Algebra & Number Theory

Volume 3

2009

No. 3

mathematical sciences publishers

Algebra & Number Theory
www.jant.org

PRODUCTION

ant@mathscipub.org

Paulo Ney de Souza, Production Manager

Silvio Levy, Senior Production Editor

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

Regular subscription rate for 2009: $200.00 a year ($140.00 electronic only).
Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94720-3840, USA.

Algebra & Number Theory, ISSN 1937-0652, at Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

PUBLISHED BY
mathematical sciences publishers
http://www.mathscipub.org
A NON-PROFIT CORPORATION
Typeset in LaTeX

On nondegeneracy of curves

Wouter Castryck and John Voight

We study the conditions under which an algebraic curve can be modeled by a Laurent polynomial that is nondegenerate with respect to its Newton polytope. We prove that every curve of genus $g \leq 4$ over an algebraically closed field is nondegenerate in the above sense. More generally, let $\mathcal{M}_g^{\mathrm{nd}}$ be the locus of nondegenerate curves inside the moduli space of curves of genus $g \geq 2$. Then we show that $\dim \mathcal{M}_g^{\mathrm{nd}} = \min(2g+1, 3g-3)$, except for $g = 7$ where $\dim \mathcal{M}_7^{\mathrm{nd}} = 16$; thus, a generic curve of genus g is nondegenerate if and only if $g \leq 4$.

Let k be a perfect field with algebraic closure $\bar{k}$. Let $f \in k[x^{\pm 1}, y^{\pm 1}]$ be an irreducible Laurent polynomial, and write $f = \sum_{(i,j)\in\mathbb{Z}^2} c_{ij} x^i y^j$. We denote by $\operatorname{supp}(f) = \{(i, j) \in \mathbb{Z}^2 : c_{ij} \neq 0\}$ the support of f, and we associate to f its Newton polytope $\Delta = \Delta(f)$, the convex hull of $\operatorname{supp}(f)$ in $\mathbb{R}^2$. We assume throughout that Δ is 2-dimensional. For a face $\tau \subset \Delta$, let $f|_\tau = \sum_{(i,j)\in\tau} c_{ij} x^i y^j$. We say that f is *nondegenerate* if, for every face $\tau \subset \Delta$ (of any dimension), the system of equations

$$f|_\tau = x\frac{\partial f|_\tau}{\partial x} = y\frac{\partial f|_\tau}{\partial y} = 0 \tag{1}$$

has no solutions in $\bar{k}^{*2}$.

From the perspective of toric varieties, the condition of nondegeneracy can be rephrased as follows. The Laurent polynomial f defines a curve $U(f)$ in the torus $\mathbb{T}_k^2 = \operatorname{Spec} k[x^{\pm 1}, y^{\pm 1}]$, and $\mathbb{T}_k^2$ embeds canonically in the projective toric surface $X(\Delta)_k$ associated to Δ over k. Let $V(f)$ be the Zariski closure of the curve $U(f)$ inside $X(\Delta)_k$. Then f is nondegenerate if and only if for every face $\tau \subset \Delta$, the intersection $V(f) \cap \mathbb{T}_\tau$ is smooth of codimension 1 in $\mathbb{T}_\tau$, where $\mathbb{T}_\tau$ is the toric component of $X(\Delta)_k$ associated to τ. (See Proposition 1.2 for alternative characterizations.)

Nondegenerate polynomials have become popular objects in explicit algebraic geometry, owing to their connection with toric geometry [Batyrev and Cox 1994]: A wealth of geometric information about $V(f)$ is contained in the combinatorics of the Newton polytope $\Delta(f)$. The notion was first used by Kouchnirenko [1976],

MSC2000: primary 14M25; secondary 14H10.

Keywords: nondegenerate curve, toric surface, Newton polytope, moduli space.

who studied nondegenerate polynomials in the context of singularity theory. Nondenegerate polynomials also emerge naturally in the theory of sparse resultants [Gel'fand et al. 1994] and admit a linear effective Nullstellensatz [Castryck et al. 2006, Section 2.3]. They make an appearance in the study of real algebraic curves in maximal position [Mikhalkin 2000] and in the problem of enumerating curves through a set of prescribed points [Mikhalkin 2003]. In the case where k is a finite field, they arise in the construction of curves with many points [Beelen and Pellikaan 2000; Kresch et al. 2002], in the p-adic cohomology theory of Adolphson and Sperber [1989], and in explicit methods for computing zeta functions of varieties over k [Castryck et al. 2006]. Despite their utility and seeming ubiquity, the *intrinsic* property of nondegeneracy has not seen detailed study, with the exception of the otherwise unpublished PhD thesis of Koelman [1991]; see Section 12 below.

We are therefore led to the central problem of this article: *Which curves are nondegenerate?* To the extent that toric varieties are generalizations of projective space, this question essentially asks us to generalize the characterization of nonsingular plane curves amongst all curves. An immediate provocation for this question was to understand the locus of curves to which the point counting algorithm of Castryck, Denef, and Vercauteren [Castryck et al. 2006] actually applies. Our results are collected in two parts.

In the first part, comprising Sections 3–7, we investigate the nondegeneracy of some interesting classes of curves (hyperelliptic, C_{ab}, and low genus curves). Our conclusions can be summarized as follows.

Theorem. *Let V be a curve of genus g over a perfect field k. Suppose that one of these conditions holds*:

(i) $g = 0$;

(ii) $g = 1$ *and* $V(k) \neq \varnothing$;

(iii) $g = 2, 3,$ *and either* $17 \leq \#k < \infty$, *or* $\#k = \infty$ *and* $V(k) \neq \varnothing$;

(iv) $g = 4$ *and* $k = \bar{k}$.

Then V is nondegenerate.

Remark. The condition $\#k \geq 17$ in (iii) ensures that k is large enough to allow nontangency to the toric boundary of $X(\Delta)_k$, but is most likely superfluous; see Remark 7.2.

In the second part, consisting of Sections 8–12, we restrict to algebraically closed fields $k = \bar{k}$ and consider the locus $\mathcal{M}_g^{\mathrm{nd}}$ of nondegenerate curves inside the coarse moduli space of all curves of genus $g \geq 2$. We prove the following:

Theorem. *We have* $\dim \mathcal{M}_g^{\mathrm{nd}} = \min(2g+1, 3g-3)$, *except for* $g = 7$ *where* $\dim \mathcal{M}_7^{\mathrm{nd}} = 16$. *In particular*, *a generic curve of genus g is nondegenerate if and only if $g \leq 4$.*

Our methods combine ideas of Bruns and Gubeladze [2002] and Haase and Schicho [2009] and are purely combinatorial — only the universal property of the coarse moduli space is used.

Conventions and notations. Throughout, $\Delta \subset \mathbb{R}^2$ will denote a polytope with $\dim \Delta = 2$. The coordinate functions on the ambient space $\mathbb{R}^2$ will be denoted by X and Y. A *facet* or *edge* of a polytope is a face of dimension 1. A *lattice polytope* is a polytope with vertices in $\mathbb{Z}^2$. Two lattice polytopes Δ and Δ' are *equivalent* if there is an affine map

$$\varphi : \mathbb{R}^2 \to \mathbb{R}^2, \quad v \mapsto Av + b$$

such that $\varphi(\Delta) = \Delta'$ with $A \in GL_2(\mathbb{Z})$ and $b \in \mathbb{Z}^2$. Two Laurent polynomials f and f' are *equivalent* if f' can be obtained from f by applying such a map to the exponent vectors. Note that equivalence preserves nondegeneracy. For a polytope $\Delta \subset \mathbb{R}^2$, we let $\operatorname{int}(\Delta)$ denote the interior of Δ. We denote the standard 2-simplex in $\mathbb{R}^2$ by $\Sigma = \operatorname{conv}(\{(0,0), (1,0), (0,1)\})$.

1. Nondegenerate Laurent polynomials

In this section, we review the geometry of nondegenerate Laurent polynomials. We retain the notation used in the introduction. In particular, k is a perfect field, $f = \sum c_{ij} x^i y^j \in k[x^{\pm 1}, y^{\pm 1}]$ is an irreducible Laurent polynomial, and Δ is its Newton polytope. Our main implicit reference on toric varieties is [Fulton 1993].

Let $k[\Delta]$ denote the graded semigroup algebra over k generated in degree d by the monomials that are supported in $d\Delta$, that is,

$$k[\Delta] = \bigoplus_{d=0}^{\infty} \langle x^i y^j t^d \mid (i,j) \in (d\Delta \cap \mathbb{Z}^2) \rangle_k .$$

Then $X = X(\Delta)_k = \operatorname{Proj} k[\Delta]$ is the projective toric surface associated to Δ over k. This surface naturally decomposes into toric components as

$$X = \bigsqcup_{\tau \subset \Delta} \mathbb{T}_\tau ,$$

where τ ranges over the faces of Δ and $\mathbb{T}_\tau \cong \mathbb{T}_k^{\dim \tau}$. The surface X is nonsingular except possibly at the zero-dimensional toric components associated to the vertices of Δ. The Laurent polynomial f defines a curve in $\mathbb{T}_k^2 \cong \mathbb{T}_\Delta \subset X$, and we denote by $V = V(f)$ its closure in X. Alternatively, if we denote $A = \Delta \cap \mathbb{Z}^2$, then X can be canonically embedded in $\mathbb{P}_k^{\#A-1} = \operatorname{Proj} k[t_{ij}]_{(i,j)\in A}$, and V is the hyperplane section $\sum c_{ij} t_{ij} = 0$ of X.

We abbreviate $\partial_x = x \frac{\partial}{\partial x}$ and $\partial_y = y \frac{\partial}{\partial y}$.

Definition 1.1. The Laurent polynomial f is *nondegenerate* if for each face $\tau \subset \Delta$, the system

$$f|_\tau = \partial_x f|_\tau = \partial_y f|_\tau = 0$$

has no solution in $\bar{k}^{*2}$.

We sometimes write that f is Δ-nondegenerate to emphasize that $\Delta(f) = \Delta$.

Proposition 1.2. *The following statements are equivalent.*

(i) *f is nondegenerate.*

(ii) *For each face $\tau \subset \Delta$, the ideal of $k[x^{\pm 1}, y^{\pm 1}]$ generated by*

$$f|_\tau,\ \partial_x f|_\tau,\ \partial_y f|_\tau$$

is the unit ideal.

(iii) *For each face $\tau \subset \Delta$, the intersection $V \cap \mathbb{T}_\tau$ is smooth of codimension* 1 *in the torus orbit $\mathbb{T}_\tau$ associated to τ.*

(iv) *The sequence of elements $f, \partial_x f, \partial_y f$* (*in degree one*) *forms a regular sequence in $k[\Delta]$.*

(v) *The quotient of $k[\Delta]$ by the ideal generated by $f, \partial_x f, \partial_y f$ is finite and of k-dimension equal to* $2\operatorname{vol}(\Delta)$.

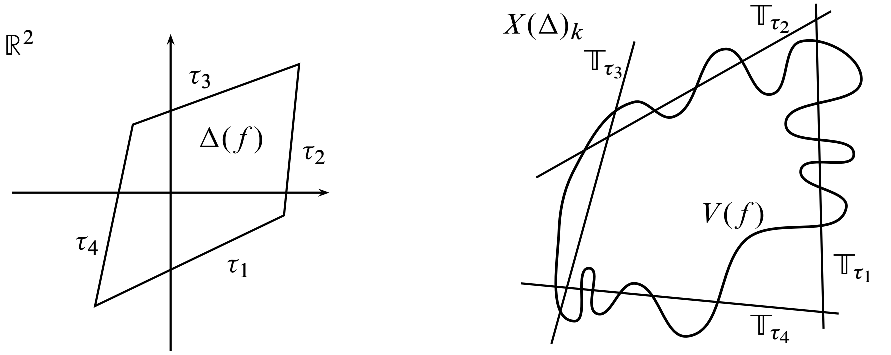
Remark 1.3. Condition (iii) can also be read that V is smooth and intersects $X \setminus \mathbb{T}_k^2$ transversally and outside the zero-dimensional toric components associated to the vertices of Δ.

Proof. See [Batyrev 1993, Section 4] for a proof of these equivalences and further discussion. □

Remark 1.4. Some authors refer to nondegenerate as Δ-*regular*, though we will not employ this term. The use of *nondegenerate* to indicate a projective variety that is not contained in a smaller projective space is unrelated to our present usage.

Example 1.5. Let $f(x, y) \in k[x, y]$ be a bivariate polynomial of degree $d \in \mathbb{Z}_{\geq 1}$ with Newton polytope $\Delta = d\Sigma = \operatorname{conv}(\{(0, 0), (d, 0), (0, d)\})$. The toric variety $X(\Delta)_k$ is the d-uple Veronese embedding of $\mathbb{P}_k^2$ in $\mathbb{P}_k^{d(d+3)/2}$, and $V(f)$ is the projective curve in $\mathbb{P}_k^2$ defined by the homogenization $F(x, y, z)$ of f. We see that $f(x, y)$ is Δ-nondegenerate if and only if $V(f)$ is nonsingular, does not contain the coordinate points $(0, 0, 1)$, $(0, 1, 0)$ and $(1, 0, 0)$, and is not tangent to any coordinate axis.

Example 1.6. The picture below illustrates nondegeneracy in case of a quadrilateral Newton polytope.



Proposition 1.7. *If $f \in k[x^{\pm 1}, y^{\pm 1}]$ is nondegenerate, then there exists a k-rational canonical divisor K_Δ on $V = V(f)$ such that $\{x^i y^j : (i, j) \in \operatorname{int}(\Delta) \cap \mathbb{Z}^2\}$ is a k-basis for the Riemann–Roch space $\mathscr{L}(K_\Delta) \subset k(V)$. In particular, the genus of V is equal to* $\#(\operatorname{int}(\Delta) \cap \mathbb{Z}^2)$.

Proof. See Khovanskiĭ [1977] or Castryck, Denef, and Vercauteren [Castryck et al. 2006, Section 2.2]. □

Remark 1.8. In general, if f is irreducible (but not necessarily nondegenerate), the geometric genus of $V(f)$ is bounded by $\#(\operatorname{int}(\Delta) \cap \mathbb{Z}^2)$. This is also known as Baker's inequality [Beelen and Pellikaan 2000, Theorem 4.2].

We conclude this section with an intrinsic definition of nondegeneracy.

Definition 1.9. A curve V over k is Δ-*nondegenerate* if V is birational over k to a curve $U \subset \mathbb{T}_k^2$ defined by a nondegenerate Laurent polynomial f with Newton polytope Δ. The curve V is *nondegenerate* if it is Δ-nondegenerate for some Δ. The curve V is *geometrically nondegenerate* if $V \times_k \bar{k}$ is nondegenerate over $\bar{k}$.

2. Moduli of nondegenerate curves

We now construct the moduli space of nondegenerate curves of given genus $g \geq 2$. Since in this article we will be concerned with dimension estimates only, we restrict to the case $k = \bar{k}$.

We denote by $\mathcal{M}_g$ the coarse moduli space of curves of genus $g \geq 2$ over k, with the property that for any flat family $\mathcal{V} \to M$ of curves of genus g, there is a (unique) morphism $M \to \mathcal{M}_g$ that maps each closed point $f \in M$ to the isomorphism class of the fiber $\mathcal{V}_f$. (See for example [Mumford 1965, Theorem 5.11].)

Let $\Delta \subset \mathbb{R}^2$ be a lattice polytope with g interior lattice points. We will construct a flat family $\mathcal{V}(\Delta) \to M_\Delta$ that parametrizes all Δ-nondegenerate curves over k. The key ingredient is provided by the following result of Gel'fand, Kapranov, and Zelevinsky [Gel'fand et al. 1994]. Let $A = \Delta \cap \mathbb{Z}^2$ and define the polynomial ring $R_\Delta = k[c_{ij}]_{(i,j) \in A}$.

Proposition 2.1. *There exists a polynomial $E_A \in R_\Delta$ with the property that for any Laurent polynomial $f \in k[x^{\pm 1}, y^{\pm 1}]$ with* $\operatorname{supp}(f) \subset \Delta$, *we have that f is Δ-nondegenerate if and only if $E_A(f) \neq 0$.*

Proof. The proof in [Gel'fand et al. 1994, Chapter 10] is over $\mathbb{C}$; however, the construction yields a polynomial over $\mathbb{Z}$ that is easily seen to characterize nondegeneracy for an arbitrary field. □

The polynomial E_A is known as the *principal A-determinant* and is given by the *A-resultant* $\operatorname{res}_A(F, \partial_1 F, \partial_2 F)$. It is homogeneous in the variables c_{ij} of degree $6 \operatorname{vol}(\Delta)$, and its irreducible factors are the *face discriminants D_τ* for faces $\tau \subset \Delta$.

Example 2.2. Consider the universal plane conic

$$F = c_{00} + c_{10}x + c_{01}y + c_{20}x^2 + c_{11}xy + c_{02}y^2,$$

associated to the Newton polytope 2Σ as in Example 1.5.

Then

$$E_A = c_{00}c_{02}c_{20}(c_{11}^2 - 4c_{02}c_{20})(c_{10}^2 - 4c_{00}c_{20})(c_{01}^2 - 4c_{00}c_{02})D_\Delta$$

where $D_\Delta = 4c_{00}c_{20}c_{02} - c_{00}c_{11}^2 - c_{10}^2c_{02} - c_{01}^2c_{20} + c_{10}c_{01}c_{11}$. The nonvanishing of the factor $c_{00}c_{02}c_{20}$ (corresponding to the discriminants of the zero-dimensional faces) ensures that the curve does not contain a coordinate point, and in particular does not have Newton polytope smaller than 2Σ; the nonvanishing of the quadratic factors (corresponding to the one-dimensional faces) ensures that the curve intersects the coordinate lines in two distinct points; and the nonvanishing of D_Δ ensures that the curve is smooth.

Let M_Δ be the complement in $\mathbb{P}_k^{\#A-1} = \operatorname{Proj} R_\Delta$ of the algebraic set defined by E_A. By the above, M_Δ parameterizes nondegenerate polynomials having Δ as Newton polytope. One can show that

$$\dim M_\Delta = \#A - 1, \tag{2}$$

which is a nontrivial statement if k is of finite characteristic (and false in general for an arbitrary number of variables), see [Castryck et al. 2006, Section 2]. Let $\mathcal{V}(\Delta)$ be the closed subvariety of

$$X(\Delta)_k \times M_\Delta \subset \operatorname{Proj} k[t_{ij}] \times \operatorname{Proj} k[c_{ij}]$$

defined by the universal hyperplane section

$$\sum_{(i,j)\in A} c_{ij}t_{ij} = 0.$$

Then the universal family of Δ-nondegenerate curves is realized by the projection map $\varphi : \mathcal{V}(\Delta) \to M_\Delta$. The fiber $\mathcal{V}(\Delta)_f$ above a nondegenerate Laurent polynomial

$f \in M_\Delta$ is precisely the corresponding curve $V(f)$, realized as the corresponding hyperplane section of $X(\Delta)_k \subset \operatorname{Proj} k[t_{ij}]$. Note that φ is indeed flat [Hartshorne 1977, Theorem III.9.9], since the Hilbert polynomial of $\mathcal{V}(\Delta)_f$ is independent of f: its degree is equal to $\deg X(\Delta)_k$ and its genus is g by Proposition 1.7.

Thus by the universal property of $\mathcal{M}_g$, there is a morphism $h_\Delta : M_\Delta \to \mathcal{M}_g$, the image of which consists precisely of all isomorphism classes containing a Δ-nondegenerate curve. Let $\mathcal{M}_\Delta$ denote the Zariski closure of the image of h_Δ. Finally, let

$$\mathcal{M}_g^{\mathrm{nd}} = \bigcup_{g(\Delta)=g} \mathcal{M}_\Delta,$$

where the union is taken over all polytopes Δ with g interior lattice points, of which there are finitely many up to equivalence; see [Hensley 1983].

The aim of Sections 8–12 is to estimate $\dim \mathcal{M}_g^{\mathrm{nd}}$. This is done by first refining the obvious upper bounds $\dim \mathcal{M}_\Delta \leq \dim M_\Delta = \#(\Delta \cap \mathbb{Z}^2) - 1$, taking into account the action of the automorphism group $\operatorname{Aut}(X(\Delta)_k)$, and then estimating the outcome in terms of g.

Remark 2.3. It follows from the fact that $\mathcal{M}_g$ is of general type for $g \geq 23$ (see for example [Harris and Morrison 1998]) that $\dim \mathcal{M}_g^{\mathrm{nd}} < \dim \mathcal{M}_g = 3g-3$ for $g \geq 23$, since each component of $\mathcal{M}_g^{\mathrm{nd}}$ is unirational. Below, we obtain much sharper results that do not rely on this deep statement.

3. Triangular nondegeneracy

In Sections 4–6, we study the nondegeneracy of certain well-known classes, such as elliptic, hyperelliptic, and C_{ab} curves. In many cases, classical constructions provide models for these curves that are supported on a triangular Newton polytope; the elementary observations in this section will allow us to prove that these models are nondegenerate when $\#k$ is not too small.

Lemma 3.1. *Let $f(x, y) \in k[x, y]$ define a smooth affine curve of genus g and suppose that $\#k > 2(g + \max(\deg_x f, \deg_y f) - 1) + \min(\deg_x f, \deg_y f)$. Then there exist $x_0, y_0 \in k$ such that the translated curve $f(x - x_0, y - y_0)$ does not contain $(0, 0)$ and is also nontangent to both the x- and the y-axis.*

Proof. Suppose $\deg_y f \leq \deg_x f$. Applying the Riemann–Hurwitz theorem to the projection map $(x, y) \mapsto x$, one sees there are at most $2(g + \deg_y f - 1)$ points with a vertical tangent. Therefore, we can find an $x_0 \in k$ such that $f(x - x_0, y)$ is nontangent to the y-axis. Subsequently, there are at most $2(g+\deg_x f - 1) + \deg_y f$ values of $y_0 \in k$ for which $f(x - x_0, y - y_0)$ is tangent to the x-axis and/or contains the point $(0, 0)$. □

Lemma 3.2. *Let $a \leq b \in \mathbb{Z}_{\geq 2}$ be such that* $\gcd(a, b) \in \{1, a\}$, *and let Δ be the triangular lattice polytope* $\operatorname{conv}(\{(0, 0), (b, 0), (0, a)\})$. *Let $f(x, y) \in k[x, y]$ be an irreducible polynomial such that*

- *f is supported on Δ, and*
- *the genus of $V(f)$ equals $g = \#(\operatorname{int}(\Delta) \cap \mathbb{Z}^2)$.*

Then $V(f)$ is Δ-nondegenerate if $\#k > 2(g + b - 1) + a$.

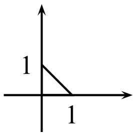
Proof. First suppose that $\gcd(a, b) = 1$. The coefficients of x^b and y^a must be nonzero, because otherwise $\#(\operatorname{int}(\Delta(f)) \cap \mathbb{Z}^2) < g$, which contradicts Baker's inequality. For the same reason, f must define a smooth affine curve: If (x_0, y_0) is a singular point (over $\bar{k}$), then $\#(\operatorname{int}(\Delta(f(x-x_0, y-y_0)) \cap \mathbb{Z}^2)) < g$. The result now follows from Lemma 3.1. Note that the nonvanishing of the face discriminant D_τ, where τ is the edge connecting $(b, 0)$ and $(0, a)$, follows automatically from the fact that τ has no interior lattice points.

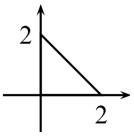
Next, suppose that $\gcd(a, b) = a$. Then we may assume that the coefficients of x^b and y^a are nonzero. Indeed, if $a < b$ then the coefficient of y^a must be nonzero. Let $g(t) \in k[t]$ be the coefficient of x^b in $f(x, y + tx^{b/a})$. It is of degree a and therefore has a nonroot $t_0 \in k$. Then substituting $y \leftarrow y + t_0 x^{b/a}$ ensures that the coefficient of x^b is nonzero as well. If $a = b$ then the coefficient of y^a might be zero, but f must contain at least one nonzero term of total degree a, and a similar argument proves the claim.

Then as above, f defines a smooth affine curve. So by applying Lemma 3.1, we may assume that the face discriminants decomposing $E_{\Delta \cap \mathbb{Z}^2}$ are nonvanishing at f, with the possible exception of D_τ, where τ is the edge connecting $(b, 0)$ and $(0, a)$. However, under the equivalence $\mathbb{R}^2 \to \mathbb{R}^2 : (X, Y) \mapsto (b - X - (b/a)Y, Y)$, the edge τ is interchanged with the edge connecting $(0, 0)$ and $(0, a)$. We obtain full nondegeneracy by applying Lemma 3.1 again. □

4. Nondegeneracy of curves of genus at most one

***Curves of genus* 0.** Let V be a curve of genus 0 over k. The anticanonical divisor embeds $V \hookrightarrow \mathbb{P}^2_k$ as a smooth conic. If $\#k = \infty$, we see that V is nondegenerate by Lemma 3.2 and Proposition 1.2. If $\#k < \infty$, then $V(k) \neq \varnothing$ by Wedderburn, and hence $V \cong \mathbb{P}^1_k$ can be embedded as a nondegenerate line in $\mathbb{P}^2$. Therefore, any curve V of genus 0 is Δ-nondegenerate, where Δ is one of the following:

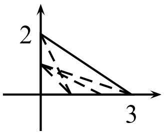




***Curves of genus* 1.** Let V be a curve of genus 1 over k. First suppose $V(k) \neq \varnothing$. Then V is an elliptic curve and hence can be defined by a nonsingular Weierstrass equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6 \tag{3}$$

with $a_i \in k$. The corresponding Newton polytope Δ is



where one of the dashed lines appears as a facet if $a_6 = 0$. By Lemma 3.2, we have V is nondegenerate if $\#k \geq 9$. With some extra work we can get rid of this condition.

For $A = \Delta \cap \mathbb{Z}^2$, the principal A-determinant has 7 or 9 face discriminants D_τ as irreducible factors. The nonvanishing of D_Δ corresponds to the fact that our curve is smooth in $\mathbb{T}^2_k$. In case τ is a vertex or a facet containing no interior lattice points, the nonvanishing of D_τ is automatic. Thus it suffices to consider the discriminants D_τ for τ a facet supported on the X-axis (denoted τ_X) or the Y-axis (denoted τ_Y). First, suppose that $\operatorname{char} k \neq 2$. After completing the square, we have $a_1 = a_3 = 0$, and the nonvanishing of D_{τ_X} follows from the fact that the polynomial $p(x) = x^3 + a_2x^2 + a_4x + a_6$ is squarefree. The nonvanishing of D_{τ_Y} (if τ_Y exists) is clear. Now suppose $\operatorname{char} k = 2$. Let δ be the number of distinct roots (over $\bar{k}$) of $p(x) = x^3 + a_2x^2 + a_4x + a_6$. If $\delta = 3$, then D_{τ_X} is nonvanishing. For the nonvanishing of D_{τ_Y}, it then suffices to substitute $x \leftarrow x + 1$ if necessary, so that a_3 is nonzero (note that not both a_1 and a_3 can be zero). If $\delta < 3$, then $p(x)$ has a root x_0 of multiplicity at least 2. Since k is perfect, this root is k-rational and after substituting $x \mapsto x + x_0$ we have $p(x) = x^3 + a_2x^2$. In particular, D_{τ_X} (if τ_X exists) and D_{τ_Y} do not vanish.

In conclusion, we have shown that every genus 1 curve V over a field k is nondegenerate, given that $V(k) \neq \varnothing$. This condition is automatically satisfied if k is a finite field (by Hasse–Weil) or if k is algebraically closed. In particular, every genus 1 curve is geometrically nondegenerate. More generally, we define the *index* of a curve V over a field k to be the least degree of an effective nonzero k-rational divisor on V (equivalently, the least extension degree of a field $L \supset k$ for which $V(L) \neq \varnothing$). We then have the following criterion.

Lemma 4.1. *A curve V of genus* 1 *is nondegenerate if and only if V has index at most* 3.

Proof. First, assume that V is nondegenerate. There are exactly 16 equivalence classes of polytopes with only 1 interior lattice point; see [Poonen and Rodriguez-Villegas 2000, Figure 2] or the appendix of this article. So we may assume that V is Δ-nondegenerate with Δ in this list. Now for every facet $\tau \subset \Delta$, the toric component $\mathbb{T}_\tau$ of $X(\Delta)_k$ cuts out an effective k-rational divisor of degree $\ell(\tau)$ on V, where $\ell(\tau)+1$ is the number of lattice points on τ. The result follows since one easily verifies that every polytope in the list contains a facet τ with $\ell(\tau) \leq 3$.

Conversely, suppose that V has index $\iota \leq 3$. If $\iota = 1$, we have shown above that V is nondegenerate. If $\iota = 2$ (respectively $\iota = 3$), using Riemann–Roch one can construct a plane model $f \in k[x, y]$ with $\Delta(f) \subset \operatorname{conv}(\{(0,0), (4,0), (0,2)\})$ (respectively $\Delta(f) \subset 3\Sigma$); see for example [Fisher 2006, Section 3] for details. Then since $V(k) = \varnothing$ and hence $\#k = \infty$, an application of Lemma 3.2 concludes the proof. □

Remark 4.2. There exist genus 1 curves of arbitrarily large index over every number field; see [Clark 2006]. Hence there exist infinitely many genus 1 curves that are not nondegenerate.

5. Nondegeneracy of hyperelliptic curves and C_{ab} curves

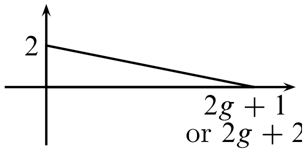
Hyperelliptic curves. A curve V over k of genus $g \geq 2$ is *hyperelliptic* if there is a nonconstant morphism $V \to \mathbb{P}^1_k$ of degree 2. The morphism is automatically separable [Hartshorne 1977, Proposition IV.2.5] and the curve can be defined by a Weierstrass equation

$$y^2 + q(x)y = p(x). \tag{4}$$

Here $p(x), q(x) \in k[x]$ satisfy

$$2 \deg q(x) \leq \deg p(x) \quad \text{and} \quad \deg p(x) \in \{2g+1, 2g+2\}.$$

The universal such curve has Newton polytope as follows:



By Lemma 3.2, if $\#k \geq 6g+5$ then V is nondegenerate. In particular, if $\#k \geq 17$ then every curve of genus 2 is nondegenerate.

If $\operatorname{char} k \neq 2$, we can drop the condition on $\#k$ by completing the square, as in the elliptic curve case. This observation immediately weakens the condition to $\#k \geq 2^{\lfloor \log_2(6g+5) \rfloor} + 1$. As a consequence, $\#k \geq 17$ is also sufficient for every hyperelliptic curve of genus 3 or 4 to be nondegenerate.

Conversely, any curve defined by a nondegenerate polynomial as in (4) is hyperelliptic. We conclude that $\dim \mathcal{M}_\Delta = \dim \mathcal{H}_g = 2g-1$; see [Hartshorne 1977, Example IV.5.5.5].

One can decide if a nondegenerate polynomial f defines a hyperelliptic curve by the following criterion, which also appears in [Koelman 1991, Lemma 3.2.9] with a more complicated proof.

Lemma 5.1. *Let $f \in k[x^\pm, y^\pm]$ be nondegenerate and let $\#\operatorname{int}(\Delta(f) \cap \mathbb{Z}^2) \geq 2$. Then $V(f)$ is hyperelliptic if and only if the interior lattice points of $\Delta(f)$ are collinear.*

Proof. We may assume that $\Delta = \Delta(f)$ has $g \geq 3$ interior lattice points, since all curves of genus 2 are hyperelliptic and any two points are collinear.

Let $L \subset k(V)$ be the subfield generated by all quotients of functions in $\mathcal{L}(K)$, where K is a canonical divisor on V. Then L does not depend on the choice of K, and L is isomorphic to the rational function field $k(\mathbb{P}^1_k)$ if and only if V is hyperelliptic.

We now show that $L \cong k(\mathbb{P}^1_k)$ if and only if the interior lattice points of Δ are collinear. We may assume $(0,0)$ is in the interior of Δ. Then from Proposition 1.7, we see that L contains all monomials of the form $x^i y^j$ for $(i,j) \in \operatorname{int}(\Delta) \cap \mathbb{Z}^2$. In particular, if the interior lattice points of Δ are not collinear then after a transformation we may assume further that $(0,1), (1,0) \in \operatorname{int}(\Delta)$, whence $L \supset k(x,y) = k(V)$; and if they are collinear, then clearly $L \cong k(\mathbb{P}^1_k)$. The result then follows. □

For this reason, we call a lattice polytope *hyperelliptic* if its interior lattice points are collinear.

A curve V over k of genus $g \geq 2$ is said to be *geometrically hyperelliptic* if $V_{\bar{k}} = V \times_k \bar{k}$ is hyperelliptic. Every hyperelliptic curve is geometrically hyperelliptic, but not conversely: If $V \to C \subset \mathbb{P}^{g-1}_k$ is the canonical morphism, then V is hyperelliptic if and only if $C \cong \mathbb{P}^1_k$. This latter condition is satisfied if and only if $C(k) \neq \varnothing$, which is guaranteed in each of the following cases: k is finite, $V(k) \neq \varnothing$, or g is even.

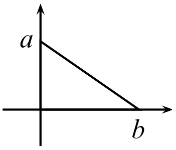
Lemma 5.2. *Let V be a geometrically hyperelliptic curve that is nonhyperelliptic. Then V is not nondegenerate.*

Proof. Suppose that V is geometrically hyperelliptic and Δ-nondegenerate for some lattice polytope Δ. Then applying Lemma 5.1 to $V_{\bar{k}}$, we see that the interior lattice points of Δ are collinear. But then again by Lemma 5.1 (now applied to V itself), V must be hyperelliptic. □

***C_{ab}* curves.** Let $a, b \in \mathbb{Z}_{\geq 2}$ be coprime. A *C_{ab} curve* is a curve having a rational place with Weierstrass semigroup $a\mathbb{Z}_{\geq 0} + b\mathbb{Z}_{\geq 0}$; see [Miura 1992]. Any C_{ab} curve is defined by a Weierstrass equation

$$f(x,y) = \sum_{\substack{i,j \in \mathbb{N} \\ ai+bj \leq ab}} c_{ij} x^i y^j = 0. \tag{5}$$

with $c_{0a}, c_{b0} \neq 0$. By Lemma 3.2, if $\#k \geq 2(g+a+b-2)$ then we may assume that this polynomial is nondegenerate with respect to its Newton polytope Δ_{ab}:



Conversely, every curve given by a Δ_{ab}-nondegenerate polynomial is C_{ab}, and the unique place dominating the point at projective infinity has Weierstrass semigroup $a\mathbb{Z}_{\geq 2} + b\mathbb{Z}_{\geq 2}$; see [Matsumoto 1998]. Note that if k is algebraically closed, the class of hyperelliptic curves of genus g coincides with the class of $C_{2,2g+1}$ curves.

The moduli space of all C_{ab} curves (for varying a and b) of fixed genus g is then a finite union of moduli spaces $\mathcal{M}_{\Delta_{ab}}$. One can show that its dimension equals $2g-1$ by an analysis of the Weierstrass semigroup, which has been done by Rim and Vitulli [1977, Corollary 6.3]. This dimension equals $\dim \mathcal{H}_g = \dim \mathcal{M}_{\Delta_{2,2g+1}}$ and in fact this is the dominating part: In Example 8.7 we will show that $\dim \mathcal{M}_{\Delta_{ab}} < 2g-1$ if $a, b \geq 3$ and $g \geq 6$.

6. Nondegeneracy of curves of genus three and four

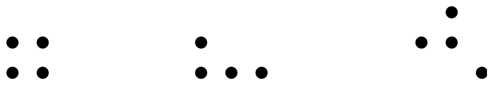
***Curves of genus* 3.** A genus 3 curve V over k is either geometrically hyperelliptic or it canonically embeds in $\mathbb{P}^2_k$ as a plane quartic.

If V is geometrically hyperelliptic, then V may not be hyperelliptic and hence (by Lemma 5.2) not nondegenerate. For example, over $\mathbb{Q}$ there exist degree 2 covers of the imaginary circle having genus 3. However, if k is finite or $V(k) \neq \varnothing$ then every geometrically hyperelliptic curve is hyperelliptic. If moreover $\#k \geq 17$ we can conclude that V is nondegenerate. See Section 5 for more details.

If V is embedded as a plane quartic, then assuming $\#k \geq 17$, we can apply Lemma 3.2 and see that V is defined by a 4Σ-nondegenerate Laurent polynomial.

***Curves of genus* 4.** Let V be a curve of genus 4 over k. If V is a geometrically hyperelliptic curve, then it is hyperelliptic, since the genus is even. Thus if $\#k \geq 17$ then V is nondegenerate (see Section 5). Assume therefore that V is nonhyperelliptic. Then it canonically embeds as a curve of degree 6 in $\mathbb{P}^3_k$ that is the complete intersection of a unique quadric surface Q and a (nonunique) cubic surface C [Hartshorne 1977, Example IV.5.2.2].

First, we note that if V is Δ-nondegenerate for some nonhyperelliptic lattice polytope $\Delta \subset \mathbb{R}^2$, then Q or C must have combinatorial origins as follows. Let $\Delta^{(1)} = \mathrm{conv}(\mathrm{int}(\Delta) \cap \mathbb{Z}^2)$. Up to equivalence, there are three possible arrangements for these interior lattice points:



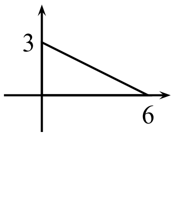
(a) (b) (c)

By Proposition 1.7, one verifies that V canonically maps to $X^{(1)} = X(\Delta^{(1)})_k \subset \mathbb{P}^3_k$. In (a), $X^{(1)}$ is nothing else but the Segre product $\mathbb{P}^1_k \times \mathbb{P}^1_k$ defined by the equation $xz = yw$ in $\mathbb{P}^3_k$, and by uniqueness it must equal Q. For (b), $X^{(1)}$ is the singular quadric cone $yz = w^2$, which again must equal Q. For (c), $X^{(1)}$ is the singular cubic $xyz = w^3$, which must be an instance of C. Note that a curve V can be Δ-nondegenerate with $\Delta^{(1)}$ as in (a) or (b), but not both: whether Q is smooth or not is intrinsic, since Q is unique. The third type (c) is special, and we leave it as an exercise to show that the locus of curves of genus 4 that canonically lie on such a singular cubic surface is a codimension ≥ 2 subspace of $\mathcal{M}_4$ (use the dimension bounds from Section 8).

With these observations in mind, we work towards conditions under which our given nonhyperelliptic genus 4 curve V is nondegenerate. Suppose first that the quadric Q has a (necessarily k-rational) singular point T; then V is called *conical*. This corresponds to the case where $V_{\bar{k}} = V \times_k \bar{k}$ has a unique g^1_3, and represents a codimension 1 subscheme of $\mathcal{M}_4$ [Hartshorne 1977, Exercise IV.5.3]. If $Q(k) = \{T\}$, then V cannot be nondegenerate with respect to any polytope with $\Delta^{(1)}$ as in (a) or (b), since then Q is not isomorphic to either of the corresponding canonical quadric surfaces $X^{(1)}$. If $Q(k) \supsetneq \{T\}$, which is guaranteed if k is finite or if $V(k) \neq \varnothing$, then after a choice of coordinates we can identify Q with the weighted projective space $\mathbb{P}(1, 2, 1)$. Our degree 6 curve V then has an equation of the form

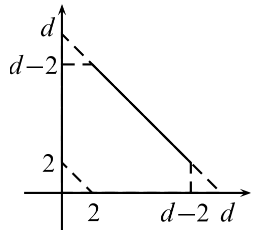
$$f(x, y, z) = y^3 + a_2(x, z)y^2 + a_4(x, z)y + a_6(x, z)$$

with $\deg a_i = i$; the equation is monic in y because $T \notin V$. By Lemma 3.2, if $\#k \geq 23$ then we may assume that $f(x, y, 1)$ is nondegenerate with respect to its Newton polytope Δ as follows:



Next, suppose that Q is smooth; then V is called *hyperboloidal*. This corresponds to the case where $V_{\bar{k}}$ has two g_3^1's, and represents a dense subscheme of $\mathcal{M}_4$ [Hartshorne 1977, Exercise IV.5.3]. If $Q \not\cong \mathbb{P}_k^1 \times \mathbb{P}_k^1$ (for example, this will be the case whenever the discriminant of Q is nonsquare), then again V cannot be nondegenerate with respect to Δ with $\Delta^{(1)}$ as in (a) or (b). Therefore suppose that k is algebraically closed. Then $Q \cong \mathbb{P}_k^1 \times \mathbb{P}_k^1$ and V can be projected to a plane quintic with 2 nodes [Hartshorne 1977, Exercise IV.5.4].

Consider the line connecting these nodes. Generically, it will intersect the nodes with multiplicity 2, that is, it will intersect all branches transversally. By Bezout, the line will then intersect the curve transversally in one other point. This observation fits within the following general phenomenon. Let $d \in \mathbb{Z}_{\geq 4}$, and consider the polytope $\Delta = d\Sigma$ with up to three of its angles pruned as follows:



(6)

Let $f \in k[x, y]$ be a nondegenerate polynomial with Newton polytope Δ. If we prune no angle of $d\Sigma$, then $X(\Delta)_k \cong \mathbb{P}_k^2$ (it is the image of the d-uple embedding) and $V(f)$ is a smooth plane curve of degree d. Pruning an angle has the effect of blowing up $X(\Delta)_k$ at a coordinate point; the image of $V(f)$ under the natural projection $X(\Delta)_k \to \mathbb{P}_k^2$ has a node at that point. If we prune $m = 2$ (respectively $m = 3$) angles, then we likewise obtain the blow-up of $\mathbb{P}_k^2$ at m points and the image of $V(f)$ in $\mathbb{P}_k^2$ has m nodes. Since f is nondegenerate, the line connecting any two of these nodes intersects the curve transversally elsewhere, and due to the shape of Δ, the intersection multiplicity at the nodes will be 2. Conversely, every projective plane curve having at most 3 nodes such that the line connecting any two nodes intersects the curve transversally (also at the nodes themselves) is nondegenerate. Indeed, after an appropriate projective transformation, it will have a Newton polytope as in (6). In particular, our hyperboloidal genus 4 curve V will be Δ-nondegenerate, where Δ equals polytope (h.1) from Section 7 below.

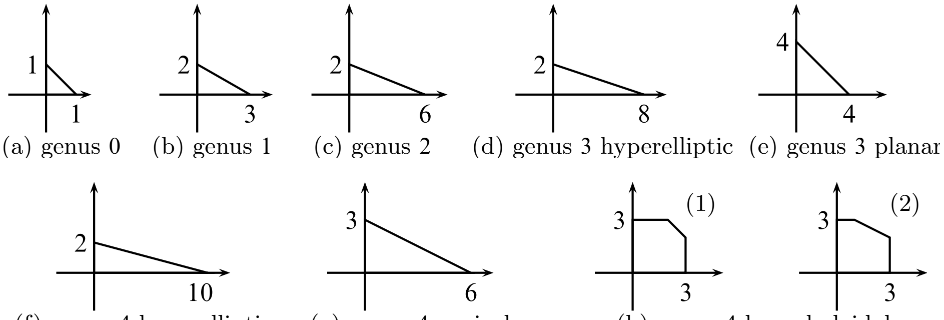
Exceptionally, the line connecting the two nodes of our quintic may be tangent to one of the branches at a node. Using a similar reasoning, we conclude that V is Δ-nondegenerate, with Δ equal to polytope (h.2) from Section 7 below.

Remark 6.1. As in Remark 4.2, an argument based on the index shows that there exist genus 4 curves that are not nondegenerate. A result by Clark [2007] states that for every $g \geq 2$, there exists a number field k and a genus g curve V over k, such that the index of V is equal to $2g - 2$, the degree of the canonical divisor.

In particular, there exists a genus 4 curve V of index 6. Such a curve cannot be nondegenerate. Indeed, for each of the above arrangements (a)–(c), $X^{(1)}$ contains the line $z = w = 0$, which cuts out an effective divisor on V of degree 3 in cases (a) and (b) and degree 2 in case (c).

