Algebra & Number Theory

Volume 8

2014

No. 8

msp

Algebra & Number Theory

msp.org/ant

PRODUCTION

production@msp.org

Silvio Levy, Scientific Editor

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

The subscription price for 2014 is US $225/year for the electronic version, and $400/year (+$55, 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 MSP.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

ALGEBRA AND NUMBER THEORY 8:8 (2014)
dx.doi.org/10.2140/ant.2014.8.1787

Relative cohomology of cuspidal forms on PEL-type Shimura varieties

Kai-Wen Lan and Benoît Stroh

We present a short proof that, for PEL-type Shimura varieties, subcanonical extensions of automorphic bundles, whose global sections over toroidal compactifications of Shimura varieties are represented by cuspidal automorphic forms, have no higher direct images under the canonical morphism to the minimal compactification, in characteristic zero or in positive characteristics greater than an explicitly computable bound.

1. Introduction

The main goal of this article is to present a short proof of Theorem 1.1 below, as an application of a certain vanishing theorem of automorphic bundles in mixed characteristics. (We refer to [Lan 2013; Lan and Suh 2012; 2013] for the precise definitions and descriptions of smooth integral models of PEL-type Shimura varieties and their various compactifications, and of the automorphic bundles and their canonical and subcanonical extensions.)

Let $\pi : \mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma} \to \mathsf{M}^{\mathrm{min}}_{\mathcal{H}}$ denote the canonical proper morphism from any projective smooth toroidal compactification to the minimal compactification of a p-integral model $\mathsf{M}_{\mathcal{H}}$ of a PEL-type Shimura variety at a neat level $\mathcal{H} \subset \mathrm{G}(\widehat{\mathbb{Z}}^p)$, where p is *good* for the integral PEL datum $(\mathcal{O}, \star, L, \langle \cdot , \cdot \rangle, h_0)$ defining $\mathsf{M}_{\mathcal{H}}$, as in [Lan

Lan is partially supported by the National Science Foundation under agreements Nos. DMS-1258962 and DMS-1352216, and by an Alfred P. Sloan Research Fellowship. Stroh is partially supported by the A.N.R. (Agence Nationale de la Recherche) under the project ANR-10-BLAN-0114 ArShiFo.
MSC2010: primary 14G35; secondary 14K10, 14G17, 32L20.
Keywords: Shimura varieties, vanishing theorem, toroidal compactification, minimal compactification, cuspidal forms.

and Suh 2013, §4.1] (and the references there). Let $W_{\nu_0,R} := W_{\nu_0,\mathbb{Z}} \otimes_{\mathbb{Z}} R$ be a representation of M_1 of weight $\nu_0 \in \mathrm{X}^+_{\mathrm{M}_1}$ over a coefficient ring R, where $W_{\nu_0,\mathbb{Z}}$ denotes a Weyl module of weight ν_0 of a split model $\mathrm{M}_{\mathrm{split}}$ of M_1 over $\mathbb{Z}$, as in [Lan and Suh 2012, §2.6]. Let $\underline{W}_{\nu_0,R} := \mathcal{E}_{\mathrm{M}_1,R}(W_{\nu_0,R})$ be the corresponding automorphic bundle over $\mathsf{M}_{\mathcal{H}}$, as in [Lan and Suh 2012, Definition 1.16 and §6.3], and let $\underline{W}^{\mathrm{sub}}_{\nu_0,R} := \mathcal{E}^{\mathrm{sub}}_{\mathrm{M}_1,R}(W_{\nu_0,R})$ be its subcanonical extension over $\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},R}$, as in [Lan and Suh 2013, Definition 4.12 and §7]. (We similarly define $W_{\nu,R}$, $\underline{W}_{\nu,R}$, and $\underline{W}^{\mathrm{sub}}_{\nu,R}$ for all $\nu \in \mathrm{X}^+_{\mathrm{M}_1}$.)

Theorem 1.1. *With the setting as above, there exists a bound $C(\nu_0)$ depending only on the integral PEL datum $(\mathcal{O}, \star, L, \langle\,\cdot\,,\,\cdot\,\rangle, h_0)$ and the weight ν_0, such that*

$$R^i \pi_* \underline{W}^{\mathrm{sub}}_{\nu_0,R} = 0 \tag{1.2}$$

for all $i > 0$ when the residue characteristics of R are zero or p greater than $C(\nu_0)$. (See Lemma 3.3 below for an explicit choice of $C(\nu_0)$.)

To help the reader understand the restriction imposed by $C(\nu_0)$, let us spell out the bound in some simple special cases. If $\nu_0 = 0$, then we can take $C(\nu_0)$ to be the relative dimension d of $\mathsf{M}_{\mathcal{H}}$ over the base scheme S_0 (see Example 3.9 below). If $\mathsf{M}_{\mathcal{H}}$ is a p-integral model of the Siegel modular variety of genus three, then the weight ν_0 is of the form $(k_1, k_2, k_3; k_0)$ for some integers k_0 and $k_1 \geq k_2 \geq k_3$, and we can take $C(\nu_0)$ to be $6 + (k_1 - k_3) + (k_2 - k_3)$ (see Example 3.10 below with $r = 3$ there). If $\mathsf{M}_{\mathcal{H}}$ is a p-integral model of a Picard modular surface, then the weight ν_0 is of the form $(k_1, k_2, k_3; k_0)$ for some integers k_0, k_1, and $k_2 \geq k_3$, and we can take $C(\nu_0)$ to be $2 + (k_2 - k_3)$ (see Example 3.12 below with $(r-q, q) = (2, 1)$ there). (In all cases, $C(\nu_0)$ is insensitive to shifting the weight ν_0 by a "parallel weight". See Section 3C below for more examples.)

We note that, when $R = \mathbb{C}$, global sections of $\underline{W}^{\mathrm{sub}}_{\nu_0,R}$ over $\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma}$ can be represented by holomorphic cuspidal automorphic forms. (See, e.g., [Harris 1990b, Proposition 5.4.2]; see also [Harris 1990a] for a survey on how the higher cohomology of $\underline{W}^{\mathrm{sub}}_{\nu_0,R}$ can be represented by nonholomorphic automorphic forms. See [Lan 2012] for the comparison between algebraic and analytic constructions hidden behind this.) Combined with the Leray spectral sequence, Theorem 1.1 allows one to identify the cohomology of $\underline{W}^{\mathrm{sub}}_{\nu_0,R}$ over $\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma}$ with the cohomology of $\pi_* \underline{W}^{\mathrm{sub}}_{\nu_0,R}$ over $\mathsf{M}^{\mathrm{min}}_{\mathcal{H}}$. Although the coherent sheaf $\pi_* \underline{W}^{\mathrm{sub}}_{\nu_0,R}$ is not locally free in general, there are reasons for $\mathsf{M}^{\mathrm{min}}_{\mathcal{H}}$ to be useful for the construction of p-adic modular forms and p-adic Galois representations.

Special cases of Theorem 1.1 have been independently proved in [Andreatta et al. 2013a; 2013b] (in the Siegel and Hilbert cases, for trivial weight ν_0) and in [Harris et al. 2013] (in the unitary case, for all weights ν_0), without any assumption on the residue characteristic p. The idea in [Harris et al. 2013] has also been carried out for all PEL-type cases in [Lan 2014]. Such results have played crucial roles in positive characteristics in [Andreatta et al. 2013a; 2013b; Emerton et al. 2013; Pilloni and Stroh 2013], and in characteristic zero in [Harris et al. 2013; Tian and Xiao 2013]. The proofs in [Andreatta et al. 2013a; 2013b] and [Harris et al. 2013; Lan 2014] directly used the toroidal and minimal boundary structures, and hence can be considered more elementary, which is why they work for all residue characteristics p; but they are lengthier and arguably more complicated. It is not easy to see from their proofs why Theorem 1.1 should be true. (It is not even clear how the two strategies in [Andreatta et al. 2013a; 2013b] and [Harris et al. 2013; Lan 2014] are related to each other.) Thus it is desirable to find a proof more closely related to other vanishing statements, at least when the residue characteristics are zero or sufficiently large.

It was first observed by the second author that this is indeed possible — in characteristic zero, the trivial weight case can be deduced from Grauert and Riemenschneider's vanishing theorem [1970]; in positive characteristics, under suitable assumptions (involving choices of projective but generally nonsmooth cone decompositions Σ for the toroidal compactification $\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma}$, whose existence is not very clearly documented in the literature), it is also possible to deduce the statement from Deligne and Illusie's [1987] and Kato's [1989] vanishing theorems. Then the first author made the observations that the assumption on cone decompositions can be relaxed by using Esnault and Viehweg's [1992] vanishing theorem as in [Lan and Suh 2011], and that (along similar lines) cases of nontrivial weights can be treated using stronger vanishing theorems in [Lan and Suh 2013]. (In the Siegel case, one can also use [Stroh 2010; 2013].)

In Section 2, we will present the proof of Theorem 1.1 and highlight the main inputs. In Section 3, we will carry out some elementary computations needed in the proof of Theorem 1.1, and find an explicit choice of $C(\nu_0)$. In Section 4, we sketch a logically simpler proof for the trivial weight case.

2. Proof of the theorem

Let $\pi : \mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma} \to \mathsf{M}^{\mathrm{min}}_{\mathcal{H}}$, $\nu_0 \in \mathrm{X}^{+}_{\mathrm{M}_1}$, and $\underline{W}^{\mathrm{sub}}_{\nu_0,R}$ be as in Section 1. Since $\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,1}$ and $\mathsf{M}^{\mathrm{min}}_{\mathcal{H},1}$ are proper over $\mathsf{S}_1 = \mathrm{Spec}(R_1)$ (see [Lan and Suh 2013, §4.1] and the references there for the notation), which are in particular separated and of finite type, for the purpose of proving Theorem 1.1 we may write R as an inductive limit over its sub-R_1-algebras and assume that R is of finite type over R_1, which is in particular noetherian. Then we may base change to R and abusively denote $\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,R} \to \mathsf{M}^{\mathrm{min}}_{\mathcal{H},R}$ by the same notation π. Our goal is to show that $R^i\pi_*\underline{W}^{\mathrm{sub}}_{\nu_0,R} = 0$ for all $i > 0$.

As in [Lan and Suh 2012, §2.6], we shall denote by $\mathrm{X}^{+,<p}_{\mathrm{M}_1}$ the subset of $\mathrm{X}^{+}_{\mathrm{M}_1}$ consisting of p-small weights, namely the weights $\nu \in \mathrm{X}^{+}_{\mathrm{M}_1}$ such that $(\nu + \rho_{\mathrm{M}_1}, \alpha) \leq p$ for all roots $\alpha \in \Phi_{\mathrm{M}_1}$, where ρ_{M_1} is the usual half sum of positive roots.

2A. ***Application of Serre's fundamental theorem.*** By [Lan and Suh 2013, Proposition 7.13], there exists some weight $\nu_1 \in \mathrm{X}^{+,<p}_{\mathrm{M}_1}$ such that $W_{\nu_1,R}$ is free of rank one as an R-module, and such that there exists an *ample line bundle* ω_{ν_1} over $\mathsf{M}^{\min}_{\mathcal{H},R}$ such that

$$\pi^*\omega_{\nu_1} \cong \underline{W}^{\mathrm{can}}_{\nu_1,R}, \tag{2.1}$$

the canonical extension $\underline{W}^{\mathrm{can}}_{\nu_1,R}$ of $\underline{W}_{\nu_1,R}$. Since (by definition)

$$\underline{W}^{\mathrm{sub}}_{\nu_0+N\nu_1,R} \cong \underline{W}^{\mathrm{sub}}_{\nu_0,R} \otimes_{\mathcal{O}_{\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,R}}} (\underline{W}^{\mathrm{can}}_{\nu_1,R})^{\otimes N} \tag{2.2}$$

for all integers N, by the projection formula [EGA 1960, 0_{I}, (5.4.10.1), p. 52] we have

$$R^i\pi_*\underline{W}^{\mathrm{sub}}_{\nu_0+N\nu_1,R} \cong (R^i\pi_*\underline{W}^{\mathrm{sub}}_{\nu_0,R})\otimes_{\mathcal{O}_{\mathsf{M}^{\min}_{\mathcal{H},R}}} \omega_{\nu_1}^{\otimes N}. \tag{2.3}$$

Then we have the following:

Lemma 2.4. *There exists some integer $N_1 \geq 0$ such that, for all integers $N \geq N_1$ and all $i \geq 0$, the sheaves $R^i\pi_*\underline{W}^{\mathrm{sub}}_{\nu_0+N\nu_1,R}$ over $\mathsf{M}^{\min}_{\mathcal{H},R}$ are generated by their global sections and satisfy $H^j(\mathsf{M}^{\min}_{\mathcal{H},R}, R^i\pi_*\underline{W}^{\mathrm{sub}}_{\nu_0+N\nu_1,R}) = 0$ for all $j > 0$.*

Proof. Since π is proper and $\mathsf{M}^{\min}_{\mathcal{H},R}$ is noetherian, by the theorem of finiteness [EGA 1961, III, Théorème (3.2.1), p. 116], the sheaves $R^i\pi_*\underline{W}^{\mathrm{sub}}_{\nu_0,R}$ are coherent over $\mathsf{M}^{\min}_{\mathcal{H},R}$ for all $i \geq 0$, and are nonzero only for finitely many i. Since ω_{ν_1} is ample over $\mathsf{M}^{\min}_{\mathcal{H},R}$, the lemma follows from (2.3) and Serre's fundamental theorem for projective schemes [EGA 1961, III, Théorème (2.2.1), p. 100]. □

2B. ***Shifting weights into the holomorphic chamber.*** Let w_0 (resp. w_1) be the longest Weyl element in $\mathrm{W}_{\mathrm{M}_1}$ (resp. $\mathrm{W}^{\mathrm{M}_1}$) (see [Lan and Suh 2012, §2.4]), so that $(-w_0)\Phi^+_{\mathrm{M}_1} = \Phi^+_{\mathrm{M}_1}$ and $W_\nu \cong W^\vee_{-w_0(\nu)}$ for all $\nu \in \mathrm{X}^{+,<p}_{\mathrm{M}_1}$ and $l(w_1) = d = \dim_{\mathsf{S}_1}(\mathsf{M}_{\mathcal{H},1})$.

Remark 2.5. When $R = \mathbb{C}$, for any $\mu \in \mathrm{X}^+_{\mathrm{G}_1}$, sections in $H^0(\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,R}, (\underline{W}^\vee_{w_1\cdot\mu,R})^{\mathrm{sub}})$ are represented by holomorphic cusp forms of weight $(-w_0)(w_1\cdot\mu) \in \mathrm{X}^+_{\mathrm{M}_1}$, which contribute via the dual BGG spectral sequence to

$$H^d_{\log-\mathrm{dR}}(\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},R}, (\underline{V}^\vee_{[\mu],R})^{\mathrm{sub}}) \cong H^d_{\mathrm{dR,c}}(\mathsf{M}_{\mathcal{H},R}, \underline{V}^\vee_{[\mu],R})$$

(compactly supported of middle degree), compatible with their contribution to the better-understood L^2 cohomology of $\mathsf{M}_{\mathcal{H},R}$. (For more explanations see [Faltings 1983, Theorem 9; Harris 1990a, §2; 1990b, Proposition 5.4.2]; see also the comparisons with transcendental results in [Lan and Suh 2012; 2013] and the references there.) Thus we consider weights of the form

$$(-w_0)(w_1\cdot\mu) = (-w_0w_1)(\mu) + (-w_0)(w_1\cdot 0)$$

holomorphic; these holomorphic weights form a translation of the dominant chamber $\mathrm{X}^+_{\mathrm{G}_1}$ because $(-w_0w_1)$ preserves $\mathrm{X}^+_{\mathrm{G}_1}$.

Proposition 2.6. *There exists an integer N_2, a positive parallel weight $\nu_2 \in \mathrm{X}^+_{\mathrm{M}_1}$, and a weight $\mu_0 \in \mathrm{X}^+_{\mathrm{G}_1}$, all of which can be explicitly determined, such that*

$$\nu_0 + N_2\nu_1 - \nu_2 = -w_0(w_1 \cdot \mu_0) \tag{2.7}$$

This proposition is elementary in nature. One can prove Proposition 2.6 using general principles that also work for all reductive groups defining Shimura varieties. However, we shall spell out a (less elegant) case-by-case argument, which has the advantage of giving explicit choices of N_2, ν_2, and μ_0 of small sizes.

We will assume Proposition 2.6 in the remainder of this section, and postpone its proof until Section 3A. In Lemma 3.3, we will give an explicit choice of $C(\mu_0)$, depending only on $(\mathcal{O}, \star, L, \langle\,\cdot\,,\cdot\,\rangle, h_0)$ and the weight ν_0, such that $C(\nu_0) \geq |\mu_0|_{\mathrm{re}}$ (see [Lan and Suh 2012, Definition 3.9]) for some triple (N_2, ν_2, μ_0) as in Proposition 2.6.

2C. *Application of automorphic vanishing.*

Corollary 2.8. *Let (N_2, ν_2, μ_0) be any triple as in Proposition 2.6. Suppose that $p > |\mu_0|_{\mathrm{re}}$ and that N is any integer satisfying $N \geq N_2$. Then we have*

$$H^i(\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,R}, \underline{W}^{\mathrm{sub}}_{\nu_0+N\nu_1,R}) = 0 \quad \textit{for every } i > 0.$$

Proof. By definition, the subset $\mathrm{X}^{+,<p}_{\mathrm{M}_1}$ of $\mathrm{X}^+_{\mathrm{M}_1}$ is preserved by translations by parallel weights. Moreover, by [Lan and Suh 2012, Remark 2.30], and by the same argument as in the proof of [Lan and Suh 2012, Lemma 7.20], we have $\nu_0 \in \mathrm{X}^{+,<p}_{\mathrm{M}_1}$ under the assumption that $p > |\mu_0|_{\mathrm{re}}$. Then the assertion $H^i(\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,R}, \underline{W}^{\mathrm{sub}}_{\nu_0+N\nu_1,R}) = 0$ follows from [Lan and Suh 2013, Theorem 8.13(2)], because $\nu := \nu_0 + N\nu_1$ and $\nu_+ := (N - N_2)\nu_1 + \nu_2$ satisfy the condition there, with $\mu(\nu - \nu_+) = \mu_0 \in \mathrm{X}^{+,<_{\mathrm{re}}p}_{\mathrm{G}_1}$ and $w(\nu) = w_1$ (so that $d - l(w(\nu)) = d - l(w_1) = 0$). □

Remark 2.9 (erratum). There are typos in [Lan and Suh 2013, Theorem 8.13]: both instances of $\mathrm{X}^{+,<_{\mathrm{W}}p}_{\mathrm{G}_1}$ there should be $\mathrm{X}^{+,<_{\mathrm{re}}p}_{\mathrm{G}_1}$, which is what was used in [Lan and Suh 2013, Corollary 7.24], on which the theorem depends.

2D. *End of the proof of Theorem 1.1.*

Let N_1 be as in Lemma 2.4, and let (N_2, ν_2, μ_0) be any triple as in Proposition 2.6 satisfying $C(\nu_0) \geq |\mu_0|_{\mathrm{re}}$ for some $C(\nu_0)$ (which will be given in Lemma 3.3 below). Suppose that $p > C(\nu_0)$ and that N is any integer satisfying $N \geq N_1$ and $N \geq N_2$. By Lemma 2.4 and by the Leray spectral sequence, and by Corollary 2.8, we have

$$H^0(\mathsf{M}^{\mathrm{min}}_{\mathcal{H},R}, R^i\pi_*\underline{W}^{\mathrm{sub}}_{\nu_0+N\nu_1,R}) \cong H^i(\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,R}, \underline{W}^{\mathrm{sub}}_{\nu_0+N\nu_1,R}) = 0 \tag{2.10}$$

for all $i > 0$. Since $R^i\pi_*\underline{W}^{\mathrm{sub}}_{\nu_0+N\nu_1,R}$ is generated by its global sections (by Lemma 2.4) it follows that

$$R^i\pi_*\underline{W}^{\mathrm{sub}}_{\nu_0+N\nu_1,R} = 0 \tag{2.11}$$

for all $i > 0$. By combining (2.3) and (2.11), we obtain the desired vanishing (1.2) for all $i > 0$ (under the assumption that $p > C(\nu_0) \geq |\mu_0|_{\mathrm{re}}$).

Suppose that the residue characteristics of R are all zero. By shrinking R and enlarging R by flat descent, we may replace the setup with a different one in which $p > C(\nu_0) \geq |\mu_0|_{\mathrm{re}}$, and obtain the desired vanishing from the above.

Thus, Theorem 1.1 follows. □

3. Elementary computations

We shall freely use the notation in [Lan and Suh 2012, §2 and §7]. The material in this section can be read without any knowledge of algebraic geometry or Shimura varieties.

3A. ***Proof of Proposition 2.6.*** We can rewrite (2.7) as

$$\nu_0 + N_2\nu_1 - \nu_2 = -w_0(w_1\mu_0 + w_1\rho - \rho) = \mu_0' + (-w_0)(w_1 \cdot 0),$$

where $\mu_0' = -(w_0 w_1)(\mu_0) \in \mathrm{X}^+_{\mathrm{G}_1}$ satisfies $V_{[\mu_0']} \cong V^\vee_{[\mu_0]}$, because $w_0 w_1$ is the longest Weyl element in $\mathrm{W}_{\mathrm{G}_1}$. Hence it suffices to find N_2 and ν_2 such that

$$\mu_0' = \nu_0 + N_2\nu_1 - \nu_2 - (-w_0)(w_1 \cdot 0) \in \mathrm{X}^+_{\mathrm{G}_1}\,. \tag{3.1}$$

Let us write $\nu_j = ((\nu_{j,\tau})_{\tau\in\Upsilon/c};\, \nu_{j,0}) = \big(((\nu_{j,\tau,i_\tau})_{1\leq i_\tau\leq r_\tau})_{\tau\in\Upsilon/c};\, \nu_{j,0}\big) \in \mathrm{X}^+_{\mathrm{M}_1}$ for $j = 0, 1, 2$. We shall also denote by ρ_τ (resp. $w_{0,\tau}$, $w_{1,\tau}$) the corresponding factors of ρ (resp. w_0, w_1). Then we need

$$\mu_{0,\tau}' = \nu_{0,\tau} + N_2\nu_{1,\tau} - \nu_{2,\tau} - (-w_{0,\tau})(w_{1,\tau} \cdot 0) \in \mathrm{X}^+_{\mathrm{G}_\tau} \tag{3.2}$$

for each factor G_τ of G_1. There are two cases:

(1) If $\tau = \tau \circ c$, then $\mathrm{G}_\tau \cong \mathrm{Sp}_{2r_\tau} \otimes_{\mathbb{Z}} R_1$ or $\mathrm{G}_\tau \cong \mathrm{O}_{2r_\tau} \otimes_{\mathbb{Z}} R_1$, and $\mathrm{M}_\tau \cong \mathrm{GL}_{r_\tau} \otimes_{\mathbb{Z}} R_1$. If $\mathrm{G}_\tau \cong \mathrm{Sp}_{2r_\tau} \otimes_{\mathbb{Z}} R_1$, set $d_\tau = \frac{1}{2}r_\tau(r_\tau + 1)$ and $r_\tau' = r_\tau + 1$. If $\mathrm{G}_\tau \cong \mathrm{O}_{2r_\tau} \otimes_{\mathbb{Z}} R_1$, set $d_\tau = \frac{1}{2}r_\tau(r_\tau - 1)$ and $r_\tau' = r_\tau$. Set $e_\tau = (1, 1, \ldots, 1)$. If $d_{[\tau]_{\mathbb{Q}}} = \sum_{\tau'\in[\tau]_{\mathbb{Q}}} d_{\tau'} = 0$, then we must have $\mathrm{G}_\tau \cong \mathrm{O}_{2r_\tau} \otimes_{\mathbb{Z}} R_1$ and $r_\tau \leq 1$, in which case (3.2) is trivially true if we take $\mu_{0,\tau}' = \nu_{0,\tau}$, any $N_2 \in \mathbb{Z}$, and $\nu_{2,\tau} = N_2\nu_{1,\tau} - (-w_{0,\tau})(w_{1,\tau} \cdot 0)$. Hence we may assume that $d_{[\tau]_{\mathbb{Q}}} > 0$. By assumption, we know that $\nu_{0,\tau,1} \geq \nu_{0,\tau,2} \geq \cdots \geq \nu_{0,\tau,r_\tau}$, and that $\nu_{1,\tau} = k_{1,\tau}e_\tau$, where $k_{1,\tau} > 0$ depends only on the equivalence class $[\tau]_{\mathbb{Q}}$ of τ (see [Lan and Suh 2012, Definition 7.12]). Also, we have $\rho_\tau = (r_\tau', r_\tau' - 1, \ldots, r_\tau' - r_\tau)$ and $(-w_{0,\tau})(w_{1,\tau} \cdot 0) = r_\tau' e_\tau$. Thus, in order for (3.2) to hold, we need

$$\nu_{0,r_\tau} + Nk_{1,\tau} - k_{2,\tau} \geq r_\tau + 1 = r_\tau' \quad \text{if } \mathrm{G}_\tau \cong \mathrm{Sp}_{2r_\tau} \otimes_{\mathbb{Z}} R_1,$$

or

$$\nu_{0,r_\tau-1} + Nk_{1,\tau} - k_{2,\tau} - r_\tau \geq |\nu_{0,r_\tau} + Nk_{1,\tau} - k_{2,\tau} - r_\tau| \quad \text{if } \mathrm{G}_\tau \cong \mathrm{O}_{2r_\tau} \otimes_{\mathbb{Z}} R_1.$$

We may take:

(a) $\mu'_{0,\tau} := \nu_{0,\tau} - \nu_{0,[\tau]_{\mathbb{Q}}} e_\tau$, where $\nu_{0,[\tau]_{\mathbb{Q}}} := \min_{\tau' \in [\tau]_{\mathbb{Q}}} (\nu_{0,\tau',r_\tau})$;

(b) $\mu_{0,\tau} := -(w_{0,\tau} w_{1,\tau})(\mu'_{0,\tau}) = \mu'_{0,\tau}$; and

(c) N_τ to be any integer satisfying $\nu_{0,[\tau]_{\mathbb{Q}}} + N_\tau k_{1,\tau} > r'_\tau$, so that

$$\nu_{0,\tau} + N\nu_{1,\tau} - \mu'_{0,\tau} - (-w_{0,\tau})(w_{1,\tau} \cdot 0) = (\nu_{0,[\tau]_{\mathbb{Q}}} + Nk_{1,\tau} - r'_\tau)\, e_\tau,$$

with a positive coefficient $\nu_{0,[\tau]_{\mathbb{Q}}} + Nk_{1,\tau} - r'_\tau > 0$ for every $N \geq N_\tau$.

(2) If $\tau \neq \tau \circ c$, then $\mathrm{G}_\tau \cong \mathrm{GL}_{r_\tau} \otimes_{\mathbb{Z}} R_1$ and $\mathrm{M}_\tau \cong (\mathrm{GL}_{q_\tau} \times \mathrm{GL}_{p_\tau}) \otimes_{\mathbb{Z}} R_1$. Set $d_\tau = p_\tau q_\tau$,

$$e_\tau = (\underbrace{1, 1, \ldots, 1}_{q_\tau}, 0, 0, \ldots, 0), \quad \text{and} \quad e'_\tau = (0, 0, \ldots, 0, \underbrace{-1, -1, \ldots, -1}_{p_\tau}).$$

If $d_{[\tau]_{\mathbb{Q}}} = \sum_{\tau' \in [\tau]_{\mathbb{Q}}/c} d_{\tau'} = 0$, then we must have $p_\tau q_\tau = 0$ for all $\tau \in [\tau]_{\mathbb{Q}}$, in which case (3.2) is trivially true if we take $\mu'_{0,\tau} = \nu_{0,\tau}$, any $N_2 \in \mathbb{Z}$, and $\nu_{2,\tau} = N_2 \nu_{1,\tau} - (-w_{0,\tau})(w_{1,\tau} \cdot 0)$. Hence we may assume that $d_{[\tau]_{\mathbb{Q}}} > 0$. By assumption, we know that

$$\nu_{0,\tau,1} \geq \nu_{0,\tau,2} \geq \cdots \geq \nu_{0,\tau,q_\tau} \quad \text{and} \quad \nu_{0,\tau,q_\tau+1} \geq \nu_{0,\tau,q_\tau+2} \geq \cdots \geq \nu_{0,\tau,r_\tau},$$

and that $\nu_{1,\tau} = k_{1,\tau} e_\tau + k_{1,\tau\circ c} e'_\tau$, where $[k_1]_\tau = k_{1,\tau} + k_{1,\tau\circ c} > 0$ depends only on the equivalence class $[\tau]_{\mathbb{Q}}$ of τ (see [Lan and Suh 2012, Proposition 7.15]). Also, we have $\rho_\tau = \frac{1}{2}(r_\tau - 1, r_\tau - 3, \ldots, -r_\tau + 1)$ and $(-w_{0,\tau})(w_{1,\tau} \cdot 0) = p_\tau e_\tau + q_\tau e'_\tau$. Thus, in order for (3.2) to hold, we need

$$\nu_{0,q_\tau} + Nk_{1,\tau} - k_{2,\tau} - p_\tau \geq \nu_{0,q_\tau+1} - Nk_{1,\tau\circ c} + k_{2,\tau\circ c} + q_\tau,$$

or equivalently

$$(\nu_{0,q_\tau} - \nu_{0,q_\tau+1}) + N[k_1]_\tau - [k_2]_\tau \geq p_\tau + q_\tau = r_\tau.$$

We may take:

(a) $\mu'_{0,\tau} := \nu_{0,\tau} - \nu_{0,[\tau]_{\mathbb{Q}}} e_\tau - (\nu'_{0,\tau,1} - \nu_{0,[\tau]_{\mathbb{Q}}})(e_\tau - e'_\tau)$, where

$$\nu_{0,[\tau]_{\mathbb{Q}}} := \min_{\tau' \in [\tau]_{\mathbb{Q}}, d_{\tau'} \neq 0} (\nu_{0,\tau',q_{\tau'}} - \nu_{0,\tau',q_{\tau'}+1}),$$
$$\nu'_{0,\tau,1} := \begin{cases} \nu_{0,\tau,1} & \text{if } q_\tau > 0, \\ \nu_{0,\tau,1} + \nu_{0,[\tau]_{\mathbb{Q}}} & \text{if } q_\tau = 0. \end{cases}$$

(b) $\mu_{0,\tau} := -(w_{0,\tau} w_{1,\tau})(\mu'_{0,\tau})$, which ends with $\mu_{0,\tau,r_\tau} = 0$ because $\mu'_{0,\tau}$ starts with $\mu'_{0,\tau,1} = 0$; and

(c) N_τ to be any integer satisfying $\nu_{0,[\tau]_{\mathbb{Q}}} + N_\tau [k_1]_\tau > r_\tau$, so that

$$\begin{aligned} &\nu_{0,\tau} + N\nu_{1,\tau} - \mu'_{0,\tau} - (-w_{0,\tau})(w_{1,\tau} \cdot 0) \\ &\qquad = (\nu_{0,\tau,1} + Nk_{1,\tau} - p_\tau)\, e_\tau + (\nu_{0,[\tau]_{\mathbb{Q}}} - \nu_{0,\tau,1} + Nk_{1,\tau\circ c} - q_\tau)\, e_\tau \end{aligned}$$

with sum of coefficients, for every $N \geq N_\tau$,

$$(\nu_{0,\tau,1}+Nk_{1,\tau}-p_\tau)+(\nu_{0,[\tau]_{\mathbb{Q}}}-\nu_{0,\tau,1}+Nk_{1,\tau\circ c}-q_\tau)=\nu_{0,[\tau]_{\mathbb{Q}}}+N[k_1]_\tau-r_\tau>0.$$

Now set:

$$\begin{aligned} N_2 &:= \max_{\tau\in\Upsilon/c}(N_\tau);\\ \mu_0 &:= ((\mu_{0,\tau})_{\tau\in\Upsilon/c};\,\mu_{0,0}) \quad \text{with any value of } \mu_{0,0};\\ \mu_0' &:= (-w_0w_1)(\mu_0);\\ \nu_2 &:= \nu_0+N_2\nu_1-\mu_0'-(-w_0)(w_1\cdot 0). \end{aligned}$$

Then the triple (N_2, ν_2, μ_0) satisfies (3.1) and hence also (2.7), as desired, because each of its factors $(N_2, \nu_{2,\tau}, \mu_{0,\tau})$ satisfies (3.2) by the above. □

3B. *Explicit choice of $C(\nu_0)$.*

Lemma 3.3. *The minimal size $|\mu_0|_{\mathrm{re}}$ (see [Lan and Suh 2012, Definition 3.9]) among all μ_0 appearing in some (N_2, ν_2, μ_0) satisfying* (2.7) *in Proposition 2.6 is smaller than or equal to*

$$C(\nu_0) := \sum_{\tau\in\Upsilon/c} C_\tau(\nu_{0,\tau}), \tag{3.4}$$

where each $C_\tau(\nu_{0,\tau})$ is defined as follows:

(1) *If $\tau = \tau\circ c$, then we set $d_\tau := \frac{1}{2}r_\tau(r_\tau+1)$ (resp. $d_\tau := \frac{1}{2}r_\tau(r_\tau-1)$) if* $\mathrm{G}_\tau \cong \mathrm{Sp}_{2r_\tau}\otimes_{\mathbb{Z}} R_1$ *(resp.* $\mathrm{G}_\tau \cong \mathrm{O}_{2r_\tau}\otimes_{\mathbb{Z}} R_1$*),* $\nu_{0,[\tau]_{\mathbb{Q}}} := \min_{\tau'\in[\tau]_{\mathbb{Q}}}(\nu_{0,\tau',r_\tau})$, *and*

$$C_\tau(\nu_{0,\tau}) := d_\tau + \sum_{1\leq i_\tau\leq r_\tau}(\nu_{0,\tau,i_\tau}-\nu_{0,[\tau]_{\mathbb{Q}}}). \tag{3.5}$$

(2) *If $\tau \neq \tau\circ c$, then we set $d_\tau := p_\tau q_\tau$,*

$$\begin{aligned} \nu_{0,[\tau]_{\mathbb{Q}}} &:= \min_{\tau'\in[\tau]_{\mathbb{Q}},d_{\tau'}\neq 0}(\nu_{0,\tau',q_{\tau'}}-\nu_{0,\tau',q_{\tau'}+1}),\\ \nu_{0,\tau,1}' &:= \begin{cases} \nu_{0,\tau,1} & \textit{if } q_\tau>0,\\ \nu_{0,\tau,1}+\nu_{0,[\tau]_{\mathbb{Q}}} & \textit{if } q_\tau=0, \end{cases} \end{aligned}$$

and

$$C_\tau(\nu_{0,\tau}) := d_\tau + \sum_{1\leq i_\tau\leq q_\tau}(\nu_{0,\tau,1}'-\nu_{0,\tau,i_\tau}) + \sum_{q_\tau<i_\tau\leq r_\tau}(\nu_{0,\tau,1}'-\nu_{0,[\tau]_{\mathbb{Q}}}-\nu_{0,\tau,i_\tau}). \tag{3.6}$$

Proof. These follow from the definition of $|\mu_0|_{\mathrm{re}} = d + \sum_{\tau\in\Upsilon/c}\bigl(\sum_{1\leq i_\tau\leq r_\tau}\mu_{0,\tau,i_\tau}\bigr)$ and the explicit choices of $\mu_{0,\tau}$ in the proof of Proposition 2.6. □

Remark 3.7. By using [Lan and Suh 2013, (7.9) and (7.11)], it is possible to reduce the proof of Theorem 1.1 to the case where the integral PEL datum is $\mathbb{Q}$-simple, and replace (3.4) with

$$C'(\nu_0) := \max_{[\tau]_{\mathbb{Q}}}\bigl(C_{[\tau]_{\mathbb{Q}}}(\nu_{0,[\tau]_{\mathbb{Q}}})\bigr), \tag{3.8}$$

where:

(1) $C_{[\tau]_{\mathbb{Q}}}(\nu_{0,[\tau]_{\mathbb{Q}}}) = 0$ if $d_{[\tau]_{\mathbb{Q}}} = \sum_{\tau'\in[\tau]_{\mathbb{Q}}/c} d_\tau \leq 1$;

(2) $C_{[\tau]_{\mathbb{Q}}}(\nu_{0,[\tau]_{\mathbb{Q}}}) = \sum_{\tau'\in[\tau]_{\mathbb{Q}}/c} C_\tau(\nu_{0,\tau})$, where $C_\tau(\nu_{0,\tau})$ are as in (3.5) and (3.6), otherwise.

We leave the details to the interested readers.

3C. ***Some examples.*** To help the reader understand the notation and formulas, we include some examples of familiar special cases.

Example 3.9 (trivial weight). If $\nu_0 = 0$, then (2.7) holds for $\mu_0 = 0$ and any sufficiently large N_2, and we have $C(\nu_0) = \sum_{\tau\in\Upsilon/c} C_\tau(\nu_{0,\tau}) = \sum_{\tau\in\Upsilon/c} d_\tau = d$ in (3.4).

Example 3.10 (Siegel case). Suppose $(\mathcal{O}, \star, L, \langle\,\cdot\,,\,\cdot\,\rangle, h_0)$ is given with $\mathcal{O} = \mathbb{Z}$ with trivial $\star$, with $(L, \langle\,\cdot\,,\,\cdot\,\rangle)$ given by $\mathbb{Z}^{\oplus 2r}$ with some standard self-dual symplectic pairing, and with any conventional choice of h_0. Then we are in the so-called *Siegel case*. There is a unique $\tau \in \Upsilon$ with $\tau = \tau \circ c$, which we can suppress in our notation, and each $\nu_0 \in \mathrm{X}^+_{\mathrm{M}_1}$ can be represented by a tuple $((\nu_{0,1}, \nu_{0,2}, \ldots, \nu_{0,r}); \nu_{0,0})$, where $\nu_{0,1} \geq \nu_{0,2} \geq \cdots \geq \nu_{0,r}$ are integers. Then μ_0 can be chosen to be

$$\nu_0 - \nu_{0,r}((1, 1, \ldots, 1, 1); 0) = ((\nu_{0,1} - \nu_{0,r}, \ldots, \nu_{0,r-1} - \nu_{0,r}, 0); \nu_{0,0})$$

(where the last entry is irrelevant), and then $C(\nu_0) = \frac{1}{2}r(r+1) + \sum_{1\leq i<r}(\nu_{0,i} - \nu_{0,r})$ (see (3.5)).

Example 3.11 ("$\mathbb{Q}$-similitude Hilbert case"). Suppose $(\mathcal{O}, \star, L, \langle\,\cdot\,,\,\cdot\,\rangle, h_0)$ is given with $\mathcal{O} = \mathcal{O}_F$ with trivial $\star$, where F is a totally real number field, with $(L, \langle\,\cdot\,,\,\cdot\,\rangle)$ given by $\mathcal{O}_F^{\oplus 2}$ with some standard symplectic pairing defined by trace, and with any conventional choice of h_0; and suppose p is any prime number unramified in $\mathcal{O}_F$. Then we are essentially in the so-called *Hilbert case*, although we only consider elements in $\mathrm{Res}_{F/\mathbb{Q}}\,\mathrm{GL}_2$ with similitudes in $\mathbb{G}_{\mathrm{m}}$ (rather than $\mathrm{Res}_{F/\mathbb{Q}}\,\mathbb{G}_{\mathrm{m}}$). There are d elements $\tau \in \Upsilon$ corresponding to the $d = [F : \mathbb{Q}]$ homomorphisms from $\mathcal{O}_F$ to an algebraic closure of $\mathbb{Q}_p$, which all satisfy $\tau = \tau \circ c$ and determine a unique equivalence class $[\tau]_{\mathbb{Q}}$ (of Galois orbits of τ), and our coefficient ring R is chosen to contain the images of all these homomorphisms, over which all linear algebraic data are split. Each $\nu_0 \in \mathrm{X}^+_{\mathrm{M}_1}$ can be represented by a tuple $((\nu_{0,\tau})_{\tau\in\Upsilon}; \nu_{0,0})$, where each $\nu_{0,\tau} = (\nu_{0,\tau,1})$ consists of just one integer $\nu_{0,\tau,1}$. Then $\nu_{0,[\tau]_{\mathbb{Q}}} = \min_{\tau\in\Upsilon}(\nu_{0,\tau,1})$, and μ_0 can be chosen to be $\nu_0 - \nu_{0,[\tau]_{\mathbb{Q}}}((1)_{\tau\in\Upsilon}; 0) = ((\nu_{0,\tau,1} - \nu_{0,[\tau]_{\mathbb{Q}}})_{\tau\in\Upsilon}; \nu_{0,0})$, and we have $C(\nu_0) = d + \sum_{\tau\in\Upsilon}(\nu_{0,\tau,1} - \nu_{0,[\tau]_{\mathbb{Q}}})$ (see (3.5)).

Example 3.12 (simplest unitary case). Suppose $(\mathcal{O}, \star, L, \langle\,\cdot\,,\cdot\,\rangle, h_0)$ is given with $\mathcal{O} = \mathcal{O}_F$, where F is an imaginary quadratic extension of $\mathbb{Q}$ with an embedding $F \hookrightarrow \mathbb{C}$, with $\star$ given by complex conjugation, with $(L, \langle\,\cdot\,,\cdot\,\rangle)$ given by a Hermitian module over $\mathcal{O}_F^{\oplus r}$ with signature $(r-q, q)$ at ∞ (using the given $F \hookrightarrow \mathbb{C}$), and with any conventional choice of h_0 (respecting the signature); and suppose p is any prime number unramified in $\mathcal{O}_F$. Then we obtain the simplest (nontrivial) *unitary case*. There is a unique representative τ of orbits in Υ/c such that $\tau \neq \tau \circ c$ and $(p_\tau, q_\tau) = (r-q, q)$, matching the signatures at ∞ and at p; hence we shall always choose this τ and suppress τ from the notation. Each $\nu_0 \in \mathrm{X}^+_{\mathrm{M}_1}$ can be represented by a tuple $((\nu_{0,1}, \nu_{0,2}, \dots, \nu_{0,q}, \nu_{0,q+1}, \dots, \nu_{0,r}); \nu_{0,0})$, where $\nu_{0,1} \geq \nu_{0,2} \geq \dots \geq \nu_{0,q}$ and $\nu_{0,q+1} \geq \dots \geq \nu_{0,r}$ are integers. If $q > 0$, then μ_0 can be chosen to be $(\nu_{0,1} - \nu_{0,q} + \nu_{0,q+1} - \nu_{0,r}, \dots, \nu_{0,1} - \nu_{0,q}, \nu_{0,1} - \nu_{0,q}, \dots, \nu_{0,1} - \nu_{0,2}, 0; \nu_{0,0})$ (note the reversed order and the repeated term $\nu_{0,1} - \nu_{0,q}$), and we have

$$C(\nu_0) = (r-q)q + \sum_{1 \leq i \leq q} (\nu_{0,1} - \nu_{0,i}) + \sum_{q < i \leq r} (\nu_{0,1} - \nu_{0,q} + \nu_{0,q+1} - \nu_{0,i}).$$

If $q = 0$, then μ_0 can be chosen to be $(\nu_{0,1} - \nu_{0,r}, \dots, \nu_{0,1} - \nu_{0,2}, 0; \nu_{0,0})$ and we have $C(\nu_0) = \sum_{1 \leq i \leq r} (\nu_{0,\tau,1} - \nu_{0,i})$; but $d = 0$ and the map π is trivial — $C(\nu_0) = 0$ suffices. (See (3.6) and Remark 3.7.)

4. Simpler proof for the trivial weight case

In this final section, we sketch a logically simpler proof for the trivial weight case $\nu_0 = 0$, which does not require the various advanced technical inputs in [Lan and Suh 2013, §§1–3] (such as the theory of F-spans in [Ogus 1994]). The key is to give a simpler proof of the vanishing statement in Corollary 2.8 when $\nu_0 = 0$ (with a suitable choice of (N_2, ν_2, μ_0)). By standard arguments, as in the proof of [Lan and Suh 2013, Theorem 8.2], we may and we shall assume that R is a perfect field extension of the residue field of R_1.

Using the extended Kodaira–Spencer isomorphism — see [Lan 2013, Theorem 6.4.1.1(4)] — and the very construction of canonical extensions of automorphic bundles using the relative Lie algebra of the universal abelian scheme, one can show that

$$\underline{W}^{\mathrm{can}}_{(-w_0)(w_1\cdot 0)} \cong (\underline{W}^{\vee}_{w_1\cdot 0})^{\mathrm{can}} \cong \Omega^d_{\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,1}/\mathsf{S}_1}(\log\infty) := \bigwedge^d (\Omega^1_{\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,1}/\mathsf{S}_1}(\log\infty))$$

as line bundles over $\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,1}$ (ignoring Tate twists). (The proof is left to the interested readers.) Moreover, the proof of Proposition 2.6 in Section 3A shows that we can take $\mu_0 = 0$ in Proposition 2.6, with some integer N_2 such that the weight $\nu_2 = N_2\nu_1 - (-w_0)(w_1 \cdot 0)$ is positive and parallel. Then we have

$$\underline{W}^{\mathrm{sub}}_{N\nu_1} \cong \underline{W}^{\mathrm{sub}}_{\nu_2} \otimes_{\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,1}} \underline{W}^{\mathrm{can}}_{(-w_0)(w_1\cdot 0)} \cong \underline{W}^{\mathrm{sub}}_{\nu_2} \otimes_{\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,1}} \Omega^d_{\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,1}/\mathsf{S}_1}(\log \mathsf{D}),$$

where D is the boundary divisor $\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,1} - \mathsf{M}_{\mathcal{H},1}$ (with reduced subscheme structure).

By [Lan and Suh 2013, Proposition 4.2(5) and Corollary 7.14], there exists a (usually nonreduced) divisor D' with $\mathsf{D}'_{\mathrm{red}} = \mathsf{D}$, and some $r_0 > 0$, such that the line bundle $(\underline{W}^{\mathrm{can}}_{\nu_2})^{\otimes r}(-\mathsf{D}')$ is ample for all integers $r \geq r_0$. (This follows from [Lan 2013, Theorem 7.3.3.4], which implies that there exists some D' as above such that $\mathcal{O}_{\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,1}}(-\mathsf{D}')$ is relatively ample over $\mathsf{M}^{\min}_{\mathcal{H},1}$.) By base change from R_1 to R, this is exactly the condition $(*)$ needed in [Esnault and Viehweg 1992, Theorem 11.5]. Then, by [Esnault and Viehweg 1992, Theorem 11.5] and by Serre duality, we obtain

$$H^i(\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,R}, \underline{W}^{\mathrm{sub}}_{N\nu_1,R}) = H^i(\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,R}, \underline{W}^{\mathrm{sub}}_{\nu_2,R} \otimes_{\mathcal{O}_{\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,1}}} \Omega^d_{\mathsf{M}^{\mathrm{tor}}_{\mathcal{H},\Sigma,1}/\mathsf{S}_1}(\log \mathsf{D})) = 0$$

for all $i > 0$. (This is the same approach taken in [Lan and Suh 2011].) This gives the desired vanishing statement in Corollary 2.8 when $\nu_0 = 0$, and we can conclude as in Section 2D. This argument does not depend on [Lan and Suh 2013, Theorem 8.13(2)], and hence not on the various advanced technical inputs in [Lan and Suh 2013, §§1–3].

Acknowledgements

The authors would like to thank the anonymous referee for valuable suggestions. Stroh would like to thank Esnault for very helpful discussions, and to thank Scholze for pointing out an error in an earlier approach.

References

[Andreatta et al. 2013a] F. Andreatta, A. Iovita, and V. Pilloni, "p-adic families of Siegel modular cuspforms", preprint, 2013, Available at http://perso.ens-lyon.fr/vincent.pilloni/AIP.pdf. To appear in *Ann. Math.* arXiv 1212.3812

[Andreatta et al. 2013b] F. Andreatta, A. Iovita, and V. Pilloni, "p-adic families of Hilbert modular forms", preprint, 2013, Available at http://perso.ens-lyon.fr/vincent.pilloni/AIP2.pdf.

[Deligne and Illusie 1987] P. Deligne and L. Illusie, "Relèvements modulo p^2 et décomposition du complexe de de Rham", *Invent. Math.* **89**:2 (1987), 247–270. MR 88j:14029 Zbl 0632.14017

[EGA 1960] A. Grothendieck, *Eléments de géométrie algébrique, I: Le langage des schémas*, Publications Mathématiques de l'I.H.E.S. **4**, Inst. Hautes Études Sci., Paris, 1960. MR 29 #1207 Zbl 0118.36206

[EGA 1961] A. Grothendieck, *Eléments de géométrie algébrique, III: Étude cohomologique des faisceaux cohérents, I*, Publications Mathématiques de l'I.H.E.S. **11**, Inst. Hautes Études Sci., Paris, 1961. MR 0163910 Zbl 0118.36206

[Emerton et al. 2013] M. Emerton, D. A. Reduzzi, and L. Xiao, "Galois representations and torsion in the coherent cohomology of Hilbert modular varieties", preprint, 2013. arXiv 1307.8003

[Esnault and Viehweg 1992] H. Esnault and E. Viehweg, *Lectures on vanishing theorems*, DMV Seminar **20**, Birkhäuser, Basel, 1992. MR 94a:14017 Zbl 0779.14003

[Faltings 1983] G. Faltings, "On the cohomology of locally symmetric Hermitian spaces", pp. 55–98 in *Séminaire d'algèbre Paul Dubreil et Marie-Paule Malliavin, 35ème année* (Paris, 1982), edited by M.-P. Malliavin, Lecture Notes in Math. **1029**, Springer, Berlin, 1983. MR 85k:22028 Zbl 0539.22008

[Grauert and Riemenschneider 1970] H. Grauert and O. Riemenschneider, "Verschwindungssätze für analytische Kohomologiegruppen auf komplexen Räumen", *Invent. Math.* **11** (1970), 263–292. MR 46 #2081 Zbl 0202.07602

[Harris 1990a] M. Harris, "Automorphic forms and the cohomology of vector bundles on Shimura varieties", pp. 41–91 in *Automorphic forms, Shimura varieties, and L-functions* (Ann Arbor, MI, 1988), vol. II, edited by L. Clozel and J. S. Milne, Perspect. Math. **11**, Academic Press, Boston, 1990. MR 91g:11063 Zbl 0716.14011

[Harris 1990b] M. Harris, "Automorphic forms of $\overline{\partial}$-cohomology type as coherent cohomology classes", *J. Differential Geom.* **32**:1 (1990), 1–63. MR 91g:11064 Zbl 0711.14012

[Harris et al. 2013] M. Harris, K.-W. Lan, R. Taylor, and J. Thorne, "On the rigid cohomology of certain Shimura varieties", preprint, Institute for Advanced Study, Princeton, NJ, 2013, Available at http://www.math.ias.edu/~rtaylor/rigcoh.pdf.

[Kato 1989] K. Kato, "Logarithmic structures of Fontaine–Illusie", pp. 191–224 in *Algebraic analysis, geometry, and number theory: proceedings of the JAMI Inaugural Conference* (Baltimore, MD, 1988), edited by J. I. Igusa, Johns Hopkins University Press, Baltimore, MD, 1989. MR 99b:14020 Zbl 0776.14004

[Lan 2012] K.-W. Lan, "Comparison between analytic and algebraic constructions of toroidal compactifications of PEL-type Shimura varieties", *J. Reine Angew. Math.* **664** (2012), 163–228. MR 2980135 Zbl 1242.14022

[Lan 2013] K.-W. Lan, *Arithmetic compactifications of PEL-type Shimura varieties*, London Mathematical Society Monographs Series **36**, Princeton University Press, 2013. MR 3186092 Zbl 1284.14004

[Lan 2014] K.-W. Lan, "Compactifications of PEL-type Shimura varieties and Kuga families with ordinary loci", preprint, 2014, Available at http://math.umn.edu/~kwlan/articles/cpt-ram-ord.pdf.

[Lan and Suh 2011] K.-W. Lan and J. Suh, "Liftability of mod p cusp forms of parallel weights", *Int. Math. Res. Not.* **2011**:8 (2011), 1870–1879. MR 2012e:11085 Zbl 1233.11042

[Lan and Suh 2012] K.-W. Lan and J. Suh, "Vanishing theorems for torsion automorphic sheaves on compact PEL-type Shimura varieties", *Duke Math. J.* **161**:6 (2012), 1113–1170. MR 2913102 Zbl 06029039

[Lan and Suh 2013] K.-W. Lan and J. Suh, "Vanishing theorems for torsion automorphic sheaves on general PEL-type Shimura varieties", *Adv. Math.* **242** (2013), 228–286. MR 3055995 Zbl 1276.11103

[Ogus 1994] A. Ogus, *F-crystals, Griffiths transversality, and the Hodge decomposition*, Astérisque **221**, Société Mathématique de France, Paris, 1994. MR 95g:14025 Zbl 0801.14004

[Pilloni and Stroh 2013] V. Pilloni and B. Stroh, "Surconvergence, ramification et modularité", preprint, 2013, Available at http://www.math.univ-paris13.fr/~stroh/Artin.pdf.

[Stroh 2010] B. Stroh, "Relèvement de formes modulaires de Siegel", *Proc. Amer. Math. Soc.* **138**:9 (2010), 3089–3094. MR 2011g:11093 Zbl 1257.11046

[Stroh 2013] B. Stroh, "Classicité en théorie de Hida", *Amer. J. Math.* **135**:4 (2013), 861–889. MR 3086063 Zbl 06203651

[Tian and Xiao 2013] Y. Tian and L. Xiao, "p-adic cohomology and classicality of overconvergent Hilbert modular forms", preprint, 2013. arXiv 1308.0779

Communicated by Richard Taylor
Received 2013-08-23 Revised 2014-08-03 Accepted 2014-10-08

kwlan@math.umn.edu *School of Mathematics, University of Minnesota,*
127 Vincent Hall, 206 Church Street SE,
Minneapolis, MN 55455, United States

stroh@math.univ-paris13.fr *C.N.R.S, Université Paris 13, LAGA, 99 avenue J.B. Clément,*
93430 Villetaneuse, France

mathematical sciences publishers msp

ALGEBRA AND NUMBER THEORY 8:8 (2014)
dx.doi.org/10.2140/ant.2014.8.1801

ℓ-modular representations of unramified p-adic U(2,1)

Robert James Kurinczuk

We construct all irreducible cuspidal ℓ-modular representations of a unitary group in three variables attached to an unramified extension of local fields of odd residual characteristic p with $\ell \neq p$. We describe the ℓ-modular principal series and show that the supercuspidal support of an irreducible ℓ-modular representation is unique up to conjugacy.

1. Introduction

The abelian category $\mathfrak{R}_R(G)$ of smooth representations of a reductive p-adic group G over an algebraically closed field R has been well studied when R has characteristic zero. The same cannot be said when R has positive characteristic ℓ; here many questions remain unanswered. In this paper, we are concerned only with the case $\ell \neq p$. We study the set $\mathrm{Irr}_R(G)$ of isomorphism classes of irreducible R-representations, eventually specialising to $G = \mathrm{U}(2, 1)$, a unitary group in three variables attached to an unramified extension F/F_0 of nonarchimedean local fields of odd residual characteristic. All R-representations henceforth considered will be smooth.

A classical strategy for the classification of irreducible R-representations is to split the problem into two steps: firstly, for any parabolic subgroup P of G with Levi decomposition $P = M \ltimes N$ and any $\sigma \in \mathrm{Irr}_R(M)$, decompose the (normalised) parabolically induced R-representation $i_P^G(\sigma)$; and, secondly, construct the irreducible R-representations which do not appear as a subquotient of an R-representation appearing in the first step, the *supercuspidal* R-representations. For any parabolic subgroup P, a supercuspidal irreducible R-representation π will have trivial Jacquet module $r_P^G(\pi) = 0$, by Frobenius reciprocity (i_P^G is right-adjoint to r_P^G). When R has characteristic zero the irreducible *cuspidal* R-representations, those whose Jacquet modules are all trivial, are all supercuspidal. However, in positive characteristic ℓ, there can exist irreducible cuspidal nonsupercuspidal R-representations.

MSC2010: 22E50.
Keywords: representations of p-adic groups, modular representations.

By transitivity of the Jacquet module and the geometric lemma — see [Vignéras 1996, II 2.19] — the *cuspidal support* of $\pi \in \mathrm{Irr}_R(G)$, that is, the set of pairs (M, σ) with M a Levi factor of a parabolic subgroup P of G and σ an irreducible cuspidal R-representation of M such that π is a subrepresentation of $i_P^G(\sigma)$, is a nonempty set consisting of a single G-conjugacy class; we say that the cuspidal support is unique up to conjugacy. By transitivity of parabolic induction, the *supercuspidal support* of $\pi \in \mathrm{Irr}_R(G)$, that is, the set of pairs (M, σ) with M a Levi factor of a parabolic subgroup P of G and σ an irreducible supercuspidal R-representation of M such that π is a subquotient of $i_P^G(\sigma)$, is nonempty. However, in general, it is not known if the supercuspidal support of an irreducible R-representation is unique up to conjugacy.

For GL_n and its inner forms, Vignéras [1996] and Mínguez and Sécherre [2014b; 2014a] showed that the supercuspidal support of an irreducible R-representation is unique up to conjugacy. The unicity of supercuspidal support is of great importance. Firstly, the unicity of supercuspidal support (up to inertia) for GL_n leads to the block decomposition of $\mathfrak{R}_R(G)$ into indecomposable summands; see [Vignéras 1998]. Secondly, it is important in Vignéras' ℓ-modular local Langlands correspondence for GL_n, which is first defined on supercuspidal elements by compatibility with the characteristic zero local Langlands correspondence and then extended to all irreducible ℓ-modular representations of GL_n. In this paper, we prove unicity of supercuspidal support for $\mathrm{U}(2, 1)$. We hope this is the first step in establishing similar results for $\mathrm{U}(2, 1)$ and in extending these to classical groups in general.

Our strategy is first to construct all irreducible cuspidal R-representations by compact induction from irreducible R-representations of compact open subgroups. The type of construction we employ has been used to great effect to construct all irreducible cuspidal R-representations in a large class of reductive p-adic groups when R has characteristic zero: [Morris 1999] for level zero R-representations of any reductive p-adic group, [Bushnell and Kutzko 1993a; 1993b] for GL_n and SL_n, [Sécherre and Stevens 2008] for inner forms of GL_n, [Yu 2001] and [Kim 2007] for arbitrary connected reductive groups under "tame" conditions, and [Stevens 2008] for classical p-adic groups with p odd. Vignéras [1996] and Mínguez and Sécherre [2014b; 2014a] adapted the characteristic zero constructions for GL_n and its inner forms to ℓ-modular representations. We perform similar adaptations to Stevens' construction to exhaust all irreducible cuspidal ℓ-modular representations of $\mathrm{U}(2, 1)$.

Theorem 5.3. *Let $G = \mathrm{U}(2, 1)$ and let π be an irreducible cuspidal R-representation of G. There exist a compact open subgroup J of G with pro-unipotent radical J^1 such that J/J^1 is a finite reductive group, an irreducible R-representation κ of J and an irreducible cuspidal R-representation σ of J/J^1 such that $\pi \simeq \mathrm{ind}_J^G(\kappa \otimes \sigma)$.*

The construction is explicit and, furthermore, all R-representations

$$\mathrm{I}_\kappa(\sigma) = \mathrm{ind}_J^G(\kappa \otimes \sigma)$$

constructed in this way are cuspidal. Moreover, we show that $\mathrm{I}_\kappa(\sigma)$ is supercuspidal if and only if σ is supercuspidal (Remark 8.2). In work in progress, joint with Stevens, we extend Stevens' construction for arbitrary classical groups to the ℓ-modular setting.

In the split case, for general linear groups all irreducible cuspidal ℓ-modular representations lift to integral ℓ-adic representations. For inner forms of GL_n, this is no longer true; some cuspidal nonsupercuspidal ℓ-modular representations do not lift. For U(2, 1) we also find cuspidal nonsupercuspidal ℓ-modular representations which do not lift (Remark 5.5). These nonlifting phenomena appear quite different. For U(2, 1) this nonlifting occurs because, in certain cases, there are ℓ-modular representations of the finite group J/J^1 which do not lift. For inner forms of GL_n, the nonlifting occurs when the normaliser of the reduction modulo ℓ of the inflation of a cuspidal ℓ-adic representation of an analogous group to J/J^1 is larger than the normaliser of all of its cuspidal lifts. We find that all supercuspidal ℓ-modular representations of U(2, 1) lift (Remark 8.2), as is the case for GL_n and its inner forms.

Secondly, by studying the corresponding Hecke algebras, we find the characters χ of the maximal diagonal torus T of U(2, 1) such that the principal series R-representation $i_B^{\mathrm{U}(2,1)}(\chi)$ is reducible. We let χ_1 denote the character of $F^\times$ given by $\chi_1(x) = \chi(\mathrm{diag}(x, \bar{x}x^{-1}, \bar{x}^{-1}))$, where $\bar{x}$ is the $\mathrm{Gal}(F/F_0)$-conjugate of x.

Theorem 6.2. *Let $G = \mathrm{U}(2, 1)$. Then $i_B^G(\chi)$ is reducible exactly in the following cases*:

(1) $\chi_1 = \nu^{\pm 2}$, *where ν is the absolute value on F*;

(2) $\chi_1 = \eta\nu^{\pm 1}$, *where η is any extension of the quadratic class field character ω_{F/F_0} to $F^\times$*;

(3) χ_1 *is nontrivial, but $\chi_1 |_{F_0^\times}$ is trivial.*

When R is of characteristic zero this is due to Keys [1984]. In our proof we need to apply his results to determine a sign. It should be possible to remove this dependency by computation using the theory of covers (cf. [Blondel 2012, Remark 3.13]). An alternative proof, when F_0 is of characteristic zero, would be to use the computations of [Keys 1984] with [Dat 2005, Proposition 8.4].

Finally, by studying the interaction of the right adjoints R_κ of the functors I_κ with parabolic induction we find cuspidal subquotients of the principal series. When cuspidal subquotients appear in the principal series we show exactly which ones from our exhaustive list do, finding that the supercuspidal support of an irreducible R-representation is unique up to conjugacy.

Theorem 8.1. *Let π be an irreducible R-representation of* $\mathrm{U}(2,1)$. *Then the supercuspidal support of π is unique up to conjugacy.*

In fact, in many cases, we obtain extra information on the irreducible quotients and subrepresentations which appear. If $\ell \neq 2$ and $\ell \mid q-1$, we show that all the principal series R-representations $i_B^{\mathrm{U}(2,1)}(\chi)$ are semisimple (Lemma 6.8). If $\ell \mid q+1$, we show that $i_B^{\mathrm{U}(2,1)}(\chi)$ has a unique irreducible subrepresentation and a unique irreducible quotient, and these are isomorphic (Lemma 6.10). A striking example of the reducibilities that occur is when $\chi = \nu^{-2}$.

Theorem (see Theorem 6.12 for more details). *Let $G = \mathrm{U}(2,1)$.*

(1) *If $\ell \nmid (q-1)(q+1)(q^2-q+1)$, then $i_B^G(\nu^{-2})$ has length two with unique irreducible subrepresentation 1_G and unique irreducible quotient* St_G.

(2) *If $\ell \neq 2$ and $\ell \mid q-1$, then $i_B^G(\nu^{-2}) = 1_G \oplus \mathrm{St}_G$ is semisimple of length two.*

(3) *If $\ell \neq 3$ and $\ell \mid q^2-q+1$, then $i_B^G(\nu^{-2})$ has length three with unique cuspidal subquotient. The unique irreducible subrepresentation is not isomorphic to the unique irreducible quotient.*

(4) *If $\ell \neq 2$ and $\ell \mid q+1$, or if $\ell = 2$ and $4 \mid q+1$, then $i_B^G(\nu^{-2})$ has length six with 1_G appearing as the unique subrepresentation and the unique quotient, and four cuspidal subquotients, one of which appears with multiplicity two. A maximal cuspidal subquotient of $i_B^G(\nu^{-2})$ is not semisimple.*

(5) *If $\ell = 2$ and $4 \mid q-1$, then $i_B^G(\nu^{-2})$ has length five with 1_G appearing as the unique subrepresentation and the unique quotient. All cuspidal subquotients of $i_B^G(\nu^{-2})$ are semisimple and the irreducible cuspidal subquotients are pairwise nonisomorphic.*

2. Notation

2A. ***Unramified unitary groups.*** Let F_0 be a nonarchimedean local field of odd residual characteristic p. Let F be an unramified quadratic extension of F_0 and $\overline{}$ a generator of $\mathrm{Gal}(F/F_0)$. If D is a nonarchimedean local field, we let $\mathfrak{o}_D$ denote the ring of integers of D, $\mathfrak{p}_D$ denote the unique maximal ideal of $\mathfrak{o}_D$, and $k_D = \mathfrak{o}_D/\mathfrak{p}_D$ denote the residue field. We let $\mathfrak{o}_0 = \mathfrak{o}_{F_0}$, $\mathfrak{p}_0 = \mathfrak{p}_{F_0}$, $k_0 = k_{F_0}$, and $q = q_0 = |k_{F_0}|$. We fix a choice of uniformiser ϖ_F of F_0.

Let V be a finite-dimensional F-vector space and $h : V \times V \to F$ a hermitian form on V, that is, a nondegenerate form which is sesquilinear (linear in the first variable and $\overline{}$-linear in the second variable) and such that $h(v_1, v_2) = \overline{h(v_2, v_1)}$ for all v_1, $v_2 \in V$. The *unitary group* $\mathrm{U}(V,h)$ is the subgroup of isometries of $\mathrm{GL}(V)$, i.e., $\mathrm{U}(V,h) = \{g \in \mathrm{GL}(V) : h(gv_1, gv_2) = h(v_1, v_2),\ v_1, v_2 \in V\}$. The form h induces an anti-involution on $\mathrm{End}_F(V)$ which we denote by $\overline{}$. Let σ denote the involution $g \mapsto \bar{g}^{-1}$ for $g \in \mathrm{GL}(V)$. We also let σ act on $\mathrm{End}_F(V)$ by $a \mapsto -\bar{a}$ for $a \in \mathrm{End}_F(V)$.

2B. ***Parahoric subgroups.*** An $\mathfrak{o}_F$*-lattice* in V is a compact open $\mathfrak{o}_F$-submodule of V. Let L be an $\mathfrak{o}_F$-lattice in V and let $\operatorname{Lat} V$ denote the set of all $\mathfrak{o}_F$-lattices in V. The $\mathfrak{o}_F$-lattice $L^\sharp = \{v \in V : h(v, L) \subseteq \mathfrak{p}_F\}$, defined relative to h, is called the *dual lattice* of L. Let $A = \operatorname{End}_F(V)$ and $\mathfrak{g} = \{X \in A : X + X^\sigma = 0\}$. An $\mathfrak{o}_F$*-lattice sequence* is a function $\Lambda : \mathbb{Z} \to \operatorname{Lat} V$ which is decreasing and periodic. Let Λ be an $\mathfrak{o}_F$-lattice sequence. The *dual* $\mathfrak{o}_F$-lattice sequence $\Lambda^\sharp$ of Λ is the $\mathfrak{o}_F$-lattice sequence defined by $\Lambda^\sharp(n) = (\Lambda(-n))^\sharp$ for all $n \in \mathbb{Z}$. We call Λ *self-dual* if there exists $k \in \mathbb{Z}$ such that $\Lambda(n) = \Lambda^\sharp(n+k)$ for all $n \in \mathbb{Z}$. If Λ is self-dual then we can always consider a translate Λ_k of Λ such that either $\Lambda_k(0) = \Lambda_k^\sharp(0)$ or $\Lambda_k(1) = \Lambda_k^\sharp(0)$.

Let Λ be an $\mathfrak{o}_F$-lattice sequence in V. For $n \in \mathbb{Z}$ define

$$\mathfrak{P}_n(\Lambda) = \{x \in A : x\Lambda(m) \subset \Lambda(m+n) \text{ for all } m \in \mathbb{Z}\},$$

which is an $\mathfrak{o}_F$-lattice in A. We let $\mathfrak{P}_n^-(\Lambda) = \mathfrak{P}_n(\Lambda) \cap \mathfrak{g}$.

If Λ is self-dual then the groups $\mathfrak{P}_n(\Lambda)$ are stable under the involution which h induces on A. In this case, define compact open subgroups of G, called *parahoric subgroups*, by

$$\begin{aligned} \mathrm{P}(\Lambda) &= \mathfrak{P}_0(\Lambda)^\times \cap G, \\ \mathrm{P}_m(\Lambda) &= (1 + \mathfrak{P}_m(\Lambda)) \cap G, \quad m \in \mathbb{N}. \end{aligned}$$

The pro-unipotent radical of $\mathrm{P}(\Lambda)$ is isomorphic to $\mathrm{P}_1(\Lambda)$. The sequence $(\mathrm{P}_m(\Lambda))_{m\in\mathbb{N}}$ is a fundamental system of neighbourhoods of the identity in G and forms a decreasing filtration of $\mathrm{P}(\Lambda)$ by normal compact open subgroups. The quotient $\mathrm{M}(\Lambda) = \mathrm{P}(\Lambda)/\mathrm{P}_1(\Lambda)$ is the k_0-points of a connected reductive group defined over k_0.

Let $\mathrm{P}_1 = \mathrm{P}(\Lambda_1)$ and $\mathrm{P}_2 = \mathrm{P}(\Lambda_2)$ be parahoric subgroups of G. Fix a set of *distinguished* double coset representatives $D_{2,1}$ for $\mathrm{P}_2\backslash G/\mathrm{P}_1$, as in [Morris 1993, §3.10]. Let $n \in D_{2,1}$; then

$$P_{\Lambda_1, n\Lambda_2} = \mathrm{P}_1^1(\mathrm{P}_1 \cap \mathrm{P}_2^n)/\mathrm{P}_1^1$$

is a parabolic subgroup of $\mathrm{M}_1 = \mathrm{P}_1/\mathrm{P}_1^1$, by [Morris 1993, Corollary 3.20]. Furthermore, the pro-p unipotent radical of $\mathrm{P}_1^1(\mathrm{P}_1 \cap \mathrm{P}_2^n)$ is $\mathrm{P}_1^1(\mathrm{P}_1 \cap (\mathrm{P}_2^n)^1)$, by [Morris 1993, Lemma 3.21]. If $D_{2,1}$ is a set of distinguished double coset representatives for $\mathrm{P}_2\backslash G/\mathrm{P}_1$, then $D_{2,1}^{-1}$ is a set of distinguished double coset representatives for $\mathrm{P}_1\backslash G/\mathrm{P}_2$. Hence

$$P_{\Lambda_2, n^{-1}\Lambda_1} = \mathrm{P}_2^1(\mathrm{P}_2 \cap {}^n\mathrm{P}_1)/\mathrm{P}_2^1$$

is a parabolic subgroup of $\mathrm{M}_2 = \mathrm{P}_2/\mathrm{P}_2^1$. Furthermore, the pro-$p$ unipotent radical of $\mathrm{P}_2^1(\mathrm{P}_2 \cap {}^n\mathrm{P}_1)$ is $\mathrm{P}_2^1(\mathrm{P}_2 \cap {}^n\mathrm{P}_1^1)$.

2C. $\mathrm{U}(2,1)(F/F_0)$. Let $x_i \in F$ for $i = 1, 2, \dots, n$. Denote by $\mathrm{diag}(x_1, \dots, x_n)$ the n-by-n diagonal matrix with entries x_i on the diagonal and by $\mathrm{adiag}(x_1, \dots, x_n)$ the n-by-n matrix $(a_{i,j})$ such that $a_{m,n+1-m} = x_{n+1-m}$ and all other entries are zero.

Let V be a three-dimensional F-vector space with standard basis $\{e_{-1}, e_0, e_1\}$ and $h : V \times V \to F$ be the nondegenerate hermitian form on V defined by, for $v, w \in V$,

$$h(v, w) = v_{-1}\overline{w_1} + v_0\overline{w_0} + v_1\overline{w_{-1}}$$

if $v = (v_{-1}, v_0, v_1)$ and $w = (w_{-1}, w_0, w_1)$ with respect to the standard basis $\{e_{-1}, e_0, e_1\}$. Let $\mathrm{U}(2, 1)(F/F_0)$ denote the unitary group attached to the hermitian space (V, h), i.e.,

$$\mathrm{U}(2, 1)(F/F_0) = \{g \in \mathrm{GL}_3(F) : gJ\bar{g}^T J = 1\},$$

where $J = \mathrm{adiag}(1, 1, 1)$ is the matrix of the form h. We let $\mathrm{U}(1, 1)(F/F_0)$ and $\mathrm{U}(2)(F/F_0)$ denote the two-dimensional unitary groups defined by the forms whose associated matrices are $\mathrm{adiag}(1, 1)$ and $\mathrm{diag}(1, \varpi_F)$ respectively. Let

$$\mathrm{U}(1)(F/F_0) = \{g \in F^\times : g\bar{g} = 1\}$$

and occasionally, for brevity, let $F^1 = \mathrm{U}(1)(F/F_0)$. We use analogous notation for unitary groups defined over extensions of F_0 and defined over finite fields.

Let B be the standard Borel subgroup of $\mathrm{U}(2, 1)(F/F_0)$ with Levi decomposition $B = T \ltimes N$, where $T = \{\mathrm{diag}(x, y, \bar{x}^{-1}) : x \in F^\times, y \in F^1\}$ and

$$N = \left\{ \begin{pmatrix} 1 & x & y \\ 0 & 1 & \bar{x} \\ 0 & 0 & 1 \end{pmatrix} : x, y \in F, y + \bar{y} = x\bar{x} \right\}.$$

The maximal F_0-split torus contained in T is $T_0 = \{\mathrm{diag}(x, 1, x^{-1}) : x \in F_0^\times\}$. The subgroup of T generated by its compact subgroups is

$$T^0 = \{\mathrm{diag}(x, y, \bar{x}^{-1}) : x \in \mathfrak{o}_F^\times, y \in F^1\}.$$

Let $T^1 = T^0 \cap \mathrm{diag}(1 + \mathfrak{p}_F, 1 + \mathfrak{p}_F, 1 + \mathfrak{p}_F)$.

Let Λ_I be the $\mathfrak{o}_F$-lattice sequence of period three given by $\Lambda_I(0) = \mathfrak{o}_F \oplus \mathfrak{o}_F \oplus \mathfrak{o}_F$, $\Lambda_I(1) = \mathfrak{o}_F \oplus \mathfrak{o}_F \oplus \mathfrak{p}_F$ and $\Lambda_I(2) = \mathfrak{o}_F \oplus \mathfrak{p}_F \oplus \mathfrak{p}_F$ with respect to the standard basis. The (standard) Iwahori subgroup of G is the parahoric subgroup

$$\mathrm{P}(\Lambda_I) = \begin{pmatrix} \mathfrak{o}_F & \mathfrak{o}_F & \mathfrak{o}_F \\ \mathfrak{p}_F & \mathfrak{o}_F & \mathfrak{o}_F \\ \mathfrak{p}_F & \mathfrak{p}_F & \mathfrak{p}_F \end{pmatrix} \cap G.$$

There are two parahoric subgroups of G which contain $\mathrm{P}(\Lambda_I)$, both of which are maximal. These correspond to the lattice sequences Λ_x of period one and Λ_y of period two with $\Lambda_x(0) = \mathfrak{o}_F \oplus \mathfrak{o}_F \oplus \mathfrak{o}_F$, $\Lambda_y(0) = \mathfrak{o}_F \oplus \mathfrak{o}_F \oplus \mathfrak{p}_F$ and $\Lambda_y(1) = \mathfrak{o}_F \oplus \mathfrak{p}_F \oplus \mathfrak{p}_F$.

Note that we have $\mathrm{M}(\Lambda_x) \simeq \mathrm{U}(2,1)(k_F/k_0)$, $\mathrm{M}(\Lambda_y) \simeq \mathrm{U}(1,1)(k_F/k_0) \times k_F^1$ and $\mathrm{M}(\Lambda_I) \simeq k_F^\times \times k_F^1$. Furthermore, $\mathrm{M}(\Lambda_I)$ is a maximal torus in $\mathrm{M}(\Lambda_x)$ and $\mathrm{P}(\Lambda_I)$ is equal to the preimage in $\mathrm{P}(\Lambda_x)$ of a Borel subgroup B_x, which we call *standard*, under the projection map $\mathrm{P}(\Lambda_x) \to \mathrm{M}(\Lambda_x)$; the same holds with y in place of x throughout.

The *affine Weyl group* $\widetilde{W} = N_G(T)/T^0$ of $\mathrm{U}(2,1)(F/F_0)$ is an infinite dihedral group generated by the cosets represented by the elements $w_x = \mathrm{adiag}(1,1,1)$ and $w_y = \mathrm{adiag}(\varpi_F, 1, \varpi_F^{-1})$. Furthermore, we have $\mathrm{P}(\Lambda_x) = \mathrm{P}(\Lambda_I) \cup \mathrm{P}(\Lambda_I) w_x \, \mathrm{P}(\Lambda_I)$ and $\mathrm{P}(\Lambda_y) = \mathrm{P}(\Lambda_I) \cup \mathrm{P}(\Lambda_I) w_y \, \mathrm{P}(\Lambda_I)$.

2D. *Reduction modulo ℓ.* Let $\overline{\mathbb{Q}}_\ell$ be an algebraic closure of the ℓ-adic numbers, $\overline{\mathbb{Z}}_\ell$ be the ring of integers of $\overline{\mathbb{Q}}_\ell$, Γ be the unique maximal ideal of $\overline{\mathbb{Z}}_\ell$, and $\overline{\mathbb{F}}_\ell = \overline{\mathbb{Z}}_\ell/\Gamma$ be the residue field of $\overline{\mathbb{Q}}_\ell$, which is an algebraic closure of the finite field with ℓ elements. Let $\mathfrak{Gr}_R(G)$ denote the *Grothendieck group* of R-representations, i.e., the free abelian group with $\mathbb{Z}$-basis $\mathrm{Irr}_R(G)$. A representation in $\mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(G)$ will be called *ℓ-adic* and a representation in $\mathfrak{R}_{\overline{\mathbb{F}}_\ell}(G)$ will be called *ℓ-modular*. We say ℓ is *banal* for G if it does not divide the pro-order of any compact open subgroup of G.

Let $(\pi, \mathcal{V})$ be a finite-length ℓ-adic representation of G. We call π *integral* if π stabilises a $\overline{\mathbb{Z}}_\ell$-lattice $\mathcal{L}$ in $\mathcal{V}$. In this case π stabilises $\Gamma\mathcal{L}$ and π induces a finite-length ℓ-modular representation on the space $\mathcal{L}/\Gamma\mathcal{L}$. In general, this depends on the choice of the lattice $\mathcal{L}$. However, due to [Vignéras 2004, Theorem 1], the semisimplification of $\mathcal{L}/\Gamma\mathcal{L}$ is independent of the lattice chosen and we define $r_\ell(\pi)$, the *reduction modulo ℓ* of π, to be this semisimple ℓ-modular representation. If π is a finite-length R-representation of G we write $[\pi]$ for the semisimplification of π in $\mathfrak{Gr}_R(G)$.

We fix choices of square roots of p in $\overline{\mathbb{Q}}_\ell^\times$ and $\overline{\mathbb{F}}_\ell^\times$ such that our chosen square root of p in $\overline{\mathbb{F}}_\ell^\times$ is the reduction modulo ℓ of our chosen square root of p in $\overline{\mathbb{Q}}_\ell^\times$, and make use of these choices in our definitions of normalised parabolic induction and the Jacquet module.

Parabolic induction preserves integrality and commutes with reduction modulo ℓ: if $P = M \ltimes N$ is a parabolic subgroup of G and σ is a finite-length integral ℓ-adic representation of M, then $r_\ell(i_P^G(\sigma)) \simeq \left[i_P^G(r_\ell(\sigma))\right]$. Furthermore, compact induction commutes with reduction modulo ℓ: if H is a closed subgroup of G and σ an integral finite-length representation of H such that $\mathrm{ind}_H^G(\sigma)$ is of finite length, then $r_\ell(\mathrm{ind}_H^G(\sigma)) = [\mathrm{ind}_H^G(r_\ell(\sigma))]$. For classical groups, due to [Dat 2005], the Jacquet module preserves integrality and commutes with reduction modulo ℓ: if $P = M \ltimes N$ is a parabolic subgroup of G and π is a finite-length integral ℓ-adic representation of G, then $r_\ell(r_P^G(\pi)) \simeq \left[r_P^G(r_\ell(\pi))\right]$. This implies that the reduction modulo ℓ of a finite-length integral cuspidal ℓ-adic representation is cuspidal.

An irreducible R-representation is admissible, due to [Vignéras 1996, II 2.8]. If π is an R-representation, we let $\tilde{\pi}$ or $\pi^\sim$ denote the contragredient representation of π.

The abelian category $\mathfrak{R}_R(G)$ has a decomposition as a direct product of full subcategories $\mathfrak{R}_R^x(G)$, consisting of all representations all of whose irreducible subquotients have level x for $x \in \mathbb{Q}_{\geqslant 0}$, which is preserved by parabolic induction and the Jacquet functor, by [Vignéras 1996, II 5.8 and 5.12].

3. Cuspidal representations of $\mathrm{U}(1, 1)(k_F/k_0)$ and $\mathrm{U}(2, 1)(k_F/k_0)$

Our description of the supercuspidal ℓ-adic representations of $\mathrm{U}(1, 1)(k_F/k_0)$ and $\mathrm{U}(2, 1)(k_F/k_0)$ and the decomposition of the ℓ-adic principal series follow from similar arguments made for $\mathrm{GL}_2(k_F)$ and $\mathrm{SL}_2(k_F)$ by Digne and Michel [1991, §15.9]. The character tables of both groups were first computed by Ennola [1963] and the ℓ-modular representations of $\mathrm{U}(2, 1)(k_F/k_0)$ were first studied by Geck [1990]. In this section, let $H = \mathrm{U}(1, 1)(k_F/k_0)$ and $G = \mathrm{U}(2, 1)(k_F/k_0)$. We can realise H and G as the fixed points of $\mathrm{GL}_2(\bar{k})$ and $\mathrm{GL}_3(\bar{k})$ under twisted Frobenius morphisms $\widetilde{F} : (a_{ij}) \mapsto (a_{ji}^q)^{-1}$, where $\bar{k}$ is an algebraic closure of k_0 containing k_F. A torus T of $\mathrm{GL}_2(\bar{k})$ (resp. $\mathrm{GL}_3(\bar{k})$) is called *minisotropic* if it is stable under the twisted Frobenius morphism $\widetilde{F}$ and is not contained in any $\widetilde{F}$-stable parabolic subgroup of $\mathrm{GL}_2(\bar{k})$ (resp. $\mathrm{GL}_3(\bar{k})$). We call a torus in H or G *minisotropic* if it is equal to the $\widetilde{F}$-fixed points of a minisotropic torus of the corresponding algebraic group.

3A. *Cuspidals of* $\mathrm{U}(1, 1)(k_F/k_0)$.

3A1. *Cuspidals.* There are $\frac{1}{2}(q^2 + q)$ irreducible ℓ-adic supercuspidal representations of H. These can be parametrised by the regular irreducible characters of the minisotropic tori of H. There is only one conjugacy class of minisotropic tori in G, which is isomorphic to $k_F^1 \times k_F^1$; hence a character of this torus corresponds to two characters of k_F^1. Furthermore, this character is regular if and only if it corresponds to two distinct characters of k_F^1. Thus the ℓ-adic supercuspidals can be parametrised by unordered pairs of distinct irreducible characters of k_F^1. Let χ_1, χ_2 be distinct ℓ-adic characters of k_F^1. Let $\sigma(\chi_1, \chi_2)$ denote the ℓ-adic supercuspidal representation parametrised by the set $\{\chi_1, \chi_2\}$.

Using Clifford Theory, the decomposition numbers for H follow from the well-known decomposition numbers of $\mathrm{SU}(1, 1)(k_F/k_0) \simeq \mathrm{SL}_2(k_0)$. We have $|H| = q(q-1)(q+1)$; hence, because q is odd, there are four cases to consider: $\ell \mid q-1$, $\ell \mid q+1$, $\ell = 2$, and ℓ is prime to (q^2-1).

All irreducible ℓ-modular cuspidal representations of H are isomorphic to the reduction modulo ℓ of an irreducible ℓ-adic supercuspidal representation. If χ is an ℓ-adic character we let $\overline{\chi}$ denote its reduction modulo ℓ. If χ_1', χ_2' are ℓ-adic characters of k_F^1, we have $r_\ell(\sigma(\chi_1, \chi_2)) = r_\ell(\sigma(\chi_1', \chi_2'))$ if and only if $\{\overline{\chi}_1, \overline{\chi}_2\} = \{\overline{\chi}_1', \overline{\chi}_2'\}$. We let $\overline{\sigma}(\overline{\chi}_1, \overline{\chi}_2) = r_\ell(\sigma(\chi_1, \chi_2))$. Furthermore, $\overline{\sigma}(\overline{\chi}_1, \overline{\chi}_2)$ is supercuspidal if and only if $|\{\overline{\chi}_1, \overline{\chi}_2\}| = 2$ and we have $\overline{\sigma}(\overline{\chi}_1, \overline{\chi}_2) = \overline{\sigma}(\overline{\chi}_2, \overline{\chi}_1)$. Hence the irreducible cuspidal nonsupercuspidal ℓ-modular representations of H are parametrised by the

ℓ-modular characters of k_F^1 and, if $\overline{\chi}$ is an ℓ-modular character of k_F^1 equal to the reduction modulo ℓ of two distinct ℓ-adic characters of k_F^1, we let $\overline{\sigma}(\overline{\chi})=\overline{\sigma}(\overline{\chi},\overline{\chi})$. When $\ell\nmid q+1$, all irreducible cuspidal ℓ-modular representations are supercuspidal.

3A2. *Cuspidal nonsupercuspidals when $\ell\mid q+1$.* Let $\ell^a\|q+1$, so that there are $(q+1)/\ell^a$ cuspidal nonsupercuspidal ℓ-modular representations denoted by $\overline{\sigma}(\overline{\chi})$; these occur as the reduction modulo ℓ of $\sigma(\chi_1,\chi_2)$ when $\overline{\chi}=\overline{\chi}_1=\overline{\chi}_2$. Let $T=\{\mathrm{diag}(x,\bar{x}^{-1}):x\in k_F^\times\}$ be the maximal diagonal torus of H and B_H be the standard Borel subgroup containing T. The principal series representations $i_{B_H}^H(\overline{\chi}\circ\xi)\simeq i_{B_H}^H(\overline{1})(\overline{\chi}\circ\det)$ are uniserial of length three with $(\overline{\chi}\circ\det)$ appearing as the unique irreducible subrepresentation and the unique irreducible quotient, and unique irreducible cuspidal subquotient $\overline{\sigma}(\overline{\chi})$.

3B. *Cuspidals of* U(2, 1)(k_F/k_0).

3B1. *ℓ-adic supercuspidals.* There are two conjugacy classes of minisotropic tori in G, which give rise to two classes of irreducible supercuspidal ℓ-adic representations coming from regular irreducible characters of these tori. Let E be an unramified cubic extension of F. One conjugacy class of the minisotropic tori has representatives isomorphic to $k_F^1\times k_F^1\times k_F^1$; the other conjugacy class has representatives isomorphic to k_E^1. However, in contrast to H, the irreducible representations parametrised by the irreducible regular characters of these tori do not constitute all the irreducible supercuspidal representations of G: additionally there exist unipotent supercuspidal representations of G. Thus we have three classes of ℓ-adic supercuspidals:

(1) There are $\frac{1}{6}(q+1)q(q-1)$ ℓ-adic supercuspidals of dimension $(q-1)(q^2-q+1)$ parametrised by the irreducible regular characters of $k_F^1\times k_F^1\times k_F^1$. An irreducible ℓ-adic character of $k_F^1\times k_F^1\times k_F^1$ is of the form $\chi_1\otimes\chi_2\otimes\chi_3$, with χ_1,χ_2,χ_3 irreducible ℓ-adic characters of k_F^1, and is regular if and only if $|\{\chi_1,\chi_2,\chi_3\}|=3$. We let $\sigma(\chi_1,\chi_2,\chi_3)$ denote the ℓ-adic supercuspidal corresponding to the set $\{\chi_1,\chi_2,\chi_3\}$.

(2) There are $\frac{1}{3}(q+1)q(q-1)$ ℓ-adic supercuspidals of dimension $(q-1)(q+1)^2$ parametrised by the irreducible regular characters of k_E^1. An irreducible ℓ-adic character ψ of k_E^1 is regular if and only if $\psi^{q+1}\neq 1$. We let $\tau(\psi)$ denote the ℓ-adic supercuspidal representation corresponding to ψ.

(3) There are $(q+1)$ unipotent ℓ-adic supercuspidals of dimension $q(q-1)$. These can be parametrised by the irreducible characters of k_F^1. We write $\nu(\chi)$ for the unipotent ℓ-adic supercuspidal representation corresponding to the irreducible ℓ-adic character χ of k_F^1.

3B2. *ℓ-modular cuspidals.* We have $|G|=q^3(q-1)(q+1)^3(q^2-q+1)$; hence there are six cases to consider: $\ell=2$, $\ell=3$ and $\ell\mid q+1$, $\ell\mid q-1$, $\ell\mid q+1$, $\ell\mid q^2-q+1$, and ℓ is prime to $(q-1)(q+1)(q^2-q+1)$. When $\ell \neq 2$, the decomposition numbers can be obtained from [Geck 1990] and [Okuyama and Waki 2002] using Clifford theory. Parabolic induction of the trivial character is completely described in [Hiss 2004, Theorem 4.1]. When $\ell \mid q-1$ or $\ell \mid q+1$, all irreducible cuspidal ℓ-modular representations lift to irreducible cuspidal ℓ-adic representations. Analogously to the two-dimensional case, we write $\overline{\nu}(\overline{\chi}) = r_\ell(\nu(\chi))$, $\overline{\tau}(\overline{\psi}) = r_\ell(\tau(\psi))$ and $\overline{\sigma}(\overline{\chi}_1, \overline{\chi}_2, \overline{\chi}_3) = r_\ell(\sigma(\chi_1, \chi_2, \chi_3))$.

When $\ell \neq 3$ and $\ell \mid q^2-q+1$, we have irreducible ℓ-modular cuspidal representations which do not lift: if ψ is an ℓ-adic character of k_E^1 such that $\psi^{q+1} \neq 1$ but $\overline{\psi}^{q+1} = \overline{1}$, then $r_\ell(\tau(\psi)) = \overline{\nu}(\overline{\chi}) \oplus \overline{\tau}^+(\overline{\chi})$, where $\overline{\chi}$ is the character of k_F^1 such that $\overline{\psi} = \overline{\chi} \circ \xi$, where $\xi(x) = x^{q-1}$, and $\overline{\tau}^+(\overline{\chi})$ does not lift. When $\ell = 2$ and $4 \mid q-1$, we also have cuspidal representations which do not lift: if ψ is an ℓ-adic character of k_E^1 such that $\psi^{q+1} \neq 1$ but $\overline{\psi}^{q+1} = \overline{1}$, then $r_\ell(\tau(\psi)) = \overline{\nu}(\overline{\chi}) \oplus \overline{\nu}(\overline{\chi}) \oplus \overline{\tau}^+(\overline{\chi})$, where $\overline{\chi}$ is the character of k_F^1 such that $\overline{\psi} = \overline{\chi} \circ \xi$, where $\xi(x) = x^{q-1}$, and $\overline{\tau}^+(\overline{\chi})$ does not lift. All other irreducible cuspidal ℓ-modular representations of G lift to ℓ-adic representations and we use the same notation as before.

3B3. *ℓ-adic principal series.* Let $T = \{\mathrm{diag}(x, y, \overline{x}^{-1}) : x \in k_F^\times,\ y \in k_F^1\}$ be the maximal diagonal torus in G and B be the standard Borel subgroup of G containing T.

Let χ_1 be an ℓ-adic character of $k_F^\times$ and χ_2 an ℓ-adic character of k_F^1. Let χ be the irreducible character of T defined by $\chi(\mathrm{diag}(x, y, x^{-q})) = \chi_1(x)\chi_2(xyx^{-q})$. The character χ is regular if and only if $\chi_1^{q+1} \neq 1$, and in this case the principal series representation $i_B^G(\chi)$ is irreducible.

If $\chi_1^{q+1} = 1$ then $\chi_1 = \chi_1' \circ \xi$, where $\xi(x) = x^{q-1}$ and χ_1' is an ℓ-adic character of k_F^1. If $\chi_1' = 1$, or equivalently $\chi_1 = 1$, then

$$i_B^G(\chi) = 1_G(\chi_2 \circ \det) \oplus \mathrm{St}_G(\chi_2 \circ \det),$$

where St_G is an irreducible q^3-dimensional representation of G. If $\chi_1' \neq 1$ then

$$i_B^G(\chi) = R_{1_H(\chi_1')}(\chi_2 \circ \det) \oplus R_{\mathrm{St}_H(\chi_1')}(\chi_2 \circ \det),$$

where $R_{1_H(\chi_1')}$ and $R_{\mathrm{St}_H(\chi_1')}$ are irreducible representations of G of dimensions q^2-q+1 and $q(q^2-q+1)$ respectively. The reducibility here comes from inducing first to the Levi subgroup $L^* = \mathrm{U}(1,1)(k_F/k_0) \times \mathrm{U}(1)(k_F/k_0)$, which is not contained in any proper rational parabolic subgroup of G. Here 1_H and St_H denote the trivial and Steinberg representations of $\mathrm{U}(1,1)(k_F/k_0)$, and R is a generalised induction from L^* to G.

3B4. *Cuspidal subquotients of ℓ-modular principal series.* If $\ell \neq 2$ and $\ell \mid q-1$, or ℓ is prime to $(q-1)(q+1)(q^2-q+1)$, then all irreducible cuspidal ℓ-modular representations are supercuspidal and the principal series representations are all semisimple.

Let $\overline{\chi}_2$ be an ℓ-modular character of k_F^1. We first describe the ℓ-modular principal series representations $i_B^G(\bar{1})(\overline{\chi}_2 \circ \det)$ in all the cases where cuspidal subquotients appear.

(1) If $\ell \neq 3$ and $\ell \mid q^2 - q + 1$, $i_B^G(\bar{1})(\overline{\chi}_2 \circ \det)$ are uniserial of length three with $(\overline{\chi} \circ \det)$ appearing as the unique irreducible subrepresentation and the unique irreducible quotient and $\overline{\tau}^+(\overline{\chi})$ as the unique irreducible cuspidal subquotient.

(2) If $\ell \neq 2$ and $\ell \mid q+1$, or $\ell = 2$ and $4 \mid q+1$, then $i_B^G(\bar{1})(\overline{\chi} \circ \det)$ have irreducible cuspidal subquotients $\overline{\nu}(\overline{\chi})$ and $\overline{\sigma}(\overline{\chi}) = \overline{\sigma}(\overline{\chi}, \overline{\chi}, \overline{\chi})$. The principal series representations $i_B^G(\bar{1})(\overline{\chi} \circ \det)$ are uniserial of length five with $(\overline{\chi} \circ \det)$ appearing as the unique irreducible subrepresentation and the unique irreducible quotient. A maximal cuspidal subquotient of $i_B^G(\bar{1})(\overline{\chi} \circ \det)$ is uniserial of length three with $\overline{\nu}(\overline{\chi})$ appearing as the unique irreducible quotient and the unique irreducible subrepresentation, and remaining subquotient $\overline{\sigma}(\overline{\chi})$.

(3) If $\ell = 2$ and $4 \mid q-1$ then $i_B^G(\bar{1})(\overline{\chi} \circ \det)$ has length four with $(\overline{\chi} \circ \det)$ appearing as the unique irreducible subrepresentation and the unique irreducible quotient, and cuspidal subquotient $\overline{\nu}(\overline{\chi}) \oplus \overline{\tau}^+(\overline{\chi})$.

Now let $\overline{\chi}_1'$ and $\overline{\chi}_2$ be ℓ-modular characters of k_F^1 with $\overline{\chi}_1'$ nontrivial and let $\overline{\chi}_1 = \overline{\chi}_1' \circ \xi$. Let $\overline{\chi}$ be the ℓ-modular character of T defined by

$$\overline{\chi}(\mathrm{diag}(x, y, x^{-q})) = \overline{\chi}_1(x)\overline{\chi}_2(xyx^{-q}).$$

If $\ell \nmid q+1$ then $i_B^G(\overline{\chi})$ does not possess any cuspidal subquotients. If $\ell \mid q+1$ then $i_B^G(\overline{\chi})$ is uniserial of length three with $\overline{R}_{\bar{1}_H(\overline{\chi}_1')}(\overline{\chi}_2 \circ \det)$ appearing as the unique irreducible subrepresentation and the unique irreducible quotient and cuspidal subquotient $\sigma(\overline{\chi}_1', \overline{\chi}_1', \overline{\chi}_2)$. This follows from [Bonnafé and Rouquier 2003, Theorem 11.8] and the principal block of H as χ corresponds to a semisimple element with centraliser $H \times k_F^1$ in the dual group.

4. Irreducible cuspidal R-representations of $\mathrm{U}(2,1)(F/F_0)$

Let $G = \mathrm{U}(2,1)(F/F_0)$. We construct all irreducible cuspidal representations of G by compact induction from certain irreducible representations of compact open subgroups. We review some general theory first and recall results of Vignéras on level zero representations. Our construction of all irreducible cuspidal representations of G then follows the outline of Stevens' construction [2008] of all irreducible cuspidal representations of classical p-adic groups in the complex case. While his construction is carried out when $R = \mathbb{C}$ the first part remains equally valid when R is any algebraically closed field of characteristic unequal to p, essentially as all groups involved are pro-p . However, when we move to defining β-extensions and beyond the subgroups we are dealing with no longer have pro-order necessarily invertible in $\overline{\mathbb{F}}_\ell$. It is here, and after, where we need to be careful and have to make nontrivial changes to the proofs of the statements of [Stevens 2008]. It turns out that, even though we have to change the proofs, the definitions and properties of β-extensions in the ℓ-modular case are completely analogous to those of complex β-extensions. We note that as we are in the special case of unramified $\mathrm{U}(2,1)(F/F_0)$, using the framework of Stevens, we can show that our β-extensions satisfy closer compatibility properties than are available in the general case of classical groups.

4A. ***Types and Hecke algebras.*** By an *R-type*, we mean a pair (K,σ) consisting of a compact open subgroup K of G and an irreducible R-representation σ of K. Given an R-type we consider the compactly induced representation $\mathrm{ind}_K^G(\sigma)$ of G, the goal being to find pairs (K,σ) such that $\mathrm{ind}_K^G(\sigma)$ is irreducible and cuspidal. Let $\pi\in\mathrm{Irr}_R(G)$; we say that π *contains* the R-type (K,σ) if π is a quotient of $\mathrm{ind}_K^G(\sigma)$.

Let (K,σ) be an R-type in G and $\mathcal{W}$ be the space of σ. The *spherical Hecke algebra* $\mathcal{H}(G,\sigma)$ of σ is the R-module consisting of the set of all functions $f: G\to\mathrm{End}_R(\mathcal{W})$ such that the support of f is a finite union of double cosets in $K\backslash G/K$ and f transforms by σ on the left and the right, i.e., for all $k_1,k_2\in K$ and all $g\in G$, $f(k_1gk_2)=\sigma(k_1)f(g)\sigma(k_2)$. The product in $\mathcal{H}(G,\sigma)$ is given by convolution: if $f_1, f_2\in\mathcal{H}(G,\sigma)$ then

$$f_1\star f_2(h)=\sum_{G/K} f_1(g)f_2(g^{-1}h).$$

The spherical Hecke algebra $\mathcal{H}(G,\sigma)$ is isomorphic to $\mathrm{End}_G(\mathrm{ind}_K^G(\sigma))$, where multiplication in $\mathrm{End}_G(\mathrm{ind}_K^G(\sigma))$ is defined by composition. For $g\in G$, let $I_g(\sigma)=\mathrm{Hom}_K(\sigma,\mathrm{ind}_{K\cap K^g}^K\sigma^g)$ and let $I_G(\sigma)=\{g\in G: I_g(\sigma)\neq 0\}$.

Let $\mathcal{M}(G,\sigma)$ denote the category of right $\mathcal{H}(G,\sigma)$-modules. Define

$$M_\sigma:\mathfrak{R}_R(G)\to\mathcal{M}(G,\sigma)$$

by $\pi\mapsto\mathrm{Hom}_G(\mathrm{ind}_K^G(\sigma),\pi)$; this is a (right) $\mathrm{End}_G(\mathrm{ind}_K^G(\sigma))$-module by precomposition. In the ℓ-adic case, if (K,σ) is a type in the sense of [Bushnell and Kutzko 1998, p. 584], M_σ induces an equivalence of categories between $\mathcal{M}(G,\sigma)$ and the full subcategory of $\mathfrak{R}_R(G)$ of representations all of whose irreducible subquotients contain (K,σ).

An R-representation $(\pi,\mathcal{V})$ of G is *quasiprojective* if, for all R-representations $(\sigma,\mathcal{W})$ of G, all surjective $\Phi\in\mathrm{Hom}_G(\mathcal{V},\mathcal{W})$ and all $\Psi\in\mathrm{Hom}_G(\mathcal{V},\mathcal{W})$, there exists $\Xi\in\mathrm{End}_G(\mathcal{V})$ such that $\Psi=\Phi\circ\Xi$.

Theorem 4.1 [Vignéras 1998, Appendix, Theorem 10]. *Let π be a quasiprojective, finitely generated R-representation of G. The map $\rho\mapsto\mathrm{Hom}_G(\pi,\rho)$ induces a bijection between the irreducible quotients of π and the simple right* $\mathrm{End}_G(\pi)$*-modules.*

Let P be a parabolic subgroup of G with Levi decomposition $P = M \ltimes N$. Let P^{op} be the opposite parabolic subgroup of P with Levi decomposition $P^{\mathrm{op}} = M \ltimes N^{\mathrm{op}}$. Let $K^+ = K \cap N$ and $K^- = K \cap N^{\mathrm{op}}$. An element z of the centre of M is called *strongly* (P, K)*-positive* if:

(1) $zK^+z^{-1} \subset K^+$ and $zK^-z^{-1} \supset K^-$.

(2) For all compact subgroups H_1, H_2 of N (resp. N^{op}), there exists a positive (resp. negative) integer m such that $z^m H_1 z^{-m} \subset H_2$.

Let (K_M, σ_M) be an R-type of M. An R-type (K, σ) is called a *G-cover* of (K_M, σ_M) relative to P if we have:

(1) $K \cap M = K_M$ and we have an Iwahori decomposition $K = K^- K_M K^+$.

(2) $\mathrm{Res}^K_{K_M}(\sigma) = \sigma_M$, $\mathrm{Res}^K_{K^+}(\sigma)$ and $\mathrm{Res}^K_{K^-}(\sigma)$ are both multiples of the trivial representation.

(3) There exists a strongly (P, K)-positive element z of the centre of M such that the double coset $Kz^{-1}K$ supports an invertible element of $\mathcal{H}_R(G, \sigma)$.

The point is that the properties of a G-cover allow one to define an injective homomorphism of algebras $j_P : \mathcal{H}(M, \sigma_M) \to \mathcal{H}(G, \sigma)$ and hence a (normalised) restriction functor $(j_P)^* : \mathcal{M}(G, \sigma) \to \mathcal{M}(M, \sigma_M)$; see [Bushnell and Kutzko 1998, p. 585] and [Vignéras 1998, II §10].

