_1, \dots, \mu_5)] = \Big(\prod \lambda_\nu\Big)^{2i} \Big(\prod \mu_\nu\Big)^{i} a_i.$$

Lemma 7.5. *The level of a critical model is at least* 1 *and equal to* 1 *if* $\mathrm{char}(k) \neq 5$.

Proof. Applying

$$[\mathrm{Diag}(1, \pi^{-1/5}, \pi^{-2/5}, \pi^{-3/5}, \pi^{-4/5}), \mathrm{Diag}(\pi^{1/5}, \pi^{2/5}, \pi^{3/5}, \pi^{4/5}, \pi)]$$

to a critical model Φ gives a model with coefficients in $\mathbb{O}_K[\pi^{1/5}]$. It follows by the isobaric property that $\pi^i \mid a_i(\Phi)$ for all i. Hence, Φ has positive level.

The model with coefficients in $\mathbb{O}_K[\pi^{1/5}]$ has reduction

$$\begin{pmatrix} 0 & \lambda_1 x_1 & \mu_2 x_2 & -\mu_3 x_3 & -\lambda_4 x_4 \\ & 0 & \lambda_3 x_3 & \mu_4 x_4 & -\mu_5 x_5 \\ & & 0 & \lambda_5 x_5 & \mu_1 x_1 \\ & & & 0 & \lambda_2 x_2 \\ & & & & 0 \end{pmatrix}$$

for some $\lambda_1, \dots, \lambda_5, \mu_1 \dots, \mu_5 \in k^\times$. The invariants of this model are

$$\begin{aligned} c_4(\lambda, \mu) &= \lambda^4 + 228\lambda^3\mu + 494\lambda^2\mu^2 - 228\lambda\mu^3 + \mu^4, \\ c_6(\lambda, \mu) &= -\lambda^6 + 522\lambda^5\mu + 10005\lambda^4\mu^2 + 10005\lambda^2\mu^4 - 522\lambda\mu^5 - \mu^6 \end{aligned}$$

and $\Delta(\lambda, \mu) = \lambda\mu(\lambda^2 - 11\lambda\mu - \mu^2)^5$, where $\lambda = \prod \lambda_i$ and $\mu = \prod \mu_i$. Computing a resultant shows that if $\operatorname{char}(k) \neq 5$, then $c_4(\lambda, \mu)$ and $\Delta(\lambda, \mu)$ have no common roots. Therefore, the critical model Φ with which we started satisfies either $v(c_4(\Phi)) = 4$ or $v(\Delta(\Phi)) = 12$. It follows that Φ has level at most 1. □

Remark 7.6. The following example of a critical model of level 2 over $K = \mathbb{Q}_5$ shows that the hypothesis $\operatorname{char}(k) \neq 5$ cannot be removed from Lemma 7.5:

$$\begin{pmatrix} 0 & x_1 & x_2 & -x_3 & -x_4 \\ & 0 & x_3 & x_4 & -x_5 \\ & & 0 & x_5 & 35x_1 \\ & - & & 0 & 5x_2 \\ & & & & 0 \end{pmatrix}.$$

We recall that the minimal level is unchanged by an unramified field extension. Replacing K by K^{sh}, we may assume for the rest of this section that K is Henselian and its residue field k is algebraically closed. To complete the proof of Theorem 7.1, we show the following:

Theorem 7.7. *If $\Phi \in X_5(\mathcal{O}_K)$ is minimal and $\mathcal{C}_\Phi(K) = \varnothing$, then Φ is $\mathcal{O}_K$-equivalent to a critical model.*

We start the proof of Theorem 7.7 with the following lemma:

Lemma 7.8. *If $\Phi \in X_5(\mathcal{O}_K)$ is minimal, then its reduction $\phi \in X_5(k)$ has the following properties*:

(i) *the 4×4 Pfaffians of ϕ are linearly independent*,

(ii) *the subscheme $\mathcal{C}_\phi \subset \mathbb{P}^4$ does not contain a plane and*

(iii) *the entries of ϕ span the space of linear forms on $\mathbb{P}^4$*.

Proof. (i) This follows by Theorem 3.2 and Lemma 3.3(i).

(ii) Suppose $\mathcal{C}_\phi$ contains the plane $\{x_4 = x_5 = 0\}$. By Lemma 3.4, we may assume the reduction mod π of $[I_5, \operatorname{Diag}(1, 1, 1, \pi, \pi)]\Phi$ takes one of the two forms given in the lemma. We decrease the level by applying either $[\operatorname{Diag}(\pi, 1, 1, 1, 1), \pi^{-1} I_5]$ or $[\operatorname{Diag}(1, 1, 1, \pi^{-1}, \pi^{-1}), B]$, where B is chosen to preserve integrality.

(iii) This is clear, as we could otherwise decrease the level by dividing one of the coordinates by π. □

Lemma 7.9. *Let $\phi \in X_5(k)$ be a genus-1 model satisfying the conclusions of Lemma 7.8. Suppose that every point on $\mathscr{C}_\phi$ is singular. Then ϕ is k-equivalent to*

$$\begin{pmatrix} 0 & 0 & x_1 & x_3 & x_4 \\ & 0 & x_2 & x_4 & x_5 \\ & & 0 & x_5 & 0 \\ & - & & 0 & 0 \\ & & & & 0 \end{pmatrix} \quad or \quad \begin{pmatrix} 0 & x_1 & 0 & x_3 & x_4 \\ & 0 & x_2 & x_4 & x_5 \\ & & 0 & x_5 & 0 \\ & - & & 0 & 0 \\ & & & & 0 \end{pmatrix} \quad or \quad \begin{pmatrix} 0 & x_1 & x_2 & x_3 & x_4 \\ & 0 & x_3 & x_4 & x_5 \\ & & 0 & x_5 & 0 \\ & - & & 0 & 0 \\ & & & & 0 \end{pmatrix}.$$

Our proof of Lemma 7.9 uses the following classification of degenerations of the twisted cubic. (Only the last sentence of the statement is needed.)

Lemma 7.10. *Let ψ be a 3×2 matrix of linear forms in $R = k[x_1, \dots, x_4]$. Suppose the 2×2 minors of ψ are linearly independent and no linear combination of them has rank 1. Then ψ is $\mathrm{GL}_2 \times \mathrm{GL}_3 \times \mathrm{GL}_4$-equivalent to one of the following*:

$$\begin{pmatrix} x_1 & x_2 \\ x_2 & x_3 \\ x_3 & x_4 \end{pmatrix}, \quad \begin{pmatrix} x_1 & x_2 \\ x_2 & x_3 \\ x_4 & 0 \end{pmatrix}, \quad \begin{pmatrix} x_1 & x_2 \\ 0 & x_3 \\ x_4 & 0 \end{pmatrix} \quad or \quad \begin{pmatrix} x_1 & 0 \\ x_2 & x_2 \\ 0 & x_3 \end{pmatrix}. \tag{9}$$

In particular, the locus of smooth points on $\Gamma = \{\operatorname{rank} \psi \le 1\} \subset \mathbb{P}^3$ spans $\mathbb{P}^3$.

Proof. We may realise Γ as the intersection of the image of the Segre embedding $\mathbb{P}^1 \times \mathbb{P}^2 \to \mathbb{P}^5$ with a linear subspace $\mathbb{P}^3$. So every component of Γ has dimension at least 1. If every component has dimension 1, then by the Buchsbaum–Eisenbud acyclicity criterion, there is a minimal free resolution

$$0 \to R(-3)^2 \xrightarrow{\psi} R(-2)^3 \xrightarrow{M} R, \tag{10}$$

where M is the vector of 2×2 minors of ψ. If in addition $\dim T_P\Gamma = 1$ for every $P \in \Gamma$, then by an argument using Serre's criterion [Eisenbud 1995, Section 18.3], the ideal in R generated by the 2×2 minors of ψ is a prime ideal. By (10), the Hilbert polynomial is

$$h(t) = \binom{t+3}{3} - 3\binom{t+1}{3} + 2\binom{t}{3} = 3t + 1.$$

Therefore, Γ is a twisted cubic and ψ is equivalent to the first of the matrices in (9).

In all other cases, $\dim T_P\Gamma > 1$ for some $P \in \Gamma$. First suppose $\operatorname{rank} \psi(P) = 1$. Moving P to $(1:0:0:0)$, we may suppose

$$\psi = \begin{pmatrix} x_1 & \alpha \\ \delta & \beta \\ \gamma & 0 \end{pmatrix},$$

where α, β, γ and δ are linear forms in x_2, x_3, x_4. Our hypotheses on the 2×2 minors ensure that α, β and γ are linearly independent; say they are x_2, x_3 and x_4.

By row and column operations (and a substitution for x_1), we may assume δ is a multiple of x_2. This gives the second and third cases in (9).

Now suppose $\operatorname{rank} \psi(P) = 0$. Let $Q \in \Gamma$ be any other point. If $\operatorname{rank} \psi(Q) = 0$, then the 2×2 minors are binary quadratic forms, and so some linear combination has rank 1. Therefore, $\operatorname{rank} \psi(Q) = 1$. If $\dim T_Q\Gamma > 1$, then our earlier analysis applies (and in fact gives a contradiction). Otherwise, we may assume

$$\psi = \begin{pmatrix} x_1 & 0 \\ \alpha & x_2 \\ \beta & x_3 \end{pmatrix},$$

where α and β are linear forms in x_2, x_3. (The 0 in the top right has been cleared by row operations.) Since $\alpha x_3 - \beta x_2$ is a rank-2 quadratic form in x_2, x_3, we can make a change of coordinates so that $\Gamma = \{x_1x_2 = x_1x_3 = x_2x_3 = 0\}$. Then ψ is equivalent to the last of the matrices in (9).

For the final statement, we note that the four cases correspond geometrically to (i) a twisted cubic, (ii) a conic and a line, (iii) three nonconcurrent lines and (iv) three concurrent lines. In each case, Γ spans $\mathbb{P}^3$, and the only singular points are the points where the components meet. □

Proof of Lemma 7.9. Let $P \in \mathscr{C}_\phi$ be a singular point. Moving P to $(1:0:0:0:0)$, we may assume ϕ takes the form

$$\begin{pmatrix} 0 & x_1 & \ell_2 & \alpha_1 & \beta_1 \\ & 0 & \ell_3 & \alpha_2 & \beta_2 \\ & & 0 & \alpha_3 & \beta_3 \\ & - & & 0 & 0 \\ & & & & 0 \end{pmatrix},$$

where ℓ_i, α_i and β_i are linear forms in $x_2, \dots, x_5$. Let ψ be the top right 3×2 submatrix, and let $\Gamma \subset \mathbb{P}^3$ be the curve defined by its 2×2 minors. Since the 2×2 minors of ψ are a subset of the 4×4 Pfaffians of ϕ, they are linearly independent. In particular, α_3 and β_3 cannot both vanish identically. Without loss of generality, α_3 is nonzero.

Suppose no linear combination of the 2×2 minors of ψ has rank 1. Then by Lemma 7.10, there is a smooth point $Q = (x_2 : x_3 : x_4 : x_5)$ on Γ with $\alpha_3(Q) \neq 0$. Solving for x_1 gives a smooth point $(x_1 : x_2 : \cdots : x_5)$ on $\mathscr{C}_\phi$. This is a contradiction. Therefore, some linear combination of the 2×2 minors of ψ has rank 1. It is then easy to see that ϕ is k-equivalent to a model of the form (5).

By properties (i) and (ii), η and x_5 are linearly independent, and γ, δ and x_5 are linearly independent. However, if η, γ, δ and x_5 were linearly independent, then taking them to be $x_2, \dots, x_5$ would give that $(0:1:0:0:0)$ is a smooth point on $\mathscr{C}_\phi$. By row and column operations, we may therefore suppose $\eta = \delta$ $(= x_4$, say).

By property (ii), β, x_4 and x_5 are linearly independent, and γ, x_4 and x_5 are linearly independent. By row and column operations (and substitutions for the x_i), we may suppose $\beta = x_3$ and $\gamma = x_2$ or x_3. If $\gamma = x_2$, then by further row and column operations (and substitutions for the x_i), we may suppose α is a multiple of x_1. The lemma now follows using property (iii). □

Proof of Theorem 7.7. Since K is Henselian, any smooth point on $\mathscr{C}_\phi$ lifts to a K-point on $\mathscr{C}_\Phi$. So we may assume ϕ takes one of the three forms in Lemma 7.9. In the first two cases, ϕ defines a pair of concurrent lines with multiplicities 2 and 3. (These cases may be distinguished by the dimension of the tangent space at the point of intersection.) In the third case, it defines a line with multiplicity 5.

We apply the transformation $[\mathrm{Diag}(1, 1, 1, 1, \pi^{-1}), \mathrm{Diag}(1, 1, 1, \pi, \pi)]$. This gives an integral model of the same level. So the reduction must again be k-equivalent to one of the three models in Lemma 7.9. We tidy up by an $\mathcal{O}_K$-equivalence that cyclically permutes the rows and columns and makes substitutions for x_4 and x_5. The reduction $\phi \in X_5(k)$ now takes the form

$$\begin{pmatrix} 0 & x_4 & x_5 & \alpha & \beta \\ & 0 & 0 & x_1 & x_3 \\ & & 0 & x_2 & 0 \\ & - & & 0 & 0 \\ & & & & 0 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0 & x_4 & x_5 & \alpha & \beta \\ & 0 & x_1 & 0 & x_3 \\ & & 0 & x_2 & 0 \\ & - & & 0 & 0 \\ & & & & 0 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0 & x_4 & x_5 & \alpha & \beta \\ & 0 & x_1 & x_2 & x_3 \\ & & 0 & x_3 & 0 \\ & - & & 0 & 0 \\ & & & & 0 \end{pmatrix},$$

where α and β are linear forms in x_1, x_2, x_3.

In the first case, $(0:0:0:1:0)$ is a point with tangent space of dimension 3, and $\mathscr{C}_\phi$ contains points not on the line $\{x_1 = x_2 = x_3 = 0\}$. So the transformation has moved us to the second case.

In the second case, we obtain a contradiction as follows. If $\alpha = x_1 + \lambda x_2 + \mu x_3$, then adding μ times the fifth row/column to the third row/column and making substitutions for x_1 and x_5, we may assume $\mu = 0$. Then $(0:0:1:0:0)$ is a smooth point on $\mathscr{C}_\phi$. Likewise, if $\beta = x_1 + \lambda x_2 + \mu x_3$, then subtracting λ times the fourth row/column from the second row/column and making substitutions for x_1 and x_4, we may assume $\lambda = 0$. Then $(0:1:0:0:0)$ is a smooth point on $\mathscr{C}_\phi$. We are forced to the conclusion that neither α nor β involves x_1. But then $\mathscr{C}_\phi$ contains the plane $\{x_2 = x_3 = 0\}$, and by Lemma 7.8, this contradicts that Φ is minimal.

In the third case, we show that if the transformation above brings us back to the third case, then the original model is critical. If $\beta = x_1 + \lambda x_2 + \mu x_3$, then adding λ times the fourth row/column to the third row/column and making substitutions for x_1 and x_5, we may assume $\lambda = 0$. Then $\mathscr{C}_\phi$ contains the lines $\{x_1 = x_2 = x_3 = 0\}$ and $\{x_1 = x_3 = x_5 = 0\}$. So if the transformation returns us to third case, then β cannot involve x_1. Since $\mathscr{C}_\phi$ does not contain a plane and the 4×4 Pfaffians of ϕ

are linearly independent, α must involve x_1 and β must involve x_2. It follows by Definition 7.2 that the original model is $\mathcal{O}_K$-equivalent to a critical model. □

8. Reduction

Let $C \subset \mathbb{P}^4$ be a genus-1 normal curve of degree 5 defined over $\mathbb{Q}$. We may represent it by a nonsingular genus-1 model $\Phi \in X_5(\mathbb{Z})$. Running the algorithm in Section 3 locally at p for all primes p dividing the discriminant $\Delta(\Phi)$, we obtain a $\mathbb{Q}$-equivalent model (still with coefficients in $\mathbb{Z}$) whose discriminant is minimal in absolute value. If C is everywhere locally soluble, then this discriminant is the minimal discriminant of $E = \operatorname{Jac}(C)$. It remains to make a $\mathrm{GL}_5(\mathbb{Z})$ change of coordinates on $\mathbb{P}^4$ so that (after running the LLL algorithm on the space of five quadrics defining the curve) the coefficients (and not just the invariants) are small. The general method, described in [Cremona et al. 2010, Section 6], is to run the LLL algorithm on the Gram matrix for the (unique) Heisenberg invariant inner product. In this section, we outline how to compute this inner product in the case $n = 5$.

We recall that the Heisenberg group is the subgroup of $\mathrm{SL}_5(\mathbb{C})$ consisting of matrices M_T that describe the action of $T \in E[5]$ on $C \subset \mathbb{P}^4$ by translation. For $T \neq 0_E$, we call the five points in $\mathbb{P}^4$ fixed by M_T a *syzygetic* 5*-tuple*. It may be shown (for example, by adapting the proof of [Fisher 2012, Proposition 4.1] or using that $H^1(\mathbb{R}, E[5])$ is trivial) that Φ is $\mathrm{SL}_5(\mathbb{R}) \times \mathrm{SL}_5(\mathbb{R})$-equivalent to a model in Hesse form:

$$\begin{pmatrix} 0 & ax_0 & bx_1 & -bx_2 & -ax_3 \\ & 0 & ax_2 & bx_3 & -bx_4 \\ & & 0 & ax_4 & bx_0 \\ & - & & 0 & ax_1 \\ & & & & 0 \end{pmatrix}. \tag{11}$$

The invariants of this model are

$$\begin{aligned} c_4 &= a^{20} + 228a^{15}b^5 + 494a^{10}b^{10} - 228a^5b^{15} + b^{20}, \\ c_6 &= -a^{30} + 522a^{25}b^5 + 10005a^{20}b^{10} + 10005a^{10}b^{20} - 522a^5b^{25} - b^{30} \end{aligned}$$

and $\Delta = D^5$, where $D = ab(a^{10} - 11a^5b^5 - b^{10})$. For a model in Hesse form, the Heisenberg group is generated by $\operatorname{Diag}(1, \zeta, \dots, \zeta^4)$, where ζ is a primitive fifth root of unity, and a cyclic permutation matrix. Since these matrices are unitary, the Heisenberg invariant inner product is the standard inner product on $\mathbb{R}^5$.

The Hessian, introduced in [Fisher 2012], is an $\mathrm{SL}_5 \times \mathrm{SL}_5$-equivariant polynomial map $H : X_5 \to X_5$ with the property that the Hessian of (11) is of the same form with a and b replaced by $-\partial D/\partial b$ and $\partial D/\partial a$.

Theorem 8.1. *Let $\Phi \in X_5(\mathbb{C})$ be a nonsingular genus-1 model with invariants c_4 and c_6. Let A be the 3×5 matrix of quadrics such that $\lambda\Phi + \mu H(\Phi)$ has 4×4 Pfaffians*

$$\{ \lambda^2 A_{1i} + \lambda\mu A_{2i} + \mu^2 A_{3i} \mid i = 1, \dots, 5 \}.$$

Then $\mathcal{X} = \{\operatorname{rank} A \leq 1\} \subset \mathbb{P}^4$ consists of thirty points, and the syzygetic 5-tuples for $\mathcal{C}_\Phi$ are the fibres of the map $\alpha : \mathcal{X} \to \mathbb{P}^2$ given by the first (or indeed any) column of A. The image of α is the set of six points $(x : y : z) \in \mathbb{P}^2$ satisfying

$$\operatorname{rank} \begin{pmatrix} 0 & 5x & y & 6c_4x + z \\ x & y & 6c_4x - z & 8c_6x \\ y & -z & 8c_6x & 9c_4^2x \end{pmatrix} \leq 2. \tag{12}$$

Proof. It suffices to prove this for Φ in Hesse form. Then $\mathcal{X}$ is defined by

$$\operatorname{rank} \begin{pmatrix} x_0^2 & x_1^2 & x_2^2 & x_3^2 & x_4^2 \\ x_1x_4 & x_0x_2 & x_1x_3 & x_2x_4 & x_0x_3 \\ x_2x_3 & x_3x_4 & x_0x_4 & x_0x_1 & x_1x_2 \end{pmatrix} \leq 1 \tag{13}$$

and by [Barth et al. 1987, Proposition 1] is a set of thirty points. Evaluating the columns of (13) at these points, we obtain $(1:0:0)$ and $(1:\zeta^i:\zeta^{-i})$ for $i = 0, \dots, 4$. These are the points $(\xi : \eta : \nu) \in \mathbb{P}^2$ satisfying

$$\operatorname{rank} \begin{pmatrix} \xi & \eta & \nu & 0 \\ \nu & \xi & 0 & -\eta \\ 0 & 0 & \eta & \nu \end{pmatrix} \leq 2. \tag{14}$$

The remaining statements follow by direct calculation. In particular, our description (12) of the image of α is checked by making the substitution

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ab & b^2 & -a^2 \\ -a(\partial D/\partial a) + b(\partial D/\partial b) & -2b(\partial D/\partial a) & -2a(\partial D/\partial b) \\ -(\partial D/\partial b)(\partial D/\partial a) & (\partial D/\partial a)^2 & -(\partial D/\partial b)^2 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \\ \nu \end{pmatrix}.$$

We note that this change of coordinates and the matrix relating the 3×3 minors of (12) and (14) each have determinant a constant times a power of D. □

After computing the Hessian exactly (using the algorithm in [Fisher 2012, Section 11]), we use Theorem 8.1 to compute the syzygetic 5-tuples numerically. We then compute a Gram matrix for the Heisenberg invariant inner product as follows.

Proposition 8.2. *Let $C \subset \mathbb{P}^4$ be a genus-1 normal curve defined over $\mathbb{R}$.*

(i) *Exactly two of the syzygetic 5-tuples for C are defined over $\mathbb{R}$, say*

$$Y = \{ y_i y_j = 0 \mid i < j \} \subset \mathbb{P}^4 \quad \text{and} \quad Z = \{ z_i z_j = 0 \mid i < j \} \subset \mathbb{P}^4,$$

where $y_0, \dots, y_4$ and $z_0, \dots, z_4$ are linear forms in $\mathbb{C}[x_0, \dots, x_4]$.

(ii) *One of the* 5*-tuples in (i) has* 5 *real points, and the other has* 1 *real point. We may therefore arrange that* $y_0, \dots, y_4$ *and* z_0 *have real coefficients and that the pairs* z_1, z_4 *and* z_2, z_3 *are complex conjugates.*

(iii) *The Heisenberg invariant quadratic form spans the* 1*-dimensional real vector space*

$$\langle y_0^2, \dots, y_4^2 \rangle \cap \langle z_0^2, z_1z_4, z_2z_3 \rangle.$$

Proof. For C in Hesse form, we may take $y_i = x_i$ and $z_i = \sum_{j=0}^{4} \zeta^{ij} x_j$. In this case, the Heisenberg invariant quadratic form is $x_0^2 + \cdots + x_4^2$. □

9. Examples

Wuthrich [2001] constructed an element of order 5 in the Tate–Shafarevich group of the elliptic curve $E/\mathbb{Q}$ with Weierstrass equation

$$y^2 + xy + y = x^3 + x^2 - 3146x + 39049.$$

His example (see also [Fisher 2008, Section 9]) is defined by the 4×4 Pfaffians of

$$\begin{pmatrix} 0 & 310x_1 + 3x_2 + 162x_5 & -34x_1 - 5x_2 - 14x_5 & 10x_1 + 28x_4 + 16x_5 & 80x_1 - 32x_4 \\ & 0 & 6x_1 + 3x_2 + 2x_5 & -6x_1 + 7x_3 - 4x_4 & -14x_2 - 8x_3 \\ & & 0 & -x_3 & 2x_2 \\ & - & & 0 & -4x_1 \\ & & & & 0 \end{pmatrix}.$$

This model has discriminant $2^{132}\Delta_E$, where Δ_E is the minimal discriminant of E. In other words, the model is minimal at all primes except $p = 2$, where the level is 11. Minimisation and reduction suggest the change of coordinates

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \leftarrow \begin{pmatrix} 0 & 4 & -8 & 4 & 8 \\ 0 & 0 & 0 & 0 & 16 \\ 0 & -4 & 4 & 0 & 12 \\ 4 & 5 & -15 & 2 & 7 \\ 4 & -12 & 20 & -12 & -8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}$$

so that Wuthrich's example simplifies to

$$\Phi = \begin{pmatrix} 0 & x_2 + x_5 & -x_5 & -x_1 + x_2 & x_4 \\ & 0 & x_2 - x_3 + x_4 & x_1 + x_2 + x_3 - x_4 - x_5 & x_1 - x_2 - x_3 - x_4 - x_5 \\ & & 0 & x_1 - x_2 + 2x_3 - x_4 - x_5 & -x_2 - x_4 + x_5 \\ & - & & 0 & -x_3 - x_4 - 2x_5 \\ & & & & 0 \end{pmatrix}.$$

Our Magma function DoubleGenusOneModel, described in [Fisher 2013], computes a genus-1 model Φ' that represents twice the class of Φ in the 5-Selmer group. This model has entries

$$\Phi'_{12}=3534132778x_1+3583651940x_2-881947110x_3-323014538x_4+3395115339x_5,$$

$$\Phi'_{13}=5079379222x_1-2965539950x_2+11022202860x_3+12821590868x_4+640276471x_5,$$

$$\Phi'_{14}=-10098238458x_1-1274966110x_2-7873816170x_3-3456923272x_4-62353929x_5,$$

$$\Phi'_{15}=-12929747724x_1-6790511810x_2-11113305270x_3-15161763156x_4 +3241937033x_5,$$

$$\Phi'_{23}=-3381247332x_1+3810679160x_2+5919634530x_3+75326852x_4-1245085426x_5,$$

$$\Phi'_{24}=-3572860258x_1-5569480730x_2-953739600x_3-2138046812x_4-858145244x_5,$$

$$\Phi'_{25}=-4674149266x_1-943631490x_2-6754488160x_3+751535046x_4+117685567x_5,$$

$$\Phi'_{34}=-1851228934x_1+5238146110x_2-165588410x_3-2070411506x_4+678105748x_5,$$

$$\Phi'_{35}=-6992835070x_1-3744630360x_2+3130208220x_3-4523781310x_4+433739425x_5,$$

$$\Phi'_{45}=780078472x_1+2039763820x_2-450062790x_3-7105731722x_4+1625466111x_5.$$

The discriminant of Φ' is Δ_E^{49}. In particular, this model is nonminimal at all bad primes of E. Minimisation and reduction suggest the change of coordinates

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \leftarrow \begin{pmatrix} 92 & -36 & -153 & 129 & -131 \\ -54 & 84 & 5 & -206 & 139 \\ -63 & -174 & -60 & -79 & 53 \\ -111 & 106 & 206 & -115 & -162 \\ 314 & -466 & 158 & -328 & -12 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}$$

so that Φ' simplifies to

$$\begin{pmatrix} 0 & -x_4+x_5 & x_3-x_4+x_5 & x_2-x_5 & x_1-x_2+x_3-x_4-2x_5 \\ & 0 & x_1+x_5 & -x_2-x_3 & -x_2+x_5 \\ & & 0 & x_4 & -x_1 \\ & - & & 0 & x_1+x_4-x_5 \\ & & & & 0 \end{pmatrix}.$$

See also [Creutz and Miller 2012, Section 7.4] for an example where our algorithms are used to help find a Mordell–Weil generator of large height.

References

[Artin et al. 2005] M. Artin, F. Rodriguez-Villegas, and J. Tate, "On the Jacobians of plane cubics", *Adv. Math.* **198**:1 (2005), 366–382. MR 2006h:14043 Zbl 1092.14054

[Barth et al. 1987] W. Barth, K. Hulek, and R. Moore, "Shioda's modular surface $S(5)$ and the Horrocks–Mumford bundle", pp. 35–106 in *Vector bundles on algebraic varieties* (Bombay, 1984), Tata Inst. Fund. Res. Stud. Math. **11**, Oxford University Press, New York, 1987. MR 88j:14027 Zbl 0676.14010

[Bosma et al. 1997] W. Bosma, J. Cannon, and C. Playoust, "The Magma algebra system, I: The user language", *J. Symbolic Comput.* **24**:3-4 (1997), 235–265. MR 1484478 Zbl 0898.68039

[Cremona et al. 2010] J. E. Cremona, T. A. Fisher, and M. Stoll, "Minimisation and reduction of 2-, 3- and 4-coverings of elliptic curves", *Algebra Number Theory* **4**:6 (2010), 763–820. MR 2012c:11120 Zbl 1222.11073

[Creutz and Miller 2012] B. Creutz and R. L. Miller, "Second isogeny descents and the Birch and Swinnerton–Dyer conjectural formula", *J. Algebra* **372** (2012), 673–701. MR 2990032

[Eisenbud 1995] D. Eisenbud, *Commutative algebra: with a view toward algebraic geometry*, Graduate Texts in Mathematics **150**, Springer, New York, 1995. MR 97a:13001 Zbl 0819.13001

[Fisher 2008] T. A. Fisher, "The invariants of a genus one curve", *Proc. Lond. Math. Soc.* (3) **97**:3 (2008), 753–782. MR 2009j:11087 Zbl 1221.11135

[Fisher 2012] T. A. Fisher, "The Hessian of a genus one curve", *Proc. Lond. Math. Soc.* (3) **104**:3 (2012), 613–648. MR 2900238 Zbl 06021282

[Fisher 2013] T. A. Fisher, "Invariant theory for the elliptic normal quintic, I: Twists of X(5)", *Math. Ann.* **356**:2 (2013), 589–616. MR 3048608 Zbl 06181632

[Kraus 1989] A. Kraus, "Quelques remarques à propos des invariants c_4, c_6 et Δ d'une courbe elliptique", *Acta Arith.* **54**:1 (1989), 75–80. MR 90j:11045 Zbl 0628.14024

[Wuthrich 2001] C. Wuthrich, "Une quintique de genre 1 qui contredit le principe de Hasse", *Enseign. Math.* (2) **47**:1-2 (2001), 161–172. MR 2002c:14037 Zbl 1064.14019

Communicated by Karl Rubin
Received 2012-02-02 Accepted 2012-08-20

T.A.Fisher@dpmms.cam.ac.uk
Department of Pure Mathematics and Mathematical Statistics,
University of Cambridge, Wilberforce Road, Cambridge,
CB3 0WB, United Kingdom
http://www.dpmms.cam.ac.uk/~taf1000/

ALGEBRA AND NUMBER THEORY 7:5 (2013)
dx.doi.org/10.2140/ant.2013.7.1207

On binary cyclotomic polynomials

Étienne Fouvry

We study the number of nonzero coefficients of cyclotomic polynomials Φ_m, where m is the product of two distinct primes.

1. Presentation of the results

Let $m \geq 1$ be an integer, and let Φ_m be the cyclotomic polynomial defined by

$$\Phi_m(X) := \prod_{\substack{j=1 \\ (j,m)=1}}^{m} (X - \exp(2\pi i j/m)).$$

This monic polynomial belongs to $\mathbb{Z}[X]$, and its degree is equal to $\varphi(m)$, the Euler function of the integer m. Let $\theta(m)$ be the number of nonzero coefficients of Φ_m. Of course, $\theta(m)$ satisfies the trivial inequalities

$$2 \leq \theta(m) \leq \varphi(m) + 1,$$

which are optimal when one considers the case $m = 1$ or $m = p$, a prime number. In these cases, all of the coefficients of Φ_m are equal to 1.

We reserve the letters p and q for prime numbers. We call an integer m *binary* if it is of the form $m = pq$, with p and q distinct. Let $\mathcal{B} = \{6, 10, 14, 15, 21, \dots\}$ be the set of binary integers. For $m \in \mathcal{B}$, we say that the associated cyclotomic polynomial Φ_m is *binary*. The coefficients of the binary cyclotomic polynomial Φ_m are equal to 0, 1 or -1. Furthermore, in that particular case, the function $\theta(m)$ has an explicit expression in terms of p and q that can be exploited by analytic number theory. More precisely:

Proposition A. *Let $m = pq$ be a binary integer with $p \neq q$. Then we have*

$$\theta(m) = 2\bar{p}_q \bar{q}_p - 1, \tag{1}$$

where $\bar{p}_q$ is the unique integer satisfying

$$\bar{p}_q p \equiv 1 \bmod q \quad \textit{and} \quad 1 \leq \bar{p}_q < q$$

and $\bar{q}_p$ is defined similarly.