7. Nondegeneracy of low genus curves: summary

We now summarize the results of the preceding sections. If k is an algebraically closed field, then every curve V of genus at most 4 over k can be modeled by a nondegenerate polynomial having one of the following as Newton polytope:



(a) genus 0 (b) genus 1 (c) genus 2 (d) genus 3 hyperelliptic (e) genus 3 planar

(f) genus 4 hyperelliptic (g) genus 4 conical (h) genus 4 hyperboloidal

Moreover, these classes are disjoint. For the polytopes (c) through (h.1), we have $\dim \mathcal{M}_\Delta = 3, 5, 6, 7, 8, 9$, respectively. All hyperelliptic curves and C_{ab} curves are nondegenerate.

For an arbitrary perfect field k, if V is not hyperboloidal and has genus at most 4, then V is nondegenerate whenever k is a sufficiently large finite field, or when k is infinite and $V(k) \neq \varnothing$; for the former, the condition $\#k \geq 23$ is sufficient but most likely superfluous (see Remark 7.2).

Remark 7.1. We can situate the nonhyperelliptic C_{ab} curves that lie in this classification. In genus 3, we have $C_{3,4}$ curves, which have a smooth model in $\mathbb{P}^2_k$, since $\Delta_{3,4}$ is nonhyperelliptic. In genus 4, we have $C_{3,5}$ curves, which are conical; this can be seen by analyzing the interior lattice points of $\Delta_{3,5}$, as in Section 6.

Remark 7.2. In case $\#k < \infty$, we proved (without further condition on $\#k$) that if V is not hyperboloidal then it can be modeled by a polynomial $f \in k[x, y]$ with Newton polytope contained in one of the polytopes (a)–(g). The condition on $\#k$ then came along with an application of Lemma 3.2 to deduce nondegeneracy. In the $g = 1$ case, we got rid of this condition by using nonlinear transformations (completing the square) and allowing smaller polytopes. Similar techniques can be used to improve (and probably even remove) the bounds on $\#k$ in genera $2 \leq g \leq 4$. For example, using naive brute force computation we have verified that in genus 2, all curves are nondegenerate whenever $\#k = 2, 4, 8$.

8. An upper bound for $\dim \mathcal{M}_g^{\mathrm{nd}}$

From now on, we assume $k = \overline{k}$. In this section, we prepare for a proof of Theorem 11.1, which gives an upper bound for $\dim \mathcal{M}_g^{\mathrm{nd}}$ in terms of g.

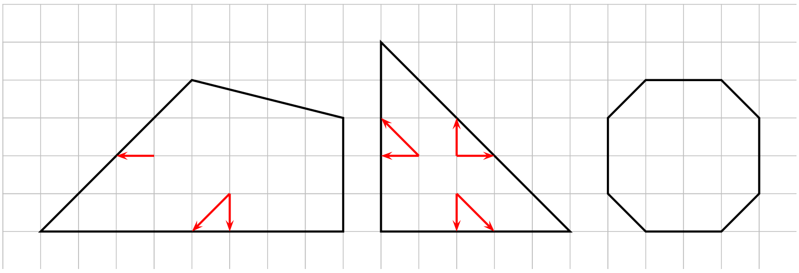
For a lattice polytope $\Delta \subset \mathbb{Z}^2$ with $g \geq 2$ interior lattice points, we sharpen the obvious upper bound $\dim \mathcal{M}_\Delta \leq \dim M_\Delta = \#(\Delta \cap \mathbb{Z}^2) - 1$ (see (2)) by incorporating the action of the automorphism group of $X(\Delta)_k$, which has been explicitly described by Bruns and Gubeladze [2002, Section 5]. In Sections 9–11 we then work towards a bound in terms of g, following ideas of Haase and Schicho [2009].

The automorphisms of $X(\Delta)_k = \operatorname{Proj} k[\Delta] \hookrightarrow \mathbb{P}^{\#(\Delta\cap\mathbb{Z}^2)-1}$ correspond to the graded k-algebra automorphisms of $k[\Delta]$, and admit a combinatorial description as follows.

Definition 8.1. A nonzero vector $v \in \mathbb{Z}^2$ is a *column vector* of Δ if there exists a facet $\tau \subset \Delta$ (the *base facet*) such that $v + ((\Delta \setminus \tau) \cap \mathbb{Z}^2) \subset \Delta$.

We denote by $c(\Delta)$ the number of column vectors of Δ.

Example 8.2. Any multiple of the standard 2-simplex Σ has 6 column vectors. The octagonal polytope below shows that a polytope may have no column vectors.



The dimension of the automorphism group $\operatorname{Aut}(X(\Delta)_k)$ is then given as follows:

Proposition 8.3 [Bruns and Gubeladze 2002, Theorem 5.3.2]. *We have*

$$\dim \operatorname{Aut}(X(\Delta)_k) = c(\Delta) + 2.$$

Proof sketch. One begins with the 2-dimensional subgroup of $\operatorname{Aut}(X(\Delta)_k)$ induced by the inclusion $\operatorname{Aut}(\mathbb{T}^2) \hookrightarrow \operatorname{Aut}(X(\Delta)_k)$. On the $k[\Delta]$-side, this corresponds to the graded automorphisms induced by $(x, y) \mapsto (\lambda x, \mu y)$ for $\lambda, \mu \in k^{*2}$.

Next, column vectors of Δ correspond to automorphisms of $X(\Delta)_k$ as follows. If v is a column vector, modulo equivalence we may assume that $v = (0, -1)$, that the base facet is supported on the X-axis, and that Δ is contained in the positive quadrant $\mathbb{R}^2_{\geq 0}$. Let $f(x, y) \in k[x, y]$ be supported on Δ. Since $v = (0, -1)$ is a column vector, the polynomial $f(x, y + \lambda)$ will again be supported on Δ for any $\lambda \in k$. Hence v induces a family of graded automorphisms $k[\Delta] \to k[\Delta]$, corresponding to a one-dimensional subgroup of $\operatorname{Aut}(X(\Delta)_k)$.

It then remains to show that these subgroups are algebraically independent from each other and from $\mathrm{Aut}(\mathbb{T}^2)$, and that together they generate $\mathrm{Aut}(X(\Delta)_k)$ (after including the finitely many automorphisms coming from $\mathbb{Z}$-affine transformations mapping Δ to itself). □

Using the fact that a curve of genus $g \geq 2$ has finitely many automorphisms we obtain the following corollary. We leave the details as an exercise.

Corollary 8.4. $\dim \mathcal{M}_\Delta \leq m(\Delta) := \#(\Delta \cap \mathbb{Z}^2) - c(\Delta) - 3.$

Remark 8.5. In order to have equality, it is sufficient that Δ is a so-called *maximal polytope* (see Section 10 for the definition). This is the main result of Koelman's thesis [1991, Theorem 2.5.12].

Example 8.6. Let $\Delta = \mathrm{conv}(\{(0,0), (2g+2,0), (0,2)\})$ as in Section 5, so that $\dim \mathcal{M}_\Delta = 2g-1$. One checks that $c(\Delta) = g+4$, so the upper bound in Corollary 8.4 reads $m(\Delta) = (3g+6) - (g+4) - 3 = 2g-1$; in this case, the bound is sharp. It is easy to verify that the bound is also sharp if $\Delta = d\Sigma$, $d \in \mathbb{Z}_{\geq 4}$; then $\dim \mathcal{M}_\Delta$ reads $(d+1)(d+2)/2 - 9 = g + 3d - 9 \leq 2g$. The latter are examples of maximal polytopes. Opposed to this, let $(d\Sigma)_0$ be obtained from $d\Sigma$ by pruning off $(0,0)$. This reduces the number of lattice points by 1 and the number of column vectors by 2. Hence our bound increases, although $d\Sigma$ and $(d\Sigma)_0$ give rise to the same moduli space. Indeed, pruning off $(0,0)$ only forces our curves in $X(d\Sigma)_k \cong \mathbb{P}^2_k$ to pass through $(0,0,1)$.

Example 8.7. We now use Corollary 8.4 to show that the dimension of the moduli space of nonhyperelliptic C_{ab} curves of genus g (where a and b may vary) has dimension strictly smaller than $2g-1 = \dim \mathcal{H}_g$ whenever $g \geq 6$. Consider $\Delta_{ab} = \mathrm{Conv}\{(0,a), (b,0), (0,0)\}$ with $a, b \in \mathbb{Z}_{\geq 3}$ coprime. Then we have

$$g = (a-1)(b-1)/2 \quad \text{and} \quad \#(\Delta \cap \mathbb{Z}^2) = g + a + b + 1,$$

and the set of column vectors is given by

$$\{(n,-1) : n = 0, \ldots, \lfloor b/a \rfloor\} \cup \{(-1,m) : m = 0, \ldots, \lfloor a/b \rfloor\}.$$

Suppose without loss of generality that $a < b$. Then a is bounded by $\sqrt{2g}+1$. Corollary 8.4 yields

$$\dim \mathcal{M}_\Delta \leq m(\Delta) = g + a + b + 1 - \left(\left\lfloor \frac{b}{a} \right\rfloor + 2\right) - 3 < a + \frac{2g-1}{a} + g - 2.$$

As a (real) function of a, this upper bound has a unique minimum at $a = \sqrt{2g-1}$. Therefore, to deduce that it is strictly smaller than $2g-1$ for all $a \in [3, \sqrt{2g}+1]$, it suffices to verify so for the boundary values $a = 3$ and $a = \sqrt{2g}+1$, which is indeed the case if $g \geq 6$.

9. A bound in terms of the genus

For the rest of this article, we will employ the following notation. Let $\Delta^{(1)}$ be the convex hull of the interior lattice points of Δ. Let r and $r^{(1)}$ denote the number of lattice points on the boundary of Δ and $\Delta^{(1)}$, respectively, and let $g^{(1)}$ denote the number of interior lattice points in $\Delta^{(1)}$, so that $g = g^{(1)} + r^{(1)}$.

We now prove the following preliminary bound.

Proposition 9.1. *If Δ has at least $g \geq 2$ interior lattice points, then we have* $\dim \mathcal{M}_\Delta \leq 2g+3$.

Proof. We may assume that Δ is nonhyperelliptic; otherwise $\dim \mathcal{M}_\Delta \leq 2g-1$ by Lemma 5.1. We may also assume that $\Delta^{(1)}$ is not a multiple of Σ, since otherwise Δ-nondegenerate curves are canonically embedded in $X(\Delta^{(1)})_k \cong \mathbb{P}^2_k$ using Proposition 1.7; then from Example 8.6 it follows that $\dim \mathcal{M}_\Delta \leq 2g$.

An upper bound for $\dim \mathcal{M}_\Delta$ in terms of g then follows from a lemma by Haase and Schicho [2009, Lemma 12], who proved that $r \leq r^{(1)} + 9$, in which equality holds if and only if $\Delta = d\Sigma$ for some $d \in \mathbb{Z}_{\geq 4}$ (a case which we have excluded). Hence

$$\#(\Delta \cap \mathbb{Z}^2) = g + r \leq g + r^{(1)} + 8 = 2g + 8 - g^{(1)}, \tag{7}$$

and thus

$$\dim \mathcal{M}_\Delta \leq m(\Delta) = \#(\Delta \cap \mathbb{Z}^2) - c(\Delta) - 3 \leq 2g + 5 - c(\Delta) - g^{(1)} \leq 2g + 5. \tag{8}$$

This bound improves to $2g+3$ if $g^{(1)} \geq 2$, so two cases remain: $g^{(1)} = 0$ and $g^{(1)} = 1$.

Suppose first that $g^{(1)} = 0$. Then by Lemma 9.2 below, any Δ-nondegenerate curve is either a smooth plane quintic (excluded), or a trigonal curve. Since the moduli space of trigonal curves has dimension $2g+1$ (a classical result, see also Section 12 below), the bound holds.

Next suppose that $g^{(1)} = 1$. Then, up to equivalence, there are only 16 possibilities for $\Delta^{(1)}$; these are listed in [Poonen and Rodriguez-Villegas 2000, Figure 2] or in the appendix below. Hence, there are only finitely many possibilities for Δ, and for each of these polytopes we find that $\#(\Delta \cap \mathbb{Z}^2) - c(\Delta) - 3 \leq 2g + 2$. □

In fact, for all but the 5 polytopes in Figure 1 (up to equivalence), we find that the stronger bound $\#(\Delta \cap \mathbb{Z}^2) - c(\Delta) - 3 \leq 2g + 1$ holds.

Lemma 9.2. *If $\Delta^{(1)}$ is a 2-dimensional polytope having no interior lattice points, then any Δ-nondegenerate curve is either trigonal or isomorphic to a smooth quintic in $\mathbb{P}^2_k$.*

Proof. Koelman's proof, from his PhD thesis [1991, Lemma 3.2.13], is based on Petri's theorem. A more combinatorial argument uses the fact that lattice polytopes

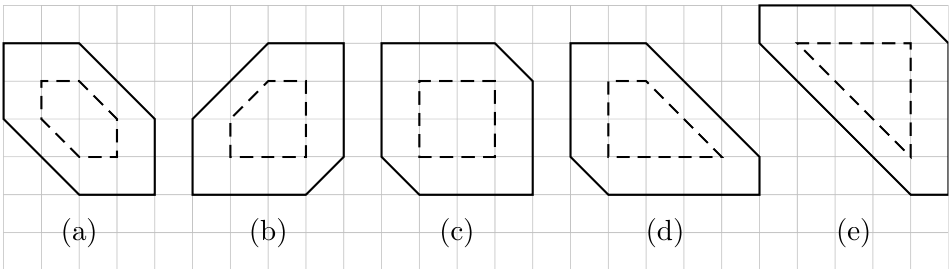


Figure 1. Polytopes with $g^{(1)} = 1$ and $\#(\Delta \cap \mathbb{Z}^2) - c(\Delta) - 3 = 2g + 2$.

of genus 0 are equivalent with either 2Σ, or with a polytope that is caught between two horizontal lines of distance 1. This was proved independently by Arkinstall, Khovanskii, Koelman, and Schicho (see the generalized statement by Batyrev and Nill [2007, Theorem 2.5]).

In the first case, Δ-nondegenerate curves canonically embed in $X(2\Sigma)_k \cong \mathbb{P}^2_k$; hence they are isomorphic to smooth plane quintics.

In the second case, it follows that Δ is caught between two horizontal lines of distance 3. This may actually fail if $\Delta^{(1)} = \Sigma$, which corresponds to smooth plane quartics. But in both situations, Δ-nondegenerate curves are trigonal. □

10. Refining the upper bound: Maximal polytopes

We further refine the bound in Proposition 9.1 by adapting the proof of the Haase–Schicho bound $r \leq r^{(1)} + 9$ in order to obtain an estimate for $r - c(\Delta)$ instead of just r. We first do this for *maximal* polytopes, and treat nonmaximal polytopes in the next section.

Definition 10.1. A lattice polytope $\Delta \subset \mathbb{Z}^n$ is *maximal* if Δ is not properly contained in another lattice polytope with the same interior lattice points, that is, for all lattice polytopes $\Delta' \supsetneq \Delta$, we have $\operatorname{int}(\Delta') \cap \mathbb{Z}^n \neq \operatorname{int}(\Delta) \cap \mathbb{Z}^n$.

We define the *relaxed polytope* $\Delta^{(-1)}$ of a lattice polytope $\Delta \subset \mathbb{Z}^2$ as follows. Let $0 \in \Delta$. To each facet $\tau \subset \Delta$ given by an inequality of the form $a_1 X + a_2 Y \leq b$ with $a_i \in \mathbb{Z}$ coprime, we define the *relaxed inequality* $a_1 X + a_2 Y \leq b + 1$ and let $\Delta^{(-1)}$ be the intersection of these relaxed inequalities. If p is a vertex of Δ given by the intersection of two such facets, we define the *relaxed vertex* $p^{(-1)}$ to be the intersection of the boundaries of the corresponding relaxed inequalities.

Lemma 10.2 ([Haase and Schicho 2009, Lemmas 9 and 10] and [Koelman 1991, Section 2.2]). *Let $\Delta \subset \mathbb{Z}^2$ be a 2-dimensional lattice polytope. Then $\Delta^{(-1)}$ is a lattice polytope if and only if $\Delta = \Delta'^{(1)}$ for some lattice polytope Δ'. Furthermore, if Δ is nonhyperelliptic, then Δ is maximal if and only if $\Delta = (\Delta^{(1)})^{(-1)}$.*

The proof of the Haase–Schicho bound $r \leq r^{(1)} + 9$ uses a theorem of Poonen and Rodriguez-Villegas [2000], which we now introduce.

A *legal move* is a pair (v, w) with $v, w \in \mathbb{Z}^2$ such that $\operatorname{conv}(\{0, v, w\})$ is a 2-dimensional triangle whose only nonzero lattice points lie on $e(v, w)$, the edge between v and w. The *length* of a legal move (v, w) is

$$\ell(v, w) = \det\begin{pmatrix} v \\ w \end{pmatrix},$$

which is of absolute value $r - 1$, where $r = \#(e(v, w) \cap \mathbb{Z}^2)$ is the number of lattice points on the edge between v and w. Note that the length can be negative.

A *legal loop* $\mathcal{P}$ is a sequence of vectors $v_1, v_2, \ldots, v_n \in \mathbb{Z}^2$ such that for all $i = 1, \ldots, n$ and indices taken modulo n,

- (v_i, v_{i+1}) is a legal move, and
- v_{i-1}, v_i, v_{i+1} are not contained in a line.

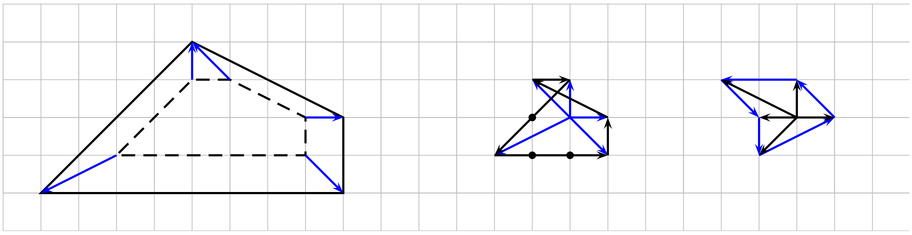
The *length* $\ell(\mathcal{P})$ of a legal loop $\mathcal{P}$ is the sum of the lengths of its legal moves.

The *winding number* of a legal loop is its winding number around 0 in the sense of algebraic topology. The *dual loop* $\mathcal{P}^\vee$ is given by $w_1, \ldots, w_n$, where $w_i = \ell(v_i, v_{i+1})^{-1} \cdot (v_{i+1} - v_i)$ for $i = 1, \ldots, n$. One can check that this is again a legal loop with the same winding number as $\mathcal{P}$ and that $\mathcal{P}^{\vee\vee} = \mathcal{P}$ after a 180° rotation.

Theorem 10.3 [Poonen and Rodriguez-Villegas 2000, Section 9.1]. *Let $\mathcal{P}$ be a legal loop with winding number w. Then $\ell(\mathcal{P}) + \ell(\mathcal{P}^\vee) = 12w$.*

Now let $\Delta \subset \mathbb{Z}^2$ be a maximal polytope with 2-dimensional interior $\Delta^{(1)}$. We associate to Δ a legal loop $\mathcal{P}(\Delta)$ as follows. By Lemma 10.2, Δ is obtained from $\Delta^{(1)}$ by relaxing the edges. Let $p_1, \ldots, p_n$ be the vertices of $\Delta^{(1)}$, enumerated counterclockwise; then $\mathcal{P}(\Delta)$ is given by the sequence $q_i = p_i^{(-1)} - p_i$, where $p_i^{(-1)}$ is the relaxed vertex of p_i.

Example 10.4. The following picture, inspired by [Haase and Schicho 2009, Figure 20], is illustrative. It shows a polytope Δ with 2-dimensional interior $\Delta^{(1)}$, the associated legal loop $\mathcal{P}(\Delta)$, and its dual $\mathcal{P}(\Delta)^\vee$. In this example, the loops satisfy $\ell(\mathcal{P}(\Delta)) = \ell(\mathcal{P}(\Delta)^\vee) = 6$.



A crucial observation is that the bold-marked lattice points of $\mathcal{P}(\Delta)$ are column vectors of Δ. This holds in general and lies at the core of our following refinement of the Haase–Schicho bound.

Lemma 10.5. *If Δ is maximal and nonhyperelliptic, then*

(a) $r - r^{(1)} = \ell(\mathcal{P}(\Delta)) \leq 9$, *and*

(b) $r - r^{(1)} - c(\Delta) \leq \min\bigl(\ell(\mathcal{P}(\Delta)), \ell(\mathcal{P}(\Delta)^\vee)\bigr) \leq 6.$

Proof. We abbreviate $\mathcal{P} = \mathcal{P}(\Delta)$.

Inequality (a) is by [Haase and Schicho 2009, Lemma 11] and works as follows. The length of the legal move (q_i, q_{i+1}) measures the difference between the number of lattice points on the facet of Δ connecting $p_i^{(-1)}$ and $p_{i+1}^{(-1)}$, and the number of lattice points on the edge of $\Delta^{(1)}$ connecting p_i and p_{i+1}. Therefore $r - r^{(1)} = \ell(\mathcal{P})$. The dual loop $\mathcal{P}^\vee$ walks (in a consistent and counterclockwise-oriented way) through the normal vectors of $\Delta^{(1)}$; therefore each move has positive length and we have $\ell(\mathcal{P}(\Delta)^\vee) \geq 3$. Since $\mathcal{P}^\vee$ has winding number 1, the statement follows from Theorem 10.3. (One can further show that equality holds if and only if Δ is a multiple of the standard 2-simplex Σ.)

To prove inequality (b), we first claim that there is a bijection between lattice points v that lie properly on a counterclockwise-oriented (positive length) legal move $q_i q_{i+1}$ of $\mathcal{P}$, and column vectors of Δ with base facet $p_i^{(-1)} p_{i+1}^{(-1)}$. Indeed, after an appropriate transformation, we may assume as we did in Proposition 8.3 that $v = (0, -1)$, that $p_i^{(-1)}$ and $p_{i+1}^{(-1)}$ lie on the X-axis, and that Δ is contained in the quadrant $\mathbb{R}^2_{\geq 0}$; after these normalizations, the claim is straightforward.

Now, since the dual loop $\mathcal{P}^\vee$ consists of counterclockwise-oriented legal moves only, it has at most $\ell(\mathcal{P}^\vee)$ vertices. Since $\mathcal{P} = \mathcal{P}^{\vee\vee}$ (after 180° rotation), $\mathcal{P}$ has at most $\ell(\mathcal{P}^\vee)$ vertices. By the claim, we have $\ell(\mathcal{P}) \leq \ell(\mathcal{P}^\vee) + c(\Delta)$, and the result follows by combining this with part (a) and Theorem 10.3. □

Corollary 10.6. *If Δ is maximal, then* $\dim \mathcal{M}_\Delta \leq 2g + 3 - g^{(1)}$. *In particular, if* $g^{(1)} \geq 2$, *then* $\dim \mathcal{M}_\Delta \leq 2g + 1$.

Proof. Lemma 10.5 gives $m(\Delta) = g + r - 3 - c(\Delta) \leq g + r^{(1)} + 3 \leq 2g + 3 - g^{(1)}$. □

Remark 10.7. Note that Lemma 10.5(a) immediately extends to nonmaximal polytopes ($r - r^{(1)}$ can only decrease), so the Haase–Schicho bound holds for arbitrary nonhyperelliptic polytopes. This we cannot conclude for part (b): If r decreases, $c(\Delta)$ may decrease more quickly so that the bound no longer holds. An example of such behavior can be found in Figure 1(c).

11. Refining the upper bound: General polytopes

We are now ready to prove the main result of Sections 8–11.

Theorem 11.1. *If $g \geq 2$, then* $\dim \mathcal{M}_g^{\mathrm{nd}} \leq 2g+1$ *except for $g=7$ where we have* $\dim \mathcal{M}_7^{\mathrm{nd}} \leq 16$.

Proof. It suffices to show that the claimed bounds hold for all polytopes Δ with g interior lattice points. By the proof of Proposition 9.1, we may assume that $\Delta^{(1)}$ is two-dimensional, that it is not a multiple of Σ, and that it has $g^{(1)} \geq 1$ interior lattice points.

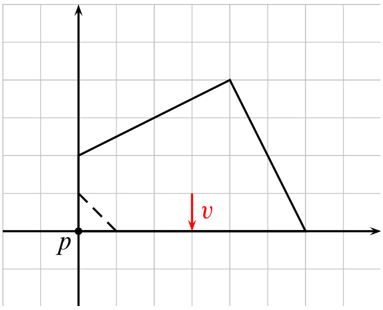
Let us first assume that $g^{(1)} \geq 2$. We will show that $\dim \mathcal{M}_\Delta \leq 2g+1$. From Corollary 10.6, we know that this is true if Δ is maximal. Therefore, suppose that Δ is nonmaximal; then it is obtained from a maximal polytope $\widetilde{\Delta}$ by taking away points on the boundary (keeping the interior lattice points intact). If two or more boundary points are taken away, then as in (8) we have

$$m(\Delta) \leq \#(\Delta \cap \mathbb{Z}^2) - 3 \leq \#(\widetilde{\Delta} \cap \mathbb{Z}^2) - 2 - 3 \leq 2g + 5 - g^{(1)} - 2 \leq 2g+1.$$

So we may assume that $\Delta = \mathrm{conv}(\widetilde{\Delta} \cap \mathbb{Z}^2 \setminus \{p\})$ for a vertex $p \in \widetilde{\Delta}$. Similarly, we may assume that $c(\Delta) < c(\widetilde{\Delta})$, for else

$$m(\Delta) = \#(\Delta \cap \mathbb{Z}^2) - c(\Delta) - 3 \leq \#(\widetilde{\Delta} \cap \mathbb{Z}^2) - c(\widetilde{\Delta}) - 3 = m(\widetilde{\Delta}) \leq 2g+1.$$

Let v be a column vector of $\widetilde{\Delta}$ that is not a column vector of $\Delta = \mathrm{conv}(\widetilde{\Delta} \cap \mathbb{Z}^2 \setminus \{p\})$. Then p must lie on the base facet τ of v. After an appropriate transformation, we may assume that $p = (0,0)$, that $v = (0,-1)$, that τ lies along the X-axis, and that $\widetilde{\Delta}$ lies in the positive quadrant, as below:



Note that $(1,1) \in \mathrm{int}(\widetilde{\Delta})$ since otherwise v would still be a column vector of Δ. But then the other facet of $\widetilde{\Delta}$ that contains p must be supported on the Y-axis, for else $(1,1)$ would no longer be in $\mathrm{int}(\Delta)$. One can now verify that if $f(x,y)$ is Δ-nondegenerate, then for all but finitely many $\lambda \in k$, the polynomial $f(x, y+\lambda)$ will have Newton polytope $\widetilde{\Delta}$ and all but finitely of those will be $\widetilde{\Delta}$-nondegenerate. Therefore, we have $\mathcal{M}_\Delta \subset \mathcal{M}_{\widetilde{\Delta}}$, and the dimension estimate follows.

Now suppose $g^{(1)} = 1$. From the finite computation in Proposition 9.1, we know that the bound $\dim \mathcal{M}_\Delta \leq 2g+1$ holds if Δ is not among the polytopes listed in Figure 1. Now in this list, the polytopes (b)–(e) are not maximal, and for these polytopes the same trick as in the $g^{(1)} \geq 2$ case applies. However, polytope (a) is maximal and contains 7 interior lattice points: Therefore, we can only prove $\dim \mathcal{M}_7^{\mathrm{nd}} \leq 16$. □

Let Δ be a nonmaximal nonhyperelliptic lattice polytope, and let $\widetilde{\Delta} = (\Delta^{(1)})^{(-1)}$ be the smallest maximal polytope that contains Δ. Let $f \in k[x^{\pm}, y^{\pm}]$ be a Δ-nondegenerate Laurent polynomial. Since $\Delta \subset \widetilde{\Delta}$, we can consider the (degree 1) locus $\widetilde{V}$ of $f = 0$ in $X(\widetilde{\Delta})_k = \operatorname{Proj} k[\widetilde{\Delta}]$. Then one can wonder whether the observation we made in the proof of Theorem 11.1 holds in general: Is there always a $\sigma \in \operatorname{Aut}(X(\widetilde{\Delta})_k)$ such that $\sigma(\widetilde{V}) \cap \mathbb{T}_k^2$ is defined by a $\widetilde{\Delta}$-nondegenerate polynomial? The answer is no, because it is easy to construct examples where the only automorphisms of $X(\widetilde{\Delta})_k$ are those induced by $\operatorname{Aut}(\mathbb{T}_k^2)$. Then $\sigma(\widetilde{V}) \cap \mathbb{T}_k^2$ is always defined by $f(\lambda x, \mu y)$ (for some $\lambda, \mu \in k^*$), which does not have $\widetilde{\Delta}$ as its Newton polytope and hence cannot be $\widetilde{\Delta}$-nondegenerate.

However, f is very close to being $\widetilde{\Delta}$-nondegenerate, and this line of thinking leads to the following observation. Let p be a vertex of $\widetilde{\Delta}$ that is not a vertex of Δ, and let q_1 and q_2 be the closest lattice points to p on the respective facets of $\widetilde{\Delta}$ containing p. The triangle spanned by p, q_1, and q_2 cannot contain any other lattice points, because otherwise removing p would affect the interior of $\widetilde{\Delta}$. Thus the volume of this triangle is equal to 1/2 by Pick's theorem, and the affine chart of $X(\widetilde{\Delta})_k$ attached to the cone at p is isomorphic to $\mathbb{A}_k^2$. In particular, $X(\widetilde{\Delta})_k$ is nonsingular in the zero-dimensional torus $\mathbb{T}_p$ corresponding to p. Then f fails to be $\widetilde{\Delta}$-nondegenerate only because $\widetilde{V}$ passes through $\mathbb{T}_p$ (that is, it passes through $(0, 0) \in \mathbb{A}_k^2$); elsewhere it fulfills the conditions of nondegeneracy: $\widetilde{V}$ is smooth, intersects the 1-dimensional tori associated to the facets of $\widetilde{\Delta}$ transversally, and does not contain the singular points of $X(\widetilde{\Delta})_k$. Now following the methods of Section 2, one could construct the bigger moduli space of curves satisfying this weaker nondegeneracy condition. Its dimension would still be bounded by $\#(\widetilde{\Delta} \cap \mathbb{Z}^2) - c(\widetilde{\Delta}) - 3$, which by Lemma 10.5 is at most $2g + 3 - g^{(1)}$ because $\widetilde{\Delta}$ is maximal. Therefore $\dim \mathcal{M}_\Delta \leq 2g + 3 - g^{(1)}$ for nonmaximal Δ, and this yields an alternative proof of Theorem 11.1. Related observations have been made by Koelman [1991, Section 2.6].

12. Trigonal curves, trinodal sextics, and sharpness of our bounds

For $g \geq 2$, we implicitly proved in Section 5 that $\dim \mathcal{M}_g^{\mathrm{nd}} \geq 2g - 1$. But already in genera 3 and 4, by the results in Section 6 we have $\dim \mathcal{M}_3^{\mathrm{nd}} = 6$ and $\dim \mathcal{M}_4^{\mathrm{nd}} = 9$, so this lower bound is an underestimation. For higher genera, we prove in this last section that the bounds given in Theorem 11.1 are sharp, mainly by investigating spaces of trigonal curves. Our main result is the following.

Theorem 12.1. *If* $g \geq 4$, *then* $\dim \mathcal{M}_g^{\mathrm{nd}} = 2g + 1$ *unless* $g = 7$, *where* $\dim \mathcal{M}_7^{\mathrm{nd}} = 16$.

Proof. It suffices to find for every genus $g \geq 5$ a lattice polytope Δ with g interior lattice points, for which $\dim \mathcal{M}_\Delta = 2g+1$ if $g \neq 7$, and $\dim \mathcal{M}_\Delta = 16$ if $g = 7$. If $g = 2h$ is even, let Δ be the rectangle

$$\operatorname{conv}\left(\{(0,0),(0,3),(h+1,3),(h+1,0)\}\right). \tag{9}$$

Note that then $\#(\Delta \cap \mathbb{Z}^2) = 2g+8$ and $c(\Delta) = 4$. If $g = 2h+1$ is odd but different from 7, let Δ be the trapezium

$$\operatorname{conv}\left(\{(0,0),(0,3),(h,3),(h+3,0)\}\right). \tag{10}$$

Again, $\#(\Delta \cap \mathbb{Z}^2) = 2g+8$ and $c(\Delta) = 4$. Finally, if $g = 7$ then, let Δ be

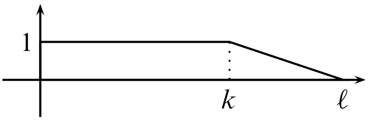
$$\operatorname{conv}\{(2,0),(0,2),(-2,2),(-2,0),(0,-2),(2,-2)\} \tag{11}$$

(that is, the polytope given in Figure 1(a)). Here, $\#(\Delta \cap \mathbb{Z}^2) = 19$ and $c(\Delta) = 0$.

We first prove that

$$\dim \mathcal{M}_\Delta = \#(\Delta \cap \mathbb{Z}^2) - 1 - \dim \operatorname{Aut}(X(\Delta)_k), \tag{12}$$

holds for the families of polytopes (9) and (10), for which the result then follows from Proposition 8.3. This can be achieved using the well-known theory of trigonal curves [Coppens 1986; Maroni 1946]. More generally, let $k, \ell \in \mathbb{Z}_{\geq 2}$ satisfy $k \leq \ell$, let $\Delta^{(1)}$ be the trapezium



and let $\Delta = \Delta^{(1)(-1)}$. Then if a curve V is Δ-nondegenerate, it is trigonal of genus $g = k + \ell + 2$. By Proposition 1.7, it can be canonically embedded in $X(\Delta^{(1)})_k$, which is the rational surface scroll $S_{k,\ell} \subset \mathbb{P}_k^{g-1}$. Then by Petri's theorem [Arbarello et al. 1985], this scroll is the intersection of all quadrics containing the canonical embedding. As a consequence, two different such canonical embeddings must differ by an automorphism of $\mathbb{P}_k^{g-1}$ that maps $X(\Delta^{(1)})_k$ to itself; in other words, any two canonical embeddings of V differ by an automorphism of $X(\Delta^{(1)})_k$.

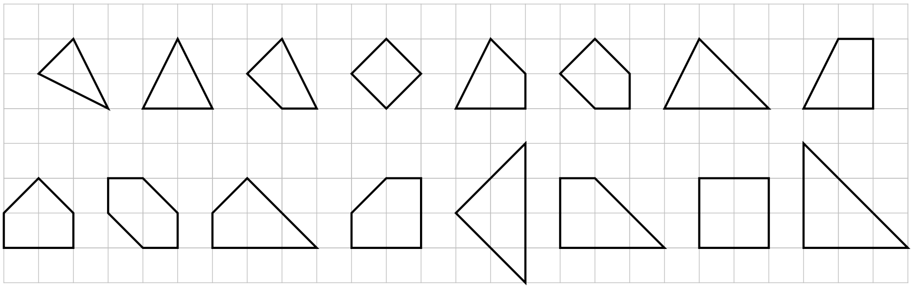
Now let $f_1, f_2 \in k[x^\pm, y^\pm]$ be Δ-nondegenerate polynomials such that $V(f_1)$ and $V(f_2)$ are isomorphic as abstract curves. Since the fans associated to Δ and $\Delta^{(1)}$ are the same, we have $X(\Delta)_k = X(\Delta^{(1)})_k$. Under this identification, $V(f_1)$ and $V(f_2)$ become canonical curves that must differ by an automorphism of $X(\Delta)_k$. Thus we can conclude (12). (Although any trigonal curve is canonically embedded in some rational normal scroll $S_{k,\ell}$ and hence in some $X(\Delta)_k$, it might fail to be nondegenerate because it can be impossible to avoid tangency to $X(\Delta)_k \setminus \mathbb{T}_k^2$.)

To conclude, suppose that Δ is as in (11). We refer to the pruned simplex (6) and the accompanying discussion; here we have $d = 6$. It follows that if f is a Δ-nondegenerate polynomial, then f gives rise to a plane sextic V with three nodes (at the coordinate points) and no other singularities. Conversely, any trinodal sextic is Δ-nondegenerate if any line connecting two nodes intersects the curve transversally elsewhere. Since the latter is an open condition, $\mathcal{M}_\Delta$ is the Zariski closure of the moduli space $\mathcal{V}_{3,6}$ of trinodal plane sextics. The variety $\mathcal{V}_{3,6}$ is in its turn the image of a Severi variety [Severi 1921], and it is a classical result that $\dim \mathcal{V}_{3,6} = 16$ — for a modern treatment, see [Sernesi 1984]. □

Remark 12.2. In his PhD thesis, Koelman [1991, Theorem 2.5.12] proves that Equation (12) holds for any polytope $\Delta \subset \mathbb{R}^2$ that is maximal and nonhyperelliptic. In fact, Koelman assumes $k = \mathbb{C}$, but his methods extend to an arbitrary algebraically closed field $k = \bar{k}$. This provides another proof of Theorem 12.1, but we are content to prove our results in the above more elementary (and classical) way.

Appendix: lattice polytopes of genus one

There are 16 equivalence classes of lattice polytopes having one interior lattice point. Polytopes representing these are drawn below in a copy of [Poonen and Rodriguez-Villegas 2000, Figure 2]. We include the list here for self-containedness. It is an essential ingredient in the proofs of Lemma 4.1 and Proposition 9.1.



Acknowledgments

Part of this paper was written while Castryck was supported by the EPSRC grant EP/C014839/1. He would also like to explicitly thank Marc Coppens for his helpful comments and for referring us to the PhD thesis of Koelman, which we only found out about during the final research phase. Voight would like to thank Bernd Sturmfels and Steven Sperber. Both authors thank the anonymous referee.

References

[Adolphson and Sperber 1989] A. Adolphson and S. Sperber, "Exponential sums and Newton polyhedra: Cohomology and estimates", *Ann. of Math.* (2) **130**:2 (1989), 367–406. MR 91e:11094 Zbl 0723.14017

[Arbarello et al. 1985] E. Arbarello, M. Cornalba, P. A. Griffiths, and J. Harris, *Geometry of algebraic curves, I*, Grundlehren der Mathematischen Wissenschaften **267**, Springer, New York, 1985. MR 86h:14019 Zbl 0559.14017

[Batyrev 1993] V. V. Batyrev, "Variations of the mixed Hodge structure of affine hypersurfaces in algebraic tori", *Duke Math. J.* **69**:2 (1993), 349–409. MR 94m:14067 Zbl 0812.14035

[Batyrev and Cox 1994] V. V. Batyrev and D. A. Cox, "On the Hodge structure of projective hypersurfaces in toric varieties", *Duke Math. J.* **75**:2 (1994), 293–338. MR 95j:14072 Zbl 0851.14021

[Batyrev and Nill 2007] V. Batyrev and B. Nill, "Multiples of lattice polytopes without interior lattice points", *Mosc. Math. J.* **7**:2 (2007), 195–207, 349. MR 2008g:52018 Zbl 1134.52020

[Beelen and Pellikaan 2000] P. Beelen and R. Pellikaan, "The Newton polygon of plane curves with many rational points", *Des. Codes Cryptogr.* **21** (2000), 41–67. MR 2003c:14024 Zbl 1005.14019

[Bruns and Gubeladze 2002] W. Bruns and J. Gubeladze, "Semigroup algebras and discrete geometry", pp. 43–127 in *Geometry of toric varieties*, edited by L. Bonavero and M. Brion, Sémin. Congr. **6**, Soc. Math. France, Paris, 2002. MR 2005e:14078 Zbl 1083.14057

[Castryck et al. 2006] W. Castryck, J. Denef, and F. Vercauteren, "Computing zeta functions of nondegenerate curves", *IMRP Int. Math. Res. Pap.* (2006), Art. ID 72017. MR 2007h:14026 Zbl 05144769

[Clark 2006] P. L. Clark, "There are genus one curves of every index over every number field", *J. Reine Angew. Math.* **594** (2006), 201–206. MR 2007b:11080 Zbl 1097.14024

[Clark 2007] P. L. Clark, "On the indices of curves over local fields", *Manuscripta Math.* **124**:4 (2007), 411–426. MR 2008m:11121 Zbl 05248329

[Coppens 1986] M. Coppens, "The Weierstrass gap sequence of the ordinary ramification points of trigonal coverings of $\mathbf{P}^1$; existence of a kind of Weierstrass gap sequence", *J. Pure Appl. Algebra* **43**:1 (1986), 11–25. MR 87j:14049 Zbl 0616.14012

[Fisher 2006] T. Fisher, "Invariants of a genus one curve", preprint, 2006. arXiv math.NT/0610318

[Fulton 1993] W. Fulton, *Introduction to toric varieties*, Annals of Mathematics Studies **131**, Princeton University Press, 1993. MR 94g:14028 Zbl 0813.14039

[Gel'fand et al. 1994] I. M. Gel'fand, M. M. Kapranov, and A. V. Zelevinsky, *Discriminants, resultants, and multidimensional determinants*, Birkhäuser, Boston, 1994. MR 95e:14045 ZBL 0827.14036

[Haase and Schicho 2009] C. Haase and J. Schicho, "Lattice polygons and the number $2i+7$", *Amer. Math. Monthly* **116**:2 (2009), 151–165. MR MR2478059

[Harris and Morrison 1998] J. Harris and I. Morrison, *Moduli of curves*, Graduate Texts in Mathematics **187**, Springer, New York, 1998. MR 99g:14031 Zbl 0913.14005

[Hartshorne 1977] R. Hartshorne, *Algebraic geometry*, Graduate Texts in Mathematics **52**, Springer, New York, 1977. MR 57 #3116 Zbl 0367.14001

[Hensley 1983] D. Hensley, "Lattice vertex polytopes with interior lattice points", *Pacific J. Math.* **105**:1 (1983), 183–191. MR 84c:52016 Zbl 0471.52006

[Khovanskiĭ 1977] A. G. Hovanskiĭ, "Newton polyhedra, and toroidal varieties", *Funkcional. Anal. i Priložen.* **11**:4 (1977), 56–64, 96. In Russian; translated in *Functional Anal. Appl.* **11**:4 (1977), 289–296. MR 57 #16291 Zbl 0445.14019

[Koelman 1991] R. Koelman, *The number of moduli of families of curves on toric surfaces*, thesis, Katholieke Universiteit te Nijmegen, 1991.

[Kouchnirenko 1976] A. G. Kouchnirenko, "Polyèdres de Newton et nombres de Milnor", *Invent. Math.* **32**:1 (1976), 1–31. MR 54 #7454 Zbl 0328.32007

[Kresch et al. 2002] A. Kresch, J. L. Wetherell, and M. E. Zieve, "Curves of every genus with many points, I: Abelian and toric families", *J. Algebra* **250**:1 (2002), 353–370. MR 2003i:11084 Zbl 1062.14027

[Maroni 1946] A. Maroni, "Le serie lineari speciali sulle curve trigonali", *Ann. Mat. Pura Appl.* (4) **25** (1946), 343–354. MR 9,463j Zbl 0061.35407

[Matsumoto 1998] R. Matsumoto, "The C_{ab} curve", preprint, 1998. Available at http://www.rmatsumoto.org/cab.ps.

[Mikhalkin 2000] G. Mikhalkin, "Real algebraic curves, the moment map and amoebas", *Ann. of Math.* (2) **151**:1 (2000), 309–326. MR 2001c:14083 Zbl 1073.14555

[Mikhalkin 2003] G. Mikhalkin, "Counting curves via lattice paths in polygons", *C. R. Math. Acad. Sci. Paris* **336**:8 (2003), 629–634. MR 2004d:14077 Zbl 1027.14026

[Miura 1992] S. Miura, "Algebraic geometric codes on certain plane curves", *Trans. IEICE* **J75-A**:11 (1992), 1735–1745.