Theorem 4.2 [Vignéras 1998, II §10.1]. *Let π be a finitely generated ℓ-modular representation of G. We have an isomorphism $(j_P)^*(M_\sigma(\pi)) \simeq M_{\sigma_M}(r^G_P(\pi))$ of representations of M.*

4B. ***Level zero ℓ-modular representations.*** An irreducible representation π of G is of level zero if it has nontrivial invariants under the pro-p unipotent radical of some maximal parahoric subgroup of G.

Let Λ be a self-dual $\mathfrak{o}_F$-lattice sequence in V and $\mathrm{P}(\Lambda)$ the associated parahoric subgroup in G. We define *parahoric induction* $\mathrm{I}_\Lambda : \mathfrak{R}_R(\mathrm{M}(\Lambda)) \to \mathfrak{R}_R(G)$ on the objects of $\mathfrak{R}_R(\mathrm{M}(\Lambda))$ by

$$\mathrm{I}_\Lambda(\sigma) = \mathrm{ind}^G_{\mathrm{P}(\Lambda)}(\sigma)$$

for σ an R-representation of $\mathrm{M}(\Lambda)$, where, by abuse of notation, we also write σ for the inflation of σ to $\mathrm{P}(\Lambda)$ by defining $\mathrm{P}_1(\Lambda)$ to act trivially. This functor has a right-adjoint, *parahoric restriction* $\mathrm{R}_\Lambda : \mathfrak{R}_R(G) \to \mathfrak{R}_R(\mathrm{M}(\Lambda))$, defined on the objects of $\mathfrak{R}_R(G)$ by

$$\mathrm{R}_\Lambda(\pi) = \pi^{\mathrm{P}_1(\Lambda)}$$

for π an R-representation of G. Parahoric induction and restriction are exact functors.

We have the following important lemma, due to Vignéras [2001]. In her paper, the statement is for a general p-adic reductive group G.

Lemma 4.3 [Vignéras 2001]. *Let* $\mathrm{P}_1 = \mathrm{P}(\Lambda_1)$ *and* $\mathrm{P}_2 = \mathrm{P}(\Lambda_2)$ *be parahoric subgroups of* G. *Let* σ *be a representation of* $\mathrm{M}(\Lambda_2)$ *and fix a set* $D_{1,2}$ *of distinguished double coset representatives of* $\mathrm{P}_1\backslash G/\,\mathrm{P}_2$. *We have an isomorphism*

$$\mathrm{R}_{\Lambda_1} \circ \mathrm{I}_{\Lambda_2}(\sigma) \simeq \bigoplus_{n\in D_{1,2}} i^{\mathrm{M}(\Lambda_1)}_{P_{\Lambda_1,n\Lambda_2}}\big(r^{\mathrm{M}(\Lambda_2)}_{P_{\Lambda_2,n^{-1}\Lambda_1}}(\sigma)\big)^n.$$

Lemma 4.4. *Let* $\mathrm{P}(\Lambda_1)$ *and* $\mathrm{P}(\Lambda_2)$ *be parahoric subgroups of* G *associated to the* $\mathfrak{o}_F$*-lattice sequences* Λ_1 *and* Λ_2 *in* V. *Suppose that* $\mathrm{P}(\Lambda_2)$ *is maximal and let* σ *be an irreducible cuspidal representation of* $\mathrm{M}(\Lambda_2)$. *We have*

$$\mathrm{R}_{\Lambda_1} \circ \mathrm{I}_{\Lambda_2}(\sigma) \simeq \begin{cases} \sigma & \textit{if } \mathrm{P}(\Lambda_1) \textit{ is conjugate to } \mathrm{P}(\Lambda_2) \textit{ in } G, \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. By cuspidality of σ, if $r^{\mathrm{M}(\Lambda_2)}_{P_{\Lambda_2,n^{-1}\Lambda_1}}(\sigma) \neq 0$, then $P_{\Lambda_2,n^{-1}\Lambda_1} = M(\Lambda_2)$. If $\mathrm{P}(\Lambda_2)$ is not conjugate to $\mathrm{P}(\Lambda_1)$ then, for all $n \in D_{1,2}$, the parabolic subgroup $P_{\Lambda_2,n^{-1}\Lambda_1}$ is a proper parabolic subgroup of $\mathrm{M}(\Lambda_2)$. Hence $\mathrm{R}_{\Lambda_2} \circ \mathrm{I}_{\Lambda_1}(\sigma) \simeq 0$ by Lemma 4.3. As $N_G(\mathrm{P}(\Lambda_2)) = \mathrm{P}(\Lambda_2)$, if there exists $n \in D_{1,2}$ such that $\mathrm{P}(n^{-1}\Lambda_1) = \mathrm{P}(\Lambda_2)$, there can be only one such n. In this case, $\mathrm{R}_{\Lambda_1} \circ \mathrm{I}_{\Lambda_2}(\sigma) \simeq \sigma$, by Lemma 4.3. □

4C. *Positive level cuspidal ℓ-modular representations.*

4C1. *Semisimple strata and characters.* Let $[\Lambda, n, r, \beta]$ be a *skew semisimple stratum* in A; see [Stevens 2008, Definition 2.8]. Associated to $[\Lambda, n, r, \beta]$ and a fixed level one character of $F_0^\times$ are:

(1) A decomposition $V = \bigoplus_{i=1}^l V_i$, orthogonal with respect to h, and a sum of field extensions $E = \bigoplus_{i=1}^l E_i$ of E such that $\Lambda = \bigoplus_{i=1}^l \Lambda_i$ with Λ_i an $\mathfrak{o}_{E_i}$-lattice sequence in V_i; we say that Λ is an $\mathfrak{o}_E$*-lattice sequence* and write Λ_E when we are considering Λ as such.

(2) The F_0-points of a product of unramified unitary groups defined over F_0, $G_E = \prod_{i=1}^l G_{E_i}$.

(3) Compact open subgroups $H(\Lambda, \beta) \subseteq J(\Lambda, \beta)$ of G with decreasing filtrations by pro-p normal compact open subgroups $H^n(\Lambda, \beta) = H(\Lambda, \beta) \cap \mathrm{P}_n(\Lambda)$ and $J^n(\Lambda, \beta) = J(\Lambda, \beta) \cap \mathrm{P}_n(\Lambda)$, $n \geqslant 1$. When Λ is fixed we write $J = J(\Lambda, \beta)$, $H = H(\Lambda, \beta)$, and use similar notation for their filtration subgroups. We have $J = P(\Lambda_E)J^1$, where $\mathrm{P}(\Lambda_E)$ is the parahoric subgroup of G_E obtained by considering Λ as an $\mathfrak{o}_E$-lattice sequence.

(4) A set of *semisimple characters* $\mathscr{C}_-(\Lambda, r, \beta)$ of $H^{r+1}(\Lambda, \beta)$. For $r = 0$, we write $\mathscr{C}_-(\Lambda, \beta) = \mathscr{C}_-(\Lambda, 0, \beta)$.

Let $[\Lambda_i, n, 0, \beta]$, $i = 1, 2$, be skew semisimple strata in A. For all $\theta_1 \in \mathscr{C}_-(\Lambda_1, \beta)$, there is a unique $\theta_2 \in \mathscr{C}_-(\Lambda_2, \beta)$ such that $1 \in I_G(\theta_1, \theta_2)$, by [Stevens 2005, Proposition 3.32]. This defines a bijection

$$\tau_{\Lambda_1,\Lambda_2,\beta} : \mathscr{C}_-(\Lambda_1, \beta) \to \mathscr{C}_-(\Lambda_2, \beta)$$

and we call $\theta_2 = \tau_{\Lambda_1,\Lambda_2,\beta}(\theta_1)$ the *transfer* of θ_1.

The skew semisimple strata in A fall into three classes:

(1) Skew simple strata $[\Lambda, n, 0, \beta]$, where E is a field.

(a) If $E = F$ we say that $[\Lambda, n, 0, \beta]$ is a scalar skew simple stratum. In this case, $J/J^1 = \mathrm{P}(\Lambda)/\mathrm{P}_1(\Lambda)$ is isomorphic to one of $\mathrm{GL}_1(k_F) \times \mathrm{U}(1)(k_F/k_0)$, $\mathrm{U}(1, 1)(k_F/k_0) \times \mathrm{U}(1)(k_F/k_0)$ or $\mathrm{U}(2, 1)(k_F/k_0)$.

(b) Otherwise, E/F is cubic and $J/J^1 \simeq \mathrm{P}(\Lambda_E)/\mathrm{P}_1(\Lambda_E) \simeq \mathrm{U}(1)(k_E/k_{E^0})$ is a finite unitary group of order $q_{E^0} + 1$, where

$$q_{E^0} = \begin{cases} q_0^3 & \text{if } E/F \text{ is unramified,} \\ q_0 & \text{if } E/F \text{ is ramified.} \end{cases}$$

(2) Skew semisimple strata $[\Lambda, n, 0, \beta] = [\Lambda_1, n_1, 0, \beta_1] \oplus [\Lambda_2, n_2, 0, \beta_2]$, not equivalent to a skew simple stratum, with $[\Lambda_i, n_i, 0, \beta_i]$ skew simple strata in $\mathrm{End}_{F_0}(V_i)$. Without loss of generality, suppose that V_1 is one-dimensional and V_2 is two-dimensional. We have $J/J^1 \simeq \prod_{i=1}^2 \mathrm{P}(\Lambda_{i,E})/\mathrm{P}_1(\Lambda_{i,E})$. If $\beta_2 \in F$ and V_2 is hyperbolic, then $G_E \simeq U(1, 1)(F/F_0) \times \mathrm{U}(1)(F/F_0)$ and $\mathrm{P}(\Lambda_{2,E})$ is a parahoric subgroup of $\mathrm{U}(1, 1)(F/F_0)$ and need not be maximal. If $\beta_2 \in F$ and V_2 is anisotropic, then $G_E \simeq \mathrm{U}(2)(F/F_0) \times \mathrm{U}(1)(F/F_0)$ is compact. If E_2/F is quadratic then it is ramified, because there is a unique unramified extension of F_0 in each degree and E_2^0/F_0 is quadratic and also fixed by the involution. Thus, if E_2/F is quadratic then $J/J^1 \simeq \mathrm{U}(1)(k_F/k_0) \times \mathrm{U}(1)(k_F/k_0)$.

(3) Skew semisimple strata $[\Lambda, n, 0, \beta] = \bigoplus_{i=1}^3 [\Lambda_i, n_i, 0, \beta_i]$, not equivalent to a skew semisimple stratum of the first two classes, with $[\Lambda_i, n_i, 0, \beta_i]$ skew simple strata in $\mathrm{End}_{F_0}(V_i)$. In this case, $J/J^1 \simeq \mathrm{U}(1)(k_F/k_0) \times \mathrm{U}(1)(k_F/k_0) \times \mathrm{U}(1)(k_F/k_0)$.

We say that π *contains* the skew semisimple stratum $[\Lambda, n, 0, \beta]$ if it contains a character $\theta \in \mathscr{C}_-(\Lambda, \beta)$.

Theorem 4.5 [Stevens 2005, Theorem 5.1]. *Let π be an irreducible cuspidal ℓ-modular representation of G. Then π contains a skew semisimple stratum $[\Lambda, n, 0, \beta]$.*

4C2. *Heisenberg representations.* Let $\theta \in \mathscr{C}_-(\Lambda, \beta)$. By [Stevens 2008, Corollary 3.29], there exists a unique irreducible representation η of $J^1(\Lambda, \beta)$ which contains θ. We call such an η a *Heisenberg representation.* Furthermore, by [Stevens

2008, Proposition 3.31],

$$\dim_R(I_g(\eta)) = \begin{cases} 1 & \text{if } g \in J^1 G_E J^1, \\ 0 & \text{otherwise.} \end{cases}$$

4C3. *β-extensions.* Assume $P(\Lambda_E)$ is maximal. A *β-extension* of a Heisenberg representation η to $J = J(\Lambda, \beta)$ is an extension κ with maximal intertwining, $I_G(\kappa) = I_G(\eta)$. By [Blasco 2002, Lemma 5.8], for all maximal skew semisimple strata which are not skew scalar simple strata, β-extensions exist in the ℓ-adic case for G, and for ℓ-modular representations we obtain β-extensions by reduction modulo ℓ from the ℓ-adic extensions. Note that the reduction modulo ℓ of an ℓ-adic β-extension $\tilde{\kappa}$ of J is irreducible: its restriction to J^1 is the reduction modulo ℓ of $\tilde{\eta} = \mathrm{Res}^J_{J^1}(\tilde{\kappa})$, reduction modulo ℓ commutes with restriction, and, as J^1 is pro-p , the reduction modulo ℓ of $\tilde{\eta}$ is irreducible. Let $[\Lambda, n, 0, \beta]$ be a scalar skew simple stratum and $\theta \in \mathscr{C}_-(\Lambda, \beta)$. Then $J^1 = H^1 = \mathrm{P}_1(\Lambda)$, $J = \mathrm{P}(\Lambda)$, and $\theta = \chi \circ \det$ for some character χ of $\mathrm{P}_1(\Lambda)$ (cf. [Bushnell and Kutzko 1993a, Definition 3.23]). The character χ extends to a character $\tilde{\chi}$ of F^1 and we define $\kappa : J \to R^\times$ by $\kappa = \tilde{\chi} \circ \det$. Then κ extends θ and is intertwined by all of G, hence is a β-extension. Hence, in the maximal case, β-extensions exist.

Let $[\Lambda, n, 0, \beta]$ be a skew semisimple stratum. Suppose $P(\Lambda_E)$ is not maximal and choose a maximal parahoric subgroup $P(\Lambda_E^m)$ of G_E associated to the $\mathfrak{o}_E$-lattice sequence Λ_E^m in V such that $P(\Lambda_E) \subset P(\Lambda_E^m)$. This implies that $P(\Lambda) \subset P(\Lambda^m)$. Note that this is the case for unramified $\mathrm{U}(2, 1)(E/F)$, but not for classical groups in general. Let $\theta \in \mathscr{C}_-(\Lambda, \beta)$ and let η be the irreducible representation of $J^1_m = J^1(\beta, \Lambda)$ which contains θ. Let $\theta_m = \tau_{\Lambda, \Lambda^m, \beta}(\theta)$ and let η_m be the irreducible representation of $J^1(\beta, \Lambda^m)$ which contains θ_m. Let κ_m be a β-extension of η_m.

Lemma 4.6. *There exists a unique extension κ of η to J such that* $\mathrm{Res}^{J_m}_{P(\Lambda_E)J^1_m}(\kappa_m)$ *and κ induce equivalent irreducible representations of* $P(\Lambda_E)P_1(\Lambda)$.

Proof. If $\mathrm{P}(\Lambda_E)$ is maximal then $\kappa_m = \kappa$ and there is nothing to prove. Let $\tilde{\kappa}_m$ be a lift of κ. By [Stevens 2008, Lemma 4.3], there exists a unique irreducible ℓ-adic representation $\tilde{\kappa}$ of J such that $\mathrm{Res}^{J_m}_{P(\Lambda_E)J^1_m}(\tilde{\kappa}_m)$ and κ induce equivalent irreducible representations of $P(\Lambda_E)P_1(\Lambda)$. By reduction modulo ℓ, we have an irreducible ℓ-modular representation $\kappa = r_\ell(\tilde{\kappa})$ which extends η such that

$$\big[\mathrm{ind}_J^{\mathrm{P}(\Lambda_E)\,\mathrm{P}_1(\Lambda)}\,\kappa\big] = \big[\mathrm{ind}^{\mathrm{P}(\Lambda_E)\,\mathrm{P}_1(\Lambda)}_{\mathrm{P}(\Lambda_E)J^1_m}\,\mathrm{Res}^{J_m}_{\mathrm{P}(\Lambda_E)J^1_m}(\kappa_m)\big].$$

By Mackey Theory,

$$\mathrm{Res}^{P(\Lambda_E)P_1(\Lambda)}_{P_1(\Lambda)}(\mathrm{ind}_J^{P(\Lambda_E)P_1(\Lambda)}\,\kappa) \simeq \mathrm{ind}_{J^1}^{P_1(\Lambda)}\,\kappa.$$

Furthermore, $J^1 \subseteq I_{P_1(\Lambda)}(\kappa) \subseteq I_{P_1(\Lambda)}(\eta) = J^1$, so $\mathrm{ind}_{J^1}^{P_1(\Lambda)}\kappa$ and hence $\mathrm{ind}_J^{\mathrm{P}(\Lambda_E)\,\mathrm{P}_1(\Lambda)}\kappa$ are irreducible. □

A *β-extension* of η is an extension κ of η to J constructed in this way. We call two β-extensions which induce equivalent representations, as in Lemma 4.6, *compatible*. With the next lemma we show we can "go backwards" and from a β-extension defined in the minimal case we define two unique compatible β-extensions in the maximal case. In this way we get a triple of *compatible* β-extensions. Let $\mathrm{P}(\Lambda_E^r)$ be a maximal parahoric subgroup of G_E containing $\mathrm{P}(\Lambda_E)$ associated to the $\mathfrak{o}_E$-lattice sequence Λ_E^r in V. Let $\theta_r = \tau_{\Lambda,\Lambda^r,\beta}(\theta)$, η_r be the irreducible representation of $J^1(\beta, \Lambda^r)$ which contains θ_r, and κ be a β-extension of η.

Lemma 4.7. *There exists a unique β-extension κ_r of η_r which is compatible with κ.*

Proof. By [Stevens 2008, Lemma 4.3], there exists a representation $\hat{\kappa}$ of $P(\Lambda_E)J_r^1$ such that κ and $\hat{\kappa}$ induce equivalent representations of $\mathrm{P}(\Lambda_E)\,\mathrm{P}_1(\Lambda)$. Let κ' be a β-extension of η_r. The restriction to $\mathrm{P}(\Lambda_E)J_r^1$ of κ' and $\hat{\kappa}$ differ by a character χ of $B_r = \mathrm{P}(\Lambda_E)/\,\mathrm{P}_1(\Lambda_E^r)$ which is trivial on the unipotent part of B_r and intertwined by the nontrivial Weyl group element w. By the Bruhat decomposition, $\mathrm{M}_r = \mathrm{M}(\Lambda_E^r) = B_r \cup B_r w B_r$; hence χ is intertwined by the whole of M_r and extends to a character of M_r. Hence $\kappa_r = \kappa \otimes \chi^{-1}$ is a β-extension of η_r which is compatible with κ. By reduction modulo ℓ, as in the proof of Lemma 4.6, we have the corresponding statement in the ℓ-modular setting. □

4C4. *κ-induction and restriction.* Fix $[\Lambda, n, 0, \beta]$ a skew semisimple stratum in A, $\theta \in \mathscr{C}_-(\Lambda, \beta)$, η the unique Heisenberg representation containing θ and κ a β-extension of η.

Let σ be an R-representation of $\mathrm{M}(\Lambda_E)$ and, by abuse of notation, we also write σ for the inflation of σ to J obtained by defining J^1 to act trivially. The functor $\mathfrak{R}_R(\mathrm{M}(\Lambda_E)) \to \mathfrak{R}_R(J)$ given by $\sigma \mapsto \kappa \otimes \sigma$ identifies $\mathfrak{R}_R(\mathrm{M}(\Lambda_E))$ with the full subcategory of η-isotypic representations of J; see [Vignéras 2001, Definition 8.1]. Define *κ-induction*, $\mathrm{I}_\kappa : \mathfrak{R}_R(\mathrm{M}(\Lambda_E)) \to \mathfrak{R}_R(G)$, by

$$\mathrm{I}_\kappa(\sigma) = \mathrm{ind}_J^G(\kappa \otimes \sigma)$$

for σ an R-representation of $\mathrm{M}(\Lambda_E)$ and defined analogously on the morphisms of $\mathfrak{R}_R(\mathrm{M}(\Lambda_E))$. This functor has a right adjoint, $\mathrm{R}_\kappa : \mathfrak{R}_R(G) \to \mathfrak{R}_R(\mathrm{M}(\Lambda_E))$, called *$\kappa$-restriction*, defined by

$$\mathrm{R}_\kappa(\pi) = \mathrm{Hom}_{J^1}(\kappa, \pi),$$

where the action of $\mathrm{M}(\Lambda_E)$ is given by: for $f \in \mathrm{Hom}_{J^1}(\kappa, \pi)$ and $m \in \mathrm{M}(\Lambda_E)$, let $j \in J$ represent the coset $m \in J/J^1$, then $m \cdot f = \pi(j) \circ f \circ \kappa(j^{-1})$.

In the level zero case, we have $J = \mathrm{P}(\Lambda)$ and we can choose κ to be trivial, thus we have $\mathrm{I}_\kappa = \mathrm{I}_\Lambda$ and $\mathrm{R}_\kappa = \mathrm{R}_\Lambda$. Hence κ-restriction and induction generalise parahoric restriction and induction. Related to $[\Lambda, n, 0, \beta]$, we also have functors of parahoric induction $\mathrm{I}_\Lambda^E : \mathfrak{R}_R(M(\Lambda_E)) \to \mathfrak{R}_R(G_E)$ and parahoric restriction $\mathrm{R}_\Lambda^E : \mathfrak{R}_R(G_E) \to \mathfrak{R}_R(M(\Lambda_E))$ obtained by considering Λ as an $\mathfrak{o}_E$-lattice sequence.

Theorem 4.8 [Kurinczuk and Stevens 2014]. *Let* $[\Lambda^i, n, 0, \beta]$, $i = 1, 2$, *be skew semisimple strata. Let* $\theta_1 \in \mathscr{C}_-(\Lambda^1, \beta)$ *and* $\theta_2 = \tau_{\Lambda^1,\Lambda^2,\beta}(\theta_1)$. *For* $i = 1, 2$, *let* η_i *be a Heisenberg extension of* θ_i, κ_i *be compatible* β*-extensions of* η_i, *and let* σ *be an* R*-representation of* $\mathrm{M}(\Lambda_E^1)$. *Then*

$$\mathrm{R}_{\kappa_2} \circ \mathrm{I}_{\kappa_1}(\sigma) \simeq \mathrm{R}_{\Lambda^2}^E \circ \mathrm{I}_{\Lambda^1}^E(\sigma).$$

The proof of Theorem 4.8 in [Kurinczuk and Stevens 2014] follows from a combination of Mackey theory, isomorphisms defined as in [Bushnell and Kutzko 1993a, Proposition 5.3.2], and the computation of the intertwining spaces $\mathrm{I}_g(\eta_1, \eta_2)$ for $g \in G$, which are one-dimensional if $g \in G_E$ and zero otherwise.

Lemma 4.9. *In the setting of Lemma 4.6, let* κ *and* κ_m *be compatible* β*-extensions. Then, for all* $\sigma \in \mathfrak{R}_R(\mathrm{M}(\Lambda_E))$, *we have*

$$\mathrm{I}_\kappa(\sigma) \simeq \mathrm{ind}_{J_m^1\,\mathrm{P}(\Lambda_E)}^G(\kappa_m \otimes \sigma)$$

and, for all R*-representations* π *of* G, *we have* $\mathrm{R}_\kappa(\pi) \simeq \mathrm{Hom}_{J_m^1\,\mathrm{P}_1(\Lambda_E)}(\kappa_m, \pi)$.

Proof. By transitivity of induction and Lemma 4.6, $\mathrm{I}_\kappa(\sigma) \simeq \mathrm{ind}_{J_m^1\,\mathrm{P}(\Lambda_E)}^G(\kappa_m \otimes \sigma)$. By reciprocity, for π an R-representation of G, $\mathrm{R}_\kappa(\pi) \simeq \mathrm{Hom}_{J_m^1\,\mathrm{P}_1(\Lambda_E)}(\kappa_m, \pi)$. □

Define $\tilde{\kappa}$*-induction* $\mathrm{I}_{\tilde{\kappa}} : \mathfrak{R}_R(\mathrm{M}(\Lambda_E)) \to \mathfrak{R}_R(G)$ by $\mathrm{I}_{\tilde{\kappa}}(\sigma) = \mathrm{ind}_J^G(\tilde{\kappa} \otimes \sigma)$ for σ an R-representation of $\mathrm{M}(\Lambda_E)$. This functor has a right adjoint, $\tilde{\kappa}$*-restriction*, $\mathrm{R}_{\tilde{\kappa}} : \mathfrak{R}_R(G) \to \mathfrak{R}_R(\mathrm{M}(\Lambda_E))$ defined by $\mathrm{R}_{\tilde{\kappa}}(\pi) = \mathrm{Hom}_{J^1}(\tilde{\kappa}, \pi)$, where the action of $\mathrm{M}(\Lambda_E)$ on $\mathrm{R}_{\tilde{\kappa}}(\pi)$ is defined analogously to κ-restriction. In fact, $\tilde{\kappa}$ is a $-\beta$-extension for the semisimple character θ^{-1} for the semisimple stratum $[\Lambda, n, 0, -\beta]$.

Lemma 4.10. *Let* π *be an* R*-representation of* G *and* σ *be an irreducible representation of* $\mathrm{M}(\Lambda_E)$. *Then* $(\mathrm{R}_\kappa(\pi))^\sim \simeq \mathrm{R}_{\tilde{\kappa}}(\tilde{\pi})$ *and, if* $\mathrm{I}_\kappa(\sigma)$ *is irreducible, then* $\mathrm{I}_\kappa(\sigma)^\sim \simeq \mathrm{I}_{\tilde{\kappa}}(\tilde{\sigma})$.

Proof. We have an isomorphism of vector spaces

$$\mathrm{Hom}_{J^1}(\kappa, \pi)^\sim \simeq \mathrm{Hom}_{J^1}(\pi, \kappa) \simeq \mathrm{Hom}_{J^1}(\tilde{\kappa}, \tilde{\pi})$$

by [Henniart and Sécherre 2014, Proposition 2.6], and checking the action of J/J^1 we have $(\mathrm{R}_\kappa(\pi))^\sim \simeq \mathrm{R}_{\tilde{\kappa}}(\tilde{\pi})$. If $\mathrm{I}_\kappa(\sigma)$ is irreducible, then it is admissible and we have $\mathrm{I}_\kappa(\sigma)^\sim \simeq \mathrm{I}_{\tilde{\kappa}}(\tilde{\sigma})$ by [Vignéras 1996, I 8.4]. □

If $\mathrm{P}(\Lambda_E)$ is not maximal, let $\kappa_T = \mathrm{Res}_{T^0}^J(\kappa)$. Define $\mathrm{R}_{\kappa_T,\Lambda} : \mathfrak{R}_R(T) \to \mathfrak{R}_R(\overline{T})$ by $\mathrm{R}_{\kappa_T,\Lambda}(\pi) = \mathrm{Hom}_{T^1}(\kappa_T, \pi)$.

5. Exhaustion of cuspidal representations

In this section, we exhaust all irreducible cuspidal ℓ-modular representations of unramified $\mathrm{U}(2,1)(F/F_0)$. To do this we construct covers. The construction we give here is a vast simplification of that of [Stevens 2008], available as we are in the special case of unramified $\mathrm{U}(2,1)$. As the covers are constructed on compact open subgroups with pro-order not necessarily invertible in $\overline{\mathbb{F}}_\ell$ it is not clear whether or not the construction will follow *mutatis mutandis* the complex construction. In fact, for the relatively simple proof we give here for $\mathrm{U}(2,1)(F/F_0)$, it does. It is only when we come to computing the parameters of associated Hecke algebras later that we have to change the complex proof, and these changes occur in computing the parameters of Hecke algebras of certain associated finite reductive groups.

5A. ***Covers.*** In the ℓ-adic case our construction of G-covers is a special case of the general results of [Stevens 2008, Propositions 7.10 and 7.13]. Let $[\Lambda, n, 0, \beta]$ be a skew semisimple stratum in A such that $\mathrm{P}(\Lambda_E)$ is not a maximal parahoric subgroup of G_E. For unramified $\mathrm{U}(2,1)(F/F_0)$, this implies $H^1(\Lambda,\beta) = J^1(\Lambda,\beta)$, which, in the notation of [ibid.], implies that $J = J_B$. Moreover, $\mathrm{P}(\Lambda_E)/\mathrm{P}_1(\Lambda_E)$ is abelian and isomorphic to $k_E^1 \times k_E^\times$. Let $\theta \in \mathscr{C}_-(\Lambda,\beta)$; then $\eta = \theta$ is the unique Heisenberg representation containing θ. Let κ be a β-extension of η and $\sigma \in \mathrm{Irr}_R(J/J^1)$. Then $\lambda = \kappa \otimes \sigma$ is a character of J. Let $\kappa_T = \mathrm{Res}^J_{T^0}(\kappa)$ and set $\lambda_T = \kappa_T \otimes \sigma$. Let $J = (J \cap \overline{N})(J \cap T)(J \cap N)$ be the Iwahori decomposition of J with respect to B. We have $\lambda(j^- j_T j^+) = \lambda_T(j_T)$ for $j^- \in (J \cap \overline{N})$, $j_T \in J \cap T$, and $j^+ \in (J \cap N)$.

Lemma 5.1. *The element w_x intertwines λ if and only if w_y intertwines λ.*

Proof. Suppose $w_x \in I_G(\lambda)$. Then, as w_x normalises T^0, w_x normalises $\mathrm{Res}^J_{T^0}(\lambda)$. For all $t \in T^0$ we have $w_x t w_x = w_y t w_y$; hence w_y normalises $\mathrm{Res}^J_{T^0}(\lambda)$. Let $j \in J \cap w_y J w_y$ be such that $j = w_y j' w_y$ with $j' \in J$. Using the Iwahori decomposition of J we have $j = j_{\overline{N}} j_T j_N$ and $j' = j'_N j'_T j'_{\overline{N}}$ with j_N, j'_N upper triangular unipotent, $j_{\overline{N}}$, $j'_{\overline{N}}$ lower triangular unipotent and j_T, j'_T in T. Thus

$$j = w_y j' w_y^{-1} = (w_y j'_N w_y)(w_y j'_T w_y)(w_y j'_{\overline{N}} w_y)$$

and, by unicity of the Iwahori decomposition, $j_{\overline{N}} = w_y j'_N w_y$, $j_T = w_y j'_T w_y$ and $j_N = w_y j'_{\overline{N}} w_y$. Therefore $w_y \in I_G(\lambda)$. □

Lemma 5.2. *Let $\lambda_T = \kappa_T \otimes \sigma$. Then (J,λ) is a G-cover of (T^0, λ_T).*

Proof. In the ℓ-modular case, it remains to show that there exists a strongly (B, J)-positive element z of the centre of T such that $Jz^{-1}J$ supports an invertible element of $\mathscr{H}(G,\lambda)$. Let $\zeta = w_x w_y$. Then ζ is strongly (B, J)-positive. For $g \in I_G(\lambda)$, because λ is a character, $I_g(\lambda) \simeq R$ and there is a unique function in

$f_g \in \mathcal{H}(G,\lambda)$ with support JgJ such that $f_g(g)=1$. We have $\zeta, \zeta^{-1} \in I_G(\lambda)$; hence $f_\zeta, f_{\zeta^{-1}} \in \mathcal{H}(G,\lambda)$.

Suppose that $w_x \notin I_G(\lambda)$, i.e., $I_G(\lambda)=JTJ$. As ζ is strongly positive,

$$J\zeta J\zeta^{-1}J = J\zeta J^{-}\zeta^{-1}J.$$

Suppose $y \in J\zeta J\zeta^{-1}J \cap JTJ$. Then we can write $y=j_1tj_2$ and $y=j_3\zeta j^{-}\zeta^{-1}j_4$ with $j_1, j_2, j_3, j_4 \in J$, $t \in T$ and $j^{-} \in J^{-}$. Thus, we can write

$$\zeta j^{-}\zeta^{-1} = jtj'$$

with $j, j' \in J$. By the Iwahori decomposition of J applied to the elements j and $(j')^{-1}$, we have

$$\zeta j^{-}\zeta^{-1} = j_{\overline{N}}j_Tj_Ntj'_Nj'_Tj'_{\overline{N}}$$

with $j_N, j'_N \in J \cap N$, $j_{\overline{N}}, j'_{\overline{N}} \in J \cap \overline{N}$ and $j_T, j'_T \in J \cap T$. Then $j_{\overline{N}}^{-1}\zeta j^{-}\zeta^{-1}(j'_{\overline{N}})^{-1} \in N$ and $j_{\overline{N}}^{-1}\zeta j^{-}\zeta^{-1}(j'_{\overline{N}})^{-1} = j_Tj_Ntj'_Nj'_T \in B$; hence $j_{\overline{N}}^{-1}\zeta j^{-}\zeta^{-1}(j'_{\overline{N}})^{-1}=1$ and $\zeta j^{-}\zeta^{-1} \in J$. Therefore, $y \in J$ and $J\zeta J\zeta^{-1}J \cap JTJ = J$. Hence $f_\zeta \star f_{\zeta^{-1}}$ is supported on the single double coset J. We have $f_\zeta \star f_{\zeta^{-1}}(1_G)=q^4$. Hence $f_{\zeta^{-1}}$ is an invertible element of $\mathcal{H}(G,\lambda)$ supported on the single double coset $J\zeta^{-1}J$.

Now, suppose that $w_x \in I_G(\lambda)$, then $w_y \in I_G(\lambda)$ by Lemma 5.1. Hence $f_{w_x}, f_{w_y} \in \mathcal{H}(G,\lambda)$. Let $s \in \{x,y\}$. The maximal parahoric subgroup $P(\Lambda_s)$ of G contains J and w_s and $P(\Lambda_s) \cap G_E$ is a maximal parahoric subgroup of G_E. Moreover, $(J \cap G_E)/(P_1(\Lambda_s) \cap G_E)$ is a Borel subgroup of $(P(\Lambda_s) \cap G_E)/(P_1(\Lambda_s) \cap G_E)$. By [Stevens 2008, Lemma 5.12], $I_G(\eta)=JG_EJ$, thus the support of $\mathcal{H}(G,\lambda)$ is contained in JG_EJ. Hence,

$$\begin{aligned}\operatorname{supp}(f_{w_s} \star f_{w_s}) &\subseteq (Jw_sJw_sJ) \cap JG_EJ\\ &\subseteq P(\Lambda_s) \cap JG_EJ = J(P(\Lambda_s) \cap G_E)J\\ &= J((J \cap G_E) \cup (J \cap G_E)w_s(J \cap G_E))J,\end{aligned}$$

by the Bruhat decomposition of $(P(\Lambda_s) \cap G_E)/(P_1(\Lambda_s) \cap G_E)$, which is a finite reductive group. Thus, $\operatorname{supp}(f_{w_s} \star f_{w_s}) \subseteq J \cup Jw_sJ$. We have that $f_{w_s} \star f_{w_s}(1_G) = [J : J \cap w_sJw_s]$ is a power of q. Let $a_s = f_{w_s} \star f_{w_s}(1_G)$ and $b_s = f_{w_s} \star f_{w_s}(w_s)$. Therefore, for $s \in \{x,y\}$, f_{w_s} is an invertible element of $\mathcal{H}(G,\lambda)$ with inverse $(1/a_s)(f_{w_s} - b_sf_1)$. By [Stevens 2008, Lemma 7.11], we have $(J \cap N)^{w_x} \subseteq J \cap N$ and $(J \cap \overline{N})^{w_y} \subseteq J \cap \overline{N}$. By the Iwahori decomposition of J,

$$\begin{aligned}Jw_yJw_xJ &= J(w_y(J \cap \overline{N})w_y)w_yw_x(w_x(J \cap T)w_x)(w_x(J \cap N)w_x)J\\ &= J(J \cap \overline{N})^{w_y}w_yw_x(J \cap T)^{w_x}(J \cap N)^{w_x}J\\ &\subseteq J(J \cap \overline{N})w_yw_x(J \cap T)(J \cap N)J = Jw_yw_xJ.\end{aligned}$$

Moreover, we clearly have $Jw_yw_xJ \subseteq Jw_yJw_xJ$. Hence $Jw_yw_xJ = Jw_yJw_xJ$.

Therefore, $f_{w_y} \star f_{w_x}$ is an invertible element of $\mathscr{H}(G, \lambda)$ supported on the single double coset $J\zeta^{-1}J$. □

5B. *Cuspidal representations.* The following theorem addresses the construction of all irreducible cuspidal ℓ-modular and ℓ-adic representations of G.

Theorem 5.3. (1) *Let* $[\Lambda, n, 0, \beta]$ *be a skew semisimple stratum in* A, $\theta \in \mathscr{C}_-(\beta, \Lambda)$, η *the unique Heisenberg representation containing* θ, κ *a* β*-extension of* η *and* σ *an irreducible cuspidal representation of* $\mathrm{M}(\Lambda_E)$. *Then* $\mathrm{I}_\kappa(\sigma)$ *is quasiprojective. Furthermore, if* $\mathrm{P}(\Lambda_E)$ *is a maximal parahoric subgroup of* G_E, *then* $\mathrm{I}_\kappa(\sigma)$ *is irreducible and cuspidal.*

(2) *Let* π *be an irreducible cuspidal representation of* G. *Then there exist a skew semisimple stratum* $[\Lambda, n, 0, \beta]$ *with* $\mathrm{P}(\Lambda_E)$ *a maximal parahoric subgroup of* G_E, $\theta \in \mathscr{C}_-(\beta, \Lambda)$, *a* β*-extension* κ *of the unique Heisenberg representation* η *which contains* θ *and an irreducible cuspidal representation* σ *of* $\mathrm{M}(\Lambda_E)$ *such that* $\pi \simeq \mathrm{I}_\kappa(\sigma)$.

Proof. (1) Quasiprojectivity follows *mutatis mutandis* the proof given in [Vignéras 2001, Proposition 6.1]. So suppose $\mathrm{P}(\Lambda_E)$ is a maximal parahoric subgroup of G_E. By Theorem 4.8 and Lemma 4.4 we have

$$\mathrm{R}_\kappa \circ \mathrm{I}_\kappa(\sigma) \simeq \mathrm{R}^E_\Lambda \circ \mathrm{I}^E_\Lambda(\sigma) \simeq \sigma.$$

The proof of irreducibility follows *mutatis mutandis* the proof given in [Vignéras 2001, Proposition 7.1].

(2) By Theorem 4.5, π contains a skew semisimple stratum $[\Lambda, n, 0, \beta]$. Suppose $\theta \in \mathscr{C}_-(\Lambda, \beta)$ is a skew semisimple character which π contains. Let κ be a β-extension of the unique Heisenberg representation η which contains θ. Then π contains $\kappa \otimes \sigma$ for some $\sigma \in \mathrm{Irr}_R(\mathrm{M}(\Lambda_E))$. We show that we may assume that σ is cuspidal. If $\mathrm{P}(\Lambda_E)$ is not maximal then σ is cuspidal, so we can suppose that $\mathrm{P}(\Lambda_E)$ is maximal. Let $B(\Lambda_E)$ be the standard Borel subgroup of $\mathrm{M}(\Lambda_E)$ and $\mathrm{P}(\Lambda'_E)$ the preimage of $B(\Lambda_E)$ under the projection map. Suppose that $r^{\mathrm{M}(\Lambda_E)}_{B(\Lambda_E)}(\sigma) \neq 0$. Then, as π contains σ, $r^{\mathrm{M}(\Lambda_E)}_{B(\Lambda_E)}(\mathrm{R}_\kappa(\pi)) \neq 0$. We have

$$r^{\mathrm{M}(\Lambda_E)}_{B(\Lambda_E)}(\mathrm{R}_\kappa(\pi)) \simeq \mathrm{Hom}_{J^1}(\kappa, \pi)^{\mathrm{P}_1(\Lambda'_E)J^1/J^1} \simeq \mathrm{Hom}_{\mathrm{P}_1(\Lambda'_E)J^1}(\kappa, \pi),$$

which, by Lemma 4.9, implies that $\mathrm{R}_{\kappa', \Lambda'}(\pi) \neq 0$ where κ' is the unique β-extension containing $\tau_{\Lambda,\Lambda',\beta}(\theta)$ compatible with κ. Hence π contains a skew semisimple stratum $[\Lambda', n, 0, \beta]$ such that $\mathrm{P}(\Lambda'_E)$ is not maximal and thus contains $\kappa' \otimes \sigma'$, with σ' a cuspidal representation of $\mathrm{M}(\Lambda'_E)$. By Theorem 4.2 and Lemma 5.2, if π contains a skew semisimple stratum $[\Lambda, n, 0, \beta]$ such that $\mathrm{P}(\Lambda_E)$ is not a maximal parahoric subgroup of G_E, then π is not cuspidal. Therefore $\mathrm{P}(\Lambda_E)$ is maximal and σ is cuspidal. □

For level zero representations we can refine the exhaustive list of irreducible cuspidal representations given in Theorem 5.3 into a classification.

Theorem 5.4. *For $i = 1, 2$, let* $\mathrm{P}(\Lambda_i)$ *be a maximal parahoric subgroup of* G *and* σ_i *an irreducible cuspidal representation of* $\mathrm{M}(\Lambda_i)$. *If* $\mathrm{Hom}_G(\mathrm{I}_{\Lambda_1}(\sigma_1), \mathrm{I}_{\Lambda_2}(\sigma_2)) \neq 0$ *then* $(\mathrm{P}(\Lambda_1), \sigma_1)$ *and* $(\mathrm{P}(\Lambda_2), \sigma_2)$ *are conjugate.*

Proof. By reciprocity and Lemma 4.3,

$$\mathrm{Hom}_G(\mathrm{I}_{\Lambda_1}(\sigma_1), \mathrm{I}_{\Lambda_2}(\sigma_2)) \simeq \bigoplus_{n \in D_{1,2}} \mathrm{Hom}_{\mathrm{M}(\Lambda_1)}\big(\sigma_1, i_{P_{\Lambda_1, n\Lambda_2}}^{\mathrm{M}(\Lambda_1)}\big(r_{P_{\Lambda_2, n^{-1}\Lambda_1}}^{\mathrm{M}(\Lambda_2)}(\sigma_2)\big)^n\big).$$

Hence

$$\mathrm{Hom}_G(\mathrm{I}_{\Lambda_1}(\sigma_1), \mathrm{I}_{\Lambda_2}(\sigma_2)) \neq 0$$

if and only if there exists $n \in D_{1,2}$ such that

$$\mathrm{Hom}_{\mathrm{M}(\Lambda_1)}\big(\sigma_1, i_{P_{\Lambda_1, n\Lambda_2}}^{\mathrm{M}(\Lambda_1)}\big(r_{P_{\Lambda_2, n^{-1}\Lambda_1}}^{\mathrm{M}(\Lambda_2)}(\sigma_2)\big)^n\big) \neq 0.$$

Assume there exists such an element n. By cuspidality of σ_2, $P_{\Lambda_2, n^{-1}\Lambda_1} = \mathrm{M}(\Lambda_2)$, so $\mathrm{P}_1(\Lambda_2)(\mathrm{P}(\Lambda_2) \cap \mathrm{P}(n^{-1}\Lambda_1)) / \mathrm{P}_1(\Lambda_2) = M(\Lambda_2)$. By cuspidality of σ_1, $P_{\Lambda_1, n\Lambda_2} = \mathrm{M}(\Lambda_1)$, so $\mathrm{P}_1(\Lambda_1)(\mathrm{P}(\Lambda_1) \cap \mathrm{P}(n\Lambda_2)) / \mathrm{P}_1(\Lambda_1) = \mathrm{M}(\Lambda_1)$. If $\mathrm{P}(\Lambda_1)$ and $\mathrm{P}(\Lambda_2)$ are not conjugate then for all $g \in G$, in particular $n \in D_{1,2}$, the group $\mathrm{P}(\Lambda_1) \cap \mathrm{P}(g\Lambda_2)$ must stabilise an edge in the building and hence is not maximal. Thus it cannot surject onto either $\mathrm{M}(\Lambda_1)$ or $\mathrm{M}(\Lambda_2)$. Hence there exists $n \in D_{1,2}$ such that $\mathrm{P}(\Lambda_1) = \mathrm{P}(n\Lambda_2)$ and $\mathrm{Hom}_{\mathrm{M}(\Lambda_1)}(\sigma_1, \sigma_2^n) \neq \{0\}$; i.e., $(\mathrm{P}(\Lambda_1), \sigma_1)$ and $(\mathrm{P}(\Lambda_2), \sigma_2)$ are conjugate. □

Remark 5.5. Let $\ell \mid (q^2 - q + 1)$. The irreducible cuspidal ℓ-modular representations $\mathrm{I}_{\Lambda_x}(\overline{\tau}^+(\overline{\chi}))$ do not lift. A lift must necessarily be cuspidal as the Jacquet functor commutes with reduction modulo ℓ. However, by Theorem 5.3, all ℓ-adic level zero irreducible cuspidal representations are of the form $\mathrm{I}_{\Lambda_x}(\sigma_x)$ or $\mathrm{I}_{\Lambda_y}(\sigma_y)$ with σ_x (resp. σ_y) an irreducible cuspidal ℓ-adic representation of $\mathrm{M}(\Lambda_x)$ (resp. $\mathrm{M}(\Lambda_y)$). Furthermore, $r_\ell(\mathrm{I}_{\Lambda_w}(\sigma_w)) = \mathrm{I}_{\Lambda_w}(r_\ell(\sigma_w))$ as compact induction commutes with reduction modulo ℓ, for $w \in \{x, y\}$. Hence, by Section 3B2, $\mathrm{I}_{\Lambda_x}(\overline{\tau}^+(\overline{\chi}))$ does not lift, but does appear in the reduction modulo ℓ of $\mathrm{I}_{\Lambda_x}(\tau(\psi))$, where $r_\ell(\mathrm{I}_{\Lambda_x}(\tau(\psi)) = \mathrm{I}_{\Lambda_x}(\overline{\nu}(\overline{\chi})) \oplus \mathrm{I}_{\Lambda_x}(\overline{\tau}^+(\overline{\chi}))$.

6. Parabolically induced representations

Let ω_{F/F_0} be the unique character of $F_0^\times$ associated to F/F_0 by local class field theory. That is, ω_{F/F_0} is defined by $\omega_{F/F_0}|_{\mathfrak{o}_{F_0}^\times} = 1$ and $\omega_{F/F_0}(\varpi_F) = -1$. All extensions of ω_{F/F_0} to $F^\times$ take values in $\overline{\mathbb{Z}}_\ell^\times$, hence are integral. Let χ_1 be a character of $F^\times$ and χ_2 be a character of F^1. Let χ be the character of T defined by

$$\chi(\mathrm{diag}(x, y, \bar{x}^{-1})) = \chi_1(x)\chi_2(x\bar{x}^{-1}y),$$

which is well-defined because $x \mapsto x\bar{x}^{-1}$ is a surjective map $F^\times \to F^1$. Every character of T appears in this way: we can recover χ_1 and χ_2 from χ by

$$\chi_1(x) = \chi(\mathrm{diag}(x, \bar{x}/x, \bar{x}^{-1})), \quad \chi_2(y) = \chi(\mathrm{diag}(1, y, 1)).$$

The character χ_2 factors through the determinant and

$$i_B^G(\chi) \simeq i_B^G(\chi_1)(\chi_2 \circ \det),$$

where χ_1 is the character $\chi_1(\mathrm{diag}(x, y, \bar{x}^{-1})) = \chi_1(x)$ of T. Hence the reducibility of $i_B^G(\chi)$ is completely determined by that of $i_B^G(\chi_1)$. The character χ is not regular if $\chi_1(x) = \chi_1(\bar{x})^{-1}$, which occurs if and only if χ_1 is an extension of 1 or ω_{F/F_0} to $F^\times$. An irreducible character χ has level zero if and only if both χ_1 and χ_2 have level zero.

Let ν be the character of T given by $\nu(\mathrm{diag}(x, y, \bar{x}^{-1})) = |x|_F$, i.e., the character with $\chi_1(x) = |x|_F$ and χ_2 trivial, where we normalise $|\cdot|_F$ so that $|\varpi|_F = 1/q$. The modulus character δ_B of B is given on T by $\delta_B = \nu^{-4}$. Because the image of ν is contained in $\overline{\mathbb{Z}}_\ell^\times$, ν and δ_B are integral. If $q^4 \equiv 1 \bmod \ell$ then δ_B is trivial.

6A. ***Hecke Algebras.*** To find the characters χ such that the induced representation $i_B^G(\chi)$ is reducible we study the algebras $\mathcal{H}(G, \lambda)$.

Theorem 6.1. *Suppose λ_T is a character of T^0. Let (J, λ) be a G-cover of (T^0, λ_T) as constructed in Lemma 5.2.*

(1) *If λ_T is regular then $\mathcal{H}(G, \lambda) \simeq R[X^{\pm 1}]$.*

(2) *If λ_T is not regular then $\mathcal{H}(G, \lambda)$ is a two-dimensional algebra generated as an R-algebra by f_{w_x} and f_{w_y} and the relations*

$$f_{w_x} \star f_{w_x} = (q^a - 1) f_{w_x} + q^a,$$
$$f_{w_y} \star f_{w_y} = (q - 1) f_{w_y} + q,$$

where $a = 3$ and $f_{w_x}(1) = f_{w_y}(1) = 1$ if λ_T is trivial on T^1 and factors through the determinant, and $a = 1$, $f_{w_x}(1) = 1/q$ and $f_{w_y}(1) = 1$ if not.

Proof. If $g \in I_G(\lambda)$ then $\mathrm{I}_g(\lambda) \simeq R$, because χ is a character. For $g \in I_G(\lambda)$, $r \in R$, we let $f_{g,r}$ denote the unique function supported on JgJ with $f_{g,r}(g) = r$. If λ_T is regular then the support of $\mathcal{H}(G, \lambda)$ is $JTJ = \bigcup_{n \in \mathbb{Z}} J\zeta^n J$ and, since each intertwining space is one-dimensional and $f_{\zeta^n,1}$ has support $J\zeta^n J$, we have an isomorphism $\mathcal{H}(G, \lambda) \simeq R[X^{\pm 1}]$ defined by $f_{\zeta,1} \mapsto X$.

Suppose $w_x \in I_G(\lambda)$. By Lemma 5.1, w_x intertwines λ if and only if w_y intertwines λ. The support of the Hecke algebra is contained in the intertwining of $\eta = \mathrm{Res}^J_{J^1}(\kappa)$, which is JG_EJ. By the semisimple intersection property [Stevens 2008, Lemma 2.6] and the Bruhat decomposition we have $JG_EJ = \bigcup_{w \in \widetilde{W}} JwJ$. As in the proof of Lemma 5.2 we have $Jw_xJw_yJ = Jw_xw_yJ$ and, similarly,

$Jw_yJw_xJ = Jw_yw_xJ$. Hence, as the intertwining spaces are one-dimensional, the support of $f_{w_x} \star f_{w_y} \star f_{w_x} \star \cdots \star f_{w_i}$ is $Jw_xw_yw_x \cdots w_iJ$. Thus, as $\widetilde{W}$ is an infinite dihedral group generated by w_x and w_y, $\mathcal{H}(G, \lambda)$ is generated by $f_{w_x,1}$ and $f_{w_y,1}$ and the quadratic relations $f_{w_x,1} \star f_{w_x,1}$ and $f_{w_y,1} \star f_{w_y,1}$. Let Λ^x and Λ^y be $\mathfrak{o}_E$-lattice sequences such that the parahoric subgroups $\mathrm{P}(\Lambda^x_E) = \mathrm{P}(\Lambda_E) \cup \mathrm{P}(\Lambda_E)w_x\,\mathrm{P}(\Lambda_E)$ and $\mathrm{P}(\Lambda^y_E) = \mathrm{P}(\Lambda_E) \cup \mathrm{P}(\Lambda_E)w_y\,\mathrm{P}(\Lambda_E)$. The parahoric subgroups $\mathrm{P}(\Lambda^x_E)$ and $\mathrm{P}(\Lambda^y_E)$ are nonconjugate, maximal and contain $\mathrm{P}(\Lambda_E)$. Let κ_x and κ_y be the β-extensions, compatible with κ, defined by Lemma 4.6 related to the skew semisimple strata $[\Lambda^x, n, 0, \beta]$ and $[\Lambda^y, n, 0, \beta]$.

For $z \in \{x, y\}$, let $\hat{\kappa}_z = \mathrm{Res}^J_{J^1(\beta,\Lambda^i)\,\mathrm{P}(\Lambda_E)}(\kappa_z)$. We have a support-preserving isomorphism

$$\mathcal{H}(G, \kappa \otimes \sigma) \simeq \mathcal{H}(G, \hat{\kappa}_z \otimes \sigma)$$

by Lemma 4.6 and transitivity of compact induction. We have a support-preserving injection of algebras

$$\mathcal{H}(\mathrm{P}(\Lambda^z_E), \sigma) \to \mathcal{H}(\mathrm{P}(\Lambda^z), \hat{\kappa}_z \otimes \sigma)$$

defined by $\Phi \mapsto \hat{\kappa}_z \otimes \Phi$, where σ is considered as a character of $\mathrm{P}(\Lambda_E)$ trivial on $\mathrm{P}_1(\Lambda_E)$.

Let B_z be the standard Borel subgroup of $\mathrm{M}(\Lambda^z_E)$. In the ℓ-adic case, by [Howlett and Lehrer 1980, Theorem 4.14], if $i_{B_z}^{\mathrm{M}(\Lambda^z_E)}(\bar{\sigma}) = \rho^z_1 \oplus \rho^z_2$ with $\dim(\rho^z_1) \geqslant \dim(\rho^z_2)$ then $\mathcal{H}(\mathrm{M}(\Lambda^z_E), \bar{\sigma})$ is generated by T^z_w, which is supported on the double coset $B_zw_xB_z$ and satisfies the quadratic relation

$$T^z_w \star T^z_w = (d_z - 1)T^z_w + d_zT^z_1,$$

where $d_z = \dim(\rho^z_1)/\dim(\rho^z_2)$ and T^z_1 is the identity of $\mathcal{H}(\mathrm{M}(\Lambda^z_E), \bar{\sigma})$. By Section 3, $d_y = q$ and

$$d_x = \begin{cases} q^3 & \text{if } \lambda_T \text{ is trivial on } T^1 \text{ and factors through the determinant,} \\ q & \text{otherwise.} \end{cases}$$

In the ℓ-modular case, we choose a lift $\hat{\bar{\sigma}}$ of $\bar{\sigma}$ such that $\hat{\bar{\sigma}}^{w_x} = \hat{\bar{\sigma}}$. Let L be a lattice in $\hat{\bar{\sigma}}$. Recall that $\hat{\bar{\sigma}}$ is called a reduction stable of $\bar{\sigma}$ if $\mathcal{H}(\mathrm{M}(\Lambda^z_E), \bar{\sigma}) = \overline{\mathbb{Z}}_\ell \otimes_{\overline{\mathbb{F}}_\ell} \mathcal{H}(\mathrm{M}(\Lambda^z_E), L)$ and $\mathcal{H}(\mathrm{M}(\Lambda^z_E), \hat{\bar{\sigma}}) = \overline{\mathbb{Q}}_\ell \otimes_{\overline{\mathbb{F}}_\ell} \mathcal{H}(\mathrm{M}(\Lambda^z_E), L)$. A basis of $\mathcal{H}(\mathrm{M}(\Lambda^z_E), \hat{\bar{\sigma}})$ is called reduction stable if it is a basis of $\mathcal{H}(\mathrm{M}(\Lambda^z_E), L)$ and $\hat{\bar{\sigma}}$ is reduction stable. By [Geck et al. 1996, Section 3.1], $\hat{\bar{\sigma}}$ is reduction stable and a basis of $\mathcal{H}(\mathrm{M}(\Lambda^z_E), \hat{\bar{\sigma}})$ is reduction stable. Hence we obtain a basis of $\mathcal{H}(\mathrm{M}(\Lambda^z_E), \bar{\sigma})$ satisfying the quadratic relations required by reduction modulo ℓ.

By inflation, T^z_w determines an element $f_{w_z,r_z} \in \mathcal{H}(\mathrm{P}(\Lambda^z_E), \sigma)$ supported on Jw_zJ. Furthermore, $f_{w_x,1} \star f_{w_x,1}(1_G) = [J : J \cap w_xJw_x] = q^3$ and $f_{w_y,1} \star f_{w_y,1}(1_G) =$

$[J : J \cap w_y J w_y] = q$ in all cases; hence $r_x = r_y = 1$ if λ_T is trivial on T^1 and factors through the determinant, and $r_x = 1/q$ and $r_y = 1$ otherwise. □

6B. ***Reducibility points.*** Suppose $i_B^G(\chi)$ is reducible and let $\lambda_T = \mathrm{Res}^T_{T^0}(\chi)$. By Theorem 6.1, λ_T is not regular. Let (J, λ) be a G-cover of (T^0, λ_T) as constructed in Lemma 5.2 with $\lambda = \kappa \otimes \sigma$. If π is an irreducible quotient of $\mathrm{I}_\kappa(\sigma)$ and an irreducible quotient of $i_B^G(\chi)$ then, by exactness of the Jacquet functor, $r_B^G(\pi)$ is one-dimensional. Hence, as $(j_B)^*(M_\lambda(\pi)) \simeq M_{\lambda_T}(r_B^G(\pi))$ by Theorem 4.2, π must correspond to a character of $\mathcal{H}(G, \lambda)$ under the bijection of Theorem 4.1. The characters of $\mathcal{H}(G, \lambda)$ are determined by their values on the generators $f_{w_x,b}$ and $f_{w_y,1}$, where we let $b = 1$ if λ_T is trivial on T^1 and factors through the determinant and $b = 1/q$ otherwise. Let a be given by Theorem 6.1. The characters of $\mathcal{H}(G, \lambda)$ are summarised as follows:

Character of $\mathcal{H}_R(G, \lambda)$	Value on $f_{w_x,b}$	Value on $f_{w_y,1}$
Ξ_{sgn}	-1	-1
Ξ_{ind}	q^a	q
Ξ_1	q^a	-1
Ξ_2	-1	q

If $q^a \neq -1 \bmod \ell$, these characters are distinct; if $q^a = -1 \bmod \ell$ but $q \neq -1 \bmod \ell$, there are two characters, $\Xi_{\mathrm{sgn}} = \Xi_1$ and $\Xi_{\mathrm{ind}} = \Xi_2$; if $q = -1 \bmod \ell$, there is a unique character $\Xi_{\mathrm{sgn}} = \Xi_1 = \Xi_{\mathrm{ind}} = \Xi_2$.

To calculate the values of χ where this reducibility occurs we study the restriction of the characters of $\mathcal{H}_R(G, \lambda)$ to $\mathcal{H}(T, \lambda_T)$ under $(j_B)^*$. The injection $j_B : \mathcal{H}(T, \lambda_T) \to \mathcal{H}(G, \lambda)$ is induced by taking the unique function $f^T_{\zeta,1} \in \mathcal{H}(T, \lambda_T)$ with support $J_T\zeta$ and $f^T_{\zeta,1}(\zeta) = 1$ to $f_{\zeta,1}$, the unique function in $\mathcal{H}(G, \lambda)$ with support $J\zeta J$ and $f_{\zeta,1}(\zeta) = 1$. Moreover, we know that $f_{\zeta,1} = \varepsilon f_{w_x,1} \star f_{w_y,1}$ for some scalar $\varepsilon \in R$. It is determining the sign of this scalar which requires work. The normalised restriction map $(j_B)^*$ is then induced by this injection and twisting by ν^{-2}. To find ε we compare the value of the characters of $\mathcal{H}(G, \lambda)$ on $f_{w_x,1} \star f_{w_y,1}$ twisted by $\varepsilon\nu^{-2}(\zeta)$ to known reducibility points.

Character χ of	$\varepsilon\nu^{-2}(\zeta)\chi(f_{w_x,1} \star f_{w_y,1})$	$\varepsilon\nu^{-2}(\zeta)\chi(f_{w_x,1} \star f_{w_y,1})$
$\mathcal{H}_R(G, \lambda)$	$a = 3,\ b = 1$	$a = 1,\ b = 1/q$
Ξ_{sgn}	$q^{-2}\varepsilon$	$q^{-1}\varepsilon$
Ξ_{ind}	$q^2\varepsilon$	$q\varepsilon$
Ξ_1	$-q\varepsilon$	$-\varepsilon$
Ξ_2	$-q^{-1}\varepsilon$	$-\varepsilon$

First consider the case when $a=3$ and $b=1/q$. As the trivial representation is an irreducible subquotient of $i_B^G(\nu^{\pm 2})$, the induced representations are reducible. Thus $\nu^{\pm 2}(\zeta)=q^{\pm 2}\in\{q^{-2}\varepsilon, q^2\varepsilon, -q\varepsilon, -q^{-1}\varepsilon\}$ and this multiset of values of the characters must be $\{q^{-2}, q^2, -q, -q^{-1}\}$. Moreover, by compatibility with the ℓ-adic case by reduction modulo ℓ, we must have $\varepsilon=1$ with $\nu^{\pm 2}$ corresponding to Ξ_{sgn} and Ξ_{ind}. The other reducibility points, corresponding to the characters Ξ_1 and Ξ_2 of $\mathcal{H}(G,\lambda)$, are the characters χ of T of the form $\chi=\eta\nu^{\pm 1}$, where η is any extension of ω_{F/F_0} to $F^\times$ which is trivial on F^1 such that $\chi\mid_{T^0}$ factors through the determinant.

Now consider the case when $a=1$ and $b=1/q$. Using an alternative method, Keys [1984] computed the ℓ-adic reducibility points. Comparing the value of a pair of these on ζ — see [Keys 1984, Section 7] — with our values in the table we must have $\varepsilon=-1$ in all other cases. This gives reducibility points the characters χ of T of the form $\chi=\eta\nu^{\pm 1}$, where η is any extension of ω_{F/F_0} to $F^\times$ not trivial on F^1, corresponding to the characters Ξ_{sgn} and Ξ_{ind} of $\mathcal{H}(G,\lambda)$, and the characters χ of T of the form χ_1 is nontrivial, but $\chi_1\mid_{F_0^\times}$ is trivial, corresponding to the characters Ξ_1 and Ξ_2 of $\mathcal{H}(G,\lambda)$.

Theorem 6.2. *Let χ be an irreducible ℓ-modular character of T. Then $i_B^G(\chi)$ is reducible exactly in the following cases*:

(1) $\chi=\nu^{\pm 2}$;

(2) $\chi=\eta\nu^{\pm 1}$, *where η is any extension of ω_{F/F_0} to $F^\times$*;

(3) χ_1 *is nontrivial, but* $\chi_1\mid_{F_0^\times}$ *is trivial.*

6C. ***Parahoric restriction and parabolic induction.*** As the parabolic functors respect the decomposition of $\mathfrak{R}_R(G)$ by level, by [Vignéras 1996, II 5.12], if χ is a level zero character of T (i.e., a character of T trivial on T^1) then all irreducible subquotients of $i_B^G(\chi)$ have level zero.

Lemma 6.3. *Let $w\in\{x,y\}$ and let χ be a level zero character of T. Then* $\mathrm{R}_{\Lambda_w}(i_B^G(\chi))\simeq i_{B_w}^{\mathrm{M}(\Lambda_w)}(\chi)$.

Proof. The proof follows by Mackey theory, as the maximal parahoric subgroups of G satisfy the Iwasawa decomposition. □

Let $[\Lambda,n,0,\beta]$ be a skew semisimple stratum in A. Let $\theta\in\mathcal{C}_-(\Lambda,\beta)$ and κ be a β-extension of the unique Heisenberg representation containing θ. Let χ be an irreducible ℓ-modular character of T which contains the R-type $(J_T,\kappa_T\otimes\sigma)$. Furthermore, suppose that $(J,\kappa\otimes\sigma)$ is a G-cover of $(J_T,\kappa_T\otimes\sigma)$ relative to B, as in Lemma 5.2. Let Λ^m be an $\mathfrak{o}_E$-lattice sequence in V such that $\mathrm{P}(\Lambda_E^m)$ is maximal and $\mathrm{P}(\Lambda_E)\subset\mathrm{P}(\Lambda_E^m)$. Let $\theta_m=\tau_{\Lambda,\Lambda^m,\beta}(\theta)$ and κ_m be the unique β-extension of the unique Heisenberg representation containing θ_m which is compatible with κ, as in Lemma 4.6. Let $B(\Lambda_E^m)$ be the Borel subgroup of $\mathrm{M}(\Lambda_E^m)$ whose preimage under the projection map $\mathrm{P}(\Lambda_E^m) \to \mathrm{M}(\Lambda_E^m)$ is equal to J. Suppose $B(\Lambda_E^m)$ has Levi decomposition $B(\Lambda_E^m) = T(\Lambda_E^m) \ltimes N(\Lambda_E^m)$.

The next theorem is a generalisation of a weakening of Lemma 6.3; precisely, it generalises the isomorphism Lemma 6.3 induces in the Grothendieck group $\mathfrak{Gr}_R(\mathrm{M}(\Lambda_w))$.

Theorem 6.4. *With the notation as above, there is an isomorphism*

$$\big[\mathrm{R}_{\kappa_m}(i_B^G(\chi))\big] \simeq \big[i_{B(\Lambda_E^m)}^{M(\Lambda_E^m)}(\mathrm{R}_{\kappa_T}(\chi))\big].$$

Proof. We prove the corresponding result in the ℓ-adic case first and deduce the ℓ-modular result by reduction modulo ℓ. The proof in the ℓ-adic case follows a similar argument made for $\mathrm{GL}_n(F)$ in [Schneider and Zink 1999]. Let $\Omega_T = [T, \rho]_T$ and $\Omega = [T, \rho]_G$ be inertial equivalence classes. Let $\mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\Omega)$ denote the full subcategory of $\mathrm{R}_{\overline{\mathbb{Q}}_\ell}(G)$ of representations all of whose irreducible subquotients have inertial support in Ω, and $\mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\Omega_T)$ denote the full subcategory of $\mathrm{R}_{\overline{\mathbb{Q}}_\ell}(T)$ of representations all of whose irreducible subquotients have inertial support in Ω_T. Let ω denote the $\mathrm{M}(\Lambda_E^m)$-conjugacy class of σ and ω_T the $T(\Lambda_E^m)$-conjugacy class of σ. Let $\mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\omega)$ be the full subcategory of $\mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\mathrm{M}(\Lambda_E^m))$ of representations all of whose irreducible subquotients have supercuspidal support in ω and $\mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\omega_T)$ be the full subcategory of $\mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(T(\Lambda_E^m))$ of representations all of whose irreducible subquotients have lie in ω_T. Let $M_\omega : \mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\omega) \to \mathcal{M}(\mathrm{M}(\Lambda_E^m), \sigma)$ be defined by $\rho \mapsto \mathrm{Hom}_{B(\Lambda_E^m)}(\sigma, \rho)$ for $\rho \in \mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\omega)$. Similarly, let $M_{\omega_T} : \mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\omega_T) \to \mathcal{M}(T(\Lambda_E^m), \sigma)$ be defined by $\rho \mapsto \mathrm{Hom}_{T(\Lambda_E^m)}(\sigma, \rho)$ for $\rho \in \mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\omega_T)$. We prove that the following diagram commutes.

$$\begin{tikzcd}
 & \mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\omega) \arrow[rr, "\simeq", "M_\omega"'] & & \mathcal{M}(\mathrm{M}(\Lambda_E^m), \sigma) & & \mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\omega) \arrow[ll, "\simeq"', "M_\omega"] & \\
\mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\Omega) \arrow[ur, "\mathrm{R}_{\kappa_m}"] \arrow[rr, "\simeq", "M_{\kappa\otimes\sigma}"'] & & \mathcal{M}(G, \kappa\otimes\sigma) \arrow[ur, "\mathrm{Res}"'] & & \mathcal{M}(T(\Lambda_E^m), \sigma) \arrow[ul, "(j_{B(\Lambda_E^m)})_*"] & & \mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\omega_T) \arrow[ll, "\simeq"', "M_{\omega_T}"] \arrow[ul, "i_{B(\Lambda_E^m)}^{\mathrm{M}(\Lambda_E^m)}"'] \\
 & \mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\Omega_T) \arrow[ul, "i_B^G"] \arrow[rr, "\simeq", "M_{\kappa_T\otimes\sigma}"'] & & \mathcal{M}(T, \kappa_T\otimes\sigma) \arrow[ul, "(j_B)_*"] \arrow[ur, "\mathrm{Res}"] & & \mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\Omega_T) \arrow[ll, "\simeq"', "M_{\kappa_T\otimes\sigma}"] \arrow[ur, "\mathrm{R}_{\kappa_T}"'] &
\end{tikzcd}$$

We have $M_\omega \circ i_{B(\Lambda_E^m)}^{\mathrm{M}(\Lambda_E^m)} \simeq (j_{B(\Lambda_E^m)})_* \circ M_{\omega_T}$ and $M_{\kappa\otimes\sigma} \circ i_B^G \simeq (j_B)_* \circ M_{\kappa_T\otimes\sigma}$ by [Bushnell and Kutzko 1998, Corollary 8.4], and $M_{\kappa\otimes\sigma}$ is an equivalences of categories by [Bushnell and Kutzko 1998, Theorems 4.3 and 8.3].

We have support-preserving injections $\alpha_1 : \mathcal{H}(\mathrm{M}(\Lambda_E^m), \sigma) \to \mathcal{H}(G, \kappa\otimes\sigma)$ and $\alpha_2 : \mathcal{H}(T(\Lambda_E^m), \sigma) \to \mathcal{H}(T, \kappa_T \otimes \sigma)$, as in the proof of Theorem 6.1, hence restriction functors $\mathcal{M}(G, \kappa\otimes\sigma)\to\mathcal{M}(\mathrm{M}(\Lambda_E^m), \sigma)$ and $\mathcal{M}(T, \kappa_T\otimes\sigma)\to\mathcal{M}(T(\Lambda_E^m), \sigma)$, denoted in the diagram by Res. Because $\mathcal{H}(T(\Lambda_E^m), \sigma)$ is one-dimensional and the injections defined are homomorphisms of algebras, we must have $j_B\circ\alpha_1\simeq j_{B(\Lambda_E^m)}\circ\alpha_2$, hence also $\mathrm{Res}\circ(j_B)_*\simeq(j_{B(\Lambda_E^m)})_*\circ\mathrm{Res}$.

We show that $M_\omega\circ\mathrm{R}_{\kappa_m}\simeq\mathrm{Res}\circ M_{\kappa\otimes\sigma}$; a similar argument shows that $M_{\omega_T}\circ\mathrm{R}_{\kappa_T}\simeq\mathrm{Res}\circ M_{\kappa_T\otimes\sigma}$. Let $\pi\in\mathfrak{R}_{\overline{\mathbb{Q}}_\ell}(\Omega)$. By Lemma 4.9 and adjointness, we have

$$\begin{aligned} M_\omega(\mathrm{R}_{\kappa_m}(\pi)) = \mathrm{Hom}_{B(\Lambda_E^m)}(\sigma, \mathrm{R}_{\kappa_m}(\pi)) &= \mathrm{Hom}_{B(\Lambda_E^m)}(\sigma, \mathrm{R}_{\kappa_m}(\pi)) \\ &\simeq \mathrm{Hom}_J(\sigma, (\mathrm{R}_{\kappa_m}(\pi))^{J_m^1\mathrm{P}_1(\Lambda_E)/J_m^1}) \\ &\simeq \mathrm{Hom}_J(\sigma, \mathrm{R}_\kappa(\pi)) \\ &\simeq \mathrm{Hom}_{J^1}(\kappa\otimes\sigma, \pi) = M_{\kappa\otimes\sigma}(\pi). \end{aligned}$$

In the ℓ-modular case, we choose lifts of κ and χ and then by the ℓ-adic isomorphism and reduction modulo ℓ we have $\big[\mathrm{R}_{\kappa_m}(i_B^G(\chi))\big]\simeq\big[i_{B(\Lambda_E^m)}^{M(\Lambda_E^m)}(\mathrm{R}_{\kappa_T}(\chi))\big]$. □

6D. ***Parabolic induction, κ-restriction, and covers.*** Let χ be an irreducible character of T. Let (T^0, λ_T) be an R-type contained in χ such that (J, λ) is a G-cover of (T^0, λ_T) relative to B as constructed in Lemma 5.2 with $\lambda=\kappa\otimes\sigma$ and $\lambda_T=\kappa_T\otimes\sigma$, where $\kappa_T=\mathrm{Res}_{T^0}^J(\kappa)$. Hence $J=\mathrm{P}(\Lambda_E)J^1$ with $\mathrm{P}(\Lambda_E)$ a nonmaximal parahoric subgroup of G_E corresponding to the $\mathfrak{o}_E$-lattice sequence Λ_E. In all cases, there are two nonconjugate maximal parahoric which contain $\mathrm{P}(\Lambda_E)$; we denote the $\mathfrak{o}_E$-lattice sequences that correspond to these by Λ_E^x and Λ_E^y. Let $m\in\{x, y\}$ and let (κ_m, Λ_E^m) be the unique pair compatible with (κ, Λ_E) as in Lemma 4.6.

Lemma 6.5. *Let π be an irreducible subrepresentation or quotient of $i_B^G(\chi)$ and $m\in\{x, y\}$. Then* $\mathrm{R}_{\kappa_m}(\pi)\neq 0$.

Proof. By the geometric lemma, $r_B^G(i_B^G(\chi))$ is filtered by χ and $\chi^{w_x}=\psi\chi$ for some unramified character ψ. Hence, by exactness of the Jacquet functor, $r_B^G(\pi)=\psi\chi$. By Theorem 4.2, π contains (J, λ) if and only if $r_B^G(\pi)$ contains (T^0, λ_T). Thus π contains (J, λ); hence $\mathrm{R}_\kappa(\pi)\neq 0$. Therefore $\mathrm{R}_{\kappa_m}(\pi)\neq 0$. □

The next lemma is crucial in our proof of unicity of supercuspidal support. It shows that parabolic induction preserves the semisimple character up to transfer.

Lemma 6.6. *Suppose that $i_B^G(\chi)$ has an irreducible cuspidal subquotient π. Then there exists $m\in\{x, y\}$ such that* $\mathrm{R}_{\kappa_m}(\pi)\neq 0$.

Proof. By Theorem 5.3 there exist a skew semisimple stratum $[\Lambda', n', 0, \beta']$ such that $\mathrm{P}(\Lambda'_{E'})$ is a maximal parahoric subgroup of $G_{E'}$, where $G_{E'}$ denotes the G-centraliser of β', a semisimple character $\theta'\in\mathscr{C}_-(\Lambda', \beta')$, a β'-extension κ' to $J'=J(\Lambda', \beta')$ of the unique Heisenberg representation η' containing θ' and a cuspidal representation $\sigma'\in\mathrm{Irr}(J'/(J')^1)$ such that $\pi\simeq\mathrm{I}_{\kappa'}(\sigma')$.

As π contains $\kappa' \otimes \sigma'$, the restriction of $i_B^G(\chi)$ to J' has $\kappa' \otimes \sigma'$ as a subquotient. We choose $\hat{\chi}$ an ℓ-adic character lifting χ such that $i_B^G(\hat{\chi})$ is reducible. Then, because restriction and parabolic induction commute with reduction modulo ℓ, the restriction of $i_B^G(\hat{\chi})$ to J' has an irreducible subquotient δ such that $r_\ell(\delta)$ contains $\kappa' \otimes \sigma'$. On restricting to $(J')^1$ we see that δ contains the unique lift $\hat{\eta}'$ of η' and, since δ is irreducible and J' normalises $\hat{\eta}'$, $\mathrm{Res}^{J'}_{(J')^1}(\delta)$ is a multiple of η'. Thus $\delta = \hat{\kappa}' \otimes \xi$, with $\hat{\kappa}'$ a lift of κ' and ξ an irreducible representation of $J'/(J')^1$ whose reduction modulo ℓ contains σ. However, ξ cannot be cuspidal, otherwise $i_B^G(\hat{\chi})$ would have a cuspidal subquotient $\mathrm{I}_{\hat{\kappa}'}(\xi)$. Hence $G_{E'}$ is not compact. Therefore $[\Lambda', n', 0, \beta']$ is either a scalar skew simple stratum or a skew semisimple stratum with splitting $V = V_1' \oplus V_2'$, with V_1' one-dimensional and V_2' two-dimensional hyperbolic. (Note that, as σ is cuspidal nonsupercuspidal, we must have $\ell \mid q+1$ or $\ell \mid q^2 - q + 1$ by Section 3.)

We continue by induction on the level $l(\pi)$ of π.

The base step is when π has level zero. If π has level zero then, as all subquotients of $i_B^G(\chi)$ have the same level as χ by [Vignéras 1996, 5.12], χ and $i_B^G(\chi)$ have level zero. Thus we can choose, and assume that we have chosen, κ', κ and κ_T to be trivial. By conjugating, we may assume $\Lambda' = \Lambda_m$ for some $m \in \{x, y\}$ and then $\kappa_m = \kappa'$ is trivial and $\mathrm{R}_{\kappa_m}(\pi) = \mathrm{R}_{\Lambda_m}(\pi) \neq 0$.

Suppose first that $[\Lambda, n, n-1, \beta]$ is equivalent to a scalar stratum $[\Lambda, n, n-1, \gamma]$. The stratum $[\Lambda, n, n-1, \gamma]$ corresponds to a character ψ_γ of $P_n(\Lambda)$ which extends to a character $\phi \circ \det$ of G. Twisting by $\phi^{-1} \circ \det$ we reduce the level of π and the level of $i_B^G(\chi)$. The stratum $[\Lambda, n, n-1, \beta - \gamma]$ is equivalent to a semisimple stratum $[\Lambda, n, n-1, \alpha]$ and the representations $\kappa(\phi^{-1} \circ \det)$, $\kappa_T(\phi^{-1} \circ \det)$ and $\kappa_m(\phi^{-1} \circ \det)$ for $m \in \{x, y\}$ are α-extensions defined on the relevant groups. Similarly, the stratum $[\Lambda', n', n'-1, \beta' - \gamma]$ is equivalent to a semisimple stratum $[\Lambda, n', n'-1, \alpha']$ and $\kappa'(\phi^{-1} \circ \det)$ is an α'-extension. Moreover, $\kappa_m(\phi^{-1} \circ \det)$ is compatible with $\kappa(\phi^{-1} \circ \det)$ for $m \in \{x, y\}$, $(\kappa \otimes \sigma)(\phi^{-1} \circ \det)$ is a G-cover of $(\kappa_T \otimes \sigma)(\phi^{-1} \circ \det)$ relative to B, $(\kappa_T \otimes \sigma)(\phi^{-1} \circ \det)$ is contained in $\chi(\phi^{-1} \circ \det)$, and $(\kappa' \otimes \sigma')(\phi^{-1} \circ \det)$ is contained in $\pi(\phi^{-1} \circ \det)$. Thus, by induction, we have

$$\mathrm{R}_{\kappa_m}(\pi) \simeq \mathrm{R}_{\kappa_m(\phi^{-1} \circ \det)}(\pi(\phi^{-1} \circ \det))$$

is nonzero for some $m \in \{x, y\}$.

Secondly, suppose that $[\Lambda', n', n' - 1, \beta']$ is equivalent to a scalar stratum $[\Lambda', n', n' - 1, \gamma']$. As in the last case, we can twist by a character to reduce the level.

Hence we may assume that both $[\Lambda, n, n-1, \beta]$ and $[\Lambda', n', n'-1, \beta']$ are not equivalent to scalar simple strata. This forces $[\Lambda, n, 0, \beta]$ (resp. $[\Lambda', n', 0, \beta']$) to be semisimple — and nonsimple — with splitting $V = V_1 \oplus V_2$ (resp. $V = V_1' \oplus V_2'$) with V_1 (resp. V_1') one-dimensional and V_2 (resp. V_2') two-dimensional hyperbolic.

Thus, by conjugation we may assume that the splitting of $[\Lambda', n', 0, \beta']$ is the same as the splitting of $[\Lambda, n, 0, \beta]$, i.e., $V_1' = V_1$ and $V_2' = V_2$. We have $E = E'$ and $G_E = G_{E'}$, and conjugating further we may assume that Λ_E' and Λ_E lie in the closure of the same chamber of the building of G_E. Moreover, Λ_E' is a vertex and Λ_E is the barycentre of the chamber.

Let

$$w = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

We have

$$J(\beta', \Lambda) = w \left(\begin{array}{c|c} \mathfrak{A}_0(\Lambda)^{11} & \mathfrak{A}_{\lfloor \frac{r'+1}{2} \rfloor}(\Lambda)^{12} \\ \hline \mathfrak{A}_{\lfloor \frac{r'+1}{2} \rfloor}(\Lambda)^{12} & \mathfrak{A}_0(\Lambda)^{22} \end{array} \right) w \cap G,$$

and

$$J(\beta, \Lambda) = w \left(\begin{array}{c|c} \mathfrak{A}_0(\Lambda)^{11} & \mathfrak{A}_{\lfloor \frac{r+1}{2} \rfloor}(\Lambda)^{12} \\ \hline \mathfrak{A}_{\lfloor \frac{r+1}{2} \rfloor}(\Lambda)^{12} & \mathfrak{A}_0(\Lambda)^{22} \end{array} \right) w \cap G,$$

where r' (resp. r) is minimal such that $[\Lambda, n', r', \beta]$ (resp. $[\Lambda, n, r, \beta]$) is equivalent to a scalar stratum. Thus, as we are now assuming that $[\Lambda, n, n-1, \beta]$ and $[\Lambda', n', n'-1, \beta']$ are not equivalent to scalar simple strata, we have $r' = n'$ and $r = n$. Furthermore, we have $l(\chi) = l(\pi)$, i.e., $n'/e(\Lambda') = n/e(\Lambda)$. We let κ'' be the unique β-extension to $J(\beta', \Lambda)$ compatible with κ' relative to a semisimple stratum $[\Lambda, n, 0, \beta']$. Therefore, $J(\beta, \Lambda) = J(\beta', \Lambda)$. Similar considerations yield $H(\beta, \Lambda) = H(\beta', \Lambda)$ and $J(\beta, \Lambda') = J(\beta', \Lambda')$.

As ξ is not cuspidal, it is a direct factor of $i_{B(\Lambda_E')}^{M(\Lambda_E')}(\hat{\tau}')$, where we choose $B(\Lambda_E')$ to be the image of $P(\Lambda_E)$ in $M(\Lambda_E')$, for some representation $\hat{\tau}'$ of $T(\Lambda_E')$. Furthermore, $i_B^G(\hat{\chi})$ contains $\hat{\kappa}'' \otimes \hat{\tau}'$ with $\hat{\kappa}''$ a lift of κ'', by Lemma 4.6 and transitivity of induction. By Lemma 5.2, $(J, \hat{\kappa}'' \otimes \hat{\tau}')$ is a G-cover of $(T^0, \hat{\kappa}_T'' \otimes \hat{\tau}')$ relative to B, where $\hat{\kappa}_T'' = \mathrm{Res}_{T^0}^J(\hat{\kappa}'')$. By [Blondel 2005, Theorem 2], $\mathrm{ind}_J^G(\hat{\kappa}'' \otimes \hat{\tau}') \simeq \mathrm{Ind}_{B^{\mathrm{op}}}^G(\mathrm{ind}_{T_0}^T(\hat{\kappa}_T'' \otimes \hat{\tau}'))$. By second adjunction of parabolic induction and parabolic restriction for ℓ-adic representations, and right adjunction of restriction with compact induction we have

$$\mathrm{Hom}_{T^0}(\hat{\kappa}_T'' \otimes \hat{\tau}', r_B^G \circ i_B^G(\hat{\chi})) \simeq \mathrm{Hom}_G(\mathrm{ind}_J^G(\hat{\kappa}'' \otimes \hat{\tau}'), i_B^G(\hat{\chi})) \neq 0.$$

We have $[r_B^G \circ i_B^G(\hat{\chi}) \mid_{T^0}] = \hat{\chi} \oplus \hat{\chi}^{w_x} \mid_{T^0} = \hat{\chi} \oplus \hat{\chi} \mid_{T^0}$. Hence $\hat{\kappa}_T'' \otimes \hat{\tau}' = \mathrm{Res}_{T^0}^T(\hat{\chi})$. Similarly if we let $\hat{\kappa}$ be a lift of κ, $\hat{\sigma}$ be a lift of σ, and $\hat{\kappa}_T = \mathrm{Res}_{T^0}^T(\hat{\kappa})$, then we have $\hat{\kappa}_T \otimes \hat{\sigma} = \mathrm{Res}_{T^0}^J(\hat{\chi})$. This implies that we have an equality of semisimple characters $\tau_{\Lambda', \Lambda, \beta'}(\hat{\theta}') = \hat{\theta}$, where $\hat{\theta}' \in \mathscr{C}_-(\beta', \Lambda')$ is contained in $\hat{\kappa}'$ and $\hat{\theta} \in \mathscr{C}(\beta, \Lambda)$ is contained in $\hat{\kappa}$.

We let $\widetilde{H}(\beta, \Lambda)$ (resp. $\widetilde{H}(\beta', \Lambda')$) denote the compact open subgroup of $\mathrm{GL}_3(F)$ defined in [Stevens 2008], which defines $H(\beta, \Lambda)$ (resp. $H(\beta', \Lambda')$) by intersecting with $\mathrm{U}(2, 1)(F/F_0)$. The Iwahori decomposition for $\widetilde{H}^1(\beta', \Lambda')$ gives $\widetilde{H}^1(\beta', \Lambda') = \widetilde{H}^1(\beta', \Lambda')^-(\widetilde{H}^1(\beta', \Lambda') \cap \widetilde{M})\widetilde{H}^1(\beta', \Lambda')^+$ where $\widetilde{H}^1(\beta', \Lambda')^-$ denotes the lower triangular unipotent matrices in $\widetilde{H}^1(\beta', \Lambda')$, $\widetilde{H}^1(\beta', \Lambda')^+$ denotes the upper triangular unipotent matrices in $\widetilde{H}^1(\beta', \Lambda')$, and $\widetilde{M}$ the subgroup of diagonal matrices. As $\widetilde{H}^1(\beta', \Lambda)$ contains $(\widetilde{H}^1(\beta', \Lambda') \cap \widetilde{M})$ and is contained in $\widetilde{H}^1(\beta', \Lambda')$, we have

$$\widetilde{H}^1(\beta', \Lambda') = \widetilde{H}^1(\beta', \Lambda')^-\big(\widetilde{H}^1(\beta', \Lambda') \cap \widetilde{H}^1(\beta', \Lambda)\big)\widetilde{H}^1(\beta', \Lambda')^+.$$

Thus a character of $\widetilde{H}^1(\beta', \Lambda')$ is determined by its values on $\widetilde{H}^1(\beta', \Lambda')^-$, $\widetilde{H}^1(\beta', \Lambda') \cap \widetilde{H}^1(\beta', \Lambda)$, and $\widetilde{H}^1(\beta', \Lambda')^+$.

The semisimple characters $\hat{\theta}$ and $\hat{\theta}'$ are equal to the restriction of semisimple characters $\tilde{\theta}$ and $\tilde{\theta}'$ of $\mathrm{GL}_3(F)$. Moreover, $\tau_{\Lambda',\Lambda,\beta'}(\tilde{\theta}') = \tilde{\theta}$ as $\tau_{\Lambda',\Lambda,\beta'}(\hat{\theta}') = \hat{\theta}$. It follows from the decomposition of $\widetilde{H}^1(\beta', \Lambda')$ given above that $\tau_{\Lambda,\Lambda',\beta}(\tilde{\theta}) = \tilde{\theta}'$; they are both trivial on $\widetilde{H}^1(\beta', \Lambda')^-$ and $\widetilde{H}^1(\beta', \Lambda')^+$, and as $\theta' = \tau_{\Lambda,\Lambda',\beta'}(\theta)$ they both agree with $\tilde{\theta}$ on $\widetilde{H}^1(\beta, \Lambda') \cap \widetilde{H}^1(\beta, \Lambda) = \widetilde{H}^1(\beta', \Lambda') \cap \widetilde{H}^1(\beta', \Lambda)$. Hence, $\tau_{\Lambda,\Lambda',\beta}(\theta) = \theta'$ by restriction and reduction modulo ℓ. As there is a unique Heisenberg representation containing θ', we have $\mathrm{R}_{\kappa_m}(\pi) \neq 0$ for some $m \in \{x, y\}$. □

Lemma 6.7. *Suppose that $i_B^G(\chi)$ is reducible with irreducible subrepresentation π_1 and quotient $\pi_2 = i_B^G(\chi)/\pi_1$. If Σ is a maximal cuspidal subquotient of $\mathrm{R}_{\kappa_m}(i_B^G(\chi))$, i.e., all subquotients of $\mathrm{R}_{\kappa_m}(i_B^G(\chi))$ not contained in Σ are not cuspidal, then $\mathrm{I}_{\kappa_m}(\Sigma)$ is a subrepresentation of π_2.*

Proof. Let Σ be a maximal cuspidal subquotient of $\mathrm{R}_{\kappa_m}(i_B^G(\chi))$. By Lemma 6.5, $\mathrm{R}_{\kappa_m}(\pi_1)$ and $\mathrm{R}_{\kappa_m}(\pi_2)$ are nonzero and must contain noncuspidal subquotients as π_1 and π_2 are not cuspidal. However, by Theorem 6.4 and Section 3, there are only two noncuspidal subquotients of $\mathrm{R}_{\kappa_m}(i_B^G(\chi))$. Thus each of $\mathrm{R}_{\kappa_m}(\pi_1)$ and $\mathrm{R}_{\kappa_m}(\pi_2)$ must have a single noncuspidal irreducible subquotient, say ρ_1 and ρ_2 respectively.

If $\mathrm{R}_{\kappa_m}(\pi_1) \neq \rho_1$ then $\mathrm{R}_{\kappa_m}(\pi_1)$ has an irreducible cuspidal subrepresentation or an irreducible cuspidal quotient. If $\mathrm{R}_{\kappa_m}(\pi_1)$ has an irreducible cuspidal subrepresentation σ then, by adjointness of R_{κ_m} and I_{κ_m}, $\mathrm{I}_{\kappa_m}(\sigma)$ is an irreducible cuspidal subrepresentation of π_1, contradicting the irreducibility and noncuspidality of π_1. If $\mathrm{R}_{\kappa_m}(\pi_1)$ has an irreducible cuspidal quotient σ then $\mathrm{R}_{\tilde{\kappa}_m}(\tilde{\pi}_1)$ has a cuspidal subrepresentation $\tilde{\sigma}$ by Lemma 4.10. Thus $\mathrm{I}_{\tilde{\kappa}_m}(\tilde{\sigma})$ is an irreducible cuspidal subrepresentation of $\tilde{\pi}_1$ by adjointness. Hence $\mathrm{I}_{\kappa_m}(\sigma)$ is an irreducible cuspidal quotient of π_1 by Lemma 4.10, contradicting the irreducibility and noncuspidality of π_1. Thus $\mathrm{R}_{\kappa_m}(\pi_1) = \rho_1$.

Similarly, if $\mathrm{R}_{\kappa_m}(\pi_2)$ has an irreducible cuspidal quotient σ, then $\mathrm{I}_{\kappa_m}(\sigma)$ is an irreducible cuspidal quotient of π_2. Hence $\mathrm{I}_{\kappa_m}(\sigma)$ is a quotient of $i_B^G(\chi)$, contradicting the cuspidality of $\mathrm{I}_{\kappa_m}(\sigma)$. Hence $\mathrm{R}_{\kappa_m}(\pi_2)$ can have no cuspidal quotients. Hence, by Section 3, Lemma 6.3 and Theorem 6.4, Σ is a subrepresentation of $\mathrm{R}_{\kappa_m}(\pi_2)$. Note that, as Theorem 6.4 only gives us an isomorphism in the Grothendieck group of finite-length representations of $\mathrm{M}(\Lambda_E)$, we have used that Σ is irreducible by Section 3 in the skew semisimple nonscalar case to imply it is a subrepresentation of $\mathrm{R}_{\kappa_m}(\pi_2)$; in all other cases we twist by a character (if necessary) and use Lemma 6.3. By reciprocity, $\mathrm{I}_{\kappa_m}(\Sigma)$ is a subrepresentation of π_2. □

By [Blondel 2005, Theorem 2] and Lemma 5.2, $\mathrm{I}_\kappa(\sigma) \simeq \mathrm{Ind}_{B^{\mathrm{op}}}^G(\mathrm{ind}_{T^0}^T(\kappa_T \otimes \sigma))$. By second adjunction (cf. [Dat 2009, Corollaire 3.9]),

$$\mathrm{Hom}_G(\mathrm{Ind}_{B^{\mathrm{op}}}^G(\mathrm{ind}_{T^0}^T(\kappa_T \otimes \sigma)), \pi) \simeq \mathrm{Hom}_T(\mathrm{ind}_{T^0}^T(\kappa_T \otimes \sigma), r_B^G(\pi)).$$

By Clifford theory, the irreducible quotients of $\mathrm{ind}_{T^0}^T(\kappa_T \otimes \sigma)$ are all the twists of χ by an unramified character. Hence, π is an irreducible quotient of $\mathrm{I}_\kappa(\sigma)$ if and only if it is an irreducible quotient of $i_B^G(\chi\psi)$ for some unramified character ψ of T.

The R-type (J, λ) is quasiprojective by Theorem 5.3; hence a simple module of $\mathscr{H}(G, \lambda)$ corresponds to an irreducible quotient of $i_B^G(\chi\psi)$ for some unramified character ψ, by the bijection of Theorem 4.1. If $i_B^G(\chi\psi)$ is reducible with proper quotient π, then the Jacquet module of π is one-dimensional by the geometric lemma. Hence, by Theorem 4.2, π must correspond to a character of $\mathscr{H}(G, \lambda)$ under the bijection of Theorem 4.1 and all characters of $\mathscr{H}(G, \lambda)$ must correspond to a proper quotient of a reducible principal series representation $i_B^G(\chi\psi)$ with ψ an unramified character of T.

Lemma 6.8. *Suppose $\ell \neq 2$ and $\ell \mid q-1$. Then $i_B^G(\chi)$ is semisimple.*

Proof. If $i_B^G(\chi)$ is irreducible then it is semisimple, so suppose $i_B^G(\chi)$ is reducible. If $i_B^G(\chi)$ has a cuspidal subquotient it is of the form $\mathrm{I}_{\kappa_m}(\sigma)$ for $m \in \{x, y\}$ and σ an irreducible cuspidal representation of $\mathrm{M}(\Lambda_E^x)$ by Theorem 5.3. By Theorem 6.4, $\mathrm{R}_{\kappa_x}(i_B^G(\chi)) = i_{B(\Lambda_E^x)}^{\mathrm{M}(\Lambda_E^x)}(\mathrm{R}_{\kappa_T}(\chi))$, and $\mathrm{R}_{\kappa_x}(\mathrm{I}_{\kappa_x}(\sigma)) = \sigma$, by Theorem 4.8 and Lemma 4.4. Hence, by exactness, σ is a cuspidal subquotient of $i_{B(\Lambda_E^x)}^{\mathrm{M}(\Lambda_E^x)}(\mathrm{R}_{\kappa_T}(\chi))$. However, by Section 3, when $\ell \mid q-1$ no such cuspidal subquotients exist; hence $i_B^G(\chi)$ has no cuspidal subquotients. Thus, by exactness of the Jacquet functor and the geometric lemma, $i_B^G(\chi)$ has length two. When $\ell \neq 2$ and $\ell \mid q-1$ there are four characters of $\mathscr{H}(G, \lambda)$, yet only two reducibility points. Hence these two reducible principal series representations must both have two nonisomorphic irreducible quotients and must be semisimple. □

Remark 6.9. If $\chi^{-1} = \chi$ then $i_B^G(\chi)$ is self-contragredient and there is a simple proof of Lemma 6.8 using the contragredient representation and avoiding the use of covers or second adjunction.

Lemma 6.10. *Let $\ell \mid q+1$. Then the unique irreducible quotient of $i_B^G(\chi)$ is isomorphic to the unique irreducible subrepresentation.*

Proof. Let π denote the unique irreducible quotient of $i_B^G(\chi)$. When $\ell \mid q+1$ there is only one character of $\mathcal{H}(G, \kappa \otimes \sigma)$. Hence π corresponds to the unique character of $\mathcal{H}(G, \lambda)$. Hence, if $\mathcal{V}$ is the space of π, $\mathrm{R}_\kappa(\mathcal{V})$ is one-dimensional and the action of J is given by σ. As δ_B is trivial, the contragredient commutes with parabolic induction: we have $(i_B^G(\chi))^\sim \simeq i_B^G(\tilde{\chi})$. Furthermore, $\tilde{\chi} = \chi^{-1}$, where χ^{-1} is the character defined by, for all $x \in F^\times$, $\chi^{-1}(x) = \chi(x^{-1})$. The character χ^{-1} is not regular and similar arguments, given for $i_B^G(\chi)$, apply to $i_B^G(\chi^{-1})$. We find that $i_B^G(\chi^{-1})$ has a unique irreducible quotient ρ which corresponds to the unique character of $\mathcal{H}(G, \tilde{\lambda})$ under the bijection of Theorem 4.1. As the contragredient is contravariant and exact, $\tilde{\rho}$ is a subrepresentation of $i_B^G(\chi)$. By Lemma 4.10, we have $(\mathrm{R}_{\tilde{\kappa}}(\rho))^\sim \simeq \mathrm{R}_\kappa(\tilde{\rho})$ which is one-dimensional and hence must be isomorphic to σ. Hence $\tilde{\rho}$ is irreducible and isomorphic to π. Thus π appears twice in the composition series of $i_B^G(\chi)$, as the unique irreducible quotient and as the unique irreducible subrepresentation. □

Remark 6.11. If $\ell \neq 3$ and $\ell \mid q^2-q+1$, then similar counting arguments show that the unique irreducible subrepresentation is not isomorphic to the unique irreducible quotient. However, in these cases we find out more information later so this argument is not necessary.

6E. *On the unramified principal series.*

6E1. *Decomposition of $i_B^G(\nu^2)$ and $i_B^G(\nu^{-2})$.* In all cases of coefficient field, the space of constant functions forms an irreducible subrepresentation of $i_B^G(\nu^{-2})$ isomorphic to 1_G. We let St_G denote the quotient of $i_B^G(\nu^{-2})$ by 1_G. Parabolic induction preserves finite-length representations; hence St_G has an irreducible quotient υ_G. By the geometric lemma, $[r_B^G \circ i_B^G(\nu^{-2})] \simeq \nu^{-2} \oplus (\nu^{-2})^{w_x}$. Considering ν^{-2} as a character of $F^\times$, we have $(\nu^{-2})^{w_x}(x) = \nu^{-2}(\bar{x}^{-1}) = \nu^2(x)$, as $\nu^{-2}(x) = \nu^{-2}(\bar{x})$. Thus $[r_B^G \circ i_B^G(\nu^{-2})] = \nu^{-2} \oplus \nu^2$. We have $r_B^G(1_G) = \nu^{-2}$, thus $r_B^G(\mathrm{St}_G) = \nu^2$ by exactness of the Jacquet functor. A quotient of a parabolically induced representation has nonzero Jacquet module; hence $r_B^G(\upsilon_G) = \nu^2$. Thus any other composition factors which occur in $i_B^G(\nu^{-2})$ must be cuspidal.

Theorem 6.12. (1) *If $\ell \nmid (q-1)(q+1)(q^2-q+1)$ then $i_B^G(\nu^{-2})$ has length two with unique irreducible subrepresentation 1_G and unique irreducible quotient St_G.*

(2) *If $\ell \neq 2$ and $\ell \mid q-1$ then $i_B^G(\nu^{-2}) = 1_G \oplus \mathrm{St}_G$ is semisimple of length two.*

(3) *If $\ell \neq 3$ and $\ell \mid q^2-q+1$ then $i_B^G(\nu^{-2})$ has length three with unique cuspidal subquotient $\mathrm{I}_{\Lambda_x}(\overline{\tau}^+(\bar{1}))$. The unique irreducible quotient υ_G is not a character.*

(4) *If $\ell \neq 2$ and $\ell \mid q+1$, or if $\ell = 2$ and $4 \mid q+1$, then $i_B^G(\nu^{-2})$ has length six with 1_G appearing as the unique subrepresentation and the unique quotient,*

and four cuspidal subquotients. Let π be a maximal proper submodule of St_G. *Then $\pi \simeq \rho \oplus \mathrm{I}_{\Lambda_y}(\bar{\sigma}(\bar{1}) \otimes \bar{1})$, where ρ is of length three with unique irreducible subrepresentation and unique irreducible quotient, both of which are isomorphic to* $\mathrm{I}_{\Lambda_x}(\bar{\nu}(\bar{1}))$, *and remaining subquotient isomorphic to* $\mathrm{I}_{\Lambda_x}(\bar{\sigma}(\bar{1}))$.

(5) *If $\ell = 2$ and $4 \mid q-1$, then $i_B^G(\nu^{-2})$ has length five with unique irreducible subrepresentation and unique irreducible quotient both isomorphic to 1_G. Let π be a maximal proper submodule of* St_G. *Then*

$$\pi \simeq \mathrm{I}_{\Lambda_x}(\bar{\nu}(\bar{1})) \oplus \mathrm{I}_{\Lambda_x}(\bar{\tau}^+(\bar{\chi})) \oplus \mathrm{I}_{\Lambda_y}(\bar{\sigma}(\bar{1}) \otimes \bar{1}).$$

Proof. By Theorem 5.3 and Lemma 6.6, if $i_B^G(\nu^{-2})$ has a cuspidal subquotient π then $\pi \simeq \mathrm{I}_{\Lambda_w}(\sigma)$ for $w \in \{x, y\}$ and σ an irreducible cuspidal representation of $\mathrm{P}(\Lambda_w)/\mathrm{P}_1(\Lambda_w)$.

If Σ_w is a maximal cuspidal subquotient of $\mathrm{R}_{\Lambda_w}(i_B^G(\nu^{-2}))$ then $\mathrm{I}_{\Lambda_w}(\Sigma_w)$ is a subrepresentation of St_G, by Lemma 6.7. Thus, we have an exact sequence

$$0 \to \mathrm{I}_{\Lambda_x}(\Sigma_x) \oplus \mathrm{I}_{\Lambda_y}(\Sigma_y) \to \mathrm{St}_G \to \upsilon_G \to 0.$$

By exactness and Section 3, we obtain composition series of $\mathrm{I}_{\Lambda_x}(\Sigma_x)$ and of $\mathrm{I}_{\Lambda_y}(\Sigma_y)$.

If $\ell \nmid (q-1)(q+1)(q^2-q+1)$, or $\ell \neq 2$ and $\ell \mid q-1$, then $\mathrm{R}_{\Lambda_x}(i_B^G(\nu^{-2}))$ and $\mathrm{R}_{\Lambda_y}(i_B^G\nu^{-2}))$ are of length two with no cuspidal subquotients, by Theorem 5.3 and Lemma 6.6. Hence, $i_B^G(\nu^{-2})$ has no cuspidal subquotients as $\mathrm{R}_{\Lambda_w}(\mathrm{I}_{\Lambda_w}(\sigma)) \simeq \sigma$ is cuspidal by Lemma 4.4. By the geometric lemma, $i_B^G(\nu^{-2})$ is of length two with 1_G as an irreducible subrepresentation and St_G as an irreducible quotient. By second adjunction,

$$\mathrm{Hom}_G(i_B^G(\nu^{-2}), 1_G) \simeq \mathrm{Hom}_T(\nu^{-2}, 1_T).$$

The character ν^{-2} is nontrivial when $\ell \nmid (q-1)(q+1)(q^2-q+1)$ and trivial when $\ell \mid q-1$. Hence 1_G is a direct factor when $\ell \neq 2$ and $\ell \mid q-1$ and $i_B^G(\nu^{-2})$ is semisimple, and $i_B^G(\nu^{-2})$ is nonsplit when $\ell \nmid (q-1)(q+1)(q^2-q+1)$.