MSC2010: primary 11N13; secondary 11L20, 11C08, 11N36, 11L07, 11N05.
Keywords: cyclotomic polynomials, exponential sums, sieves.

For a proof of this basic result, see [Carlitz 1966, Theorem; Bzdęga 2012], and for an interesting characterization of the nonzero coefficients of Φ_{pq}, see [Lam and Leung 1996] for instance.

Recently Bzdęga [2012] started the study of the distribution function of the map

$$m \in \mathcal{B} \mapsto \theta(m).$$

Let us review his results. Let γ and x be real numbers satisfying $0 < \gamma < \frac{1}{2}$ and $x \geq 6$, and let $H_\gamma(x)$ be the counting function

$$H_\gamma(x) := \#\big\{\, m : m \in \mathcal{B}, m \leq x, \theta(m) \leq m^{\frac{1}{2}+\gamma} \,\big\} \tag{2}$$

(because of the inequality (12) below, it is useless to study H_γ for $\gamma \leq 0$). With these conventions, Bzdęga [2012, Theorem] proved the following:

Theorem A. *For every* $0 < \gamma < \frac{1}{2}$ *and every* $\epsilon > 0$, *there exist* $C(\gamma)$, $c(\epsilon, \gamma) > 0$ *and* $x_0 = x_0(\epsilon, \gamma)$ *such that for* $x \geq x_0$ *one has the inequalities*

$$c(\epsilon, \gamma) x^{\frac{1}{2}+\gamma-\epsilon} \leq H_\gamma(x) \leq C(\gamma) x^{\frac{1}{2}+\gamma}. \tag{3}$$

The idea of Bzdęga is to relate the integers $m = pq$ contributing to $H_\gamma(x)$ to the solutions of the equations

$$\ell q - np = 1, \tag{4}$$

where ℓ and n are integers satisfying some inequalities depending on p, q and γ. Write $t = np$. By (4) and by ingenious considerations, he is led to counting integers t such that t and $t+1$ both have a large prime factor. Appealing to a deep result of Hildebrand [1985] on p-stable subsets of integers, Bzdęga deduces the inequalities (3).

Our plan is to study (4) in the context of prime number theory and to get three different types of results according to the size of γ. These results suggest that this investigation becomes more and more intricate as γ decreases to 0. The first result gives an asymptotic formula when γ is large. Its proof is mainly based on bounds for Kloosterman–Ramanujan sums over primes (see Lemmas 2 and 3 below) and on the Bombieri–Vinogradov theorem (see Lemma 5).

Theorem 1. *For* $0 < \gamma < \frac{1}{2}$, *let*

$$C(\gamma) := \frac{2}{1+2\gamma} \log \frac{1+2\gamma}{1-2\gamma}. \tag{5}$$

Then for every $\gamma_0 > 0$, *uniformly for* γ *satisfying* $\frac{12}{25} + \gamma_0 \leq \gamma \leq \frac{1}{2} - \gamma_0$, *we have*

$$H_\gamma(x) \sim C(\gamma) \frac{x^{\frac{1}{2}+\gamma}}{\log x}$$

as $x \to \infty$.

The second result produces a universal upper bound for $H_\gamma(x)$ and is a rather direct consequence of the two-dimensional sieve (see Lemma 4).

Theorem 2. *For every* $\gamma_0 > 0$, *there exists* $C^+(\gamma_0)$ *such that, for every* γ *satisfying* $\gamma_0 \le \gamma \le \frac{1}{2} - \gamma_0$ *and for every* $x \ge 6$, *the following inequality holds*:

$$H_\gamma(x) \le C^+(\gamma_0) \frac{x^{\frac{1}{2}+\gamma}}{\log x}.$$

The last result is a lower bound when γ is large enough. Judging by the tools involved, it is certainly the deepest of our three results (see Lemma 7).

Theorem 3. *For every* $\gamma_0 > 0$, *there exist* $C^-(\gamma_0) > 0$ *and* $x(\gamma_0)$ *such that, for every* γ *satisfying* $\frac{15}{98} + \gamma_0 \le \gamma \le \frac{1}{2} - \gamma_0$ *and for every* $x \ge x(\gamma_0)$, *the following inequality holds*:

$$H_\gamma(x) \ge C^-(\gamma_0) \frac{x^{\frac{1}{2}+\gamma}}{\log x}.$$

When $\gamma = \frac{1}{2}$, $H_\gamma(x)$ counts the number of binary integers less than x, and this number is asymptotic to $x(\log\log x)(\log x)^{-1}$. This explains why the asymptotic formula in Theorem 1 cannot be uniform for $\gamma < \frac{1}{2}$. Finally, we postpone to Section 7 a discussion on a conjectural value of $H_\gamma(x)$.

2. Tools

2.1. *Notation.*

- We reserve the letters p and q for distinct prime numbers. For brevity, we replace the symbols $\bar{p}_q$ and $\bar{q}_p$ (defined in Proposition A) by $\bar{p}$ and $\bar{q}$.
- For $x \ge 1$, $\mathscr{L}$ denotes $\log 2x$, and $\xi := 1 + \mathscr{L}^{-1}$.
- For $N \ge 1$, the notation $n \sim N$ and $n \approx N$ respectively replaces the conditions $N < n \le 2N$ and $N < n \le \xi N$.
- For $N \ge 1$, the notation $n \asymp N$ means that n satisfies $c_1 N < n \le c_2 N$, where $0 < c_1 < c_2$ are absolute constants that are useless to specify.
- For $x \ge 1$, $\pi(x)$ is the number of primes less than x.
- For integers r and s, $\pi(x; r, s)$ is the number of ps less than x and congruent to s modulo r.
- For a real number t, $e(t)$ is the additive character $\exp(2\pi i t)$.
- The number of positive divisors of the integer n is denoted by $\tau(n)$.

2.2. ***Trigonometric sums.*** To detect the oscillations of the fractional part of the quotient $\bar{q}/p$, we shall appeal to the following well known lemma of Vinogradov, which is stated in different ways in the literature:

Lemma 1 [Vinogradov 1954, Lemma 12, page 32]. *Let* $r \geq 1$ *be an integer, and let* β *and* Δ *be real numbers satisfying* $0 < \Delta < \beta/2 < 1/4$. *Then there exist two functions* $\psi^{\pm}$ *with period* 1 *satisfying*

$$\begin{cases} \psi^{+}(t) = 1 & \text{for } 0 \leq t \leq \beta, \\ 0 \leq \psi^{+}(t) \leq 1 & \text{for } -\Delta \leq t \leq 0 \text{ or } \beta \leq t \leq \beta + \Delta, \\ \psi^{+}(t) = 0, & \text{if } t \pmod 1 \notin [-\Delta, \beta + \Delta], \end{cases} \tag{6}$$

$$\begin{cases} \psi^{-}(t) = 1 & \text{for } \Delta \leq t \leq \beta - \Delta, \\ 0 \leq \psi^{-}(t) \leq 1 & \text{for } 0 \leq t \leq \Delta \text{ or } \beta - \Delta \leq t \leq \beta, \\ \psi^{-}(t) = 0, & \text{if } t \pmod 1 \notin [0, \beta], \end{cases} \tag{7}$$

and

$$\psi^{\pm}(t) = \sum_{m=-\infty}^{\infty} c_m^{\pm} e(mt) \quad \text{for every real } t. \tag{8}$$

The coefficients $c_m^{\pm}$ *satisfy the equalities* $c_0^{\pm} = \beta \pm \Delta$ *and the inequalities*

$$|c_m^{\pm}| \leq 2 \min\Big\{\beta \pm \Delta, \frac{1}{\pi|m|}, \frac{1}{\pi|m|}\Big(\frac{r}{\pi|m|\Delta}\Big)^{r}\Big\}, \quad m \neq 0.$$

2.3. ***Kloosterman–Ramanujan sums over primes.*** For real $y \geq x \geq 1$ and for a a nonzero integer, we introduce the following trigonometric sum over primes:

$$S_p(a; x, y) := \sum_{x<q<y} e\Big(a\frac{\bar{q}}{p}\Big). \tag{9}$$

This sum differs from a classical Kloosterman–Ramanujan sum by the fact that the summation is restricted to prime values. We will benefit from oscillations of the function $q \mapsto e(a(\bar{q}/p))$ under the form of the two following lemmas extracted from [Fouvry and Shparlinski 2011]. The proofs of these two lemmas are based on the method of Garaev [2010]. For more general results on sums of this type, see [Fouvry and Michel 1998].

The first of these two lemmas considers the case where p is small compared with x and y.

Lemma 2 [Fouvry and Shparlinski 2011, Theorem 3.2]. *The bound*

$$S_p(a; x, y) \ll p^{-\frac{1}{2}} x \mathscr{L}^2 + p^{\frac{1}{4}} x^{\frac{4}{5}} \mathscr{L}^{\frac{3}{2}}$$

holds uniformly for every prime $p \geq 2$, *for every integer* a *not divisible by* p *and for every* $1 \leq x \leq y \leq 2x$.

This bound is interesting for $p \leq x^{\frac{4}{5}}$ only. We will have to deal with sums $S_p(a; x, y)$ for p slightly less than x. Still based on the method of Garaev, we have the following average bound of this sum, which is Theorem 3.3 of [Fouvry and Shparlinski 2011] for the choices $x_p = x$ and $x'_p = 2x$; the extension to the statement given is straightforward.

Lemma 3. *For every $\epsilon > 0$, the inequality*

$$\sum_{p \sim P} \max_{(a,p)=1} \left| S_p(a; x_p, x'_p) \right| \ll_\epsilon \left(x^{\frac{3}{5}} P^{\frac{13}{10}} + x^{\frac{5}{6}} P^{\frac{13}{12}} \right) P^\epsilon$$

holds uniformly for $P^{\frac{3}{2}} \geq x \geq 1$ and for any sequences of integers $(x_p)_{p \sim P}$ and $(x'_p)_{p \sim P}$ satisfying $x \leq x_p \leq x'_p \leq 2x$.

2.4. ***The two-dimensional sieve.*** The following lemma can be obtained by Brun's sieve and will be used in the proof of Theorem 2 since it produces an upper bound for the number of solutions to (4) with a large uniformity over ℓ and n:

Lemma 4 [Friedlander and Iwaniec 2010, Proposition 6.22]. *Let a, b and h be positive integers satisfying*

$$(a, b) = (ab, h) = 1 \quad \text{and} \quad 2 \mid abh.$$

Let $N_{abh}(x, z)$ be the number of pairs of positive integers m and n satisfying $am \leq x$, $(mn, h) = 1$, $am + h = bn$ and mn has no prime factors less than z. Then, for $z \geq 2$ and

$$x \geq \tau(h) abz (\log z)^4, \tag{10}$$

we have the inequality

$$N_{abh}(x, z) \ll \frac{hx}{\varphi(abh)} (\log z)^{-2},$$

where the implied constant is absolute.

2.5. ***The Bombieri–Vinogradov theorem.*** We now recall this cornerstone of current analytic number theory. It gives the average behavior of the function $\pi(x; r, s)$ and replaces the assumption of the Generalized Riemann Hypothesis for Dirichlet L-functions in many applications. Among the numerous possible references, we give here the version in [Iwaniec and Kowalski 2004, Theorem 17.1, (17.24)].

Lemma 5. *For every $A \geq 0$, there exists $C(A)$ such that, for every $x \geq 1$ and for $R := x^{\frac{1}{2}} \mathscr{L}^{-2A-6}$, one has the inequality*

$$\sum_{r \leq R} \max_{(s,r)=1} \left| \pi(x; r, s) - \frac{\pi(x)}{\varphi(r)} \right| \leq C(A) x \mathscr{L}^{-A-1}.$$

2.6. ***A variant of the Brun–Titchmarsh theorem.*** The proof of Theorem 3 heavily depends on lower bounds for the function $\pi(x; r, s)$ in cases that are not covered by Lemma 5, which means r is larger than $x^{\frac{1}{2}}$. We first recall the original statement of Mikawa [2001, Theorem].

Lemma 6. *Let $L > \frac{32}{17}$ and $A, B > 0$ be given. Let s be an integer and R be large with $0 < |s| \leq (\log R)^B$. Then, except possibly for $O(R(\log R)^{-A})$ integers r satisfying $(r, s) = 1$ and $r \sim R$, we have*

$$\inf\{\, p : p \equiv s \mod r \,\} \ll r^L,$$

where the implied constants depend only on A, B and L.

This result can be interpreted as an average version of Linnik's famous theorem concerning the least prime in an arithmetic progression. Actually, Mikawa's proof gives more. For instance, it instantly gives a lower bound with the correct order of magnitude for the function $\pi(r^L; r, s)$ for almost all r as above. Due to the value of L, this result can be viewed as a lower bound of the function $\pi(x; r, s)$ for almost all r coprime with s and slightly larger than $\sqrt{x}$. As far as we know, the first result of that type was due to Rousselet [1988] following techniques of Fouvry [1985], who was dealing with upper bounds of the function $\pi(x; r, s)$ (Brun–Titchmarsh theorem on average). The problem of giving both upper and lower bounds for $\pi(x; r, s)$ for almost r in the interval $[x^{\frac{1}{2}}, x^{\frac{1}{2}+\delta}]$, where δ is a small positive constant, was then treated in several remarkable papers [Bombieri et al. 1987; 1989; Baker and Harman 1996].

We give an improved version of Lemma 6 where we count primes in the interval $]x, 2x]$ with some uniformity over the congruence class $\bar{s} \bmod r$ (as above, $\bar{s}$ is the multiplicative inverse of $s \bmod r$). Such a generalization is necessary for our application and is possible by the structure of the proof of Lemma 6 based on bounds for Kloosterman sums on average (see [Habsieger and Sivak-Fischler 2010, Theorem 1.5] for another reference where this extension is made).

Lemma 7. *For every $K < \frac{17}{32}$, there exist $\alpha_K > 0$, $\beta_K > 0$ and x_K such that for every $x > x_K$, every R satisfying $2 \leq R < x^K$ and every s such that $1 \leq |s| \leq x^{\beta_K}$, the inequality*

$$\pi(2x; r, \bar{s}) - \pi(x; r, \bar{s}) \geq \alpha_K \frac{x}{\varphi(r) \log x},$$

holds for every $r \sim R$ coprime with s with at most $R(\log R)^{-2}$ exceptions.

Remark. Of course, in this lemma, we can suppose that the functions $K \mapsto \alpha_K$ and $K \mapsto \beta_K$ are decreasing and $K \mapsto x_K$ is increasing.

3. Basic transformations

3.1. ***Properties of the function θ.*** We first write the expression of $\theta(pq)$ given by Proposition A in an asymmetrical way. Actually, Bézout's identity and the inequalities $1 \le \bar{p} < q$ and $1 \le \bar{q} < p$ lead to the equality

$$p\bar{p} + q\bar{q} = 1 + pq,$$

which transforms (1) into

$$\theta(pq) = 2pq \cdot \frac{\bar{q}}{p}\Big(1 + \frac{1}{pq} - \frac{\bar{q}}{p}\Big) - 1. \tag{11}$$

Now suppose that $p < q$. From the trivial inequalities

$$\frac{1}{p} \le \frac{\bar{q}}{p} \le 1 - \frac{1}{p}$$

and from the properties of the function $t \mapsto t((1 + 1/pq) - t)$, we deduce

$$\theta(pq) \ge q > (pq)^{\frac{1}{2}}, \tag{12}$$

which implies that $H_\gamma(x) = 0$ for $\gamma \le 0$.

We now want to translate in an efficient manner the inequality

$$\theta(pq) \le (pq)^{\frac{1}{2}+\gamma}.$$

In order to control uniformity aspects, we will frequently assume that we have

$$\gamma_0 \le \gamma \le \tfrac{1}{2} - \gamma_0, \tag{13}$$

where γ_0 is a fixed positive number.

For $t \ge T(\gamma_0)$, let $0 < \theta_0(t) < 1 - \theta_1(t) < 1$ be the solutions of the polynomial equation of degree 2 in the unknown X

$$2tX\Big(1 + \frac{1}{t} - X\Big) - 1 = t^{\frac{1}{2}+\gamma}.$$

For simplicity, we omit in the sequel the dependency on the parameter γ.

Lemma 8. *We suppose that (13) holds. Let $m = pq$ be a binary integer with $p < q$ and $m \ge T(\gamma_0)$. Then*

$$\theta(m) \le m^{\frac{1}{2}+\gamma} \iff 0 < \frac{\bar{q}}{p} \le \theta_0(m) \text{ or } 1 - \theta_1(m) \le \frac{\bar{q}}{p} < 1. \tag{14}$$

The functions $t \mapsto \theta_0(t), \theta_1(t)$ are decreasing for $t > T(\gamma_0)$, are of $\mathscr{C}^\infty$-class and satisfy

$$\theta_0(t),\ \theta_1(t) = \frac{t^{\gamma-\frac{1}{2}}}{2} + O(t^{2\gamma-1}),$$

where the implied constant depends on γ_0 only.

Proof. The proof of (14) is easy; it is only a transcription of (11). Finally, the asymptotic behaviors of the functions $\theta_i(t)$ are consequences of the exact formula

$$\theta_0(t),\, 1-\theta_1(t) = \frac{1+\frac{1}{t}\mp\sqrt{\left(1+\frac{1}{t}\right)^2-2\frac{t^{1/2+\gamma}+1}{t}}}{2}. \qquad \square$$

3.2. ***Decomposition of $H_\gamma(x)$.*** We always suppose that (13) is true. Let $T(\gamma_0)$ be defined as in Lemma 8. We use (14) to split the set contributing to $H_\gamma(x)$

$$\left\{ (p,q) : p<q,\, T(\gamma_0) \le pq \le x,\, \theta(pq) \le (pq)^{\frac{1}{2}+\gamma} \right\}$$

into two disjoint subsets corresponding to $0<\bar{q}/p \le \theta_0(pq)$ or $1-\theta_1(pq) \le \bar{q}/p < 1$. Let $H_\gamma^0(x)$ and $H_\gamma^1(x)$ be the corresponding cardinalities, which give the equality

$$H_\gamma(x) = H_\gamma^0(x) + H_\gamma^1(x) + O(T(\gamma_0)). \tag{15}$$

We shall concentrate our study on the case of $H_\gamma^0(x)$ since the case of $H_\gamma^1(x)$ is quite similar because the functions θ_0 and θ_1 play the same role (see Lemma 8).

To control the order of magnitude of the variables p and q, we consider, for P, $Q \ge 2$ such that $PQ \ge T(\gamma_0)$, the counting functions

$$R_\gamma(P,Q) := \#\left\{ (p,q) : p<q,\, pq \le x,\, p \approx P,\, q \approx Q,\, 0 < \frac{\bar{q}}{p} \le \Theta_0 \right\}, \tag{16}$$

where

$$\Theta_0 = \theta_0(PQ). \tag{17}$$

Since the function θ_0 is decreasing, we obtain the inequality

$$H_\gamma^0(x) \le \sum_P \sum_Q R_\gamma(P,Q), \tag{18}$$

where the sum is over pairs (P, Q), where P and Q are of the form $2\cdot\xi^k$ for $k = 0, 1, 2, \ldots$ and satisfy the inequalities

$$T(\gamma_0) \le PQ \le x \quad \text{and} \quad P \le \xi Q. \tag{19}$$

Finally note that (12) implies that we can even restrict the summation to the cases

$$4(PQ)^{\frac{1}{2}+\gamma} \ge Q \tag{20}$$

since otherwise $R_\gamma(P,Q) = 0$. Combining (19) and (20), we deduce that P and Q satisfy the inequalities

$$P \le \xi Q \quad \text{and} \quad \kappa_0 Q^{\frac{1-2\gamma}{1+2\gamma}} \le P \le x Q^{-1} \quad \text{with } \kappa_0 = 4^{-\frac{2}{1+2\gamma}}. \tag{21}$$

The inequality (18) can be easily transformed into a lower bound on $H^0_\gamma(x)$ if one replaces Θ_0 by Θ_0' with $\Theta_0' := \theta_0(\xi^2 PQ)$ in the definition (16) of $R_\gamma(P, Q)$. We note that

$$\Theta_0' - \Theta_0 = O(\Theta_0 \mathscr{L}^{-1}), \tag{22}$$

as a result of Lemma 8 and the fineness of the cutting of the sum $H^0_\gamma(x)$ (see (18)).

4. Proof of Theorem 1

The first purpose of this section is to prove the following:

Proposition 1. *Let $\gamma_0 > 0$. Then uniformly for γ satisfying*

$$\tfrac{12}{25} + \gamma_0 \le \gamma \le \tfrac{1}{2} - \gamma_0 \tag{23}$$

and for (P, Q) satisfying the conditions (21), one has the equality

$$R_\gamma(P, Q) = \tfrac{1}{2}(PQ)^{\gamma - \frac{1}{2}}(1 + O(\mathscr{L}^{-1}))\Bigg(\sum_{\substack{p \approx P \\ pq \le x}} \sum_{\substack{q \approx Q \\ p < q}} 1\Bigg) + O(x^{\frac{1}{2}+\gamma}\mathscr{L}^{-6}) + O(Q\mathscr{L}^{-4}).$$

Our proof depends on the size of P compared with Q.

4.1. ***When P is small.*** Let $\mathscr{E}(p, \Theta_0)$ denote the set of congruence classes $s \bmod p$ such that $0 < \bar{s}/p \le \Theta_0$. Of course, $\bar{s}$ is the multiplicative inverse of $s \bmod p$. By the definition (17) and by Lemma 8, its cardinality satisfies

$$\#\mathscr{E}(p, \Theta_0) = (\tfrac{1}{2} + O(\mathscr{L}^{-1}))P^{\frac{1}{2}+\gamma}Q^{\gamma - \frac{1}{2}} + O(1). \tag{24}$$

Let

$$y_p := \max(Q, p) \quad \text{and} \quad z_p := \min(\xi Q, x/p). \tag{25}$$

With this definition, we have the equality

$$R_\gamma(P, Q) = \sum_{\substack{p \approx P \\ y_p \le z_p}} \sum_{s \in \mathscr{E}(p, \Theta_0)} \big(\pi(z_p; p, s) - \pi(y_p; p, s)\big). \tag{26}$$

For (P, Q) satisfying (21), the trivial estimate

$$\big(\pi(z_p; p, s) - \pi(y_p; p, s)\big) \le Q/p + 1 \ll Q/p$$

inserted in (26) gives the bound

$$R_\gamma(P, Q) \ll (PQ)^{\frac{1}{2}+\gamma} + Q \ll (PQ)^{\frac{1}{2}+\gamma} \tag{27}$$

by (20). Hence, for the proof of Proposition 1, we may add the extra condition

$$PQ \ge x\mathscr{L}^{-12}. \tag{28}$$

The equalities (24) and (26) and Lemma 5 allow us to improve (27) by

$$R_\gamma(P,Q)=\bigl[(\tfrac12+O(\mathscr{L}^{-1}))P^{\frac12+\gamma}Q^{\gamma-\frac12}+O(1)\bigr]\biggl(\sum_{\substack{p\approx P\\ pq\le x}}\sum_{\substack{q\approx Q\\ p<q}}\frac{1}{\varphi(p)}\biggr)$$
$$+O((PQ)^{\frac12+\gamma}\mathscr{L}^{-6})+O(Q\mathscr{L}^{-6})\quad(29)$$

provided

$$P\le Q^{\frac12}\mathscr{L}^{-100}.\qquad(30)$$

The contribution of the $O(1)$-term to the right-hand side of (29) is bounded by $Q\mathscr{L}^{-4}$, up to a multiplicative constant. Recalling the restriction (28), we see that the proof of Proposition 1 is complete in the particular case

$$P\le x^{\frac13}\mathscr{L}^{-100}.\qquad(31)$$

4.2. ***Medium values of P.*** We apply Lemma 1 with the choices

$$\beta=\Theta_0,\quad \Delta=\Theta_0\mathscr{L}^{-3},\quad r=4.$$

We then have the inequalities

$$\sum_{\substack{p\approx P\\ pq\le x}}\sum_{\substack{q\approx Q\\ p<q}}\psi^-\Bigl(\frac{\bar q}{p}\Bigr)\le R_\gamma(P,Q)\le\sum_{\substack{p\approx P\\ pq\le x}}\sum_{\substack{q\approx Q\\ p<q}}\psi^+\Bigl(\frac{\bar q}{p}\Bigr).\qquad(32)$$

We only study the upper bound of $R_\gamma(P,Q)$ in (32). We recall the definitions (9) and (25). We apply Lemma 1 (in a slightly weaker form) and decompose the sums according to the values of m and whether p and m are coprime. This gives

$$\sum_{\substack{p\approx P\\ pq\le x}}\sum_{\substack{q\approx Q\\ p<q}}\psi^+\Bigl(\frac{\bar q}{p}\Bigr)$$
$$\le(\beta+\Delta)\sum_{\substack{p\approx P\\ pq\le x}}\sum_{\substack{q\approx Q\\ p<q}}1$$
$$+2\sum_{p\approx P}\biggl\{\sum_{\substack{1\le|m|\le\Delta^{-1}\\ p\nmid m}}\frac{1}{\pi|m|}+\sum_{\substack{|m|>\Delta^{-1}\\ p\nmid m}}\frac{256}{\pi^5|m|^5\Delta^4}$$
$$+\sum_{\substack{1\le|m|\le\Delta^{-1}\\ p\mid m}}\frac{2}{\pi|m|}+\sum_{\substack{|m|>\Delta^{-1}\\ p\mid m}}\frac{256}{\pi^5|m|^5\Delta^4}\biggr\}\bigl|S_p\bigl(m;y_p,z_p\bigr)\bigr|.\qquad(33)$$

It remains to apply Lemma 2 when $p\nmid m$, or the trivial inequality $|S_p|\le Q$ otherwise, and to sum over m to obtain the inequality

$$\sum_{\substack{p\approx P\\ pq\le x}}\sum_{\substack{q\approx Q\\ p<q}}\psi^+\Big(\frac{\bar{q}}{p}\Big)\le(\beta+\Delta)\sum_{\substack{p\approx P\\ pq\le x}}\sum_{\substack{q\approx Q\\ p<q}}1$$
$$+O\bigg(\sum_{p\approx P}\big\{(\mathscr{L}+1)(p^{-\frac{1}{2}}Q+p^{\frac{1}{4}}Q^{\frac{4}{5}})\mathscr{L}^2+(p^{-1}\mathscr{L}+p^{-1})Q\big\}\bigg). \quad (34)$$

Using the upper bound $\sum_{p\approx P}1\ll_{\gamma_0}P\mathscr{L}^{-2}$, we see that the error term satisfies

$$\text{error term}\ll_{\gamma_0}\big(P^{\frac{1}{2}}Q+P^{\frac{5}{4}}Q^{\frac{4}{5}}\big)\mathscr{L}. \quad (35)$$

By Lemma 8 and (28), we have the equality

$$\Theta_0=\tfrac{1}{2}(PQ)^{\gamma-\frac{1}{2}}+O((PQ)^{2\gamma-1})=\tfrac{1}{2}(PQ)^{\gamma-\frac{1}{2}}(1+O(\mathscr{L}^{-3})),$$

which, combined with (32), (34) and (35) gives the inequality

$$R_\gamma(P,Q)$$
$$=\tfrac{1}{2}(PQ)^{\gamma-\frac{1}{2}}(1+O(\mathscr{L}^{-3}))\bigg(\sum_{\substack{p\approx P\\ pq\le x}}\sum_{\substack{q\approx Q\\ p<q}}1\bigg)+O\big((P^{\frac{1}{2}}Q+P^{\frac{5}{4}}Q^{\frac{4}{5}})\mathscr{L}\big). \quad (36)$$

Recalling the restrictions (21), we see that (36) implies Proposition 1 as soon as P satisfies the inequalities

$$P\ge x^{1-2\gamma}\mathscr{L}^{14}\quad\text{and}\quad P\le x^{\frac{20}{9}\gamma-\frac{2}{3}}\mathscr{L}^{-16}. \quad (37)$$

4.3. ***Large values of P.*** Actually, in (33) we may benefit from the summation over $p\approx P$ by appealing to Lemma 3 instead of Lemma 2. By the same technique as in Section 4.2, we arrive at the equality

$$R_\gamma(P,Q)=\Theta_0(1+O(\mathscr{L}^{-3}))\bigg(\sum_{\substack{p\approx P\\ pq\le x}}\sum_{\substack{q\approx Q\\ p<q}}1\bigg)+O_\epsilon\big((P^{\frac{13}{10}}Q^{\frac{3}{5}}+P^{\frac{13}{12}}Q^{\frac{5}{6}})x^\epsilon\big) \quad (38)$$

provided $P^{\frac{3}{2}}\ge Q$ and ϵ is an arbitrary positive number. Hence, by (21) and (28), we see that (38) implies Proposition 1 as soon as P satisfies the extra conditions

$$P\ge x^{\frac{2}{5}},\qquad P\le x^{\frac{10}{7}\gamma-\frac{1}{7}-2\epsilon}\quad\text{and}\quad P\le x^{4\gamma-\frac{4}{3}-5\epsilon}. \quad (39)$$

Suppose now that γ satisfies (23) and that P satisfies $1\le P\le 2\sqrt{x}$. Then we see that P satisfies at least one of the sets of conditions (31), (37) or (39). This completes the proof of Proposition 1. □

4.4. ***Conclusion of the proof of Theorem 1.*** We insert the expansion of $R_\gamma(P,Q)$ given in Proposition 1 in the right-hand side of (18) and sum over (P,Q) satisfying (21). Recall that the numbers P and Q are of the shape $2\cdot\xi^k$. We first consider the contribution of the term $O(Q\mathscr{L}^{-4})$. By (21), this contribution satisfies

$$\begin{aligned}
O(Q\mathscr{L}^{-4})\text{ term} &\ll \mathscr{L}^{-4}\sum_Q Q \sum_{\kappa_0 Q^{\frac{1-2\gamma}{1+2\gamma}}\le P<xQ^{-1}} 1\\
&\ll \mathscr{L}^{-3}\sum_{Q\le(\frac{x}{\kappa_0})^{\frac12+\gamma}} Q\Bigl(\log\Bigl(\frac{x}{\kappa_0}Q^{-\frac{2}{1+2\gamma}}\Bigr)+1\Bigr)\\
&\ll \mathscr{L}^{-3}\Bigl\{\sum_{Q\le(\frac{x}{\kappa_0})^{\frac12+\gamma}\mathscr{L}^{-1}} Q\mathscr{L}+\log\mathscr{L}\sum_{(\frac{x}{\kappa_0})^{\frac12+\gamma}\mathscr{L}^{-1}\le Q\le(\frac{x}{\kappa_0})^{\frac12+\gamma}} Q\Bigr\}\\
&\ll x^{\frac12+\gamma}\mathscr{L}^{-\frac32}.
\end{aligned}$$

Since the number of (P,Q) satisfying (21) is $O(\mathscr{L}^4)$, the contribution of the term $O(x^{\frac12+\gamma}\mathscr{L}^{-6})$ (coming from Proposition 1) is $O(x^{\frac12+\gamma}\mathscr{L}^{-2})$. From the above considerations, we deduce the inequality

$$H^0_\gamma(x)\le(\tfrac12+o(1))\sum_P\sum_Q(PQ)^{\gamma-\frac12}\Bigl(\sum_{\substack{p\approx P\\ pq\le x}}\sum_{\substack{q\approx Q\\ p<q}}1\Bigr)+O(x^{\frac12+\gamma}\mathscr{L}^{-2}),$$

where P and Q satisfy (21). We now want to drop the dissection parameters P and Q. To do so, we remark that $(PQ)^{\gamma-\frac12}=(1+o(1))(pq)^{\gamma-\frac12}$ for $p\approx P$ and $q\approx Q$. We gather the rectangles of summation $]P,\xi P]\times]Q,\xi Q]$ to deduce the inequality

$$H^0_\gamma(x)\le(\tfrac12+o(1))\Bigl(\sum\sum_{p<q\le x/p}(pq)^{\gamma-\frac12}\Bigr)+O(x^{\frac12+\gamma}\mathscr{L}^{-2}).\tag{40}$$

By the prime number theorem, we have

$$\sum\sum_{p<q\le x/p}(pq)^{\gamma-\frac12}\sim\int_{x^{\frac12}}^{x^{\frac12+\gamma}}\frac{y^{\gamma-\frac12}}{\log y}\,dy\int_3^{xy^{-1}}\frac{z^{\gamma-\frac12}}{\log z}\,dz\quad(x\to\infty).$$

Write $y:=x^u$ and $z:=x^v$ to deduce

$$\begin{aligned}
\sum\sum_{p<q\le x/p}(pq)^{\gamma-\frac12}&\sim\int_{\frac12}^{\frac12+\gamma}\frac{x^{u(\gamma+\frac12)}}{u}\,du\int_{\frac{\log 3}{\log x}}^{1-u}\frac{x^{v(\gamma+\frac12)}}{v}\,dv\\
&\sim\int_{\frac12}^{\frac12+\gamma}\frac{x^{u(\gamma+\frac12)}}{u}\cdot\frac{x^{(1-u)(\gamma+\frac12)}}{(1-u)\bigl(\gamma+\frac12\bigr)\log x}du\sim C(\gamma)\frac{x^{\frac12+\gamma}}{\log x},
\end{aligned}\tag{41}$$

where $C(\gamma)$ is defined in (5).