[Mumford 1965] D. Mumford, *Geometric invariant theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete (N.S.) **34**, Springer, Berlin, 1965. MR 35 #5451 Zbl 0147.39304

[Poonen and Rodriguez-Villegas 2000] B. Poonen and F. Rodriguez-Villegas, "Lattice polygons and the number 12", *Amer. Math. Monthly* **107**:3 (2000), 238–250. MR 2001b:52022 Zbl 0988.52024

[Rim and Vitulli 1977] D. S. Rim and M. A. Vitulli, "Weierstrass points and monomial curves", *J. Algebra* **48**:2 (1977), 454–476. MR 56 #15652 Zbl 0412.14002

[Sernesi 1984] E. Sernesi, "On the existence of certain families of curves", *Invent. Math.* **75**:1 (1984), 25–57. MR 85e:14035 Zbl 0541.14024

[Severi 1921] F. Severi, *Vorlesungen über algebraische Geometries*, Teubner, Leipzig, 1921. JFM 48.0687.01

Communicated by Joseph Gubeladze
Received 2008-04-11 Revised 2008-12-22 Accepted 2009-02-17

wouter.castryck@esat.kuleuven.be
Katholieke Universiteit Leuven,
Departement Elektrotechniek (ESAT),
Afdeling SCD – COSIC, Kasteelpark Arenberg 10,
B-3001 Leuven (Heverlee), Belgium

jvoight@gmail.com
Department of Mathematics and Statistics,
University of Vermont, 16 Colchester Ave,
Burlington, VT 05401, United States
http://www.cems.uvm.edu/~voight/

Self-points on elliptic curves

Christian Wuthrich

Let $E/\mathbb{Q}$ be an elliptic curve of conductor N and let p be a prime. We consider trace-compatible towers of modular points in the noncommutative division tower $\mathbb{Q}(E[p^\infty])$. Under weak assumptions, we can prove that all these points are of infinite order and determine the rank of the group they generate. Also, we use Kolyvagin's construction of derivative classes to find explicit elements in certain Tate–Shafarevich groups.

1. Introduction

1.1. ***Definition of self-points.*** Let $E/\mathbb{Q}$ be an elliptic curve and write N for its conductor. As proved in [Breuil et al. 2001], there exists a modular parametrisation

$$\varphi_E : X_0(N) \to E$$

that is a surjective morphism defined over $\mathbb{Q}$ and maps the cusp ∞ on the modular curve $X_0(N)$ to O. The open subvariety $Y_0(N)$ in $X_0(N)$ is a moduli space for the set of pairs (A, C), where A is an elliptic curve and C is a cyclic subgroup in A of order N. More precisely, if $k/\mathbb{Q}$ is a field, then $Y_0(N)(k)$ is in bijection with the set of such pairs (A, C) with A and C defined over k, up to isomorphism over the algebraic closure $\bar{k}$.

In particular, we may consider the pairs $x_C = (E, C)$ for any given cyclic subgroup C of order N in E as a point in $Y_0(N)(\mathbb{C})$. Its image $P_C = \varphi_E(x_C)$ under the modular parametrisation is called a *self-point* of E. The field of definition of the point P_C on E is the same as the field of definition $\mathbb{Q}(C)$ of C. The compositum of all $\mathbb{Q}(C)$ will be denoted by K_N; it is the smallest field K such that the Galois group $\mathrm{Gal}(\overline{K}/K)$ acts by scalars on $E[N]$.

More generally, for any integer m we define a number field K_m as follows. There is a Galois representation attached to the m-torsion points on E, given by

$$\bar{\rho}_m : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \longrightarrow \mathrm{Aut}(E[m]) \cong \mathrm{GL}_2(\mathbb{Z}/m\mathbb{Z}) \longrightarrow \mathrm{PGL}_2(\mathbb{Z}/m\mathbb{Z}).$$

MSC2000: primary 11G05; secondary 11G18, 11G40.

Keywords: elliptic curves, modular point, modular curves.

The author was supported by a fellowship of the Swiss National Science Foundation.

The field K_m is the field fixed by the kernel of $\bar{\rho}_m$. The Galois group of the extension $K_m/\mathbb{Q}$ can be viewed via $\bar{\rho}_m$ as a subgroup of $\mathrm{PGL}_2(\mathbb{Z}/m\mathbb{Z})$.

We will call *higher self-point* the image under φ_E of any pair (A, C) in which A is an elliptic curve that is isogenous to E over $\overline{\mathbb{Q}}$, though the most interesting case of higher self-points is the case when the isogeny between E and A is of degree a prime power p^n. In particular, this prime p is allowed to divide the conductor N.

This construction imitates the definition of Heegner points, where one uses pairs (A, C) with A having complex multiplication. More generally, modular points on elliptic curves were considered earlier by Harris [1979] without any restriction on A. This article is a sequel to the articles [Delaunay and Wuthrich 2008] and [Wuthrich 2007] on self-points, where we have emphasised already that the theory of self-points differs from the well-known theory of Heegner points. For instance, there does not seem to be a link between the root numbers and the question of whether the self-points are of infinite order.

We present here not only a generalisation of the previous results on self-points, but also we introduce the construction of derivative classes à la Kolyvagin. Indeed, Kolyvagin [1990] was able to find upper bounds on certain Selmer groups by constructing cohomology classes starting from Heegner points. We propose here to do the analogue for self-points. But the situation is radically different as the Galois groups involved are noncommutative; rather than finding upper bounds of Selmer groups over the base field, we will find *lower* bounds on Selmer groups over certain number fields.

1.2. ***The results for self-points.*** The main question that arises first is whether we can determine if the self-points are of infinite order in the Mordell–Weil group $E(\mathbb{Q}(C))$. It was shown in [Delaunay and Wuthrich 2008] that the self-points are always of infinite order if the conductor is a prime number. We extend here the method and provide a framework to treat the general case. In Section 5.2 we will prove the following.

Theorem 12. *Let $E/\mathbb{Q}$ be a semistable elliptic curve of conductor $N \neq 30$ or 210. Then all the self-points are of infinite order.*

But the methods are more general and we are able to prove that they are of infinite order in most cases. In fact, we conjecture that this holds whenever E does not admit complex multiplication. In Section 6.2 we will give a self-point of finite order on a curve with complex multiplication. In the largest generality, we are able to prove in Theorem 2 that there is at least one self-point of infinite order under the assumption that $j(E) \notin \frac{1}{2}\mathbb{Z}$.

Next we address the question of the rank of the group generated by self-points in $E(K_N)$. If N is prime, we saw that the only relation among the self-points is that the sum of all of them is a torsion point in $E(\mathbb{Q})$. For a general conductor, we find that for all proper divisors d of N and all cyclic subgroups B in E of order d, the sum of all self-points P_C with $C \supset B$ is torsion. This is proved in Proposition 4 as a consequence of the existence of the degeneracy maps on modular curves. For a lot of semistable curves, the following result shows that these are the only relations among self-points.

Theorem 14. *Let $E/\mathbb{Q}$ be a semistable elliptic curve. Suppose that N is not equal to* 30 *or* 210*. Suppose that for each prime $p \mid N$ such that $\bar{\rho}_p$ is not surjective, there is a prime $\ell \mid N$ such that the Tamagawa number c_ℓ is not divisible by p. Then the group generated by the self-points is of rank N.*

We think that this may hold more generally.

Conjecture. *Let $E/\mathbb{Q}$ be an elliptic curve without complex multiplication. Then all the self-points are of infinite order and the only relations among them are produced by the degeneracy maps as described in Proposition 4. In particular, the rank of the group generated by self-points is equal to*

$$\delta(N) = \prod_{p \mid N} \lceil (1 - p^{-2}) \cdot p^{\operatorname{ord}_p(N)} \rceil,$$

where $\lceil x \rceil$ denotes the smallest integer no less than x.

The expression $\delta(N)$ in the conjecture is equal to N if and only if N is square-free.

1.3. ***The results for higher self-points.*** We are particularly interested in higher self-points that are modular points coming from a pair (E', C') in which E' has an isogeny to E of degree a power of a prime p. We treat two cases: when p is a prime of good reduction and when p is a prime of multiplicative reduction.

For simplicity we only sketch the results for the good case here, that is, $p \nmid N$. See Section 7 for more details.

We fix now a cyclic subgroup C in E of order N; the following construction depends on this choice, but our notation will not reflect this. Let D be a cyclic subgroup of E of order p^{n+1} and let $E' = E/D$. Given any self-point P_C, we may consider the image C' of C under the isogeny $E \to E'$. The higher self-point Q_D is defined to be the image of $(E', C') \in Y_0(N)$ under the modular parametrisation φ_E. It is a point in the Mordell–Weil group of E over the field $\mathbb{Q}(C, D)$, which is contained in $K_{p^{n+1}N}$. In Corollary 20, we are able to prove that the higher self-points are all of infinite order in some cases.

Theorem 1. *Let $E/\mathbb{Q}$ be a semistable curve of conductor N not equal to* 30 *or* 210*. Suppose that p is a prime such that $p > N$, and such that the Galois representation $\bar{\rho}_p : \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{PGL}_2(\mathbb{F}_p)$ is surjective. Let s be the rank of the group generated by the self-points in $E(K_N)$. Then the higher self-points in $E(K_{p^{n+1}N})$ generate a group of rank $s \cdot (p+1) \cdot p^n$.*

If one assumes that the prime is of ordinary reduction for E, one can weaken the condition on the bad reduction substantially.

Furthermore these higher self-points are trace-compatible in the following sense. Let D be a cyclic subgroup of order p^{n+1}, and let a_p be the p-th Fourier coefficient of the modular form associated to the isogeny class of E. Then we have

$$\sum_{D'\supset D} Q_{D'} = a_p \cdot Q_D,$$

where the sum runs over all cyclic subgroups D' of order p^{n+2} containing D. For any number field F, we will write

$$\rho_{F,p} : \mathrm{Gal}(\bar{F}/F) \longrightarrow \mathrm{Aut}(T_pE) \cong \mathrm{GL}_2(\mathbb{Z}_p) \longrightarrow \mathrm{PGL}_2(\mathbb{Z}_p)$$

for the representation of $\mathrm{Gal}(\bar{F}/F)$ on the Tate module T_pE. If the Galois representation $\rho_{K_N,p}$ is surjective, then we can reformulate the above relation by saying that the trace of $Q_{D'}$ from its field of definition to the field of definition of Q_D is equal to $a_p \cdot Q_D$. This trace compatibility reminds one of the definition of an Euler system. However, the field $\mathbb{Q}(C, D)$ is not Galois over $\mathbb{Q}$ and the Galois closure is not an abelian extension, and worse not even a solvable extension.

The higher self-points are the only known towers of points of infinite order in the division tower $\mathbb{Q}(E[p^\infty])$ of E. Nevertheless the growth of the rank of the Mordell–Weil group should often be faster than the lower bound $(p+1)p^n$ that we establish here in many cases. This is due to changing signs in the functional equations and the corresponding parity results on the corank of Selmer groups. See [Coates et al. 2009; Mazur and Rubin 2008]. These results predict, under the assumption of the finiteness of the Tate–Shafarevich group, that there should be more points of infinite order in the division tower that are not accounted for by higher self-points. Furthermore the higher self-points do not seem to be linked in any obvious way to root numbers. Also it is completely unknown if there is a relation to L-functions (or to noncommutative p-adic L-functions as in [Coates et al. 2005]) in analogy to the Gross–Zagier formula for Heegner points.

1.4. ***Derivatives.*** Kolyvagin [1990] has used Heegner points of infinite order to construct cohomology classes that obstruct the existence of further points of infinite order. We aim to use a similar construction to build cohomology classes from higher self-points of infinite order.

Let p be a prime of either good ordinary reduction or of multiplicative reduction. If p does not divide the conductor N, define $F_n = K_{p^{n+1}N}$; otherwise let $F_n = K_{p^nN}$. Put $F = F_{-1}$. If we suppose that $\rho_{F,p} : \mathrm{Gal}(\bar{F}/F) \to \mathrm{PGL}_2(\mathbb{Z}_p)$ is surjective, then $\mathrm{Gal}(F_n/F) = \mathrm{PGL}_2(\mathbb{Z}/p^{n+1}\mathbb{Z})$. We are interested in a particular cyclic subgroup A in $\mathrm{Gal}(F_n/F)$. Choosing a $\mathbb{Z}_p$-basis of the quadratic unramified extension $\mathbb{O}$ of $\mathbb{Z}_p$

gives a map

$$\mathbb{O}^\times \to \mathrm{GL}_2(\mathbb{Z}_p) \to \mathrm{PGL}_2(\mathbb{Z}_p) \to \mathrm{PGL}_2(\mathbb{Z}/p^{n+1}\mathbb{Z}),$$

whose image is a cyclic group A_n of order $(p+1)\cdot p^n$. By a slight abuse of notation we will denote the subfield of F_n fixed by A_n by F_n^A.

The construction of derivatives provides us with a map

$$\partial_n : \mathrm{H}^1(A_n, S) \to Ш(E/F_n^A).$$

The source is a cohomology group of the saturated group S of higher self-points (see Section 8 for the definitions). Although we do not know its exact structure, we can prove that it contains at least p^n elements. It seems plausible to think that the map ∂_n is very often injective, but we do have no means to prove this in a single case. Nevertheless, we are able to show the existence of points of infinite order in $E(F_n^A)$ whenever the map is not injective. Here is the final result:

Theorem 21. *Let $E/\mathbb{Q}$ be an elliptic curve. Suppose E does not have potentially good supersingular reduction for any prime of additive reduction. Let p be a prime of either good ordinary or multiplicative reduction. Assume that $\rho_{F,p}$ is surjective and that K_N contains a self-point of infinite order. Then we have*

$$\#\,\mathrm{Sel}_{p^n}(E/F_n^A) \geqslant p^n.$$

The construction of derivatives relies on a property of modular representation theory. The higher self-points generate in the Mordell–Weil group a copy of the irreducible Steinberg representation. More precisely, if H_n denotes $\mathrm{Gal}(F_n/F)$, there is a certain $\mathbb{Q}[H_n]$-module in $E(F_n)\otimes\mathbb{Q}$ that is irreducible, but this module is no longer irreducible over $\mathbb{F}_\ell[H_n]$ when ℓ divides $(p+1)\cdot p^n$. Perhaps the idea of using modular representation theory to study Selmer groups, which was developed in [Greenberg 2008], could shed new light on these derivatives.

2. The fundamental theorem

Theorem 2. *Let $E/\mathbb{Q}$ be an elliptic curve of conductor N. If the j-invariant of E is not in $\frac{1}{2}\mathbb{Z}$, then there is at least one self-point P_C of infinite order in $E(K_N)$.*

Proof. Let p be a prime that divides the denominator of the j-invariant of E. If possible, we avoid $p=2$. Note that p^2 may divide N, but we know that E acquires multiplicative reduction over some extension of $\mathbb{Q}$ at p.

First we fix an embedding of $\overline{\mathbb{Q}}$ into $\overline{\mathbb{Q}}_p$. We consider the modular parametrisation over $\overline{\mathbb{Z}}_p$. The modular curve $X_0(N)$ over $\overline{\mathbb{Z}}_p$ has a neighbourhood of the cusp ∞ consisting of pairs (A, C) of a Tate curve of the form $A = \overline{\mathbb{Q}}_p^\times/q^{\mathbb{Z}}$, together with a cyclic subgroup C of order N generated by the N-th root of unity. The parameter q is a p-adic analytic uniformiser at ∞, so that the $\mathrm{Spf}\,\overline{\mathbb{Z}}_p[\![q]\!]$ is the formal completion of $X_0(N)/\overline{\mathbb{Z}}_p$ at the cusp ∞; see [Katz and Mazur 1985, Chapter 8].

Let $f_E = \sum a_n q^n$ be the normalised newform associated to E. Then $f_E/q \cdot dq$ is the associated differential. Let c_E be the Manin constant (of the not necessarily strong Weil curve E), that is, the number such that $\varphi_E^*(\omega_E) = c_E \cdot f_E/q \cdot dq$, where ω_E is the invariant differential on E. The rigid analytic map induced by φ_E on the completion can now be characterised as

$$\log_E(\varphi_E(q)) = \int_O^{\varphi_E(q)} \omega_E = c_E \cdot \int_0^q f_E \frac{dq}{q} = c_E \cdot \sum_{n\geqslant 1} \frac{a_n}{n} \cdot q^n. \tag{1}$$

Here $\log_E$ denotes the formal logarithm associated to E from the formal group $\hat{E}(\overline{\mathfrak{m}})$ to the maximal ideal $\hat{\mathbb{G}}_a(\overline{\mathfrak{m}}) = \overline{\mathfrak{m}}$ of $\overline{\mathbb{Z}}_p$. We deduce from this description the following lemma that will be useful later. Write $|\cdot|_p$ for the normalised absolute value such that $|p|_p = p^{-1}$.

Lemma 3. *Let (A, C) be a point in $Y_0(N)(\overline{\mathbb{Q}}_p)$ such that A is isomorphic to the Tate curve with parameter $q_0 \neq 0$ and C is isomorphic to the Galois module of N-th roots of unity $\mu[N]$. If $|q_0|_p < p^{-1/(p-1)}$, then $\varphi_E(A, C)$ is a point of infinite order on $E(\overline{\mathbb{Q}}_p)$.*

Proof. Under the condition on the absolute value of q_0, we know that the sum on the right-hand side of (1) converges. We consider the sum

$$z = c_E \cdot \sum_{n\geqslant 1} \frac{a_n}{n} \cdot q_0^n.$$

Since the Manin constant is known to be an integer (see [Edixhoven 1991]), the absolute value of the right-hand side is

$$|z|_p = |c_E|_p \cdot \left| q_0 + \frac{a_p}{p} q_0^p \right|_p$$

as these are the terms of large absolute value. However note that the condition on q_0 implies that the second term on the right side is actually slightly smaller that the first, and hence the absolute value of the sum is bounded by

$$|z|_p = |c_E|_p \cdot |q_0|_p < p^{-1/(p-1)}.$$

Therefore the value of z lies in the domain of convergence of the p-adic elliptic exponential $\exp_E$, and we obtain that $\varphi_E(A, C) = \exp_E(z)$. Since we know that $|z|_p \neq 0$, we can deduce that $\exp_E(z)$ is not a torsion point in $E(\overline{\mathbb{Q}}_p)$. □

We may now finish proving the theorem. Since E has multiplicative reduction over $\overline{\mathbb{Z}}_p$, exactly one of the $x_C = (E, C)$ if in the neighbourhood of ∞ on $X_0(N)$;

it is represented by the p-adic Tate parameter q_E associated to E together with the group C isomorphic to $\mu[N]$. If $p \neq 2$, then we know that

$$|q_E|_p = |j(E)|_p^{-1} \leqslant p^{-1} < p^{-1/(p-1)},$$

and if p had to be chosen to be equal to 2 in the beginning then we know that

$$|q_E|_2 = |j(E)|_2^{-1} \leqslant p^{-2} < p^{-1/(p-1)}.$$

Hence in any case, the lemma applies and provides us with a point of infinite order among the self-points. □

If the chosen prime p is such that p^2 does not divide N, then q_E lies in $p^{\upsilon}\mathbb{Z}_p$, where $\upsilon = -\operatorname{ord}_p(j(E))$. Hence the proof's point P_C will be defined over $\mathbb{Q}_p$.

The restriction at $p = 2$ seems unnecessary. Often one can deduce the result of the theorem by hand for curves whose j-invariant is an odd integer divided by 2. We present here an easy example. For the curve 2450o1 in Cremona's tables [1997] with j-invariant $-189/2$, the 2-adic Tate parameter is equal to $2+2^2+2^4+\mathbf{O}(2^9)$ and the newform is $f_E = q - q^2 + q^4 + \mathbf{O}(q^8)$. From this one concludes that $\log_E(P_C) = 2^3 + \mathbf{O}(2^5)$. So P_C is of infinite order. Nevertheless we do not see any easy argument to prove that $P_C \neq O$ for a general curve with $j(E) \in \frac{1}{2}\mathbb{Z} \setminus \mathbb{Z}$, as it seems that the 2-adic valuation of $\log_E(P_C)$ can be arbitrarily large.

2.1. ***A torsion self-point.*** We believe that this theorem is still valid if E is a curve with integral j-invariant as long as the curve does not admit complex multiplication. But not all self-points are of infinite order. We present here a surprisingly easy example of a self-point that is torsion.

The curve 27a2 admits a cyclic isogeny of degree 27 defined over $\mathbb{Q}$ to the curve 27a4. Let E be either of the two curves. Then E has exactly one cyclic subgroup of order 27 defined over $\mathbb{Q}$, that is, E admits a self-point in $E(\mathbb{Q})$. Since the rank of $E(\mathbb{Q})$ is zero, the self-point has to be of finite order. Note that these curves have complex multiplication. See Section 6.2 for more detailed computations on these self-points.

3. Relations

In [Delaunay and Wuthrich 2008] it is shown that the self-points on a curve of prime conductor satisfy exactly one relation. What kind of relations could occur among the self-points for a curve of conductor N? Here is a first part of an answer. First, we need some more notation. The Galois group $G = G_N = \operatorname{Gal}(K_N/\mathbb{Q})$ was identified with a subgroup of $\mathrm{PGL}_2(\mathbb{Z}/N\mathbb{Z})$. For any divisor d of N, we define the image of G_N under the projection $\mathrm{PGL}_2(\mathbb{Z}/N\mathbb{Z}) \to \mathrm{PGL}_2(\mathbb{Z}/d\mathbb{Z})$ as G_d and by K_d its fixed field in K_N. In other words, K_d is the smallest number field for which the absolute Galois groups acts by scalars on $E[d]$.

Proposition 4. *The sum of all self-points is a torsion point defined over $\mathbb{Q}$. If $d \neq N$ is an integer dividing N, then there are relations of the form*

$$R_B : \quad \sum_{C \supset B} P_C \text{ is torsion in } E(K_d),$$

where B is any given cyclic subgroup of order d and C runs through all cyclic groups of order N containing B.

Proof. The degeneracy map $\pi : X_0(N) \to X_0(d)$ induces $\pi^* : J_0(d) \to J_0(N)$ on Jacobians. Given a cyclic subgroup B of order d on E, we may consider the point $x_B = (E, B)$ on $X_0(d)$. The divisor class

$$\pi^*[(x_B) - (\infty)] = \sum_{C \supset B} [(x_C)] - \pi^*[(\infty)]$$

is in the image of π^* in $J_0(N)$ and hence in the kernel of the map $\varphi_E : J_0(N) \to E$ because N is the exact conductor of E. This gives the relation R_B.

Taking $d = 1$ gives the result that the sum of all self-points is a torsion point. Since this sum is fixed by the Galois group, it has to be a rational point. □

4. The Steinberg representations

The aim is to describe certain irreducible representations that will appear in the study of self-points. Let $N > 1$ be an integer. We are interested in the group $P = \mathrm{PGL}_2(\mathbb{Z}/N\mathbb{Z})$. We will decompose the $\mathbb{Q}[P]$-module V whose basis $\{e_C\}$ as a $\mathbb{Q}$-vector space is in bijection with the projective line $\mathbb{P}^1(\mathbb{Z}/N\mathbb{Z})$ and on which the P-action is given by the usual permutation on the basis. So it can be written as

$$V = \bigoplus_{C \in \mathbb{P}^1(\mathbb{Z}/N\mathbb{Z})} \mathbb{Q} e_C = \mathrm{Ind}_B^P(\mathbb{1}_B),$$

where B is a Borel subgroup of P and $\mathbb{1}_B$ is its trivial representation.

Theorem 5. *The $\mathbb{Q}[\mathrm{PGL}_2(\mathbb{Z}/N\mathbb{Z})]$-module V splits into the sum $V = \bigoplus_{D \mid N} W_D$ of irreducible $\mathbb{Q}[\mathrm{PGL}_2(\mathbb{Z}/N\mathbb{Z})]$-modules W_D, where D runs through all divisors of N. Let $D = \prod_p p^{d_p}$ be the prime decomposition of a divisor D of N. Define*

$$\delta_p = \lceil p^{d_p} - p^{d_p - 2} \rceil = \begin{cases} 1 & \textit{if } d_p = 0, \\ p & \textit{if } d_p = 1, \\ p^{d_p} - p^{d_p-2} & \textit{if } d_p > 1. \end{cases}$$

Then W_D has dimension $\delta(D) = \prod_{p \mid D} \delta_p$ as a $\mathbb{Q}$-vector space.

Proof. We split the proof into three parts according to whether N is a prime, a prime power or any integer. The first two cases could also be treated by invoking [Silberger 1970, Theorem 3.3 on page 58], but, since we need the explicit description of W_D later on, we prefer to prove this theorem in detail. Since the proof is inductive on N, we will now write P_N for $\mathrm{PGL}_2(\mathbb{Z}/N\mathbb{Z})$ and V_N for its V.

Case: N is prime. Write $p = N$. The claim is simply that the $\mathbb{Q}[P]$-module V_p splits into two irreducible components $W_1 \oplus W_p$. We define W_1 to be the 1-dimensional subspace of V generated by the vector $v_1 = \sum_C e_C$, where the sum runs over all C in $\mathbb{P}^1(\mathbb{F}_p)$. Of course, $W_1 = V_p^P$ is an irreducible $\mathbb{Q}[P]$-submodule of V_p and the space

$$W_p = \left\{\sum a_C \cdot e_C \;\middle|\; \sum a_C = 0\right\}$$

is a complement to it. It remains to show that W_p is irreducible. Let g be an element of order p in P, such as the class of $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$. On $V_p \otimes \mathbb{C}$ the element g acts with eigenvalues $1, 1, \zeta, \zeta^2, \ldots, \zeta^{p-1}$, where ζ is a primitive p-th root of unity. Hence on W_p every p-th root of unity appears exactly once as an eigenvalue. So the only possibility for W_p to split up into two $\mathbb{Q}[P]$-submodules would have to involve a 1-dimensional and a $(p-1)$-dimensional submodule.

As we can see from the fact that $\mathrm{PSL}_2(\mathbb{F}_p)$ is a simple group when $p > 3$ and by direct calculations for $p = 2$ and 3, there are only two one-dimensional representations of $\mathrm{PGL}_2(\mathbb{F}_p)$: the trivial representation and the one with kernel $\mathrm{PSL}_2(\mathbb{F}_p)$ of index 2. Since $\mathrm{PSL}_2(\mathbb{F}_p)$ acts transitively on $\mathbb{P}^1(\mathbb{F}_p)$, the one-dimensional subrepresentations of V_p must be contained in $V_p^{\mathrm{PSL}_2(\mathbb{F}_p)} = W_1$.

Case: N is a prime power. We write $N = p^k$ with p prime. We prove the statement by induction on k. The case $k = 1$ has been treated already; thus we may assume that $k \geqslant 2$. The claim is that V_{p^k} splits as $\bigoplus W_{p^m}$, where m runs from 0 to k.

There is a reduction map $\alpha : \mathbb{P}^1(\mathbb{Z}/p^k\mathbb{Z}) \to \mathbb{P}^1(\mathbb{Z}/p^{k-1}\mathbb{Z})$ that is surjective and all of whose fibres contain p elements. Define

$$V' = \left\{\sum a_C e_C \;\middle|\; a_C = a_{C'} \text{ whenever } \alpha(C) = \alpha(C')\right\}.$$

It is easy to see that V' is canonically isomorphic to $V_{p^{k-1}}$ as a vector space, so we will identify them. The action of P_{p^k} factors through the quotient $P_{p^k} \to P_{p^{k-1}}$ induced by reduction. By induction, V' splits as a $\mathbb{Q}[P_{p^{k-1}}]$-module into the sum $V' = \bigoplus_{m=0}^{k-1} W_{p^m}$; this also decomposes V' into irreducible $\mathbb{Q}[P_{p^k}]$-modules. As a complement to V', we define

$$W_{p^k} = \left\{\sum a_C e_C \;\middle|\; \sum_{\alpha(C)=D} a_C = 0 \text{ for all } D \text{ in } \mathbb{P}^1(\mathbb{Z}/p^{k-1}\mathbb{Z})\right\}.$$

It is clear that W_{p^k} is a $\mathbb{Q}[P_{p^k}]$-submodule of V_{p^k}. If $k > 1$ then its dimension is equal to

$$\begin{aligned}\dim_{\mathbb{Q}} W_{p^k} &= \#\mathbb{P}^1(\mathbb{Z}/p^k\mathbb{Z}) - \#\mathbb{P}^1(\mathbb{Z}/p^{k-1}\mathbb{Z})\\ &= (p+1)\cdot p^{k-1} - (p+1)\cdot p^{k-2} = p^k - p^{k-2}.\end{aligned}$$

It remains to show that W_{p^k} is irreducible.

Let ∞ be any point in $\mathbb{P}^1(\mathbb{F}_p)$ and write U^∞ for the preimage of ∞ under the reduction map $\mathbb{P}^1(\mathbb{Z}/p^k\mathbb{Z}) \to \mathbb{P}^1(\mathbb{F}_p)$. Within V, we define a linear subspace

$$V^\infty = \left\{\sum a_C e_C \;\middle|\; a_C = 0 \text{ if } C \in U^\infty\right\}$$

of dimension p^k and let $W^\infty = W_{p^k} \cap V^\infty$ and $V'^\infty = V' \cap V^\infty$. Let g be an element of P_{p^k} of order p^k whose fixed points lie in U^∞. If ∞ is $(0:1)$, then we may take the class of the matrix $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$. The element g acts on $V^\infty \otimes \mathbb{C}$ such that every p^k-th root of unity appears exactly once. The eigenvalues of g on the subspace V'^∞ are all p^{k-1}-st roots of unity. Hence on W^∞ every primitive p^k-th root of unity appears exactly once as an eigenvalue. So W^∞ is an irreducible $\mathbb{Q}[\langle g\rangle]$-module, and so if W_{p^k} splits as a $\mathbb{Q}[P_{p^k}]$-module, then W^∞ has to be completely contained in one of the summands. But for any two distinct points ∞ and ∞' in $\mathbb{P}^1(\mathbb{F}_p)$ the spaces W^∞ and $W^{\infty'}$ span the whole of W_{p^k}. Hence W_{p^k} cannot be reducible.

The general case follows fairly easily from the previous cases. Let $N = \prod p^{n_p}$ be the prime decomposition of N. We may suppose that N is not a prime power, since we have treated this case already. Now the group P_N splits as

$$P_N = \mathrm{PGL}_2(\mathbb{Z}/N\mathbb{Z}) = \prod_{p\mid N} \mathrm{PGL}_2(\mathbb{Z}/p^{n_p}\mathbb{Z}) = \prod_{p\mid N} P_{p^{n_p}}$$

by the Chinese remainder theorem. Similarly, we have

$$\mathbb{P}^1(\mathbb{Z}/N\mathbb{Z}) = \prod_{p\mid N} \mathbb{P}^1(\mathbb{Z}/p^{n_p}\mathbb{Z}) \quad \text{and so} \quad V_N = \bigotimes_{p\mid N} V_{p^{n_p}}$$

as a $\mathbb{Q}[P_N]$-module. Now we use the previous case to rewrite

$$V_N = \bigotimes_{p\mid N} \bigoplus_{m=0}^{n_p} W_{p^m}.$$

Let D be any divisor of N and $\prod p^{d_p}$ its prime factorisation. Then define

$$W_D = \bigotimes_{p\mid D} W_{p^{d_p}}.$$

It is clear from the representation theory of direct products that W_D is irreducible. Rearranging the above decomposition of V_N, we arrive at the desired expression $V_N = \bigoplus_{D|N} W_D$. □

Proposition 6. *Let p be a prime. Let G be a subgroup of a Borel subgroup of $\mathrm{PGL}_2(\mathbb{F}_p)$ acting on $V_p = \bigoplus \mathbb{Q}e_C$. Suppose that the class of $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$ belongs to G. Then V_p decomposes into irreducible $\mathbb{Q}[G]$-modules as $W_1 \oplus W_1' \oplus W_p'$, where W_p' is an irreducible $\mathbb{Q}[G]$-module of dimension $p-1$.*

Proof. Let C_0 be the element of $\mathbb{P}^1(\mathbb{F}_p)$ fixed by the Borel group containing G. By assumption, we know that C_0 is the only fixed point of G acting on $\mathbb{P}^1(\mathbb{F}_p)$. Hence V_p contains two linearly independent vectors that are fixed by G, namely e_{C_0} and $v_0 = \sum_{C \neq C_0} e_C$. The $\mathbb{Q}[G]$-submodule

$$W_p' = \Big\{ \sum_{C \neq C_0} a_C \cdot e_C \;\Big|\; \sum_{C \neq C_0} a_C = 0 \Big\}$$

is a complement to V_p^G. Now use the class g of the matrix $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$ as before to show that W_p' is irreducible since the eigenvalues of g on W_p' are exactly the set of all primitive p-th roots of unity. □

In fact one can show that Theorem 5 holds even for the complex representation $V \otimes \mathbb{C}$ as $\mathbb{C}[\mathrm{PGL}_2(\mathbb{Z}/N\mathbb{Z})]$-modules. On the other hand, Proposition 6 really relies on the fact that we are only considering decompositions as $\mathbb{Q}[G]$-modules. For instance, we may well take G to be the cyclic group generated by the matrix $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$; then of course $W_p' \otimes \mathbb{C}$ will split into 1-dimensional representations. But since the p-th roots of unity are not all defined over $\mathbb{Q}$, at least if $p > 2$, this decomposition does not hold in general for W_p'.

We can now reformulate the statement of Proposition 4 as follows. There is a G-equivariant map $\iota : V_N \to E(K_N) \otimes \mathbb{Q}$, defined by sending e_C to P_C. It has a kernel containing all submodules W_d for $d \neq N$ dividing N. So it induces a map $\iota : W_N \to E(K_N) \otimes \mathbb{Q}$ that is G-equivariant. By the fundamental Theorem 2, this morphism is nontrivial if $j \notin \frac{1}{2}\mathbb{Z}$. Hence we can deduce the following corollary.

Corollary 7. *The self-points generate a group of rank at most $\delta(N)$ inside $E(K_N)$. If W_N is an irreducible $\mathbb{Q}[G_N]$-module and the j-invariant is not in $\frac{1}{2}\mathbb{Z}$, then the self-points generate a group of rank $\delta(N)$ and the Galois group acts like the Steinberg representation W_N on it.*

5. Self-points on semistable curves

We will suppose in this section that the curve $E/\mathbb{Q}$ is semistable. In particular, the j-invariant cannot belong to $\frac{1}{2}\mathbb{Z}$ since all primes dividing N must appear in the denominator of $j(E)$ and there is no curve of conductor 2. Hence the fundamental Theorem 2 applies to E.

5.1. ***Some lemmata.*** Recall that K_m was defined to be the field fixed by the kernel of $\bar{\rho}_m$. We denote the Galois group $\mathrm{Gal}(K_m/\mathbb{Q})$ by G_m and think of it as a subgroup in $\mathrm{PGL}_2(\mathbb{Z}/m\mathbb{Z})$.

In what follows, we often have to split up the primes dividing N into two groups. Let s, standing for "surjective", be the product of all primes p dividing N such that the representation $\bar{\rho}_p$ is surjective. Let m, standing for "méchant", be the product of the remaining primes dividing N. Note that there are not many choices for m as described in the following lemma.

Lemma 8. *We have $m \in \{1, 2, 3, 4, 5, 6, 7, 10\}$. If $p \mid m$, then G_p is contained in a Borel group of $\mathrm{PGL}_2(\mathbb{F}_p)$ and hence is either a cyclic or a metacyclic[1] group.*

Proof. Let $p \mid m$. By a theorem of Serre [1996], the curve admits a p-isogeny $E \to E'$ defined over $\mathbb{Q}$, and either E or E' must have a point of order p defined over $\mathbb{Q}$. Then by Mazur's theorem [1978] on torsion points on elliptic curves over $\mathbb{Q}$, and we know now that $p \leqslant 7$ and that $m \leqslant 10$. □

Lemma 9. *Let $E/\mathbb{Q}$ be a semistable elliptic curve. Then the largest prime p dividing N is such that the representation $\bar{\rho}_p$ is surjective, and $p - 1 > m$ unless N is 30 or 210.*

Proof. If N is divisible by a prime $p \geqslant 13$, then the largest prime p dividing N cannot divide m and satisfies $p - 1 > m$ because $m \leqslant 10$ by the previous lemma. Hence we are left with a finite list of possible N to check. This can be done easily; to illustrate it we show in Table 1 the list of curves of square-free conductors N whose prime divisors are among $\{2, 3, 5, 7\}$. For the full proof, we would need to list also conductors divisible by 11, but then the list will be far too long to be included here. However the only three exceptional isogeny classes can already be seen in this table.

To each isogeny class, we give the number i of isogenous curves, the maximal degree d of an isogeny among them, the value of m, and the largest $p \mid N$ such that $\bar{\rho}_p$ is surjective. This ends the proof. □

Lemma 10. *Let $E/\mathbb{Q}$ be a semistable elliptic curve with $6 \mid N$ and such that the representation $\bar{\rho}_2$ is surjective onto $\mathrm{PGL}_2(\mathbb{F}_2)$. If there exists a prime $p \mid N$ such that $3 \nmid c_p$, then K_2 cannot be contained in K_3.*

Proof. We wish to derive a contradiction from the assumption that K_2 is contained in K_3. By assumption, the Galois group G_2 of the extension $K_2/\mathbb{Q}$ is $\mathrm{PGL}_2(\mathbb{F}_2)$, which is isomorphic to the symmetric group on three letters $\mathfrak{S}_3$. The Galois group

[1] metacyclic: a semidirect product of cyclic groups

N	14a	15a	21a	**30a**	35a	42a	70a	105a	**210a**	**210b**	210c	210d	210e
i	6	8	6	**8**	3	6	4	4	**8**	**8**	6	4	8
d	18	16	8	**12**	9	8	4	4	**12**	**12**	8	4	16
m	2	1	1	**6**	1	2	2	1	**6**	**6**	2	2	2
p	7	5	7	**5**	7	7	7	7	**7**	**7**	7	7	7

Table 1. Some of the “evil curves” to be treated separately in Lemma 9.

G_3 is contained in $\mathrm{PGL}_3(\mathbb{F}_3) = \mathfrak{S}_4$. Therefore the Galois group $\mathrm{Gal}(K_3/K_2)$ is contained in the Klein group V_4 of $\mathfrak{S}_4$.

Suppose first that the reduction of E at p is split multiplicative. Let q_E be the Tate parameter of E over $\mathbb{Q}_p$. Choose a place v above p in K_2 and a place w above v in K_3. Then the completion $K_{3,w}$ is equal to $\mathbb{Q}_p(\zeta_3, \sqrt[3]{q_E})$ and $K_{2,v}$ is equal to $\mathbb{Q}_p(\sqrt{q_E})$. Since 3 does not divide $c_p \geqslant 1$, we know that q_E cannot be a cube. Therefore the degree of $K_{3,w}/K_{2,v}$ is divisible by 3. This is impossible since the degree of K_3/K_2 must be a power of 2.

If the reduction is nonsplit multiplicative at p, one can do the same argument but transposed to the extension L of $\mathbb{Q}_p$ over which E acquires split multiplicative reduction. As $L/\mathbb{Q}_p$ is of degree 2, we still find that the degree of $K_{3,w}/K_{2,v}$ must be a multiple of 3. □

Lemma 11. *Let $E/\mathbb{Q}$ be a semistable elliptic curve. For* (ii) *and* (iii) *below, we assume that if* $2 \mid N$ *and* $3 \mid N$, *then there is a prime $p \mid N$ such that* $3 \nmid c_p$.

(i) *G_s acts transitively on the set $\mathbb{P}^1(\mathbb{Z}/s\mathbb{Z})$ of cyclic subgroups of order s in E.*

(ii) *The Steinberg representation W_s is irreducible as a $\mathbb{Q}[G_s]$-module.*

(iii) *Suppose W_m decomposes into irreducible $\mathbb{Q}[G_m]$-modules as $U_1 \oplus \cdots \oplus U_k$. Then W_N decomposes into irreducible $\mathbb{Q}[G_N]$-modules as*

$$W_N = \bigoplus_{i=1}^{k} (U_i \otimes W_s).$$

Proof. We will first prove by induction the statement in (ii) with s replaced by any of its divisors r, assuming the additional hypothesis. If $r = p$ is prime then $G_p = \mathrm{PGL}_2(\mathbb{F}_p)$ and Theorem 5 shows that W_p is irreducible as a $\mathbb{Q}[G_p]$-module. Let p be the largest prime factor of r. We may suppose that r is composite and so $p > 2$. Put $t = r/p \geqslant 2$. We assume that W_t is an irreducible $\mathbb{Q}[G_t]$-module. We wish to prove that W_r is an irreducible $\mathbb{Q}[G_r]$-module.

$$\begin{tikzcd}
& K_r & \\
K_t \arrow[ur, dash, "H_p"] \arrow[dr, dash] \arrow[ddr, dash, "G_t"'] & & K_p \arrow[ul, dash] \arrow[dl, dash, "H_p"'] \arrow[ddl, dash, "G_p"] \\
& K_t \cap K_p \arrow[d, dash] & \\
& \mathbb{Q} &
\end{tikzcd}$$

The Galois group $H_p = \mathrm{Gal}(K_r/K_t)$ is isomorphic to that of the extension $K_p/K_t \cap K_p$. Hence H_p is a normal subgroup of $G_p = \mathrm{PGL}_2(\mathbb{F}_p)$. We use the fact that $\mathrm{PSL}_2(\mathbb{F}_p)$ is simple for $p > 3$. So H_p is either all of G_p, $\mathrm{PSL}_2(\mathbb{F}_p)$, the trivial group or, in the case $p = 3$, the Klein group V_4 in $\mathrm{PGL}_2(\mathbb{F}_3) = \mathfrak{S}_4$. Treating the four cases separately, we will prove that W_p is an irreducible $\mathbb{Q}[H_p]$-module.

If H_p is all of G_p, then W_p is irreducible as a $\mathbb{Q}[H_p]$-module by Theorem 5. If H_p is equal to $\mathrm{PSL}_2(\mathbb{F}_p)$, then W_p could split at most into two subspace of equal dimension since $\mathrm{PSL}_2(\mathbb{F}_p)$ has index 2 in $\mathrm{PGL}_2(\mathbb{F}_p)$. But the dimension of W_p is odd unless $p = 2$, which we excluded. Hence W_p is irreducible.

Next, we will exclude the case when H_p is trivial. If it were so, then there is a surjective map from G_t onto $G_p = \mathrm{PGL}_2(\mathbb{F}_p)$. The group G_t is contained in $\mathrm{PGL}_2(\mathbb{Z}/t\mathbb{Z})$, whose order is

$$\prod_{\ell \mid t} \ell \cdot (\ell + 1) \cdot (\ell - 1).$$

So the order of G_t cannot be divisible by p since p is larger than any of the ℓ, unless $p = 3$ and $t = 2$. It is also impossible that there is a surjective map from $\mathrm{PGL}_2(\mathbb{F}_2)$ onto $\mathrm{PGL}_2(\mathbb{F}_3)$. So H_p is not trivial.

Finally, we treat the case when H_p is the Klein group in $\mathrm{PGL}_2(\mathbb{F}_3)$. Since $p = 3$, we have $t = 2$. As $G_2 = \mathrm{PGL}_2(\mathbb{F}_2) = \mathfrak{S}_3$, the only possibility for this case is when K_2 is contained in K_3. But it was shown in Lemma 10 that this is not possible under our additional hypothesis.

Let X be a sub-$\mathbb{Q}[G_r]$-module of $W_r = W_p \otimes W_t$. As H_p acts trivially on W_t, we deduce that there is a subspace Z of W_t such that $X = W_p \otimes Z$. By the induction hypothesis, we know that W_t is irreducible as a $\mathbb{Q}[G_t]$-module. Hence $Z = W_t$ and we have shown that W_r is $\mathbb{Q}[G_r]$-irreducible.

Now we will prove (i). If the additional hypothesis is verified, then W_s is an irreducible $\mathbb{Q}[G_s]$-module by (ii); hence G_s acts transitively on $\mathbb{P}^1(\mathbb{Z}/s\mathbb{Z})$. But the only place where we used the additional hypothesis in the proof of (ii) is when we excluded the possibility that H_p is the Klein group in $\mathrm{PGL}_2(\mathbb{F}_3)$. But since the

Klein group acts transitively on $\mathbb{P}^1(\mathbb{F}_3)$, we can prove directly the truth of (i) in general.