In all other cases, $i_B^G(\nu^{-2})$ has cuspidal subquotients. Thus 1_G cannot be a direct factor. Therefore $i_B^G(\nu^{-2})$ has a unique irreducible quotient υ_G and a unique irreducible subrepresentation 1_G. When $\ell \mid q+1$ the unique irreducible quotient is isomorphic to the unique irreducible subrepresentation by Lemma 6.10; hence $\upsilon_G \simeq 1_G$. When $\ell \neq 3$ and $\ell \mid q^2-q+1$, the representation $\mathrm{R}_{\Lambda_y}(i_B^G(\nu^{-2}))$ has noncuspidal subquotients 1_{M_y} and St_{M_y}. By exactness, $\mathrm{R}_{\Lambda_y}(\upsilon_G) \simeq \mathrm{St}_{M_y}$; hence 1_G is not isomorphic to υ_G, which is not a character. □

Note that $i_B^G(\nu^2) \simeq i_B^G(\nu^{-2})^{\sim}$; hence decompositions of $i_B^G(\nu^2)$ can be obtained from Theorem 6.12.

6E2. *Decomposition of unramified $i_B^G(\eta\nu)$ and $i_B^G(\eta\nu^{-1})$.* Let η be the unique unramified character of $F^\times$ extending ω_{F/F_0}. If $\ell \mid q+1$, then $\omega_{F/F_0}\nu^{-1} = \omega_{F/F_0}\nu = 1$;

hence we refer to Theorem 6.12. When $\ell \mid q^2+q+1$ we have $\nu^2=\eta\nu^{-1}$ and $\nu^{-2}=\eta\nu$; hence once more we refer to Theorem 6.12. When $\ell \mid q-1$, ν is trivial, hence $\eta\nu=\eta\nu^{-1}=\eta$. Thus $i_B^G(\eta)$ is self-contragredient. By Lemma 6.8, $i_B^G(\eta)$ has length two and is semisimple.

6F. ***Cuspidal subquotients of the ramified level zero principal series.*** We describe the reducible principal series $i_B^G(\chi)$ which have length greater than two when χ is a level zero character of T which does not factor through the determinant map. We twist by a character that factors through the determinant map so that we can assume $\chi_2=1$. Then $\chi^{q+1}=1$ and $\chi=\overline{\psi}\circ\xi$ for $\overline{\psi}$ a nontrivial character of k_F^1.

When $\ell \nmid q+1$, because $\mathrm{R}_{\Lambda_x}(i_B^G(\chi))$ and $\mathrm{R}_{\Lambda_y}(i_B^G(\chi))$ have no cuspidal subquotients, $i_B^G(\chi)$ is of length two.

Theorem 6.13. *Let $\ell \mid q+1$. The representation $i_B^G(\chi)$ has length four with a unique irreducible subrepresentation and a unique irreducible quotient, and cuspidal subquotient isomorphic to $\mathrm{I}_{\Lambda_x}(\overline{\sigma}(\overline{\psi},\overline{\psi},\overline{1}))\oplus\mathrm{I}_{\Lambda_y}(\overline{\sigma}(\overline{\psi})\otimes\overline{1})$. Furthermore, the unique irreducible subrepresentation is isomorphic to the unique irreducible quotient.*

Proof. The proof is similar to the proof of Theorem 6.12. □

7. Cuspidal subquotients of positive level principal series

In this section, suppose that χ_1 is a positive level character of $F^\times$ trivial on $F_0^\times$ and χ is the character of T given by χ_1 and $\chi_2=1$. We assume we are in the same setting as Section 6D with (T^0,λ_T) an R-type contained in χ, (J,λ) a G-cover of (T^0,λ_T) relative to B with $\lambda=\kappa\otimes\sigma$, and (κ_m,Λ^m) compatible with (κ,Λ) for $m\in\{x,y\}$. We have $\mathrm{M}(\Lambda_E^m)\simeq\mathrm{U}(1,1)(k_F/k_0)\times\mathrm{U}(1)(k_F/k_0)$. When $\ell\nmid q+1$, there are no cuspidal subquotients of $\mathrm{U}(1,1)(k_F/k_0)$, and hence no cuspidal subquotients of $i_B^G(\chi)$, by Lemma 6.6. Thus it remains to look at the case when $\ell \mid q+1$. Let $\overline{\psi}=(\chi\kappa_T^{-1})^{T^1}$ and $\overline{\chi}$ the character of k_F^1 such that $\overline{\psi}=\overline{\chi}\circ\xi$.

Theorem 7.1. *Suppose $\ell \mid q+1$. The representation $i_B^G(\chi)$ has length four with unique irreducible subrepresentation and unique irreducible quotient which are isomorphic, and cuspidal subquotient isomorphic to $\mathrm{I}_{\kappa_x}(\sigma(\overline{\chi})\otimes\overline{1})\oplus\mathrm{I}_{\kappa_y}(\sigma(\overline{\chi})\otimes\overline{1})$.*

Proof. The proof is similar to the proof of Theorem 6.12. □

8. Supercuspidal support

Theorem 8.1. *Let G be an unramified unitary group in three variables and π an irreducible ℓ-modular representation of G. Then the supercuspidal support of π is unique up to conjugacy.*

Proof. Suppose π is not cuspidal. Then the supercuspidal support of π is equal to the cuspidal support of π and is thus unique up to conjugacy. If π is cuspidal nonsupercuspidal then it appears in one of the decompositions given in Theorems 6.12, 6.13, and 7.1, or is a twist of such a representation by a character that factors through the determinant map, and we see that the supercuspidal support of π is unique up to conjugacy. □

Remark 8.2. Let $\mathrm{I}_{\kappa'}(\sigma')$ be an irreducible cuspidal representation of G as constructed in Theorem 5.3. After the decomposition of the parabolically induced representations given in Theorems 6.12, 6.13, and 7.1, we see that $\mathrm{I}_{\kappa'}(\sigma')$ is supercuspidal if and only if σ' is supercuspidal. Hence all supercuspidal representations of G lift, by Section 3.

Acknowledgements

Part of this work was included in my doctoral thesis. I would like to thank my supervisor, Shaun Stevens, for his support and encouragement both during my doctorate and after. I would like to thank Vincent Sécherre and Alberto Mínguez for many useful conversations. Particularly, I would like to thank Vincent Sécherre for some technical suggestions regarding κ-induction and κ-restriction. I would like to thank the referee for their careful reading and helpful suggestions. This work was supported by the Engineering and Physical Sciences Research Council (EP/H00534X/1).

References

[Blasco 2002] L. Blasco, "Description du dual admissible de U(2, 1)(F) par la théorie des types de C. Bushnell et P. Kutzko", *Manuscripta Math.* **107**:2 (2002), 151–186. MR 2003d:22017 Zbl 1108.22011

[Blondel 2005] C. Blondel, "Quelques propriétés des paires couvrantes", *Math. Ann.* **331**:2 (2005), 243–257. MR 2005k:20117 Zbl 1062.22035

[Blondel 2012] C. Blondel, "Représentation de Weil et β-extensions", *Ann. Inst. Fourier* (*Grenoble*) **62**:4 (2012), 1319–1366. MR 3025745 Zbl 1252.22009

[Bonnafé and Rouquier 2003] C. Bonnafé and R. Rouquier, "Catégories dérivées et variétés de Deligne–Lusztig", *Publ. Math. Inst. Hautes Études Sci.* 97 (2003), 1–59. MR 2004i:20079 Zbl 1054.20024

[Bushnell and Kutzko 1993a] C. J. Bushnell and P. C. Kutzko, *The admissible dual of* GL(N) *via compact open subgroups*, Annals of Mathematics Studies **129**, Princeton University Press, 1993. MR 94h:22007 Zbl 0787.22016

[Bushnell and Kutzko 1993b] C. J. Bushnell and P. C. Kutzko, "The admissible dual of SL(N), I", *Ann. Sci. École Norm. Sup.* (4) **26**:2 (1993), 261–280. MR 94a:22033 Zbl 0787.22017

[Bushnell and Kutzko 1998] C. J. Bushnell and P. C. Kutzko, "Smooth representations of reductive p-adic groups: Structure theory via types", *Proc. London Math. Soc.* (3) **77**:3 (1998), 582–634. MR 2000c:22014 Zbl 0911.22014

[Dat 2005] J.-F. Dat, "v-tempered representations of p-adic groups, I: l-adic case", *Duke Math. J.* **126**:3 (2005), 397–469. MR 2005m:22014 Zbl 1063.22017

[Dat 2009] J.-F. Dat, "Finitude pour les représentations lisses de groupes p-adiques", *J. Inst. Math. Jussieu* **8**:2 (2009), 261–333. MR 2010e:22007 Zbl 1158.22020

[Digne and Michel 1991] F. Digne and J. Michel, *Representations of finite groups of Lie type*, London Mathematical Society Student Texts **21**, Cambridge University Press, 1991. MR 92g:20063 Zbl 0815.20014

[Ennola 1963] V. Ennola, "On the characters of the finite unitary groups", *Ann. Acad. Sci. Fenn. Ser. A I No.* **323** (1963), 35. MR 28 #143 Zbl 0109.26001

[Geck 1990] M. Geck, "Irreducible Brauer characters of the 3-dimensional special unitary groups in nondefining characteristic", *Comm. Algebra* **18**:2 (1990), 563–584. MR 91b:20016 Zbl 0696.20011

[Geck et al. 1996] M. Geck, G. Hiss, and G. Malle, "Towards a classification of the irreducible representations in non-describing characteristic of a finite group of Lie type", *Math. Z.* **221**:3 (1996), 353–386. MR 98a:20017 Zbl 0858.20008

[Henniart and Sécherre 2014] G. Henniart and V. Sécherre, "Types et contragrédientes", *Canad. J. Math.* **66**:6 (2014), 1287–1304. MR 3270784

[Hiss 2004] G. Hiss, "Hermitian function fields, classical unitals, and representations of 3-dimensional unitary groups", *Indag. Math.* (*N.S.*) **15**:2 (2004), 223–243. MR 2005c:20080 Zbl 1139.05308

[Howlett and Lehrer 1980] R. B. Howlett and G. I. Lehrer, "Induced cuspidal representations and generalised Hecke rings", *Invent. Math.* **58**:1 (1980), 37–64. MR 81j:20017 Zbl 0435.20023

[Keys 1984] D. Keys, "Principal series representations of special unitary groups over local fields", *Compositio Math.* **51**:1 (1984), 115–130. MR 85d:22031 Zbl 0547.22009

[Kim 2007] J.-L. Kim, "Supercuspidal representations: An exhaustion theorem", *J. Amer. Math. Soc.* **20**:2 (2007), 273–320. MR 2008c:22014 Zbl 1111.22015

[Kurinczuk and Stevens 2014] R. Kurinczuk and S. Stevens, "ℓ-modular cuspidal representations of classical p-adic groups", preprint, 2014.

[Mínguez and Sécherre 2014a] A. Mínguez and V. Sécherre, "Représentations lisses modulo ℓ de $\mathrm{GL}_m(D)$", *Duke Math. J.* **163**:4 (2014), 795–887. MR 3178433 Zbl 1293.22005

[Mínguez and Sécherre 2014b] A. Mínguez and V. Sécherre, "Types modulo ℓ pour les formes intérieures de $\mathrm{GL}(n)$ sur un corps local non archimédien", *Proc. London Math. Soc.* (3) **109**:4 (2014), 823–891.

[Morris 1993] L. Morris, "Tamely ramified intertwining algebras", *Invent. Math.* **114**:1 (1993), 1–54. MR 94g:22035 Zbl 0854.22022

[Morris 1999] L. Morris, "Level zero **G**-types", *Compositio Math.* **118**:2 (1999), 135–157. MR 2000g:22029 Zbl 0937.22011

[Okuyama and Waki 2002] T. Okuyama and K. Waki, "Decomposition numbers of $\mathrm{SU}(3, q^2)$", *J. Algebra* **255**:2 (2002), 258–270. MR 2003h:20026 Zbl 1023.20005

[Schneider and Zink 1999] P. Schneider and E.-W. Zink, "K-types for the tempered components of a p-adic general linear group", *J. Reine Angew. Math.* **517** (1999), 161–208. MR 2001f:22029 Zbl 0934.22021

[Sécherre and Stevens 2008] V. Sécherre and S. Stevens, "Représentations lisses de $\mathrm{GL}_m(D)$, IV: Représentations supercuspidales", *J. Inst. Math. Jussieu* **7**:3 (2008), 527–574. MR 2009d:22023 Zbl 1140.22014

[Stevens 2005] S. Stevens, "Semisimple characters for p-adic classical groups", *Duke Math. J.* **127**:1 (2005), 123–173. MR 2006a:22017 Zbl 1063.22018

[Stevens 2008] S. Stevens, "The supercuspidal representations of p-adic classical groups", *Invent. Math.* **172**:2 (2008), 289–352. MR 2010e:22008 Zbl 1140.22016

[Vignéras 1996] M.-F. Vignéras, *Représentations l-modulaires d'un groupe réductif p-adique avec $l \neq p$*, Progress in Mathematics **137**, Birkhäuser, Boston, 1996. MR 97g:22007 Zbl 0859.22001

[Vignéras 1998] M.-F. Vignéras, "Induced R-representations of p-adic reductive groups", *Selecta Math.* (*N.S.*) **4**:4 (1998), 549–623. MR 99k:22026 Zbl 0943.22017

[Vignéras 2001] M.-F. Vignéras, "Irreducible modular representations of a reductive p-adic group and simple modules for Hecke algebras", pp. 117–133 in *European Congress of Mathematics, I* (Barcelona, 2000), edited by C. Casacuberta et al., Progr. Math. **201**, Birkhäuser, Basel, 2001. MR 2003g:22023 Zbl 1024.22011

[Vignéras 2004] M.-F. Vignéras, "On highest Whittaker models and integral structures", pp. 773–801 in *Contributions to automorphic forms, geometry, and number theory*, edited by H. Hida et al., Johns Hopkins Univ. Press, Baltimore, MD, 2004. MR 2006b:11048 Zbl 1084.11023

[Yu 2001] J.-K. Yu, "Construction of tame supercuspidal representations", *J. Amer. Math. Soc.* **14**:3 (2001), 579–622. MR 2002f:22033 Zbl 0971.22012

Communicated by Marie-France Vignéras
Received 2013-10-21 Revised 2014-06-24 Accepted 2014-09-08

robkurinczuk@gmail.com
Department of Mathematics,
Heilbronn Institute for Mathematical Research,
University of Bristol, Bristol, BS8 1TW, United Kingdom

ALGEBRA AND NUMBER THEORY 8:8 (2014)
dx.doi.org/10.2140/ant.2014.8.1839

McKay natural correspondences on characters

Gabriel Navarro, Pham Huu Tiep and Carolina Vallejo

Let G be a finite group, let p be an odd prime, and let $P \in \mathrm{Syl}_p(G)$. If $\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$, then there is a canonical correspondence between the irreducible complex characters of G of degree not divisible by p belonging to the principal block of G and the linear characters of P. As a consequence, we give a characterization of finite groups that possess a self-normalizing Sylow p-subgroup or a p-decomposable Sylow normalizer.

1. Introduction

The McKay conjecture, one of the main problems in the representation theory of finite groups, asserts that if G is a finite group and P is a Sylow p-subgroup of G, then $|\mathrm{Irr}_{p'}(G)| = |\mathrm{Irr}_{p'}(\boldsymbol{N}_G(P))|$, where $\mathrm{Irr}_{p'}(G)$ is the set of the irreducible complex characters of G that have degree not divisible by p. It is well known that, in general, no choice-free correspondence can exist between $\mathrm{Irr}_{p'}(G)$ and $\mathrm{Irr}_{p'}(\boldsymbol{N}_G(P))$. (On the other hand, the existence of certain type of bijections between these two sets is the idea on which a possible solution of the McKay conjecture is nowadays based [Isaacs et al. 2007].)

A key case to consider and understand in the McKay conjecture is when P is self-normalizing or, even, when $\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$. It is not often that a natural correspondence of characters is found.

Theorem A. *Let G be a finite group, let p be odd, and let $P \in \mathrm{Syl}_p(G)$. Suppose that $\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$. If $\chi \in \mathrm{Irr}_{p'}(G)$ lies in the principal block, then*

$$\chi_{\boldsymbol{N}_G(P)} = \chi^* + \Delta,$$

where $\chi^* \in \mathrm{Irr}(\boldsymbol{N}_G(P))$ is linear in the principal block and Δ is either zero or a character whose irreducible constituents all have degree divisible by p. Furthermore,

The research of Navarro and Vallejo is partially supported by the Spanish Ministerio de Educación y Ciencia proyecto MTM2013-40464-P. Tiep gratefully acknowledges the support of the NSF (grants DMS-0901241 and DMS-1201374) and the Simons Foundation Fellowship 305247.

MSC2010: primary 20C15; secondary 20C20.

Keywords: McKay conjecture, self-normalizing Sylow subgroup, p-decomposable Sylow normalizer.

the map $\chi \mapsto \chi^*$ *is a bijection*

$$\mathrm{Irr}_{p'}(B_0(G)) \to \mathrm{Irr}_{p'}\big(B_0(\boldsymbol{N}_G(P))\big),$$

where $\mathrm{Irr}_{p'}(B_0(G))$ *is the set of irreducible characters in the principal block of* G *of degree not divisible by* p.

For $p = 2$, Theorem A is unfortunately false, as shown, for instance, by S_5. (To prove the McKay conjecture for $p = 2$ for groups with a self-normalizing Sylow p-subgroup is still a challenge.) For p odd, Theorem A is also not true for p-blocks of maximal defect, as shown by the following example: $G = \mathrm{SL}_2(27) \cdot C_3$ has a rational, faithful, irreducible character χ of degree 26 that belongs to the unique nonprincipal 3-block of maximal defect of G, and $\chi_{\boldsymbol{N}_G(P)}$ contains two linear characters as irreducible constituents.

Theorem A yields the following immediate consequence.

Corollary B. *Let* G *be a finite group, let* p *be odd, and let* $P \in \mathrm{Syl}_p(G)$. *Suppose that* $\boldsymbol{N}_G(P) = P$. *Then there is a natural bijection* $\chi \mapsto \chi^*$ *between* $\mathrm{Irr}_{p'}(G)$ *and the linear characters of* P. *In fact, if* $\chi \in \mathrm{Irr}_{p'}(G)$ *and* $\lambda \in \mathrm{Irr}(P)$ *is linear, then* χ *and* λ *correspond under the bijection if and only if*

$$\chi_P = \lambda + \Delta,$$

where Δ *is either zero or a character whose irreducible constituents all have degree divisible by* p. *This happens if and only if*

$$\lambda^G = \chi + \Xi,$$

where Ξ *is either zero or a character whose irreducible constituents all have degree divisible by* p.

Corollary B was proved in [Navarro 2003] for p-solvable groups (although a different proof was later given in [Isaacs and Navarro 2008]).

We now mention several applications. A not very well-known consequence of the Galois version of the McKay conjecture [Navarro 2004] states that whenever G is a finite group and p is an odd prime, then $\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$ if and only if the principal character 1_G is the unique p-rational p'-degree character in the principal block of G. If $\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$, it follows by Theorem A that the fields of values of the p'-degree nontrivial irreducible characters in the principal block are cyclotomic fields $\mathbb{Q}_{p^a}$ for $a > 0$, which implies one half of the statement above. (The other half will be treated separately in another paper.)

A consequence of Theorem A and Corollary B is the following (perhaps surprising) characterization of finite groups that possess a self-normalizing Sylow p-subgroup or a p-decomposable Sylow normalizer (i.e., $\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$), for a given odd prime p.

Corollary C. *Let G be a finite group, let p be odd, and let $P \in \mathrm{Syl}_p(G)$.*

(a) *$\boldsymbol{N}_G(P) = P$ if and only if*

$$(1_P)^G = 1_G + \Xi,$$

where Ξ is either zero or a character whose irreducible constituents all have degree divisible by p.

(b) *$\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$ if and only if 1_G is the only irreducible constituent of $(1_{P\boldsymbol{C}_G(P)})^G$ that has p'-degree and belongs to the principal p-block of G.*

It is remarkable that Corollary C(a) gives the exact opposite of a recent result by G. Malle and Navarro [2012]: a finite group G has a normal Sylow p-subgroup if and only if all irreducible constituents of $(1_P)^G$ have degree not divisible by p.

Corollary C is false for $p = 2$, as shown again by $G = \mathsf{S}_5$: in this case $(1_P)^G$ contains the trivial character of G and an irreducible character of degree 5.

Now, we come back to Theorem A and natural correspondences. Although it is entirely possible that, under the hypotheses of Theorem A, a natural correspondence exists between *all* the characters in $\mathrm{Irr}_{p'}(G)$ and $\mathrm{Irr}_{p'}(\boldsymbol{N}_G(P))$ (not only the characters in the principal blocks), we have not been able to find it, except for p-solvable groups. In this case, our correspondence in Theorem D below extends the Glauberman correspondence (and the correspondence in Theorem A).

Theorem D. *Let G be a finite p-solvable group, and let $P \in \mathrm{Syl}_p(G)$. Suppose that $\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$, and let $N = \boldsymbol{O}_{p'}(G)$. Let $\mathrm{Irr}_P(N)$ be the set of P-invariant characters $\theta \in \mathrm{Irr}(N)$. Then, for every $\theta \in \mathrm{Irr}_P(N)$ and linear $\lambda \in \mathrm{Irr}(P/P')$, there is a canonically defined character*

$$\lambda \star \theta \in \mathrm{Irr}_{p'}(G).$$

Furthermore, the map

$$\mathrm{Irr}(P/P') \times \mathrm{Irr}_P(N) \to \mathrm{Irr}_{p'}(G)$$

given by $(\lambda, \theta) \mapsto \lambda \star \theta$ is a bijection. As a consequence, $\boldsymbol{N}_G(P) = P \times \boldsymbol{C}_N(P)$, and if $\theta^* \in \mathrm{Irr}(\boldsymbol{C}_N(P))$ is the Glauberman correspondent of $\theta \in \mathrm{Irr}_P(N)$, then the map

$$\lambda \times \theta^* \mapsto \lambda \star \theta$$

is a natural bijection $\mathrm{Irr}_{p'}(\boldsymbol{N}_G(P)) \to \mathrm{Irr}_{p'}(G)$. Also, if $\theta = 1_N$ and $\lambda \in \mathrm{Irr}(P/P')$, then $\lambda \times \theta^*$ is the unique linear constituent of $(\lambda \star \theta)_{\boldsymbol{N}_G(P)}$.

Theorem D suggests studying the blocks B of finite groups with defect group D satisfying $\boldsymbol{N}_G(D, b_D) = D\boldsymbol{C}_G(D)$, where (D, b_D) is a *root* of B. However, we will leave this for another place.

2. An extension theorem

We begin with a well-known lemma. If $N \lhd G$ and $\theta \in \mathrm{Irr}(N)$, then $I_G(\theta)$ denotes the stabilizer of θ in G. In general, we follow the notation of [Isaacs 2006] for characters. If G is a finite group, $\mathrm{Irr}_{p'}(G)$ denotes the set of the irreducible complex characters of G whose degree $\chi(1)$ is not divisible by the prime p.

Lemma 2.1. *Let G be a finite group, let p be a prime, let $P \in \mathrm{Syl}_p(G)$, and let $\chi \in \mathrm{Irr}_{p'}(G)$. Assume that $L \lhd G$. Then χ_L has a P-invariant irreducible constituent θ, and all such constituents are $\boldsymbol{N}_G(P)$-conjugate. In particular, if $\boldsymbol{N}_{G/L}(PL/L) = PL/L$, then θ is unique.*

Proof. Let $\eta \in \mathrm{Irr}(L)$ be any irreducible constituent, and let T be the inertia subgroup of η in G. By the Clifford correspondence, $|G : T|$ is not divisible by p, and therefore $P^{g^{-1}} \leq T$ for some $g \in G$, and thus P fixes $\eta^g = \theta$. If P fixes θ^x, then $P^{x^{-1}}$ and P are Sylow p-subgroups of $I = I_G(\theta)$, and $P^y = P^{x^{-1}}$ for some $y \in I$. Hence $yx \in \boldsymbol{N}_G(P)$ and $\theta^x = \theta^{yx}$. The second part easily follows. □

Lemma 2.2. *Let G be a finite group, let p be prime, and let $P \in \mathrm{Syl}_p(G)$. Let $L \lhd G$, and assume that $\boldsymbol{N}_{G/L}(PL/L) = PL/L$. Let $\theta \in \mathrm{Irr}(L)$ be P-invariant, let $T = I_G(\theta)$ be the stabilizer of θ in G and assume that $\psi \in \mathrm{Irr}(T \mid \theta)$ has p'-degree. Then*

$$(\psi^G)_{LP} = \psi_{LP} + \Delta,$$

where either $\Delta = 0$ or every irreducible constituent of Δ has degree divisible by p.

Proof. Let

$$G = \bigcup_{x \in D} TxP$$

be a disjoint union of double cosets with $1 \in D$. Then, by Mackey's formula, we have that

$$(\psi^G)_{LP} = \psi_{LP} + \sum_{1 \neq x \in D} ((\psi^x)_{T^x \cap LP})^{LP}.$$

Suppose that some irreducible constituent α of $((\psi^x)_{T^x \cap LP})^{LP}$ has degree not divisible by p for $1 \neq x \in D$. Hence $\alpha_L \in \mathrm{Irr}(L)$ by Corollary (11.29) of [Isaacs 2006]. Thus the irreducible character $\alpha_{T^x \cap LP}$ lies under ψ^x. However $(\psi^x)_L = d\theta^x$, so we conclude that $\theta^x = \alpha_L$ is P-invariant. Then by Lemma 2.1, we have that $\theta^x = \theta$ and therefore $x \in T$. But this is impossible since $1 \neq x \in D$ is a representative of the double cosets of T and P in G. □

The following theorem is key in this paper, and follows from deep results in [Navarro and Späth 2014] and [Späth 2013] on the McKay conjecture. Despite many efforts, we have been unable to find an elementary proof of it. Recall that a finite simple group X is *involved* in a finite group G if there exist $K \lhd H \leq G$ such that $X \cong H/K$. The so-called *inductive Alperin–McKay condition* is defined in Definition 7.2 of [Späth 2013]. For character triples, see Chapter 11 of [Isaacs 2006].

Theorem 2.3. *Let G be a finite group, and let p be a prime. Let $P \in \mathrm{Syl}_p(G)$ and assume that $P = N_G(P)$. Let $L \lhd G$ and let $\theta \in \mathrm{Irr}(L)$ be P-invariant of p'-degree. Suppose that $L \lhd H$ with H/L a p'-group. Assume that all nonabelian simple groups of order divisible by p involved in L satisfy the inductive Alperin–McKay condition for p. If θ is H-invariant, then θ extends to H. In particular, this holds if p is odd.*

Proof. We argue by induction on $|G|$. Let $Q = P \cap L$. We are going to use Theorem 7.1 of [Navarro and Späth 2014]. The notation $\mathrm{Irr}_0(L \mid Q)$ in that theorem is defined in Notation 2.1 of the same article, and since $Q \in \mathrm{Syl}_p(L)$, we have that $\mathrm{Irr}_0(L \mid Q) = \mathrm{Irr}_{p'}(L)$ in this case. Theorem 7.1 of [loc. cit.] implies now that there is a $N_G(Q)$-equivariant bijection

$$\Pi_Q : \mathrm{Irr}_{p'}(L) \to \mathrm{Irr}_{p'}(N_L(Q))$$

such that the character triples (T, L, θ) and $(N_T(Q), N_L(Q), \theta')$ are isomorphic, where $\theta' = \Pi_Q(\theta)$ and $T = I_G(\theta)$. (In Section 3 of [loc. cit.] the reader will find the appropriate definitions involved in Theorem 7.1 there.) Since Π_Q is $N_G(Q)$-equivariant, we have that $N_T(Q) = I_{N_G(Q)}(\theta')$. Since $P \leq N_G(Q)$, we have that θ' is P-invariant. By character triple isomorphisms, we have that θ extends to H if and only if θ' extends to $N_H(Q)$. Also $N_H(Q)/N_L(Q)$ is a p'-group, so, arguing by induction, it is no loss to assume that $Q \lhd G$. Since $N_G(P) = P$, it follows that $C_{L/Q}(P) = 1$. Let $\eta \in \mathrm{Irr}(Q)$ be P-invariant under θ, by Theorem (13.27) of [Isaacs 2006]. Let $I = I_G(\eta)$. Since θ is H-invariant, we have that $H = L(I \cap H)$ by using Clifford's theorem. Let $\tau \in \mathrm{Irr}(I \cap L \mid \eta)$ be the Clifford correspondent of θ over η. By the uniqueness in the Clifford correspondence, we have that τ is $I \cap H$-invariant. If $I < G$, then by induction we have that τ has an extension $\rho \in \mathrm{Irr}(I \cap H)$. Now,

$$(\rho^H)_L = (\rho_{I \cap L})^L = \epsilon^L = \theta,$$

and we are also done in this case. So we may assume that η is G-invariant. Since $C_{L/Q}(P) = 1$, by Problem (13.10) of [loc. cit.] θ is the unique P-invariant constituent of η^L. Now, we have that η has an extension $\hat{\eta} \in \mathrm{Irr}(H)$ by Corollary (8.16) of [loc. cit.]. Since $(\hat{\eta})_L$ is P-invariant and lies over η, it coincides with θ by uniqueness. Hence θ extends to H, as required.

If p is odd, then by Theorem A of [Guralnick et al. 2004], we have that all nonabelian composition factors of G of order divisible by p are $\mathrm{PSL}_2(3^{3^a})$ with $a \geq 1$ and that $p = 3$. By elementary general group theory, if X is a simple group involved in G, then X is involved in a composition factor of G. By using the classification of the subgroups of $\mathrm{PSL}_2(p^f)$ in Satz II.8.27 of [Huppert 1967], we have that the only simple groups involved in G of order divisible by p are $\mathrm{PSL}_2(3^{3^b})$ (with $p = 3$

and $b \geq 1$). Now, the proof of Theorem 8.4 of [Späth 2013] shows that the simple groups $\mathrm{PSL}_2(3^{3^b})$ with $b \geq 1$ satisfy the inductive Alperin–McKay condition. □

Corollary 2.4. *Let G be a finite group, p any prime, $P \in \mathrm{Syl}_p(G)$, and assume that $P = \mathbf{N}_G(P)$. Let $L \lhd G$. Let $\chi \in \mathrm{Irr}_{p'}(G)$, and let $\theta \in \mathrm{Irr}(L)$ be P-invariant under χ. Assume that all nonabelian simple groups of order divisible by p involved in L satisfy the inductive Alperin–McKay condition for p. Then θ extends to $I_G(\theta)$. In particular, this holds if p is odd.*

Proof. We may assume that θ is G-invariant. Now, χ_{PL} has some irreducible constituent $\xi \in \mathrm{Irr}(PL)$ such that p does not divide $\xi(1)$. Then $\xi_L = \theta$, by Corollary (11.29) of [Isaacs 2006]. Suppose now that $q \neq p$ is another prime, and let Q/L be a Sylow q-subgroup of G/L. Then θ extends to Q by Theorem 2.3. Hence, we have that θ extends to G by Corollary (11.30) of [loc. cit.]. □

3. A group theoretical result

Our aim in this Section is to prove Theorem 3.2 below. We start with the following lemma, whose parts (ii) and (iii) will be used in the proof of Theorem A.

Lemma 3.1. *Let $S := \mathrm{PSL}_2(q)$ with $q = 3^{3^a}$ for some $a \geq 1$, $P \in \mathrm{Syl}_3(\mathrm{Aut}(S))$, and $Q := P \cap S \in \mathrm{Syl}_3(S)$.*

(i) *Assume that Y is a $3'$-subgroup of* $\mathrm{Aut}(S)$ *that centralizes Q. Then $Y = 1$.*

(ii) *Assume that $Q \leq R \leq P$ and $\mathbf{C}_{\mathbf{N}_S(Q)/Q}(R) = 1$. Then $R = P$.*

(iii) $\mathrm{Irr}(S)$ *contains exactly four P-invariant characters: the principal character 1_S, two irreducible Weil characters $\eta^{\pm}$ of degree $(q-1)/2$, and the Steinberg character of degree q. If $\alpha \in \{1_S, \eta^+, \eta^-\}$, then α_Q contains a unique P-invariant irreducible constituent α^*, which occurs with multiplicity one. Finally, the map $\alpha \mapsto \alpha^*$ is a bijection between $\{1_S, \eta^+, \eta^-\}$ and the set of P-invariant irreducible characters of Q.*

Proof. (i) Recall that $\mathrm{Aut}(S) \cong \mathrm{PGL}_2(q) \cdot C_{3^a}$. Since Y is a $3'$-group, it embeds in $\mathbf{C}_H(Q)$ for $H := \mathrm{PGL}_2(q)$. But $\mathbf{C}_H(Q) = Q$, hence the claim follows.

(ii) Without loss we may assume that Q is the image of the subgroup

$$\left\{ [x] := \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \;\middle|\; x \in \mathbb{F}_q \right\}$$

in $\mathrm{PSL}_2(q)$ and $P = \langle Q, \sigma \rangle$, where σ acts on S as the field automorphism raising every entry y of any matrix in $\mathrm{SL}_2(q)$ to y^3. Then the maximal subgroup $\langle Q, \sigma^3 \rangle$ of P centralizes a subgroup of order 13 of $\mathbf{N}_S(Q)/Q$, namely, the one induced by $\{\mathrm{diag}(z, z^{-1}) \mid z \in \mathbb{F}_{27}^{\times}\}$. Hence the claim follows.

(iii) We keep the notation of (ii). The character table of S is given, for instance, in [Digne and Michel 1991, Table 2]. Now, it is straightforward to check that 1_S, $\eta^{\pm}$, and the Steinberg character (of degree q) are the only P-invariant irreducible characters of S. Next,

$$\mathrm{Irr}(Q) = \left\{\lambda_a : [x] \to \omega^{\mathrm{tr}_{\mathbb{F}_q/\mathbb{F}_3}(ax)} \mid a \in \mathbb{F}_q\right\},$$

where $\omega \in \mathbb{C}$ is a fixed primitive cubic root of unity. Since σ acts on $\mathrm{Irr}(Q)$ via $\lambda_a \mapsto \lambda_{a^3}$, the only P-invariant irreducible characters of Q are $1_Q = \lambda_0$, λ_1, and λ_{-1}. Relabeling η^+ and η^- if necessary, we have that

$$(\eta^+)_Q = \sum_{a \in \mathbb{F}_q^{\times 2}} \lambda_a, \quad (\eta^-)_Q = \sum_{a \in \mathbb{F}_q^{\times} \setminus \mathbb{F}_q^{\times 2}} \lambda_a.$$

Hence λ_1 and λ_{-1} are the only P-invariant irreducible constituents of $(\eta^+)_Q$ and $(\eta^-)_Q$, respectively, each occurring with multiplicity one. □

Theorem 3.2. *Let G be a finite group, let p be a prime, and let $P \in \mathrm{Syl}_p(G)$. Suppose that $\boldsymbol{N}_G(P) = P \times X$. If p is odd or G is p-solvable, then $X \le \boldsymbol{O}_{p'}(G)$. In particular, if $\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$, then $\boldsymbol{O}_{p'}(\boldsymbol{N}_G(P)) \le \boldsymbol{O}_{p'}(G)$.*

Proof. We argue by induction on $|G|$. If $N \lhd G$, then

$$\boldsymbol{N}_{G/N}(PN/N) = \boldsymbol{N}_G(P)N/N = PN/N \times XN/N.$$

Hence, if $N > 1$, then we have that $XN/N \le \boldsymbol{O}_{p'}(G/N)$. In particular, we may assume that $\boldsymbol{O}_{p'}(G) = 1$. Now, suppose that $N = \boldsymbol{O}_p(G) > 1$. Then we conclude that $X \le \boldsymbol{O}_{pp'}(G) = M$. Since $[X, P] = 1$, then $[X, N] = 1$. However, using that $\boldsymbol{O}_{p'}(G) = 1$, we have that $\boldsymbol{C}_M(N) = \boldsymbol{Z}(N) \times \boldsymbol{O}_{p'}(M) = \boldsymbol{Z}(N)$, and we conclude that $X = 1$, in this case. Hence, we may assume that G is not p-solvable, and that p is odd.

Now, let N be a minimal normal subgroup of G. By [Guralnick et al. 2004], we have that $N = S_1 \times \cdots \times S_k$, where $\{S_1, \dots, S_k\}$ are transitively permuted by G and $S_1 = S = \mathrm{PSL}_2(3^{3^a})$. Now $P \cap N = (P \cap S_1) \times \cdots \times (P \cap S_k)$. Fix some index i. Since $[P, X] = 1$, we have $[Q_i, X] = 1$, where $1 < Q_i = P \cap S_i \in \mathrm{Syl}_3(S_i)$. Now, if $x \in X$, then we have that $(S_i)^x = S_j$ for some j. However $Q_i^x \le S_i^x \cap S_i = S_j \cap S_i$, so we conclude that $X \le \boldsymbol{N}_G(S_i)$ for all i with $[X, Q_i] = 1$. Let $Y_i = X\boldsymbol{C}_G(S_i)/\boldsymbol{C}_G(S_i)$, which is a $3'$-subgroup of $\mathrm{Aut}(S_i)$ centralizing the Sylow 3-subgroup Q_i of S_i. By Lemma 3.1(i), $Y_i = 1$, whence $X \le \boldsymbol{C}_G(S_i)$ for all i. Thus $X \le \boldsymbol{C}_G(N)$ for every minimal normal subgroup. Since $\boldsymbol{F}(G) = 1$, we have $\boldsymbol{F}^*(G) = \boldsymbol{E}(G) = E$. Since $\boldsymbol{Z}(E) = 1$, we have that E is semisimple and $\boldsymbol{C}_G(E) = 1$ by Theorems 9.7 and 9.8 of [Isaacs 2008]. Now, E is a direct product of nonabelian simple groups K_i and the normal closure of K_i is a minimal normal subgroup of G (by Lemma 9.17 of [Isaacs 2008], for instance), and we conclude that $X \le \boldsymbol{C}_G(E) = 1$, as desired.

Finally, since $\boldsymbol{C}_G(P) = \boldsymbol{Z}(P) \times \boldsymbol{O}_{p'}(\boldsymbol{N}_G(P))$ (by the Schur–Zassenhaus theorem), it follows that if $\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$, then $\boldsymbol{N}_G(P) = P \times \boldsymbol{O}_{p'}(\boldsymbol{N}_G(P))$, and we apply the first part of the theorem. □

Note that Theorem 3.2 is not true for $p = 2$: If $G = E_6(11)$ and $P \in \mathrm{Syl}_2(G)$, then $\boldsymbol{N}_G(P) = P \times C_5$; see [Kondratiev and Mazurov 2003, Theorem 6(c)].

4. Proof of main theorem

We will also need the following result.

Lemma 4.1 [Navarro et al. 2007]. *Suppose that a finite p-group P acts on a finite group G, stabilizing $N \lhd G$. Suppose that $Q/N \in \mathrm{Syl}_p(G/N)$ is P-invariant, and assume that*

$$G/N = T_1/N \times \cdots \times T_a/N,$$

where the subgroups $T_1, \dots, T_a$ are permuted by P. Let $Q_1 = Q \cap T_1$, and let P_1 be the stabilizer of T_1 in P. If $\boldsymbol{C}_{\boldsymbol{N}_G(Q)/Q}(P) = 1$, then $\boldsymbol{C}_{\boldsymbol{N}_{T_1}(Q_1)/Q_1}(P_1) = 1$.

Proof. This follows by applying Lemma 4.1 of [Navarro et al. 2007] to each of the P-orbits on $\{T_1, \dots, T_a\}$. □

The proof of the following lemma is a trivial consequence of the fact that $\boldsymbol{O}_{p'}(G)$ is contained in the kernel of all the irreducible characters in the principal block of G. (See, for instance, Theorem (6.10) of [Navarro 1998].)

Lemma 4.2. *Suppose that N is a normal subgroup of H, with $N \leq \boldsymbol{O}_{p'}(H)$. Suppose that $H = NU$ for some $U \leq H$. Then restriction defines a bijection between the characters of the principal block of H and of the principal block of $U \cap N$.*

We are finally ready to prove the main result of this paper. The only way we have found to prove it is to use a strong induction using normal subgroups. Theorem A of the introduction will be recovered by letting $L = 1$ in the next result.

Theorem 4.3. *Let G be a finite group, and let p be an odd prime. Let $P \in \mathrm{Syl}_p(G)$, and suppose that $\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$. Let $L \lhd G$. Let $\chi \in \mathrm{Irr}_{p'}(G)$ lie in the principal block of G. Then $\chi_{L\boldsymbol{N}_G(P)} = \chi^* + \Delta$, where $\chi^* \in \mathrm{Irr}_{p'}(L\boldsymbol{N}_G(P))$ lies in the principal block of $L\boldsymbol{N}_G(P)$, and Δ is either zero or a character of $L\boldsymbol{N}_G(P)$ whose irreducible constituents all have degree divisible by p. Furthermore, the map $\chi \mapsto \chi^*$ is a bijection*

$$\mathrm{Irr}_{p'}(B_0(G)) \to \mathrm{Irr}_{p'}(B_0(L\boldsymbol{N}_G(P))).$$

Proof. (I) Let (G, L) be a counterexample to the first part of the theorem with $|G| \cdot |G/L|$ as small as possible.

(a) Here we show that $\boldsymbol{O}_{p'}(G) = 1$ and $\boldsymbol{N}_G(P) = P$. To this end, using Theorem 3.2 we can write $\boldsymbol{N}_G(P) = P \times X$, where $X \leq N := \boldsymbol{O}_{p'}(G)$. Write $\overline{G} = G/N$ and use the bar convention. Hence $\overline{L} = LN/N$, $\overline{P} = PN/N$ and $\boldsymbol{N}_{\overline{G}}(\overline{P}) = \overline{P} \times \overline{X} = \overline{P}$, by elementary group theory. Now, $N \leq \ker(\chi)$. If $N > 1$, then, considering χ as a character of $\overline{G}$, by induction we have that

$$\chi_{\overline{L}\boldsymbol{N}_{\overline{G}}(\overline{P})} = \chi_{\overline{L}\,\overline{P}} = \chi^* + \Delta,$$

where χ^* is an irreducible character of p'-degree in the principal block of $\overline{L}\,\overline{P} = LPN/N$ and either $\Delta = 0$ or Δ is a character of LPN/N such that every irreducible constituent of Δ has degree divisible by p. Now, Lemma 4.2 applies, and we are done in this case. So we have that $N = 1$ and that $\boldsymbol{N}_G(P) = P$. Hence, every p'-degree character of every subgroup H with $P \leq H \leq G$ (or of every quotient G/K of G) lies in the principal block of H (of G/K) by the first main theorem of Brauer (Theorem (4.17) of [Navarro 1998]).

(b) Next we show that $L = 1$. By Lemma 2.1, let $\theta \in \mathrm{Irr}(L)$ be P-invariant under χ. Let $T = I_G(\theta)$ be the stabilizer of θ in G, and let $\psi \in \mathrm{Irr}(T \mid \theta)$ be the Clifford correspondent of χ over θ. Assume that $T < G$. By the choice of G, we have that

$$\psi_{LP} = \psi^* + \Xi,$$

where ψ^* has p'-degree and either $\Xi = 0$ or the irreducible constituents of Ξ have degree divisible by p. Now, we use Lemma 2.2 to conclude that we may assume that θ is G-invariant. By Corollary 2.4, we then have that θ has an extension $\tilde{\theta} \in \mathrm{Irr}(G)$. Now, by Gallagher's corollary [Isaacs 2006, Corollary (6.17)], we have that $\chi = \beta\tilde{\theta}$, for some $\beta \in \mathrm{Irr}(G/L)$. Now if $L \neq 1$, then the theorem holds for G/L, whence we have that β_{PL} is the sum of a p'-degree irreducible character β^* of PL/L (and hence linear) plus some character Δ of PL/L such that all of its irreducible constituents have degree divisible by p, or $\Delta = 0$. Then

$$\chi_{LP} = (\beta^*)\tilde{\theta}_{LP} + \Delta\tilde{\theta}_{LP},$$

and, using Gallagher's corollary, we see that we are done again. Hence $L = 1$, as desired.

(c) Now we can show that $p = 3$, $E := \boldsymbol{F}^*(G) = \boldsymbol{E}(G) = S_1 \times \cdots \times S_n$ with $S_i \cong \mathrm{PSL}_2(q_i)$ for some $q_i = 3^{3^{a_i}}$, a_i, $n \geq 1$, and $E \lhd G \leq \mathrm{Aut}(E)$. Indeed, suppose that $K := \boldsymbol{O}_p(G) \neq 1$. Since $|G/K| < |G| = |G/L|$, the first statement of the theorem holds for (G, K) and so for (G, L) as well (since $KP = P$), contradicting the choice of (G, L). Thus $\boldsymbol{O}_p(G) = 1$. Since $\boldsymbol{O}_{p'}(G) = 1$, we have now that $\boldsymbol{F}(G) = 1$ and so $E = \boldsymbol{F}^*(G) = \boldsymbol{E}(G)$. Next, $\boldsymbol{Z}(E) \leq \boldsymbol{F}(G) = 1$, whence $E = S_1 \times \cdots \times S_n$ is a direct product of nonabelian simple groups and $\boldsymbol{C}_G(E) = 1$, yielding $E \lhd G \leq \mathrm{Aut}(E)$.

Since $\boldsymbol{N}_G(P) = P$ and p is odd, we have that $p = 3$ and $S_i \cong \mathrm{PSL}_2(q_i)$ with $q_i = 3^{3^{a_i}}$ by the main result of [Guralnick et al. 2004].

(d) Let $Q := P \cap E \in \mathrm{Syl}_p(E)$ and write $Q = Q_1 \times \cdots \times Q_n$ with $Q_i \in \mathrm{Syl}_p(S_i)$. Since P is self-normalizing in EP, by [Navarro et al. 2007, Lemma 2.1(ii)], $\boldsymbol{C}_{\boldsymbol{N}_E(Q)/Q}(P) = 1$. This in turn implies by Lemma 4.1 that $\boldsymbol{C}_{\boldsymbol{N}_{S_i}(Q_i)/Q_i}(P_i) = 1$ for $P_i := \boldsymbol{N}_P(S_i)$. It follows by Lemma 3.1(ii) that P_i must induce the full subgroup $C_{3^{a_i}}$ of field automorphisms of S_i. Applying Lemma 3.1(iii) to S_i, we see that the P_i-invariant irreducible characters of p'-degree of S_i are $\alpha_i := 1_{S_i}$ and the two Weil characters $\eta_i^{\pm}$ of degree $(q_i - 1)/2$. Furthermore, for each $\alpha \in \{\alpha_i, \eta_i^{\pm}\}$, P_i fixes a unique irreducible constituent α^* of α_{Q_i}, occurring with multiplicity one. Moreover, the map $\alpha \mapsto \alpha^*$ is a bijection between the set of irreducible P_i-invariant characters of p'-degree of S_i and that of Q_i.

(e) Since the theorem holds for (G, E),

$$\chi_{EP} = \chi^* + \Delta,$$

where $\chi^* \in \mathrm{Irr}_{p'}(EP)$ and all the irreducible constituents of Δ (if any) have degree divisible by p. In particular, $\theta := (\chi^*)_E$ is irreducible. Write

$$\theta = \theta_1 \times \cdots \times \theta_n,$$

with $\theta_i \in \mathrm{Irr}_{p'}(S_i)$. Since θ is P-invariant, it follows that θ_i is P_i-invariant of p'-degree, and so $\theta_i \in \{\alpha_i, \eta_i^{\pm}\}$ by (d). As mentioned above,

$$(\theta_i)_{Q_i} = \theta_i^* + \delta_i,$$

where $\theta_i^* \in \mathrm{Irr}(Q_i)$ is P_i-invariant and δ_i is a sum of non-P_i-invariant irreducible characters of Q_i. Setting

$$\tilde{\theta} := \theta_1^* \times \cdots \times \theta_n^*,$$

we see that each irreducible constituent of $\theta_Q - \tilde{\theta}$ is non-P-invariant and so must lie under an irreducible character of P of degree divisible by p. But $p \nmid \theta(1)$. Hence θ_P contains a unique linear constituent which lies above $\tilde{\theta}$. Denote this constituent by θ^*. We have shown that every irreducible constituent of $\theta_P - \theta^* = (\chi^*)_P - \theta^*$ is of degree divisible by p, whereas $\theta^*(1) = 1$.

(f) It remains to show that every irreducible constituent of Δ_P has degree divisible by p. Assume the contrary: Δ_P contains a linear constituent λ, and write

$$\lambda_Q = \lambda_1 \times \cdots \times \lambda_n,$$

with $\lambda_i \in \mathrm{Irr}(Q_i)$. Let $\gamma \in \mathrm{Irr}(EP)$ be an irreducible constituent of Δ that contains λ upon restriction to P. Also, let

$$\beta = \beta_1 \times \cdots \times \beta_n \in \mathrm{Irr}(E)$$

lie under γ and above λ_Q. Since $E \lhd G$ and both θ and β are irreducible constituents of χ_E, β is G-conjugate to θ. Recall that $\theta = \theta_1 \times \cdots \times \theta_n$, with $\theta_i \in \{\alpha_i, \eta_i^{\pm}\}$. Also, note that the set $\{\alpha_i, \eta_i^{\pm}\}$ is $\mathrm{Aut}(S_i)$-invariant. It follows that $\beta_i \in \{\alpha_i, \eta_i^{\pm}\}$. As mentioned in (d), $(\beta_i)_{Q_i}$ contains a unique P_i-invariant irreducible constituent β_i^*, and each irreducible constituent of $(\beta_i)_{Q_i} - \beta_i^*$ is non-P_i-invariant. Denoting

$$\tilde{\beta} := \beta_1^* \times \cdots \times \beta_n^*,$$

we see that no irreducible constituent of $\beta_Q - \tilde{\beta}$ can be invariant under P. But λ_Q lies under β_Q and is P-invariant. Hence $\lambda_Q = \tilde{\beta}$ and $\lambda_i = \beta_i^*$.

(g) Now we consider two cases.

Case 1: β is not P-invariant. In this case, there is some $g \in P$ such that $\beta^g \neq \beta$. Then β^g lies above $(\lambda_Q)^g = \lambda_Q$ and under γ. Writing $\beta^g = \beta_1' \times \cdots \times \beta_n'$ and arguing as in (f), we see that $\beta_i' \in \{\alpha_i, \eta_i^{\pm}\}$ and, moreover,

$$\beta_i^* = \lambda_i = (\beta_i')^*.$$

As mentioned in (d), the map $\alpha \mapsto \alpha^*$ is a bijection. It follows that $\beta_i = \beta_i'$ and so $\beta = \beta^g$, a contradiction.

Case 2: β is P-invariant. Then, by Corollary 2.4, β extends to $\hat{\beta} \in \mathrm{Irr}(EP)$. Since γ lies above β, by Gallagher's corollary we have that $\gamma = \hat{\beta}\mu$, where $\mu \in \mathrm{Irr}(P/Q)$ is considered as a character of EP/E. Note that $p \mid \gamma(1)$, as γ is an irreducible constituent of Δ. On the other hand, $p \nmid \hat{\beta}(1) = \beta(1)$. It follows that $p \mid \mu(1)$. As shown in (f), no irreducible constituent of

$$\hat{\beta}_Q - \lambda_Q = \beta_Q - \lambda_Q$$

can be P-invariant. Hence λ is the unique linear constituent of $\hat{\beta}_P$. Certainly, $\mu\lambda$ is irreducible over P and nonlinear. Furthermore, again as shown in (f), every irreducible constituent of

$$(\gamma_P - \mu\lambda)_Q = \mu(1) \cdot (\beta_Q - \lambda_Q)$$

is non-P-invariant and so must lie under an irreducible P-character of degree divisible by p. Thus the degree of every irreducible constituent of $\gamma_P - \mu\lambda$ is divisible by p, and the same is true for $\mu\lambda \in \mathrm{Irr}(P)$. Consequently, the linear character λ cannot be a constituent of γ_P, again a contradiction.

Thus we have completed the proof of the first statement of the theorem.

(II) Now we prove that our map $\chi \mapsto \chi^*$ is a bijection. Recall that $\boldsymbol{O}_{p'}(\boldsymbol{N}_G(P)) \leq \boldsymbol{O}_{p'}(G)$ by Theorem 3.2 and that $\boldsymbol{O}_{p'}(G)$ is contained in the kernel of any $\psi \in \mathrm{Irr}(B_0(G))$. Modding out by $\boldsymbol{O}_{p'}(G)$, we may assume that $\boldsymbol{O}_{p'}(G) = 1$ and so $\boldsymbol{N}_G(P) = P$. Hence the principal block is the only block of maximal defect of G, and the same is true for LP. Since all the nonabelian composition factors of G

of order divisible by p are $\mathrm{PSL}_2(3^{3^a})$ with $a \geq 1$, we know by [Isaacs et al. 2007, Theorem A] that the McKay conjecture is true for G and for LP. Hence

$$|\operatorname{Irr}_{p'}(G)| = |\operatorname{Irr}_{p'}(\boldsymbol{N}_G(P))| = |\operatorname{Irr}(P/P')| = |\operatorname{Irr}_{p'}(\boldsymbol{N}_{LP}(P))| = |\operatorname{Irr}_{p'}(LP)|.$$

Now, if $\delta \in \operatorname{Irr}_{p'}(LP)$, then some irreducible constituent χ of δ^G has p'-degree. Therefore χ_{LP} contains δ and, by the first statement of the theorem, we necessarily have that $\chi^* = \delta$. Thus the map $\chi \mapsto \chi^*$ is surjective, and therefore injective. □

The proof of Corollary B, which we restate below, is now immediate.

Corollary B. *Let G be a finite group, let p be odd, and let $P \in \operatorname{Syl}_p(G)$. Suppose that $\boldsymbol{N}_G(P) = P$. Then there is a natural bijection $\chi \mapsto \chi^*$ between $\operatorname{Irr}_{p'}(G)$ and the linear characters of P. In fact, if $\chi \in \operatorname{Irr}_{p'}(G)$ and $\lambda \in \operatorname{Irr}(P)$ is linear, then χ and λ correspond under the bijection if and only if*

$$\chi_P = \lambda + \Delta,$$

where Δ is either zero or a character whose irreducible constituents all have degree divisible by p. This happens if and only if

$$\lambda^G = \chi + \Xi,$$

where Ξ is either zero or a character whose irreducible constituents all have degree divisible by p.

Proof. If $\boldsymbol{N}_G(P) = P$, then the principal block is the unique block of maximal defect by Brauer's first main theorem. Hence, the first part of the corollary follows from Theorem 4.3 by letting $L = 1$. For the second part, if $\lambda \in \operatorname{Irr}(P)$ is linear, then λ^G has degree not divisible by p, and therefore λ^G has a constituent $\chi \in \operatorname{Irr}_{p'}(G)$. Then $[\chi_P, \lambda] \neq 0$ and it follows that necessarily $\lambda = \chi^*$. It also follows that χ is unique, because our map is injective. □

Next is Corollary C.

Corollary C. *Let G be a finite group, let p be odd, and let $P \in \operatorname{Syl}_p(G)$.*

(a) *$\boldsymbol{N}_G(P) = P$ if and only if*

$$(1_P)^G = 1_G + \Xi,$$

where Ξ is either zero or a character whose irreducible constituents all have degree divisible by p.

(b) *$\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$ if and only 1_G is the only irreducible constituent of $(1_{P\boldsymbol{C}_G(P)})^G$ that belongs to $\operatorname{Irr}_{p'}(B_0(G))$.*

Proof. Both proofs are very similar. We start with (a). One implication follows from Corollary B. Assume now that

$$(1_P)^G = 1_G + \Xi,$$

where Ξ is either zero or a character whose irreducible constituents all have degree divisible by p, but $N := \boldsymbol{N}_G(P) > P$. Then there exists a nonprincipal character $\gamma \in \mathrm{Irr}(N/P)$, which can be viewed as an N-character. Since γ has p'-degree (because N/P is a p'-group), it follows that γ^G possesses an irreducible constituent $\chi \in \mathrm{Irr}_{p'}(G)$. Now, χ lies over $\gamma \neq 1_N$ and therefore $1_G \neq \chi$ lies over 1_P, a contradiction.

Next, we prove (b). Write $C = \boldsymbol{C}_G(P)$. One implication follows from Theorem 4.3. Assume now that 1_G is the unique irreducible constituent of $(1_{PC})^G$ that belongs to $\mathrm{Irr}_{p'}(B_0(G))$ and that $N := \boldsymbol{N}_G(P) > PC$. Then there exists a nonprincipal character $\gamma \in \mathrm{Irr}(N/PC)$, which can be viewed as an N-character. Since N is p-solvable, and $\boldsymbol{O}_{p'}(N) \leq C \leq \ker\gamma$, it follows that γ lies in the principal block of N by [Navarro 1998, Theorem (10.20)]. Also, γ has p'-degree, because N/PC is a p'-group. If b is now the principal block of N, we know that $b^G = B = B_0(G)$ is the principal block of G, by Brauer's third main theorem [loc. cit., Theorem (6.7)]. Write

$$(\gamma^G)_B = \sum_{\chi \in \mathrm{Irr}(B)} [\gamma^G, \chi]\chi.$$

(This is called the *B-part* of γ^G; see page 72 of [loc. cit.].) Now, by [loc. cit., Corollary (6.4)], we have that

$$1 = (\gamma^G(1))_p = ((\gamma^G)_B(1))_p,$$

where n_p is the largest power of p dividing the integer n. It then follows that some irreducible constituent χ of γ^G lies in $\mathrm{Irr}_{p'}(B)$. We now have that χ lies over γ and therefore over 1_{PC}. Since $\gamma \neq 1_N$, it follows that $\chi \neq 1_G$, and this is a contradiction. □

5. p-solvable groups

Our proof of Theorem D is short but uses deep character theory of p-solvable groups. We assume that the reader is familiar with π-special characters (i.e., the characters of π-degree whose subnormal irreducible constituents have determinantal π-order; see [Gajendragadkar 1979]).

Lemma 5.1. *Suppose that $L \lhd G$, $P \in \mathrm{Syl}_p(G)$ and $\boldsymbol{N}_{G/L}(PL/L) = PL/L$. Assume that G/L is p-solvable. Let $\theta \in \mathrm{Irr}(L)$ be P-invariant and p'-special. Then there exists a unique $\hat{\theta} \in \mathrm{Irr}(G \mid \theta)$ such that $\hat{\theta}$ is p'-special.*

Proof. We argue by induction on $|G:L|$. Let K/L be a chief factor of G, and notice that G/K has a self-normalizing Sylow p-subgroup, by elementary group theory. Assume first that K/L is a p-group, and let $\eta \in \mathrm{Irr}(K \mid \theta)$ be the unique p'-special character lying over θ, by using Proposition 4.3 of [Gajendragadkar 1979]. By uniqueness, η is P-invariant, and by induction, there is a unique p'-special character $\hat{\eta} \in \mathrm{Irr}(G)$ that lies over η (and therefore over θ). Now, if $\hat{\theta}$ is any other p'-special character of G lying over θ and $\psi \in \mathrm{Irr}(K)$ lies under $\hat{\theta}$ and over θ, we have that ψ is p'-special by Proposition 4.1 of [Gajendragadkar 1979], and therefore $\psi = \eta$, by uniqueness. But in this case, $\hat{\theta} = \hat{\eta}$, by using the inductive hypothesis.

Suppose finally that K/L is a p'-group. Then $\boldsymbol{C}_{K/L}(PL/L) = 1$, using that PL/L is self-normalizing. Hence, by Problem (13.10) of [Isaacs 2006], there exists a unique P-invariant $\tau \in \mathrm{Irr}(K \mid \theta)$. Also, τ is p'-special by Lemma 4.4 of [Gajendragadkar 1979]. By induction, there exists a unique p'-special character $\hat{\tau}$ lying over τ (and therefore over θ). Suppose now that $\gamma \in \mathrm{Irr}(G)$ is any other p'-special character lying over θ. By Lemma 2.1, let $\phi \in \mathrm{Irr}(K)$ be P-invariant under γ, and, by Theorem (13.27) of [Isaacs 2006], let $\rho \in \mathrm{Irr}(L)$ be P-invariant under ϕ. Then ρ and θ are P-invariant and lie under γ, so $\rho = \theta$ by Lemma 2.1. Then $\phi = \tau$ by the uniqueness of τ, and hence $\gamma = \hat{\tau}$ by induction. □

We restate Theorem D for the reader's convenience.

Theorem D. *Let G be a finite p-solvable group, and let $P \in \mathrm{Syl}_p(G)$. Suppose that $\boldsymbol{N}_G(P) = P\boldsymbol{C}_G(P)$, and let $N = \boldsymbol{O}_{p'}(G)$. Let $\mathrm{Irr}_P(N)$ be the set of P-invariant $\theta \in \mathrm{Irr}(N)$. Then for every $\theta \in \mathrm{Irr}_P(N)$ and linear $\lambda \in \mathrm{Irr}(P/P')$, there is a canonically defined character*

$$\lambda \star \theta \in \mathrm{Irr}_{p'}(G).$$

Furthermore, the map

$$\mathrm{Irr}(P/P') \times \mathrm{Irr}_P(N) \to \mathrm{Irr}_{p'}(G)$$

given by $(\lambda, \theta) \mapsto \lambda \star \theta$ is a bijection. As a consequence, $\boldsymbol{N}_G(P) = P \times \boldsymbol{C}_N(P)$, and if $\theta^* \in \mathrm{Irr}(\boldsymbol{C}_N(P))$ is the Glauberman correspondent of $\theta \in \mathrm{Irr}_P(N)$, then the map

$$\lambda \times \theta^* \mapsto \lambda \star \theta$$

is a natural bijection $\mathrm{Irr}_{p'}(\boldsymbol{N}_G(P)) \to \mathrm{Irr}_{p'}(G)$. Also, if $\theta = 1_N$ and $\lambda \in \mathrm{Irr}(P/P')$, then $\lambda \times \theta^*$ is the unique linear constituent of $(\lambda \star \theta)_{\boldsymbol{N}_G(P)}$.

Proof. By Theorem 3.2, we can write $\boldsymbol{N}_G(P) = P \times X$, where $X := \boldsymbol{C}_N(P)$. Let $\lambda \in \mathrm{Irr}(P)$ be linear and let $\theta \in \mathrm{Irr}_P(N)$. Since $P \cap N = 1$, we trivially have that λ extends to PN. Now, by Theorem 2.1 of [Isaacs and Navarro 2008] (or see Corollary 2.2 of [Isaacs and Navarro 2001] for a self-contained proof), there exists a maximal subgroup $P \subseteq W \subseteq G$ such that λ extends to W. Hence $PN \subseteq W$. Now, by elementary character theory, let $\hat{\lambda} \in \mathrm{Irr}(W)$ be the unique linear character of p-power order that extends λ. Now, $\boldsymbol{N}_{W/N}(PN/N) = PN/N$, and by Lemma 5.1, there exists a unique p'-special $\hat{\theta} \in \mathrm{Irr}(W)$ lying over θ. Now, by Theorem 2.2 of [Isaacs and Navarro 2008] and Theorem C of [Navarro 1997] we have that

$$\lambda \star \theta := (\hat{\theta}\hat{\lambda})^G \in \mathrm{Irr}(G).$$

Notice that $\lambda \star \theta$ has p'-degree, because $\hat{\theta}$ has p'-degree and $|G : W|$ is not divisible by p. (We notice for the record that $(\lambda \star \theta)_W$ contains $\hat{\theta}\hat{\lambda}$, and therefore, when restricted to N, we have that $(\lambda \star \theta)$ lies over θ. It is not in general true that $\lambda \star \theta$ lies over λ, on the other hand.)

We have now defined a map

$$\mathrm{Irr}(P/P') \times \mathrm{Irr}_P(N) \to \mathrm{Irr}_{p'}(G)$$

given by $(\lambda, \theta) \mapsto \lambda \star \theta$.

Next we show that our map is surjective. Let $\chi \in \mathrm{Irr}_{p'}(G)$. By Theorem 3.6 of [Isaacs and Navarro 2008], we have that χ is a *satellite* of some $\psi \in B_p(G)$ of p'-degree (see Section 3 of [Isaacs and Navarro 2008] for the necessary definitions). In other words, this means that there is some linear character $\delta \in \mathrm{Irr}(P)$ and a p'-special character $\alpha \in \mathrm{Irr}(U)$, where U is the maximal subgroup of G to which δ extends, such that

$$\chi = (\hat{\delta}\alpha)^G,$$

where the order of $\hat{\delta}$ is a p-power and $\hat{\delta}$ extends δ. Now, α_N contains a (unique) P-invariant character $\mu \in \mathrm{Irr}_P(N)$ by Lemma 2.1, and it follows that α is the unique p'-special character of U lying over μ by Lemma 5.1. It follows then that $\chi = \delta \star \mu$, and, therefore, that our map is surjective.

Recall that the Glauberman correspondence [Isaacs 2006, Theorem (13.1)] provides a natural bijection

$$\mathrm{Irr}_P(N) \to \mathrm{Irr}(\boldsymbol{C}_N(P)).$$

Since the McKay conjecture is true for p-solvable groups (see for instance [Isaacs et al. 2007]) we have that

$$|\mathrm{Irr}_{p'}(G)| = |\mathrm{Irr}_{p'}(\boldsymbol{N}_G(P))| = |\mathrm{Irr}(P/P')||\mathrm{Irr}(\boldsymbol{C}_N(P))| = |\mathrm{Irr}(P/P')||\mathrm{Irr}_P(N)|.$$

It then follows that our map is bijective.

In the case where $\theta = 1_N$, the second part of the theorem easily follows from Theorem 3.1 of [Isaacs and Navarro 2008] applied in the group G/N. □

Under the hypothesis of the previous theorem, we notice that the blocks with defect group $P \in \mathrm{Syl}_p(G)$ of G can be parametrized by the P-invariant irreducible characters of $N = \boldsymbol{O}_{p'}(G)$. The fact that in this case $\mathrm{Irr}_{p'}(G \mid \theta)$ and $\mathrm{Irr}_{p'}(\boldsymbol{N}_G(P) \mid \theta^*)$

have the same cardinality is a consequence of [Okuyama and Wajima 1980]. Our hypothesis, however, allows us to obtain a canonical bijection in our case.

6. A non-*p*-solvable example

To finish the paper, it might be interesting to show the reader how to construct a natural bijection $\mathrm{Irr}_{p'}(G) \to \mathrm{Irr}_{p'}(\boldsymbol{N}_G(P))$ in the paradigmatic case where $G = \mathrm{SL}_2(3^{3^a}) \cdot C_{3^a}$ and $p = 3$, with $P \in \mathrm{Syl}_p(G)$ as usual.

Let $S := [G, G] = \mathrm{SL}_2(q)$ with $q = 3^{3^a}$, and let σ denote the field automorphism of S of order $t := 3^a$, so that $G = S \rtimes \langle\sigma\rangle$. Using [Digne and Michel 1991, Table 2], it is easy to check that $\mathrm{Irr}_{p'}(S)$ contains exactly six σ-invariant characters: 1_S, two Weil characters $\eta_{1,2}$ of degree $(q-1)/2$ (denoted $\chi_b^{\pm}$ in [Isaacs et al. 2007, §15]), two Weil characters $\xi_{1,2}$ of degree $(q+1)/2$ (denoted $\chi_a^{\pm}$ in that work), and a unique rational-valued character ψ of degree $q-1$. Here, the three former characters are nonfaithful, and the three latter ones are faithful. Furthermore, one can label $\xi_{1,2}$ such that

$$\xi_i(x) = \eta_i(x) + 1$$

for any element $x \in S$ of order 3 and $i = 1, 2$. Since G/S is cyclic (and generated by σ), it follows that all these 6 characters extend to G, and the $6t$ extensions are precisely the characters in $\mathrm{Irr}_{p'}(G)$. In particular, 1_S extends to λ_j, $1 \leq j \leq t$, with $\lambda_1 = 1_G$. Next, we will single out a "canonical" extension for each of the remaining five characters of S. As shown in [Navarro and Tiep 2014, §3], G embeds in $H := \mathrm{Sp}_{2t}(3)$ in such a way that η_i extends to a Weil character of H that takes value 1 at σ. We will denote the restriction of this character of H to G by $\tilde{\eta}_i$, so that

$$\tilde{\eta}_i(\sigma) = 1, \quad i = 1, 2.$$

Likewise, ξ_i extends to a Weil character of H that takes value 2 at σ, and we will denote the restriction of this character of H to G by $\tilde{\xi}_i$, so that

$$\tilde{\xi}_i(\sigma) = 2, \quad i = 1, 2.$$

Finally, by [Navarro and Tiep 2008, Corollary 2.2], there is a unique rational-valued extension $\tilde{\psi}$ of ψ to G.

Let 1_Z and ν denote the two linear characters of $Z := \boldsymbol{Z}(G) \cong C_2$. For any $\gamma \in \mathrm{Irr}(Z)$, let $\mathrm{Irr}_{p'}(G \mid \gamma)$ denote the set of characters $\chi \in \mathrm{Irr}_{p'}(G)$ that lie above γ, and similarly for $N := \boldsymbol{N}_G(P) = P \times Z$. Now we see that

$$\begin{aligned}\mathrm{Irr}_{p'}(G \mid 1_Z) &= \{\lambda_j, \tilde{\eta}_i\lambda_j \mid 1 \leq i \leq 2, 1 \leq j \leq t\},\\ \mathrm{Irr}_{p'}(G \mid \nu) &= \{\tilde{\psi}\lambda_j, \tilde{\xi}_i\lambda_j \mid 1 \leq i \leq 2, 1 \leq j \leq t\}.\end{aligned}$$

Moreover, the first set is contained in the principal 3-block $B_0(G)$ of G and the second set is contained in the other 3-block of maximal defect $B_1(G)$ of G. Theorem A yields a natural correspondence $\mathrm{Irr}_{p'}(B_0(G)) \to \mathrm{Irr}_{p'}(B_0(N))$. To get a natural correspondence $\mathrm{Irr}_{p'}(B_1(G)) \to \mathrm{Irr}_{p'}(B_1(N))$, it therefore suffices to define a natural correspondence between $\mathrm{Irr}_{p'}(G \mid 1_Z) = \mathrm{Irr}_{p'}(B_0(G))$ and $\mathrm{Irr}_{p'}(G \mid \nu) = \mathrm{Irr}_{p'}(B_1(G))$, which can be given by

$$\lambda_j \mapsto \tilde{\psi}\lambda_j, \quad \tilde{\eta}_i\lambda_j \mapsto \tilde{\xi}_i\lambda_j,$$

and a natural correspondence between $\mathrm{Irr}_{p'}(N \mid 1_Z) = \mathrm{Irr}_{p'}(B_0(N))$ and $\mathrm{Irr}_{p'}(N \mid \nu) = \mathrm{Irr}_{p'}(B_1(N))$, which can be given by

$$\mu \times 1_Z \mapsto \mu \times \nu$$

for all $\mu \in \mathrm{Irr}(P/P')$.

Note that an equivariant bijection $\pi : \mathrm{Irr}_{p'}(S) \to \mathrm{Irr}_{p'}(\boldsymbol{N}_S(P \cap S))$ was constructed in [Isaacs et al. 2007, (15F)]. Choosing $\pi(\chi_a^{\pm})$ and $\pi(\chi_b^{\pm})$ suitably, one can check that π extends (from S to G) to our bijection $\mathrm{Irr}_{p'}(G) \to \mathrm{Irr}_{p'}(\boldsymbol{N}_G(P))$.

Acknowledgements

The authors are grateful to the referee for careful reading and helpful comments that greatly improved the exposition of the paper.

References

[Digne and Michel 1991] F. Digne and J. Michel, *Representations of finite groups of Lie type*, London Mathematical Society Student Texts **21**, Cambridge University Press, 1991. MR 92g:20063 Zbl 0815.20014

[Gajendragadkar 1979] D. Gajendragadkar, "A characteristic class of characters of finite π-separable groups", *J. Algebra* **59**:2 (1979), 237–259. MR 82b:20012 Zbl 0426.20007

[Guralnick et al. 2004] R. M. Guralnick, G. Malle, and G. Navarro, "Self-normalizing Sylow subgroups", *Proc. Amer. Math. Soc.* **132**:4 (2004), 973–979. MR 2004m:20043 Zbl 1049.20010

[Huppert 1967] B. Huppert, *Endliche Gruppen, I*, Die Grundlehren der Mathematischen Wissenschaften **134**, Springer, Berlin-New York, 1967. MR 37 #302 Zbl 0217.07201

[Isaacs 2006] I. M. Isaacs, *Character theory of finite groups*, AMS Chelsea, Providence, RI, 2006. MR 2270898 Zbl 1119.20005

[Isaacs 2008] I. M. Isaacs, *Finite group theory*, Graduate Studies in Mathematics **92**, Amer. Math. Soc., Providence, RI, 2008. MR 2009e:20029 Zbl 1169.20001

[Isaacs and Navarro 2001] I. M. Isaacs and G. Navarro, "Characters of p'-degree of p-solvable groups", *J. Algebra* **246**:1 (2001), 394–413. MR 2002m:20011 Zbl 0998.20008

[Isaacs and Navarro 2008] I. M. Isaacs and G. Navarro, "Character sums and double cosets", *J. Algebra* **320**:10 (2008), 3749–3764. MR 2009i:20019 Zbl 1189.20012

[Isaacs et al. 2007] I. M. Isaacs, G. Malle, and G. Navarro, "A reduction theorem for the McKay conjecture", *Invent. Math.* **170**:1 (2007), 33–101. MR 2008h:20016 Zbl 1138.20010

[Kondratiev and Mazurov 2003] A. S. Kondratiev and V. D. Mazurov, "2-signalizers of finite simple groups", *Algebra Logika* **42**:5 (2003), 594–623. In Russian; translated in *Algebra Logic* **42**:5 (2003), 333–348. MR 2004j:20024 Zbl 1067.20015

[Malle and Navarro 2012] G. Malle and G. Navarro, "Characterizing normal Sylow p-subgroups by character degrees", *J. Algebra* **370** (2012), 402–406. MR 2966846 Zbl 1277.20010

[Navarro 1997] G. Navarro, "New properties of the π-special characters", *J. Algebra* **187**:1 (1997), 203–213. MR 97m:20008 Zbl 0890.20010

[Navarro 1998] G. Navarro, *Characters and blocks of finite groups*, London Mathematical Society Lecture Note Series **250**, Cambridge University Press, 1998. MR 2000a:20018 Zbl 0903.20004

[Navarro 2003] G. Navarro, "Linear characters of Sylow subgroups", *J. Algebra* **269**:2 (2003), 589–598. MR 2004m:20019 Zbl 1037.20007

[Navarro 2004] G. Navarro, "The McKay conjecture and Galois automorphisms", *Ann. of Math.* (2) **160**:3 (2004), 1129–1140. MR 2005m:20022 Zbl 1079.20010

[Navarro and Späth 2014] G. Navarro and B. Späth, "On Brauer's height zero conjecture", *J. Eur. Math. Soc.* (*JEMS*) **16**:4 (2014), 695–747. MR 3191974 Zbl 06293942

[Navarro and Tiep 2008] G. Navarro and P. H. Tiep, "Rational irreducible characters and rational conjugacy classes in finite groups", *Trans. Amer. Math. Soc.* **360**:5 (2008), 2443–2465. MR 2008k:20014 Zbl 1137.20009

[Navarro and Tiep 2014] G. Navarro and P. H. Tiep, "Brauer characters and rationality", *Math. Z.* **276**:3-4 (2014), 1101–1112. MR 3175172 Zbl 06304104

[Navarro et al. 2007] G. Navarro, P. H. Tiep, and A. Turull, "p-rational characters and self-normalizing Sylow p-subgroups", *Represent. Theory* **11** (2007), 84–94. MR 2008a:20018 Zbl 1146.20005

[Okuyama and Wajima 1980] T. Okuyama and M. Wajima, "Character correspondence and p-blocks of p-solvable groups", *Osaka J. Math.* **17**:3 (1980), 801–806. MR 82a:20013 Zbl 0446.20003

[Späth 2013] B. Späth, "A reduction theorem for the Alperin–McKay conjecture", *J. Reine Angew. Math.* **680** (2013), 153–189. MR 3100954 Zbl 1283.20006

Communicated by David Benson
Received 2014-01-31 Revised 2014-06-17 Accepted 2014-08-25

gabriel.navarro@uv.es *Departamento de Álgebra, Universitat de València,*
Doctor Moliner, 50, 46100 Burjassot, València, Spain

tiep@math.arizona.edu *Department of Mathematics, University of Arizona, 617*
North Santa Rita Avenue, Tucson, AZ 85721, United States

carolina.vallejo@uv.es *Departamento de Álgebra, Universitat de València,*
Doctor Moliner, 50, 46100 Burjassot, València, Spain

ALGEBRA AND NUMBER THEORY 8:8 (2014)
dx.doi.org/10.2140/ant.2014.8.1857

Quantum matrices by paths

Karel Casteels

We study, from a combinatorial viewpoint, the *quantized coordinate ring of $m \times n$ matrices* $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$ over an infinite field $\mathbb{K}$ (often simply called *quantum matrices*).The first part of this paper shows that $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$, which is traditionally defined by generators and relations, can be seen as a subalgebra of a quantum torus by using paths in a certain directed graph. Roughly speaking, we view each generator of $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$ as a sum over paths in the graph, each path being assigned an element of the quantum torus. The $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$ relations then arise naturally by considering intersecting paths. This viewpoint is closely related to Cauchon's deleting derivations algorithm.

The second part of this paper applies the above to the theory of torus-invariant prime ideals of $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$. We prove a conjecture of Goodearl and Lenagan that all such prime ideals, when the quantum parameter q is a non-root of unity, have generating sets consisting of quantum minors. Previously, this result was known to hold only when $\text{char}(\mathbb{K}) = 0$ and with q transcendental over $\mathbb{Q}$. Our strategy is to prove the stronger result that the quantum minors in a given torus-invariant ideal form a Gröbner basis.

1. Introduction

The purpose of this paper is to introduce a "combinatorial model" of $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$, the quantized coordinate ring of $m \times n$ matrices over a field $\mathbb{K}$ (simply called *quantum matrices*). We demonstrate the utility of this model by using it to study the prime spectrum of $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$.

MSC2010: primary 16T20; secondary 16T30.

Keywords: quantum algebra, quantum groups, quantum matrices, combinatorial noncommutative algebra, noncommutative Gröbner bases.

Quantum matrices have generated a good deal of interest since their discovery during the initial development of quantum group theory in the 1980s. This is because not only do quantum matrices underlie many of the traditional quantum groups such as the quantum special and general linear groups, but there are also interesting connections with topics such as braided tensor categories and knot theory. See [Takeuchi 2002] for a brief survey. More recently, it has been observed [Goodearl et al. 2011a; 2011b; Launois and Lenagan 2009] that the prime spectrum of quantum matrices is deeply related to the theory of totally nonnegative matrices and the *totally nonnegative grassmannian* in the sense of [Postnikov 2006].

Since the late 1990s, much effort has been expended toward understanding the structure of the prime and primitive spectra of various quantum algebras. Quantum matrices have received particular attention since, while this algebra has a seemingly simple structure (for example, it is an iterated Ore extension over the field $\mathbb{K}$), many problems have proven difficult to resolve. In particular, the machinery employed to analyze $\mathrm{Spec}(\mathbb{O}_q(\mathcal{M}_{m,n}(\mathbb{K})))$ has tended to use fairly sophisticated viewpoints from noncommutative ring theory and representation theory, and even then often requires extra restrictions on the base field $\mathbb{K}$ and choice of quantum parameter q.

The $\mathcal{H}$-stratification theory of [Goodearl and Letzter 2000] (see also [Brown and Goodearl 2002]) is an important advancement toward understanding the prime and primitive spectra of some quantum algebras. Briefly, many noncommutative rings support a rational action of a torus $\mathcal{H}$ which allows one to partition the prime spectrum of the ring into finitely many *$\mathcal{H}$-strata*, each $\mathcal{H}$-stratum being homeomorphic (with respect to the usual Zariski topology) to the prime spectrum of a Laurent polynomial ring in finitely many commuting indeterminates, and each containing a unique $\mathcal{H}$-invariant prime ideal. Moreover, the primitive ideals of the algebra are precisely those that are maximal within their $\mathcal{H}$-stratum. For these reasons, an important first step towards understanding the prime and primitive spectra is to first study the $\mathcal{H}$-invariant prime ideals, called *$\mathcal{H}$-primes.*

The deleting derivations algorithm of [Cauchon 2003a; 2003b] has also proven quite useful. Roughly speaking, this procedure shows that when the $\mathcal{H}$-stratification theory applies to a given quantum algebra, one can often embed the set of $\mathcal{H}$-primes into the set of $\mathcal{H}$-primes of a *quantum affine space.* This is convenient since quantum affine spaces are typically easy to handle thanks to results of Goodearl and Letzter [1998]. The strategy then is to reverse the deleting derivations procedure in order to transfer (more easily obtained) information about the quantum affine space back to information about the quantum algebra.

The $\mathcal{H}$-stratification and the deleting derivations theories both apply to quantum matrices in the generic case, i.e., when the parameter q is a non-root of unity, and so a natural problem is to find generating sets for the $\mathcal{H}$-primes. For 2×2 quantum matrices, this problem is fairly straightforward, yet even the 3×3 case required a significant amount of work by Goodearl and Lenagan [2002; 2003]. However, in all cases their generating sets consisted of *quantum minors*, and so it was conjectured that this held true in general. Launois [2004a; 2004b] was the first to prove this conjecture under the constraints $\mathbb{K} = \mathbb{C}$ and q transcendental over $\mathbb{Q}$. This was later extended to any $\mathbb{K}$ of characteristic zero [Goodearl et al. 2011a].

An important part of Cauchon's results is a parametrization of the $\mathcal{H}$-primes of quantum matrices using what are now known in the quantum algebra community as *Cauchon diagrams*. It turns out that a Cauchon diagram encodes fundamental information about the corresponding $\mathcal{H}$-stratum. For example, the Krull dimension can be easily calculated from the Cauchon diagram using the main result of [Bell et al. 2012]. Launois also described an algorithm to find the generators of a given $\mathcal{H}$-prime from its Cauchon diagram, but the calculations involved very quickly become unwieldy. A graph-theoretic interpretation of Launois' algorithm provided in [Casteels 2011] forms the starting point for some of the results presented below. In fact, much of Section 3.1 may be seen as a combinatorial interpretation of the deleting derivations algorithm.

It is notable that Cauchon diagrams arose independently in work of Postnikov [2006] in his investigations of the totally nonnegative Grassmannian. In this context, Cauchon diagrams are called ⅃-diagrams (also Le-diagrams) and have been investigated by several authors (see [Lam and Williams 2008] and [Talaska 2011] in particular). The connections between these two areas and Poisson geometry have been explored by Goodearl, Launois and Lenagan [Goodearl et al. 2011b; 2011a].

Finally, let us also mention that Yakimov [2010; 2013] has developed representation-theoretic methods with great success. In particular, he independently verified (and generalized) Goodearl and Lenagan's conjecture, but again, only under the constraint that $\mathrm{char}(\mathbb{K}) = 0$ and q transcendental over $\mathbb{Q}$. Furthermore, the generating sets obtained are actually smaller than Launois' in general. It is unclear how Yakimov's work relates to the viewpoint presented in this paper; however, recent work of Geiger and Yakimov [2014] explores the connections between Yakimov's work and Cauchon's, and so there is quite possibly a close relationship.

As will be reviewed in Section 2, the usual description of $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$ is by generators and relations. Our approach to $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$ is the focus of Section 3, where we begin by giving a directed graph and then assign elements ("weights") of a quantum torus to directed paths. We then discuss various subalgebras of the quantum torus generated by sums over path weights. In particular, Corollary 3.2.5 shows that quantum matrices can be so obtained. One nice aspect of this is that the quantum matrix relations naturally arise by considering intersecting paths (see the proofs of Theorem 3.1.12 and Theorem 3.2.3).

While at first it may appear that the description of quantum matrices "by paths" is a mere curiosity, it is in fact an indispensable tool in the bulk of this paper, Section 4.

Here, the Goodearl–Lenagan conjecture is an immediate corollary to a stronger result, Theorem 4.4.1, which states that for *any* infinite field $\mathbb{K}$ and non-root of unity $q \in \mathbb{K}^*$, the quantum minors in a given $\mathcal{H}$-prime form a Gröbner basis with respect to a certain term ordering. The difficulty with this approach is that for a given $\mathcal{H}$-prime of $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$, a priori we do not know any generating sets at all to which we can apply Buchberger's algorithm, so we must check that the minors form a Gröbner basis by direct verification of the definition. The way we do this is by using the strategy noted above for the deleting derivations algorithm. That is, we transfer an (easily obtained) Gröbner basis for an $\mathcal{H}$-prime in a quantum affine space to a Gröbner basis for an $\mathcal{H}$-prime in quantum matrices.

Finally, many nonstandard terms and notation have been invented for use in this paper. A combined index and glossary is provided in a List of terms and notation to assist the reader in more easily locating the definitions, should the need arise.

2. Quantum matrices

Let us first set some data, notation and conventions that are to be used throughout this paper:

- Fix an infinite field $\mathbb{K}$, integers $m, n \geq 2$, and a nonzero, non-root of unity $q \in \mathbb{K}$.
- For a positive integer k, we set $[k] = \{1, 2, \dots, k\}$.
- The set of $m \times n$ matrices with integer entries is denoted by $\mathcal{M}_{m,n}(\mathbb{Z})$. The set of $m \times n$ matrices with nonnegative integer entries is denoted by $\mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$.
- The (i, j)-entry of $N \in \mathcal{M}_{m,n}(\mathbb{Z})$ is denoted by $(N)_{i,j}$, and (i, j) is called the *coordinate* of this entry. In view of this, the elements of $[m] \times [n]$ are called coordinates.
- We often describe relative positions of coordinates using the usual meaning of terms such as north, northwest, etc. For example, (i, j) is *northwest* of (r, s) if $i < r$ and $j < s$, and *north* if $i < r$ and $j = s$.

The restriction $m, n \geq 2$ is made simply to avoid some inconveniences in various definitions that would occur if $m = 1$ or $n = 1$. Fortunately, it is already known that all results presented in this paper hold when $m = 1$ or $n = 1$, since in these cases all algebras in this paper reduce to quantum affine spaces, and such algebras can be dealt with using results of [Goodearl and Letzter 1998].

2.1. *The algebras $R^{(t)}$.*

Definition 2.1.1. The *lexicographic order* on $[m] \times [n]$ is the total order $<$ obtained by setting

$$(i, j) < (k, \ell) \iff i < k, \text{ or } i = k \text{ and } j < \ell.$$

If $(i, j) \in [m] \times [n]$, then $(i, j)^-$ denotes the largest element less than (i, j) with respect to the lexicographic order.

Note 2.1.2. Any reference in this paper relating to an ordering of the coordinates $[m] \times [n]$ is with respect to the lexicographic order.

The algebras in the next definition each have a set of generators indexed by $[m] \times [n]$. It is natural to place these generators as the entries of an $m \times n$ matrix that we call the *matrix of generators*.

Definition 2.1.3. Let $t \in [mn]$ and set (r, s) to be the t-th smallest coordinate. Define $R^{(t)}$ to be the $\mathbb{K}$-algebra with the $m \times n$ matrix of generators $X = [x_{i,j}]$ subject to the following relations. If

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is any 2×2 submatrix of X, then:

(1) $ab = qba,\ cd = qdc$;

(2) $ac = qca,\ bd = qdb$;

(3) $bc = cb$;

(4) $ad = \begin{cases} da & \text{if } d = x_{k,\ell} \text{ and } (k, \ell) > (r, s); \\ da + (q - q^{-1})bc & \text{if } d = x_{k,\ell} \text{ and } (k, \ell) \leq (r, s). \end{cases}$

Example 2.1.4. If $m = 2$, $n = 3$ and $t = 5$, then $(r, s) = (2, 2)$ and $R^{(5)}$ has matrix of generators

$$\begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \end{bmatrix}.$$

The relations corresponding to Part (4) of Definition 2.1.3 are

$$\begin{aligned} x_{1,1}x_{2,2} &= x_{2,2}x_{1,1} + (q - q^{-1})x_{1,2}x_{2,1}, \\ x_{1,1}x_{2,3} &= x_{2,3}x_{1,1}, \\ x_{1,2}x_{2,3} &= x_{2,3}x_{1,2}. \end{aligned}$$

The two extremities in the collection of $R^{(t)}$ are of the most interest to us.

Notation 2.1.5. With respect to the notation in Definition 2.1.3:

(1) If $t = 1$, then in Definition 2.1.3(4) we always have

$$ad = da.$$

We call this algebra $m \times n$ *quantum affine space*, denoted $\mathcal{O}_q(\mathbb{K}^{m \times n})$. The entries of the matrix of generators of $\mathcal{O}_q(\mathbb{K}^{m \times n})$ will often be labeled by $t_{i,j}$ for $(i, j) \in [m] \times [n]$.

(2) If $t = mn$, then in Definition 2.1.3(4) we always have

$$ad = da + (q - q^{-1})bc.$$

This algebra is the *quantized coordinate ring of $m \times n$ matrices over $\mathbb{K}$*, denoted by $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$ and simply referred to as the $(m \times n)$ *quantum matrices.*

(3) The localization of $R^{(1)} = \mathcal{O}_q(\mathbb{K}^{m\times n})$ with respect to the multiplicative set generated by the standard generators $t_{i,j}$ is called the $(m \times n)$ *quantum torus* $\mathcal{O}_q((\mathbb{K}^{\times})^{m\times n})$.

(4) Two elements $y, z \in R^{(t)}$ will be said to *q^*-commute* if there is an integer r such that $yz = q^r zy$. Note that commuting elements q^*-commute.

In later sections, we work intimately with monomials in the generators of $R^{(t)}$, so we here set some notation in this area. For the remainder of this section, fix $t \in [mn]$, and let $[x_{i,j}]$ be the matrix of generators for $R^{(t)}$.

Notation 2.1.6. If $N \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$, then we write

$$\boldsymbol{x}^N = x_{1,1}^{(N)_{1,1}} x_{1,2}^{(N)_{1,2}} \cdots x_{m,n}^{(N)_{m,n}} \in R^{(t)},$$

written so that the indices obey the lexicographic order from smallest to largest as one goes from left to right. We call such a monomial a *lexicographic term.* Similar notation will be used both for the quantum torus (where $N \in \mathcal{M}_{m,n}(\mathbb{Z})$), and, if (r, s) is the t-th smallest coordinate, for $R^{(t)}[x_{r,s}^{-1}]$ (where all entries of N are nonnegative except possibly the (r, s)-entry).

It is not difficult to check that each $R^{(t)}$ may be written as an iterated Ore extension, which immediately yields the following:

Theorem 2.1.7. *The following properties hold for every $t \in [mn]$:*

(1) *$R^{(t)}$ is a Noetherian domain.*

(2) *As a $\mathbb{K}$-vector space, $R^{(t)}$ has a basis consisting of the lexicographic terms $\boldsymbol{x}^N$ with $N \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$. The same properties also hold for the $m \times n$ quantum torus* (*but with $N \in \mathcal{M}_{m,n}(\mathbb{Z})$*). □

Definition 2.1.8. The *lexicographic expression* of $a \in R^{(t)}$ is the unique linear combination $a = \sum_{N\in\mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})} \alpha_N \boldsymbol{x}^N$ of distinct lexicographic terms with $\alpha_N \neq 0$. A lexicographic term in this expression will be called a *lex term* of a.

For $R^{(1)} = \mathcal{O}_q(\mathbb{K}^{m\times n})$, we will require a slight extension of Theorem 2.1.7. Observe that any monomial $\boldsymbol{t} = t_{i_1,j_1} t_{i_2,j_2} \cdots t_{i_\ell,j_\ell}$ in the standard generators of $R^{(1)}$ may be written as $\boldsymbol{t} = q^{\ell} \boldsymbol{t}^{M^{\text{lex}}}$ for some integer ℓ and lexicographic term $\boldsymbol{t}^{M^{\text{lex}}}$. Since $q^{\ell} \neq 0$, the next result follows easily.

Proposition 2.1.9. *For any coordinate (r,s), the set of lexicographic monomials of $\mathcal{O}_q(\mathbb{K}^{m\times n})$ involving only $t_{i,j}$ with $(i,j) > (r,s)$ is linearly independent over the subalgebra generated by the $t_{i,j}$ with $(i,j) \le (r,s)$. Moreover, for a set $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \dots, \boldsymbol{t}_\ell\}$ of monomials in the standard generators of $\mathcal{O}_q(\mathbb{K}^{m\times n})$, the following are equivalent:*

(1) *The set $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \dots, \boldsymbol{t}_\ell\}$ is linearly independent over $\mathbb{K}$.*

(2) *The set $\{\boldsymbol{t}_1^{M_1^{\mathrm{lex}}}, \boldsymbol{t}_2^{M_2^{\mathrm{lex}}}, \dots, \boldsymbol{t}_\ell^{M_\ell^{\mathrm{lex}}}\}$ is linearly independent over $\mathbb{K}$.*

(3) *The matrices $M_1^{\mathrm{lex}}, \dots, M_\ell^{\mathrm{lex}}$ are distinct.*

A similar set of statements hold for the $m \times n$ quantum torus. □

We conclude this section by noting that $R^{(t)}$ has a natural $\mathbb{Z}_{\ge 0}^{m+n}$-grading that will be very much exploited in the proof of Theorem 4.4.1. If

$$\boldsymbol{s} = (r_1, r_2, \dots, r_m, c_1, c_2, \dots, c_n) \in (\mathbb{Z}_{\ge 0})^{m+n},$$

then the homogeneous component of degree $\boldsymbol{s}$ is the subspace of $R^{(t)}$ spanned by the lexicographic monomials of the form $\boldsymbol{x}^N$, where N satisfies

$$\sum_{j=1}^{n} (N)_{i,j} = r_i \quad \text{for all } i \in [m],$$

$$\sum_{i=1}^{m} (N)_{i,j} = c_j \quad \text{for all } j \in [n].$$

In other words, the sum of all entries in row i of N equals r_i, and the sum of all entries in column j of N equals c_j. All references in this paper to a grading on $R^{(t)}$ will be with respect to this grading.

2.2. ***The deleting derivations algorithm.*** The relation between $R^{(t)}$ and $R^{(t-1)}$ has been studied by Cauchon [2003b] as a special case of the more general theory developed in [Cauchon 2003a]. Here, we review his results as they apply to these algebras. For each result in this section, we fix $t \in [mn]$ with $t \ne 1$, let (r,s) denote the t-th smallest coordinate, and let $[x_{i,j}]$ be the matrix of generators of $R^{(t)}$ and $[y_{i,j}]$ the matrix of generators for $R^{(t-1)}$.

Theorem 2.2.1 [Cauchon 2003a, Lemme 2.1 and Théorème 3.2.1].

(1) *The multiplicative set generated by $x_{r,s}$ is a left and right Ore set for $R^{(t)}$, and the multiplicative set generated by $y_{r,s}$ is a left and right Ore set for $R^{(t-1)}$.*

(2) *There is an injective homomorphism*

$$\overrightarrow{\cdot} : R^{(t-1)} \to R^{(t)}[x_{r,s}^{-1}]$$

defined on the standard generators by

$$\overrightarrow{y_{i,j}} = \begin{cases} x_{i,j} - x_{i,s}x_{r,s}^{-1}x_{r,j} & \text{if } i < r \text{ and } j < s; \\ x_{i,j} & \text{otherwise.} \end{cases}$$

(3) *There is an injective homomorphism*

$$\overleftarrow{\cdot} : R^{(t)} \to R^{(t-1)}[y_{r,s}^{-1}]$$

defined on the standard generators by

$$\overleftarrow{x_{i,j}} = \begin{cases} y_{i,j} + y_{i,s}y_{r,s}^{-1}y_{r,j} & \text{if } i < r \text{ and } j < s; \\ y_{i,j} & \text{otherwise.} \end{cases}$$

(4) $R^{(t)}[x_{r,s}^{-1}] = R^{(t-1)}[y_{r,s}^{-1}]$. □

The homomorphism in Theorem 2.2.1(2) is called the *deleting derivations map*. We call the homomorphism in Theorem 2.2.1(3) the *adding derivations map*. (This map is called the "reverse deleting derivations map" in [Launois 2004a], and a step of the "restoration" algorithm in [Goodearl et al. 2011b].)

The strategy of Cauchon's theory is to use these maps to iteratively transfer information between $R^{(1)} = \mathcal{O}_q(\mathbb{K}^{m\times n})$ and $R^{(mn)} = \mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$. For example, to embed the prime spectrum of the latter algebra into the prime spectrum of the former.

As usual, for an algebra A, denote by $\text{Spec}(A)$ the set of prime ideals, equipped with the Zariski topology. We may partition $\text{Spec}(R^{(t)})$ as

$$\text{Spec}(R^{(t)}) = \text{Spec}^{\not\in}(R^{(t)}) \cup \text{Spec}^{\in}(R^{(t)}),$$

where

$$\text{Spec}^{\not\in}(R^{(t)}) = \{P \in \text{Spec}(R^{(t)}) \mid x_{r,s} \not\in P\},$$

and

$$\text{Spec}^{\in}(R^{(t)}) = \{P \in \text{Spec}(R^{(t)}) \mid x_{r,s} \in P\}.$$

Theorem 2.2.2 [Cauchon 2003b, Section 3.1]. *There exists an injective map*

$$\phi_t : \text{Spec}(R^{(t)}) \to \text{Spec}(R^{(t-1)})$$

satisfying the following properties:

(1) *Restricted to* $\text{Spec}^{\not\in}(R^{(t)})$, ϕ_t *is bijective, sending* $P \in \text{Spec}^{\not\in}(R^{(t)})$ *to*

$$\phi_t(P) = \overleftarrow{P}[y_{r,s}^{-1}] \cap R^{(t-1)}.$$

If $Q \in \text{Spec}^{\not\in}(R^{(t-1)})$, *then*

$$\phi_t^{-1}(Q) = \overrightarrow{Q}[x_{r,s}^{-1}] \cap R^{(t)}.$$

(2) *Restricted to* $\operatorname{Spec}^{\in}(R^{(t)})$, ϕ_t *is injective, sending* $P \in \operatorname{Spec}^{\in}(R^{(t)})$ *to*

$$\phi_t(P) = g^{-1}(P/\langle x_{r,s}\rangle),$$

where $g : R^{(t-1)} \to R^{(t)}/\langle x_{r,s}\rangle$ *is the unique homomorphism that maps the standard generators as* $y_{i,j} \mapsto x_{i,j} + \langle x_{r,s}\rangle$. □

2.3. ***$\mathcal{H}$-stratification.*** For many quantum algebras, including the $R^{(t)}$, the structure of the prime spectrum may be understood by first understanding the prime ideals that are invariant under a rational action of an algebraic torus $\mathcal{H}$. For $R^{(t)}$ with matrix of generators $[x_{i,j}]$, let $\mathcal{H} = (\mathbb{K}^*)^{m+n}$ and note that every $h = (\rho_1, \dots, \rho_m, \gamma_1, \dots, \gamma_n) \in \mathcal{H}$ induces an automorphism of $R^{(t)}$ by

$$h \cdot x_{i,j} = \rho_i \gamma_j x_{i,j}.$$

Definition 2.3.1. An *$\mathcal{H}$-prime* is a prime ideal $K \in \operatorname{Spec}(R^{(t)})$ such that $h \cdot K = K$ for all $h \in \mathcal{H}$. The set of all $\mathcal{H}$-primes of $R^{(t)}$ is denoted $\mathcal{H}\text{-}\operatorname{Spec}(R^{(t)})$. The *$\mathcal{H}$-stratum* associated to an $\mathcal{H}$-prime K is the set

$$\operatorname{Spec}_K(R^{(t)}) = \left\{ P \in \operatorname{Spec}(R^{(t)}) \;\middle|\; \bigcap_{h\in\mathcal{H}} h \cdot P = K \right\}.$$

Theorem 2.3.2 [Goodearl and Letzter 2000; Brown and Goodearl 2002, Part II]. *For every* $t \in [mn]$, *there are finitely many $\mathcal{H}$-primes in* $\mathcal{H}\text{-}\operatorname{Spec}(R^{(t)})$, *and*

$$\operatorname{Spec}(R^{(t)}) = \bigsqcup_{K\in\mathcal{H}\text{-}\operatorname{Spec}(R^{(t)})} \operatorname{Spec}_K(R^{(t)}).$$ □

Remark 2.3.3. Theorem 2.2.1 and Theorem 2.3.2 are where it is necessary to require q to be a nonzero, non-root of unity. We also note here that the $\mathcal{H}$-primes are well known to be homogeneous ideals.

The $\mathcal{H}$-primes of $R^{(1)} = \mathcal{O}_q(\mathbb{K}^{m\times n})$ have generating sets of a simple form.

Theorem 2.3.4 [Goodearl and Letzter 1998, Section 2.1(ii)]. *A prime ideal* $K \in \operatorname{Spec}(R^{(1)})$ *is an $\mathcal{H}$-prime if and only if there exists a* $B \subseteq [m] \times [n]$ *such that*

$$K = \langle t_{i,j} \mid (i,j) \in B\rangle.$$ □

It is convenient to describe these $\mathcal{H}$-primes by using diagrams.

Definition 2.3.5. An $m \times n$ *diagram* is an $m \times n$ grid of squares, each square colored either black or white.

We index the squares of a diagram as one would the entries of an $m \times n$ matrix. If

$$K = \langle t_{i,j} \mid (i,j) \in B\rangle \in \mathcal{H}\text{-}\operatorname{Spec}(R^{(1)})$$

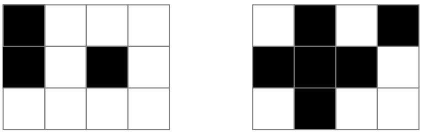


Figure 1. Two 3 × 4 diagrams.

for some $B \subseteq [m] \times [n]$, then the diagram corresponding to K is that in which the black squares are precisely those $(i, j) \in B$. Conversely, any diagram defines a subset $B \subseteq [m] \times [n]$ corresponding to the indices of the black squares, and therefore a corresponding $K \in \mathcal{H}\text{-}\mathrm{Spec}(R^{(1)})$. We henceforth identify a diagram with the corresponding subset $B \subseteq [m] \times [n]$. Figure 1 presents two diagrams, the left one corresponding to the $\mathcal{H}$-prime $\langle t_{1,1}, t_{2,1}, t_{2,3}\rangle \in \mathcal{H}\text{-}\mathrm{Spec}(\mathbb{O}_q(\mathbb{K}^{3\times 4}))$.

The deleting derivations map behaves nicely with respect to $\mathcal{H}$-primes.

Theorem 2.3.6 [Cauchon 2003b, Section 3.1]. *For every $t \in [mn]$, $t \neq 1$, the map ϕ_t injects $\mathcal{H}\text{-}\mathrm{Spec}(R^{(t)})$ into $\mathcal{H}\text{-}\mathrm{Spec}(R^{(t-1)})$. Consequently, the composition*

$$\phi = \phi_2 \circ \cdots \circ \phi_{mn}$$

is an injection of $\mathcal{H}\text{-}\mathrm{Spec}(\mathbb{O}_q(\mathcal{M}_{m,n}(\mathbb{K})))$ into $\mathcal{H}\text{-}\mathrm{Spec}(\mathbb{O}_q(\mathbb{K}^{m\times n}))$. □

In view of the strategy mentioned in Section 2.2, a natural problem is to identify the diagrams of those $\mathcal{H}$-primes in $\mathcal{H}\text{-}\mathrm{Spec}(R^{(1)})$ that are the image of an $\mathcal{H}$-prime in $\mathcal{H}\text{-}\mathrm{Spec}(R^{(mn)})$ under ϕ. We call these *Cauchon diagrams*

Definition 2.3.7. A diagram is a Cauchon diagram if, for any given black square, either every square to the left or every square above is also black.

The right diagram in Figure 1 is an example of a Cauchon diagram, while the left is not a Cauchon diagram since the black square in position $(2, 3)$ has a white square both above and to its left.