The study of $H^1_\gamma(x)$ defined in (15) is similar to the study of $H^0_\gamma(x)$. Combining (15), (40) and (41), we finally arrive at the inequality

$$H_\gamma(x) \le (1+o(1))C(\gamma)\frac{x^{\frac{1}{2}+\gamma}}{\log 2x}. \tag{42}$$

To produce a lower bound for $H^0_\gamma(x)$, we follow the idea presented at the end of Section 3.2, which consists of replacing the constant Θ_0 by Θ'_0 in the definition of $R_\gamma(P, Q)$. By (22), we also obtain the inequalities

$$H^0_\gamma(x),\, H^1_\gamma(x) \ge (1-o(1))\frac{C(\gamma)}{2}\cdot\frac{x^{\frac{1}{2}+\gamma}}{\log 2x}$$

as x tends to infinity. Summing these two inequalities, we arrive at

$$H_\gamma(x) \ge (1-o(1))C(\gamma)\frac{x^{\frac{1}{2}+\gamma}}{\log 2x}.$$

Combining with (42), this completes the proof of Theorem 1. □

5. Proof of Theorem 2

We still suppose that (13) is satisfied and that PQ is large enough, which means $PQ \ge T(\gamma_0)$, where $T(\gamma_0)$ is defined in Lemma 8. Since we are searching for an upper bound, it is useless to work with a very thin cutting up as in (16). So let

$$S^0_\gamma(P, Q) := \#\Big\{(p, q) : p \sim P, q \sim Q, p < q, 0 < \frac{\bar{q}}{p} \le \Theta_0\Big\}, \tag{43}$$

$$S^1_\gamma(P, Q) := \#\Big\{(p, q) : p \sim P, q \sim Q, p < q, 1-\Theta_1 < \frac{\bar{q}}{p} < 1\Big\}, \tag{44}$$

where Θ_0 is still defined by (17) and $\Theta_1 = \theta_1(PQ)$. We then have the inequality

$$H_\gamma(x) \le \sum_{(P,Q)}\sum S^0_\gamma(P, Q) + \sum_{(P,Q)}\sum S^1_\gamma(P, Q) + O(T(\gamma_0)), \tag{45}$$

where P and Q are powers of 2 and satisfy $P \le 2Q$ and $T(\gamma_0) \le PQ \le x$. We will focus our study on the case of $S^0_\gamma(P, Q)$.

Define

$$L := P^{\gamma+\frac{1}{2}}Q^{\gamma-\frac{1}{2}}. \tag{46}$$

If (p, q) contributes to $S^0_\gamma(P, Q)$, then we have the equality (4) for some ℓ satisfying $1 \le \ell \ll L$. Hence, we have the inequality

$$S^0_\gamma(P, Q) \le \sum_{1\le\ell\ll L}\ \sum_{n\asymp \ell Q/P} F(\ell, n, P, Q), \tag{47}$$

where

- the constants implicit in the symbols $\ll$ and $\asymp$ depend on γ_0 only and
- $F(\ell, n, P, Q)$ is the number of solutions of the equation $\ell q - np = 1$ in primes $p \sim P$ and $q \sim Q$.

By Lemma 4, we have the inequality

$$F(\ell, n, P, Q) \ll \frac{\ell Q}{\varphi(\ell n)} \cdot \log^{-2} z \tag{48}$$

provided $z \leq P^{\frac{1}{2}}$ and $\ell Q \geq \ell n z \log^4 z$. By the order of magnitude of the parameters, this last condition reduces to

$$P \gg \ell z \log^4 z.$$

However, since we have $\ell \ll L$, this inequality is satisfied as soon as

$$(PQ)^{\frac{1}{2}-\gamma} \gg z^2.$$

Choose $z := (PQ)^{\frac{1}{6}-\frac{\gamma}{3}}$. With this choice of z inserted in (48) and by (47), we obtain the inequality

$$S_\gamma^0(P, Q) \ll_{\gamma_0} \frac{Q}{\log^2(PQ)} \sum_{1 \leq \ell \ll L} \ell \sum_{n \asymp \ell Q/P} \frac{1}{\varphi(\ell n)}. \tag{49}$$

Recall the inequality $\varphi(\ell n) \geq \varphi(\ell)\varphi(n)$ and the bound $\sum_{t \sim T} \varphi^{-1}(t) \ll 1$, which is uniform in $T \geq 1$. Then summing over ℓ and n in (49), we deduce the inequality

$$S_\gamma^0(P, Q) \ll_{\gamma_0} LQ \log^{-2}(PQ) \ll_{\gamma_0} (PQ)^{\gamma+\frac{1}{2}} \log^{-2}(PQ).$$

This bound also holds for $S_\gamma^1(P, Q)$. Inserting this bound in (45) and summing over (P, Q) such that $PQ \leq x$, we conclude the proof of Theorem 2. □

6. Proof of Theorem 3

We now suppose that

$$\tfrac{15}{98} + \gamma_0 \leq \gamma \leq \tfrac{13}{27}$$

since the case where γ takes large values is covered by Theorem 1. Define also

$$K_0 := \frac{17 - 49\gamma_0}{32 - 4\gamma_0} \quad (< \tfrac{17}{32}).$$

To deal with the lower bound of $H_\gamma(x)$, we consider

$$T_\gamma^0(P, Q) := \#\big\{\, (p, q) : p \sim P, q \sim Q, 0 < \frac{\bar{q}}{p} \leq \Theta_0^\dagger \,\big\} \tag{50}$$

with

$$\Theta_0^\dagger := \theta_0(4PQ),$$

where θ_0 is defined in Lemma 8. We have the inequality

$$H_\gamma(x) \geq H_\gamma^0(x) \geq \sum_P \sum_Q T_\gamma^0(P, Q), \tag{51}$$

where $H_\gamma^0(x)$ is defined in (15) and the sum is over the pairs (P, Q) of the form $(2^k, 2^\ell)$ with

$$P \leq Q^{K_0}, \qquad x/16 \leq PQ \leq x/4, \qquad P \leq Q/2 \quad \text{and} \quad 1 \leq L \leq Q^{\beta_{K_0}}, \tag{52}$$

where L is defined in (46) and β_K is the constant introduced in Lemma 7. If the triple (ℓ, p, q) is such that $1 \leq \ell \ll L$, $p \sim P$ and $q \sim Q$ and satisfies $\ell q - np = 1$ for some integer n, then it contributes to $T_\gamma^0(P, Q)$. This leads to the inequality

$$T_\gamma^0(P, Q) \geq \sum_{p \sim P} \sum_{1 \leq \ell \ll L} \left(\pi(2Q; p, \bar{\ell}) - \pi(Q; p, \bar{\ell})\right).$$

Thanks to (52), we can apply Lemma 7, giving

$$T_\gamma^0(P, Q) \geq \alpha_{K_0} \sum_{p \sim P} L \cdot \frac{Q}{\varphi(p) \log 2Q} - O\left(\frac{P}{\log^2 2P} \cdot L \cdot \frac{Q}{P \log 2Q}\right),$$

which simplifies into

$$T_\gamma^0(P, Q) \geq \frac{\alpha_{K_0}}{2} \cdot \frac{LQ}{\log 2P \log 2Q} \tag{53}$$

for $x \geq x_0$ and (P, Q) satisfying (52).

In terms of P, the conditions (52) and $L \gg 1$ reduce to

$$P \ll x^{\frac{K_0}{1+K_0}} \quad \text{and} \quad x^{\frac{1}{2}-\gamma} \ll P \ll x^{(\frac{1}{2}+\beta_{K_0}-\gamma)/(1+\beta_{K_0})}. \tag{54}$$

The definition of K_0 implies the inequality

$$\frac{K_0}{1+K_0} - \left(\frac{1}{2} - \gamma\right) \geq \frac{K_0}{1+K_0} - \left(\frac{17}{49} - \gamma_0\right) \gg_{\gamma_0} 1.$$

Combining with the inequality $\beta_{K_0} > 0$, we see that there are $\gg_{\gamma_0} \mathscr{L}$ values of P of the form $P = 2^k$ satisfying (54). Since we also have $x/(16P) \leq Q \leq x/(4P)$, we deduce that there are $\gg_{\gamma_0} \mathscr{L}$ pairs (P, Q) satisfying (52). It remains to insert the lower bound (53) in (51) and to sum over the suitable (P, Q) to deduce

$$H_\gamma^0(x) \gg_{\gamma_0} x^{\frac{1}{2}+\gamma} \mathscr{L}^{-1}.$$

This completes the proof of Theorem 3. □

Remark. Not using Lemma 7 but only Lemma 5, one proves Theorem 3 but under the more restrictive condition $\frac{1}{6} + \gamma_0 \leq \gamma \leq 1 - \gamma_0$.

7. A conjectural formula

One may conjecture that for every $\gamma_0 > 0$, one has

$$H_\gamma(x) \sim C(\gamma)\frac{x^{\frac{1}{2}+\gamma}}{\log x} \tag{55}$$

as $x \to \infty$ uniformly under the condition (13). This conjecture, if true, would be an important extension of Theorem 1. However, (55) is a consequence of the Elliott–Halberstam Conjecture (see [Friedlander and Iwaniec 2010, page 406] for instance).

Conjecture 1. *For any $\epsilon > 0$ and any $A > 0$, one has*

$$\sum_{r \le x^{1-\epsilon}} \max_{(s,r)=1} \left|\pi(x; r, s) - \frac{\pi(x)}{\varphi(r)}\right| = O_{\epsilon,A}(x\mathscr{L}^{-A}). \tag{56}$$

This conjecture can be interpreted as a considerable improvement of Lemma 5 since it gives the average behavior of the function $\pi(x; r, s)$ for almost all $r \le x^{1-\epsilon}$.

We now give some indications on how to deduce (55) from Conjecture 1. First of all, one applies the formula (56) to evaluate $R_\gamma(P, Q)$ as written in (26). This shows that (29) is true uniformly for $P \le Qx^{-\epsilon}$ (compare with (30)). Summing over all these (P, Q), we see that their contribution to $H_\gamma(x)$ is $\sim (C(\gamma) - O(\epsilon))x^{\frac{1}{2}+\gamma}\mathscr{L}^{-1}$ by a computation analogous to (41) and (42) with uniformity given by (13).

For the remaining (P, Q) (those that satisfy $Qx^{-\epsilon} \le P \le \xi \cdot Q$), we apply the two-dimensional sieve as in Section 5. Then one shows that their contribution to $H_\gamma(x)$ is $O_{\gamma_0}(\epsilon x^{\frac{1}{2}+\gamma}\mathscr{L}^{-1})$. Summing up these two contributions and letting ϵ tend to 0, we get (55).

Acknowledgements

The author warmly thanks I. Shparlinski for indicating the work of B. Bzdęga and for generously sharing his ideas about this problem. We also thank F. Jouve for his comments on a preliminary version of that work and the referee for his judicious remarks, which clarified some parts of the present paper.

References

[Baker and Harman 1996] R. C. Baker and G. Harman, "The Brun–Titchmarsh theorem on average", pp. 39–103 in *Analytic number theory, Vol. 1* (Allerton Park, IL, 1995), edited by B. C. Berndt et al., Progr. Math. **138**, Birkhäuser, Boston, MA, 1996. MR 97h:11096 Zbl 0853.11078

[Bombieri et al. 1987] E. Bombieri, J. B. Friedlander, and H. Iwaniec, "Primes in arithmetic progressions to large moduli, II", *Math. Ann.* **277**:3 (1987), 361–393. MR 88f:11085 Zbl 0625.10036

[Bombieri et al. 1989] E. Bombieri, J. B. Friedlander, and H. Iwaniec, "Primes in arithmetic progressions to large moduli, III", *J. Amer. Math. Soc.* **2**:2 (1989), 215–224. MR 89m:11087 Zbl 0674.10036

[Bzdęga 2012] B. Bzdęga, "Sparse binary cyclotomic polynomials", *J. Number Theory* **132**:3 (2012), 410–413. MR 2875347 Zbl pre06005607

[Carlitz 1966] L. Carlitz, "The number of terms in the cyclotomic polynomial $F_{pq}(x)$", *Amer. Math. Monthly* **73** (1966), 979–981. MR 34 #2517 Zbl 0146.26704

[Fouvry 1985] É. Fouvry, "Théorème de Brun–Titchmarsh: application au théorème de Fermat", *Invent. Math.* **79**:2 (1985), 383–407. MR 86g:11052 Zbl 0557.10035

[Fouvry and Michel 1998] É. Fouvry and P. Michel, "Sur certaines sommes d'exponentielles sur les nombres premiers", *Ann. Sci. École Norm. Sup.* (4) **31**:1 (1998), 93–130. MR 98m:11088 Zbl 0915.11045

[Fouvry and Shparlinski 2011] É. Fouvry and I. E. Shparlinski, "On a ternary quadratic form over primes", *Acta Arith.* **150**:3 (2011), 285–314. MR 2842959 Zbl 1243.11093

[Friedlander and Iwaniec 2010] J. Friedlander and H. Iwaniec, *Opera de cribro*, American Mathematical Society Colloquium Publications **57**, American Mathematical Society, Providence, RI, 2010. MR 2011d:11227 Zbl 1226.11099

[Garaev 2010] M. Z. Garaev, "An estimate for Kloosterman sums with primes and its application", *Mat. Zametki* **88**:3 (2010), 365–373. In Russian; translated in *Math. Notes* **88**:3 (2010), 330–337. MR 2882176 Zbl pre05980675

[Habsieger and Sivak-Fischler 2010] L. Habsieger and J. Sivak-Fischler, "An effective version of the Bombieri–Vinogradov theorem, and applications to Chen's theorem and to sums of primes and powers of two", *Arch. Math.* (*Basel*) **95**:6 (2010), 557–566. MR 2011m:11190 Zbl 05833792

[Hildebrand 1985] A. Hildebrand, "On a conjecture of Balog", *Proc. Amer. Math. Soc.* **95**:4 (1985), 517–523. MR 87c:11001 Zbl 0597.10056

[Iwaniec and Kowalski 2004] H. Iwaniec and E. Kowalski, *Analytic number theory*, American Mathematical Society Colloquium Publications **53**, American Mathematical Society, Providence, RI, 2004. MR 2005h:11005 Zbl 1059.11001

[Lam and Leung 1996] T. Y. Lam and K. H. Leung, "On the cyclotomic polynomial $\Phi_{pq}(X)$", *Amer. Math. Monthly* **103**:7 (1996), 562–564. MR 97h:11150 Zbl 0868.11016

[Mikawa 2001] H. Mikawa, "On primes in arithmetic progressions", *Tsukuba J. Math.* **25**:1 (2001), 121–153. MR 2002c:11116 Zbl 1017.11049

[Rousselet 1988] B. Rousselet, "Inégalités de type Brun–Titchmarsh en moyenne", pp. 91–123 in *Groupe de travail en théorie analytique et élémentaire des nombres, 1986–1987*, Publ. Math. Orsay **88**, Univ. Paris XI, Orsay, 1988. MR 89g:11076 Zbl 0669.10067

[Vinogradov 1954] I. M. Vinogradov, *The method of trigonometrical sums in the theory of numbers*, Interscience Publishers, New York, 1954. Zbl 0055.27504

Communicated by Philippe Michel
Received 2012-02-20 Revised 2012-06-29 Accepted 2012-08-13

etienne.fouvry@math.u-psud.fr *Laboratoire de Mathématique,*
Campus d'Orsay, Université de Paris-Sud,
Bâtiment 425 UMR 8628, 91405 Orsay Cedex, France
http://www.math.u-psud.fr/~fouvry/

ALGEBRA AND NUMBER THEORY 7:5 (2013)
dx.doi.org/10.2140/ant.2013.7.1225

Local and global canonical height functions for affine space regular automorphisms

Shu Kawaguchi

In memory of Professor Masaki Maruyama

Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism defined over a number field K. For each place v of K, we construct the v-adic Green functions $G_{f,v}$ and $G_{f^{-1},v}$ (i.e., the v-adic canonical height functions) for f and f^{-1}. Next we introduce for f the notion of good reduction at v, and using this notion, we show that the sum of v-adic Green functions over all v gives rise to a canonical height function for f that satisfies a Northcott-type finiteness property. Using an earlier result, we recover results on arithmetic properties of f-periodic points and non-f-periodic points. We also obtain an estimate of growth of heights under f and f^{-1}, which was independently obtained by Lee by a different method.

Introduction

Height functions are one of the basic tools in diophantine geometry. On abelian varieties defined over a number field, there exist Néron–Tate canonical height functions that behave well relative to the n-th power map. Tate's elegant construction is via a global method using a relation of an ample divisor relative to the n-th power map. Néron's construction is via a local method and gives deeper properties of the canonical height functions. Both constructions are useful in studying arithmetic properties of abelian varieties.

In [Kawaguchi 2006], we showed the existence of canonical height functions for affine plane polynomial automorphisms of dynamical degree at least 2. Our construction was via a global method using the effectiveness of a certain divisor on a certain rational surface. In this paper, we use a local method to construct a canonical height function for affine space regular automorphisms $f : \mathbb{A}^N \to \mathbb{A}^N$, which coincides with the one in [Kawaguchi 2006] when $N = 2$. We note that arithmetic properties of polynomial automorphisms over number fields have been

Supported in part by MEXT Grant-in-Aid for Young Scientists (B) 21740018.
MSC2010: primary 37P30; secondary 11G50, 37P05, 37P20.
Keywords: canonical height, local canonical height, regular polynomial automorphism.

studied, for example, by Silverman [1994], Denis [1995], Marcello [2000; 2003], and the author [Kawaguchi 2006].

We recall the definition of regular polynomial automorphisms. Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a polynomial automorphism of degree $d \geq 2$ defined over a field, and let $\overline{f} : \mathbb{P}^N \dashrightarrow \mathbb{P}^N$ denote its birational extension to $\mathbb{P}^N$. We write f^{-1} for the inverse of f, d_- for the degree of f^{-1}, and $\overline{f^{-1}}$ for its birational extension to $\mathbb{P}^N$. Then f is said to be *regular* if the intersection of the set of indeterminacy of $\overline{f}$ and that of $\overline{f^{-1}}$ is empty over an algebraic closure of the field (see Definition 2.1 and Remark 2.2). Over $\mathbb{C}$, dynamical properties of affine space regular polynomial automorphisms f are deeply studied, in which the Green function for f plays a pivotal role; see [Sibony 1999, §2].

In Sections 1 and 2, we construct a Green function (a local canonical height function) for f over an algebraically closed field Ω with nontrivial nonarchimedean absolute value $|\cdot|$. For $x = (x_1, \dots, x_N) \in \Omega^N$, we set $\|x\| = \max_{1\leq i\leq N}\{|x_i|\}$. Our results are put together as follows.

Theorem A (see Proposition 1.1, Lemma 1.3, and Theorem 2.3). *Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism of degree $d \geq 2$ defined over Ω.*

(1) *For all $x \in \mathbb{A}^N(\Omega)$, the limits*

$$\lim_{n\to+\infty} \frac{1}{d^n} \log\max\{\|f^n(x)\|, 1\} \quad \textit{and} \quad \lim_{n\to+\infty} \frac{1}{d_-^n} \log\max\{\|f^{-n}(x)\|, 1\}$$

exist and are nonnegative. We respectively write $G_f(x) \geq 0$ and $G_{f^{-1}}(x) \geq 0$ for the limits, which we call Green functions for f and f^{-1}. They satisfy the functional equations $G_f(f(x)) = dG_f(x)$ and $G_{f^{-1}}(f^{-1}(x)) = d_-G_{f^{-1}}(x)$.

(2) *There are constants $c_f, c_{f^{-1}} \in \mathbb{R}$ such that, on $\mathbb{A}^N(\Omega)$,*

$$G_f(\,\cdot\,) \leq \log\max\{\|\cdot\|, 1\} + c_f,$$
$$G_{f^{-1}}(\,\cdot\,) \leq \log\max\{\|\cdot\|, 1\} + c_{f^{-1}}$$

(3) *There are subsets V^+ and V^- of $\mathbb{A}^N(\Omega)$ with $V^+ \cup V^- = \mathbb{A}^N(\Omega)$ and constants $c^+, c^- \in \mathbb{R}$ such that*

$$G_f(\,\cdot\,) \geq \log\max\{\|\cdot\|, 1\} + c^+ \quad \textit{on } V^+,$$
$$G_{f^-}(\,\cdot\,) \geq \log\max\{\|\cdot\|, 1\} + c^- \quad \textit{on } V^-.$$

Over $\mathbb{C}$, Green functions are constructed using compactness arguments [Sibony 1999, §2]. Here we use more algebraic arguments based on Hilbert's Nullstellensatz. Our construction of $V^\pm$ and $c^\pm$ is rather delicate with a choice of two parameters ε and δ, which behaves well when we work over number fields in Sections 6 and 7. We note that over $\mathbb{C}$, our construction gives a different proof of the existence of Green functions with more explicit estimates (see Section 5). In Section 3, we continue to study some basic properties of regular polynomial automorphisms f over Ω, characterizing the set of the points with unbounded orbit by G_f and showing a filtration property for f.

Now we turn our attention to number fields. Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a polynomial automorphism defined over a number field K. For each place v of K, let K_v denote the completion of K with respect to v and $\overline{K}_v$ an algebraic closure of K_v. Then f induces a regular polynomial automorphism over $\overline{K}_v$, so we have Green functions $G_{f,v}$ and $G_{f^{-1},v}$ and estimates with $c_{f,v}$, $c_{f^{-1},v}$, and $c_v^{\pm}$ as in Theorem A. (Here we use the suffix v to indicate that we work over $\overline{K}_v$. See Section 5 when v is archimedean.)

We want to define the canonical height functions $\hat{h}_f^+$ and $\hat{h}_f^-$ for f as the sum of $G_{f,v}$ and $G_{f^{-1},v}$ over all the places v of K. To this end, we introduce the notion of good reduction at a nonarchimedean place v of K. Let $\overline{R}_v$ denote the ring of integers of $\overline{K}_v$ and $\tilde{k}_v$ the residue field. Recall that the notion of good reduction for an endomorphism φ of $\mathbb{P}^1$ over $\overline{K}_v$ is introduced in [Morton and Silverman 1994], which means that φ extends to a morphism over $\overline{R}_v$ and the induced morphism $\tilde{\varphi}$ over $\tilde{k}_v$ has the same degree as φ. Here we say that a regular polynomial automorphism $f : \mathbb{A}^N \to \mathbb{A}^N$ has *good reduction* at v if f extends to an automorphism over $\overline{R}_v$ and the induced morphism $\tilde{f}$ over $\tilde{k}_v$ is again a regular polynomial automorphism such that the degrees of $\tilde{f}$ and $\tilde{f}^{-1}$ are the same as the degrees of f and f^{-1}, respectively (see Definition 4.1 for the precise definition).

Using the notion of good reduction, we show the existence of canonical height functions. Let $h : \mathbb{A}^N(\overline{K}) \to \mathbb{R}$ denote the usual logarithmic Weil height function.

Theorem B (see Proposition 6.2 and Theorem 6.3). *Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism of degree $d \geq 2$ over a number field K. Let $d_- \geq 2$ denote the degree of f^{-1}.*

(1) *Then f has good reduction at v except for finitely many places. Further, if this is the case, we can take the constants $c_{f,v} = c_{f^{-1},v} = c_v^{\pm} = 0$ in Theorem A, so*

$$G_f(\,\cdot\,) = \log\max\{\|\cdot\|, 1\} \quad \textit{on } V^+,$$
$$G_{f^{-1}}(\,\cdot\,) = \log\max\{\|\cdot\|, 1\} \quad \textit{on } V^-.$$

(2) *For all $x \in \mathbb{A}^N(\overline{K})$, the limits*

$$\hat{h}_f^+(x) := \lim_{n\to+\infty} \frac{1}{d^n} h(f^n(x)) \quad \textit{and} \quad \hat{h}_f^-(x) := \lim_{n\to+\infty} \frac{1}{d_-^n} h(f^{-n}(x)) \tag{0-1}$$

exist. Further, we have the decomposition into the sum of local Green functions

$$\hat{h}_f^+(x) = \sum_{v\in M_K} n_v G_{f,v}(x) \quad \textit{and} \quad \hat{h}_f^-(x) = \sum_{v\in M_K} n_v G_{f^{-1},v}(x).$$

(3) *We define* $\hat{h}_f : \mathbb{A}^N(\overline{K}) \to \mathbb{R}$ *by* $\hat{h}_f := \hat{h}_f^+ + \hat{h}_f^-$. *Then* $\hat{h}_f$ *satisfies* $\hat{h}_f \gg\ll h$ *and*

$$\frac{1}{d}\hat{h}_f \circ f + \frac{1}{d_-}\hat{h}_f \circ f^{-1} = \Big(1 + \frac{1}{dd_-}\Big)\hat{h}_f.$$

Further, for $x \in \mathbb{A}^N(\overline{K})$ *we have*

$$\hat{h}_f(x) = 0 \iff \hat{h}_f^+(x) = 0 \iff \hat{h}_f^-(x) = 0 \iff x \textit{ is } f\textit{-periodic.}$$

In [Kawaguchi 2006] we have defined $\hat{h}_f^+(x)$ as $\limsup_{n\to\infty} \frac{1}{d^n}h(f^n(x))$, and similarly for $\hat{h}_f^-$. Theorem B shows that $\big\{\frac{1}{d^n}h(f^n(x))\big\}_{n=0}^{+\infty}$ and $\big\{\frac{1}{d_-^n}h(f^{-n}(x))\big\}_{n=0}^{+\infty}$ are in fact convergent sequences, i.e., lim sup can be replaced by lim as in (0-1).

Using estimates on local Green functions over all places, we obtain the following estimate on global height functions for all $N \geq 2$ [Kawaguchi 2006, §4; Silverman 2006, Conjecture 3; 2007, Conjecture 7.18]. This result has been independently proved by Chong Gyu Lee [2013]. His proof uses a global method and is based on the effectiveness of a certain divisor (as was done for $N = 2$ in [Kawaguchi 2006]).

Corollary C (see Theorem 7.1). *Let* $f : \mathbb{A}^N \to \mathbb{A}^N$ *be a regular polynomial automorphism over a number field* K. *With the notation as above, there exists a constant* $c \geq 0$ *such that*

$$\frac{1}{d}h(f(x)) + \frac{1}{d_-}h(f^{-1}(x)) \geq \Big(1 + \frac{1}{dd_-}\Big)h(x) - c \tag{0-2}$$

for all $x \in \mathbb{A}^N(\overline{K})$. *Further, we have*

$$\liminf_{\substack{x \in \mathbb{A}^N(\overline{K}) \\ h(x)\to\infty}} \frac{\frac{1}{d}h(f(x)) + \frac{1}{d_-}h(f^{-1}(x))}{h(x)} = 1 + \frac{1}{dd_-}.$$

Since (0-2) holds, by the argument of [Kawaguchi 2006] we recover the results on f-periodic points and refine the results on non-f-periodic points in [Silverman 1994; Denis 1995; Marcello 2000; 2003]. For $x \in \mathbb{A}^N(\overline{K})$, let $O_f(x) := \{\, f^n(x) \mid n \in \mathbb{Z}\,\}$ denote the f-orbit of x. If $O_f(x)$ is infinite, we have the canonical height $\hat{h}(O_f(x))$ of $O_f(x)$ (see Equation (7-6)).

Corollary D (see Equation (7-6) and Corollary 7.4). *Let* $f : \mathbb{A}^N \to \mathbb{A}^N$ *be a regular polynomial automorphism over a number field* K. *With the notation as above,*

(1) *the set of* f*-periodic points in* $\mathbb{A}^N(\overline{K})$ *is a set of bounded height and*

(2) *for any infinite orbit* $O_f(x)$,

$$\#\{\, y \in O_f(x) \mid h(y) \leq T\,\} = \Big(\frac{1}{\log d} + \frac{1}{\log d_-}\Big)\log T - \hat{h}(O_f(x)) + O(1)$$

as $T \to +\infty$, *where* $O(1)$ *is independent of* T *and* x *but depends on* f.

1. Nonarchimedean Green functions for polynomial maps

Let Ω be an algebraically closed field with nontrivial nonarchimedean absolute value $|\cdot|$ and R its ring of integers. For a point $x = (x_1, \dots, x_N) \in \mathbb{A}^N(\Omega)$, the norm of x is defined by $\|x\| = \max_{i=1,\dots,N}\{|x_i|\}$. We set $\log^+(a) := \log\max\{a, 1\}$ for $a \in \mathbb{R}_{\geq 0}$ as usual so that $\log^+\|x\| = \log\max\{\|x\|, 1\} = \log\|(x, 1)\|$.

Let $f = (f_1, \dots, f_N) : \mathbb{A}^N \to \mathbb{A}^N$ be a polynomial map of degree $d \geq 2$ defined over Ω, where $f_1(X), \dots, f_N(X)$ are polynomials in $\Omega[X_1, \dots, X_N]$ such that $d = \max_{i=1,\dots,N}\{\deg f_i\}$. We write $F_i(X, T) := T^d f_i(X/T) \in \Omega[X_1, \dots, X_N, T]$ for homogenization of f_i. Let $\bar{f} = (F_1 : \cdots : F_N : T^d) : \mathbb{P}^N \dashrightarrow \mathbb{P}^N$ denote the extension of f to $\mathbb{P}^N$. We put $F := (F_1, \dots, F_N, T^d) : \mathbb{A}^{N+1} \to \mathbb{A}^{N+1}$, which is a lift of $\bar{f}$.

For the composition $f^n = f \circ \cdots \circ f$, we write $f^n = (f_1^n, \dots, f_N^n)$. Similarly, for the composition $F^n = F \circ \cdots \circ F$, we write $F^n = (F_1^n, \dots, F_N^n, T^{d^n})$. Let d_n denote the degree of f^n, and let $F_{ni}(X, T) = T^{d_n} f_i^n(X/T) \in \Omega[X_1, \dots, X_N, T]$ be homogenization of f_i^n. Since $F_i^n(X, 1) = f_i^n(X) = F_{ni}(X, 1)$, counting degrees gives $F_i^n(X, T) = T^{d^n - d_n} F_{ni}(X, T)$.