Finally we must prove (iii). We follow again along the same lines as the proof of (ii). Of course, we may assume that $m > 1$. Let $1 \leqslant i \leqslant k$, and let $r \mid s$. We will prove by induction that $U_i \otimes W_r$ is an irreducible $\mathbb{Q}[G_{rm}]$-module. Let p be the largest prime dividing r and let $t = r/p$. By induction, we may suppose that $U_i \otimes W_t$ is G_{tm}-irreducible. Let $H_p = \mathrm{Gal}(K_{rm}/K_{tm}) \subset \mathrm{PGL}_2(\mathbb{F}_p)$. As before, if we can prove that W_p is an irreducible $\mathbb{Q}[H_p]$-module, then we know that $U_i \otimes W_r = U_i \otimes W_t \otimes W_p$ is G_{rm}-irreducible. Once again we must exclude only the possibility that H_p is trivial or equal to the Klein group V_4 in $\mathrm{PGL}_2(\mathbb{F}_3)$.

Suppose first that $p = 2$. By maximality of p, we must have $t = 1$. If H_p is trivial, then there is a surjective map from G_m to $\mathrm{PGL}_2(\mathbb{F}_2)$. Running through all the possible odd m in Lemma 8, we find that only $m = 3$ can be possible. Moreover in this case we must have $K_2 = K_3$. Again we use Lemma 10 to exclude this possibility.

We treat now the case that $p = 3$. Then $t = 1$ or $t = 2$. Suppose that H_p is trivial. There must be a surjective map from G_{tm} to $\mathrm{PGL}_2(\mathbb{F}_3) \cong \mathfrak{S}_4$. We can check that if $t = 1$, then we must have $m = 7$ since otherwise $\#G_m$ will not be a multiple of 3. But $\#G_7$ is not divisible by 24. If $t = 2$, then m can only be 5 or 7. Again it cannot be 7. So we must have $G_{tm} \subset \mathfrak{S}_3 \times (\mathbb{Z}/4\mathbb{Z} \ltimes \mathbb{Z}/5\mathbb{Z})$, and it is easy to check that the latter group does not have a subquotient isomorphic to $\mathfrak{S}_4$.

Continuing with the case $p = 3$, we suppose now that H_p is the Klein group in $\mathrm{PGL}_2(\mathbb{F}_3)$. This time we have a surjection of G_{tm} onto $\mathfrak{S}_3$. If $t = 1$, we can again check that there is no possibility for G_m. So suppose that $t = 2$. Then G_{tm} is contained in $\mathfrak{S}_3 \times G_m$. Then the only possibility for the surjection is that G_m lies in its kernel and $\mathrm{PGL}_2(\mathbb{F}_2)$ maps isomorphically onto $\mathfrak{S}_3$. In this case we would have that K_2 is contained in K_3. Once again Lemma 10 excludes this.

The very last step is to assume that $p > 3$ and that H_p is trivial. Then there is a surjective map from G_{tm} to $\mathrm{PGL}_2(\mathbb{F}_p)$. By the maximality of p, we know that $\#\,\mathrm{PGL}_2(\mathbb{Z}/t\mathbb{Z})$ is not divisible by p. Therefore $p \neq m$ must divide $\#G_m$. Running through the list of possible groups in Lemma 8, we find that this is not possible. □

5.2. *Results for semistable curves.*

Theorem 12. *Let $E/\mathbb{Q}$ be a semistable elliptic curve of conductor N with N not equal to* 30 *or* 210. *Then all the self-points P_C are of infinite order in $E(\mathbb{Q}(C))$.*

Proof. By Lemma 9, we may choose a prime p dividing N such that $\bar{\rho}_p$ is surjective and such that $p - 1 > m$.

Any cyclic subgroup C of order N may be written as $C = A \oplus B$, with A of order m and B of order $s = N/m$. Now we use the previous lemma. For any fixed A, the group G_N acts transitively on the set $\{A \oplus B\}_B$ as B runs over all cyclic subgroups of order s in E. Hence all self-points $\{P_C\}$ with the m-part A fixed are conjugate in $E(K_N)$. In particular, if $m = 1$, then all self-points are conjugate and the fundamental Theorem 2 proves the theorem. So suppose now that $m > 1$.

Now we use the p-adic proof of Theorem 2. We identify the curve $E/\overline{\mathbb{Q}}_p$ with the Tate curve $\overline{\mathbb{Q}}_p^\times/q_E^{\mathbb{Z}}$. Fix a cyclic subgroup A of order m in E, and let $B = \mu[s]$ and $C = A \oplus B$. Since any self-point is conjugate to such a point, it is sufficient to prove that P_C is of infinite order.

For each $\ell \mid m$, let A_ℓ be the ℓ-torsion part of A. Write A'' for the direct sum of all A_ℓ such that A_ℓ is generated by the ℓ-th roots of unities $\mu[\ell]$ in $E(\overline{\mathbb{Q}}_p)$. Write A' for the sum of all other A_ℓ. So $A = A' \oplus A''$. Denote the order of A' by m' and likewise the order of A'' by m''. Now we consider the isogeny ψ with kernel A', given by

$$\begin{tikzcd} 0 \arrow[r] & A' \arrow[r] & E \arrow[r, "\psi"] & E' \arrow[r] & 0. \end{tikzcd}$$

If $\hat{A}'$ is the kernel of the dual isogeny $\hat{\psi} : E' \to E$, then we may consider the point

$$x'_C = (E', \hat{A}' \oplus \psi(A'') \oplus \psi(B)) \in X_0(N)(\overline{\mathbb{Q}}_p),$$

which is nothing other than the Atkin–Lehner involution $w_{m'}$ applied to the point $x_C = (E, C)$. We know already that $\psi(B) = \mu[k]$ and $\psi(A'') = \mu[m'']$, but we also see that the group $\hat{A}'$ is isomorphic to $\mu[m']$. Hence the point x'_C lies now close to the cusp ∞ and its Tate-parameter will be a certain m'-th root u of q_E. Since

$$|u|_p = (|q_E|_p)^{1/m'} = p^{-c_p/m'} < p^{-1/(p-1)}$$

because $m' \leqslant m < p-1$, we can apply Lemma 3 to show that $\varphi_E(x'_C)$ is of infinite order. But the Atkin–Lehner involutions w_ℓ act like multiplication by $-a_\ell \in \{\pm 1\}$ for all primes ℓ dividing N, as shown in [Atkin and Lehner 1970]. Therefore $P_C = \varphi_E(x_C) = \pm\varphi_E(x'_C) + T$, where T is a point of finite order, and hence P_C is of infinite order. □

As remarked earlier we have a G_N-equivariant map

$$\iota : W_N \to E(K_N) \otimes \mathbb{Q}$$

Part (ii) of Lemma 11 shows this:

Theorem 13. *Let $E/\mathbb{Q}$ be a semistable elliptic curve with N not equal to* 30 *or* 210*. Suppose all the representations $\bar{\rho}_p$ for all primes $p \mid N$ are surjective. Then the group generated by the self-points is of rank N and the Galois group acts like the irreducible Steinberg representation W_N on it.*

We prove now an extension of this theorem to the case when $m \neq 1$. In particular W_N might not be irreducible anymore. Unfortunately we cannot prove that the rank is N in general for a semistable curve since we have to exclude the possibility that the curve has two distinct isogenies of the same degree defined over $\mathbb{Q}$: If the curve has two isogenies of degree p over $\mathbb{Q}$, then in the decomposition of W_N into irreducible $\mathbb{Q}[G]$-modules, there will be a representation that appears with multiplicity 2. The second hypothesis in the following theorem excludes this possibility, but it is also needed elsewhere to be able to apply the lemmata from the previous section.

Theorem 14. *Let $E/\mathbb{Q}$ be a semistable elliptic curve. Suppose that N is not equal to* 30 *or* 210*. Suppose that for each prime $p \mid N$ such that $\bar{\rho}_p$ is not surjective, there is a prime $\ell \mid N$ such that the Tamagawa number c_ℓ is not divisible by p. Then the group generated by the self-points is of rank N.*

Proof. As a consequence of the second hypothesis, we know that for each $p \mid N$ there is an element of order p in G_p. See the appendix of [Serre 1968]. Since either G_p is all of $\mathrm{PGL}_2(\mathbb{F}_p)$ or it is contained in the Borel subgroup, we conclude that either G_p acts transitively on $\mathbb{P}^1(\mathbb{F}_p)$ or it has one single fixed point, which we will call $C_p \in \mathbb{P}^1(\mathbb{F}_p)$.

Let $p \mid m$. Then by Proposition 6, the $\mathbb{Q}[G_p]$-module W_p decomposes as the sum of the trivial part W'_1 and an irreducible part W'_p of dimension $p-1$. If m is not prime it can only be either $2 \cdot 3$ or $2 \cdot 5$ by Mazur's theorem. If $m = 6$, then W_6 decomposes as $W'_1 \oplus W'_2 \oplus W'_3 \oplus W'_6$, where $W'_6 = W'_2 \otimes W'_3$. To see that the latter is also irreducible one needs only to note that the dimension of W'_2 is 1. In the same way, for $m = 10$, we have an irreducible component W'_{10}.

Using Lemma 11, we know now that W_N decomposes as

$$W_N = \bigoplus_{d \mid m} (W'_d \otimes W_s)$$

into irreducible $\mathbb{Q}[G_N]$-modules. We must now prove that none of the components belongs to the kernel of the map $\iota : W_N \to E(K) \otimes \mathbb{Q}$.

First recall the definition of $W'_d \otimes W_s$. It contains all elements

$$\sum_{C \in \mathbb{P}^1(\mathbb{Z}/N\mathbb{Z})} a_C e_C \in \bigoplus_{C \in \mathbb{P}^1(\mathbb{Z}/N\mathbb{Z})} \mathbb{Q} e_C,$$

subject to the following three conditions.

- For all $N \neq b \mid N$ and all cyclic subgroups B of order b, the sum $\sum_{C \supset B} a_C$ vanishes.
- For all primes $p \mid d$ and all $C \supset C_p$, we have $a_C = 0$.
- For all primes $p \mid (m/d)$ and all $C \not\supset C_p$, we have $a_C = 0$.

Let $d \mid m$. Define A to be the direct sum of C_p for all $p \mid (m/d)$. So A is a cyclic group of order m/d. The map $\imath$ on $W'_d \otimes W_s$ is induced from the map

$$\imath_d : \bigoplus_D \mathbb{Q} e_{A\oplus D} \longrightarrow E(K) \otimes \mathbb{Q},$$

where D runs through all the cyclic subgroups D in E of order $d \cdot s$ such that D does not contain any of the C_p with $p \mid d$. Since this map sends $e_{A\oplus D}$ to the self-point $P_{A\oplus D}$, it follows from Theorem 12 that the map $\imath_d$ is not trivial.

Now we use the relations in Proposition 4 to see that, for all $b \mid ds$ and all cyclic groups B of order b not containing any of the C_p, we have $\sum_{D \supset B} e_{A\oplus D} \in \ker \imath_d$. Hence the only irreducible part of the domain of $\imath_d$ that does not lie in the kernel is $W'_d \otimes W_s$ Hence $\imath_d$ induces an injection $W'_d \otimes W_s \to E(K) \otimes \mathbb{Q}$. □

The hypothesis in this last theorem is fulfilled for the very large part of semistable curves. We could not find a strong Weil curve with $N < 10{,}000$ for that the theorem would not apply. The first curve which does not satisfy the hypothesis with $p = 3$ is 651e2 since it has $G_3 = \mathbb{Z}/2\mathbb{Z}$, and the Tamagawa numbers are $c_3 = 3$, $c_7 = 3$, and $c_{31} = 3$. For $p = 2$, the examples that do not satisfy the hypothesis are exactly those that have all 2-torsion points defined over $\mathbb{Q}$, as for instance 30a2.

6. Examples

Table 2 shows some computations done for the optimal curves (with one exception) of smallest conductor. We do not give the complete explanation of how one obtains these results. For more detail, we refer the reader to [Delaunay and Wuthrich 2008] and [Wuthrich 2007]. We will consider two curves in more detail later.

N	11a1	14a1	15a1	17a1	19a1	20a1	21a1	24a1	26a1
torsion	5	$2\cdot 3$	$2\cdot 4$	4	3	$2\cdot 3$	$2\cdot 4$	$2\cdot 4$	3
isogeny	25	18	16	4	9	6	8	8	9
W_N	1	2	1	1	1	2	1	4	1
rank	11	14	15	17	19	15	21	18*	26

N	26b1	27a**2**	30a1	32a1	33a1	34a1	35a1	37a1	38a1
torsion	7	3	$2\cdot 3$	4	$2\cdot 2$	$2\cdot 3$	3	1	3
isogeny	7	27	12	4	4	6	9	1	9
W_N	1	5	4	?	1	2	1	1	1
rank	26	**20**	30*	**12***	33	34	35	37	38

Table 2. The ranks of the group generated by self-points for some curves.

We label the curves as in Cremona's tables [1997]. The first line of our table shows the structure of the torsion group over $\mathbb{Q}$; for example, $2 \cdot 4$ means that $E(\mathbb{Q})_{\text{tors}} = \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/4\mathbb{Z}$. The next line indicates the largest degree of a cyclic isogeny defined over $\mathbb{Q}$ on E. The last two lines are those containing information about self-points: First we counted the number of irreducible $\mathbb{Q}[G_N]$-modules in W_N, and finally we computed the rank of the group generated by self-points in $E(K_N)$.

The two values in bold face are lower than the usual conjectured rank, which is no surprise since these two curves have complex multiplication. When there is no $*$ sign next to the rank, the value is proved using the results in the previous section. The sign $*$ indicates that we have only empirically computed the rank using the following method.

Using high precision computation we may find a very good approximation to the values of

$$z_C = \int_{x_C}^{\infty} f_E(q)\frac{dq}{q}$$

as elements of $\mathbb{C}$, where C runs over all cyclic subgroups of order N in E. Hence z_C maps to P_C under $\mathbb{C} \to \mathbb{C}/\Lambda_E \to E(\mathbb{C})$, where $\Lambda_E = \mathbb{Z}\omega_1 \oplus \mathbb{Z}\omega_2$ is the period lattice of E. Let t be the order of the torsion subgroup of E over $\mathbb{Q}$. Consider the abelian group spanned by $\frac{1}{t}\omega_1$, $\frac{1}{t}\omega_2$ and all the z_C in a complex vector space of dimension $2 + \#\mathbb{P}^1(\mathbb{Z}/N\mathbb{Z})$. Using the LLL algorithm, we find small vectors in this lattice. These are likely to give relations

$$b_1\omega_1 + b_2\omega_2 + \sum_C a_C z_C = 0$$

with b_1, b_2, and a_C all integers. This yields a probable relation among the self-points. Unfortunately we might not catch those relations involving torsion points on E not defined over $\mathbb{Q}$. So to increase the likelihood of finding all relations we multiply t by a product of small primes. For all cases for which we were able to determine the rank, this empirical computation gave the same answer. In principle these computations could be made rigorous by considering exact estimates for the error terms.

6.1. ***Conductor 24.*** We present here an example of a curve where we are unable to determine the rank of the group generated by self-points. The Mordell–Weil group of the curve 24a1, given by the equation

$$E: \quad y^2 = x^3 - x^2 - 4 \cdot x + 4,$$

is $E(\mathbb{Q}) = \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/4\mathbb{Z}$. The situation is rather complicated and we do not explain all computations here. The field K_4 turns out to be $\mathbb{Q}(i, \sqrt{3})$, which happens to be equal to $\mathbb{Q}(E[4])$. There is are two nontrivial Galois-orbits of 4-torsion points, one over $\mathbb{Q}(\sqrt{3})$ and the other over $\mathbb{Q}(\sqrt{-3})$. Hence the representation V_4 splits as

$$V_4 = \mathbb{1} \oplus \mathbb{1} \oplus \mathbb{1} \oplus \mathbb{1} \oplus \mathbb{1}(\sqrt{3}) \oplus \mathbb{1}(\sqrt{-3})\,,$$

where $\mathbb{1}(\sqrt{d})$ is the one-dimensional representation corresponding to the Dirichlet character associated to $\mathbb{Q}(\sqrt{d})$. Now the field K_8 can be computed too; it coincides with $\mathbb{Q}(E[8])$ in this case. It is a degree 16 extension of discriminant $2^{36} \cdot 3^{12}$, and contains the extension $\mathbb{Q}(i, \sqrt{2}, \sqrt{3})$. The subextension K_4 is fixed by the centre of the Galois group G_8. The group G_8 admits two irreducible 2-dimensional representations, one of which we call Z_2. Then the representation V_8 splits in many components and we find that

$$W_8 = \mathbb{1}(\sqrt{2}) \oplus \mathbb{1}(\sqrt{-2}) \oplus Z_2 \oplus Z_2.$$

The first two factors correspond to two pairs of lines in $E[8]$ defined over $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt{-2})$ respectively. The other lines are defined over fields of degree 4.

Using that the field K_3 intersects K_8 in $\mathbb{Q}(\sqrt{-3})$, we find that W_{24} splits into 4 irreducible factors $W_{24} = W_3(\sqrt{2}) \oplus W_3(\sqrt{-2}) \oplus Z_6 \oplus Z_6$. Here $Z_6 = W_3 \otimes Z_2$ is an irreducible representation of dimension 6. In particular, this representation appears with multiplicity 2. So the usual proof that there are no further relations among self-points will not work.

The cyclic subgroup of order 8 in E that corresponds to $\mu[8]$ over $\mathbb{Q}_3$ contains the rational 4-torsion point. So one of the two factors of dimension 3 in W_{24} certainly appears in $E(K_N) \otimes \mathbb{Q}$. But we are unable to show that any other self-points are of infinite order by means of Theorem 12.

Though we can only conclude that the rank r of the group generated by the self-points satisfies $3 \leqslant r \leqslant 18$, we strongly believe that $r = 18$, as suggested by the empirical computations.

6.2. ***Conductor 27.*** There are four curves of conductor 27 forming the isogeny graph

$$27a2 \leftarrow 27a1 \leftarrow 27a3 \leftarrow 27a4$$

The isogenies $\leftarrow$ are all of degree 3, and in the sense that they are drawn here, the kernels are $\mathbb{Z}/3\mathbb{Z}$ while the dual isogenies have kernel $\mu[3]$. Over the field $F = \mathbb{Q}(\sqrt{-3}) = \mathbb{Q}(\zeta)$, with ζ a third root of unity, the curves 27a1 and 27a3 become isomorphic, the same holds for the curves 27a2 and 27a4. The first pair has complex multiplication by the maximal order $\mathbb{Z}[\zeta]$, while the second pair has complex multiplication by $\mathbb{Z}[3\zeta]$.

Let E be the curve 27a2 defined by $y^2 + y = x^3 - 270 \cdot x - 1708$.

Theorem 15. *The self-points on the curve* 27*a*2 *generate a group of rank* 20 *in* $E(K_{27})$. *There are exactly two linearly independent self-points defined over* $K_3 = \mathbb{Q}(\sqrt[6]{-3})$, *and they generate a subgroup of finite index in* $E(K_3)$.

The proof is contained in the following explanations, but we do omit certain computations.

The field K_3 is equal to $\mathbb{Q}(\sqrt[6]{-3})$, and the Galois group G_3 is a dihedral group of order 6. In fact some 3-torsion points are defined over $F = \mathbb{Q}(\sqrt{-3})$, some others are over $\mathbb{Q}(\sqrt[3]{-3})$, and we have $V_3 = \mathbb{1} \oplus \mathbb{1}(\sqrt{-3}) \oplus Z_2$, where Z_2 is the unique irreducible 2-dimensional representation of G_3.

In order to determine the structure of V_{27}, we need to use the theory of complex multiplication. Let H_{27} be the subgroup $\mathrm{Gal}(K_{27}/F)$ inside G_{27}. We know that the representation $\bar{\rho}_{27,F}$ now maps to

$$\bar{\rho}_{27,F} : H_{27} \rightarrowtail \frac{\mathrm{Aut}_{\mathcal{O}/27\mathcal{O}}(E[27])}{(\mathbb{Z}/27\mathbb{Z})^\times} = \frac{(\mathcal{O}/27\mathcal{O})^\times}{(\mathbb{Z}/27\mathbb{Z})^\times}$$
$$\xrightarrow{\cong} \left\{\left(\begin{smallmatrix}1 & * \\ 0 & 1\end{smallmatrix}\right) \in \mathrm{PGL}_2(\mathbb{Z}/27\mathbb{Z})\right\} \xrightarrow{\cong} \mathbb{Z}/27\mathbb{Z},$$

where $\mathcal{O} = \mathbb{Z}[3\,\zeta]$ is the ring of endomorphisms of E/F. It is possible to verify that H_{27} is equal to this group, and hence G_{27} is a dihedral group of order 54 generated by $h = \left(\begin{smallmatrix}1 & 1 \\ 0 & 1\end{smallmatrix}\right)$ and $s = \left(\begin{smallmatrix}1 & 0 \\ 0 & -1\end{smallmatrix}\right)$. The computation of V_{27} is now easy and one finds

$$W_{27} = \mathbb{1} \oplus \mathbb{1}(\sqrt{-3}) \oplus Z_2 \oplus Z_2 \oplus Z_{18}.$$

Here Z_2 is the unique 2-dimensional irreducible $\mathbb{Q}[G_{27}]$-module (the action of h has trace -1), and Z_{18} is the unique irreducible 18-dimensional $\mathbb{Q}[G_{27}]$-module (it splits over $\mathbb{C}$ into six 2-dimensional representations). Since the curve 27a2 is not the strong Weil curve in the isogeny class, the modular parametrisation φ_E from the elliptic curve $X_0(27)$ to E is not an isomorphism but an isogeny of degree 3. The curve $X_0(27)$ has six cusps represented by the classes $\{\infty, 0, \frac{1}{3}, \frac{2}{3}, \frac{2}{9}, \frac{4}{9}\}$. The group $X_0(27)(\mathbb{Q})$ contains the cusps ∞ and 0 and the self-point obtained from the isogeny 27a2→27a4. They form exactly the kernel of φ_E. The other cusps are mapped to the 3-torsion points defined over F on E. In fact $E(F) = \mathbb{Z}/3\mathbb{Z}$ and $E(K_3)_{\mathrm{tors}} = \mathbb{Z}/3\mathbb{Z} \oplus \mathbb{Z}/3\mathbb{Z}$. A two-descent over K_3 shows that the 2-Selmer group of E/K_3 has two copies of $\mathbb{Z}/2\mathbb{Z}$ in it.

The trivial factor in W_{27} corresponds to the self-point obtained from the 27-isogeny defined over $\mathbb{Q}$ on 27a2. We know that it is the point O in $E(\mathbb{Q})$. The factor $\mathbb{1}(\sqrt{-3})$ in W_{27} must also belong to the kernel of $\iota : W_{27} \to E(K_{27}) \otimes \mathbb{Q}$ since the Mordell–Weil group $E(F)$ is of rank 0. Of the factors Z_2 at least one must be in the kernel since the rank of $E(K_3)$ is bounded by 2 from above. It is not hard to check by looking at traces of Frobenii that the torsion subgroup of $E(K_{27})$ only contains nine 3-torsion points. Since the degree of φ_E is 3, there are at most 27 points in $X_0(27)(K_{27})$ that map to torsion points in $E(K_{27})$ under φ_E. Since there are 36 points x_C, we conclude that at least 9 self-points are of infinite order.

Looking at the decomposition of W_{27}, we see that Z_{18} cannot belong to the kernel of $\imath$.

Finally we have to show that there is a self-point of infinite order in $E(K_3)$. This will show that the second copy of Z_2 does not belong to the kernel of $\imath$. This can be done numerically. The point $\tau_C = \frac{1}{6} \cdot (-1 + \sqrt{-3})$ in the upper half plane corresponds to a point x_C in $X_0(27)$. We find that

$$-\tfrac{1}{8}(36 \cdot s^5 + 15 \cdot s^4 - 45 \cdot s^3 - 18 \cdot s^2 + 69 \cdot s + 99) \quad \text{with } s = \sqrt[6]{-3}$$

is the x-coordinate of the self-point P_C in $E(K_3)$. Its canonical height is 1.5191 and hence P_C is of infinite order. This point P_C and its conjugates over F will generate a group of rank 2 in $E(K_3)$. Since we have computed the 2-Selmer group earlier, we conclude that the rank of $E(K_3)$ is as claimed equal to 2.

It seems plausible that this P_C can also be constructed as an "exotic Heegner point" using the construction of Bertolini, Darmon and Prasanna [$\geq$ 2009], but the authors exclude there explicitly the case of conductor $N = 27$.

7. Higher self-points

In this section, we investigate three particular cases of higher self-points. Let $E/\mathbb{Q}$ be an elliptic curve of conductor N. For any cyclic subgroup D in E we may consider the isogenous curve E/D with a suitable choice of a cyclic subgroup of order N in it. In the first case, we use subgroups D defined over $\mathbb{Q}$ to construct new points and for the two other cases we use subgroups D of prime-power order p^n, first when p divides the conductor and then when it does not divide the conductor.

7.1. ***Self-points via rational isogenies.*** Let D be a cyclic subgroup in E defined over $\mathbb{Q}$. Suppose for simplicity that the order of D is prime to N. Then for any cyclic subgroup C of order N on E,

$$Q_D = \varphi_E(E/D, (C + D)/D)$$

is a higher self-point defined over the same field as P_C. It would be interesting to know in general when P_C and Q_D are linearly independent. For instance this can be shown on the curves of conductor 11: There are 3 curves in the isogeny class, and hence we find, for any fixed C, one self-point and two higher self-points on E defined over $\mathbb{Q}(C)$. Using the canonical height pairing, we can prove the linear independence of these three points computed explicitly on E. So the rank of $E(\mathbb{Q}(C))$ will have to be at least 3. See [Delaunay and Wuthrich 2008] and [Wuthrich 2007] for more details on this example.

In some cases the method of the proof of Theorem 12 can be used to show that Q_D is also of infinite order. But the methods of the proof of Theorem 14 will not be sufficient to prove the independence of P_C and Q_D.

7.2. ***The multiplicative case.*** Let now p be a prime dividing N exactly once, that is, E has multiplicative reduction at p. Let M be such that $N = p \cdot M$. As a base-field we will consider here the number field $F = K_M$, the smallest field such that its absolute Galois group acts as scalars on $E[M]$. In the particular situation when $N = p$ is prime then $F = \mathbb{Q}$; the same is true for instance if E is a curve of conductor 14 and $p = 7$.

For any $n \geqslant 0$, we define now F_n to be the field $K_{p^n N}$ and H_n to be the Galois group of F_n/F. Via the Galois representation

$$\rho_{F,p} : \mathrm{Gal}(\bar{F}/F) \longrightarrow \mathrm{Aut}(T_p E) \cong \mathrm{GL}_2(\mathbb{Z}_p) \longrightarrow \mathrm{PGL}_2(\mathbb{Z}_p),$$

the group H_n identifies with a subgroup of $\mathrm{PGL}_2({}^{\mathbb{Z}}/_{p^{n+1}\mathbb{Z}})$.

Fix a subgroup B of order M in E. Let $n \geqslant 0$, and let D be a cyclic subgroup of order p^{n+1} in E. Let $A = D[p]$ and $C = A \oplus B$, which is a cyclic subgroup of order N. Write ψ for the isogeny $E \to E'$ of kernel D and $\hat{\psi}$ for its dual. Define

$$C' = \ker(\hat{\psi})[p] \oplus \psi(B),$$

which is a cyclic subgroup of E' of order $M \cdot p = N$. The image of the point $y_D = (E', C') \in Y_0(N)$ through the map φ_E will be denoted by Q_D. It is by definition a higher self-point. We will say that "Q_D lies over P_C" or "over B".

In particular, if $n = 0$, then $D = A$ is a cyclic subgroup of order p. From the construction above, we see that the point y_D is nothing but $w_p(x_C)$, where w_p is the Atkin–Lehner involution on $X_0(N)$. Hence we have that $Q_D = -a_p \cdot P_C + T$ for some 2-torsion point T defined over $\mathbb{Q}$. Here $a_p = \pm 1$ is, as before, the Hecke eigenvalue of the newform f_E attached to the isogeny class of E.

Let D be a cyclic subgroup of E of order p^{n+1}. By the definition of the Hecke operator T_p on $J_0(N)$, we have $T_p((y_D) - (\infty)) = \sum_{D' \supset D}((y_{D'}) - (\infty))$, where the sum runs over all cyclic subgroups D' in E of order p^{n+2} containing D. This gives us the relation

$$a_p \cdot Q_D = \sum_{D' \supset D} Q_{D'}. \tag{2}$$

Hence by induction, we know that Q_D is of infinite order if the self-point P_C is.

Lemma 16. *Let B be a fixed subgroup of order M in E, and let $n \geqslant 0$. Then $\sum_D Q_D$ is a torsion point in $E(F)$, where the sum is over all cyclic subgroups D of E of order p^{n+1}.*

Proof. Suppose first that $n = 0$. Then we sum over all cyclic subgroups $D = A$ of order p, which gives

$$\sum_D Q_D = \sum_{C \supset B} (-a_p P_C + T) = (p+1) \cdot T - a_p \sum_{C \supset B} P_C.$$

The first term on the right side is clearly torsion and the second term contains exactly one of the relations from Proposition 4. Now by induction, we assume that the statement holds for n. But then $\sum_{D'} Q_{D'}$, with the sum running over all cyclic subgroups D' of order p^{n+2}, is, by (2), equal to $a_p \cdot \sum_D Q_D$, with the sum now running over cyclic subgroups of order p^{n+1}. □

The $\mathbb{Q}$-vector space with basis $\{e_D\}_D$ in bijection with $\mathbb{P}^1(\mathbb{Z}/_{p^{n+1}}\mathbb{Z})$ is a natural $\mathbb{Q}[H_n]$-module. Define

$$V'_{(n)} = \frac{\bigoplus_A \mathbb{Q}e_D}{\mathbb{Q}(\sum_D e_D)},$$

which is a vector space of dimension $p^{n+1} + p^n - 1$.

Fix a cyclic subgroup B of order M in E. By the previous lemma, there is a morphism of $\mathbb{Q}[H_n]$-modules given by

$$\iota_n = \iota_{B,n} : V'_{(n)} \longrightarrow E(F_n) \otimes \mathbb{Q}, \quad e_D \longmapsto Q_D$$

We assume that the Galois representation $\rho_{F,p}$ is surjective onto $\mathrm{PGL}_2(\mathbb{Z}_p)$. So H_n is isomorphic to $\mathrm{PGL}_2(\mathbb{Z}/_{p^{n+1}}\mathbb{Z})$ and the $\mathbb{Q}[H_n]$-module $V'_{(n)}$ is the Steinberg representation, which was denoted by V_{p^n}/W_1 in Section 4.

Theorem 17. *Suppose $E/\mathbb{Q}$ is an elliptic curve and p a prime of multiplicative reduction. Suppose that $\rho_{F,p}$ is surjective and that there is a self-point P_C of infinite order in $E(F_0)$. Then for all $n \geqslant 0$ and all cyclic subgroups D of order p^{n+1} with $D[p] \subset C$, the point Q_D is of infinite order. They generate in $E(F_n) \otimes \mathbb{Q}$ a $\mathbb{Q}[H_n]$-module isomorphic to the representation $V'_{(n)}$ of dimension $p^{n+1} + p^n - 1$.*

As a special case, we recover [Delaunay and Wuthrich 2008, Theorem 8] in the case when $N = p$ is prime and $F = \mathbb{Q}$.

Proof. We only have to show that ι_n is injective. Suppose $n \geqslant 0$ is the smallest value such that ι_n is not injective. Since $V'_{(n)} = W_{p^{n+1}} \oplus V'_{(n-1)}$ if $n > 0$ and $V'_{(0)} = W_p$, this means that ι_n induced on $W_{p^{n+1}}$ is not injective. Since this is an irreducible $\mathbb{Q}[H_n]$-module when $\rho_{F,p}$ is surjective, this means that ι_n is trivial on $W_{p^{n+1}}$. This is impossible since we have shown that all Q_D above P_C are of infinite order. □

7.3. ***The good case.*** Let p be a prime not dividing N, that is, of good reduction for E. Let F be a number field such that $E(F)$ contains a self-point P_C of infinite order. We fix the corresponding cyclic subgroup C of order N in E.

For any $n \geqslant 0$, let F_n be the smallest Galois extension of F such that the absolute Galois group $\mathrm{Gal}(\bar{F}/F)$ acts via scalars on $E[p^{n+1}]$; hence $F_n = F \cdot K_{p^{n+1}}$. Define H_n to be the Galois group $\mathrm{Gal}(F_n/F)$, which will be considered as a subgroup of $\mathrm{PGL}_2(\mathbb{Z}/_{p^{n+1}}\mathbb{Z})$.

For any $n \geqslant 0$ and any cyclic subgroup D of order p^{n+1}, we construct a higher self-point Q_D in $E(F_n)$ as follows. Let $\psi : E \to E/D$ be the isogeny associated to D. Put $y_D = (E/D,\ \psi(C)) \in Y_0(N)$ and $Q_D = \varphi_E(y_D)$. This is a higher self-point "above P_C".

Again we may use the definition of the Hecke operator T_p to prove that, for all $n \geqslant 0$ and D as before,

$$a_p \cdot Q_D = \sum_{D' \supset D} Q_{D'}, \tag{3}$$

where the sum runs over all cyclic subgroups D' of order p^{n+2} in E containing D. Furthermore we have

$$a_p \cdot P_C = \sum_{D} Q_D, \tag{4}$$

with the sum running over all cyclic subgroups D of order p in E.

Let $V_{(n)} = V_{p^{n+1}}$ be the $\mathbb{Q}[H_n]$-module whose basis $\{e_D\}_D$ as a vector space over $\mathbb{Q}$ is in bijection with $\mathbb{P}^1(\mathbb{Z}/_{p^{n+1}\mathbb{Z}})$. We have a H_n-morphism defined by

$$\iota_n = \iota_{C,n} : V_{(n)} \longrightarrow E(F_n) \otimes \mathbb{Q}, \quad e_D \longmapsto Q_D$$

Theorem 18. *Let $E/\mathbb{Q}$ be an elliptic curve of conductor N. Let p be a prime of good and ordinary reduction for E. Let F be a number field such that $E(F)$ contains a self-point P_C of infinite order. Suppose that the representation $\rho_{F,p}$ is surjective. Then all higher self-points Q_D constructed above are of infinite order and they generate a group of rank $p^n \cdot (p+1)$.*

Proof. By induction on n, using the formulae (3) and (4) and the hypothesis that p is ordinary to guarantee that $a_p \neq 0$. □

The above easy proof of the theorem breaks down if E has supersingular reduction at p, for a_p is then almost always equal to 0.

Theorem 19. *Let $E/\mathbb{Q}$ be a semistable elliptic curve of conductor N not equal to* 30 *or* 210*. Let $p > N$ be a supersingular prime for E. Let $F = K_N$. Suppose that the representation $\rho_{F,p}$ is surjective. Then all higher self-points Q_D above a given self-point P_C are of infinite order, and they generate a group of rank $p^n \cdot (p+1)$.*

Proof. We follow the proof of Theorem 12. Let $\ell > 2$ be a prime dividing N. We proved that the self-points are of infinite order by showing that when a certain Atkin–Lehner involution is applied to one of the conjugates of x_C, one obtains a point ℓ-adically close to the cusp ∞ on $X_0(N)(\overline{\mathbb{Q}}_\ell)$.

Let Q_D be a higher self-point above the self-point P_C. Since $\rho_{F,p}$ is surjective, the point Q_D will be conjugate over K_N to all other higher self-points above the same self-point. Therefore without loss of generality we may assume that the cyclic subgroup D on E corresponds to $\mu[p^{n+1}]$ in $E(\overline{\mathbb{Q}}_\ell)$. Then the point $y_D = (E', C')$ is represented by a Tate curve over $\overline{\mathbb{Q}}_\ell$ with parameter $q_{E'}$ equal to the p^{n+1}-st power of q_E.

Let r be a divisor of N such that $w_r(y_D)$ is the pair $(E'', \mu[N])$, with E'' the Tate curve with parameter $q_{E'}^{1/r}$. Using that $p > N \geqslant r$, we find that

$$|q_{E'}^{1/r}|_\ell = |q_E|_\ell^{p^{n+1}/r} \leqslant \ell^{-(p/r)\cdot p^n} \leqslant \ell^{-1} < \ell^{-1/(\ell-1)},$$

and hence Lemma 3 shows that $\varphi_E(E'', \mu[N])$ is of infinite order. Then as usual Q_D differs from $\pm\varphi_E(w_r(y_D))$ by a torsion point. So Q_D is of infinite order.

Since the representation W_{p^n} is irreducible for $\mathrm{PGL}_2(\mathbb{Z}/p^{n+1}\mathbb{Z})$, we can show by induction that the rank of the group generated by higher self-points is $\dim(V_{(n)}) = p^n \cdot (p+1)$. □

Putting the previous two results together, we are able to show a corollary that holds for all but finitely many primes p.

Corollary 20. *Suppose $E/\mathbb{Q}$ is a semistable curve of conductor N not equal to* 30 *or* 210. *Let p be a prime such that $p > N$,* (*so it is of good reduction*), *and such that $\bar\rho_p : \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{PGL}_2(\mathbb{F}_p)$ is surjective. Let s be the rank of the group generated by self-points in $E(K_N)$. Then the higher self-points in $E(K_{p^{n+1}N})$ generate a group of rank $s \cdot (p+1) \cdot p^n$.*

Proof. Take $F = K_N$ in the previous theorems. We only have to show the condition that $\rho_{F,p}$ is surjective. It is enough to show that $\bar\rho_{F,p} : \mathrm{Gal}(\bar F/F) \to \mathrm{PGL}_2(\mathbb{F}_p)$ has all of $\mathrm{PSL}_2(\mathbb{F}_p)$ in its image, since the representation V_{p^n} will still have the same decomposition.

Let H_p be the group $\mathrm{Gal}(K_{Np}/K_N)$, that is, the image of $\bar\rho_{F,p}$. It is equal to the normal subgroup in $\mathrm{Gal}(K_p/\mathbb{Q}) \cong \mathrm{PGL}_2(\mathbb{F}_p)$ corresponding to the subextension $K_p/K_N \cap K_p$. Since $p > 11$ when $p > N$, we have that $\mathrm{PGL}_2(\mathbb{F}_p)$ has only three normal subgroups, namely itself, $\mathrm{PSL}_2(\mathbb{F}_p)$ and $\{1\}$. By the remark above, we only have to exclude that H_p is not trivial.

If H_p was trivial, then p, dividing the order of $\mathrm{PGL}_2(\mathbb{F}_p)$, would have to divide the order of G_N, which is a subgroup of $\mathrm{PGL}_2(\mathbb{Z}/N\mathbb{Z})$. But if $p > N$, then p cannot divide the order of $\mathrm{PGL}_2(\mathbb{Z}/N\mathbb{Z})$, except when $p = 3$ and $N = 2$, which cannot occur as a conductor. □

8. Derivatives

Let $E/\mathbb{Q}$ be an elliptic curve of conductor N. Let p be an *odd* prime of *ordinary*, either good or multiplicative, reduction. To treat the cases of higher self-points discussed in the Sections 7.2 and 7.3 simultaneously, we choose now a base field F. If E has good ordinary reduction at p, then F is any number field such that $E(F)$ contains a self-point P_C of infinite order. If p divides N, then F is a number field such that the absolute Galois group of F acts by scalars on $E[N/p]$.

We will suppose from now on that $\rho_{F,p} : \mathrm{Gal}(\bar F/F) \to \mathrm{PGL}_2(\mathbb{Z}_p)$ is surjective.

We suppose that F_n is the smallest extension of F such that the Galois group $H_n = \mathrm{Gal}(F_n/F)$ acts by scalars on $E[p^{n+1}]$. By assumption the map $\rho_{F,p}$ induces an isomorphism from H_n to $\mathrm{PGL}_2(\mathbb{Z}/p^{n+1}\mathbb{Z})$. Also, this implies that $E(F_n)$ has no p-torsion elements.

Let $\mathbb{O}$ be the ring of integers in the unramified quadratic extension of $\mathbb{Q}_p$. Choosing a basis of $\mathbb{O}$ over $\mathbb{Z}_p$ and viewing each element $u \in \mathbb{O}^\times$ as the ($\mathbb{Z}_p$-linear) multiplication by u on $\mathbb{O}$, we get a homomorphism

$$\Psi : \mathbb{O}^\times \to \mathrm{GL}_2(\mathbb{Z}_p) \to \mathrm{PGL}_2(\mathbb{Z}_p),$$

whose kernel is $\mathbb{Z}_p^\times$. The image of the composition

$$\mathbb{O}^\times \to \mathrm{PGL}_2(\mathbb{Z}_p) \to \mathrm{PGL}_2(\mathbb{Z}/p^{n+1}\mathbb{Z}) \to H_n$$

will be denoted by A_n. This is a cyclic group of order $(p+1)\cdot p^n = \#\mathbb{P}^1(\mathbb{Z}/p^{n+1}\mathbb{Z})$. It is the projective version of the nonsplit Cartan group in $\mathrm{GL}_2(\mathbb{Z}/p^{n+1}\mathbb{Z})$. To simplify notation, we will write F_n^A for the subfield of F_n fixed by A_n.

Theorem 21. *Let $E/\mathbb{Q}$ be an elliptic curve. Suppose E does not have potentially good supersingular reduction for any prime of additive reduction. Let p be a prime of either good ordinary or multiplicative reduction. Let F be the number field as above and assume that $\rho_{F,p}$ is surjective. Then we have $\#\,\mathrm{Sel}_{p^n}(E/F_n^A) \geqslant p^n$, where A is any nonsplit Cartan group in* $\mathrm{PGL}_2(\mathbb{Z}_p)$.

The proof of this theorem will be completed in Section 8.3.

Since there are no p-torsion points in $E(F_n)$, as $\rho_{F,p}$ is assumed to be surjective, there is an isomorphism $\mathrm{H}^1(F_n^A, E[p^k]) \to \mathrm{H}^1(F_n, E[p^k])^{A_n}$ induced by the restriction map. This implies that the map

$$\mathrm{Sel}_{p^n}(E/F_n^A) \to \mathrm{Sel}_{p^n}(E/F_n)^{A_n}$$

is injective. We conjecture that the elements in the Selmer group constructed in Theorem 21 do not lie in the image of the Kummer map, but represent nontrivial elements in the Tate–Shafarevich group Ш(E/F_n^A). If so, these classes in the Tate–Shafarevich group will capitulate in the extension F_n/F_n^A, since the elements of the Selmer group in the theorem restrict to elements in the image of the higher self-points inside $\mathrm{Sel}_{p^n}(E/F_n)$. It would be very interesting to verify this conjecture in some cases, but even for the smallest cases like $p = 11$ it seems completely impossible to compute the classes explicitly. Nevertheless it is natural to make this conjecture when comparing it to Kolyvagin's conjecture on the nontriviality of derivative classes of Heegner points (as investigated in [Jetchev et al. 2007]).

8.1. *The field extension.*

Lemma 22. *The cyclic group A_n intersects trivially any Borel subgroup in H_n.*

Proof. We prove the statement that the image of Ψ in $\mathrm{PGL}_2(\mathbb{Z}_p)$ intersects trivially any of its Borel subgroups B. Let L be the $\mathbb{Z}_p$-line in $\mathbb{O}$ such that B is the stabiliser under the action of $\mathrm{PGL}_2(\mathbb{Z}_p)$ on $\mathbb{P}^1(\mathbb{Z}_p)$ viewed as the set of $\mathbb{Z}_p$-modules in $\mathbb{O}$ generated by a unit. Let $\alpha \in \mathbb{O}^\times$ be any element with a nontrivial image under Ψ. Then $\alpha \notin \mathbb{Z}_p^\times$ cannot fix L. □

This implies in particular that any generator α_n of A_n acts simply transitively on the set $\mathbb{P}^1(\mathbb{Z}/p^{n+1}\mathbb{Z})$.

Lemma 23. *Let v be either a place of ordinary reduction above p or an infinite place or a place of potentially multiplicative reduction. Then the image of*

$$\bar\rho_{F_v,p} : \mathrm{Gal}(\bar F_v/F_v) \to \mathrm{PGL}_2(\mathbb{Z}/p^{n+1}\mathbb{Z})$$

lies in a Borel subgroup of $\mathrm{PGL}_2(\mathbb{Z}/p^{n+1}\mathbb{Z})$.