Theorem 2.3.8 [Cauchon 2003b, Théorème 3.2.2]. *A diagram is a Cauchon diagram if and only if the corresponding $\mathcal{H}$-prime in $\mathcal{H}\text{-}\mathrm{Spec}(R^{(1)})$ is the image under ϕ of an $\mathcal{H}$-prime in $\mathcal{H}\text{-}\mathrm{Spec}(R^{(mn)})$.* □

3. Quantum matrices by paths

3.1. ***Graphs and paths.*** Let B be a Cauchon diagram and, by Theorem 2.3.8, consider the corresponding $\mathcal{H}$-prime K of $\mathbb{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$. With the notation of Section 2.3, the image of K under the composition $\phi_{t+1} \circ \cdots \circ \phi_{mn}$ is an $\mathcal{H}$-prime K_t of $R^{(t)}$. The goal of this section is to explain how $R^{(t)}/K_t$ is isomorphic to a subalgebra $A_B^{(t)}$ of the quantum torus $\mathbb{O}_q((\mathbb{K}^\times)^{m\times n})$ defined by considering paths in a directed graph that is defined using B. In particular, when $B = \varnothing$, we obtain a combinatorial description of $\mathbb{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$.

Definition 3.1.1. For a Cauchon diagram B, construct a directed graph $G_B^{m\times n}$, called the *Cauchon graph*,[1] as follows. The vertex set consists of *white vertices*

$$W = ([m]\times[n]) \setminus B,$$

together with *row vertices* $R = [m]$ and *column vertices*[2] $C = [n]$. The set of directed edges E consists precisely of those in the following list:

(1) If $(i, j), (i, j') \in W$ are distinct white vertices with $j > j'$ such that there is no white vertex (i, j'') for any $j' < j'' < j$, then we make an edge from (i, j) to (i, j').

(2) If $(i, j), (i', j) \in W$ are distinct white vertices with $i < i'$ such that there is no white vertex (i'', j) for any $i < i'' < i'$, then we make an edge from (i, j) to (i', j).

(3) For $i \in R$, we make an edge from i to (i, j), where j is the largest integer such that $(i, j) \in W$ (if such a j exists).

(4) For $j \in C$, we make an edge from (i, j) to j where i is the largest integer such that $(i, j) \in W$ (if such an i exists).

Note 3.1.2. There is a natural way to embed a Cauchon graph in the plane by placing it "on top" of the Cauchon diagram B as follows. The white vertices are placed at the center of the corresponding white squares, the row vertices to the right of the corresponding diagram row, and the column vertices underneath the corresponding diagram column. An example is illustrated in Figure 2. We call this the *standard embedding* and always assume a given Cauchon graph is equipped with it. Hence, without confusion we can refer to aspects of a Cauchon graph using common directional or geometric terms.[3] That a diagram is a Cauchon diagram easily implies that the corresponding Cauchon graph has the following important property.

Proposition 3.1.3. *The standard embedding of a Cauchon graph is planar.* □

Definition 3.1.4. A *path* in $G_B^{m\times n}$ is a sequence $P = (v_0, v_1, \dots, v_k)$ of distinct vertices such that[4] for all $i \in [k]$, there exists an edge in $G_B^{m\times n}$ directed from v_{i-1} to v_i. Naturally, we say that P *starts* at v_0 and *ends* at v_k and write $P : v_0 \to v_k$.

We consider a directed edge e from v to w to be a path and write $e : v \to w$. If e is the edge between two consecutive vertices in a path P, then we abuse notation

[1] Cauchon graphs already appear in [Postnikov 2006], where they are called Γ-graphs. We call them Cauchon graphs here to be consistent with the Cauchon diagrams from which they derive.

[2] There is ambiguity between labels of the row and column vertices, but the type of vertex we mean will always be explicitly stated.

[3] For example, horizontal, vertical, above, below, northwest, etc.

[4] Strictly speaking, we are defining a *directed* path, but we will never have use for nondirected paths in this paper.

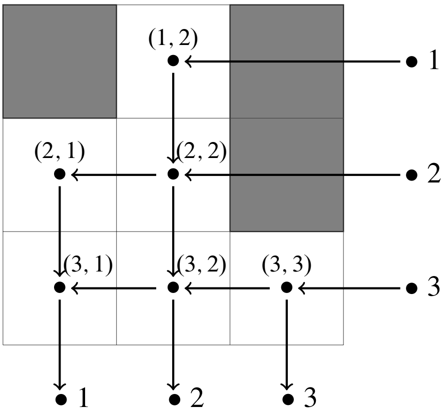


Figure 2. The graph $G_B^{3\times 3}$, embedded on top of the 3×3 Cauchon diagram $B = \{(1,1),(1,3),(2,3)\}$.

by writing $e \in P$. Finally, if $P : u \to v$, $Q : v \to w$, then we write $P \cup Q$ to denote the concatenation of P and Q. To a path in a Cauchon graph we will assign an element of the quantum torus as follows:

Definition 3.1.5. Let $G_B^{m\times n}$ be a Cauchon graph. Define the function

$$w : E \to \mathbb{O}_q((\mathbb{K}^\times)^{m\times n})$$

as follows, where the numbering and notation correspond to the edge types of Definition 3.1.1:

(1) $w(e : (i,j) \to (i,j')) = t_{i,j}^{-1} t_{i,j'}$;

(2) $w(e : (i,j) \to (i',j)) = 1$;

(3) $w(e : i \to (i,j)) = t_{i,j}$;

(4) $w(e : (i,j) \to j) = 1$.

The image $w(e)$ of an edge e is called the *weight* of e.

If $P = (v_0, v_1, \dots v_k)$ is a path, and $e_i : v_{i-1} \to v_i$, then the weight of P is defined to be

$$w(P) = w(e_1)w(e_2)\cdots w(e_k).$$

Example 3.1.6. Figure 3 illustrates the graph of Figure 2 with edges labeled by their weights. The weight of the path

$$P = (1,(1,2),(2,2),(2,1),(3,1),1)$$

is

$$w(P) = (t_{1,2})(1)(t_{2,2}^{-1}t_{2,1})(1)(1) = t_{1,2}t_{2,2}^{-1}t_{2,1}.$$

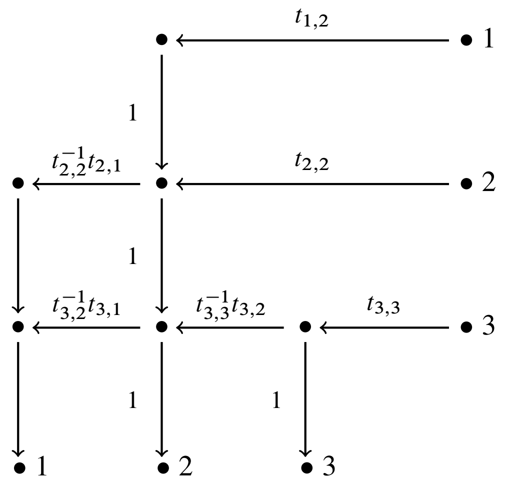


Figure 3. The graph $G_B^{3\times 3}$, with $B = \{(1,1),(1,3),(2,3)\}$, and edges labeled by their weights. (Labels of white vertices omitted.)

It is convenient to observe that for a row vertex i and a column vertex j, the weight of a path $P : i \to j$ can be computed by looking at the sequence of "turns".

Definition 3.1.7. Let $P = (v_0, v_1, \dots, v_{k-1}, v_k)$ be a path in a Cauchon graph starting from row vertex $i = v_0$ and ending at column vertex $j = v_k$.

- A Γ-*turn* in P is a white vertex $v_i \in P$ such that the edge from v_{i-1} to v_i is horizontal, and the edge from v_i to v_{i+1} is vertical.
- A ⅃-*turn* in P is a white vertex $v_i \in P$ such that the edge from v_{i-1} to v_i is vertical and the edge from v_i to v_{i+1} is horizontal.

The next proposition follows easily using the definitions of edge and path weights.

Proposition 3.1.8. *Let $P : i \to j$ be a path in a Cauchon graph, where i is a row vertex and j is a column vertex. If $(v_{i_1}, v_{i_2}, \dots, v_{i_t}) \subset P$ is the subsequence consisting of all Γ-turns and ⅃-turns, then*

$$w(P) = t_{v_{i_1}} t_{v_{i_2}}^{-1} t_{v_{i_3}} \cdots t_{v_{i_{t-1}}}^{-1} t_{v_{i_t}}.$$

Example 3.1.9. For the path P in Example 3.1.6, the vertex $(1,2)$ is a Γ-turn, $(2,2)$ is a ⅃-turn, and $(2,1)$ is a Γ-turn, so that $w(P) = (t_{1,2})(t_{2,2}^{-1})(t_{2,1})$. This, of course, agrees with Example 3.1.6.

Parts (1) and (2) of the next result are Lemmas 3.5 and 3.6 respectively in [Casteels 2011]. Part (3) is proven similarly.

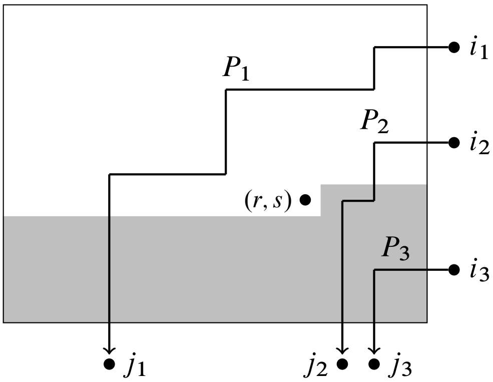


Figure 4. The shaded area represents all white vertices greater than the t-th smallest coordinate (r,s). (This convention will be repeated in later illustrations.) In this example, $P_1 \in \Gamma_B^{(t)}(i_1, j_1)$ and $P_3 \in \Gamma_B^{(t)}(i_3, j_3)$ but $P_2 \notin \Gamma_B^{(t)}(i_2, j_2)$.

Lemma 3.1.10. *In a Cauchon graph $G_B^{m\times n}$, let (a,b) be a white vertex, i and k row vertices with $i<k$, and j and ℓ column vertices with $j<\ell$.*

(1) *If $P : i \to (a,b)$ and $Q : (a,b) \to \ell$ are paths in $G_B^{m\times n}$ with only (a,b) in common, then*

$$w(P)w(Q) = \begin{cases} w(Q)w(P) & \text{if } b=\ell, \text{ i.e., } Q \text{ has only vertical edges,} \\ q^{-1}w(Q)w(P) & \text{otherwise.} \end{cases}$$

(2) *If $P : (a,b) \to j$ and $Q : (a,b) \to \ell$ are paths in $G_B^{m\times n}$ with only (a,b) in common, then*

$$w(P)w(Q) = \begin{cases} w(Q)w(P) & \text{if } b=\ell, \text{ i.e., } Q \text{ has only vertical edges,} \\ qw(Q)w(P) & \text{otherwise.} \end{cases}$$

(3) *If $P : i \to (a,b)$ and $Q : k \to (a,b)$ are paths in $G_B^{m\times n}$ with only (a,b) in common, then*

$$w(P)w(Q) = qw(Q)w(P).$$

For the remainder of this section, fix $t \in [mn]$ and let (r,s) be the t-th smallest coordinate.

Notation 3.1.11. For a row vertex i and a column vertex j of $G_B^{m\times n}$, let $\Gamma_B^{(t)}(i,j)$ denote the set of all paths $P : i \to j$ in $G_B^{m\times n}$ for which no vertex larger than (r,s) is a $\lrcorner$-turn.

Figure 4 is meant to clarify Notation 3.1.11, and while we have drawn a vertex (r,s) in this figure, it will not exist if $(r,s) \in B$. The main theorem of this section is the following:

Theorem 3.1.12. *Let $G_B^{m\times n}$ be a Cauchon graph, let i, k be row vertices with $i < k$, and let j, ℓ be column vertices.*

(1) *If $j < \ell$, then there exists a permutation of $\Gamma_B^{(t)}(i, j) \times \Gamma_B^{(t)}(i, \ell)$ sending $(P, Q) \mapsto (\widetilde{P}, \widetilde{Q})$, where*

$$w(P)w(Q) = qw(\widetilde{Q})w(\widetilde{P}).$$

(2) *If $j = \ell$, then there exists a permutation of $\Gamma_B^{(t)}(i, j) \times \Gamma_B^{(t)}(k, j)$ sending $(P, Q) \mapsto (\widetilde{P}, \widetilde{Q})$, where*

$$w(P)w(Q) = qw(\widetilde{Q})w(\widetilde{P}).$$

(3) *If $j > \ell$, then there exists a permutation of $\Gamma_B^{(t)}(i, j) \times \Gamma_B^{(t)}(k, \ell)$ sending $(P, Q) \mapsto (\widetilde{P}, \widetilde{Q})$, where*

$$w(P)w(Q) = w(\widetilde{Q})w(\widetilde{P}).$$

(4) *If $j < \ell$, then*:

(a) *If $P \in \Gamma_B^{(t)}(i, j)$, $Q \in \Gamma_B^{(t)}(k, \ell)$ and $P \cap Q = \varnothing$, then*

$$w(P)w(Q) = w(Q)w(P).$$

(b) *There exists a bijection from the subset of $\Gamma_B^{(t)}(i, j) \times \Gamma_B^{(t)}(k, \ell)$ consisting of those (P, Q) with $P \cap Q \neq \varnothing$ to $\Gamma_B^{(t)}(i, \ell) \times \Gamma_B^{(t)}(k, j)$ sending (P, Q) to $(\widetilde{P}, \widetilde{Q})$, where*

$$w(P)w(Q) = qw(\widetilde{Q})w(\widetilde{P}).$$

Proof. Part (1): Let $(P, Q) \in \Gamma_B^{(t)}(i, j) \times \Gamma_B^{(t)}(i, \ell)$. Since $j < \ell$, P and Q have a last (white) vertex in common, say (a, b). See Figure 5. Therefore, we may write $P = P_1 \cup P_2$, where $P_1 : i \to (a, b)$ and $P_2 : (a, b) \to j$, and $Q = Q_1 \cup Q_2$, where $Q_1 : k \to (a, b)$ and $Q_2 : (a, b) \to \ell$. Define $\widetilde{P} = Q_1 \cup P_2$ and $\widetilde{Q} = P_1 \cup Q_2$. We have $(\widetilde{P}, \widetilde{Q}) \in \Gamma_B^{(t)}(i, j) \times \Gamma_B^{(t)}(i, \ell)$ and that $\widetilde{\widetilde{P}} = P$ and $\widetilde{\widetilde{Q}} = Q$, i.e., the map $(P, Q) \mapsto (\widetilde{P}, \widetilde{Q})$ is an involution and so a permutation.

Finally, we apply Lemma 3.1.10 to make our final conclusion as follows. If Q_2 has only vertical edges, then (using the cited part of the lemma at each line)

$$\begin{aligned} w(P)w(Q) &= w(P_1)w(P_2)w(Q_1)w(Q_2) \\ &\overset{(1)}{=} qw(P_1)w(Q_1)w(P_2)w(Q_2) \\ &\overset{(1,3)}{=} qw(P_1)w(Q_2)w(Q_1)w(P_2) \\ &= qw(\widetilde{Q})w(\widetilde{P}). \end{aligned}$$

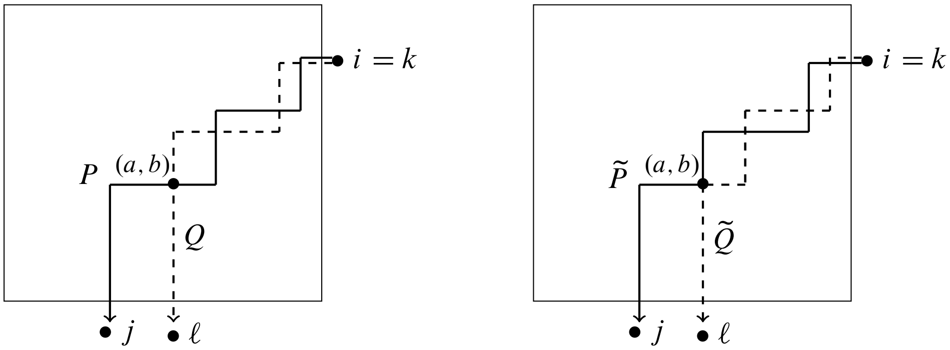


Figure 5. Illustration of Part (1) in the proof of Theorem 3.1.12. The left figure shows paths P (solid) and Q (dashed). The right figure shows paths $\widetilde{P}$ (solid) and $\widetilde{Q}$ (dashed).

If Q_2 has a horizontal edge, then

$$\begin{aligned} w(P)w(Q) &= w(P_1)w(P_2)w(Q_1)w(Q_2) \\ &\overset{(1)}{=} q^{-1}qw(P_1)w(Q_2)w(P_2)w(Q_1) \\ &\overset{(1,3)}{=} qw(P_1)w(Q_2)w(Q_1)w(P_2) \\ &= qw(\widetilde{Q})w(\widetilde{P}). \end{aligned}$$

Part (2): Let $(P,Q) \in \Gamma_B^{(t)}(i,j) \times \Gamma_B^{(t)}(k,j)$. In this case, P and Q have a first common vertex, say (a,b). Therefore, we may write $P = P_1 \cup P_2$, where $P_1 : i \to (a,b)$ and $P_2 : (a,b) \to j$, and $Q = Q_1 \cup Q_2$, where $Q_1 : k \to (a,b)$ and $Q_2 : (a,b) \to \ell$. Define $\widetilde{P} = P_1 \cup Q_2$ and $\widetilde{Q} = Q_1 \cup P_2$. We again have $(\widetilde{P},\widetilde{Q}) \in \Gamma_B^{(t)}(i,j) \times \Gamma_B^{(t)}(k,j)$ and that the map $(P,Q) \mapsto (\widetilde{P},\widetilde{Q})$ is a permutation. The remainder of the proof for Part (2) proceeds as in Part (1) and by using (1) and (2) of Lemma 3.1.10.

Part (3): Let $(P,Q) \in \Gamma_B^{(t)}(i,j) \times \Gamma_B^{(t)}(k,\ell)$, where $i < k$ and $j > \ell$. In this case, P and Q have a first common vertex (a,b) and a last common vertex (a',b'). We can write $P = P_1 \cup P_2 \cup P_3$, where $P_1 : i \to (a,b)$, $P_2 : (a,b) \to (a',b')$ and $P_3 : (a',b') \to j$. Similarly $Q = Q_1 \cup Q_2 \cup Q_3$, where $Q_1 : k \to (a,b)$, $Q_2 : (a,b) \to (a',b')$ and $Q_3 : (a',b') \to \ell$. Define $\widetilde{P} = P_1 \cup Q_2 \cup P_3$ and $\widetilde{Q} = Q_1 \cup P_2 \cup Q_3$.

We again have $(\widetilde{P},\widetilde{Q}) \in \Gamma_B^{(t)}(i,j) \times \Gamma_B^{(t)}(k,\ell)$ and that the map $(P,Q) \mapsto (\widetilde{P},\widetilde{Q})$ is a permutation. To prove the final conclusion concerning the weights relation, we must consider several possibilities, according to whether or not any of P_2, P_3 and Q_2 consists only of vertical edges, or no edges at all (the other paths here always have a horizontal edge). Here, we discuss only the case that P_2, P_3 and Q_2 each have a horizontal edge, the other possibilities being dealt with similarly. Before we begin, we should mention that, strictly speaking, P_2 and Q_2 do not begin nor end at a row or column vertex, and so Lemma 3.1.10 does not directly apply. In order to use the lemma, we identify P_2 and Q_2 respectively with the paths obtained by adding the vertical path from (a', b') to b' and the horizontal path from a to (a, b). We can do this since in either case these latter paths have the same weight as $w(P_2)$ or $w(P_3)$ respectively, by Proposition 3.1.8.

We have

$$\begin{aligned}
w(P)w(Q) &= w(P_1)w(P_2)w(P_3)w(Q_1)w(Q_2)w(Q_3)\\
&\overset{(1)}{=} qw(P_1)w(P_2)w(Q_1)w(Q_2)w(P_3)w(Q_3)\\
&\overset{(2)}{=} w(P_1)w(P_2)w(Q_1)w(Q_2)w(Q_3)w(P_3)\\
&\overset{(1)}{=} q^{-1}w(P_1)w(P_2)w(Q_1)w(Q_3)w(Q_2)w(P_3)\\
&\overset{(1)}{=} w(P_1)w(Q_1)w(P_2)w(Q_3)w(Q_2)w(P_3)\\
&\overset{(3,1)}{=} w(Q_1)w(P_2)w(P_1)w(Q_3)w(Q_2)w(P_3),
\end{aligned}$$

where each line again uses the respective part of Lemma 3.1.10 and the third line is applying the cited part to P_2 and $Q_1 \cup Q_2$. That the last line is equal to $w(\widetilde{Q})w(\widetilde{P})$ is now implied by the fact that $w(P_1)$ and $w(Q_3)$ commute. Indeed, we have

$$\begin{aligned}
w(P_1)w(Q_3) &= w(P_1)w(Q_2)^{-1}w(Q_2)w(Q_3)\\
&\overset{(1)}{=} qw(Q_2)^{-1}w(P_1)w(Q_2)w(Q_3)\\
&\overset{(1)}{=} w(Q_2)^{-1}w(Q_2)w(Q_3)w(P_1)\\
&= w(Q_3)w(P_1),
\end{aligned}$$

where the third line is applying the cited lemma to P_3 and $Q_2 \cup Q_3$.

Part (4a): Lemma 3.4 in [Casteels 2011] shows that the weight of any edge not sharing a vertex with Q commutes with $w(Q)$. Since this is the case for all edges of P we immediately have $w(P)w(Q) = w(Q)w(P)$.

Part (4b): As in Part (1), we let (a, b) be the last common vertex in a nondisjoint pair of paths $(P, Q) \in \Gamma_B^{(t)}(i, j) \times \Gamma_B^{(t)}(k, \ell)$. We then "switch" the tails of P and Q at (a, b) to obtain a $\widetilde{P} : i \to \ell$ and a $\widetilde{Q} : k \to j$. The remainder of the proof is as in Part (1). □

3.2. *The algebras $A_B^{(t)}$.* In this section we introduce for each $t \in [mn]$ and Cauchon diagram B a subalgebra $A_B^{(t)}$ of $\mathbb{O}_q((\mathbb{K}^\times)^{m\times n})$. When $B = \varnothing$, we will see that $A_\varnothing^{(t)} \simeq R^{(t)}$. Throughout this section we fix $t \in [mn]$ and let (r, s) be the t-th smallest coordinate.

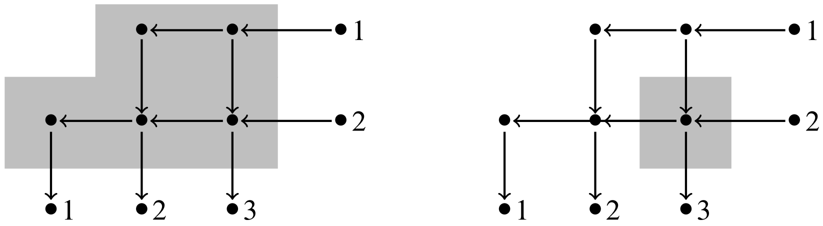


Figure 6. Two copies of the graph $G^{2\times 3}_{\{(1,1)\}}$ of Example 3.2.2. The left picture is shaded to assist the definition of $A^{(1)}_B$, the right picture shaded to assist that of $A^{(5)}_B$.

Definition 3.2.1. We define $A^{(t)}_B$ to be the subalgebra of $\mathbb{O}_q((\mathbb{K}^\times)^{m\times n})$ with the $m \times n$ matrix of generators $[x_{i,j}]$ where, for each coordinate (i, j),

$$x_{i,j} = \sum_{P\in\Gamma^{(t)}_B(i,j)} w(P).$$

When $B = \varnothing$ we write $A^{(t)} = A^{(t)}_{\varnothing}$.

Example 3.2.2. Consider the 2×3 Cauchon diagram $B = \{(1, 1)\}$. Figure 6 presents two copies of the corresponding Cauchon graph, where we continue our illustrative convention that no path may contain a ⅃-turn in the shaded region. For each $t \in [6]$, we denote by $[x^{(t)}_{i,j}]$ the matrix of generators for $A^{(t)}_B$.

The left graph of Figure 6 corresponds to $t = 1$. In this case, any path from row vertex 1 to column vertex 1 necessarily contains a ⅃-turn in the shaded region. Therefore, $A^{(1)}_B$ has matrix of generators

$$\begin{bmatrix} x^{(1)}_{1,1} & x^{(1)}_{1,2} & x^{(1)}_{1,3} \\ x^{(1)}_{2,1} & x^{(1)}_{2,2} & x^{(1)}_{2,3} \end{bmatrix} = \begin{bmatrix} 0 & t_{1,2} & t_{1,3} \\ t_{2,1} & t_{2,2} & t_{2,3} \end{bmatrix}.$$

One may check that $A^{(1)}_B = A^{(2)}_B = A^{(3)}_B = A^{(4)}_B$. For $t = 5$, the Cauchon graph is illustrated on the right in Figure 6. In this case, there exists a unique path in $\Gamma^{(5)}_B(1, 1)$, so that the matrix of generators for $A^{(5)}_B$ is

$$\begin{bmatrix} x^{(5)}_{1,1} & x^{(5)}_{1,2} & x^{(5)}_{1,3} \\ x^{(5)}_{2,1} & x^{(5)}_{2,2} & x^{(5)}_{2,3} \end{bmatrix} = \begin{bmatrix} t_{1,2}t^{-1}_{2,2}t_{2,1} & t_{1,2} & t_{1,3} \\ t_{2,1} & t_{2,2} & t_{2,3} \end{bmatrix}.$$

Finally, one may check that $A^{(6)}_B$ has matrix of generators

$$\begin{bmatrix} x^{(6)}_{1,1} & x^{(6)}_{1,2} & x^{(6)}_{1,3} \\ x^{(6)}_{2,1} & x^{(6)}_{2,2} & x^{(6)}_{2,3} \end{bmatrix} = \begin{bmatrix} t_{1,2}t^{-1}_{2,2}t_{2,1} + t_{1,3}t^{-1}_{2,3}t_{2,1} & t_{1,2} + t_{1,3}t^{-1}_{2,3}t_{2,2} & t_{1,3} \\ t_{2,1} & t_{2,2} & t_{2,3} \end{bmatrix}.$$

Theorem 3.1.12 implies some commutation relations between the generators of $A_B^{(t)}$. Compare the following result with Definition 2.1.3:

Theorem 3.2.3. *If $X = [x_{i,j}]$ is the matrix of generators for $A_B^{(t)}$, and*

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is any 2×2 submatrix of X, then:

(1) $ab = qba, cd = qdc$;

(2) $ac = qca, bd = qdb$;

(3) $bc = cb$;

(4) $ad = \begin{cases} da & \text{if } d = x_{k,\ell} \text{ and } (k,\ell) > (r,s); \\ da + (q - q^{-1})bc & \text{if } d = x_{k,\ell} \text{ and } (k,\ell) \leq (r,s). \end{cases}$

Proof. First note that for any coordinates (i, j) and (i', j'),

$$\begin{aligned} x_{i,j}x_{i',j'} &= \sum_{\substack{P \in \Gamma_B^{(t)}(i,j) \\ Q \in \Gamma_B^{(t)}(i',j')}} w(P)w(Q) \\ &= \sum_{\substack{P,Q \\ P \cap Q = \varnothing}} w(P)w(Q) + \sum_{\substack{P,Q \\ P \cap Q \neq \varnothing}} w(P)w(Q). \end{aligned} \tag{3-1}$$

Let

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} x_{i,j} & x_{i,\ell} \\ x_{k,j} & x_{k,\ell} \end{bmatrix}$$

be a 2×2 submatrix of X.

First, consider $x_{i,j}$ and $x_{i,\ell}$. In this case the first sum in (3-1) is necessarily empty, since any pair $(P, Q) \in \Gamma_B^{(t)}(i, j) \times \Gamma_B^{(t)}(i, \ell)$ has row vertex i in common. Part (1) of Theorem 3.1.12 shows that for any such pair, there is a unique pair $(\widetilde{P}, \widetilde{Q}) \in \Gamma_B^{(t)}(i, j) \times \Gamma_B^{(t)}(i, \ell)$ such that $w(P)w(Q) = qw(\widetilde{Q})w(\widetilde{P})$. Hence, (3-1) implies $x_{i,j}x_{i,\ell} = qx_{i,\ell}x_{i,j}$ The relations between $x_{k,j}$ and $x_{k,\ell}$; $x_{i,j}$ and $x_{k,j}$; $x_{i,\ell}$ and $x_{k,\ell}$; and $x_{i,j}$ and $x_{k,j}$ are all obtained similarly.

Now consider $x_{i,j}$ and $x_{k,\ell}$. If $(r, s) < (k, \ell)$, then

$$\Gamma_B^{(t)}(k, \ell) = \{Q = (k, (k, \ell), \ell)\},$$

and any $P \in \Gamma_B^{(t)}(i, j)$ is disjoint from Q by definition of $\Gamma_B^{(t)}(i, j)$. Hence $x_{i,j}x_{k,\ell} = x_{k,\ell}x_{i,j}$ by Part (4a) of Theorem 3.1.12. If $(k, \ell) \leq (r, s)$, then by (3-1) and Part (4b) of Theorem 3.1.12, we obtain

$$x_{i,j}x_{k,\ell} = qx_{i,\ell}x_{k,j} + \sum_{\substack{P\in\Gamma_B^{(t)}(i,j)\\ Q\in\Gamma_B^{(t)}(i,j)\\ P\cap Q=\varnothing}} w(P)w(Q).$$

Since the weights of disjoint paths commute by Part (4a) of Theorem 3.1.12, it follows that $x_{i,j}x_{k,\ell} - x_{k,\ell}x_{i,j} = (q-q^{-1})x_{i,\ell}x_{k,j}$. □

The intuition behind these algebras is that one obtains $A_B^{(t)}$ from $A_B^{(t-1)}$ by "allowing more paths". To be more precise, let $[x_{i,j}]$ be the matrix of generators for $A_B^{(t)}$ and $[y_{i,j}]$ that for $A_B^{(t-1)}$. We have

$$x_{i,j} = y_{i,j} + \sum w(P) \tag{3-2}$$

as elements of $\mathbb{O}_q((\mathbb{K}^\times)^{m\times n})$, where the sum is over all paths $P : i \to j$ for which (r,s) is a ⅃-turn in P. If $i \geq r$, $j \geq s$, or $(r,s) \in B$, then no such P exists and

$$x_{i,j} = y_{i,j}.$$

On the other hand, if $(r,s) \notin B$ and both $i < r$ and $j < s$, suppose $P : i \to j$ is a path with a ⅃-turn at (r,s). Consider $w(P)w(Q)$, where $Q = (r,(r,s),s)$. As in the proof of Theorem 3.1.12, we may form paths $\widetilde{P} : i \to s$ and $\widetilde{Q} : r \to j$ by "switching tails" at (r,s). Since $w(P)w(Q) = qw(\widetilde{Q})w(\widetilde{P})$, multiplying (3-2) through by $y_{r,s} = x_{r,s} = w(Q)$ gives

$$x_{i,j}x_{r,s} = y_{i,j}y_{r,s} + \sum w(P)y_{r,s} = y_{i,j}y_{r,s} + qy_{i,s}y_{r,j}.$$

One may easily check that $t_{r,s} = x_{r,s} = y_{r,s}$ generates a left and right Ore set for $A_B^{(t)}$ and $A_B^{(t-1)}$. (For $x_{r,s}$, this follows from the observation that $x_{i,j}x_{r,s}^{m+1} = x_{r,s}^m a$ for some $a \in A_B^{(t)}$ when $x_{i,j} \neq 0$ and (i,j) is northwest of (r,s).) Hence, we have just proved Parts (1) and (2) of the following result. Part (3) follows from these, and Part (4) is trivial.

Compare the following theorem with [Cauchon 2003a, Proposition 5.4.2]:

Theorem 3.2.4. (1) *If* $(r,s) \notin B$, *then* $A_B^{(t-1)}$ *is a subalgebra of*

$$A_B^{(t)}[x_{r,s}^{-1}],$$

where

$$y_{i,j} = \begin{cases} x_{i,j} - x_{i,s}(x_{r,s})^{-1}x_{r,j} & \text{if } i < r \text{ and } j < s;\\ x_{i,j} & \text{otherwise.}\end{cases}$$

(2) *If* $(r,s) \notin B$, *then* $A_B^{(t)}$ *is a subalgebra of*

$$A_B^{(t-1)}[y_{r,s}^{-1}]$$

where

$$x_{i,j} = \begin{cases} y_{i,j} + y_{i,s}(y_{r,s})^{-1}y_{r,j} & \text{if } i < r \text{ and } j < s; \\ y_{i,j} & \text{otherwise.} \end{cases}$$

(3) *If* $(r,s) \notin B$, *then* $A_B^{(t)}[x_{r,s}^{-1}] = A_B^{(t-1)}[y_{r,s}^{-1}]$.

(4) *If* $(r,s) \in B$, *then* $A_B^{(t)} = A_B^{(t-1)}$. □

In view of Theorem 2.2.1, we conclude the following when $B = \varnothing$.

Corollary 3.2.5. *For every* $t \in [mn]$ *we have* $R^{(t)} \simeq A^{(t)}$, *where* $R^{(t)}$ *are the algebras of* Definition 2.1.3, *and where the standard generator of* $R^{(t)}$ *with coordinate* (i,j) *maps to the generator of* $A^{(t)}$ *with coordinate* (i,j).

Hence, $A^{(1)} \simeq \mathcal{O}_q(\mathbb{K}^{m\times n})$, $A^{(mn)} \simeq \mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$ and both the deleting derivations and $\mathcal{H}$-stratification theories apply to $A^{(t)}$. Moreover, we follow the arrow notation introduced in Section 2.2 to distinguish a generator $x_{i,j}$ of $A_B^{(t)}$ from its image $\overleftarrow{x_{i,j}}$ in $A_B^{(t-1)}$, and a generator $y_{i,j}$ of $A_B^{(t-1)}[y_{r,s}^{-1}]$ from its image $\overrightarrow{y_{i,j}}$ in $A_B^{(t)}[x_{r,s}^{-1}]$.

3.3. *$\mathcal{H}$-primes as kernels.*

Fix $t \in [mn]$ and a Cauchon diagram B. Denote the matrix of generators for $A^{(t)}$ by $[x_{i,j}]$ and the matrix of generators for $A_B^{(t)}$ by $[x_{i,j}^B]$.

Definition 3.3.1. For $t \in [mn]$ and a Cauchon diagram B, let $\sigma_B^{(t)} : A^{(t)} \to A_B^{(t)}$ be defined on the standard generators by

$$\sigma_B^{(t)}(x_{i,j}) = x_{i,j}^B.$$

Section 3.1 of [Cauchon 2003b] implies the following two results:

Proposition 3.3.2. *The map* $\sigma_B^{(t)}$ *extends to a well-defined surjective homomorphism.*

Theorem 3.3.3. *One has*

$$\ker(\sigma_B^{(t)}) \in \mathcal{H}\text{-}\operatorname{Spec}(A^{(t)}).$$

Moreover, if $t > 1$,

$$\ker(\sigma_B^{(t-1)}) = \phi_t(\ker(\sigma_B^{(t)})),$$

where ϕ_t *is as in* Theorem 2.2.2.

We conclude this short section with a technical lemma. For $M \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$, write $M = M_0 + M_1$, where

$$(M_0)_{i,j} = \begin{cases} (M)_{i,j} & \text{if } (i,j) \leq (r,s); \\ 0 & \text{if } (i,j) > (r,s); \end{cases}$$

and $M_1 = M - M_0$. Now, let $K_t = a \in \ker\big(\sigma_B^{(t)}\big)$. Let $\mathcal{M}$ denote the set of $M \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$ for which $\boldsymbol{x}^M$ is a lex term of a. Hence, for some $\alpha_M \in \mathbb{K}^*$, we have

$$a = \sum_{M \in \mathcal{M}} \alpha_M \boldsymbol{x}^M = \sum_{M \in \mathcal{M}} \alpha_M \boldsymbol{x}^{M_0} \boldsymbol{x}^{M_1} = \sum_{N \in \mathcal{M}_{m,n}(\mathbb{Z})} \Bigg(\sum_{\substack{M \in \mathcal{M} \\ M_1 = N_1}} \alpha_M \boldsymbol{x}^{M_0} \Bigg) \boldsymbol{x}^{N_1}.$$

Consider

$$\sigma_B^{(t)}(a) = \sum_{N \in \mathcal{M}_{m,n}(\mathbb{Z})} \Bigg(\sum_{\substack{M \in \mathcal{M} \\ M_1 = N_1}} \alpha_M \sigma_B^{(t)}(\boldsymbol{x}^{M_0}) \Bigg) \sigma_B^{(t)}(\boldsymbol{x}^{N_1}) = 0 \tag{3-3}$$

Let $N \in \mathcal{M}_{m,n}(\mathbb{Z})$. If there is a coordinate $(i, j) > (r, s)$ with both $(i, j) \in B$ and $(N)_{i,j} \geq 1$, then $\boldsymbol{x}^{N_1} \in K_t$ since $x_{i,j} = t_{i,j}$ and $\sigma_B^{(t)}(x_{i,j}) = 0$. Otherwise, $\boldsymbol{x}^{N_1} \neq 0$, and the coefficient of $\sigma_B^{(t)}(\boldsymbol{x}^{N_1})$ must be 0 by Proposition 2.1.9; i.e., we have that

$$\sum_{\substack{M \in \mathcal{M} \\ M_1 = N_1}} \alpha_M \boldsymbol{x}^{M_0} \in K_t.$$

Lemma 3.3.4. *With the notation of the preceding two paragraphs, we have that if $a \in K_t$, then*

$$a = a' + \sum_{\substack{N \in \mathcal{M}_{m,n}(\mathbb{Z}) \\ \boldsymbol{x}^{N_1} \notin K_t}} a_N \boldsymbol{x}^{N_1},$$

where in the second summand each a_N is in K_t, and $a' \in K_t$ has the property that every lex term $\boldsymbol{x}^L$ of a' satisfies $\boldsymbol{x}^{L_1} \in K_t$, i.e., $(L)_{i,j} \geq 1$ for some $(i, j) > (r, s)$ and $(i, j) \in B$.

4. Generators of $\mathcal{H}$-primes

The goal of this section is the proof of Theorem 4.4.1, where we show that an $\mathcal{H}$-prime in $\mathcal{H}$-Spec$(\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K})))$ has, as a right ideal, a Gröbner basis consisting of the quantum minors it contains. That these elements also form a Gröbner basis as a left ideal can be shown similarly.

We begin by defining quantum minors in Section 4.1 and recalling Theorem 4.4 in [Casteels 2011], which shows that a q-analogue of Lindström's classic lemma [1973] holds in the context of Cauchon graphs. We follow this by reviewing the notions of Gröbner bases as applied to the algebras $A^{(t)}$, and finally prove the main result in Section 4.4.

4.1. ***Quantum minors.*** Throughout this section, we fix a Cauchon diagram B and a $t \in [mn]$. Set (r, s) to be the t-th smallest coordinate and $[x_{i,j}]$ to be the matrix of generators for $A_B^{(t)}$.

Definition 4.1.1. Let $I=\{i_1<i_2<\cdots<i_k\}\subseteq[m]$ and $J=\{j_1<j_2<\cdots<j_k\}\subseteq[n]$ be nonempty subsets of the same cardinality. The *quantum minor* associated to I and J is the element of $A_B^{(t)}$ defined by

$$[I\mid J]_B^{(t)}=\sum_{\sigma\in S_k}(-q)^{\ell(\sigma)}x_{i_1,j_{\sigma(1)}}\cdots x_{i_k,j_{\sigma(k)}},$$

where S_k is the set of permutations of $[k]$ and $\ell(\sigma)$ is the number of inversions of $\sigma\in S_k$, i.e., the number of pairs $i,i'\in[k]$ with $i<i'$ but $\sigma(i)>\sigma(i')$.

Remark 4.1.2. The defining expression for $[I\mid J]_B^{(t)}$ is its lexicographic expression. More precisely, for $\sigma\in S_k$, let P_σ be the $m\times n$ matrix whose submatrix indexed by (I,J) equals the standard $k\times k$ permutation matrix corresponding to σ, and where all other entries of P_σ are zero. We can then write

$$[I\mid J]_B^{(t)}=\sum_{\sigma\in S_k}(-q)^{\ell(\sigma)}\boldsymbol{x}^{P_\sigma}.$$

We will often write $[I\mid J]^{(t)}$ for $[I\mid J]_\varnothing^{(t)}$. However, for the remainder of this section, we write $[I\mid J]=[I\mid J]_B^{(t)}$. For the remainder of this paper we shorten "quantum minor" to just "minor".

Definition 4.1.3. For $I=\{i_1<i_2<\cdots<i_k\}\subseteq[m]$ and $J=\{j_1<j_2<\cdots<j_k\}\subseteq[n]$, each (i_ℓ,j_ℓ) is called a *diagonal coordinate* of $[I\mid J]$. Moreover, (i_k,j_k) is the *maximum coordinate* of $[I\mid J]$.

As elements of $\mathcal{O}_q((\mathbb{K}^\times)^{m\times n})$, each minor whose maximum coordinate is at most (r,s) reduces to a particularly nice form via a q-analogue of Lindström's lemma. To explain, we first need to set some notation. At this point, the reader may wish to recall some of the notation defined in Section 3.1.

Definition 4.1.4. Let $I=\{i_1,\dots,i_k\}\subseteq[m]$ and $J=\{j_1,\dots,j_k\}\subseteq[n]$ be such that $|I|=|J|=k$.

(1) A *vertex-disjoint path system* from the row vertices I to the column vertices J in $G_B^{m\times n}$ is a set of k mutually disjoint paths $(P_1,\dots,P_k)$ where $P_r\in\Gamma_B^{(t)}(i_r,j_r)$ for each $r\in[k]$. We write

$$\Gamma_B^{(t)}(I\mid J)=\{\text{all vertex-disjoint path systems from } I \text{ to } J \text{ in } G_B^{m\times n}\}.$$

(2) If $\mathcal{P}=(P_1,\dots,P_k)\in\Gamma_B^{(t)}(I\mid J)$, then the *weight* of $\mathcal{P}$ is the product

$$w(\mathcal{P})=w(P_1)w(P_2)\cdots w(P_k)\in\mathcal{O}_q((\mathbb{K}^\times)^{m\times n}).$$

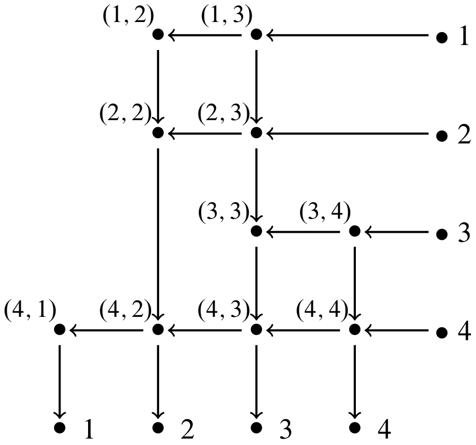


Figure 7. A Cauchon graph.

Notation 4.1.5. If we wish to explicitly write out the elements of I and J in either $[I \mid J]$ or $\Gamma_B^{(t)}(I \mid J)$, we will omit the braces. For example, we write

$$[I \mid J] = [\{i_1, \dots, i_k\} \mid \{j_1, \dots, j_k\}] = [i_1, \dots, i_k \mid j_1, \dots, j_k].$$

Example 4.1.6. For the Cauchon graph shown in Figure 7, the path system $\mathcal{P} = (P_1, P_2, P_3)$, where

$$\begin{aligned} P_1 &= \bigl(1, (1,3), (1,2), (2,2), (4,2), (4,1), 1\bigr), \\ P_2 &= \bigl(2, (2,3), (3,3), (4,3), 3\bigr), \\ P_3 &= \bigl(4, (4,4), 4\bigr), \end{aligned}$$

is a vertex-disjoint path system in $\Gamma_B^{(16)}(1,2,3 \mid 1,3,4)$. In fact, it is the unique such vertex-disjoint path system, and

$$w(\mathcal{P}) = (t_{1,2}t_{4,2}^{-1}t_{4,1})(t_{2,3})(t_{3,4}).$$

The reader may verify that the set $\Gamma_B^{(16)}(1,2 \mid 1,2)$ is empty.

The following is the q-analogue of a special case of Lindström’s lemma:

Theorem 4.1.7 [Casteels 2011, Theorem 4.4]. *If $[I \mid J]$ has maximum coordinate at most (r, s), then, as an element of $\mathcal{O}_q((\mathbb{K}^\times)^{m\times n})$,*

$$[I \mid J] = \sum_{\mathcal{P}\in\Gamma_B^{(t)}(I \mid J)} w(\mathcal{P}). \qquad \square$$

The proof in [Casteels 2011] deals with the case $t = mn$ and uses a technique similar to the “tail-switching” method of Theorem 3.1.12. The same proof is valid here due to the assumption that the maximum coordinate of the minor is at most (r, s).

Example 4.1.8. In the Cauchon graph of Figure 7, say with $t = 16$, there is no vertex-disjoint path system from $\{1,2\}$ to $\{1,2\}$. Theorem 4.1.7 tells us that $[1,2 \mid 1,2] = 0$. This may be verified directly:

$$\begin{aligned}[1,2 \mid 1,2] &= x_{1,1}x_{2,2} - qx_{1,2}x_{2,1} \\ &= (t_{1,2}t_{4,2}^{-1}t_{4,1} + t_{1,3}t_{2,3}^{-1}t_{2,2}t_{4,2}^{-1}t_{4,1} + t_{1,3}t_{4,3}^{-1}t_{4,1})(t_{2,2} + t_{2,3}t_{4,3}^{-1}t_{4,2}) \\ &\qquad - q(t_{1,2} + t_{1,3}t_{2,3}^{-1}t_{1,3}t_{4,3}^{-1}t_{4,2})(t_{2,2}t_{4,2}^{-1}t_{4,1} + t_{2,3}t_{4,3}^{-1}t_{4,1}) \\ &= 0.\end{aligned}$$

Similarly, it may be checked that

$$\begin{aligned}[1,2,3 \mid 1,3,4] &= x_{1,1}x_{2,3}x_{3,4} - qx_{1,1}x_{2,4}x_{3,3} - qx_{1,3}x_{2,1}x_{3,4} - q^3x_{1,4}x_{2,3}x_{3,1} \\ &\qquad + q^2x_{1,3}x_{2,4}x_{3,1} + q^2x_{1,4}x_{2,1}x_{3,3} \\ &= w(P_1)w(P_2)w(P_3) \\ &= (t_{1,2}t_{4,2}^{-1}t_{4,1})(t_{2,3})(t_{3,4}),\end{aligned}$$

where P_1, P_2 and P_3 are as in Example 4.1.6.

Before moving on, a quick application of Theorem 4.1.7 is worth mentioning: the well-known fact that in $\mathcal{O}_q(\mathcal{M}_{n,n}(\mathbb{K}))$ the *quantum determinant*

$$D_q = [1,2\dots,n \mid 1,2,\dots,n]$$

is central. Indeed, it is easy to see that there is exactly one vertex-disjoint path system from $[n]$ to $[n]$ in $G_{\varnothing}^{n\times n}$, namely $\mathcal{P} = (P_1,\dots,P_n)$, where $P_i = (i,(i,i),i)$ for each $i \in [n]$. Hence,

$$D_q = t_{1,1}t_{2,2}\cdots t_{n,n}.$$

Centrality of D_q follows from the observation that the right-hand side commutes with every generator $t_{i,j}^{\pm 1}$ of $\mathcal{O}_q((\mathbb{K}^\times)^{m\times n})$.

The next result was given as Theorem 4.5 in [Casteels 2011], but under the additional assumption that q is transcendental over $\mathbb{Q}$. We here provide a proof for when q is a nonzero non-root of unity.

Theorem 4.1.9. *A quantum minor $[I \mid J]$ with maximum coordinate at most (r,s) equals zero if and only if there does* not *exist a vertex-disjoint path system from I to J; i.e., if and only if $\Gamma_B^{(t)}(I \mid J) = \varnothing$.*

Proof. If $\Gamma_B^{(t)}(I \mid J) = \varnothing$, then Theorem 4.1.7 implies that $[I \mid J] = 0$.

Now suppose $\Gamma_B^{(t)}(I \mid J) \neq \varnothing$, i.e., there is at least one vertex-disjoint path system from I to J. The weight of a vertex-disjoint path system $\mathcal{P}$ is equal to

$q^{\alpha}\boldsymbol{t}^{M_{\mathcal{P}}} \in \mathcal{O}_q((\mathbb{K}^{\times})^{m\times n})$ for some integer α, where

$$(M_{\mathcal{P}})_{i,j} = \begin{cases} 1 & \text{if there is a path in } \mathcal{P} \text{ with a } \Gamma\text{-turn at } (i,j); \\ -1 & \text{if there is a path in } \mathcal{P} \text{ with a } \lrcorner\text{-turn at } (i,j); \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, if for any distinct $\mathcal{P}, \mathcal{Q} \in \Gamma_B^{(t)}(I \mid J)$ one has $M_{\mathcal{P}} \neq M_{\mathcal{Q}}$, then, by Theorem 4.1.7 and Proposition 2.1.9, we may conclude that $[I \mid J] \neq 0$.

Suppose $\mathcal{P} = (P_1, \dots, P_k)$ and $\mathcal{Q} = (Q_1, \dots, Q_k)$ are two vertex-disjoint path systems from I to J and that $M_{\mathcal{P}} = M_{\mathcal{Q}}$, i.e., a path in $\mathcal{P}$ has a Γ-turn or a $\lrcorner$-turn at (i, j) if and only if a path in Q does. We aim to show that $\mathcal{P} = \mathcal{Q}$. First, consider the paths P_k and Q_k. Let (i_k, ℓ) be the first vertex where P_k turns and (i_k, ℓ') the first vertex where Q_k turns. If $\ell > \ell'$, then Q_k goes straight through (i_k, ℓ'). However, since $\mathcal{Q}$ contains some path Q that turns at (i, ℓ), this implies (since B is a Cauchon diagram) that Q and Q_k intersect, contradicting the choice of $\mathcal{Q}$ as a vertex-disjoint path system. The symmetric case shows that $\ell \not< \ell'$ and hence $\ell = \ell'$. A similar argument can then be applied to the remainder of the turning vertices (if any) in P_k and Q_k, from which we conclude that $P_k = Q_k$. Repeating the argument with P_{k-1} and Q_{k-1}, etc., we see that $\mathcal{P} = \mathcal{Q}$, as desired. □

Recall the map $\sigma_B^{(t)} : A^{(t)} \to A_B^{(t)}$ of Section 3.3.

Corollary 4.1.10. *A quantum minor $[I \mid J]^{(t)} \in A^{(t)}$ with maximum coordinate at most (r, s) is in $\ker(\sigma_B^{(t)})$ if and only if there does not exist a vertex-disjoint path system from I to J in $G_B^{m\times n}$, i.e., $\Gamma_B^{(t)}(I \mid J) = \varnothing$.* □

We conclude this section by showing how one may construct new vertex-disjoint path systems from I to J from old ones. First, suppose i is a row vertex and j is a column vertex in $G_B^{m\times n}$, and consider two paths $P : i \to j$ and $Q : i \to j$. Let $(i = v_0, \dots, v_k = j)$ be the subsequence of all vertices that P and Q have in common. For each $a \in [k]$, let P_a and Q_a denote the subpaths of P and Q, respectively, starting at v_{a-1} and ending at v_a. If $P_a \neq Q_a$, then the first edge of P_a is perpendicular to the first edge of Q_a. If the first edge of P_a is horizontal, let us say that P_a is *above* Q_a, otherwise P_a is *below* Q_a. Now consider the paths

$$U_a = \begin{cases} P_a & \text{if } P_a = Q_a; \\ P_a & \text{if } P_a \text{ is above } Q_a; \\ Q_a & \text{if } Q_a \text{ is above } P_a; \end{cases}$$

and

$$L_a = \begin{cases} P_a & \text{if } P_a = Q_a; \\ P_a & \text{if } P_a \text{ is below } Q_a; \\ Q_a & \text{if } Q_a \text{ is below } P_a. \end{cases}$$

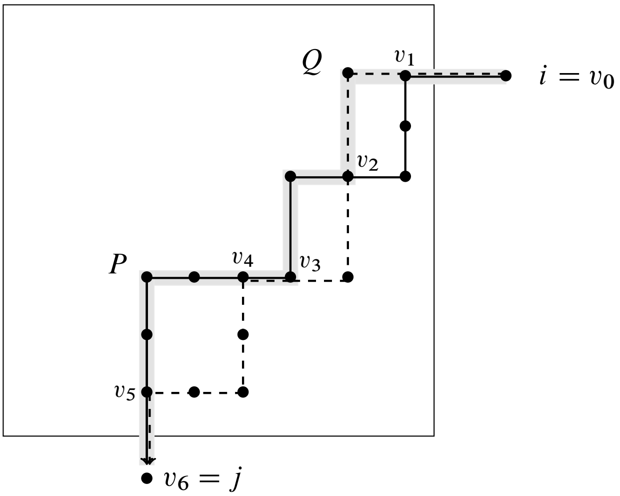


Figure 8. P is the solid path; Q is the dashed path; $U(P, Q)$ is the shadowed path.

Definition 4.1.11. With notation as in the preceding paragraph, we let $U(P, Q) : i \to j$ be the path

$$U(P, Q) = U_1 \cup U_2 \cup \cdots \cup U_k$$

and $L(P, Q) : i \to j$ the path

$$L(P, Q) = L_1 \cup L_2 \cup \cdots \cup L_k.$$

Example 4.1.12. With respect to Figure 8, U_1 is the solid path from $i = v_0$ to v_1, U_2 is the dashed path from v_1 to v_2, U_3 is the solid path from v_2 to v_3, etc. On the other hand, L_1 is the solid path from $i = v_0$ to v_1, L_2 is the solid path from v_1 to v_2, L_3 is the solid path from v_2 to v_3, etc.

The following lemma states the key property of $U(P, Q)$ that we require:

Lemma 4.1.13. *For a row vertex i and column vertex j in $G_B^{m \times n}$, consider two paths $P : i \to j$ and $Q : i \to j$. Suppose that $R : i' \to j'$ is a path with $i' > i$. If R is disjoint from either P or Q, then R is disjoint from $U(P, Q)$.*

Proof. With respect to P and Q, we use the notation of the paragraph just prior to Example 4.1.12. Without loss of generality, suppose P and R are disjoint.

If R and $U(P, Q)$ have a vertex w in common, then $w \in Q$ and there exists an a such that w is in the subpath Q_a of Q. Since $w \in U(P, Q)$, we have $U_a = Q_a$ for this a and so Q_a is above P_a. On the other hand, since $i' > i$, R must intersect the Jordan curve formed by P_a and Q_a. Since $G_B^{m \times n}$ is planar, the intersection occurs at a vertex of P, a contradiction. □

Corollary 4.1.14. *Let $i < i'$ be two row vertices and $j < j'$ two column vertices in $G_B^{m\times n}$. Suppose $P : i \to j$ and $P' : i' \to j'$ are disjoint paths and $Q : i \to j$ and $Q' : i' \to j'$ are disjoint paths. Then $U(P, Q)$ and $U(P', Q')$ are disjoint.*

Proof. By two applications of Lemma 4.1.13, $U(P, Q)$ is disjoint from both P' and Q'. Since $U(P', Q')$ consists only of subpaths coming from either P' or Q', we have that $U(P, Q)$ and $U(P', Q')$ are disjoint as well. □

Repeated application of Corollary 4.1.14 immediately gives the following result.

Corollary 4.1.15. *Let $\mathcal{P} = (P_1, \dots, P_k)$ and $\mathcal{Q} = (Q_1, \dots, Q_k)$ be vertex-disjoint path systems from I to J. Then*

$$U(\mathcal{P}, \mathcal{Q}) = (U(P_1, Q_1), \dots, U(P_k, Q_k))$$

is a vertex-disjoint path system from I to J. □

Now, if $\Gamma_B^{(t)}(I \mid J)$ is nonempty, then repeated application of Corollary 4.1.15 to the finitely many path systems in $\Gamma_B^{(t)}(I \mid J)$ shows that the next definition is sensible.

Definition 4.1.16. If $\Gamma_B^{(t)}(I \mid J) \neq \varnothing$, then the *supremum* of $\Gamma_B^{(t)}(I \mid J)$ is the (unique) vertex-disjoint path system $(Q_1, \dots, Q_k) \in \Gamma_B^{(t)}(I \mid J)$ such that for any $\mathcal{P} = (P_1, \dots, P_k) \in \Gamma_B^{(t)}(I \mid J)$ one has, for each $i \in [k]$,

$$U(Q_i, P_i) = Q_i.$$

For $L(P, Q)$, it is clear that results similar to Lemma 4.1.13, Corollary 4.1.14 and Corollary 4.1.15 hold. We omit their explicit statements here, but note that the next definition is also sensible.

Definition 4.1.17. If $\Gamma_B^{(t)}(I \mid J) \neq \varnothing$, then the *infimum* of $\Gamma_B^{(t)}(I \mid J)$ is the (unique) vertex-disjoint path system $(Q_1, \dots, Q_k) \in \Gamma_B^{(t)}(I \mid J)$ such that for any $\mathcal{P} = (P_1, \dots, P_k) \in \Gamma_B^{(t)}(I \mid J)$ one has, for each $i \in [k]$,

$$L(Q_i, P_i) = Q_i.$$

Example 4.1.18. Once again, consider the Cauchon graph shown in Figure 7. The supremum of $\Gamma_B^{(16)}(1, 3 \mid 1, 3)$ is the path system $(\tilde{Q}_1, \tilde{Q}_2)$, where

$$\begin{aligned}\tilde{Q}_1 &= \big(1, (1, 3), (1, 2), (2, 2), (4, 2), (4, 1), 1\big),\\ \tilde{Q}_2 &= \big(3, (3, 4), (3, 3), (4, 3), 3\big),\end{aligned}$$

while the infimum of $\Gamma_B^{(16)}(1, 3 \mid 1, 3)$ is the path system (Q_1, Q_2), where

$$\begin{aligned}Q_1 &= \big(1, (1, 3), (2, 3), (2, 2), (4, 2), (4, 1), 1\big),\\ Q_2 &= \big(3, (3, 4), (4, 4), (4, 3), 3\big).\end{aligned}$$

4.2. ***Gröbner bases.*** Gröbner basis theory is well known in commutative algebra, and fortunately many of its key aspects transfer easily to quantum matrices and the algebras $R^{(t)} \simeq A^{(t)}$. For a more general and detailed account of Gröbner basis theory for noncommutative algebras, we refer the reader to the book of Bueso, Gómez-Torrecillas and Verschoren [Bueso et al. 2003].

Throughout this section, we fix $t \in [mn]$, let (r, s) be the t-th smallest coordinate, and denote the matrix of generators of $A^{(t)}$ by $[x_{i,j}]$. We now define a total order of the lexicographic monomials in $A^{(t)}$.

Definition 4.2.1. The *matrix lexicographic order* $\prec$ on $\mathcal{M}_{m,n}(\mathbb{Z})$ is defined as follows. If $M \neq N \in \mathcal{M}_{m,n}(\mathbb{Z})$, let (k, ℓ) be the least coordinate in which M and N differ. Then we set

$$M \prec N \iff (M)_{k,\ell} < (N)_{k,\ell}$$

and say that "$M \prec N$ *at* (k, ℓ)".

If $M \prec N$ are both in $\mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$, then the matrix lexicographic order induces a total order (that we also call matrix lexicographic) on the lexicographic monomials of $A^{(t)}$, by setting

$$\boldsymbol{x}^M \prec \boldsymbol{x}^N \iff M \prec N.$$

By allowing the (r, s)-entry in M and N to be negative, this terminology extends to a total order on the lexicographic monomials of $A^{(t)}[x_{r,s}^{-1}]$.

For example, under the matrix lexicographic order, we have

$$x_{i,j} \prec x_{k,\ell} \iff (i, j) > (k, \ell).$$

If $(i, j), (k, \ell) \leq (r, s)$, and (i, j) is northwest of (k, ℓ), then we have the relation

$$x_{k,\ell}x_{i,j} = x_{i,j}x_{k,\ell} - (q - q^{-1})x_{i,\ell}x_{k,j}.$$

On the other hand, we also have

$$x_{i,\ell}x_{k,j} \prec x_{i,j}x_{k,\ell}.$$

Essentially by repeated application of these facts and the other relations amongst the standard generators, we obtain the following, which is a special case of the more general Proposition 2.4 in [Bueso et al. 2003]:

Proposition 4.2.2. *For $M, N \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$, the lexicographic expression of $\boldsymbol{x}^M\boldsymbol{x}^N$ is*

$$\boldsymbol{x}^M\boldsymbol{x}^N = q^{\alpha}\boldsymbol{x}^{M+N} + \sum_{L \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})} \alpha_L \boldsymbol{x}^L,$$

for some integer α and where for every $\alpha_L \neq 0$ one has $L \prec M + N$. □

Definition 4.2.3. Let $M, N \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$. We say that $\boldsymbol{x}^M$ *divides* $\boldsymbol{x}^N$ if

$$(M)_{i,j} \leq (N)_{i,j} \quad \text{for all } (i, j) \in [m] \times [n].$$

Using this terminology, we will use Proposition 4.2.2 in the following way:

Corollary 4.2.4. *Let $M, N \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$. If $\boldsymbol{x}^M$ divides $\boldsymbol{x}^N$, then there exists an integer α, matrices $L \prec N$, and scalars $\alpha_L \in \mathbb{K}^*$ such that*

$$\boldsymbol{x}^N = q^\alpha \boldsymbol{x}^M \boldsymbol{x}^{N-M} + \sum_L \alpha_L \boldsymbol{x}^L. \qquad \square$$

Remark 4.2.5. Proposition 4.2.2, Definition 4.2.3 and Corollary 4.2.4 extend to $A^{(t)}[x_{r,s}^{-1}]$ by allowing the (r, s)-entry in each matrix to be negative.

Definition 4.2.6. Let $a \in A^{(t)}$ with lexicographic expression

$$a = \sum_L \alpha_L \boldsymbol{x}^L.$$

The *leading term* of a is the maximum lex term of a with respect to the matrix lexicographic order. We denote the leading term of a by $\ell t(a)$.

We are now ready to give the definition of a Gröbner basis for a right ideal.

Definition 4.2.7. Let J be a right ideal of $A^{(t)}$, and let

$$G = \{g_1, g_2, \dots, g_k\} \subseteq J.$$

We say that G is a *Gröbner basis* for J if for every $a \in J$ there exists a $g_i \in G$ such that $\ell t(g_i)$ divides $\ell t(a)$.

If one has a Gröbner basis $\{g_1, g_2, \dots, g_k\}$ for a right ideal J, then one may find an expression for any $a \in J$ as a combination of the g_i recursively. If $\ell t(a)$ is divided by $\ell t(g_i)$, then by Corollary 4.2.4 we may write

$$a = g_i a' + b,$$

where $\ell t(b) \prec \ell t(a)$. Since $b \in J$, we can repeat the process if $b \neq 0$. As there are only finitely many lexicographic terms smaller than $\ell t(a)$, this will end after finitely many steps. Thus, the elements of the Gröbner basis generate J.

We will eventually deal with quantum minors, and in this context require the following, more refined version of Corollary 4.2.4.

Lemma 4.2.8. *Let $[I \mid J]^{(t)} \in A^{(t)}$ be a minor with maximum coordinate (i_k, j_k). Recalling* Remark 4.1.2, *if we write*

$$[I \mid J]^{(t)} = \sum_{\sigma \in S_k} (-q)^{\ell(\sigma)} \boldsymbol{x}^{P_\sigma},$$

then:

(1) *One has* $\ell t([I \mid J]^{(t)}) = \boldsymbol{x}^{P_{\mathrm{id}}}$, *where* id *is the identity permutation.*

(2) *If* $\boldsymbol{x}^{P_{\mathrm{id}}}$ *divides* $\boldsymbol{x}^{M}$ *for some* $M \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$, *then*

$$\boldsymbol{x}^{M} = q^{\alpha}[I \mid J]^{(t)}\boldsymbol{x}^{M-P_{\mathrm{id}}} + w \tag{4-1}$$

for some integer α *and* $w \in A^{(t)}$, *where if* $\ell t(w) = \boldsymbol{x}^{K}$, *then* $K \prec M$ *at a coordinate northwest of* (i_k, j_k).

The first part of Lemma 4.2.8 is a trivial observation. The justification for the second part is fairly technical, but its heart is the following auxiliary lemma. For this lemma we set $E_{k,\ell}$ to be the $m \times n$ matrix with 1 in coordinate (k, ℓ) and 0 elsewhere.

Lemma 4.2.9. *If* $(i, j) \in [m] \times [n]$ *and* $\boldsymbol{x}^{M} \in A^{(t)}$ *is such that all entries of* M *in coordinates larger than* (a, b) *are zero, then we may write*

$$\boldsymbol{x}^{M} x_{i,j} = q^{\alpha} x_{i,j} \boldsymbol{x}^{M} + w,$$

where $\alpha \in \mathbb{Z}$ *and if* $w \neq 0$ *and* $\boldsymbol{x}^{K}$ *is a lex term of* w, *then* M *and* K *are equal in all entries northeast of* (i, j). *Moreover, if* $\ell t(w) = \boldsymbol{x}^{L}$, *then* $L \prec M + E_{i,j}$ *at a coordinate northwest of* (i, j)

Proof. We proceed by induction on j, starting with the easy observation that for $j = 1$, $x_{i,j}$ and $\boldsymbol{x}^{M}$ q^{*}-commute.

Now, fix $j > 1$. Consider the process of commuting $x_{i,j}$ to the left of $\boldsymbol{x}^{M}$, and define *step* (a, b) to be the point in this process just before we commute $x_{i,j}$ past $x_{a,b}^{(M)_{a,b}}$. For a given (a, b), let $M_0 \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$ be equal to M in all entries with coordinate less than (a, b), and let $M_1 = M - M_0$. Suppose we are at step (a, b) and we have an expression of the form

$$\boldsymbol{x}^{M} x_{i,j} = q^{\alpha} \boldsymbol{x}^{M_0} x_{a,b}^{(M)_{a,b}} x_{i,j} \boldsymbol{x}^{M_1} + w,$$

where $\alpha \in \mathbb{Z}$ and $w \in A^{(t)}$ is such that $\ell t(w) \prec M + E_{i,j}$, and if $w \neq 0$ and $\boldsymbol{x}^{K}$ is a lex term of w, then M and K are equal in all entries northeast of (i, j). We claim that there is such an expression for step $(a, b)^{-}$. Note that, once proven, repeated applications of this claim proves the inductive step, and hence the lemma.

If $x_{a,b}$ and $x_{i,j}$ q^{*}-commute, then the claim is trivial, so suppose $x_{a,b}x_{i,j} = x_{i,j}x_{a,b} + (q - q^{-1})x_{i,b}x_{a,j}$. Thus $b < j$ and, as is easily shown by induction on $(M)_{a,b}$, there is a $c \in \mathbb{K}$ such that

$$x_{a,b}^{(M)_{a,b}} x_{i,j} = x_{i,j} x_{a,b}^{(M)_{a,b}} + c x_{i,b} x_{a,b}^{(M)_{a,b}-1} x_{a,j}.$$

From this, we obtain

$$\begin{aligned} q^{\alpha} \boldsymbol{x}^{M_0} x_{a,b}^{(M)_{a,b}} x_{i,j} \boldsymbol{x}^{M_1} + w = {} & q^{\alpha} \boldsymbol{x}^{M_0} x_{i,j} x_{a,b}^{(M)_{a,b}} \boldsymbol{x}^{M_1} \\ & + c q^{\alpha} \boldsymbol{x}^{M_0} x_{i,b} x_{a,b}^{(M)_{a,b}-1} x_{a,j} \boldsymbol{x}^{M_1} + w. \end{aligned}$$

Notice that the claim is established if we can show that any lex term $\boldsymbol{x}^K$ of $\boldsymbol{x}^{M_0}x_{i,b}x_{a,b}^{(M)_{a,b}-1}x_{a,j}\boldsymbol{x}^{M_1}$ is such that K equals M northeast of (i,j).

As M_1 is zero in all entries with coordinates less than (a,b), there is a $\beta\in\mathbb{Z}$ with $x_{a,b}^{(M)_{a,b}-1}x_{a,j}\boldsymbol{x}^{M_1}=q^{\beta}\boldsymbol{x}^{M'_1}$, where $M'_1=M_1+((M)_{a,b}-1)E_{a,b}+E_{a,j}$. Since $b<j$, we apply the induction hypothesis for b to obtain

$$x_{i,b}\boldsymbol{x}^{M'_1}=q^{\gamma}\boldsymbol{x}^{M'_1+E_{i,b}}-w',$$

for some integer γ and $w'\in A^{(t)}$, where any lex term $\boldsymbol{x}^{K'}$ of w' is such that $K'\prec M'_1$ and K' equals M'_1 in all entries northeast of (i,b), and so in particular northeast of (i,j). Moreover, since $K'\prec M'_1$, we know that K' can only be zero in all entries with coordinate less than (a,b). For this reason, $\boldsymbol{x}^{M_0}\boldsymbol{x}^{K'}=\boldsymbol{x}^{M_0+K'}$, where M_0+K' is equal to M in all entries northeast of (i,j). As $M'_1+E_{i,b}$ also equals M in all entries northeast of (i,j), we have established the claimed expression at step $(a,b)^-$.

Finally, from the above procedure we also get $L\prec M+E_{i,j}$, where $\ell t(w)=\boldsymbol{x}^L$. Furthermore, since the commutation relations are homogeneous with respect to the grading introduced at the end of Section 2.1, we in fact have that $L\prec M+E_{i,j}$ at a coordinate northwest of (i,j). □

Lemma 4.2.9 roughly says that as we commute $x_{i,j}$ to the left of $\boldsymbol{x}^M$ and find the lexicographic expression of any new terms, one never needs to "create or destroy" any generator with coordinate northeast of (i,j).

Proof of Lemma 4.2.8, Part (2). By applying Lemma 4.2.9 to the generators corresponding to $\boldsymbol{x}^{P_{\mathrm{id}}}$ in $\boldsymbol{x}^M$, we find that there is an integer α and a $w\in A^{(t)}$ such that

$$\boldsymbol{x}^M=q^{\alpha}\boldsymbol{x}^{P_{\mathrm{id}}}\boldsymbol{x}^{M-P_{\mathrm{id}}}+w',$$

where $w'\in A^{(t)}$ and if $\ell t(w')=\boldsymbol{x}^K$, then $K\prec M$ at a coordinate northwest of (i_k,j_k). On the other hand, notice that if $\sigma\in S_k$ with $\sigma\neq\mathrm{id}$, then

$$\boldsymbol{x}^{P_\sigma}\boldsymbol{x}^{M-P_{\mathrm{id}}}=\boldsymbol{x}^{M-P_{\mathrm{id}}+P_\sigma}+w'',$$

where $\boldsymbol{x}^{M-P_{\mathrm{id}}+P_\sigma}$ is the leading term of the right side and $M-P_{\mathrm{id}}+P_\sigma\prec M$ at a coordinate northwest of (i_k,j_k). Our desired equation

$$\boldsymbol{x}^M=q^{\alpha}[I\mid J]^{(t)}\boldsymbol{x}^{M-P_{\mathrm{id}}}+w,$$

follows for some integer α and $w\in A^{(t)}$, where, if $\ell t(w)=\boldsymbol{x}^K$, then $K\prec M$ at a coordinate northwest of (i_k,j_k). □

4.3. ***Adding derivations and lexicographic expressions.*** Throughout this section, we fix $t\in[mn], t\neq 1$ and let (r,s) be the t-th smallest coordinate. Let $[x_{i,j}]$ be the matrix of generators for $A^{(t)}$ and $[y_{i,j}]$ the matrix of generators for $A^{(t-1)}$.

The proof of the main theorem requires a somewhat detailed understanding of the effect of the adding derivations map on the lexicographic expressions of an element $a \in A^{(t)}$ and its image $\overleftarrow{a} \in A^{(t-1)}[y_{r,s}^{-1}]$. This short section provides this information.

Recall from Section 2.2 that the adding derivations map is the homomorphism

$$\overleftarrow{\cdot} : A^{(t)} \to A^{(t-1)}[y_{r,s}^{-1}]$$

defined on the standard generators by

$$\overleftarrow{x_{i,j}} = \begin{cases} y_{i,j} + y_{i,s}y_{r,s}^{-1}y_{r,j} & \text{if } (i,j) \text{ is northwest of } (r,s); \\ y_{i,j} & \text{otherwise;} \end{cases}$$

or, equivalently, by

$$\overleftarrow{x_{i,j}} = \begin{cases} y_{i,j} + qy_{i,s}y_{r,j}y_{r,s}^{-1} & \text{if } (i,j) \text{ is northwest of } (r,s); \\ y_{i,j} & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{x}^M \in A^{(t)}$ and write

$$\boldsymbol{x}^M = x_{i_1,j_1}x_{i_2,j_2}\cdots x_{i_p,j_p},$$

where for each $k \in [p-1]$, $(i_k, j_k) \leq (i_{k+1}, j_{k+1})$. Let $\mathfrak{D}$ be the set of all k such that (i_k, j_k) is northwest of (r,s). Then we may write

$$\overleftarrow{\boldsymbol{x}^M} = \sum_{C \subseteq \mathfrak{D}} q^{|C|}\, \overset{C}{\overleftarrow{x_{i_1,j_1}}}\,\overset{C}{\overleftarrow{x_{i_2,j_2}}} \cdots \overset{C}{\overleftarrow{x_{i_p,j_p}}}, \tag{4-2}$$

where, for $C \subseteq \mathfrak{D}$,

$$\overset{C}{\overleftarrow{x_{i_k,j_k}}} = \begin{cases} y_{i_k,s}y_{r,j_k}y_{r,s}^{-1} & \text{if } k \in C; \\ y_{i_k,j_k} & \text{if } k \notin C. \end{cases}$$

Lemma 4.3.1. *With notation as in the preceding discussion, let $z \in A^{(t-1)}[y_{r,s}^{-1}]$ be a summand on the right side of* (4-2), *so that for some $C \subseteq D$,*

$$z = \overset{C}{\overleftarrow{x_{i_1,j_1}}}\,\overset{C}{\overleftarrow{x_{i_2,j_2}}} \cdots \overset{C}{\overleftarrow{x_{i_p,j_p}}}\,.$$

Then in the lexicographic expression of z, written as

$$z = \sum_{L_C \in \mathcal{M}_{m,n}(\mathbb{Z})} \alpha_{L_C}\, \boldsymbol{y}^{L_C}$$

where $\alpha_{L_C} \in \mathbb{K}^*$, the following hold:

(1) *For each L_C,*

$$(L_C)_{r,s} = (M)_{r,s} - |C|.$$

(2) *If $C \neq \varnothing$, then for every L_C, we have $L_C \prec M$ at the least (i_k, j_k) for which $k \in C$.*

(3) *For each L_C and for each $i \in [m] \setminus r$,*

$$(L_C)_{i,s} = (M)_{i,s} + |\{k \in C \mid i_k = i\}|.$$

(4) *If (i, j) is northwest of (r, s) and if*

$$(L_C)_{i,j} > (M)_{i,j} - |\{k \in C \mid (i_k, j_k) = (i, j)\}|,$$

then there is a coordinate (i, j') with $1 \leq j' < j$ such that

$$(L_C)_{i,j'} < (M)_{i,j'} - |\{k \in C \mid (i_k, j_k) = (i, j')\}|.$$

(5) *For each L_C, the entries in coordinates* not *north, west or northwest of (r, s) are equal to the corresponding entries in M.*

Proof. First, let us split the summand z by row indices, i.e., write

$$z = \left(\overleftarrow{x_{1,j_{1,1}}}^{C}\overleftarrow{x_{1,j_{1,2}}}^{C}\cdots\overleftarrow{x_{1,j_{1,p_1}}}^{C}\right)\cdots\left(\overleftarrow{x_{m,j_{m,1}}}^{C}\overleftarrow{x_{m,j_{1,2}}}^{C}\cdots\overleftarrow{x_{m,j_{1,pm}}}^{C}\right),$$

where, for each $i \in [m]$, the generators appearing in the monomial

$$\overleftarrow{x_{i,j_{i,1}}}^{C}\overleftarrow{x_{i,j_{i,2}}}^{C}\cdots\overleftarrow{x_{i,j_{i,p_i}}}^{C}$$

have indices

$$(a, b) \in \{(i, j) \mid j \in [n]\} \cup \{(r, j) \mid j \in [s]\}.$$

Moreover, if $y_{r,j}$ appears with $j \neq s$, then $y_{r,j}$ is to the right of any $y_{i,j'}$ with $j' < j$. In other words, such a $y_{r,j}$ q^*-commutes with every generator appearing to its right. Also, in $A^{(t-1)}$, we have that $y_{r,s}$ actually q^*-commutes with every generator of $A^{(t-1)}$. Thus $y_{r,s}^{-1}$ q^*-commutes with every generator in $A^{(t-1)}[y_{r,s}^{-1}]$ and we may write

$$\overleftarrow{x_{i,j_{i,1}}}^{C}\overleftarrow{x_{i,j_{i,2}}}^{C}\cdots\overleftarrow{x_{i,j_{i,p_1}}}^{C} = q^{\alpha}\,\boldsymbol{y}^{M_i}\,\boldsymbol{y}^{R_i}\,y_{r,s}^{-\beta},$$

where $\alpha \in \mathbb{Z}$, β is the number of occurrences of $y_{r,s}^{-1}$ in the left monomial, $M_i \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$ is the matrix defined by

$$(M_i)_{a,b} = \begin{cases} 0 & \text{if } a \neq i; \\ (M)_{i,b} - |\{k \in C \mid (i_k, j_k) = (i, b)\}| & \text{if } a = i \text{ and } 1 \leq b < s; \\ (M)_{i,s} + |\{k \in C \mid i_k = i\}| & \text{if } a = i \text{ and } b = s; \\ (M)_{i,b} & \text{if } s < b \leq n; \end{cases}$$

and R_i is a matrix whose nonzero entries appear only in coordinates between $(r, 1)$ and $(r, s - 1)$.

It follows that we may write

$$z = q^{\alpha'} \boldsymbol{y}^{M_1} \boldsymbol{y}^{R_1} \boldsymbol{y}^{M_2} \boldsymbol{y}^{R_2} \cdots \boldsymbol{y}^{M_{r-1}} \boldsymbol{y}^{R_{r-1}} \boldsymbol{y}^{R_r} y_{r,s}^{-|C|} \boldsymbol{y}^{L} \tag{4-3}$$

for some $\alpha' \in \mathbb{Z}$, where the entries of R_r equal those of M at coordinates between $(r, 1)$ and $(r, s-1)$ and are zero elsewhere, and where entries of L equal those of M at all coordinates greater than (r, s).

Next, let $y_{r,j}$ be a generator with $1 \leq j < s$, and consider $y_{r,j} \boldsymbol{y}^{M_i}$ for some $1 \leq i < r$. Recall that, for $j' < j$, we have the relation

$$y_{r,j} y_{i,j'} = y_{i,j'} y_{r,j} - (q - q^{-1}) y_{i,j} y_{r,j'}.$$

Repeated application of this relation implies that

$$y_{r,j} \boldsymbol{y}^{M_i} = \boldsymbol{y}^{M_i} y_{r,j} + \sum_{\ell} \alpha_\ell \boldsymbol{y}^{M_i^\ell} \boldsymbol{y}^{R^\ell},$$

for nonzero scalars α_ℓ and where:

(1) Every $M_i^\ell \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$ satisfies $M_i^\ell \prec M_i$, and the entries of each M_i^ℓ differ from those in M_i only between coordinates $(i, 1)$ and $(i, s-1)$;

(2) Each $R^\ell \in \mathcal{M}_{m,n}(\mathbb{Z}_{\geq 0})$ has nonzero entries only between coordinates $(r, 1)$ and $(r, s-1)$.

In particular, when finding the lexicographic expression of the monomial z written in the form of (4-3), we never create or destroy any of the generators $y_{i,s}$, $y_{r,s}^{\pm 1}$, nor any generator with coordinates *not* north, west or northwest. Parts (1), (3) and (5) of the lemma follow. It also follows that for every L_C and $i \in [r-1]$, if the entries in L_C and M with coordinates between $(i, 1)$ and $(i, s-1)$ differ, then the first different entry is smaller in L_C. This implies Part (4). Finally, Part (2) comes from the fact that each term in the lexicographic expression of z must start with $y_{i_1,j_1} \cdots y_{i_{k-1},j_{k-1}}$ since no subsequent relation produces a generator $y_{a,b}$ with $(a, b) < (i_k, j_k)$. □

Corollary 4.3.2. *If $a \in A^{(t)}$ and $\ell t(a) = \boldsymbol{x}^M$, then $\ell t(\overleftarrow{a}) = \boldsymbol{y}^M$.*

Proof. If $C \neq \varnothing$, then each term $\boldsymbol{y}^{L_C}$ in the resulting lexicographic expression satisfies $\boldsymbol{y}^{L_C} \prec \boldsymbol{y}^M$ by Part (2) of Lemma 4.3.1. On the other hand,

$$\boldsymbol{y}^M = \overset{\varnothing}{\overleftarrow{x_{i_1,j_1}}} \overset{\varnothing}{\overleftarrow{x_{i_2,j_2}}} \cdots \overset{\varnothing}{\overleftarrow{x_{i_p,j_p}}}. \qquad \square$$

4.4. ***Generators of $\mathscr{H}$-primes.*** We come to the main theorem of this paper. It is fairly straightforward to modify the proof and some of the above definitions to obtain the analogous result for left ideals. We remind the reader that the terms and notation used in the following proof can be found in the List of terms and notation.

Theorem 4.4.1. *Fix the following data: a Cauchon diagram B; $t \in [mn]$; (r, s) the t-th smallest coordinate; $[x_{i,j}]$ the matrix of generators for $A^{(t)}$; and the sequence of $\mathcal{H}$-primes $(K_1, \dots, K_{mn})$ where*

$$K_t = \ker\bigl(\sigma_B^{(t)}\bigr).$$

Let G_t be the set of all $x_{i,j}$ with $(i, j) > (r, s)$ and $(i, j) \in B$, together with all quantum minors in K_t whose maximum coordinate is at most (r, s). Then G_t is a Gröbner basis for K_t as a right ideal.

Proof. First, note that $B = \varnothing$ if and only if $K_1 = \langle 0\rangle$. On the other hand, in view of Theorem 2.2.1, we have $K_t = \langle 0\rangle$ for some $t \in [mn]$ if and only if $K_t = \langle 0\rangle$ for every $t \in [mn]$. Since the empty set generates $\langle 0\rangle$, we are done in the case $B = \varnothing$. From now on, we suppose $B \neq \varnothing$ and proceed by induction on t.

If $t = 1$, then the only minor in $A^{(1)} = \mathcal{O}_q(\mathbb{K}^{m\times n})$ whose maximum coordinate is $(1, 1)$ is

$$[1 \mid 1]^{(1)} = t_{1,1}.$$

Since $t_{1,1} \in K_1$ if and only if $(1, 1) \in B$, we see that G_1 is precisely the set of generators $t_{i,j}$ with $(i, j) \in B$. On the other hand, these $t_{i,j}$ generate K_1 by Theorem 2.3.4 and so Proposition 2.1.9 implies G_1 is indeed a Gröbner basis.

So now suppose $t \neq 1$ and that G_{t-1} is a Gröbner basis for K_{t-1}. Let $[y_{i,j}]$ be the matrix of generators for $A^{(t-1)}$. There are two cases to consider, according to whether or not $(r, s) \in B$.

If $(r, s) \in B$, then, as elements of $\mathcal{O}_q((\mathbb{K}^\times)^{m\times n})$, we have for each coordinate (i, j) that

$$\sigma_B^{(t)}(x_{i,j}) = \sigma_B^{(t-1)}(y_{i,j}).$$

Therefore,

$$a = \sum_L \alpha_L \boldsymbol{x}^L \in K_t$$

if and only if

$$a' = \sum_L \alpha_L \boldsymbol{y}^L \in K_{t-1}.$$

Hence, if $\boldsymbol{y}^M$ divides $\ell t(a')$, then $\boldsymbol{x}^M$ divides $\ell t(a)$.

Now, the previous paragraph also implies that if $[I \mid J]^{(t-1)} \in K_{t-1}$ with maximum coordinate at most $(r, s)^-$, then $[I \mid J]^{(t)} \in K_t$ with maximum coordinate strictly less than (r, s), so that $[I \mid J]^{(t)} \in G_t$. Also, if $(i, j) > (r, s)$ is such that $(i, j) \in B$, then $x_{i,j} \in K_t$. Finally, since $(r, s) \in B$,

$$[r \mid s]^{(t)} = x_{r,s} \in K_t.$$

It now follows that since G_{t-1} is a Gröbner basis for K_{t-1}, G_t is a Gröbner basis[5] for K_t.

Now assume $(r, s) \notin B$, i.e., $x_{r,s} \notin K_t$, and that G_{t-1} is a Gröbner basis for K_{t-1}. In the following we aim to verify that G_t satisfies Definition 4.2.7 for K_t, but this requires some effort. The strategy we employ is as follows. Suppose a nonzero $a \in K_t$ is chosen such that $\ell t(a) = \boldsymbol{x}^M$ is not divisible by the leading term of a member of G_t. Using the full power of the paths viewpoint developed above, we deduce in Claims 1 and 2 some structural properties of M. Using the information so obtained, we then find a term $\boldsymbol{y}^{N_C} \in A^{(t-1)}$ that *is not* divisible by the leading term of any member of G_{t-1} (Claim 3) yet *is* the leading term of an element of K_{t-1} (Claims 4 and 5). Of course, these opposing properties contradict the induction hypothesis.

Fix a nonzero, monic $a \in K_t$ with lexicographic expression

$$a = \boldsymbol{x}^M + \sum_L \alpha_L \boldsymbol{x}^L,$$

where $\ell t(a) = \boldsymbol{x}^M$. Furthermore, we may assume that a is homogeneous with respect to the grading introduced at the end of Section 2.1, i.e., that for each $i \in [m]$, the i-th row sum of every L and M are equal, and for every $j \in [n]$, the j-th column sum of M and every L are equal.

If there exists an $(i, j) \in B$ with $(i, j) > (r, s)$ and $(M)_{i,j} \geq 1$, then $x_{i,j} \in G_t$ divides $\ell t(a)$, and we are done. So we may assume no such (i, j) exists. In fact by Lemma 3.3.4 we may further assume that M and every L have the same values in each coordinate $(i, j) > (r, s)$, and, without loss of generality, that these entries are all zero, i.e., $(M)_{i,j} = 0 = (L)_{i,j}$ for all $(i, j) > (r, s)$.

Since $(r, s) \notin B$, we have

$$K_t = \overrightarrow{K_{t-1}}[x_{r,s}^{-1}] \cap A^{(t)},$$

and so there exists a $b \in K_{t-1}$ and a nonnegative integer h with

$$a = \overrightarrow{b}\, x_{r,s}^{-h}.$$

Then $b = \overleftarrow{a}\, y_{r,s}^h$, and, by Corollary 4.3.2,

$$\ell t(b) = \boldsymbol{y}^M y_{r,s}^h.$$

We henceforth call a minor in G_{t-1} whose leading term divides $\ell t(b)$ *critical*. Note that since the maximum coordinate of a critical minor is at most $(r, s)^-$, its leading term actually divides $\boldsymbol{y}^M$. By induction, there exists at least one critical

[5] In general we have actually shown that a subset of G_t is a Gröbner basis for K_t, but nothing is lost by adding the extra minors in K_t with maximum coordinates equal to (r, s).

minor. Now, if $[I \mid J]^{(t-1)}$ is critical and $[I \mid J]^{(t)} \in K_t$, then, since the maximum coordinate of $[I \mid J]^{(t)}$ is strictly less than (r, s), we have found an element of G_t whose leading term divides $\ell t(a)$, and we are done. *From now on, we assume that if $[I \mid J]^{(t-1)}$ is critical, then $[I \mid J]^{(t)} \notin K_t$.*

Claim 1. *If $[I \mid J]^{(t-1)}$ is critical, where $I = (i_1 < i_2 < \cdots < i_k)$ and $J = (j_1 < j_2 < \cdots < j_k)$, then we may assume the following*:

(1) *The set $\Gamma_B^{(t)}(I \mid J)$ is nonempty and every vertex-disjoint path system in it contains a path with a ⅃-turn at (r, s).*

(2) *If $(i_{k'}, j_{k'})$ is the largest diagonal coordinate northwest of (r, s), then*

$$[i_1, \ldots, i_{k'} \mid j_1, \ldots, j_{k'}]^{(t-1)}$$

is critical.

(3) *If (i_k, j_k) is northwest of (r, s), then for every (i, j) with $i_k < i \leq r$ and $j_k < j \leq s$, one has $(M)_{i,j} = 0$.*

Proof of Claim 1. Part (1): This is simply restating the assumption preceding the claim, since otherwise there is a vertex-disjoint path system in $\Gamma_B^{(t-1)}(I \mid J)$, i.e.,

$$[I \mid J]^{(t-1)} \notin K_{t-1}.$$

Part (2): By Part (1), there exists a ⅃-turn at (r, s) in any vertex-disjoint path system in $\Gamma_B^{(t)}(I \mid J)$. Hence $r \notin I$ (in particular, $i_k < r$), $s \notin J$ and at least (i_1, j_1) is northwest of (r, s). Therefore (i_k, j_k) is either northwest or northeast of (r, s).

If (i_k, j_k) is northwest of (r, s), then there is nothing to prove, so suppose (i_k, j_k) is northeast of (r, s). If $[I \setminus i_k \mid J \setminus j_k]^{(t-1)} \in K_{t-1}$, then replace $[I \mid J]^{(t-1)}$ with $[I \setminus i_k \mid J \setminus j_k]^{(t-1)}$ and restart this argument. So assume that (i_k, j_k) is northeast of (r, s) and $[I \setminus i_k \mid J \setminus j_k]^{(t-1)} \notin K_{t-1}$, i.e, there exists a vertex-disjoint path system

$$\mathcal{P} = (P_1, \ldots, P_{k-1}) \in \Gamma_B^{(t-1)}(I \setminus i_k \mid J \setminus j_k).$$

Let

$$\mathcal{Q} = (Q_1, \ldots, Q_k) \in \Gamma_B^{(t)}(I \mid J).$$

From Part (1), there exists a $Q_\alpha : i_\alpha \to j_\alpha$ containing (r, s) as a ⅃-turn. Clearly, we must have $\alpha = k'$, and $k' \neq k$ since (i_k, j_k) is northeast of (r, s). Recalling Corollary 4.1.15, consider the vertex-disjoint path system

$$\mathcal{R} = U(\mathcal{P}, \mathcal{Q} \setminus Q_k) \in \Gamma_B^{(t-1)}(I \setminus i_k \mid J \setminus j_k).$$

See Figure 9. Since $P_{k'}$ does not contain a ⅃-turn at (r, s), the path $U(P_{k'}, Q_{k'})$ does not contain a ⅃-turn at (r, s). Moreover, by Corollary 4.1.14, $\mathcal{R}$ is disjoint from Q_k. Hence, $\mathcal{R} \cup Q_k$ is a vertex-disjoint path system in the empty set $\Gamma_B^{(t-1)}(I \mid J)$, an impossibility.

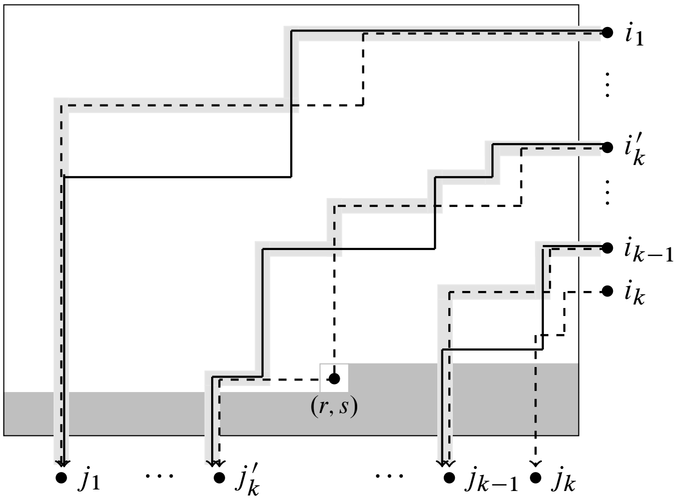


Figure 9. Illustration of the idea used to prove Part (2) of Claim 1. The dashed paths represent $\mathcal{Q} \in \Gamma_B^{(t)}(I \mid J)$. The solid paths represent $\mathcal{P} \in \Gamma_B^{(t-1)}(I \setminus i_k \mid J \setminus j_k)$. The shaded paths represent $U(\mathcal{P}, \mathcal{Q} \setminus Q_k)$.

Part (3): If $(i, j) = (r, s)$ and $(M)_{r,s} \geq 1$, then $[I \cup r \mid J \cup s]^{(t)}$ is a minor whose leading term divides $\boldsymbol{x}^M$ with maximum coordinate (r, s). The only path in $\Gamma_B^{(t)}(r, s)$ is $(r, (r, s), s)$. Hence, if $\Gamma_B^{(t)}(I \cup r \mid J \cup s)$ is nonempty, then any path system in this set would have a subpath system from I to J not using (r, s). But this is a vertex-disjoint path system in the empty set $\Gamma_B^{(t-1)}(I \mid J)$, an impossibility. Thus, $[I \cup r \mid J \cup s]^{(t)} \in G_t$ with leading term dividing $\boldsymbol{x}^M = \ell t(a)$, and there is nothing left to prove. So we may assume $(M)_{r,s} = 0$.

If $(i, j) \neq (r, s)$ but $(M)_{i,j} \geq 1$, then the leading term of $[I \cup i \mid J \cup j]^{(t-1)}$ divides $\boldsymbol{y}^M$. Since $[I \mid J]^{(t-1)} \in K_{t-1}$, there is no vertex-disjoint path system in $\Gamma_B^{(t-1)}(I \mid J)$ and so certainly no vertex-disjoint path system in $\Gamma_B^{(t-1)}(I \cup i \mid J \cup j)$. Thus, $[I \cup i \mid J \cup j]^{(t-1)}$ is critical and so there exists a $\mathcal{P} \in \Gamma_B^{(t)}(I \cup i \mid J \cup j)$. By Part (1) and vertex-disjointness, the path $P : i \to j \in \mathcal{P}$ is necessarily the path with a ⅃-turn at (r, s). But then $\mathcal{P} \setminus \{P\}$ is a vertex-disjoint path system in the empty set $\Gamma_B^{(t-1)}(I \mid J)$, an impossibility. This completes the proof of Claim 1. □

We now say that a coordinate (i, j) is *critical* if (i, j) is northwest of (r, s) and there exists a critical minor with (i, j) as its maximum coordinate.

Claim 2. *If (i, j) is critical, then every (i, j') for $j < j' < s$ with $(M)_{i,j'} \geq 1$ is critical, and every (i', j) for $i < i' < r$ with $(M)_{i',j} \geq 1$ is critical.*

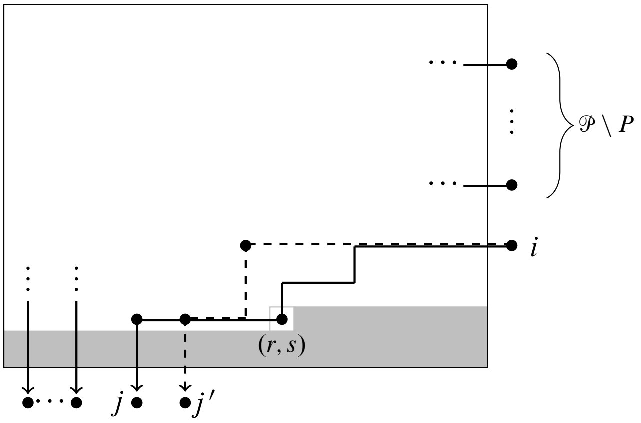


Figure 10. Illustration of the idea used in proving Claim 2. In the notation of that proof, the dashed line represents Q and the solid line represents P. The other vertices and partial paths represent $\mathcal{P} \setminus P = \mathcal{Q} \setminus Q$.

Proof of Claim 2 . Suppose $[I \mid J]^{(t-1)}$ is a critical minor whose maximum coordinate is (i, j). Notice that the leading term of

$$[I \mid J \setminus j \cup j']^{(t-1)}$$

divides y^M and its maximum coordinate is (i, j'), so it remains to show that this minor is in K_{t-1}.

Since $[I \mid J]^{(t-1)}$ is critical, we may consider the supremum $\mathcal{P} \in \Gamma_B^{(t)}(I \mid J) \neq \varnothing$, which, by Part (1) of Claim 1, contains a path $P : i \to j$ with a ⅃-turn at (r, s). Notice that P must have a horizontal subpath from (r, s) to (r, j), followed by a Γ-turn at (r, j), and then vertically down to the column vertex j. In particular, (r, j) is a white vertex. See Figure 10.

Suppose that $[I \mid J \setminus j \cup j']^{(t-1)} \notin K_{t-1}$, i.e., there exists a vertex-disjoint path system $\mathcal{Q}$ from I to $J \setminus j \cup j'$ in $\Gamma_B^{(t-1)}(I \mid J \setminus j \cup j')$. Therefore, the path $Q : i \to j'$ in $\mathcal{Q}$ does *not* use vertex (r, s). By considering the appropriate supremums, we may assume without loss of generality that $\mathcal{Q} \setminus Q = \mathcal{P} \setminus P$. Now, since $j' > j$, Q must intersect P in order to end at j'. Since Q cannot have a ⅃-turn at a (r, s) or any larger vertex, the Cauchon condition implies that (r, j') is a white vertex. On the other hand, $\mathcal{P} \setminus P$ is disjoint from both Q and P. If we let R be the path starting at i, equal to Q up to (r, j'), and then equal to P until the column vertex j, then R is a path from i to j that does not contain (r, s). Now $(\mathcal{P} \setminus P) \cup R$ is a vertex-disjoint path system in $\Gamma_B^{(t-1)}(I \mid J)$, a contradiction. That a coordinate (i', j) with $i < i' < r$ with $(M)_{i',j} \geq 1$ is critical is proven similarly. This completes the proof of Claim 2. □

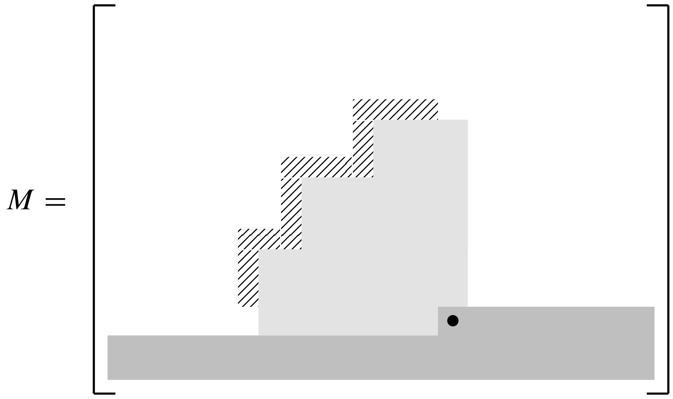


Figure 11. Structure of M: The bullet represents coordinate (r, s). All critical coordinates lie in the striped region. All entries in the two regions shaded solid gray are 0.

To summarize the discussion so far, we have shown that it suffices to assume the following:

- If $[I \mid J]^{(t-1)}$ is a critical minor, then $\Gamma_B^{(t)}(I \mid J) \neq \varnothing$ and every vertex-disjoint path system contains a path with a ⅃-turn at (r, s) (by Part (1) of Claim 1).
- Every critical minor contains a critical coordinate (by Part (2) of Claim 1).
- For each critical coordinate (i, j), there is a critical minor whose maximum coordinate is (i, j) (by definition).
- For each critical coordinate (i, j) (of which there exists at least one), $(M)_{k,\ell} = 0$ for all $i < k \leq r$ and $j < \ell \leq s$ (by Part (3) of Claim 1). In particular, no critical coordinate is northwest of another critical coordinate and so any critical minor contains a *unique* critical coordinate. See Figure 11.
- If (i, j) is northwest of (r, s) and (i, j) is *not* a critical coordinate, then no coordinate above or to its left is critical (by Claim 2).

The remainder of this proof will show that the above list of assumptions leads to a contradiction to the induction hypothesis.

Recalling the notation in Section 4.3, let

$$\ell t(a) = \boldsymbol{x}^M = x_{i_1,j_1} x_{i_2,j_2} \cdots x_{i_p,j_p},$$

and set

$$C = \{k \in [p] \mid (i_k, j_k) \text{ is critical}\},$$

where C is nonempty (since, by induction, there exists at least one critical minor, which in turn contains a critical coordinate). Consider the monomial

$$\overset{\xleftarrow{C}}{x}_{i_1,j_1}\overset{\xleftarrow{C}}{x}_{i_2,j_2}\cdots\overset{\xleftarrow{C}}{x}_{i_p,j_p}\,y_{r,s}^{h}.$$

By the assumptions just established, Lemma 4.3.1 and Proposition 4.2.2, the lexicographic expression of this monomial equals

$$q^{\alpha}\boldsymbol{y}^{N_C}+\sum_{L_C\in\mathcal{M}_{m,n}(\mathbb{Z})}\alpha_{L_C}\boldsymbol{y}^{L_C},\tag{4-4}$$

for some integer α and with every $L_C \prec N_C$, where

$$(N_C)_{i,j}=\begin{cases}0 & \text{if } (i,j) \text{ is critical;}\\ (M)_{i,j} & \text{if } i\neq r,\ j\neq s \text{ and } (i,j) \text{ not critical;}\\ (M)_{i,s}+\sum_{j'}(M)_{i,j'} & \text{if } i\neq r \text{ and } j=s;\\ (M)_{r,j}+\sum_{i'}(M)_{i',j} & \text{if } i=r \text{ and } j\neq s;\\ h-|C| & \text{if } i=r \text{ and } j=s;\end{cases}$$

and where the sum in the case that $i\neq r$ and $j=s$ is over all j' with (i,j') critical, and the sum in the case that $i=r$ and $j\neq s$ is over all i' with (i',j) critical. With respect to Figure 11, the entries in the striped region are 0 in N_C, while entries above (r,s) (respectively to the left of (r,s)) may become nonzero if there is a critical coordinate to the left (respectively above).

Claim 3. *The term $\boldsymbol{y}^{N_C}$ is not divisible by the leading term of any element of G_{t-1}. Consequently, $\boldsymbol{y}^{N_C}$ is not the leading term of any element of K_{t-1}.*

Proof of Claim 3. To the contrary, suppose that $\boldsymbol{y}^{N_C}$ is divisible by the leading term of some element in G_{t-1}. Since $(N_C)_{i,j}=(M)_{i,j}=0$ for every $(i,j)\geq(r,s)$, this element is a minor

$$[I\mid J]^{(t-1)},$$

where, say,

$$I=(i_1<\cdots<i_z)\quad\text{and}\quad J=(j_1<\cdots<j_z).$$

Now, $[I\mid J]^{(t-1)}$ does not contain a critical coordinate, since $(N_C)_{i,j}=0$ for all critical coordinates (i,j). Moreover, we may in this way conclude that $\boldsymbol{y}^{M}$ is not divisible by the leading term of $[I\mid J]^{(t-1)}$. By the structure of the entries of N_C compared to M, we then must have that $[I\mid J]^{(t-1)}$ contains a coordinate (i_k,j_k) in which $(N_C)_{i_k,j_k}>0$ while $(M)_{i,j}=0$, and so there are only two possibilities: either $(i_k,j_k)=(i_k,s)$, where (i_k,j'_k) is critical for some j'_k, or $(i_k,j_k)=(r,j_k)$, where (i'_k,j_k) is critical for some i'_k. We here show that the former possibility leads to a contradiction. The latter case is dealt with similarly.