Proposition 1.1. *Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a polynomial map of degree $d \geq 2$ defined over Ω. Then for all $x \in \mathbb{A}^N(\Omega)$, $\frac{1}{d^n}\log^+\|f^n(x)\|$ converges to a nonnegative real number as $n \to +\infty$.*

Proof. We take an $r \in R$ so that $rF_i \in R[X, T]$ for all $i = 1, \dots, N$. We set

$$a_n := \frac{1}{d^n}\log^+\|f^n(x)\|, \quad b_n := \frac{1}{d^n}\log\|F^n(x, 1)\|, \quad c_n := \frac{1}{d^n}\log\|(rF)^n(x, 1)\|,$$

where $rF = (rF_1, \dots, rF_N, rT^d)$. We claim that

$$a_n = b_n = c_n - \frac{1 - d^{-n}}{d - 1}\log|r|. \tag{1-1}$$

Indeed, the first equality follows from $(f^n(x), 1) = (F^n(x, 1))$. The second equality follows from $(rF)^n = r^{1+d+\cdots+d^{n-1}}F^n = r^{(d^n-1)/(d-1)}F^n$. It follows from $\|(rF)(x, 1)\| \leq \|(x, 1)\|^d$ that

$$\frac{1}{d^n}\log\|(rF)^n(x, 1)\| \leq \frac{1}{d^n}\log\|(rF)^{n-1}(x, 1)\|^d = \frac{1}{d^{n-1}}\log\|(rF)^{n-1}(x, 1)\|.$$

In other words, $\{c_n\}_{n=1}^{+\infty}$ is a nonincreasing sequence. Equation (1-1) implies that $\{c_n\}_{n=1}^{+\infty}$ is bounded from below. Indeed, since a_n is nonnegative and $|r| \leq 1$, we have $c_n \geq a_n + \frac{1}{d-1}\log|r| \geq \frac{1}{d-1}\log|r|$. Thus, $\lim_{n\to+\infty} c_n$ exists. Equation (1-1) then gives the existence of $\lim_{n\to+\infty} a_n$, which is nonnegative from the definition. □

Proposition 1.1 allows the following definition:

Definition 1.2. For a polynomial map $f : \mathbb{A}^N \to \mathbb{A}^N$ defined over Ω, we define the nonnegative function $G_f : \mathbb{A}^N(\Omega) \to \mathbb{R}$ by

$$G_f(x) := \lim_{n\to+\infty} \frac{1}{d^n} \log^+ \| f^n(x) \| \quad \text{for } x \in \mathbb{A}^N(\Omega)$$

and call it the *Green function* for f.

Lemma 1.3. *Let C'_f be the maximum of the absolute value of all the coefficients of $f_i(X)$ for $1 \le i \le N$, and we set*

$$c_f = \frac{1}{d-1} \log \max\{C'_f, 1\}.$$

Then

$$G_f(\cdot) \le \log^+ \| \cdot \| + c_f \quad \text{on } \mathbb{A}^N(\Omega).$$

Proof. We take $r \in R$ such that $|r| = 1/\max\{C'_f, 1\}$. Then $rF_i \in R[X, T]$ for all $i = 1, \dots, N$. From the proof of Proposition 1.1, we have

$$G_f(x) \le \lim_{n\to+\infty} c_n - \frac{1}{d-1}\log|r| \le c_0 - \frac{1}{d-1}\log|r| = \log^+\|x\| - \frac{1}{d-1}\log|r|.$$

Hence, we get the assertion. □

Lemma 1.4 below shows that for some polynomial maps f, G_f is not interesting. However, we will see in the next section that G_f enjoys nice properties for regular polynomial automorphisms f (see Definition 2.1 and Theorem 2.3).

To state Lemma 1.4, we recall that a polynomial map f is said to be *algebraically stable* if $d_n = d^n$ for all $n \ge 1$ [Sibony 1999, §1.4].

Lemma 1.4. *If f is not algebraically stable, then $G_f(x) = 0$ for all $x \in \mathbb{A}^N(\Omega)$.*

Proof. We take n_0 such that $d_{n_0} < d^{n_0}$, and we put $g = f^{n_0}$. Proposition 1.1 tells us that $(1/d_{n_0}^m)\log^+\|g^m(x)\|$ converges to a nonnegative number as $m \to +\infty$. Hence,

$$\frac{1}{d^{n_0 m}} \log^+ \|f^{n_0 m}(x)\| = \left(\frac{d_{n_0}}{d^{n_0}}\right)^m \frac{1}{d_{n_0}^m} \log^+\|g^m(x)\| \to 0 \quad \text{as } m \to +\infty.$$

From Proposition 1.1, we get $G_f(x) = 0$. □

2. Nonarchimedean Green functions for regular automorphisms

In this section, we consider polynomial automorphisms. Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a polynomial automorphism of degree $d \ge 2$ defined over an algebraically closed field Ω with nontrivial nonarchimedean absolute value.

As before, let $\overline{f} = (F_1(X, T) : \cdots : F_N(X, T) : T^d) : \mathbb{P}^N \dashrightarrow \mathbb{P}^N$ denote the extension of f to $\mathbb{P}^N$. We denoted by d_- the degree of the inverse $f^{-1} : \mathbb{A}^N \to \mathbb{A}^N$ of f. The integer $d_- \ge 2$ may be different from d. We denote the extension of f^{-1} to $\mathbb{P}^N$ by $\overline{f^{-1}} = (G_1(X, T) : \cdots : G_N(X, T) : T^{d_-}) : \mathbb{P}^N \dashrightarrow \mathbb{P}^N$.

Let I_+ and I_- denote the set of indeterminacy of $\overline{f}$ and $\overline{f^{-1}}$, respectively:

$$I_+ = \{\, (x:0) \in \mathbb{P}^N(\Omega) \mid F_1(x,0) = \cdots = F_N(x,0) = 0 \,\},$$
$$I_- = \{\, (x:0) \in \mathbb{P}^N(\Omega) \mid G_1(x,0) = \cdots = G_N(x,0) = 0 \,\}.$$

Definition 2.1 [Sibony 1999, §2.2]. A polynomial automorphism $f : \mathbb{A}^N \to \mathbb{A}^N$ is called *regular* if $I_+ \cap I_- = \varnothing$.

Remark 2.2. The definition of regular polynomial automorphisms works over any algebraically closed field.

The purpose of this section is to prove the following theorem, which says that the Green functions for regular automorphisms exhibit nice properties:

Theorem 2.3. *Let Ω be an algebraically closed field with nontrivial nonarchimedean valuation and $f : \mathbb{A}^N \to \mathbb{A}^N$ a regular polynomial automorphism over Ω. Then there are open subsets V^+ and V^- of $\mathbb{A}^N(\Omega)$ with respect to the topology induced from the valuation on Ω and constants $c^+, c^- \in \mathbb{R}$ with the properties*

(i) $G_f(\,\cdot\,) \geq \log^+\|\cdot\| + c^+$ *on* V^+,

(ii) $G_{f^{-1}}(\,\cdot\,) \geq \log^+\|\cdot\| + c^-$ *on* V^-, *and*

(iii) $V^+ \cup V^- = \mathbb{A}^N(\Omega)$.

Remark 2.4. Over $\mathbb{C}$, corresponding results (and much more) were established by Sibony [1999, §2.2]. Here since $\mathbb{A}^N(\Omega)$ is not locally compact in general, we give a different proof that is more algebraic in nature based on Hilbert's Nullstellensatz. We also give V^+, V^-, c^+, and c^- with precise estimates so that they work well when we introduce the notion of good reduction in Section 4.

Before proving Theorem 2.3, we will need several lemmas. We begin by introducing some notation. Since $I_+ \cap I_-$ is empty, $F_1(X,0), \dots, F_N(X,0)$ and $G_1(X,0), \dots, G_N(X,0)$ have no solutions in common other than 0. Thus, for each $1 \leq i \leq N$, there are polynomials $P_{ij}(X)$, $Q_{ij}(X) \in \Omega[X]$ for $1 \leq j \leq N$ such that

$$\sum_{j=1}^{N} P_{ij}(X) F_j(X,0) + \sum_{j=1}^{N} Q_{ij}(X) G_j(X,0) = X_i^m \tag{2-1}$$

with some $m \geq 1$. Hence, there is a polynomial $R_i(X,T) \in \Omega[X,T]$ such that

$$\sum_{j=1}^{N} P_{ij}(X) F_j(X,T) + \sum_{j=1}^{N} Q_{ij}(X) G_j(X,T) + T R_i(X,T) = X_i^m. \tag{2-2}$$

Here we may and do assume that m is independent of i. Replacing $P_{ij}(X)$ by its homogeneous part with degree $m-d$, $Q_{ij}(X)$ by its homogeneous part with degree $m-d_-$, and $R_i(X,T)$ by its homogeneous part with degree $m-1$, we may and do assume that the $P_{ij}(X)$, $Q_{ij}(X)$, and $R_i(X,T)$ are homogeneous polynomials with degree $m-d$, $m-d_-$, and $m-1$, respectively.

Let C' be the maximum of the absolute value of all the coefficients of $P_{ij}(X)$, $Q_{ij}(X)$, and $R_i(X,T)$ for $1 \le i \le N$ and $1 \le j \le N$. We set

$$C = \max\{C', 1\}. \tag{2-3}$$

We fix real numbers $\varepsilon > 0$ and $\delta > 0$ as follows. First we choose δ to satisfy $\delta \le \frac{1}{C}$. Then choose ε to satisfy

$$\varepsilon \le \min\left\{\frac{\delta^{1/d}}{C}, \frac{\delta^{1/d_-}}{C}\right\}.$$

This ensures $\varepsilon \le \frac{1}{C}$, so in particular, $\varepsilon \le 1$. To sum up, we have

$$\varepsilon \le \tfrac{1}{C}, \quad \delta \le \tfrac{1}{C}, \quad (\varepsilon C)^d \le \delta, \quad \text{and} \quad (\varepsilon C)^{d_-} \le \delta. \tag{2-4}$$

For example,

$$\varepsilon = \frac{1}{C^{\min\{d,d_-\}}} \quad \text{and} \quad \delta = \frac{1}{C^{\min\{d,d_-\}(\min\{d,d_-\}-1)}} \tag{2-5}$$

satisfy (2-4).

We define $N^+_{\delta,\varepsilon}$ and $V^+_{\delta,\varepsilon}$ by

$$\begin{aligned} N^+_{\delta,\varepsilon} &:= \{\, x \in \mathbb{A}^N(\Omega) \mid 1 < \varepsilon\|x\| \text{ and } \|f(x)\| < \delta\|x\|^d \,\},\\ V^+_{\delta,\varepsilon} &:= \mathbb{A}^N(\Omega) \setminus N^+_{\delta,\varepsilon} = \{\, x \in \mathbb{A}^N(\Omega) \mid \|x\| \le \tfrac{1}{\varepsilon} \text{ or } \|f(x)\| \ge \delta\|x\|^d \,\}. \end{aligned} \tag{2-6}$$

Intuitively, points in $N^+_{\delta,\varepsilon}$ are near to the hyperplane $\{(x:0) \in \mathbb{P}^N(\Omega)\}$ at infinity (measured by ε) and also near to I_+ in "the direction of x" (measured by δ). We note that both $N^+_{\delta,\varepsilon}$ and $V^+_{\delta,\varepsilon}$ are open and closed with respect to the topology induced from the valuation of Ω.

Remark 2.5. We set

$$\overline{N}^+_{\delta,\varepsilon} = \left\{ (x:t) \in \mathbb{P}^N(\Omega) \;\middle|\; |t| < \varepsilon\|x\| \text{ and } \|(F(x,t), t^d)\| < \delta\|(x,t)\|^d \right\}.$$

Then $N^+_{\delta,\varepsilon} = \overline{N}^+_{\delta,\varepsilon} \cap \mathbb{A}^N(\Omega)$. If $(x:t) \in I_+$, then $t=0$ and $F(x,t)=0$. Thus, $|t|=0$ and $\|(F(x,t),t^d)\| = 0$, so we have

$$I_+ \subseteq \overline{N}^+_{\delta,\varepsilon}.$$

The next lemma says that if a point is not too close to I_+, then f maps it to a point that is also not very close to I_+ and that the measurement of "closeness" is uniform with respect to the point.

Lemma 2.6. *We have* $f(V^+_{\delta,\varepsilon}) \subseteq V^+_{\delta,\varepsilon}$.

Proof. Taking the complement, it suffices to show that

$$f^{-1}(N^{+}_{\delta,\varepsilon}) \subseteq N^{+}_{\delta,\varepsilon}.$$

Suppose $x = (x_1, \dots, x_N) \in N^{+}_{\delta,\varepsilon}$. Without loss of generality, we assume $|x_1| = \|x\|$. We note $f(x) = (F_1(x, 1), \dots, F_N(x, 1))$ and $f^{-1}(x) = (G_1(x, 1), \dots, G_N(x, 1))$. Since $\varepsilon \le 1$, we have $\|x\| > 1$. Then the definition of $N^{+}_{\delta,\varepsilon}$ gives

$$\tfrac{1}{\varepsilon} < \|x\|, \tag{2-7}$$

$$\|f(x)\| < \delta\|x\|^{d}. \tag{2-8}$$

We need to show that $f^{-1}(x) \in N^{+}_{\delta,\varepsilon}$, which is equivalent to

$$1 < \varepsilon\|f^{-1}(x)\|, \tag{2-9}$$

$$\|x\| < \delta\|f^{-1}(x)\|^{d}. \tag{2-10}$$

First we show (2-9). To derive a contradiction, we assume that $\|f^{-1}(x)\| \le \frac{1}{\varepsilon}$. Let $\lambda > 0$ be any small number. We have

$$\begin{aligned}
&\biggl|\sum_{j=1}^{N} P_{1j}(x)F_j(x, 1) + \sum_{j=1}^{N} Q_{1j}(x)G_j(x, 1) + R_1(x, 1)\biggr| \\
&\qquad < \max\{C\|x\|^{m-d} \cdot \delta\|x\|^{d}, (C+\lambda)\|x\|^{m-d_-}\tfrac{1}{\varepsilon}, (C+\lambda)\|x\|^{m-1}\} \\
&\qquad \le \max\{C\delta\|x\|^{m}, (C+\lambda)\|x\|^{m-d_-+1}, (C+\lambda)\|x\|^{m-1}\} \qquad \text{(from (2-7))} \\
&\qquad \le \max\{C\delta\|x\|^{m}, (C+\lambda)\|x\|^{m-1}\} \qquad (\text{since } d_- \ge 2).
\end{aligned}$$

Since $\lambda > 0$ is arbitrary, (2-2) and the assumption that $|x_1| = \|x\|$ then gives either $\|x\|^{m} \le C\|x\|^{m-1}$ or $\|x\|^{m} < C\delta\|x\|^{m}$. Equivalently, we have either $\|x\| \le C$ or $1 < C\delta$. However, the former contradicts (2-4) and (2-7) while the latter contradicts (2-4). Hence, we get (2-9).

Next we show (2-10). To derive a contradiction, we assume the contrary, i.e., $\|x\| \ge \delta\|f^{-1}(x)\|^{d}$. Letting $\lambda > 0$ be any small number, we have

$$\begin{aligned}
&\biggl|\sum_{j=1}^{N} P_{1j}(x)F_j(x, 1) + \sum_{j=1}^{N} Q_{1j}(x)G_j(x, 1) + R_1(x, 1)\biggr| \\
&\quad < \max\bigl\{C\|x\|^{m-d} \cdot \delta\|x\|^{d}, (C+\lambda)\|x\|^{m-d_-} \cdot (\tfrac{1}{\delta})^{1/d}\|x\|^{1/d}, (C+\lambda)\|x\|^{m-1}\bigr\} \\
&\quad \le \max\{C\delta\|x\|^{m}, (C+\lambda)(\tfrac{1}{\delta})^{1/d}\|x\|^{m-d_-+1/d}, (C+\lambda)\|x\|^{m-1}\} \\
&\quad \le \max\{C\delta\|x\|^{m}, (C+\lambda)(\tfrac{1}{\delta})^{1/d}\|x\|^{m-1}\} \qquad \Bigl(\text{since } d_- - \frac{1}{d} \ge 1\Bigr).
\end{aligned}$$

Since $\lambda > 0$ is arbitrary, (2-2) and the assumption that $|x_1| = \|x\|$ gives this time

$$\text{either} \quad \|x\| \le (\tfrac{1}{\delta})^{1/d}C \quad \text{or} \quad 1 < C\delta.$$

However, the former contradicts (2-4) and (2-7) while the latter contradicts (2-4). Hence, we get (2-10), which completes the proof. □

Lemma 2.7. *Set* $C^+_{\delta,\varepsilon} := \min\{\delta, \varepsilon^d\}$. *Then*

$$\max\{\|f(x)\|, 1\} \geq C^+_{\delta,\varepsilon} \cdot \max\{\|x\|^d, 1\} \quad \textit{for all } x \in V^+_{\delta,\varepsilon}.$$

Proof. For $x \in V^+_{\delta,\varepsilon}$, the definition of $V^+_{\delta,\varepsilon}$ gives

$$\text{either} \quad \|x\| \leq \tfrac{1}{\varepsilon} \quad \text{or} \quad \max\{\|f(x)\|, 1\} \geq \delta \max\{\|x\|^d, 1\}.$$

If the latter holds, then we get the assertion since $\delta \geq C^+_{\delta,\varepsilon}$. If the former holds, then $C^+_{\delta,\varepsilon}\|x\|^d \leq 1$. We get $\max\{\|f(x)\|, 1\} \geq 1 \geq C^+_{\delta,\varepsilon} \cdot \max\{\|x\|^d, 1\}$ noting that $C^+_{\delta,\varepsilon} \leq 1$. □

Lemma 2.8. *Set* $c^+_{\delta,\varepsilon} := \frac{1}{d-1} \log C^+_{\delta,\varepsilon}$. *Then*

$$G_f(x) \geq \log^+\|x\| + c^+_{\delta,\varepsilon} \quad \textit{for all } x \in V^+_{\delta,\varepsilon}.$$

Proof. Suppose $x \in V^+_{\delta,\varepsilon}$. It follows from Lemma 2.6 that $f^n(x) \in V^+_{\delta,\varepsilon}$ for all $n \geq 1$. Then Lemma 2.7 gives

$$\log^+\|f^n(x)\| \geq d \log^+\|f^{n-1}(x)\| + \log C^+_{\delta,\varepsilon}.$$

The usual telescoping argument tells us that

$$\begin{aligned} G_f(x) &= \lim_{n\to+\infty} \frac{1}{d^n} \log^+\|f^n(x)\| \\ &= \log^+\|x\| + \sum_{n=1}^{\infty} \frac{1}{d^n}(\log^+\|f^n(x)\| - d\log^+\|f^{n-1}(x)\|) \\ &\geq \log^+\|x\| + c^+_{\delta,\varepsilon}. \end{aligned}$$ □

With f^{-1} in place of f, we define $N^-_{\delta,\varepsilon}$ and $V^-_{\delta,\varepsilon}$ by

$$\begin{aligned} &N^-_{\delta,\varepsilon} := \left\{ x \in \mathbb{A}^N(\Omega) \;\middle|\; 1 < \varepsilon\|x\| \text{ and } \max\{\|f^{-1}(x)\|, 1\} < \delta \max\{\|x\|^{d_-}, 1\} \right\}, \\ &V^-_{\delta,\varepsilon} := \mathbb{A}^N(\Omega) \setminus N^-_{\delta,\varepsilon}. \end{aligned} \tag{2-11}$$

Then setting $c^-_{\delta,\varepsilon} := \frac{1}{d_- -1} \log \min\{\delta, \varepsilon^{d_-}\}$, we have

$$G_{f^{-1}}(x) \geq \log^+\|x\| + c^-_{\delta,\varepsilon} \quad \text{for all } x \in V^-_{\delta,\varepsilon}. \tag{2-12}$$

The next lemma may be seen as a quantified version of the fact that a point cannot be too close to both I^+ and I^- since $I^+ \cap I^- = \varnothing$.

Lemma 2.9. $V^+_{\delta,\varepsilon} \cup V^-_{\delta,\varepsilon} = \mathbb{A}^N(\Omega)$, *or equivalently*, $N^+_{\delta,\varepsilon} \cap N^-_{\delta,\varepsilon} = \varnothing$.

Proof. Taking the complement, it suffices to show that $N^+_{\delta,\varepsilon} \cap N^-_{\delta,\varepsilon} = \varnothing$. To derive a contradiction, we assume that there is an $x \in N^+_{\delta,\varepsilon} \cap N^-_{\delta,\varepsilon}$. Then we have

$$\|x\| > \tfrac{1}{\varepsilon}, \tag{2-13}$$

$$\|f(x)\| < \delta\|x\|^d, \tag{2-14}$$

$$\|f^{-1}(x)\| < \delta\|x\|^{d_-}. \tag{2-15}$$

Without loss of generality, we assume that $|x_1| = \|x\|$. Let $\lambda > 0$ be any small number. By (2-13)–(2-15), we have

$$\begin{aligned}\left|\sum_{j=1}^N P_{1j}(x)F_j(x,1) + \sum_{j=1}^N Q_{1j}(x)G_j(x,1) + R_1(x,1)\right| \\ &< \max\{C\|x\|^{m-d}\cdot\delta\|x\|^d,\, C\|x\|^{m-d_-}\cdot\delta\|x\|^{d_-},\, (C+\lambda)\|x\|^{m-1}\} \\ &\le \max\{C\delta\|x\|^m,\, (C+\lambda)\|x\|^{m-1}\}.\end{aligned}$$

Since λ is arbitrary, it follows from (2-2) that $\|x\|^m < C\delta\|x\|^m$ or $\|x\|^m \le C\|x\|^{m-1}$. Hence, we get

$$\text{either} \quad 1 < C\delta \quad \text{or} \quad \|x\| \le C.$$

However, the former contradicts (2-4) while the latter contradicts (2-4) and (2-13). Thus, we have $N^+_{\delta,\varepsilon} \cap N^-_{\delta,\varepsilon} = \varnothing$. □

Proof of Theorem 2.3. Let δ and ε be constants satisfying (2-4). Then Theorem 2.3 holds with $V^\pm = V^\pm_{\delta,\varepsilon}$ and $c^\pm = c^\pm_{\delta,\varepsilon}$. Indeed, the condition (i) follows from Lemma 2.8 and the condition (ii) from (2-12) while the condition (iii) follows from Lemma 2.9. □

3. Nonarchimedean Green functions and the set of escaping points

In this section, we continue to study basic properties of regular polynomial automorphisms defined over Ω. We keep the notation and the assumption of Section 2. In particular, $f : \mathbb{A}^N \to \mathbb{A}^N$ denotes a regular polynomial automorphism of degree $d \ge 2$ defined over Ω.

In analogy with the field of complex numbers, we define the set W^+ of escaping points and the set $\mathcal{K}^+$ of nonescaping points by

$$\begin{aligned} W^+ &:= \{\, x \in \mathbb{A}^N(\Omega) \mid \|f^n(x)\| \to +\infty \ (n \to +\infty)\,\},\\ \mathcal{K}^+ &:= \{\, x \in \mathbb{A}^N(\Omega) \mid \{f^n(x)\}_{n=0}^{+\infty} \text{ is bounded with respect to } \|\cdot\|\,\}.\end{aligned}$$

Then the following theorem holds, which is a nonarchimedean version of the results of [Bedford and Smillie 1991, §2 and §3; Sibony 1999, §2]:

Theorem 3.1. *Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism over Ω, and let G_f be the Green function for f.*

(1) *The set $\mathcal{K}^+$ is exactly the set of points where G_f vanish*:

$$\mathcal{K}^+ = \{ x \in \mathbb{A}^N(\Omega) \mid G_f(x) = 0 \}.$$

(2) $\mathbb{A}^N(\Omega) = W^+ \amalg \mathcal{K}^+$ (*disjoint union*).

To prove Theorem 3.1, we need the following two lemmas. Recall that δ and ε are fixed constants satisfying (2-4).

Lemma 3.2. *For any $x \in N^+_{\delta,\varepsilon/2}$, one has $\|x\| \le \frac{1}{2}\|f^{-1}(x)\|$.*

Proof. It follows from $x \in N^+_{\delta,\varepsilon/2}$ that

$$\|x\| > \tfrac{2}{\varepsilon} \quad \text{and} \quad \|f(x)\| < \delta\|x\|^d. \tag{3-1}$$

To derive a contradiction, we assume that $\|x\| > \frac{1}{2}\|f^{-1}(x)\|$. Without loss of generality, we assume that $|x_1| = \|x\|$. Then (we take $\lambda = C$ here)

$$\begin{aligned} &\left|\sum_{j=1}^N P_{1j}(x)F_j(x,1) + \sum_{j=1}^N Q_{1j}(x)G_j(x,1) + R_1(x,1)\right| \\ &\qquad < \max\{C\|x\|^{m-d}\cdot\delta\|x\|^d, C\|x\|^{m-d_-}\cdot 2\|x\|, 2C\|x\|^{m-1}\} \\ &\qquad \le \max\{C\delta\|x\|^m, 2C\|x\|^{m-1}\}. \end{aligned}$$

Using (2-2), we get

$$\text{either} \quad 1 < C\delta \quad \text{or} \quad \|x\| < 2C.$$

However, the former contradicts (2-4). If the latter holds, then Equation (3-1) implies $1 < C\varepsilon$, contradicting (2-4). This completes the proof. □

Lemma 3.3. *For any $x \in \mathbb{A}^N(\Omega)$, one has $f^n(x) \in V^+_{\delta,\varepsilon/2}$ for all sufficiently large n.*

Proof. Note that $\frac{\varepsilon}{2}$ and δ satisfy (2-4) with $\frac{\varepsilon}{2}$ in place of ε. Thus, if $x \in V^+_{\delta,\varepsilon/2}$, then Lemma 2.6 gives $f^n(x) \in V^+_{\delta,\varepsilon/2}$ for all $n \ge 0$.

Suppose now that $x \in N^+_{\delta,\varepsilon/2}$. We take a positive integer n_0 so that $\|x\| \le 2^{n_0+1}/\varepsilon$. We claim that $f^{n_0}(x) \in V^+_{\delta,\varepsilon/2}$. Indeed, if we assume the contrary, then Lemma 3.2 applied to $x, \dots, f^{n_0}(x) \in N^+_{\delta,\varepsilon/2}$ gives

$$\tfrac{2}{\varepsilon} < \|f^{n_0}(x)\| \le \tfrac{1}{2}\|f^{n_0-1}(x)\| \le \cdots \le \frac{1}{2^{n_0}}\|x\|,$$

which contradicts our choice of n_0. Thus, $f^n(x) \in V^+_{\delta,\varepsilon/2}$ for all $n \ge n_0$. □

Proof of Theorem 3.1. (1) We get $\mathcal{K}^+ \subseteq \{x \in \mathbb{A}^N(\Omega) \mid G_f(x) = 0\}$ from Definition 1.2. To show the other inclusion, we assume that $G_f(x) = 0$. Then $G_f(f^n(x)) = d^n G_f(x) = 0$ for all $n \geq 0$. By Lemma 3.3, we take n_0 such that $f^{n_0}(x) \in V^+_{\delta,\varepsilon/2}$. It follows from Lemmas 2.6 and 2.8 (applied to $\frac{\varepsilon}{2}$ in place of ε) that

$$G_f(f^n(x)) \geq \log^+ \| f^n(x) \| + c^+_{\delta,\varepsilon/2}$$

for all $n \geq n_0$. Combined with $G_f(f^n(x)) = 0$, we see that $\|f^n(x)\| \leq \exp(-c^+_{\delta,\varepsilon/2})$ for all $n \geq n_0$. Thus, $\{x \in \mathbb{A}^N(\Omega) \mid G_f(x) = 0\} \subseteq \mathcal{K}^+$.

(2) If $x \notin \mathcal{K}^+$, then $G_f(x) > 0$ by (1). Definition 1.2 then gives $\|f^n(x)\| \to +\infty$ as $n \to +\infty$. □

With f^{-1} in place of f, we put

$$W^- := \{x \in \mathbb{A}^N(\Omega) \mid \|f^{-n}(x)\| \to +\infty \ (n \to +\infty)\},$$
$$\mathcal{K}^- := \{x \in \mathbb{A}^N(\Omega) \mid \{f^{-n}(x)\}_{n=0}^{+\infty} \text{ is bounded with respect to } \|\cdot\|\}.$$

Then we have $\mathbb{A}^N(\Omega) = W^- \amalg \mathcal{K}^-$ as in Theorem 3.1.

In the rest of this section, we give filtrations of $\mathbb{A}^N$ relative to f over nonarchimedean fields as in [Bedford and Smillie 1991, §2.2; Shafikov and Wolf 2003, §3] over $\mathbb{C}$.

We set

$$B_\varepsilon = \{x \in \mathbb{A}^N(\Omega) \mid \|x\| \leq \tfrac{1}{\varepsilon}\},$$
$$U^+_{\delta,\varepsilon} = \{x \in \mathbb{A}^N(\Omega) \mid \|x\| > \tfrac{1}{\varepsilon} \text{ and } \|f(x)\| \geq \delta \|x\|^d\},$$

where δ and ε are constants satisfying (2-4).

Since $\varepsilon \leq 1$ and $\delta/\varepsilon^d \geq C^d \geq 1$ by (2-4), we have

$$U^+_{\delta,\varepsilon} = \left\{x \in \mathbb{A}^N(\Omega) \;\middle|\; \|x\| > \tfrac{1}{\varepsilon} \text{ and } \max\{\|f(x)\|, 1\} \geq \delta \max\{\|x\|, 1\}^d \right\}$$

so that $B_\varepsilon \amalg U^+_{\delta,\varepsilon} = V^+_{\delta,\varepsilon}$.

Proposition 3.4. *We assume that* ε *and* δ *satisfy*

$$\varepsilon^{d-1} \leq \delta \quad \textit{and} \quad \varepsilon^{d_- - 1} \leq \delta \tag{3-2}$$

in addition to (2-4) (*for example, if we take* ε *and* δ *as* (2-5), *then they also satisfy* (3-2)). *Then we have the following*:

(1) $\mathbb{A}^N(\Omega) = B_\varepsilon \amalg U^+_{\delta,\varepsilon} \amalg N^+_{\delta,\varepsilon}$ (*disjoint union*),

(2) $f(U^+_{\delta,\varepsilon}) \subseteq U^+_{\delta,\varepsilon}$ *and* $f(B_\varepsilon \amalg U^+_{\delta,\varepsilon}) \subseteq B_\varepsilon \amalg U^+_{\delta,\varepsilon}$, *and*

(3) $f^{-1}(N^+_{\delta,\varepsilon}) \subseteq N^+_{\delta,\varepsilon}$ *and* $f^{-1}(B_\varepsilon \amalg N^+_{\delta,\varepsilon}) \subseteq B_\varepsilon \amalg N^+_{\delta,\varepsilon}$.

Proof. (1) This is obvious from the definition.

(2) Since $B_\varepsilon \amalg U^+_{\delta,\varepsilon} = V^+_{\delta,\varepsilon}$, we have $f(B_\varepsilon \amalg U^+_{\delta,\varepsilon}) \subseteq B_\varepsilon \amalg U^+_{\delta,\varepsilon}$ by Lemma 2.6.

Suppose that $x \in U^+_{\delta,\varepsilon}$. Then

$$\|f(x)\| \geq \delta\|x\|^d > \frac{\delta}{\varepsilon^d} \geq \tfrac{1}{\varepsilon}, \tag{3-3}$$

where we have used (3-2) in the last inequality. Also since $x \in U^+_{\delta,\varepsilon} \subseteq V^+_{\delta,\varepsilon}$, we have $f(x) \in V^+_{\delta,\varepsilon}$ by Lemma 2.6. Since $f(x) \notin B_\varepsilon$ by (3-3), we get $f(x) \in V^+_{\delta,\varepsilon} \setminus B_\varepsilon = U^+_{\delta,\varepsilon}$. Hence, $f(U^+_{\delta,\varepsilon}) \subseteq U^+_{\delta,\varepsilon}$.