Proof. First suppose that v divides p. Since E is of ordinary reduction at v, there is a cyclic subgroup of $E[p^{n+1}]$ of order p^{n+1} that is fixed by the Galois group $\mathrm{Gal}(\bar F_v/F_v)$. This subgroup consists of all elements of $E[p^{n+1}]$ with trivial reduction over $\bar F_v$. Therefore the image of $\bar\rho_{F_v,p}$ is contained in the stabiliser of this point in $\mathbb{P}^1(\mathbb{Z}/p^{n+1}\mathbb{Z})$, which is a Borel subgroup.

Now, let v be a place of split multiplicative reduction for E. From the description of E as a Tate curve over F_v, we see that there is subgroup isomorphic to $\mu[p^{n+1}]$ inside $E[p^{n+1}]$. As before $\mathrm{Gal}(\bar F_v/F_v)$ will fix this subgroup and hence the image of $\bar\rho_{F_v,p}$ is contained in a Borel subgroup.

Next, we suppose that v is a place of bad reduction, but not of split multiplicative type. Then by hypothesis, E has either nonsplit multiplicative or additive and potentially multiplicative reduction. In both cases there exists a quadratic extension L of F_v, unramified in the first case and ramified in the second, such that E has split multiplicative reduction over L; see [Serre 1972, page 312]. Hence $E[p^{n+1}]$ can be described as the set of $\zeta^i \cdot a^j$, with ζ a primitive p^{n+1}-st root of unity, a a p^{n+1}-st root of the Tate-parameter q and $0 \leqslant i, j < p^{n+1}$; the action of $\sigma \in \mathrm{Gal}(\bar F_v/F_v)$ is given by $\sigma * (\zeta^i \cdot a^j) = \chi_L(\sigma)\cdot\sigma(\zeta)^i \cdot \sigma(a)^j$, where χ_L is the quadratic character associated to L/F_v. Therefore the subgroup generated by ζ is still fixed under $\mathrm{Gal}(\bar F_v/F_v)$.

Finally, we have to treat the case when v is an infinite place. But for any p, there is a cyclic subgroup of order p^{n+1} in $E(\mathbb{R})$; hence the image is contained in a Borel subgroup. □

Remark: We used here in a crucial way the assumption that p is a prime of ordinary reduction. Certainly it will not hold for places of additive reduction that are potentially supersingular.

Proposition 24. *Suppose that none of the primes of additive reduction for E are potentially good supersingular. Then then extension F_n/F_n^A is nowhere ramified. Moreover all places above ∞, p, and N split completely in this extension.*

Proof. Since F_n is a subfield of $F(E[p^\infty])$, it is unramified outside ∞, p, and N. By the previous lemma, the decomposition group of a place υ dividing $\infty \cdot p \cdot N$ in F inside H_n is contained in a Borel. Since any Borel intersects $A_n = \mathrm{Gal}(F_n/F_n^A)$ trivially by Lemma 22, the places above $\infty \cdot p \cdot N$ in F_n^A split completely. □

8.2. *The A-cohomology of the Steinberg representation.*

Let

$$V_n' = \{ f : \mathbb{P}^1(\mathbb{Z}/p^{n+1}\mathbb{Z}) \to \mathbb{Q} \mid \textstyle\sum_D f(D) = 0 \}$$

be the $\mathbb{Q}[H_n]$-module considered earlier in Section 7.2. It is a $\mathbb{Q}$-vector space of dimension $m-1$ with $m = (p+1) \cdot p^n$. There is a natural lattice T_n' in V_n' that is fixed by H_n, defined by

$$T_n' = \{ f : \mathbb{P}^1(\mathbb{Z}/p^{n+1}\mathbb{Z}) \to \mathbb{Z} \mid \textstyle\sum_D f(D) = 0 \}.$$

Lemma 25. $\mathrm{H}^1(A_n, T_n') = \mathbb{Z}/m\mathbb{Z}$.

Proof. The A_n-fixed part of V_n' is trivial, since A_n acts transitively on $\mathbb{P}^1(\mathbb{Z}/p^{n+1}\mathbb{Z})$: A function $f : \mathbb{P}^1(\mathbb{Z}/p^{n+1}\mathbb{Z}) \to \mathbb{Q}$ that is fixed by A_n would necessarily be constant, but then $\sum_D f(D) = 0$ implies that $f = 0$. Consider now the exact sequence

$$0 \to T_n' \to V_n' \to V_n'/T_n' \to 0$$

of H_n-modules, which induces an isomorphism $(V_n'/T_n')^{A_n} \to \mathrm{H}^1(A_n, T_n')$ since $\mathrm{H}^1(H_n, V_n') = 0$ as V_n' is divisible. So we are looking to determine the A_n-fixed functions in

$$V_n'/T_n' = \{ f : \mathbb{P}^1(\mathbb{Z}/p^{n+1}\mathbb{Z}) \to \mathbb{Q}/\mathbb{Z} \mid \textstyle\sum_D f(D) = 0 \}.$$

Such a function must be constant, since A_n acts transitively. Say $f(D) = f_0$. Then $m \cdot f_0 = 0$, so $f_0 \in (1/m)\mathbb{Z}$ gives the result. □

Proposition 26. *If U is any lattice in V_n' fixed by H_n, then $\#\mathrm{H}^1(A_n, U) = m$.*

Proof. The lattice U is contained in a scaled version of T_n' with finite index, say $0 \to U \to T_n' \to Z \to 0$. Since the Herbrand quotient[2] satisfies $h(A_n, Z) = 1$ for the finite A_n-module Z, we have $\#\mathrm{H}^1(A_n, U) = h(A_n, U) = h(A_n, T_n') = \#\mathrm{H}^1(A_n, T_n') = m$. □

It is not true in general that $\mathrm{H}^1(A_n, U)$ is cyclic. For $n = 0$, it can have up to three cyclic factors.

[2] We set $h(G, A) = \#\mathrm{H}^1(G, A)/\#\mathrm{H}^2(G, A)$ for a finite cyclic group G acting on a G-module A.

8.3. *Proof of Theorem 21.* We have an injection

$$\imath : V_n' \to E(F_n) \otimes \mathbb{Q}, \quad f \mapsto \textstyle\sum_D f(D) \cdot Q_D,$$

where Q_D is the higher self-point constructed in Sections 7.2 and 7.3. Let S_n be the saturated group generated by the higher self-points in $E(F_n)$, that is,

$$S_n = \bigl\{ P \in E(F_n) \bigm| \text{there is a } k > 0 \text{ such that } k \cdot P \in \mathbb{Z}[H_n] \cdot Q_D \bigr\}.$$

By definition all torsion points in $E(F_n)$ belong to S_n; moreover we have

$$0 \to E(F_n)_{\mathrm{tors}} \to S_n \to U_n \to 0,$$

where U_n can be identified as a H_n-stable lattice in the image of $\imath$. Because there are no A_n-fixed elements in U_n, we find

$$\begin{aligned} 0 \longrightarrow \mathrm{H}^1(A_n, E(F_n)_{\mathrm{tors}}) \longrightarrow \mathrm{H}^1(A_n, S_n) \longrightarrow \mathrm{H}^1(A_n, U_n) \\ \longrightarrow \mathrm{H}^2(A_n, E(F_n)_{\mathrm{tors}}) \longrightarrow \mathrm{H}^2(A_n, S_n) \longrightarrow 0. \end{aligned}$$

Since the Herbrand quotient $h(A_n, E(F_n)_{\mathrm{tors}})$ is trivial, we find

$$\begin{aligned} \#\mathrm{H}^1(A_n, S_n) &= \#\mathrm{H}^1(A_n, U_n) \cdot \#\mathrm{H}^1(A_n, S_n) \\ &\geqslant \#\mathrm{H}^1(A_n, U_n) = m = (p+1) \cdot p^n \end{aligned}$$

by Proposition 26. Note also that since $E(F_n)$ has no p-torsion points, we know that $\#\mathrm{H}^1(A_n, S_n)[p^n] = \#\mathrm{H}^1(A_n, U_n)[p^n] = p^n$. Consider the natural inclusion of S_n into $E(F_n)$. The cokernel of this inclusion Y_n is a free $\mathbb{Z}$-module. The long exact sequence

$$0 \longrightarrow E(F_n^A)_{\mathrm{tors}} \longrightarrow E(F_n^A) \longrightarrow {Y_n}^{A_n} \longrightarrow \mathrm{H}^1(A_n, S_n) \longrightarrow \mathrm{H}^1(A_n, E(F_n)) \quad (5)$$

shows that ${Y_n}^{A_n}$ has the same rank as $E(F_n^A)$.

Composing the last map in the above sequence with the inflation map will be called the *derivation map*

$$\partial_n : \mathrm{H}^1(A_n, S_n) \longrightarrow \mathrm{H}^1(A_n, E(F_n)) \xrightarrow{\mathrm{inf}} \mathrm{H}^1(F_n^A, E).$$

Since S_n has no p-torsion elements, we can identify the p^n-torsion part of the source with

$$\Bigl(\frac{S_n}{p^n\, S_n}\Bigr)^{A_n} \xrightarrow{\cong} \mathrm{H}^1(A_n, S_n)[p^n] \ ,$$

and therefore we call the image of ∂_n the *derived classes* of higher self-points.

Lemma 27. *The image of ∂_n is contained in* $Ш(E/F_n^A)$.

Proof. Let κ be the lift of an element in the image of ∂_n under the map

$$\mathrm{H}^1(F_n^A, E[m']) \longrightarrow \mathrm{H}^1(F_n^A, E)[m']$$

for a sufficiently large m'. Since the extension F_n/F_n^A is nonramified at a place v outside the set Σ of places in F_n^A above p, N or ∞, the restriction of κ to $\mathrm{H}^1(F_{n,v}^A, E[m'])$ will lie in $\mathrm{H}^1_f(F_n^A, E[m'])$. Now for any place v in Σ, the place v splits completely in extension F_n/F_n^A by Proposition 24. Therefore the restriction of κ to $\mathrm{H}^1(F_{n,v}^A, E)[m']$ is trivial since it comes from the inflation

$$\mathrm{H}^1(F_n/F_n^A, E(F_n)) \longrightarrow \mathrm{H}^1(F_n^A, E).$$

Hence κ belongs to the Selmer group within $\mathrm{H}^1(F_n^A, E[m'])$. □

We can now end the proof of Theorem 21. Denote by s the minimal number of generators of the kernel of ∂_n. From the long exact sequence (5), we see that the rank of ${Y_n}^{A_n}$ is at least s. So, if ∂_n is not injective, then $\mathrm{rank}(E(F_n^A))$ is positive. So either the image of ∂, lifted to the Selmer group, will contribute p^n elements or else $E(F_n^A)$ will give rise to a copy of $\mathbb{Z}/_{p^n\mathbb{Z}}$ in $\mathrm{Sel}_{p^n}(E/F_n^A)$. □

We add here a comment on the case when E has supersingular reduction at p. It turns out that construction of derivative classes in $\mathrm{H}^1(F_n^A, E)$ using higher self-points works the same, provided that the higher self-points are of infinite order. The main difference is that the cohomology classes do not belong to the Tate–Shafarevich group. In fact, under the assumption that the derivative map is not trivial, they will provide classes that are orthogonal to elements from the Selmer group and could be used to bound the Selmer group from above, just like Kolyvagin's classes built from Heegner points. Unfortunately we do not know a way of proving the assumption; hence these derivative classes cannot be used to say something about the Selmer group.

8.4. ***Derivative of self-points.*** Besides constructing derivative classes of higher self-points, we can also produce cohomology classes from self-points. We only sketch here the results whose proofs are similar to the previous sections.

Let $E/\mathbb{Q}$ be an elliptic curve of conductor N. Assume for simplicity that $N = p$ is prime. Put $K = K_p$. It is known that ρ_p is surjective; for more details see [Delaunay and Wuthrich 2008]. So the Galois group $G = \mathrm{Gal}(K/\mathbb{Q})$ is isomorphic to $\mathrm{PGL}_2(\mathbb{F}_p)$. Let A be any cyclic subgroup of order $p+1$ in G.

Theorem 28. *There is map ∂ to the Tate–Shafarevich group $Ш(E/K^A)$ from a group of order at least $p+1$. If r is the difference of the rank of $E(\mathbb{Q}(C))$ and $E(\mathbb{Q})$, then*

$$\#\,\mathrm{Sel}_{p+1}(E/K^A) \geqslant (p+1)^r \cdot \#E(\mathbb{Q})[p+1].$$

As before we consider the saturation of the self-points S in $E(K)$. We know that S modulo its torsion part is a lattice U in the Steinberg representation of $\mathrm{PGL}_2(\mathbb{F}_p)$. As we have seen in Section 8.2, the cohomology group $\mathrm{H}^1(A, U)$ will have $p+1$ elements. In [Delaunay and Wuthrich 2008, Section 4], we computed the torsion subgroup of $E(K)$. Using this we obtain that $E(K^A)_{\mathrm{tors}} = E(\mathbb{Q})_{\mathrm{tors}}$ and

$$\mathrm{H}^1(A, E(K)_{\mathrm{tors}}) = \mathrm{H}^2(A, E(K)_{\mathrm{tors}}) = \begin{cases} \mathbb{Z}/2\mathbb{Z}, \\ 0, \end{cases}$$

the nontrivial case occurring exactly when E is one of the curves 17a2, 17a3, 17a4 or any Neumann–Setzer curve. As before, this shows that $\mathrm{H}^1(A, S)$ has either $p+1$ or $2\,(p+1)$ elements. The derivative map is again

$$\partial : \mathrm{H}^1(A, S) \longrightarrow \mathrm{H}^1(A, E(K)) \longrightarrow \mathrm{H}^1(K^A, E),$$

and its image is in the Tate–Shafarevich group $Ш(E/K^A)$. Denote by Y the quotient of $E(K)$ by S. Then $\ker \partial$ is the quotient of Y^A by $E(K^A)$. If this map ∂ is not injective, then there is a $y \in Y^G$, lifting to a point of infinite order $Q \in E(K)$, such that Q does not belong to $E(K^A)$ but a nonzero multiple of it does. So either ∂ is surjective or there are points of infinite order defined over K^A that only become divisible in $E(K)$.

We should add that the control theorem for the Selmer group is not necessarily perfect; the kernel of $\mathrm{Sel}_{p+1}(E/K^A) \to \mathrm{Sel}_{p+1}(E/K)$ can be of order 1 or 2.

It is also worth adding another particular property of K^A: the L-series of E over K^A is the product of $\prod_\rho L(E, \rho, s)$, where ρ runs over all distinct irreducible representations of $\mathrm{PGL}_2(\mathbb{F}_p)$ except the Steinberg representation and the nontrivial 1-dimensional representation. It is not known whether this L-series admits analytic continuation.

Acknowledgments

It is a pleasure to thank John Coates, Henri Darmon, Christophe Delaunay, Ralph Greenberg, Masato Kurihara, and Marusia Rebolledo for helpful and interesting discussions concerning self-points.

References

[Atkin and Lehner 1970] A. O. L. Atkin and J. Lehner, "Hecke operators on $\Gamma_0(m)$", *Math. Ann.* **185** (1970), 134–160. MR 42 #3022 Zbl 0177.34901

[Bertolini, Darmon and Prasanna ≥ 2009] M. Bertolini, H. Darmon, and K. Prasanna, "Exotic Heegner points", in preparation.

[Breuil et al. 2001] C. Breuil, B. Conrad, F. Diamond, and R. Taylor, "On the modularity of elliptic curves over $\mathbb{Q}$: Wild 3-adic exercises", *J. Amer. Math. Soc.* **14**:4 (2001), 843–939. MR 2002d:11058 Zbl 0982.11033

[Coates et al. 2005] J. Coates, T. Fukaya, K. Kato, R. Sujatha, and O. Venjakob, "The GL_2 main conjecture for elliptic curves without complex multiplication", *Publ. Math. Inst. Hautes Études Sci.* 101 (2005), 163–208. MR 2007b:11172

[Coates et al. 2009] J. Coates, T. Fukaya, K. Kato, and R. Sujatha, "Root numbers, Selmer groups, and non-commutative Iwasawa theory", *J. Algebraic Geom.* (2009).

[Cremona 1997] J. E. Cremona, *Algorithms for modular elliptic curves*, 2nd ed., Cambridge University Press, 1997. MR 99e:11068 Zbl 0872.14041

[Delaunay and Wuthrich 2008] C. Delaunay and C. Wuthrich, "Self-points on elliptic curves of prime conductor", preprint, 2008. To appear in *Internat. J. Number Theory*.

[Edixhoven 1991] B. Edixhoven, "On the Manin constants of modular elliptic curves", pp. 25–39 in *Arithmetic algebraic geometry* (Texel, 1989), edited by G. van der Geer et al., Progr. Math. **89**, Birkhäuser, Boston, 1991. MR 92a:11066 Zbl 0749.14025

[Greenberg 2008] R. Greenberg, "Iwasawa theory, projective modules, and modular representations", preprint, 2008, Available at http://www.math.washington.edu/~greenber/NewMod.pdf.

[Harris 1979] M. Harris, "Systematic growth of Mordell–Weil groups of abelian varieties in towers of number fields", *Invent. Math.* **51**:2 (1979), 123–141. MR 80i:14015 Zbl 0429.14013

[Jetchev et al. 2007] D. Jetchev, K. Lauter, and W. Stein, "Explicit Heegner points: Kolyvagin's conjecture and nontrivial elements in the Shafarevich–Tate group", preprint, 2007. arXiv 0707.0032

[Katz and Mazur 1985] N. M. Katz and B. Mazur, *Arithmetic moduli of elliptic curves*, Annals of Mathematics Studies **108**, Princeton University Press, 1985. MR 86i:11024 Zbl 0576.14026

[Kolyvagin 1990] V. A. Kolyvagin, "Euler systems", pp. 435–483 in *The Grothendieck Festschrift, II*, edited by P. Cartier et al., Progr. Math. **87**, Birkhäuser, 1990. MR 92g:11109 Zbl 0742.14017

[Mazur 1978] B. Mazur, "Rational isogenies of prime degree", *Invent. Math.* **44**:2 (1978), 129–162. MR 80h:14022 Zbl 0386.14009

[Mazur and Rubin 2008] B. Mazur and K. Rubin, "Growth of Selmer rank in nonabelian extensions of number fields", *Duke Math. J.* **143**:3 (2008), 437–461. MR 2423759 Zbl 1151.11023

[Serre 1968] J.-P. Serre, *Abelian l-adic representations and elliptic curves*, Benjamin, New York, 1968. MR 41 #8422 Zbl 0186.25701

[Serre 1972] J.-P. Serre, "Propriétés galoisiennes des points d'ordre fini des courbes elliptiques", *Invent. Math.* **15**:4 (1972), 259–331. MR 52 #8126 Zbl 0235.14012

[Serre 1996] J.-P. Serre, "Travaux de Wiles (et Taylor, ...), I", pp. 319–332 in *Séminaire Bourbaki, 1994/95*, Astérisque **237,** Exposé 803, Société Math. de France, Paris, 1996. MR 97m:11076 Zbl 0957.11027

[Silberger 1970] A. J. Silberger, PGL_2 *over the p-adics: its representations, spherical functions, and Fourier analysis*, Lecture Notes in Mathematics **166**, Springer, Berlin, 1970. MR 44 #2891 Zbl 0204.44102

[Wuthrich 2007] C. Wuthrich, "Self-points on an elliptic curve of conductor 14", pp. 189–195 in *Proceedings of the Symposium on Algebraic Number Theory and Related Topics*, edited by K. Hashimoto, RIMS Kôkyûroku Bessatsu **B4**, Res. Inst. Math. Sci., Kyoto, 2007. MR 2402010 Zbl 05258116

Communicated by Karl Rubin
Received 2008-06-10 Revised 2009-02-23 Accepted 2009-02-24

christian.wuthrich@nottingham.ac.uk
School of Mathematical Sciences, University of Nottingham,
Nottingham NG7 2RD, United Kingdom

ALGEBRA AND NUMBER THEORY 3:3(2009)

Weyl groupoids of rank two and continued fractions

Michael Cuntz and István Heckenberger

We present a relationship between continued fractions and Weyl groupoids of Cartan schemes of rank two. This allows one to decide easily if a given Cartan scheme of rank two admits a finite root system. We obtain obstructions and sharp bounds for the entries of the Cartan matrices.

1. Introduction

Root systems and crystallographic Coxeter groups are key tools in the study of semisimple Lie algebras [Bourbaki 1968]. In the structure theory of pointed Hopf algebras [Montgomery 1993] a similar role is expected to be played by Weyl groupoids and their root systems. Let us give some hints towards this claim. The most striking results on pointed Hopf algebras rely on the lifting method of Andruskiewitsch and Schneider [1998]. Based on it, many new examples of finite-dimensional pointed Hopf algebras have been detected, and fairly general classification results were achieved [Andruskiewitsch and Schneider 2005; Heckenberger 2009]. The first step in the lifting method is the determination of finite-dimensional Nichols algebras of finite group type. The upper triangular part of a small quantum group, also called Frobenius–Lusztig kernel, is a prominent example. A very natural symmetry object of Nichols algebras of finite group type is the Weyl groupoid. This was observed first in [Heckenberger 2006] for Nichols algebras of diagonal type, and then in [Andruskiewitsch et al. 2008] in a very general setting.

An axiomatic approach to Weyl groupoids and their root systems, without referring to Nichols algebras, was initiated in [Heckenberger and Yamane 2008]. The theory includes and extends the theory of crystallographic Coxeter groups, but contains even such examples which do not seem to be related to Nichols algebras of diagonal type. In this paper we use the language and some structural and classification results achieved in [Cuntz and Heckenberger 2008]; see Section 2 for the most essential definitions and facts.

MSC2000: primary 20F55; secondary 11A55, 16W30.

Keywords: Cartan matrix, continued fraction, Nichols algebra, Weyl groupoid.

Heckenberger is supported by the German Research Foundation (DFG) via a Heisenberg fellowship.

For the classification of Nichols algebras of diagonal type it is crucial to be able to decide whether a given Cartan scheme (a categorical generalization of the notion of a generalized Cartan matrix; see Definition 2.1) admits a finite root system. Because of the large variety of examples, this seems to be a difficult task. In our paper, we present a very efficient method for Cartan schemes of rank two. It relies on a relationship between Cartan schemes of rank two and continued fractions [Perron 1929]. Instead of giving a complete list of Cartan schemes of rank two admitting a finite root system (which is then unique by a result in [Cuntz and Heckenberger 2008]), we present an algorithm in Theorem 6.19. It works with very elementary operations on sequences of positive integers and transforms any Cartan scheme into another one, for which the answer is known. The algorithm is based on various observations: on the introduction and study of coverings of Cartan schemes in Section 3, on an old theorem of Stern, Pringsheim, and Tietze, and a variation of a transformation formula for continued fractions (Section 4 and Lemma 5.2) on the characterization of simple connected Cartan schemes admitting a finite root system in terms of certain sequences of positive integers (Proposition 6.5 and Theorem 6.6), and on the description of Cartan schemes with object change diagram a cycle using characteristic sequences (Definition 6.9). As an application, in Section 7 we give obstructions for the entries of the Cartan matrices in a Cartan scheme admitting a finite root system. We present the power of our method on a small example at the end of Section 6.

We are confident that a suitable generalization of our method to Cartan schemes and Weyl groupoids of higher rank would have a deep impact on the classification of Nichols algebras, and consider it as a great challenge for the future.

2. Cartan schemes, root systems, and their Weyl groupoids

If not stated otherwise, we follow the notation in [Cuntz and Heckenberger 2008]. Let us start by recalling the main definitions.

Let I be a nonempty finite set and $\{\alpha_i \mid i \in I\}$ the standard basis of $\mathbb{Z}^I$. By [Kac 1990, §1.1] a generalized Cartan matrix $C = (c_{ij})_{i,j\in I}$ is a matrix in $\mathbb{Z}^{I\times I}$ such that

(M1) $c_{ii} = 2$ and $c_{jk} \leq 0$ for all $i, j, k \in I$ with $j \neq k$,

(M2) if $i, j \in I$ and $c_{ij} = 0$, then $c_{ji} = 0$.

Definition 2.1. Let A be a nonempty set, $\rho_i : A \to A$ a map for all $i \in I$, and $C^a = (c^a_{jk})_{j,k\in I}$ a generalized Cartan matrix in $\mathbb{Z}^{I\times I}$ for all $a \in A$. The quadruple

$$\mathcal{C} = \mathcal{C}(I, A, (\rho_i)_{i\in I}, (C^a)_{a\in A})$$

is called a *Cartan scheme* if

(C1) $\rho_i^2 = \mathrm{id}$ for all $i \in I$,

(C2) $c_{ij}^a = c_{ij}^{\rho_i(a)}$ for all $a \in A$ and $i, j \in I$.

Remark 2.2. The preceding definition of a Cartan scheme has the striking advantage to be very simple, but sufficiently powerful to admit the definition of a Weyl groupoid, as we will see below. For some investigations it can be of advantage to consider more general axioms (for example by allowing the maps ρ_i to be partially defined) or to impose additional restrictions (like (C3) below, or other for example to exclude the existence of associated roots which are neither positive nor negative). We will mostly consider Cartan schemes admitting a root system. This restriction still gives many more examples than those coming from contragredient Lie superalgebras and Nichols algebras of diagonal type with finite root system. Nevertheless, up to now no further axioms on Cartan schemes are known which keep this property.

Two Cartan schemes

$$\mathcal{C} = \mathcal{C}(I, A, (\rho_i)_{i\in I}, (C^a)_{a\in A}) \quad \text{and} \quad \mathcal{C}' = \mathcal{C}'(I', A', (\rho_i')_{i\in I'}, (C'^a)_{a\in A'})$$

are termed *equivalent* if there are bijections $\varphi_0 : I \to I'$ and $\varphi_1 : A \to A'$ such that

$$\varphi_1(\rho_i(a)) = \rho'_{\varphi_0(i)}(\varphi_1(a)), \qquad c^{\varphi_1(a)}_{\varphi_0(i)\varphi_0(j)} = c^a_{ij} \tag{2-1}$$

for all $i, j \in I$ and $a \in A$.

Let $\mathcal{C} = \mathcal{C}(I, A, (\rho_i)_{i\in I}, (C^a)_{a\in A})$ be a Cartan scheme. For all $i \in I$ and $a \in A$ define $\sigma_i^a \in \mathrm{Aut}(\mathbb{Z}^I)$ by

$$\sigma_i^a(\alpha_j) = \alpha_j - c_{ij}^a \alpha_i \qquad \text{for all } j \in I. \tag{2-2}$$

The *Weyl groupoid of* $\mathcal{C}$ is the category $\mathcal{W}(\mathcal{C})$ such that $\mathrm{Ob}(\mathcal{W}(\mathcal{C})) = A$ and the morphisms are generated by the maps $\sigma_i^a \in \mathrm{Hom}(a, \rho_i(a))$ with $i \in I$, $a \in A$. In this paper, we will always denote the set of all morphisms of $\mathcal{W}(\mathcal{C})$ by $\mathrm{Hom}(\mathcal{W}(\mathcal{C}))$. Formally, for $a, b \in A$ the set $\mathrm{Hom}(a, b)$ consists of the triples (b, f, a), where

$$f = \sigma_{i_n}^{\rho_{i_{n-1}}\cdots\rho_{i_1}(a)} \cdots \sigma_{i_2}^{\rho_{i_1}(a)} \sigma_{i_1}^a$$

and $b = \rho_{i_n} \cdots \rho_{i_2}\rho_{i_1}(a)$ for some $n \in \mathbb{N}_0$ and $i_1, \dots, i_n \in I$. The composition is induced by the group structure of $\mathrm{Aut}(\mathbb{Z}^I)$:

$$(a_3, f_2, a_2) \circ (a_2, f_1, a_1) = (a_3, f_2 f_1, a_1)$$

for all $(a_3, f_2, a_2), (a_2, f_1, a_1) \in \mathrm{Hom}(\mathcal{W}(\mathcal{C}))$. By abuse of notation we will write $f \in \mathrm{Hom}(a, b)$ instead of $(b, f, a) \in \mathrm{Hom}(a, b)$.

The cardinality of I is termed the *rank of* $\mathcal{W}(\mathcal{C})$. A Cartan scheme is called *connected* if its Weyl groupoid is connected, that is, if for all $a, b \in A$ there exists $w \in \mathrm{Hom}(a, b)$.

In many cases it will be natural to assume that a Cartan scheme satisfies the following additional property.

(C3) If $a, b \in A$ and $(b, \mathrm{id}, a) \in \mathrm{Hom}(a, b)$, then $a = b$.

Definition 2.3. Let $\mathcal{C} = \mathcal{C}(I, A, (\rho_i)_{i\in I}, (C^a)_{a\in A})$ be a Cartan scheme. For all $a \in A$ let $R^a \subset \mathbb{Z}^I$, and define $m^a_{i,j} = |R^a \cap (\mathbb{N}_0\alpha_i + \mathbb{N}_0\alpha_j)|$ for all $i, j \in I$ and $a \in A$. We say that

$$\mathcal{R} = \mathcal{R}(\mathcal{C}, (R^a)_{a\in A})$$

is a *root system of type* $\mathcal{C}$ if it satisfies the following axioms.

(R1) $R^a = R^a_+ \cup -R^a_+$, where $R^a_+ = R^a \cap \mathbb{N}^I_0$, for all $a \in A$.

(R2) $R^a \cap \mathbb{Z}\alpha_i = \{\alpha_i, -\alpha_i\}$ for all $i \in I$, $a \in A$.

(R3) $\sigma^a_i(R^a) = R^{\rho_i(a)}$ for all $i \in I$, $a \in A$.

(R4) If $i, j \in I$ and $a \in A$ such that $i \neq j$ and $m^a_{i,j}$ is finite, then $(\rho_i\rho_j)^{m^a_{i,j}}(a) = a$.

If $\mathcal{R}$ is a root system of type $\mathcal{C}$, then we say that $\mathcal{W}(\mathcal{R}) = \mathcal{W}(\mathcal{C})$ is the *Weyl groupoid of* $\mathcal{R}$. Further, $\mathcal{R}$ is called *connected* if $\mathcal{C}$ is a connected Cartan scheme. If $\mathcal{R} = \mathcal{R}(\mathcal{C}, (R^a)_{a\in A})$ is a root system of type $\mathcal{C}$ and $\mathcal{R}' = \mathcal{R}'(\mathcal{C}', (R'^a)_{a\in A'})$ is a root system of type $\mathcal{C}'$, then we say that $\mathcal{R}$ and $\mathcal{R}'$ are *equivalent* if $\mathcal{C}$ and $\mathcal{C}'$ are equivalent Cartan schemes given by maps $\varphi_0 : I \to I'$, $\varphi_1 : A \to A'$ as in Definition 2.1, and if the map $\varphi_0^* : \mathbb{Z}^I \to \mathbb{Z}^{I'}$ given by $\varphi_0^*(\alpha_i) = \alpha_{\varphi_0(i)}$ satisfies $\varphi_0^*(R^a) = R'^{\varphi_1(a)}$ for all $a \in A$.

There exist many interesting examples of root systems of type $\mathcal{C}$ related to semisimple Lie algebras, Lie superalgebras and Nichols algebras of diagonal type, respectively. For further details and results we refer to [Heckenberger and Yamane 2008] and [Cuntz and Heckenberger 2008].

Convention 2.4. In connection with Cartan schemes, upper indices usually refer to elements of A. Often, these indices will be omitted if they are uniquely determined by the context.

Remark 2.5. If $\mathcal{C}$ is a Cartan scheme and there exists a root system of type $\mathcal{C}$, then $\mathcal{C}$ satisfies (C3) by [Heckenberger and Yamane 2008, Lemma 8(iii)].

Definition 4.3 of [Cuntz and Heckenberger 2008] introduced the concept of an *irreducible* root system of type $\mathcal{C}$. By Proposition 4.6 of the same paper, if $\mathcal{C}$ is a connected Cartan scheme and $\mathcal{R}$ is a finite root system of type $\mathcal{C}$, then $\mathcal{R}$ is irreducible if and only if the generalized Cartan matrix C^a is indecomposable for one (equivalently, for all) $a \in A$.

Here is a fundamental result about Weyl groupoids.

Theorem 2.6 [Heckenberger and Yamane 2008, Theorem 1]. *Let $\mathcal{C} = \mathcal{C}(I, A, (\rho_i)_{i\in I}, (C^a)_{a\in A})$ be a Cartan scheme and $\mathcal{R} = \mathcal{R}(\mathcal{C}, (R^a)_{a\in A})$ a root system of type $\mathcal{C}$. Let $\mathcal{W}$ be the abstract groupoid with* $\mathrm{Ob}(\mathcal{W}) = A$ *such that* $\mathrm{Hom}(\mathcal{W})$ *is generated by abstract morphisms $s_i^a \in \mathrm{Hom}(a, \rho_i(a))$, where $i \in I$ and $a \in A$, satisfying the relations*

$$s_i s_i 1_a = 1_a, \quad (s_j s_k)^{m^a_{j,k}} 1_a = 1_a, \qquad a \in A,\ i, j, k \in I,\ j \neq k$$

(see Convention 2.4). *Here 1_a is the identity of the object a, and $(s_j s_k)^\infty 1_a$ is understood to be 1_a. The functor $\mathcal{W} \to \mathcal{W}(\mathcal{R})$, which is the identity on the objects, and on the set of morphisms is given by $s_i^a \mapsto \sigma_i^a$ for all $i \in I$, $a \in A$, is an isomorphism of groupoids.*

Definition 2.7. Let $\mathcal{C} = \mathcal{C}(I, A, (\rho_i)_{i\in I}, (C^a)_{a\in A})$ be a Cartan scheme. Let Γ be a nondirected graph such that the vertices of Γ correspond to the elements of A. Assume that for all $i \in I$ and $a \in A$ with $\rho_i(a) \neq a$ there is precisely one edge between the vertices a and $\rho_i(a)$ with label i, and all edges of Γ are given in this way. The graph Γ is called the *object change diagram* of $\mathcal{C}$. If $\mathcal{R} = \mathcal{R}(\mathcal{C}, (R^a)_{a\in A})$ is a root system of type $\mathcal{C}$, then we also say that Γ is the object change diagram of $\mathcal{R}$.

3. Coverings of Cartan schemes, Weyl groupoids, and root systems

Two Cartan schemes can be related to each other in different ways. In this section we analyze coverings of Cartan schemes. The definition is motivated by the corresponding notion in topology.

Definition 3.1. Let

$$\mathcal{C} = \mathcal{C}(I, A, (\rho_i)_{i\in I}, (C^a)_{a\in A}) \quad \text{and} \quad \mathcal{C}' = \mathcal{C}'(I, A', (\rho'_i)_{i\in I}, (C'^a)_{a\in A'})$$

be connected Cartan schemes. Let $\pi : A' \to A$ be a map such that $C^{\pi(a)} = C'^a$ for all $a \in A'$ and the diagrams

$$\begin{tikzcd} A' \arrow[r, "\rho'_i"] \arrow[d, "\pi"'] & A' \arrow[d, "\pi"] \\ A \arrow[r, "\rho_i"'] & A \end{tikzcd} \tag{3-1}$$

commute for all $i \in I$. We say that $\pi : \mathcal{C}' \to \mathcal{C}$ is a *covering*, and that $\mathcal{C}'$ is a *covering of* $\mathcal{C}$.

The composition of two coverings is again one. For any covering $\pi : \mathscr{C}' \to \mathscr{C}$ of Cartan schemes $\mathscr{C}', \mathscr{C}$, the map $\pi : A' \to A$ is surjective by (3-1), since A' is nonempty and $\mathscr{C}$ is connected.

Remark 3.2. Many of the following results can be formulated without assuming that $\mathscr{C}$ and/or $\mathscr{C}'$ in Definition 3.1 are connected Cartan schemes. In that case one should assume that π is a surjective map. However, in the applications we are interested in, all Cartan schemes are connected, and hence we prefer the above definition in order to simplify the terminology.

Any covering $\pi : \mathscr{C}' \to \mathscr{C}$ of Cartan schemes $\mathscr{C}', \mathscr{C}$ induces a covariant functor $F_\pi : \mathscr{W}(\mathscr{C}') \to \mathscr{W}(\mathscr{C})$ by letting

$$F_\pi(a') = \pi(a'), \quad F_\pi(\sigma_i^{a'}) = \sigma_i^{\pi(a')} \qquad \text{for all } i \in I,\ a' \in A'.$$

In this case the Weyl groupoid $\mathscr{W}(\mathscr{C}')$ is termed a *covering of* $\mathscr{W}(\mathscr{C})$, and the functor F_π a covering of Weyl groupoids.

First we need a technical result.

Lemma 3.3. *Let $\pi : \mathscr{C}' \to \mathscr{C}$ be a covering, and assume that $\mathscr{C}'$ satisfies Axiom* (C3).

(1) *$\mathscr{C}$ satisfies* (C3).

(2) *Let $a \in A$ and $a', a'' \in A'$ such that $\pi(a') = \pi(a'') = a$. If there exists $w' \in$* $\mathrm{Hom}(a', a'')$ *such that $F_\pi(w') \in F_\pi(\mathrm{End}(a'))$, then $a' = a''$.*

Proof. (1) Let $a \in A$. If $k \in \mathbb{N}_0$ and $i_1, \dots, i_k \in I$, then Definition 3.1 gives that $\sigma_{i_1} \cdots \sigma_{i_{k-1}} \sigma_{i_k}^a = \sigma_{i_1} \cdots \sigma_{i_{k-1}} \sigma_{i_k}^{a'}$ in $\mathrm{Aut}(\mathbb{Z}^I)$ for all $a' \in A'$ with $\pi(a') = a$. Assume now that $\sigma_{i_1} \cdots \sigma_{i_{k-1}} \sigma_{i_k}^a = \mathrm{id}$. Then $\rho'_{i_1} \cdots \rho'_{i_k}(a') = a'$ for all $a' \in A'$ with $\pi(a') = a$, since $\mathscr{C}'$ satisfies (C3). Hence $\rho_{i_1} \cdots \rho_{i_k}(a) = a$ by (3-1). This yields the claim.

(2) Let $w'' \in \mathrm{End}(a')$ with $F_\pi(w'') = F_\pi(w')$. Then $F_\pi(w' w''^{-1}) = \mathrm{id}_a$, and hence $w' w''^{-1} = \mathrm{id}$ in $\mathrm{Aut}(\mathbb{Z}^I)$. Since $\mathscr{C}'$ satisfies (C3), it follows that $w' w''^{-1} = \mathrm{id}_{a'}$, and hence $a' = a''$. □

Let $\mathscr{C} = \mathscr{C}(I, A, (\rho_i)_{i \in I}, (C^a)_{a \in A})$ be a connected Cartan scheme, $\mathscr{W}(\mathscr{C})$ its Weyl groupoid, and $a \in A$. Coverings of $\mathscr{C}$ can be parametrized by subgroups of $\mathrm{End}(a) \subset \mathrm{Hom}(\mathscr{W}(\mathscr{C}))$ (up to conjugation).

Proposition 3.4. (1) *Let $\mathscr{C}'$ be a connected Cartan scheme and assume that $\pi : \mathscr{C}' \to \mathscr{C}$ is a covering. Let $a' \in A'$ with $\pi(a') = a$.*

(a) *The group homomorphism $F_\pi : \mathrm{End}(a') \to \mathrm{End}(a)$ is injective.*

(b) *For each $b' \in A'$ with $\pi(b') = a$ the subgroup $F_\pi(\mathrm{End}(b'))$ of* $\mathrm{End}(a)$ *is conjugate to $F_\pi(\mathrm{End}(a'))$.*

(c) *If U' is a subgroup of* $\mathrm{End}(a)$ *conjugate to $F_\pi(\mathrm{End}(a'))$, then there exists $b' \in A'$ with $\pi(b') = a$ and $F_\pi(\mathrm{End}(b')) = U'$.*

(2) *Suppose that* $U \subset \mathrm{End}(a)$ *is a subgroup. There exists a covering* $\pi : \mathscr{C}' \to \mathscr{C}$ *and* $b' \in A'$ *such that*

$$F_\pi(\mathrm{End}(b')) = U, \tag{3-2}$$

$$|\pi^{-1}(b)| = [\mathrm{End}(a) : U] \quad \textit{for all } b \in A. \tag{3-3}$$

If $\mathscr{C}$ *satisfies Axiom* (C3), *then up to equivalence there is a unique covering* $\mathscr{C}'$ *satisfying* (3-2) *and Axiom* (C3). *For this covering* (3-3) *holds.*

Proof. (1A) Each element $w' \in \mathrm{End}(a')$ is a product of $\sigma_i^{b'}$ for some $i \in I$ and $b' \in A'$. Moreover, w' can be naturally regarded as an element in $\mathrm{Aut}(\mathbb{Z}^I)$. The same is true for $w \in \mathrm{End}(a)$. Since $C'^{b'} = C^{\pi(b')}$ for all $b' \in A'$, $F_\pi(w')$ identifies with the same element of $\mathrm{Aut}(\mathbb{Z}^I)$ as w'.

(1B) Let $b' \in A'$. Since $\mathscr{C}'$ is connected, there exists $w' \in \mathrm{Hom}(a', b')$. Then $\mathrm{End}(b') = w' \,\mathrm{End}(a') w'^{-1}$. Since F_π is a functor,

$$F_\pi(\mathrm{End}(b')) = F_\pi(w') F_\pi(\mathrm{End}(a')) F_\pi(w')^{-1}.$$

(1C) Assume that $w \in \mathrm{End}(a)$ such that $U' = w F_\pi(\mathrm{End}(a'))w^{-1}$. Then $w = \sigma_{i_1} \cdots \sigma_{i_{k-1}} \sigma_{i_k}^a$ for some $k \in \mathbb{N}_0$ and $i_1, \dots, i_k \in I$. Let $w' = \sigma_{i_1} \cdots \sigma_{i_{k-1}} \sigma_{i_k}^{a'}$ and $b' = \rho'_{i_1} \cdots \rho'_{i_k}(a')$. Then $\mathrm{End}(b') = w' \,\mathrm{End}(a') w'^{-1}$, and hence $F_\pi(\mathrm{End}(b')) = w F_\pi(\mathrm{End}(a')) w^{-1} = U'$.

(2) We construct $\mathscr{C}'$ explicitly. Let

$$A' = \mathrm{Hom}(\mathscr{W}(\mathscr{C}))/U = \{gU \subset \mathrm{Hom}(a, b) \mid b \in A, g \in \mathrm{Hom}(a, b)\}$$

be the set of left cosets. For all $i \in I$ and $gU \in A'$ with $g \in \mathrm{Hom}(a, b)$, where $b \in A$, define $C'^{gU} = C^b$ and $\rho'_i(gU) = \sigma_i^b gU$. Then $\rho'_i : A' \to A'$ satisfies (C1) since $\sigma_i^{\rho_i(b)} \sigma_i^b = \mathrm{id}$ and $\rho_i^2 = \mathrm{id}$, and $\mathscr{C}'$ fulfills (C3), since $\mathscr{C}$ does. Since $\mathscr{C}$ is connected, $\mathscr{C}' = \mathscr{C}'(I, A', (\rho'_i)_{i \in I}, (C'^{a'})_{a' \in A'})$ is a connected Cartan scheme. Define $\pi : A' \to A$ by $\pi(gU) = b$ for all $b \in A$, $g \in \mathrm{Hom}(a, b)$. Then $F_\pi(\mathrm{End}(1_a U)) = U$ and $|\pi^{-1}(a)| = [\mathrm{End}(a) : U]$. Since $\mathscr{C}'$ is connected, $|\pi^{-1}(b)| = |\pi^{-1}(a)|$ for all $b \in A$.

Assume that $\mathscr{C}$ satisfies (C3). We show that $\mathscr{C}'$ satisfies (C3). For $l \in \{1, 2\}$ let $a_l \in A$ and $g_l \in \mathrm{Hom}(a, a_l)$ such that $(g_1 U, \mathrm{id}, g_2 U) \in \mathrm{Hom}(\mathscr{W}(\mathscr{C}'))$. Then there exist $k \in \mathbb{N}_0$ and $i_1, \dots, i_k \in I$ such that $\sigma_{i_1} \cdots \sigma_{i_{k-1}} \sigma_{i_k}^{a_2} g_2 U = g_1 U$ and that $\sigma_{i_1} \cdots \sigma_{i_{k-1}} \sigma_{i_k}^{a_2} = \mathrm{id}$ in $\mathrm{Aut}(\mathbb{Z}^I)$. Since $\mathscr{C}$ fulfills (C3), we obtain that $a_1 = a_2$, and hence $g_2 U = g_1 U$. Therefore $\mathscr{C}'$ satisfies (C3).