Before we begin, we simplify our presentation slightly by further assuming that $(i_k, j_k) = (i_k, s)$ is the maximum coordinate of $[I \mid J]^{(t-1)}$, i.e., that $z = k$. The general case is obtained by simply adding in $i_{k+1}, \dots, i_z$ and $j_{k+1}, \dots, j_z$ to the respective index sets of every minor we consider below.

As $\boldsymbol{y}^M$ is divisible by the leading term of $[I \setminus i_k \mid J \setminus s]^{(t-1)}$ (a minor with no critical coordinate), we have $[I \setminus i_k \mid J \setminus s]^{(t-1)} \notin K_{t-1}$. So it is well-defined to set

$$\widetilde{\mathcal{Q}} = (\widetilde{Q}_1, \widetilde{Q}_2, \dots, \widetilde{Q}_{k-1})$$

to be the supremum and

$$\mathcal{Q} = (Q_1, Q_2, \dots, Q_{k-1})$$

to be the infimum of $\Gamma_B^{(t-1)}(I \setminus i_k \mid J \setminus s)$.

Because (i_k, j'_k) is critical for some j'_k, there exists, by Claim 1, a critical quantum minor $[I' \mid J']^{(t-1)}$ where, for a (possibly nonpositive) integer α, we write

$$I' = (i'_\alpha < i'_{\alpha+1} < \cdots < i'_k = i_k) \quad \text{and} \quad J' = (j'_\alpha < j'_{\alpha+1} < \cdots < j'_k).$$

Set

$$\widetilde{\mathcal{P}} = (\widetilde{P}_\alpha, \dots, \widetilde{P}_k)$$

to be the supremum and

$$\mathcal{P} = (P_\alpha, \dots, P_k)$$

to be the infimum of $\Gamma_B^{(t)}(I' \mid J')$. By Claim 1, P_k is a path from i'_k to j'_k in which (r, s) is a ⅃-turn.

The constructions to follow will show that if $\alpha \leq 1$, then we can construct a vertex-disjoint path system

$$\mathcal{R}_1 \in \Gamma_B^{(t-1)}(I \mid J),$$

or, if $\alpha > 1$, a vertex-disjoint path system

$$\mathcal{R}'_\alpha \in \Gamma_B^{(t-1)}(I' \mid J').$$

As both $\Gamma_B^{(t-1)}(I \mid J)$ and $\Gamma_B^{(t-1)}(I' \mid J')$ were assumed to be empty sets, either case will establish a contradiction and so complete the proof of Claim 3. The construction is fairly intricate, so we first give an indication on how we plan to proceed. For $\ell \in [k]$, let $I_\ell = (i_\ell < \cdots < i_k)$ and $J_\ell = (j_\ell < \cdots j_k)$. Define I'_ℓ and J'_ℓ for $\alpha \leq \ell \leq k$ similarly. The first step is to build a vertex-disjoint path system $\mathcal{R}_k \in \Gamma_B^{(t-1)}(I_k \mid J_k)$ using $\mathcal{Q}$. If $k = 1$, then we are done. Otherwise, we use $\mathcal{R}_k$ to build $\mathcal{R}'_k \in \Gamma_B^{(t-1)}(I'_k \mid J'_k)$. Again, if $\alpha = k$, then we are done. Now suppose we have found $\mathcal{R}_{\ell+1} \in \Gamma_B^{(t-1)}(I_{\ell+1} \mid J_{\ell+1})$ and $\mathcal{R}'_{\ell+1} \in \Gamma_B^{(t-1)}(I_{\ell+1} \mid J_{\ell+1})$ and that $\ell + 1 > \max(1, \alpha)$. We will show how to construct $\mathcal{R}_\ell \in \Gamma_B^{(t-1)}(I_\ell \mid J_\ell)$ using $\mathcal{R}_{\ell+1}$ and $\mathcal{R}'_{\ell+1}$. If $\ell = 1$ we are done. Otherwise, we construct $\mathcal{R}'_\ell \in \Gamma_B^{(t-1)}(I'_\ell \mid J'_\ell)$

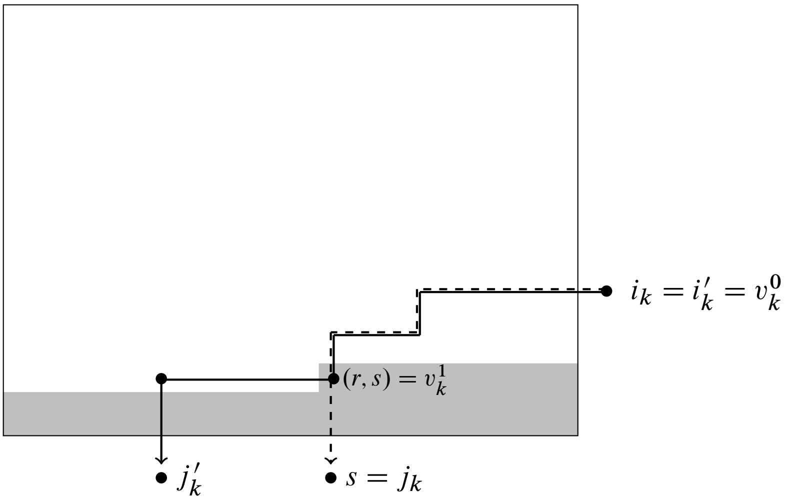


Figure 12. Construction of Q_k (dashed) from P_k (solid) in the proof of Claim 3.

using $\mathcal{R}'_{\ell+1}$ and the just-constructed $\mathcal{R}_\ell$. If $\ell = \alpha$ we are then done; otherwise we repeat the above, eventually ending with the desired vertex-disjoint path systems.

Now we give the promised details of the previous paragraph, beginning with the construction $\mathcal{R}_k$. Recall that $P_k \in \mathcal{P}$ has a subpath starting at row vertex $i'_k = i_k$ and ending at vertex (r,s). Define Q_k to be this subpath followed by the vertical path from (r,s) to column vertex s. For the purposes of the construction, set $v^0_k = i_k$, $v^1_k = (r,s)$, and note that v^0_k is the first vertex that P_k and Q_k have in common, while v^1_k is the last vertex they have in common. If one sets $R_k = Q_k$, then note that we (trivially) have $R_k = Q_k$ from i_k to v^0_k, $R_k = U(P_k, Q_k)$ from v^0_k to v^1_k, and $R_k = Q_k$ from v^1_k to $j_k = s$. See Figure 12.

Set $\mathcal{R}_k = (R_k)$. Of course, $\mathcal{R}_k$ is a vertex-disjoint path system from i_k to j_k in $\Gamma^{(t-1)}_B(I_k \mid J_k)$. If $k = 1$, then we are done, so we may assume $k > 1$.

In order to construct $\mathcal{R}'_k$, we first need to prove that $j_{k-1} \geq j'_k$. To the contrary, suppose $j_{k-1} < j'_k$, and consider

$$[I \mid J \setminus s \cup j'_k]^{(t-1)}.$$

If $[I \mid J \setminus s \cup j'_k]^{(t-1)} \in K_{t-1}$, then it is critical and so there exists a vertex-disjoint path system from I to $J \setminus s \cup j'_k$ with the path from i_k to j'_k containing a ⅃-turn at (r,s). But just as in the construction of Q_k above, we may replace this path with a path from i_k to s, thereby producing a vertex-disjoint path system from I to J in the empty set $\Gamma^{(t-1)}_B(I \mid J)$, which is absurd. Next, suppose $[I \mid J \setminus s \cup j'_k]^{(t-1)} \notin K_{t-1}$, so that there does exist a vertex-disjoint path system from I to $J \setminus s \cup j'_k$, where the path $Q' : i_k \to j'_k$ does not contain a ⅃-turn at (r,s). We may take this path

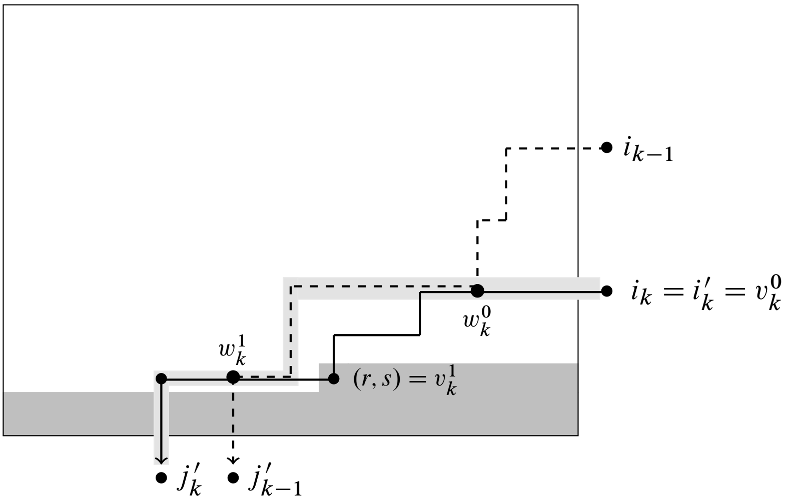


Figure 13. Q_{k-1} is the dashed path, P_k is the solid path, R'_k is the shadowed path.

system to be

$$(\widetilde{Q}_1, \ldots, \widetilde{Q}_{k-1}, Q').$$

Now $\widetilde{Q}_{k-1}$ is disjoint from Q', and so disjoint from $L(Q', P_k)$ by the lemma that is analogous to Corollary 4.1.14. But this latter path contains (r, s) (since P_k does) and so we may replace Q' with a path from i_k to s, thereby again impossibly producing a vertex-disjoint path system in the empty set $\Gamma^{(t-1)}_B(I \mid J)$. We can therefore conclude that $j_{k-1} \geq j'_k$.

As $k > 1$, consider Q_{k-1}, which, in particular, does not contain (r, s). Now, Q_{k-1} must intersect Q_k at a vertex coming before (r, s) on Q_k, as otherwise $\mathfrak{Q} \cup Q_k \in \Gamma^{(t-1)}_B(I \mid J)$. Let w^0_k be the first such common vertex. On the other hand, since $j_{k-1} \geq j'_k$ and Q_{k-1} goes above (r, s), Q_{k-1} must also share with P_k at least one vertex after (r, s). Let w^1_k be the last vertex that Q_{k-1} and P_k share. See Figure 13.

Define R'_k to be the path that equals P_k from i'_k to w^0_k, then equals $U(Q_{k-1}, P_k)$ from w^0_k to w^1_k, and then equals P_k from w^1_k to j'_k. Observe that R'_k does not contain (r, s), so that

$$\mathcal{R}'_k = (R'_k)$$

is a vertex-disjoint path system in $\Gamma^{(t-1)}_B(I'_k \mid J'_k)$. If $k = \alpha$, then again we have obtained the desired contradiction, and so we may assume $\alpha < k$.

Now let ℓ be an integer with $\max(\alpha, 1) \leq \ell < k$. Assume that $i_{\ell+1} \leq i'_{\ell+1}$, $j_\ell \geq j'_{\ell+1}$ and that we have the following data:

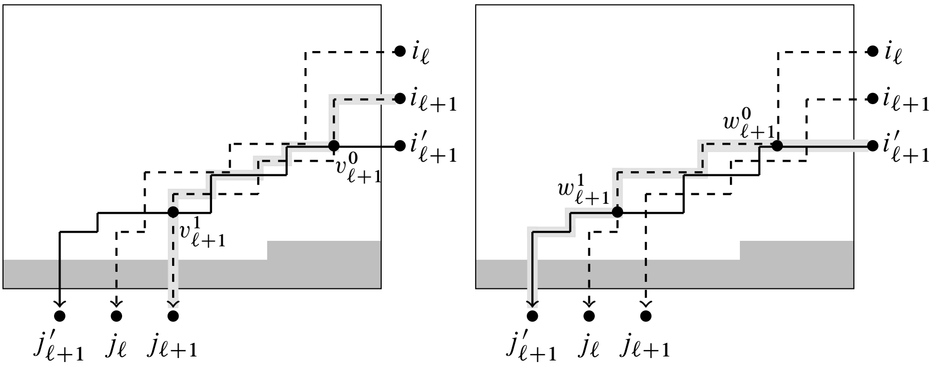


Figure 14. $R_{\ell+1}$ is the shaded path on the left diagram; $R'_{\ell+1}$ is the shaded path on the right diagram.

- We have a $\mathcal{R}_{\ell+1} = (R_{\ell+1}, \dots, R_k) \in \Gamma_B^{(t-1)}(I_{\ell+1} \mid J_{\ell+1})$. Moreover, there exists a vertex $v^0_{\ell+1}$ which is the first vertex that $P_{\ell+1}$ and $Q_{\ell+1}$ have in common, a vertex $v^1_{\ell+1}$ which is the last vertex that $P_{\ell+1}$ and $Q_{\ell+1}$ have in common, and $R_{\ell+1}$ equals $Q_{\ell+1}$ from $i_{\ell+1}$ to $v^0_{\ell+1}$, equals $U(P_{\ell+1}, Q_{\ell+1})$ from $v^0_{\ell+1}$ to $v^1_{\ell+1}$, and equals $Q_{\ell+1}$ from $v^1_{\ell+1}$ to $j_{\ell+1}$.
- We have a $\mathcal{R}'_{\ell+1} = (R'_{\ell+1}, \dots, R'_k) \in \Gamma_B^{(t-1)}(I'_{\ell+1} \mid J'_{\ell+1})$. Moreover, there exists a vertex $w^0_{\ell+1}$ which is the first vertex that $P_{\ell+1}$ and Q_ℓ have in common, a vertex $w^1_{\ell+1}$ which is the last vertex that $P_{\ell+1}$ and Q_ℓ have in common, and $R'_{\ell+1}$ equals $P_{\ell+1}$ from $i'_{\ell+1}$ to $w^0_{\ell+1}$, equals $U(P_{\ell+1}, Q_\ell)$ from $w^0_{\ell+1}$ to $w^1_{\ell+1}$, and equals $P_{\ell+1}$ from $w^1_{\ell+1}$ to $j'_{\ell+1}$.

We will construct a path $R_\ell : i_\ell \to j_\ell$ disjoint from $R_{\ell+1}$, but first we need to show that $i_\ell \le i'_\ell$. Suppose that $i_\ell > i'_\ell$. Since $j_\ell \ge j'_{\ell+1} > j'_\ell$, we may consider the minor

$$[I'' \mid J'']^{(t-1)} = [i'_\alpha, \dots, i'_\ell, i_\ell, \dots, i_{k-1} \mid j'_\alpha, \dots, j'_\ell, j_\ell, \dots, j_{k-1}]^{(t-1)}.$$

Note that this minor does not contain a critical coordinate since $[I \mid J]^{(t-1)}$ doesn't and (i_k, j_k) is the unique critical coordinate in $[I' \mid J']^{(t-1)}$. But as $\boldsymbol{y}^M$ is divisible by the leading term of $[I'' \mid J'']^{(t-1)}$, we know that $[I'' \mid J'']^{(t-1)}$ is not in K_{t-1}, i.e., $\Gamma_B^{(t-1)}(I'' \mid J'')$ is nonempty.

Indeed, $(\widetilde{P}_1, \dots, \widetilde{P}_\ell, Q_\ell, \dots, Q_{k-1}) \in \Gamma_B^{(t-1)}(I'' \mid J'')$, since for any path system in $\Gamma_B^{(t-1)}(I'' \mid J'')$ we choose, the subpath system from $\{i'_1, \dots, i'_\ell\}$ to $\{j'_1, \dots, j'_\ell\}$ may be replaced with the supremum of

$$\Gamma_B^{(t-1)}(i'_1, \dots, i'_\ell \mid j'_1, \dots, j'_\ell),$$

and the subpath system from $\{i_\ell, \dots, i_{k-1}\}$ to $\{j_\ell, \dots, j_{k-1}\}$ with the infimum of

$$\Gamma_B^{(t-1)}(i_\ell, \dots, i_{k-1} \mid j_\ell, \dots, j_{k-1}).$$

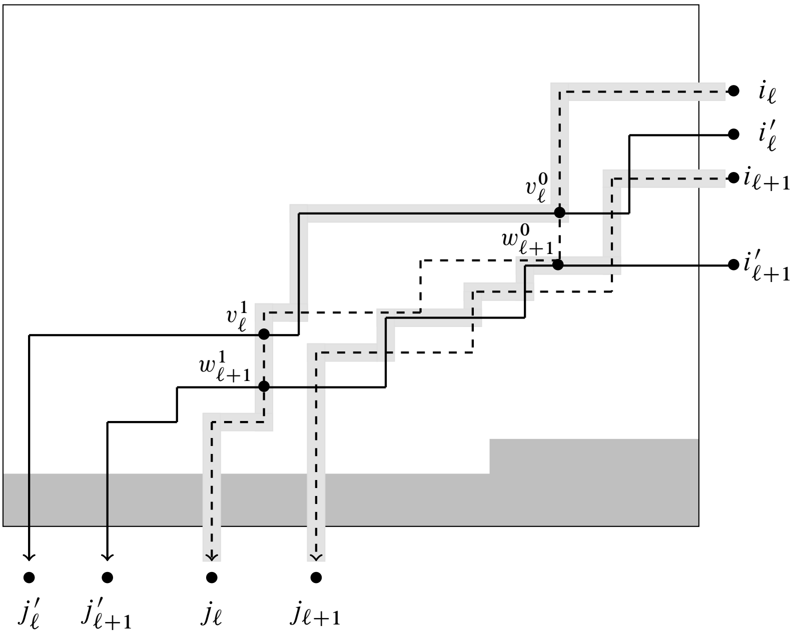


Figure 15. Constructing R_ℓ (upper shaded path). Note that it is disjoint from $R_{\ell+1}$ (lower shaded path).

These two sets are, of course, $(\widetilde{P}_1, \dots \widetilde{P}_\ell)$ and $(Q_\ell, \dots, Q_{k-1})$ respectively. In particular, this implies $\widetilde{P}_\ell$ is disjoint from both Q_ℓ. But $\widetilde{P}_\ell$ is also disjoint from $P_{\ell+1}$. By the construction of $R'_{\ell+1}$, it follows that $\widetilde{P}_\ell$ and $R'_{\ell+1}$ are also disjoint, so that

$$\{\widetilde{P}_1, \dots, \widetilde{P}_\ell\} \cup \mathcal{R}'_{\ell+1}$$

forms a vertex-disjoint path system in the empty set $\Gamma_B^{(t-1)}(I \mid J)$. Since this is an impossibility, it must be the case that $i_\ell \le i'_\ell$.

Next, we construct $\mathcal{R}_\ell$. Recall that $R'_{\ell+1}$ has a first vertex $w^0_{\ell+1}$ that is common to $P_{\ell+1}$ and Q_ℓ. On the other hand, since P_ℓ and $P_{\ell+1}$ are disjoint and $i_\ell \le i'_\ell < i'_{\ell+1}$, it must be the case that $P_{\ell+1}$ intersects Q_ℓ. Let v^0_ℓ be the first vertex they have in common and note that v^0_ℓ comes before $w^0_{\ell+1}$ on Q_ℓ. See Figure 15 for an example.

Next, observe that P_ℓ must also intersect Q_ℓ at a vertex coming after $w^0_{\ell+1}$. This is the case since otherwise P_ℓ is disjoint from $R'_{\ell+1}$ after $w^0_{\ell+1}$. But by the construction of $R'_{\ell+1}$, we would then have $(P_1, \dots, P_\ell) \cup \mathcal{R}'_\ell$, a vertex-disjoint path system in the empty set $\Gamma_B^{(t-1)}(I' \mid J')$. So, let v^1_ℓ be the last vertex that Q_ℓ and P_ℓ have in common. Define R_ℓ as the path equal to Q_ℓ from i_ℓ to v^0_ℓ, equal to $U(P_\ell, Q_\ell)$ from v^0_ℓ to v^1_ℓ, and then equal to Q_ℓ from v^1_ℓ to j_ℓ. Since Q_ℓ is disjoint from $Q_{\ell+1}$ up to v^0_ℓ and after v^1_ℓ, and $U(P_\ell, Q_\ell)$ is disjoint from $U(P_{\ell+1}, Q_{\ell+1})$, we see that R_ℓ is disjoint from $R_{\ell+1}$, and so

$$\mathcal{R}_\ell = \mathcal{R}_{\ell+1} \cup R_\ell \in \Gamma_B^{(t-1)}(i_\ell, \dots, i_k \mid j_\ell, \dots, j_k).$$

If $\ell = 1$, then we have obtained the required path system, completing the proof of this claim.

Assume $\ell > 1$. To construct $\mathcal{R}'_\ell$, we first must show that $j_{\ell-1} \geq j'_\ell$. To the contrary, suppose that $j_{\ell-1} < j'_\ell$. Now, $i_{\ell-1} < i_\ell \leq i'_\ell$, so we may consider the minor

$$[I''' \mid J''']^{(t-1)} = [i_1, \dots, i_{\ell-1}, i'_\ell, \dots, i'_k \mid j_1, \dots, j_{\ell-1}, j'_\ell, \dots, j'_k]^{(t-1)}.$$

Since y^M is divisible by the leading term of $[I''' \mid J''']^{(t-1)}$, there are two possibilities. If $[I''' \mid J''']^{(t-1)}$ is in K_{t-1}, then it is a critical minor, and so there is a vertex-disjoint path system in

$$\Gamma_B^{(t)}(i_1, \dots, i_{\ell-1}, i'_\ell, \dots, i'_k \mid j_1, \dots, j_{\ell-1}, j'_\ell, \dots, j'_k),$$

which we may take to be

$$(\widetilde{Q}_1, \dots, \widetilde{Q}_{\ell-1}, P_\ell, \dots, P_k).$$

Therefore, $\widetilde{Q}_{\ell-1}$ is disjoint from both P_ℓ and Q_ℓ, and so disjoint from R_ℓ by the latter path's construction. Hence, $(\widetilde{Q}_1, \dots, \widetilde{Q}_{\ell-1}) \cup \mathcal{R}_\ell$ is a vertex-disjoint path system in the empty set $\Gamma_B^{(t-1)}(I \mid J)$, an impossibility. The other possibility is that $[I''' \mid J''']^{(t-1)}$ is not in K_{t-1}. This possibility is dealt with in a manner similar to the above, where we justified the inequality $j_{k-1} \geq j'_k$. It follows that $j_{\ell-1} \geq j'_\ell$.

We now describe the construction of R'_ℓ. Since $\ell > 1$, consider $Q_{\ell-1}$. This path is disjoint from Q_ℓ. If $Q_{\ell-1}$ does not intersect P_ℓ at a vertex between v_ℓ^0 and v_ℓ^1, then $Q_{\ell-1}$ is disjoint from R_ℓ, so that $(Q_1, \dots, Q_{\ell-1}) \cup \mathcal{R}_\ell$ is a vertex-disjoint path system in the empty set $\Gamma_B^{(t-1)}(I \mid J)$, an impossibility. So we may let w_ℓ^0 be the first vertex that $Q_{\ell-1}$ shares with P_ℓ. Now, since $j'_\ell \leq j_{\ell-1} < j_\ell$, and the two subpaths of P_ℓ and Q_ℓ starting at v_ℓ^1 together with the line from j'_ℓ to j_ℓ form a closed curve in the plane, $Q_{\ell-1}$ must intersect P_ℓ at a vertex after v_ℓ^1. Let w_ℓ^1 be their last common vertex after v_ℓ^1. We now take R'_ℓ to be the path equal to P_ℓ from i'_ℓ to w_ℓ^0, equal to $U(P_\ell, Q_{\ell-1})$ from w_ℓ^0 to w_ℓ^1, and equal to P_ℓ from w_ℓ^1 to j'_ℓ. See Figure 16 for an example. That R'_ℓ is disjoint from $R'_{\ell+1}$ is seen similarly to when we showed that R_ℓ and $R_{\ell+1}$ are disjoint.

Of course, we now take

$$\mathcal{R}'_\ell = \mathcal{R}'_{\ell+1} \cup R'_\ell \in \Gamma_B^{(t-1)}(i'_\ell, \dots, i'_k \mid j'_\ell, \dots, j'_k).$$

If $\ell = \alpha$, then we are done. Otherwise continue as above. As this process ends when $\ell = \max(\alpha, 1)$, we eventually construct a vertex-disjoint path system in either the empty set $\Gamma_B^{(t-1)}(I \mid J)$ or the empty set $\Gamma_B^{(t-1)}(I' \mid J')$. This contradiction completes the proof of Claim 3. □

Claim 4. *The term* y^{N_C} *from Expression* (4-4) *is a lex term of* $b = \overleftarrow{a}\, y_{r,s}^h$.

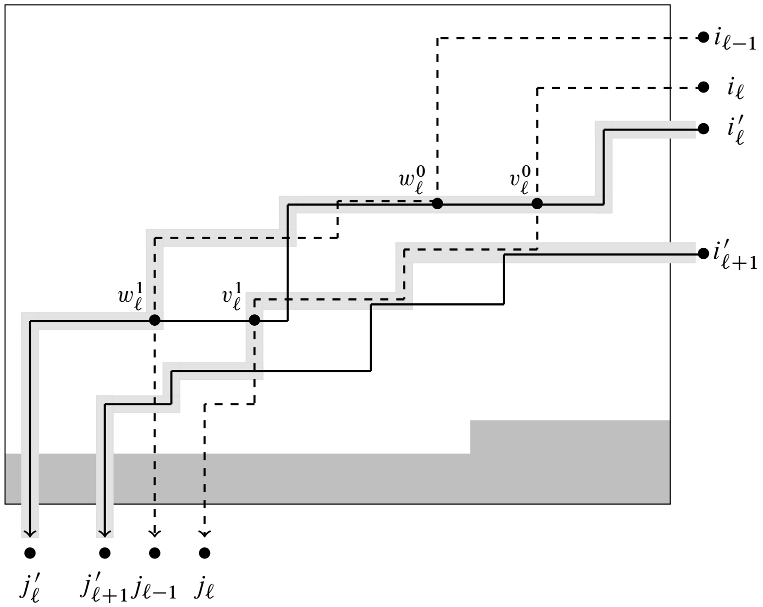


Figure 16. Constructing R'_ℓ (upper shaded path). Note that it is disjoint from $R'_{\ell+1}$ (lower shaded path).

Proof of Claim 4. Recall that a lexicographic term is said to be a *lex term* of an element of $A^{(t-1)}$ or $A^{(t)}$ if it has a nonzero coefficient in the lexicographic expression of that element.

We have already seen that $\boldsymbol{y}^{N_C}$ is a lex term of

$$\overset{C}{\overleftarrow{x}}_{i_1,j_1}\overset{C}{\overleftarrow{x}}_{i_2,j_2}\cdots\overset{C}{\overleftarrow{x}}_{i_p,j_p}\; y^h_{r,s}.$$

We will show that this is, in fact, the unique appearance of $\boldsymbol{y}^{N_C}$ in (the lexicographic expression of) any summand of

$$b = \overleftarrow{a}\, y^h_{r,s} = \overleftarrow{\boldsymbol{x}^M}\, y^h_{r,s} + \sum_L \alpha_L \overleftarrow{\boldsymbol{x}^L}\, y^h_{r,s},$$

and so is a lex term of b.

To start, consider in $\overleftarrow{\boldsymbol{x}^M}\, y^h_{r,s}$ the lexicographic expression of some

$$\overset{C'}{\overleftarrow{x}}_{i_1,j_1}\overset{C'}{\overleftarrow{x}}_{i_2,j_2}\cdots\overset{C'}{\overleftarrow{x}}_{i_p,j_p}\; y^h_{r,s} = \sum_{L_{C'}\in\mathcal{M}_{m,n}(\mathbb{Z})} \alpha_{L_{C'}}\boldsymbol{y}^{L_{C'}},$$

where $C' \neq C$. Suppose C' is chosen so that there is an $L_{C'}$ equal to N_C.

Now, by Lemma 4.3.1, each term $\boldsymbol{y}^{L_{C'}}$ satisfies $(L_{C'})_{r,s} = h - |C'|$. Since $(N_C)_{r,s} = h - |C|$, we must have if $|C'| = |C| > 0$. But, since $C \neq C'$, there must exist $k \in C'$ such that (i_k, j_k) is not a critical coordinate. Since (i_k, j_k) is not critical, we should have

$$(L_{C'})_{i_k,j_k} = (N_C)_{i_k,j_k} = (M)_{i_k,j_k} > (M)_{i_k,j_k} - |\{k' \in C' \mid (i_{k'}, j_{k'}) = (i_k, j_k)\}|.$$

By Part (4) of Lemma 4.3.1, there is a coordinate (i_k, j) with $j < j_k$ and

$$(L_{C'})_{i_k,j} < (M)_{i_k,j} = (N_C)_{i_k,j},$$

where the equality follows from the fact that since (i_k, j_k) is not critical, neither is (i_k, j) by Claim 2. Hence, $L_{C'}$ cannot be equal to N_C since their entries differ in coordinate (i_k, j). This is a contradiction and so we conclude that $\boldsymbol{y}^{N_C}$ is a lex term of $\overleftarrow{\boldsymbol{x}^M}\, y^h_{r,s}$.

Next, suppose

$$\boldsymbol{x}^L = x_{a_1,b_1} \cdots x_{a_t,b_t}$$

appears in a, where $(a_k, b_k) \le (a_{k+1}, b_{k+1})$ for each $k \in [t-1]$ and where $L \prec M$ at coordinate (i, j). With the notation of Section 4.3, consider

$$\overleftarrow{\boldsymbol{x}^L}\, y^h_{r,s} = \sum_D q^{|D|}\, \overset{D}{\overleftarrow{x_{a_1,b_1}}}\overset{D}{\overleftarrow{x_{a_2,b_2}}} \cdots \overset{D}{\overleftarrow{x_{a_t,b_t}}}\; y^h_{r,s}.$$

Suppose that $\boldsymbol{y}^{N_C}$ appears in

$$\overset{D}{\overleftarrow{x_{a_1,b_1}}}\overset{D}{\overleftarrow{x_{a_2,b_2}}} \cdots \overset{D}{\overleftarrow{x_{a_t,b_t}}}\; y^h_{r,s} = \sum_{L_D} \alpha_{L_D} \boldsymbol{y}^{L_D}.$$

By Lemma 4.3.1(5), every entry in an L_D with coordinates not northwest, north or west of (r, s) must equal the corresponding entry in L. Since we also require $L_D = N_C$ for some D, this implies that those entries are equal to the corresponding entry in M as well. Thus, (i, j) can only be north, west or northwest of (r, s). On the other hand, if $j = s$, then all entries in L and M in row i except coordinate (i, j) are equal. By homogeneity, this means that we must also have $(L)_{i,j} = (M)_{i,j}$, a contradiction. Hence (i, j) is not north of (r, s), and by similar reasoning (i, j) is not west of (r, s). Therefore, we may assume that $L \prec M$ at a coordinate (i, j) northwest of (r, s).

There are two cases to consider. First, suppose (i, j) is not a critical coordinate. In this case,

$$(N_C)_{i,j} = (M)_{i,j} > (L)_{i,j},$$

and so we may proceed as above by applying Part (4) of Lemma 4.3.1 to see that in order to have $(L_D)_{i,j} = (N_C)_{i,j}$, we would require an entry with coordinate (i, j') with $j' < j$ to satisfy

$$(L_D)_{i,j'} < (L)_{i,j'} = (M)_{i,j'} = (N_C)_{i,j'}.$$

Hence we cannot have $N_C = L_D$ in this case.

Next, suppose (i, j) is critical. Let (i, j_0) be the least critical coordinate in row i. Notice that no $(i, j') = (a_k, b_k)$ with $j' < j_0$ has $k \in D$, for reasons similar to the previous paragraph. Now, consider j' with $j_0 < j' \leq s$. By Part (3) of Claim 1 applied to (i, j_0), we know that every entry of M south of (i, j') is equal to zero. Hence, the sum of the entries in column j' of M is equal to $\sum_{i'=1}^{i} (M)_{i',j'}$. By homogeneity, this is equal to the sum of the entries in column j' of L. On the other hand, the entries north of (i, j') in L are equal to the corresponding entries in M. Since all entries of L are nonnegative, we see that

$$(L)_{i,j'} \leq (M)_{i,j'}$$

for every $j_0 < j' \leq s$. Also, since the entries of L and M are equal prior to (i, j_0) and $L \prec M$, we must also have $(L)_{i,j_0} \leq (M)_{i,j_0}$. But, since we know that $(L)_{i,j} < (M)_{i,j}$, applying Part (3) of Lemma 4.3.1 gives

$$\begin{aligned}(L_D)_{i,s} &= (L)_{i,s} + |\{k \in D \mid i_k = i\}| \\ &\leq (L)_{i,s} + \sum_{j'=j_0}^{s} (L)_{i,j'} < (M)_{i,s} + \sum_{j'=j_0}^{s} (M)_{i,j'} = (N_C)_{i,s}.\end{aligned}$$

Hence, we cannot have $L_D = N_C$ in this case either, and so this completes the proof of Claim 4. □

Claim 5. *There exists an element of K_{t-1} for which $\boldsymbol{y}^{N_C}$ is the leading term.*

Note that Claims 3 and 5 are incompatible, thus providing the required contradiction to the assumptions on the entries of M and completing the proof of Theorem 4.4.1.

Proof of Claim 5 . By Lemma 3.3.4, we may write

$$b = \sum_{i=0}^{\infty} b_i y_{r,s}^{i},$$

where finitely many $b_i \neq 0$ and each $b_i \in K_{t-1}$ with lexicographic expression using only generators with coordinates less than (r, s).

By Claim 4, $\boldsymbol{y}^{N_C}$ is a lex term of b and so, since $(N_C)_{r,s} = h - |C|$, it is a lex term of

$$z_0 = b_{h-|C|} y_{r,s}^{h-|C|}.$$

Suppose for a positive integer k that we have constructed an element $z_{k-1} \in K_{t-1}$ in which $\boldsymbol{y}^{N_C}$ is a lex term. Moreover, suppose any lex term of z_{k-1} that is greater than $\boldsymbol{y}^{N_C}$ also is a lex term of z_0. If $\ell t(z_{k-1}) = \boldsymbol{y}^{N_C}$, then we have found the required element of K_{t-1}. Otherwise, we construct in the following an element

$z_k \in K_{t-1}$ with the same properties as z_{k-1}, but in which there are fewer lex terms greater than $\boldsymbol{y}^{N_C}$. Since there are only finitely many lex terms of z_0 that are greater than $\boldsymbol{y}^{N_C}$, this process must end after finitely many steps, resulting in an element of K_{t-1} whose leading term is $\boldsymbol{y}^{N_C}$, as required.

Let

$$\ell t(z_{k-1}) = \boldsymbol{y}^L \succ \boldsymbol{y}^{N_C},$$

so that for some $\gamma_L, \gamma_{N_C} \in \mathbb{K}^*$ we may write

$$z_{k-1} = \gamma_L \boldsymbol{y}^L + \gamma_{N_C} \boldsymbol{y}^{N_C} + z'_{k-1}.$$

In particular, observe that in z'_{k-1} there are fewer lex terms greater than $\boldsymbol{y}^{N_C}$ than in z_{k-1}. Also, $\boldsymbol{y}^L \prec \boldsymbol{y}^M y_{r,s}^h$ since the latter term is the leading term of b but $\boldsymbol{y}^L \in b_{h-|C|} y_{r,s}^{h-|C|} \neq b_h y_{r,s}^h$ since $|C| > 0$. Finally, for $i \in [r-1]$, let C_i denote the critical coordinates in row i.

Let i_0 be the least index such that $C_i = C_{i_0}$ is nonempty. Let (c_0, d_0) be the least coordinate in C_{i_0}. Since $\boldsymbol{y}^{N_C} \prec \boldsymbol{y}^L \prec \boldsymbol{y}^M y_{r,s}^h$ and the entries of N_C and M at coordinates prior to (c_0, d_0) are equal, we have that the entries of L, M and N_C are equal prior to (c_0, d_0) as well.

Suppose $(c_0, d) \in C_{i_0}$ is such that $(L)_{c_0,d} > 0$. In this case, we proceed as follows. Since (c_0, d) is a critical coordinate, there is a critical minor $[I \mid J]^{(t-1)} \in K_{t-1}$ with maximum coordinate (c_0, d) whose leading term divides $\boldsymbol{y}^M$, and so divides $\boldsymbol{y}^L$ by the previous paragraph. By Lemma 4.2.8, we have

$$\boldsymbol{y}^L = q^\alpha [I \mid J]^{(t-1)} \boldsymbol{y}^{L-P_{\mathrm{id}}} + w,$$

where $w \in A^{(t-1)}$ has the property that if $\ell t(w) = \boldsymbol{y}^K$, then $K \prec L$ at an entry northwest of (c_0, d). Since all entries of L northwest of (c_0, d) are equal to those of N_C and M, we have that $\ell t(w) \prec \boldsymbol{y}^{N_C}$ as well.

Hence,

$$\begin{aligned} z_{k-1} &= \gamma_L \boldsymbol{y}^L + \gamma_{N_C} \boldsymbol{y}^{N_C} + z'_{k-1} \\ &= \gamma_L (q^\alpha [I \mid J]^{(t-1)} \boldsymbol{y}^{L-P_{\mathrm{id}}} + w) + \gamma_{N_C} \boldsymbol{y}^{N_C} + z'_{k-1}, \end{aligned}$$

so that if we define

$$z_k = z_{k-1} - \gamma_L q^\alpha [I \mid J]^{(t-1)} \boldsymbol{y}^{L-P_{\mathrm{id}}} = \gamma_{N_C} \boldsymbol{y}^{N_C} + \gamma_L w + z'_{k-1},$$

then we have $z_k \in K_{t-1}$ satisfying the desired properties described above.

Now, suppose each coordinate $(c_0, d) \in C_{i_0}$ is such that $(L)_{c_0,d} = 0$. Thus, L and N_C are equal in all entries prior to (c_0, s). Also, since $\boldsymbol{y}^L$ is a lex term of b, there must be a lex term $\boldsymbol{x}^{L'}$ of a so that $\boldsymbol{y}^L$ is a lex term of $\boldsymbol{x}^{\overleftarrow{L'}} y_{r,s}^h$. We also have $\boldsymbol{x}^{L'} \preceq \boldsymbol{x}^M$, and it follows by Part (2) of Lemma 4.3.1 that the entries in L' and M are equal prior to (c_0, d_0).

Now, as in the proof of Claim 4, we may apply homogeneity to conclude that $(L')_{c_0,d} \leq (M)_{c_0,d}$ for each $(c_0, d) \in C_{i_0}$, and if any of these inequalities are strict, then $(L)_{i_0,s} < (N_C)_{i_0,s}$, contradicting the assumption that $N_C \prec L$. Hence, L' and M have equal entries prior to (c_0, s).

Now, let i_1 be the second least index such that C_{i_1} is nonempty, and consider coordinates from (c_0, s) to $(c_1, d_1)^-$, where (c_1, d_1) is the least coordinate in C_{i-1}. Since $y^{N_C} \prec y^L$, we know that if any entry in L and N_C in these coordinates differ, then the first differing entry is larger in L than in N_C. On the other hand, the entries of N_C and M are equal in this range of coordinates. Thus, if the first differing entry is larger in L than in N_C, then this entry in L' is larger than in M, yet every entry prior in L' is equal to that in M, implying that $y^M \prec y^{L'}$, a contradiction. Hence, the entries in this range of coordinates are equal in N_C, M, L and L'.

Since all entries northwest of a critical coordinate are equal in M, N_C, L and L', we may now repeat the above arguments with the coordinates in C_{i_1}, and subsequent C_i if necessary. Eventually we must find a critical coordinate with a positive entry in L, as otherwise we would find that $N_C = L$, contradicting the assumption that $y^{N_C} \prec y^L$. Hence, we can always construct the required z_k and, eventually, an element of K_{t-1} with leading term y^{N_C}. This completes the proof of Claim 5 and the theorem. □

4.5. *Conclusions.* The motivating goal of this work was to demonstrate the conjecture of Goodearl and Lenagan that when $q \in \mathbb{K}^*$ is a non-root of unity, an $\mathcal{H}$-prime of $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$ is generated by the set of quantum minors it contains. That this is true is already immediate corollary of the $t = mn$ case of our Theorem 4.4.1. However, the theorem actually implies a sharper result, since we may consider a *minimal* Gröbner basis for the $\mathcal{H}$-prime. The idea here is simple: if G is a Gröbner basis for an ideal and if $g_1, g_2 \in G$ are such that $\ell t(g_1)$ is divisible by $\ell t(g_2)$, then $G \setminus g_1$ remains a Gröbner basis for the ideal. With respect to $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$, this means the following. Suppose $[I \mid J]^{(mn)} = [I \mid J]$ is a minor with $I = \{i_1 < i_2 < \cdots < i_k\}$ and $J = \{j_1 < j_2 < \cdots j_k\}$. If $L \subsetneq [k]$, $I' = I \cap \{i_\ell \mid \ell \in L\}$ and $J' = J \cap \{j_\ell \mid \ell \in L\}$, then call $[I' \mid J']$ a *diagonal subminor* of $[I \mid J]$. From the $t = mn$ case of Theorem 4.4.1 we find the following:

Corollary 4.5.1. *If $q \in \mathbb{K}^*$ is a non-root of unity, then every $\mathcal{H}$-prime K of $\mathcal{O}_q(\mathcal{M}_{m,n}(\mathbb{K}))$, is generated, as a right ideal, by those quantum minors in K with no diagonal subminor in K. These quantum minors form a minimal Gröbner basis for K with respect to the matrix lexicographic order.*

In the statement of Corollary 4.5.1, "right ideal" can be replaced by "left ideal" after proving the left ideal version of Theorem 4.4.1.

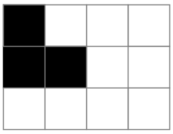


Figure 17. The Cauchon diagram corresponding to K.

Example 4.5.2. Let K be the $\mathcal{H}$-prime of $\mathcal{O}_q(\mathcal{M}_{3,4}(\mathbb{K}))$ corresponding to the Cauchon diagram in Figure 17. By using Corollary 4.1.10, we find that the quantum minors in K are

$$\{[123|123], [123|124], [12|12], [13|12], [23|12], [23|13], [23|23]\}.$$

Theorem 4.4.1 says that these form a Gröbner basis for K. However, $[12|12]$ is a diagonal subminor of $[123|123]$ and $[123|124]$. Therefore,

$$\{[12|12], [13|12], [23|12], [23|13], [23|23]\}$$

is a minimal (in fact reduced) Gröbner basis for K.

List of terms and notation

To assist in the reading of this paper, in particular the proof of Theorem 4.4.1, we below provide an index of some terms and notation used throughout this paper.

Coordinates	Beginning of Section 2.
Lexicographic order	Definition 2.1.1.
$(r,s)^-$	Definition 2.1.1.
Cauchon Diagram	Definitions 2.3.5 and 2.3.7.
$G_B^{m\times n}$	(Cauchon graph) Definition 3.1.1.
$\Gamma_B^{(t)}(I \mid J)$	Definition 4.1.4.
$U(P,Q)$	Definition 4.1.11.
$L(P,Q)$	Definition 4.1.11.
$U(\mathcal{P},\mathcal{Q})$ *(Supremum)*	Definition 4.1.16.
$L(\mathcal{P},\mathcal{Q})$ *(Infimum)*	Definition 4.1.17.
$A^{(t)}, A_B^{(t)}$	Definition 3.2.1.
$\boldsymbol{x}^N$	Notation 2.1.6.
Lexicographic expression	Definition 2.1.8.
Lex term of	Definition 2.1.8.
$\sigma_B^{(t)}$	Definition 3.3.1.
$\overrightarrow{a}$	Theorem 2.2.1.
$\overleftarrow{a}$	Theorem 2.2.1.

$\overset{C}{\overleftarrow{x_{i,j}}}$	Lemma 4.3.1 and preceding paragraph.
(Quantum) Minor $[I \mid J]^{(t)}_B, [I \mid J]^{(t)}, [I \mid J]$	Definition 4.1.1.
Diagonal coordinate (of a minor)	Definition 4.1.3.
Maximum coordinate (of a minor)	Definition 4.1.3.
$\prec$	Definition 4.2.1.
$\ell t(a)$ *(leading term of* $a \in A^{(t)}$*)*	Definition 4.2.6.
Gröbner Basis	Definition 4.2.7.
N_C	See Expression (4-4) just prior to Claim 3 in proof of Theorem 4.4.1.
Critical Minor	A minor in K_{t-1} whose leading term divides $\ell t(b) = \ell t(\overleftarrow{a}\, y^h_{r,s})$.
Critical Coordinate	A coordinate (i, j) that is northwest of (r, s) such that there exists a critical minor with (i, j) as its maximum coordinate.

Acknowledgements

Most of this work was completed while the author was a visiting assistant professor at the University of California, Santa Barbara. He remains grateful for the hospitality he received there. The author also thanks Ken Goodearl for many helpful discussions, and Stéphane Launois and Milen Yakimov for thoughtful comments. The anonymous referee made a careful, and very much appreciated, reading of this manuscript, and their suggestions resulted in a highly improved paper. A portion of this work was performed with the support of a Marie Skłodowska-Curie International Incoming Fellowship.

References

[Bell et al. 2012] J. Bell, K. Casteels, and S. Launois, "Enumeration of $\mathscr{H}$-strata in quantum matrices with respect to dimension", *J. Combin. Theory Ser. A* **119**:1 (2012), 83–98. MR 2012j:20150 Zbl 1232.05238

[Brown and Goodearl 2002] K. A. Brown and K. R. Goodearl, *Lectures on algebraic quantum groups*, Birkhäuser, Basel, 2002. MR 2003f:16067 Zbl 1027.17010

[Bueso et al. 2003] J. L. Bueso, J. Gómez-Torrecillas, and A. Verschoren, *Algorithmic methods in non-commutative algebra: applications to quantum groups*, Mathematical Modelling: Theory and Applications **17**, Kluwer, Dordrecht, 2003. MR 2005c:16069 Zbl 1063.16054

[Casteels 2011] K. Casteels, "A graph theoretic method for determining generating sets of prime ideals in quantum matrices", *J. Algebra* **330** (2011), 188–205. MR 2012c:20137 Zbl 1273.17016

[Cauchon 2003a] G. Cauchon, "Effacement des dérivations et spectres premiers des algèbres quantiques", *J. Algebra* **260**:2 (2003), 476–518. MR 2004g:16044 Zbl 1017.16017

[Cauchon 2003b] G. Cauchon, "Spectre premier de $O_q(M_n(k))$: image canonique et séparation normale", *J. Algebra* **260**:2 (2003), 519–569. MR 2004g:16045 Zbl 1024.16001

[Geiger and Yakimov 2014] J. Geiger and M. Yakimov, "Quantum Schubert cells via representation theory and ring theory", *Michigan Math. J.* **63**:1 (2014), 125–157. MR 3189471 Zbl 06293416

[Goodearl and Lenagan 2002] K. R. Goodearl and T. H. Lenagan, "Prime ideals invariant under winding automorphisms in quantum matrices", *Int. J. Math.* **13**:5 (2002), 497–532. MR 2003h:20091 Zbl 1054.16030

[Goodearl and Lenagan 2003] K. R. Goodearl and T. H. Lenagan, "Winding-invariant prime ideals in quantum 3×3 matrices", *J. Algebra* **260**:2 (2003), 657–687. MR 2004g:20069 Zbl 1059.16035

[Goodearl and Letzter 1998] K. R. Goodearl and E. S. Letzter, "Prime and primitive spectra of multiparameter quantum affine spaces", pp. 39–58 in *Trends in ring theory* (Miskolc, 1996), edited by V. Dlab and L. Márki, CMS Conf. Proc. **22**, American Mathematical Society, Providence, RI, 1998. MR 99h:16045 Zbl 0904.16001

[Goodearl and Letzter 2000] K. R. Goodearl and E. S. Letzter, "The Dixmier–Moeglin equivalence in quantum coordinate rings and quantized Weyl algebras", *Trans. Amer. Math. Soc.* **352**:3 (2000), 1381–1403. MR 2000j:16040 Zbl 0978.16040

[Goodearl et al. 2011a] K. R. Goodearl, S. Launois, and T. H. Lenagan, "Torus-invariant prime ideals in quantum matrices, totally nonnegative cells and symplectic leaves", *Math. Z.* **269**:1-2 (2011), 29–45. MR 2012h:16071 Zbl 1234.16018

[Goodearl et al. 2011b] K. R. Goodearl, S. Launois, and T. H. Lenagan, "Totally nonnegative cells and matrix Poisson varieties", *Adv. Math.* **226**:1 (2011), 779–826. MR 2012e:53162 Zbl 1210.14055

[Lam and Williams 2008] T. Lam and L. Williams, "Total positivity for cominuscule Grassmannians", *New York J. Math.* **14** (2008), 53–99. MR 2008m:05307 Zbl 1144.20029

[Launois 2004a] S. Launois, "Generators for $\mathcal{H}$-invariant prime ideals in $O_q(\mathcal{M}_{m,p}(\mathbb{C}))$", *Proc. Edinb. Math. Soc.* (2) **47**:1 (2004), 163–190. MR 2005c:20083 Zbl 1072.16036

[Launois 2004b] S. Launois, "Les idéaux premiers invariants de $O_q(\mathcal{M}_{m,p}(\mathbb{C}))$", *J. Algebra* **272**:1 (2004), 191–246. MR 2005a:20074 Zbl 1042.16033

[Launois and Lenagan 2009] S. Launois and T. H. Lenagan, "From totally nonnegative matrices to quantum matrices and back, via Poisson geometry", preprint, 2009. To appear in the *Proceedings of the Belfast Workshop on Algebra, Combinatorics and Dynamics*. arXiv 0911.2990

[Lindström 1973] B. Lindström, "On the vector representations of induced matroids", *Bull. London Math. Soc.* **5** (1973), 85–90. MR 49 #95 Zbl 0262.05018

[Postnikov 2006] A. Postnikov, "Total positivity, Grassmannians, and networks", preprint, 2006. arXiv math/0609764

[Takeuchi 2002] M. Takeuchi, "A short course on quantum matrices", pp. 383–435 in *New directions in Hopf algebras*, edited by S. Montgomery and H.-J. Schneider, Math. Sci. Res. Inst. Publ. **43**, Cambridge University Press, 2002. MR 2003i:16059 Zbl 1014.17015

[Talaska 2011] K. Talaska, "Combinatorial formulas for Γ-coordinates in a totally nonnegative Grassmannian", *J. Combin. Theory Ser. A* **118**:1 (2011), 58–66. MR 2012b:05070 Zbl 1232.05047

[Yakimov 2010] M. Yakimov, "Invariant prime ideals in quantizations of nilpotent Lie algebras", *Proc. Lond. Math. Soc.* (3) **101**:2 (2010), 454–476. MR 2011f:20109 Zbl 1229.17020

[Yakimov 2013] M. Yakimov, "A proof of the Goodearl–Lenagan polynormality conjecture", *Int. Math. Res. Not.* **2013**:9 (2013), 2097–2132. MR 3053415

Communicated by J. Toby Stafford
Received 2014-02-12 Revised 2014-08-06 Accepted 2014-09-12

K.L.Casteels@kent.ac.uk *School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent, CT2 7NF, United Kingdom*

ALGEBRA AND NUMBER THEORY 8:8 (2014)
dx.doi.org/10.2140/ant.2014.8.1913

Twisted Bhargava cubes

Wee Teck Gan and Gordan Savin

In his reinterpretation of Gauss's composition law for binary quadratic forms, Bhargava determined the integral orbits of a prehomogeneous vector space which arises naturally in the structure theory of the split group Spin_8. We consider a twisted version of this prehomogeneous vector space which arises in quasisplit Spin_8^E, where E is an étale cubic algebra over a field F. We classify the generic orbits over F by twisted composition F-algebras of E-dimension 2.

1. Introduction

The seminal work of Bhargava [2004a; 2004b; 2004c] has extended Gauss's composition law for binary quadratic forms to far more general situations. The key step in his extension is the investigation of the integral orbits of a group over $\mathbb{Z}$ on a lattice in a prehomogeneous vector space. The prehomogeneous vector space which plays a role in elucidating the nature of the classical Gauss's composition arises from a simply connected Chevalley group G of type D_4. More precisely, let $P = MN$ be a maximal parabolic subgroup of G corresponding to the branching point of the Dynkin diagram of type D_4. As it is readily seen from the Dynkin diagram,

MSC2010: primary 11S90; secondary 17A75, 17C40.
Keywords: Bhargava's cubes, twisted composition algebras.

the derived group M_{der} of the Levi factor M is isomorphic to SL_2^3. The unipotent radical N is 9-dimensional, and is a two-step nilpotent group with 1-dimensional center Z. The adjoint action of M_{der} on the abelian quotient N/Z is isomorphic to $V = V_2 \otimes V_2 \otimes V_2$, where V_2 is the standard 2-dimensional representation of SL_2. Since Bhargava regards an element of $(\mathbb{Z}^2)^{\otimes 3}$ as a cube whose vertices are labeled by elements of $\mathbb{Z}$, we shall refer to the prehomogeneous vector space V or its elements as Bhargava's cubes.

One of Bhargava's achievements is the determination of the corresponding integral orbits, i.e., $\mathrm{SL}_2(\mathbb{Z})^3$-orbits on $\mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$. In particular, he discovered that generic orbits are in bijection with isomorphism classes of tuples (A, I_1, I_2, I_3), where A is an order in an étale quadratic $\mathbb{Q}$-algebra and I_1, I_2 and I_3 are elements in the narrow class group of A, i.e., invertible fractional ideals, such that $I_1 \cdot I_2 \cdot I_3 = 1$. More precisely, to every cube Bhargava attaches three pairs (A_i, B_i), $i = 1, 2, 3$, of 2×2 matrices by slicing the cube in the three possible ways. In this way he obtains three binary quadratic forms

$$Q_i(x, y) = -\det(A_i x + B_i y).$$

A remarkable fact, discovered by Bhargava, is that the three forms have the same discriminant Δ. It is a degree-4 polynomial on V, invariant under the action of M_{der}. The cube is generic if $\Delta \neq 0$. In this case, the ring A is the unique quadratic order of discriminant Δ and the three fractional ideals I_i correspond to the three quadratic forms Q_i by a dictionary that essentially goes back to Gauss.

We now consider the group G over a field F of characteristic different from 2 and 3. The group G is exceptional in the sense that its outer automorphism group is isomorphic to S_3: no other absolutely simple linear algebraic group has such a large outer automorphism group. In particular, since S_3 is also the group of automorphisms of the split étale cubic F-algebra $F \times F \times F$, we see that every étale cubic F-algebra E determines a quasisplit form G_E. Fixing an épinglage of G defines a splitting of the outer automorphism group S_3 to $\mathrm{Aut}(G)$, so that S_3 acts on V by a group of symmetries of the cube, fixing two opposite vertices. Then the quasisplit group G_E contains a maximal parabolic subgroup $P_E = M_E N_E$, which is a twisted form of the parabolic P mentioned above. The derived group $M_{E,\mathrm{der}}$ of M_E is isomorphic to $\mathrm{Res}_{E/F}\,\mathrm{SL}_2$. The adjoint action of $M_{E,\mathrm{der}}$ on N_E/Z_E, where Z_E is the center of N_E, is isomorphic to a twisted form V_E of V. We shall call $V_E(F)$ (or its elements) the E-twisted Bhargava cube.

Since the action of S_3 on V permutes the three pairs (A_i, B_i) of 2×2 matrices obtained by slicing a cube in three different ways, it follows by Galois descent that Δ gives rise to a degree-4 polynomial invariant on V_E, denoted by Δ_E. It is a quasi-invariant for M_E. More precisely, if $v \in V_E(F)$ and $g \in M_E(F)$, then

$$\Delta_E(gv) = \chi(v)^2 \cdot \Delta_E(v),$$

where χ is a character of M_E given by the adjoint action on Z_E. An $M_E(F)$-orbit $\mathbb{O} \subset V_E(F)$ is called generic if $\Delta_E(v) \neq 0$ for one and hence for all $v \in \mathbb{O}$. If $\mathbb{O}$ is generic, then the quadratic algebra $K = F\big(\sqrt{\Delta_E(v)}\big)$ is étale. It is an invariant of the generic orbit.

The purpose of this paper is to classify the generic $M_E(F)$-orbits on $V_E(F)$. The main result is:

Theorem 1.1. *Let F be a field of characteristic different from 2 or 3. Fix an étale cubic F-algebra E.*

(i) *There are natural bijections between the following sets*:

(a) *Generic $M_E(F)$-orbits $\mathbb{O}$ on the E-twisted Bhargava cube.*

(b) *E-isomorphism classes of E-twisted composition algebras (C, Q, β) over F which are of E-dimension 2.*

(c) *E-isomorphism classes of pairs (J, i), where J is a Freudenthal–Jordan algebra over F of dimension 9 and*

$$i : E \hookrightarrow J$$

is an F-algebra homomorphism. Here an E-isomorphism from (J, i) to (J', i') is a commutative diagram

$$\begin{tikzcd} E \arrow[r, "i"] \arrow[d] & J \arrow[d] \\ E \arrow[r, "i'"] & J' \end{tikzcd}$$

where the first vertical arrows is the identity, while the second is an F-isomorphism of J and J'.

(ii) *The bijections in* (i) *identify*

$$\mathrm{Stab}_{M_E}(\mathbb{O}) \cong \mathrm{Aut}_E(C, Q, \beta) \cong \mathrm{Aut}_E(i : E \hookrightarrow J).$$

(iii) *Let $K = F\big(\sqrt{\Delta_E(v)}\big)$ be the étale quadratic algebra K attached to a generic orbit $\mathbb{O}$ containing v. Let $L = E \otimes_F K$. The group* $\mathrm{Stab}_{M_E}(\mathbb{O})$ *in* (ii) *sits in a short exact sequence of algebraic groups*

$$1 \longrightarrow T_{E,K} \longrightarrow \mathrm{Stab}_{M_E}(\mathbb{O}) \longrightarrow \mathbb{Z}/2\mathbb{Z} \longrightarrow 1,$$

where

$$T_{E,K}(F) = \{x \in L^\times : N_{L/E}(x) = 1 = N_{L/K}(x)\}$$

is a 2-dimensional torus and where the conjugation action of the nontrivial element of $\mathbb{Z}/2\mathbb{Z}$ on $T_{E,K}$ is given by $x \mapsto x^{-1}$.

The reader is probably not familiar with some terminology in the theorem, so an explanation is necessary. In order to define twisted composition algebras, recall that the algebra E carries a natural cubic form, the norm N_E. The norm defines a quadratic map $x \mapsto x^\#$ from E to E such that $x \cdot x^\# = N_E(x)$. For example, if $E = F^3$, then

$$N_E(x_1, x_2, x_3) = x_1 x_2 x_3 \text{ and } (x_1, x_2, x_3)^\# = (x_2 x_3, x_3 x_1, x_1 x_2).$$

Now, an E-twisted composition algebra (or simply twisted composition algebra) of E-dimension 2 is a triple (C, Q, β) where:

- C is an E-vector space of dimension 2.
- $Q : C \longrightarrow E$ is a quadratic form.
- $\beta : C \longrightarrow C$ is a quadratic map such that, for every $v \in C$ and $x \in E$,
  $$\beta(xv) = x^\# \cdot \beta(v) \quad \text{and} \quad Q(\beta(v)) = Q(v)^\#.$$
- If b_Q is the bilinear form associated to Q, then $b_Q(v, \beta(v)) \in F$ for every $v \in C$.

This definition is due to Springer, as is the bijection of the sets (b) and (c). More precisely, suppose we have an algebra embedding $i : E \hookrightarrow J$. Then we have a decomposition

$$J = E \oplus C,$$

where C is defined as the orthogonal complement to E with respect to the trace form on J. The upshot is that the Jordan algebra J determines the structure of a twisted composition algebra on C, and vice versa.

Our contribution is the bijection between the sets (a) and (b). Starting with a twisted cube, we define a twisted composition algebra. In fact, the construction works over $\mathbb{Z}$, and can be tied to Bhargava's description as follows. Let (I_1, I_2, I_3) be a triple of ideals in a quadratic order A such that $I_1 \cdot I_2 \cdot I_3 = A$. Let $N(I)$ denote the norm of the ideal I and $z \mapsto \bar{z}$ denote the action of the nontrivial automorphism of the étale quadratic $\mathbb{Q}$-algebra containing A. Let

$$C = I_1 \oplus I_2 \oplus I_3.$$

Then C is a twisted composition algebra with quadratic form $Q : C \to \mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$ defined by

$$Q(z_1, z_2, z_3) = \left(\frac{N(z_1)}{N(I_1)}, \frac{N(z_2)}{N(I_2)}, \frac{N(z_3)}{N(I_3)} \right)$$

and quadratic map $\beta : C \to C$ defined by

$$\beta(z_1, z_2, z_3) = (\bar{z}_2 \bar{z}_3 N(I_1), \bar{z}_3 \bar{z}_1 N(I_2), \bar{z}_1 \bar{z}_2 N(I_3)).$$

The key parts of the paper are as follows. In order to prove the correspondence of generic $M_E(F)$-orbits and twisted composition algebras, we give a Galois cohomological argument in Theorem 8.3, based on the observation that the stabilizer of a distinguished cube is isomorphic to the automorphism group of a distinguished twisted composition algebra. This gives a conceptual explanation for the existence of the bijection. However, for arithmetic applications (such as Bhargava's), it is essential to have an explicit description of the bijection. This is done in two steps. Firstly, after reviewing the theory of twisted composition algebras, we prove in Proposition 3.5 that every twisted composition algebra C of E-dimension 2 has a reduced basis, i.e., a basis of the form $\{v, \beta(v)\}$ for some $v \in C$. Secondly, by reinterpreting Bhargava's work in the framework of twisted composition algebras in Section 10, we attach to every generic E-twisted cube a twisted composition algebra together with a good basis. In this correspondence, changing the cube by another in the same $M_E(F)$-orbit corresponds to changing the good basis. Since reduced bases are good, every twisted composition algebra is obtained in this construction.

We also consider $\widetilde{M} = M \rtimes S_3$ and its twisted form $\widetilde{M}_E$. In this case, generic $\widetilde{M}_E(F)$-orbits correspond to the F-isomorphism classes of objects in (b) and (c). The isomorphisms of the stabilizer groups in (ii) lead us to another description of $T_{E,K}$, which we view as an *exceptional Hilbert 90* theorem. This is the topic of Section 11. We conclude the paper by illustrating the main results in the case where F is a local field.

2. Étale cubic algebras

Let F be a field of characteristic different from 2 and 3. Let $\overline{F}$ be a separable closure of F, with absolute Galois group $\mathrm{Gal}(\overline{F}/F)$.

2-1. ***Étale cubic algebras.*** An étale cubic algebra is an F-algebra E such that $E \otimes_F \overline{F} \cong \overline{F}^3$. More concretely, an étale cubic F-algebra is of the form

$$E = \begin{cases} F \times F \times F; \\ F \times K, \text{ where } K \text{ is a quadratic field extension of } F; \\ \text{a cubic field.} \end{cases}$$

Since the split algebra $F \times F \times F$ has automorphism group S_3 (the symmetric group on 3 letters), the isomorphism classes of étale cubic algebras E over F are naturally classified by the pointed cohomology set $H^1(F, S_3)$, or more explicitly by the set of conjugacy classes of homomorphisms

$$\rho_E : \mathrm{Gal}(\overline{F}/F) \longrightarrow S_3.$$

2-2. ***Discriminant algebra of*** $\boldsymbol{E}$**.** By composing the homomorphism ρ_E with the sign character of S_3, we obtain a quadratic character (possibly trivial) of $\mathrm{Gal}(\overline{F}/F)$

which corresponds to an étale quadratic algebra K_E. We call K_E the *discriminant algebra* of E. To be concrete,

$$K_E = \begin{cases} F \times F & \text{if } E = F^3 \text{ or a cyclic cubic field;} \\ K & \text{if } E = F \times K; \\ \text{the unique quadratic subfield in the Galois closure of } E & \text{otherwise.} \end{cases}$$

2-3. ***Twisted form of S_3.*** Fix an étale cubic F-algebra E. Then, via the associated homomorphism ρ_E, $\mathrm{Gal}(\overline{F}/F)$ acts on S_3 (by inner automorphisms) and thus defines a twisted form S_E of the finite constant group scheme S_3. For any commutative F-algebra A, we have

$$S_E(A) = \mathrm{Aut}_A(E \otimes_F A).$$

2-4. ***Quadratic map #.*** Given an étale cubic F-algebra, let $N_E : E \longrightarrow F$ be the norm map on E and let $\mathrm{Tr}_E : E \longrightarrow F$ be the trace map. Then N_E is a cubic form and Tr_E is a linear form on E. There is a quadratic map

$$\# : E \longrightarrow E$$

such that

$$a^\# \cdot a = a \cdot a^\# = N_E(a) \quad \text{for } a \in E.$$

It has an associated symmetric bilinear map

$$a \times b := (a+b)^\# - a^\# - b^\#.$$

For the split algebra F^3, we have:

$$N(a_1, a_2, a_3) = a_1 a_2 a_3, \quad \mathrm{Tr}(a_1, a_2, a_3) = a_1 + a_2 + a_3,$$
$$(a_1, a_2, a_3)^\# = (a_2 a_3, a_3 a_1, a_1 a_2).$$

We note the following identity in E:

$$(f \times y)y + fy^\# = \mathrm{Tr}_{E/F}(fy^\#). \tag{2.1}$$

This curious identity can be checked in $E \otimes_F \overline{F} \cong \overline{F}^3$; we leave it as an interesting exercise for the reader.

3. Twisted composition algebras

In this section, we introduce the E-twisted composition algebra of dimension 2 over E. This notion was introduced by Springer, and the two standard (perhaps only) references, covering many topics of this paper, are [Knus et al. 1998] and [Springer and Veldkamp 2000]. Twisted composition algebras are treated in Chapter VIII, §36 of the former and Chapter 4 of the latter.

3-1. ***Twisted composition algebras.*** A twisted composition algebra over F is a quadruple (E, C, Q, β), where:

- E is an étale cubic F-algebra.
- C is a free E-module equipped with a nondegenerate quadratic form Q, with associated symmetric bilinear form $b_Q(v_1, v_2) = Q(v_1+v_2)-Q(v_1)-Q(v_2)$.
- $\beta : C \longrightarrow C$ is a quadratic map such that
$$\beta(av) = a^{\#} \cdot \beta(v) \quad \text{and} \quad Q(\beta(v)) = Q(v)^{\#}$$
for every $a \in E$ and $v \in C$.
- If we set
$$N_C(v) := b_Q(v, \beta(v)),$$
then $N_C(v) \in F$ for every $v \in C$.

For a fixed E, we shall call (C, Q, β) an E-twisted composition algebra (over F), and the cubic form N_C the norm form of C. Frequently, for ease of notation, we shall simply denote this triple by C, suppressing the mention of Q and β.

3-2. ***Morphisms.*** An F-morphism of twisted composition algebras (E, C, Q, β) and (E', C', Q', β') is a pair $(\phi, \sigma) \in \mathrm{Hom}_F(C, C') \times \mathrm{Hom}_F(E, E')$ such that
$$\phi(av) = \sigma(a) \cdot \phi(v)$$
for $v \in C$ and $a \in E$, and
$$\phi \circ \beta = \beta' \circ \phi \quad \text{and} \quad \sigma \circ Q = Q' \circ \phi.$$
In particular, we have the automorphism group $\mathrm{Aut}_F(E, C, Q, \beta)$. The second projection gives a natural homomorphism
$$\mathrm{Aut}_F(E, C, Q, \beta) \to S_E.$$
The kernel of this map is the subgroup $\mathrm{Aut}_E(C, Q, \beta)$ consisting of those ϕ which are E-linear; we shall call these E-morphisms.

3-3. **$\mathrm{Aut}_F(E, C)$*-action and isomorphism classes.*** Let us fix an E-vector space C and let $\mathrm{Aut}_E(C)$ be the automorphism group of C as an E-vector space. Let
$$\mathrm{Aut}_F(E, C) = \{(g, \sigma) \in \mathrm{Aut}_F(C) \times \mathrm{Aut}_F(E) : g \circ \lambda = \sigma(\lambda) \cdot g \text{ for all } \lambda \in E\}.$$
This is the group of E-sesquilinear automorphisms of C. The second projection induces a short exact sequence
$$1 \longrightarrow \mathrm{Aut}_E(C) \longrightarrow \mathrm{Aut}_F(E, C) \longrightarrow S_E \longrightarrow 1.$$

This short exact sequence is split. Indeed, the choice of an E-basis for C gives a splitting, with S_E acting on the coordinates with respect to the basis.

Now if (C, Q, β) is an E-twisted composition algebra, then for any $(g, \sigma) \in \mathrm{Aut}_F(E, C)$, the triple

$$(C', Q', \beta') = (C, \sigma \circ Q \circ g^{-1}, g \circ \beta \circ g^{-1})$$

is also an E-twisted composition algebra. The norm forms are related by

$$N_{C'} = N_C \circ g^{-1}.$$

Moreover, we have

$$(g, \sigma) \in \mathrm{Hom}_F((E, C, Q, \beta), (E, C', Q', \beta')).$$

Thus the map $(Q, \beta) \mapsto (Q', \beta')$ defines an action of $\mathrm{Aut}_F(E, C)$ on the set of pairs (Q, β) which define an E-twisted composition algebra structure on C. The orbits of such pairs under $\mathrm{Aut}_F(E, C)$ are precisely the F-isomorphism classes of E-twisted composition algebras of a given E-dimension $\dim_E C$, and the stabilizer of a given pair (Q, β) is precisely the automorphism group $\mathrm{Aut}_F(E, C, Q, \beta)$. Similarly, the set of orbits under $\mathrm{Aut}_E(C)$ is the set of E-isomorphism classes of such E-twisted composition algebras, and the stabilizer of a particular (Q, β) is $\mathrm{Aut}_E(C, Q, \beta)$.

3-4. ***Dimension-2 case.*** It is known, by Corollary 36.4 in [Knus et al. 1998], that for any E-twisted composition algebra (C, Q, β), $\dim_E C = 1, 2, 4$ or 8. We shall only be interested in the case when $\dim_E C = 2$.

We give an example that will feature prominently in this paper. We set $C_E = E \oplus E$, and define Q and β by

$$Q(x, y) = x \cdot y \quad \text{and} \quad \beta(x, y) = (y^\#, x^\#)$$

for every $(x, y) \in E \oplus E$. It is easy to check that this defines an E-twisted composition algebra over F, with norm form

$$N_C(x, y) = N_E(x) + N_E(y).$$

The group of automorphisms of this E-twisted composition algebra is easy to describe. Let E^1 be the set of elements e in E such that $N(e) = e \cdot e^\# = 1$. For every element $e \in E^1$, we have an E-automorphism i_e defined by $i_e(x, y) = (ex, e^\# y)$. We also have an E-automorphism w defined by $w(x, y) = (y, x)$. The group of E-automorphisms is

$$\mathrm{Aut}_E(C_E, Q, \beta) = E^1 \rtimes \mathbb{Z}/2\mathbb{Z}$$

and the group of F-automorphisms is

$$\mathrm{Aut}_F(C_E, Q, \beta) = (E^1 \rtimes \mathbb{Z}/2\mathbb{Z}) \rtimes S_E = E^1 \rtimes (\mathbb{Z}/2\mathbb{Z} \times S_E).$$

If $E = F \times F \times F$, we denote the corresponding twisted composition algebra by $C_0 = (C_0, Q_0, \beta_0)$ and refer to it as the split twisted composition algebra. In this case, E^1 consists of (t_1, t_2, t_3) such that $t_1 t_2 t_3 = 1$, so that

$$\mathrm{Aut}_E(C_0, Q_0, \beta_0) \cong \mathbb{G}_m^2 \rtimes \mathbb{Z}/2\mathbb{Z}.$$

Observe that there is a natural splitting

$$S_3 \times \mathbb{Z}/2\mathbb{Z} \longrightarrow \mathrm{Aut}_F(C_0, Q_0, \beta_0). \tag{3.1}$$

3-5. ***Identities.*** It follows by [Knus et al. 1998, Proposition 36.3] that if (E, C, Q, β) is a twisted composition algebra over F, then $C \otimes_F \overline{F}$ is isomorphic to $C_0 \otimes_F \overline{F}$. This fact is useful for verifying polynomial identities in C. Indeed, any polynomial identity in C may be verified over $\overline{F}$ and thus just needs to be checked in C_0. In the following lemma, we list some useful identities which may be checked in this manner.

Lemma 3.2. *Let (E, C, Q, β) be a twisted composition algebra over F. Then*

$$\beta^2(v) = N_C(v)v - Q(v)\beta(v) \tag{3.3}$$

and

$$\beta(xv + y\beta(v)) = (y^\# N_C(v) - (-Q(v)x) \times y) \cdot v + (x^\# - Q(v)y^\#) \cdot \beta(v) \tag{3.4}$$

for any $v \in C$ and $x, y, \in E$.

It follows from (3.3) that Q is in fact determined by β in a twisted composition algebra. The proof of these identities can be found in [Springer and Veldkamp 2000, Lemmas 4.1.3 and 4.2.7]. We note that (3.4) looks slightly different from its counterpart there (Lemma 4.2.7), but the two are equivalent by the identity (2.1).

3-6. ***Reduced basis.*** If $\dim_E C = 2$, we call an E-basis of C of the form $\{v, \beta(v)\}$ a reduced basis of C. We note:

Proposition 3.5. *Let (C, Q, β) be an E-twisted composition algebra.*

(i) *For $v \in C$, let*

$$\Delta_C(v) = N_C(v)^2 - 4 \cdot N_E(Q(v)) \in F.$$

Then $\{v, \beta(v)\}$ is an E-basis of C if and only if $\Delta_C(v) \neq 0$.

(ii) *The degree-6 homogeneous polynomial Δ_C factors over F as*

$$\Delta_C = a \cdot P^2,$$

with $a \in F^\times$ and P an absolutely irreducible homogeneous polynomial of degree 3 *over F. The square class of a is uniquely determined, and for any $g \in \mathrm{Aut}_E(C, Q, \beta)$,*

$$P(gv) = \begin{cases} P(v) & \text{if } g \in \mathrm{Aut}_E(C, Q, \beta)^0, \\ -P(v) & \text{if } g \notin \mathrm{Aut}_E(C, Q, \beta)^0. \end{cases} \tag{3.6}$$

(iii) *The algebra (C, Q, β) has a reduced basis.*

(iv) *Let $\{v', \beta(v')\}$ be another reduced basis of C. Let $g \in \mathrm{Aut}_E(C)$ be such that $g(v) = v'$ and $g(\beta(v)) = \beta(v')$. Then* $\det(g) \in F^\times$.

Proof. (i) The set $\{v, \beta(v)\}$ is a basis if and only if the matrix of the symmetric bilinear form b_Q with respect to $\{v, \beta(v)\}$ has determinant in $E^\times$. Since

$$b_Q(v, v) = 2Q(v), \quad b_Q(\beta(v), \beta(v)) = 2Q(v)^\# \quad \text{and} \quad b_Q(v, \beta(v)) = N_C(v),$$

it follows that the determinant is $-\Delta_C(v)$.

(ii) We first work over $\overline{F}$, in which case we may assume that $C = E^2$, with $E = \overline{F}^3$, $Q(x, y) = xy$ and $\beta(x, y) = (y^\#, x^\#)$. Then $N_C(x, y) = N_E(x) + N_E(y)$. So

$$\Delta_C(x, y) = (N_E(x) + N_E(y))^2 - 4N_E(x)N_E(y) = (N_E(x) - N_E(y))^2.$$

The cubic polynomial $P_0(x, y) = N_E(x) - N_E(y) = x_1x_2x_3 - y_1y_2y_3$ (with $x = (x_1, x_2, x_3) \in \overline{F}^3$) is easily seen to be irreducible over $\overline{F}$.

To descend back to F, we note that for any $\sigma \in \mathrm{Gal}(\overline{F}/F)$, $\sigma(P_0) = \pm P_0$ by unique factorization of polynomials over $\overline{F}$. Thus there is a quadratic character χ_K of $\mathrm{Gal}(\overline{F}/F)$ such that $\sigma(P_0) = \chi_K(\sigma) \cdot P_0$. If K is the quadratic étale F-algebra associated to χ_K, represented by $a \in F^\times$, then we see that $P = \sqrt{a}^{-1} \cdot P_0$ is defined over F and $\Delta_C = a \cdot P^2$.

It is clear that the square class of a is uniquely determined. Equation (3.6) can be checked over $\overline{F}$; we leave it to the reader.

(iii) Since F has more than 3 elements (as we assumed that char$(F) \neq 2$ or 3), there exists $v \in C$ such that $P(v) \neq 0$. Hence $\Delta_C(v) \neq 0$ by (ii) and $\{v, \beta(v)\}$ is a reduced basis by (i).

(iv) If $v' = xv + y\beta(v)$, then $\beta(v')$ is given by (3.4). So the transition matrix between the bases $\{v, \beta(v)\}$ and $\{v', \beta(v')\}$ is given by

$$g = \begin{pmatrix} x & y^\# N_C(v) - (-Q(v)x) \times y \\ y & x^\# - Q(v)y^\# \end{pmatrix}.$$

Hence

$$\begin{aligned} \det(g) &= N_E(x) - N_E(y)N_C(v) + (-Q(v)xy^\# + ((-Q(v)x) \times y)y) \\ &= N_E(x) - N_E(y)N_C(v) - \mathrm{Tr}_E(Q(v)xy^\#) \in F \end{aligned}$$

where the second equality follows by applying (2.1). □

We note that Proposition 3.5(i) and (iii) are contained in [Springer and Veldkamp 2000, Lemma 4.2.12], but (ii) seems to be new; at least we are not able to find it in [Springer and Veldkamp 2000] or [Knus et al. 1998]. The results of the proposition will be used later in the paper.

3-7. ***The quadratic algebra K_C.*** An immediate consequence of the proposition is that to every twisted composition algebra (E, C, Q, β) with $\dim_E C = 2$, we can associate an étale quadratic algebra K_C which is given by the square-class of $\Delta_C(v) \in F^\times$ as in the proof of Proposition 3.5(ii). Thus we have a map

$$\{\text{twisted composition } F\text{-algebras with } E\text{-rank } 2\} \longrightarrow \{\text{étale quadratic } F\text{-algebras}\}. \tag{3.7}$$

For example, if C_E is the twisted composition algebra introduced in Section 3-4, then

$$\Delta_C(x, y) = (N_E(x) - N_E(y))^2$$

and the quadratic algebra associated to C_E is the split algebra $F \times F$.

3-8. ***Cohomological description.*** We come now to the classification of twisted composition algebras C of rank 2 over E. Since every such C is isomorphic to C_0 over $\overline{F}$, the set of isomorphism classes of twisted composition algebras over F is classified by the pointed cohomology set

$$H^1(F, \mathrm{Aut}_F(F^3, C_0, Q_0, \beta_0)).$$

We have seen that $\mathrm{Aut}_F(F^3, C_0, Q_0, \beta_0) \cong \mathbb{G}_m^2 \rtimes (\mathbb{Z}/2\mathbb{Z} \times S_3)$, and so there is a natural map

$$H^1(F, \mathrm{Aut}_F(F^3, C_0, Q_0, \beta_0)) \longrightarrow H^1(F, \mathbb{Z}/2\mathbb{Z}) \times H^1(F, S_3). \tag{3.8}$$

Composing this with the first or second projections, we obtain natural maps

$$H^1(F, \mathrm{Aut}_F(F^3, C_0, Q_0, \beta_0)) \longrightarrow H^1(F, \mathbb{Z}/2\mathbb{Z}) = \{\text{étale quadratic } F\text{-algebras}\} \tag{3.9}$$

and

$$H^1(F, \mathrm{Aut}_F(F^3, C_0, Q_0, \beta_0)) \longrightarrow H^1(F, S_3). \tag{3.10}$$

All these projection maps are surjective, because of the natural splitting in (3.1). Indeed, (3.1) endows each fiber of the maps in (3.8), (3.9) and (3.10) with a distinguished point. We shall see in a moment that the map in (3.9) is the map defined in (3.7).

For an étale cubic F-algebra E with associated cohomology class $[E] \in H^1(F, S_3)$, the fiber of (3.10) over $[E]$ is precisely the set of F-isomorphism classes of E-twisted composition algebras. Moreover, a Galois descent argument shows that the distinguished point in this fiber furnished by the splitting (3.1) is none other than the E-twisted composition algebra C_E constructed in Section 3-4.

Using C_E as the base point, the fiber in question is identified naturally with the set $H^1(F, \mathrm{Aut}_E(C_E, Q, \beta))$ modulo the natural action of $S_E(F)$ (by conjugation). The cohomology set $H^1(F, \mathrm{Aut}_E(C_E, Q, \beta))$ classifies the E-isomorphism classes of E-twisted composition algebras C over F, and the action of $S_E(F)$ is given by

$$\sigma : (C, Q, \beta) \mapsto (C \otimes_{E,\sigma} E, \sigma \circ Q, \beta)$$

for $\sigma \in S_E(F)$.

Lemma 3.11. *The maps defined by* (3.7) *and* (3.9) *are the same.*

Proof. We fix the cubic algebra E and let $C_E = (E^2, Q, \beta)$ be the distinguished E-twisted composition algebra introduced in Section 3-4. Let $\Delta_C = P^2$ be the homogeneous polynomials as given in Proposition 3.5(ii).

Any E-twisted composition algebra C' is given by a pair of tensors (Q', β') on E^2, and there is an element $g \in \mathrm{GL}_2(E \otimes_F \overline{F})$ such that $g \cdot (Q, \beta) = (Q', \beta')$. A 1-cocycle associated to (Q', β') is given by

$$a_\sigma = g^{-1}\sigma(g) \in \mathrm{Aut}_{\overline{F}^3}(E^2, Q, \beta) \qquad \text{for } \sigma \in \mathrm{Gal}(\overline{F}/F).$$

The corresponding $\Delta_{C'}$ is related to Δ_C by

$$\Delta_{C'}(v) = \Delta_C(g^{-1}v).$$

Now, the quadratic algebra associated to C' by (3.9) corresponds to the quadratic character

$$\chi : \sigma \mapsto [a_\sigma] \in \pi_0(\mathrm{Aut}_{\overline{F}^3}(E^2, Q, \beta)) = \mathbb{Z}/2\mathbb{Z}$$

of $\mathrm{Gal}(\overline{F}/F)$. By (3.6), we thus have

$$P(a_\sigma^{-1}v) = \chi(\sigma) \cdot P(v)$$

for any $v \in (E \otimes_F \overline{F})^2$.

On the other hand, the quadratic algebra associated to C' by (3.7) is defined by $\sqrt{\Delta_{C'}(v)}$ for any $v \in E^2$ such that $\Delta_{C'}(v) \neq 0$. Since

$$\sqrt{\Delta_{C'}(v)} = \sqrt{\Delta_C(g^{-1}v)} = P(g^{-1}v),$$

we need to show that

$$\sigma(P(g^{-1}v)) = \chi(\sigma) P(g^{-1}v).$$

But we have

$$\sigma(P(g^{-1}v)) = P(\sigma(g)^{-1}v) = P(a_\sigma^{-1}g^{-1}v) = \chi(\sigma)\cdot P(g^{-1}v),$$

as desired. □

3-9. ***Tits construction.*** Given an element

$$([E],[K]) \in H^1(F,S_3)\times H^1(F,\mathbb{Z}/2\mathbb{Z}),$$

we describe the composition algebras in the fiber of (3.8) over $([E],[K])$. Note that by (3.1), we have a distinguished point in this fiber. Now, we have:

Proposition 3.12. *If C is an E-twisted composition algebra, with associated étale quadratic algebra K, then we may identify C with $E\otimes_F K$, such that*

$$Q(x) = e\cdot N_{E\otimes_F K/E}(x) \quad \textit{for some } e\in E^\times$$

and

$$\beta(x) = \bar{x}^{\#}\cdot e^{-1}\cdot\bar{\nu} \quad \textit{for some } \nu\in K$$

where $x\mapsto\bar{x}$ is induced by the nontrivial automorphism of K over F. Moreover, we have:

$$N_{E/F}(e) = N_{K/F}(\nu).$$

The distinguished point in the fiber of (3.8) *over* $([E],[K])$ *corresponds to taking* $(e,\nu)=(1,1)$.

Proof. The proof of Proposition 3.5(i) shows that the quadratic discriminant algebra associated to Q is $E\otimes_F K$. Hence, we may identify C with $E\otimes_F K$ with Q given by $e\cdot N_{E\otimes_F K/E}$ for some $e\in E^\times$. On the other hand, we claim that for $x\in E\otimes_F K$ and $x_0\in C$, one has

$$\beta(x\cdot x_0) = \bar{x}^{\#}\cdot\beta(x_0).$$

Indeed, one can check this by going to $\overline{F}$, where one is reduced to checking this identity in the split algebra C_0, which is straightforward. This shows that β is determined by $\beta(1)=\bar{\nu}\cdot e^{-1}$ for some $\nu\in E\otimes_F K$. However, the identity

$$Q(1)^{\#} = Q(\beta(1))$$

implies that

$$\nu\cdot\bar{\nu} = N_{E\otimes_F K/E}(\nu) = N_{E/F}(e)\in F.$$

The requirement that $N(x)\in F$ for all $x\in E\otimes_F K$ implies that

$$\mathrm{Tr}_{E\otimes_F K/E}(\bar{\nu}\cdot N_{E\otimes_F K/K}(x))\in F.$$

In particular, taking $x = 1$ and then a trace-zero element $\delta \in K$ one obtains, respectively,

$$\nu + \bar{\nu} \in F \quad \text{and} \quad \nu\delta + \overline{\nu\delta} \in F.$$

All these conditions imply that $\nu \in K$.

Finally, it is easy to see by Galois descent that the distinguished point in the fiber over $([E], [K])$ corresponds to $(e, \nu) = (1, 1)$. □

The description of twisted composition algebras given in the above proposition is sometimes referred to as a Tits construction (though usually this terminology is reserved for the Jordan algebra associated to the above twisted composition algebra by Springer's construction, which is the subject matter of the next section).

3-10. ***Automorphism group.*** Using Proposition 3.12, it is not difficult to determine the automorphism group of any twisted composition algebra C. Indeed, if $C \cong E \otimes_F K$ as in the proposition, then the special orthogonal group

$$SO(C, Q) = \{\lambda \in E \otimes_F K : N_{E\otimes K/E}(\lambda) = 1\}$$

acts $E \otimes K$-linearly on C by multiplication and preserves Q. An element $\lambda \in SO(C, Q)$ preserves β if and only if

$$\bar{\lambda}^{\#} = \lambda.$$

But $\lambda^{\#} = \lambda^{-1}$ since $N_{E\otimes K/E}(\lambda) = \lambda \cdot \lambda^{\#} = 1$. So

$$\mathrm{Aut}_E(C, Q, \beta) \cap SO(C, Q) = \{\lambda \in L = E \otimes K : N_{L/E}(\lambda) = 1 = N_{L/K}(\lambda)\} = T_{E,K},$$

which is a 2-dimensional torus. Since we know the automorphism group of the split twisted composition algebra (C_0, Q_0, β_0), we see that

$$\mathrm{Aut}_E(C, Q, \beta)^0 = T_{E,K}$$

and $\mathrm{Aut}_E(C, Q, \beta)$ sits in short exact sequences of algebraic groups as in (iii) of Theorem 1.1.

3-11. ***Cohomology of $T_{E,K}$.*** Using Proposition 3.12 and the above description of $\mathrm{Aut}_E(C, Q, \beta)^0$, we can describe the fiber of the natural map

$$H^1(F, \mathrm{Aut}_F(F^3, C_0, Q_0, \beta_0)) \longrightarrow H^1(F, \mathbb{Z}/2\mathbb{Z}) \times H^1(F, S_3)$$

over the element $([K], [E]) \in H^1(F, \mathbb{Z}/2\mathbb{Z}) \times H^1(F, S_3)$. Indeed, this fiber is equal to

$$H^1(F, T_{E,K}) \text{ modulo the action of } S_E(F) \times \mathbb{Z}/2\mathbb{Z}.$$

The cohomology group $H^1(T_{E,K})$ classifies twisted composition algebras with fixed E and K, up to $E \otimes_F K$-linear isomorphism. With $L = E \otimes_F K$, one has a short exact sequence of algebraic tori

$$1 \longrightarrow T_{E,K} \longrightarrow L^\times \xrightarrow{N_{L/E} \times N_{L/K}} (E^\times \times K^\times)^0 \longrightarrow 1,$$

where

$$(E \times K)^0 = \{(e, \nu) \in E^\times \times K^\times : N_{E/F}(e) = N_{K/F}(\nu)\}.$$

The associated long exact sequence gives

$$H^1(F, T_{E,K}) \cong (E^\times \times K^\times)^0 / \operatorname{Im} L^\times. \tag{3.13}$$

This isomorphism is quite evident in the context of Proposition 3.12. Indeed, Proposition 3.12 tells us that any twisted composition algebra C with invariants (E, K) is given by an element $(e, \nu) \in (E^\times \times K^\times)^0$. Any L-linear map from C to another twisted composition algebra C' with associated pair (e', ν') is given by multiplication by an element $a \in L^\times$, and this map is an isomorphism of twisted composition algebras if and only if

$$(e, \nu) = (e' \cdot N_{L/E}(a), \nu' \cdot N_{L/K}(a)).$$

This is precisely what (3.13) expresses.

4. Springer's construction

We can now relate twisted composition algebras to Freudenthal–Jordan algebras. This construction is due to Springer. Our exposition follows [Knus et al. 1998, §38A, p. 522].

4-1. ***Freudenthal–Jordan algebra of dimension* 9.** A Freudenthal–Jordan algebra J of dimension 9 over F is a Jordan algebra which is isomorphic over $\overline{F}$ to the Jordan algebra J_0 associated to the associative algebra $M_3(F)$ of 3×3-matrices, with Jordan product

$$a \circ b = \tfrac{1}{2} \cdot (ab + ba).$$

An element $a \in J$ satisfies a characteristic polynomial

$$X^3 - T_J(a)X^2 + S_J(a)X - N_J(a) \in F[X].$$

The maps T_J and N_J are called the trace and norm maps of J respectively. The element

$$a^\# = a^2 - T_J(a)a + S_J(a)$$

is called the adjoint of a. It satisfies $a \cdot a^\# = N_J(a)$. The cross product of two elements $a, b \in J$ is defined by

$$a \times b = (a+b)^\# - a^\# - b^\#.$$

4-2. ***Cohomological description.*** The automorphism group of J_0 is $\mathrm{PGL}_3 \rtimes \mathbb{Z}/2\mathbb{Z}$, with $g \in \mathrm{PGL}_3$ acting by conjugation and the nontrivial element of $\mathbb{Z}/2\mathbb{Z}$ acting by the transpose $a \mapsto a^t$. Thus, the isomorphism classes of Freudenthal–Jordan algebras of dimension 9 are parametrized by the pointed set $H^1(F, \mathrm{PGL}_3 \rtimes \mathbb{Z}/2\mathbb{Z})$, and there is an exact sequence of pointed sets

$$\begin{tikzcd}
H^1(F,\mathrm{PGL}_3) \arrow[r, "f"] & H^1(F,\mathrm{PGL}_3 \rtimes \mathbb{Z}/2\mathbb{Z}) \arrow[r, "\pi"] & H^1(F,\mathbb{Z}/2\mathbb{Z}) \arrow[d, equal] \\
& & \{\text{étale quadratic } F\text{-algebras}\}.
\end{tikzcd}$$

The map π is surjective and the fiber of π over the split quadratic algebra F^2 is the image of f. By [Serre 2002, Proposition 39(ii) and Corollary 1, p. 52], the image of f is $H^1(F, \mathrm{PGL}_3)$ modulo a natural action of $\mathbb{Z}/2\mathbb{Z}$. Now the set $H^1(F, \mathrm{PGL}_3)$ parametrizes the set of central simple F-algebras B of degree 3, and the $\mathbb{Z}/2\mathbb{Z}$ action in question is $B \mapsto B^{\mathrm{op}}$. Then the map f sends B to the associated Jordan algebra.

In general, for any étale quadratic F-algebra K, an element in the fiber of π over $[K] \in H^1(F, \mathbb{Z}/2\mathbb{Z})$ is the Jordan algebra $J_3(K)$ of 3×3-Hermitian matrices with entries in K. The automorphism group of $J_3(K)$ is an adjoint group $PGU_3^K \rtimes \mathbb{Z}/2\mathbb{Z}$. Using $J_3(K)$ as the base point, the fiber of π over $[K]$ can then be identified with $H^1(F, PGU_3^K)$ modulo the action of $\mathbb{Z}/2\mathbb{Z}$ (by [Serre 2002, pp. 50 and 52]). By [Knus et al. 1998, p. 400], $H^1(F, PGU_3^K)$ has an interpretation as the set of isomorphism classes of pairs (B_K, τ) where

- B_K is a central simple K-algebra of degree 3,
- τ is an involution of the second kind on B_K.

Moreover, the action of the nontrivial element $\tau_K \in \mathrm{Aut}(K/F) = \mathbb{Z}/2\mathbb{Z}$ is via the Galois twisting action $B \mapsto B \otimes_{K,\tau_K} K$, so that

$$H^1(F, PGU_3^K)/\mathbb{Z}/2\mathbb{Z} \longleftrightarrow \{F\text{-isomorphism classes of } (B_K, \tau)\}.$$

Then the map f sends (B_K, τ) to the Jordan algebra B_K^τ of τ-symmetric elements in B_K.

If J is a Freudenthal–Jordan algebra of dimension 9, we will write K_J for the étale quadratic algebra corresponding to $\pi(J)$.

4-3. ***Relation with twisted composition algebras.*** Fix an étale cubic F-algebra E and a Freudenthal–Jordan algebra J. Suppose we have an algebra embedding

$$i : E \hookrightarrow J.$$

Then, with respect to the trace form T_J, we have an orthogonal decomposition

$$J = i(E) \oplus C,$$

where $C = i(E)^{\perp}$. We shall identify E with its image under i. Then for $e \in E$ and $v \in C$, one can check that $e \times v \in C$. Thus, setting

$$e \circ v := -e \times v$$

equips C with the structure of an E-vector space. Moreover, writing

$$v^{\#} = (-Q(v), \beta(v)) \in E \oplus C = J$$

for $Q(v) \in E$ and $\beta(v) \in V$, we obtain a quadratic form Q on C and a quadratic map β on C. Then, by Theorem 38.6 in [Knus et al. 1998], the triple (C, Q, β) is an E-twisted composition algebra over F.

Conversely, given an E-twisted composition algebra C over F, the same theorem says that the space $E \oplus C$ can be given the structure of a Freuthendal–Jordan algebra over F. In particular, we have described the bijective correspondence between the objects in (b) and (c) of the main theorem:

$$\{E\text{-twisted composition algebras over } F\}$$
$$\updownarrow$$
$$\{i : E \longrightarrow J \text{ with } J \text{ Freudenthal–Jordan of dimension } 9\}.$$

It is also clear that under this identification, one has

$$\mathrm{Aut}_F(i : E \to J) = \mathrm{Aut}_F(i(E)^{\perp}).$$

4-4. ***Example.*** Let K be an étale quadratic F-algebra and consider the Jordan algebra $J_3(K)$ of 3×3 Hermitian matrices with entries in K. Let $E = F \times F \times F$ be the subalgebra of $J_3(K)$ consisting of diagonal matrices. Then C consists of matrices

$$v = \begin{pmatrix} 0 & \bar{z}_3 & z_2 \\ z_3 & 0 & \bar{z}_1 \\ \bar{z}_2 & z_1 & 0 \end{pmatrix}.$$

Thus $C = K \times K \times K$, and one checks that

$$Q(z_1, z_2, z_3) = (z_1\bar{z}_1, z_2\bar{z}_2, z_3\bar{z}_3) \quad \text{and} \quad \beta(z_1, z_2, z_3) = (\bar{z}_2\bar{z}_3, \bar{z}_3\bar{z}_1, \bar{z}_1\bar{z}_2).$$

The algebra C is the distinguished point in the fiber of $([F^3],[K])$, in the sense of Proposition 3.12. The automorphism group of C is given by

$$\mathrm{Aut}_F(C,Q,\beta) = (K^1 \times K^1 \times K^1)^0 \rtimes (\mathbb{Z}/2\mathbb{Z} \times S_3),$$

where K^1 denotes the torus of norm-1 elements in K and $(K^1 \times K^1 \times K^1)^0$ denotes the subgroup of triples (t_1,t_2,t_3) such that $t_1t_2t_3 = 1$.

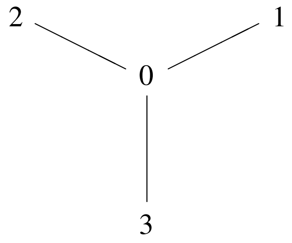
4-5. ***The quadratic algebra associated to $i : E \to J$.*** If an E-twisted composition algebra C corresponds to a conjugacy class of embeddings $i : E \longrightarrow J$, then we may ask how the quadratic algebra K_C associated to C can be described in terms of $i : E \longrightarrow J$. In this case, $C = E^{\perp}$ is an E-twisted composition algebra, and so $C = E \otimes K_C$ for a quadratic algebra K_C as in Proposition 3.12. On the other hand, we know that J is associated to a pair (B_{K_J},τ), where B_{K_J} is a central simple algebra over an étale quadratic F-algebra K_J and τ is an involution of the second kind. Now, Examples (5) and (6) on page 527 in [Knus et al. 1998] show that

$$[K_C]\cdot[K_E]\cdot[K_J] = 1 \in H^1(F,\mathbb{Z}/2\mathbb{Z}) = F^{\times}/F^{\times 2}.$$

5. Quasisplit groups of type D_4

In this section, we shall introduce the E-twisted Bhargava's cube by way of the quasisplit groups of type D_4.

5-1. ***Root system.*** Let Ψ be a root system of type D_4 and $\Pi = \{\alpha_0,\alpha_1,\alpha_2,\alpha_3\}$ a set of simple roots such that the corresponding Dynkin diagram is



The group of diagram automorphisms $\mathrm{Aut}(\Pi)$ is identified with S_3, the group of permutations of $\{1,2,3\}$. We denote the highest root by $\beta_0 = \alpha_1+\alpha_2+\alpha_3+2\alpha_0$.

5-2. ***Quasisplit groups of type D_4.*** Let G be a split, simply connected Chevalley group of type D_4. We fix a maximal torus T contained in a Borel subgroup B defined over F. The group G is then generated by root groups $U_\alpha \cong \mathbb{G}_a$, where $\alpha \in \Psi$. Steinberg showed that one can pick the isomorphisms $x_\alpha : \mathbb{G}_a \to U_\alpha$ such that

$$[x_\alpha(u), x_{\alpha'}(u')] = x_{\alpha+\alpha'}(\pm uu')$$

whenever $\alpha+\alpha'$ is a root. Fixing such a system of isomorphisms fixes an épinglage (or pinning) for G. As Kac noted, a choice of signs corresponds to an orientation of the Dynkin diagram. Since one can pick an orientation of the Dynkin diagram which is invariant under $\mathrm{Aut}(\Pi)$, the group of automorphisms of Π can be lifted to a group of automorphisms of G. Thus, we have a semidirect product

$$\widetilde{G} = G \rtimes \mathrm{Aut}(\Pi) = G \rtimes S_3,$$

where the action of S_3 permutes the root subgroups U_α and the isomorphisms x_α.

Since the outer automorphism group S_3 of G is also the automorphism group of the split étale cubic F-algebra F^3, we see that every cubic étale algebra E defines a simply connected quasisplit form G_E of G, whose outer automorphism group is the finite group scheme S_E. Thus,

$$\widetilde{G}_E = G_E \rtimes S_E$$

is a form of $\widetilde{G}$, and it comes equipped with a pair $B_E \supset T_E$, consisting of a Borel subgroup B_E containing a maximal torus T_E, both defined over F, as well as a Chevalley–Steinberg system of épinglage relative to this pair.

5-3. ***G_2 root system.*** The subgroup of G_E fixed pointwise by S_E is isomorphic to the split exceptional group of type G_2.

Observe that $B = G_2 \cap B_E$ is a Borel subgroup of G_2 and $T = T_E \cap G_2$ is a maximal split torus of G_2. Via the adjoint action of T on G_E, we obtain the root system Ψ_{G_2} of G_2, so that

$$\Psi_{G_2} = \Psi|_T.$$

We denote the short simple root of this G_2 root system by α and the long simple root by β. Then

$$\beta = \alpha_0|_T \quad \text{and} \quad \alpha = \alpha_1|_T = \alpha_2|_T = \alpha_3|_T.$$

Thus, the short root spaces have dimension 3, whereas the long root spaces have dimension 1. For each root $\gamma \in \Psi_{G_2}$, the associated root subgroup U_γ is defined over F and the Chevalley–Steinberg system of épinglage gives isomorphisms

$$U_\gamma \cong \begin{cases} \mathrm{Res}_{E/F}\,\mathbb{G}_a & \text{if } \gamma \text{ is short,} \\ \mathbb{G}_a & \text{if } \gamma \text{ is long.} \end{cases}$$

5-4. ***The parabolic subgroup P_E.*** The G_2 root system gives rise to two parabolic subgroups of G_E. One of these is a maximal parabolic $P_E = M_E N_E$ known as the Heisenberg parabolic. Its unipotent radical N_E is a Heisenberg group with center $Z_E = U_{\beta_0}$; see Section 2 in [Gan et al. 2002]. Moreover,

$$N_E/Z_E = U_\beta \times U_{\beta+\alpha} \times U_{\beta+2\alpha} \times U_{\beta+3\alpha} \cong \mathbb{G}_a \times \mathrm{Res}_{E/F}\,\mathbb{G}_a \times \mathrm{Res}_{E/F}\,\mathbb{G}_a \times \mathbb{G}_a$$

and

$$\widetilde{M}_E = M_E \rtimes S_E \cong \mathrm{GL}_2(E)^0 \rtimes S_E,$$

where

$$\mathrm{GL}_2(E)^0 = \{g \in \mathrm{GL}_2(E) : \det(g) \in F^\times\}.$$

We shall fix the isomorphism $M_E \rtimes S_E \cong \mathrm{GL}_2(E)^0 \rtimes S_E$ as follows. We first consider the case when $E = F^3$ is split. The pinning gives us an identification

$$M_{\mathrm{der}}(F) \cong \mathrm{SL}_2(F)^3$$

such that

$$\alpha_1^\vee(t) = \left(\begin{pmatrix} t & \\ & t^{-1} \end{pmatrix}, 1, 1 \right) \in \mathrm{SL}_2(F)^3,$$

while $\alpha_2^\vee(t)$ and $\alpha_3^\vee(t)$ are defined analogously by cyclically permuting the entries of $\alpha_1^\vee(t)$. We extend this identification to $M(F)$ by

$$\alpha_0^\vee(t) = \left(\begin{pmatrix} 1 & \\ & t \end{pmatrix}, \begin{pmatrix} 1 & \\ & t \end{pmatrix}, \begin{pmatrix} 1 & \\ & t \end{pmatrix} \right) \in (\mathrm{GL}_2(F)^3)^0.$$

Note that, under the identification,

$$\beta_0^\vee(t) = \left(\begin{pmatrix} t & \\ & t \end{pmatrix}, \begin{pmatrix} t & \\ & t \end{pmatrix}, \begin{pmatrix} t & \\ & t \end{pmatrix} \right) \in (\mathrm{GL}_2(F)^3)^0.$$

Finally, since the pinning is invariant under the action of $\mathrm{Aut}(\Pi) \cong S_3$, it follows that

$$\widetilde{M}(F) \cong (\mathrm{GL}_2(F)^3)^0 \rtimes S_3,$$

where S_3 acts on $(\mathrm{GL}_2(F)^3)^0$ by permuting the components. For general E, one obtains the desired isomorphism by a Galois descent argument.