(3) We put

$$\begin{aligned} U^-_{\delta,\varepsilon} &= \{ x \in \mathbb{A}^N(\Omega) \mid \|x\| > \tfrac{1}{\varepsilon} \text{ and } \|f^{-1}(x)\| \geq \delta\|x\|^{d_-} \} \qquad (3\text{-}4)\\ &= \left\{ x \in \mathbb{A}^N(\Omega) \;\middle|\; \|x\| > \tfrac{1}{\varepsilon} \text{ and } \max\{\|f^{-1}(x)\|, 1\} \geq \delta \max\{\|x\|, 1\}^{d_-} \right\}, \end{aligned}$$

where the second equality follows from $\delta/\varepsilon^d_- \geq C^{d_-} \geq 1$ by (2-4). Then as in (2), we have $f^{-1}(U^-_{\delta,\varepsilon}) \subseteq U^-_{\delta,\varepsilon}$. Since $B_\varepsilon \amalg U^-_{\delta,\varepsilon} = V^-_{\delta,\varepsilon}$, Lemma 2.9 implies $N^+_{\delta,\varepsilon} \subseteq U^-_{\delta,\varepsilon}$.

Suppose that $x \in N^+_{\delta,\varepsilon}$. Then

$$f^{-1}(x) \in f^{-1}(N^+_{\delta,\varepsilon}) \subseteq f^{-1}(U^-_{\delta,\varepsilon}) \subseteq U^-_{\delta,\varepsilon}.$$

In particular, $\|f^{-1}(x)\| > \frac{1}{\varepsilon}$ so that $f^{-1}(x) \notin B_\varepsilon$. On the other hand, since $x \notin U^+_{\delta,\varepsilon}$ and $f(U^+_{\delta,\varepsilon}) \subseteq U^+_{\delta,\varepsilon}$, we get $f^{-1}(x) \notin U^+_{\delta,\varepsilon}$, so $f^{-1}(x) \in N^+_{\delta,\varepsilon} = \mathbb{A}^N(\Omega) \setminus (B_\varepsilon \amalg U^+_{\delta,\varepsilon})$. We conclude that $f^{-1}(N^+_{\delta,\varepsilon}) \subseteq N^+_{\delta,\varepsilon}$.

Next we show $f^{-1}(B_\varepsilon \amalg N^+_{\delta,\varepsilon}) \subseteq B_\varepsilon \amalg N^+_{\delta,\varepsilon}$. Since $U^+_{\delta,\varepsilon} = \mathbb{A}^N(\Omega) \setminus (B_\varepsilon \amalg N^+_{\delta,\varepsilon})$, it suffices to show that $f^{-1}(U^+_{\delta,\varepsilon}) \supseteq U^+_{\delta,\varepsilon}$, which is obvious from $f(U^+_{\delta,\varepsilon}) \subseteq U^+_{\delta,\varepsilon}$. □

Proposition 3.5. *We assume that* ε *and* δ *satisfy*

$$\varepsilon^{d-1} < \delta \quad \textit{and} \quad \varepsilon^{d_- -1} < \delta \tag{3-5}$$

in addition to (2-4). *Then we have*

(1) $\bigcup_{n=0}^{+\infty} f^{-n}(U^+_{\delta,\varepsilon}) = W^+$ *and*

(2) $\bigcup_{n=0}^{+\infty} f^{n}(N^+_{\delta,\varepsilon}) = W^-$.

Proof. (1) We set $r := \delta/\varepsilon^{d-1} > 1$. We first show that $U^+_{\delta,\varepsilon} \subseteq W^+$. Indeed, if $x \in U^+_{\delta,\varepsilon}$, then

$$\|f(x)\| \geq \delta\|x\|^d > \frac{\delta}{\varepsilon^{d-1}}\frac{1}{\varepsilon} = r\frac{1}{\varepsilon}.$$

Since $f(U^+_{\delta,\varepsilon}) \subseteq U^+_{\delta,\varepsilon}$, we inductively get $\|f^n(x)\| > r^{(d^n-1)/(d-1)}\frac{1}{\varepsilon}$ for all $n \geq 0$. Hence, $x \in W^+$. This completes the proof of $U^+_{\delta,\varepsilon} \subseteq W^+$. Since $f^{-1}(W^+) = W^+$, we get $f^{-n}(U^+_{\delta,\varepsilon}) \subseteq W^+$ for all $n \geq 0$ so that $\bigcup_{n=0}^{+\infty} f^{-n}(U^+_{\delta,\varepsilon}) \subseteq W^+$.

To show the inclusion $\bigcup_{n=0}^{+\infty} f^{-n}(U^+_{\delta,\varepsilon}) \supseteq W^+$, suppose that $x \notin \bigcup_{n=0}^{+\infty} f^{-n}(U^+_{\delta,\varepsilon})$. We need to show that $x \in \mathcal{K}^+$. Since $f^n(x) \notin U^+_{\delta,\varepsilon}$, we have either $f^n(x) \in B_\varepsilon$ or $f^n(x) \in N^+_{\delta,\varepsilon}$.

Case 1. Suppose there is an $n_0 \geq 0$ such that $f^{n_0}(x) \in B_\varepsilon$. Then $f^{n_0+1}(x) \in B_\varepsilon \amalg U^+_{\delta,\varepsilon}$ by Proposition 3.4(2). Since $f^{n_0+1}(x) \notin U^+_{\delta,\varepsilon}$, we obtain $f^{n_0+1}(x) \in B_\varepsilon$. Inductively, $f^n(x) \in B_\varepsilon$ for all $n \geq n_0$, so we conclude that $x \in \mathcal{K}^+$.

Case 2. Suppose that $f^n(x) \in N^+_{\delta,\varepsilon}$ for all $n \geq 0$. By Lemma 3.3, there is an $n_0 \geq 0$ such that $f^n(x) \in V^+_{\delta,\varepsilon/2}$ for all $n \geq n_0$. Then for all $n \geq n_0$, we have

$$f^n(x) \in V^+_{\delta,\varepsilon/2} \cap N^+_{\delta,\varepsilon} \subseteq \{\, y \in \mathbb{A}^N(\Omega) \mid \tfrac{1}{\varepsilon} < \|y\| \leq \tfrac{2}{\varepsilon} \,\}.$$

Hence, $x \in \mathcal{K}^+$.

In both cases, we have $x \in \mathcal{K}^+$, so we get $\bigcup_{n=0}^{+\infty} f^{-n}(U^+_{\delta,\varepsilon}) \supseteq W^+$.

(2) Let $U^-_{\delta,\varepsilon}$ be the set defined by (3-4). Then $\bigcup_{n=0}^{+\infty} f^n(U^-_{\delta,\varepsilon}) = W^-$ by the argument in (1), and so $\bigcup_{n=0}^{+\infty} f^n(N^+_{\delta,\varepsilon}) \subseteq W^-$. To show the other inclusion, suppose that $x \notin \bigcup_{n=0}^{+\infty} f^n(N^+_{\delta,\varepsilon})$. Then we have either $f^{-n}(x) \in B_\varepsilon$ or $f^{-n}(x) \in U^+_{\delta,\varepsilon}$.

Case 1. If there is an $n_0 \geq 0$ such that $f^{-n_0}(x) \in B_\varepsilon$, then the argument of Case 1 of (1) together with Proposition 3.4(3) gives $f^{-n}(x) \in B_\varepsilon$ for all $n \geq n_0$.

Case 2. Suppose that $f^{-n}(x) \in U^+_{\delta,\varepsilon}$ for all $n \geq 0$. Then the argument of Case 2 of (1) together with Lemma 3.3 with f^{-1} in place of f gives $\frac{1}{\varepsilon} < \|x\| < \frac{2}{\varepsilon}$ for sufficiently large n.

In both cases, we get $x \in \mathcal{K}^-$. Hence, $\bigcup_{n=0}^{+\infty} f^n(N^+_{\delta,\varepsilon}) \supseteq W^-$. □

Remark 3.6. If we take

$$0 < \varepsilon < \frac{1}{C^{\min\{d,d_-\}}} \quad \text{and} \quad \delta = \frac{1}{C^{\min\{d,d_-\}(\min\{d,d_-\}-1)}},$$

then they satisfy both (2-4) and (3-5).

4. Regular automorphisms having good reduction

Morton and Silverman [1994] introduced the notion of having good reduction for endomorphisms of $\mathbb{P}^1$ over Ω, which has been useful in studying endomorphisms of $\mathbb{P}^1$ over a global field. For endomorphisms of $\mathbb{P}^N$ having good reduction, see for example [Kawaguchi and Silverman 2007, Remark 12; 2009]. In this section, we introduce the notion of having good reduction for regular polynomial automorphisms of $\mathbb{A}^N$ over Ω. This notion will be useful in studying regular polynomial automorphisms over a global field in Sections 6 and 7.

As in Section 1, R denotes the ring of integers of Ω. Let M be the maximal ideal of R and $\tilde{k} := R/M$ the residue field. Note that $\tilde{k}$ is algebraically closed since Ω is algebraically closed.

Definition 4.1 (Good reduction). Let $f = (f_1, \dots, f_N) : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism over an algebraically closed field Ω with nontrivial nonarchimedean absolute value, and let $f^{-1} = (g_1, \dots, g_N) : \mathbb{A}^N \to \mathbb{A}^N$ denote its inverse. We write d and d_- for the degrees of f and f^{-1}, respectively. We say that f *has good reduction* if the following three conditions are satisfied:

(i) We have that f extends to the polynomial automorphism $f : \mathbb{A}^N_R \to \mathbb{A}^N_R$ over R, so both $f_1(X), \dots, f_N(X)$ and $g_1(X), \dots, g_N(X)$ are in $R[X_1, \dots, X_N]$.

(ii) Let $\tilde{f} = (\widetilde{f_1}, \dots, \widetilde{f_N}) : \mathbb{A}^N_{\tilde{k}} \to \mathbb{A}^N_{\tilde{k}}$ and $\widetilde{f^{-1}} = (\widetilde{g_1}, \dots, \widetilde{g_N}) : \mathbb{A}^N_{\tilde{k}} \to \mathbb{A}^N_{\tilde{k}}$ be the induced polynomial automorphisms over $\tilde{k}$. Then the degrees of $\tilde{f}$ and $\widetilde{f^{-1}}$ are equal to d and d_-, respectively.

(iii) We have that $\tilde{f}$ is regular (see Remark 2.2).

We give some equivalent conditions for regular polynomial automorphisms f to have good reduction. As in Section 1, let $F_i(X, T)$ and $G_j(X, T)$ be the homogenization of $f_i(X)$ and $g_j(X)$. If $F_i(X, T)$ and $G_j(X, T)$ are defined over R, $\widetilde{F_i}(X, T)$ and $\widetilde{G_j}(X, T)$ denote their reductions to $\tilde{k}$. Let $\rho : R \to \tilde{k}$ be the natural map.

Proposition 4.2. *Let f be a regular polynomial automorphism of $\mathbb{A}^N$ over Ω. Assume that f satisfies the conditions (i) and (ii) of Definition 4.1. Then the following are equivalent*:

(1) *We have that f has good reduction*, *i.e.*, *f also satisfies Definition 4.1(iii).*

(2) *As ideals in $R[X_1, \dots, X_N, T]$*, *one has*

$$(X_1, \dots, X_N, T)^k \subseteq (F_1(X, T), \dots, F_N(X, T), G_1(X, T), \dots, G_N(X, T), T)$$

for some integer $k \geq 1$.

(3) *As ideals in $R[X_1, \dots, X_N]$*, *one has*

$$(X_1, \dots, X_N)^\ell \subseteq (F_1(X, 0), \dots, F_N(X, 0), G_1(X, 0), \dots, G_N(X, 0))$$

for some integer $\ell \geq 1$.

Proof. (1) $\Longrightarrow$ (3). It suffices to show that

$$(X_1, \dots, X_N)^\ell \subseteq (F_1(X, 0)^{d_-}, \dots, F_N(X, 0)^{d_-}, G_1(X, 0)^d, \dots, G_N(X, 0)^d) \tag{4-1}$$

for some $\ell \geq 1$. We set

$$I = \left\{ r \in R \;\middle|\; \begin{array}{c} \text{there is an } \ell \geq 1 \text{ such that } r(X_1, \dots, X_N)^\ell \subseteq \\ (F_1(X, 0)^{d_-}, \dots, F_N(X, 0)^{d_-}, G_1(X, 0)^d, \dots, G_N(X, 0)^d) \end{array} \right\}.$$

Since f is regular, I is a nonzero ideal of R.

We claim that $\rho(I) \neq 0$. Indeed, suppose that $\rho(I) = 0$. Then elimination theory tells us that there is a point $x = (x_1 : \cdots : x_n) \in \mathbb{P}^{N-1}(\tilde{k})$ such that $\widetilde{F}_i(x, 0) = 0$ and $\widetilde{G}_j(x, 0) = 0$ for all i and j; see [Kawaguchi and Silverman 2007, Theorem 6]. Since f satisfies condition (ii), $\widetilde{F}_i(X, T)$ and $\widetilde{G}_j(X, T)$ are the homogenizations of $\widetilde{f}_i$ and $\widetilde{g}_j$, respectively. Then the existence of such an $x \in \mathbb{P}^{N-1}(\tilde{k})$ contradicts condition (iii), which yields the claim.

Since $\rho(I) \neq 0$, there is an $r \in I$ such that $r \in R^{\times} = R \setminus M$. Then $I = R$, and we obtain Equation (4-1).

(3) $\Longrightarrow$ (1). The assumption of (3) gives, as ideals in $\tilde{k}[X]$,

$$(X_1, \ldots, X_N)^{\ell} \subseteq (\rho(F_1(X,0)), \ldots, \rho(F_N(X,0)), \rho(G_1(X,0)), \ldots, \rho(G_N(X,0))).$$

Since $\rho(F_i(X, 0)) = \widetilde{F}_i(X, 0)$ and $\rho(G_j(X, 0)) = \widetilde{G}_j(X, 0)$, we obtain that $\tilde{f}$ is regular.

(2) $\Longrightarrow$ (3). We have only to put $T = 0$.

(3) $\Longrightarrow$ (2). It suffices to show that for any $\alpha = 1, \ldots, N$, there are an integer $k \geq 1$ and polynomials $P_i(X, T)$, $Q_j(X, T)$, and $R(X, T)$ defined over R such that

$$X_\alpha^k = \sum_{i=1}^{N} P_i(X, T)F(X, T) + \sum_{j=1}^{N} Q_j(X, T)G_j(X, T) + TR(X, T). \tag{4-2}$$

By the assumption of (iii), there are an integer $\ell \geq 1$ and polynomials $P_i(X)$ and $Q_j(X)$ defined over R such that

$$X_\alpha^\ell = \sum_{i=1}^{N} P_i(X)F(X, 0) + \sum_{j=1}^{N} Q_j(X)G_j(X, 0).$$

We set $k := \ell$, $P_i(X, T) := P_i(X)$, and $Q_j(X, T) := Q_j(X)$. Then

$$X_\alpha^k - \sum_{i=1}^{N} P_i(X, T)F(X, T) - \sum_{j=1}^{N} Q_j(X, T)G_j(X, T)$$

is a polynomial in $R[X, T]$ that is divisible by T. Hence, there is a polynomial $R(X, T)$ in $R[X, T]$ satisfying Equation (4-2). □

Suppose now that a regular polynomial automorphism f has good reduction. By Proposition 4.2, for each $1 \leq i \leq N$ there are polynomials $P_{ij}(X)$ and $Q_{ij}(X)$ in $R[X]$ that satisfy (2-1). Then the polynomial $R_i(X, T)$ in (2-2) is also defined over R, and the constant C in (2-3) is equal to 1. This means that $\varepsilon = 1$ and $\delta = 1$ satisfy (2-4). It follows that when f has good reduction, G_f and $\log^+\|\cdot\|$ are related simply.

Proposition 4.3. *Suppose that f has good reduction.*

(1) $G_f(\cdot) \le \log^+\|\cdot\|$ *and* $G_{f^{-1}}(\cdot) \le \log^+\|\cdot\|$ *on* $\mathbb{A}^N(\Omega)$.

(2) $\log^+\|\cdot\| = G_f(\cdot)$ *on* $V^+_{1,1}$ *and* $\log^+\|\cdot\| = G_{f^{-1}}(\cdot)$ *on* $V^-_{1,1}$. *Moreover,* $\mathbb{A}^N(\Omega) = V^+_{1,1} \cup V^-_{1,1}$.

Proof. (1) Since the $f_i(X)$ are defined over R, in the proof of Lemma 1.3 we may take $r = 1$ so that $c_f = 0$. Thus, $G_f(\cdot) \le \log^+\|\cdot\|$ on $\mathbb{A}^N(\Omega)$. The estimate for $G_{f^{-1}}$ is similar.

(2) Since $\varepsilon = 1$ and $\delta = 1$ satisfy (2-4), Lemma 2.9 gives $\mathbb{A}^N(\Omega) = V^+_{1,1} \cup V^-_{1,1}$. The constant $c^+_{1,1}$ in Lemma 2.8 is equal to 0, and thus, combined with (1), we have $\log^+\|x\| = G_f(x)$ for all $x \in V^+_{1,1}$. The estimate for $G_{f^{-1}}$ is similar. □

5. Green functions for regular automorphisms over $\mathbb{C}$

In this section, we remark that the proof of Theorem 2.3 gives a different proof (more explicit and without compactness arguments) of the corresponding estimates of Green functions over $\mathbb{C}$.

We write the usual absolute value of $\mathbb{C}$ for $|\cdot|_\infty$, and we set $\|x\|_\infty := \max_i\{|x_i|_\infty\}$ for $x = (x_1, \dots, x_N) \in \mathbb{A}^N(\mathbb{C})$.

Let $f = (f_1, \dots, f_N) : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism of degree $d \ge 2$ defined over $\mathbb{C}$. Then the Green function for f is defined by [Sibony 1999, §2]

$$G_f(x) := \lim_{n\to+\infty} \frac{1}{d^n} \log^+\|f^n(x)\| \quad \text{for } x \in \mathbb{A}^N(\mathbb{C}). \tag{5-1}$$

Let $\|f\|_\infty$ be the maximum of the absolute values of all the coefficients of $f_i(X)$ for $1 \le i \le N$, and set $c_{f,\infty} = \frac{1}{d-1}\log\max\left\{\binom{N+d-1}{d}\|f\|_\infty, 1\right\}$. Note that $\binom{N+d-1}{d}$ is the number of monomials of degree d in the ring of homogeneous polynomials in N variables. Since

$$\log^+\|f(x)\| \le d\log^+\|x\| + \log\max\left\{\binom{N+d-1}{d}\|f\|_\infty, 1\right\}, \tag{5-2}$$

we get

$$G_f(x) \le \log^+\|x\| + c_{f,\infty} \quad \text{for any } x \in \mathbb{A}^N(\mathbb{C}). \tag{5-3}$$

Let $P_{ij}(X), Q_{ij}(X) \in \mathbb{C}[X]$ and $R(X, T) \in \mathbb{C}[X, T]$ be polynomials satisfying (2-2). As before, we may and do assume that the $P_{ij}(X)$, $Q_{ij}(X)$, and $R_i(X, T)$ are homogeneous polynomials with degree $m - d$, $m - d_-$, and $m - 1$, respectively. We write $\|P\|_\infty$ for the maximum of the absolute values of all the coefficients of $P_{ij}(X)$ for $1 \le i \le N$ and $1 \le j \le N$, and we write $\|Q\|_\infty$ and $\|R\|_\infty$ similarly. We set

$$C'_\infty = \max\left\{\binom{N+m-d-1}{m-d}\|P\|_\infty, \binom{N+m-d_--1}{m-d_-}\|Q\|_\infty, \binom{N+m}{m-1}\|R\|_\infty, 1\right\}.$$

We note the above formula for C'_∞ is not as explicit as in the nonarchimedean case since it involves the coefficients of P, Q, and R and not only those of F and G. However, $\|P\|_\infty$, $\|Q\|_\infty$, and $\|R\|_\infty$ can be expressed in terms of F and G via an effective version of Hilbert's Nullstellensatz (see [Masser and Wüstholz 1983, Chapter 4] for example).

We put

$$C_\infty = (2N+1)C'_\infty.$$

Fix real numbers $\varepsilon > 0$ and $\delta > 0$ satisfying (2-4) with C_∞ in place of C. We define $N^{\pm}_{\delta,\varepsilon}$ and $V^{\pm}_{\delta,\varepsilon}$ by (2-6) and (2-11) with $\mathbb{C}$ in place of Ω. Then exactly as in Theorem 2.3, we have the following:

Theorem 5.1. *Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism over $\mathbb{C}$.*

(i) $G_f(\,\cdot\,) \geq \log^+\|\cdot\| + c^+_{\delta,\varepsilon}$ *on* $V^+_{\delta,\varepsilon}$.

(ii) $G_{f^{-1}}(\,\cdot\,) \geq \log^+\|\cdot\| + c^-_{\delta,\varepsilon}$ *on* $V^-_{\delta,\varepsilon}$.

(iii) $V^+_{\delta,\varepsilon} \cup V^-_{\delta,\varepsilon} = \mathbb{A}^N(\Omega)$.

6. Global theory of regular automorphisms

In this section, we turn our attention to regular automorphisms over a number field.

Let K be a number field and O_K its ring of integers. We fix an embedding $K \subset \overline{K}$ into an algebraic closure. Let M_K be the set of absolute values on K. We extend the absolute values on K to those on $\overline{K}$.

Let L be a finite extension field of K. For $x \in \mathbb{A}^N(L)$, we define

$$h(x) = \sum_{v \in M_K} n_v \log^+\|x\|_v, \tag{6-1}$$

where $n_v = [L_v : K_v]/[L : K]$. This gives rise to the logarithmic Weil *height function*

$$h : \mathbb{A}^N(\overline{K}) \to \mathbb{R}.$$

For more details on height functions, we refer the reader to [Bombieri and Gubler 2006; Hindry and Silverman 2000; Lang 1983].

Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism over $\overline{K}$ (see Remark 2.2). If the coefficients of f are all defined over K, then we say that f is a regular polynomial automorphism over K.

Lemma 6.1. *If $f : \mathbb{A}^N \to \mathbb{A}^N$ is a polynomial automorphism over K, then the coefficients of f^{-1} are also all defined over K.*

Proof. We take a finite Galois extension field L of K such that the coefficients of f^{-1} are elements of L. For every $\sigma \in \mathrm{Gal}(L/K)$, the uniqueness of the inverse gives $(f^{-1})^\sigma = f^{-1}$. Thus, the coefficients of f^{-1} are in fact elements of K. □

In [Kawaguchi 2006], we constructed (global) canonical height functions $\hat{h}_f^+$ and $\hat{h}_f^-$ for polynomial automorphisms f over K under the assumption that there exists a constant $c \geq 0$ such that

$$\frac{1}{d}h(f(x)) + \frac{1}{d_-}h(f^{-1}(x)) \geq \Big(1 + \frac{1}{dd_-}\Big)h(x) - c \tag{6-2}$$

for all $x \in \mathbb{A}^N(\overline{K})$, where d and d_- denote the degrees of f and f^{-1}. (We showed in op. cit. that (6-2) holds for regular polynomial automorphisms in dimension $N = 2$ by a global method, i.e., a method using the effectiveness of a certain divisor on a certain rational surface.)

In the following, using properties of local Green functions studied in the previous sections, we will first construct in Theorem 6.3 (global) canonical height functions h_f^+ and h_f^- for regular polynomial automorphisms. Indeed, we will construct h_f^+ and h_f^- as appropriate sums of local Green functions. Then we show local versions of (6-2) for all places v, and summing them up, we will obtain (6-2) for regular polynomial automorphisms in any dimension $N \geq 2$ in Theorem 7.1.

For a finite subset S of M_K that contains all the archimedean absolute values of K, we let $O_{K,S}$ denote the ring of S-integers:

$$O_{K,S} = \{\, x \in K \mid \|x\|_v \leq 1 \text{ for all } v \notin S \,\}.$$

Proposition 6.2. *Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism of degree $d \geq 2$ over a number field K. Then there exists a finite subset S of M_K that contains all the archimedean absolute values of K with the following property*: *for all $v \notin S$, f induces a regular polynomial automorphism over $\overline{K}_v$ that has good reduction.*

Proof. We write $f = (f_1, \dots, f_N)$ and let $F_i(X, T) \in K[X, T]$ be the homogenization of f_i. Let d_- denote the degree of $f^{-1} = (g_1, \dots, g_N)$, and in virtue of Lemma 6.1, let $G_j(X, T) \in K[X, T]$ be the homogenization of g_j. Then there are an integer m and homogeneous polynomials $P_{ij}(X) \in K[X]$ of degree $m - d$, $Q_{ij}(X) \in K[X]$ of degree $m - d_-$, and $R_i(X, T) \in K[X, T]$ of degree $m - 1$ such that (2-2) holds as polynomials in $K[X, T]$.

We take a finite subset S of M_K that contains all the archimedean absolute values of K with the following properties:

(i) The coefficients of $F_i(X, T)$, $G_j(X, T)$, $P_{ij}(X)$, $Q_{ij}(X)$, and $R_i(X, T)$ are all in $O_{K,S}$.

(ii) For $v \notin S$, we let $\rho_v : O_{K,S} \to \tilde{k}_v$ denote the natural map, where $\tilde{k}_v$ is the residue field of $(O_K)_v$. Then $\deg(f) = \deg(\rho_v(f))$ and $\deg(f^{-1}) = \deg(\rho_v(f^{-1}))$.

Then for any $v \notin S$, $f \times_K \overline{K}_v : \mathbb{A}^N_{\overline{K}_v} \to \mathbb{A}^N_{\overline{K}_v}$ satisfies the properties (i) and (ii) of Definition 4.1 and (3) of Proposition 4.2. Hence, $f \times_K \overline{K}_v$ has good reduction. □

Theorem 6.3. *Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism of degree $d \geq 2$ over a number field K. Let $d_- \geq 2$ denote the degree of f^{-1}.*

(1) *For all $x \in \mathbb{A}^N(\overline{K})$, the limits*

$$\lim_{n\to+\infty} \frac{1}{d^n} h(f^n(x)) \quad \textit{and} \quad \lim_{n\to+\infty} \frac{1}{d_-^n} h(f^{-n}(x))$$

exist. We write $\hat{h}_f^+(x)$ and $\hat{h}_f^-(x)$ for the limits, respectively.

(2) *(Global-to-local decomposition) For each place $v \in M_K$, let $G_{f,v}$ and $G_{f^{-1},v}$ be the Green functions for f and f^{-1} at v, respectively. Then for all $x \in \mathbb{A}^N(\overline{K})$,*

$$\hat{h}_f^+(x) = \sum_{v\in M_K} n_v G_{f,v}(x) \quad \textit{and} \quad \hat{h}_f^-(x) = \sum_{v\in M_K} n_v G_{f^{-1},v}(x).$$

(3) *We define $\hat{h}_f : \mathbb{A}^N(\overline{K}) \to \mathbb{R}$ by*

$$\hat{h}_f := \hat{h}_f^+ + \hat{h}_f^-.$$

Then $\hat{h}_f$ satisfies the following two conditions:

(3i) $\frac{1}{d}\hat{h}_f \circ f + \frac{1}{d_-}\hat{h}_f \circ f^{-1} = (1 + \frac{1}{dd_-})\hat{h}_f$ *on* $\mathbb{A}^N(\overline{K})$ *and*

(3ii) $h + O(1) \leq \hat{h}_f \leq 2h + O(1)$ *on* $\mathbb{A}^N(\overline{K})$.

(4) *The function $\hat{h}_f$ has the following uniqueness property: if $h' : \mathbb{A}^N(\overline{K}) \to \mathbb{R}$ is a function satisfying the condition (3i) such that $h' = \hat{h}_f + O(1)$, then $h' = \hat{h}_f$.*

(5) *The functions $\hat{h}_f^+$, $\hat{h}_f^-$, and $\hat{h}_f$ are nonnegative. Further, for $x \in \mathbb{A}^N(\overline{K})$ we have*

$$\hat{h}_f(x) = 0 \iff \hat{h}_f^+(x) = 0 \iff \hat{h}_f^-(x) = 0 \iff x \textit{ is } f\textit{-periodic.}$$

Proof. For each $v \in M_K$, we have estimates of Green functions for f at v as in Lemmas 1.3 and 2.8. We use the suffix v when we work over the absolute value $v \in M_K$. For example, the Green function for f at v is denoted $G_{f,v}$ and constants c_f and $c_{\varepsilon,\delta}^{\pm}$ in Lemmas 1.3 and 2.8 and (2-12) are denoted $c_{f,v}$ and $c_{\varepsilon,\delta,v}^{\pm}$, respectively.

Let S be a finite subset of M_K as in Proposition 6.2.

(1)(2) We fix $x \in \mathbb{A}^N(\overline{K})$. We will show the existence of $h_f^+(x)$ and the decomposition $h_f^+(x) = \sum_{v\in M_K} n_v G_{f,v}(x)$. The existence and decomposition for $h_f^-(x)$ are shown similarly.

For $v \in M_K$ and $n \geq 0$, we set

$$G_{v,n}^+(x) := \frac{1}{d^n} \log^+ \| f^n(x) \|_v.$$

Then the following are true:

- We have $0 \le G^+_{v,n}(x) \le \log^+\|x\|_v + c_{f,v}$ for all $v \in M_K$ and $n \ge 0$ from Proposition 1.1, Lemma 1.3, and Equations (5-2) and (5-3). Indeed, if v is nonarchimedean, then with r in the proof of Proposition 1.1, we have only to set $c_{f,v} = -\frac{1}{d-1}\log|r|$. If v is archimedean, then by (5-2) we have only to set $c_{f,v} = \frac{1}{d-1}\log\max\bigl\{\binom{N+d-1}{d}\|f\|_\infty, 1\bigr\}$.
- We have $\lim_{n\to+\infty} G^+_{v,n}(x) = G_{f,v}(x)$ from Definition 1.2 and Equation (5-1).
- We have $\frac{1}{d^n}h(f^n(x)) = \sum_{v\in M_K} n_v G^+_{v,n}(x)$ from Equation (6-1).
- We may take $c_{f,v} = 0$ for any $v \notin S$ from Propositions 4.3 and 6.2.
- We have $\sum_{v\in M_K} n_v(\log^+\|x\|_v + c_{f,v}) = h(x) + \sum_{v\in S} n_v c_{f,v} < +\infty$.

Lebesgue's dominated convergence theorem then implies that $\sum_{v\in M_K} n_v G^+_{v,n}(x)$ converges as $n \to +\infty$ and that

$$\begin{aligned}\lim_{n\to+\infty}\frac{1}{d^n}h(f^n(x)) &= \lim_{n\to+\infty}\sum_{v\in M_K} n_v G^+_{v,n}(x)\\ &= \sum_{v\in M_K}\lim_{n\to+\infty} n_v G^+_{v,n}(x) = \sum_{v\in M_K} n_v G_{f,v}(x).\end{aligned}$$

This completes the proof of (1) and (2)

(3)(4)(5) First we have

$$\begin{aligned}\hat{h}_f(x) &= \sum_{v\in M_K} n_v G_{f,v}(x) + \sum_{v\in M_K} n_v G_{f^{-1},v}(x) \\ &\le \sum_{v\in M_K} n_v(2\log^+\|x\|_v + c_{f,v} + c_{f^{-1},v}) = 2\hat{h}(x) + \sum_{v\in S} n_v(c_{f,v} + c_{f^{-1},v}).\end{aligned} \tag{6-3}$$

On the other hand, we have

- $\min\{c^+_{\varepsilon,\delta,v}, c^-_{\varepsilon,\delta,v}\} + \log^+\|x\| \le G_{f,v}(x) + G_{f^{-1},v}(x)$ from Lemma 2.8, (2-12), and Theorem 5.1 and
- for any $v \notin S$, we may take $\varepsilon = 1$ and $\delta = 1$ and $\min\{c^+_{1,1,v}, c^-_{1,1,v}\} = 0$ from Propositions 4.3 and 6.2.