Finally, let $\pi : \mathscr{C}' \to \mathscr{C}$ and $\pi'' : \mathscr{C}'' \to \mathscr{C}$ be coverings of $\mathscr{C}$ satisfying (C3), and assume that there exist $b' \in A'$, $b'' \in A''$ such that $\pi(b') = \pi''(b'') = a$ and $F_\pi(\mathrm{End}(b')) = F_{\pi''}(\mathrm{End}(b'')) = U$. We have to show that $\mathscr{C}'$ and $\mathscr{C}''$ are equivalent Cartan schemes. Define $\phi : A' \to A''$ by

$$\phi(\rho'_{i_1} \cdots \rho'_{i_k}(b')) = \rho''_{i_1} \cdots \rho''_{i_k}(b'') \qquad \text{for all } k \in \mathbb{N}_0, i_1, \dots, i_k \in I.$$

Then ϕ is well-defined: Assume that $\rho'_{i_1}\cdots\rho'_{i_k}(b')=b'$. Then $\sigma_{i_1}\cdots\sigma_{i_k}^{b'}\in\mathrm{End}(b')$, and hence an application of π and F_π gives $\rho_{i_1}\cdots\rho_{i_k}(a)=a$, $\sigma_{i_1}\cdots\sigma_{i_k}^{a}\in U$. Thus $F_{\pi''}(\sigma_{i_1}\cdots\sigma_{i_k}^{b''})\in U$, and hence Lemma 3.3(2) gives that $\rho''_{i_1}\cdots\rho''_{i_k}(b'')=b''$. The compatibility of ϕ with ρ', ρ'', $C'^{b'}$, $C''^{b''}$ is fulfilled by Definition 3.1 and by definition of ϕ. Further, $\phi: A'\to A''$ is a bijection, the construction of ϕ^{-1} being analogous. Hence ϕ gives rise to an equivalence of the Cartan schemes $\mathscr{C}'$ and $\mathscr{C}''$. □

Definition 3.5. We say that a Cartan scheme $\mathscr{C}$ is *simply connected* if $\mathrm{End}(a)$ is the trivial group for all $a\in A$.

Corollary 3.6. *Let $\mathscr{C}$ be a connected Cartan scheme satisfying (C3). Then up to equivalence there exists a unique covering $\mathscr{C}'$ of $\mathscr{C}$ which is simply connected and satisfies (C3).*

As usual, this simply connected covering of $\mathscr{C}$ is called the *universal covering*.

Proof. The claim follows from Proposition 3.4(2) by setting $U=\{1\}$. □

Proposition 3.7. *Let $\mathscr{C}$, $\mathscr{C}'$ be connected Cartan schemes and $\pi:\mathscr{C}'\to\mathscr{C}$ a covering.*

(1) *If there exists a root system $\mathscr{R}'$ of type $\mathscr{C}'$, then the equations*

$$R^a=\bigcap_{\substack{a'\in A'\\ \pi(a')=a}} R'^{a'} \qquad \textit{for all } a\in A \tag{3-4}$$

define a root system $\mathscr{R}$ of type $\mathscr{C}$.

(2) *If there exists a root system $\mathscr{R}$ of type $\mathscr{C}$, and $\mathscr{C}'$ satisfies (C3), then the equations*

$$R'^{a'}=R^{\pi(a')} \qquad \textit{for all } a'\in A' \tag{3-5}$$

define a root system $\mathscr{R}'$ of type $\mathscr{C}'$.

Proof. (1) By Definition 3.1 and Axioms (R1)–(R4) for $\mathscr{R}'$, the Axioms (R1)–(R4) are fulfilled for $\mathscr{R}$.

(2) Since Axioms (R1)–(R3) hold for $\mathscr{R}$, they also hold for $\mathscr{R}'$. Suppose that $i,j\in I$ and $a'\in A'$ such that $i\neq j$ and that $m_{i,j}^{a'}=m_{i,j}^{a}$ is finite, where $a=\pi(a')$. Then $(\sigma_i\sigma_j)^{m_{i,j}^a}1_a=\mathrm{id}_a$ by Theorem 2.6. Hence $(\sigma_i\sigma_j)^{m_{i,j}^a}1_{a'}=\mathrm{id}$, and (C3) for $\mathscr{C}'$ implies that $(\rho'_i\rho'_j)^{m_{i,j}^{a'}}(a')=a'$. Thus (R4) holds for $\mathscr{R}'$ and hence $\mathscr{R}'$ is a root system of type $\mathscr{C}'$. □

4. Continued fractions

Continued fractions are related to Weyl groupoids of Cartan schemes of rank two. We recall some basic facts about continued fractions and formulate the facts we will use in our study.

A *continued fraction* is a sequence of indeterminates $a_1, a_2, a_3, \dots, b_0, b_1, \dots$ written in the form

$$b_0 + \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \dots = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \ddots}} \tag{4-1}$$

(see [Perron 1929] for an introduction). We assume the a_i and b_i are integers. The *convergents* of (4-1) are the numbers

$$\frac{A_\nu}{B_\nu} = b_0 + \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \dots + \frac{a_\nu|}{|b_\nu},$$

for $\nu \in \mathbb{N}$, also given by the recursion

$$\begin{pmatrix} B_0 & A_0 \\ B_{-1} & A_{-1} \end{pmatrix} = \begin{pmatrix} 1 & b_0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} b_\nu & a_\nu \\ 1 & 0 \end{pmatrix} \begin{pmatrix} B_{\nu-1} & A_{\nu-1} \\ B_{\nu-2} & A_{\nu-2} \end{pmatrix} = \begin{pmatrix} B_\nu & A_\nu \\ B_{\nu-1} & A_{\nu-1} \end{pmatrix}. \tag{4-2}$$

One says that the continued fraction (4-1) is *convergent* if, for some $\nu_0 \in \mathbb{N}$, the sequence $(A_\nu/B_\nu)_{\nu \geq \nu_0}$ is well-defined and converges in $\mathbb{R}$.

The case where all a_ν are 1 is the most important one and well understood. However, we will be interested in a different case: From now on, let $a_\nu = -1$, $b_\nu \in \mathbb{N}$ for all ν and assume that the sequence $b_1, b_2, \dots$ is periodic. For any $i \in \mathbb{Z}$, let

$$\eta(i) = \begin{pmatrix} i & -1 \\ 1 & 0 \end{pmatrix} \in \mathrm{SL}(2, \mathbb{Z}). \tag{4-3}$$

We will often need the following equations, which hold for all $i, j, k \in \mathbb{Z}$.

$$\eta(i)^{-1} = \begin{pmatrix} 0 & 1 \\ -1 & i \end{pmatrix}, \tag{4-4}$$

$$\eta(i)\eta(j) = \begin{pmatrix} ij-1 & -i \\ j & -1 \end{pmatrix}, \tag{4-5}$$

$$\eta(i)\eta(j)\eta(k) = \begin{pmatrix} (ij-1)k-i & -(ij-1) \\ jk-1 & -j \end{pmatrix}, \tag{4-6}$$

$$\tau\eta(i)\tau = \eta(i)^{-1}, \quad \tau\eta(i)^{-1}\tau = \eta(i), \tag{4-7}$$

where

$$\tau = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{4-8}$$

By (4-2),

$$\begin{pmatrix} B_n \\ B_{n-1} \end{pmatrix} = \eta(b_n)\cdots\eta(b_1)\begin{pmatrix} B_0 \\ B_{-1} \end{pmatrix}.$$

The product $\eta(b_n)\cdots\eta(b_1)$ will appear in the study of Weyl groupoids of rank two. In particular, we will need to know for which sequences $b_n, \dots, b_1$ this product has finite order. If it has finite order, then, since $B_{-1} = 0$, there exists $\nu \in \mathbb{N}$ such that $B_\nu = 0$.

The following fact is well-known. Variations of it were considered for example by Stern, Pringsheim, and Tietze; see respectively Satz 15 (§51), Satz 24 (§53), and Satz 1 (§35) in [Perron 1929].

Theorem 4.1. *If $a_\nu = -1$ and $b_\nu \geq 2$ for all $\nu \in \mathbb{N}$, then the continued fraction $\frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \cdots$ is convergent.*

Thus we get:

Corollary 4.2. *Let $n \in \mathbb{N}$ and $b_1, \dots, b_n \in \mathbb{Z}$. If $b_i \geq 2$ for all $i \in \{1, \dots, n\}$, then $\eta(b_1)\cdots\eta(b_n)$ does not have finite order.*

Proof. Assume $b_i \geq 2$ for all $i \in \{1, \dots, n\}$. If $\eta(b_1)\cdots\eta(b_n)$ had finite order, then the periodic continued fraction

$$\frac{-1|}{|b_n} + \frac{-1|}{|b_{n-1}} + \cdots + \frac{-1|}{|b_1} + \frac{-1|}{|b_n} + \frac{-1|}{|b_{n-1}} + \cdots + \frac{-1|}{|b_1} + \frac{-1|}{|b_n} + \cdots$$

would have infinitely many convergents with denominator 0. This is a contradiction to Theorem 4.1. □

One can also prove Corollary 4.2 without Theorem 4.1, using for example [Heckenberger 2008, Lemma 9].

5. Distinguished finite sequences of integers

We now study a special class of finite sequences of positive integers. They correspond to a class of continued fractions which are not convergent. Later we will use these sequences to classify finite root systems of type $\mathcal{C}$ and rank two. Recall the definition of the map $\eta : \mathbb{Z} \to \mathrm{SL}(2, \mathbb{Z})$ from (4-3).

Definition 5.1. Let $\mathcal{A}$ denote the set of finite sequences $(c_1, \dots, c_n)$ of integers such that $n \geq 1$ and $\eta(c_1)\cdots\eta(c_n) = -\mathrm{id}$. Let $\mathcal{A}^+$ be the subset of $\mathcal{A}$ formed by those $(c_1, \dots, c_n) \in \mathcal{A}$, for which $c_i \geq 1$ for all $i \in \{1, \dots, n\}$ and the entries in the first column of $\eta(c_1)\cdots\eta(c_i)$ are nonnegative for all $i < n$.

The following lemma will be crucial for our analysis of $\mathcal{A}^+$. It is related to a well-known transformation formula for continued fractions [Perron 1929, §37, Equations (1), (2)].

Lemma 5.2. *Let $n \geq 3$ and $c = (c_1, 1, c_3, c_4, \dots, c_n)$ such that $c_i \in \mathbb{Z}$ for all $i \in \{1, \dots, n\}$. Let $c' = (c_1 - 1, c_3 - 1, c_4, \dots, c_n)$.*

(1) *$c' \in \mathcal{A}$ if and only if $c \in \mathcal{A}$.*

(2) *$c' \in \mathcal{A}^+$ if and only if $c \in \mathcal{A}^+$, $c_1, c_3 \geq 2$.*

(3) *If $c \in \mathcal{A}^+$, then either $n = 3$, $c_1 = c_3 = 1$ or $n > 3$, $c_1, c_3 \geq 2$.*

Proof. If $i, k \in \mathbb{Z}$, then

$$\eta(i)\eta(1)\eta(k) = \begin{pmatrix} ik - i - k & 1 - i \\ k - 1 & -1 \end{pmatrix} = \eta(i-1)\eta(k-1)$$

by (4-5) and (4-6). This gives (1). By (4-5), the first column of $\eta(c_1)\eta(1)$ contains only nonnegative integers if and only if $c_1 \geq 1$. Thus (2) holds. Let $c \in \mathcal{A}^+$ such that $c_1 = 1$ or $c_3 = 1$. Then (4-6) gives that the upper left entry of $\eta(c_1)\eta(1)\eta(c_3)$ is -1, and hence $n = 3$. Then $c \in \mathcal{A}$ implies that $c_1 = c_3 = 1$. Hence (3) is proven. □

Proposition 5.3. *Let $n \in \mathbb{N}$ and $(c_1, \dots, c_n) \in \mathcal{A}^+$.*

(1) *Let $i, j \in \{1, \dots, n\}$ with $i \leq j$ and $(i, j) \neq (1, n)$. Then*

$$\eta(c_i)\eta(c_{i+1}) \cdots \eta(c_j) \in \mathrm{SL}(2, \mathbb{Z})$$

such that the first column contains only nonnegative and the second only nonpositive integers.

(2) *Let $i \in \{1, \dots, n\}$. Then $(c_i, c_{i+1}, \dots, c_n, c_1, \dots, c_{i-1}) \in \mathcal{A}^+$.*

(3) *$(c_n, c_{n-1}, \dots, c_2, c_1) \in \mathcal{A}^+$.*

(4) *If $n \leq 3$ then $(c_1, \dots, c_n) = (1, 1, 1)$.*

Proof. (1) We proceed by induction on the lexicographically ordered pairs (i, j).

If $i = j$ then we are done, since the matrix $\eta(c_i)$ satisfies the claim.

Let $i, j \in \{1, \dots, n\}$ with $i < j$ and $(i, j) \neq (1, n)$. Assume that the claim holds for all pairs $(i', j') \in \{1, \dots, n\}$ such that $i' \leq j'$ and either $i' < i$ or $i' = i$, $j' < j$. Let

$$\eta(c_i) \cdots \eta(c_j) = \begin{pmatrix} a & -b \\ c & -d \end{pmatrix}$$

with $a, b, c, d \in \mathbb{Z}$. Clearly, $-ad + bc = 1$ since $\eta(k) \in \mathrm{SL}(2, \mathbb{Z})$ for all $k \in \mathbb{Z}$. Moreover, (4-4) gives that

$$\eta(c_i) \cdots \eta(c_{j-1}) = \begin{pmatrix} a & -b \\ c & -d \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & c_j \end{pmatrix} = \begin{pmatrix} b & -(bc_j - a) \\ d & -(dc_j - c) \end{pmatrix}.$$

Hence $b, d \geq 0$ by induction hypothesis.

If $i = 1$, then $a, c \geq 0$ by definition of $\mathcal{A}^+$ and the assumption $(i, j) \neq (1, n)$, and hence we are done. Otherwise

$$\eta(c_{i-1}) \cdots \eta(c_j) = \begin{pmatrix} c_{i-1} & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a & -b \\ c & -d \end{pmatrix} = \begin{pmatrix} c_{i-1}a - c & d - c_{i-1}b \\ a & -b \end{pmatrix},$$

and hence $a > 0$ by induction hypothesis. Since $a, b, d \geq 0$, we get $bc = 1+ad \geq 1$, and hence $c > 0$, which proves the claim.

(2) It suffices to prove the claim for $i = 2$. If $\eta(c_1) \cdots \eta(c_n) = -\mathrm{id}$, then clearly $\eta(c_2) \cdots \eta(c_n)\eta(c_1) = -\mathrm{id}$. Let $j \in \{2, \dots, n\}$. Then the entries in the first column of $\eta(c_2) \cdots \eta(c_j)$ are nonnegative by part (1) of the proposition. This gives (2).

(3) Recall the definition of τ in (4-8). Then (4-7) gives that

$$\eta(c_n)\eta(c_{n-1}) \cdots \eta(c_1) = \tau\, \eta(c_n)^{-1} \eta(c_{n-1})^{-1} \cdots \eta(c_1)^{-1} \tau = -\mathrm{id}$$

since $\eta(c_1) \cdots \eta(c_n) = -\mathrm{id}$. Therefore $(c_n, c_{n-1}, \dots, c_1) \in \mathcal{A}$.

Let $2 \leq i \leq n$ and assume that

$$\eta(c_i)\eta(c_{i+1}) \cdots \eta(c_n) = \begin{pmatrix} a & -b \\ c & -d \end{pmatrix}$$

for some $a, b, c, d \in \mathbb{Z}$. Then $a, b, c, d \geq 0$ and $bc - ad = 1$ by part (1) of the proposition. We obtain that

$$\begin{aligned} \eta(c_n) \cdots \eta(c_i) &= \tau\, \eta(c_n)^{-1} \cdots \eta(c_i)^{-1} \tau \\ &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} -d & b \\ -c & a \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} a & -c \\ b & -d \end{pmatrix}. \end{aligned}$$

Thus $(c_n, c_{n-1}, \dots, c_1) \in \mathcal{A}^+$.

(4) Equations $\eta(c_1) = -\mathrm{id}$, $\eta(c_1)\eta(c_2) = -\mathrm{id}$ have no solutions with $c_1, c_2 \in \mathbb{N}$ by (4-3), (4-5). Let now $n = 3$ and $c_1, c_2, c_3 \in \mathbb{N}$. If $c_1, c_2, c_3 \geq 2$, then $(c_1, c_2, c_3) \notin \mathcal{A}$ by Corollary 4.2. Otherwise $c_1 = c_2 = c_3 = 1$ by Lemma 5.2(3) and part (2) of the proposition. Relation $(1, 1, 1) \in \mathcal{A}^+$ holds by (4-5) with $i = j = 1$. □

By Proposition 5.3(2) and (3), the dihedral group $\mathbb{D}_n$ of $2n$ elements, where $n \in \mathbb{N}$, acts on sequences of length n in $\mathcal{A}^+$ by cyclic permutation of the entries and by reflections. This action gives rise to an equivalence relation $\sim$ on $\mathcal{A}^+$ by taking the orbits of the action as equivalence classes. For brevity we will usually not distinguish between elements of $\mathcal{A}^+$ and $\mathcal{A}^+/\sim$. By Proposition 5.3(4) there is precisely one element of $\mathcal{A}^+/\sim$ of length 3.

Lemma 5.2 suggests to introduce a further equivalence relation $\approx$ on $\mathcal{A}^+$. Let $n, m \in \mathbb{N}$ with $m \geq n$, and let $c = (c_1, \dots, c_n)$, $d = (d_1, \dots, d_m) \in \mathcal{A}^+$. We write $c \approx' d$ if and only if $m = n$, $c \sim d$ or $m = n+1$, $d = (c_1+1, 1, c_2+1, c_3, c_4, \dots, c_n)$.

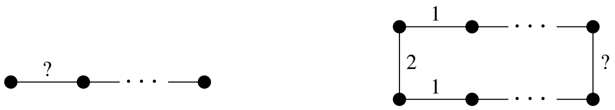


Figure 1. Left: chain diagram. Right: cycle diagram.

Definition 5.4. Let $c, d \in \mathcal{A}^+$. Write $c \approx d$ if and only if there exists $k \in \mathbb{N}$ and a sequence $c = e_1, e_2, \dots, e_k = d$ of elements of $\mathcal{A}^+$, such that $e_i \approx' e_{i+1}$ or $e_{i+1} \approx' e_i$ for all $i \in \{1, 2, \dots, k-1\}$.

Clearly, $\approx$ is an equivalence relation on $\mathcal{A}^+$. We are interested in the equivalence classes of $\mathcal{A}^+/\approx$.

Theorem 5.5. *The only element of* $\mathcal{A}^+/\approx$ *is* $(1, 1, 1)$.

Proof. Let $n \geq 1$ and $c = (c_1, \dots, c_n) \in \mathcal{A}^+$. By Proposition 5.3(4) it suffices to prove that if $n \geq 4$, then $c \approx c'$ for some $c' = (c'_1, c'_2, \dots, c'_{n-1}) \in \mathcal{A}^+$.

Assume that $n \geq 4$. By Corollary 4.2 there exists $i \in \{1, \dots, n\}$ such that $c_i = 1$. By Proposition 5.3(2) and the definition of $\approx$ we may assume that $c_2 = 1$. Now apply Lemma 5.2(2) and (3) to obtain the desired $c' \in \mathcal{A}^+$. □

Corollary 5.6. *If* $n \in \mathbb{N}$, $(c_1, \dots, c_n) \in \mathcal{A}^+$, *then* $\sum_{i=1}^n c_i = 3(n-2)$.

Proof. The expression $\sum_{i=1}^n c_i - 3(n-2)$ is zero for $c = (1, 1, 1)$ and is an invariant of $\approx$. □

6. Connected root systems of rank two

Throughout this section I will denote a two-element set, A a finite set, and $\mathcal{C} = \mathcal{C}(I, A, (\rho_i)_{i \in I}, (C^a)_{a \in A})$ a connected Cartan scheme. Since $\rho_i^2 = \mathrm{id}$ for all $i \in I$, and $\mathcal{C}$ is connected, the object change diagram of $\mathcal{C}$ is either a chain (if ρ_i has a fixed point for some $i \in I$) or a cycle; see Figure 1.

Recall that an element $w \in \mathrm{Hom}(\mathcal{W}(\mathcal{C}))$ is called *even* if $\det(w) = 1$.

Lemma 6.1. *The object change diagram of* $\mathcal{C}$ *is a cycle if and only if* $\mathrm{End}(a)$ *contains only even elements* (*for all* $a \in A$).

Proof. If the object change diagram of $\mathcal{C}$ is a cycle, then for all $a \in A$, $\mathrm{End}(a)$ consists of the elements $(\sigma_i \sigma_j)^{k|A|/2} 1_a$, where $k \in \mathbb{Z}$ and $I = \{i, j\}$. These are all even. Otherwise the object change diagram of $\mathcal{C}$ is a chain, and there exists $a \in A$ and $i \in I$ such that $\rho_i(a) = a$. Then $\mathrm{End}(a)$ is generated by σ_i^a and $(\sigma_j \sigma_i)^{|A|-1} \sigma_j^a$ which are odd. □

Assume that $\mathcal{C}$ admits a finite root system. The next proposition explains the relationship between the $m_{i,j}^a$ and the number $|A|$ of objects. For this, we need the following standard lemma.

Lemma 6.2. *Let $M \in \mathrm{GL}(2,\mathbb{Z})$. If the order e of M is finite, then*

$$-2 \le \mathrm{tr}(M) \le 2, \qquad e \in \{1,2,3,4,6\}.$$

Proposition 6.3. *Assume that $I = \{i, j\}$ and that $\mathscr{C}$ admits a finite root system. Then $m^a_{i,j} = m^a_{j,i} = |R^a_+|$ for all objects a. If the object change diagram is a cycle, then $m^a_{i,j} = \frac{1}{2}m\,|A|$ for some $m \in \{1,2,3,4,6\}$. If it is a chain, then $m^a_{i,j} = m\,|A|$ with the same possibilities for m.*

Proof. We have $m^a_{i,j} = m^a_{j,i} = |R^a_+|$ by Definition 2.3 for all objects a. Axiom (R3) from the same definition implies that $m^a_{i,j}$ does not depend on a. Let $d = |A|$ if the object change diagram is a chain and $d = |A|/2$ if it is a cycle. Then $(\sigma_j\sigma_i)^k 1_a \in \mathrm{End}(a)$, $k \in \mathbb{N}_0$, if and only if $k \in \mathbb{N}_0 d$. Theorem 2.6 and Lemma 6.2 give that $md = m^a_{i,j}$ for some $m \in \{1,2,3,4,6\}$. □

We are going to give a characterization of finite connected irreducible root systems of type $\mathscr{C}$. First we analyze root systems with simply connected Cartan schemes.

Lemma 6.4. *Assume that $\mathscr{C}$ is simply connected and that $\mathscr{R}$ is a finite root system of type $\mathscr{C}$. Then the object change diagram of $\mathscr{C}$ is a cycle with $|R^a|$ vertices, where $a \in A$.*

Proof. Since $\mathscr{C}$ is simply connected, $\mathrm{End}(a) = \{1\}$ for all $a \in A$. By Lemma 6.1 the object change diagram of $\mathscr{C}$ is a cycle. Now

$$\left|\mathrm{Hom}(\mathscr{W}(\mathscr{C}))1_a\right| = |A| \cdot \left|\mathrm{End}(a)\right|$$

since $\mathscr{C}$ is connected. Again, $\mathscr{C}$ is simply connected, hence $|A| = \left|\mathrm{Hom}(\mathscr{W}(\mathscr{C}))1_a\right|$. This is equal to $2|R^a_+|$ by Theorem 2.6, since $|I| = 2$. □

Proposition 6.5. *Assume that $I = \{i, j\}$ and that $\mathscr{R}$ is a finite irreducible root system of type $\mathscr{C}$. Let $a \in A$ and $n = |R^a_+|$. Let $a_1, a_2, \dots, a_{2n} \in A$ and $c_1, c_2, \dots, c_{2n} \in \mathbb{Z}$ such that*

$$\begin{aligned} a_{2r-1} &= (\rho_j\rho_i)^{r-1}(a), & a_{2r} &= \rho_i(\rho_j\rho_i)^{r-1}(a), \\ c_{2r-1} &= -c^{a_{2r-1}}_{ij}, & c_{2r} &= -c^{a_{2r}}_{ji} \end{aligned} \tag{6-1}$$

for all $r \in \{1,2,\dots,n\}$. Then $(c_1, c_2, \dots, c_n) \in \mathscr{A}^+$, $c_{n+r} = c_r$ for all $r \in \{1,2,\dots,n\}$, and $\rho_j(a_{2n}) = a$.

Proof. For all $r \in \mathbb{Z}$ let $i_r \in I$ such that $i_r = i$ for r odd and $i_r = j$ for r even. Let $\theta_{2r-1} = \sigma_i^{a_{2r-1}}\tau$, $\theta_{2r} = \tau\sigma_j^{a_{2r}} \in \mathrm{SL}(2,\mathbb{Z})$ for all $r \in \{1,\dots,n\}$. Then $\theta_r = \eta(c_r)$ for all $r \in \{1,\dots,2n\}$. Since $\mathscr{R}$ is irreducible, $c_r > 0$ for all r. By Lemmas 4 and 7 of [Heckenberger and Yamane 2008], $\ell(\sigma^{a_n}_{i_n}\cdots\sigma^{a_2}_{i_2}\sigma^{a}_{i_1}) = n$. Hence

$$\sigma^{a_n}_{i_n}\cdots\sigma^{a_2}_{i_2}\sigma^{a}_{i_1}(\{\alpha_1,\alpha_2\}) = \{-\alpha_1,-\alpha_2\}$$

by Lemma 8(iii) of the same work. Thus $\theta_n \cdots \theta_2\theta_1(\{\alpha_1, \alpha_2\}) = \{-\alpha_1, -\alpha_2\}$, and since $\det \theta_r = 1$ for all r, we conclude that $\theta_n \cdots \theta_2\theta_1 = -\mathrm{id}$. Hence $(c_n, \dots, c_2, c_1)$ lies in $\mathscr{A}$.

Clearly, if $2 \le r \le n$, then the first column of $\theta_n \cdots \theta_{r+1}\theta_r$ has nonnegative entries if and only if $\sigma_{i_n} \cdots \sigma_{i_{r+1}}\sigma_{i_r}^{a_r}(\alpha_{i_{r-1}})$ is a positive root. The latter is true by [Heckenberger and Yamane 2008, Lemma 4], and hence $(c_n, \dots, c_2, c_1) \in \mathscr{A}^+$. Then $(c_1, c_2, \dots, c_n) \in \mathscr{A}^+$ by Proposition 5.3(3).

Replacing in the construction a by a_2 and i by j, we find that $(c_2, \dots, c_n, c_{n+1})$ lies in $\mathscr{A}^+$. Then $\eta(c_1)^{-1} = -\eta(c_2) \cdots \eta(c_n) = \eta(c_{n+1})^{-1}$, and hence $c_1 = c_{n+1}$. Thus $c_{n+r} = c_r$ for all $r \in \{1, 2, \dots, n\}$ by induction on r. Finally, $\rho_j(a_{2n}) = (\rho_j\rho_i)^n(a) = a$ by (R4). □

The construction in Proposition 6.5 associates to any pair $(i, a) \in I \times A$ a sequence $(c_1, c_2, \dots, c_n) \in \mathscr{A}^+$. This defines a map

$$\Phi : I \times A \to \mathscr{A}^+.$$

Proposition 6.5 gives immediately, that

$$\Phi(j, a) = (c_n, c_{n-1}, \dots, c_1), \quad \Phi(j, \rho_i(a)) = (c_2, c_3, \dots, c_n, c_1). \tag{6-2}$$

Thus, by definition of $\sim$, the induced map $\Phi : I \times A \to \mathscr{A}^+/\sim$ is constant. But we can say more.

Theorem 6.6. *Let $n \in \mathbb{N}$ and $c = (c_1, c_2, \dots, c_n) \in \mathscr{A}^+$. Then there is a unique (up to equivalence) finite connected irreducible root system $\mathscr{R}$ with simply connected Cartan scheme of rank two such that $c \in \operatorname{Im} \Phi$.*

Proof. Assume that $c \in \mathscr{A}^+$, $\mathscr{R}$ is a connected irreducible root system of rank two, $i \in I$, and $a \in A$ such that $\Phi(i, a) = c$. If the Cartan scheme of $\mathscr{R}$ is simply connected, then by Lemma 6.4 and Proposition 6.5 the object change diagram of $\mathscr{R}$ is a cycle and $|A| = 2n$. The Cartan matrices C^a and the sets R^a, where $a \in A$, are then uniquely determined by the construction in Proposition 6.5. Thus $\mathscr{R}$ is uniquely determined. We describe $\mathscr{R}$ explicitly.

Let $I = \{i, j\}$ and let $A = \{a_1, \dots, a_{2n}\}$ be a set with $2n$ elements. Define $\rho_i, \rho_j : A \to A$ such that

$$\begin{aligned} \rho_i(a_{2r-1}) &= a_{2r}, & \rho_i(a_{2r}) &= a_{2r-1}, \\ \rho_j(a_{2r}) &= a_{2r+1}, & \rho_j(a_{2r+1}) &= a_{2r} \end{aligned} \tag{6-3}$$

for all $r \in \{1, 2, \dots, n\}$, where $a_{2n+1} = a_1$. Then $\rho_i^2 = \rho_j^2 = \mathrm{id}$. Let $c_{ln+r} = c_r$ for all $r \in \{1, 2, \dots, n\}$ and $l \in \mathbb{Z}$, and define

$$C^{a_{2r-1}} = \begin{pmatrix} 2 & -c_{2r-1} \\ -c_{2r-2} & 2 \end{pmatrix}, \quad C^{a_{2r}} = \begin{pmatrix} 2 & -c_{2r-1} \\ -c_{2r} & 2 \end{pmatrix} \tag{6-4}$$

for all $r \in \{1, 2, \ldots, n\}$. Since $c_r \in \mathbb{N}$ for all $r \in \{1, 2, \ldots, 2n\}$, the matrices C^{a_r} satisfy (M1) and (M2). Since also (C1) and (C3) hold, $\mathcal{C} = \mathcal{C}(I, A, (\rho_i, \rho_j), (C^a)_{a \in A})$ is a connected Cartan scheme.

Now define

$$R^{a_{2r-1}} = \bigl\{\pm\eta(c_{2r-1})\eta(c_{2r}) \cdots \eta(c_{2r-2+l})\tbinom{1}{0} \bigm| 0 \le l \le n-1\bigr\},$$
$$R^{a_{2r}} = \bigl\{\pm\tau\eta(c_{2r})\eta(c_{2r+1}) \cdots \eta(c_{2r+l-1})\tbinom{1}{0} \bigm| 0 \le l \le n-1\bigr\},$$

for all $r \in \{1, 2, \ldots, n\}$. Note that $|R^a_+| = n$ for all $a \in A$. Indeed, otherwise $\eta(c_r)\eta(c_{r+1}) \cdots \eta(c_{r+l-1})\tbinom{1}{0} = \tbinom{1}{0}$ for some $r \in \{1, \ldots, 2n\}$ and $l \in \{1, \ldots, n-1\}$. Then

$$\eta(c_{r+1})\eta(c_{r+2}) \cdots \eta(c_{r+l-1})\tbinom{1}{0} = \eta(c_r)^{-1}\tbinom{1}{0} = \tbinom{0}{-1},$$

a contradiction to Proposition 5.3(1) and (2).

Axiom (R1) is fulfilled by Proposition 5.3(2). Let $r \in \{1, 2, \ldots, 2n\}$. Equation $\eta(c_r)\eta(c_{r+1}) \cdots \eta(c_{r+n-1}) = -\mathrm{id}$ implies that

$$\eta(c_r)\eta(c_{r+1}) \cdots \eta(c_{r+n-2}) = -\eta(c_{r+n-1})^{-1},$$

and hence $\pm\alpha_1, \pm\alpha_2 \in R^{a_r}$. Since $\tau, \eta(l) \in \mathrm{SL}(2, \mathbb{Z})$ for all $l \in \mathbb{Z}$, we get (R2). (R4) holds by (6-3), since $|R^a_+| = n$ for all $a \in A$.

Now we prove (R3). Let $r \in \{1, 2, \ldots, 2n\}$. Then $\sigma_i^{a_r} = \eta(-c_{ij}^{a_r})\tau = \tau\eta(-c_{ij}^{a_r})^{-1}$ by (6-4), (4-7). If r is odd, then

$$\sigma_i^{a_r}(R^{a_r}) = \tau\eta(c_r)^{-1}\bigl(\bigl\{\pm\eta(c_r)\eta(c_{r+1}) \cdots \eta(c_{r+l-1})\tbinom{1}{0} \bigm| 0 \le l \le n-1\bigr\}\bigr) \subset R^{a_{r+1}} = R^{\rho_i(a_r)},$$

and if r is even, then

$$\sigma_i^{a_r}(R^{a_r}) = \eta(c_{r-1})\tau\bigl(\bigl\{\pm\tau\eta(c_r)\eta(c_{r+1}) \cdots \eta(c_{r+l-1})\tbinom{1}{0} \bigm| 0 \le l \le n-1\bigr\}\bigr) \subset R^{a_{r-1}} = R^{\rho_i(a_r)}.$$

Similarly, $\sigma_j^{a_r} = \tau\eta(c_{r-1})$ for odd r and $\sigma_j^{a_r} = \eta(c_r)^{-1}\tau$ for even r. Hence $\sigma_j^{a_r}(R^{a_r}) \subset R^{\rho_j(a_r)}$, (R3) holds, and $\mathcal{R}$ is a finite irreducible root system of type $\mathcal{C}$. The Cartan scheme $\mathcal{C}$ is simply connected, since $\bigl|\mathrm{Hom}(\mathcal{W}(\mathcal{C}))1_{a_1}\bigr| = 2n = |A|$ and $(c_1, \ldots, c_n) \in \mathcal{A}$. Finally, $\Phi(i, a_1) = (c_1, \ldots, c_n)$ because of (6-1), (6-3), and (6-4). □

Corollary 6.7. *Assume that there is a finite root system $\mathcal{R}$ of type $\mathcal{C}$. Then there are $a \in A$ and $i, j \in I$ with $i \neq j$ such that $c^a_{ij} = 0$ or $c^a_{ij} = -1$.*

Proof. If $\mathcal{R}$ is not irreducible, then $C^a_{ij} = 0$ for all $a \in A$ and $i, j \in I$ with $i \neq j$; see the end of Section 2. Otherwise Proposition 6.5 gives that the negatives of the entries of the Cartan matrices of $\mathcal{C}$ give rise to a sequence $(c_1, \dots, c_n) \in \mathcal{A}^+$. By Corollary 4.2, this sequence has an entry 1, and the corollary is proven. □

Remark 6.8. The assumption in Corollary 6.7 can be weakened for example by requiring only that $\mathcal{W}(\mathcal{C})$ is finite. We don't work out the details, since we are mainly interested in Cartan schemes admitting (finite) root systems.

We are going to give a very effective algorithm to decide if our given connected Cartan scheme $\mathcal{C}$ admits a finite irreducible root system. The central notions towards this will be the characteristic sequences and centrally symmetric Cartan schemes. Our algorithm can also be used to get a more precise classification of root systems of rank two, for example in form of explicit lists for a given number of objects.

Definition 6.9. Assume that the object change diagram of $\mathcal{C}$ is a cycle. Let $i \in I$, $a \in A$, and define $a_1, \dots, a_{|A|} \in A$ and $c_1, \dots, c_{|A|} \in \mathbb{N}_0$ by

$$\begin{aligned} a_{2k-1} &= (\rho_j \rho_i)^{k-1}(a), & a_{2k} &= (\rho_i \rho_j)^{k-1}\rho_i(a), \\ c_{2k-1} &= -c_{ij}^{a_{2k-1}}, & c_{2k} &= -c_{ji}^{a_{2k}} \end{aligned}$$

for all $k \in \{1, 2, \dots, |A|/2\}$, where $I = \{i, j\}$. Then $(c_1, c_2, \dots, c_{|A|})$ is called the *characteristic sequence of* $\mathcal{C}$ with respect to i and a. The Cartan scheme $\mathcal{C}$ is termed *centrally symmetric* if $c_k = c_{k+|A|/2}$ for all $k \in \{1, 2, \dots, |A|/2\}$. In this case we write also $(c_1, c_2, \dots, c_{|A|/2})^2$ for $(c_1, c_2, \dots, c_{|A|})$.

Remark 6.10. Let $(c_1, c_2, \dots, c_{|A|})$ be the characteristic sequence of $\mathcal{C}$ with respect to i and a. Then the characteristic sequences with respect to j and a and i and $\rho_i(a)$, respectively, are $(c_{|A|}, c_{|A|-1}, \dots, c_1)$ and $(c_1, c_{|A|}, c_{|A|-1}, \dots, c_3, c_2)$, respectively. Thus if $\mathcal{C}$ is centrally symmetric with respect to i and a, it is also centrally symmetric with respect to j and a and i and $\rho_i(a)$, respectively. Since $\mathcal{C}$ is connected, this means that $\mathcal{C}$ being centrally symmetric is independent of the choice of $i \in I$ and $a \in A$.

Remark 6.11. Characteristic sequences must not be confused with elements of $\mathcal{A}$ or $\mathcal{A}^+$. Their precise relationship will not be needed in the sequel, so we don't work it out in detail.

Remark 6.12. Let $n \in \mathbb{N}$ and let $c = (c_1, c_2, \dots, c_{2n})$ be a sequence of positive integers. By axioms (M1) and (C3) there is a unique (up to equivalence) connected Cartan scheme $\mathcal{C}$ with object change diagram a cycle, such that the characteristic sequence of $\mathcal{C}$ (with respect to some $i \in I$ and $a \in A$) is c.

Remark 6.13. Assume that $\mathcal{C}$ is simply connected, and that there exists a finite irreducible root system of type $\mathcal{C}$. Then $\mathcal{C}$ is centrally symmetric by Lemma 6.4 and Proposition 6.5.

Remark 6.14. Assume that the object change diagram of $\mathcal{C}$ is a cycle. By Lemma 6.1 and Proposition 3.4 the object change diagram of an n-fold covering $\mathcal{C}'$ of $\mathcal{C}$, where $n \in \mathbb{N}$, is a cycle. The characteristic sequence of $\mathcal{C}'$ is just the n-fold repetition of the characteristic sequence of $\mathcal{C}$. Thus an n-fold covering of $\mathcal{C}$ is centrally symmetric if and only if $\mathcal{C}$ is centrally symmetric or n is even.

Lemma 6.15. *Assume that there exists a finite irreducible root system of type $\mathcal{C}$. Suppose that the object change diagram of $\mathcal{C}$ is a chain. Then there is a unique double covering $\mathcal{C}'$ of $\mathcal{C}$ and a finite irreducible root system of type $\mathcal{C}'$ such that the object change diagram of $\mathcal{C}'$ is a cycle.*

Proof. By assumption there exists $a \in A$ and $i \in I$ such that $\rho_i(a) = a$. Then $\mathrm{End}(a)$ is generated by σ_i^a and $\tau^a = (\sigma_j\sigma_i)^{|A|-1}\sigma_j^a$, where $I = \{i, j\}$. Since σ_i^a, τ^a are reflections, for the subgroup $U = \langle \sigma_i^a \tau^a \rangle \subset \mathrm{End}(a)$ we obtain that $[\mathrm{End}(a) : U] = 2$, and U consists of even elements. By Proposition 3.4(2) there exists a unique double covering $\mathcal{C}'$ of $\mathcal{C}$ satisfying Axiom (C3) such that $\mathrm{End}(a') \simeq U$ for all $a' \in A'$. By Lemma 6.1 the object change diagram of $\mathcal{C}'$ is a cycle. The uniqueness of $\mathcal{C}'$ holds, since U is the unique subgroup of $\mathrm{End}(a)$ consisting of even elements and satisfying $[\mathrm{End}(a) : U] = 2$. The existence of a finite irreducible root system of type $\mathcal{C}'$ follows from Proposition 3.7(2). □

Remark 6.16. If $\mathcal{C}'$ is a Cartan scheme with object change diagram a cycle, then $\mathcal{C}'$ is the double covering of a Cartan scheme with object change diagram a chain if and only if there exist $i \in I'$, $a \in A'$, such that the characteristic sequence of $\mathcal{C}'$ with respect to i and a is of the form $(c_1, \dots, c_n, c_{n+1}, c_n, c_{n-1}, \dots, c_2)$ with $n = |A'|/2$ and $c_1, \dots, c_{n+1} \in \mathbb{N}_0$.

Lemma 6.17. *Assume that there exists a finite irreducible root system of type $\mathcal{C}$. Suppose that the object change diagram of $\mathcal{C}$ is a cycle, and that $\mathcal{C}$ is not centrally symmetric. Then there is a unique double covering $\mathcal{C}'$ of $\mathcal{C}$ which admits a* (*finite irreducible*) *root system. The Cartan scheme $\mathcal{C}'$ is centrally symmetric.*

Proof. Since the object change diagram of $\mathcal{C}$ is a cycle, $\mathrm{End}(a)$ is cyclic for all $a \in A$. The universal covering of $\mathcal{C}$ is centrally symmetric by Remark 6.13. Since $\mathcal{C}$ is not centrally symmetric, $|\mathrm{End}(a)|$ is even by Remark 6.14 and Proposition 3.4(2). By Proposition 3.4(2) there is a unique double covering $\mathcal{C}'$ of $\mathcal{C}$ satisfying (C3). It admits a finite irreducible root system of type $\mathcal{C}'$ by Proposition 3.7(2). All coverings of $\mathcal{C}$ admitting a root system fulfill (C3). Hence $\mathcal{C}'$ is the only double covering of $\mathcal{C}$ admitting a root system. This $\mathcal{C}'$ is centrally symmetric by Remark 6.14. □

Remark 6.18. Let $\mathcal{C}'$ be a Cartan scheme with object change diagram a centrally symmetric cycle, and $n = |A'|$. Then $\mathcal{C}'$ is the double covering of a Cartan scheme with object change diagram a not centrally symmetric cycle if and only if $n \in 4\mathbb{N}$, and with respect to one (equivalently, all) pair $(i', a') \in I' \times A'$ the characteristic sequence of $\mathcal{C}'$ is not of the form

$$(c_1, c_2, \dots, c_{n/4}, c_1, c_2, \dots, c_{n/4})^2,$$

where $c_1, \dots, c_{n/4} \in \mathbb{N}_0$.

In order to decide if a given connected Cartan scheme admits a finite root system, Lemmas 6.15 and 6.17 allow us to concentrate on centrally symmetric Cartan schemes. Further, since the classification of finite root systems with at most three objects is known [Cuntz and Heckenberger 2008], we may assume that the Cartan scheme has at least 4 objects.

For any matrix C, let C^{t} denote the transpose of C.

Theorem 6.19. *Let $\mathcal{C} = \mathcal{C}(I, A, (\rho_i)_{i\in I}, (C^a)_{a\in A})$ be a connected centrally symmetric Cartan scheme with $|A| \geq 4$.*

(1) *Assume that the characteristic sequence of $\mathcal{C}$ contains 0. Then $c^a_{ij} = 0$ for all $a \in A$ and $i, j \in I$ with $i \neq j$. Moreover, $\mathcal{C}$ admits a finite root system if and only if $|A| = 4$.*

(2) *If all entries of the characteristic sequence of $\mathcal{C}$ are at least two, then $\mathcal{C}$ does not admit a finite root system.*

(3) *Assume that the characteristic sequence of $\mathcal{C}$ is of the form*

$$c = (c_1, 1, c_3, c_4, \dots, c_{|A|/2})^2.$$

(Thus $0 \notin \{c_1, \dots, c_{|A|/2}\}$ by (1).) If $c_1 = 1$ or $c_3 = 1$, then there is a finite root system of type $\mathcal{C}$ if and only if $|A| = 6$ and $c_1 = c_3 = 1$. If $c_1 > 1$ and $|A| = 4$, then there is a finite root system of type $\mathcal{C}$ if and only if $c_1 \in \{2, 3\}$. If $c_1 > 1$, $c_3 > 1$, and $|A| \geq 6$, then there is a finite root system of type $\mathcal{C}$ if and only if the Cartan scheme with object change diagram a cycle with $|A| - 2$ edges and with characteristic sequence

$$(c_1 - 1, c_3 - 1, c_4, \dots, c_{|A|/2})^2 \tag{6-5}$$

admits a finite root system.