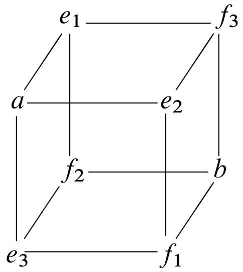
6. Bhargava's cube

In this section, we shall examine the split case, where the pinning for G gives a $\mathbb{Z}$-structure on N/Z; for more details see Section 4 in [Gan et al. 2002].

6-1. ***Bhargava's cube.*** Let V_2 be the standard representation of SL_2. Recall that we have identified M_{der} with SL_2^3 and M with $(\mathrm{GL}_2^3)^0$. Under this identification, the representation of M_{der} on N/Z is isomorphic to the representation of SL_2^3 on $V = V_2 \otimes V_2 \otimes V_2$. Since $\beta_0^\vee(t)$ acts on N/Z as multiplication by t, it follows that $(\mathrm{GL}_2^3)^0$ acts on V by the standard action *twisted* by $\det^{-1}$. The group $S_3 \cong \mathrm{Aut}(\Pi)$ acts on $V_2 \otimes V_2 \otimes V_2$ by permuting the three factors.

Since V is an absolutely irreducible SL_2^3-module, the isomorphism of N/Z and V is unique up to a nonzero scalar. Since $\beta_0^\vee(t)$ acts on N/Z as multiplication by t, the bijection between M-orbits on N/Z and M-orbits on V does not depend on the choice of the isomorphism. If we demand that the isomorphism preserves $\mathbb{Z}$-structures, i.e., that it gives an isomorphism of $(N/Z)(\mathbb{Z})$ and $\mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$, then it is unique up to a sign.

An element $v \in V(F)$ is represented by a cube



where $a, \dots, b \in F$ and the vertices correspond to the standard basis in $F^2 \otimes F^2 \otimes F^2$. More precisely, we fix this correspondence so that

$$\begin{pmatrix}1\\0\end{pmatrix} \otimes \begin{pmatrix}1\\0\end{pmatrix} \otimes \begin{pmatrix}1\\0\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}0\\1\end{pmatrix} \otimes \begin{pmatrix}0\\1\end{pmatrix} \otimes \begin{pmatrix}0\\1\end{pmatrix}$$

correspond to the vertices marked with letters a and b, respectively. We note that elementary matrices in $\mathrm{SL}_2(F)^3$ act on the space of cubes by the following three types of "row-column" operations on cubes:

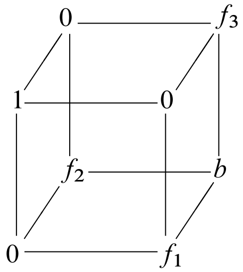
- add or subtract the front face from the rear face of the cube, and vice-versa;
- add or subtract the top face from the bottom face of the cube, and vice-versa;
- add or subtract the right face from the left face of the cube, and vice-versa.

The group $S_3 \cong \mathrm{Aut}(\Pi)$ acts as the group of symmetries of the cube fixing the two vertices marked a and b. We shall often write the cube as a quadruple

$$(a, e, f, b),$$

where $e = (e_1, e_2, e_3)$ and $f = (f_1, f_2, f_3) \in F^3$.

6-2. ***Reduced and distinguished cube.*** It is not hard to see that, using the action of $M(F)$, every cube can be transformed into a cube of the form $(1, 0, f, b)$:



We shall call such a cube a *reduced cube*. In particular, we call the cube $v_0 = (1,0,0,-1)$ the *distinguished cube*.

6-3. ***Stabilizer of distinguished cube.*** Let $\mathrm{Stab}_M(v_0)$ and $\mathrm{Stab}_{\widetilde{M}}(v_0)$ be the respective stabilizers in M and $\widetilde{M}$ of the distinguished cube $v_0 \in V$. Since $\mathrm{Aut}(\Pi)$ stabilizes v_0, the group $\mathrm{Stab}_{\widetilde{M}}(v_0)$ is a semidirect product of $\mathrm{Stab}_M(v_0)$ and $\mathrm{Aut}(\Pi)$. We shall now compute $\mathrm{Stab}_M(v_0)$. Let $g = (g_1, g_2, g_3) \in M(F)$, where

$$g_i = \begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix}.$$

Since

$$v_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and

$$g \cdot v_0 = \det(g)^{-1} \cdot \begin{pmatrix} a_1 \\ c_1 \end{pmatrix} \otimes \begin{pmatrix} a_2 \\ c_2 \end{pmatrix} \otimes \begin{pmatrix} a_3 \\ c_3 \end{pmatrix} - \det(g)^{-1} \cdot \begin{pmatrix} b_1 \\ d_1 \end{pmatrix} \otimes \begin{pmatrix} b_2 \\ d_2 \end{pmatrix} \otimes \begin{pmatrix} b_3 \\ d_3 \end{pmatrix},$$

$g \cdot v_0 = v_0$ if and only if eight equations hold. Six of these equations are homogeneous. They are

$$a_1c_2a_3 = b_1d_2b_3, \quad a_1c_2c_3 = b_1d_2d_3,$$

with the additional four obtained by cyclically permuting the indices. If we multiply the first equation by d_3, the second by b_3, and subtract them, then

$$0 = a_1c_2a_3d_3 - a_1c_2c_3b_3 = a_1c_2(a_3d_3 - c_3b_3).$$

Since $a_3d_3 - c_3b_3 \neq 0$, we have $a_1c_2 = 0$. A similar manipulation of these two equations gives $b_1d_2 = 0$. By permuting the indices, we have $a_ic_j = b_id_j = 0$ for all $i \neq j$. This implies that all the g_i are simultaneously diagonal or off-diagonal. Now it is easy to see that the remaining two equations imply that $\mathrm{Stab}_M(v_0)$ has two connected components, and the identity component consists of $g = (g_1, g_2, g_3)$ such that g_i are diagonal matrices, $a_id_i = 1$, and $a_1a_2a_3 = 1$. The other component of $\mathrm{Stab}_M(v_0)$ contains an element $w = (w_1, w_2, w_3)$ of order 2, where

$$w_i = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

We now have a complete description of $\mathrm{Stab}_M(v_0)$ (and of $\mathrm{Stab}_{\widetilde{M}}(v_0)$):

$$\mathrm{Stab}_M(v_0) \cong \{(a_1, a_2, a_3) \in \mathbb{G}_m^3 : a_1a_2a_3 = 1\} \rtimes \mathbb{Z}/2\mathbb{Z} \cong \mathbb{G}_m^2 \rtimes \mathbb{Z}/2\mathbb{Z}.$$

In particular, we have shown:

Proposition 6.1. *The stabilizer* $\mathrm{Stab}_{\widetilde{M}}(v_0)$ *in* $\widetilde{M}$ *of the distinguished cube* $v_0 = (1,0,0,-1)$ *is isomorphic to the group of* F*-automorphisms of the split twisted composition algebra* C_0. *Indeed, they give identical subgroups of* $(\mathrm{GL}_2(F)^3)^0 \rtimes S_3$, *where we fix the isomorphism* $M(F) \cong (\mathrm{GL}_2(F)^3)^0$ *as above.*

6-4. ***Three quadratic forms.*** One key observation in [Bhargava 2004a] is that one can slice the cube (given in the picture in Section 6-1) in three different ways, giving three pairs of matrices:

$$A_1 = \begin{pmatrix} a & e_2 \\ e_3 & f_1 \end{pmatrix}, \quad B_1 = \begin{pmatrix} e_1 & f_3 \\ f_2 & b \end{pmatrix},$$
$$A_2 = \begin{pmatrix} a & e_3 \\ e_1 & f_2 \end{pmatrix}, \quad B_2 = \begin{pmatrix} e_2 & f_1 \\ f_3 & b \end{pmatrix},$$
$$A_3 = \begin{pmatrix} a & e_1 \\ e_2 & f_3 \end{pmatrix}, \quad B_3 = \begin{pmatrix} e_3 & f_2 \\ f_1 & b \end{pmatrix}.$$

Note that the pairs (A_2, B_2) and (A_3, B_3) are obtained by rotating the pair (A_1, B_1) about the axis passing through a and b. For each pair (A_i, B_i), Bhargava defines a quadratic binary form by

$$Q_i = -\det(A_i x + B_i y).$$

Proposition 6.2. *Given a cube* v, *the three forms* Q_1, Q_2 *and* Q_3 *have the same discriminant* $\Delta = \Delta(v)$.

Proof. We may assume the cube is reduced. Now an easy computation show that the three forms are

$$\begin{cases} Q_1(x,y) = -f_1x^2 - bxy + f_2f_3y^2, \\ Q_2(x,y) = -f_2x^2 - bxy + f_3f_1y^2, \\ Q_3(x,y) = -f_3x^2 - bxy + f_1f_2y^2. \end{cases}$$

These forms have the same discriminant $\Delta = b^2 + 4f_1f_2f_3$. □

6-5. ***Quartic invariant.*** To every cube $v \in V$, the discriminant $\Delta(v)$ described in the previous proposition is a homogeneous quartic polynomial in v, which is invariant under the action of $\mathrm{SL}_2(F)^3$. This describes the quartic invariant of the prehomogeneous vector space V. An explicit computation gives the formula

$$\Delta = a^2b^2 - 2ab(e_1f_1 + e_2f_2 + e_3f_3) + e_1^2f_1^2 + e_2^2f_2^2 + e_3^2f_3^2 \\ + 4af_1f_2f_3 + 4be_1e_2e_3 - 2(e_1e_2f_1f_2 + e_2e_3f_2f_3 + e_3e_1f_3f_1).$$

If v is reduced, then this simplifies to $\Delta(v) = b^2 + 4f_1f_2f_3$. It is easy to check that for $g \in M$, one has

$$\Delta(g \cdot v) = \det(g)^2 \cdot \Delta(v).$$

Thus, we see that Δ gives a well-defined map

$$\Delta : \{\text{generic } \widetilde{M}(F)\text{-orbits on } V(F)\} \longrightarrow F^\times/F^{\times 2} = \{\text{étale quadratic } F\text{-algebras}\}.$$

7. E-twisted Bhargava cube

Now we can extend the discussion of the previous section to the case of general E, where $V_E = F \oplus E \oplus E \oplus F$ and $\widetilde{M}_E = \mathrm{GL}_2(E)^0 \rtimes S_E$, via a Galois descent using a cocycle in the class of

$$[E] \in H^1(F, \mathrm{Aut}(\Pi)) = H^1(F, S_3).$$

A cube is a quadruple $v = (a, e, f, b)$, where $e, f \in E$. As in the split case, we shall call cubes of the form $v = (1, 0, f, b)$ reduced, and the vector $v_{0,E} = (1, 0, 0, -1)$ the E-distinguished cube.

7-1. ***Quartic invariant.*** By Galois descent, we see that the basic polynomial invariant Δ_E is given by

$$\Delta_E(a, e, f, b) = a^2b^2 - 2ab\,\mathrm{Tr}_{E/F}(ef) + \mathrm{Tr}_{E/F}(e^2f^2) \\ + 4aN_{E/F}(f) + 4bN_{E/F}(e) - 2\,\mathrm{Tr}_{E/F}(e^\# f^\#).$$

If v is reduced, then this simplifies to

$$\Delta_E(1, 0, f, b) = b^2 + 4 \cdot N_{E/F}(f).$$

7-2. ***Group action.*** It is useful to note the action of certain elements of $\mathrm{GL}_2(E)^0$ on V_E. Specifically, $\sigma \in S_E$ acts by $\sigma(a, e, f, b) = (a, \sigma(e), \sigma(f), b)$. Moreover, the diagonal torus elements

$$t_{\alpha,\beta} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} \quad \text{with } \alpha\beta \in F^\times$$

act by

$$(a, e, f, b) \mapsto (\alpha^\# \beta^{-1} a,\ \alpha^\# \alpha^{-1} e,\ \beta^\# \beta^{-1} f,\ \beta^\# \alpha^{-1} b).$$

It is easy to check that

$$\Delta_E(t_{\alpha,\beta} \cdot v) = (\alpha\beta)^2 \cdot \Delta_E(v).$$

Since the actions of $\mathrm{SL}_2(E)$ and S_E preserve Δ_E, we see that

$$\Delta_E(g \cdot v) = (\det g)^2 \cdot \Delta_E(v),$$

so that Δ_E induces a map

$$\{\widetilde{M}_E\text{-orbits on } V_E\} \longrightarrow F^\times/F^{\times 2} = \{\text{étale quadratic algebras}\}.$$

In addition, the standard Weyl group element

$$w=\begin{pmatrix}0&1\\1&0\end{pmatrix}\in \mathrm{GL}_2(E)^0$$

acts by

$$w:(a,e,f,b)\mapsto(-b,-f,-e,-a).$$

7-3. ***Stabilizer of distinguished E-cube.*** We can readily determine the stabilizer of the E-distinguished cube . Namely, under the action described in Section 7-2, it is easy to see that the subgroup

$$E^1=\left\{\begin{pmatrix}\alpha&\\&\alpha^{-1}\end{pmatrix}:\alpha\in E^1\right\}\subset \mathrm{SL}_2(E)$$

fixes the E-distinguished cube $v_{0,E}$. So does the Weyl group element w. Thus, we see that

$$\mathrm{Stab}_{M_E}(v_{0,E})\cong E^1\rtimes\mathbb{Z}/2\mathbb{Z}\quad\text{and}\quad \mathrm{Stab}_{\widetilde{M}_E}(v_{0,E})=E^1\rtimes(\mathbb{Z}/2\mathbb{Z}\times S_E).$$

In particular, we have shown:

Proposition 7.1. *The stabilizer* $\mathrm{Stab}_{\widetilde{M}_E}(v_{0,E})$ *in* $\widetilde{M}_E$ *of the E-distinguished cube* $(1,0,0,-1)$ *is isomorphic to the group of F-automorphisms of the twisted composition algebra C_E introduced in* Section 3-4. *Indeed, they are identical as subgroups of* $\mathrm{GL}_2(E)^0\rtimes S_E(F)$ *under the fixed isomorphism* $M_E(F)\cong\mathrm{GL}_2(E)^0$.

8. Generic orbits

We come now to the main result of this paper: the determination of the generic $\widetilde{M}_E(F)$-orbits in $V_E(F)$.

8-1. ***A commutative diagram.*** We have the following commutative diagram

$$\begin{tikzcd} H^1(F,\mathrm{Stab}_{\widetilde{M}}(v_0)) \arrow[r] \arrow[d] & H^1(F,\widetilde{M}) \arrow[d] \\ H^1(F,\mathrm{Aut}_F(C_0,Q_0,\beta_0)) \arrow[r] & H^1(F,S_3) \end{tikzcd} \tag{8.1}$$

We make several observations about this commutative diagram.

Lemma 8.2. (i) *The first vertical arrow is bijective.*
(ii) *The second vertical arrow is bijective.*
(iii) *The horizontal arrows are surjective.*

Proof. (i) This follows by Proposition 6.1.

(ii) Let the second vertical arrow be denoted by ψ. Since $\widetilde{M}$ is a semidirect product of M and S_3, the map ψ is surjective. For injectivity, we shall use the exact sequence of pointed sets

$$1 \longrightarrow H^1(F, M) \longrightarrow H^1(F, \widetilde{M}) \longrightarrow H^1(F, S_3) \longrightarrow 1.$$

Let $c \in H^1(F, S_3)$ and let E be the étale cubic algebra corresponding to c. Then M_E is the twist of M by c. In order to prove that $\psi^{-1}(c)$ consists of one element, it suffices to show that $H^1(F, M_E)$ is trivial, by the twisting argument on page 50 of [Serre 2002]. We have an exact sequence of algebraic groups

$$1 \longrightarrow M_{E,\mathrm{der}} \longrightarrow M_E \longrightarrow \mathrm{GL}_1 \longrightarrow 1,$$

where $M_{E,\mathrm{der}} \cong \mathrm{Res}_{E/F}\,\mathrm{SL}_2$. By Hilbert's theorem 90, $H^1(F, \mathrm{GL}_1)$ is trivial. Since

$$H^1(F, \mathrm{Res}_{E/F}\,\mathrm{SL}_2) = H^1(E, \mathrm{SL}_2) = 0$$

(see [Serre 2002, p. 130]), it follows that $H^1(F, M_E)$ is trivial.

(iii) This follows because $\mathrm{Stab}_{\widetilde{M}}(v_0) = \mathrm{Stab}_M(v_0) \rtimes \mathrm{Aut}(\Pi)$, hence

$$H^1(F, \mathrm{Stab}_{\widetilde{M}}(v_0)) \to H^1(F, \mathrm{Aut}(\Pi))$$

has a natural splitting. □

8-2. *Determination of orbits.* We can now determine the generic $\widetilde{M}_E(F)$-orbits on $V_E(F)$.

Theorem 8.3. *Fix an étale cubic F-algebra E.*

(i) *The generic $\widetilde{M}_E(F)$-orbits on $V_E(F)$ are in bijective correspondence with the set of F-isomorphism classes of E-twisted composition algebras over F, with the orbit of $v_{0,E} = (1, 0, 0, 1)$ corresponding to the twisted composition algebra C_E introduced in* Section 3-4.

(ii) *The generic $M_E(F)$-orbits on $V_E(F)$ are in bijective correspondence with the set of E-isomorphism classes of E-twisted composition algebras over F.*

(iii) *There is a commutative diagram*

$$\begin{tikzcd}
\{E\text{-twisted composition algebras}\} \arrow[r] \arrow[d] & \{\text{étale quadratic } F\text{-algebras}\} \arrow[d] \\
\{\text{generic } \widetilde{M}_E\text{-orbits on } V_E\} \arrow[r] & F^\times/F^{\times 2}
\end{tikzcd}$$

where the bottom arrow is the map induced by Δ_E (see Section 7-2).

Proof. (i) Given a cohomology class $[E] \in H^1(F, S_3)$ corresponding to an étale cubic F-algebra, we consider the fibers of the two horizontal arrows in the commutative diagram (8.1) over $[E]$. Since the map $\mathrm{Stab}_{\widetilde{M}}(v_0) \longrightarrow S_3$ splits, the fiber of the second horizontal arrow has a distinguished element which corresponds to the twisted composition algebra C_E. Similarly, the fiber over $[E]$ of the first horizontal arrow has a distinguished point which corresponds to the orbit of $v_{0,E} = (1, 0, 0, -1)$. Moreover, these two distinguished point correspond under the first vertical arrow.

By the twisting argument [Serre 2002, p. 50] we see that both fibers in question are naturally identified with

$$\mathrm{Ker}(H^1(F, \mathrm{Stab}_{\widetilde{M}_E}(v_{0,E})) \longrightarrow H^1(F, \widetilde{M}_E)).$$

Thus, the fiber of the first horizontal map over $[E]$ are the generic $\widetilde{M}_E$-orbits in V_E, while the fibers of the second map are F-isomorphism classes of E-twisted composition algebras.

(ii) The bijection follows because both sets are in natural bijection with the set $H^1(F, \mathrm{Stab}_{M_E}(v_{0,E}) = H^1(F, \mathrm{Aut}_E(C_E))$.

(iii) Suppose an E-twisted composition algebra is represented by a cocycle

$$(a_\sigma) \in H^1(F, \mathrm{Stab}_{M_E}(v_{0,E})).$$

Then the associated étale quadratic F-algebra K corresponds to the group homomorphism

$$\eta_K : \mathrm{Gal}(\bar{F}/F) \longrightarrow \mathrm{Stab}_{M_E}(v_{0,E})(\bar{F}) \longrightarrow \mathbb{Z}/2\mathbb{Z}$$

given by $\sigma \mapsto a_\sigma \mapsto \pi(a_\sigma)$, where $\pi : \mathrm{Stab}_{M_E}(v_{0,E}) \to \mathbb{Z}/2\mathbb{Z}$ is the natural projection. In fact, regarding $\mathrm{Stab}_{M_E}(v_{0,E}) \subseteq M_E$ as described in Section 7-3, we see that the map π is simply given by the determinant map on $M_E = \mathrm{GL}_2(E)^0$.

On the other hand, the cocycle splits in $H^1(F, M_E) = 0$, so that we may write

$$a_\sigma = g^{-1} \cdot \sigma(g) \quad \text{for some } g \in M_E(\bar{F}).$$

Then the $\widetilde{M}_E$-orbit associated to (a_σ) is that of $g \cdot v_{0,E}$. Now, we have

$$\Delta_E(g \cdot v_{0,E}) = \det(g)^2 \cdot \Delta_E(v_{0,E}) = \det(g)^2$$

and

$$\eta_K(\sigma) = \det(a_\sigma) = \det(g)^{-1} \cdot \sigma(\det(g))$$

for any $\sigma \in \mathrm{Gal}(\bar{F}/F)$. This shows that $\det(g)$ is a trace-zero element in K, so that K is represented by the square class of $\det(g)^2 \in F^\times$, as desired. □

In particular, we have established Theorem 1.1. However, the bijection between the generic $\widetilde{M}_E(F)$-orbits on $V_E(F)$ and the F-isomorphism classes of twisted composition algebras is obtained by a Galois cohomological argument, which is quite formal and not at all explicit. For applications, it is necessary to have an explicit description of the bijection. We shall arrive at such an explicit description in the following sections.

9. Reinterpreting Bhargava

In this section, revisiting the case when $E = F^3$ is split, we shall reinterpret [Bhargava 2004a] in the framework of twisted composition algebras, leading to an explicit recipe for the bijection in Theorem 8.3.

9-1. ***Bhargava's result.*** We first review briefly Bhargava's results and, following him, we shall work over $\mathbb{Z}$. Note that we have an action of the group $\mathrm{SL}_2(\mathbb{Z})^3$ on the set of integer-valued cubes, by the "row-column" operations described in Section 6-1.

In order to state the main result of Bhargava, we need a couple of definitions. Fix a discriminant Δ. Let $K = \mathbb{Q}(\sqrt{\Delta})$ and R the unique order of discriminant Δ. A module M is a full lattice in K. In particular, it is a $\mathbb{Z}$-module of rank 2. We shall write $M = \{u, v\}$ if u and v span M. For example,

$$R = \left\{1, \frac{\Delta + \sqrt{\Delta}}{2}\right\}.$$

By fixing this basis of R, we have also fixed a preferred orientation of bases of modules. An oriented module is a pair (M, ϵ), where ϵ is a sign. If $M = \{u, v\}$, then M becomes an oriented module (M, ϵ), where $\epsilon = 1$ if and only if the orientation of $\{u, v\}$ is preferred. The norm of an oriented module (M, ϵ) is $N(M) = \epsilon \cdot [R : M]$.

Then:

- A triple of oriented modules (M_1, M_2, M_3), with R as the multiplier ring, is said to be *colinear* if there exists $\delta \in K^\times$ such that the product of the three oriented modules is a principal oriented ideal $((\delta), \epsilon)$, where $\epsilon = \mathrm{sign}(N(\delta))$, i.e., $M_1 M_2 M_3 = (\delta)$, as ordinary modules, and $N(M_1)N(M_2)N(M_3) = N(\delta)$.
- A cube is *projective* of discriminant Δ if the three associated forms are primitive and have discriminant Δ.
- Two triples of oriented modules (M_1, M_2, M_3) and (M_1', M_2', M_3') are equivalent if there exist μ_1, μ_2, μ_3 in $K^\times$ with $M_i' = \mu_i M_i$ and $\epsilon_i' = \mathrm{sign}(N(\mu_i))\epsilon_i$ for $i = 1, 2, 3$.

Then, Bhargava [2004a] showed:

Theorem 9.1. *There is a bijection, to be described in the proof, between the equivalence classes of oriented colinear triples of discriminant Δ and the $SL_2(\mathbb{Z})^3$-equivalence classes of projective cubes of discriminant Δ.*

Sketch of proof. Let v be a projective cube. Again, without any loss of generality we can assume that the cube is reduced and that the numbers f_1, f_2 and f_3 are nonzero. Define three modules by

$$M_1=\left\{1,\frac{b-\sqrt{\Delta}}{2f_1}\right\},\quad M_2=\left\{1,\frac{b-\sqrt{\Delta}}{2f_2}\right\}\quad\text{and}\quad M_3=\left\{1,\frac{b-\sqrt{\Delta}}{2f_3}\right\}.$$

The norms of the three modules are $-1/f_1, -1/f_2$ and $-1/f_3$, respectively, if we take the given bases to be proper. For δ, we shall take

$$\delta=-\frac{2}{b+\sqrt{\Delta}},$$

which has the correct norm $-1/(f_1f_2f_3)$.

The modules M_i, with given oriented bases, correspond to the quadratic forms Q_i. More precisely, if

$$z_i=x_i+y_i\frac{b-\sqrt{\Delta}}{2f_i}\in M_i,$$

then

$$-f_iN(z_i)=Q_i(x_i,y_i)=-f_ix_i^2-bx_iy_i+f_i^\# y_i^2,$$

where $f^\#=(f_2f_3,f_3f_1,f_1f_2)$. □

9-2. ***Integral twisted composition algebras.*** We can now give a reinterpretation of Bhargava's results, in particular of Bhargava's triples (M_1,M_2,M_3), in the framework of twisted composition algebras. Assume the notation from the previous subsection, so that $M_1M_2M_3=(\delta)$. Set

$$C=M_1\oplus M_2\oplus M_3.$$

We shall define a pair of tensors (Q,β) on C as follows:

- Define a quadratic form $Q:C\to\mathbb{Z}\times\mathbb{Z}\times\mathbb{Z}$ by

$$Q(z_1,z_2,z_3)=(-f_1N(z_1),-f_2N(z_2),-f_3N(z_3))=-f\cdot(N(z_1),N(z_2),N(z_3)).$$

- Define a quadratic map $\beta:C\to C$ by

$$\beta(z_1,z_2,z_3)=\delta(f_2f_3\bar z_2\bar z_3,f_3f_1\bar z_3\bar z_1,f_1f_2\bar z_1\bar z_2)=\delta\cdot f^\#\cdot(\bar z_1,\bar z_2,\bar z_3)^\#.$$

The relations $M_1M_2M_3=(\delta)$ and $M\bar{M}=N(M)$ imply that β is well defined. Moreover, using $N(\delta)=-1/(f_1f_2f_3)$, one checks that

$$Q(\beta(z_1,z_2,z_3))=Q(z_1,z_2,z_3)^\#$$

and

$$N_C(z_1,z_2,z_3)=\operatorname{Tr}\left(\frac{z_1z_2z_3}{\delta}\right).$$

Thus the triple (C,Q,β) is a twisted composition algebra over $\mathbb{Z}$.

In terms of the coordinates (x_i, y_i) given by

$$z_i = x_i + y_i \frac{b - \sqrt{\Delta}}{2f_i},$$

we have seen in the sketch proof of Theorem 9.1 that

$$Q_i(z_i) = -f_i N(z_i) = -f_i x_i^2 - b x_i y_i + f_i^{\#} y_i^2.$$

We shall now do the same for β. Write $\beta(z_1, z_2, z_3) = (z_1', z_2', z_3')$, and let (x_i', y_i') be the coordinates of z_i'. A short calculation shows that

$$x_1' = -\begin{pmatrix} x_3 & y_3 \end{pmatrix} \begin{pmatrix} 0 & f_3 \\ f_2 & b \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \quad \text{and} \quad y_1' = \begin{pmatrix} x_3 & y_3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & f_1 \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix},$$

while the expressions for (x_2', y_2') and (x_3, y_3) are obtained by cyclically permuting the indices.

There are two important observations to be made here:

- Firstly, these formulas make sense for any triple (f_1, f_2, f_3) and any b, i.e., the f_i can be zero. The axioms of twisted composition algebra are satisfied for formal reasons. For example, if $(f_1, f_2, f_3) = (0, 0, 0)$ and $b = -1$, we get the split algebra C_0.
- Secondly, the two matrices are two opposite faces of the cube. This gives a hint how to directly associate a composition algebra to any cube in general (i.e., not just a reduced cube).

9-3. ***From cubes to twisted composition algebras.*** The above discussion suggests an explicit recipe for associating a twisted composition algebra over $F \times F \times F$ to any cube $v \in V(F)$.

Let $C = F^2 \times F^2 \times F^2$. An element $z \in C$ is a triple (z_1, z_2, z_3) of column vectors $z_i = \binom{x_i}{y_i}$. Slice a cube into three pairs of 2×2-matrices (A_i, B_i), as before, and let

$$Q_i(z_i) = -\det(A_i x_i + B_i y_i).$$

Then, we set:

- $Q : C \to F \times F \times F$, defined by

$$Q(z_1, z_2, z_3) = (Q_1(z_1), Q_2(z_2), Q_3(z_3)).$$

- $\beta : C \to C$, defined by

$$\beta(z_1, z_2, z_3) = (z_1', z_2', z_3'),$$

where $z_i' = \binom{x_i'}{y_i'}$,

$$x_1' = -z_3^\top B_1 z_2, \quad x_2' = -z_1^\top B_2 z_3, \quad x_3' = -z_2^\top B_3 z_1$$

and

$$y_1' = z_3^\top A_1 z_2, \quad y_2' = z_1^\top A_2 z_3, \quad y_3' = z_2^\top A_3 z_1.$$

Thus, starting from a cube v, we have defined a pair of tensors (Q, β) on $C = F^2 \times F^2 \times F^2$. Let

$$\tilde{\phi} : V(F) \longrightarrow \{\text{tensors } (Q, \beta) \text{ on } C\}$$

be the resulting map. We may express this map using the coordinates (a, e, f, b) of a cube. A short calculation gives

$$Q(x, y) = (e^\# - af)x^2 + (-ab - 2ef + \mathrm{Tr}(ef))xy + (f^\# - be)y^2,$$
$$\beta(x, y) = (-ex^\# - by^\# - (fx) \times y, ax^\# + fy^\# + (ey) \times x).$$

In the next section, we shall study the properties of the map $\tilde{\phi}$; for example, we shall show that a (Q, β) in the image of $\tilde{\phi}$ does define a twisted composition algebra on C.

10. Explicit parametrization

Using the results of the previous section, we can now give an explicit description of the bijection between $\widetilde{M}_E(F)$-orbits of nondegenerate cubes and F-isomorphism classes of E-twisted composition algebras.

10-1. ***Definition of*** $\boldsymbol{\tilde{\phi}}$**.** Let us write $C = E \cdot e_1 \oplus E \cdot e_2$. Motivated by the case where $E = F^3$, studied in the previous section, we define the map

$$\tilde{\phi} : V_E(F) \longrightarrow \{\text{tensors } (Q, \beta) \text{ on } C\}$$

using the coordinates $v = (a, e, f, b)$ of a cube, with $a, b \in F$ and $e, f \in E$, by

$$\begin{aligned} Q(x, y) &= (e^\# - af)x^2 + (-ab - 2ef + \mathrm{Tr}(ef))xy + (f^\# - be)y^2, \\ \beta(x, y) &= (-ex^\# - by^\# - (fx) \times y, ax^\# + fy^\# + (ey) \times x). \end{aligned} \tag{10.1}$$

In particular, for a reduced cube $(1, 0, f, b)$, one has

$$\begin{aligned} Q(x, y) &= -fx^2 - bxy + f^\# y^2, \\ \beta(x, y) &= (-by^\# - (fx) \times y, x^\# + fy^\#). \end{aligned} \tag{10.2}$$

Thus, the image of the distinguished cube $v_{E,0} = (1, 0, 0, -1)$ is the algebra C_E. Observe also that one has

$$\beta(1, 0) = (0, 1) \quad \text{and} \quad \beta(0, 1) = (-b, f). \tag{10.3}$$

Thus, the standard basis $\{e_1, e_2\}$ is a reduced basis with respect to (Q, β), in the sense of Section 3-6.

Proposition 10.4. (i) *The map $\tilde{\phi}$ is injective.*

(ii) *For $g \in \mathrm{GL}_2(E)^0$ and $\sigma \in S_E(F)$, one has*

$$\tilde{\phi}(g \cdot v) = {}^t g^{-1} \cdot \tilde{\phi}(v) \quad \text{and} \quad \tilde{\phi}(\sigma \cdot v) = \sigma \cdot \tilde{\phi}(v)$$

for any $v \in V_E(F)$.

Thus, the map $\tilde{\phi}$ is $\mathrm{GL}_2(E)^0 \rtimes S_E$-equivariant, with respect to the outer automorphism $(g, \sigma) \mapsto ({}^t g^{-1}, \sigma)$ of $\mathrm{GL}_2(E)^0 \rtimes S_E$, and where the action of $\mathrm{GL}_2(E)^0 \rtimes S_E$ on the set of (Q, β) is given as in Section 3-3.

(iii) *For any nondegenerate cube v, $\tilde{\phi}(v) = (Q, \beta)$ defines a twisted composition algebra on C.*

Proof. (i) If $\tilde{\phi}(a, e, f, b) = (Q, \beta)$, then

$$\beta(1, 0) = (-e, a) \quad \text{and} \quad \beta(0, 1) = (-b, f).$$

Hence the cube (a, e, f, b) is uniquely determined by β.

(ii) We can verify this equivariance property over $\overline{F}$; thus we only need to check it for $E = F^3$. For the central element $(t, t, t) \in \mathrm{GL}_2(E)^0$ or the element $\sigma \in S_E$, the desired equivariance property is clear. Thus, it remains to verify it for elementary matrices such as

$$g = (E_u, 1, 1) = \left(\begin{pmatrix} 1 & u \\ 0 & 1 \end{pmatrix}, 1, 1 \right) \in (\mathrm{GL}_2(F) \times \mathrm{GL}_2(F) \times \mathrm{GL}_2(F))^0.$$

Now, if the cube v has a pair of faces (A_1, B_1), then the corresponding pair for $g \cdot v$ is

$$(A_1', B_1') = (A_1 + u B_1, B_1).$$

Slicing the cube in the other two ways, we obtain

$$(A_2', B_2') = (E_u A_2, E_u B_2) \quad \text{and} \quad (A_3', B_3') = (A_3 E_u^t, B_3 E_u^t).$$

Hence, if $\tilde{\phi}(g \cdot v) = (Q', \beta')$, then β' is given on $(z_1, z_2, z_3) \in F^2 \times F^2 \times F^2$ by

$$\begin{pmatrix} x_1' & x_2' & x_3' \\ y_1' & y_2' & y_3' \end{pmatrix} = \begin{pmatrix} -z_3^t B_1 z_2 & -z_1^t E_u B_2 z_3 & -z_2^t B_3 E_u^t z_1 \\ z_3^t (A_1 + u B_1) z_2 & z_1^t E_u A_2 z_3 & z_2^t A_3 E_u^t z_1 \end{pmatrix}.$$

On the other hand, ${}^tg^{-1}$ acts on β by precomposing by $({}^tg^{-1})^{-1} = g^t$, and postcomposing by ${}^tg^{-1}$:

$$\begin{aligned}
{}^tg^{-1}\cdot\beta(g^t(z_1,z_2,z_3)) &= {}^tg^{-1}\cdot\beta(E_u^t z_1, z_2, z_3)\\
&= {}^tg^{-1}\cdot\begin{pmatrix} -z_3^tB_1z_2 & -z_1^tE_uB_2z_3 & -z_2^tB_3E_u^tz_1\\ z_3^tA_1z_2 & z_1^tE_uA_2z_3 & z_2^tA_3E_u^tz_1\end{pmatrix}\\
&= \begin{pmatrix} -z_3^tB_1z_2 & -z_1^tE_uB_2z_3 & -z_2^tB_3E_u^tz_1\\ z_3^t(A_1+uB_1)z_2 & z_1^tE_uA_2z_3 & z_2^tA_3E_u^tz_1\end{pmatrix}\\
&= \beta'(z_1,z_2,z_3).
\end{aligned}$$

(iii) Again, we may work over $\overline{F}$, and hence we may assume that $E = F^3$. If v is a reduced cube, we have seen in Section 9-2 that (Q,β) defines a twisted composition algebra on E^2. Since every $\widetilde{M}(F)$-orbit contains a reduced cube, the result follows by (ii). □

The occurrence of the outer automorphism $g \mapsto {}^tg^{-1}$ is natural here. Indeed, assume that $E = F^3$ and regard $\mathrm{GL}_2(F)$ as $\mathrm{GL}(V)$ for a 2-dimensional F-vector space V. Then the quadratic map β is an element of $(V^*)^{\oplus 3}\otimes_F (V^*)^{\oplus 3}\otimes_F V^{\oplus 3}$, whereas its associated cube is an element in $V\otimes_F V\otimes_F V\otimes_F \det(V)^{-1}$. Thus scaling a cube by $t\in F^\times$ corresponds to scaling β by t^{-1}.

10-2. ***Reduced cubes and bases.*** To describe the image of $\tilde{\phi}$, we examine the case of reduced cubes more carefully.

Proposition 10.5. *Suppose that the pair (Q,β) defines a twisted composition algebra structure on E^2 such that the standard basis $\{e_1,e_2\}$ is reduced (i.e., $\beta(e_1)=e_2$). Then (Q,β) is the image under $\tilde{\phi}$ of the reduced cube*

$$v = (1, 0, -Q(e_1), -N_{Q,\beta}(e_1)).$$

Moreover, $\Delta_E(v) = \Delta_{Q,\beta}(e_1)$ (where the Δ on the left side is the quasi-invariant form on the space V_E of cubes while the one on the right is defined in Proposition 3.5*).*

Proof. We need to show that Q and β are uniquely determined by $f = -Q(e_1)$ and $b = -N_{Q,\beta}(e_1)$. Since

$$Q(e_2) = Q(\beta(e_1)) = f^\# \quad\text{and}\quad b_Q(e_1,e_2) = b_Q(e_1,\beta(e_1)) = N(e_1) = -b,$$

we see that Q is uniquely determined. Then $\beta(xe_1+ye_2)$ is uniquely determined by (3.4) in Lemma 3.2. Finally, observe that

$$\Delta_E(v) = \Delta_{Q,\beta}(e_1) = b^2 + 4N_E(f). \qquad □$$

10-3. ***Good bases.*** We call a basis of C a *good basis* if it is in the $\mathrm{Aut}_E(C)^0 \cong \mathrm{GL}_2(E)^0$-orbit of a reduced basis. By Proposition 3.5(iv), this notion is independent of the choice of the reduced basis. Similarly, since the action of S_E preserves the set of reduced cubes, the notion of good bases does not depend on whether one uses $\mathrm{Aut}_E(C)^0$ or $\mathrm{Aut}_F(E, C)^0 \cong \mathrm{GL}_2(E)^0 \rtimes S_E$.

As a consequence of the proposition, we have:

Corollary 10.6. (i) *The map $\tilde{\phi}$ gives a bijection between the set of reduced* (*nondegenerate*) *cubes and the set of* (Q, β) *on* E^2 *such that the standard basis* $\{e_1, e_2\}$ *is reduced.*

(ii) *The image of* $\tilde{\phi}$ *consists precisely of those* (Q, β) *such that the standard basis* $\{e_1, e_2\}$ *of* $C = E^2$ *is a good basis for* (Q, β).

The definition we have given for a good basis $\{e_1, e_2\}$ may not seem very satisfactory. It would have been more satisfactory if one defines a good basis for (C, Q, β) using purely the forms (Q, β) rather than using the action of $\mathrm{Aut}_E(C)^0$. Indeed, it will not be easy to check that a given basis is good by our definition. However, by Corollary 10.6, one knows *a posteriori* that a basis $\{e_1, e_2\}$ is good for (C, Q, β) if and only if $\beta(xe_1 + ye_2)$ has the form given in (10.1) with $a, b \in F$. We would have taken this as a definition, but it would have seemed completely unmotivated without the results of this section!

10-4. ***A commutative diagram.*** As a summary of the above discussion, we have the following refinement and explication of Theorem 8.3:

Theorem 10.7. (i) *The bijective map* $\tilde{\phi}$ *descends to give a commutative diagram*

$$\begin{tikzcd}
V_E(F)^0 = \{c \in V_E(F) : \Delta_E(c) \neq 0\} \arrow[r] \arrow[d, "\tilde{\phi}"'] & \tilde{M}_E(F)\text{-orbits on } V_E(F)^0 \arrow[d, "\phi"] \\
\{\text{pairs } (Q,\beta) \text{ on } E^2\text{: standard basis is good}\} \arrow[r] \arrow[d] & \{\mathrm{GL}_2(E)^0 \rtimes S_E(F)\text{-orbits of } (Q,\beta)\} \arrow[d] \\
\{F\text{-isomorphism classes of pairs } (C,b)\} \arrow[r] & \{F\text{-isomorphism classes of } C\}
\end{tikzcd}$$

where all vertical arrows are $\mathrm{GL}_2(E)^0 \rtimes S_E(F)$*-equivariant bijections and, in the last row,* C *denotes an* E*-twisted composition algebra and* b *denotes a good basis of* C. *Moreover the action of* $\mathrm{GL}_2(E)^0 \rtimes S_E(F)$ *on a pair* $(C, \{e_1, e_2\})$ *is given as follows*: $g \in \mathrm{GL}_2(E)^0$ *sends the pair to* $(C, \{e_1', e_2'\})$, *where*

$$\begin{pmatrix} e_1' \\ e_2' \end{pmatrix} = g \cdot \begin{pmatrix} e_1 \\ e_2 \end{pmatrix},$$

whereas $\sigma \in S_E$ *sends the pair to* $(E \otimes_{E,\sigma} C, \{e_1, e_2\})$.

(ii) *The bijection* ϕ *agrees with the one given in Theorem 8.3.*

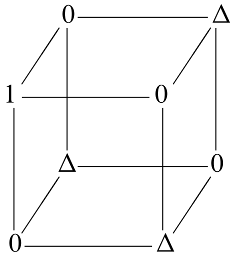
Proof. (i) Our discussion above already shows that $\tilde{\phi}$ is bijective and descends to give the map ϕ. It remains to show that the induced map ϕ is bijective. The surjectivity of ϕ follows from Proposition 3.5(iii) and (iv) and Corollary 10.6(i). The injectivity of ϕ follows from Proposition 10.4(i) and (ii). We leave the bijection and the equivariance of the lower half of the diagram to the reader.

(ii) The map $\tilde{\phi}$ sends the distinguished cube $v_{E,0} = (1,0,0,-1)$ to the pair (Q_0, β_0) on E^2, which defines the algebra C_E. Moreover, $\tilde{\phi}$ is equivariant with respect to the automorphism $g \mapsto {}^tg^{-1}$ of $\mathrm{GL}_2(E)$, which preserves the subgroup $\mathrm{Stab}_{\mathrm{GL}_2(E)^0}(v_{E,0}) = \mathrm{Aut}_E(Q_0, \beta_0) \subset \mathrm{GL}_2(E)^0$. Finally, since $\tilde{\phi}$ is algebraic, it is Galois-equivariant with respect to base field extension. All these imply that we have a commutative diagram

$$\begin{tikzcd}
\{\mathrm{GL}_2(E)^0\text{-orbits on } V_E(F)^0\} \arrow[r] \arrow[d, "\phi"] & H^1(F, \mathrm{Stab}_{\mathrm{GL}_2(E)^0}(v_{E,0})) \arrow[d, "g \mapsto {}^tg^{-1}"] \\
\{E\text{-isomorphism classes of twisted composition algebras}\} \arrow[r] & H^1(F, \mathrm{Aut}_E(C_E))
\end{tikzcd}$$

Since the map $g \mapsto {}^tg^{-1}$ of $\mathrm{Stab}_{\mathrm{GL}_2(E)^0}(v_{E,0}) = \mathrm{Aut}_E(Q_0, \beta_0)$ is given by conjugation by the element $w \in \mathrm{Aut}_E(Q_0, \beta_0)(F)$, we see that the induced map on H^1 is trivial. Hence ϕ agrees with the bijection given in Theorem 8.3 by a Galois cohomological argument. □

10-5. ***An example.*** As an example, assume that $K = F(\sqrt{\Delta})$ and consider the composition algebra given by the example in Section 4-4. (This is the distinguished point in the fiber of $([F^3], [K])$.) Then $v = (\sqrt{\Delta}, \sqrt{\Delta}, \sqrt{\Delta})$ and $\beta(v) = (\Delta, \Delta, \Delta)$ is a reduced basis. The corresponding reduced cube is



10-6. ***Relation with Tits' construction.*** If $f \in E^{\times}$, we can relate the construction of $\tilde{\phi}$ attached to the reduced cube $(1, 0, -f, b)$ to Proposition 3.12. Identify $E \oplus E$ with $E \otimes K$ using the E-linear isomorphism given by

$$(x, y) \mapsto x \otimes 1 + \frac{y}{f} \otimes \frac{b - \sqrt{\Delta}}{2} = x + y\frac{b - \sqrt{\Delta}}{2f},$$

where, in the last expression, we omitted tensor product signs for readability. Then Q can be written as

$$Q\left(x+y\frac{b-\sqrt{\Delta}}{2f}\right)=-f\cdot N_{E\otimes K/E}\left(x+y\frac{b-\sqrt{\Delta}}{2f}\right)$$

and β as

$$\beta\left(x+y\frac{b-\sqrt{\Delta}}{2f}\right)=-\frac{2}{b+\sqrt{\Delta}}\cdot f^{\#}\cdot\left(x+y\frac{b+\sqrt{\Delta}}{2f}\right)^{\#}.$$

Indeed, if $E=F^3$, these formulae are exactly the same as those in Section 9-2 . Let

$$e=-f \quad\text{and}\quad \nu=-\frac{b+\sqrt{\Delta}}{2}.$$

Using $e^{-1}\cdot\bar{\nu}=\nu^{-1}\cdot e^{\#}$ (since $N_{E/F}(e)=N_{K/F}(\nu)$) this composition algebra is the algebra attached to the pair (e,ν), as in Proposition 3.12. Conversely, a composition algebra given by a pair (e,ν), as in Proposition 3.12, arises from the cube $(1,0,-e,b)$ where $b=-\operatorname{Tr}_{K/F}(\nu)$.

11. Exceptional Hilbert 90

Assume that E is an étale cubic F-algebra with corresponding étale quadratic discriminant algebra K_E, and let K be an étale quadratic F-algebra. Recall that

$$T_{E,K}=\{x\in E\otimes_F K : N_{E/F}(x)=1=N_{K/F}(x)\}.$$

Suppose, for example, that $[K_E]=[K]=1$, so E is a Galois extension, and $T_{E,K}$ is the group of norm-one elements in $E^\times$. Let σ be a generator of the Galois group $G_{E/F}$. Then Hilbert's theorem 90 states that the map

$$x\mapsto\sigma(x)/\sigma^2(x)$$

induces an isomorphism of $E^\times/F^\times$ and $T_{E,K}(F)$. Our goal in this section is to generalize this statement to all tori $T_{E,K}$, thus obtaining an exceptional Hilbert's theorem 90. As an application, we give an alternative description of $H^1(F,T_{E,K})$.

11-1. ***The torus $T_{E,K}$.*** We first describe the torus $T_{E,K}$ by Galois descent. Over $\bar{F}$, we have the identification

$$T_{E,K}(\bar{F})=\{(\underline{a},\underline{b})\in\bar{F}^3\otimes\bar{F}^2 : a_ib_i=1 \text{ for all } i \text{ and } a_1a_2a_3=1\}.$$

The F-structure is given by the twist of the Galois action on coordinates by the cocycle

$$\rho_E\times\rho_K:\operatorname{Gal}(\bar{F}/F)\longrightarrow\operatorname{Aut}(\bar{F}^3)\times\operatorname{Aut}(\bar{F}^2)\cong S_3\times\mathbb{Z}/2\mathbb{Z},$$

where S_3 and $\mathbb{Z}/2\mathbb{Z}$ act on $\mathbb{Z}^3$ and $\mathbb{Z}^2$, respectively, by permuting the coordinates.

We may describe $T_{E,K}$ using its cocharacter lattice X. We have

$$X = \{(\underline{a}, -\underline{a}) \in \mathbb{Z}^3 \otimes \mathbb{Z}^2 : a_1 + a_2 + a_3 = 0\},$$

equipped with the Galois action given by

$$\rho_E \otimes \rho_K : \mathrm{Gal}(\overline{F}/F) \longrightarrow S_3 \times \mathbb{Z}/2\mathbb{Z}.$$

11-2. *The torus $T'_{E,K}$.* Now we introduce another torus $T'_{E,K}$ over F. Let K_J be the étale quadratic F-algebra such that $[K_J] \cdot [K] \cdot [K_E] = 1$ in $H^1(F, \mathbb{Z}/2\mathbb{Z})$. We define the tori

$$\widetilde{T}'_{E,K} = \{x \in E \otimes_F K_J : N_{E \otimes K_J/E}(x) \in F^\times\}$$

and

$$T'_{E,K} = \widetilde{T}'_{E,K}/K_J^\times,$$

where the last quotient is taken in the sense of algebraic groups. If $J = B^\tau$, where B is a degree-3 central simple K_J-algebra with an involution τ of the second kind, and $E \to J$ is an F-embedding or, equivalently, $E \otimes_F K_J \to B$ is a K_J-embedding such that τ pulls back to the nontrivial element of $\mathrm{Aut}(E \otimes_F K_J/E)$, then $T'_{E,K}$ acts naturally as a group of automorphisms of the embedding $E \to J$.

We may again describe these tori by Galois descent. Over $\overline{F}$, we may identify

$$\widetilde{T}'_{E,K}(\overline{F}) = \{(\underline{a}, \underline{b}) \in (\overline{F}^\times)^3 \otimes (\overline{F}^\times)^2 : a_1 b_1 = a_2 b_2 = a_3 b_3\},$$

and $T'_{E,K}(\overline{F})$ is the quotient of this by the subgroup consisting of the elements $(a \cdot \underline{1}, b \cdot \underline{1})$. The action of $\mathrm{Gal}(\overline{F}/F)$ which gives the F-structure of $\widetilde{T}'_{E,K}$ is then described as follows. Let $\rho_E : \mathrm{Gal}(\overline{F}/F) \longrightarrow S_3$ be the cocycle associated to E, so that $\mathrm{sign} \circ \rho_E : \mathrm{Gal}(\overline{F}/F) \longrightarrow \mathbb{Z}/2\mathbb{Z}$ is the homomorphism associated to K_E. On the other hand, we let ρ_K be the homomorphism associated to K, so that

$$(\mathrm{sign} \circ \rho_E) \cdot \rho_K : \mathrm{Gal}(\overline{F}/F) \longrightarrow \mathbb{Z}/2\mathbb{Z}$$

is the homomorphism associated to K_J. Now the action of $\mathrm{Gal}(\overline{F}/F)$ on $\overline{F}^3 \otimes \overline{F}^2$ is the twist of the action on coordinates by the cocycle

$$\rho_E \times (\mathrm{sign} \circ \rho_E) \cdot \rho_K : \mathrm{Gal}(\overline{F}/F) \longrightarrow S_3 \times \mathbb{Z}/2\mathbb{Z}.$$

As before, we may describe the tori $\widetilde{T}'_{E,K}$ and $T'_{E,K}$ by their cocharacter lattice. The cocharacter lattice $\widetilde{Y}$ of $\widetilde{T}'_{E,K}$ is given by

$$\widetilde{Y} = \{(\underline{a}, \underline{b}) \in \mathbb{Z}^3 \otimes \mathbb{Z}^2 : a_1 + b_1 = a_2 + b_2 = a_3 + b_3\},$$

equipped with the Galois action given by

$$\rho_E \times (\mathrm{sign} \circ \rho_E) \cdot \rho_K : \mathrm{Gal}(\overline{F}/F) \longrightarrow S_3 \times \mathbb{Z}/2\mathbb{Z}.$$

This contains the Galois-stable sublattice

$$Z = (1,1,1) \otimes \mathbb{Z}^2,$$

so that $Y = \widetilde{Y}/Z$ is the cocharacter lattice of $T'_{E,K}$.

11-3. ***A homomorphism.*** We are going to construct a morphism of tori from $\widetilde{T}'_{E,K}$ to $T_{E,K}$. We shall first define this morphism over $\overline{F}$ and then shows that it descends to F.

Now we may define a morphism over $\overline{F}$,

$$f : \widetilde{T}'_{E,K}(\overline{F}) \longrightarrow T_{E,K}(\overline{F}),$$

by

$$f : \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \mapsto \begin{pmatrix} a_2/a_3 & a_3/a_1 & a_1/a_2 \\ b_2/b_3 & b_3/b_1 & b_1/b_2 \end{pmatrix}.$$

It is easy to see that this defines an $\overline{F}$-isomorphism of tori

$$f : T'_{E,K}(\overline{F}) \cong T_{E,K}(\overline{F}).$$

Moreover, if $\sigma \in S_e(\overline{F}) = S_3$ is the cyclic permutation

$$(a_1, a_2, a_3) \mapsto (a_2, a_3, a_1),$$

then the map f is given by

$$f(x) = \sigma(x)/\sigma^2(x).$$

Now the morphism f induces a map

$$f_* : \widetilde{Y} \longrightarrow X,$$

given by

$$\begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \mapsto \begin{pmatrix} a_2 - a_3 & a_3 - a_1 & a_1 - a_2 \\ b_2 - b_3 & b_3 - b_1 & b_1 - b_2 \end{pmatrix}.$$

This induces an isomorphism of $\mathbb{Z}$-modules $Y \cong X$.

11-4. ***Exceptional Hilbert 90.*** The main result of this section is:

Theorem 11.1. *The isomorphism $f : T'_{E,K} \times_F \overline{F} \longrightarrow T_{E,K} \times_F \overline{F}$ is defined over F, and thus gives an isomorphism of tori*

$$T'_{E,K} \longrightarrow T_{E,K}$$

given by

$$x \mapsto \sigma(x)/\sigma^2(x).$$

Proof. It remains to prove that f is defined over F. For this, we may work at the level of cocharacter lattices, and we need to show that f_* is Galois-equivariant. For this, regard $\mathbb{Z}^3 \otimes \mathbb{Z}^2$ as a $S_3 \times \mathbb{Z}/2\mathbb{Z}$-module with the permutation of the coordinates in $\mathbb{Z}^3$ and $\mathbb{Z}^2$. Then observe that f_* is not equivariant with respect to $S_3 \times \mathbb{Z}/2\mathbb{Z}$. On the other hand, we have the automorphism of $S_3 \times \mathbb{Z}/2\mathbb{Z}$ given by

$$(g, h) \mapsto (g, \operatorname{sign}(g) \cdot h)$$

If we twist the $S_3 \times \mathbb{Z}/2\mathbb{Z}$-module structure on the domain of f_* by this automorphism, then f_* is easily seen to be equivariant. Together with our description of the $\operatorname{Gal}(\overline{F}/F)$-actions on the domain and codomain of f_*, the desired $\operatorname{Gal}(\overline{F}/F)$-equivariance follows. □

11-5. *Cohomology of $T_{E,K}$.* As an application of the exceptional Hilbert 90, we may give an alternative description of the cohomology group $H^1(F, T_{E,K})$, which classifies twisted composition algebras with fixed invariants (E, K) up to $E \otimes_F K$-linear isomorphisms.

In order to state our results, we need additional notation. For every quadratic extension K_J of F, let $\operatorname{Res}^1_{K_J/F} \mathbb{G}_m$ be the 1-dimensional torus defined by the short exact sequence of algebraic tori

$$1 \longrightarrow \operatorname{Res}^1_{K_J/F} \mathbb{G}_m \longrightarrow \operatorname{Res}_{K_J/F} \mathbb{G}_m \longrightarrow \mathbb{G}_m \longrightarrow 1.$$

By the classical Hilbert theorem 90, the associated long exact sequence gives the exact sequence

$$1 \longrightarrow H^2(F, \operatorname{Res}^1_{K_J/F} \mathbb{G}_m) \longrightarrow H^2(K_J, \mathbb{G}_m) \longrightarrow H^2(F, \mathbb{G}_m),$$

where the last map is the corestriction. By a theorem of Albert, Riehm, and Scharlau [Knus et al. 1998, Theorem 3.1], the kernel of the corestriction map is the set of Brauer equivalence classes of central simple algebras over K_J that admit an involution of the second kind, and so we can view $H^2(F, \operatorname{Res}^1_{K_J/F} \mathbb{G}_m)$ as the set of Brauer equivalence classes of such algebras.

Proposition 11.2. *Let K_J be an étale quadratic algebra with $[K_J] \cdot [K] \cdot [K_E] = 1$, and set $M = E \otimes_F K_J$.*

(i) *If K_J is a field, then we have an exact sequence*

$$\begin{aligned}1 \longrightarrow E^\times / F^\times N_{M/E}(M^\times) \longrightarrow H^1(F, T_{E,K})& \\ \longrightarrow H^2(F, \operatorname{Res}^1_{K_J/F} \mathbb{G}_m) \longrightarrow H^2(E, \operatorname{Res}^1_{M/E} \mathbb{G}_m)&.\end{aligned}$$

The image of $H^1(F, T_{E,K})$ consists of those central simple algebras over K_J which contain M as a K_J-subalgebra and which admit an involution of the second kind fixing E (or equivalently, restricting to the nontrivial automorphism of M over E).

(ii) *If $K_J = F^2$, then we have a simplified version of the above sequence:*

$$H^1(F, T_{E,K}) = \mathrm{Ker}(H^2(F, \mathbb{G}_m) \longrightarrow H^2(E, \mathbb{G}_m)).$$

Proof. (i) By the exceptional Hilbert theorem 90, we have a short exact sequence of algebraic tori

$$1 \longrightarrow \mathrm{Res}^1_{K_J/F}\,\mathbb{G}_m \longrightarrow \mathrm{Res}_{E/F}\,\mathrm{Res}^1_{M/E}\,\mathbb{G}_m \longrightarrow T_{E,K} \longrightarrow 1.$$

Now, (i) follows from the associated long exact sequence, using

$$H^1(F, \mathrm{Res}^1_{K_J/F}\,\mathbb{G}_m) = F^\times/N_{K_J/F}K_J^\times,$$
$$H^1(E, \mathrm{Res}^1_{M/E}\,\mathbb{G}_m) = E^\times/N_{M/E}M^\times.$$

(ii) One argues as above, except that since $K_J = F^2$, we have

$$1 \longrightarrow \mathbb{G}_m \longrightarrow \mathrm{Res}_{E/F}\,\mathbb{G}_m \longrightarrow T_{E,K} \longrightarrow 1.$$

Thus the long exact sequence gives

$$1 \longrightarrow H^1(F, T_{E,K}) \longrightarrow H^2(F, \mathbb{G}_m) \longrightarrow H^2(E, \mathbb{G}_m). \qquad \square$$

11-6. ***Interpretation.*** The above description of $H^1(F, T_{E,K})$ fits beautifully with the correspondence between E-twisted composition algebras and conjugacy classes of embeddings $E \hookrightarrow J$, where J is a Freudenthal–Jordan algebra of dimension 9.

More precisely, Proposition 11.2 exhibits $H^1(F, T_{E,K})$ as the set of isomorphism classes of triples (B, τ, i), where:

- B is a central simple K_J-algebra of degree 3.
- τ is an involution of the second kind on B.
- $i : E \longrightarrow B^\tau$ is an F-algebra embedding, or equivalently a K_J-algebra embedding $i : M = E \otimes_F K_J \longrightarrow B$ such that τ pulls back to the nontrivial element of $\mathrm{Aut}(M/E)$.

The map $\pi : H^1(F, T_{E,K}) \to H^2(F, \mathrm{Res}^1_{K_J/F}\,\mathbb{G}_m)$ sends (B, τ, i) to B. For a fixed

$$[B] \in \mathrm{Ker}(H^2(F, \mathrm{Res}^1_{K_J/F}\,\mathbb{G}_m) \longrightarrow H^2(E, \mathrm{Res}^1_{M/E}\,\mathbb{G}_m)),$$

so that B contains $M = E \otimes_F K_J$ as an K_J-subalgebra, the fiber of π over $[B]$ is the set of $\mathrm{Aut}_{K_J}(B)$-conjugacy classes of pairs (τ, i). The Skolem–Noether theorem says that any two embeddings $M \hookrightarrow B$ are conjugate, and on fixing an embedding $i : M \hookrightarrow B$, the fiber of π over $[B]$ is then the set of $\mathrm{Aut}_{K_J}(B, i)$-conjugacy classes of involutions of the second kind on B which restricts to the nontrivial automorphism of M over E. Therefore, the exact sequence in Proposition 11.2(i) says that the set of such $\mathrm{Aut}_{K_J}(B, i)$-conjugacy classes of involutions is identified with $E^\times/F^\times N_{M/E}(M^\times)$. One has a natural map on the fiber $\pi^{-1}([B])$ sending a

$\mathrm{Aut}_{K_J}(B,i)$-conjugacy class of involutions to its $\mathrm{Aut}_{K_J}(B)$-conjugacy class. This is the surjective map described in Corollary 19.31 in [Knus et al. 1998].

On the other hand, the map sending the triple (B,τ,i) to the pair (B,τ) is the natural map

$$H^1(F,T_{E,K})\longrightarrow H^1(F,PGU_3^{K_J})$$

induced by the map $T_{E,K}\hookrightarrow PU_3^{K_J}$ where $PGU_3^{K_J}$ is the identity component of the automorphism group of the Freuthendal–Jordan algebra associated to the distinguished twisted composition algebra with invariants (E,K).

12. Local fields

In this section, we specialize and explicate the main result in the case of local fields.

12-1. ***Local fields.*** Let F be a local field, E an étale cubic F-algebra, and K_E the corresponding discriminant algebra. Let K be an étale quadratic F-algebra. We consider

$$\widetilde{\Omega}_{E,K}=\{\text{generic } \widetilde{M}_E\text{-orbits on } V_E \text{ with associated quadratic algebra } K\},$$
$$\Omega_{E,K}=\{\text{generic } M_E\text{-orbits on } V_E \text{ with associated quadratic algebra } K\}.$$

We have seen that $\widetilde{\Omega}_{E,K}$ has a distinguished element: this is the distinguished point of $H^1(T_{E,K})$ which is fixed by $S_E(F)\times\mathbb{Z}/2\mathbb{Z}$. Moreover, by Galois cohomological arguments,

$$\widetilde{\Omega}_{E,K}=H^1(F,T_{E,K})/S_E(F)\times\mathbb{Z}/2\mathbb{Z}\quad\text{and}\quad\Omega_{E,K}=H^1(F,T_{E,K})/\mathbb{Z}/2\mathbb{Z}.$$

We would like to explicate the sets $\widetilde{\Omega}_{E,K}$ and $\Omega_{E,K}$.

12-2. ***Cohomology of tori.*** Recall that in (3.13), we have shown

$$H^1(F,T_{E,K})=(E^\times\times K^\times)^0/\operatorname{Im}(L^\times),$$

where $L=E\otimes_F K$,

$$(E^\times\times K^\times)^0=\{(e,\nu)\in E^\times\times K^\times : N_{E/F}(e)=N_{K/F}(\nu)\}$$

and the map from $L^\times$ to $(E^\times\times K^\times)^0$ is given by

$$a\mapsto (N_{L/E}(a),N_{L/K}(a)).$$

This description of $H^1(F,T_{E,K})$ is natural but may not be so explicit. When F is a local field, we can further explicate this description.

Since the case when E or K is not a field is quite simple, we consider the case when E and K are both fields. In that case, the norm map induces an isomorphism

$$E^\times/N_{L/E}(L^\times)\longrightarrow F^\times/N_{K/F}(K^\times)\cong\mathbb{Z}/2\mathbb{Z},$$

so that any $(e, \nu) \in (E^\times \times K^\times)^0$ has $e = N_{L/E}(a)$ for some $a \in L^\times$. Hence any element in $H^1(F, T_{E,K})$ is represented by $(1, \nu)$ for $\nu \in K^1 = \{\nu \in K^\times : N_{K/F}(\nu) = 1\}$. We thus deduce that, with $L^1 = \{a \in L^\times : N_{L/E}(a) = 1\}$,

$$H^1(F, T_{E,K}) = K^1/N_{L/K}(L^1) \cong K^\times/F^\times N_{L/K}(L^\times),$$

where the last isomorphism is induced by the usual Hilbert theorem 90. Using this last expression, we easily see that

$$H^1(F, T_{E,K}) = \begin{cases} 1 & \text{if } K \neq K_E, \\ \mathbb{Z}/3\mathbb{Z} & \text{if } K = K_E. \end{cases}$$

Exchanging the roles of E and K in the above argument, one also has

$$H^1(F, T_{E,K}) = E^1/N_{L/E}(L_1),$$

where now $L_1 = \{a \in L^\times : N_{L/K}(a) = 1\}$. If E/F is Galois (and K is a field), it follows by the usual Hilbert theorem 90 that

$$H^1(F, T_{E,K}) = E^1/N_{L/E}(L_1) \cong E^\times/F^\times N_{L/E}(E^\times) = 1,$$

thus partially recovering the result of the last section.

Alternatively, we could use Proposition 11.2 to compute $H^1(F, T_{E,K})$. If K_J is a field, then the only central simple K_J-algebra which admits an involution of the second kind is the split algebra $M_3(K_J)$. Thus we deduce from Proposition 11.2(i) that

$$H^1(F, T_{E,K}) \cong E^\times/F^\times N_{M/E}(M^\times),$$

where $M = E \otimes_F K_J$. On the other hand, if K_J is split, Proposition 11.2(ii) gives

$$H^1(F, T_{E,K}) \cong \operatorname{Ker}(H^2(F, \mathbb{G}_m) \longrightarrow H^2(E, \mathbb{G}_m)),$$

which is $\mathbb{Z}/3\mathbb{Z}$ when E is a field.

12-3. *Fibers.* With the various computations of $H^1(F, T_{E,K})$ given above, it is not difficult to show the following proposition which determines $|\widetilde{\Omega}_{E,K}|$ and $|\Omega_{E,K}|$.

Proposition 12.1. *We have*

E	K	$T_{E,K}$	$H^1(F, T_{E,K})$	$\lvert\widetilde{\Omega}_{E,K}\rvert$	$\lvert\Omega_{E,K}\rvert$
$F \times K_E$	$K = K_E$	$K^\times$	1	1	1
$F \times K_E$, K_E *a field*	*field* $\neq K_E$	$(K \otimes K_E)^\times/K_E^\times$	$\mathbb{Z}/2\mathbb{Z}$	2	2
$F \times K_E$, K_E *a field*	$F \times F$	$K_E^\times$	1	1	1
F^3	*field*	$K^\times/F^\times \times K^\times/F^\times$	$\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$	2	4
field	$K = K_E$	$E^\times/F^\times$	$\mathbb{Z}/3\mathbb{Z}$	2	2
field	$K \neq K_E$		1	1	1

Here, the difference in the last two columns reflects the fact that $S_E(F)$ acts trivially on $H^1(F, T_{E,K})$ except when $E = F^3$ and K is a field.

12-4. ***Embeddings into J.*** The main theorem says that the elements of $\Omega_{E,K}$ are in bijection with the conjugacy classes of embeddings

$$E \hookrightarrow J,$$

where J is a 9-dimensional Freudenthal–Jordan algebra associated to a pair (B, τ), where B is a central simple algebra over the quadratic algebra K_J and τ is an involution of the second kind on B. We now describe the elements of $\Omega_{E,K}$ in terms of such embeddings.

- When F is p-adic and $K = K_E$, so that $K_J = F \times F$ is split, then

$$(B, \tau) = (D \times D^{\mathrm{op}}, \mathrm{sw})$$

where D is a central simple F-algebra of degree 3 and sw denotes the involution which switches the two factors. Thus, there are two possible J in this case: the Jordan algebra J^+ attached to $M_3(F)$ or the Jordan algebra J^- attached to a cubic division F-algebra (and its opposite). In either case, the set of embeddings $E \longrightarrow J$ is either empty or a single conjugacy class, and it is empty if and only if $J = J^-$ and E is not a field. Thus when $K = K_E$, we have

$$\widetilde{\Omega}_{E,K} = \Omega_{E,K} = \begin{cases} \{E \to J^+, E \to J^-\} & \text{if } E \text{ is a field;} \\ \{E \to J^+\} & \text{if } E \text{ is not a field.} \end{cases}$$

On the other hand, when K_J is a field, then $B = M_3(K_J)$, and there is a unique isomorphism class of involution of the second kind on B, given by conjugation by a nondegenerate hermitian matrix, so that J is isomorphic to the Jordan algebra of 3×3-Hermitian matrices with entries in K_J. According to the proposition, there is a unique conjugacy class of embedding $E \hookrightarrow J$ unless $E = F \times K_E$ and K is a field with $K \neq K_E$. In the exceptional case, there are two subalgebras $E \subset J$ up to conjugacy. We may write down the 2 non-F-isomorphic twisted composition algebras corresponding to these. The twisted composition algebra can be realized on

$$E \otimes_F K = K \times (K_E \otimes K).$$

Let $\{1, \alpha\}$ denote representatives of $F^\times / NK^\times$. Then the two twisted composition algebras correspond to

$$(e, \nu) = ((1, 1), 1) \quad \text{or} \quad ((1, \alpha), \alpha) \in (F \times K_E)^\times \times K^\times.$$

We see that these two twisted composition algebras are not isomorphic because they are not isomorphic as quadratic spaces over E (even allowing for twisting by $S_E(F)$).

Further, when $E = F^3$, there are in fact four conjugacy classes of embeddings $E \hookrightarrow J$. This corresponds to the fact that the F-isomorphism class of the twisted composition algebras associated to $((1,\alpha),\alpha)$ above breaks into three E-isomorphism classes. These are associated to

$$(e_1,\nu_1) = ((1,\alpha,\alpha),\alpha), \quad (e_2,\nu_2) = ((\alpha,1,\alpha),\alpha), \quad (e_3,\nu_3) = ((\alpha,\alpha,1),\alpha).$$

• When $F = \mathbb{R}$, then $E = \mathbb{R}^3$ or $\mathbb{R} \times \mathbb{C}$. When $K_J = \mathbb{R}^2$ is split, then there is a unique J, namely the one associated to $M_3(\mathbb{R})$, and there is a unique conjugacy class of embeddings $E \hookrightarrow J$.

When $K_J = \mathbb{C}$, then there are two possible J, associated to $B = M_3(\mathbb{C})$ and the involution τ given by the conjugation action of two Hermitian matrices with signature $(1,2)$ and $(3,0)$. We denote these two Jordan algebras by $J_{1,2}$ and $J_{3,0}$.

When $E = \mathbb{R}^3$ and $K = \mathbb{C}$, we have $|\Omega_{E,K}| = 2$. However, the two elements in question correspond to embeddings

$$\mathbb{R}^3 \hookrightarrow J_{3,0} \quad \text{and} \quad \mathbb{R}^3 \hookrightarrow J_{1,2}.$$

Thus, we see that these subalgebras are unique up to conjugacy. When $E = \mathbb{R} \times \mathbb{C}$ and $K = \mathbb{R}^2$, we have $|\Omega_{E,K}| = 1$. This reflects the fact that there is no embedding $\mathbb{R} \times \mathbb{C} \hookrightarrow J_{3,0}$, and there is a unique conjugacy class of embeddings $\mathbb{C} \hookrightarrow J_{1,2}$.

Acknowledgment

This work began when both authors participated in the program “Branching Laws” at the Institute for Mathematical Sciences at the National University of Singapore in March 2012, and was completed during Savin’s stay at the Hong Kong University of Science and Technology in May 2013. Both authors thank Chengbo Zhu for his invitation to the IMS program. Savin would like to thank Jianshu Li for the invitation, and HKUST for excellent working environment.

Gan is partially supported by AcRF Tier One grant R-146-000-155-112, and Savin is supported by a National Science Foundation grant DMS-0852429.

References

[Bhargava 2004a] M. Bhargava, “Higher composition laws, I: A new view on Gauss composition, and quadratic generalizations”, *Ann. of Math.* (2) **159**:1 (2004), 217–250. MR 2005f:11062a Zbl 1072.11078

[Bhargava 2004b] M. Bhargava, “Higher composition laws, II: On cubic analogues of Gauss composition”, *Ann. of Math.* (2) **159**:2 (2004), 865–886. MR 2005f:11062b Zbl 1169.11044

[Bhargava 2004c] M. Bhargava, “Higher composition laws, III: The parametrization of quartic rings”, *Ann. of Math.* (2) **159**:3 (2004), 1329–1360. MR 2005k:11214 Zbl 1169.11045

[Gan et al. 2002] W. T. Gan, B. Gross, and G. Savin, “Fourier coefficients of modular forms on G_2”, *Duke Math. J.* **115**:1 (2002), 105–169. MR 2004a:11036 Zbl 1165.11315

[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**, Amer. Math. Soc., Providence, RI, 1998. MR 2000a:16031 Zbl 0955.16001

[Serre 2002] J.-P. Serre, *Galois cohomology*, Springer, Berlin, 2002. MR 2002i:12004 Zbl 1004.12003

[Springer and Veldkamp 2000] T. A. Springer and F. D. Veldkamp, *Octonions, Jordan algebras and exceptional groups*, Springer, Berlin, 2000. MR 2001f:17006 Zbl 1087.17001

Communicated by Efim Zelmanov
Received 2014-02-24 Revised 2014-07-23 Accepted 2014-07-26

matgwt@nus.edu.sg *Department of Mathematics, National University of Singapore,*
Block S17, 10 Lower Kent Ridge Road, Ridge Road,
Singapore 119076, Singapore

savin@math.utah.edu *Department of Mathematics, University of Utah,*
Salt Lake City, UT 84112, United States

ALGEBRA AND NUMBER THEORY 8:8 (2014)
dx.doi.org/10.2140/ant.2014.8.1959

Proper triangular $\mathbb{G}_a$-actions on $\mathbb{A}^4$ are translations

Adrien Dubouloz, David R. Finston and Imad Jaradat

We describe the structure of geometric quotients for proper locally triangulable $\mathbb{G}_a$-actions on locally trivial $\mathbb{A}^3$-bundles over a nœtherian normal base scheme X defined over a field of characteristic 0. In the case where $\dim X = 1$, we show in particular that every such action is a translation with geometric quotient isomorphic to the total space of a vector bundle of rank 2 over X. As a consequence, every proper triangulable $\mathbb{G}_a$-action on the affine four space $\mathbb{A}^4_k$ over a field of characteristic 0 is a translation with geometric quotient isomorphic to $\mathbb{A}^3_k$.

Introduction

The study of algebraic actions of the additive group $\mathbb{G}_a = \mathbb{G}_{a,\mathbb{C}}$ on complex affine spaces $\mathbb{A}^n = \mathbb{A}^n_{\mathbb{C}}$ has a long history which began in 1968 with a pioneering result of Rentschler [1968] who established that every such action on the plane $\mathbb{A}^2$ is triangular in a suitable polynomial coordinate system. Consequently, every fixed point free $\mathbb{G}_a$-action on $\mathbb{A}^2$ is a translation, in the sense that the geometric quotient $\mathbb{A}^2/\mathbb{G}_a$ is isomorphic to $\mathbb{A}^1$ and that $\mathbb{A}^2$ is equivariantly isomorphic to $\mathbb{A}^2/\mathbb{G}_a \times \mathbb{G}_a$ where $\mathbb{G}_a$ acts by translations on the second factor.

Arbitrary $\mathbb{G}_a$-actions turn out to be no longer triangulable in higher dimensions [Bass 1984]. But the question whether a fixed point free $\mathbb{G}_a$-action on $\mathbb{A}^3$ is a translation or not was settled affirmatively, first for triangulable actions in [Snow 1988], then in [Deveney and Finston 1994] under the additional assumption that the action is proper and then in general in [Kaliman 2004]. The argument for triangulable actions depends on their explicit form in an appropriate coordinate system which is used to check that the algebraic quotient $\pi : \mathbb{A}^3 \to \mathbb{A}^3/\!/\mathbb{G}_a = \operatorname{Spec}(\Gamma(\mathbb{A}^3, \mathcal{O}_{\mathbb{A}^3})^{\mathbb{G}_a})$ is a geometric quotient and that $\mathbb{A}^3/\!/\mathbb{G}_a$ is isomorphic to $\mathbb{A}^2$. For proper actions, the properness implies that the geometric quotient $\mathbb{A}^3/\mathbb{G}_a$,

Research supported in part by NSF Grant OISE-0936691 and ANR Grant "BirPol" ANR-11-JS01-004-01.

MSC2010: primary 14L30; secondary 14R10, 14R20, 14R25.

Keywords: proper additive group actions, geometric quotients, principal homogeneous bundles, affine fibrations.

which a priori only exists as an algebraic space, is separated whence a scheme by virtue of Chow's Lemma. This means equivalently that the $\mathbb{G}_a$-action is not only locally equivariantly trivial in the étale topology but in fact locally trivial in the Zariski topology, that is, that $\mathbb{A}^3$ is covered by invariant Zariski affine open subsets of the form $V_i = U_i \times \mathbb{G}_a$ on which $\mathbb{G}_a$ acts by translations on the second factor. Since $\mathbb{A}^3$ is factorial, the open subsets V_i can even be chosen to be principal, which implies in turn that $\mathbb{A}^3/\mathbb{G}_a$ is a quasiaffine scheme, in fact an open subset of $\mathbb{A}^3/\!/\mathbb{G}_a \simeq \mathbb{A}^2$ with at most finite complement. The equality $\mathbb{A}^3/\mathbb{G}_a = \mathbb{A}^3/\!/\mathbb{G}_a$ ultimately follows by comparing Euler characteristics. Kaliman's general proof proceeds along a completely different approach, drawing on topological arguments to show directly that the algebraic quotient morphism $\pi : \mathbb{A}^3 \to \mathbb{A}^3/\!/\mathbb{G}_a$ is a locally trivial $\mathbb{A}^1$-bundle. Similar topological methods have been also applied by Kaliman and Saveliev [2004] to conclude more generally that every fixed point free $\mathbb{G}_a$-action on a smooth complex contractible affine threefold X is a translation in the broader sense that X has the structure of a trivial $\mathbb{G}_a$-bundle over its geometric quotient $X/\mathbb{G}_a$, which is a smooth contractible affine surface.

Kaliman's result can be reinterpreted as the striking fact that the topological contractiblity of $\mathbb{A}^3$ is a strong enough constraint to guarantee that a fixed point free $\mathbb{G}_a$-action on it is automatically proper. This implication fails completely in higher dimensions where nonproper fixed point free $\mathbb{G}_a$-actions abound, even in the case of triangular actions on $\mathbb{A}^4$ as illustrated by Deveney, Finston and Gehrke [Deveney et al. 1994]. And starting from dimension 5, properness is known to be no longer enough to imply global equivariant triviality as illustrated by examples of proper triangular actions on $\mathbb{A}^5$ with strictly quasiaffine geometric quotients constructed by Winkelmann [1990].

On the other hand, a general characterization claimed by Fauntleroy and Magid [1976] asserted that proper $\mathbb{G}_a$-actions on factorial affine varieties were always locally equivariantly trivial in the Zariski topology, with quasiaffine geometric quotients. But counterexamples were constructed latter on by Deveney and Finston [1995] in the form of proper triangular actions on $\mathbb{A}^5$ whose geometric quotients exists only as separated algebraic spaces. So the question whether a proper $\mathbb{G}_a$-action on $\mathbb{A}^4$ is a translation or is at least locally equivariantly trivial in the Zariski topology is essentially the last unsettled problem concerning proper $\mathbb{G}_a$-actions on affine spaces, and very little progress had been made on the subject during the last decades.

The only existing partial results so far concern triangular actions: Deveney, van Rossum and Finston [2004] established that a Zariski locally equivariantly trivial triangular $\mathbb{G}_a$-action on $\mathbb{A}^4$ is a translation. The proof depends on the finite generation of the ring of invariants for such actions established by Daigle and Freudenburg [2001] and exploits the very particular structure of these rings.

Incidentally, it is known in general that local triviality for a proper action on $\mathbb{A}^n$ follows from the finite generation and regularity of the ring of invariants. But even knowing the former for triangular actions on $\mathbb{A}^4$, a direct proof of the latter condition remains elusive. The second positive result concerns a special type of triangular $\mathbb{G}_a$-actions generated by derivations of $\mathbb{C}[x, y, z, u]$ of the form $r(x)\partial_y + q(x, y)\partial_z + p(x, y)\partial_u$ where $r(x) \in \mathbb{C}[x]$ and $p(x, y), q(x, y) \in \mathbb{C}[x, y,]$. To insist on the fact that $p(x, y)$ belongs to $\mathbb{C}[x, y]$ and not only to $\mathbb{C}[x, y, z]$ as it would be the case for a general triangular situation, these derivations (and the $\mathbb{G}_a$-actions they generate) were named *twin-triangular* in [Deveney and Finston 2000]. The case where $r(x)$ has simple roots was first settled there by explicitly computing the invariant ring $\mathbb{C}[x, y, z, u]^{\mathbb{G}_a}$ and investigating the structure of the algebraic quotient morphism $\mathbb{A}^4 \to \mathbb{A}^4/\!/\mathbb{G}_a = \mathrm{Spec}(\mathbb{C}[x, y, z_1, z_2]^{\mathbb{G}_a})$. The simplicity of the roots of $r(x)$ was crucial to achieve the computation, and the generalization of the result to arbitrary twin-triangular actions obtained in 2012 by the first two authors [Dubouloz and Finston 2014] required completely different methods which focused more on the nature of the corresponding geometric quotients $\mathbb{A}^4_{\mathbb{C}}/\mathbb{G}_a$. The latter a priori exist only as separated algebraic spaces and the crucial step in *loc. cit.* was to show that for twin-triangular actions they are in fact schemes, or, equivalently that proper twin-triangular $\mathbb{G}_a$-actions on $\mathbb{A}^4$ are not only locally equivariantly trivial in the étale topology but also in the Zariski topology. This enabled in turn the use of the aforementioned result of Deveney, Finston, and van Rossum to conclude that such actions are indeed translations.

In this article, we reconsider proper triangular actions on $\mathbb{A}^4$ in broader framework and we develop new techniques which enable to completely solve the question of global equivariant triviality for such actions. The triangularity assumption is of course a restriction, and it might look quite artificial from a geometric point of view. But its main consequence is to reduce an a priori four-dimensional problem to a relative three-dimensional one over a parameter space, a reduction which is crucial for our argument and turns out to be the natural context in which to interpret the aforementioned counterexamples to global or Zariski local equivariant triviality. A second more technical benefit is that it enables an effective characterization of the properness of a $\mathbb{G}_a$-action in terms of its associated locally nilpotent derivation, a problem which is in general much more delicate to handle than deciding the weaker property of being fixed point free.

The existence of smooth factorial affine hypersurfaces of $\mathbb{A}^5$ on which the proper triangular $\mathbb{G}_a$-actions constructed by Deveney and Finston [1995] restrict to proper actions whose geometric quotients exist only as separated algebraic spaces shows that even under appropriate triangularity assumptions, the question whether a proper $\mathbb{G}_a$-action on $\mathbb{A}^4$ is Zariski locally equivariantly trivial remains a subtle problem. It also indicates that in order to weaken these appropriate hypotheses, additional algebrogeometric properties of $\mathbb{A}^4$ beyond factoriality, such as for instance topological contractibility, should play a role in the problem. But on the other hand, the existence of smooth contractible complex affine threefolds nonisomorphic to $\mathbb{A}^3$ shows that topological methods are not sufficient to infer that a given proper $\mathbb{G}_a$-action on $\mathbb{A}^4$ is a translation from its local or even global equivariant triviality. In particular, knowing that every such action is a translation would solve the Zariski Cancellation Problem in dimension three, for if X is a variety such that $X \times \mathbb{A}^1 \simeq \mathbb{A}^4$, the $\mathbb{G}_a$-action by translations on the second factor of $X \times \mathbb{A}^1$ is obviously proper.

In this article we embed the study of proper triangular $\mathbb{G}_a$-actions on $\mathbb{A}^4$ into the following more general setup: given a nœtherian normal scheme X defined over a field of characteristic zero, we consider Zariski locally trivial $\mathbb{A}^3$-bundles $\pi : E \to X$ equipped with proper locally triangulable actions of the additive group scheme $\mathbb{G}_{a,X}$. The local triangularity assumption means roughly that X can be covered by affine open subsets $U = \operatorname{Spec}(A)$ over which the restriction of E is equivariantly isomorphic to $\mathbb{A}^3_U = \operatorname{Spec}(A[y, z, u])$ equipped with the $\mathbb{G}_{a,U}$-action induced by a triangular A-derivation of $A[y, z, u]$. Our main result then is this:

Theorem. *Let X be a nœtherian normal scheme defined over a field of characteristic zero, let $\pi : E \to X$ be a Zariski locally trivial $\mathbb{A}^3$-bundle equipped with a proper locally triangulable $\mathbb{G}_{a,X}$-action and let $\mathrm{p} : \mathfrak{X} = E/\mathbb{G}_{a,X} \to X$ be the geometric quotient taken in the category of algebraic X-spaces. Then there exists an open subscheme U of X with $\operatorname{codim}_X(X \setminus U) \geq 2$ such that $\mathfrak{X}_U = \mathrm{p}^{-1}(U) \to U$ has the structure of a Zariski locally trivial $\mathbb{A}^2$-bundle.*

The conclusion of this theorem is essentially optimal. Indeed, in the example due to Winkelmann [1990], one has $X = \operatorname{Spec}(\mathbb{C}[x, y])$, $\pi = \mathrm{pr}_{x,y} : \mathbb{A}^3_X = \operatorname{Spec}(\mathbb{C}[x, y][u, v, w]) \to X$ equipped with the proper triangular $\mathbb{G}_{a,X}$-action generated by the $\mathbb{C}[x, y]$-derivation $\partial = x\partial_u + y\partial_v + (1+xv-yu)\partial_w$ of $\mathbb{C}[x, y][u, v, w]$, and the geometric quotient $\mathrm{p} : \mathfrak{X} = \mathbb{A}^3_X/\mathbb{G}_{a,X} \to X$ is the strictly quasiaffine complement of the closed subset $\{x = y = z = 0\}$ in the 4-dimensional smooth affine quadric $Q \subset \mathbb{A}^3_X$ with equation $xt_2 + yt_1 = z(z + 1)$. The structure morphism $\mathrm{p} : \mathfrak{X} \to X$ is easily seen to be an $\mathbb{A}^2$-fibration, which restricts to a locally trivial $\mathbb{A}^2$-bundle over the open subset $U = X \setminus \{(0, 0)\}$. However, there is no Zariski or étale open neighborhood of the origin $(0, 0) \in X$ over which $\mathrm{p} : \mathfrak{X} \to X$ restricts to a trivial $\mathbb{A}^2$-bundle for otherwise $\mathrm{p} : \mathfrak{X} \to X$ would be an affine morphism and so $\mathfrak{X}$ would be an affine scheme. The situation for the $\mathbb{C}[x, y]$-derivation $\partial = x\partial_u + y\partial_v + (1+xv^2)\partial_w$ of $\mathbb{C}[x, y][u, v, w]$ constructed by Deveney and Finston [1995] is very similar: here the geometric quotient $\mathfrak{X} = \mathbb{A}^3_X/\mathbb{G}_{a,X}$ is a separated algebraic space which is not a scheme and the structure morphism $\mathrm{p} : \mathfrak{X} \to X$ is again an $\mathbb{A}^2$-fibration restricting to a Zariski locally trivial $\mathbb{A}^2$-bundle over $U = X \setminus \{(0, 0)\}$ but whose restriction to any Zariski or étale open neighborhood of the origin $(0, 0) \in X$ is nontrivial.

In contrast, in the case of a 1-dimensional affine base, we can immediately derive the following corollaries:

Corollary. *Let* $\pi : E \to S$ *be a rank* 3 *vector bundle over an affine Dedekind scheme* $S = \mathrm{Spec}(A)$ *defined over a field k of characteristic* 0. *Then every proper locally triangulable* $\mathbb{G}_{a,S}$*-action on* E *is equivariantly trivial with geometric quotient* $E/\mathbb{G}_{a,S}$ *isomorphic to a vector bundle of rank* 2 *over* S, *stably isomorphic to* E.

Proof. By the previous theorem, the geometric quotient $\mathrm{p} : E/\mathbb{G}_{a,S} \to S$ has the structure of a Zariski locally trivial $\mathbb{A}^2$-bundle, hence is a vector bundle of rank 2 by [Bass et al. 1977]. In particular, $E/\mathbb{G}_{a,S}$ is affine, which implies in turn that $\rho : E \to E/\mathbb{G}_{a,S}$ is a trivial $\mathbb{G}_{a,S}$-bundle. So E is isomorphic to $E/\mathbb{G}_{a,S} \times_S \mathbb{A}^1_S$ as vector bundles over S. □

Corollary. *Let* $S = \mathrm{Spec}(A)$ *be an affine Dedekind scheme defined over a field of characteristic* 0. *Then every proper triangular* $\mathbb{G}_{a,S}$*-action on* $\mathbb{A}^3_S$ *is a translation.*

Proof. By the previous corollary, $\mathbb{A}^3_S/\mathbb{G}_{a,S}$ is a stably trivial vector bundle of rank 2 over S, whence is isomorphic to the trivial bundle $\mathbb{A}^2_S$ over S by virtue of [Bass 1968, Chapter IV, Corollary 3.5]. □

Coming back to the question of proper triangular $\mathbb{G}_a$-actions on $\mathbb{A}^4$, the observation that such actions preserve a variable in a appropriate coordinate system and hence can be considered as proper triangular actions of the additive group scheme $\mathbb{G}_{a,S}$ on the affine 3-space $\mathbb{A}^3_S$ over the affine Dedekind base $S = \mathbb{A}^1$ suffices to settle the problem:

Corollary. *If k is a field of characteristic* 0, *then every proper triangular* $\mathbb{G}_{a,k}$*-action on* $\mathbb{A}^4_k$ *is a translation.*

It is worth mentioning that our Main Theorem and an appeal to the aforementioned result [Deveney et al. 2004] would already be enough to conclude that every proper triangular $\mathbb{G}_{a,k}$-action on $\mathbb{A}^4_k$ is a translation, but our results do actually eliminate the need for *loc. cit.* hence the a priori dependency on the fact that the corresponding rings of invariants are finitely generated.

Let us now briefly explain the general philosophy behind the proof. After localizing at codimension 1 points of X, the Main Theorem reduces to the statement that a proper $\mathbb{G}_{a,S}$-action $\sigma : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S \to \mathbb{A}^3_S$ on the affine space $\mathbb{A}^3_S = \mathrm{Spec}(A[y, z, u])$ over the spectrum of a discrete valuation ring, generated by a triangular A-derivation $\partial = a\partial_y + q(y)\partial_z + p(y, z)\partial_u$ of $A[y, z, u]$, where $a \in A \setminus \{0\}$, $q(y) \in A[y]$ and $p(y, z) \in A[y, z]$, is a translation. Triangularity immediately implies that the restriction of σ to the generic fiber of $\mathrm{pr}_S : \mathbb{A}^3_S \to S$ is a translation with $a^{-1}y$ as a global slice. This reduces the problem to the study of neighborhoods of points of the geometric quotient $\mathfrak{X} = \mathbb{A}^3_S/\mathbb{G}_{a,S}$ supported on the closed fiber of the structure morphism $\mathrm{p} : \mathfrak{X} \to S$. A second feature of triangularity is that σ commutes with the action $\tau:\mathbb{G}_{a,S}\times_S\mathbb{A}^3_S\to\mathbb{A}^3_S$ generated by the A-derivation ∂_u which therefore descends to a $\mathbb{G}_{a,S}$-action $\bar{\tau}$ on the geometric quotient $\mathfrak{X}=\mathbb{A}^3_S/\mathbb{G}_{a,S}$. On the other hand, σ descends via the projection $\mathrm{pr}_{y,z}:\mathbb{A}^3_S\to\mathbb{A}^2_S=\mathrm{Spec}(A[y,z])$ to the action $\bar{\sigma}$ on $\mathbb{A}^2_S$ generated by the A-derivation $\bar{\partial}=a\partial_y+q(y)\partial_z$ of $A[y,z]$. Even though $\bar{\sigma}$ and $\bar{\tau}$ are no longer fixed point free in general, if we take the quotient of $\mathbb{A}^2_S$ by the action $\bar{\sigma}$ as an algebraic stack $[\mathbb{A}^2_S/\mathbb{G}_{a,S}]$ we obtain a cartesian square

$$\begin{tikzcd}
\mathbb{A}^3_S \arrow[r] \arrow[d, "\mathrm{pr}_{y,z}"'] & \mathfrak{X}=\mathbb{A}^3_S/\mathbb{G}_{a,S} \arrow[d] \\
\mathbb{A}^2_S \arrow[r] & {[\mathbb{A}^2_S/\mathbb{G}_{a,S}]}
\end{tikzcd}$$

which simultaneously identifies the quotient stacks $[\mathbb{A}^2_S/\mathbb{G}_{a,S}]$ for the action $\bar{\sigma}$ and $[\mathfrak{X}/\mathbb{G}_{a,S}]$ for the action $\bar{\tau}$ with the quotient stack of $\mathbb{A}^3_S$ for the $\mathbb{G}^2_{a,S}$-action defined by the commuting actions σ and τ. In this setting, the global equivariant triviality of the action σ becomes equivalent to the statement that a separated algebraic S-space $\mathfrak{X}$ admitting a $\mathbb{G}_{a,S}$-action whose algebraic stack quotient $[\mathfrak{X}/\mathbb{G}_{a,S}]$ is isomorphic to that of a triangular $\mathbb{G}_{a,S}$-action on $\mathbb{A}^2_S$ is an affine scheme.

While a direct proof of this reformulation seems totally out of reach with existing methods, it turns out that its conclusion holds over a certain $\mathbb{G}_{a,S}$-invariant principal open subset V of $\mathbb{A}^2_S$ which dominates S and for which the algebraic stack quotient $[V/\mathbb{G}_{a,S}]$ is in fact represented by a locally separated algebraic subspace of $[\mathbb{A}^2_S/\mathbb{G}_{a,S}]$. This provides at least an affine open subscheme $V\times_S\mathbb{A}^1_S/\mathbb{G}_{a,S}$ of $\mathfrak{X}$ dominating S, and leaves us with a closed subset of codimension at most 2 of $\mathfrak{X}$, supported on the closed fiber of $\mathrm{p}:\mathfrak{X}\to S$, in a neighborhood of which no further information is a priori available to decide even the schemeness of $\mathfrak{X}$. But similar to the argument in [Dubouloz and Finston 2014], this situation can be rescued for twin-triangular actions: the fact that for such actions $\partial u=p(y,z)$ is actually a polynomial in y only enables the same reasoning with respect to the other projection $\mathrm{pr}_{y,u}:\mathbb{A}^3_S\to\mathbb{A}^2_S=\mathrm{Spec}(A[y,u])$, yielding a second affine open subscheme $V'\times_S\mathbb{A}^1_S/\mathbb{G}_{a,S}$ of $\mathfrak{X}$ dominating S. This implies at least the schemeness of $\mathfrak{X}$, provided that the open subsets V and V' can be chosen so that the union of the corresponding open subschemes of $\mathfrak{X}$ covers the closed fiber of $\mathrm{p}:\mathfrak{X}\to S$.

The scheme of the article is the following. The first two sections recall basic notions and discuss a couple of preliminary technical reductions. The third section is devoted to establishing an effective criterion for nonproperness of fixed point free triangular actions from which we deduce the intermediate fact that every proper triangular action is twin-triangulable. Then in the next section, we establish that proper twin-triangular actions are indeed translations. Here, in contrast with the proof for the complex case given in [Dubouloz and Finston 2014], our argument is independent of finite generation of rings of invariants and reduces the systematic study of algebraic spaces quotients to a minimum thanks to an appropriate "Sheshadri cover trick" [Seshadri 1972].

1. Recollection on proper, fixed point free and locally triangulable $\mathbb{G}_a$-actions

1A. ***Proper versus fixed point free actions.*** Recall that an action $\sigma : \mathbb{G}_{a,S} \times_S E \to E$ of the additive group scheme $\mathbb{G}_{a,S} = \mathrm{Spec}_S(\mathbb{O}_S[t]) = S \times_{\mathbb{Z}} \mathrm{Spec}(\mathbb{Z}[t])$ on an S-scheme E is called proper if the morphism $\Phi = (\mathrm{pr}_2, \sigma) : \mathbb{G}_{a,S} \times_S E \to E \times_S E$ is proper.

1A1. If S is moreover defined over a field k of characteristic zero, then the fact that $\mathbb{G}_{a,k}$ is affine and has no nontrivial algebraic subgroups implies that properness is equivalent to Φ being a closed immersion. In particular, a proper $\mathbb{G}_{a,S}$-action is in this case fixed point free and as such, is equivariantly locally trivial in the étale topology on E. That is, there exists an affine S-scheme U and a surjective étale morphism $f : V = U \times_S \mathbb{G}_{a,S} \to E$ which is equivariant for the action of $\mathbb{G}_{a,S}$ on $U \times_S \mathbb{G}_{a,S}$ by translations on the second factor. This implies in turn the existence of a geometric quotient $\rho : E \to \mathfrak{X} = E/\mathbb{G}_{a,S}$ in the form of an étale locally trivial principal $\mathbb{G}_{a,S}$-bundle over an algebraic S-space $\mathrm{p} : \mathfrak{X} \to S$ (see, for example, [Laumon and Moret-Bailly 2000, Corollary 10.4]). Informally, $\mathfrak{X}$ is the quotient of U by the étale equivalence relation which identifies two points $u, u' \in U$ whenever there exists $t, t' \in \mathbb{G}_{a,S}$ such that $f(u,t) = f(u',t')$.

1A2. Conversely, a fixed point free $\mathbb{G}_{a,S}$-action is proper if and only if the geometric quotient $\mathfrak{X} = E/\mathbb{G}_{a,S}$ is a separated S-space. Indeed, by definition $\mathrm{p} : \mathfrak{X} \to S$ is separated if and only if the diagonal morphism $\Delta : \mathfrak{X} \to \mathfrak{X} \times_S \mathfrak{X}$ is a closed immersion, a property which is local on the target with respect to the fpqc topology [Knutson 1971, II, Extension 3.8; SGA1 1971, VIII, Corollaire 5.5]. Since $\rho : E \to \mathfrak{X}$ is a $\mathbb{G}_{a,S}$-bundle, taking the fpqc base change by $\rho \times \rho : E \times_S E \to \mathfrak{X} \times_S \mathfrak{X}$ yields a cartesian square

$$\begin{tikzcd}
\mathbb{G}_{a,S} \times_S E \arrow[r, "\Phi"] \arrow[d, "\rho\circ\mathrm{pr}_2"'] & E \times_S E \arrow[d, "\rho\times\rho"] \\
\mathfrak{X} \arrow[r, "\Delta"] & \mathfrak{X} \times_S \mathfrak{X}
\end{tikzcd}$$

from which we see that Δ is a closed immersion if and only if Φ is.

1B. ***Locally triangulable actions.*** Given an affine scheme $S = \mathrm{Spec}(A)$ defined over a field of characteristic zero, an action $\sigma : \mathbb{G}_{a,S} \times_S \mathbb{A}^n_S \to \mathbb{A}^n_S$ generated by a locally nilpotent A-derivation ∂ of $\Gamma(\mathbb{A}^n_S, \mathbb{O}_{\mathbb{A}^n_S})$ is called *triangulable* if there exists an isomorphism of A-algebras $\tau : \Gamma(\mathbb{A}^n_A, \mathbb{O}_{\mathbb{A}^n_A}) \xrightarrow{\sim} A[x_1, \dots, x_n]$ such that the conjugate $\delta = \tau \circ \partial \circ \tau^{-1}$ of ∂ is triangular with respect to the ordered coordinate system $(x_1, \dots, x_n)$, that is, has the form

$$\delta = p_0 \frac{\partial}{\partial x_1} + \sum_{i=1}^{n} p_{i-1}(x_1, \dots, x_{i-1}) \frac{\partial}{\partial x_i}$$

where $p_0 \in A$ and for every $i = 1, \dots, n$, $p_{i-1}(x_1, \dots, x_{i-1}) \in A[x_1, \dots, x_{i-1}] \subset A[x_1, \dots, x_n]$. By localizing this notion over the base S, we arrive at the following definition:

Definition 1.1. Let X be a scheme defined over a field of characteristic zero and let $\pi : E \to X$ be a Zariski locally trivial $\mathbb{A}^n$-bundle over X. An action $\sigma : \mathbb{G}_{a,X} \times_X E \to E$ of $\mathbb{G}_{a,X}$ on E is called *locally triangulable* if there exists a covering of $\operatorname{Spec}(A)$ by affine open subschemes $S_i = \operatorname{Spec}(A_i)$, $i \in I$, such that $E|_{S_i}$ is isomorphic to $\mathbb{A}^n_{S_i}$ and such that the $\mathbb{G}_{a,S_i}$-action $\sigma_i : \mathbb{G}_{a,S_i} \times_{S_i} \mathbb{A}^n_{S_i} \to \mathbb{A}^n_{S_i}$ on $\mathbb{A}^n_{S_i}$ induced by σ is triangulable.

A Zariski locally trivial $\mathbb{A}^1$-bundle $\pi : E \to X$ equipped with a fixed point free $\mathbb{G}_{a,X}$-action is nothing but a principal $\mathbb{G}_{a,X}$-bundle. As mentioned in the introduction, the nature of fixed point free locally triangulable $\mathbb{G}_{a,X}$-actions on Zariski locally trivial $\mathbb{A}^2$-bundles $\pi : E \to X$ is classically known. Namely, we have the following generalization of the main theorem of [Snow 1988]:

Proposition 1.2. *Let X be a nœtherian normal scheme defined over a field of characteristic* 0 *and let $\pi : E \to X$ be a Zariski locally trivial $\mathbb{A}^2$-bundle equipped with a fixed point free locally triangulable $\mathbb{G}_{a,X}$-action. Then the geometric quotient* $\mathrm{p} : E/\mathbb{G}_{a,X} \to X$ *has the structure of a Zariski locally trivial $\mathbb{A}^1$-bundle over X.*

Proof. The assertion being local on the base X, we may assume that $X = \operatorname{Spec}(A)$ is the spectrum of a normal local domain containing a field of characteristic 0 and that $E = \mathbb{A}^2_X = \operatorname{Spec}(A[y, z])$ is equipped with the $\mathbb{G}_{a,X}$-action generated by a triangular derivation $\partial = a\partial_y + q(y)\partial_z$ of $A[y, z]$, where $a \in A$ and $q(y) \in A[y]$. The fixed point freeness hypothesis is equivalent to the property that a and $q(y)$ generate the unit ideal in $A[y, z]$. So $q(y)$ has the form $q(y) = b + c\tilde{q}(y)$ where $b \in A$ is relatively prime with a, $c \in \sqrt{aA}$ and $\tilde{q}(y) \in A[y]$. Letting $Q(y) = \int_0^y q(\tau)\, d\tau = by + c \int_0^y \tilde{q}(\tau)\, d\tau$, the polynomial $v = az - Q(y) \in A[y, z]$ belongs to the kernel $\operatorname{Ker} \partial$ of ∂ hence defines a $\mathbb{G}_{a,X}$-invariant morphism $v : E \to \mathbb{A}^1_X = \operatorname{Spec}(A[t])$. Since a and b generate the unit ideal in A, it follows from the Jacobian criterion that $v : E \to \mathbb{A}^1_X$ is a smooth morphism. Furthermore, the fibers of v coincide precisely with the $\mathbb{G}_{a,X}$-orbits on E. Indeed, over the principal open subset $X_a = \operatorname{Spec}(A_a)$ of X, ∂ admits $a^{-1}y$ as a slice and we have an equivariant isomorphism $E|_{X_a} \simeq \operatorname{Spec}(A[a^{-1}v, a^{-1}y]) \simeq \mathbb{A}^1_{X_a} \times_X \mathbb{G}_{a,X}$ where $\mathbb{G}_{a,X}$ acts by translations on the second factor. On the other hand, the restriction $E|_Z$ of E over the closed subset $Z \subset X$ with defining ideal $\sqrt{aA} \subset A$ is equivariantly isomorphic to $\mathbb{A}^2_Z$ equipped with the $\mathbb{G}_{a,Z}$-action generated by the derivation $\bar{\partial} = \bar{b}\partial_z$ of $(A/\sqrt{aA})[y,z]$, where $\bar{b} \in (A/\sqrt{aA})^*$ denotes the residue class of b. The restriction of v to $E|_Z$ coincides via this isomorphism to the morphism $\mathbb{A}^2_Z \to \mathbb{A}^1_Z$ defined by the polynomial $\bar{v} = \bar{b}y \in (A/\sqrt{aA})[y,z]$ which is obviously a geometric quotient. The above properties imply that the morphism $\tilde{v} : E/\mathbb{G}_{a,X} \to \mathbb{A}^1_X$ induced by v is smooth and bijective. Since it admits étale quasisections, $\tilde{v}$ is then an isomorphism locally in the étale topology on $\mathbb{A}^1_X$ whence an isomorphism. □

2. Preliminary reductions

2A. ***Reduction to a local base.*** The statement of the Main Theorem can be rephrased equivalently as the fact that a proper locally triangulable $\mathbb{G}_{a,S}$-action on a Zariski locally trivial $\mathbb{A}^3$-bundle $\pi : E \to S$ is a translation in codimension 1. This means that for every point $s \in S$ of codimension 1 with local ring $\mathcal{O}_{S,s}$, the fiber product $E \times_S S' \simeq \mathbb{A}^3_{S'}$ of $E \to S$ with the canonical immersion $S' = \operatorname{Spec}(\mathcal{O}_{S,s}) \hookrightarrow S$ equipped with the induced proper triangular action of $\mathbb{G}_{a,S'} = \mathbb{G}_{a,S} \times_S S'$ is equivariantly isomorphic to the trivial bundle $\mathbb{A}^2_{S'} \times_{S'} \mathbb{G}_{a,S'}$ over S' equipped with the action of $\mathbb{G}_{a,S'}$ by translations on the second factor.