Then

$$\begin{aligned}\hat{h}_f(x) &= \sum_{v\in M_K} n_v G_{f,v}(x) + \sum_{v\in M_K} n_v G_{f^{-1},v}(x) \\ &\ge \sum_{v\in M_K} n_v(\log^+\|x\|_v + \min\{c^+_{\varepsilon,\delta,v}, c^-_{\varepsilon,\delta,v}\}) = \hat{h}_{nv}(x) + \sum_{v\in S} n_v \min\{c^+_{\varepsilon,\delta,v}, c^-_{\varepsilon,\delta,v}\}.\end{aligned} \tag{6-4}$$

Equations (6-3) and (6-4) give (3ii). For the rest of the proof, see [Kawaguchi 2006, Theorem 4.2(2–4)]. □

Remark 6.4. Theorem 6.3(1) shows that $\left\{\frac{1}{d^n}h(f^n(x))\right\}_{n=0}^{+\infty}$ and $\left\{\frac{1}{d_-^n}h(f^{-n}(x))\right\}_{n=0}^{+\infty}$ are convergent sequences, which gives an improvement of [Kawaguchi 2006] since we replace lim sup by lim in the definition of $\hat{h}_f^{\pm}$.

We now introduce another function

$$\tilde{h}_f(x) := \sum_{v\in M_K} n_v \max\{G_{f,v}(x),\, G_{f^{-1},v}(x)\} \tag{6-5}$$

for $x \in \mathbb{A}^N(\overline{K})$. The next proposition shows that $\tilde{h}_f$ also behaves well relative to f.

Proposition 6.5. (1) *On $\mathbb{A}^N(\overline{K})$, $\tilde{h}_f = h + O(1)$.*

(2) *For $x \in \mathbb{A}^N(\overline{K})$, we have $\tilde{h}_f(x) = 0$ if and only if $\hat{h}_f(x) = 0$.*

Proof. (1) We use the notation of the proof of Theorem 6.3. By Lemmas 1.3 and 2.8, Equations (2-12) and (5-3), and Theorem 5.1, we have

$$\log^+\|x\|_v + \min\{c^+_{\varepsilon,\delta,v}, c^-_{\varepsilon,\delta,v}\} \le \max\{G_{f,v}(x), G_{f^{-1},v}(x)\} \le \log^+\|x\|_v + \max\{c_{f,v}, c_{f^{-1},v}\}.$$

Summing up over all places v, we get

$$h(x) + \sum_{v\in M_K} n_v \min\{c^+_{\varepsilon,\delta,v}, c^-_{\varepsilon,\delta,v}\} \le \tilde{h}_f(x) \le h(x) + \sum_{v\in M_K} n_v \max\{c_{f,v}, c_{f^{-1},v}\}.$$

Since we have $c_{f,v} = c_{f^{-1},v} = c^+_{\varepsilon,\delta,v} = c^-_{\varepsilon,\delta,v} = 0$ except for finitely many v (indeed for every $v \notin S$), this gives the assertion.

(2) Since $G_{f,v}$ and $G_{f^{-1},v}$ are nonnegative functions, we see that $\tilde{h}_f(x) = 0$ if and only if $G_{f,v}(x) = G_{f^{-1},v}(x) = 0$ for every $v \in M$ if and only if $\hat{h}_f(x) = 0$. □

7. Arithmetic properties of regular polynomial automorphisms

In this section, we give some applications of local and global canonical height functions. The first application is the following theorem on the usual height function [Kawaguchi 2006, §4; Silverman 2006, Conjecture 3; 2007, Conjecture 7.18], which is independently obtained by Lee [2013] via a different method (global method based on the effectiveness of a certain divisor as in the case of $N = 2$ in [Kawaguchi 2006]).

Theorem 7.1. *Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism over a number field K. Let d and d_- be the degrees of f and f^{-1}.*

(1) *There exists a constant $c \ge 0$ such that*

$$\frac{1}{d}h(f(x)) + \frac{1}{d_-}h(f^{-1}(x)) \ge \left(1 + \frac{1}{dd_-}\right)h(x) - c$$

for all $x \in \mathbb{A}^N(\overline{K})$.

(2) *We have*

$$\liminf_{\substack{x\in\mathbb{A}^N(\overline{K})\\ h(x)\to\infty}} \frac{\frac{1}{d}h(f(x))+\frac{1}{d_-}h(f^{-1}(x))}{h(x)} = 1+\frac{1}{dd_-}. \tag{7-1}$$

Proof. (1) We set

$$\widetilde{G}_{f,v} := \max\{G_{f,v},\, G_{f^{-1},v}\}.$$

Claim 7.1.1. *For all $x\in\mathbb{A}^N(\overline{K})$, we have*

$$\frac{1}{d}\widetilde{G}_{f,v}(f(x))+\frac{1}{d_-}\widetilde{G}_{f,v}(f^{-1}(x)) \geq \Bigl(1+\frac{1}{dd_-}\Bigr)\widetilde{G}_{f,v}(x). \tag{7-2}$$

We first show that Claim 7.1.1 implies (1). Indeed, we assume Claim 7.1.1. Then summing up over all v, we have

$$\frac{1}{d}\tilde{h}(f(x))+\frac{1}{d_-}\tilde{h}(f^{-1}(x)) \geq \Bigl(1+\frac{1}{dd_-}\Bigr)\tilde{h}(x). \tag{7-3}$$

Since $\tilde{h}_f = h+O(1)$ by Proposition 6.5(1), Equation (7-3) yields (1).

To show Claim 7.1.1, for notational convenience let $A=G_{f,v}(x)$, $B=G_{f^{-1},v}(x)$, and $\gamma=\frac{1}{dd_-}$. Then the definition of $\widetilde{G}_{f,v}$ and $\widetilde{G}_{f^{-1},v}$ and the functional equation of $G_{f,v}(x)$ and $G_{f^{-1},v}(x)$ show that the equality (7-2) is equivalent to

$$\max\{A,\gamma B\}+\max\{\gamma A, B\} \geq (1+\gamma)\max\{A, B\}. \tag{7-4}$$

But the left-hand side of (7-4) is

$$\max\{(1+\gamma)A,\, A+B,\, \gamma(A+B),\, (1+\gamma)B\},$$

which is clearly greater than or equal to the right-hand side of (7-4). This completes the proof of Claim 7.1.1 and hence the proof of Theorem 7.1(1).

(2) From (1), we obtain

$$\liminf_{\substack{x\in\mathbb{A}^N(\overline{K})\\ h(x)\to\infty}} \frac{\frac{1}{d}h(f(x))+\frac{1}{d_-}h(f^{-1}(x))}{h(x)} \geq 1+\frac{1}{dd_-}.$$

On the other hand, it is shown in [Kawaguchi 2006, Proposition 4.4] that for any polynomial automorphism $f:\mathbb{A}^N\to\mathbb{A}^N$, one has

$$\liminf_{\substack{x\in\mathbb{A}^N(\overline{K})\\ h(x)\to\infty}} \frac{\frac{1}{d}h(f(x))+\frac{1}{d_-}h(f^{-1}(x))}{h(x)} \leq 1+\frac{1}{dd_-}. \tag{7-5}$$

Combining these two inequalities gives the assertion. □

Remark 7.2. It is shown in [Kawaguchi 2006, Theorem 4.4] that the equality (7-1) holds in dimension $N=2$ for regular polynomial automorphisms. Theorem 7.1(2) asserts that the equality holds in any dimension $N \geq 2$ for regular polynomial automorphisms.

Theorem 6.3 recovers the following theorem on f-periodic points.

Corollary 7.3 [Marcello 2000]. *Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism over a number field K. Then the set of f-periodic points in $\mathbb{A}^N(\overline{K})$ is a set of bounded height. In particular, for any integer $D \geq 1$ the set*

$$\{\, x \in \mathbb{A}^N(\overline{K}) \mid x \textit{ is } f\textit{-periodic},\ [K(x):K] \leq D \,\}$$

is finite.

Proof. By Theorem 6.3(3ii)(5), $\hat{h}_f$ satisfies $\hat{h}_f \gg\ll h$, and a point $x \in \mathbb{A}^N(\overline{K})$ is f-periodic if and only if $\hat{h}_f(x) = 0$. Thus, we get the assertion. □

For a non-f-periodic point x, let $O_f(x) := \{\, f^n(x) \mid n \in \mathbb{Z} \,\}$ denote the f-orbit of x. We define the canonical height of the orbit $O_f(x)$ by

$$\hat{h}_f(O_f(x)) = \frac{\log \hat{h}_f^+(x)}{\log d} + \frac{\log \hat{h}_f^-(x)}{\log d_-}. \tag{7-6}$$

We note that for any integer n, Theorem 6.3 implies that

$$\begin{aligned}\frac{\log \hat{h}_f^+(f^n(x))}{\log d} + \frac{\log \hat{h}_f^-(f^n(x))}{\log d_-} &= \frac{\log d^n \hat{h}_f^+(x)}{\log d} + \frac{\log d_-^{-n} \hat{h}_f^-(x)}{\log d_-} \\ &= \frac{\log \hat{h}_f^+(x)}{\log d} + \frac{\log \hat{h}_f^-(x)}{\log d_-}.\end{aligned}$$

Thus, the value $\hat{h}_f(O_f(x))$ depends only on the orbit $O_f(x)$ and not the particular choice of the point x in the orbit. The next corollary gives a refinement of [Marcello 2003, Corollary B].

Corollary 7.4. *Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a regular polynomial automorphism over a number field K. Let d and d_- be the degrees of f and f^{-1}. Then for any infinite orbit $O_f(x)$,*

$$\#\{\, y \in O_f(x) \mid h(y) \leq T \,\} = \left(\frac{1}{\log d} + \frac{1}{\log d_-}\right) \log T - \hat{h}(O_f(x)) + O(1)$$

as $T \to +\infty$. Here the $O(1)$ bound depends only f, independent of the orbit $O_f(x)$.

Proof. Since f satisfies (7-3), we apply [Kawaguchi 2006, Theorem 5.2]. □

In the rest of this section, we consider some global-to-local arithmetic properties. Suppose that f is a regular polynomial automorphism. By Theorem 6.3(2)(5), $x \in \mathbb{A}^N(\overline{K})$ is f-periodic if and only if $G_{f,v}(x) = 0$ for all $v \in M_K$. By Theorem 3.1 for nonarchimedean v and [Sibony 1999, §2] for archimedean v, $G_{f,v}(x) = 0$ is equivalent to $\{f^n(x)\}_{n=0}^{+\infty}$ being bounded with respect to $\|\cdot\|_v$. Thus, we see that $x \in \mathbb{A}^N(\overline{K})$ is f-periodic if and only if $\{f^n(x)\}_{n=0}^{+\infty}$ is bounded with respect to $\|\cdot\|_v$ for all $v \in M_K$.

This actually holds for any polynomial map f, replacing f-periodic points by f-preperiodic points (see [Call and Goldstine 1997, Corollary 6.3] for $N = 1$).

Proposition 7.5. *Let $f : \mathbb{A}^N \to \mathbb{A}^N$ be a polynomial map over a number field K. For $x \in \mathbb{A}^N(\overline{K})$, the following are equivalent*:

(i) *x is f-preperiodic and*

(ii) *for every $v \in M_K$, $\{f^n(x)\}_{n=0}^{+\infty}$ is bounded with respect to the v-adic topology.*

Proof. Taking a finite extension field of K over which x is defined if necessary, we may assume that x is defined over K. It is obvious that (i) implies (ii). We assume (ii) and show (i). We take a finite subset S of M_K containing the set of all archimedean absolute values such that x and f are defined over $O_{K,S}$. Then for any $v \notin S$, we have

$$\|f^n(x)\|_v \leq 1 \quad \text{for all } n \geq 0.$$

Since we assume (ii), there is a constant C_v for each $v \in S$ such that

$$\|f^n(x)\|_v \leq C_v \quad \text{for all } n \geq 0.$$

Then we have

$$h(f^n(x)) = \sum_{v \in M_K} n_v \log^+ \|f^n(x)\| \leq \sum_{v \in S} n_v C_v \quad \text{for all } n \geq 0.$$

Then

$$\{\, f^n(x) \mid n \geq 0 \,\} \subseteq \left\{ y \in \mathbb{A}^N(K) \;\middle|\; h(y) \leq \sum_{v \in S} n_v C_v \right\}.$$

Since the latter set is finite, the set $\{f^n(x)\}_{n \geq 0}$ is finite, so x is f-preperiodic. □

Acknowledgments

My deep thanks go to Professor Joseph H. Silverman for his encouragement and discussions. Some preliminary work was done while I was staying at the IMJ in Paris and the CRM in Barcelona from October of 2005 to July of 2006. I thank Professors Tien-Cuong Dinh, Charles Favre, and Nessim Sibony for helpful conversations and with whom I had opportunities to talk during my stay in Paris. Finally, I would like to dedicate this paper in memory of Professor Masaki Maruyama, who was my thesis adviser. I often recall his smiling face and loud voice. I thank the referee for carefully reading the manuscript and giving many kind and helpful comments.

References

[Bedford and Smillie 1991] E. Bedford and J. Smillie, "Polynomial diffeomorphisms of $\mathbb{C}^2$: currents, equilibrium measure and hyperbolicity", *Invent. Math.* **103**:1 (1991), 69–99. MR 92a:32035 Zbl 0721.58037

[Bombieri and Gubler 2006] E. Bombieri and W. Gubler, *Heights in Diophantine geometry*, New Mathematical Monographs **4**, Cambridge University Press, 2006. MR 2007a:11092 Zbl 1115.11034

[Call and Goldstine 1997] G. S. Call and S. W. Goldstine, "Canonical heights on projective space", *J. Number Theory* **63**:2 (1997), 211–243. MR 98c:11060 Zbl 0895.14006

[Denis 1995] L. Denis, "Points périodiques des automorphismes affines", *J. Reine Angew. Math.* **467** (1995), 157–167. MR 96m:14018 Zbl 0836.11036

[Hindry and Silverman 2000] M. Hindry and J. H. Silverman, *Diophantine geometry: An introduction*, Graduate Texts in Mathematics **201**, Springer, New York, 2000. MR 2001e:11058 Zbl 0948.11023

[Kawaguchi 2006] S. Kawaguchi, "Canonical height functions for affine plane automorphisms", *Math. Ann.* **335**:2 (2006), 285–310. MR 2007a:11093 Zbl 1101.11019

[Kawaguchi and Silverman 2007] S. Kawaguchi and J. H. Silverman, "Dynamics of projective morphisms having identical canonical heights", *Proc. Lond. Math. Soc.* (3) **95**:2 (2007), 519–544. MR 2008j:11076 Zbl 1130.11035

[Kawaguchi and Silverman 2009] S. Kawaguchi and J. H. Silverman, "Nonarchimedean Green functions and dynamics on projective space", *Math. Z.* **262**:1 (2009), 173–197. MR 2010g:37172 Zbl 1161.32009

[Lang 1983] S. Lang, *Fundamentals of Diophantine geometry*, Springer, New York, 1983. MR 85j:11005 Zbl 0528.14013

[Lee 2013] C. Lee, "An upper bound for the height for regular affine automorphisms of $\mathbb{A}^n$", *Math. Ann.* **355** (2013), 1–16.

[Marcello 2000] S. Marcello, "Sur les propriétés arithmétiques des itérés d'automorphismes réguliers", *C. R. Acad. Sci. Paris Sér. I Math.* **331**:1 (2000), 11–16. MR 2001d:11072 Zbl 1044.11056

[Marcello 2003] S. Marcello, "Sur la dynamique arithmétique des automorphismes de l'espace affine", *Bull. Soc. Math. France* **131**:2 (2003), 229–257. MR 2004d:11053 Zbl 1048.11052

[Masser and Wüstholz 1983] D. W. Masser and G. Wüstholz, "Fields of large transcendence degree generated by values of elliptic functions", *Invent. Math.* **72**:3 (1983), 407–464. MR 85g:11060 Zbl 0516.10027

[Morton and Silverman 1994] P. Morton and J. H. Silverman, "Rational periodic points of rational functions", *Internat. Math. Res. Notices* **2004**:2 (1994), 97–110. MR 95b:11066 Zbl 0819.11045

[Shafikov and Wolf 2003] R. Shafikov and C. Wolf, "Filtrations, hyperbolicity, and dimension for polynomial automorphisms of $\mathbb{C}^n$", *Michigan Math. J.* **51**:3 (2003), 631–649. MR 2004i:37096 Zbl 1053.37024

[Sibony 1999] N. Sibony, "Dynamique des applications rationnelles de $\mathbf{P}^k$", pp. 97–185 in *Dynamique et géométrie complexes* (Lyon, 1997), Panor. Synthèses **8**, Soc. Math. France, Paris, 1999. MR 2001e:32026 Zbl 1020.37026

[Silverman 1994] J. H. Silverman, "Geometric and arithmetic properties of the Hénon map", *Math. Z.* **215**:2 (1994), 237–250. MR 95f:14040 Zbl 0807.58021

[Silverman 2006] J. H. Silverman, “Height bounds and preperiodic points for families of jointly regular affine maps”, *Pure Appl. Math. Q.* **2**:1, part 1 (2006), 135–145. MR 2007a:11095 Zbl 1154.11328

[Silverman 2007] J. H. Silverman, *The arithmetic of dynamical systems*, Graduate Texts in Mathematics **241**, Springer, New York, 2007. MR 2008c:11002 Zbl 1130.37001

Communicated by Joseph Silverman
Received 2012-04-10 Revised 2012-08-06 Accepted 2012-09-04

kawaguch@math.kyoto-u.ac.jp *Department of Mathematics, Kyoto University, Oiwake-cho Kitashirakawa, Sakyo-ku, Kyoto-shi 606-8502, Japan*

ALGEBRA AND NUMBER THEORY 7:5 (2013)
dx.doi.org/10.2140/ant.2013.7.1253

On the ranks of the 2-Selmer groups of twists of a given elliptic curve

Daniel M. Kane

Swinnerton-Dyer considered the proportion of twists of an elliptic curve with full 2-torsion that have 2-Selmer group of a particular dimension. Swinnerton-Dyer obtained asymptotic results on the number of such twists using an unusual notion of asymptotic density. We build on this work to obtain similar results on the density of twists with particular rank of 2-Selmer group using the natural notion of density.

1. Introduction

Let c_1, c_2 and c_3 be distinct rational numbers. Let E be the elliptic curve defined by the equation

$$y^2 = (x - c_1)(x - c_2)(x - c_3).$$

We make the additional technical assumption that none of the $(c_i - c_j)(c_i - c_k)$ are squares. This is equivalent to saying that E is an elliptic curve over $\mathbb{Q}$ with complete 2-torsion and no cyclic subgroup of order 4 defined over $\mathbb{Q}$. For b a square-free number, let E_b be the twist defined by the equation

$$y^2 = (x - bc_1)(x - bc_2)(x - bc_3).$$

Let S be a finite set of places of $\mathbb{Q}$ including 2, ∞ and all of the places at which E has bad reduction. Let D be a positive integer divisible by 8 and by the primes in S. Let $S_2(E_b)$ denote the 2-Selmer group of the curve E_b. We will be interested in how the rank varies with b and in particular in the asymptotic density of b's such that $S_2(E_b)$ has a given rank.

The parity of $\dim(S_2(E_b))$ depends only on the class of b as an element of $\prod_{\nu \in S} \mathbb{Q}_\nu^*/(\mathbb{Q}_\nu^*)^2$. We claim that for exactly half of these values this dimension is odd and exactly half of the time it is even. In particular, we make the following claim, which will be proved in Section 4:

MSC2010: 11G05.
Keywords: Selmer group, elliptic curve, density.

Lemma 1. *There exists a set S consisting of exactly half of the classes c in $(\mathbb{Z}/D)^*/((\mathbb{Z}/D)^*)^2$ such that for any positive integer b relatively prime to D we have that* $\dim(S_2(E_b))$ *is even if and only if b represents a class in S.*

Let $b = p_1 p_2 \cdots p_n$, where p_i are distinct primes relatively prime to D. In [Swinnerton-Dyer 2008], the rank of $S_2(E_b)$ is shown to depend only on the images of the p_i in $(\mathbb{Z}/D)^*/((\mathbb{Z}/D)^*)^2$ and upon which p_i are quadratic residues modulo which p_j. There are $2^{n|S|+\binom{n}{2}}$ possible sets of values for these. Let $\pi_d(n)$ be the fraction of this set of possibilities that cause $S_2(E_b)$ to have rank exactly d. Then the main theorem of [Swinnerton-Dyer 2008] together with Lemma 1 implies:

Theorem 2. *Let $\alpha_0 = \alpha_1 = 0$ and $\alpha_{n+2} = \dfrac{2^n}{\prod_{j=1}^{n}(2^j - 1)\prod_{j=0}^{\infty}(1+2^{-j})}$. Then*

$$\lim_{n\to\infty} \pi_d(n) = \alpha_d.$$

The actual theorem proved in [Swinnerton-Dyer 2008] says that if, in addition, the class of b in $\prod_{v\in S} \mathbb{Q}_v^*/(\mathbb{Q}_v^*)^2$ is fixed, then the analogous $\pi_d(n)$ either converge to $2\alpha_d$ for d even and 0 for d odd or to $2\alpha_d$ for d odd and 0 for d even.

This tells us information about the asymptotic density of twists of E whose 2-Selmer group has a particular rank. Unfortunately, this asymptotic density is taken in a somewhat awkward way by letting the number of primes dividing b go to infinity. In this paper, we prove the following more natural version of Theorem 2:

Theorem 3. *Let E be an elliptic curve over $\mathbb{Q}$ with full 2-torsion defined over $\mathbb{Q}$ and such that*

$$\lim_{n\to\infty} \pi_d(n) = \alpha_d$$

with α_d as given in Theorem 2. Then

$$\lim_{N\to\infty} \frac{\#\{b \le N : b \text{ square-free}, (b, D) = 1 \text{ and } \dim(S_2(E_b)) = d\}}{\#\{b \le N : b \text{ square-free and } (b, D) = 1\}} = \alpha_d.$$

Applying this to twists of E by divisors of D and noting that twists by squares do not affect the Selmer rank, we obtain:

Corollary 4.
$$\lim_{N\to\infty} \frac{\#\{b \le N : \dim(S_2(E_b)) = d\}}{N} = \alpha_d.$$

Corollary 5.
$$\lim_{N\to\infty} \frac{\#\{-N \le b \le N : \dim(S_2(E_b)) = d\}}{2N} = \alpha_d.$$

Our technique is fairly straightforward. Our goal will be to prove that the average moments of the size of the Selmer groups will be as expected. As it turns out, this along with Lemma 1 will be enough to determine the probability of seeing a given rank. In order to analyze the Selmer groups, we follow the method described in

[Swinnerton-Dyer 2008]. Here the 2-Selmer group of E_b can be expressed as the intersection of two Lagrangian subspaces, U and W, of a particular symplectic space, V, over $\mathbb{F}_2$. Although U, V and W all depend on b, once the number of primes dividing b has been fixed along with its congruence class modulo D, these spaces can all be written conveniently in terms of the primes, p_i, dividing b, which we think of as formal variables. Using the formula $|U \cap W| = (1/\sqrt{|V|}) \sum_{u \in U,\, w \in W} (-1)^{u \cdot w}$, we reduce our problem to bounding the size of the "characters" $(-1)^{u \cdot w}$ when averaged over b. These "characters" turn out to be products of Dirichlet characters of the p_i and Legendre symbols of pairs of the p_i. The bulk of our analytic work is in proving these bounds. These bounds will allow us to discount the contribution from most of the terms in our sum (in particular the ones in which Legendre symbols show up in a nontrivial way) and allow us to show that the average of the remaining terms is roughly what should be expected from Swinnerton-Dyer's result.

We should point out the connections between our work and that of [Heath-Brown 1994], where our main result is proved for the particular curve

$$y^2 = x^3 - x.$$

We employ techniques similar to those of Heath-Brown, but the algebra behind them is organized significantly differently. His overall strategy is again to compute the average sizes of moments of $|S_2(E_b)|$ and use these to get at the ranks. He computes $|S_2(E_b)|$ using a different formula than ours. Essentially what he does is use some tricks specific to his curve to deal with the conditions relating to primes dividing D, and instead of considering each prime individually, he groups them based on how they occur in u and w. He lets D_i be the product of all primes dividing b that relate in a particular way (indexed by i). He then gets a formula for $|S_2(E_b)|$ that's a sum over ways of writing b as a product, $b = \prod D_i$, of some term again involving characters of the D_i and Legendre symbols. Using techniques similar to ours, he shows that terms in this sum where the Legendre symbols have a nonnegligible contribution (are not all trivial due to one of the D_i being 1) can be ignored. He then uses some algebra to show that the average of the remaining terms is the desired value. This step differs from our technique where we merely make use of Swinnerton-Dyer's result to compute our average. Essentially, we show that the algebra and the analysis for this problem can be done separately and use [Swinnerton-Dyer 2008] to take care of the algebra. Finally, Heath-Brown uses some techniques from linear algebra to show that the moment bounds imply the correct densities of ranks while we use techniques from complex analysis.

We also note the work of Yu [2005]. In this paper, Yu shows that for a wide family of curves of full 2-torsion that the average size of the 2-Selmer group of a twist is equal to 12. This work uses techniques along the lines of Heath-Brown's, though has some added complication in order to deal with the greater generality.

One advantage of our technique over these others is that we can, to some degree, separate the algebra involved in analyzing the sizes of these Selmer groups from the analysis. When considering the distribution of ranks of Selmer groups of twists of an elliptic curve, there are two types of density estimates that have come up in the literature. The first is to use the natural notion of density over some obvious ordering of twist parameter. The other is to use some notion similar to that of Swinnerton-Dyer, which can be thought of as letting the number of primes dividing the twist parameter go to infinity. Although one is usually interested in natural densities, the Swinnerton-Dyer–type results are often easier to prove as they tend to be essentially algebraic in nature while results about natural density will generally require some tricky analytic work. The techniques of this paper show how asymptotics of the Swinnerton-Dyer–type can be upgraded to results for natural density. Although we have only managed to carry out this procedure for the family of curves used in Theorem 2, there is hope that this procedure might have greater applicability. For example, if someone were to obtain a Swinnerton-Dyer–type result for twists of an elliptic curve with full 2-torsion over $\mathbb{Q}$ that *has* a rational 4-isogeny, it is almost certain that the techniques from this paper would allow one to obtain a result for the same curve using the natural density. Additionally, in [Klagsbrun et al. 2013], Klagsbrun, Mazur and Rubin consider the ranks of twists of an elliptic curve with $\mathrm{Gal}(K(E[2])/K) \simeq S_3$ and obtain Swinnerton-Dyer–type density results. It is possible that ideas in this paper may be adapted to improve these results to work with a more natural notion of density as well. Unfortunately, working in this extended context will likely complicate the analytic aspects of the argument considerably. For example, while we make important use of the fact that the rank of $S_2(E_b)$ depends only on congruence classes of primes dividing b and Legendre symbols between them, it is shown in [Friedlander et al. 2013] that, for curves with cyclic cubic field of 2-torsion, the Selmer rank can depend on more complicated algebraic objects (such as what they term the spin of a prime).

In Section 2, we introduce some basic concepts that will be used throughout. In Section 3, we will prove the necessary character bounds. We use these bounds in Section 4 to establish the average moments of the size of the Selmer groups. Finally, in Section 5, we explain how these results can be used to prove our main theorem.

2. Preliminaries

2.1. ***Asymptotic notation.*** Throughout the rest of this paper, we will make extensive use of O and similar asymptotic notation. In our notation, $O(X)$ will denote a quantity that is at most $H \cdot X$ for some *absolute* constant H. If we need asymptotic notation that depends on some parameters, we will use $O_{a,b,c}(X)$ to denote a quantity that is at most $H(a, b, c) \cdot X$, where H is some function depending only on a, b and c.

2.2. ***Number of prime divisors.*** In order to make use of Swinnerton-Dyer's result, we will need to consider twists of E by integers $b \le N$ with a specific number of prime divisors. For an integer m, we let $\omega(m)$ be the number of prime divisors of m. In our analysis, we will need to have estimates on the number of such b with a particular number of prime divisors. We define

$$\Pi_n(N) = \#\{\text{primes } p \le N \text{ such that } \omega(p) = n\}.$$

Lemma 6 [Hardy and Ramanujan 1917, Lemma A]. *There exist absolute constants C and K such that for any ν and x*

$$\Pi_{\nu+1}(x) \le \frac{Kx}{\log x} \frac{(\log\log x + C)^\nu}{\nu!}.$$

By maximizing the above in terms of ν, it is easy to see:

Corollary 7. *We have*

$$\Pi_n(N) = O\left(\frac{N}{\sqrt{\log\log N}}\right).$$

It is also easy to see from the above that most integers of size roughly N have about $\log\log N$ prime factors. In particular:

Corollary 8. *There is a constant $c > 0$ such that for all N, the number of $b \le N$ with $|\omega(b) - \log\log N| > (\log\log N)^{3/4}$ is at most*

$$2N \exp\bigl(-c\sqrt{\log\log N}\bigr).$$

In particular, the fraction of $b \le N$ with $|\omega(b) - \log\log N| < (\log\log N)^{3/4}$ goes to 1 as N goes to infinity.

We will use Corollary 8 to restrict our attention only to twists by b with an appropriate number of prime divisors.

3. Character bounds

Our main purpose in this section will be to prove the following propositions:

Proposition 9. *Fix positive integers D, n and N with $4 \mid D$, $\log\log N > 1$ and $(\log\log N)/2 < n < 2\log\log N$, and let $c > 0$ be a real number. Let $d_{i,j}, e_{i,j} \in \mathbb{Z}/2$ for $i, j = 1, \ldots, n$ with $e_{i,j} = e_{j,i}$, $d_{i,j} = d_{j,i}$ and $e_{i,i} = d_{i,i} = 0$ for all i and j. Let χ_i be a quadratic character with modulus dividing D for $i = 1, \ldots, n$. Let m be the number of indices i such that at least one of the following holds:*

- *$e_{i,j} = 1$ for some j or*
- *χ_i has modulus not dividing 4 or*
- *χ_i has modulus exactly 4 and $d_{i,j} = 0$ for all j.*

Let $\epsilon(p) = (p-1)/2$. Then if $m > 0$,

$$\left| \frac{1}{n!} \sum_{S_{N,n,D}} \prod_i \chi_i(p_i) \prod_{i<j} (-1)^{\epsilon(p_i)\epsilon(p_j)d_{i,j}} \prod_{i<j} \left(\frac{p_i}{p_j}\right)^{e_{i,j}} \right| = O_{c,D}(Nc^m), \qquad (1)$$

where $S_{N,n,D}$ is the set of n-tuples of distinct primes $p_1, \dots, p_n$ such that $b = p_1 \cdots p_n$ is relatively prime to D and of size at most N.

Note that m is the number of indices i such that, no matter how we fix the values of p_j for the $j \neq i$, the summand on the left-hand side of (1) still depends on p_i. The index set $S_{N,n,D}$ above is a way of indexing (up to overcounting by a factor of $n!$) the set of integers $b \leq N$ that are square-free, relatively prime to D and have $\omega(b) = n$. This notation will be used throughout the rest of the paper. The sum in (1) can be thought of as a sum over such b (the $1/n!$ term accounts for the overcounting) of a "character" defined by the χ_i, $d_{i,j}$ and $e_{i,j}$. Proposition 9 will allow us to show that the "characters" in which the Legendre symbols make a nontrivial appearance add a negligible contribution to our moments.