Proof. (1) follows from (M2), (C3), and (R4), and (2) from Corollary 6.7.

(3) If $c_1 = 1$ or $c_3 = 1$, then there exists $a \in A$ such that $c^a_{ij} = c^a_{ji} = -1$, where $I = \{i, j\}$. Then Lemma 4.8 of [Cuntz and Heckenberger 2008] gives that $m^a_{i,j} = 3$ and $c_r = 1$ for all $r \in \{1, 3, 4, \dots, |A|/2\}$. By (R4) we get $|A| = 6$.

Assume next that $c_1 > 1$ and $|A| = 4$. Then $C^a = C^b$ for all $a, b \in A$, and hence $\mathcal{C}$ admits a finite root system if and only if C^a is of finite type and (R4) holds (see Theorem 3.3 of the same reference), that is, $c_1 \in \{2, 3\}$.

Finally, assume that $c_1 > 1$, $c_3 > 1$, $|A| \geq 6$, and $\mathcal{C}$ admits a finite root system. By Proposition 3.7, the universal covering $\mathcal{C}'$ of $\mathcal{C}$ admits a finite root system. Hence A' is finite by (C1) and (R4). Therefore $\operatorname{End}(a) \subset \operatorname{Hom}(\mathcal{W}(\mathcal{C}))$ is finite for all $a \in A$ by (3-3). Let $m = |\operatorname{End}(a)|$. Remark 6.14 and Lemma 6.4 tell that the object change diagram of $\mathcal{C}'$ is a centrally symmetric cycle, and the characteristic sequence of $\mathcal{C}'$ is an m-fold repetition of c. Let

$$\tilde{c} = (c_1, 1, c_3, c_4, \dots, c_{|A|/2}).$$

By Proposition 6.5 the m-fold repetition of $\tilde{c}$ is an element of $\mathcal{A}^+$. Since $|A| \geq 6$, Lemma 5.2(2) gives that the m-fold repetition of

$$\tilde{c}' = (c_1 - 1, c_3 - 1, c_4, \dots, c_{|A|/2})$$

is in $\mathcal{A}^+$. Let $\mathcal{C}''$ be the connected simply connected Cartan scheme which corresponds to the m-fold repetition of $\tilde{c}'$ via Theorem 6.6. It admits a finite root system. Now $\mathcal{C}''$ is the m-fold covering of a Cartan scheme $\mathcal{C}'''$ with characteristic sequence given in (6-5). Hence Proposition 3.7 gives that $\mathcal{C}'''$ admits a finite root system.

We have shown that if $\mathcal{C}$ admits a finite root system, then also $\mathcal{C}'''$. The proof of the converse goes in the same way, and we are done. □

Example 6.20. Consider the connected Cartan scheme $\mathcal{C}$ of rank two with 4 objects, object change diagram a cycle and characteristic sequence $(5, 1, 2, 2)$. To check that $\mathcal{C}$ admits a finite root system, consider the double covering $\mathcal{C}'$ corresponding to the characteristic sequence $(5, 1, 2, 2)^2$. By Proposition 3.7, $\mathcal{C}$ admits a finite root system if and only if $\mathcal{C}'$ does. Theorem 6.19(3) allows one to replace $\mathcal{C}'$ by the Cartan scheme with characteristic sequence $(4, 1, 2)^2$ and then $(3, 1)^2$. Thus $\mathcal{C}$ admits a finite root system.

If we start with the characteristic sequence $(5, 1, 2, 3)$ for $\mathcal{C}$, then the analogous arguments produce the characteristic sequences $(5, 1, 2, 3)^2$, $(4, 1, 3)^2$ and $(3, 2)^2$, and then $\mathcal{C}$ does not admit a finite root system by Theorem 6.19(2).

Example 6.21. Theorem 6.19 also enables us to list all connected centrally symmetric Cartan schemes which admit a finite root system to a fixed number of objects. For example if $|A| = 4$, then there are 3 such schemes and they belong to the characteristic sequences $(0, 0)^2$, $(1, 2)^2$, $(1, 3)^2$. Therefore by Theorem 6.19(2) and (3), the only connected centrally symmetric Cartan schemes (up to equivalence) which have 6 objects and admit a finite root system are those that correspond to the characteristic sequences $(1, 1, 1)^2$, $(2, 1, 3)^2$ and $(2, 1, 4)^2$, and, if $|A| = 8$, then we obtain $(2, 1, 2, 1)^2$, $(3, 1, 2, 3)^2$, $(2, 2, 1, 4)^2$, $(3, 1, 4, 1)^2$, $(3, 1, 2, 4)^2$, $(2, 2, 1, 5)^2$ and $(3, 1, 5, 1)^2$. Similarly, we have 15, 47, 136 connected centrally symmetric Cartan schemes up to equivalence with 10, 12, 14 objects, respectively, admitting a finite root system.

According to Lemma 6.17 and the above lists for $|A| = 4$ and $|A| = 8$, the complete list of all characteristic sequences to irreducible Cartan schemes which admit a finite root system, with object change diagram a cycle and 4 objects is thus: $(1, 2, 1, 2)$, $(1, 3, 1, 3)$, $(3, 1, 2, 3)$, $(2, 2, 1, 4)$, $(3, 1, 4, 1)$, $(3, 1, 2, 4)$, $(2, 2, 1, 5)$, $(3, 1, 5, 1)$.

Remark 6.16 and the list for $|A| = 8$ also supports us with Cartan schemes with 4 objects which admit a finite root system and have a chain as object change diagram. The symmetry property mentioned in Remark 6.16 is fulfilled for the sequences $(2, 1, 2, 1)^2$ (also in the form $(1, 2, 1, 2)^2$), $(3, 1, 4, 1)^2$ (also in the form $(4, 1, 3, 1)^2$), and $(3, 1, 5, 1)^2$ (also in the form $(5, 1, 3, 1)^2$). This yields the following 6 Cartan schemes with 4 objects (the Cartan matrices represent the objects).

$$\begin{tikzcd}
\begin{pmatrix} 2 & -2 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -2 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "1"] & \begin{pmatrix} 2 & -2 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -2 \\ -1 & 2 \end{pmatrix} \\
\begin{pmatrix} 2 & -1 \\ -2 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -1 \\ -2 & 2 \end{pmatrix} \arrow[r, dash, "1"] & \begin{pmatrix} 2 & -1 \\ -2 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -1 \\ -2 & 2 \end{pmatrix} \\
\begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -4 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "1"] & \begin{pmatrix} 2 & -4 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} \\
\begin{pmatrix} 2 & -4 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "1"] & \begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -4 \\ -1 & 2 \end{pmatrix} \\
\begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -5 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "1"] & \begin{pmatrix} 2 & -5 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} \\
\begin{pmatrix} 2 & -5 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "1"] & \begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix} \arrow[r, dash, "2"] & \begin{pmatrix} 2 & -5 \\ -1 & 2 \end{pmatrix}
\end{tikzcd}$$

To complete the classification of all connected Cartan schemes with finite root system and 4 objects, it remains to calculate all connected Cartan schemes with finite root system and 8 objects with object change diagram a not centrally symmetric cycle, and then to apply Remark 6.16 to them, as indicated above, to get all chains with 4 objects. This is certainly an easy task for a computer but there are too many such Cartan schemes to list them here.

7. Bounds

Let $\mathcal{C} = \mathcal{C}(I, A, (\rho_i)_{i\in I}, (C^a)_{a\in A})$ be a connected Cartan scheme of rank two admitting a finite irreducible root system of type $\mathcal{C}$. Then A is finite by (C1) and (R4). Let $-q = -q(\mathcal{C})$ denote the sum of all nondiagonal entries of the Cartan matrices of $\mathcal{C}$, and $h = |\mathrm{End}(a)|$ for an $a \in A$. Then $|\mathrm{End}(b)| = h$ for all $b \in A$, since $\mathcal{C}$ is connected.

Theorem 7.1. *We have* $h(6|A| - q) = 24$ *and*

$$|R^a_+| = \frac{h|A|}{2} = \frac{12|A|}{6|A| - q}.$$

Proof. The universal covering $\mathcal{C}'$ of $\mathcal{C}$ has $h|A|$ objects by (3-3), and $q(\mathcal{C}')/4 = 3(h|A|/2 - 2)$ by Proposition 6.5 and Corollary 5.6. Since $q(\mathcal{C}') = hq(\mathcal{C})$, we obtain that $hq = 6(h|A| - 4)$. Hence $h(6|A| - q) = 24$. Lemma 6.4 tells that $|R^a_+| = h|A|/2$. This yields the claim. □

Remark 7.2. Proposition 6.3 and Theorem 7.1 give that $h \in \{1, 2, 3, 4, 6\}$ if the object change diagram of $\mathcal{C}$ is a cycle, and $h/2 \in \{1, 2, 3, 4, 6\}$ if it is a chain. But this result could have been obtained much easier. Nevertheless, Theorem 7.1 gives a restriction for $q = 6|A| - 24/h$ for given number $|A|$ of objects in a finite irreducible root system.

Next we give sharp bounds for the entries of the Cartan matrices.

Proposition 7.3. *Assume that* $|A| \geq 2$. *Let* $c \leq 0$ *be an entry of* C^a *for some* $a \in A$. *If the object change diagram is a cycle* (*chain*), *then* $|c| \leq |A| + 1$ ($|c| \leq 2|A| + 1$).

Proof. Assume first that the object change diagram of $\mathcal{C}$ is a cycle. If $|A| \geq 4$ and $\mathcal{C}$ is centrally symmetric, then Theorem 6.19(2) and (3) yields by induction on $|A|$, that $|c| \leq |A|/2 + 1$. If $\mathcal{C}$ is not centrally symmetric, then by Lemma 6.17 there exists a double covering of $\mathcal{C}$ which is centrally symmetric. Hence $|c| \leq |A| + 1$.

If the object change diagram of $\mathcal{C}$ is a chain, then by Lemma 6.15 there exists a double covering of $\mathcal{C}$ which has a cycle as object change diagram. Hence $|c| \leq 2|A| + 1$. □

Proposition 7.4. *For all* $n \geq 1$ *there exist finite connected irreducible root systems* $\mathcal{R}$ *of rank two with* $|A| = 2n$ *and object change diagram a cycle or* $|A| = n$ *and object change diagram a chain such that* $-(2n + 1)$ *is an entry in a Cartan matrix* C^a, $a \in A$.

Proof. For $n = 1$ the claim follows from [Cuntz and Heckenberger 2008, Proposition 5.2].

Theorem 6.19 tells that for all $n \geq 2$ the Cartan scheme $\mathcal{C}_n$ with $4n$ objects, object change diagram a cycle, and characteristic sequence

$$(3, \underbrace{2, 2, \ldots, 2}_{n-2 \text{ times}}, 1, 2n + 1, 1, \underbrace{2, 2, \ldots, 2}_{n-2 \text{ times}})^2 \tag{7-1}$$

admits a finite irreducible root system with $|A| = 4n$. Indeed, if $n = 2$, then using Theorem 6.19(3) we can transform the sequence $(3, 1, 5, 1)^2$ first to $(2, 4, 1)^2$. By changing the reference object, the latter is equivalent to $(4, 1, 2)^2$, and using the same result we may reduce it to $(3, 1)^2$. If $n > 2$, then using Theorem 6.19(3) we may transform the sequence in (7-1) in two steps, first to

$$(3, \underbrace{2, 2, \dots, 2}_{n-3 \text{ times}}, 1, 2n, 1, \underbrace{2, 2, \dots, 2}_{n-2 \text{ times}})^2,$$

and then to

$$(3, \underbrace{2, 2, \dots, 2}_{n-3 \text{ times}}, 1, 2n-1, 1, \underbrace{2, 2, \dots, 2}_{n-3 \text{ times}})^2.$$

By induction on n we obtain that $\mathcal{C}_n$ admits a finite irreducible root system. By Remark 6.18, $\mathcal{C}_n$ is the double covering of a Cartan scheme $\mathcal{C}'_n$ with $2n$ objects, object change diagram a cycle, and characteristic sequence

$$(3, \underbrace{2, 2, \dots, 2}_{n-2 \text{ times}}, 1, 2n+1, 1, \underbrace{2, 2, \dots, 2}_{n-2 \text{ times}}).$$

By Proposition 3.7, $\mathcal{C}'_n$ admits a finite irreducible root system $\mathcal{R}'$, and $\mathcal{R}'$ is such a root system we are looking for. By Remark 6.16, $\mathcal{C}'_n$ is the double covering of a Cartan scheme $\mathcal{C}''_n$ with n objects and object change diagram a chain. By Proposition 3.7, $\mathcal{C}''_n$ admits a finite irreducible root system $\mathcal{R}''$, and the proposition is proven. □

Corollary 7.5. *Any $c \in \mathbb{N}$ occurs as the negative of an entry of a Cartan matrix of a finite connected irreducible root system of rank two.*

Proof. For even c use the appropriate intermediate step in the proof of Proposition 7.4. □

Corollary 7.6. *For $r, n \in \mathbb{N}$, there are only finitely many finite root systems $\mathcal{R}$ of rank r with n objects.*

Proof. Let I, A be finite sets with $|I| = r$ and $|A| = n$, and let $\mathcal{R}$ be a finite root system of rank r with object set A. For all $i, j \in I$ with $i \neq j$ the restriction $\mathcal{R}|_{\{i,j\}}$ (see [Cuntz and Heckenberger 2008, Definition 4.1]) is a finite root system of rank two. Hence the entries of the Cartan matrices of $\mathcal{R}$ are bounded by $2|A| + 1$ by Proposition 7.3. Since for all $i \in I$, ρ_i is one of finitely many permutations of A, and since finite root systems are uniquely determined by their Cartan scheme, the claim is proven. □

References

[Andruskiewitsch and Schneider 1998] N. Andruskiewitsch and H.-J. Schneider, "Lifting of quantum linear spaces and pointed Hopf algebras of order p^3", *J. Algebra* **209**:2 (1998), 658–691. MR 99k:16075 Zbl 55.0262.09

[Andruskiewitsch and Schneider 2005] N. Andruskiewitsch and H.-J. Schneider, "On the classification of finite-dimensional pointed Hopf algebras", Preprint, 2005. To appear in *Ann. Math.* arXiv math.QA/0502157

[Andruskiewitsch et al. 2008] N. Andruskiewitsch, I. Heckenberger, and H.-J. Schneider, "The Nichols algebra of a semisimple Yetter–Drinfeld module", Preprint, 2008. arXiv 0803.2430

[Bourbaki 1968] N. Bourbaki, *Groupes et algèbres de Lie. ch. 4, 5 et 6*, Actualités Scientifiques et Industrielles **1337**, Hermann, Paris, 1968. MR 39 #1590 Zbl 0186.33001

[Cuntz and Heckenberger 2008] M. Cuntz and I. Heckenberger, "Weyl groupoids with at most three objects", Preprint, 2008. arXiv 0805.1810

[Heckenberger 2006] I. Heckenberger, "The Weyl groupoid of a Nichols algebra of diagonal type", *Invent. Math.* **164**:1 (2006), 175–188. MR 2007e:16047 Zbl 05027328

[Heckenberger 2008] I. Heckenberger, "Rank 2 Nichols algebras with finite arithmetic root system", *Algebr. Represent. Theory* **11**:2 (2008), 115–132. MR 2009a:16080 Zbl 05250098

[Heckenberger 2009] I. Heckenberger, "Classification of arithmetic root systems", *Adv. Math.* **220**:1 (2009), 59–124. MR 2462836 Zbl 05376870

[Heckenberger and Yamane 2008] I. Heckenberger and H. Yamane, "A generalization of Coxeter groups, root systems, and Matsumoto's theorem", *Math. Z.* **259** (2008), 255–276. MR 2009e:20087 Zbl 05267026

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

[Montgomery 1993] S. Montgomery, *Hopf algebras and their actions on rings*, CBMS Regional Conference Series in Mathematics **82**, Amer. Math. Soc., 1993. MR 94i:16019 Zbl 0793.16029

[Perron 1929] O. Perron, *Die Lehre von den Kettenbrüchen*, Teubner, Leipzig, 1929.

Communicated by Susan Montgomery
Received 2008-07-01 Revised 2009-01-19 Accepted 2009-03-06

cuntz@mathematik.uni-kl.de *Fachbereich Mathematik, Universität Kaiserslautern,
Postfach 3049, D-67653 Kaiserslautern, Germany*
http://www.mathematik.uni-kl.de/~cuntz/en/index.html

i.heckenberger@googlemail.com *Mathematisches Institut, Ludwig-Maximilians-Universität
München, Theresienstr. 39, D-80333 München, Germany*
http://www.mi.uni-koeln.de/~iheckenb/istvane.html

The semigroup of Betti diagrams

Daniel Erman

The recent proof of the Boij–Söderberg conjectures reveals new structure about Betti diagrams of modules, giving a complete description of the cone of Betti diagrams. We begin to expand on this new structure by investigating the semigroup of such diagrams. We prove that this semigroup is finitely generated, and answer several other fundamental questions about it.

1. Introduction

Recent work of a number of authors [Boij and Söderberg 2008b; 2008a; Eisenbud et al. 2007; Eisenbud and Schreyer 2009], completely characterizes the structure of Betti diagrams of graded modules, but only if one is allowed to take arbitrary rational multiples of the diagrams. This Boij–Söderberg theory shows that the rational cone of Betti diagrams is a simplicial fan whose rays and facet equations have a remarkably simple description.[1]

In this note, we consider the integral structure of Betti diagrams from the perspective of Boij–Söderberg theory, and we begin to survey this new landscape. In particular, we replace the cone by the semigroup of Betti diagrams (see Definition 1.1 below) and answer several fundamental questions about the structure of this semigroup.

We first use the results of Boij–Söderberg theory to draw conclusions about the semigroup of Betti diagrams. This comparison leads to Theorem 1.3, that the semigroup of Betti diagrams is finitely generated.

We then seek conditions which prevent a diagram from being the Betti diagram of an actual module. Using these conditions, we build families of diagrams which are *not* the Betti diagram of any module. For instance, consider the family

$$E_\alpha := \begin{pmatrix} 2+\alpha & 3 & 2 & - \\ - & 5+6\alpha & 7+8\alpha & 3+3\alpha \end{pmatrix}, \quad \alpha \in \mathbb{N}.$$

MSC2000: primary 13D02; secondary 13D25.

Keywords: Boij–Söderberg Theory, Betti diagrams, Betti tables, minimal free resoultions.

The author was partially supported by an NDSEG fellowship.

[1] See [Boij and Söderberg 2008b] for the original conjecture, [Eisenbud and Schreyer 2009] for the Cohen–Macaulay case, and [Boij and Söderberg 2008a] for the general case. The introduction of [Eisenbud and Schreyer 2009] includes a particularly clear exposition of the main results.

We will use the theory of Buchsbaum–Eisenbud multiplier ideals to conclude that no member of this family can be the Betti diagram of a module. Yet each E_α belongs to the cone of Betti diagrams, and in fact, if we were to multiply any diagram E_α by 3, then the result *would* equal the Betti diagram of a module.

We produce further examples of obstructed diagrams by using properties of the Buchsbaum–Rim complex. Based on our examples, we establish several negative results about the semigroup of Betti diagrams. These negative results are summarized in Theorem 1.6.

To state our results more precisely, we introduce notation. Let S be the polynomial ring $S = k[x_1, \dots, x_n]$ where k is any field. If M is any finitely generated graded S-module, we can take a minimal free resolution

$$0 \to F_p \to \cdots \to F_1 \to F_0 \to M \to 0$$

with $F_i = \bigoplus_j S(-j)^{\beta_{i,j}(M)}$. We write $\beta(M)$ for the Betti diagram of M, thought of as an element of the vector space $\bigoplus_{j=-\infty}^{\infty} \bigoplus_{i=0}^{p} \mathbb{Q}$ with coordinates $\beta_{i,j}(M)$. The set of graded S-modules is a semigroup under the operation of direct sum, and the vector space is a semigroup under addition. By observing that $\beta(M \oplus M') = \beta(M) + \beta(M')$, we can think of β as a map of semigroups:

$$\{\text{ finitely generated graded } S\text{-modules}\} \xrightarrow{\beta} \bigoplus_{j=-\infty}^{\infty} \bigoplus_{i=0}^{p} \mathbb{Q}.$$

The image of this map is thus a semigroup. Furthermore, if we restrict β to any subsemigroup of S-modules, then the image of the restricted map is also a semigroup.

A *degree sequence* will mean an integral sequence $d = (d_0, \dots, d_p) \in \mathbb{N}^{p+1}$ where $d_i < d_{i+1}$. If there exists a Cohen–Macaulay module M of codimension p with all Betti numbers equal to zero except for $\beta_{i,d_i}(M)$, then we say that $\beta(M)$ is a pure diagram of type d. It was first shown in [Herzog and Kühl 1984] that any two pure diagrams of type d would be scalar multiples of one another. The existence of modules whose Betti diagrams are pure diagrams of type d was conjectured by [Boij and Söderberg 2008b] and proved by [Eisenbud et al. 2007] in characteristic 0 and by [Eisenbud and Schreyer 2009] in arbitrary characteristic. These pure diagrams play a central role in the Boij–Söderberg theorems.

Fix two degree sequences $\underline{d}$ and $\overline{d}$ of length p and such that $\underline{d}_i \leq \overline{d}_i$ for all i. Consider the semigroup $\mathfrak{X}$ of graded modules M such that

- M has projective dimension $\leq p$, and
- the Betti number $\beta_{i,j}(M)$ is nonzero only if $i \leq p$ and $\underline{d}_i \leq j \leq \overline{d}_i$.

Our choice of $\mathfrak{X}$ is meant to match the simplicial structure of the cone of Betti diagrams. We may now define our main objects of study.

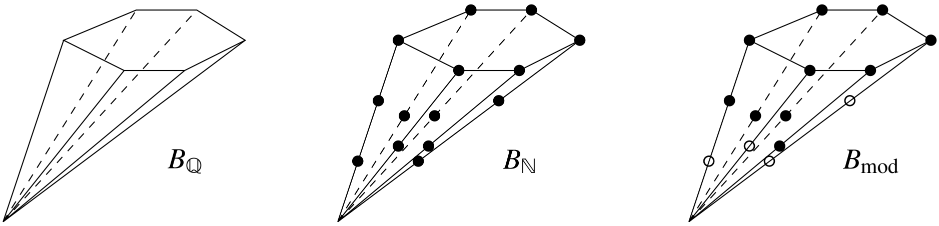


Figure 1. The cone of Betti diagrams $B_{\mathbb{Q}}$ is a simplicial fan which is described explicitly in [Eisenbud and Schreyer 2009] and [Boij and Söderberg 2008a]. This description can be used to understand the integral structure of the semigroup of virtual Betti diagrams $B_{\mathbb{N}}$. The semigroup of Betti diagrams B_{mod} is more mysterious.

Definition 1.1. The *semigroup of Betti diagrams* B_{mod} is defined as

$$B_{\mathrm{mod}} = B_{\mathrm{mod}}(\underline{d}, \bar{d}) := \operatorname{im} \beta|_{\mathfrak{X}}.$$

To study this object, it will be useful to consider two related ones:

Definition 1.2. The *cone of Betti diagrams* $B_{\mathbb{Q}}$ is the positive rational cone over the semigroup of Betti diagrams. The *semigroup of virtual Betti diagrams* $B_{\mathbb{N}}$ is the semigroup of lattice points in $B_{\mathbb{Q}}$.

One could define a cone of Betti diagrams without restricting which Betti numbers can be nonzero. This is the choice that [Eisenbud and Schreyer 2009] make, and our cone of Betti diagrams equals their big cone restricted to an interval. We choose to work with a finite dimensional cone in order to discuss the finiteness properties of B_{mod}.

A naive hope would be that the semigroups $B_{\mathbb{N}}$ and B_{mod} are equal. But a quick search yields virtual Betti diagrams which cannot equal the Betti diagram of module. Take for example the following pure diagram of type $(0, 1, 3, 4)$:

$$D_1 := \pi_{(0,1,3,4)} = \begin{pmatrix} 1 & 2 & - & - \\ - & - & 2 & 1 \end{pmatrix}. \tag{1}$$

This diagram belongs to the semigroup of virtual Betti diagrams. However, D_1 cannot equal the Betti diagram of an actual module as the two first syzygies would satisfy a linear Koszul relation which does not appear in the diagram D_1.

It is thus natural to compare B_{mod} and $B_{\mathbb{N}}$, and we will consider some questions about the semigroup of Betti diagrams:

(Q1) Is B_{mod} finitely generated?

(Q2) Does $B_{\mathrm{mod}} = B_{\mathbb{N}}$ in some special cases?

(Q3) Is B_{mod} a saturated semigroup?

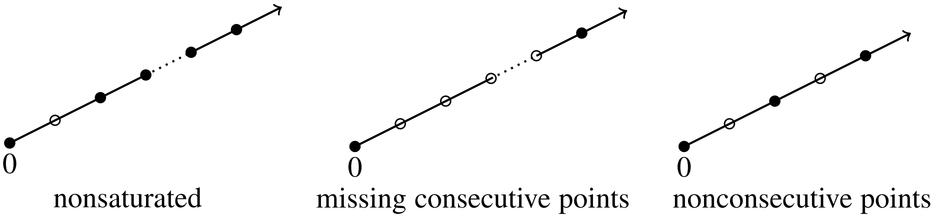


Figure 2. There exist rays that exhibit each of these behaviors.

(Q4) Is $B_{\mathbb{N}} \setminus B_{\text{mod}}$ a finite set?

(Q5) On a single ray, can we have consecutive points of $B_{\mathbb{N}}$ which fail to belong to B_{mod}? Nonconsecutive points?

In Section 2, we answer (Q1) affirmatively:

Theorem 1.3. *The semigroup of Betti diagrams B_{mod} is finitely generated.*

Sections 3 and 4 of this paper develop obstructions which prevent a virtual Betti diagram from being the diagram of some module. These obstructions are our tools for answering the other questions above. In Section 5, we consider (Q2), and prove the following:

Proposition 1.4. $B_{\mathbb{N}} = B_{\text{mod}}$ *for projective dimension* 1 *and for projective dimension* 2 *level modules.*

Our proof of Proposition 1.4 rests heavily on [Söderberg 2006], which shows the existence of level modules of embedding dimension 2 and with a given Hilbert function by constructing these modules as quotients of monomial ideals.

In [Erman $\geq$ 2009] we verify that, in a certain sense, projective dimension 2 diagrams generated in a single degree are "unobstructed." This leads us to:

Conjecture 1.5. $B_{\mathbb{N}} = B_{\text{mod}}$ for projective dimension 2 diagrams.

In the final section, we will consider questions (Q3)–(Q5). Here we show that the semigroup of Betti diagrams can have rather complicated behavior (see also Figure 2):

Theorem 1.6. *Each of the following occurs in the semigroup of Betti diagrams*:

(1) B_{mod} *is not necessarily a saturated semigroup.*

(2) *The set* $|B_{\mathbb{N}} \setminus B_{\text{mod}}|$ *is not necessarily finite.*

(3) *There exist rays of* B_{mod} *missing at least* $\dim S - 2$ *consecutive lattice points.*

(4) *There exist rays of* $B_{\mathbb{N}}$ *where the points of* B_{mod} *are nonconsecutive lattice points.*

Remark 1.7. Almost nothing in this paper would be changed if we swapped the semigroup $\mathfrak{X}$ for some subsemigroup of $\mathfrak{X}$ which respects the simplicial structure of $B_{\mathbb{Q}}$. For instance, we could consider the subsemigroup of Cohen–Macaulay modules of codimension e. The analogous statements of Theorems 1.3 and 1.6 and Proposition 1.4 all remain true in the Cohen–Macaulay case; one can even use the same proofs.

This paper is organized as follows. In Section 2, we prove that the semigroup of Betti diagrams is finitely generated. Sections 3 and 4 introduce obstructions for a virtual Betti diagram to be the Betti diagram of some module. The obstructions in Section 3 are based on properties of the Buchsbaum–Rim complex, and the obstruction in Section 4 focuses on the linear strand of a resolution and is based on the properties of Buchsbaum–Eisenbud multiplier ideals. Section 5 deals with the semigroup of Betti diagrams for small projective dimension, and contains the proof of Proposition 1.4. In Section 6 we prove Theorem 1.6 by constructing explicit examples based on our obstructions. Section 7 offers some open questions.

2. Finite generation of the semigroup of Betti diagrams

We fix a pair of degree sequences $\overline{d}, \underline{d} \in \mathbb{N}^{p+1}$ and work with the corresponding semigroup of Betti diagrams B_{mod}. Our proof of the finite generation of the semigroup of Betti diagrams uses the structure of the cone of Betti diagrams, so we begin by reviewing the relevant results. This structure was first proved in [Eisenbud and Schreyer 2009] for the Cohen–Macaulay case; the general case is similar, and was worked out in [Boij and Söderberg 2008a].

If d is any degree sequence then we set π_d to be the first lattice point on the ray corresponding to d. As illustrated in Figure 3, the cone $B_{\mathbb{Q}}$ is a rational simplicial fan whose defining rays correspond to rays of pure diagrams. To describe the simplicial structure, we recall the following partial ordering on degree sequences, introduced in [Boij and Söderberg 2008a]:

Definition 2.1. Let $d \in \mathbb{N}^{t+1}$ and $d' \in \mathbb{N}^{u+1}$. Then $d \leq d'$ if $t \geq u$ and $d_i \leq d'_i$ for all $i \leq u$.

The simplices of the fan $B_{\mathbb{Q}}$ correspond to maximal chains

$$d^0 < d^1 < \cdots < d^{s-1} < d^s$$

of degree sequences, where if $d^j \in \mathbb{N}^{t+1}$ then $\underline{d}_i \leq d^j_i \leq \overline{d}_i$ for all $i \leq t$. There are thus $s+1$ positions which may be nonzero for a Betti diagram in B_{mod} [Boij and Söderberg 2008a, Example 1]. In particular, $s+1 = \sum_{i=0}^{p} \overline{d}_i - \underline{d}_i + 1$.

Before proving Theorem 1.3, we first prove a simpler analog for the semigroup of virtual Betti diagrams $B_{\mathbb{N}}$.

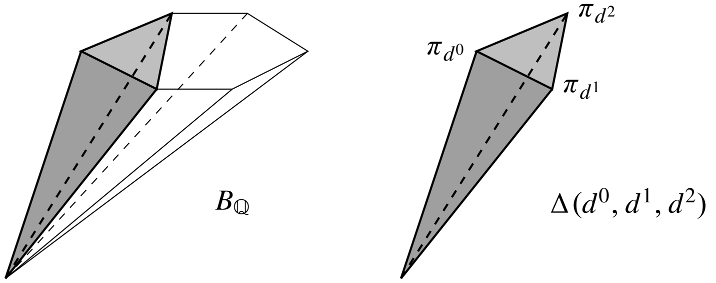


Figure 3. The cone $B_{\mathbb{Q}}$ is a simplicial fan. The simplex corresponding to a maximal sequence d^0, d^1, d^2 is highlighted in gray. The extremal rays of a simplex correspond to pure diagrams.

Lemma 2.2. *The semigroup $B_{\mathbb{N}}$ is finitely generated. There exists an integer m such that every virtual Betti diagram can be written as a $(1/m)\mathbb{N}$-combination of pure diagrams.*

Proof. Since $B_{\mathbb{N}}$ consists of the lattice points of the simplicial fan $B_{\mathbb{Q}}$, it is sufficient to prove this lemma after restricting to a single simplex Δ. Let $\pi_{d^0}, \dots, \pi_{d^s}$ be the pure diagrams defining Δ. Then the semigroup $B_{\mathbb{N}} \cap \Delta$ is generated by pure diagrams spanning Δ and by the lattice points inside the fundamental parallelepiped of Δ. This proves the first claim.

For the second claim of the lemma, let $P_1, \dots, P_N$ be the minimal generators of $B_{\mathbb{N}} \cap \Delta$. Every generator can be written as a positive rational sum:

$$P_i = \sum_j \frac{p_{ij}}{q_{ij}} \pi_{d^j}, \qquad p_{ij}, q_{ij} \in \mathbb{N}.$$

We set m_Δ to be the least common multiple of all the q_{ij}. Then we set m to be the least common multiple of the m_Δ for all Δ. □

We refer to m_Δ as a *universal denominator* for $B_{\mathbb{N}} \cap \Delta$. The existence of this universal denominator is central to our proof of the finite generation of B_{mod}.

Proof of Theorem 1.3. It is sufficient to prove the theorem for $B_{\text{mod}} \cap \Delta$ where Δ is a simplex of $B_{\mathbb{Q}}$. Let $\pi_{d^0}, \dots, \pi_{d^s}$ be the pure diagrams defining Δ, and let m_Δ be the universal denominator for $B_{\mathbb{N}} \cap \Delta$.

For $i = 0, \dots, s$, let $c_i \in \mathbb{N}$ be minimal such that $c_i \pi_{d^i}$ belongs to B_{mod}. The existence of such a c_i is guaranteed by Theorems 0.1 and 0.2 of [Eisenbud et al. 2007] and Theorem 5.1 of [Eisenbud and Schreyer 2009]. Let $\mathcal{S}_1$ be the semigroup generated by the pure diagrams $c_i \pi_{d^i}$. Let $\mathcal{S}_0$ be the semigroup generated by the pure diagrams $(1/m_\Delta)\pi_{d^i}$. Then we have the inclusions of semigroups

$$\mathcal{S}_1 \subseteq (B_{\text{mod}} \cap \Delta) \subseteq (B_{\mathbb{N}} \cap \Delta) \subseteq \mathcal{S}_0.$$

Passing to semigroup rings gives

$$k[\mathcal{S}_1] \subseteq k[B_{\mathrm{mod}} \cap \Delta] \subseteq k[B_{\mathbb{N}} \cap \Delta] \subseteq k[\mathcal{S}_0].$$

Observe that $k[\mathcal{S}_1]$ and $k[\mathcal{S}_0]$ are both polynomial rings of dimension $s+1$, and that $k[\mathcal{S}_1] \subseteq k[\mathcal{S}_0]$ is a finite extension of rings. This implies that $k[\mathcal{S}_1] \subseteq k[B_{\mathrm{mod}} \cap \Delta]$ is also a finite extension, and hence $k[B_{\mathrm{mod}} \cap \Delta]$ is a finitely generated k-algebra. We conclude that $B_{\mathrm{mod}} \cap \Delta$ is a finitely generated semigroup. □

Computing generators of $B_{\mathbb{N}}$. Minimal generators of $B_{\mathbb{N}} \cap \Delta$ can be computed explicitly as the generators of the $\mathbb{N}$-solutions to a certain linear $\mathbb{Z}$-system defined by the π_{d^i} and by m_Δ. For an overview of relevant algorithms, see the introduction of [Pisón-Casares and Vigneron-Tenorio 2004]. The following example illustrates the method.

Consider $S = k[x, y], \underline{d} = (0, 1, 4), \bar{d} = (0, 3, 4)$. The corresponding cone of Betti diagrams has several simplices and we choose the simplex Δ spanned by the maximal chain of degree sequences

$$(0) > (0, 3) > (0, 3, 4) > (0, 2, 4) > (0, 1, 4).$$

The corresponding pure diagrams are

$$\begin{pmatrix} 1 & - & - \\ - & - & - \\ - & - & - \end{pmatrix}, \begin{pmatrix} 1 & - & - \\ - & - & - \\ - & 1 & - \end{pmatrix}, \begin{pmatrix} 1 & - & - \\ - & - & - \\ - & 4 & 3 \end{pmatrix}, \begin{pmatrix} 1 & - & - \\ - & 2 & - \\ - & - & 1 \end{pmatrix}, \begin{pmatrix} 3 & 4 & - \\ - & - & - \\ - & - & 1 \end{pmatrix}. \tag{2}$$

First we must compute m_Δ. To do this, we consider the square matrix Φ whose columns correspond to the pure diagrams above:

$$\Phi = \begin{pmatrix} 1 & 1 & 1 & 1 & 3 \\ 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 1 & 4 & 0 & 0 \\ 0 & 0 & 3 & 1 & 1 \end{pmatrix}. \tag{3}$$

The columns of Φ are indexed by the pure diagrams in (2) and the rows of Φ are indexed by the Betti numbers $\beta_{0,0}, \beta_{1,1}, \beta_{1,2}, \beta_{1,3}$ and $\beta_{2,4}$ respectively. Since the columns of Φ are $\mathbb{Q}$-linearly independent, it follows that the cokernel of Φ is entirely torsion. Note that each minimal generator of $B_{\mathbb{N}} \cap \Delta$ is either a pure diagram or corresponds to a unique nonzero torsion element of $\operatorname{coker}(\Phi)$. The annihilator of $\operatorname{coker} \Phi$ is thus the universal denominator for Δ. A computation in Macaulay2 shows that $m_\Delta = 12$ in this case.

We next compute minimal generators of the $\mathbb{N}$-solutions of the linear $\mathbb{Z}$-system

$$\mathbb{Z}^{10} \xrightarrow{\begin{pmatrix} -12 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 3 \\ 0 & -12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & -12 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & -12 & 0 & 0 & 1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & -12 & 0 & 0 & 3 & 1 & 1 \end{pmatrix}} \mathbb{Z}^5.$$

The $\mathbb{N}$-solutions of the above system correspond to elements of $B_{\mathbb{N}} \cap \Delta$ under the correspondence

$$\begin{aligned} &(b_1, b_2, b_3, b_4, b_5, a_1, a_2, a_3, a_4, a_5) \\ &\qquad\mapsto \frac{a_1}{12}\pi_{(0)} + \frac{a_2}{12}\pi_{(0,3)} + \frac{a_3}{12}\pi_{(0,3,4)} + \frac{a_4}{12}\pi_{(0,2,4)} + \frac{a_5}{12}\pi_{(0,1,4)}. \end{aligned}$$

Computation yields that $B_{\mathbb{N}} \cap \Delta$ has 14 minimal semigroup generators.[2] These consist of the 5 pure diagrams from line (2) plus the following 9 diagrams:

$$\begin{pmatrix} 1 & 1 & - \\ - & - & - \\ - & 1 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 2 & - \\ - & 1 & - \\ - & - & 1 \end{pmatrix}, \begin{pmatrix} 1 & - & - \\ - & 1 & - \\ - & 2 & 2 \end{pmatrix}, \begin{pmatrix} 1 & - & - \\ - & - & - \\ - & 2 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 2 & - \\ - & - & - \\ - & 1 & 1 \end{pmatrix},$$

$$\begin{pmatrix} 3 & 3 & - \\ - & - & - \\ - & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & - & - \\ - & - & - \\ - & 3 & 2 \end{pmatrix}, \begin{pmatrix} 2 & 1 & - \\ - & 1 & - \\ - & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & - & - \\ - & 1 & - \\ - & 1 & 1 \end{pmatrix}.$$

It is not difficult to verify that each of these generators is the Betti diagram of some module. Thus in this case we have $B_{\mathbb{N}} \cap \Delta = B_{\text{mod}} \cap \Delta$.

Remark 2.3. We can easily bound the number of generators of $B_{\mathbb{N}} \cap \Delta$ from above. Let Δ be a simplex spanned by $d^0, \dots, d^s$. Let Φ be the square matrix

$$\Phi : \mathbb{Z}^{s+1} \to \bigoplus_{i=0}^{n} \bigoplus_{j=\underline{d}_i}^{\bar{d}_i} \mathbb{Z},$$

which sends the ℓth generator to the pure diagram π_{d^ℓ}. As in line (3), the cokernel of Φ will be entirely torsion (this follows from [Boij and Söderberg 2008a, Proposition 1].) Each minimal generator of $B_{\mathbb{N}} \cap \Delta$ will correspond to either a pure diagram or a unique nonzero element of $\operatorname{coker} \Phi$. Since the order of $\operatorname{coker} \Phi$ equals the determinant of Φ, the number of generators of $B_{\mathbb{N}} \cap \Delta$ is bounded above by $\det(\Phi) + s$.

We know of no effective upper bound for the number of generators of $B_{\text{mod}} \cap \Delta$.

[2] We use [Sturmfels 1993, Algorithm 2.7.3] for this computation. Also, see [Pisón-Casares and Vigneron-Tenorio 2004] for other relevant algorithms.

Remark 2.4. Although the semigroup $B_{\mathbb{N}}$ is saturated, the map $k[B_{\text{mod}}] \to k[B_{\mathbb{N}}]$ may not be the normalization map. For instance, if there is a ray r such that $r \cap B_{\text{mod}}$ only contains every other lattice point, then the saturation of $r \cap B_{\text{mod}}$ will not equal $r \cap B_{\mathbb{N}}$. Eisenbud et al. [2007] conjecture that there are no rays corresponding to pure diagrams which have this property .

3. Buchsbaum–Rim obstructions to existence of Betti diagrams

In Proposition 3.1 we illustrate obstructions which prevent a virtual Betti diagram from being the Betti diagram of an actual module. To yield information not contained in the main results of [Eisenbud and Schreyer 2009; Boij and Söderberg 2008a], these obstructions must be sensitive to scalar multiplication of diagrams. For simplicity, we restrict to the case that M is generated in degree 0, though all of these obstructions can be extended to the general case.

We say that a diagram D is a *Betti diagram* if D equals the Betti diagram of some module M, and we say that D is a *virtual* Betti diagram if D belongs to the semigroup of virtual Betti diagrams $B_{\mathbb{N}}$. Many properties of modules (for example, codimension, Hilbert function) can be computed directly from the Betti diagram. We extend such properties to virtual diagrams in the obvious way. Proposition 3.1 only involves quantities which can be determined entirely from the Betti diagram; thus we may easily test whether an arbitrary virtual Betti diagram is "obstructed" in the sense of this proposition.

Proposition 3.1 (Buchsbaum–Rim obstructions). *Let M a graded module of codimension $e \geq 2$ with minimal presentation*

$$\bigoplus_{\ell=1}^{b} S(-j_\ell) \xrightarrow{\phi} S^a \longrightarrow M \longrightarrow 0.$$

Assume that $j_1 \leq j_2 \leq \cdots \leq j_b$. Then we have the following obstructions, which are independent of one another, and each of which occurs for some virtual Betti diagram.

(1) *Second syzygy obstruction*:

$$\underline{d}_2(M) \leq \sum_{\ell=1}^{a+1} j_\ell.$$

(2) *Codimension obstruction*:

$$b = \sum_j \beta_{1,j}(M) \geq e + a - 1.$$

If we have equality, $\beta(M)$ must equal the Betti diagram of the Buchsbaum–Rim complex of ϕ.

(3) *Regularity obstruction in Cohen–Macaulay case*: *If M is Cohen–Macaulay*,

$$\operatorname{reg}(M)+e=\bar{d}_e(M)\leq\sum_{\ell=b-e-a+2}^{b}j_\ell.$$

Note that both the weak and strong versions of the Buchsbaum–Eisenbud–Horrocks rank conjecture about minimal Betti numbers (see [Buchsbaum and Eisenbud 1977] or [Charalambous et al. 1990] for a description) would lead to similar obstructions. Since each Buchsbaum–Eisenbud–Horrocks conjecture imposes a condition on each column of the Betti diagram, the corresponding obstruction would greatly strengthen part (2) of Proposition 3.1.

Remark 3.2. For D a diagram, let $D^\vee$ be the diagram obtained by rotating D by 180 degrees. When D is the Betti diagram of a Cohen–Macaulay module M of codimension e, then $D^\vee$ is the Betti diagram of some twist of $M^\vee:=\operatorname{Ext}^e_S(M,S)$, which is also a Cohen–Macaulay module of codimension e. Thus, in the Cohen–Macaulay case, we may apply these obstructions to D or to $D^\vee$.