2A1. So we are reduced to the case where S is the spectrum of a discrete valuation ring A containing a field of characteristic 0, say with maximal ideal $\mathfrak{m}$ and residue field $\kappa = A/\mathfrak{m}$, and where $\pi = \mathrm{pr}_S : E = \mathbb{A}^3_S = \operatorname{Spec}(A[y,z,u]) \to S = \operatorname{Spec}(A)$ is equipped with a proper triangulable $\mathbb{G}_{a,S}$-action $\sigma : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S \to \mathbb{A}^3_S$. Letting $x \in \mathfrak{m}$ be uniformizing parameter, every such action is equivalent to one generated by an A-derivation ∂ of $A[y,z,u]$ of the form

$$\partial = x^n\partial_y + q(y)\partial_z + p(y,z)\partial_u$$

where $n \geq 0$, $q(y) \in A[y]$ and $p(y,z) = \sum_{r=0}^{\ell} p_r(y)z^r \in A[y,z]$, the fixed point freeness of σ being equivalent to the property that x^n, $q(y)$ and $p(y,z)$ generate the unit ideal in $A[y,z,u]$.

2B. ***Reduction to proving the affineness of the geometric quotient.*** With the notation of Section 2A1, we can already observe that if $n = 0$ then y is an obvious global slice for ∂ and hence that the action is globally equivariantly trivial with geometric quotient $\mathfrak{X} = \mathbb{A}^3_S/\mathbb{G}_{a,S} \simeq \mathbb{A}^2_S$. Similarly, if the residue class of $q(y)$ in $\kappa[y]$ is a nonzero constant then the action σ is a translation. Indeed, in this case, the $\mathbb{G}_{a,S}$-action $\bar{\sigma} : \mathbb{G}_{a,S} \times_S \mathbb{A}^2_S \to \mathbb{A}^2_S$ on $\mathbb{A}^2_S = \operatorname{Spec}(A[y,z])$ generated by the A-derivation $\bar{\partial} = x^n\partial_y + q(y)\partial_z$ of $A[y,z]$ is fixed point free hence globally equivariantly trivial with geometric quotient $\mathbb{A}^2_S/\mathbb{G}_{a,S} \simeq \mathbb{A}^1_S$ by virtue of Proposition 1.2. On the other hand, the $\mathbb{G}_{a,S}$-equivariant projection $\mathrm{pr}_{y,z} : \mathbb{A}^3_S \to \mathbb{A}^2_S$ descends to a locally trivial $\mathbb{A}^1$-bundle between the geometric quotients $\mathbb{A}^3_S/\mathbb{G}_{a,S}$ and $\mathbb{A}^2_S/\mathbb{G}_{a,S}$, and since $\mathbb{A}^2_S/\mathbb{G}_{a,S} \simeq \mathbb{A}^1_S$ is affine and factorial, it follows that $\mathbb{A}^3_S/\mathbb{G}_{a,S} \simeq \mathbb{A}^2_S/\mathbb{G}_{a,S} \times_S \mathbb{A}^1_S \simeq \mathbb{A}^2_S$. The affineness of $\mathbb{A}^2_S$ implies in turn that the quotient morphism $\mathbb{A}^3_S \to \mathbb{A}^3_S/\mathbb{G}_{a,S}$ is the trivial $\mathbb{G}_{a,S}$-bundle whence that $\sigma : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S \to \mathbb{A}^3_S$ is a translation. Alternatively, one can observe that a global slice $s \in A[y, z]$ for the action $\bar{\sigma}$ is also a global slice for σ via the inclusion $A[y, z] \subset A[y, z, u]$

More generally, the following lemma reduces the question of global equivariant triviality with geometric quotient $\mathfrak{X} = \mathbb{A}^3_S/\mathbb{G}_{a,S}$ isomorphic to $\mathbb{A}^2_S$ to showing that $\mathfrak{X}$, which a priori only exists as an algebraic S-space, is an affine S-scheme:

Lemma 2.1. *A fixed point free triangular action $\sigma : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S \to \mathbb{A}^3_S$ is a translation if and only if its geometric quotient $\mathfrak{X} = \mathbb{A}^3_S/\mathbb{G}_{a,S}$ is an affine S-scheme.*

Proof. One direction is clear, so assume that $\mathfrak{X}$ is an affine S-scheme. It suffices to show that the structure morphism $\mathrm{p} : \mathfrak{X} \to S$ is an $\mathbb{A}^2$-fibration, that is, a faithfully flat morphism with all its fibers isomorphic to affine planes over the corresponding residue fields. Indeed, if so, the affineness of $\mathfrak{X}$ implies on the one hand that $\mathfrak{X}$ is isomorphic to the trivial $\mathbb{A}^2$-bundle $\mathbb{A}^2_S$ by virtue of [Sathaye 1983] and on the other hand that $\rho : \mathbb{A}^3_S \to \mathfrak{X}$ is isomorphic to the trivial $\mathbb{G}_{a,S}$-bundle $\mathfrak{X} \times_S \mathbb{G}_{a,S}$ over S, which yields $\mathbb{G}_{a,S}$-equivariant isomorphisms $\mathbb{A}^3_S \simeq \mathfrak{X} \times_S \mathbb{G}_{a,S} \simeq \mathbb{A}^2_S \times_S \mathbb{G}_{a,S}$.

To see that $\mathrm{p} : \mathfrak{X} \to S$ is an $\mathbb{A}^2$-fibration, recall that $\mathrm{pr}_S : \mathbb{A}^3_S \to S$ and the quotient morphism $\rho : \mathbb{A}^3_S \to \mathfrak{X} = \mathbb{A}^3_S/\mathbb{G}_{a,S}$ are both faithfully flat, so that $\mathrm{p} : \mathfrak{X} \to S$ is faithfully flat too [Knutson 1971, II.3.2; EGA 1965, IV$_2$, Corollaire 2.2.13(iii)]. Letting $\mathfrak{m}$ and ξ be the closed and generic points of S respectively, the fibers $\mathrm{pr}_S^{-1}(\mathfrak{m}) \simeq \mathbb{A}^3_\kappa$ and $\mathrm{pr}_S^{-1}(\xi) \simeq \mathbb{A}^3_{\kappa(\xi)}$ coincide with the total spaces of the restriction of the $\mathbb{G}_{a,S}$-bundle $\rho : \mathbb{A}^3_S \to \mathfrak{X}$ over the fibers $\mathfrak{X}_\mathfrak{m} = \mathrm{p}^{-1}(\mathfrak{m})$ and $\mathfrak{X}_\xi = \mathrm{p}^{-1}(\xi)$ respectively. Since the $\mathbb{G}_{a,\kappa(\xi)}$-action induced by σ on $\mathrm{pr}_S^{-1}(\xi)$ admits $x^{-n}y$ as a global slice, it is a translation with geometric quotient $\mathbb{A}^3_{\kappa(\xi)}/\mathbb{G}_{a,\kappa(\xi)} \simeq \mathbb{A}^2_{\kappa(\xi)}$ and so $\mathfrak{X}_\xi \simeq \mathbb{A}^2_{\kappa(\xi)}$. On the other hand, we may assume in view of the above discussion that $n \geq 1$ so that the $\mathbb{G}_{a,\kappa}$-action on $\mathrm{pr}_S^{-1}(\mathfrak{m}) \simeq \mathbb{A}^3_\kappa$ induced by σ coincides with the fixed point free action generated by the $\kappa[y]$-derivation $\bar{\partial} = \bar{q}(y)\partial_z + \bar{p}(y, z)\partial_u$ of $\kappa[y][z, u]$, where $\bar{q}(y)$ and $\bar{p}(y, z)$ denote the respective residue classes of $q(y)$ and $p(y, z)$ modulo x. By virtue of Proposition 1.2, the geometric quotient $\mathbb{A}^3_\kappa/\mathbb{G}_{a,\kappa}$ has the structure of a Zariski locally trivial $\mathbb{A}^1$-bundle over $\mathbb{A}^1_\kappa = \mathrm{Spec}(\kappa[y])$ hence is isomorphic to $\mathbb{A}^2_\kappa$. This implies that $\mathfrak{X}_\mathfrak{m} \simeq \mathbb{A}^3_\kappa/\mathbb{G}_{a,\kappa} \simeq \mathbb{A}^2_\kappa$, as desired. □

Note that the above characterization holds independently of the a priori knowledge that the corresponding rings of invariants are finitely generated. But on the other hand, by exploiting the more general fact that arbitrary $\mathbb{G}_{a,S}$-actions on the affine 3-space $\mathbb{A}^3_S$ over the spectrum S of a discrete valuation ring A containing a field of characteristic 0 have finitely generated rings of invariants [Bhatwadekar and Daigle 2009], one can derive the following stronger alternative:

Proposition 2.2. *A fixed point free action $\sigma : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S \to \mathbb{A}^3_S$ is either a translation or its geometric quotient $\mathfrak{X} = \mathbb{A}^3_S / \mathbb{G}_{a,S}$ is an algebraic space which is not a scheme.*

Proof. Indeed, the quotient morphism $\rho : \mathbb{A}^3_S \to \mathfrak{X}$ is again an $\mathbb{A}^2$-fibration thanks to [Daigle and Kaliman 2009, Theorem 3.2] which asserts that for every field κ of characteristic 0 a fixed point free action of $\mathbb{G}_{a,\kappa}$-action on $\mathbb{A}^3_\kappa$ is a translation, and so the assertion is equivalent to the fact that a Zariski locally equivariantly trivial action σ has affine geometric quotient $\mathfrak{X}$. This can be seen in a similar way as in the proof of [Deveney et al. 2004, Theorem 2.1]. Namely, by hypothesis we can find an open covering of $\mathbb{A}^3_S$ by finitely many invariant affine open subsets U_i on which the induced $\mathbb{G}_{a,S}$-action is a translation with affine geometric quotient $U_i/\mathbb{G}_{a,S}$, $i = 1, \dots, n$. Since U_i and $\mathbb{A}^3_S$ are affine, $\mathbb{A}^3_S \setminus U_i$ is a $\mathbb{G}_{a,S}$-invariant Weil divisor on $\mathbb{A}^3_S$ which is in fact principal as A, whence $A[y, z, u]$, is factorial. It follows that there exists invariant regular functions $f_i \in A[y, z, u]^{\mathbb{G}_a} \simeq \Gamma(\mathfrak{X}, \mathcal{O}_{\mathfrak{X}})$ such that $U_i = \operatorname{Spec}(A[x, y, z]_{f_i})$ coincides with the inverse image by the quotient morphism $\rho : \mathbb{A}^3_S \to \mathfrak{X}$ of the principal open subset $\mathfrak{X}_{f_i}$ of $\mathfrak{X}$, $i = 1, \dots, n$. Since $\rho : \mathbb{A}^3_S \to \mathfrak{X}$ is a $\mathbb{G}_{a,S}$-bundle and U_i is isomorphic to $U_i/\mathbb{G}_{a,S} \times_S \mathbb{G}_{a,S}$ by assumption, we conclude that $\mathfrak{X}$ is covered by the principal affine open subsets $\mathfrak{X}_{f_i} \simeq U_i/\mathbb{G}_{a,S}$, $i = 1, \dots, n$, whence is quasiaffine. Now since by the aforementioned result [Bhatwadekar and Daigle 2009], $A[y, z, u]^{\mathbb{G}_a}$ is an integrally closed finitely generated A-algebra, it is enough to check that the canonical open immersion $j : \mathfrak{X} \to X = \operatorname{Spec}(\Gamma(\mathfrak{X}, \mathcal{O}_{\mathfrak{X}})) \simeq \operatorname{Spec}(A[y, z, u]^{\mathbb{G}_a})$ is surjective. The surjectivity over the generic point of S follows immediately from the fact the kernel of a locally nilpotent derivation of a polynomial ring in three variables over a field K of characteristic 0 is isomorphic to a polynomial ring in two variables over K (see, for example, [Miyanishi 1986]). So it remains to show that the induced open immersion $j_{\mathfrak{m}} : \mathfrak{X}_m \simeq \mathbb{A}^2_\kappa \hookrightarrow X_{\mathfrak{m}} = \operatorname{Spec}(A[y, z, u]^{\mathbb{G}_a} \otimes_A A/\mathfrak{m})$ between the corresponding fibers over the closed point $\mathfrak{m}$ of S is surjective, in fact, an isomorphism. Since $x \in A[y, z, u]^{\mathbb{G}_a}$ is prime, $X_{\mathfrak{m}} \simeq \operatorname{Spec}(A[y, z, u]^{\mathbb{G}_a}/(x))$ is an integral κ-scheme of finite type and [Bhatwadekar and Daigle 2009, Corollary 4.10] can be interpreted more precisely as the fact that $X_{\mathfrak{m}}$ is isomorphic to $C \times_\kappa \mathbb{A}^1_\kappa$ for a certain 1-dimensional affine κ-scheme C. This implies in turn that $j_{\mathfrak{m}}$ is an isomorphism. Indeed, since C is dominated via $j_{\mathfrak{m}}$ by a general affine line $\mathbb{A}^1_\kappa \subset \mathbb{A}^2_\kappa$, its normalization $\tilde{C}$ is isomorphic to $\mathbb{A}^1_\kappa$ and so $j_{\mathfrak{m}}$ factors through an open immersion $\tilde{j}_{\mathfrak{m}} : \mathbb{A}^2_\kappa \hookrightarrow \tilde{C} \times_\kappa \mathbb{A}^1_\kappa \simeq \mathbb{A}^2_\kappa$. The latter is surjective for otherwise the complement of its image would be of pure codimension 1 hence a principal divisor $\operatorname{div}(f)$ for a nonconstant regular function f on $\tilde{C} \times_\kappa \mathbb{A}^1_\kappa$. But then f would restrict to a nonconstant invertible function on the image of $\mathbb{A}^2_\kappa$ which is absurd. Thus

$\tilde{j}_{\mathfrak{m}} : \mathbb{A}^2_\kappa \hookrightarrow \tilde{C} \times_\kappa \mathbb{A}^1_\kappa \simeq \mathbb{A}^2_\kappa$ is an isomorphism and since the normalization morphism $\tilde{C} \times_\kappa \mathbb{A}^1_\kappa \to C \times_\kappa \mathbb{A}^1_\kappa$ is finite whence closed it follows that $j_{\mathfrak{m}} : \mathbb{A}^2_\kappa \hookrightarrow C \times_\kappa \mathbb{A}^1_\kappa$ is an open and closed immersion hence an isomorphism. □

2C. ***Reduction to extensions of irreducible derivations.*** In view of the discussion at the beginning of Section 2B, we may assume for the A-derivation

$$\partial = x^n \partial_y + q(y)\partial_z + p(y,z)\partial_u$$

that $n > 0$ and that the residue class of $q(y)$ in $\kappa[y]$ is either zero or not constant. In the first case, $q(y) \in \mathfrak{m}A[y]$ has the form $q(y) = x^\mu q_0(y)$ where $\mu > 0$ and where $q_0(y) \in A[y]$ has nonzero residue class modulo $\mathfrak{m}$, so that the derivation $\bar{\partial} = x^n \partial_y + q(y)\partial_z$ induced by ∂ on the subring $A[y,z]$ is reducible. On the other hand, the fixed point freeness of the $\mathbb{G}_{a,S}$-action σ generated by ∂ implies that up to multiplying u by an invertible element in A, one has $p(y,z) = 1 + x^\nu p_0(y,z)$ for some $\nu > 0$ and $p_0(y,z) \in A[y,z]$.

If $\mu \geq n$, then letting $Q_0(y) = \int_0^y q_0(\tau)\, d\tau \in A[y]$, the $\mathbb{G}_{a,S}$-invariant polynomial $z_1 = z - x^{\mu-n} Q_0(y)$ is a variable of $A[y,z,u]$ over $A[y,u]$, and so ∂ is conjugate to the derivation $x^n \partial_y + p(y, z_1 + x^{\mu-n} Q_0(y))\partial_u$ of the polynomial ring in two variables $A[z_1][y,u]$ over $A[z_1]$. Since σ is fixed point free, Proposition 1.2 implies that it is equivariantly trivial with geometric quotient isomorphic to the total space of the trivial $\mathbb{A}^1$-bundle over $\mathbb{A}^1_S = \operatorname{Spec}(A[z_1])$ whence to $\mathbb{A}^2_S$.

Otherwise, if $\mu < n$, then the $\mathbb{G}_{a,S}$-action $\tilde{\sigma} : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S \to \mathbb{A}^3_S$ on $\mathbb{A}^3_S = \operatorname{Spec}(A[\tilde{y}, \tilde{z}, \tilde{u}])$ generated by the A-derivation

$$\tilde{\partial} = x^{n-\mu} \partial_{\tilde{y}} + q_0(\tilde{y})\partial_{\tilde{z}} + (1 + x^\nu p_0(\tilde{y}, \tilde{z}))\partial_{\tilde{u}}$$

is again fixed point free, hence admits a geometric quotient $\tilde{\rho} : \mathbb{A}^3_S \to \tilde{\mathfrak{X}} = \mathbb{A}^3_S / \mathbb{G}_{a,S}$ in the form of an étale locally trivial $\mathbb{G}_{a,S}$-bundle over a certain algebraic S-space $\tilde{\mathfrak{X}}$.

Lemma 2.3. *The quotient spaces $\mathfrak{X} = \mathbb{A}^3_S / \mathbb{G}_{a,S}$ and $\tilde{\mathfrak{X}} = \mathbb{A}^3_S / \mathbb{G}_{a,S}$ for the $\mathbb{G}_{a,S}$-actions σ and $\tilde{\sigma}$ on $\mathbb{A}^3_S$ generated by ∂ and $\tilde{\partial}$ respectively are isomorphic. In particular σ is proper* (*resp. equivariantly trivial*) *if and only if $\tilde{\sigma}$ is proper* (*resp. equivariantly trivial*).

Proof. Letting $\tilde{\rho}_i : V_i = \mathbb{A}^3_S \to \tilde{\mathfrak{X}}_i = V_i / \mathbb{G}_{a,S}$, $i = 0, \ldots, \mu$, denote the geometric quotient of $V_i = \operatorname{Spec}(A[\tilde{y}_i, \tilde{z}_i, \tilde{u}_i])$ for the fixed point free $\mathbb{G}_{a,S}$-action $\tilde{\sigma}_i$ generated by the A-derivation

$$\tilde{\partial}_i = (1 + x^\nu p_0(\tilde{y}_i, \tilde{z}_i))\partial_{\tilde{u}_i} + x^{\mu-i} q_0(\tilde{y}_i)\partial_{\tilde{z}_i} + x^{n-i}\partial_{\tilde{y}_i},$$

the first assertion will follow from the more general fact that $\tilde{\mathfrak{X}}_i \simeq \tilde{\mathfrak{X}}_{i+1}$ for every $i = 0, \ldots, \mu - 1$. Indeed, we first observe that since $\tilde{u}_i$ is a slice for $\tilde{\partial}_i$ modulo x,

$\tilde{\mathfrak{X}}_{i,\mathfrak{m}} = \tilde{\mathfrak{X}}_i \times_S \operatorname{Spec}(\kappa)$ is isomorphic to $\mathbb{A}^2_\kappa = \operatorname{Spec}((A/\mathfrak{m})[\tilde{y}_i, \tilde{z}_i])$ and the restriction of $\tilde{\rho}_i$ over $\tilde{\mathfrak{X}}_{i,\mathfrak{m}}$ is isomorphic to the trivial bundle $\mathrm{pr}_1 : \tilde{\mathfrak{X}}_{i,\mathfrak{m}} \times_\kappa \operatorname{Spec}(\kappa[\tilde{u}_i]) \to \tilde{\mathfrak{X}}_{i,\mathfrak{m}}$. Now let $\beta_i : V_{i+1} \to V_i$ be the affine modification of the total space of $\tilde{\rho}_i : \mathbb{A}^3_S \to \tilde{\mathfrak{X}}_i$ with center at the zero section of the induced bundle $\mathrm{pr}_1 : \tilde{\mathfrak{X}}_{i,\mathfrak{m}} \times_\kappa \operatorname{Spec}(\kappa[\tilde{u}_i]) \to \tilde{\mathfrak{X}}_{i,\mathfrak{m}}$ and with principal divisor x. In view of the previous description, $\beta_i : V_{i+1} \to V_i$ coincides with the affine modification of $\operatorname{Spec}(A[\tilde{y}_i, \tilde{z}_i, \tilde{u}_i])$ with center at the ideal $(x, \tilde{u}_i)$ and principal divisor x, that is, with the birational S-morphism induced by the homomorphism of A-algebra

$$\begin{aligned} \beta_i^* : A[\tilde{y}_{i+1}, \tilde{z}_{i+1}, \tilde{u}_{i+1}] &\to A[\tilde{y}_i, \tilde{z}_i, \tilde{u}_i], \\ (\tilde{y}_{i+1}, \tilde{z}_{i+1}, \tilde{u}_{i+1}) &\mapsto (\tilde{y}_i, \tilde{z}_i, x\tilde{u}_i). \end{aligned}$$

By construction, β_i is equivariant for the $\mathbb{G}_{a,S}$-actions $\tilde{\sigma}_{i+1}$ and $\bar{\sigma}_i$ generated respectively by the locally nilpotent A-derivations $\tilde{\partial}_{i+1}$ of $A[\tilde{y}_{i+1}, \tilde{z}_{i+1}, \tilde{u}_{i+1}]$ and $\bar{\partial}_i = x\tilde{\partial}_i$ of $A[\tilde{y}_i, \tilde{z}_i, \tilde{u}_i]$. Furthermore, since $\tilde{\rho}_i : V_i \to \tilde{\mathfrak{X}}_i$ is also $\mathbb{G}_{a,S}$-invariant for the action $\bar{\sigma}_i$, the morphism $\tilde{\rho}_i \circ \beta_i : V_{i+1} \to \tilde{\mathfrak{X}}_i$ is $\mathbb{G}_{a,S}$-invariant, whence descends to a morphism $\tilde{\beta}_i : \tilde{\mathfrak{X}}_{i+1} \to \tilde{\mathfrak{X}}_i$. Since the latter restricts to an isomorphism over the generic point of S, it remains to check that it is also an isomorphism in a neighborhood of every point $p \in \tilde{\mathfrak{X}}_i$ lying over the closed point $\mathfrak{m}$ of S. Let $f : U = \operatorname{Spec}(B) \to \tilde{\mathfrak{X}}_i$ be an affine étale neighborhood of such a point $p \in \tilde{\mathfrak{X}}_i$ over which $\tilde{\rho}_i : V_i \to \tilde{\mathfrak{X}}_i$ becomes trivial, say $V_i \times_{\tilde{\mathfrak{X}}_i} U$ is isomorphic to $\mathbb{A}^1_U = \operatorname{Spec}(B[\tilde{v}_i])$. The $\mathbb{G}_{a,S}$-action on V_i generated by $\bar{\partial}_i$ lifts to the $\mathbb{G}_{a,U}$-action on $\mathbb{A}^1_U$ generated by the locally nilpotent B-derivation $x\partial_{\tilde{v}_i}$ and since $\beta_i : V_{i+1} \to V_i$ is the affine modification of V_i with center at the zero section of the restriction of $\tilde{\rho}_i : V_i \to \tilde{\mathfrak{X}}_i$ over the closed point of S, we have a commutative diagram

$$\begin{tikzcd}
& V_{i+1} \arrow[dl, "\beta_i"'] \arrow[dd, "\tilde{\rho}_{i+1}"] & & \mathbb{A}^1_U \arrow[ll] \arrow[dl, "\delta_i"'] \arrow[dd, "\mathrm{pr}_U"] \\
V_i \arrow[dd, "\tilde{\rho}_i"'] & & \mathbb{A}^1_U \arrow[ll] \arrow[dd, "\mathrm{pr}_U"] & \\
& \tilde{\mathfrak{X}}_{i+1} \arrow[dl, "\tilde{\beta}_i"'] & & U \arrow[ll] \arrow[dl, equal] \\
\tilde{\mathfrak{X}}_i & & U \arrow[ll, "f"] &
\end{tikzcd}$$

in which the top and front squares are cartesian, and where the morphism $\delta_i : \mathbb{A}^1_U = \operatorname{Spec}(B[\tilde{v}_{i+1}]) \to \mathbb{A}^1_U = \operatorname{Spec}(B[\tilde{v}_i])$ is defined by the B-algebras homomorphism $B[\tilde{v}_i] \to B[\tilde{v}_{i+1}]$, $\tilde{v}_i \mapsto x\tilde{v}_{i+1}$. The latter is equivariant for the action on $\operatorname{Spec}(B[\tilde{v}_{i+1}])$ generated by the locally nilpotent B-derivation $\partial_{\tilde{v}_{i+1}}$ and we conclude that $\mathrm{pr}_2 : \mathbb{A}^1_U \simeq \mathbb{A}^1_U \times_{V_i} V_{i+1} \to V_{i+1}$ is an étale trivialization of the $\mathbb{G}_{a,S}$-action induced by $\tilde{\sigma}_{i+1}$ on the open subscheme $(\tilde{\rho}_i \circ \beta_i)^{-1}(f(U))$ of V_{i+1}. This implies in turn that $U \times_{\tilde{\mathfrak{X}}_i} \tilde{\mathfrak{X}}_{i+1} \simeq U$, whence that $\tilde{\beta}_i : \tilde{\mathfrak{X}}_{i+1} \to \tilde{\mathfrak{X}}_i$ is an isomorphism in a neighborhood of $p \in \tilde{\mathfrak{X}}_i$ as desired.

The second assertion is a direct consequence of the fact that properness and global equivariant triviality of σ and $\tilde{\sigma}$ are respectively equivalent to the separatedness and the affineness of the geometric quotients $\mathfrak{X} \simeq \tilde{\mathfrak{X}}$. □

2C1. Summing up, we are now reduced to proving that a proper $\mathbb{G}_{a,S}$-action on $\mathbb{A}^3_S$ generated by an A-derivation

$$\partial = x^n \partial_y + q(y)\partial_z + p(y,z)\partial_u$$

of $A[y,z,u]$, such that $n > 0$ and $q(y) \in A[y]$ has nonconstant residue class in $\kappa[y]$, has affine geometric quotient $\mathfrak{X} = \mathbb{A}^3_S/\mathbb{G}_{a,S}$. This will be done in two steps in the next sections: we will first establish that a proper $\mathbb{G}_{a,S}$-action as above is conjugate to one generated by a special type of A-derivation called *twin-triangular.* Then we will prove in Section 4 that proper twin-triangular $\mathbb{G}_{a,S}$-actions on $\mathbb{A}^3_S$ do indeed have affine geometric quotients.

3. Reduction to twin-triangular actions

We keep the same notation as in Section 2A1 above, namely A is a discrete valuation ring containing a field of characteristic 0, with maximal ideal $\mathfrak{m}$, residue field $\kappa = A/\mathfrak{m}$, and uniformizing parameter $x \in \mathfrak{m}$. We let again $S = \operatorname{Spec}(A)$.

We call an A-derivation ∂ of $A[y,z,u]$ *twin-triangulable* if there exists a coordinate system (y, z_+, z_-) of $A[y,z,u]$ over $A[y]$ in which the conjugate of ∂ is *twin-triangular*, that is, has the form $x^n\partial_y + p_+(y)\partial_{z_+} + p_-(y)\partial_{z_-}$ for certain polynomials $p_\pm(y) \in A[y]$. This section is devoted to the proof of the following intermediate characterization of proper triangular $\mathbb{G}_{a,S}$-actions:

Proposition 3.1. *With the notation above, let ∂ be an A-derivation of $A[y,z,u]$ of the form*

$$\partial = x^n \partial_y + q(y)\partial_z + p(y,z)\partial_u$$

where $n > 0$ and where $q(y) \in A[y]$ has nonconstant residue class in $\kappa[y]$. If the $\mathbb{G}_{a,S}$-action on $\mathbb{A}^3_S = \operatorname{Spec}(A[y,z,u])$ generated by ∂ is proper, then ∂ is twin-triangulable.

The proof given below proceeds in two steps: we first construct a coordinate $\tilde{u}$ of $A[y,z,u]$ over $A[y,z]$ with the property that $\partial\tilde{u} = \tilde{p}(y,z)$ is either a polynomial in y only or its leading term $\tilde{p}_\ell(y)$ as a polynomial in z has a very particular form. In the second case, we exploit the properties of $\tilde{p}_\ell(y)$ to show that the $\mathbb{G}_{a,S}$-action generated by ∂ is not proper.

3A. ***The $\sharp$-reduction of a triangular A-derivation.*** The conjugate of an A-derivation $\partial = x^n\partial_y + q(y)\partial_z + p(y,z)\partial_u$ of $A[y,z,u]$, as in Proposition 3.1, by an isomorphism of $A[y,z]$-algebras $\psi : A[y,z][\tilde{u}] \xrightarrow{\sim} A[y,z][u]$ is again triangular of the form

$$\psi^{-1}\partial\psi = x^n\partial_y + q(y)\partial_z + \tilde{p}(y,z)\partial_{\tilde{u}}$$

for some polynomial $\tilde{p}(y,z) \in A[y,z]$. In particular, we may choose from the very beginning a coordinate system of $A[y,z,u]$ over $A[y,z]$ with the property that the degree of $\partial u \in A[y,z]$ with respect to z is minimal among all possible conjugates $\psi^{-1}\partial\psi$ of ∂ as above. In what follows, we will say for short that such a derivation ∂ is $\sharp$-*reduced* with respect to the coordinate system (y,z,u). Letting $Q(y) = \int_0^y q(\tau)\,d\tau \in A[y]$, this property can be characterized effectively as follows:

Lemma 3.2. *Let $\partial = x^n\partial_y + q(y)\partial_z + p(y,z)\partial_u$ be a $\sharp$-reduced derivation of $A[y,z,u]$ as in Proposition 3.1. If ∂ is not twin-triangular (i.e. $p(y,z) = p_0(y) \in A[y]$) then the leading term $p_\ell(y)$, $\ell \geq 1$, of $p(y,z)$ as a polynomial in z is not congruent modulo x^n to a polynomial of the form $q(y)f(Q(y))$ for some $f(\tau) \in A[\tau]$.*

Proof. Suppose that $p(y,z) = \sum_{r=0}^{\ell} p_r(y)z^r$ with $\ell \geq 1$ and that

$$p_\ell(y) = q(y)f(Q(y)) + x^n g(y)$$

for some polynomials $f(\tau), g(\tau) \in A[\tau]$. Then letting $G(y) = \int_0^y g(\tau)\,d\tau$ and

$$\tilde{u} = u - G(y)z^\ell - \sum_{k=0}^{\deg f} \frac{(-1)^k}{\prod_{j=0}^{k}(\ell+1+j)} f^{(k)}(Q(y))x^{kn}z^{\ell+1+k},$$

one checks by direct computation that

$$\partial\tilde{u} = \sum_{r=0}^{\ell-2} p_r(y)z^r + (p_{\ell-1}(y) - G(y)q(y))z^{\ell-1}.$$

Thus $(y,z,\tilde{u})$ is a coordinate system of $A[y,z,u]$ over $A[y,z]$ in which the image of $\tilde{u}$ by the conjugate of ∂ has degree $\leq \ell-1$, a contradiction to the $\sharp$-reducedness of ∂. □

To prove Proposition 3.1, it remains to show that a proper $\mathbb{G}_{a,S}$-action on $\mathbb{A}^3_S$ generated by a $\sharp$-reduced A-derivation of $A[y,z,u]$ is twin-triangular. This is done in the next subsection.

3B. ***A nonvaluative criterion for nonproperness.*** To disprove the properness of an algebraic action $\sigma : \mathbb{G}_{a,S} \times_S E \to E$ of $\mathbb{G}_{a,S}$ on an S-scheme E, it suffices in principle to check that the image of $\Phi = (\mathrm{pr}_2, \sigma) : \mathbb{G}_a \times_S E \to E \times_S E$ is not closed. However, this image turns out to be complicated to determine in general, and it is more convenient for our purpose to consider the following auxiliary construction: letting $j : \mathbb{G}_{a,S} \simeq \operatorname{Spec}(\mathbb{O}_S[t]) \hookrightarrow \mathbb{P}^1_S = \operatorname{Proj}(\mathbb{O}_S[w_0, w_1])$, $t \mapsto [t : 1]$ be the natural open immersion, the properness of the projection $\mathrm{pr}_{E\times_S E} : \mathbb{P}^1_S \times_S E \times_S E \to E \times_S E$ implies that (p_2, σ) is proper if and only if $\varphi = (j \circ \mathrm{pr}_1, \mathrm{pr}_2, \sigma) : \mathbb{G}_{a,S} \times_S E \to \mathbb{P}^1_S \times_S E \times_S E$ is proper, hence a closed immersion. Therefore the nonproperness of σ is equivalent to the fact that the closure of $\operatorname{Im}(\varphi)$ in $\mathbb{P}^1_S \times_S E \times_S E$ intersects the "boundary" $\{w_1 = 0\}$ in a nontrivial way.

3B1. Now let $\sigma : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S \to \mathbb{A}^3_S$ be the $\mathbb{G}_{a,S}$-action generated by a non-twin-triangular $\sharp$-reduced A-derivation $\partial = x^n \partial_y + q(y)\partial_z + p(y,z)\partial_u$ of $A[y,z,u]$ and let

$$\varphi = (j \circ \mathrm{pr}_1, \mathrm{pr}_2, \mu) : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S = \operatorname{Spec}(A[t][y,z,u]) \to \mathbb{P}^1_S \times_S \mathbb{A}^3_S \times_S \mathbb{A}^3_S$$

be the corresponding immersion. To disprove the properness of σ, it is enough to check that the image by φ of the closed subscheme $H = \{z = 0\} \simeq \operatorname{Spec}(A[t][y,u])$ of $\mathbb{G}_{a,S} \times_S \mathbb{A}^3_S$ is not closed in $\mathbb{P}^1_S \times_S \mathbb{A}^3_S \times_S \mathbb{A}^3_S$. After identifying $A[y,z,u] \otimes_A A[y,z,u]$ with the polynomial ring $A[y_1, y_2, z_1, z_2, u_1, u_2]$ in the obvious way, the image of H by $(\mathrm{pr}_1, \mathrm{pr}_2, \sigma) : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S \to \mathbb{A}^1_S \times_S \mathbb{A}^3_S \times_S \mathbb{A}^3_S$ is equal to the closed subscheme of $\operatorname{Spec}(A[t][y_1, y_2, z_1, z_2, u_1, u_2])$ defined by the following system of equations:

$$\begin{aligned}
&y_2 = y_1 + x^n t,\\
&z_1 = 0,\\
&z_2 = x^{-n}(Q(y_1 + x^n t) - Q(y_1)) = (y_1 - y_2)^{-1}(Q(y_2) - Q(y_1))t,\\
&u_2 = u_1 + x^{-n} \textstyle\int_0^t p(y_1 + x^n\tau)(Q(y_1 + x^n\tau) - Q(y_1))\, d\tau.
\end{aligned}$$

Letting $p(y,z) = \sum_{r=0}^{\ell} p_r(y) z^r$ with $\ell \geq 1$ and

$$\Gamma_r(y_1, y_2) = \int_{y_1}^{y_2} p_r(\xi)(Q(\xi) - Q(y_1))^r \, d\xi \in A[y_1, y_2], \quad r = 0, \dots, \ell,$$

the last equality can be rewritten modulo the first ones in the form

$$\begin{aligned}
u_2 &= u_1 + \sum_{r=0}^{\ell} x^{-nr} \int_0^t p_r(y_1 + x^n\tau)(Q(y_1 + x^n\tau) - Q(y_1))^r \, d\tau\\
&= u_1 + t(y_2 - y_1)^{-1} \sum_{r=0}^{\ell} x^{-nr} \int_{y_1}^{y_2} p_r(\xi)(Q(\xi) - Q(y_1))^r \, d\xi\\
&= u_1 + \sum_{r=0}^{\ell} \big((y_2 - y_1)^{-r-1} \Gamma_r(y_1, y_2)\big) t^{r+1}.
\end{aligned}$$

It follows that the closure V of $\varphi(H)$ is contained in the closed subscheme W of $\mathbb{P}^1_S \times_S \mathbb{A}^3_S \times_S \mathbb{A}^3_S$ defined by the equations $z_1 = 0$ and

$$\begin{aligned}(y_2 - y_1)w_1 - x^n w_0 &= 0,\\ w_1 z_2 - (y_2 - y_1)^{-1}(Q(y_2) - Q(y_1))w_0 &= 0,\\ w_1^{\ell+1}(u_2 - u_1) - \textstyle\sum_{r=0}^{\ell}\big((y_2 - y_1)^{-r-1}\Gamma_r(y_1, y_2)\big)w_0^{r+1}w_1^{\ell-r} &= 0.\end{aligned}$$

We further observe that W is irreducible, whence equal to V, given that $\Gamma_\ell(y_1, y_2) \in A[y_1, y_2]$ does not belong to the ideal generated by x^n and $Q(y_2) - Q(y_1)$. If so, then $W = V$ intersects $\{w_1 = 0\}$ along a closed subscheme Z isomorphic to the spectrum of the algebra

$$\big(A[y_1, y_2]/(x^n, (y_2-y_1)^{-1}(Q(y_2)-Q(y_1)), (y_2-y_1)^{-\ell-1}\Gamma_\ell(y_1, y_2))\big)[z_2, u_1, u_2].$$

By virtue of the $\sharp$-reducedness assumption $p_\ell(y)$ is not of the form $q(y)f(Q(y)) + x^n g(y)$, so the nonproperness of $\sigma : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S \to \mathbb{A}^3_S$ is then a consequence of the following lemma which guarantees precisely that $\Gamma_\ell(y_1, y_2) \notin (x^n, Q(y_2) - Q(y_1))A[y_1, y_2]$ and that Z is not empty.

Lemma 3.3. *Let $q(y) \in A[y]$ be a polynomial with nonconstant residue class in $\kappa[y]$ and let $Q(y) = \int_0^y q(\tau)\,d\tau$. For a polynomial $p(y) \in A[y]$ and an integer $\ell \geq 1$, the following holds*:

(a) *The polynomial $\Gamma_\ell(y_1, y_2) = \int_{y_1}^{y_2} p(y)(Q(y) - Q(y_1))^\ell\,dy$ belongs to the ideal $(x^n, Q(y_2) - Q(y_1))$ if and only if $p(y)$ can be written in the form $q(y)f(Q(y)) + x^n g(y)$ for certain polynomials $f(\tau), g(\tau) \in A[\tau]$.*

(b) *The polynomial $(y_2 - y_1)^{-\ell-1}\Gamma_\ell(y_1, y_2)$ is not invertible modulo the ideal $(x^n, (y_2 - y_1)^{-1}(Q(y_2) - Q(y_1)))$.*

Proof. For the first assertion, a sequence of ℓ successive integrations by parts shows that

$$\begin{aligned}\Gamma_\ell(y_1, y_2) &= \big[E_1(y)(Q(y) - Q(y_1))^\ell\big]_{y_1}^{y_2} - \ell\int_{y_1}^{y_2} E_1(y)q(y)(Q(y) - Q(y_1))^{\ell-1}\,dy\\ &= S(y_1, y_2) + (-1)^\ell \ell! \int_{y_1}^{y_2} E_\ell(y)q(y)\,dy\\ &= S(y_1, y_2) + (-1)^\ell \ell!(E_{\ell+1}(y_2) - E_{\ell+1}(y_1)),\end{aligned}$$

where E_k is defined recursively by $E_1(y) = \int_0^y p(\tau)\,d\tau$, $E_{k+1}(y) = \int_0^y E_k(\tau)q(\tau)\,d\tau$, and where $S(y_1, y_2) \in (Q(y_2) - Q(y_1))A[y_1, y_2]$. So $\int_{y_1}^{y_2} p(y)(Q(y) - Q(y_1))^r\,dy$ belongs to $(x^n, Q(y_2) - Q(y_1))A[y_1, y_2]$ if and only if $E_{\ell+1}(y_2) - E_{\ell+1}(y_1)$ belongs to this ideal.

Since the residue class of $Q(y) \in A[y]$ in $\kappa[y]$ is not constant, it follows from the local criterion for flatness that $A[y]$ is a faithfully flat algebra over $A[Q(y)]$.

By faithfully flat descent, this implies in turn that the sequence

$$A[Q(y)] \hookrightarrow A[y] \overset{\cdot\otimes 1 - 1\otimes\cdot}{\longrightarrow} A[y] \otimes_{A[\tau]} A[y]$$

is exact whence, with the natural identification

$$A[y] \otimes_{A[\tau]} A[y] \simeq A[y_1, y_2]/(Q(y_2) - Q(y_1)),$$

that a polynomial $F \in A[y]$ with $F(y_2) - F(y_1)$ belonging to the ideal

$$(Q(y_2) - Q(y_1))A[y_1, y_2]$$

has the form $F(y) = G(Q(y))$ for a certain polynomial $G(\tau) \in A[\tau]$. Thus $E_{\ell+1}(y_2) - E_{\ell+1}(y_1)$ belongs to $(x^n, Q(y_2) - Q(y_1))A[y_1, y_2]$, if and only if $E_{\ell+1}(y)$ is of the form $G(Q(y)) + x^n R_{\ell+1}(y)$ for some $G(\tau), R_{\ell+1}(\tau) \in A[\tau]$. This implies in turn that $E_\ell(y)q(y) = G'(Q(y))q(y) + x^n R'_{\ell+1}(y)$ whence, since $q(y) \in A[y] \setminus \mathfrak{m}A[y]$ is not a zero divisor modulo x^n, that $E_\ell(y) = G'(Q(y)) + x^n R_\ell(y)$ for a certain $R_\ell(\tau) \in A[\tau]$. We conclude by induction that $E_1(y) = G^{(\ell+1)}(Q(y)) + x^n R_1(y)$ and finally that $p(y) = G^{(\ell+2)}(Q(y))q(y) + x^n R(y)$ for a certain $R(\tau) \in A[\tau]$. This proves (a).

The second assertion is clear in the case where $p(y) \in \mathfrak{m}A[y]$. Otherwise, if $p(y) \in A[y] \setminus \mathfrak{m}A[y]$ then reducing modulo x and passing to the algebraic closure $\bar{\kappa}$ of κ, it is enough to show that if $q(y) \in \bar{\kappa}[y]$ is not constant and $p(y) \in \bar{\kappa}[y]$ is a nonzero polynomial then for every $\ell \geq 1$, the affine curves C and D in $\mathbb{A}^2_{\bar{\kappa}} = \operatorname{Spec}(\bar{\kappa}[y_1, y_2])$ defined by the vanishing of the polynomials $\Theta(y_1, y_2) = (y_2 - y_1)^{-\ell-1} \int_{y_1}^{y_2} p(y)(Q(y) - Q(y_1))^\ell \, dy$ and $R(y_1, y_2) = (y_2 - y_1)^{-1} \int_{y_1}^{y_2} q(y) \, dy$ respectively always intersect each other. Suppose on the contrary that $C \cap D = \varnothing$ and let $m = \deg q \geq 1$ and $d = \deg p \geq 0$. Then the closures $\overline{C}$ and $\overline{D}$ of C and D respectively in $\mathbb{P}^2_{\bar{\kappa}} = \operatorname{Proj}(\bar{\kappa}[y_1, y_2, y_3])$ intersect each others along a closed subscheme Y of length $\deg \overline{C} \cdot \deg \overline{D} = m(d + \ell m)$ supported on the line $\{y_3 = 0\} \simeq \operatorname{Proj}(\bar{\kappa}[y_1, y_2])$. By definition, up to multiplication by a nonzero scalar, the top homogeneous components of R and Θ have the form $\prod_{i=1}^m (y_2 - \zeta^i y_1)$, where $\zeta \in \bar{\kappa}$ is a primitive $(m+1)$-th root of unity, and $(y_2 - y_1)^{\ell-1} \int_{y_1}^{y_2} y^d (y^{m+1} - y_1^{m+1})^\ell \, dy$ respectively. But on the other hand, we have for every $i = 1, \ldots, m$

$$\begin{aligned}\bar{\kappa}[y_2] \Big/ \left(y_2 - \zeta^i, (y_2 - 1)^{-r-1} \int_1^{y_2} y^d (y^{m+1} - 1)^r \, dy \right)& \\ \simeq \bar{\kappa}[y_2] \Big/ \left(y_2 - \zeta^i, (\zeta^i - 1)^{-r-1} \int_1^{\zeta^i} \tau^d (\tau^{m+1} - 1)^r \, d\tau \right)&,\end{aligned}$$

and hence the length of the above algebra is either 1 or 0 depending on whether $\int_1^{\zeta^i} \tau^d (\tau^{m+1} - 1) \, d\tau \in \bar{\kappa}$ is zero or not. This implies that the length of Y is at most equal to m and so the only possibility would be that $d = 0$ and $\ell = m = 1$, in other words C and D are parallel lines in $\mathbb{A}^2_{\bar{\kappa}}$. But since $\int_1^{-1}(\tau^2-1)\,d\tau \neq 0$, this last possibility is also excluded. □

4. Global equivariant triviality of twin-triangular actions

By virtue of Proposition 3.1, every proper triangular $\mathbb{G}_{a,S}$-action on $\sigma\colon\mathbb{G}_{a,S}\times_S\mathbb{A}^3_S\to\mathbb{A}^3_S$ on $\mathbb{A}^3_S$ is conjugate to one generated by a twin-triangular A-derivation ∂ of $A[y,z_+,z_-]$ of the form

$$\partial = x^n\partial_y + p_+(y)\partial_{z_+} + p_-(y)\partial_{z_-}$$

for certain polynomials $p_\pm(y)\in A[y]$. So to complete the proof of the Main Theorem, it remains to show the following generalization of the main result in [Dubouloz and Finston 2014]:

Proposition 4.1. *Let S be the spectrum of discrete valuation ring A containing a field of characteristic* 0. *Then a proper twin-triangular $\mathbb{G}_{a,S}$-action on $\mathbb{A}^3_S$ has affine geometric quotient $\mathfrak{X}=\mathbb{A}^3_S/\mathbb{G}_{a,S}$.*

4A1. The principle of the proof given below is the following: we exploit the twin triangularity to construct two $\mathbb{G}_{a,S}$-invariant principal open subsets W_{Γ_+} and W_{Γ_-} in $\mathbb{A}^3_S$ with the property that the union of corresponding principal open subspaces $\mathfrak{X}_{\Gamma_\pm}=W_{\Gamma_\pm}/\mathbb{G}_{a,S}$ of $\mathfrak{X}$ covers the closed fiber of the structure morphism $\mathrm{p}:\mathfrak{X}\to S$. We then show that $\mathfrak{X}_{\Gamma_+}$ and $\mathfrak{X}_{\Gamma_-}$ are in fact affine subschemes of $\mathfrak{X}$. On the other hand, since ∂ admits $x^{-n}y$ as a global slice over A_x, the generic fiber of p is isomorphic to the affine plane over the function field A_x of S. So it follows that $\mathfrak{X}$ is covered by three principal affine open subschemes $\mathfrak{X}_{\Gamma_+}$, $\mathfrak{X}_{\Gamma_-}$ and $\mathfrak{X}_x$ corresponding to regular functions x, Γ_+, Γ_- which generate the unit ideal in $\Gamma(\mathfrak{X},\mathbb{O}_{\mathfrak{X}})\simeq A[y,z_+,z_-]^{\mathbb{G}_{a,S}}\subset A[y,z_+,z_-]$, whence is an affine scheme.

4A2. The fact that the affineness of $\mathrm{p}:\mathfrak{X}=\mathbb{A}^3_S/\mathbb{G}_{a,S}\to S=\mathrm{Spec}(A)$ is a local property with respect to the fpqc topology on S [SGA1 1971, VIII, Corollaire 5.6] enables a reduction to the case where the discrete valuation ring A is Henselian or complete. Since it contains a field of characteristic zero, an elementary application of Hensel's Lemma implies that a maximal subfield of such a local ring A is a field of representatives, that is, a subfield which is mapped isomorphically by the quotient projection $A\mapsto A/\mathfrak{m}$ onto the residue field $\kappa=A/\mathfrak{m}$. This is in fact the only property of A that we will use in the sequel. So from now on, $(A,\mathfrak{m},\kappa)$ is a discrete valuation ring containing a field κ of characteristic 0 and with residue field $A/\mathfrak{m}\simeq\kappa$.

4B. ***Twin-triangular actions in general position and associated invariant covering.*** Here we construct a pair of principal $\mathbb{G}_{a,S}$-invariant open subsets $W_\pm=W_{\Gamma_\pm}$ of $\mathbb{A}^3_S$ associated with a twin-triangular A-derivation of $A[y,z_+,z_-]$ whose geometric quotients will be studied in the next subsection. We begin with a technical condition which will be used to guarantee that the union of W_+ and W_- covers the closed fiber of the projection $\mathrm{pr}_S : \mathbb{A}^3_S \to S$.

Definition 4.2. Let $(A, \mathfrak{m}, \kappa)$ be a discrete valuation ring containing a field of characteristic 0 and let $x \in \mathfrak{m}$ be a uniformizing parameter. A twin-triangular A-derivation $\partial = x^n \partial_y + p_+(y)\partial_{z_+} + p_-(y)\partial_{z_-}$ of $A[y, z_+, z_-]$ is said to be in *general position* if it satisfies the following properties:

(a) The residue classes $\bar{p}_\pm \in \kappa[y]$ of the polynomials $p_\pm \in A[y]$ modulo $\mathfrak{m}$ are both nonzero and relatively prime.

(b) There exist integrals $\bar{P}_\pm \in A[y]$ of $\bar{p}_\pm$ with respect to y for which the inverse images of the branch loci of the morphisms $\bar{P}_+ : \mathbb{A}^1_\kappa \to \mathbb{A}^1_\kappa$ and $\bar{P}_- : \mathbb{A}^1_\kappa \to \mathbb{A}^1_\kappa$ are disjoint.

Lemma 4.3. *With the notation above, every twin-triangular A-derivation ∂ of $A[y, z_+, z_-]$ generating a fixed point free $\mathbb{G}_{a,S}$-action on $\mathbb{A}^3_S$ is conjugate to one in general position.*

Proof. A twin-triangular derivation $\partial = x^n \partial_y + p_+(y)\partial_{z_+} + p_-(y)\partial_{z_-}$ generates a fixed point free $\mathbb{G}_{a,S}$-action if and only if x^n, $p_+(y)$ and $p_-(y)$ generate the unit ideal in $A[y, z_+, z_-]$. So the residue classes $\bar{p}_+$ and $\bar{p}_-$ of p_+ and p_- are relatively prime and at least one of them, say $\bar{p}_-$, is nonzero. If $\bar{p}_+ = 0$ then p_- is necessarily of the form $p_-(y) = c + x\tilde{p}_-(y)$ for some $c \in A^*$ and so changing z_+ for $z_+ + z_-$ yields a twin-triangular derivation conjugate to ∂ for which the corresponding polynomials $p_\pm(y)$ both have nonzero residue classes modulo x. More generally, changing z_- for $az_- + bz_+$ for general $a \in A^*$ and $b \in A$ yields a twin-triangular derivation conjugate to ∂ and still satisfying condition (a) in Definition 4.2. So it remains to show that up to such a coordinate change, condition (b) in the definition can be achieved.

This can be seen as follows : we consider $\mathbb{A}^2_\kappa$ embedded in $\mathbb{P}^2_\kappa = \mathrm{Proj}(\kappa[u, v, w])$ as the complement of the line $L_\infty = \{w = 0\}$ so that the coordinate system (u, v) on $\mathbb{A}^2$ is induced by the projections from the κ-rational points $[0 : 1 : 0]$ and $[1 : 0 : 0]$ respectively. We let C be the closure in $\mathbb{P}^2$ of the image of the morphism $j = (\bar{P}_+, \bar{P}_-) : \mathbb{A}^1_\kappa = \mathrm{Spec}(\kappa[y]) \to \mathbb{A}^2_\kappa$ defined by the residue classes $\bar{P}_+$ and $\bar{P}_-$ in $\kappa[y]$ of integrals $P_\pm(y) \in A[y]$ of $p_\pm(y)$, and we denote by $Z \subset C$ the image by j of the inverse image of the branch locus of $\bar{P}_+ : \mathbb{A}^1_\kappa \to \mathbb{A}^1_\kappa$. Note that Z is a finite subset of C defined over k, and therefore the set of lines in $\mathbb{P}^2_k$ passing through a point of Z and tangent to a local analytic branch of C at some point is finite. This follows from the fact that the set of lines in $\mathbb{P}^2_k$ intersecting transversely a fixed curve is Zariski open. Therefore, the complement of the finitely many intersection points of these lines with L_∞ is a Zariski open subset U of L_∞

with the property that for every $q \in U$, the line through q and every arbitrary point of Z intersects every local analytic branch of C transversally at every point. By construction, the rational projections from $[0:1:0]$ and an arbitrary κ-rational point in $U \setminus \{[0:1:0]\}$ induce a new coordinate system on $\mathbb{A}^2_\kappa$ of the form $(u, av+bu)$, $a \neq 0$, with the property that Z is not contained in the inverse image of the branch locus of the morphism $a\bar{P}_- + b\bar{P}_+ : \mathbb{A}^1_\kappa \to \mathbb{A}^1_\kappa$. Changing z_- for $az_- + bz_+$ for a pair $(a,b) \in \kappa^* \times \kappa \subset A^* \times A$ corresponding to a general point in U yields a twin-triangular derivation conjugate to ∂ and satisfying simultaneously conditions (a) and (b) in Definition 4.2. □

4B1. Now let $\partial = x^n\partial_y + p_+(y)\partial_{z_+} + p_-(y)\partial_{z_-}$ be a twin-triangular A-derivation of $A[y, z_+, z_-]$ generating a proper whence fixed point free $\mathbb{G}_{a,S}$-action $\sigma : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S \to \mathbb{A}^3_S$. By virtue of Lemma 4.3 above, we may assume up to a coordinate change preserving twin-triangularity that ∂ is in general position. Property (a) in Definition 4.2 then guarantees in particular that the triangular derivations $\partial_\pm = x^n\partial_y + p_\pm(y)\partial_{z_\pm}$ of $A[y, z_\pm]$ are both irreducible. Furthermore, given any integral $P_\pm(y) \in A[y]$ of $p_\pm(y)$, the morphism $\bar{P}_\pm : \mathbb{A}^1_\kappa \to \mathbb{A}^1_\kappa$ obtained by restricting $P_\pm : \mathbb{A}^1_S = \operatorname{Spec}(A[y]) \to \mathbb{A}^1_S = \operatorname{Spec}(A[t])$ to the closed fiber of $\operatorname{pr}_S : \mathbb{A}^3_S \to S$ is not constant. The branch locus of $\bar{P}_\pm$ is then a principal divisor $\operatorname{div}(\alpha_\pm(t))$ for a certain polynomial $\alpha_\pm(t) \in \kappa[t] \subset A[t]$ generating the kernel of the homomorphism $\kappa[t] \to \kappa[y]/(\bar{p}_\pm(y))$, $t \mapsto \bar{P}_\pm(y) + (\bar{p}_\pm(y))$. Property (b) in Definition 4.2 guarantees that we can choose P_+ and P_- in such a way that the polynomial $\alpha_+(\bar{P}_+(y))$ and $\alpha_-(\bar{P}_-(y))$ generate the unit ideal in $\kappa[y]$. Up to a diagonal change of coordinates on $A[y, z_+, z_-]$, we may further assume without loss of generality that $\bar{P}_+$ and $\bar{P}_-$ are monic.

4B2. We let $R_\pm = A[t]_{\alpha_\pm}$ and we let $U_\pm = \operatorname{Spec}(R_\pm)$ be the principal open subset of $\mathbb{A}^1_S = \operatorname{Spec}(A[t])$ where $\alpha_\pm$ does not vanish. The polynomial $\Phi_\pm = -x^n z_\pm + P_\pm(y) \in A[y, z_+, z_-]$ belongs to the kernel of ∂ hence defines a $\mathbb{G}_{a,S}$-invariant morphism $\Phi_\pm : \mathbb{A}^3_S = \operatorname{Spec}(A[y, z_+, z_-]) \to \mathbb{A}^1_S = \operatorname{Spec}(A[t])$. We let

$$W_\pm = \Phi_\pm^{-1}(U_\pm) \simeq \operatorname{Spec}\bigl(R_\pm[y, z_+, z_-]/(-x^n z_\pm + P_\pm(y) - t)\bigr)$$

Note that $W_\pm$ is a $\mathbb{G}_{a,S}$-invariant open subset of $\mathbb{A}^3_S$ which can be identified with the principal open subset where the $\mathbb{G}_{a,S}$-invariant regular function $\Gamma_\pm = \alpha_\pm \circ \Phi_\pm$ does not vanish. Since $\alpha_+(\bar{P}_+(y))$ and $\alpha_-(\bar{P}_-(y))$ generate the unit ideal in $\kappa[y]$, it follows that the union of W_+ and W_- covers the closed fiber of the projection $\operatorname{pr}_S : \mathbb{A}^3_S \to S$.

4C. ***Affineness of geometric quotients.*** With the notation of Section 4B2 above, the geometric quotient $\mathfrak{X}_\pm = W_\pm/\mathbb{G}_{a,S}$ for the action induced by $\sigma : \mathbb{G}_{a,S} \times_S \mathbb{A}^3_S \to \mathbb{A}^3_S$ can be identified with the principal open subspace $\mathfrak{X}_{\Gamma_\pm}$ of $\mathfrak{X} = \mathbb{A}^3_S/\mathbb{G}_{a,S}$ where the invariant function $\Gamma_\pm \in A[y, z_+, z_-]^{\mathbb{G}_{a,S}} \simeq \Gamma(\mathfrak{X}, \mathcal{O}_\mathfrak{X})$ does not vanish. The properness of σ implies that $\mathfrak{X}$, whence $\mathfrak{X}_+$ and $\mathfrak{X}_-$, are separated algebraic spaces, and the construction of W_+ and W_- guarantees that the closed fiber of the structure morphism $\mathrm{p} : \mathfrak{X} \to S$ is contained in the union of $\mathfrak{X}_+$ and $\mathfrak{X}_-$. So to complete the proof of Proposition 4.1, it remains to show that $\mathfrak{X}_\pm$ is an affine scheme. In fact, since $\mathfrak{X}_\pm$ is by construction an algebraic space over the affine scheme $U_\pm = \mathrm{Spec}(R_\pm)$, its affineness is equivalent to that of the structure morphism $q_\pm : \mathfrak{X}_\pm \to U_\pm$, a property which can be checked locally with respect to the étale topology on $U_\pm$.

4C1. In our situation, there is a natural finite étale base change $\varphi_\pm : \tilde{U}_\pm \to U_\pm$ which is obtained as follows: By construction, see Section 4B1 above, the morphism $\bar{P}_\pm : \mathbb{A}^1_\kappa = \mathrm{Spec}(\kappa[y]) \to \mathrm{Spec}(\kappa[t])$, restricts to a finite étale covering $h_{0,\pm} : C_{1,\pm} = \mathrm{Spec}(\kappa[y]_{\alpha_\pm(\bar{P}_\pm(y))}) \to C_\pm = \mathrm{Spec}(\kappa[t]_{\alpha_\pm(t)})$ of degree $r_\pm = \deg_y(\bar{P}_\pm(y))$. Letting $\tilde{C}_\pm = \mathrm{Spec}(B_\pm)$ be the normalization of $C_\pm$ in the Galois closure $L_\pm$ of the field extension $i_\pm : \kappa(t) \hookrightarrow \kappa(y)$, the induced morphism $h_\pm : \tilde{C}_\pm \to C_\pm$ is an étale Galois cover with Galois group $G_\pm = \mathrm{Gal}(L_\pm/\kappa(t))$, which factors as

$$h_\pm : \tilde{C}_\pm = \mathrm{Spec}(B_\pm) \xrightarrow{h_{1,\pm}} C_{1,\pm} = \mathrm{Spec}\big(\kappa[y]_{\alpha_\pm(\bar{P}_\pm(y))}\big) \xrightarrow{h_{0,\pm}} C_\pm = \mathrm{Spec}(\kappa[t]_{\alpha_\pm(t)})$$

where $h_{1,\pm} : \tilde{C}_\pm \to C_{1,\pm}$ is an étale Galois cover for a certain subgroup $H_\pm$ of $G_\pm$ of index $r_\pm$. Letting $\tilde{R}_\pm = A \otimes_\kappa B_\pm \simeq A[t]_{\alpha_\pm(t)} \otimes_{\kappa[t]_{\alpha_\pm(t)}} B_\pm$ and $\tilde{U}_\pm = \mathrm{Spec}(\tilde{R}_\pm)$, the morphism $\varphi_\pm = \mathrm{pr}_1 : \tilde{U}_\pm \simeq U_\pm \times_{C_\pm} \tilde{C}_\pm \to U_\pm$ is an étale Galois cover with Galois group $G_\pm$, in particular a finite morphism. Since $\mathfrak{X}_\pm$ is separated, the algebraic space $\tilde{\mathfrak{X}}_\pm = \mathfrak{X}_\pm \times_{U_\pm} \tilde{U}_\pm$ is separated and, by construction, isomorphic to the geometric quotient of the scheme

$$\tilde{W}_\pm = W_\pm \times_{U_\pm} \tilde{U}_\pm \simeq \mathrm{Spec}\big(\tilde{R}_\pm[y, z_+, z_-]/(-x^n z_\pm + P_\pm(y) - t)\big)$$

by the proper $\mathbb{G}_{a,\tilde{U}_\pm}$-action generated by the locally nilpotent $\tilde{R}_\pm$-derivation $x^n \partial_y + p_+(y)\partial_{z_+} + p_-(y)\partial_{z_-}$ of $\tilde{R}_\pm[y, z_+, z_-]//(-x^n z_\pm + P_\pm(y) - t)$, which commutes with the action of $G_\pm$. The following lemma completes the proof of Proposition 4.1 whence of the Main Theorem.

Lemma 4.4. *The geometric quotient $\tilde{\mathfrak{X}}_\pm = \tilde{W}_\pm/\mathbb{G}_{a,\tilde{U}_\pm}$ is an affine $\tilde{U}_\pm$-scheme.*

Proof. Since $\tilde{U}_\pm$ is affine, the assertion is equivalent to the affineness of $\tilde{\mathfrak{X}}_\pm$. From now on, we only consider the case of $\tilde{\mathfrak{X}}_+ = \tilde{W}_+/\mathbb{G}_{a,\tilde{U}_+}$, the case of $\tilde{\mathfrak{X}}_-$ being similar. To simplify the notation, we drop the corresponding subscript "+", writing simply $\tilde{W} = \mathrm{Spec}(\tilde{R}[y, z, z_-]/(-x^n z + P(y) - t))$. We denote $x \otimes 1 \in \tilde{R} = A \otimes_\kappa B$ by x and we further identify B with a sub-κ-algebra of $\tilde{R}$ via the homomorphism $1 \otimes \mathrm{id}_B : B \to \tilde{R}$ and with the quotient $\tilde{R}/x\tilde{R}$ via the composition $1 \otimes \mathrm{id}_B : B \to A \otimes_\kappa B \to A \otimes_\kappa B/((x \otimes 1)A \otimes_\kappa B) = \kappa \otimes_\kappa B \simeq B$.

By construction of B, the monic polynomial $\bar{P}(y) - t \in B[y]$ splits as $\bar{P}(y) - t = \prod_{\bar{g} \in G/H}(y - t_{\bar{g}})$ for certain elements $t_{\bar{g}} \in B$, $\bar{g} \in G/H$, on which the Galois group G

acts by permutation $g' \cdot t_{\bar{g}} = t_{\overline{(g')^{-1} \cdot g}}$. Furthermore, since $h_0 : C_1 \to C$ is étale, it follows that for distinct $\bar{g}, \bar{g}' \in G/H$, $t_{\bar{g}} - t_{\bar{g}'} \in B$ is an invertible regular function on $\tilde{C}$ whence on $\tilde{U} = S \times_{\mathrm{Spec}(\kappa)} \tilde{C}$ via the identifications made above. This implies in turn that there exists a collection of elements $\sigma_{\bar{g}} \in \tilde{R}$ with respective residue classes $t_{\bar{g}} \in B = \tilde{R}/x\tilde{R}$ modulo x, $\bar{g} \in G/H$, on which G acts by permutation, a G-invariant polynomial $S_1 \in \tilde{R}[y]$ with invertible residue class modulo x and a G-invariant polynomial $S_2 \in \tilde{R}[y]$ such that in $\tilde{R}[y]$ one can write

$$P(y) - t = S_1(y) \prod_{\bar{g} \in G/H} (y - \sigma_{\bar{g}}) + x^n S_2(y).$$

Concretely, the elements $\sigma_{\bar{g}} = \sigma_{\bar{g},n-1} \in \tilde{R}$, $\bar{g} \in G/H$, can be constructed by induction via a sequence of elements $\sigma_{\bar{g},m} \in \tilde{R}$, $\bar{g} \in G/H$, $m = 0, \dots, n-1$, starting with $\sigma_{\bar{g},0} = t_{\bar{g}} \in B \subset \tilde{R}$ and culminating in $\sigma_{\bar{g},n-1} = \sigma_{\bar{g}}$, and characterized by the property that for every $m = 0, \dots, n-1$, there exists $\mu_{\bar{g},m} \in \tilde{R}$ such that $P(\sigma_{\bar{g},m}) - t = x^{m+1}\mu_{\bar{g},m}$, $\bar{g} \in G/H$. Indeed, writing $P(y) - t = \prod_{\bar{g} \in G/H}(y - t_{\bar{g}}) + x\tilde{P}(y)$ for a certain $\tilde{P}(y) \in \tilde{R}[y]$ and assuming that the $\sigma_{\bar{g},m}$, $\bar{g} \in G/H$, have been constructed up to a certain index $m < n-1$, we look for elements $\sigma_{\bar{g},m+1} \in \tilde{R}$ written in the form $\sigma_{\bar{g},m} + x^{m+1}\lambda_{\bar{g}}$ for some $\lambda_{\bar{g}} \in \tilde{R}$. For a fixed $\bar{g}_0 \in G/H$, the conditions impose that

$$\begin{aligned} P(\sigma_{\bar{g}_0,m+1}) - t &= \prod_{\bar{g} \in G/H} (\sigma_{\bar{g}_0,m} + x^{m+1}\lambda_{\bar{g}_0} - t_{\bar{g}}) + x\tilde{P}(\sigma_{\bar{g}_0,m} + x^{m+1}\lambda_{\bar{g}_0}) \\ &= x^{m+1}\lambda_{\bar{g}_0} \prod_{\bar{g} \in (G/H)\setminus\{\bar{g}_0\}} (t_{\bar{g}_0} - t_{\bar{g}}) + P(\sigma_{\bar{g}_0,m}) - t + x^{m+2}\nu_{\bar{g}_0,m} \\ &= x^{m+1}\lambda_{\bar{g}_0} \prod_{\bar{g} \in (G/H)\setminus\{\bar{g}_0\}} (t_{\bar{g}_0} - t_{\bar{g}}) + x^{m+1}\mu_{\bar{g}_0,m} + x^{m+2}\nu_{\bar{g}_0,m} \end{aligned}$$

for some $\nu_{\bar{g}_0,m} \in \tilde{R}$, and since $\prod_{\bar{g} \in (G/H)\setminus\{\bar{g}_0\}}(t_{\bar{g}_0} - t_{\bar{g}}) \in \tilde{R}^*$, we conclude that

$$\lambda_{\bar{g}_0} = \frac{\mu_{\bar{g}_0,m}}{\prod_{\bar{g} \in (G/H)\setminus\{\bar{g}_0\}}(t_{\bar{g}_0} - t_{\bar{g}})} \quad \text{and} \quad \mu_{\bar{g}_0,m+1} = \nu_{\bar{g}_0,m}.$$

A direct computation shows further that $g' \cdot \sigma_{\bar{g},m+1} = \sigma_{\overline{(g')^{-1} \cdot g},m+1}$ and that $g' \cdot \mu_{\bar{g},m+1} = \mu_{\overline{(g')^{-1} \cdot g},m+1}$. Iterating this procedure $n-1$ times yields the desired collection of elements $\sigma_{\bar{g}} = \sigma_{\bar{g},n-1} \in \tilde{R}$. By construction, $\prod_{\bar{g} \in G/H}(y - \sigma_{\bar{g}}) \in \tilde{R}[y]$ is then an invariant polynomial which divides $P(y) - t$ modulo $x^n\tilde{R}$, which implies in turn the existence of the G-invariant polynomials $S_1(y), S_2(y) \in \tilde{R}[y]$.

The closed fiber of the induced morphism $\tilde{W} \to S$ consists of a disjoint union of closed subschemes $D_{\bar{g}} \simeq \mathrm{Spec}(\tilde{R}[z, z_-]) \simeq \mathbb{A}^2_{\tilde{C}}$ with defining ideals $(x, y - \sigma_{\bar{g}})$, $\bar{g} \in G/H$. The open subscheme $\tilde{W}_{\bar{g}} = \tilde{W} \setminus \bigcup_{\bar{g}' \in (G/H)\setminus\{\bar{g}\}} D_{\bar{g}'}$ of $\tilde{W}$ is $\mathbb{G}_{a,\tilde{U}}$-invariant and one checks using the above expression for $P(y)-t$ that the rational map

$$\tilde{W} \dashrightarrow \operatorname{Spec}(\tilde{R}[u_{\bar g}, z_-]),$$

$$(y,z,z_-)\mapsto (u_{\bar g}, z_-) = \left(\frac{y-\sigma_{\bar g}}{x^n} = \frac{z-S_2(y)}{S_1(y)\prod_{\bar g'\in (G/H)\setminus\{\bar g\}}(y-\sigma_{\bar g'})}, z_-\right)$$

induces a $\mathbb{G}_{a,\tilde U}$-equivariant isomorphism $\tau_g : \tilde W_{\bar g} \xrightarrow{\sim} \mathbb{A}^2_{\tilde U} = \operatorname{Spec}(\tilde R[u_{\bar g}, z_-])$ for the $\mathbb{G}_{a,\tilde U}$-action on $\mathbb{A}^2_{\tilde U}$ generated by the locally nilpotent $\tilde R$-derivation $\partial_{u_{\bar g}} + p_-(x^n u_{\bar g}+\sigma_{\bar g})\partial_{z_-}$ of $\tilde R[u_{\bar g}, z_-]$. The latter is a translation with $u_{\bar g}$ as a global slice and with geometric quotient $\tilde W_{\bar g}/\mathbb{G}_{a,\tilde U}$ isomorphic to $\operatorname{Spec}(\tilde R[v_{\bar g}])$ where

$$v_{\bar g} = z_- - x^{-n}(P_-(x^n u_{\bar g}+\sigma_{\bar g}) - P_-(\sigma_{\bar g})) \in \tilde R[u_{\bar g}, z_-]^{\mathbb{G}_{a,\tilde U}}.$$

By construction, for distinct $\bar g, \bar g' \in G/H$, the rational functions $\tau^*_{\bar g} v_{\bar g}$ and $\tau^*_{\bar g'} v_{\bar g'}$ on $\tilde W$ differ by the addition of the element

$$f_{\bar g,\bar g'} = x^{-n}(P_-(\sigma_{\bar g}) - P_-(\sigma_{\bar g'})) \in \tilde R_x \in \Gamma(\tilde W_{\bar g}\cap \tilde W_{\bar g'}, \mathcal{O}_{\tilde W}).$$

This implies that $\tilde{\mathfrak{X}} = \tilde W/\mathbb{G}_{a,\tilde U}$ is isomorphic to the $\tilde U$-scheme obtained by gluing r copies $\tilde{\mathfrak{X}}_g = \operatorname{Spec}(\tilde R[v_{\bar g}])$ of $\mathbb{A}^1_{\tilde U}$ along the principal open subsets $\tilde{\mathfrak{X}}_{\bar g,x} \simeq \operatorname{Spec}(\tilde R_x[v_{\bar g}])$ via the isomorphisms induced by the $\tilde R_x$-algebra isomorphisms

$$\xi^*_{\bar g,\bar g'} : \tilde R_x[v_{\bar g}] \to \tilde R_x[v_{\bar g'}],\ v_{\bar g} \mapsto v_{\bar g'} + f_{\bar g,\bar g'}, \quad \bar g, \bar g' \in G/H,\ \bar g \neq \bar g'.$$

Since by assumption $\tilde{\mathfrak{X}}$ is separated, it follows from [EGA 1960, I, Proposition (5.5.6)] that for every pair of distinct elements $\bar g, \bar g' \in G/H$, the subring $\tilde R[v_{\bar g'}, f_{\bar g,\bar g'}]$ of $\tilde R_x[v_{\bar g'}]$ generated by the union of $\tilde R[v_{\bar g'}]$ and $\xi^*_{\bar g,\bar g'}(\tilde R[v_{\bar g}])$ is equal to $\tilde R_x[v_{\bar g'}]$. This holds if and only if $\tilde R[f_{\bar g,\bar g'}] = \tilde R_x$ whence if and only if $f_{\bar g,\bar g'} \in \tilde R_x$ has the form $f_{\bar g,\bar g'} = x^{-m_{\bar g,\bar g'}} F_{\bar g,\bar g'}$ for a certain $m_{\bar g,\bar g'} > 1$ and an element $F_{\bar g,\bar g'} \in \tilde R$ with invertible residue class modulo x.

This additional information enables a proof of the affineness of $\tilde{\mathfrak{X}}$ by induction on r as follows: given a pair of distinct elements $\bar g, \bar g' \in G/H$ such that $m_{\bar g,\bar g'} = m > 0$ is maximal, we let $\theta_{\bar g} = 0$ and $\theta_{\bar g''} = x^{m-m_{\bar g,\bar g''}} F_{\bar g,\bar g''} \in \tilde R$ for every $\bar g'' \in (G/H)\setminus\{\bar g\}$. The choice of the elements $\theta_{\bar g''} \in \tilde R$ guarantees that the local sections

$$\psi_{\bar g''} = x^m v_{\bar g''} + \theta_{\bar g''} \in \Gamma(\tilde{\mathfrak{X}}_{\bar g''}, \mathcal{O}_{\tilde{\mathfrak{X}}}), \quad \bar g'' \in G/H,$$

glue to a global regular function $\psi \in \Gamma(\tilde{\mathfrak{X}}, \mathcal{O}_{\tilde{\mathfrak{X}}})$. Since $\theta_{\bar g'} = F_{\bar g,\bar g'}$ is invertible modulo x, the regular functions x, ψ and $\psi - \theta_{\bar g'}$ generate the unit ideal in $\Gamma(\tilde{\mathfrak{X}}, \mathcal{O}_{\tilde{\mathfrak{X}}})$. The principal open subset $\tilde{\mathfrak{X}}_x$ of $\tilde{\mathfrak{X}}$ is isomorphic to $\tilde{\mathfrak{X}}_{\bar g,x} \simeq \operatorname{Spec}(\tilde R_x[v_{\bar g}])$ for every $\bar g \in G/H$, hence is affine. On the other hand, $\tilde{\mathfrak{X}}_\psi$ and $\tilde{\mathfrak{X}}_{\psi-\theta_{\bar g'}}$ are contained respectively in the open subschemes $\tilde{\mathfrak{X}}(\bar g)$ and $\tilde{\mathfrak{X}}(\bar g')$ obtained by gluing only the $r-1$ open subsets $\tilde{\mathfrak{X}}_{\bar g''}$ corresponding to the elements $\bar g''$ in $(G/H)\setminus\{\bar g\}$ and $(G/H)\setminus\{\bar g'\}$ respectively. By the induction hypothesis, the latter are both affine and hence $\tilde{\mathfrak{X}}_{\psi}$ and $\tilde{\mathfrak{X}}_{\psi-\theta_{\bar{g}'}}$ are affine as well. This shows that $\tilde{\mathfrak{X}}$ is an affine scheme and completes the proof. □

References

[Bass 1968] H. Bass, *Algebraic K-theory*, W. A. Benjamin, New York, 1968. MR 40 #2736 Zbl 0174.30302

[Bass 1984] H. Bass, "A nontriangular action of $\mathbb{G}_a$ on $\mathbb{A}^3$", *J. Pure Appl. Algebra* **33**:1 (1984), 1–5. MR 85j:14086 Zbl 0555.14019

[Bass et al. 1977] H. Bass, E. H. Connell, and D. L. Wright, "Locally polynomial algebras are symmetric algebras", *Invent. Math.* **38**:3 (1977), 279–299. MR 55 #5613 Zbl 0371.13007

[Bhatwadekar and Daigle 2009] S. M. Bhatwadekar and D. Daigle, "On finite generation of kernels of locally nilpotent R-derivations of $R[X, Y, Z]$", *J. Algebra* **322**:9 (2009), 2915–2926. MR 2011b:13088 Zbl 1234.13027

[Daigle and Freudenburg 2001] D. Daigle and G. Freudenburg, "Triangular derivations of $k[X_1, X_2, X_3, X_4]$", *J. Algebra* **241**:1 (2001), 328–339. MR 2002g:13058 Zbl 1018.13013

[Daigle and Kaliman 2009] D. Daigle and S. Kaliman, "A note on locally nilpotent derivations and variables of $k[X, Y, Z]$", *Canad. Math. Bull.* **52**:4 (2009), 535–543. MR 2011j:14125 Zbl 1185.14056

[Deveney and Finston 1994] J. K. Deveney and D. R. Finston, "$\mathbb{G}_a$ actions on $\mathbb{C}^3$ and $\mathbb{C}^7$", *Comm. Algebra* **22**:15 (1994), 6295–6302. MR 95j:13004 Zbl 0867.13002

[Deveney and Finston 1995] J. K. Deveney and D. R. Finston, "A proper $\mathbb{G}_a$ action on $\mathbb{C}^5$ which is not locally trivial", *Proc. Amer. Math. Soc.* **123**:3 (1995), 651–655. MR 95j:14065 Zbl 0832.14036

[Deveney and Finston 2000] J. K. Deveney and D. R. Finston, "Twin triangular derivations", *Osaka J. Math.* **37**:1 (2000), 15–21. MR 2001f:14088 Zbl 0968.14025

[Deveney et al. 1994] J. K. Deveney, D. R. Finston, and M. Gehrke, "$\mathbb{G}_a$ actions on $\mathbb{C}^n$", *Comm. Algebra* **22**:12 (1994), 4977–4988. MR 95e:14038 Zbl 0817.14029

[Deveney et al. 2004] J. K. Deveney, D. R. Finston, and P. van Rossum, "Triangular $\mathbb{G}_a$ actions on $\mathbb{C}^4$", *Proc. Amer. Math. Soc.* **132**:10 (2004), 2841–2848. MR 2005d:14064 Zbl 1077.14093

[Dubouloz and Finston 2014] A. Dubouloz and D. R. Finston, "Proper twin-triangular $\mathbb{G}_a$-actions on $\mathbb{A}^4$ are translations", *Proc. Amer. Math. Soc.* **142**:5 (2014), 1513–1526. MR 3168459 Zbl 06269648 arXiv 1109.6302

[EGA 1960] A. Grothendieck and J. Dieudonné, "Éléments de géométrie algébrique, I: Le langage des schémas", *Inst. Hautes Études Sci. Publ. Math.* **4** (1960), 5–228. MR 29 #1207 Zbl 0118.36206

[EGA 1965] A. Grothendieck and J. Dieudonné, "Éléments de géométrie algébrique, IV: Étude locale des schémas et des morphismes de schémas, II", *Inst. Hautes Études Sci. Publ. Math.* **24** (1965), 5–231. MR 33 #7330 Zbl 0135.39701

[Fauntleroy and Magid 1976] A. Fauntleroy and A. R. Magid, "Proper $\mathbb{G}_a$-actions", *Duke Math. J.* **43**:4 (1976), 723–729. MR 54 #5254 Zbl 0351.14026

[Kaliman 2004] S. Kaliman, "Free $\mathbb{C}_+$-actions on $\mathbb{C}^3$ are translations", *Invent. Math.* **156**:1 (2004), 163–173. MR 2005b:14102 Zbl 1058.14076

[Kaliman and Saveliev 2004] S. Kaliman and N. Saveliev, "$\mathbb{C}_+$-actions on contractible threefolds", *Michigan Math. J.* **52**:3 (2004), 619–625. MR 2005h:14145 Zbl 1067.14067

[Knutson 1971] D. Knutson, *Algebraic spaces*, Lecture Notes in Math. **203**, Springer, Berlin, 1971. MR 46 #1791 Zbl 0221.14001

[Laumon and Moret-Bailly 2000] G. Laumon and L. Moret-Bailly, *Champs algébriques*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) **39**, Springer, Berlin, 2000. MR 2001f:14006 Zbl 0945.14005

[Miyanishi 1986] M. Miyanishi, "Normal affine subalgebras of a polynomial ring", pp. 37–51 in *Algebraic and topological theories* (Kinosaki, 1984), edited by M. Nagata et al., Kinokuniya, Tokyo, 1986. MR 1102251 Zbl 0800.14018

[Rentschler 1968] R. Rentschler, "Opérations du groupe additif sur le plan affine", *C. R. Acad. Sci. Paris Sér. A* **267** (1968), 384–387. MR 38 #1093 Zbl 0165.05402

[Sathaye 1983] A. Sathaye, "Polynomial ring in two variables over a DVR: a criterion", *Invent. Math.* **74**:1 (1983), 159–168. MR 85j:14098 Zbl 0538.13006

[Seshadri 1972] C. S. Seshadri, "Quotient spaces modulo reductive algebraic groups", *Ann. of Math.* (2) **95** (1972), 511–556; errata, ibid. (2) **96**:3 (1972), 599. MR 46 #9044 Zbl 0241.14024

[SGA1 1971] A. Grothendieck et al., *Revêtements étales et groupe fondamental*, Lecture Notes in Math. **224**, Springer, Berlin, 1971. MR 50 #7129 Zbl 1039.14001

[Snow 1988] D. M. Snow, "Triangular actions on $\mathbb{C}^3$", *Manuscripta Math.* **60**:4 (1988), 407–415. MR 89e:32043 Zbl 0644.14018

[Winkelmann 1990] J. Winkelmann, "On free holomorphic $\mathbb{C}$-actions on $\mathbb{C}^n$ and homogeneous Stein manifolds", *Math. Ann.* **286**:1-3 (1990), 593–612. MR 90k:32094 Zbl 0708.32004

Communicated by Hubert Flenner
Received 2014-04-23 Accepted 2014-09-10

adrien.dubouloz@u-bourgogne.fr *CNRS, Institut de Mathématiques de Bourgogne,
Université de Bourgogne, 9 Avenue Alain Savary,
BP 47870, 21078 Dijon, France*

dfinston@nmsu.edu *Department of Mathematical Sciences, New Mexico
State University, Las Cruces, NM 88003, United States*

imad_jar@nmsu.edu *Department of Mathematical Sciences, Jordan University of
Science and Technology, P.O.Box 3030, Irbid 22110, Jordan*

ALGEBRA AND NUMBER THEORY 8:8 (2014)
dx.doi.org/10.2140/ant.2014.8.1985

Multivariate Apéry numbers and supercongruences of rational functions

Armin Straub

One of the many remarkable properties of the Apéry numbers $A(n)$, introduced in Apéry's proof of the irrationality of $\zeta(3)$, is that they satisfy the two-term supercongruences

$$A(p^r m) \equiv A(p^{r-1} m) \pmod{p^{3r}}$$

for primes $p \geqslant 5$. Similar congruences are conjectured to hold for all Apéry-like sequences. We provide a fresh perspective on the supercongruences satisfied by the Apéry numbers by showing that they extend to all Taylor coefficients $A(n_1, n_2, n_3, n_4)$ of the rational function

$$\frac{1}{(1-x_1-x_2)(1-x_3-x_4)-x_1x_2x_3x_4}.$$

The Apéry numbers are the diagonal coefficients of this function, which is simpler than previously known rational functions with this property.

Our main result offers analogous results for an infinite family of sequences, indexed by partitions λ, which also includes the Franel and Yang–Zudilin numbers as well as the Apéry numbers corresponding to $\zeta(2)$. Using the example of the Almkvist–Zudilin numbers, we further indicate evidence of multivariate supercongruences for other Apéry-like sequences.

1. Introduction

The *Apéry numbers*

$$A(n) = \sum_{k=0}^{n} \binom{n}{k}^2 \binom{n+k}{k}^2 \tag{1}$$

played a crucial role in R. Apéry's proof [Apéry 1979; van der Poorten 1979] of the irrationality of $\zeta(3)$, and have inspired much further work. Among many other interesting properties, they satisfy congruences with surprisingly large moduli,

MSC2010: primary 11A07; secondary 11B83, 11B37, 05A10.

Keywords: Apéry numbers, supercongruences, diagonals of rational functions, Almkvist–Zudilin numbers.

referred to as *supercongruences*, a term coined by F. Beukers [1985]. For instance, for all primes $p \geqslant 5$ and all positive integers r,

$$A(p^r m) \equiv A(p^{r-1} m) \pmod{p^{3r}}. \tag{2}$$

The special case $m = 1$, $r = 1$ was conjectured by S. Chowla, J. Cowles and M. Cowles [Chowla et al. 1980], who established the corresponding congruence modulo p^2. The case $r = 1$ was subsequently shown by I. Gessel [1982] and Y. Mimura [1983], while the general case was proved by M. Coster [1988]. The proof is an adaption of Beukers' [1985] proof of the related congruence

$$A(p^r m - 1) \equiv A(p^{r-1} m - 1) \pmod{p^{3r}}, \tag{3}$$

again valid for all primes $p \geqslant 5$ and all positive integers r. That congruence (3) can be interpreted as an extension of (2) to negative integers is explained in Remark 1.3. For further congruence properties of the Apéry numbers, we refer to [Cowles 1980; Beukers 1987; Ahlgren and Ono 2000; Kilbourn 2006].

Given a series

$$F(x_1, \ldots, x_d) = \sum_{n_1, \ldots, n_d \geqslant 0} a(n_1, \ldots, n_d) x_1^{n_1} \cdots x_d^{n_d}, \tag{4}$$

its *diagonal coefficients* are the coefficients $a(n, \ldots, n)$ and the *diagonal* is the ordinary generating function of the diagonal coefficients. For our purposes, F will always be a rational function. It is well-known (see, for instance, [Lipshitz and van der Poorten 1990, Theorem 5.2]) that the diagonal of a rational function satisfies a Picard–Fuchs linear differential equation, and as such "comes from geometry". In particular, the diagonal coefficients satisfy a linear recurrence with polynomial coefficients.

Many sequences of number-theoretic interest can be represented as the diagonal coefficients of rational functions. In particular, it is known [Christol 1984; Lipshitz and van der Poorten 1990] that the Apéry numbers are the diagonal coefficients of the rational function

$$\frac{1}{(1 - x_1)\big[(1 - x_2)(1 - x_3)(1 - x_4)(1 - x_5) - x_1 x_2 x_3\big]}. \tag{5}$$

Several other rational functions of which the Apéry numbers are the diagonal coefficients are given in [Bostan et al. 2013], where it is also discussed how these can be obtained from the representation of the Apéry numbers as the binomial sum (1). However, all of these rational function involve at least five variables and, in each case, the polynomial in the denominator factors. Our first result shows that in fact the Apéry numbers are the diagonal coefficients of a simpler rational function in only four variables.

Theorem 1.1. *The Apéry numbers* $A(n)$, *defined in* (1), *are the diagonal coefficients of*

$$\frac{1}{(1-x_1-x_2)(1-x_3-x_4)-x_1x_2x_3x_4}. \tag{6}$$

Representing a sequence as the diagonal of a rational function has certain benefits. For instance, asymptotic results can be obtained directly and explicitly from this rational function. This is the subject of *multivariate asymptotics*, as developed in [Pemantle and Wilson 2002]. For details and a host of worked examples we refer to [Pemantle and Wilson 2008]. As a second example, the rational generating function provides a means to compute the sequence modulo a fixed prime power. Indeed, the diagonal of a rational function with integral Taylor coefficients, such as (6), is algebraic modulo p^α for any α [Lipshitz and van der Poorten 1990]. A recent demonstration that this can be done very constructively is given in [Rowland and Yassawi 2013], where the values modulo p^α of sequences such as the Apéry numbers are, equivalently, encoded as finite automata.

We note that a statement such as Theorem 1.1 is more or less automatic to prove once discovered. For instance, given a rational function, we can always repeatedly employ a binomial series expansion to represent the Taylor coefficients as a nested sum of hypergeometric terms. In principle, creative telescoping [Petkovšek et al. 1996] will then obtain a linear recurrence satisfied by the diagonal coefficients, in which case it suffices to check that the alternative expression satisfies the same recurrence and agrees for sufficiently many initial values.

For the rational function $F(\boldsymbol{x})$ given in (6), we can gain considerably more insight. Indeed, for all the Taylor coefficients $A(\boldsymbol{n})$, defined by

$$F(x_1,x_2,x_3,x_4) = \sum_{n_1,n_2,n_3,n_4 \geqslant 0} A(n_1,n_2,n_2,n_4)\, x_1^{n_1}x_2^{n_2}x_3^{n_3}x_4^{n_4}, \tag{7}$$

we find, for instance by applying MacMahon's master theorem [1915, pp. 93–98] as detailed in Section 4, the explicit formula

$$A(\boldsymbol{n}) = \sum_{k\in\mathbb{Z}} \binom{n_1}{k}\binom{n_3}{k}\binom{n_1+n_2-k}{n_1}\binom{n_3+n_4-k}{n_3}, \tag{8}$$

of which Theorem 1.1 is an immediate consequence.

An instance of our main result is the observation that the supercongruence (2) for the Apéry numbers generalizes to all coefficients (8) of the rational function (6) in the following sense:

Theorem 1.2. *Let* $\boldsymbol{n} = (n_1, n_2, n_3, n_4) \in \mathbb{Z}^4$. *The coefficients* $A(\boldsymbol{n})$, *defined in* (7) *and extended to negative integers by* (8), *satisfy, for primes* $p \geqslant 5$ *and positive* *integers r, the supercongruences*

$$A(p^r\boldsymbol{n}) \equiv A(p^{r-1}\boldsymbol{n}) \pmod{p^{3r}}. \tag{9}$$

Note that the Apéry numbers are $A(n) = A(n, n, n, n)$, so that (9) indeed generalizes (2). Our reason for allowing negative entries in $\boldsymbol{n}$ is that by doing so, we also generalize Beukers' supercongruence (3). Indeed, as explained in Remark 1.3 below, $A(n-1) = A(-n, -n, -n, -n)$. Theorem 1.2 is a special case of our main result, Theorem 3.2, in which we prove such supercongruences for an infinite family of sequences. This family includes other Apéry-like sequences such as the Franel and Yang–Zudilin numbers, as well as the Apéry numbers corresponding to $\zeta(2)$.

We therefore review Apéry-like sequences in Section 2. Though no uniform reason is known, each Apéry-like sequence appears to satisfy a supercongruence of the form (2), some of which have been proved [Beukers 1985; Coster 1988; Chan et al. 2010; Osburn and Sahu 2011; 2013; Osburn et al. 2014] while others remain open [Osburn et al. 2014]. A major motivation for this note is to work towards an understanding of this observation. Our contribution to this question is the insight that, at least for several Apéry-like sequences, these supercongruences generalize to all coefficients of a rational function. Our main result, which includes the case of the Apéry numbers outlined in this introduction, is given in Section 3. In that section, we also record two further conjectural instances of this phenomenon. Finally, we provide proofs for our results in Sections 4 and 5.

Remark 1.3. Let us indicate that congruence (3) can be interpreted as the natural extension of (2) to the case of negative integers m. To see this, generalize the definition (1) of the Apéry numbers $A(n)$ to all integers n by setting

$$A(n) = \sum_{k\in\mathbb{Z}} \binom{n}{k}^2 \binom{n+k}{k}^2. \tag{10}$$

Here, we assume the values of the binomial coefficients to be defined as the (limiting) values of the corresponding quotient of gamma functions, that is,

$$\binom{n}{k} = \lim_{z\to 0} \frac{\Gamma(z+n+1)}{\Gamma(z+k+1)\Gamma(z+n-k+1)}.$$

Since $\Gamma(z+1)$ has no zeros, and poles only at negative integers z, one observes that the binomial coefficient $\binom{n}{k}$ is finite for all integers n and k. Moreover, the binomial coefficient with integer entries is nonzero only if $k \geqslant 0$ and $n-k \geqslant 0$, or if $n < 0$ and $k \geqslant 0$, or if $n < 0$ and $n-k \geqslant 0$. Note that in each of these cases $k \geqslant 0$ or $n-k \geqslant 0$, so that the symmetry $\binom{n}{k} = \binom{n}{n-k}$ allows us to compute these binomial coefficients in the obvious way. For instance, $\binom{-3}{-5} = \binom{-3}{2} = (-3)(-4)/2! = 6$. As carefully shown in [Sprugnoli 2008], for all integers n and k, we have the negation rule

$$\binom{n}{k} = \operatorname{sgn}(k)(-1)^k \binom{-n+k-1}{k}, \tag{11}$$

where $\operatorname{sgn}(k) = 1$ for $k \geqslant 0$ and $\operatorname{sgn}(k) = -1$ for $k < 0$. Applying (11) to the sum (10), we find that

$$A(-n) = A(n-1).$$

In particular, the congruence (3) is equivalent to (2) with $-m$ in place of m.

Remark 1.4. The proof of formula (8) in Section 4 shows that the coefficients can be expressed as

$$A(n_1, n_2, n_3, n_4) = \operatorname{ct} \frac{(x_1+x_2+x_3)^{n_1}(x_1+x_2)^{n_2}(x_3+x_4)^{n_3}(x_2+x_3+x_4)^{n_4}}{x_1^{n_1}x_2^{n_2}x_3^{n_3}x_4^{n_4}},$$

representing them as the constant terms of Laurent polynomials. In particular, the Apéry numbers (1) are the constant term of powers of a Laurent polynomial. Namely,

$$A(n) = \operatorname{ct}\left[\frac{(x_1+x_2)(x_3+1)(x_1+x_2+x_3)(x_2+x_3+1)}{x_1x_2x_3}\right]^n.$$

Since the Newton polyhedron of this Laurent polynomial has the origin as its only interior integral point, the results of [Samol and van Straten 2009; Mellit and Vlasenko 2013] apply to show that $A(n)$ satisfies the Dwork congruences

$$A(p^r m + n)A(\lfloor n/p \rfloor) \equiv A(p^{r-1}m + \lfloor n/p \rfloor)A(n) \pmod{p^r}$$

for all primes p and all integers $m, n \geqslant 0$, $r \geqslant 1$. In particular,

$$A(p^r m) \equiv A(p^{r-1}m) \pmod{p^r}, \tag{12}$$

which is a weaker version of (2) that holds for the large class of sequences represented as the constant term of powers of a Laurent polynomial, subject only to the condition on the Newton polyhedron. This gives another indication why congruence (2) is referred to as a supercongruence. It would be of considerable interest to find similarly well-defined classes of sequences for which supercongruences of the form (12) but modulo p^{kr} for $k > 1$ hold. Let us note that the case $r = 1$ of the Dwork congruences implies the Lucas congruences

$$A(n) \equiv A(n_0)A(n_1)\cdots A(n_\ell) \pmod{p},$$

where $n_0, \dots, n_\ell \in \{0, 1, \dots, p-1\}$ are the p-adic digits of $n = n_0 + n_1 p + \cdots + n_\ell p^\ell$. It is shown in [Rowland and Yassawi 2013] that Lucas congruences hold for all Taylor coefficients of certain rational functions. Additional divisibility properties in this direction are obtained in [Delaygue 2013] for Apéry-like numbers as well as for constant terms of powers of certain Laurent polynomials. Finally, we note that an extension of Dwork congruences to the multivariate setting has been considered in [Krattenthaler and Rivoal 2011]. In contrast to our approach, where, for instance, the Apéry numbers appear as the diagonal (multivariate) Taylor coefficients of a multivariate function $F(\boldsymbol{x})$, the theory developed in [Krattenthaler and Rivoal 2011] is concerned with functions $G(\boldsymbol{x}) = G(x_1, \dots, x_d)$ for which, say, the Apéry numbers are the (univariate) Taylor coefficients of the specialization $G(x, \dots, x)$.

2. Review of Apéry-like numbers

The Apéry numbers $A(n)$ are characterized by the 3-term recurrence

$$(n+1)^3 u_{n+1} = (2n+1)(an^2+an+b)u_n - n(cn^2+d)u_{n-1}, \tag{13}$$

where $(a, b, c, d) = (17, 5, 1, 0)$, together with the initial conditions

$$u_{-1} = 0, \quad u_0 = 1. \tag{14}$$

As explained in [Beukers 2002], the fact that in the recursion (13) we divide by $(n+1)^3$ at each step means that we should expect the denominator of u_n to grow like $(n!)^3$. While this is what happens for generic choice of the parameters (a, b, c, d), the Apéry numbers have the, from this perspective, exceptional property of being integral. Initiated by Beukers [2002], systematic searches have therefore been conducted for recurrences of this kind, which share the property of having an integer solution with initial conditions (14). This was done by D. Zagier [2009] for recurrences of the form

$$(n+1)^2 u_{n+1} = (an^2+an+b)u_n - cn^2 u_{n-1}, \tag{15}$$

by G. Almkvist and W. Zudilin [2006] for recurrences of the form (13) with $d = 0$ and, more recently, by S. Cooper [2012] for recurrences of the form (13). In each case, apart from degenerate cases, only finitely many sequences have been discovered. For details and a possibly complete list of the sequences, we refer to [Zagier 2009; Almkvist and Zudilin 2006; Almkvist et al. 2011; Cooper 2012].

Remarkably, and still rather mysteriously, all of these sequences, often referred to as *Apéry-like*, share some of the interesting properties of the Apéry numbers. For instance, they all are the coefficients of modular forms expanded in terms of a corresponding modular function. In the case of the Apéry numbers $A(n)$, for instance, it was shown by Beukers [1987] that

$$\sum_{n\geqslant 0} A(n) \left(\frac{\eta(\tau)\eta(6\tau)}{\eta(2\tau)\eta(3\tau)} \right)^{12n} = \frac{\eta^7(2\tau)\eta^7(3\tau)}{\eta^5(\tau)\eta^5(6\tau)}, \tag{16}$$

where $\eta(\tau)$ is the Dedekind eta function $\eta(\tau) = e^{\pi i \tau/12} \prod_{n\geqslant 1} (1 - e^{2\pi i n \tau})$. The modular function and the modular form appearing in (16) are modular with respect to the congruence subgroup $\Gamma_0(6)$ of level 6 (in fact, they are modular with respect to a slightly larger group). While this relation with modular forms can be proven in each individual case, no conceptual explanation is available, in the sense that if an additional Apéry-like sequence was found we would not know *a priori* that its generating function has a modular parametrization such as (16).

As a second example, it is conjectured and in some cases proven [Osburn et al. 2014] that each Apéry-like sequence satisfies a supercongruence of the form (2). Again, no uniform explanation is available and, the known proofs [Gessel 1982; Mimura 1983; Beukers 1985; Coster 1988] of the supercongruences (2) and (3) all rely on the explicit binomial representation (1) of the Apéry numbers. However, not all Apéry-like sequences have a comparably effective binomial representation so that, for instance, for the *Almkvist–Zudilin numbers* [Almkvist et al. 2011, Sequence (4.12)(δ); Chan and Zudilin 2010; Chan et al. 2010]

$$Z(n) = \sum_{k=0}^{n} (-3)^{n-3k} \binom{n}{3k} \binom{n+k}{n} \frac{(3k)!}{k!^3}, \tag{17}$$

which solve (13) with $(a, b, c, d) = (-7, -3, 81, 0)$, the supercongruence

$$Z(p^r m) \equiv Z(p^{r-1} m) \pmod{p^{3r}} \tag{18}$$

for primes $p \geqslant 3$ is conjectural only.

It would therefore be of particular interest to find alternative approaches to proving supercongruences. In this paper, we provide a new perspective on supercongruences of the form (18) by showing that they hold, at least for several Apéry-like sequences, for all coefficients $C(\boldsymbol{n})$ of a corresponding rational function, which has the sequence of interest as its diagonal coefficients. In such a case, one may then hope to use properties of the rational function to prove, for some $k > 1$, the supercongruence

$$C(p^r \boldsymbol{n}) \equiv C(p^{r-1} \boldsymbol{n}) \pmod{p^{kr}}.$$

For instance, for fixed p^r, these congruences can be proved, at least in principle, by computing the multivariate generating functions of both $C(p^r \boldsymbol{n})$ and $C(p^{r-1} \boldsymbol{n})$, which are rational functions because they are multisections of a rational function, and comparing them modulo p^{kr}.

Let us note that, in Example 3.9 below, we give a characterization of the Almkvist–Zudilin numbers (17) as the diagonal of a surprisingly simple rational function, and conjecture that the supercongruences (18), which themselves have not been proved yet, again extend to all coefficients of this rational function. We hope that the simplicity of the rational function might help inspire a proof of these supercongruences.

3. Main result and examples

We now generalize what we have illustrated in the introduction for the Apéry numbers $A(n)$ to an infinite family of sequences $A_{\lambda,\varepsilon}(n)$, indexed by partitions λ and $\varepsilon \in \{-1, 1\}$, which includes other Apéry-like numbers such as the Franel and Yang–Zudilin numbers as well as the sequence used by Apéry in relation with $\zeta(2)$. Our main theorem is Theorem 3.2, in which we prove (multivariate) supercongruences for this family of sequences, thus unifying and extending a number of known supercongruences. To begin with, the sequences we are concerned with are introduced by the following extension of formula (8). Here, $\boldsymbol{x}^{\boldsymbol{n}}$ is short for $x_1^{n_1} x_2^{n_2} \cdots x_d^{n_d}$.