Proposition 10. *Let n, N and D be positive integers satisfying $\log\log N > 1$ and $(\log\log N)/2 < n < 2\log\log N$. Let*

$$G = \big((\mathbb{Z}/D)^*/((\mathbb{Z}/D)^*)^2\big)^n.$$

Let $f : G \to \mathbb{C}$ be a function with $|f|_\infty \leq 1$. Then

$$\frac{1}{n!} \sum_{S_{N,n,D}} f(p_1, \dots, p_n) = \left(\frac{1}{|G|} \sum_{g\in G} f(g)\right)\left(\frac{|S_{N,n,D}|}{n!}\right) + O_D\left(\frac{N \log\log\log N}{\log\log N}\right). \qquad (2)$$

(*Here $f(p_1, \dots, p_n)$ is really f applied to the vector of their reductions modulo D.*)

This proposition says that the average of f over such $S_{N,n,D}$ is roughly equal to the average of f over G. This will allow us to show that the average value of the remaining terms in our moment calculation equals what we would expect given Swinnerton-Dyer's result.

We begin with a proposition that gives a more precise form of Proposition 9 in the case when the $e_{i,j}$ are all 0.

Proposition 11. *Let D, n and N be integers with $4 \mid D$ and $\log\log N > 1$. Let $C > 0$ be a real number. Let $d_{i,j} \in \mathbb{Z}/2$ for $i, j = 1, \dots, n$ with $d_{i,j} = d_{j,i}$ and $d_{i,i} = 0$. Let χ_i be a quadratic character of modulus dividing D for $i = 1, \dots, n$. Suppose that no Dirichlet character of modulus dividing D has an associated Siegel zero larger than $1 - \beta^{-1}$. Let*

$$B = \max(e^{(C+2)\beta \log\log N}, e^{K(C+2)^2(\log D)^2(\log\log(DN))^2}, n\log^{C+2}(N))$$

for K a sufficiently large absolute constant. Suppose that $B^n < \sqrt{N}$. Let m be the number of indices i such that either

- *χ_i does not have modulus dividing 4 or*
- *χ_i has modulus exactly 4 and $d_{i,j} = 0$ for all j.*

Then

$$\left|\frac{1}{n!}\sum_{S_{N,n,D}}\prod_i \chi_i(p_i)\prod_{i<j}(-1)^{\epsilon(p_i)\epsilon(p_j)d_{i,j}}\right| = O\left(\frac{N}{\sqrt{\log\log N}}\right)\left(O\left(\frac{\log\log B}{n}\right)^m + (\log N)^{-C}\right). \quad (3)$$

Note once again that m is the number of i such that if the values of p_j for $j \neq i$ are all fixed, the resulting summand will still depend on p_i.

The basic idea of the proof will be by induction on m. If $m = 0$, we can bound by the number of terms in our sum, giving a bound of $\Pi_n(N)$, which we bound using Corollary 7. If $m > 0$, there is some p_i such that no matter how we set the other p_j, our character still depends on p_i. We split into cases based on whether $p_i > B$. If $p_i > B$, we fix the values of the other p_j and use bounds on character sums. For $p_i \leq B$, we note that this happens for only about a $(\log\log B)/n$ fraction of the terms in our sum and for each possible value of p_i inductively bound the remaining sum. To deal with the first case, we prove the following:

Lemma 12. *Let K be a sufficiently large constant. Take χ any nontrivial Dirichlet character of modulus at most D and with no Siegel zero more than $1-\beta^{-1}$, constants $N, C > 0$ and X any integer with*

$$X > \max(e^{(C+2)\beta\log\log N}, e^{K(C+2)^2(\log D)^2(\log\log(DN))^2}).$$

Then,

$$\left|\sum_{p\leq X}\chi(p)\right| \leq O(X\log^{-C-2}(N)),$$

where the sum is over primes $p \leq X$.

Proof. Theorem 5.27 of [Iwaniec and Kowalski 2004] implies that, for any Y, for some constant $c > 0$,

$$\sum_{n\leq Y}\chi(n)\Lambda(n) = Y\cdot O\left(Y^{-\beta^{-1}} + \exp\left(\frac{-c\sqrt{\log Y}}{\log D}\right)(\log D)^4\right).$$

Note that the contribution to the above coming from n a power of a prime is $O(\sqrt{Y})$. Using Abel summation to reduce this to a sum over p of $\chi(p)$ rather than

$\chi(p)\log p$, we find that

$$\sum_{p\le X}\chi(p)\le X\cdot O\left(X^{-\beta^{-1}}+\exp\left(\frac{-c\sqrt{\log X}}{\log D}\right)(\log D)^4\right)+O(\sqrt{X}).$$

The former term is sufficiently small since by assumption $X>e^{(C+2)\beta\log\log N}$. The latter term is small enough since $X>e^{K(C+2)^2(\log D)^2(\log\log(DN))^2}$. The last term is small enough since clearly $X>\log^{2C+4}(N)$. □

For positive integers n, N and D and S a set of prime numbers, denote by $Q(n,N,D,k,S)$ the maximum possible absolute value of a sum of the form given in (3) with $m\ge k$ with the added restriction that none of the p_i lie in S. In particular, a sum of the form

$$\frac{1}{n!}\sum_{S_{N,n,D'}}\prod_i\chi_i(p_i)\prod_{i<j}(-1)^{\epsilon(p_i)\epsilon(p_j)d_{i,j}},$$

where χ_i are characters of modulus dividing D, $d_{i,j}\in\{0,1\}$ and

$$D'=D\cdot\prod_{p\in S}p.$$

We write the inductive step for our main bound as follows.

Lemma 13. *Consider integers n, D, N, M, C and B with*

$$B>\max(e^{(C+2)\beta\log\log M},e^{K(C+2)^2(\log D)^2(\log\log(DM))^2},n\log^{C+2}(M)),$$

where $1-\beta^{-1}$ is the largest Siegel zero of a Dirichlet character whose modulus divides D, and K is a large enough constant. Then, if $1\le k\le n$ and S is a set of primes not exceeding B, the quantity $Q(n,N,D,k,S)$ defined above is at most

$$O(N\log N\log^{-C-2}(M))+\frac{1}{n}\sum_{\substack{p<B\\ p\notin S}}Q(n-1,N/p,D,k-1,S\cup\{p\}).$$

Proof. Since $k\ge 1$, there must be an i such that either χ_i has modulus bigger than 4 or has modulus exactly 4 and all of the $d_{i,j}$ are 0. Without loss of generality, n is such an index. We split our sum into cases depending on whether $p_n\ge B$. For $p_n\ge B$, we proceed by fixing all of the p_j for $j\ne n$ and summing over p_n. Letting $P=\prod_{i=1}^{n-1}p_i$, we have

$$\sum_{P=1}^{N/B}\frac{1}{n!}\sum_{\substack{P=p_1\cdots p_{n-1}\\ p_i\text{ distinct}\\ p_i\notin S,\ (D,P)=1}}a\sum_{\substack{B\le p_n\le N/P\\ p_n\ne p_j}}\chi(p_n),$$

where a is some constant of norm 1 depending on $p_1 \cdots p_{n-1}$ and χ is a nontrivial character of modulus dividing D, perhaps also depending on $p_1, \dots, p_{n-1}$. The condition that $p_n \neq p_j$ alters the value of the inner sum by at most n. With this condition removed, we may bound the inner sum by applying Lemma 12 (taking the difference of the terms with $X = N/P$ and $X = B$). Hence, the value of the inner sum is at most $O(N/P \log^{-C-2}(M) + n)$. Since

$$N/P \geq B \geq n \log^{C+2}(M),$$

this is just $O(N/P \log^{-C-2}(M))$. Note that for each P, there are at most $(n-1)!$ ways of writing it as a product of $n-1$ primes (since the primes will be unique up to ordering). Hence, ignoring the extra $1/n$ factor, the sum above is at most

$$\sum_{P=1}^{N/B} O(N/P \log^{-C-2}(M)) = O(N \log N \log^{-C-2} M).$$

For $p_n < B$, we fix p_n and consider the sum over the remaining p_i. We note that for p a prime not in S and relatively prime to D, this sum is $\pm 1/n$ times a sum of the type bounded by $Q(n-1, N/p, D, k-1, S \cup \{p\})$. In particular, we note that, since by assumption the value of m for our original sum was at least k, upon fixing this value of p_n, the value of m for the resulting sum is at least $k-1$ and is thus bounded by $Q(n-1, N/p, D, k-1, S \cup \{p\})$. □

Proof of Proposition 11. We prove by induction on k that for n, N, D, C, M, β and B as above with

$$B > \max(e^{(C+2)\beta \log\log M}, e^{K(C+2)^2(\log D)^2(\log\log(DM))^2}, n \log^{C+2}(M))$$

and S a set of primes less than or equal to B and c a sufficiently large constant,

$$Q(n, N, D, k, S) \leq c\left(\frac{N}{\sqrt{\log\log(N/B^n)}}\right)\left(\frac{c \log\log B}{n}\right)^k + cN \log N \log^{-C-2}(M) \sum_{a=0}^{k-1} \left(\frac{c \log\log B}{n}\right)^a. \quad (4)$$

Plugging in $M = N$, $k = m$, $S = \varnothing$ and

$$B = \max(e^{(C+2)\beta \log\log N}, e^{K(C+2)^2(\log D)^2(\log\log(DN))^2}, n \log^{C+2}(N))$$

yields the necessary result.

We prove (4) by induction on k. For $k = 0$, the sum is at most the sum over $b = p_1 \cdots p_n$ with appropriate conditions of $1/n!$. Since each such b can be written as such a product in at most $n!$ ways, this is at most $\Pi_n(N)$, which by Corollary 7 is at most $c(N/\sqrt{\log\log N})$ for some constant c, as desired.

For larger values of k, we use the inductive hypothesis and Lemma 13 to bound $Q(n, N, D, k, S)$ by

$$\begin{aligned}
cN \log N \log^{-C-2}(M) + \frac{1}{n}\sum_{p<B} Q(n-1, N/p, D, k-1, S')& \\
&\leq cN \log N \log^{-C-2}(M) \\
&\quad + \frac{1}{n}\sum_{p<B}\frac{1}{p}c\left(\frac{N}{\sqrt{\log\log(N/pB^{n-1})}}\right)\left(\frac{c\log\log B}{n-1}\right)^{k-1} \\
&\quad + \frac{1}{n}\sum_{p<B}\frac{1}{p}cN\log N\log^{-C-2}(M)\sum_{a=0}^{k-2}\left(\frac{c\log\log B}{n-1}\right)^{a} \\
&\leq cN \log N \log^{-C-2}(M) \\
&\quad + c\left(\frac{N}{\sqrt{\log\log(N/B^{n})}}\right)\left(\frac{c\log\log B}{n}\right)^{k} \\
&\quad + cN\log N\log^{-C-2}(M)\sum_{a=0}^{k-2}\left(\frac{c\log\log B}{n}\right)^{a+1} \\
&\leq c\left(\frac{N}{\sqrt{\log\log(N/B^{n})}}\right)\left(\frac{c\log\log B}{n}\right)^{k} \\
&\quad + cN\log N\log^{-C-2}(M)\sum_{a=0}^{k-1}\left(\frac{c\log\log B}{n}\right)^{a}.
\end{aligned}$$

Above we use that

$$\frac{1}{n}\left(\frac{1}{n-1}\right)^{a}\sum_{p<B}\frac{1}{p} \leq c\log\log B\left(\frac{1}{n}\right)^{a+1}$$

for all $a \leq n$ if c is sufficiently large. This completes the inductive hypothesis, proving (4) and completing the proof. □

Proof of Proposition 10. First note that we can assume that $4 \mid D$. This is because if that is not the case, we can split our sum up into two cases, one where none of the p_i are 2 and one where one of the p_i is 2. In either case, we get a sum of the same form but now can assume that D is divisible by 4. We assume this so that we can use Proposition 11.

It is clear that the difference between the left-hand side of (2) and the main term on the right-hand side is

$$\frac{1}{|G|}\left(\sum_{\chi\in\widehat{G}\backslash\{1\}}\left(\frac{1}{n!}\sum_{S_{N,n,D}}\chi(p_1,\dots,p_n)\right)\left(\sum_{g\in G} f(g)\chi(g)\right)\right).$$

Using Cauchy–Schwarz, we find that this is at most

$$\frac{1}{|G|}\sqrt{|G|}\,|f|_2\Bigg(\sum_{\chi\in\widehat{G}\backslash\{1\}}\bigg|\frac{1}{n!}\sum_{S_{N,n,D}}\chi(p_1,\ldots,p_n)\bigg|^2\Bigg)^{1/2}.$$

We note that $|f|_2 \le \sqrt{|G|}$ and hence that $(1/|G|)\sqrt{|G|}\,|f|_2 \le 1$. Bounding the character sum using Proposition 11 (using the minimal possible value of B), we get $O(N^2/\log\log N)$ times

$$\sum_{\chi\in\widehat{G}\backslash\{1\}} O_D\bigg(\frac{\log\log\log N}{\log\log N}\bigg)^{2s},$$

where above s is the number of components on which χ (thought of as a product of characters of $(\mathbb{Z}/D\mathbb{Z})^*$) is nontrivial. Since each component of χ can either be trivial or have one of finitely many nontrivial values (each of which contributes $O_D((\log\log\log N)^2/(\log\log N)^2))$ and this can be chosen independently for each component, the inner sum is

$$\begin{aligned}\bigg(1+O_D\bigg(\frac{\log\log\log N}{\log\log N}\bigg)^2\bigg)^n-1 &= \exp\bigg(O_D\bigg(\frac{(\log\log\log N)^2}{\log\log N}\bigg)\bigg)-1\\ &= O_D\bigg(\frac{(\log\log\log N)^2}{\log\log N}\bigg).\end{aligned}$$

Hence, the total error is at most

$$\frac{1}{|G|}\sqrt{|G|}\sqrt{|G|}O_D\Bigg(\bigg(\frac{N^2\log\log\log^2(N)}{\log\log^2(N)}\bigg)^{1/2}\Bigg)=O_D\bigg(\frac{N\log\log\log N}{\log\log N}\bigg). \quad \square$$

The proof of Proposition 9 is along the same lines as the proof of Proposition 11. Again we induct on m. This time, we use Lemma 13 as our base case (when all of the $e_{i,j}$ are 0). If some $e_{i,j}$ is nonzero, we break into cases based on whether p_i and p_j are larger than some integer A (which will be some power of $\log N$). If both p_i and p_j are large, then fixing the remaining primes and summing over p_i and p_j gives a relatively small result. Otherwise, fixing one of these primes at a small value, we are left with a sum of a similar form over the other primes. Unfortunately, doing this will increase our D by a factor of p_i and may introduce characters with bad Siegel zeroes. To counteract this, we will begin by throwing away all terms in our sum where $D\prod_i p_i$ is divisible by the modulus of the worst Siegel zero in some range and use standard results to bound the badness of other Siegel zeroes.

We begin with some lemmas that will allow us to bound sums of Legendre symbols of p_i and p_j as they vary over primes.

Lemma 14. *Let Q and N be positive integers with $Q^2 \geq N$. Let a be a function $\{1, 2, \ldots, N\} \to \mathbb{C}$, supported on square-free numbers. Then we have*

$$\sum_{\chi} \left| \sum_{n=1}^{N} a_n \chi(n) \right|^2 = O(Q\sqrt{N}\|a\|^2). \tag{5}$$

where the outer sum ranges over quadratic characters whose modulus does not exceed Q and is either a prime or four times a prime, and where $\|a\|^2 = \sum_{n=1}^{N} |a_n|^2$ is the squared L^2 norm.

Note the similarity between this and Lemma 4 of [Heath-Brown 1994].

Proof. Let M be the largest positive integer such that $Q^2 \leq NM^2 \leq 4Q^2$. Let $b : \{1, 2, \ldots, M^2\} \to \mathbb{C}$ be the function $b_{n^2} = 1/M$ and $b = 0$ on nonsquares. Let $c = a * b$ be the multiplicative convolution of a and b. Note that, since a is supported on square-free numbers and b supported on squares, $\|c\|^2 = \|a\|^2\|b\|^2 = \|a\|^2/M$. Applying the multiplicative large sieve inequality (see [Iwaniec and Kowalski 2004, Theorem 7.13]) to c,

$$\sum_{q \leq Q} \frac{q}{\phi(q)} \sideset{}{^*}\sum_{\chi \bmod q} \left| \sum_{n} c_n \chi(n) \right|^2 \leq (Q^2 + NM^2 - 1)\|c\|^2. \tag{6}$$

The right-hand side is easily seen to be

$$O(Q^2)\|a\|^2/M = O(Q^2\|a\|^2/(\sqrt{Q^2/N})) = O(Q\sqrt{N}\|a\|^2).$$

For the left-hand side, we may note that it only becomes smaller if we remove the $q/\phi(q)$ or ignore the characters that are not quadratic or do not have moduli either a prime or 4 times a prime. For such characters χ, note that

$$\sum_{n} c_n \chi(n) = \left(\sum_{n} a_n \chi(n) \right) \left(\sum_{n} b_n \chi(n) \right) = \Omega\left(\sum_{n} a_n \chi(n) \right),$$

where the last equality above follows from the fact that χ is 1 on squares not dividing its modulus and noting that, since its modulus divides 4 times a prime, the latter case only happens at even numbers of multiples of p. Hence, the left side of (6) is at least a constant multiple of the left side of (5). This completes the proof. □

Lemma 15. *Let $A \leq X$ be positive numbers, and let $a, b : \mathbb{Z} \to \mathbb{C}$ be functions such that $|a(n)|, |b(n)| \leq 1$ for all n. Denoting by $\left(-\right)$ the Legendre symbol, we have*

$$\left| \sum_{\substack{p_1, p_2 \text{ prime and } \geq A \\ p_1 p_2 \leq X}} a(p_1) b(p_2) \left(\frac{p_1}{p_2} \right) \right| = O(X \log(X) A^{-1/8}).$$

Proof. We first bound the sum of the terms for which $p_1 \leq \sqrt{X}$.

We begin by partitioning $[A, \sqrt{X}]$ into $O(A^{1/4} \log X)$ intervals of the form $[Y, Y(1 + A^{-1/4}))$. We break up our sum based on which of these intervals p_1 lies in. We throw away the terms for which $p_2 \geq X/(Y(1 + A^{-1/4}))$ once such an interval is fixed. We note that for such terms $p_1 p_2 \geq X(1 + A^{-1/4})^{-1}$. Therefore, the number of such terms in our original sum is at most $O(XA^{-1/4})$, and thus, throwing these away introduces an error of at most $O(XA^{-1/4})$.

The sum of the remaining terms is at most

$$\sum_{A \leq p_2 \leq X/(Y(1+A^{-1/4}))} \left| \sum_{Y \leq p_1 \leq Y(1+A^{-1/4})} a(p_1)\left(\frac{p_1}{p_2}\right)\right|.$$

By Cauchy–Schwarz, this is at most

$$\sqrt{X/Y}\left(\sum_{A \leq p_2 \leq X/(Y(1+A^{-1/4}))} \left| \sum_{Y \leq p_1 \leq Y(1+A^{-1/4})} a(p_1)\left(\frac{p_1}{p_2}\right)\right|^2\right)^{1/2}.$$

In the evaluation of the above, we may restrict the support of a to primes between Y and $Y(1 + A^{-1/4})$. Therefore, by Lemma 14, the above is at most

$$\sqrt{X/Y} \cdot O(\sqrt{(X/Y)Y^{1/2}(YA^{-1/4})}) = O(XY^{-1/4}A^{-1/8}) = O(XA^{-3/8}).$$

Hence, summing over the $O(A^{1/4} \log X)$ such intervals, we get a total contribution of $O(X \log(X) A^{-1/8})$.

We get a similar bound on the sum of terms for which $p_2 \leq \sqrt{X}$. Finally, we need to subtract off the sum of terms where both p_1 and p_2 are at most $\sqrt{X}$. This is

$$\sum_{A \leq p_1 \leq \sqrt{X}} \sum_{A \leq p_2 \leq \sqrt{X}} a(p_1)b(p_2)\left(\frac{p_1}{p_2}\right).$$

This is at most

$$\sum_{A \leq p_2 \leq \sqrt{X}} \left| \sum_{A \leq p_1 \leq \sqrt{X}} a(p_1)\left(\frac{p_1}{p_2}\right)\right|.$$

By Cauchy–Schwarz and Lemma 14, this is at most

$$\sqrt{X^{1/2}}O(\sqrt{X^{1/2}X^{1/4}X^{1/2}}) = O(X^{7/8}) = O(XA^{-1/8}).$$

Hence, all of our relevant factors are $O(X \log(X) A^{-1/8})$, thus proving our bound. □

As mentioned above, in proving Proposition 9, we are going to want to deal separately with the terms in which $D \prod_i p_i$ is divisible by a particular bad Siegel zero. In particular, for $X \leq Y$, let $q(X, Y)$ be the modulus of the Dirichlet character with the worst (closest to 1) Siegel zero of any Dirichlet character with modulus between X and Y. In analogy with the Q defined in the proof of Proposition 11, for integers n, N, D, k, X and Y and a set S of primes, we define $Q(n, N, D, k, X, Y, S)$ to be the largest possible value of

$$\left|\frac{1}{n!}\sum_{S'_{N,n,D}}\prod_i \chi_i(p_i)\prod_{i<j}(-1)^{\epsilon(p_i)\epsilon(p_j)d_{i,j}}\prod_{i<j}\left(\frac{p_i}{p_j}\right)^{e_{i,j}}\right|. \tag{7}$$

Above, $S'_{N,n,D}$ is the subset of $S_{N,n,D}$ such that none of the p_i are in S and such that $q(X, Y)$ does not divide $D\prod p_i$ and where the χ_i are Dirichlet characters of modulus dividing D, $e_{i,j}, d_{i,j} \in \{0, 1\}$ and k is at most the number of indices i such that

- $e_{i,j} = 1$ for some j or
- χ_i has modulus not dividing 4 or
- χ_i has modulus exactly 4 and $d_{i,j} = 0$ for all j.

We wish to prove an inductive bound on Q. In particular, we show:

Lemma 16. *Let n, N, D, k, X and Y be as above. Let β be a real number so that the worst Siegel zero of a Dirichlet series of modulus at most D other than $q(X, Y)$ is at most $1-\beta^{-1}$. Let M, A, B and C be integers such that*

$$B > \max(e^{(C+2)\beta\log\log M}, e^{K(C+2)^2(\log D)^2(\log\log(DM))^2}, n\log^{C+2}(M), A)$$

for a sufficiently large constant K. Then for S a set of primes less than or equal to A, we have that $Q(n, N, D, k, X, Y, S)$ is at most the maximum of

$$N\left(O\left(\frac{\log\log B}{n}\right)^k + O(\log N\log^{-C-2}(M))\sum_{a=0}^{k-1}O\left(\frac{\log\log B}{n}\right)^a\right)$$

and

$$O(N\log^2(N)A^{-1/8}) + \frac{2}{n}\sum_{p<A}Q(n-1, N/p, Dp, k-1, X, Y, S\cup\{p\})$$
$$+\frac{1}{n(n-1)}\sum_{p_1,p_2<A}Q(n-2, N/p_1p_2, Dp_1p_2, k-2, X, Y, S\cup\{p_1, p_2\}).$$

Proof. We consider a sum of the form given in (7). If all of the $e_{i,j}$ are 0, we have a form of the type handled in the proof of Proposition 11, and our sum is bounded by the first of our two expressions by (4).

Otherwise, some $e_{i,j}$ is 1. Without loss of generality, this is $e_{n-1,n}$. We can also assume that $d_{n-1,n} = 0$ since adding or removing the appropriate term is equivalent to reversing the Legendre symbol. We split our sum into parts based on which of p_{n-1} and p_n are at least A. In particular, we take the sum of terms with both at least A plus the sum of terms where $p_{n-1} < A$ plus the sum of terms with $p_n < A$ minus the sum of terms with both less than A.

First, consider the case where $p_{n-1}, p_n \geq A$. Fixing the values of $p_1, \ldots, p_{n-2}$ and letting $P = \prod_{i=1}^{n-2} p_i$, we consider the remaining sum over p_{n-1} and p_n. We have

$$\frac{\pm 1}{n!} \sum_{\substack{A \leq p_{n-1}, p_n, \\ p_{n-1} \neq p_n, \\ (p_i, DP)=1, \\ Q \nmid DPp_{n-1}p_n, \\ p_{n-1}p_n \leq N/P}} a(p_{n-1}) b(p_n) \left(\frac{p_{n-1}}{p_n}\right),$$

where a and b are some functions $\mathbb{Z} \to \mathbb{C}$ such that $|a(x)|, |b(x)| \leq 1$ for all x. We note that the condition that $(p_i, DP) = 1$ can be expressed by setting a and b equal to 0 for some appropriate set of primes. We note that the condition that $q(X, Y)$ not divide $DPp_{n-1}p_n$ is only relevant if DP is missing only one or two primes of $q(X, Y)$. In the former case, it is equivalent to making one more value illegal for the p_i. In the latter case, it eliminates at most two terms. The condition that the p_i are distinct removes at most $\sqrt{N/P}$ terms from our sum. Therefore, perhaps after setting a and b to 0 on some set of primes, the above is

$$\frac{\pm 1}{n!} \left(O(\sqrt{N/P}) + \sum_{\substack{A \leq p_{n-1}, p_n, \\ p_{n-1}p_n \leq N/P}} a(p_{n-1}) b(p_n) \left(\frac{p_{n-1}}{p_n}\right) \right).$$

By Lemma 15, this is at most

$$\frac{1}{n!} O(N/P \log(N) A^{-1/8}).$$

Now for each $P \leq N$, it can be written in at most $(n-2)!$ ways; hence, the sum over all $p_{n-1}, p_n \geq A$ is at most

$$\sum_{P=1}^{N} O(N/P \log(N) A^{-1/8}) = O(N \log^2(N) A^{-1/8}).$$

Next, we consider the case where $p_n < A$. We deal with this case by setting p_n to each possible value of size at most A individually. It is easy to check that after setting p_n to such a value p, the sum over the remaining p_i is $1/n$ times a sum of the form bounded by $Q(n-1, N, Dp, k-1, X, Y, S \cup \{p\})$. Hence, the sum over all terms with $p_n < A$ is at most

$$\frac{1}{n} \sum_{p<A} Q(n-1, N/p, Dp, k-1, X, Y, S \cup \{p\}).$$

The sum of the terms with $p_{n-1} < A$ has the same bound, and the sum of terms with both less than A is similarly seen to be at most

$$\frac{1}{n(n-1)} \sum_{p_1, p_2 < A} Q(n-2, N/p_1p_2, Dp_1p_2, k-2, X, Y, S \cup \{p_1, p_2\}). \qquad \square$$

We now use Lemma 16 to prove an inductive bound on Q.

Lemma 17. *Let n, N, D, k, X, Y, S, M, A, B, C and β be as above. Assume furthermore that $Y \geq DA^n$,*

$$B > \max(e^{(C+2)\beta \log\log M}, e^{K(C+2)^2(\log Y)^2(\log\log(YM))^2}, n\log^{C+2} M, A)$$

and S contains only elements of size at most A. Let $L = n-k$. Then the quantity $Q(n, N, D, k, X, Y, S)$ is at most

$$N\biggl(O\Bigl(\frac{\log\log B}{L}\Bigr)^k \\ + O(\log^2(N)A^{-1/8} + \log(N)\log^{-C-2} M)\sum_{a=0}^{k-1} O\Bigl(\frac{\log\log B}{L}\Bigr)^a\biggr).$$

Note that we will wish to apply this lemma with n about $\log\log N$, D a constant, A polylog N, X polylog N, $M = N$, $Y = DA^n$ and B its minimum possible value.

Proof. We proceed by induction on k. In particular, we show that for a sufficiently large constant c that $Q(n, N, D, k, X, Y, S)$ is at most

$$cN\biggl(\Bigl(\frac{c\log\log B}{L}\Bigr)^k + \bigl(\log^2(N)A^{-1/8} + \log(N)\log^{-C-2} M\bigr)\sum_{a=0}^{k-1}\Bigl(\frac{c\log\log B}{L}\Bigr)^a\biggr).$$

We bound Q inductively by Lemma 16. Our base case is when Q is equal to

$$N\biggl(O\Bigl(\frac{\log\log B}{n}\Bigr)^k + O\bigl(\log N\log^{-C-2} M\bigr)\sum_{a=0}^{k-1} O\Bigl(\frac{\log\log B}{n}\Bigr)^a\biggr)$$

(which must happen if $k = 0$). In this case, our desired bound holds assuming that c is sufficiently large.

Otherwise, $Q(n, N, D, k, X, Y, S)$ is bounded by

$$O(N\log^2(N)A^{-1/8}) + \frac{2}{n}\sum_{p<A} Q(n-1, N/p, Dp, k-1, X, Y, S\cup\{p\}) \\ + \frac{1}{n(n-1)}\sum_{p_1,p_2<A} Q(n-2, N/p_1p_2, Dp_1p_2, k-2, X, Y, S\cup\{p_1, p_2\}).$$

Notice that the parameters of Q in the above also satisfy our hypothesis, so we may bound them inductively. Note also that, for the above values of Q, the value of L is the same. Letting $U = (c\log\log B)/L$ and

$$E = c(\log^2(N)A^{-1/8} + \log N\log^{-C-2} M),$$

then for c sufficiently large the above is easily seen to be at most

$$N\left(E+\frac{U}{2}\left(U^{k-1}+E\sum_{a=0}^{k-2}U^a\right)+\frac{U^2}{2}\left(U^{k-2}+E\sum_{a=0}^{k-3}U^a\right)\right)\leq N\left(U^k+E\sum_{a=0}^{k-1}U^a\right).$$

This completes our inductive step and finishes the proof. □

Proof of Proposition 9. The basic idea will be to compare the sum in question to the quantity $Q(n, N, D, k, X, Y, \varnothing)$ for appropriate settings of the parameters. We begin by fixing the constant c in the proposition statement. We let C be a constant large enough that $c^n > \log^{-C}(N)$ (recall that n was $O(\log\log N)$). We set A to $\log^{8C+16}(N)$, X to $\log^C(N)$ and Y to $DA^n = \exp(O_D(C(\log\log N)^2))$. We let $M = N$.

We note that β comes from either the worst Siegel zero of modulus less that X or the second worst Siegel zero of modulus less than Y. By Theorem 5.28 of [Iwaniec and Kowalski 2004], β is at most $O_\epsilon(X^\epsilon)$ in the former case and at most $O(\log Y)$ in the latter case. Hence (changing ϵ by a factor of C), we have unconditionally that $\beta = O_\epsilon(\log^\epsilon(N))$ for any $\epsilon > 0$. We next let

$$B = \max(e^{(C+2)\beta\log\log M}, e^{K(C+2)^2(\log Y)^2(\log\log(YM))^2}, n\log^{C+2}(M), A).$$

Hence, for sufficiently large N (in terms of ϵ and D),

$$\log\log B < \epsilon\log\log N.$$

Finally, we pick k so that $n/2 \geq k \geq m/2$. Thus, $L = n-k > n/2 = \Omega(\log\log N)$. Noting that we satisfy the hypothesis of Lemma 16, we have that, for N sufficiently large relative to ϵ and D, $Q(n, N, D, k, X, Y, \varnothing)$ is at most

$$N\left(O(\epsilon)^{m/2} + O(\log^2(N)\log^{-C-2}(N) + \log N\log^{-C-1}(N))\sum_{a=0}^{k}O(\epsilon)^a\right).$$

If ϵ is small enough that the term $O(\epsilon)$ is at most $1/2$, this is at most

$$N(O(\epsilon)^{m/2} + \log^{-C}(N)).$$

If additionally the $O(\epsilon)$ term is less than c^2, this is

$$O(Nc^m).$$

Hence, for N sufficiently large relative to c and D,

$$Q(n, N, D, k, X, Y, \varnothing) = O(Nc^m).$$

Therefore, unequivocally,

$$Q(n, N, D, k, X, Y, \varnothing) = O_{c,D}(Nc^m).$$

Finally, we note that the difference between $Q(n, N, D, k, X, Y, \varnothing)$ and the term that we are trying to bound is exactly the sum over such terms where $p_1 \cdots p_n$ is divisible by $q(X, Y)/\gcd(q(X, Y), D)$. Since $q(X, Y) \geq X$, there are only $O_D(N \log^{-C}(N))$ such products. Since each product can be obtained in at most $n!$ ways, each contributing at most $1/n!$, this difference is $O_D(N \log^{-C}(N)) = O(Nc^m)$ at most. Therefore, the thing we wish to bound is $O_{c,D}(Nc^m)$. □

4. Average sizes of Selmer groups

Here we use the results from the previous section to prove the following:

Proposition 18. *Let E be an elliptic curve satisfying the conditions of Theorem 3 (and in particular by Theorem 2, for any E with full 2-torsion defined over $\mathbb{Q}$ and no cyclic 4-isogeny defined over $\mathbb{Q}$). Let S be a finite set of places containing $2, \infty$ and all of the places where E has bad reduction. Let x be either -1 or a power of 2. Let $\omega(m)$ denote the number of prime factors of m. Say that $(m, S) = 1$ if m is an integer not divisible by any of the finite places in S. For positive integers N, let $\mathcal{S}_N$ denote the set of integers $b \leq N$ square-free with $|\omega(b) - \log\log N| \leq (\log\log N)^{3/4}$ and $(b, S) = 1$. Then*

$$\lim_{N\to\infty} \frac{\sum_{\mathcal{S}_N} x^{\dim(S_2(E_b))}}{|\mathcal{S}_N|} = \sum_n x^n \alpha_n.$$

This says that the k-th moment of $|S_2(E_b)|$ averaged over $b \leq N$ with

$$|\omega(b) - \log\log N| \leq (\log\log N)^{3/4}$$

is what you would expect given Theorem 2. Furthermore, Proposition 18 says that, averaged over the same set of bs, the rank of the Selmer group is odd half of the time. The latter part of the proposition follows from Lemma 1.