Given any map $\widetilde{\phi}$ between free modules F and G, we can construct the Buchsbaum–Rim complex on this map, which we denote as $\operatorname{Buchs}_\bullet(\widetilde{\phi})$. The Betti table of the complex $\operatorname{Buchs}_\bullet(\widetilde{\phi})$ will depend only on the Betti numbers of F and G, and it can be thought of as an approximation of the Betti diagram of the cokernel of $\widetilde{\phi}$.

As in the statement of Proposition 3.1, let M be a graded S-module of codimension ≥ 2 with minimal presentation

$$F_1:=\bigoplus_{\ell=1}^{b}S(-j_\ell)\xrightarrow{\phi}S^a\longrightarrow M\longrightarrow 0.$$

We will consider free submodules $\widetilde{F_1}\subseteq F_1$, the induced map $\widetilde{\phi}:\widetilde{F_1}\to S^a$, and the Buchsbaum–Rim complex on $\widetilde{\phi}$. By varying $\widetilde{\phi}$ we will produce the obstructions listed in Proposition 3.1.

To prove the first obstruction, we introduce some additional notation. Let the first syzygies of M be $\sigma_1,\dots,\sigma_b$ with degrees $\deg(\sigma_\ell)=j_\ell$. The first stage of the Buchsbaum–Rim complex on ϕ is the complex

$$\bigwedge^{a+1}F_1\xrightarrow{\epsilon}F_1\to S^a.$$

A basis of $\bigwedge^{a+1}F_1$ is given by $e_{I'}$ where I' is a subset $I'\subseteq\{1,\dots,b\}$ with $|I'|=a+1$. Let $\det(\phi_{I'\setminus\{i\}})$ be the maximal minor corresponding to the columns $I'\setminus\{i\}$. Then the map ϵ sends $e_{I'}\mapsto\sum_{i\in I'}e_i\det(\phi_{I'\setminus\{i\}})$. We refer to $\epsilon(e_{I'})$ as a *Buchsbaum–Rim second syzygy*, and we denote it by $\rho_{I'}$. There are $\binom{b}{a+1}$

Buchsbaum–Rim second syzygies. It may happen that one of these syzygies specializes to 0 in the case of ϕ. But as we now prove, if $\rho_{I'}$ specializes to 0 then we can find another related syzygy in lower degree.

Lemma 3.3. *Let $I' = \{i_1, \dots, i_{a+1}\} \subseteq \{1, \dots, b\}$, and assume that $\rho_{I'}$ is a trivial second syzygy. Then M has a second syzygy of degree strictly less than $\sum_{i \in I'} j_i$ and supported on a subset of the columns corresponding to I'.*

Proof. Let A be an $a \times b$-matrix representing ϕ. Let $C = \{1, \dots, b\}$ index the columns of A, and let $W = \{1, \dots, a\}$ index the rows of A. If $I \subseteq C$ and $J \subseteq W$ then we write $A_{I,J}$ for the corresponding submatrix.

The Buchsbaum–Rim syzygy $\rho_{I'}$ is trivial if and only if all the $a \times a$ minors of $A_{I',W}$ are zero. Let $a' = \operatorname{rank} A_{I',W}$ which by assumption is strictly less than a. We may assume that the upper left $a' \times a'$ minor of $A_{I',W}$ is nonzero. We set $I'' = \{i_1, \dots, i_{a'+1}\}$ and $J'' = \{1, \dots, a'\}$. Let τ be the Buchsbaum–Rim syzygy of $A_{I'',J''}$. Then $\tau \neq 0$ because $\det(A_{I'' \setminus \{a'+1\}, J''}) \neq 0$. Also $(A_{I'',J''}) \cdot \tau = 0$. Thus,

$$\big(A_{I'',W}\big) \cdot \tau = \begin{pmatrix} A_{I'',J} \\ A_{I'',W-J''} \end{pmatrix} \cdot \tau = \begin{pmatrix} 0 \\ * \end{pmatrix}.$$

There exists an invertible matrix $B \in GL_a(k(x_1, \dots, x_n))$ such that

$$B \cdot A_{I'',W} = \begin{pmatrix} A_{I'',J''} \\ 0 \end{pmatrix}.$$

This gives

$$0 = (B \cdot A_{I'',W}) \cdot \tau = B \cdot (A_{I'',W} \cdot \tau).$$

Since B is invertible over $k(x_1, \dots, x_n)$, we conclude that $A_{I'',W} \cdot \tau = 0$. Thus τ is a syzygy on the columns of A indexed by I'', and therefore τ represents a second syzygy of M. The degree of τ is $\sum_{i \in I''} j_i$ which is strictly less than $\sum_{i \in I'} j_i$. □

We may now prove the second syzygy obstruction and the codimension obstruction.

Proof of the second syzygy obstruction in Proposition 3.1. Apply Lemma 3.3, choosing $I' = \{1, \dots, a+1\}$. □

Proof of codimension obstruction in Proposition 3.1. Recall that the module M has minimal presentation

$$\bigoplus_{\ell=1}^{b} S(-j_\ell) \xrightarrow{\phi} S^a \longrightarrow M \longrightarrow 0.$$

Let $\operatorname{Buchs}_\bullet(\phi)$ be the Buchsbaum–Rim complex of ϕ. Then we have

$$\operatorname{codim} M \leq \operatorname{pdim} M \leq \operatorname{pdim} \operatorname{Buchs}_\bullet(\phi) = b - a + 1 = \sum_j \beta_{1,j}(M) - a + 1.$$

Since M has codimension e, we obtain the desired inequality. In the case of equality, the maximal minors of ϕ contain a regular sequence of length e, so we may conclude that

$$\beta(M) = \beta(\mathrm{Buchs}_\bullet(\phi)). \qquad \square$$

Proof of regularity obstruction in Proposition 3.1. Since M is Cohen–Macaulay of codimension e, we may assume by Artinian reduction that M is finite length. Recall that $b = \sum_j \beta_{1,j}(M)$ and let ϕ as in the proof of the codimension obstruction. If $b = e + a - 1$, then

$$\mathrm{reg}(M) = \mathrm{reg}(\mathrm{Buchs}_\bullet(\phi)) = \sum_{\ell=1}^{b} j_\ell .$$

We are left with the case that $b > e + a - 1$. Recall that $\sigma_1, \dots, \sigma_b$ is a basis of the syzygies of M. We may change bases on the first syzygies by sending $\sigma_i \mapsto \sum p_{i\ell}\sigma_\ell$ where $\deg(p_{i\ell}) = \deg\sigma_i - \deg\sigma_\ell = j_i - j_\ell$, and where the matrix $(p_{i\ell})$ is invertible over the polynomial ring. We choose a generic $(p_{i\ell})$ which satisfies these conditions. Let $\widetilde{\phi}$ be the map defined by $\sigma_b, \sigma_{b-1}, \dots, \sigma_{b-e-a+2}$. Define $M' := \mathrm{coker}\,\widetilde{\phi}$. By construction, M' has finite length, $\beta(M') = \beta(\mathrm{Buchs}_\bullet(\widetilde{\phi}))$, and M' surjects onto M. Thus we have

$$\sum_{\ell=b-e-a+2}^{f} j_\ell = \mathrm{reg}(M') \geq \mathrm{reg}(M) = \bar{d}_n(M),$$

where the inequality follows from Corollary 20.19 of [Eisenbud 1995]. $\square$

Proof of independence of obstructions in Proposition 3.1. To show that the obstructions of Proposition 3.1 are independent, we construct an explicit example of a virtual Betti diagram with precisely one of the obstructions.

For Proposition 3.1(1), consider

$$2\cdot\pi_{(0,1,5,6,7,8)} + \pi_{(0,5,6,7,8,9)} = \begin{pmatrix} 3 & 4 & - & - & - & - \\ - & - & - & - & - & - \\ - & - & - & - & - & - \\ - & 70 & 252 & 336 & 200 & 45 \end{pmatrix}.$$

Then $\underline{d}_2 = 5 > 4$ so this diagram has a Buchsbaum–Rim second syzygy obstruction.

For Proposition 3.1(2), consider

$$\pi_{(0,1,3,4)} = \begin{pmatrix} 1 & 2 & - & - \\ - & - & 2 & 1 \end{pmatrix}.$$

In this case $\sum \beta_{1,j}(\pi_{(0,1,3,4)}) = 2 < 3+1-1 = 3$. More generally, the pure diagram $\pi_{(0,1,\alpha,\alpha+1)}$ has a codimension obstruction for any $\alpha \geq 3$.

For the case of equality in Proposition 3.1(2), consider

$$\pi_{(0,1,6,10)} = \begin{pmatrix} 6 & 8 & - & - \\ - & - & - & - \\ - & - & - & - \\ - & - & 3 & - \\ - & - & - & - \\ - & - & - & - \\ - & - & - & - \\ - & - & - & 1 \end{pmatrix}.$$

Since we have $\sum \beta_{1,j}(\pi_{(0,1,6,10)}) = 8 = 3+6-1$, the diagram $\pi_{(0,1,6,10)}$ should equal the Betti table of the Buchsbaum–Rim complex on a map: $\phi : R(-1)^8 \to R^6$. This is not the case.

For Proposition 3.1(3), consider

$$2 \cdot \pi_{(0,1,4,9,10)} = \begin{pmatrix} 6 & 10 & - & - & - \\ - & - & - & - & - \\ - & - & 6 & - & - \\ - & - & - & - & - \\ - & - & - & - & - \\ - & - & - & - & - \\ - & - & - & 6 & 4 \end{pmatrix}.$$

Here we have $\bar{d}_4 = 10 > 9 = \sum_{j=1}^{9} 1$.

□

4. A linear strand obstruction in projective dimension 3

In this section, we build obstructions based on one of Buchsbaum and Eisenbud's structure theorems about free resolutions in the special case of codimension 3 [Buchsbaum and Eisenbud 1974]. The motivation of this section is to explain why the following virtual Betti diagrams do not belong to B_{mod}:

$$D = \begin{pmatrix} 2 & 4 & 3 & - \\ - & 3 & 4 & 2 \end{pmatrix}, \quad D' = \begin{pmatrix} 3 & 6 & 4 & - \\ - & 4 & 6 & 3 \end{pmatrix}, \quad D'' = \begin{pmatrix} 2 & 3 & 2 & - \\ - & 5 & 7 & 3 \end{pmatrix}. \tag{4}$$

Note that these diagrams do not have any of the Buchsbaum–Rim obstructions. In fact, there are virtual Betti diagrams similar to each of these which are Betti diagrams of modules. For instance, all of the following variants of D *are* Betti diagrams of modules:

$$\begin{pmatrix} 2 & 4 & 1 & - \\ - & 1 & 4 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 4 & 2 & - \\ - & 2 & 4 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 4 & 3 & 1 \\ - & 3 & 5 & 2 \end{pmatrix}, \quad \begin{pmatrix} 4 & 8 & 6 & - \\ - & 6 & 8 & 4 \end{pmatrix}.$$

The problem with D must therefore relate to the fact that it has too many linear second syzygies to *not* contain a Koszul summand. Yet whatever obstruction exists for D must disappear upon scaling from D to $2 \cdot D$. Incidentally, the theory of matrix pencils could be used to show that D and D'' are not Betti diagrams. We do not approach this problem via matrix pencils because we seek an obstruction which does not depend on the fact that $\beta_{0,0} = 2$.

Let $S = k[x, y, z]$ and let M be a graded S-module M of finite length. Further, let M be generated in degree 0 and with regularity 1, so that

$$\beta(M) = \begin{pmatrix} a & b & c & d \\ - & b' & c' & d' \end{pmatrix}.$$

Let T_i be the maps along the top row of the resolution of M so that we have a complex

$$0 \longrightarrow S(-3)^d \xrightarrow{(T_3)} S(-2)^c \xrightarrow{(T_2)} S(-1)^b \xrightarrow{(T_1)} S^a \longrightarrow 0.$$

Similarly, let U_j stand for matrices which give the maps along the bottom row of the resolution of M. Observe that each T_i and U_j consists entirely of linear forms, and that $U_1 = 0$. If $d \neq 0$, then the minimal resolution of M contains a copy of the Koszul complex as a free summand. Since we may split off this summand, we assume that $d = 0$.

Proposition 4.1 (Maximal minor, codimension 3 obstruction). *Let M be as defined above, and continue with the same notation. Then*

$$b' - a + \operatorname{rank} T_1 + \operatorname{rank} U_3 \leq c'.$$

Equivalently $c - d' + \operatorname{rank} T_1 + \operatorname{rank} U_3 \leq b$.

Proof. By assumption, M has a minimal free resolution given by

$$0 \longrightarrow S(-4)^{d'} \xrightarrow{\begin{pmatrix} Q_3 \\ U_3 \end{pmatrix}} S(-2)^c \oplus S(-3)^{c'} \xrightarrow{\begin{pmatrix} T_2 & Q_2 \\ 0 & U_2 \end{pmatrix}} S(-1)^b \oplus S(-2)^{b'} \xrightarrow{\begin{pmatrix} T_1 & Q_1 \end{pmatrix}} S^a \longrightarrow M.$$

Each Q_i stands for a matrix of degree 2 polynomials. By [Buchsbaum and Eisenbud 1974] we know that each maximal minor of the middle matrix is the product of a corresponding maximal minor from the first matrix and a corresponding maximal minor from the third matrix.

Let $\tau = \operatorname{rank} T_1$ and $\mu = \operatorname{rank} U_3$. Since $\operatorname{codim} M \neq 0$, the rank of the matrix $\begin{pmatrix} T_1 & Q_1 \end{pmatrix}$ equals a. By thinking of this matrix over the quotient field $k(x, y, z)$, we may choose a basis of the column space which contains τ columns from T_1 and $a - \tau$ columns from Q_1. Let Δ_1 be the determinant of the resulting $a \times a$ submatrix, and observe that Δ_1 is nonzero. Similarly, we may construct a $d' \times d'$

minor Δ_3 from the last matrix such that Δ_3 is nonzero and involves μ rows from U_3 and $d' - \mu$ rows from Q_3.

Now consider the middle matrix

$$\begin{array}{cc} & \begin{array}{cc} c & c' \end{array} \\ \begin{array}{c} b \\ b' \end{array} & \begin{pmatrix} T_2 & Q_2 \\ 0 & U_2 \end{pmatrix}. \end{array}$$

Note that the columns of this matrix are indexed by the rows of the third matrix, and the rows of this matrix are indexed by the columns of the first matrix. Choose the unique maximal submatrix such that the columns repeat none of the choices from Δ_3 and such that the rows repeat none of the choices from Δ_1. We obtain a matrix of the shape

$$\begin{array}{cc} & \begin{array}{cc} c-d'+\mu & c'-\mu \end{array} \\ \begin{array}{c} b-\tau \\ b'-a+\tau \end{array} & \begin{pmatrix} * & * \\ 0 & * \end{pmatrix}. \end{array}$$

Since M has finite length, the Herzog–Kühl conditions [1984] imply that $c' + c - d' = b + b' - a$, and thus this is a square matrix. If Δ_2 is the determinant of the matrix constructed above, then $\Delta_2 = \Delta_1 \Delta_3$ by [Buchsbaum and Eisenbud 1974]. Since $\Delta_1 \neq 0$ and $\Delta_3 \neq 0$, this implies that the $(b' - a + \tau \times c - d' + \mu)$ block of zeroes in the lower left corner cannot be too large. In particular,

$$b' - a + \tau + c - d' + \mu \leq b' + b - a.$$

By applying the Herzog–Kühl equality $c' + c - d' = b + b' - a$, we obtain the desired results. □

We now prove a couple of lemmas which will allow us to use this obstruction to rule out the virtual Betti diagrams from (4). We continue with the same notation, but without the assumption that $d = 0$.

Definition 4.2. A matrix T is *decomposable* if there exists a change of coordinates on the source and target of T such that T becomes block diagonal or such that T contains a column or row of all zeroes. If T is not decomposable then we say that T is *indecomposable*.

Lemma 4.3. *If the Betti diagram*

$$\begin{pmatrix} a & b & c & d \\ - & b' & c' & d' \end{pmatrix}$$

is Cohen–Macaulay and is a minimal generator of B_{mod}, then T_1 is indecomposable or $b = 0$.

Proof. If we project the semigroup B_{mod} onto its linear strand via

$$\begin{pmatrix} a & b & c & d \\ - & b' & c' & d' \end{pmatrix} \mapsto \begin{pmatrix} a & b & c & d \end{pmatrix},$$

then the image equals the semigroup of linear strands in B_{mod}. By the Herzog–Kühl equations, the linear strand

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

of such a Cohen–Macaulay module determines the entire Betti diagram. Hence the projection induces an isomorphism between the subsemigroup of Cohen–Macaulay modules of codimension 3 in B_{mod} and the semigroup of linear strands in B_{mod}. The modules with T_1 decomposable and $b \neq 0$ cannot be minimal generators of the semigroup of linear strands in B_{mod}. □

Lemma 4.4. *Let the notation be as above.*

(a) *If there exists a free submodule $F \subseteq S(-1)^b$ such that $F \cong S(-1)^3$ and such that the restricted map $T_1|_F$ has rank* 1, *then the minimal resolution of M contains a copy of the Koszul complex as a direct summand.*

(b) *If $a = 2$, $b \geq 3$, and T_1 is indecomposable then T_1 has rank* 2.

Proof. (a) Given the setup of the lemma, we have that $T_1|_F$ is an $a \times 3$ matrix of rank 1 with linearly independent columns over k. All matrices of linear forms of rank 1 are compression spaces by [Eisenbud and Harris 1988]. Since the columns of $T_1|_F$ are linearly independent, this means that we may choose bases such that

$$T_1|_F = \begin{pmatrix} x & y & z \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix}. \tag{5}$$

The result follows immediately.

(b) Assume that T_1 has rank 1 and apply part (a) with F any free submodule isomorphic to $S(-1)^3$. We may then assume that the first three columns of T_1 look like (5), and whether $b = 3$ or $b > 3$, it quickly follows that T_1 is decomposable. □

Proposition 4.5. *The virtual Betti diagrams*

$$D = \begin{pmatrix} 2 & 4 & 3 & - \\ - & 3 & 4 & 2 \end{pmatrix}, \quad D' = \begin{pmatrix} 3 & 6 & 4 & - \\ - & 4 & 6 & 3 \end{pmatrix}, \quad D'' = \begin{pmatrix} 2 & 3 & 2 & - \\ - & 5 & 7 & 3 \end{pmatrix}$$

do not belong to B_{mod}.

Proof. Assuming D were a Betti diagram, Lemma 4.3 implies that the corresponding matrices T_1 and U_3 are indecomposable. Lemma 4.4(b) implies that for D as in (5), we have $\operatorname{rank} T_1 = \operatorname{rank} U_3 = 2$. Observe that D now has a maximal minor obstruction, as $c - d' + \tau + \mu = 5$ while $b = 4$.

Next we consider D'. If D' were a Betti diagram, the corresponding T_1 and U_3 would both have to be indecomposable. If also T_1 had rank 2, then Theorem 1.1 of [Eisenbud and Harris 1988] would imply that it is a compression space. In particular, T_1 would have one of the following forms:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & * & * \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 & 0 & * \\ * & * & * & * & * & * \end{pmatrix}, \quad \text{or} \quad \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{pmatrix}.$$

The matrix forms on the left and right fail to be indecomposable. The middle form could not have linearly independent columns, since each $*$ stands for a linear form, and we are working over $k[x, y, z]$. Thus T_1 and U_3 both have rank 3, and it follows that D' has a maximal minor obstruction.

In the case of D'', similar arguments show that the ranks of T_1 and U_3 must equal 2 and 3 respectively. Thus D'' also has a maximal minor obstruction. □

Example 4.6. The diagram $2 \cdot D$ belongs to B_{mod}. In fact, if $N = k[x, y, z]/(x, y, z)^2$ and $N^\vee = Ext^3(N, S)$, then

$$\beta(N \oplus N^\vee(4)) = \begin{pmatrix} 1 & - & - & - \\ - & 6 & 8 & 3 \end{pmatrix} + \begin{pmatrix} 3 & 8 & 6 & - \\ - & - & - & 1 \end{pmatrix} = \begin{pmatrix} 4 & 8 & 6 & - \\ - & 6 & 8 & 4 \end{pmatrix} = 2 \cdot D.$$

This diagram does not have a maximal minor obstruction as $\operatorname{rank} T_1 = \operatorname{rank} U_3 = 3$.

Conversely, up to isomorphism the direct sum $N \oplus N^\vee(4)$ is the only module M whose Betti diagram equals $2 \cdot D$. The key observation is that for M to avoid having a maximal minor obstruction, we must have that $\operatorname{rank} T_1 + \operatorname{rank} U_3 \leq 6$. Thus we may assume that M is determined by a 4×8 matrix of linear forms which has rank at most 3. Such matrices are completely classified in [Eisenbud and Harris 1988], and an argument such as that in Proposition 4.5 can rule out all possibilities except that $M \cong N \oplus N^\vee(4)$.

In the proof of Theorem 1.6(4), we will show that $3 \cdot D$ does not belong to B_{mod}.

5. Special cases when $B_{\mathbb{N}} = B_{\mathrm{mod}}$

In this section we prove Proposition 1.4 in two parts. We first deal with projective dimension 1.

Proposition 5.1. *Let $S = k[x]$ and fix $\underline{d} \leq \bar{d}$. Then $B_{\mathbb{N}} = B_{\mathrm{mod}}$. The semigroup B_{mod} is minimally generated by pure diagrams.*

Proof. Let $D \in B_{\mathbb{N}}$ be a virtual Betti diagram of projective dimension 1. We may assume that D is a Cohen–Macaulay diagram of codimension 1. Then the Herzog–Kühl conditions [1984] imply that D has the same number of generators and first syzygies. List the degrees of the generators of D in increasing order $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_s$, and list the degrees of the syzygies of D in increasing order $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_s$. Then $D \in B_{\mathbb{N}}$ if and only if we have

$$\alpha_i + 1 \leq \gamma_i$$

for $i = 1, \dots, s$. Choose M to be a direct sum of the modules

$$M_i := \operatorname{coker}(\phi_i : R(-\gamma_i) \to R(-\alpha_i)),$$

where ϕ_i is represented by any element of degree $\gamma_i - \alpha_i$ in R. Note that $\beta(M_i)$ equals the pure diagram $\pi_{(\alpha_i,\gamma_i)}$. Thus $D \in B_{\mathrm{mod}}$ and $D = \beta(M) = \sum_i \pi_{(\alpha_i,\gamma_i)}$. □

Definition 5.2 [Boij 2000]. A graded module M is a *level module* if its generators are concentrated in a single degree and its socle is concentrated in a single degree.

We now show that in the case of projective dimension 2 level modules, the semigroups $B_{\mathbb{N}}$ and B_{mod} are equal.

Proposition 5.3. *Let $S = k[x, y]$ and fix $\underline{d} \leq \bar{d}$ such that $\underline{d}_0 = \bar{d}_0$ and $\underline{d}_2 = \bar{d}_2$. Then $B_{\mathbb{N}} = B_{\mathrm{mod}}$.*

Proof. We may assume that $\underline{d}_0 = 0$, and then we are considering the semigroup of level modules of projective dimension 2 with socle degree $(\underline{d}_2 - 2)$. Let $D \in B_{\mathbb{N}}$ and let c be a positive integer such that $cD \in B_{\mathrm{mod}}$. Let $\vec{h}(D) = (h_0, h_1, \dots)$ be the Hilbert function of D. The main result of [Söderberg 2006] shows that $\vec{h}(D)$ is the Hilbert function of some level module of embedding dimension 2 if and only if $h_{i-1} - 2h_i + h_i \leq 0$ for all $i \leq \underline{d}_2 - 2$.

Since $cD \in B_{\mathrm{mod}}$, we know that $\vec{h}(cD) = c\vec{h}(D)$ is the Hilbert function of a level module. Thus

$$ch_{i-1} - 2ch_i + ch_i \leq 0.$$

The same holds when we divide by c, and thus $\vec{h}(D)$ is the Hilbert function of some level module M. Since M is also a level module, its Betti diagram must equal D. □

Remark 5.4. We conjectured above that $B_{\mathbb{N}} = B_{\mathrm{mod}}$ in general in projective dimension 2. Some evidence for this conjecture is provided by computations by Erman [≥ 2009] which prove that all virtual Betti diagrams of projective dimension 2 and generated in a single degree are "unobstructed" in the sense of Proposition 3.1.

6. The structure of $B_{\mathbb{N}} \setminus B_{\mathrm{mod}}$

We are now prepared to prove Theorem 1.6 and thus show that for projective dimension greater than 2, the semigroups $B_{\mathbb{N}}$ and B_{mod} diverge.

The various pieces of the theorem follow from a collection of obstructed virtual Betti diagrams.

Proof of Theorem 1.6(1): B_{mod} is not necessarily a saturated semigroup. We will show that on the ray corresponding to

$$D_1 = \begin{pmatrix} 1 & 2 & - & - \\ - & - & 2 & 1 \end{pmatrix},$$

every lattice point except D_1 itself belongs to B_{mod}. We have seen in (1) that $D_1 \notin B_{\mathrm{mod}}$. Certainly $2 \cdot D_1 \in B_{\mathrm{mod}}$ as $2 \cdot D$ is the Buchsbaum–Rim complex on a generic 2×4 matrix of linear forms. We claim that $3 \cdot D_1$ also belongs to B_{mod}. In fact, if we set $S = k[x, y, z]$ and

$$M := \operatorname{coker} \begin{pmatrix} x & y & z & 0 & 0 & 0 \\ 0 & 0 & x & y & z & 0 \\ x+y & 0 & 0 & x & y & z \end{pmatrix},$$

then the Betti diagram of M is $3 \cdot D_1$. □

Proof of Theorem 1.6(2): $|B_{\mathbb{N}} \setminus B_{\mathrm{mod}}|$ may be infinite. We will show that for all $\alpha \in \mathbb{N}$, the virtual Betti diagram

$$E_\alpha := \begin{pmatrix} 2+\alpha & 3 & 2 & - \\ - & 5+6\alpha & 7+8\alpha & 3+3\alpha \end{pmatrix}$$

does not belong to B_{mod}.

Note that $E_0 \notin B_{\mathrm{mod}}$ by Proposition 4.5. Imagine now that $\beta(M) = E_\alpha$ for some α. Let T_1 be the linear part of the presentation matrix of M so that T_1 is an $(\alpha + 2) \times 3$ matrix of linear forms. Let T_2 be the (3×2) matrix of linear second syzygies and write

$$T_1 \cdot T_2 = \begin{pmatrix} l_{1,1} & l_{1,2} & l_{1,3} \\ l_{2,1} & l_{2,2} & l_{2,3} \\ \vdots & \vdots & \vdots \end{pmatrix} \cdot \begin{pmatrix} s_{1,1} & s_{1,2} \\ s_{2,1} & s_{2,2} \\ s_{3,1} & s_{3,2} \end{pmatrix}.$$

By Lemma 4.4(a), the rank of T_1 must be at least 2. Let T_1' be the top two rows of T_1, and by shuffling the rows of T_1, we may assume that the rank of T_1' equals 2. So then may assume that $l_{1,1}$ and $l_{2,2}$ are nonzero. Since each column of T_2 has at least 2 nonzero entries, it follows that the syzygies represented by T_2 remain nontrivial syzygies on the columns of T_1'.

It is possible however that columns of T_1' are not k-linearly independent. But since the rank of T_1' equals 2, we know that at least two of the columns are linearly independent. Let C be the cokernel of T_1', and let $M' := C_{\leq 1}$ be the truncation of C in degrees greater than 1. Then we would have

$$\beta(M') = \begin{pmatrix} 2 & 3 & 2 & - \\ - & 5 & 7 & 3 \end{pmatrix} \text{ or } = \begin{pmatrix} 2 & 2 & 2 & - \\ - & * & * & * \end{pmatrix}.$$

The first case is impossible by Proposition 4.5, and the second case does not even belong to $B_{\mathbb{N}}$. □

Proof of Theorem 1.6(3): A ray of B_{mod} *can miss* $\dim S - 2$ *consecutive lattice points.* Fix some prime $P \geq 2$ and let

$$S = k[x_1, \dots, x_{P+1}].$$

Consider the degree sequence

$$d = (0, 1, P+1, P+2, \dots, 2P).$$

We will show that the first $P-1$ lattice points of the ray r_d have a codimension obstruction.

Let $\overline{\pi}_d$ be the pure diagram of type d where we fix $\beta_{0,0}(\overline{\pi}_d) = 1$. We claim that

- $\beta_{1,1}(\overline{\pi}_d) = 2$, and
- all the entries of $\beta(\overline{\pi}_d)$ are positive integers.

For both claims we use the formula $\beta_{i,d_i}(\overline{\pi}_d) = \prod_{k \neq i} \frac{d_k}{(-1)^k (d_i - d_k)}$. We first compute

$$\beta_{1,1}(\overline{\pi}_d) = \frac{(P+1) \cdot \dots \cdot (2P-1) \cdot (2P)}{(P \cdot (P+1) \dots (2P-1))} = \frac{2P}{P} = 2.$$

For the other entries of $\overline{\pi}_d$ we compute

$$\begin{aligned} \beta_{i,d_i}(\overline{\pi}_d) &= \frac{2P \cdot (2P-1) \cdot \dots \cdot (P+1)}{(i-2)!\,(P-i+1)!} \cdot \frac{1}{P+i-1} \cdot \frac{1}{P+i-2} \\ &= \frac{1}{P} \binom{P+i-3}{i-2} \binom{2P}{P-i+1}. \end{aligned}$$

Note that $\binom{2P}{P-i+1}$ is divisible by P for all $i \geq 2$ and thus $\beta_{i,d_i}(\overline{\pi}_d)$ is an integer as claimed.

Since $\beta_{0,0} = 1$ and $\beta_{1,1} = 2$, the diagram $c \cdot \overline{\pi}_d$ has a codimension obstruction for $c = 1, \dots, P-1$. Thus the first $P-1$ lattice points of the ray of π_d do not correspond to Betti diagrams. □

Proof of Theorem 1.6(4): There exist rays of $B_{\mathbb{N}}$ where the points of B_{mod} are nonconsecutive lattice points. Consider the ray corresponding to

$$D_2 = \begin{pmatrix} 2 & 4 & 3 & - \\ - & 3 & 4 & 2 \end{pmatrix}.$$

Proposition 4.5 shows that D_2 does not belong to B_{mod}. In Example 4.6 we showed that $2 \cdot D_2$ does belong to B_{mod}. Thus, it will be sufficient to show that

$$3 \cdot D_2 = \begin{pmatrix} 6 & 12 & 9 & - \\ - & 9 & 12 & 6 \end{pmatrix}$$

does not belong to B_{mod}.

We assume for a contradiction that there exists M such that $\beta(M) = 3 \cdot D_2$. Then the minimal free resolution of M is

$$0 \longrightarrow R(-4)^6 \xrightarrow{\binom{Q_3}{U_3}} R(-2)^9 \oplus R(-3)^{12} \xrightarrow{\begin{pmatrix} T_2 & Q_2 \\ 0 & U_2 \end{pmatrix}} R(-1)^{12} \oplus R(-2)^9 \xrightarrow{(T_1 \; Q_1)} R^6 \quad (6)$$

where T_1, T_2, U_2 and U_3 are matrices of linear forms. By Proposition 4.1 we have that $\operatorname{rank} T_1 + \operatorname{rank} U_3 \leq 9$. Since the diagram $3 \cdot D_2$ is Cohen–Macaulay and symmetric, we may use Remark 3.2 to assume that $\operatorname{rank} T_1 \leq 4$.

We next use the fact that, after a change of coordinates, T_2 contains a second syzygy which involves only 2 of the variables of S. This is proved in Lemma 6.1 below. Change coordinates so that the first column of T_2 represents this second syzygy and equals

$$\begin{pmatrix} y \\ -x \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Since T_1 must be indecomposable, we may put T_1 into the form

$$T_1 = \begin{pmatrix} x & y & z & 0 & \dots & 0 \\ 0 & 0 & * & * & \dots & * \\ \vdots & & & & & \vdots \\ 0 & 0 & * & * & \dots & * \end{pmatrix}. \quad (7)$$

Now set $\widetilde{T}_1$ to be the lower right corner of $*$'s in T_1. Since $\operatorname{rank} T_1 \leq 4$ we have that $\operatorname{rank} \widetilde{T}_1 \leq 3$. Matrices of rank ≤ 3 are fully classified, and by applying Corollary 1.4 of [Eisenbud and Harris 1988] we conclude that $\widetilde{T}_1$ is a compression space. We can rule out the compression spaces cases where $\widetilde{T}_1$ has a column or a row equal to zero, or else T_1 would have been decomposable. Thus $\widetilde{T}_1$ is equivalent to one of the two following forms:

$$\begin{pmatrix} 0&0&0&0&0&0&0&0&0&*\\ 0&0&0&0&0&0&0&0&0&*\\ 0&0&0&0&0&0&0&0&0&*\\ 0&0&0&0&0&0&0&0&0&*\\ *&*&*&*&*&*&*&*&*&*\\ *&*&*&*&*&*&*&*&*&* \end{pmatrix} \quad\text{or}\quad \begin{pmatrix} 0&0&0&0&0&0&0&0&*&*\\ 0&0&0&0&0&0&0&0&*&*\\ 0&0&0&0&0&0&0&0&*&*\\ 0&0&0&0&0&0&0&0&*&*\\ 0&0&0&0&0&0&0&0&*&*\\ *&*&*&*&*&*&*&*&*&* \end{pmatrix}.$$

If we substitute the matrix on the left into the form for T_1 from (7), then we see that T_1 would have 8 k-linearly independent columns which are supported on only the bottom two rows. Since all entries of T_1 are linear forms in $k[x, y, z]$, this is impossible. We can similarly rule out the possibility of the matrix on the right. □

Lemma 6.1. *If there exists a minimal resolution as in* (6), *then the matrix* T_2 *contains a second syzygy involving only* 2 *variables of* S.

Proof. Assume that this is not the case and quotient by the variable z. Then the quotient matrices $\overline{T}_1$ and $\overline{T}_2$ still multiply to 0. It is possible that after quotienting, some of the columns of T_1 are dependent. However this is not possible for T_2. For if some combination went to 0 after quotienting by z, then there would exist a column of T_2, that is, a second syzygy of M, which involves only the variable z. This is clearly impossible. Thus the columns of $\overline{T}_2$ are linearly independent.

Nevertheless, we know that the columns of a 6×12 matrix of linear forms over $k[x, y]$ can satisfy at most 6 independent linear syzygies. By changing coordinates we may arrange that 3 of the columns of $\overline{T}_2$ are *trivial* syzygies on $\overline{T}_1$. By trivial syzygy, we mean a column of $\overline{T}_2$ where the nonzero entries of that columns multiply with zero entries of $\overline{T}_1$. For an example of how a nontrivial syzygy over $k[x, y, z]$ can become trivial after quotienting by z, consider

$$\begin{pmatrix} x & z & 0\\ y & 0 & z \end{pmatrix}\begin{pmatrix} z\\ -x\\ -y \end{pmatrix} \to \begin{pmatrix} x & 0 & 0\\ y & 0 & 0 \end{pmatrix}\begin{pmatrix} 0\\ -x\\ -y \end{pmatrix}.$$

Change coordinates so that the first 3 columns of $\overline{T}_2$ represent the trivial syzygies and are in Kronecker normal form. By assumption, each column of $\overline{T}_2$ involves both x and y, so these first 3 columns must consist of combinations of the Kronecker blocks

$$B_1 = \begin{pmatrix} x\\ y \end{pmatrix},\quad B_2 = \begin{pmatrix} x & 0\\ y & x\\ 0 & y \end{pmatrix},\quad B_3 = \begin{pmatrix} x & 0 & 0\\ y & x & 0\\ 0 & y & x\\ 0 & 0 & y \end{pmatrix}.$$

Since each nonzero entry in the trivial part of $\overline{T}_2$ must multiply with a 0 from $\overline{T}_1$, this forces certain columns of $\overline{T}_1$ to equal 0. More precisely, the number of nonzero rows in the trivial part of $\overline{T}_2$ is a lower bound for the number of columns of $\overline{T}_1$ which are identically zero. The block decomposition shows that the trivial part of $\overline{T}_2$ has at least 4 nonzero rows, and thus $\overline{T}_1$ has at least 4 columns which are identically zero.

But now the nonzero part of $\overline{T}_1$ is a 6×8 matrix of linear forms, and this can satisfy at most 4 linear syzygies. This forces two *additional* columns of $\overline{T}_2$ to be trivial syzygies which in turn forces more columns of $\overline{T}_1$ to equal zero, and so on.

Working through this iterative process, we eventually conclude that $\overline{T}_1$ contains 8 columns which are identically zero. This means that T_1 must have contained 8 columns which involved only z. But since T_1 is a 6×12 matrix of linear forms with linearly independent columns, this is impossible. □

Remark 6.2. Consider the diagram

$$D = \frac{a}{2}\pi_{(0,1,2,4)} + \frac{b}{2}\pi_{(0,2,3,4)} = \begin{pmatrix} \frac{3a+b}{2} & 4a & 3a & - \\ - & 3b & 4b & \frac{a+3b}{2} \end{pmatrix}.$$

Clearly $D \in B_{\mathbb{N}}$ if and only if $a+b$ is even. By an argument analogous to that in the proof of Theorem 1.6(2), one can show that $D \notin B_{\mathrm{mod}}$ if $a = 1$ or $b = 1$.

Recent unpublished work of Eisenbud and Schreyer uses this example to greatly strengthen parts (2) and (4) of Theorem 1.6. They show that $D \notin B_{\mathrm{mod}}$ whenever a is odd (or equivalently whenever b is odd). Furthermore, they show that if M is any module such that

$$\beta(M) = a'\pi_{(0,1,2,4)} + b'\pi_{(0,2,3,4)},$$

then the module M splits into a direct sum of the pure pieces. Namely, $M \cong M' \oplus M''$ where $\beta(M') = a'\pi_{(0,1,2,4)}$ and $\beta(M'') = b'\pi_{(0,2,3,4)}$. Similar results are shown to hold in codimension greater than 3.

Based on a generalization of Eisenbud and Schreyer's methods, we have recently computed all generators for B_{mod} when $\underline{d} = (0, 1, 2, 3)$ and $\overline{d} = (1, 2, 3, 4)$. This computation will appear in [Erman $\geq$ 2009].

7. Further questions

An ambitious question is whether we can find a better description of B_{mod} or compile a complete list of obstructions. Here are several more specific questions. A further list of questions is compiled in [Erman et al. 2008].

(1) Bounds on B_{mod}: Can we bound the number of generators of the semigroup of Betti diagrams? Can we bound the size of a minimal generator of the semigroup of Betti diagrams?

(2) The behavior of single rays: Given a degree sequence d, what is the minimal c_d such that $c_d \pi_d$ is the Betti diagram of some module? In many cases where computation is feasible, it is known that the examples produced by Eisenbud et al. [2007] and Eisenbud and Schreyer [2009] do not represent the first element of B_{mod} on the ray. In some other cases, it is known that π_d itself does not belong to B_{mod} so that c_d is greater than 1. Can we find better lower and upper bounds for the integer c_d?

(3) Dependence on the characteristic: Schreyer's conjecture that the semigroup of Betti diagrams depends on the characteristic of k has recently been proved by Kunte [2008, Corollary 2.4.10]. In particular, Kunte shows that the virtual Betti diagram

$$\begin{pmatrix} 1 & - & - & - & - & - \\ - & 10 & 16 & - & - & - \\ - & - & - & 16 & 10 & - \\ - & - & - & - & - & 1 \end{pmatrix}$$

is not the Betti diagram of a finite length algebra when the characteristic of k equals 2. It was previously known that this is a Betti diagram when the characteristic of k equals 0. To what extent does B_{mod} depend on the characteristic? Can we find obstructions which only live in specific characteristics?

Acknowledgments

I thank David Eisenbud for his suggestions and encouragement throughout my work on this project. I also thank Marc Chardin, Frank Schreyer, and the participants in the 2008 Minimal Resolutions conference at Cornell University for many useful discussions. Finally, I thank the makers of Macaulay2 and Singular.

References

[Boij 2000] M. Boij, "Artin level modules", *J. Algebra* **226**:1 (2000), 361–374. MR 2001e:13024 Zbl 0998.13001

[Boij and Söderberg 2008a] M. Boij and J. Söderberg, "Betti numbers of graded modules and the Multiplicity Conjecture in the non-Cohen-Macaulay case", preprint, 2008. arXiv 0803.1645

[Boij and Söderberg 2008b] M. Boij and J. Söderberg, "Graded Betti numbers of Cohen–Macaulay modules and the Multiplicity Conjecture", *J. Lond. Math. Soc.* (2) **78** (2008), 85–106. MR 2427053

[Buchsbaum and Eisenbud 1974] D. A. Buchsbaum and D. Eisenbud, "Some structure theorems for finite free resolutions", *Advances in Math.* **12** (1974), 84–139. MR 49 #4995 Zbl 0297.13014

[Buchsbaum and Eisenbud 1977] D. A. Buchsbaum and D. Eisenbud, "Algebra structures for finite free resolutions, and some structure theorems for ideals of codimension 3", *Amer. J. Math.* **99**:3 (1977), 447–485. MR 56 #11983 Zbl 0373.13006

[Charalambous et al. 1990] H. Charalambous, E. G. Evans, and M. Miller, "Betti numbers for modules of finite length", *Proc. Amer. Math. Soc.* **109**:1 (1990), 63–70. MR 90j:13021 Zbl 0703.13014

[Eisenbud 1995] D. Eisenbud, *Commutative algebra: with a view toward algebraic geometry*, Graduate Texts in Mathematics **150**, Springer, New York, 1995. MR 97a:13001 Zbl 0819.13001

[Eisenbud and Harris 1988] D. Eisenbud and J. Harris, "Vector spaces of matrices of low rank", *Adv. in Math.* **70**:2 (1988), 135–155. MR 89j:14010 Zbl 0657.15013

[Eisenbud and Schreyer 2009] D. Eisenbud and F.-O. Schreyer, "Betti numbers of graded modules and cohomology of vector bundles", **22**:3 (2009), 859–888. arXiv 0712.1843

[Eisenbud et al. 2007] D. Eisenbud, G. Fløystad, and J. Weyman, "The existence of pure free resolutions", preprint, 2007. arXiv 0712.1843

[Erman $\geq$ 2009] D. Erman, *Extensions and applications of Boij–Söderberg theory*, Ph.D. Thesis, University of California, Berkeley. In preparation.

[Erman et al. 2008] D. Erman, F. Moore, and J. Nilson (editors), "Open questions related to the Boij–Söderberg theorems", preprint, 2008, Available at math.berkeley.edu/~derman/Papers/questions.pdf.

[Herzog and Kühl 1984] J. Herzog and M. Kühl, "On the Betti numbers of finite pure and linear resolutions", *Comm. Algebra* **12**:13-14 (1984), 1627–1646. MR 85e:13021 Zbl 0543.13008

[Kunte 2008] M. Kunte, *Gorenstein modules of finite length*, Ph.D. thesis, Universität des Saarlandes, 2008. arXiv 0807.2956

[Pisón-Casares and Vigneron-Tenorio 2004] P. Pisón-Casares and A. Vigneron-Tenorio, "$\mathbb{N}$-solutions to linear systems over $\mathbb{Z}$", *Linear Algebra Appl.* **384** (2004), 135–154. MR 2005i:13039 Zbl 1126.13020

[Söderberg 2006] J. Söderberg, "Artinian level modules of embedding dimension two", *J. Pure Appl. Algebra* **207**:2 (2006), 417–432. MR 2007i:13017 Zbl 1102.13013

[Sturmfels 1993] B. Sturmfels, *Algorithms in invariant theory*, Springer, Vienna, 1993. MR 94m:13004 Zbl 0802.13002

Communicated by Craig Huneke
Received 2008-11-09 Revised 2009-01-22 Accepted 2009-02-20

derman@math.berkeley.edu *Department of Mathematics, University of California, Berkeley, CA 94720-3840, United States*
http://math.berkeley.edu/~derman

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the ANT website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in ANT are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@mathscipub.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 3 No. 3 2009

1937-0652(2009)3:3;1-B