Theorem 3.1. *Let $\alpha \in \mathbb{C}$ and $\lambda = (\lambda_1, \dots, \lambda_\ell) \in \mathbb{Z}_{>0}^\ell$ with $d = \lambda_1 + \cdots + \lambda_\ell$, and set $s(j) = \lambda_1 + \cdots + \lambda_{j-1}$. Then the Taylor coefficients of the rational function*

$$\left(\prod_{j=1}^{\ell} \left[1 - \sum_{r=1}^{\lambda_j} x_{s(j)+r} \right] - \alpha x_1 x_2 \cdots x_d \right)^{-1} = \sum_{\boldsymbol{n} \in \mathbb{Z}_{\geqslant 0}^d} A_{\lambda,\alpha}(\boldsymbol{n}) \boldsymbol{x}^{\boldsymbol{n}} \tag{19}$$

are given by

$$A_{\lambda,\alpha}(\boldsymbol{n}) = \sum_{k \in \mathbb{Z}} \alpha^k \prod_{j=1}^{\ell} \binom{n_{s(j)+1} + \cdots + n_{s(j)+\lambda_j} - (\lambda_j - 1)k}{n_{s(j)+1} - k, \dots, n_{s(j)+\lambda_j} - k, k}. \tag{20}$$

The proof of this elementary but crucial result will be given in Section 4. Observe that the multivariate Apéry numbers $A(\boldsymbol{n})$, defined in (8), are the special case $A_{(2,2),1}(\boldsymbol{n})$.

Our main result, of which Theorem 1.2 is the special case $\lambda = (2, 2)$ and $\varepsilon = 1$, follows next. Note that, if $\boldsymbol{n} \in \mathbb{Z}_{\geqslant 0}^d$, then the sum (20) defining $A_{\lambda,\alpha}(\boldsymbol{n})$ is finite and runs over $k = 0, 1, \dots, \min(n_1, \dots, n_d)$. On the other hand, if $\max(\lambda_1, \dots, \lambda_\ell) \geqslant 2$, then $A_{\lambda,\alpha}(\boldsymbol{n})$ is finite for any $\boldsymbol{n} \in \mathbb{Z}^d$.

Theorem 3.2. *Let $\varepsilon \in \{-1, 1\}$, $\lambda = (\lambda_1, \dots, \lambda_\ell) \in \mathbb{Z}_{>0}^\ell$, and assume that $\boldsymbol{n} \in \mathbb{Z}^d$, $d = \lambda_1 + \cdots + \lambda_\ell$ is such that $A_{\lambda,\varepsilon}(\boldsymbol{n})$, as defined in* (20), *is finite.*

(a) *If $\ell \geqslant 2$, then, for all primes $p \geqslant 3$ and integers $r \geqslant 1$,*

$$A_{\lambda,\varepsilon}(p^r \boldsymbol{n}) \equiv A_{\lambda,\varepsilon}(p^{r-1} \boldsymbol{n}) \pmod{p^{2r}}. \tag{21}$$

If $\varepsilon = 1$, then these congruences also hold for $p = 2$.

(b) *If $\ell \geqslant 2$ and* $\max(\lambda_1, \dots, \lambda_\ell) \leqslant 2$, *then, for primes $p \geqslant 5$ and integers $r \geqslant 1$,*

$$A_{\lambda,\varepsilon}(p^r \boldsymbol{n}) \equiv A_{\lambda,\varepsilon}(p^{r-1} \boldsymbol{n}) \pmod{p^{3r}}. \tag{22}$$

A proof of Theorem 3.2 is given in Section 5. One of the novel features of the proof, which is based on the approach of [Gessel 1982] and [Beukers 1985], is that it proceeds in a uniform fashion for all $\boldsymbol{n} \in \mathbb{Z}^d$. As outlined in Remark 1.3, this allows us to also conclude, and to a certain extent explain, the shifted supercongruences (3), which, among Apéry-like numbers, are special to the Apéry numbers as well as their version (23) related to $\zeta(2)$. In cases where $\boldsymbol{n}$ has negative entries, the summation (20), while still finite, may include negative values for k (see Remark 1.3). We therefore extend classical results, such as Jacobsthal's binomial congruences, to the case of binomial coefficients with negative entries.

Example 3.3. For $\lambda = (2)$, the numbers (20) specialize to the *Delannoy numbers*

$$A_{(2),1}(\boldsymbol{n}) = \sum_{k\in\mathbb{Z}} \binom{n_1}{k}\binom{n_1+n_2-k}{n_1},$$

which, for $n_1, n_2 \geqslant 0$, count the number of lattice paths from $(0, 0)$ to (n_1, n_2) with steps $(1, 0)$, $(0, 1)$ and $(1, 1)$. The Delannoy numbers do not satisfy (21) or (22), thus demonstrating the necessity of the condition $\ell \geqslant 2$ in Theorem 3.2. They do satisfy (21) modulo p^r, by virtue of Remark 1.4.

Example 3.4. The Apéry-like sequence

$$B(n) = \sum_{k\in\mathbb{Z}} \binom{n}{k}^2\binom{n+k}{k}, \tag{23}$$

which satisfies recurrence (15) with $(a, b, c) = (11, 3, -1)$, was introduced by Apéry [Apéry 1979; van der Poorten 1979] along with (1) and used to (re)prove the irrationality of $\zeta(2)$. By Theorem 3.1 with $\lambda = (2, 1)$ and $\varepsilon = 1$, the numbers $B(n)$ are the diagonal coefficients of the rational function

$$\frac{1}{(1-x_1-x_2)(1-x_3)-x_1x_2x_3} = \sum_{\boldsymbol{n}\in\mathbb{Z}^3_{\geqslant 0}} B(\boldsymbol{n})\boldsymbol{x}^{\boldsymbol{n}}. \tag{24}$$

In addition to the binomial sum for $B(\boldsymbol{n})$ given by Theorem 3.1, MacMahon's master theorem (Theorem 4.1) shows that $B(n_1, n_2, n_3)$ is the coefficient of $x_1^{n_1}x_2^{n_2}x_3^{n_3}$ in the product $(x_1+x_2+x_3)^{n_1}(x_1+x_2)^{n_2}(x_2+x_3)^{n_3}$. An application of Theorem 3.2 shows that, for $\boldsymbol{n} \in \mathbb{Z}^3$ and integers $r \geqslant 1$, the supercongruences

$$B(p^r\boldsymbol{n}) \equiv B(p^{r-1}\boldsymbol{n}) \pmod{p^{3r}} \tag{25}$$

hold for all primes $p \geqslant 5$. In the diagonal case $n_1 = n_2 = n_3$, this result was first proved by Coster [1988].

Proceeding as in Remark 1.3, and using the curious identity

$$\sum_{k=0}^{n}\binom{n}{k}^2\binom{n+k}{k} = \sum_{k=0}^{n}(-1)^{n+k}\binom{n}{k}\binom{n+k}{k}^2, \tag{26}$$

we find that $B(-n) = (-1)^{n-1}B(n-1)$ for $n > 0$. Consequently, (25) implies the shifted supercongruences $B(p^r m - 1) \equiv B(p^{r-1}m - 1)$, which hold modulo p^{3r} for all primes $p \geqslant 5$ and were first proved in [Beukers 1985], along with (3). We observe that, among the known Apéry-like numbers, the sequence $B(n)$ and the Apéry numbers (1) are the only ones to satisfy shifted supercongruences of the form (3) in addition to the supercongruences of the form (2).

Example 3.5. As a result of Theorem 3.1 with $\lambda = (3, 1)$ and $\varepsilon = 1$, the numbers

$$C(n) = \sum_{k=0}^{n} \binom{n}{k}^2 \binom{n+k}{k} \binom{n+2k}{k}$$

are the diagonal coefficients of the rational function

$$\frac{1}{(1 - x_1 - x_2 - x_3)(1 - x_4) - x_1x_2x_3x_4}.$$

By Theorem 3.2, it follows that $C(p^r n) \equiv C(p^{r-1}n)$ modulo p^{2r}, for all primes p. We note that this congruence does not, in general, hold modulo a larger power of p, as is illustrated by $C(5) = 4{,}009{,}657 \not\equiv 7 = C(1)$ modulo 5^3. This demonstrates that in Theorem 3.2(a) the modulus p^{2r} of the congruences cannot, in general, be replaced with p^{3r}, even for $p \geqslant 5$.

Example 3.6. Next, we consider the sequences

$$Y_d(n) = \sum_{k=0}^{n} \binom{n}{k}^d. \tag{27}$$

The numbers $Y_3(n)$ satisfy the recurrence (15) with $(a, b, c) = (7, 2, -8)$ and are known as *Franel numbers* [1894], while the numbers $Y_4(n)$, corresponding to $(a, b, c, d) = (6, 2, -64, 4)$ in (13), are sometimes referred to as *Yang–Zudilin numbers* [Chan et al. 2010]. It follows from Theorem 3.1 with $\lambda = (1, 1, \dots, 1)$ and $\varepsilon = 1$ that

$$\frac{1}{(1-x_1)(1-x_2)\cdots(1-x_d) - x_1x_2\cdots x_d} = \sum_{\boldsymbol{n}\in\mathbb{Z}^d_{\geqslant 0}} Y_d(\boldsymbol{n})\boldsymbol{x}^{\boldsymbol{n}}, \tag{28}$$

where

$$Y_d(\boldsymbol{n}) = \sum_{k\geqslant 0} \binom{n_1}{k}\binom{n_2}{k}\cdots\binom{n_d}{k}. \tag{29}$$

It is proved in [Chan et al. 2010] that $Y_d(pn) \equiv Y_d(n)$ modulo p^3 for primes $p \geqslant 5$ if $d \geqslant 2$. These congruences are generalized to the multivariate setting by Theorem 3.2, which shows that, if $d \geqslant 2$, then, for $\boldsymbol{n} \in \mathbb{Z}^d_{\geqslant 0}$ and integers $r \geqslant 1$,

$$Y_d(p^r\boldsymbol{n}) \equiv Y_d(p^{r-1}\boldsymbol{n}) \pmod{p^{3r}} \tag{30}$$

for primes $p \geqslant 5$. Note that

$$Y_2(\boldsymbol{n}) = \sum_{k \in \mathbb{Z}} \binom{n_1}{k}\binom{n_2}{k} = \binom{n_1+n_2}{n_1}.$$

Hence, congruence (30) includes, in particular, the appealing binomial congruence

$$\binom{pa}{pb} \equiv \binom{a}{b} \pmod{p^3},$$

which is attributed to W. Ljunggren [Granville 1997] and which generalizes the classical congruences by C. Babbage, J. Wolstenholme and J. W. L. Glaisher. It is further refined by E. Jacobsthal's binomial congruence, which we review in Lemma 5.1 and which the proof of Theorem 3.2 crucially depends on.

Let us conclude this section with two conjectural examples, which suggest that our results are not an isolated phenomenon.

Example 3.7. As noted in the introduction for the Apéry numbers, there is no unique rational function of which a given sequence is the diagonal. For instance, the Franel numbers $Y_3(n)$ are also the diagonal coefficients of the rational function

$$\frac{1}{1-(x_1+x_2+x_3)+4x_1x_2x_3}. \tag{31}$$

A rational function $F(\boldsymbol{x})$ is said to be *positive* if its Taylor coefficients (4) are all positive. The Askey–Gasper rational function (31), whose positivity is proved in [Askey and Gasper 1977] and [Gillis et al. 1983], is an interesting instance of a rational function on the boundary of positivity (if the 4 is replaced by $4+\varepsilon$, for any $\varepsilon > 0$, then the resulting rational function is not positive). The present work was, in part, motivated by the observation [Straub and Zudilin 2014] that for several of the rational functions, which have been shown or conjectured to be on the boundary of positivity, the diagonal coefficients are arithmetically interesting sequences with links to modular forms. Note that the Askey–Gasper rational function (31) corresponds to the choice $\lambda = (3)$ and $\alpha = -4$ in Theorem 3.1, which makes its Taylor coefficients $G(\boldsymbol{n}) = A_{(3),-4}(\boldsymbol{n})$ explicit. We also note that an application of MacMahon's master theorem (Theorem 4.1) shows that $G(n_1, n_2, n_3)$ is the coefficient of $x_1^{n_1}x_2^{n_2}x_3^{n_3}$ in the product $(x_1-x_2-x_3)^{n_1}(x_2-x_1-x_3)^{n_2}(x_3-x_1-x_2)^{n_3}$. Although it is unclear how one might adjust the proof of Theorem 3.2, numerical evidence suggests that the coefficients $G(\boldsymbol{n})$ satisfy supercongruences modulo p^{3r} as well.

Conjecture 3.8. The coefficients $G(\boldsymbol{n})$ of the rational function (31) satisfy, for primes $p \geqslant 5$ and integers $r \geqslant 1$,

$$G(p^r \boldsymbol{n}) \equiv G(p^{r-1}\boldsymbol{n}) \pmod{p^{3r}}.$$

Example 3.9. Remarkably, the previous example has a four-variable analog, which involves the Almkvist–Zudilin numbers $Z(n)$, introduced in (17). Namely, the numbers $Z(n)$ are the diagonal coefficients of the unexpectedly simple rational function

$$\frac{1}{1-(x_1+x_2+x_3+x_4)+27x_1x_2x_3x_4}, \tag{32}$$

as can be deduced from Theorem 3.1 with $\lambda = (4)$ and $\alpha = -27$. Again, numerical evidence suggests that the coefficients $Z(\boldsymbol{n})$ of (32) satisfy supercongruences modulo p^{3r}. This is particularly interesting, since even the univariate congruences (18) are conjectural at this time.

Conjecture 3.10. The coefficients $Z(\boldsymbol{n})$ of the rational function (32) satisfy, for primes $p \geqslant 5$ and integers $r \geqslant 1$,

$$Z(p^r\boldsymbol{n}) \equiv Z(p^{r-1}\boldsymbol{n}) \pmod{p^{3r}}.$$

Remark 3.11. The rational functions (31) and (32) involved in the previous examples make it natural to wonder whether supercongruences might similarly exist for the family of rational functions given by

$$\frac{1}{1-(x_1+x_2+\cdots+x_d)+(d-1)^{d-1}x_1x_2\cdots x_d}.$$

This does not, however, appear to be the case for $d \geqslant 5$. In fact, no value $b \neq 0$ in

$$\frac{1}{1-(x_1+x_2+\cdots+x_d)+bx_1x_2\cdots x_d}$$

appears to give rise to supercongruences (by computing coefficients, we have ruled out supercongruences modulo p^{2r} for integers $|b| < 100{,}000$ and $d \leqslant 25$).

4. The Taylor coefficients

This section is devoted to proving Theorem 3.1. Before we give a general proof, we offer an alternative approach based on MacMahon's master theorem, to which we refer at several occasions in this note and which offers additional insight into the Taylor coefficients by expressing them as coefficients of certain polynomials (see also Remark 1.4). This approach, which we apply here to prove formula (8), is based on the following result of P. MacMahon [1915], coined by himself "a master theorem in the Theory of Permutations". Here, $[\boldsymbol{x}^{\boldsymbol{m}}]$ denotes the coefficient of $x_1^{m_1}\cdots x_n^{m_n}$ in the expansion of what follows.

Theorem 4.1. *For $\boldsymbol{x}=(x_1,\dots,x_n)$, matrices $A\in\mathbb{C}^{n\times n}$ and $\boldsymbol{m}=(m_1,\dots,m_n)\in\mathbb{Z}^n_{\geqslant 0}$,*

$$[\boldsymbol{x}^{\boldsymbol{m}}]\prod_{i=1}^{n}\biggl(\sum_{j=1}^{n}A_{i,j}x_j\biggr)^{m_i}=[\boldsymbol{x}^{\boldsymbol{m}}]\frac{1}{\det(I_n-AX)},$$

where X is the diagonal $n\times n$ matrix with entries $x_1,\dots,x_n$.

Proof of formula (8). We note that

$$\frac{1}{(1-x_1-x_2)(1-x_3-x_4)-x_1x_2x_3x_4}=\frac{1}{\det(I_4-MX)},$$

where M and X are the matrices

$$M=\begin{pmatrix}1&1&1&0\\1&1&0&0\\0&0&1&1\\0&1&1&1\end{pmatrix},\quad X=\begin{pmatrix}x_1&&&\\&x_2&&\\&&x_3&\\&&&x_4\end{pmatrix}.$$

An application of MacMahon's master theorem therefore shows that the coefficients $A(\boldsymbol{n})$, with $\boldsymbol{n}=(n_1,n_2,n_3,n_4)$, are given by

$$A(\boldsymbol{n})=[\boldsymbol{x}^{\boldsymbol{n}}](x_1+x_2+x_3)^{n_1}(x_1+x_2)^{n_2}(x_3+x_4)^{n_3}(x_2+x_3+x_4)^{n_4}.$$

In order to extract the requisite coefficient, we expand the right-hand side as

$$\begin{aligned}&(x_1+x_2+x_3)^{n_1}(x_1+x_2)^{n_2}(x_3+x_4)^{n_3}(x_2+x_3+x_4)^{n_4}\\&=\sum_{k_1,k_4}\binom{n_1}{k_1}\binom{n_4}{k_4}x_2^{n_4-k_4}x_3^{n_1-k_1}(x_1+x_2)^{k_1+n_2}(x_3+x_4)^{n_3+k_4}\\&=\sum_{k_1,k_2,k_3,k_4}\binom{n_1}{k_1}\binom{n_4}{k_4}\binom{k_1+n_2}{k_2}\binom{n_3+k_4}{k_3}x_1^{k_1+n_2-k_2}x_2^{n_4-k_4+k_2}x_3^{n_1-k_1+k_3}x_4^{n_3+k_4-k_3}.\end{aligned}$$

The summand contributes to $x_1^{n_1}x_2^{n_2}x_3^{n_3}x_4^{n_4}$ if and only if $n_i-k_i=n_j-k_j$ for all $i,j=1,\dots,4$. Writing $k=n_i-k_i$ for the common value, we obtain

$$A(n_1,n_2,n_3,n_4)=\sum_{k\in\mathbb{Z}}\binom{n_1}{k}\binom{n_4}{k}\binom{n_1-k+n_2}{n_2-k}\binom{n_3+n_4-k}{n_3-k},$$

which is equivalent to the claimed (8). □

Proof of Theorem 3.1. Recall the elementary formula

$$\frac{1}{(1-x)^{k+1}}=\sum_{n\geqslant 0}\binom{n+k}{k}x^n,$$

for integers $k \geqslant 0$. Combined with an application of the multinomial theorem, it implies that

$$\frac{1}{(1-x_1-\cdots-x_\rho)^{k+1}} = \sum_{n_1\geqslant 0}\cdots\sum_{n_\rho\geqslant 0}\binom{n_1+\cdots+n_\rho+k}{n_1,\ldots,n_\rho,k}x_1^{n_1}\cdots x_\rho^{n_\rho},$$

and hence

$$\frac{(x_1\cdots x_\rho)^k}{(1-x_1-\cdots-x_\rho)^{k+1}} = \sum_{n_1\geqslant 0}\cdots\sum_{n_\rho\geqslant 0}\binom{n_1+\cdots+n_\rho-(\rho-1)k}{n_1-k,\ldots,n_\rho-k,k}x_1^{n_1}\cdots x_\rho^{n_\rho}. \quad (33)$$

Here, we used that the multinomial coefficient vanishes if $k > \min(n_1,\ldots,n_\rho)$. Geometrically expanding the left-hand side of (19), we find that

$$\left(\prod_{j=1}^{\ell}\left[1-\sum_{r=1}^{\lambda_j}x_{s(j)+r}\right]-\alpha x_1x_2\cdots x_d\right)^{-1} = \sum_{k\geqslant 0}\alpha^k\prod_{j=1}^{\ell}\frac{(x_{s(j)+1}\cdots x_{s(j)+\lambda_j})^k}{\left[1-\sum_{r=1}^{\lambda_j}x_{s(j)+r}\right]^{k+1}},$$

which we further expand using (33) to get

$$\sum_{k\geqslant 0}\alpha^k\sum_{\boldsymbol{n}\in\mathbb{Z}_{\geqslant 0}^d}\boldsymbol{x}^{\boldsymbol{n}}\prod_{j=1}^{\ell}\binom{n_{s(j)+1}+\cdots+n_{s(j)+\lambda_j}-(\lambda_j-1)k}{n_{s(j)+1}-k,\ldots,n_{s(j)+\lambda_j}-k,k} = \sum_{\boldsymbol{n}\in\mathbb{Z}_{\geqslant 0}^d}A_{\lambda,\alpha}(\boldsymbol{n})\boldsymbol{x}^{\boldsymbol{n}},$$

with $A_{\lambda,\alpha}(\boldsymbol{n})$ as in (20). □

5. The supercongruences

Our proof of Theorem 3.2, which generalizes the supercongruence in Theorem 1.2, builds upon the respective proofs in [Gessel 1982] and [Beukers 1985].

We need a number of lemmas in preparation. To begin with, we prove the following extension of Jacobsthal's binomial congruence [Gessel 1983; Granville 1997] to binomial coefficients which are allowed to have negative entries (see Remark 1.3).

Lemma 5.1. *For all primes p and all integers a, b,*

$$\binom{ap}{bp}\Big/\binom{a}{b} \equiv \varepsilon \pmod{p^q}, \quad (34)$$

where q is the power of p dividing $p^3ab(a-b)/12$ and where $\varepsilon = 1$, unless $p = 2$ and $(a, b) \equiv (0, 1)$ modulo 2, in which case $\varepsilon = -1$.

Proof. Congruence (34), for nonnegative a, b, is proved in [Gessel 1983] (alternatively, a proof for $p \geqslant 5$ is given in [Granville 1997]). We therefore only indicate how to extend (34) to negative values of a or b. Note that, for all $a, b \in \mathbb{Z}$ with $b \neq 0$,

$$\binom{a}{b} = \frac{a}{b}\binom{a-1}{b-1},$$

and hence

$$\binom{ap}{bp}\Big/\binom{a}{b}=\binom{ap-1}{bp-1}\Big/\binom{a-1}{b-1}.$$

We claim that the extension of (34) to the case $a<0$ and $b<0$ therefore follows from

$$\binom{a}{b}=\binom{-b-1}{-a-1}(-1)^{a-b}\operatorname{sgn}(a-b), \tag{35}$$

where sgn is defined as in Remark 1.3. This is clear for $p\geqslant 3$. Write $\varepsilon(a,b)=-1$ if $(a,b)\equiv(0,1)$ modulo 2 and $\varepsilon(a,b)=1$ otherwise. It is straightforward to check that

$$(-1)^{a-b}\varepsilon(-b,-a)=\varepsilon(a,b),$$

which shows the case $p=2$.

Similarly, if $a<0$ and $b>0$, then we may apply

$$\binom{a}{b}=\binom{b-a-1}{-a-1}(-1)^{b+1}\operatorname{sgn}(a-b)\operatorname{sgn}(-a-1)$$

as well as

$$(-1)^{b}\varepsilon(b-a,-a)=\varepsilon(a,b).$$

A derivation of the above binomial identities, which are valid for all $a,b\in\mathbb{Z}$, may be found in [Sprugnoli 2008]. □

Much simpler and better known is the following congruence:

Lemma 5.2. *Let $p\geqslant 5$ be a prime, and $\varepsilon\in\{-1,1\}$. Then, for all integers $r\geqslant 0$,*

$$\sum_{k=1,\,p\nmid k}^{p^r-1}\frac{\varepsilon^k}{k^2}\equiv 0 \pmod{p^r}. \tag{36}$$

Proof. Let α be an odd integer, not divisible by p, such that $\alpha^2\not\equiv 1$ modulo p (take, for instance, $\alpha=3$). Then,

$$\frac{1}{\alpha^2}\sum_{k=1,\,p\nmid k}^{p^r-1}\frac{\varepsilon^k}{k^2}=\sum_{k=1,\,p\nmid k}^{p^r-1}\frac{\varepsilon^k}{(\alpha k)^2}\equiv\sum_{k=1,\,p\nmid k}^{p^r-1}\frac{\varepsilon^k}{k^2}\pmod{p^r},$$

since the second and third sum run over the same residues modulo p^r (note that $\varepsilon^{\alpha k}=\varepsilon^k$ since α is odd). As α^2 is not divisible by p, the congruence (36) follows. □

The next lemmas establish properties of the summands of the numbers $A_{\lambda,\varepsilon}(\boldsymbol{n})$ as introduced in (20), which will be needed in our proof of Theorem 3.2. Throughout this section, we fix the notation of Theorem 3.2, letting $\lambda=(\lambda_1,\dots,\lambda_\ell)\in\mathbb{Z}_{>0}^{\ell}$ with $d=\lambda_1+\cdots+\lambda_\ell$, and setting $s(j)=\lambda_1+\cdots+\lambda_{j-1}$.

Lemma 5.3. *Let* $\boldsymbol{n} \in \mathbb{Z}^d$, $k \in \mathbb{Z}$, *and define*

$$A_\lambda(\boldsymbol{n}; k) = \prod_{j=1}^{\ell} \binom{n_{s(j)+1} + \cdots + n_{s(j)+\lambda_j} - (\lambda_j - 1)k}{n_{s(j)+1} - k, \ldots, n_{s(j)+\lambda_j} - k, k}. \tag{37}$$

(a) *If* $\ell \geqslant 2$, *then, for all primes* p *and integers* $r \geqslant 1$,

$$A_\lambda(p^r \boldsymbol{n}; pk) \equiv A_\lambda(p^{r-1}\boldsymbol{n}; k) \pmod{p^{2r}}. \tag{38}$$

(b) *If* $\ell \geqslant 2$ *and* $\max(\lambda_1, \ldots, \lambda_\ell) \leqslant 2$, *then, for primes* $p \geqslant 5$ *and integers* $r \geqslant 1$,

$$A_\lambda(p^r \boldsymbol{n}; pk) \equiv A_\lambda(p^{r-1}\boldsymbol{n}; k) \pmod{p^{3r}}. \tag{39}$$

Proof. We show (38) and (39) by proving that for integers $r, s \geqslant 1$ and k such that $p \nmid k$,

$$A_\lambda(p^r \boldsymbol{n}; p^s k) \equiv A_\lambda(p^{r-1}\boldsymbol{n}; p^{s-1}k) \pmod{p^{\alpha r}}, \tag{40}$$

where $\alpha = 2$ or $\alpha = 3$ depending on whether $\max(\lambda_1, \ldots, \lambda_\ell) \leqslant 2$.

Let us first consider the case $\ell \geqslant 2$ and $\max(\lambda_1, \ldots, \lambda_\ell) \leqslant 2$. Then each factor of (37) is a single binomial, if $\lambda_j = 1$, or of the form

$$\binom{m_1}{k}\binom{m_1+m_2-k}{m_1},$$

if $\lambda_j = 2$. Let p be a prime such that $p \geqslant 5$. It follows from Jacobsthal's congruence (34) that

$$\binom{p^r m_1}{p^s k} \Big/ \binom{p^{r-1} m_1}{p^{s-1} k} \equiv 1 \pmod{p^{r+s+\min(r,s)}}$$

as well as

$$\binom{p^r(m_1+m_2) - p^s k}{p^r m_1} \Big/ \binom{p^{r-1}(m_1+m_2) - p^{s-1} k}{p^{r-1} m_1} \equiv 1 \pmod{p^{r+2\min(r,s)}}.$$

Consequently,

$$A_\lambda(p^r \boldsymbol{n}; p^s k) = c A_\lambda(p^{r-1}\boldsymbol{n}; p^{s-1}k) \tag{41}$$

with $c \equiv 1$ modulo $p^{r+2\min(r,s)}$. If $s \geqslant r$, this proves congruence (40) with $\alpha = 3$. On the other hand, suppose $s \leqslant r$. Since $p \nmid k$, we have

$$\binom{p^r n}{p^s k} = p^{r-s} \frac{n}{k} \binom{p^r n - 1}{p^s k - 1} \equiv 0 \pmod{p^{r-s}}.$$

Since $\ell \geqslant 2$, it follows that $p^{2(r-s)}$ divides $A_\lambda(p^r \boldsymbol{n}; p^s k)$. Since $(r+2s)+2(r-s) = 3r$, the congruence (40), with $\alpha = 3$, now follows from (41). This shows (b).

Let us now turn to the proof of (a). Assume that $\ell \geqslant 2$. For any positive integer ρ,

$$\binom{m_1 + \cdots + m_\rho - (\rho-1)k}{m_1 - k, \ldots, m_\rho - k, k} = \binom{m_1}{k}\binom{m_1 + (m_2-k) + \cdots + (m_\rho - k)}{m_1, m_2-k, \ldots, m_\rho - k},$$

so that, as in the previous case, $p^{\ell(r-s)}$ divides $A_\lambda(p^r\boldsymbol{n};\, p^s k)$ if $r \geqslant s$.

Initially, assume that $p \geqslant 3$. By further unraveling the multinomial coefficient as a product of binomial coefficients and applying Jacobsthal's congruence (34) as above, we find that

$$A_\lambda(p^r\boldsymbol{n};\, p^s k) = cA_\lambda(p^{r-1}\boldsymbol{n};\, p^{s-1}k)$$

with $c \equiv 1$ modulo $p^{3\min(r,s)-\delta}$, and $\delta = 0$ if $p \geqslant 5$ and $\delta = 1$ if $p = 3$. In light of $p^{2(r-s)}$ dividing $A_\lambda(p^r\boldsymbol{n};\, p^s k)$ if $r \geqslant s$, we conclude congruence (40) with $\alpha = 2$.

Now, consider $p = 2$. If $r \geqslant 2$ and $s \geqslant 2$, then the sign ε in Jacobsthal's congruence (34) is always $+1$ when applying the above approach, and we again find that (40) holds with $\alpha = 2$. On the other hand, if $r = 1$, then it suffices to use the (combinatorial) congruence

$$\binom{pa}{pb} \equiv \binom{a}{b} \pmod{p^2},$$

which holds for all primes p. It remains to consider the case $r \geqslant 2$ and $s = 1$. Applying the approach employed for $p \geqslant 3$, we find that

$$A_\lambda(p^r\boldsymbol{n};\, p^s k) = cA_\lambda(p^{r-1}\boldsymbol{n};\, p^{s-1}k), \tag{42}$$

where $c \equiv \pm 1$ modulo $p^{3\min(r,s)-2} = 2$. If $\max(\lambda_1, \dots, \lambda_\ell) \leqslant 2$, then we, in fact, have $c \equiv (-1)^\ell$ modulo $p^{r+2\min(r,s)-2} = 2^r$. Since $A_\lambda(p^r\boldsymbol{n};\, p^s k)$ is divisible by $p^{\ell(r-1)}$, congruence (40) trivially holds with $\alpha = 2$ if $\ell \geqslant 3$. Hence, we may assume that $\ell = 2$. If $\max(\lambda_1, \lambda_2) \leqslant 2$, then $c \equiv 1$ modulo 2^r in (42) and, since both sides of (42) are divisible by 2^{2r-2}, congruence (40) with $\alpha = 2$ again follows. Finally, suppose that there is j such that $\lambda_j \geqslant 3$. Then the factor corresponding to j in (37) is of the form

$$\binom{m_1}{k}\binom{m_1+m_2-k}{m_1}\binom{m_1+m_2+m_3-2k}{m_3-k}\binom{m_1+\cdots+m_\rho-(\rho-1)k}{m_1+m_2+m_3-2k,\, m_4-k, \dots}.$$

Note that for even m_1, m_2, m_3 and odd k, the third binomial in this product is even. Hence, $A_\lambda(p^r\boldsymbol{n};\, p^s k)$ is divisible by $2^{2(r-1)+1} = 2^{2r-1}$. In light of (42), this proves congruence (40) with $\alpha = 2$. □

The next congruence, with $k \geqslant 0$, has been used in [Beukers 1985]. For our present purpose, we extend it to the case of negative k.

Lemma 5.4. *For primes p, integers m, k and integers $r \geqslant 1$,*

$$\binom{p^r m-1}{k}(-1)^k \equiv \binom{p^{r-1}m-1}{[k/p]}(-1)^{[k/p]} \pmod{p^r}. \tag{43}$$

Proof. First, assume that $k \geqslant 0$. Following [Beukers 1985, Lemma 2], we split the defining product of the binomial coefficient, according to whether the index is divisible by p or not, to obtain

$$\binom{p^r m-1}{k} = \prod_{j=1}^{k} \frac{p^r m - j}{j} = \prod_{j=1, p\nmid j}^{k} \frac{p^r m - j}{j} \prod_{\lambda=1}^{[k/p]} \frac{p^{r-1}m - \lambda}{\lambda}$$
$$= \binom{p^{r-1}m-1}{[k/p]} \prod_{j=1,\, p\nmid j}^{k} \frac{p^r m - j}{j}.$$

Congruence (43), with $k \geqslant 0$, follows upon reducing modulo p^r.

On the other hand, assume $k < 0$. Since (43) is trivial if $m > 0$, we let $m \leqslant 0$. We use the basic symmetry relation

$$\binom{p^r m-1}{k} = \binom{p^r m-1}{p^r m-k-1}$$

and note that, since $k < 0$, the binomials are zero unless $p^r m - k - 1 \geqslant 0$. Observe that for all integers k, m,

$$[(p^r m - k - 1)/p] = p^{r-1}m + [-(k+1)/p] = p^{r-1}m - [k/p] - 1. \qquad (44)$$

Thus, assuming $p^r m - 1 - k \geqslant 0$, we may apply (43) to find

$$\binom{p^r m-1}{k}(-1)^k = \binom{p^r m-1}{p^r m-k-1}(-1)^k$$
$$\equiv \binom{p^{r-1}m-1}{p^{r-1}m-[k/p]-1}(-1)^{[k/p]}(-1)^{p^r m+p^{r-1}m}$$
$$= \binom{p^{r-1}m-1}{[k/p]}(-1)^{[k/p]}(-1)^{p^r m+p^{r-1}m} \pmod{p^r}.$$

It only remains to note that $p^r m + p^{r-1}m = p^{r-1}(p+1)m$ is even unless $p = 2$ and $r = 1$. Hence, in all cases, $(-1)^{p^r m+p^{r-1}m} \equiv 1$ modulo p^r. □

Lemma 5.5. *For primes p, integers m_1, m_2, k and integers $r \geqslant 1$,*

$$\binom{p^r m_1+p^r m_2-k-1}{p^r m_1} \equiv \binom{p^{r-1}m_1+p^{r-1}m_2-[k/p]-1}{p^{r-1}m_1} \pmod{p^r}.$$

Proof. By an application of (11),

$$\binom{m_1+m_2-k-1}{m_1} = \operatorname{sgn}(m_2-k-1)(-1)^{m_2-k-1}\binom{-m_1-1}{m_2-k-1}.$$

Since for all $a \in \mathbb{Z}$, $\operatorname{sgn}(a) = \operatorname{sgn}([a/p])$, the claimed congruence therefore follows from (44) and Lemma 5.4. □

The following generalizes [Beukers 1985, Lemma 3] to our needs:

Lemma 5.6. *Let p be a prime and $\boldsymbol{n} \in \mathbb{Z}^d$.*

- *Let $a_k \in \mathbb{Z}_p$, with $k \in \mathbb{Z}$, be such that, for all $l, s \in \mathbb{Z}$ with $s \geqslant 0$,*
$$\sum_{[k/p^s]=l} a_k \equiv 0 \pmod{p^s}.$$
- *Let $C(\boldsymbol{n}; k)$ be such that, for all $k, r \in \mathbb{Z}$ with $r \geqslant 0$,*
$$C(p^r \boldsymbol{n}; k) \equiv C(p^{r-1}\boldsymbol{n}; [k/p]) \pmod{p^r}. \tag{45}$$

Then, for all $r, l \in \mathbb{Z}$ with $r \geqslant 0$,
$$\sum_{[k/p^r]=l} a_k C(p^r \boldsymbol{n}; k) \equiv 0 \pmod{p^r}. \tag{46}$$

Proof. The claim is trivial for $r = 0$. Fix $r > 0$ and assume, for the purpose of induction on r, that the congruence (46) holds for the exponent $r - 1$ in place of r. By the assumption (45) on $C(\boldsymbol{n}; k)$, we have that, modulo p^r,
$$\begin{aligned}\sum_{[k/p^r]=l} a_k C(p^r \boldsymbol{n}; k) &\equiv \sum_{[k/p^r]=l} a_k C(p^{r-1}\boldsymbol{n}; [k/p]) \\ &= \sum_{[m/p^{r-1}]=l} \left(\sum_{[k/p]=m} a_k \right) C(p^{r-1}\boldsymbol{n}; m) \\ &= p \sum_{[m/p^{r-1}]=l} b_m C(p^{r-1}\boldsymbol{n}; m),\end{aligned}$$
where b_m is the sequence
$$b_m = \frac{1}{p} \sum_{[k/p]=m} a_k.$$

We note that, for all $s, l \in \mathbb{Z}$ with $s \geqslant 0$,
$$\sum_{[m/p^s]=l} b_m = \frac{1}{p} \sum_{[m/p^s]=l} \sum_{[k/p]=m} a_k = \frac{1}{p} \sum_{[k/p^{s+1}]=m} a_k \equiv 0 \pmod{p^s},$$
so that we may apply our induction hypothesis (46) with $r - 1$ to conclude
$$\sum_{[k/p^r]=l} a_k C(p^r \boldsymbol{n}; k) = p \sum_{[m/p^{r-1}]=l} b_m C(p^{r-1}\boldsymbol{n}; m) \equiv 0 \pmod{p^r}.$$
The claim therefore follows by induction. □

We are now in a comfortable position to prove Theorem 3.2.

Proof of Theorem 3.2. In terms of the numbers $A_{\lambda,\varepsilon}(\boldsymbol{n};k)$, defined in (37), we have

$$A_{\lambda,\varepsilon}(\boldsymbol{n}) = \sum_{k\geqslant 0} \varepsilon^k A_\lambda(\boldsymbol{n};k) = \sum_{s\geqslant 0} G_s(\boldsymbol{n}),$$

where

$$G_s(\boldsymbol{n}) = \sum_{p\nmid k} \varepsilon^{p^s k} A_\lambda(\boldsymbol{n}; p^s k).$$

Suppose that $\ell \geqslant 2$. Further, suppose that $p \geqslant 3$, or that $p = 2$ and $\varepsilon = 1$. Then $\varepsilon^{p^s k} = \varepsilon^{p^{s-1}k}$, and it follows from Lemma 5.3 that, for $s \geqslant 1$,

$$G_s(p^r\boldsymbol{n}) \equiv G_{s-1}(p^{r-1}\boldsymbol{n}) \pmod{p^{2r}}.$$

In order to prove that $A_{\lambda,\varepsilon}(p^r\boldsymbol{n}) \equiv A_{\lambda,\varepsilon}(p^{r-1}\boldsymbol{n})$ modulo p^{2r}, it therefore remains only to show that $G_0(p^r\boldsymbol{n}) \equiv 0$ modulo p^{2r}. This, however, is immediate because, as observed in the proof of Lemma 5.3, $A_\lambda(p^r\boldsymbol{n};k)$, with $p\nmid k$, is divisible by $p^{\ell r}$. This proves congruence (21).

Now, suppose that $\ell \geqslant 2$ and $\max(\lambda_1,\dots,\lambda_\ell) \leqslant 2$. Let p be a prime such that $p \geqslant 5$. It again follows from $\varepsilon^{p^s k} = \varepsilon^{p^{s-1}k}$ and Lemma 5.3 that, for $s \geqslant 1$,

$$G_s(p^r\boldsymbol{n}) \equiv G_{s-1}(p^{r-1}\boldsymbol{n}) \pmod{p^{3r}}.$$

To prove $A_{\lambda,\varepsilon}(p^r\boldsymbol{n}) \equiv A_{\lambda,\varepsilon}(p^{r-1}\boldsymbol{n})$ modulo p^{3r}, we have to show that $G_0(p^r\boldsymbol{n}) \equiv 0$ modulo p^{3r}. As in the previous case, this is trivial if $\ell \geqslant 3$. We thus assume $\ell = 2$.

Note that, since $\max(\lambda_1,\dots,\lambda_\ell) \leqslant 2$, each factor of $A_\lambda(\boldsymbol{n};k)$ is of the form

$$\binom{m_1}{k} \quad \text{or} \quad \binom{m_1}{k}\binom{m_1+m_2-k}{m_1}.$$

Using the basic identity

$$\binom{m_1}{k} = \frac{m_1}{k}\binom{m_1-1}{k-1},$$

it is clear that the numbers

$$B_\lambda(\boldsymbol{n};k) = \frac{k^2}{n_1 n_{1+\lambda_1}} A_\lambda(\boldsymbol{n};k)$$

are integers. Moreover, it follows from Lemmas 5.4 and 5.5, and the fact that $\ell = 2$, that the integers $C_\lambda(\boldsymbol{n};k) = B_\lambda(\boldsymbol{n};k+1)$ satisfy, for all $k, r \in \mathbb{Z}$ with $r \geqslant 0$,

$$C(p^r\boldsymbol{n};k) \equiv C(p^{r-1}\boldsymbol{n};[k/p]) \pmod{p^r}.$$

If $p \nmid k$ then $[(k-1)/p] = [k/p]$ so that, in particular,

$$C(p^r\boldsymbol{n},k-1) \equiv C(p^r\boldsymbol{n},[k/p]) \equiv C(p^r\boldsymbol{n};k) \pmod{p^r}.$$

By construction,

$$G_0(p^r\boldsymbol{n}) = p^{2r} n_1 n_{1+\lambda_1} \sum_{p\nmid k} \frac{\varepsilon^k}{k^2} C(p^r\boldsymbol{n}; k-1),$$

so that in order to show that $G_0(p^r\boldsymbol{n}) \equiv 0$ modulo p^{3r}, it suffices to prove

$$\sum_{p\nmid k} \frac{\varepsilon^k}{k^2} C(p^r\boldsymbol{n}; k) \equiv 0 \pmod{p^r}. \tag{47}$$

Define $a_k = \varepsilon^k/k^2$, if $p \nmid k$, and $a_k = 0$ otherwise. Since $p \geqslant 5$, it follows from Lemma 5.2 that, for all $l, s \in \mathbb{Z}$ with $s \geqslant 0$,

$$\sum_{[k/p^s]=l} a_k = \sum_{k=1,p\nmid k}^{p^s-1} \frac{\varepsilon^{lp^s+k}}{(lp^s+k)^2} \equiv \varepsilon^l \sum_{k=1,p\nmid k}^{p^s-1} \frac{\varepsilon^k}{k^2} \equiv 0 \pmod{p^s}.$$

Hence, the conditions of Lemma 5.6 are met, allowing us to conclude that

$$\sum_{p\nmid k} \frac{\varepsilon^k}{k^2} C(p^r\boldsymbol{n}; k) = \sum_{l} \sum_{[k/p^r]=l} a_k C(p^r\boldsymbol{n}; k) \equiv 0 \pmod{p^r}.$$

This shows (47) and completes our proof. □

Acknowledgements

I wish to thank Robert Osburn and Wadim Zudilin for interesting and motivating discussions on Apéry-like numbers and supercongruences. Much appreciated comments on an earlier version of this note have also been sent by Bruce Berndt and Brundaban Sahu. Moreover, I am very grateful to the referee for several suggestions that improved this paper. Finally, I thank the Max Planck Institute for Mathematics in Bonn, where most of this work was completed, for providing wonderful working conditions.

References

[Ahlgren and Ono 2000] S. Ahlgren and K. Ono, "A Gaussian hypergeometric series evaluation and Apéry number congruences", *J. Reine Angew. Math.* **518** (2000), 187–212. MR 2001c:11057 Zbl 0940.33002

[Almkvist and Zudilin 2006] G. Almkvist and W. Zudilin, "Differential equations, mirror maps and zeta values", pp. 481–515 in *Mirror symmetry, V*, edited by N. Yui et al., AMS/IP Stud. Adv. Math. **38**, Amer. Math. Soc., Providence, RI, 2006. MR 2008j:14073 Zbl 1118.14043

[Almkvist et al. 2011] G. Almkvist, D. van Straten, and W. Zudilin, "Generalizations of Clausen's formula and algebraic transformations of Calabi–Yau differential equations", *Proc. Edinb. Math. Soc.* (2) **54**:2 (2011), 273–295. MR 2012e:33016 Zbl 1223.33007

[Apéry 1979] R. Apéry, "Irrationalité de $\zeta(2)$ et $\zeta(3)$", pp. 11–13 in *Journées arithmétiques de Luminy: Colloque international du Centre national de la recherche scientifique* (Luminy, 1978), Astérisque **61**, Société Mathématique de France, Paris, 1979. Zbl 0401.10049

[Askey and Gasper 1977] R. Askey and G. Gasper, "Convolution structures for Laguerre polynomials", *J. Analyse Math.* **31** (1977), 48–68. MR 58 #6393 Zbl 0347.33006

[Beukers 1985] F. Beukers, "Some congruences for the Apéry numbers", *J. Number Theory* **21**:2 (1985), 141–155. MR 87g:11032 Zbl 0571.10008

[Beukers 1987] F. Beukers, "Another congruence for the Apéry numbers", *J. Number Theory* **25**:2 (1987), 201–210. MR 88b:11002 Zbl 0614.10011

[Beukers 2002] F. Beukers, "On Dwork's accessory parameter problem", *Math. Z.* **241**:2 (2002), 425–444. MR 2003i:12013 Zbl 1023.34081

[Bostan et al. 2013] A. Bostan, S. Boukraa, G. Christol, S. Hassani, and J.-M. Maillard, "Ising n-fold integrals as diagonals of rational functions and integrality of series expansions", *J. Phys. A* **46**:18 (2013), 185202. MR 3055667 Zbl 1267.82021

[Chan and Zudilin 2010] H. H. Chan and W. Zudilin, "New representations for Apéry-like sequences", *Mathematika* **56**:1 (2010), 107–117. MR 2011c:11067 Zbl 1275.11035

[Chan et al. 2010] H. H. Chan, S. Cooper, and F. Sica, "Congruences satisfied by Apéry-like numbers", *Int. J. Number Theory* **6**:1 (2010), 89–97. MR 2011e:11004 Zbl 05687127

[Chowla et al. 1980] S. Chowla, J. Cowles, and M. Cowles, "Congruence properties of Apéry numbers", *J. Number Theory* **12**:2 (1980), 188–190. MR 81k:10021 Zbl 0428.10008

[Christol 1984] G. Christol, "Diagonales de fractions rationnelles et equations différentielles", exposé no. 18 in *Study group on ultrametric analysis,* 1982/83, vol. 2, edited by Y. Amice et al., Inst. Henri Poincaré, Paris, 1984. MR 85k:12003 Zbl 0534.12018

[Cooper 2012] S. Cooper, "Sporadic sequences, modular forms and new series for $1/\pi$", *Ramanujan J.* **29**:1-3 (2012), 163–183. MR 2994096 Zbl 06120451

[Coster 1988] M. J. Coster, *Supercongruences*, Ph.D. thesis, Universiteit Leiden, 1988.

[Cowles 1980] J. Cowles, "Some congruence properties of three well-known sequences: two notes", *J. Number Theory* **12**:1 (1980), 84–86. MR 81g:10015 Zbl 0425.10033

[Delaygue 2013] E. Delaygue, "Arithmetic properties of Apéry-like numbers", preprint, 2013. arXiv 1310.4131

[Franel 1894] J. Franel, "Réponse 42 [à une question de Laisant]", *L'Intermédiaire des Mathématiciens* **1** (1894), 45–47.

[Gessel 1982] I. Gessel, "Some congruences for Apéry numbers", *J. Number Theory* **14**:3 (1982), 362–368. MR 83i:10012 Zbl 0482.10003

[Gessel 1983] I. M. Gessel, "Some congruences for generalized Euler numbers", *Canad. J. Math.* **35**:4 (1983), 687–709. MR 85f:11013 Zbl 0493.10014

[Gillis et al. 1983] J. Gillis, B. Reznick, and D. Zeilberger, "On elementary methods in positivity theory", *SIAM J. Math. Anal.* **14**:2 (1983), 396–398. MR 84i:42017 Zbl 0599.42500

[Granville 1997] A. Granville, "Arithmetic properties of binomial coefficients, I: Binomial coefficients modulo prime powers", pp. 253–276 in *Organic mathematics* (Burnaby, BC, 1995), edited by J. Borwein et al., CMS Conf. Proc. **20**, Amer. Math. Soc., Providence, RI, 1997. MR 99h:11016 Zbl 0903.11005

[Kilbourn 2006] T. Kilbourn, "An extension of the Apéry number supercongruence", *Acta Arith.* **123**:4 (2006), 335–348. MR 2007e:11049 Zbl 1170.11008

[Krattenthaler and Rivoal 2011] C. Krattenthaler and T. Rivoal, "Multivariate p-adic formal congruences and integrality of Taylor coefficients of mirror maps", pp. 241–269 in *Arithmetic and Galois theories of differential equations*, edited by L. Di Vizio and T. Rivoal, Sémin. Congr. **23**, Soc. Math. France, Paris, 2011. MR 3076084 Zbl 06308142

[Lipshitz and van der Poorten 1990] L. Lipshitz and A. J. van der Poorten, "Rational functions, diagonals, automata and arithmetic", pp. 339–358 in *Number theory* (Banff, AB, 1988), edited by R. A. Mollin, de Gruyter, Berlin, 1990. MR 93b:11095 Zbl 0694.10008

[MacMahon 1915] P. A. MacMahon, *Combinatory analysis*, vol. I, Cambridge Univ. Press, 1915.

[Mellit and Vlasenko 2013] A. Mellit and M. Vlasenko, "Dwork's congruences for the constant terms of powers of a Laurent polynomial", preprint, 2013. arXiv 1306.5811

[Mimura 1983] Y. Mimura, "Congruence properties of Apéry numbers", *J. Number Theory* **16**:1 (1983), 138–146. MR 84h:10016 Zbl 0504.10007

[Osburn and Sahu 2011] R. Osburn and B. Sahu, "Supercongruences for Apéry-like numbers", *Adv. in Appl. Math.* **47**:3 (2011), 631–638. MR 2822206 Zbl 1244.11006

[Osburn and Sahu 2013] R. Osburn and B. Sahu, "A supercongruence for generalized Domb numbers", *Funct. Approx. Comment. Math.* **48**:part 1 (2013), 29–36. MR 3086958 Zbl 06157208

[Osburn et al. 2014] R. Osburn, B. Sahu, and A. Straub, "Supercongruences for sporadic sequences", preprint, 2014. To appear in *Proc. Edinburgh Math. Soc.* arXiv 1312.2195

[Pemantle and Wilson 2002] R. Pemantle and M. C. Wilson, "Asymptotics of multivariate sequences, I: Smooth points of the singular variety", *J. Combin. Theory Ser. A* **97**:1 (2002), 129–161. MR 2003a:05015 Zbl 1005.05007

[Pemantle and Wilson 2008] R. Pemantle and M. C. Wilson, "Twenty combinatorial examples of asymptotics derived from multivariate generating functions", *SIAM Rev.* **50**:2 (2008), 199–272. MR 2009b:05023 Zbl 1149.05003

[Petkovšek et al. 1996] M. Petkovšek, H. S. Wilf, and D. Zeilberger, $A = B$, A K Peters, Ltd., Wellesley, MA, 1996. MR 97j:05001 Zbl 0848.05002

[van der Poorten 1979] A. van der Poorten, "A proof that Euler missed. . .Apéry's proof of the irrationality of $\zeta(3)$", *Math. Intelligencer* **1**:4 (1979), 195–203. An informal report. MR 80i:10054 Zbl 0409.10028

[Rowland and Yassawi 2013] E. Rowland and R. Yassawi, "Automatic congruences for diagonals of rational functions", preprint, 2013. arXiv 1310.8635v1

[Samol and van Straten 2009] K. Samol and D. van Straten, "Dwork congruences and reflexive polytopes", preprint, 2009. arXiv 0911.0797

[Sprugnoli 2008] R. Sprugnoli, "Negation of binomial coefficients", *Discrete Math.* **308**:22 (2008), 5070–5077. MR 2009g:05009 Zbl 1158.05003

[Straub and Zudilin 2014] A. Straub and W. Zudilin, "Positivity of rational functions and their diagonals", *J. Appr. Theory* (online publication May 2014).

[Zagier 2009] D. Zagier, "Integral solutions of Apéry-like recurrence equations", pp. 349–366 in *Groups and symmetries*, edited by J. Harnad and P. Winternitz, CRM Proc. Lecture Notes **47**, Amer. Math. Soc., Providence, RI, 2009. MR 2010h:11069 Zbl 1244.11042

Communicated by Andrew Granville
Received 2014-05-07 Revised 2014-09-10 Accepted 2014-10-19

astraub@illinois.edu Department of Mathematics, University of Illinois at Urbana–Champaign, Urbana, IL 61801, United States

ALGEBRA AND NUMBER THEORY 8:8 (2014)
dx.doi.org/10.2140/ant.2014.8.2009

The image of Carmichael's λ-function

Kevin Ford, Florian Luca and Carl Pomerance

We show that the counting function of the set of values of Carmichael's λ-function is $x/(\log x)^{\eta+o(1)}$, where $\eta = 1-(1+\log\log 2)/(\log 2) = 0.08607\ldots$.

1. Introduction

Euler's function φ assigns to a natural number n the order of the group of units of the ring of integers modulo n. It is of course ubiquitous in number theory, as is its close cousin λ, which gives the exponent of the same group. Already appearing in Gauss's *Disquisitiones Arithmeticae*, λ is commonly referred to as Carmichael's function, after R. D. Carmichael, who studied it about a century ago. (A *Carmichael number n* is composite but nevertheless satisfies $a^n \equiv a \pmod n$ for all integers a, just as primes do. Carmichael discovered these numbers, which are characterized by the property that $\lambda(n) \mid n-1$.)

It is interesting to study φ and λ as functions. For example, how easy is it to compute $\varphi(n)$ or $\lambda(n)$ given n? It is indeed easy if we know the prime factorization of n. Interestingly, we know the converse. By [Miller 1976], given either $\varphi(n)$ or $\lambda(n)$, it is easy to find the prime factorization of n.

Within the realm of "arithmetic statistics" one can also ask for the behavior of φ and λ on typical inputs n, and ask how far this varies from their values on average. For φ, this type of question goes back to the dawn of the field of probabilistic number theory with the seminal paper of Schoenberg [1928], while some results in this vein for λ are found in [Erdős et al. 1991].

One can also ask about the value sets of φ and λ. That is, what can one say about the integers which appear as the order or exponent of the groups $(\mathbb{Z}/n\mathbb{Z})^*$?

These are not new questions. Let $V_\varphi(x)$ denote the number of positive integers $n \leqslant x$ for which $n=\varphi(m)$ for some m. Pillai [1929] showed $V_\varphi(x) \leqslant x/(\log x)^{c+o(1)}$ as $x\to\infty$, where $c=(\log 2)/e$. On the other hand, since $\varphi(p)=p-1$, $V_\varphi(x)$ is at least $\pi(x+1)$ (the number of primes in $[1, x+1]$), and so $V_\varphi(x) \geqslant (1+o(1))x/\log x$.

Ford was supported in part by National Science Foundation grant DMS-1201442. Pomerance was supported in part by NSF grant DMS-1001180.

MSC2010: primary 11N64; secondary 11A25, 11N25.

Keywords: Carmichael's function, Carmichael's lambda function.

In one of his earliest papers, Erdős [1935] showed that the lower bound is closer to the truth: we have $V_\varphi(x) = x/(\log x)^{1+o(1)}$ as $x \to \infty$. This result has since been refined by a number of authors, including Erdős and Hall, Maier and Pomerance, and Ford; see [Ford 1998] for the current state of the art.

Essentially the same results hold for the sum-of-divisors function σ, but only recently were we able to show that there are infinitely many numbers that are simultaneously values of φ and of σ [Ford et al. 2010] , thus settling an old problem of Erdős.

In this paper, we address the range problem for Carmichael's function λ. From the definition of $\lambda(n)$ as the exponent of the group $(\mathbb{Z}/n\mathbb{Z})^*$, it is immediate that $\lambda(n) \mid \varphi(n)$ and that $\lambda(n)$ is divisible by the same primes as $\varphi(n)$. We also have

$$\lambda(n) = \operatorname{lcm}[\lambda(p^a) : p^a \parallel n],$$

where $\lambda(p^a) = p^{a-1}(p-1)$ for odd primes p with $a \geqslant 1$ or $p = 2$ and $a \in \{1, 2\}$. Further, $\lambda(2^a) = 2^{a-2}$ for $a \geqslant 3$. Put $V_\lambda(x)$ for the number of integers $n \leqslant x$ with $n = \lambda(m)$ for some m. Note that since $p - 1 = \lambda(p)$ for all primes p, it follows that

$$V_\lambda(x) \geqslant \pi(x+1) = (1+o(1))\frac{x}{\log x} \quad (x \to \infty), \tag{1-1}$$

as with φ. In fact, one might suspect that the story for λ is completely analogous to that of φ. As it turns out, this is not the case.

It is fairly easy to see that $V_\varphi(x) = o(x)$ as $x \to \infty$, since most numbers n are divisible by many different primes, so most values of $\varphi(n)$ are divisible by a high power of 2. This argument fails for λ, and in fact it is not immediately obvious that $V_\lambda(x) = o(x)$ as $x \to \infty$. Such a result was first shown in [Erdős et al. 1991], where it was established that there is a positive constant c with $V_\lambda(x) \ll x/(\log x)^c$. In [Friedlander and Luca 2007], a value of c in this result was computed. It was shown there that, as $x \to \infty$,

$$V_\lambda(x) \leqslant \frac{x}{(\log x)^{\alpha+o(1)}} \quad \text{holds with} \quad \alpha = 1 - e(\log 2)/2 = 0.057913\dots. \tag{1-2}$$

The exponents on the logarithms in the lower and upper bounds (1-1) and (1-2) were brought closer in the recent paper [Luca and Pomerance 2014], where it was shown that, as $x \to \infty$,

$$\frac{x}{(\log x)^{0.359052}} < V_\lambda(x) \leqslant \frac{x}{(\log x)^{\eta+o(1)}} \text{ with } \eta = 1 - \frac{1+\log\log 2}{\log 2} = 0.08607\dots.$$

In Section 2.1 of that paper, a heuristic was presented suggesting that the correct exponent of the logarithm should be the number η. In the present paper, we confirm the heuristic from [Luca and Pomerance 2014] by proving the following theorem:

Theorem 1. *We have* $V_\lambda(x) = x(\log x)^{-\eta+o(1)}$ *as* $x \to \infty$.

Just as results on $V_\varphi(x)$ can be generalized to similar multiplicative functions, such as σ, we would expect our result to be generalizable to functions similar to λ enjoying the property $f(mn) = \mathrm{lcm}[f(m), f(n)]$ when m, n are coprime.

Since the upper bound in Theorem 1 was proved in [Luca and Pomerance 2014], we need only show that $V_\lambda(x) \geqslant x/(\log x)^{\eta+o(1)}$ as $x \to \infty$. We remark that in our lower bound argument we will count only squarefree values of λ.

The same number η in Theorem 1 appears in an unrelated problem. As shown by Erdős [1960], the number of distinct entries in the multiplication table for the numbers up to n is $n^2/(\log n)^{\eta+o(1)}$ as $n \to \infty$. Similarly, the asymptotic density of the integers with a divisor in $[n, 2n]$ is $1/(\log n)^{\eta+o(1)}$ as $n \to \infty$. See [Ford 2008a; 2008b] for more on these kinds of results. As explained in the heuristic argument presented in [Luca and Pomerance 2014], the source of η in the λ-range problem comes from the distribution of integers n with about $(1/\log 2)\log\log n$ prime divisors: the number of these numbers $n \in [2, x]$ is $x/(\log x)^{\eta+o(1)}$ as $x \to \infty$. Curiously, the number η arises in the same way in the multiplication table problem: most entries in an n-by-n multiplication table have about $(1/\log 2)\log\log n$ prime divisors (a heuristic for this is given in the introduction of [Ford 2008a]).

We mention two related unsolved problems. Several papers [Banks et al. 2004; Banks and Luca 2011; Freiberg 2012; Pollack and Pomerance 2014] have discussed the distribution of numbers n such that n^2 is a value of φ; in [Pollack and Pomerance 2014] it was shown that the number of such $n \leqslant x$ is between $x/(\log x)^{c_1}$ and $x/(\log x)^{c_2}$, where $c_1 > c_2 > 0$ are explicit constants. Is the count of the form $x/(\log x)^{c+o(1)}$ for some number c? The numbers c_1, c_2 in [Pollack and Pomerance 2014] are not especially close. The analogous problem for λ is wide open. In fact, it seems that a reasonable conjecture (from [Pollack and Pomerance 2014]) is that asymptotically all even numbers n have n^2 in the range of λ. On the other hand, it has not been proved that there is a lower bound of the shape $x/(\log x)^c$ with some positive constant c for the number of such numbers $n \leqslant x$.

2. Lemmas

Here we present some estimates that will be useful in our argument. To fix notation, for a positive integer q and an integer a, we let $\pi(x; q, a)$ be the number of primes $p \leqslant x$ in the progression $p \equiv a \pmod q$, and put

$$E^*(x; q) = \max_{y \leqslant x} \left| \pi(y; q, 1) - \frac{\mathrm{li}(y)}{\varphi(q)} \right|,$$

where $\mathrm{li}(y) = \int_2^y \mathrm{d}t/\log t$.

We also let $P^+(n)$ and $P^-(n)$ denote the largest and smallest prime factors of n, respectively, with the convention that $P^-(1) = \infty$ and $P^+(1) = 0$. Let $\omega(m)$ be the number of distinct prime factors of m, and let $\tau_k(n)$ be the k-th divisor function;

that is, the number of ways to write $n = d_1 \cdots d_k$ with $d_1, \dots, d_k$ positive integers. Let μ denote the Möbius function.

First, we present an estimate for the sum of reciprocals of integers with a given number of prime factors.

Lemma 2.1. *Suppose x is large. Uniformly for $1 \leqslant h \leqslant 2 \log\log x$,*

$$\sum_{\substack{P^+(b)\leqslant x \\ \omega(b)=h}} \frac{\mu^2(b)}{b} \asymp \frac{(\log\log x)^h}{h!}.$$

Proof. The upper bound follows very easily from

$$\sum_{\substack{P^+(b)\leqslant x \\ \omega(b)=h}} \frac{\mu^2(b)}{b} \leqslant \frac{1}{h!}\biggl(\sum_{p\leqslant x} \frac{1}{p}\biggr)^h = \frac{(\log\log x + O(1))^h}{h!} \asymp \frac{(\log\log x)^h}{h!}$$

upon using Mertens' theorem and the given upper bound on h. For the lower bound, we have

$$\sum_{\substack{P^+(b)\leqslant x \\ \omega(b)=h}} \frac{\mu^2(b)}{b} \geqslant \frac{1}{h!}\biggl(\sum_{p\leqslant x} \frac{1}{p}\biggr)^h \biggl[1 - \binom{h}{2}\biggl(\sum_{p\leqslant x} \frac{1}{p}\biggr)^{-2} \sum_{p} \frac{1}{p^2}\biggr].$$

Again, the sums of $1/p$ are each $\log\log x + O(1)$. The sum of $1/p^2$ is smaller than 0.46, hence for large enough x the bracketed expression is at least 0.08, and the desired lower bound follows. □

Next, we recall (see e.g., [Davenport 2000, Chapter 28]) the well-known theorem of Bombieri and Vinogradov, and then we prove a useful corollary.

Lemma 2.2. *For any number $A > 0$ there is a number $B > 0$ so that for $x \geqslant 2$*

$$\sum_{q\leqslant\sqrt{x}(\log x)^{-B}} E^*(x;q) \ll_A \frac{x}{(\log x)^A}.$$

Corollary 1. *For any integer $k \geqslant 1$ and number $A > 0$ we have for all $x \geqslant 2$ that*

$$\sum_{q\leqslant x^{1/3}} \tau_k(q) E^*(x;q) \ll_{k,A} \frac{x}{(\log x)^A}.$$

Proof. Apply Lemma 2.2 with A replaced by $2A + k^2$, Cauchy's inequality, the trivial bound $|E^*(x;q)| \ll x/q$ and the easy bound

$$\sum_{q\leqslant y} \frac{\tau_k^2(q)}{q} \ll_k (\log y)^{k^2} \tag{2-1}$$

to get

$$\Big(\sum_{q\leqslant x^{1/3}} \tau_k(q)E^*(x;q)\Big)^2 \leqslant \Big(\sum_{q\leqslant x^{1/3}} \tau_k(q)^2|E^*(x;q)|\Big)\Big(\sum_{q\leqslant x^{1/3}} |E^*(x;q)|\Big)$$
$$\ll_{k,A} x\Big(\sum_{q\leqslant x^{1/3}} \frac{\tau_k(q)^2}{q}\Big)\frac{x}{(\log x)^{2A+k^2}}$$
$$\ll_{k,A} \frac{x^2}{(\log x)^{2A}},$$

which leads to the desired conclusion. □

Finally, we need a lower bound from sieve theory.

Lemma 2.3. *There are absolute constants $c_1 > 0$ and $c_2 \geqslant 2$ so that for $y \geqslant c_2$, $y^3 \leqslant x$, and any even positive integer b, we have*

$$\sum_{\substack{n\in(x,2x] \\ bn+1 \text{ prime} \\ P^-(n)>y}} 1 \geqslant \frac{c_1 bx}{\varphi(b)\log(bx)\log y} - 2\sum_{m\leqslant y^3} 3^{\omega(m)}E^*(2bx;bm).$$

Proof. We apply a standard lower bound sieve to the set

$$\mathcal{A} = \Big\{\frac{\ell-1}{b} : \ell \text{ prime},\ \ell\in(bx+1,2bx],\ \ell\equiv 1 \pmod{b}\Big\}.$$

Letting $\mathcal{A}_d$ be the set of elements of $\mathcal{A}$ divisible by a squarefree integer d, we have $|\mathcal{A}_d| = Xg(d)/d + r_d$, where

$$X = \frac{\mathrm{li}(2bx)-\mathrm{li}(bx+1)}{\varphi(b)}, \quad g(d) = \prod_{\substack{p\,|\,d \\ p\nmid b}} \frac{p}{p-1}, \quad |r_d| \leqslant 2E^*(2bx;db).$$

It follows that for $2\leqslant v < w$,

$$\sum_{v\leqslant p<w} \frac{g(p)}{p}\log p = \log\frac{w}{v} + O(1),$$

the implied constant being absolute. Apply [Halberstam and Richert 1974, Theorem 8.3] with $q=1$, $\xi = y^{3/2}$ and $z=y$, observing that the condition $\Omega_2(1,L)$ on page 142 of that work holds with an absolute constant L. With the function $f(u)$ as defined on pages 225–227 there, we have $f(3) = \frac{2}{3}e^{\gamma}\log 2 > \frac{4}{5}$. Then with B_{19} the absolute constant in Theorem 8.3 of that work, we have

$$f(3) - B_{19}\frac{L}{(\log\xi)^{1/14}} \geqslant \frac{1}{2}$$

for large enough c_2. We obtain the bound

$$\begin{aligned}\#\{x < n \leqslant 2x : bn+1 \text{ prime}, P^-(n) > y\}&\\ &\geqslant \frac{X}{2}\prod_{p\leqslant y}\Bigl(1-\frac{g(p)}{p}\Bigr) - \sum_{m\leqslant \xi^2} 3^{\omega(m)}|r_m|\\ &\geqslant \frac{c_1 bx}{\varphi(b)\log(bx)\log y} - 2\sum_{m\leqslant y^3} 3^{\omega(m)} E^*(2bx; bm). \quad \square\end{aligned}$$

3. The set-up

If $n = \lambda(p_1 p_2 \cdots p_k)$, where $p_1, p_2, \dots, p_k$ are distinct primes, then we have $n = \operatorname{lcm}[p_1 - 1, p_2 - 1, \dots, p_k - 1]$. If we further assume that n is squarefree and consider the Venn diagram of the sets $S_1, \dots, S_k$ of the prime factors of $p_1 - 1, \dots, p_k - 1$, respectively, then this equation gives an ordered factorization of n into $2^k - 1$ factors (some of which may be the trivial factor 1). Here we "see" the shifted primes $p_i - 1$ as products of certain subsequences of 2^{k-1} of these factors. Conversely, given n and an ordered factorization of n into $2^k - 1$ factors, we can ask how likely it is for those k products of 2^{k-1} factors to all be shifted primes. Of course, this is not likely at all, but if n has many prime factors, and so many factorizations, the odds that there is at least one such "good" factorization improve. For example, when $k = 2$, we factor a squarefree number n as $a_1 a_2 a_3$, and we ask for $a_1 a_2 + 1 = p_1$ and $a_2 a_3 + 1 = p_2$ to both be prime. If so, we would have $n = \lambda(p_1 p_2)$. The heuristic argument from [Luca and Pomerance 2014] was based on this idea. In particular, if a squarefree n is even and has at least $\theta_k \log\log n$ odd prime factors (where $\theta_k > k/\log(2^k - 1)$ is fixed and $\theta_k \to 1/\log 2$ as $k \to \infty$), then there are so many factorizations of n into $2^k - 1$ factors that it becomes likely that n is a λ-value. The lower bound proof from [Luca and Pomerance 2014] concentrated just on the case $k = 2$, but here we attack the general case. As in that work, we let $r(n)$ be the number of representations of n as the λ of a number with k primes. To see that $r(n)$ is often positive, we show that its average value is large, and that the average value of $r(n)^2$ is not much larger. Our conclusion will follow from Cauchy's inequality.

Let $k \geqslant 2$ be a fixed integer, let x be sufficiently large (in terms of k), and put

$$y = \exp\left\{\frac{\log x}{200k\log\log x}\right\}, \qquad l = \left\lfloor \frac{k}{(2^k - 1)\log(2^k - 1)}\log\log y\right\rfloor. \tag{3-1}$$

For $n \leqslant x$, let $r(n)$ be the number of representations of n of the form

$$n = \prod_{i=0}^{k-1} a_i \prod_{j=1}^{2^k-1} b_j, \tag{3-2}$$

where $P^+(b_j) \leqslant y < P^-(a_i)$ for all i and j, where $2 \mid b_{2^k-1}$, where $\omega(b_j) = l$ for each j, where $a_i > 1$ for all i, and where furthermore $a_i B_i + 1$ is prime for all i, where

$$B_i = \prod_{\lfloor j/2^i \rfloor \text{ odd}} b_j. \tag{3-3}$$

Observe that each B_i is even since it is a multiple of b_{2^k-1} (because $\lfloor (2^k-1)/2^i \rfloor = 2^{k-i} - 1$ is odd), each B_i is the product of 2^{k-1} of the numbers b_j, and that every b_j divides $B_0 \cdots B_{k-1}$. Also, if n is squarefree and $r(n) > 0$, then the primes $a_i B_i + 1$ are all distinct, and it follows that

$$n = \lambda\biggl(\prod_{i=0}^{k-1}(a_i B_i + 1)\biggr);$$

therefore such $n \leqslant x$ are counted by $V_\lambda(x)$. We count how often $r(n) > 0$ using Cauchy's inequality in the following standard way:

$$\#\bigl\{2^{-2k}x < n \leqslant x : \mu^2(n) = 1,\ r(n) > 0\bigr\} \geqslant \frac{S_1^2}{S_2}, \tag{3-4}$$

where

$$S_1 = \sum_{2^{-2k}x < n \leqslant x} \mu^2(n) r(n), \qquad S_2 = \sum_{2^{-2k}x < n \leqslant x} \mu^2(n) r^2(n).$$

Our application of Cauchy's inequality is rather sharp, as we will show below that $r(n)$ is approximately 1 on average over the kind of integers we are interested in, both in mean and in mean-square. More precisely, in the next section, we prove

$$S_1 \gg \frac{x}{(\log x)^{\beta_k} (\log\log x)^{O_k(1)}}, \tag{3-5}$$

and in the final section we prove

$$S_2 \ll \frac{x (\log\log x)^{O_k(1)}}{(\log x)^{\beta_k}}, \tag{3-6}$$

where

$$\beta_k = 1 - \frac{k}{\log(2^k - 1)}(1 + \log\log(2^k - 1) - \log k). \tag{3-7}$$

Together, the inequalities (3-4), (3-5) and (3-6) imply that

$$V_\lambda(x) \gg \frac{x}{(\log x)^{\beta_k} (\log\log x)^{O_k(1)}}.$$

We deduce the lower bound of Theorem 1 by noting that $\lim_{k\to\infty} \beta_k = \eta$.

Throughout, constants implied by the symbols O, $\ll$, $\gg$, and $\asymp$ may depend on k, but not on any other variable.

4. The lower bound for S_1

For convenience, when using the sieve bound in Lemma 2.3, we consider a slightly larger sum S_1' than S_1, namely

$$S_1' := \sum_{n\in\mathcal{N}} r(n),$$

where $\mathcal{N}$ is the set of $n \in (2^{-2k}x, x]$ of the form $n = n_0 n_1$ with $P^+(n_0) \leqslant y < P^-(n_1)$ and n_0 squarefree. That is, in S_1' we no longer require the numbers $a_0, \dots, a_{k-1}$ in (3-2) to be squarefree. The difference between S_1 and S_1' is very small; indeed, putting $h = 2^k + k - 1$, note that $r(n) \leqslant \tau_h(n)$, so that we have by (3-2) the estimate

$$\begin{aligned} S_1' - S_1 &\leqslant \sum_{\substack{n\leqslant x\\ \exists p>y:p^2\mid n}} \tau_h(n) \leqslant \sum_{p>y} \sum_{\substack{n\leqslant x\\ p^2\mid n}} \tau_h(n) \leqslant \sum_{p>y} \tau_h(p^2) \sum_{m\leqslant x/p^2} \tau_h(m) \\ &\leqslant \sum_{p>y} \tau_h(p^2) \frac{x}{p^2} \sum_{m\leqslant x} \frac{\tau_h(m)}{m} \ll \frac{x(\log x)^h}{y}. \end{aligned} \tag{4-1}$$

Here we have used the inequality $\tau_h(uv) \leqslant \tau_h(u)\tau_h(v)$, as well as the easy bound

$$\sum_{m\leqslant x} \frac{\tau_h(m)}{m} \ll (\log x)^h, \tag{4-2}$$

which is similar to (2-1). By (3-2), the sum S_1' counts the number of $(2^{k-1}+k)$-tuples $(a_0, \dots, a_{k-1}, b_1, \dots, b_{2^k-1})$ satisfying

$$2^{-2k}x < a_0 \cdots a_{k-1} b_1 \cdots b_{2^k-1} \leqslant x \tag{4-3}$$

and with $P^+(b_j) \leqslant y < P^+(a_i)$ for every i and j, $b_1 \cdots b_{2^k-1}$ squarefree, $2 \mid b_{2^k-1}$, $\omega(b_j) = l$ for every j, $a_i > 1$ for every i, and $a_i B_i + 1$ prime for every i, where B_i is defined in (3-3). Fix numbers $b_1, \dots, b_{2^k-1}$. Then

$$b_1 \cdots b_{2^k-1} \leqslant y^{(2^k-1)l} \leqslant y^{2\log\log x} = x^{1/100k}. \tag{4-4}$$

In the above, we used the fact that $k \leqslant 2\log(2^k-1)$. Fix also $A_0, \dots, A_{k-1}$, each a power of 2 exceeding $x^{1/2k}$, such that

$$\frac{x}{2b_1 \cdots b_{2^k-1}} < A_0 \cdots A_{k-1} \leqslant \frac{x}{b_1 \cdots b_{2^k-1}}. \tag{4-5}$$

Then (4-3) holds whenever $A_i/2 < a_i \leqslant A_i$ for each i. By Lemma 2.3, using the facts that $B_i/\varphi(B_i) \geqslant 2$ (because B_i is even) and $A_i B_i \leqslant x$ (a consequence of (4-5)), we deduce that the number of choices for each a_i is at least

$$\frac{c_1 A_i}{\log x \log y} - 2 \sum_{m \leqslant y^3} 3^{\omega(m)} E^*(A_i B_i; mB_i).$$

Using the elementary inequality

$$\prod_{j=1}^{k} \max(0, x_j - y_j) \geqslant \prod_{j=1}^{k} x_j - \sum_{i=1}^{k} y_i \prod_{j \neq i} x_j,$$

valid for any nonnegative real numbers x_j, y_j, we find that the number of admissible k-tuples $(a_0, \ldots, a_{k-1})$ is at least

$$\frac{c_1^k A_0 \cdots A_{k-1}}{(\log x \log y)^k} - \frac{2c_1^{k-1} A_0 \cdots A_{k-1}}{(\log x \log y)^{k-1}} \sum_{i=0}^{k-1} \frac{1}{A_i} \sum_{m \leqslant y^3} 3^{\omega(m)} E^*(A_i B_i; mB_i) \\ = M(\boldsymbol{A}, \boldsymbol{b}) - R(\boldsymbol{A}, \boldsymbol{b}),$$

say. By symmetry and (4-5),

$$\sum_{\boldsymbol{A}, \boldsymbol{b}} R(\boldsymbol{A}, \boldsymbol{b}) \\ \ll \frac{x}{(\log x \log y)^{k-1}} \sum_{\boldsymbol{b}} \frac{1}{b_1 \cdots b_{2^k-1}} \sum_{\boldsymbol{A}} \frac{1}{A_0} \sum_{m \leqslant y^3} 3^{\omega(m)} E^*(A_0 B_0; mB_0), \quad (4\text{-}6)$$

where the sum on $\boldsymbol{b}$ is over all $(2^k - 1)$−tuples satisfying $b_1 \cdots b_{2^k-1} \leqslant x^{1/100k}$. Write $b_1 \cdots b_{2^k-1} = B_0 B_0'$, where $B_0' = b_2 b_4 \cdots b_{2^k-2}$. Given B_0 and B_0', the number of corresponding tuples $(b_1, \ldots, b_{2^k-1})$ is at most $\tau_{2^{k-1}}(B_0)\tau_{2^{k-1}-1}(B_0')$. Suppose $D/2 < B_0 \leqslant D$, where D is a power of 2. Since $E^*(x; q)$ is an increasing function of x, $E^*(A_0 B_0; mB_0) \leqslant E^*(A_0 D; mB_0)$. Also, $3^{\omega(m)} \leqslant \tau_3(m)$ and

$$\sum_{B_0' \leqslant x} \frac{\tau_{2^{k-1}-1}(B_0')}{B_0'} \ll (\log x)^{2^{k-1}-1}$$

(this is (4-2) with h replaced by $2^{k-1} - 1$). We therefore deduce that

$$\sum_{\boldsymbol{A}, \boldsymbol{b}} R(\boldsymbol{A}, \boldsymbol{b}) \\ \ll \frac{x (\log x)^{2^{k-1}-1}}{(\log x \log y)^{k-1}} \sum_{\boldsymbol{A}} \frac{1}{A_0} \sum_{D} \frac{1}{D} \sum_{\substack{D/2 < B_0 \leqslant D \\ m \leqslant y^3}} \tau_3(m) \tau_{2^{k-1}}(B_0) E^*(A_0 D; mB_0),$$

with the sum taken over $(A_0, \ldots, A_{k-1}, D)$, each a power of 2, $D \leqslant x^{1/100k}$, $A_i \geqslant x^{1/2k}$ for each i and $A_0 \cdots A_{k-1} D \leqslant x$. With A_0 and D fixed, the number of choices for $(A_1, \dots, A_{k-1})$ is $\ll (\log x)^{k-1}$. Writing $q = mB_0$, we obtain

$$\begin{aligned}\sum_{\boldsymbol{A},\boldsymbol{b}} R(\boldsymbol{A}, \boldsymbol{b})& \\ &\ll x \frac{(\log x)^{2^{k-1}-1}}{(\log y)^{k-1}} \sum_{D \leqslant x^{1/100k}} \sum_{x^{1/2k} < A_0 \leqslant x/D} \frac{1}{A_0 D} \sum_{q \leqslant y^3 x^{1/100k}} \tau_{2^{k-1}+3}(q) E^*(A_0 D; q) \\ &\ll \frac{x}{(\log x)^{\beta_k + 1}},\end{aligned}$$

where we used Corollary 1 in the last step, with $A = 2^{k-1} - k + 4 + \beta_k$.

For the main term, by (4-5), given any $b_1, \dots, b_{2^{k-1}}$, the product $A_0 \cdots A_{k-1}$ is determined (and larger than $\frac{1}{2} x^{1-1/100k}$ by (4-4)), so there are $\gg (\log x)^{k-1}$ choices for the k-tuple $A_0, \dots, A_{k-1}$. Hence,

$$\sum_{\boldsymbol{A},\boldsymbol{b}} M(\boldsymbol{A}, \boldsymbol{b}) \gg \frac{x}{(\log y)^k \log x} \sum_{\boldsymbol{b}} \frac{1}{b_1 \cdots b_{2^k - 1}}.$$

Let $b = b_1 \cdots b_{2^k-1}$. Given an even, squarefree integer b, the number of ordered factorizations of b as $b = b_1 \cdots b_{2^k-1}$, where each $\omega(b_i) = l$ and b_{2^k-1} is even, is equal to

$$\frac{((2^k-1)l)!}{(2^k-1)(l!)^{2^k-1}}.$$

Let $b' = b/2$, so $h := \omega(b') = (2^k-1)l - 1 = k(\log\log y)/\log(2^k-1) + O(1)$. Applying Lemma 2.1, Stirling's formula and the fact that $(2^k-1)l = h + O(1)$ produces

$$\begin{aligned}\sum_{\boldsymbol{b}} \frac{1}{b_1 \cdots b_{2^k-1}} &\geqslant \frac{((2^k-1)l)!}{2(2^k-1)(l!)^{2^k-1}} \sum_{\substack{P^+(b') \leqslant y \\ \omega(b') = h}} \frac{\mu^2(b')}{b'} \\ &\gg \frac{((2^k-1)l)!}{(l!)^{2^k-1}} \frac{(\log\log y)^h}{h!} = \frac{(\log\log y)^h}{(l!)^{2^k-1}} (\log\log x)^{O(1)} \\ &= \left[\frac{(2^k-1)e\log(2^k-1)}{k}\right]^{(2^k-1)l} (\log\log x)^{O(1)} \\ &= (\log y)^{\frac{k}{\log(2^k-1)} \log\left[\frac{(2^k-1)e\log(2^k-1)}{k}\right]} (\log\log x)^{O(1)} \\ &= (\log y)^{k-\beta_k+1} (\log\log x)^{O(1)}.\end{aligned}$$

Invoking (3-1), we obtain that

$$\sum_{\boldsymbol{A},\boldsymbol{b}} M(\boldsymbol{A}, \boldsymbol{b}) \geqslant \frac{x}{(\log x)^{\beta_k} (\log\log x)^{O(1)}}. \tag{4-7}$$

Inequality (3-5) now follows from estimate (4-7) and our earlier estimates (4-1) of $S_1' - S_1$ and (4-6) of $\sum_{\boldsymbol{A},\boldsymbol{b}} R(\boldsymbol{A}, \boldsymbol{b})$.

5. A multivariable sieve upper bound

Here we prove an estimate from sieve theory that will be useful in our treatment of the upper bound for S_2.

Lemma 5.1. *Suppose that*:

- $y, x_1, \dots, x_h$ *are reals with* $3 < y \leqslant 2\min\{x_1, \dots, x_h\}$.
- $I_1, \dots, I_k$ *are nonempty subsets of* $\{1, \dots, h\}$.
- $b_1, \dots, b_k$ *are positive integers such that if* $I_i = I_j$, *then* $b_i \neq b_j$.

For $\boldsymbol{n} = (n_1, \dots, n_h)$ *a vector of positive integers and for* $1 \leqslant j \leqslant k$, *let* $N_j = N_j(\boldsymbol{n}) = \prod_{i \in I_j} n_i$. *Then*

$$\#\big\{\boldsymbol{n} : x_i < n_i \leqslant 2x_i \ (1 \leqslant i \leqslant h),\ P^-(n_1 \cdots n_h) > y,\ b_j N_j + 1 \text{ prime } (1 \leqslant j \leqslant k)\big\} \\ \ll_{h,k} \frac{x_1 \cdots x_h}{(\log y)^{h+k}} (\log\log(3b_1 \cdots b_k))^k.$$

Proof. Throughout this proof, all Vinogradov symbols $\ll$ and $\gg$ as well as the Landau symbol O depend on both h and k. Without loss of generality, suppose that $y \leqslant (\min(x_i))^{1/(h+k+10)}$. Since $n_i > x_i \geqslant y^{h+k+10}$ for every i, we see that the number of h-tuples in question does not exceed

$$S := \#\{\boldsymbol{n} : x_i < n_i \leqslant 2x_i \ (1 \leqslant i \leqslant h),\ P^-(n_1 \cdots n_h(b_1N_1+1)\cdots(b_kN_k+1)) > y\}.$$

We estimate S in the usual way with sieve methods, although this is a bit more general than the standard applications and we give the proof in some detail (the case $h = 1$ being completely standard). Let $\mathscr{A}$ denote the multiset

$$\mathscr{A} = \left\{ n_1 \cdots n_h \prod_{j=1}^{k} (b_j N_j + 1) \ :\ x_j < n_j \leqslant 2x_j \ (1 \leqslant j \leqslant h) \right\}.$$

For squarefree $d \leqslant y^2$ composed of primes $\leqslant y$, we have by a simple counting argument

$$|\mathscr{A}_d| := \#\{a \in \mathscr{A} : d \mid a\} = \frac{\nu(d)}{d^h} X + r_d,$$

where $X = x_1 \cdots x_h$, $\nu(d)$ is the number of solution vectors $\boldsymbol{n}$ modulo d of the congruence

$$n_1 \cdots n_h \prod_{j=1}^{k} (b_j N_j + 1) \equiv 0 \pmod{d},$$

and the remainder term satisfies, for $d \leqslant \min(x_1, \dots, x_h)$,

$$|r_d| \leqslant \nu(d) \sum_{i=1}^{h} \prod_{\substack{1\leqslant l\leqslant h\\ l\neq i}} \left(\left\lfloor \frac{x_l}{d} \right\rfloor + 1 \right) \leqslant \nu(d) \sum_{i=1}^{h} \frac{(x_1+d)\cdots(x_h+d)}{(x_i+d)d^{h-1}}$$
$$\ll \frac{\nu(d)X}{d^{h-1}\min(x_i)}.$$

The function $\nu(d)$ is clearly multiplicative and satisfies the global upper bound $\nu(p) \leqslant (h+k)p^{h-1}$ for every p. If $\nu(p) = p^h$ for some $p \leqslant y$, then clearly $S = 0$. Otherwise, the hypotheses of [Halberstam and Richert 1974, Theorem 6.2] (Selberg's sieve) are clearly satisfied, with $\kappa = h+k$, and we deduce that

$$S \ll X \prod_{p\leqslant y} \left(1 - \frac{\nu(p)}{p^h}\right) + \sum_{\substack{d\leqslant y^2\\ P^+(d)\leqslant y}} \mu^2(d) 3^{\omega(d)} |r_d|.$$

By our initial assumption about the size of y,

$$\sum_{d\leqslant y^2} \mu^2(d) 3^{\omega(d)} |r_d| \ll \frac{X}{\min(x_i)} \sum_{d\leqslant y^2} (3k+3h)^{\omega(d)} \ll \frac{Xy^3}{\min(x_i)} \ll \frac{X}{y}.$$

For the main term, consideration only of the congruence $n_1 \cdots n_h \equiv 0 \pmod p$ shows that

$$\nu(p) \geqslant h(p-1)^{h-1} = hp^{h-1} + O(p^{h-2})$$

for all p. On the other hand, suppose that $p \nmid b_1 \cdots b_k$ and furthermore that $p \nmid (b_i - b_j)$ whenever $I_i = I_j$. Each congruence $b_j N_j + 1 \equiv 0 \pmod p$ has $p^{h-1} + O(p^{h-2})$ solutions with $n_1 \dots n_h \not\equiv 0 \pmod p$, and any two of these congruences have $O(p^{h-2})$ common solutions. Hence, $\nu(p) = (h+k)p^{h-1} + O(p^{h-2})$. In particular,

$$\frac{h}{p} + O\left(\frac{1}{p^2}\right) \leqslant \frac{\nu(p)}{p^h} \leqslant \frac{h+k}{p} + O\left(\frac{1}{p^2}\right). \tag{5-1}$$

Further, writing $E = b_1 \cdots b_k \prod_{i\neq j} |b_i - b_j|$, the upper bound (5-1) above is in fact an equality except when $p \mid E$. We obtain

$$\prod_{p\leqslant y}\left(1 - \frac{\nu(p)}{p^h}\right) \ll \prod_{p\leqslant y}\left(1 - \frac{1}{p}\right)^{k+h} \prod_{p\mid E}\left(1 - \frac{1}{p}\right)^{-k} \ll \frac{(E/\varphi(E))^k}{(\log y)^{h+k}} \ll \frac{(\log\log 3E)^k}{(\log y)^{h+k}}$$

and the desired bound follows. □

6. The upper bound for S_2

Here, S_2 is the number of solutions of

$$n = \prod_{i=0}^{k-1} a_i \prod_{j=1}^{2^k-1} b_j = \prod_{i=0}^{k-1} a_i' \prod_{j=1}^{2^k-1} b_j', \tag{6-1}$$

with $2^{-2k}x < n \leqslant x$, n squarefree,

$$P^+(b_1b_1' \cdots b_{2^k-1}b_{2^k-1}') \leqslant y < P^-(a_0a_0' \cdots a_{k-1}a_{k-1}'),$$

$\omega(b_j) = \omega(b_j') = l$ for every j, $a_i > 1$ for every i, $2 \mid b_{2^k-1}$, $2 \mid b_{2^k-1}'$, and $a_iB_i + 1$ and $a_i'B_i' + 1$ prime for $0 \leqslant i \leqslant k-1$, where B_i' is defined analogously to B_i (see (3-3)). Trivially, we have

$$a := \prod_{i=0}^{k-1} a_i = \prod_{i=0}^{k-1} a_i', \qquad b := \prod_{j=1}^{2^k-1} b_j = \prod_{j=1}^{2^k-1} b_j'. \tag{6-2}$$

We partition the solutions of (6-1) according to the number of the primes a_iB_i+1 that are equal to one of the primes $a_j'B_j' + 1$, a number which we denote by m. By symmetry (that is, by appropriate permutation of the vectors $(a_0, \dots, a_{k-1})$, $(a_0', \dots, a_{k-1})$, $(b_1, \dots, b_{2^k-1})$ and $(b_1', \dots, b_{2^k-1}')$[1]), without loss of generality we may suppose that $a_iB_i = a_i'B_i'$ for $0 \leqslant i \leqslant m-1$ and that

$$a_iB_i \neq a_jB_j \quad (i \geqslant m, j \geqslant m). \tag{6-3}$$

Consequently,

$$a_i = a_i' \quad \text{and} \quad B_i = B_i' \quad (0 \leqslant i \leqslant m-1). \tag{6-4}$$

Now fix m and all the b_j and b_j'. For $0 \leqslant i \leqslant m-1$, place a_i into a dyadic interval $(A_i/2, A_i]$, where A_i is a power of 2. The primality conditions on the remaining variables are now coupled with the condition

$$a_m \cdots a_{k-1} = a_m' \cdots a_{k-1}'.$$

[1] The permutations may be described explicitly. Suppose that $m \leqslant k-1$ and that we wish to permute $(b_1, \dots, b_{2^k-1})$ such that $B_{i_1}, \dots, B_{i_m}$ become $B_0, \dots, B_{m-1}$, respectively. Let $S_i = \{1 \leqslant j \leqslant 2^k - 1 : \lfloor j/2^i \rfloor \text{ odd}\}$. The Venn diagram for the sets $S_{i_1}, \dots, S_{i_m}$ has $2^m - 1$ components of size 2^{k-m-1} and one component of size $2^{k-m-1} - 1$, and we map the variables b_j with j in a given component to the variables whose indices are in the corresponding component of the Venn diagram for $S_0, \dots, S_{m-1}$.

To aid the bookkeeping, let $\alpha_{i,j} = \gcd(a_i, a'_j)$ for $m \leqslant i, j \leqslant k-1$. Then

$$a_i = \prod_{j=m}^{k-1} \alpha_{i,j}, \qquad a'_j = \prod_{i=m}^{k-1} \alpha_{i,j}. \tag{6-5}$$

As each $a_i > 1, a'_j > 1$, each product above contains at least one factor that is greater than 1. Let I denote the set of pairs of indices (i, j) such that $\alpha_{i,j} > 1$, and fix I. For $(i, j) \in I$, place $\alpha_{i,j}$ into a dyadic interval $(A_{i,j}/2, A_{i,j}]$, where $A_{i,j}$ is a power of 2 and $A_{i,j} \geqslant y$. By the assumption on the range of n, we have

$$A_0 \cdots A_{m-1} \prod_{(i,j)\in I} A_{i,j} \asymp \frac{x}{b}. \tag{6-6}$$

For $0 \leqslant i \leqslant m-1$, we use Lemma 5.1 (with $h = 1$) to deduce that the number of a_i with $A_i/2 < a_i \leqslant A_i$, $P^-(a_i) > y$ and $a_i B_i + 1$ prime is

$$\ll \frac{A_i \log\log B_i}{\log^2 y} \ll \frac{A_i (\log\log x)^3}{\log^2 x}. \tag{6-7}$$

Counting the vectors $(\alpha_{i,j})_{(i,j)\in I}$ subject to the conditions

- $A_{i,j}/2 < \alpha_{i,j} \leqslant A_{i,j}$ and $P^-(\alpha_{i,j}) > y$ for $(i, j) \in I$;
- $a_i B_i + 1$ prime $(m \leqslant i \leqslant k-1)$;
- $a'_j B'_j + 1$ prime $(m \leqslant j \leqslant k-1)$;
- condition (6-5)

is also accomplished with Lemma 5.1, this time with $h = |I|$ and with $2(k-m)$ primality conditions. The hypothesis in the lemma concerning identical sets I_i, which may occur if $\alpha_{i,j} = a_i = a'_j$ for some i and j, is satisfied by our assumption (6-3), which implies in this case that $B_i \neq B'_j$. The number of such vectors is at most

$$\ll \frac{\prod_{(i,j)\in I} A_{i,j} (\log\log x)^{2k-2m}}{(\log y)^{|I|+2k-2m}} \ll \frac{\prod_{(i,j)\in I} A_{i,j} (\log\log x)^{|I|+4k-4m}}{(\log x)^{|I|+2k-2m}}. \tag{6-8}$$

Combining the bounds (6-7) and (6-8), and recalling (6-6), we see that the number of possibilities for the $2k$-tuple $(a_0, \dots, a_{k-1}, a'_0 \dots, a'_{k-1})$ is at most

$$\ll \frac{x(\log\log x)^{O(1)}}{b(\log x)^{|I|+2k}}.$$

With I fixed, there are $O((\log x)^{|I|+m-1})$ choices for $A_0, \dots, A_{m-1}$ and $A_{i,j}$ subject to (6-6), and there are $O(1)$ possibilities for I. We infer that with m and all of the

b_j, b'_j fixed, the number of possible $(a_0, \dots, a_{k-1}, a'_0 \dots, a'_{k-1})$ is at most

$$\ll \frac{x(\log\log x)^{O(1)}}{b(\log x)^{2k+1-m}}.$$

We next prove that the identities in (6-4) imply that

$$B_{\boldsymbol{v}} = B'_{\boldsymbol{v}} \quad (\boldsymbol{v} \in \{0, 1\}^m), \tag{6-9}$$

where $B_{\boldsymbol{v}}$ is the product of all b_j where the m least significant base-2 digits of j are given by the vector $\boldsymbol{v}$, and $B'_{\boldsymbol{v}}$ is defined analogously. Fix $\boldsymbol{v} = (v_0, \dots, v_{m-1})$. For $0 \leqslant i \leqslant m-1$, let $C_i = B_i$ if $v_i = 1$ and $C_i = b/B_i$ if $v_i = 0$, and define C'_i analogously. By (3-3), each number b_j where the last m base-2 digits of j are equal to $\boldsymbol{v}$ divides every C_i, and no other b_j has this property. By (6-4), $C_i = C'_i$ for each i and thus

$$C_0 \cdots C_{m-1} = C'_0 \cdots C'_{m-1}.$$

As the numbers b_j are pairwise coprime, in the above equality the primes having exponent m on the left are exactly those dividing $B_{\boldsymbol{v}}$, and similarly the primes on the right side having exponent m are exactly those dividing $B'_{\boldsymbol{v}}$. This proves (6-9).

Say b is squarefree. We count the number of dual factorizations of b compatible with both (6-2) and (6-9). Each prime dividing b first "chooses" which $B_{\boldsymbol{v}} = B'_{\boldsymbol{v}}$ to divide. Once this choice is made, there is the choice of which b_j to divide and also which b'_j. For the $2^m - 1$ vectors $\boldsymbol{v} \neq \boldsymbol{0}$, $B_{\boldsymbol{v}} = B'_{\boldsymbol{v}}$ is the product of 2^{k-m} numbers b_j and also the product of 2^{k-m} numbers b'_j. Similarly, $B_{\boldsymbol{0}}$ is the product of $2^{k-m} - 1$ numbers b_j and $2^{k-m} - 1$ numbers b'_j. Thus, ignoring that $\omega(b_j) = \omega(b'_j) = l$ for each j and that b_{2^k-1} and b'_{2^k-1} are even, the number of dual factorizations of b is at most

$$\bigl((2^m-1)(2^{k-m})^2 + (2^{k-m}-1)^2\bigr)^{\omega(b)} = (2^{2k-m} - 2^{k+1-m} + 1)^{\omega(b)}. \tag{6-10}$$

Again, let

$$h = \omega(b) = (2^k - 1)l = \frac{k}{\log(2^k-1)} \log\log y + O(1),$$

as in Section 4. Lemma 2.1 and Stirling's formula give

$$\sum_{\substack{P^+(b) \leqslant y \\ \omega(b) = h}} \frac{\mu^2(b)}{b} \ll \frac{(\log\log y)^h}{h!} \ll (e\log(2^k-1)/k)^h.$$

Combined with our earlier bound (6-10) for the number of admissible ways to dual factor each b, we obtain

$$S_2 \ll \frac{x(\log\log x)^{O(1)}}{\log x}(e\log(2^k-1)/k)^h \times \sum_{m=0}^{k}(\log y)^{m-2k+\frac{k}{\log(2^k-1)}\log(2^{2k-m}-2^{k+1-m}+1)}. \quad (6\text{-}11)$$

For real $t \in [0,k]$, let $f(t) = k\log(2^{2k-t} - 2^{k+1-t} + 1) - (2k-t)\log(2^k-1)$. We have $f(0) = f(k) = 0$ and

$$f''(t) = \frac{k(\log 2)^2(2^{2k} - 2^{k+1})2^{-t}}{(2^{2k-t} - 2^{k+1-t} + 1)^2} > 0.$$

Hence, $f(t) < 0$ for $0 < t < k$. Thus, the sum on m in (6-11) is $O(1)$, and (3-6) follows.

Theorem 1 is therefore proved.

Acknowledgements

Part of this work was done while Ford and Luca visited Dartmouth College in Spring 2013. They thank the people there for their hospitality. Pomerance gratefully acknowledges a helpful conversation with Andrew Granville in which the heuristic argument behind our proof first arose. The authors also thank one of the referees for constructive comments which improved the paper.

References

[Banks and Luca 2011] W. D. Banks and F. Luca, “Power totients with almost primes”, *Integers* **11**:3 (2011), 307–313. MR 2988064 Zbl 1268.11141

[Banks et al. 2004] W. D. Banks, J. B. Friedlander, C. Pomerance, and I. E. Shparlinski, “Multiplicative structure of values of the Euler function”, pp. 29–47 in *High primes and misdemeanours: lectures in honour of the 60th birthday of Hugh Cowie Williams*, edited by A. van der Poorten and A. Stein, Fields Inst. Commun. **41**, Amer. Math. Soc., Providence, RI, 2004. MR 2005f:11217 Zbl 1099.11055

[Davenport 2000] H. Davenport, *Multiplicative number theory*, 3rd ed., Graduate Texts in Mathematics **74**, Springer, New York, 2000. MR 2001f:11001 Zbl 1002.11001

[Erdős 1935] P. Erdős, “On the normal number of prime factors of $p-1$ and some related problems concerning Euler’s φ-function”, *Q. J. Math., Oxf. Ser.* **6** (1935), 205–213. Zbl 0012.14905

[Erdős 1960] P. Erdős, “Об одном асимптотическом неравенстве в теории чисел (An asymptotic inequality in number theory)”, *Vestnik Leningrad. Univ.* **15**:13 (1960), 41–49. MR 23 #A3720 Zbl 0104.26804

[Erdős et al. 1991] P. Erdős, C. Pomerance, and E. Schmutz, “Carmichael’s lambda function”, *Acta Arith.* **58**:4 (1991), 363–385. MR 92g:11093 Zbl 0734.11047

[Ford 1998] K. Ford, “The distribution of totients”, *Ramanujan J.* **2**:1-2 (1998), 67–151. Updated version on the author’s web page. MR 99m:11106 Zbl 0914.11053

[Ford 2008a] K. Ford, "The distribution of integers with a divisor in a given interval", *Ann. of Math.* (2) **168**:2 (2008), 367–433. MR 2009m:11152 Zbl 1181.11058

[Ford 2008b] K. Ford, "Integers with a divisor in $(y, 2y]$", pp. 65–80 in *Anatomy of integers*, edited by J.-M. De Koninck et al., CRM Proc. Lecture Notes **46**, Amer. Math. Soc., Providence, RI, 2008. MR 2009i:11113 Zbl 1175.11053

[Ford et al. 2010] K. Ford, F. Luca, and C. Pomerance, "Common values of the arithmetic functions ϕ and σ", *Bull. Lond. Math. Soc.* **42**:3 (2010), 478–488. MR 2011m:11191 Zbl 1205.11010

[Freiberg 2012] T. Freiberg, "Products of shifted primes simultaneously taking perfect power values", *J. Aust. Math. Soc.* **92**:2 (2012), 145–154. MR 2999152 Zbl 06124076

[Friedlander and Luca 2007] J. B. Friedlander and F. Luca, "On the value set of the Carmichael λ-function", *J. Aust. Math. Soc.* **82**:1 (2007), 123–131. MR 2008c:11124 Zbl 1146.11046

[Halberstam and Richert 1974] H. Halberstam and H.-E. Richert, *Sieve methods*, London Mathematical Society Monographs **4**, Academic Press, London, New York, 1974. MR 54 #12689 Zbl 0298.10026

[Luca and Pomerance 2014] F. Luca and C. Pomerance, "On the range of Carmichael's universal-exponent function", *Acta Arith.* **162**:3 (2014), 289–308. MR 3173026 Zbl 1292.11109

[Miller 1976] G. L. Miller, "Riemann's hypothesis and tests for primality", *J. Comput. System Sci.* **13**:3 (1976), 300–317. MR 58 #470a Zbl 0349.68025

[Pillai 1929] S. S. Pillai, "On some functions connected with $\phi(n)$", *Bull. Amer. Math. Soc.* **35**:6 (1929), 832–836. MR 1561819 Zbl 55.0710.02

[Pollack and Pomerance 2014] P. Pollack and C. Pomerance, "Square values of Euler's function", *Bull. Lond. Math. Soc.* **46**:2 (2014), 403–414. MR 3194758 Zbl 1297.11125

[Schoenberg 1928] I. Schoenberg, "Über die asymptotische Verteilung reeller Zahlen mod 1", *Math. Z.* **28**:1 (1928), 171–199. MR 1544950 Zbl 54.0212.02

Communicated by Andrew Granville
Received 2014-06-23 Revised 2014-09-04 Accepted 2014-10-09

ford@math.uiuc.edu
Department of Mathematics,
University of Illinois at Urbana–Champaign,
1409 West Green Street, Urbana, IL 61801, United States

florian.luca@wits.ac.za
School of Mathematics, University of the Witwatersrand,
P.O. Box Wits 2050, Johannesburg, South Africa

Instituto de Matématicas, UNAM Juriquilla,
Santiago de Querétaro, 76230, Querétaro de Arteaga, México

carl.pomerance@dartmouth.edu
Mathematics Department, Dartmouth College, Kemeny Hall,
Hanover, NH 03755, 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.

Length There is no a priori limit on the length of an *ANT* article, but *ANT* considers long articles only if the significance-to-length ratio is appropriate. Very long manuscripts might be more suitable elsewhere as a memoir instead of a journal article.

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 8 No. 8 2014

1937-0652(2014)8:8;1-3