Proof of Lemma 1. First we replace E by a twist such that $c_i - c_j$ are pairwise relatively prime integers. It is now the case that E has everywhere good or multiplicative reduction, and we are now concerned with $\dim(S_2(E_{db}))$ for some constant $d \mid D$. By [Mazur and Rubin 2010, Theorem 2.7; Kramer 1981, Corollary 1], we have that $\dim(S_2(E_{bd})) \equiv \dim(S_2(E)) \bmod 2$ if and only if $(-1)^x \chi_{bd}(-N) = 1$ where $x = \omega(d)$, N is the product of the primes not dividing d at which E has bad reduction and χ_{bd} is the quadratic character corresponding to the extension $\mathbb{Q}(\sqrt{bd})$. From this, the lemma follows immediately. □

In order to prove the rest of Proposition 18, we will need a concrete description of the Selmer groups of twists of E. We follow the treatment given in [Swinnerton-Dyer 2008]. Let $b = p_1 \cdots p_n$ where p_i are distinct primes relatively prime to S (we leave which primes unspecified for now). Let $B = S \cup \{p_1, \ldots, p_n\}$. For $\nu \in B$, let V_ν be the subspace of $(u_1, u_2, u_3) \in (\mathbb{Q}_\nu^*/(\mathbb{Q}_\nu^*)^2)^3$ such that $u_1u_2u_3 = 1$. Note that V_ν has a symplectic form given by $(u_1, u_2, u_3) \cdot (v_1, v_2, v_3) = \prod_{i=1}^{3}(u_i, v_i)_\nu$, where $(u_i, v_i)_\nu$ is the Hilbert symbol. Let $V = \prod_{\nu \in B} V_\nu$ be a symplectic $\mathbb{F}_2$-vector space of dimension $2M$.

There are two important Lagrangian subspaces of V. The first, which we call U, is the image in V of $(\mathbb{Z}_B^*/(\mathbb{Z}_B^*)^2)_1^3$. The other, which we call W, is given as the product of W_ν over $\nu \in B$, where W_ν consists of points of the form $(x - bc_1, x - bc_2, x - bc_3)$ for $(x, y) \in E_b$. Note that we can write $W = W_S \times W_b$ where $W_S = \prod_{\nu \in S} W_\nu$ and $W_b = \prod_{\nu | b} W_\nu$. The Selmer group is given by

$$S_2(E_b) = U \cap W.$$

As written, U, W and V all depend on the primes dividing b. Fortunately, as we will see, there are natural spaces U' and W' that depend very little on b with convenient isomorphisms to U and W. It would also be possible to similarly parametrize V, but this will prove to be unnecessary as we intend to compute the size of the intersection of U and W solely in terms of the restriction of the symplectic pairing on V to $U \times W$.

Let U' be the $\mathbb{F}_2$-vector space generated by the symbols ν and ν' for $\nu \in S$ and p_i and p_i' for $1 \le i \le n$. There exists an isomorphism $f : U' \to U$ given by $f(\infty) = (-1, -1, 1)$, $f(\infty') = (1, -1, -1)$, $f(p) = (p, p, 1)$ and $f(p') = (1, p, p)$.

Note also that W_{p_i} is generated by $((c_1 - c_2)(c_1 - c_3), b(c_1 - c_2), b(c_1 - c_3))$ and $(b(c_3 - c_1), b(c_3 - c_2), (c_3 - c_1)(c_3 - c_2))$. If we define W' to be the $\mathbb{F}_2$-vector space generated by the symbols p_i and p_i' for $1 \le i \le n$, then there is an isomorphism $g : W' \to W_b$ given by $g(p_i) = ((c_1 - c_2)(c_1 - c_3), b(c_1 - c_2), b(c_1 - c_3)) \in W_{p_i}$ and $g(p_i') = (b(c_3 - c_1), b(c_3 - c_2), (c_3 - c_1)(c_3 - c_2)) \in W_{p_i}$.

Let $G = \prod_{\nu \in S} \mathfrak{o}_\nu^*/(\mathfrak{o}_\nu^*)^2$ (here $\mathfrak{o}_\nu^*$ are the units in the ring of integers of k_ν). Note that W_S is determined by the restriction of b to G. So for $c \in G$, let $W_{S,c}$ be W_S for such b. Let $W_c' = W_{S,c} \times W'$. Then we have a natural map $g_c : W_c' \to V$ that is an isomorphism between W_c' and W if b restricts to c.

Proof of Proposition 18. For $x = -1$, this proposition just says that the parity is odd half of the time, which follows from Lemma 1. For $x = 2^k$, this says something about the expected value of $|S_2(E_b)|^k$. For $x = 2^k$, we will show that, for each $n \in (\log\log N - (\log\log N)^{3/4}, \log\log N + (\log\log N)^{3/4})$,

$$\sum_{S_{N,n,D}} |S_2(E_b)|^k = |S_{N,n,D}| \left(\sum_m \alpha_m (2^k)^m + \delta(n, N) \right) + O_{E,k}\left(\frac{N(\log\log\log N)^2}{\log\log N} \right),$$

where $\delta(n, N)$ is some function such that $\lim_{N\to\infty} \delta(n, N) = 0$. Summing over n and noting that there are $\Omega(N)$ values of $b \le N$ square-free with $(b, S) = 1$ and $|\omega(b) - \log\log N| < (\log\log N)^{3/4}$ gives us our desired result.

In order to do this, we need to better understand $|S_2(E_b)| = |U \cap W|$. For $v \in V$, we have, since U is Lagrangian of size 2^M,

$$\frac{1}{2^M}\sum_{u\in U}(-1)^{u\cdot v} = \begin{cases} 1 & \text{if } v \in U^{\perp}, \\ 0 & \text{else,} \end{cases}$$
$$= \begin{cases} 1 & \text{if } v \in U, \\ 0 & \text{else.} \end{cases}$$

Hence,

$$\begin{aligned} |S_2(E_b)| &= |U \cap W| \\ &= \#\{w \in W : w \in U\} \\ &= \sum_{w\in W} \frac{1}{2^M}\sum_{u\in U}(-1)^{u\cdot w} \\ &= \frac{1}{2^M}\sum_{u\in U,\ w\in W}(-1)^{u\cdot w} \\ &= \frac{1}{2^M}\sum_{u\in U',\ w\in W'_b}(-1)^{f(u)\cdot g_b(w)}. \end{aligned}$$

If we extend f and g_c to $f^k : (U')^k \to U^k$ and $g_c^k : (W'_c)^k \to V^k$ and extend the inner product on V to an inner product on V^k,

$$|S_2(E_b)|^k = \frac{1}{2^{kM}}\sum_{\substack{u\in(U')^k \\ w\in(W'_b)^k}}(-1)^{f^k(u)\cdot g_b^k(w)}$$

and therefore that

$$|S_2(E_b)|^k = \frac{1}{2^{kM}|G|}\sum_{\substack{c\in G,\ \chi\in\widehat{G} \\ u\in(U')^k \\ w\in(W'_c)^k}}\chi(bc^{-1})(-1)^{f^k(u)\cdot g_c^k(w)}. \tag{8}$$

Notice that once we fix values of c, χ, u and w in (8), the summand (when treated as a function of $p_1, \ldots, p_n$) is of the same form as the "characters" studied in Section 3.

We want to take the sum over $S_{N,n,D}$ of $|S_2(E_b)|^k$. If we let D be 8 times the product of the finite odd primes in S, we note that each such b can be expressed exactly $n!$ ways as a product $b = p_1 \cdots p_n$ with p_i distinct and $(p_i, D) = 1$. Therefore, this sum equals

$$\frac{1}{n!}\sum_{S_{N,n,D}}\frac{1}{2^{kM}|G|}\sum_{\substack{c\in G,\ \chi\in\widehat{G}, \\ u\in(U')^k,\ w\in(W'_c)^k}}\prod_i \chi(p_i)\overline{\chi}(c)(-1)^{f^k(u)\cdot g_c^k(w)}.$$

Interchanging the order of summation gives us

$$\frac{1}{2^{kM}|G|}\sum_{S_{N,n,D}}\frac{\bar{\chi}(c)}{n!}\sum_{\substack{p_1,\dots,p_n\\ \text{distinct primes},\\ (D,p_i)=1,\\ \prod_i p_i\le N}}\left(\prod_i \chi(p_i)\right)(-1)^{f^k(u)\cdot g_c^k(w)}.$$

Now the inner sum is exactly of the form studied in Proposition 9.

We first wish to bound the contribution from terms where this inner sum has terms of the form $\left(\frac{p_i}{p_j}\right)$ or in the terminology of Proposition 9 for which not all of the $e_{i,j}$ are 0. In order to do this, we will need to determine how many of these terms there are and how large their values of m are. Notice that terms of the form $\left(\frac{p_i}{p_j}\right)$ show up here when we are evaluating the Hilbert symbols of the form $(p, b(c_a - c_b))_p$, $(p, b(c_a - c_b))_q$, $(q, b(c_a - c_b))_p$ and $(q, b(c_a - c_b))_q$ and in no other places.

Let $U_i \subset U'$ be the subspace generated by $p_i = (p_i, p_i, 1)$ and $p_i' = (1, p_i, p_i)$. For $u \in U'$, let u_i be its component in U_i in the obvious way. Let $W_i \subset W'$ be W_{p_i}. For $w \in W_c'$, let w_i be its component in W_i. It is not hard to see that the power of $\left(\frac{p_i}{p_j}\right)$ appearing in $(-1)^{f^k(u)\cdot g_c^k(w)}$ depends only on the projections of u and w onto $U_i \times U_j$ and $W_i \times W_j$, respectively. Our analysis of these exponents will be simplified considerably by noting that the U_i and W_i have convenient isomorphisms to fixed spaces, which we call U_0 and W_0. In particular, let U_0 be the $\mathbb{F}_2$-vector space with formal generators p and p'. We have a natural isomorphism between U_0 and U_i sending p to p_i and p' to p_i'. We will hence often think of u_i as an element of U_0. Similarly, let W_0 be the $\mathbb{F}_2$-vector space with formal generators $((c_1-c_2)(c_1-c_3)$, $b(c_1-c_2), b(c_1-c_3))$ and $(b(c_3-c_1), b(c_3-c_2), (c_3-c_1)(c_3-c_2))$. We similarly have natural isomorphisms between W_i and W_0 and will often consider w_i as an element of W_0 instead of W_i.

Additionally, we have a bilinear form $U_0 \times W_0 \to \mathbb{F}_2$ defined by

$$\begin{aligned}
&p\cdot((c_1-c_2)(c_1-c_3), b(c_1-c_2), b(c_1-c_3))\\
&\qquad = p'\cdot(b(c_3-c_1), b(c_3-c_2), (c_3-c_1)(c_3-c_2))\\
&\qquad = 1,\\
&p'\cdot((c_1-c_2)(c_1-c_3), b(c_1-c_2), b(c_1-c_3))\\
&\qquad = p\cdot(b(c_3-c_1), b(c_3-c_2), (c_3-c_1)(c_3-c_2))\\
&\qquad = 0.
\end{aligned}$$

Notice that if $u \in U'$ and $w \in W_c'$, then the exponent of $\left(\frac{p_i}{p_j}\right)$ that appears in $(-1)^{f(u)\cdot g_c(w)}$ is $(u_i + u_j)\cdot(w_i + w_j)$. Similarly, if $u \in (U')^k$ and $w \in (W_c')^k$, the exponent of $\left(\frac{p_i}{p_j}\right)$ that appears in $(-1)^{f^k(u)\cdot g_c^k(w)}$ is $(u_i + u_j)\cdot(w_i + w_j)$, where u_*

and w_* are thought of as elements of U_0^k and W_0^k, and the inner product is extended to $U_0^k \times W_0^k$ as $(x_1, \dots, x_k) \cdot (y_1, \dots, y_k) = \sum_{i=1}^k x_i \cdot y_i$.

Let $T = U_0^k \times W_0^k$. We define by $\langle (u, w), (u', w') \rangle = u \cdot w' + u' \cdot w$ a symplectic form on T. Also define a quadratic form q on T by $q(u, w) = u \cdot w$. We claim, given some sequence of elements, $t_x = (u_x, w_x) \in T$ for $x \in I$, that $(u_x + u_y) \cdot (w_x + w_y) = 0$ for all pairs $x, y \in I$ only if all of the t_x lie in a translate of a Lagrangian subspace of T under the symplectic form $\langle -, - \rangle$. To show this, we note that, for $t = (u, w)$ and $t' = (u', w')$, $(u + u') \cdot (w + w') = \langle t, t' \rangle + q(t) + q(t')$. We need to show that, for all $x, y, z \in I$, $\langle (t_x + t_y), (t_x + t_z) \rangle = 0$. This is true because

$$\begin{aligned}
&\langle (t_x + t_y), (t_x + t_z) \rangle \\
&\quad = \langle t_x, t_x \rangle + \langle t_x, t_z \rangle + \langle t_y, t_x \rangle + \langle t_y, t_z \rangle \\
&\quad = \langle t_x, t_z \rangle + \langle t_y, t_x \rangle + \langle t_y, t_z \rangle \\
&\quad = \langle t_x, t_z \rangle + \langle t_y, t_x \rangle + \langle t_y, t_z \rangle + 2q(t_x) + 2q(t_y) + 2q(t_z) \\
&\quad = (\langle t_y, t_x \rangle + q(t_x) + q(t_y)) + (\langle t_x, t_z \rangle + q(t_x) + q(t_z)) + (\langle t_y, t_z \rangle + q(t_y) + q(t_z)) \\
&\quad = 0.
\end{aligned}$$

Given $u = (u_1, \dots, u_n) \in \prod_{i=1}^n U_i^k$ and $w = (w_1, \dots, w_n) \in \prod_{i=1}^n W_i^k$, suppose that we have a set of l indices in $\{1, 2, \dots, n\}$, which we call *active* indices, such that $(-1)^{f^k(u) \cdot g^k(w)}$ has terms of the form $\left(\frac{p_i}{p_j}\right)$ only if i and j are both active, and suppose furthermore that each active index shows up as either i or j in at least one such term. Let $t_i = (u_i, w_i) \in T$ (where we have identified u_i and w_i as elements of U_0^k and W_0^k, respectively). We claim that t_i takes fewer than 4^k different values on nonactive indices, i. We note that our notion of active indices is similar to the notion in [Heath-Brown 1994] of linked indices.

Since $\langle t_i, t_j \rangle + q(t_i) + q(t_j) = 0$ for any two nonactive indices t_i and t_j, all of these must lie in a translate of some Lagrangian subspace of T. Therefore, t_i can take at most 4^k values on nonactive indices. Suppose for sake of contradiction that all of these values are actually assumed by some nonactive index. Then consider t_j for j an active index. The t_i for i either nonactive or equal to j must similarly lie in a translate of a Lagrangian subspace. Since such a space is already determined by the nonactive indices and since all elements of this affine subspace are already occupied, t_j must equal t_i for some nonactive i. But this means that every t_j is assumed by some nonactive index, which implies that no terms of the form $\left(\frac{p_i}{p_j}\right)$ survive, yielding a contradiction.

Now consider the number of such u and w so that there are $l \geq 1$ active indices. Once we fix the values t_i that are allowed to be taken by the nonactive indices (which can only be done in finitely many ways), there are $\binom{n}{l}$ ways to choose the active indices, at most $2^k - 1$ ways to pick t_i for each nonactive index and at most 2^{2k} ways for each active index. Hence, the total number of such u and w with exactly l active indices is

$$O\left(\binom{n}{l}(4^k-1)^{n-l}(4^{2k})^l\right).$$

The value of the inner sum for such a (u,w) is at most $O_{E,k}(N(2^{-2k-1})^l)$ by Proposition 9. Hence, summing over all $l>0$ and recalling the 2^{-Mk} out front, we get a contribution of at most

$$\begin{aligned} N4^{-nk}O_{E,k}\left(\sum_l \binom{n}{l}(4^k-1)^{n-l}\left(\frac{1}{2}\right)^l\right) &= N4^{-nk}O_{E,k}((4^k-1/2)^n) \\ &= N\,O_{E,k}((1-4^{-k-1})^n) \\ &= N\,O_{E,k}\left((\log N)^{-4^{-k-2}}\right). \end{aligned}$$

Therefore, we may safely ignore all of the terms in which a $\left(\frac{p_i}{p_j}\right)$ shows up. This is our analogue of Lemma 6 in [Heath-Brown 1994].

Notice that, by the above analysis, the number of remaining terms must be $O_{k,E}(2^{Mk})$. Additionally, for these terms, we may apply Proposition 10. Therefore, each term, up to an error of $O_E((\log\log\log N)^2/\log\log N)$, equals $|S_{N,n,D}|$ times the average of its summand over all possible conjugacy classes of $p_1,\dots,p_n$ modulo $4D$. Since there are $O_{k,E}(2^{Mk})$ such terms and since there is an outer factor of 2^{-kM}, we reach two conclusions. Firstly, the sum in question is bounded by $O_{k,E}(|S_{N,n,D}|)$. Secondly, $1/n!$ times the sum over $S_{N,n,D}$ of $|S_2(E_b)|^k$ is, to within an error of $O_{E,k}((\log\log\log N)^2/\log\log N)$ equal to $|S_{N,n,D}|$ times the average over $b=p_1\cdots p_n$ over all possible values of p_i modulo $4D$ and Legendre symbols $\left(\frac{p_i}{p_j}\right)$ of $|S_2(E_b)|^k$. By definition, this latter average is simply

$$\sum_d \pi_d(n)2^{kd}.$$

Using the fact that this is bounded for $k+1$ independently of n, we find that $\pi_d(n)=O_{k,E}(2^{-(k+1)d})$. In order to complete the proof of our proposition, we need to show that

$$\lim_{n\to\infty}\sum_d(\pi_d(n)-\alpha_d)2^{kd}=0.$$

But this follows from the fact that

$$\sum_{d>X}(\pi_d(n)-\alpha_d)2^{kd}=O_{E,k}\left(\sum_{d>X}2^{-d}\right)=O_{E,k}(2^{-X})$$

and that $\pi_d(n)\to\alpha_d$ for all d by assumption. □

5. From sizes to ranks

In this section, we turn Proposition 18 into a proof of Theorem 3. This section is analogous to Section 8 of [Heath-Brown 1994] although our techniques are significantly different. We begin by doing some computations with the α_i.

Note that

$$\alpha_{n+2} = \left(\frac{1}{\prod_{j=0}^{\infty}(1+2^{-j})}\right)2^{-\binom{n}{2}}\prod_{j=1}^{n}(1-2^{-j})^{-1}.$$

Now $\prod_{j=1}^{n}(1-2^{-j})^{-1}$ is the sum over partitions, P, into parts of size at most n of $2^{-|P|}$. Equivalently, taking the transpose, it is the sum over partitions P with at most n parts of $2^{-|P|}$. Multiplying by $2^{-\binom{n}{2}}$, we get the sum over partitions P with n distinct parts (possibly a part of size 0) of $2^{-|P|}$. Therefore,

$$F(x) = \sum_{n=0}^{\infty}\alpha_n x^n = \frac{x^2\prod_{j=0}^{\infty}(1+2^{-j}x)}{\prod_{j=0}^{\infty}(1+2^{-j})}$$

since the x^{d+2} coefficient of $F(x)$ is also the sum over partitions, P, into exactly d distinct parts (perhaps one of which is 0) of $2{-}|P|$ divided by $\prod_{j=0}^{\infty}(1+2^{-j})$. This implies in particular that $\sum_{n=0}^{\infty}\alpha_n$ equals 1 as it should.

Let T_N be the set of square-free $b \le N$ with $|\omega(b) - \log\log N| < (\log\log N)^{3/4}$ and $(b, D) = 1$. Let $C_d(N)$ be

$$\frac{\#\{b \in T_N : \dim(S_2(E_b)) = d\}}{|T_N|}.$$

Let $C(N) = (C_0(N), C_1(N), \dots) \in [0, 1]^{\omega}$. Theorem 3 is equivalent to showing that

$$\lim_{N\to\infty} C(N) = (\alpha_0, \alpha_1, \dots).$$

Lemma 19. *Suppose that some subsequence of the $C(N)$ converges to some sequence $(\beta_0, \beta_1, \dots) \in [0, 1]^{\omega}$ in the product topology. Let $G(x) = \sum_n \beta_n x^n$. Then $G(x)$ has infinite radius of convergence and $F(x) = G(x)$ for $x = -1$ or x equals a power of 2. Also $\beta_0 = \beta_1 = 0$.*

This lemma says that, if the $C(N)$ have some limit, the naïve attempt to compute moments of the Selmer groups from this limit would succeed.

Proof. The last claim follows from the fact that since E_b has full 2-torsion, its 2-Selmer group always has rank at least 2. Notice that $\sum_d C_d(N)x^d$ is equal to the average size of $x^{\dim(S_2(E_b))}$ over $b \le N$ square-free, relatively prime to D with $|\omega(b) - \log\log N| < (\log\log N)^{3/4}$. This has limit $F(x)$ as $N \to \infty$ by Proposition 18 if x is -1 or a power of 2. In particular, it is bounded. Therefore, there exists an R_k such that

$$\sum_d C_d(N)2^{kd} \leq R_k$$

for all N. Hence, $C_d(N) \leq R_k 2^{-kd}$ for all d and N, which implies $\beta_d \leq R_k 2^{-kd}$. Therefore, G has infinite radius of convergence.

Furthermore, if we pick a subsequence, $N_i \to \infty$, such that $C_d(N_i) \to \beta_d$ for all d,

$$\begin{aligned} F(2^k) &= \lim_{i\to\infty} \sum_d C_d(N_i)2^{dk} \\ &= \lim_{i\to\infty} \sum_{d\leq X} C_d(N_i)2^{dk} + O\biggl(\sum_{d>X} R_{k+1}2^{-d}\biggr) \\ &= \lim_{i\to\infty} \sum_{d\leq X} C_d(N_i)2^{dk} + O(R_{k+1}2^{-X}) \\ &= \sum_{d\leq X} \beta_d 2^{dk} + O(R_{k+1}2^{-X}). \end{aligned}$$

So

$$\lim_{X\to\infty} \sum_{d\leq X} \beta_d 2^{dk} = F(2^k).$$

Thus, $G(2^k) = F(2^k)$. For $x = -1$, the argument is similar but comes from the equidistribution of parity rather than expectation of size. □

Lemma 20. *Suppose that $G(x) = \sum_n \beta_n x^n$ is a Taylor series with infinite radius of convergence. Suppose also that $\beta_n \in [0, 1]$ for all n and that $G(x) = F(x)$ for x equal to -1 or a power of 2. Suppose also that $\beta_0 = \beta_1 = 0$. Then $\beta_n = \alpha_n$ for all n.*

Proof. First we wish to prove a bound on the size of the coefficients of G. Note that

$$F(2^k) = \frac{2^{2k}(1+2^k)(1+2^{k-1})\cdots}{(1+2^0)(1+2^{-1})\cdots} = 2^{2k}\prod_{j=1}^{k}(1+2^k) = O(2^{2k+k(k+1)/2}).$$

Now

$$2^{nk}\beta_n \leq G(2^k) = F(2^k) = O(2^{2k+k(k+1)/2}).$$

Therefore,

$$\beta_n = O(2^{2k+k(k+1)/2-kn}).$$

Setting $k = n$, we find that

$$\beta_n = O(2^{-n^2/2+5n/2}) = O\bigl(2^{-\binom{n-2}{2}}\bigr).$$

The same can be said for F. Now consider $F - G$. This is an entire function whose x^n coefficient is bounded by $O(2^{-\binom{n-2}{2}})$. Furthermore, $F - G$ vanishes to order at least 2 at 0 and order at least 1 at -1 and at powers of 2. The bounds on coefficients imply that

$$|F(x)-G(x)| \le O\biggl(\sum_n 2^{-\binom{n-2}{2}}|x|^n\biggr).$$

The terms in the above sum clearly decay rapidly for n on either side of $\log_2(|x|)$. Hence,

$$\begin{aligned}|F(x)-G(x)| &= O\bigl(2^{(-\log_2(|x|)^2+5\log_2(|x|))/2+\log_2(|x|)^2}\bigr)\\ &= O\bigl(2^{(\log_2(|x|)^2+5\log_2(|x|))/2}\bigr).\end{aligned}$$

In particular, $F-G$ is a function of order less than 1. Hence, it must equal

$$Cx^{2+t}\prod_\rho(1-x/\rho),$$

where the product is over nonzero roots ρ of $F-G$ and t is some nonnegative integer. On the other hand, Jensen's theorem tells us that if $C \neq 0$ the average value of $\log_2(|F-G|)$ on a circle of radius R is

$$\log_2|C| + (2+t)\log_2 R + \sum_{|\rho|<R}\log_2(R/|\rho|).$$

Setting $R = 2^k$ and noting the contributions from $\rho = -1$ and $\rho = 2^j$ for $j < k$,

$$O(1)+3k+\sum_{j<k}(k-j) = O(1)+3k+\binom{k+1}{2} = O(1)+\frac{k^2+7k}{2} > \frac{k^2+5k}{2},$$

which is larger than $\log_2(|F-G|)$ can be at this radius, providing a contradiction. □

Proof of Theorem 3. Suppose that $C(N)$ does not have limit $(\alpha_0, \alpha_1, \dots)$. Then there is some subsequence N_i such that $C(N_i)$ avoid some neighborhood of $(\alpha_0, \alpha_1, \dots)$. By compactness, $C(N_i)$ must have some subsequence with a limit $(\beta_0, \beta_1, \dots)$. By Lemmas 19 and 20, $(\alpha_0, \alpha_1, \dots) = (\beta_0, \beta_1, \dots)$. This is a contradiction.

Therefore, $\lim_{N\to\infty} C(N) = (\alpha_0, \alpha_1, \dots)$. Hence, $\lim_{N\to\infty} C_d(N) = \alpha_d$ for all d. The theorem follows immediately from this and the fact the fraction of $b \le N$ square-free with $(b, D) = 1$ that have $|\omega(b) - \log\log N| < (\log\log N)^{3/4}$ approaches 1 as $N \to \infty$. □

It should be noted that our bounds on the rate of convergence in Theorem 3 are noneffective in two places. One is our treatment in this last section. We assume that we do not have an appropriate limit and proceed to find a contradiction. This is not a serious obstacle, and if techniques similar to those of [Heath-Brown 1994] were used instead, it could be overcome. The more serious problem comes in our proof of Proposition 9, where we make use of noneffective bounds on the size of Siegel zeroes. In particular, the rate of convergence depends on the function

$Z(\epsilon)$, which is the largest modulus q of a Dirichlet character with a Siegel zero larger than $1-q^{\epsilon}$ (or 1 if no such q exists). It should then be the case that, if for a sufficiently large constant K and integer $m>d$ we have that $N>\exp(Z(K^{-m})^{K})$ and $N>\exp(\exp(e^{Kd}))$, then

$$\left|\frac{\#\{b\le N:\dim(S_2(E_b))=d\}}{N}-\alpha_d\right|\le O_E\bigl(2^{-\binom{d}{2}}(\log\log N)^{-1/8}+2^{-\binom{d}{2}-m^2}\bigr).$$

References

[Friedlander et al. 2013] J. B. Friedlander, H. Iwaniec, B. Mazur, and K. Rubin, "The spin of prime ideals", *Invent. Math.* (2013). arXiv 1110.6331

[Hardy and Ramanujan 1917] G. H. Hardy and S. Ramanujan, "The normal number of prime factors of a number", *Quart. J. Math.* **48** (1917), 76–92. JFM 46.0262.03

[Heath-Brown 1994] D. R. Heath-Brown, "The size of Selmer groups for the congruent number problem, II", *Invent. Math.* **118**:2 (1994), 331–370. MR 95h:11064 Zbl 0815.11032

[Iwaniec and Kowalski 2004] H. Iwaniec and E. Kowalski, *Analytic number theory*, American Mathematical Society Colloquium Publications **53**, American Mathematical Society, Providence, RI, 2004. MR 2005h:11005 Zbl 1059.11001

[Klagsbrun et al. 2013] Z. Klagsbrun, B. Mazur, and K. Rubin, "Disparity in Selmer ranks of quadratic twists of elliptic curves", *Ann. of Math.* **178**:1 (2013), 287–320. arXiv 1111.2321

[Kramer 1981] K. Kramer, "Arithmetic of elliptic curves upon quadratic extension", *Trans. Amer. Math. Soc.* **264**:1 (1981), 121–135. MR 82g:14028 Zbl 0471.14020

[Mazur and Rubin 2010] B. Mazur and K. Rubin, "Ranks of twists of elliptic curves and Hilbert's tenth problem", *Invent. Math.* **181**:3 (2010), 541–575. MR 2012a:11069 Zbl 1227.11075

[Swinnerton-Dyer 2008] P. Swinnerton-Dyer, "The effect of twisting on the 2-Selmer group", *Math. Proc. Cambridge Philos. Soc.* **145**:3 (2008), 513–526. MR 2010d:11059 Zbl 1242.11041

[Yu 2005] G. Yu, "On the quadratic twists of a family of elliptic curves", *Mathematika* **52**:1-2 (2005), 139–154. MR 2007f:11058 Zbl 1105.14044

Communicated by Karl Rubin
Received 2012-06-27 Revised 2013-01-04 Accepted 2013-01-06

dankane@math.stanford.edu Stanford University, Department of Mathematics,
Building 380, Sloan Hall, Stanford, CA, 94305, United States

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the ANT website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in ANT are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. 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.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

Algebra & Number Theory

Volume 7 No. 5 2013

Sharp upper bounds for the Betti numbers of a given Hilbert polynomial 1019
GIULIO CAVIGLIA and SATOSHI MURAI

Comparing numerical dimensions 1065
BRIAN LEHMANN

Some consequences of a formula of Mazur and Rubin for arithmetic local constants 1101
JAN NEKOVÁŘ

Quantized mixed tensor space and Schur–Weyl duality 1121
RICHARD DIPPER, STEPHEN DOTY and FRIEDERIKE STOLL

Weakly commensurable S-arithmetic subgroups in almost simple algebraic groups of types B and C 1147
SKIP GARIBALDI and ANDREI RAPINCHUK

Minimisation and reduction of 5-coverings of elliptic curves 1179
TOM FISHER

On binary cyclotomic polynomials 1207
ÉTIENNE FOUVRY

Local and global canonical height functions for affine space regular automorphisms 1225
SHU KAWAGUCHI

On the ranks of the 2-Selmer groups of twists of a given elliptic curve 1253
DANIEL M. KANE